TELECOM
PARIS
école nationale
supérieure des
télécommunications

THESE

présentée pour obtenir le grade de

Docteur de l'Ecole Nationale Supérieure des Télécommunications

Spécialité: Signal et Image

Mahdi TRIKI

Quelques Contributions en Traitement Statistique du Signal et Applications au Débruitage Audio et à la Localisation des Mobiles

Thèse soutenu le 15 Juin 2007, devant le jury composé de :

Gaël RICHARD	Président
Pierre COMON	Rapporteur
Philippe FORSTER	Rapporteur
Meriem JAIDANE	Examinateur
Walter KELLERMANN	Examinateur
Dirk SLOCK	Directeur de thèse

Thèse réalisée au sein de l'Institut Eurecom

TELECOM
PARIS
école nationale
supérieure des
télécommunications

THESIS

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
from Ecole Nationale Supérieure des Télécommunications

Specialization: Signal and Image

Mahdi TRIKI

Some Contributions to Statistical Signal Processing and Applications to Audio Enhancement and Mobile Localization

Defended on June 15, 2007, before the committee composed by:

Gaël RICHARD	President
Pierre COMON	Reader
Philippe FORSTER	Reader
Meriem JAIDANE	Examiner
Walter KELLERMANN	Examiner
Dirk SLOCK	Thesis supervisor

à qui je pense quand ça ne va pas bien...

Acknowledgements

First, I would like to thank Prof. Dirk Slock, my thesis supervisor, for his guidance, encouragements, continuous help and availability.

I am grateful to Prof. Richard for doing me the honor of presiding the jury of my thesis, as well as Prof. Comon, Prof. Forster, Prof. Jaidane, and Prof. Kellermann for accepting to be members of the jury.

I would also like to extend my warmest thanks to my colleagues and friends at Eurecom Institute for the enjoyable time we had around a cup of coffee, with a soccer ball, or in a sushi restaurant...

My gratitude goes to my parents and my family. To them I dedicate this thesis.

Abstract

A random or stochastic process is a mathematical model for a phenomenon that evolves in time in an unpredictable manner from the viewpoint of the observer. It may be unpredictable due to the effect of the interference or noise in a communication link or storage medium, or it may be an information-bearing signal (deterministic from the viewpoint of the observer at the transmitter but random to an observer at the receiver). If prior information on the signal structure or statistics is available, the accuracy of the statistical signal processing significantly increases by an appropriate exploitation of such prior. In this thesis, we investigate three kinds of prior: spectral, spatial, and statistical information; and we consider applications to audio enhancement and mobile localization.

First, we investigate the structural representation of audio signals. The proposed model exploits both the sparsity and the time-frequency correlation of the audio signal. We have considered the application of our model to audio enhancement and underdetermined audio separation. Experimental results reveal that the proposed approach is suitable for the analysis of music and speech signals, and produces good auditive synthetic results. Simulations show also that the proposed scheme outperforms the classic matching pursuit schemes in terms of separation accuracy and robustness.

Then, we investigate blind dereverberation of audio signals. A multichannel linear prediction based equalizer is proposed, exploiting spatial, temporal, and spectral diversities. Simulations show that the proposed Delay-&-Predict equalizer outperforms the classic Delay-&-Sum beamformer.

The last part of the thesis focuses on Bayesian parameter estimation. Classical Bayesian approaches lead to useful MSE reduction, but they also introduce a bias (often annoying for several applications). We introduce the concept of Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation, in which unbiasedness is forced for one parameter at a time. In such a way, every parameter in turn is treated as deterministic while the other parameters are treated as Bayesian. The more general introduction of the CWCU concept is motivated by LMMSE channel estimation, for which the implications of the concept are illustrated in various ways. Application to mobile localization is investigated in more details.

Résumé

Un processus aléatoire ou stochastique est un modèle mathématique qui décrit un phénomène qui évolue dans le temps d'une façon imprévisible. Il peut être imprévisible dû à l'effet de l'interférence ou au bruit dans un support de transmission ou de stockage, ou dû à une manque d'information (déterministe du point de vue d'un observateur à l'émetteur mais aléatoire à un observateur au récepteur). Si des informations a priori, sur la structure ou les statistiques du signal, sont disponibles les performances du traitement statistique augmentent d'une manière significative en exploitant un tel a priori. Dans cette thèse, nous étudions trois types d'a priori : spectrale, spatiale, et statistique; et nous considérons particulièrement des applications au débuitage audio et à la localisation des mobiles.

D'abord, nous étudions la représentation structurale du signal audio. Le modèle proposé exploite l'espacement et les corrélations tempofréquentielles du signal audio. Nous avons appliqué notre modèle au débruitage audio la séparation audio sous-déterminée. Les résultats expérimentaux montrent que l'approche proposée convient à l'analyse des signaux de musiques et de la parole, et produisent de bons résultats auditive. Les simulations prouvent également que le schéma proposé surpasse les schéma "matching pursuit" classiques en termes d'exactitude et de robustesse de séparation.

Ensuite, nous étudions le dereverberation aveugle des signaux audio. Nous proposons un égaliseur basé sur la prédiction linéaire multicanale, exploitant les diversités spatiales, temporelles, et spectrales. Les simulations prouvent que l'égaliseur proposé (Delay-&-Predict) surpasse le filtre spatial classique (Delay-&-Sum).

La dernière partie de la thèse se concentre sur l'estimation bayésienne des paramètres. Les approches bayésiennes classiques produisent une réduction utile du MSE, mais en dépit d'un biais non nul (souvent gênant dans plusieurs applications). Nous introduisons le concept d'estimation conditionnellement non-biaisé par morceau, pour laquelle la contrainte du biais concerne un paramètre à la fois. De cette manière, chaque paramètre est traité comme déterministe tandis que les autres paramètres sont traités comme bayésiens. Une introduction plus générale du concept est motivée par l'estimation LMMSE des canaux, pour laquelle les implications du concept sont illustrées dans diverses manières. L'application à la localisation des mobiles est étudiée en détails.

Contents

II CWCU Estimation and Application to Mobile Localization 151

List of Figures

List of Tables

Acronyms

AIR	Acoustic Impulse Response
AoA	Angle of Arrival
AR	AutoRegressive
B-CRB	Biased Cramer-Rao Bound
BCWCU	Block Component-Wise Conditionally Unbiased
BLUE	Best Linear Unbiased Estimator
BS	Base Station
BSS	Blind Source Separation
CC	Cross-Correlation
CDMA	Code Division Multiple Access
CIR	Channel Impulse Response
CPICH	Common PIlot Channel
CRB	Cramer-Rao Bound
CW	Component-Wise
CWCU	Component-Wise Conditionally Unbiased
DL	Down-Link
DoA	Direction of Arrival
DPD	Direct Position Determination
DRR	Direct to Reverberant energy Ratio
D-&-S	Delay and Sum
D-&-P	Delay and Predict
EM	Expectation Maximization
EPSD	Exactly Periodic Subspace Decomposition
EPS	Exactly Periodic Subspace
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Techniques
ESS	Enhanced Signal Strength
FIM	Fisher Information Matrix
FIR	Finite Impulse Response

GCC	Generalized Cross-Correlation
HCRB	Hybrid Cramer-Rao Bound
HMP	Harmonic Matching Pursuit
HR	High-Resolution
HRMP	High Resolution Matching Pursuit
ICA	Independent Component Analysis
IPDL	Idel Period-Down Link Transmission
ISIC	Iterated Successive Interference Cancellation
LF	Location Fingerprinting
LMMSE	Linear Minimum Mean Squared Error
LP	Linear Prediction
LS	Least-Squares
LoS	Line of Sight
M3P	Meta Molecular Matching Pursuit
MA	MovingAverage
MAP	Maximum a posteriori
MCMC	Markov Chain Monto Carlo
MCRB	Modified Cramer-Rao Bound
ME	Maximum Entropy
MF	Matching Filtering
MFB	Matched Filter Bound
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
ML	Maximum Likelihood
MLSE	Maximum Likelihood Symbol Detection
MMI	Minimum Mutual Information
MMP	Molecular Matching Pursuit
MMSE	Minimum Mean Squared Error
MP	Matching Pursuit
MS	Mobile Station
MSE	Mean Squared Error
MUSIC	MULtiple Signal Classification
NLoS	Non Line of Sight
PDP	Power Delay Profile
PDP-F	Power Delay Profile Fingerprinting
PESQ	Perceptual Evaluation of Speech Quality
PHS	Personal Handy phone System
PSD	Power Spectral Density

PSDF	Power Spectral Density Function
PSDP	Power Space Delay Profile
PSDP-F	Power Space Delay Profile Fingerprinting
RMSE	Root Mean Squared Error
QPSE	Quasi-Periodic Signal Extraction
SIC	Successive Interference Cancellation
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
SRA	Statistical Room Acoustics
STFT	Short-Time Fourier Transform
TDE	Time-Delay Estimation
TDoA	Time Difference of Arrival
ToA	Time of Arrival
TRINICON	TRIple-N ICA for CONvolutive mixtures
UCRB	Uniform Cramer-Rao Bound
UMTS	Universal Mobile Telecommunications System
WLAN	Wireless Local Area Network
WT	Wavelet Transform

Notations

Here is a list of the main notations and symbols used in this document. We have tried to keep consistent notations throughout the document, but some symbols have different definitions depending on when they occur in the text.

a	Scalar variable
$\mathbf{a}$	Vector variable
$\mathbf{A}$	Matrix variable
$\mathbf{I}_M$	$M \times M$ dimensional identity matrix
$a(t),\ a(\tau)$	Continuous-time function of the temporal variable t or τ
$a(n)$	Discrete-time function of the temporal variable $a(n) = a(nT)$ for a given sampling period T
$\{a_n\}_n$	Finite or Infinite sequence of elements a_n.
$A(z)$	z-transform of $\mathbf{A}(n)$
$A^{\dagger}(z)$	$A^H\left(1/z^*\right)$: the matched filter associated to $A(z)$
$A(f)$	$A\left(e^{j2\pi f}\right)$: Fourier Transform of $\mathbf{A}(n)$
$j = \sqrt{-1}$	The unitary imaginary number
$\Re(x)$	Real part of x
$\Im(x)$	Imaginary part of x
$\log(x)$	Logarithm of x
e^x	Exponential of x
$(.)^*$	Complex conjugate
$(.)^T$	Transpose
$(.)^H$	Hermitian conjugate
$\det(\mathbf{A})$	Determinant of the matrix $\mathbf{A}$
$\mathrm{tr}\ (\mathbf{A})$	Trace of the matrix $\mathbf{A}$

diag $(\mathbf{A})$	Diagonal matrix of the diagonal elements of the matrix $\mathbf{A}$
bdiag $(\mathbf{A})$	Block-diagonal matrix of the block-diagonal elements of the matrix $\mathbf{A}$
$\mathcal{P}_A$	$\mathbf{A}\left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H$: Projection on the column space of $\mathbf{A}$
$E\{.\}$	Statistical expectation operator
$\langle . \rangle$	Spatial expectation operator
δ_{ij}	Kronecker delta
$\delta(t)$	Dirac delta function
$O(N)$	Of the order of N
$\mathbb{R}$	Real number set
$\mathbb{N}$	Positive natural number set

Chapter 1

Introduction

Audio enhancement aims at improving the performance of audio communication systems in noisy environments. Typically, the quality of an audio signal captured in real-world environments is invariably degraded by acoustic interference (figure 1.1).

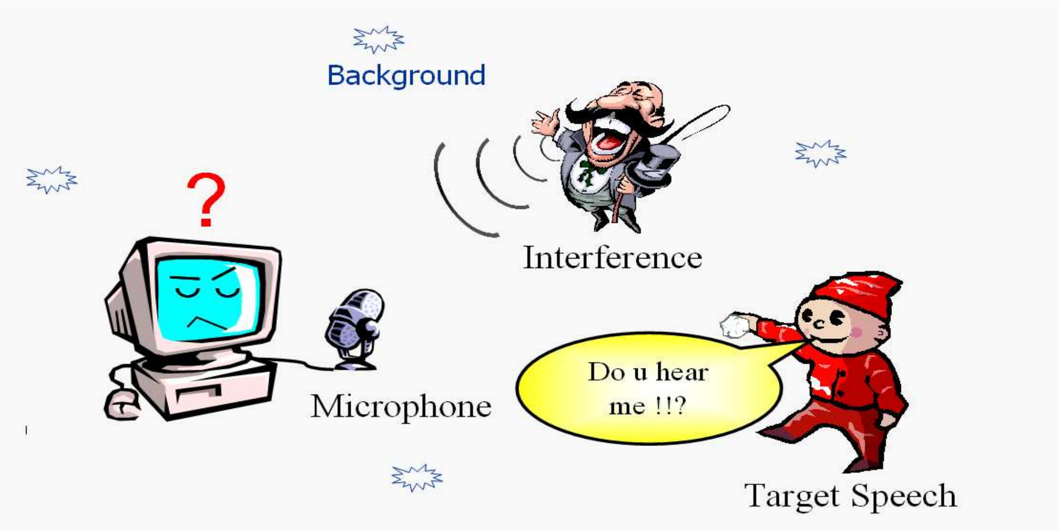

Figure 1.1: Audio signal captured in a real-world environment.

In fact, the audio signal either originates from some noisy location or is affected by distortion or noise over the channel or at the receiving end. This interference can be broadly classified into two distinct categories: additive and convolutive:

- The additive noise originates from surrounding sounds coming from other ”competing” speakers, music, background noise, etc. In fact, the *principle of superposition*, which applied to linear systems, states that the total response at a given place and time caused by two or more sources propagating in the same space is the *sum* of the separate responses which would have been produced by the individual sources.

- The convolutive interference (commonly referred to as reverberation) is due to sound wave reflections from surrounding walls and objects. Indeed, a sound wave in an enclosed or semi-enclosed environment will be broken up as it is bounced back and forth among the reflecting surfaces. Reverberation is the result of a multiplicity of echoes whose speed of repetition is too quick for them to be perceived as separate from one another.

The contaminated audio signal is typically captured by a set of microphones. Audio signal enhancement and restoration aims, by combining this set of observations and using adequate digital processing tools, to recover as well as possible the original audio signal (see figure 1.2)

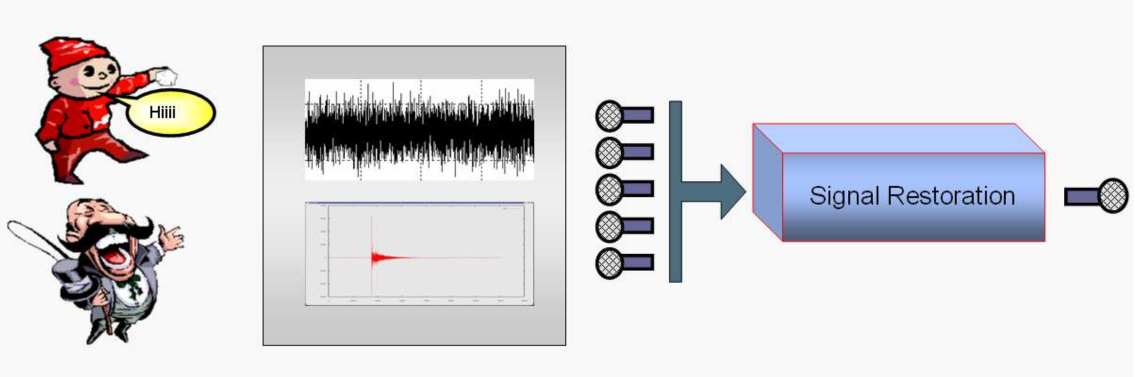

Figure 1.2: Acoustical signal enhancement and restoration.

1.1 Thesis Overview & Outline

Audio enhancement is considered a difficult problem. In fact, the nature and the characteristics of the audio and the noise sources can change in time and from application to application. On the other hand, spectral, spatial, and statistical prior information is available. In this thesis, we investigate the exploitation of such priors. In chapters 2 and 3, we propose time domain approaches taking into account the time-frequency structure of the audio signal. In chapter 4, spatial diversity prior information is exploited to design a dereverberation scheme. In part II, we focus on the use of prior statistical information. We introduce a new estimation scheme (the Component-Wise Conditionally Unbiased (CWCU) estimation), in which we impose conditional unbiasedness on one component at a time. Despite the fact that the application of the CWCU concept to audio processing seems natural, it is not considered in the context of this work. In this thesis, we consider applications to digital communication such as supervised channel and direction of arrival estimation (chapter V), and application to mobile terminal positioning (chapter VI).

This thesis is composed of two parts. The first one deals with audio signal enhancement and restoration; the second with component-wise conditionally unbiased estimation and application to mobile localization. A brief overview of the general framework of this thesis and of each part is given in this section. An abstract and an introduction are provided at the beginning of each chapter. Other research are carried out during the thesis period, including analysis of adaptive filtering convergence [192], and tracking capabilities [193]. Being out of the focus of this thesis, this work is not reported here.

1.1.1 Part I: Audio Signal Enhancement and Restoration

As we have seen previously, at a given microphone the signal of interest is possibly affected by four different kinds of perturbations:

- Noise (non-coherent noise, ambient noise): is generally inevitable and is everywhere at all time.
- Acoustic echo: occurs due to the coupling between the loudspeakers and microphones.

- Reverberation: is a result of the effect of convolutive interference on the desired signal.
- Interference (coherent noise): originates from concurrent sound sources.

Combating these perturbations leads to the development of diverse acoustic signal processing techniques; including noise reduction, echo cancellation, speech dereverberation, and source separation, each of which is a rich subject of research. For many of those problems and applications, the number of inputs and outputs of the acoustic system has been found to be crucial for the choice of the algorithms and their complexity [82]. In this thesis, we consider three extreme acoustic configurations:

- The first is a single-input single-output (SISO) system. From an algebraic point of view, this leads to an exactly determined problem (the number of unknowns equals the number of observations). Thus, the enhancement can be performed using linear processing.
- The second is a multiple-input single output (MISO) system. From an algebraic point of view, this leads to an underdetermined problem (the number of unknowns exceeds the number of observations). Classically, we refer to this problem as underdetermined source separation.
- The third is a single-input multiple-output (SIMO) system. From an algebraic point of view, this leads to an overdetermined problem (the number of unknowns is lower than the number of observations). We consider this configuration to investigate the speech dereverberation problem.

In chapter 2, we consider the classic noise reduction problem. Audio enhancement aims at improving the performance of audio communication systems in the presence of additive noise. We exploit the prior harmonic structure to identify and enhance the audio components. We model an audio signal as a periodic signal with (slow) global variation of amplitude (reflecting attack, sustain, decay) and phase. The global phase variation is modeled as a piecewise linear part (interpreted in terms of global time-warping), and an excess part (assumed to have small magnitude). The bandlimited variation of the global amplitude and phase gets expressed through a subsampled representation and parametrization of the corresponding signals.

In chapter 3, we exploit the temporal and harmonic structure of audio signals to perform underdetermined source separation. The periodic signals are assumed to have distinct periodicity (sparse time-frequency mixture), and/or to arrive at a set of sensors with different amplitude and delay (with different spatiotemporal signature). We propose a separation technique that takes into account simultaneously the source signal structure and the propagation environment parameters (time of arrival, signal attenuation).

In chapter 4, we consider the blind multichannel dereverberation problem for a single source. The multichannel reverberation impulse response is assumed to be stationary enough to allow estimation of the correlations it induces from the received signals. It is well-known that a single-input multi-output filter can be equalized blindly by applying multichannel linear prediction to its output when the input is white. When the input is colored, the multichannel linear prediction will both equalize the reverberation filter and whiten the source. We exploit the spatiotemporal channel diversity, and the speech signal non-stationarity to estimate the averaged source correlation structure, which can hence be used to determine a source whitening filter. Multichannel linear prediction is then applied to the sensor signals filtered by the source whitening filter, to obtain source dereverberation. Particular attention is paid to the blind estimation of the source color (via the optimization of the AR coefficients and order). We also investigate the robustness of the scheme to the presence of additive noise.

1.1.2 Part II: CWCU Estimation and Application to Mobile Localization

Generally, estimator designs are subject to a tradeoff between bias and variance. If prior information on the parameter statistics is available, classic Bayesian estimation theory allows the exploitation of such prior to reduce the number of the effective parameters to be estimated. Nevertheless, Bayesian estimation leads to (conditionally) biased estimation. Note that in a Bayesian context, unbiasedness normally refers to unconditional bias. Unconditional unbiasedness (unbiasedness on the average) is a weak constraint that is e.g. obtained by LMMSE estimation in the case of zero-mean variables. Conditional unbiasedness, which is normally considered in deterministic parameter estimation, is a much stronger constraint and is e.g. not attained by LMMSE

estimation.
This conditional bias is detrimental for a number of applications. Particularly, this bias is annoying in audio applications (which justifies the fact that Bayesian estimation is rarely used for such applications). This motivates the introduction and the investigation of Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation.

In chapter 5, we introduce the general concept of CWCU Bayesian estimation. Instead of constraining the parameter vector estimate to be jointly unbiased, we impose conditional unbiasedness on one parameter component at a time. In such a way, every parameter in turn is treated as deterministic while the others are being treated as Bayesian. If the parameters are transmitted symbols, the CWCU approach introduced here corresponds to the familiar unbiased symbol detection whereas joint deterministic unbiasedness leads to the zero-forcing approach.
The more general introduction of the CWCU concept was also motivated by LMMSE channel estimation, for which the implications of the concept are illustrated in various ways, including the effect on angle of arrival estimation, repercussion for trained channel estimation etc. Motivated by the channel tracking application, we also introduce CWCU Kalman filtering.

Non-Line-of-Sight and multipath propagation conditions pose significant problems for most mobile terminal positioning approaches. In contrast, Power Delay Profile (PDP) fingerprinting thrives on multipath propagation. This multipath extension of T(D)oA is based on matching an estimated PDP from one or several base stations with a memorized PDP map for a given cell. It is obvious that the overall location accuracy depends strongly of the quality of the PDP estimation. In chapter 6, we propose exploiting the prior structural information of the channel (parametric decomposition) to enhance the PDP estimation accuracy, and increase the localization accuracy. Although the Bayesian channel estimation allows better noise suppression, the bias introduced is very annoying for the parametric channel decomposition (as it leads to a modification of the pulse-shape structure). This fact motivates the use of CWCU estimation for such application. We propose to exploit the prior knowledge on the channel structure to enhance the PDP estimation and increase the localization accuracy. Using Bayesian and deterministic frameworks, we introduce one and two-step PDP-fingerprinting based localization approaches. In the case of multi-antenna reception/transmission, we introduce an extension of PDP-fingerprinting exploiting the additional spatial information.

1.2 List of contributions

The major contributions of this thesis are summarized as follows:

- Audio signal description using a periodic model with global amplitude and phase modulation.
- Audio signal extraction using global amplitude modulation and global time-warping.
- Audio signal extraction using global amplitude and phase modulation.
- Application of the global modulation based models to musical signals analysis and enhancement.
- Application of the global modulation based models to speech segmentation and enhancement.
- Application of the global modulation based models to underdetermined convolutive source separation.
- Multichannel Linear prediction (LP) based approach for blind speech dereverberation.
- Parametric blind estimation of input correlation based on spatiotemporal diversity of SIMO acoustic channels.
- Order selection for the parametric blind estimation of the input color.
- Per Channel time-delay compensation for multichannel LP-based equalization, leading to Delay-&-Predict equalization as opposed to Delay-&-Sum beamforming.
- MISO post processing of MIMO LP for blind MMSE-ZF multichannel equalization with delay.
- Component-Wise Conditionally Unbiased (CWCU) Bayesian parameters estimation.

- CWCU Kalman filtering.
- CWCU multichannel Wiener filtering.
- Analysis of the interplay between the joint bias and the prior covariance rank in Block-CWCU-LMMSE.
- Application of CWCU concept to multiple channel estimation.
- Power Delay Profile (PDP) fingerprinting with and without time reference.
- Reproducible localization validation using ray tracing multipath in a box.
- Deterministic and Bayesian parametric PDP estimation and application for PDP fingerprinting.
- Power Space Delay Profile (PSDP) for SIMO/MISO/MIMO transmission.
- PSDP fingerprinting for mobile localization.

1.3 List of publications

- Mahdi Triki and Dirk T.M. Slock, “Bridging Classical Localization and Fingerprinting Techniques for NLoS Scenarios,” manuscript in preparation.
- Mahdi Triki and Dirk T.M. Slock, “Delay and Predict Equalization For Blind Speech Dereverberation,” manuscript in preparation.
- Mahdi Triki and Dirk T.M. Slock, “Periodic Signal Extraction with Global Amplitude and Phase Modulation for Audio Source Separation,” manuscript in preparation.
- Mahdi Triki, Tayeb Sadiki, and Dirk T.M. Slock, “Window Optimization Issues in Recursive Least-Squares Adaptive Filtering and Tracking,” manuscript in preparation.

- Mahdi Triki and Dirk T.M. Slock, “Mobile Localization for NLoS Propagation”, *In Proc. of IEEE Int. Symp. on Personal Indoor and Mobile Radio Communication (PIMRC)*, Sept. 2007.

- M. Triki and D.T.M. Slock, “AR Source Modeling Based on Spatiotemporally Diverse Multichannel Outputs and Application to Multimicrophone Dereverberation,” *In Proc. of Int. Conf. on Digital Signal Processing (DSP)*, July 2007.

- M. Triki and D.T.M. Slock, “Multivariate LP Based MMSE-ZF Equalizer Design Considerations and Application to MultiMicrophone Dereverberation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2007.

- Mahdi Triki, and Dirk T.M. Slock, "Investigation of Some Bias and MSE Issues in Block-Component-Wise Conditionally Unbiased LMMSE," *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2006.

- Mahdi Triki, Dirk T.M. Slock, Vincent Rigal, and Pierrick François, "Mobile Terminal Positioning via Power Delay Profile Fingerprinting: Reproducible Validation Simulations," *In Proc. of IEEE Vehicular Technology Conf.*, Sept. 2006.

- M. Triki and D.T.M. Slock, “Iterated Delay and Predict Equalization for Blind Speech Dereverberation,” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006

- Mahdi Triki and Dirk T.M. Slock, “A Novel Voiced Speech Enhancement Approach Based on Modulated Periodic Signal Extraction,” *In Proc. of European Signal Processing Conf. (EUSIPCO)*, Sept. 2006.

- M. Triki and D.T.M. Slock, “Delay and Predict Equalization For Blind Speech Dereverberation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

- Mahdi Triki and Dirk T.M. Slock, “Music Source Separation via Sparsified Dictionaries vs. Parametric Models,” *In Proc. of Int. Sym. on Communications, Control, and Signal Processing (ISCCSP)*, March 2006.

- Mahdi Triki, and Dirk T.M. Slock, "Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation," *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2005.

- M. Triki and D.T.M. Slock, "Blind Dereverberation of Quasi-periodic Sources Based on Multichannel Linear Prediction," *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2005.

- Mahdi Triki, Salah Abdellatif, and Dirk T.M. Slock, "Interference Cancellation with Bayesian Channel Models and Application to TDOA/IPDL," *In Proc. of Int. Symp. on Signal Processing and its Applications (ISSPA)*, August 2005.

- Mahdi Triki and Dirk T.M. Slock, "Multi-channel mono-path periodic signal extraction with global amplitude and phase modulation for music and speech signal analysis," *In Proc. of IEEE Work. on Statistical Signal Processing (SSP)*, pp.77-82, July 2005.

- Mahdi Triki and Dirk T.M. Slock, "Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition," *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp.233-236, March 2005.

- Tayeb Sadiki, Mahdi Triki, and Dirk T.M. Slock, "Window Optimization Issues in Recursive Least-Squares Adaptive Filtering and Tracking," *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2004.

- Mahdi Triki and Dirk T.M. Slock, "The Instrumental Variable Multichannel FAP-RLS and FAP Algorithms," *In Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, July 2004.

Part I

Acoustical Signal Enhancement and Restoration

Chapter 2

Audio Signal Enhancement

Audio enhancement aims at improving the performance of audio communication systems in noisy environments. Harmonic structural information is one of the key ingredients to identify and enhance the audio components. We model an audio signal as a periodic signal with (slow) global variation of amplitude (reflecting attack, sustain, decay) and phase. The global phase variation is decomposed into a piece-linear part (interpreted in terms of global time-warping), and an excess part (assumed to have small magnitude). The bandlimited variation of the global amplitude and phase gets expressed through a subsampled representation and parametrization of the corresponding signals. Assuming additive white Gaussian noise, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Particular attention is paid to the estimation of the basic periodic signal, which can have a non-integer period. Simulation results reveal that the proposed approach is appropriate for musical note modeling and extraction; and for voiced frame speech identification and enhancement.

2.1 Introduction

Audio enhancement aims at improving the performance of audio communication systems in noisy environments. The need for enhancing audio signals arises in many situations in which the audio signal either originates from some noisy location or is affected by the noise over the channel or at the receiving end. In the presence of background noise, the human auditory system is capable of employing effective mechanisms to reduce the effect of noise on speech perception. Although such mechanisms are not well understood at the present state of knowledge (to allow the design of speech enhancement systems based on auditory principles), several practical methods for speech enhancement have already been developed. Several reviews can be found in the literature [106, 129, 46].
From a signal processing point of view, additive noise is easier to deal with than convolutive noise or nonlinear disturbances. Moreover, due to the bursty nature of audio signals (speech, music...), it is possible to observe the noise by itself during audio signal pauses, which can be of great value. On the other hand, audio enhancement is still a difficult problem for two reasons. First, the nature and characteristics of the noise source can change in time and from application to application. It is therefore difficult to find flexible schemes that work in different practical environments. Second, the human ear, the final judge, does not believe in a simple mathematical error criterion; and optimizing this mathematical criterion does not lead necessarily to the enhancement of the output audio quality. In fact, the enhancement can be measured by using two perceptual criteria: quality and intelligibility. The "quality" of the enhanced signal measures its clarity, distortion and the level of residual noise. The "quality" is a subjective measure that is indicative of the extent to which the listener is comfortable with the enhanced signal. The second criterion measures the "intelligibility" of the enhanced signal. This is an objective measure which provides the percentage of words that could be correctly identified by listeners. The two performance measures are uncorrelated: a signal may be of good quality and poor intelligibility and vice versa. It is very difficult to satisfy both at the same time. Most speech enhancement systems improve the quality of the signal at the expense of reducing its intelligibility.

In this chapter, we consider a Single Input Single Output (SISO) framework: a single source signal is captured (together with additive noise) using a single microphone (figure 2.1). From an algebraic point of view, this leads

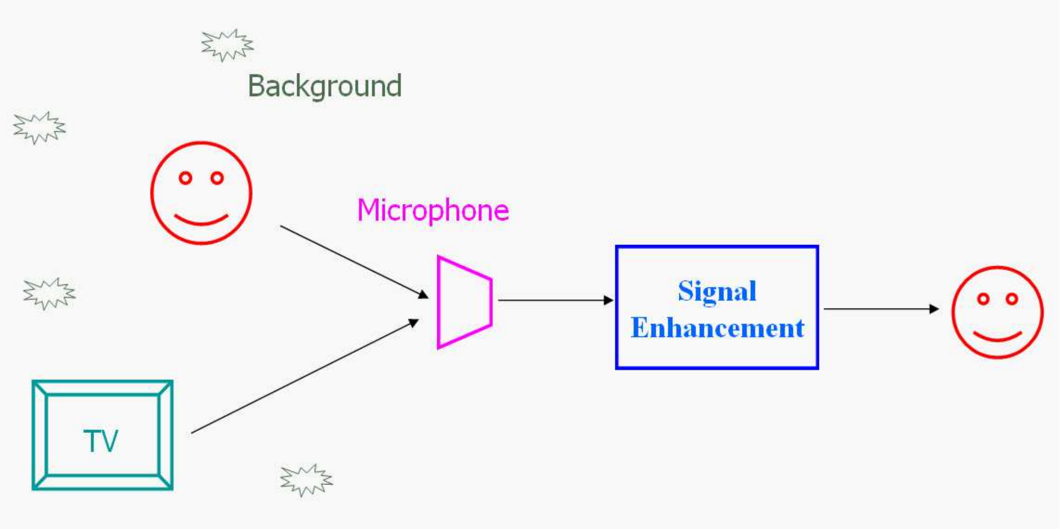

Figure 2.1: Audio enhancement: problem statement.

to an exactly determined problem (number of unknowns equals number of observations). Thus, enhancement can be performed using linear processing. For instance, a frequency-domain Wiener filter could be computed and used to enhance the received audio signal. This leads to the well-known spectral subtraction technique [128]. The spectral subtraction method is by far the most popular and most used in real word applications. However, this approach introduces some artifacts referred to as musical noise, due to the spectral estimation problem.

In this thesis, we propose exploiting the harmonic structure of the audio signal (voicing in the glottal source signal, string and wind oscillation in musical instruments...). We model an audio signal as a periodic signal with (slow) global variation of amplitude (characterizing the evolution of the signal power) and phase (emphasizing the harmonic structure of the audio signal). The global phase variation is decomposed into a piece-linear part (interpreted in terms of global time-warping), and an excess part (assumed to have small magnitude). The bandlimited variation of global amplitude and phase gets expressed through a subsampled representation and parametrization of the corresponding signals.

This chapter is organized as follows. In section 2.2, a brief overview on audio signal representation and enhancement is presented. Then, the global modulation models and the associated audio signal extraction procedures are presented in sections 2.3 and 2.4,. Finally, applications to music and speech signal enhancement are investigated in section 2.6.

2.2 Audio signal enhancement: a brief overview

In the framework of audio signal analysis and representation, there have been recent significant advances in two directions: sparse and structured representations. In fact, audio signals contain superimposed structures, such as transients and stationary parts or multiple notes and instruments, and have been shown to have sparse decompositions in a variety of time-frequency dictionaries. A second point of view tries to exploit the harmonic structure of the audio signal. In fact, the audio signal energy is almost concentrated around the fundamental frequency and the partials. Moreover, the different harmonics are correlated. Sparse and structured decompositions of audio signals are shown to be effective, and appear to be extremely useful in many signal processing applications: compression, source separation, noise reduction...

2.2.1 Sparse representation of audio signals

There has been a considerable interest in the last decade in developing flexible decompositions of nonstationary signals. In particular, atomic decompositions are shown to be effective for representing signal components whose localizations in time and frequency vary widely. In fact, as many signals display both oscillatory phenomena (for which time-frequency methods are efficient) and transients (for which time-scale techniques are better adapted), atomic decompositions were developed using redundant families of atoms that can independently characterize scale and frequency. Such decompositions are similar to a text written with a small vocabulary. Although this vocabulary might be sufficient to express all ideas, it requires using circumvolutions that replace unavailable words by full sentences [108].

Time-Frequency atomic decomposition

Decomposition of signals over a family of functions that are well localized both in time and frequency has found many applications in signal processing and harmonic analysis. Depending upon the choice of time-frequency atoms, the decomposition might have very different properties. Generally, The family $D = \{g_\gamma(t)\}_{\gamma \in \Gamma}$ is extremely redundant. To represent efficiently any function $f(t)$, we must select an appropriate countable subset of atoms $\{g_{\gamma_n}(t)\}_{n \in I\!N}$ so that $f(t)$ can be written as

$$f(t) = \sum_{m=-\infty}^{+\infty} a_m g_{\gamma_m}(t)$$

Depending upon the choice of the atoms $g_{\gamma_m}(t)$, the expansion coefficients $\{a_m\}_{m \in I\!N}$ give explicit information on certain types of properties of $f(t)$.
A general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in L^2(I\!R)$. For any scale $s > 0$, frequency modulation ξ and translation u, we denote $\gamma = (s, u, \xi)$ and we define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{j2\pi\xi t}. \tag{2.1}$$

The index γ is an element of the set $\Gamma = I\!R^+ \times I\!R^2$.
A special and useful choice for audio decomposition is a Gaussian window, i.e., $g(t) = \mathcal{N}(t) = \frac{1}{\sqrt{2\pi}}\ exp\left(-\frac{t^2}{2}\right)$. This window choice defines the Gabor family, i.e.

$$g_\gamma(t) = \frac{1}{\sqrt{s}}\, \mathcal{N}\left(\frac{t-u}{s}\right) e^{i\xi t} = \frac{1}{\sqrt{2\pi(s\sigma)}}\ .\ e^{\left(-\frac{(t-u)^2}{2(s\sigma)^2}\right)}\ .\ e^{j2\pi\xi t} \tag{2.2}$$

The function $g_\gamma(t)$ is centered at time u and its energy is concentrated in a neighborhood of u, whose size is proportional to s. Its Fourier transform is centered at the frequency $f = \xi$, and its energy is concentrated in a neighborhood of ξ, whose size is proportional to $1/s$. Short scale atoms almost correspond to "clicks", whereas large scale atoms are nearly pure sine waves. This dictionary is thus likely to comply with the representation of transients structures as well as of stationary features.
Another useful atom family is called the "Chirp dictionary". The chirp dictionary is an extension of the Gabor dictionary: every chirp atom is obtained from the Gaussian window $\mathcal{N}(t)$ by dilation, translation, frequency, and chirp modulation[67]. It can thus be described with its index $\gamma = (s, u, \xi, c)$

$$g_\gamma(t) = \frac{1}{\sqrt{s}}\, \mathcal{N}\left(\frac{t-u}{s}\right) e^{j2\pi\left(\xi(t-u)+\frac{c}{2}(t-u)^2\right)} \tag{2.3}$$

As a result, $g_{(s,u,\xi,c)}$ is localized at time u with a temporal dispersion proportional to its scale s. The Wigner-Ville distribution $WV\left[g_{(s,u,\xi,c)}\right](t,f)$ of a chirp atom defines a quadratic time-frequency energy distribution. It is localized around the line of instantaneous frequency $f(t)=\xi+c(t-u)$. Its dispersion is proportional to $1/s$ in the direction of $f(t)$.

One can also remark that atomic decomposition is a general framework that includes classic signal decomposition techniques, for instance Fourier and wavelet transforms. In fact, windowed Fourier transform and wavelet transform correspond to special families of time-frequency atoms, that are frames or bases of $L^2(I\!R)$:

* In a windowed Fourier transform, all the atoms g_{γ_m} have a constant scale $s_m = s_0$ and are thus mainly localized over an interval whose size is proportional to s_0. If the main signal structures are localized over a time-scale of the order of s_0, the expansion coefficients a_m give important insights on their localization and frequency content. However, a windowed Fourier transform is not well adapted to describe structures that are much smaller or much larger than s_0.

* On the other hand, the wavelet transform decomposes signals over time-frequency atoms of varying scales, called wavelets. A wavelet family $\left(g_{\gamma_m}(t)\right)_{m\in I\!N}$ is built by relating the frequency parameter ξ_m to the scale s_m with $\xi_m = \frac{\xi_0}{s_m}$, where ξ_0 is a constant. The resulting family is composed of dilations and translations of a single function, multiplied by complex phase parameter. The expansion coefficients a_m of a function over wavelet families characterize the scaling behavior of signal structures. This is important for the analysis of fractals and singular behaviors. However, expansion coefficients in a wavelet frame do not provide precise estimates of the frequency content of waveforms whose Fourier transform is well localized, especially at high frequencies. This is due to the restriction on the frequency parameter ξ_m, that remains inversely proportional to the scale s_m.

Matching Pursuit (MP) algorithm

Let us consider a family of vectors $D=\left(g_\gamma\right)_{\gamma\in\Gamma}$ included in the Hilbert space H (for example $L^2\left(I\!R\right)$) with a unit norm $\|g_\gamma\|=1$. For a given $f\in H$, getting the best M^{th} order approximate, i.e.

$$\widehat{f}_M = \sum_{m=1}^{M} c_m g_{\gamma_m} = \arg\min_{c_m,\gamma_m} \left\| f - \sum_{m=1}^{M} c_m g_{\gamma_m} \right\| \tag{2.4}$$

is an NP-hard problem. The matching pursuit [108] is a greedy strategy to decompose iteratively a signal into a linear combination of atoms chosen among the dictionary D. It defines an m^{th} order residual $R^{m-1}f$ (starting with $R^0 f = f$) in the following way.

1. Compute for all $\gamma \in \Gamma$
$$\left|\left\langle R^{m-1}f, g_\gamma \right\rangle\right|^2 \tag{2.5}$$

2. Choose an element $g_{\gamma_m} \in D$ which "closely" matches the residual $R^{m-1}f$ in the sense that
$$\left|\left\langle R^{m-1}f, g_{\gamma_m} \right\rangle\right|^2 = \sup_{\gamma\in\Gamma} \left|\left\langle R^{m-1}f, g_\gamma \right\rangle\right|^2 \tag{2.6}$$

3. Compute the new residual by removing the component along the selected atom
$$R^m f = R^{m-1}f - \left\langle R^{m-1}f, g_{\gamma_m} \right\rangle g_{\gamma_m} \tag{2.7}$$

The error $\left\|R^M f\right\|$ is proved to decay to zero [68]. Thus, MP provides an atomic decomposition of the signal

$$f = \sum_{m=1}^{\infty} \left\langle R^{m-1}f, g_{\gamma_m} \right\rangle g_{\gamma_m}$$

Generally, we consider sampled signals (we denote by N the signal length). Thus, the signal space H has a finite dimension N. In such case, the MP has specific properties that are studied in [109]. For instance, it was proven that the norm of the residues decays *exponentially.* On the other hand, due to the limitations of the sampling rate and the signal size, Gabor and Chirp dictionaries have a finite number of elements; and the MP algorithm can be implemented with a total complexity:

- $O\left(N\log(N)\right)$ per iteration: using the Gabor dictionary [108].
- $O\left(N^2\log(N)\right)$ per iteration: using the Chirp dictionary [67].

High Resolution Matching Pursuit (HRMP) algorithm

As we have seen in the previous section, the matching pursuit is a greedy algorithm in the sense that it optimizes at each step the amount of the signal energy it grasps. This often leads to a choice of features which globally fits the signal structures but is not best adapted to its local structures.

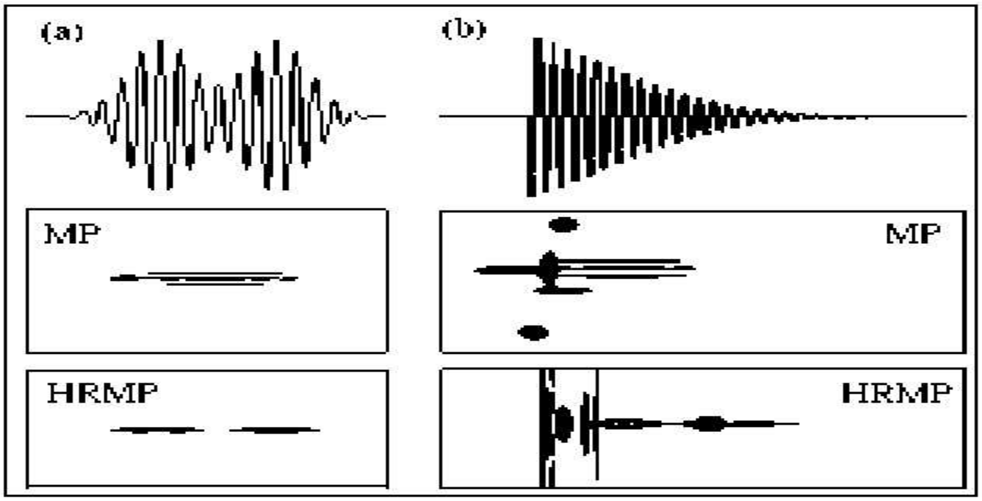

Figure 2.2: Time-Freq distributions of signals (top) obtained with MP (middle) and HRMP (bottom): (a) two close bumps, with a four atom decomposition (b) an attack pattern, with a ten atom decomposition.

For instance, a signal composed of two bumps modulated by a sinusoidal wave at frequency ξ (Figure 2.2-a) is first decomposed into a large atom at frequency ξ (middle horizontal line on Figure 2.2-a-MP) that covers the time support of both bumps. Then, in order to remove the energy created between the two bumps by this first atom, MP chooses two atoms of the same size as the first one, with frequencies $\xi + \Delta\xi$(upper line) and $\xi - \Delta\xi$ (lower line). Moreover, we observe that MP does not keep a good localization of attack patterns (Figure 2.2-b-MP), which leads to a little, but still audible, pre-echo at re-synthesis stage. This is due to the atom selection criterion that allows the creation of energy where there was none previously.

To avoid this problem, Donoho and Chen[33] introduced the Basis Pursuit, which makes a full optimization, by minimizing the sparsity measure $\sum_{m \in I\!N} |a_m|$ over all possible decompositions $f = \sum_{m \in I\!N, \gamma_m \in \Gamma} a_m g_{\gamma_m}$. The major drawback of the proposed technique is its high computation complexity, since the optimization leads to large scale linear-programming problem.

Another solution is proposed by Gribonval et al. [71, 72]. In these references, the authors introduce a modified MP scheme called High Resolution Matching Pursuit (HRMP) algorithm. The HRMP keeps the fast algorithm structure of MP, while using a different correlation function that allows the pursuit to emphasize local fit over global fit at each step. The additional complexity concerns only the computation of the correlation function. The proposed correlation function $C(f, g_\gamma)$ maximizes the amount of signal energy that the pursuit can grasp, when choosing the atom g_γ:

$$C(f, g_\gamma) = \epsilon \min_{\gamma_i \in I_\gamma} \frac{|\langle f, g_{\gamma_i} \rangle|}{|\langle g_\gamma, g_{\gamma_i} \rangle|}$$

where I_γ is a subset of indices such that atoms $g_{\gamma_i}, \gamma_i \in I_\gamma$ have a time support included in the support of g_γ, and are modulated at the same frequency of g_γ . ϵ is evaluated as follows: if $\langle f, g_{\gamma_i} \rangle$ have the same sign for all$\gamma_i \in I_\gamma$, then ϵ is this common sign; else $\epsilon = 0$.
Intuitively, in the matching pursuit, the inner-product (used as a correlation function between the time-frequency atom and the audio signal) disregards whether the signal contains energy on the whole time-frequency support of the chosen atom. On the contrary, the new correlation function avoids creating energy at time locations where there was none. It can thus distinguish close time features as shown in Figure 2.2-a-HRMP. Moreover it can avoid pre-echo effects (Figure 2.2-b-HRMP). Remark that as the atoms chosen for the decomposition have a smaller time support than with a usual MP decomposition, they also have a larger frequency support. Thus, the HRMP performs higher time resolution decomposition than the classic MP, but at the expense of decreased frequency resolution.

Harmonic Matching Pursuit (HMP) algorithm

Audio signals contain superimposed structures such as transients and stationary parts. Moreover, the energy of the signal is mostly located around fundamental frequency and partials. Given this strong harmonic content, Gribonval and Bacry propose decomposing audio signals using a dictionary of harmonic atoms [69]

$$h(t) = \sum_{k=1}^{K} c_k g_{s,u,\xi_k}(t) \qquad \|h\|^2 = \int |h(t)|^2 \, dt = 1$$

where $g_{s,u,\xi_k}(t)$ is a Gabor atom, K is the number of Gabor atoms extracted at each step, and $\{c_k\}_{k \in I\!R}$ represents the weighting coefficients of the different atoms. To emphasis the harmonic structure of the audio signal, the extracted atoms are selected to be harmonically related, i.e.,

$$\xi_k \approx k \; \xi_0 \quad 1 \leq k \leq K \tag{2.8}$$

The (almost) harmonicity $\xi_k \approx k\xi_0$ can be defined as

$$|\xi_k - k\xi_0| \leq A/s \quad A \approx 1$$

One can also remark that the Gabor atoms are special cases of harmonic atoms, with $K = 1$. Moreover, local-cosine atoms are also harmonic atoms, with $K = 2$, and $\xi_1 = -\xi_2$.

Using the harmonic dictionary, the MP algorithm becomes

1. Compute $\left\|P_{\nu_\gamma} R^{m-1}\right\|$ for all $\gamma \in \Gamma_h = \{(s, u, \xi_1, \cdots, \xi_K) \quad / \quad \xi_k \approx k\xi_0\}$.
2. Select the "best" harmonic subspace ν_{γ_m}

$$\gamma_m = \arg\max_{\gamma \in \Gamma_h} \left\|P_{\nu_\gamma} R^{m-1}\right\|$$

3. Compute the "best" harmonic atom and the new residual

$$\begin{aligned} h^m(t) &= \frac{P_{\nu_{\gamma_m}} R^{m-1}}{\left\|P_{\nu_{\gamma_m}} R^{m-1}\right\|} \\ \left\langle R^{m-1}, h^m \right\rangle &= \left\|P_{\nu_{\gamma_m}} R^{m-1}\right\| \end{aligned}$$

Note that no exhaustive search over the parameters $\{c_k\}_k$ is needed for the optimization of harmonic atoms. However, at each step, we need to compute $\left\|P_{\nu_{\gamma_m}} R^{m-1}\right\|$ for every subspace ν_γ, as well as the exact projection $P_{\nu_{\gamma_m}} R^{m-1}$

for the selected subspace. For general harmonic dictionaries, this computation is time demanding and makes the standard matching pursuit unusable. This fact further motivates faster greedy algorithms ([69, 98]). The basic observation is the quasiorthogonality of the Gabor atoms. In fact, due to the localization of the Gabor atoms in the frequency domain, we have

$$\left\langle g_{s,u,\xi_k}(t), g_{s,u,\xi_{k'}}(t)\right\rangle \;\approx\; \delta_{k,k'} \tag{2.9}$$

where $\delta_{k,k'} = \begin{cases} 1 & \text{if } k = k' \\ 0 & \text{else} \end{cases}$ is the delta function.
Thus, we have

$$\left\| P_{\nu_\gamma} R^{m-1} \right\|^2 \approx \sum_{k=1}^{K} \left| \left\langle R^{m-1}, g_{s,u,\xi_k}(t) \right\rangle \right|^2 \tag{2.10}$$

Finally, an efficient implementation of the Harmonic Matching Pursuit can be done by peaking, at each iteration, the best harmonically related K-atoms. The modified scheme is thus defined as follows:

1. Compute for all $\gamma = (s, u, \xi_0) \in \Gamma$.

$$C(R^{m-1}, g_\gamma) = \sum_{k=1}^{K} \sup_{|\xi_k - k\xi_0| \leq A/s} \left| \left\langle R^{m-1}, g_{(s,u,\xi_k)} \right\rangle \right|^2$$

2. Select the "best" harmonic subspace ν_{γ_m}

$$\gamma_m = \arg\max_{\gamma \in \Gamma} C(R^{m-1}, g_\gamma) \qquad (\gamma_m \in \Gamma_h)$$

3. Compute the new residual

$$R^m = R^{m-1} - \sum_{k=1}^{K} \left\langle R^{m-1}, g_{(s,u,\xi_k)} \right\rangle g_{(s,u,\xi_k)}$$

Therefore, we define a fast Harmonic MP algorithm. The complexity of the scheme is $O\left(KN\right)$ per iteration [69]. The decomposition method was shown to be efficient on audio analysis; specially in decomposing musical signal into harmonic structures of different duration and harmonic content.

Molecular Matching Pursuit (MMP) algorithm

Despite the MP scheme can be implemented in a fast fashion, its algorithmic complexity is still too high to be used on high-dimensional signals such as audio signals. To reduce computational complexity, the audio signal structure should be taken into account. In fact, atoms are not randomly located, but forms structures, or clusters, in the parameter plane. For instance, HMP reduces the MP complexity by exploiting the harmonic structure, which seems well adapted to signals composed of constant-frequency partials such as piano recordings. On such signals, it gives a meaningful decomposition into harmonic structures with very few harmonic molecules that seem to fit the notes. However, on sound signals with more frequency modulation such as bowed strings or trumpet, harmonic molecules are too "rigid" to represent what one could consider as elementary sound objects.

The Molecular Matching Pursuit (MMP) algorithm provides a flexible approach to exploit the signal structure. The main idea is to track the local structures that appear in the time-frequency domain[41]. As a matter of fact, it defines sound molecules as groups of *neighboring* atoms. In other words, each step of MMP consists in selecting a group of atoms $\{g_{s,u_k,\xi_k}\}_{k=1:K}$ (t) with s a fixed single scale, to form a tonal molecule corresponding to a partial (horizontal line in the time-scale plane). Thus, the algorithm is computationally very efficient (compared to MP), as a large number of coefficients (typically between 5 and 50) is selected at every iteration.

An extension of the MMP is proposed in [98], called Meta Molecular Matching Pursuit (M3P) algorithm. M3P allows selecting harmonically related tonal molecules, subsequently grouped in harmonic combs (called meta-molecules). The major difference with HMP is that, at each iteration m, the duration and shape of a harmonic molecule is not defined a priori by the model: these parameters result a posteriori from the grouping of Short Time Fourier Transform (STFT) atoms. From an algorithmic point of view, the molecules are not optimized through the direct maximization of a correlation criterion. The M3P algorithm defines a method to group short time, single scale atoms in molecules that fit the harmonic structures of the signal.
Comparing to the HMP, the M3P provides more flexible tool to represent frequency modulated harmonic objects. The "meta molecule" chains together atoms admitting a single scale. The better accuracy and increased flexibility

of M3P for the decomposition of sounds into frequency-modulated objects comes however with a price, since a greater number of parameters is required to describe the objects (plus possible partial misses).

2.2.2 Structured representation of audio signals

In the previous section, we have seen that taking into account the harmonic structure of the audio signals leads to an improvement in the enhancement performance and the computational complexity. Several authors propose exploiting further this harmonic structure. In fact, the signal energy is concentrated around the fundamental frequency and the partials. Moreover, the different harmonics are correlated. Consider, for example, a sound played by an acoustic guitar (see figure 2.10). The output signal is the convolution of the string response and the box response. On the other hand, the string response depends mainly on the fundamental frequency and the power of the attack, while the box response is time invariant. Consequently, the output signal is a function of some *time invariant* parameters and a *few* time variant parameters. Thus, we should not treat the harmonics separately as a simple filter bank or a classic MP approach would do.

Sinusoidal representation of audio signal

Sinusoidal model based music analysis/synthesis has received considerable interest in the computer music community [62, 63, 42]. The sinusoidal transform, originally developed by Quatieri and McAulay [113], represents a signal as a sum of discrete time-varying sinusoids or partials:

$$s(t) = \sum_{k=0}^{P} A_k(t) \cos\left(\theta_k(t)\right) \quad . \tag{2.11}$$

The estimation of the model parameters is typically carried out using a STFT with a fixed analysis frame size and a fixed stride between frames. The sinusoids are extracted by peak-picking in the STFT magnitude spectrum. Intermediate values are obtained by linear interpolation. A fundamental problem faced by the traditional sinusoidal-model based techniques, and which arises due to the STFT, is smearing of the frequency response [105, 144]. In fact, over the period of a single analysis frame, the algorithm estimates the amplitude, frequency and phase of any sinusoids it believes to be present. Because of the near logarithmic scale of pitch perception, we need very long windows in order to accurately estimate the pitch of low frequency partials. On the other hand, the time resolution of these parameters is only as fine as the window length, itself. And, since the music signal is strongly non-stationary , it is not always possible to find a good tradeoff between time and frequency resolution. Also, determining the sinusoid parameters from the STFT peak amplitude and phase only works well for high frequency resolution, high Signal-to-Noise Ratio (SNR) and in the absence of modulation. Another drawback of these techniques is that they ignore the harmonic structure of the music signal and the correlation between different frames.

To overcome the resolution limit of the Fourier transform (due to windowing), non-linear interpolation [145, 91, 77] and dichotomy [211, 1] based approaches were suggested to better localize the peak in the STFT domain. High-Resolution (HR) methods are also proposed to overcome the STFT resolution limit and to provide more accurate estimates of the signal parameters. The estimation of the model parameters is achieved in two steps: first, the frequencies are computed using a high-resolution method, from which the amplitudes and the initial phases are deduced by minimizing a Least-Squares (LS) criterion. The foundation of HR methods dates back from the work by Prony [147] which estimate a sum of exponentials via Linear Prediction (LP). This approach was further investigated by Pisarenko [140] for estimating sinusoids in noise. More recently, HR techniques rely on subspace signal analysis, for instance the MUltiple Signal Classification (MUSIC) algorithm [159, 19, 100], and the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [153, 154, 213, 16].

Another drawback of the classic sinusoidal techniques is that they treat separately (independently) the different time frames, ignoring the correlation between these frames. To exploit this interframe correlations, Virtanen and Klapuri [198] suggest a post-processing step tracking the model parameters. In fact, once amplitude and frequency of each peak are estimated at each frame, the detected peaks are tracked together in inter-frame trajectories, which leads to a set of sinusoidal trajectories with time-varying frequencies and amplitudes (see fig. 2.3).

A common drawback of all the previous techniques is that they ignore the harmonic structure of the audio signal. In fact, they consider the signal as a mixture of a finite number of arbitrary sinusoids, ignoring the harmonic

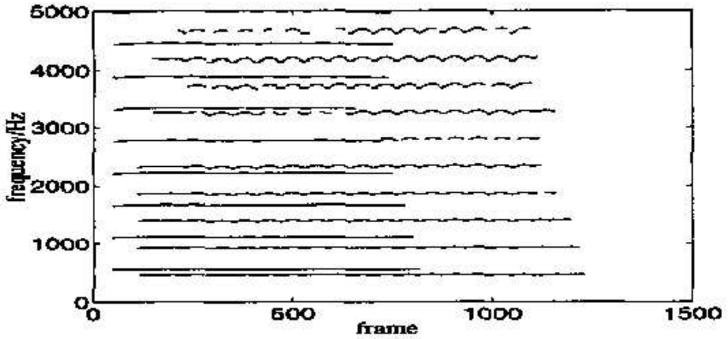

Figure 2.3: Sinusoidal trajectories of a signal consisting of oboe and violin sounds [198].

(periodic) content of the signal.

Periodic representation of audio signal

For treating periodic signals, the state of the art is limited to the estimation of pure periodic signals with period equal to an integer number of samples [134, 135]. In these references, the authors propose a Maximum Likelihood (ML) approach to analyze pure periodic signals. They show that the resulting procedure can be interpreted as a signal projection onto suitable subspaces. In [135], the audio record is assumed as a mixture of unknown periodic signals in white Gaussian noise of unknown intensity. Authors propose a pitch detection scheme based on a ML formulation. The proposed scheme searches for the M largest periodicities, in the sense of the norm of projection onto the subspace spanned by signals of period T.
In [134], Muresan and Parks extend the previous approach. They introduce the concept of exact periodicity. A signal exactly period T is not exactly period $2T$, $3T$, etc... Although, it continues to be of period $2T$, $3T$, etc... The Exactly Periodic Subspace Decomposition (EPSD) is performed using projection on the "orthogonal" subspaces spanned by the exactly periodic signals. Then, EPSD is done by projecting the signal onto orthogonal Exactly Periodic Subspace (EPS), which is similar to Fourier and wavelet decompositions. The difference, however, is that unlike the Fourier or wavelet decomposition, which take projections onto orthogonal vectors, the EPSD takes projections

onto orthogonal subspaces. Viewing the problem in the frequency domain reveals a simple algorithm for computing the EPSD. Indeed, any signal of period T can be written as a linear combination of the harmonics at frequencies $0, \left(\frac{1}{T}\right), \cdots, \left(\frac{T-1}{T}\right)$ (see figure (2.4)).

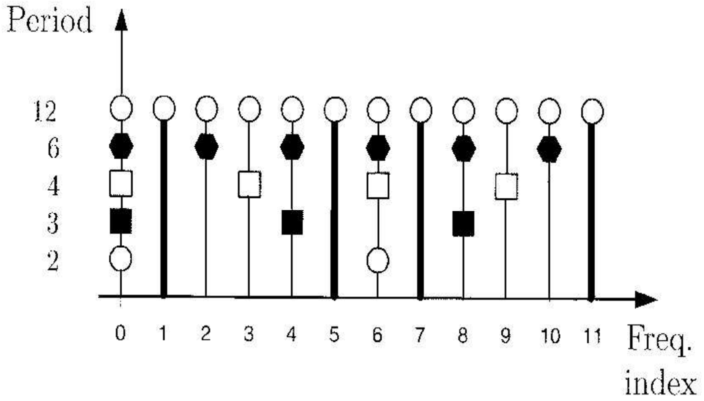

Figure 2.4: Frequency content of a periodic signal (with period $T = 12$.

In general, the basis functions for the subspaces of exactly period T signals is spanned by harmonics that are multiples of the fundamental frequency $\frac{1}{T}$ but not multiples of any frequency $\frac{1}{\overline{T}}$ (with $\overline{T}/T$). For example, the subspace of exactly periodic signal with period 12 is the collection of harmonics that belong only to period 12 (i.e., the harmonics with k=1, 5, 7, and 11).

The decomposition of audio signal into periodic features was reconsidered by De Cheveigné and Slama [164]. The authors introduce (for a given candidate period T) the periodic-aperiodic decomposition:

$$x^{+T} = [x(t) + x(t-T)]/2$$
$$x^{-T} = [x(t) - x(t-T)]/2$$

These processes corresponds to a time-domain comb-filtered version of x. They can be interpreted as the "periodic" and "aperiodic" parts of x. Indeed, if x is periodic with period T, the $x^{+T} = x$, and $x^{-T} = 0$. The authors suggest using periodic-aperiodic power decomposition for the acoustic scene analysis, and they apply the proposed approach for extraction of periodic source separation (periods should be integer multiples of the sampling frequency).

2.3 QPSE with global amplitude modulation and global timewarping

In the previous section, we have seen that the major drawback of the sinusoidal modelling based techniques is that they consider the signal as a mixture of a finite number of arbitrary sinusoids, ignoring the harmonic structure of the audio signal. On the other hand, periodic modelling seems to be too rigid to model real signals. Motivated by the previous observation, we propose merging the periodic signal analysis and sinusoidal modeling in order to give more flexibility to the periodic signal analysis, and to impose more structure on sinusoidal modelling.

In the sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids:

$$s(n) = \sum_{k=0}^{P} a_k(n) \cos\left(\theta_k(n)\right) \tag{2.12}$$

where $\theta_k(n)$ represents the instantaneous phase of the k^{th} partial. Since the audio signal is almost harmonic (quasi-periodic), $\theta_k(n)$ can be decomposed into

$$\theta_k(n) = 2\pi k n f_0 \ + \ 2\pi\varphi_k(n) \tag{2.13}$$

where $\varphi_k(n)$ characterizes the evolution of the instantaneous phases around the k^{th} harmonic, and can be assumed to be slowly time varying.
The global modulation assumption implies that all harmonic amplitudes evolve proportionally in time, and that the instantaneous frequency of each harmonic is proportional to the harmonic index, i.e.,

$$\left\{ \begin{array}{l} a_k(n) = a_k \ a(n) \\ 2\pi\varphi_k(n) = 2\pi k \ \varphi(n) \ + \ \Phi_k \end{array} \right. . \tag{2.14}$$

In summary, we model an audio signal as the superposition of harmonic components with a global amplitude modulation and global time warping:

$$\begin{aligned} y(n) \;&=\; s(n) \;+\; v(n) \\ &= \textstyle\sum_k a_k(n) \;\cos\left(2\pi k n f_0 + 2\pi \varphi_k(n)\right) + v(n) \\ &= a(n) \textstyle\sum_k a_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right) + v(n) \\ &= a(n)\; \theta\left(n + \frac{\varphi(n)}{f_0}\right) \;+\; v(n) \end{aligned} \tag{2.15}$$

where

- $v(n)$ is additive white Gaussian noise.
- $a(n)$ represents the amplitude modulating signal
- $\varphi(n)$ denotes the phase modulating signal (that can be interpreted in terms of time warping).
- $\theta(n) = \sum_k a_k \cos\left(2\pi k f_0 n + \Phi_k\right)$ is a $T = \frac{1}{f_0}$ is the basic periodic signal.

Thus, the audio signal is modeled as a periodic signal with global amplitude and phase modulation. The periodic signal $\theta(n)$ characterizes the spectrum envelope of the audio source. It can be considered as a signature for instrument classification and recognition applications. Whereas the amplitude and phase modulated signals ($a(n)$ and $\varphi(n)$) represent respectively the time evolution of the note power and pitch. Remark also that the global phase modulation can be interpreted in terms of dynamic time-warping: it "warps" (stretches or compresses in time) the basic periodic signal $\theta(n)$ to fit the received signal $s(n)$.

Global phase/frequency modulation

Consider, first, the case of a pure periodic signal:

$$s(n) = \sum_k a_k \cos\left(2\pi k n f_0 + \Phi_k\right) \quad . \tag{2.16}$$

If the fundamental period $T = \frac{1}{f_0}$ is an integer, then $\boldsymbol{\theta} = [\theta(1) \cdots \theta(T)]^T$, the signal over one period, is sufficient to describe the totality of the periodic signal $\mathbf{s} = [s(1) \cdots s(N)]^T$ [134, 135]:

$$\mathbf{s} = \begin{bmatrix} I_T \\ I_T \\ \vdots \end{bmatrix} \theta = \mathbf{F}\boldsymbol{\theta} \tag{2.17}$$

where the column space of $\mathbf{F}$ corresponds to the signal subspace for a periodic signal of period T. When T is not integer, we shall take the vector $\boldsymbol{\theta}$ of size $\lceil T \rceil$ (and not longer, to minimize identifiability problems). $\boldsymbol{\theta}$ contains a set of sufficient statistics to describe the whole periodic signal. And, it exists an optimal interpolation matrix $\mathbf{F}$ that generate exactly the periodic signal vector $\mathbf{s}$ (according to (2.17)). However, this interpolation matrix is quite complex to compute and to handle. Therefore, we consider an approximated interpolation and we work with FIR interpolation filters. We introduce an $\mathbf{F}$ of similar structure as in (2.17), but with the identity matrices replaced by banded blocks corresponding to the (time-varying) FIR filter. For simplicity, we shall assume that a certain degree of oversampling has been performed so that simple linear interpolation (corresponding to a triangular interpolation filter) produces good results. In that case, the matrix $\mathbf{F}$ is given by:

$$\mathbf{F} \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 & \\ \beta & 1{-}\beta & \cdots & 0 & \\ & \ddots & & & \\ & & \ddots & & \\ 0 & & \beta\lfloor T \rfloor & 1{-}\beta\lfloor T \rfloor & \\ 1{-}\beta\lceil T \rceil & & \cdots & \beta\lceil T \rceil & \\ & \vdots & & & \end{bmatrix} \tag{2.18}$$

where $\beta = 1 - \frac{T}{\lceil T \rceil}$. This is a banded matrix with only two consecutive (in a modulo sense) non-zero elements per-row , providing a convex combination of two available samples to approximate an intermediately positioned sample.

The same approach can be used to take into account a given time warping by considering $f(n) = f_0 + \psi(n)$, being a piecewise constant function of time (see figure 2.5). As a result, the phase becomes a piecewise linear function of time. The time period T_1 over which the instantaneous frequency is supposed to be constant is chosen such that $\frac{1}{T_1}$ exceeds (well) the (assumed) bandwidth of variation of the instantaneous frequency. As a result, the frequency and

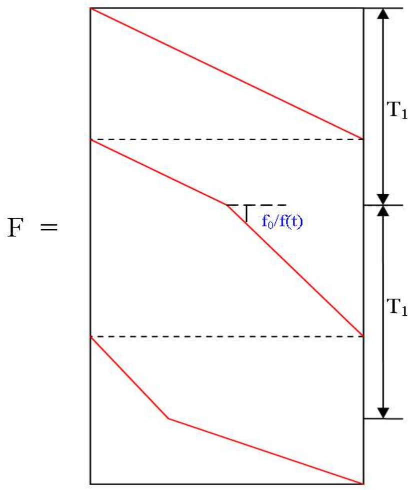

Figure 2.5: Interpolation matrix structure.

hence phase variation gets parameterized by the subsampled values at rate $\frac{1}{T_1}$. The way figure 2.5 should be interpreted is that the line indicates for each row of the matrix the point for which an interpolation value has to be provided. In the case of simple linear interpolation the two matrix elements on the row surrounding the intersection of the line with the row will correspond to an appropriate convex combination.

Global amplitude modulation

While time warping focuses on the time evolution of the instantaneous frequency and allows the modeling of several musical phenomena (vibrato, glissando ...), the global amplitude-modulating signal allows an evolution of the note power, reflecting attack, sustain, and decay. The amplitude signal is assumed to be a non-negative low-pass signal. Hence it can also be represented as an interpolated version of a subsampled signal (see further). So the audio signal can be written as:

$$\mathbf{y} = \underbrace{\mathbf{A}\ \mathbf{F}\boldsymbol{\theta}}_{=\,\mathbf{s}} + \mathbf{v} \tag{2.19}$$

where :
- $\mathbf{y} = [y(1) \cdots y(N)]^T$, represents the observation vector
- $\mathbf{s} = [s(1) \cdots s(N)]^T$, represents the signal of interest
- $\mathbf{v} = [v(1) \cdots v(N)]^T$, denotes the noise vector
- $\boldsymbol{\theta} = [\theta(1) \cdots \theta(\lceil T \rceil)]$, characterizes the harmonic signature over essentially one period
- $\mathbf{A} = diag[A(1) \cdots A(N)]$, represents the global amplitude modulation signal
- $\mathbf{F}$ is an $N \times \lceil T \rceil$ interpolation matrix characterizing the time warping.

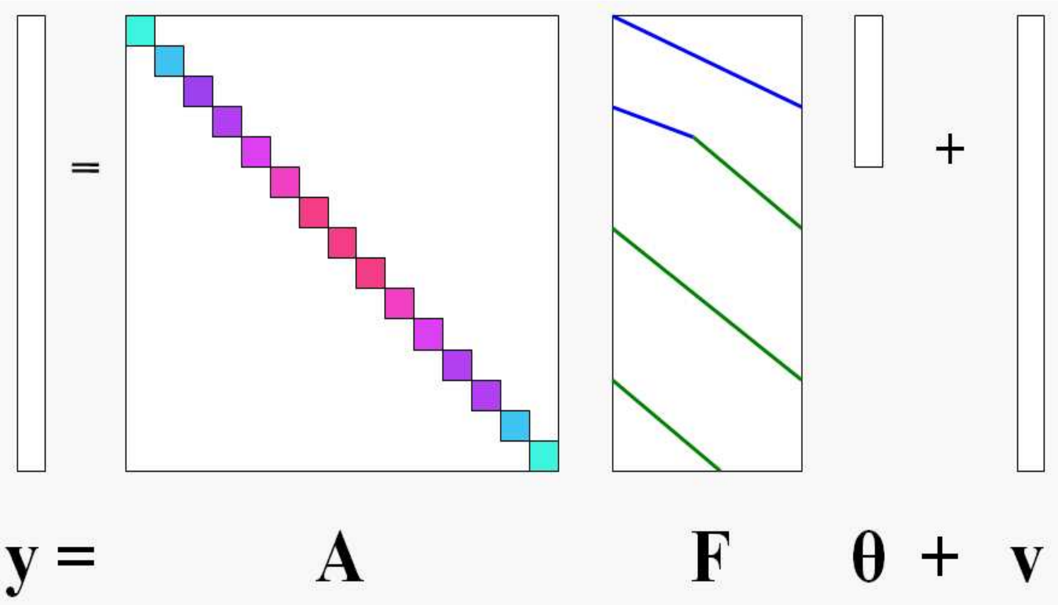

Figure 2.6: QPSE with global amplitude modulation and global timewarping.

2.3.1 Periodic signal extraction procedure

The previous model is linear in $\boldsymbol{\theta}$, $\mathbf{A}$, or $\mathbf{F}$ (separately), $\mathbf{F}$ being parameterized nonlinearly. Trying to estimate all factors jointly is a difficult nonlinear problem. Indeed, as the noise is assumed to be a white Gaussian signal, the ML approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|\mathbf{y} - \mathbf{A}\,\mathbf{F}\,\boldsymbol{\theta}\|^2 \tag{2.20}$$

where $\mathbf{A}$ and $\mathbf{F}$ are parameterized in terms of subsamples. However, the estimation can easily be performed iteratively.

Periodic signature estimation

If we assume that the matrices $\widehat{\mathbf{A}}, \widehat{\mathbf{F}}$ are given, the periodic signature $\boldsymbol{\theta}$ can be isolated as

$$\mathbf{y} = \widehat{\mathbf{A}}\,\widehat{\mathbf{F}}\;\boldsymbol{\theta} + \mathbf{v} = \mathbf{G}_\theta\;\boldsymbol{\theta} + \mathbf{v} \tag{2.21}$$

Then minimizing (2.20) w.r.t. θ leads to

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{G}_\theta{}^T\mathbf{G}_\theta\right)^{-1}\mathbf{G}_\theta{}^T\mathbf{y}\ \ . \tag{2.22}$$

Hence the periodic signature gets estimated by using the data over the whole note duration.

Instantaneous amplitude estimation

The amplitude signal could similarly be estimated from (2.20) by isolating $\widehat{\mathbf{F}}$ and $\widehat{\boldsymbol{\theta}}$. However, such an estimation procedure would not guarantee positive values for the estimated amplitude signal. Alternatively, consider performing the square of the note signal:

$$\begin{aligned} s^2(n) &= a^2(n)\ \left(\textstyle\sum_k a_k\cos(2\pi k n f_0 + 2\pi k\varphi(n) + \Phi_k)\right)^2 \\ &= a^2(n)\,\overline{\theta^2}\ +\ (\text{high}\ \ \text{freq.}\ \ \text{terms}) \end{aligned} \tag{2.23}$$

where $\overline{\theta^2} = \dfrac{1}{2}\displaystyle\sum_k a_k^2 = \dfrac{1}{\lceil T\rceil}\sum_{n=1}^{\lceil T\rceil}\widehat{\theta}^2(n) = \dfrac{1}{\lceil T\rceil}\|\widehat{\boldsymbol{\theta}}\|^2$ denotes the power of the periodic signal. Taking into account additive noise leads to:

$$y^2(n) - v^2(n) = \underbrace{a^2(n)\,\overline{\theta^2}}_{\text{signal}} + \underbrace{2s(n)v(n) + (\text{high}\ \ \text{freq.}\ \ \text{terms})}_{\text{noise}}$$

from which we shall estimate $a^2(n)$ via least-squares. We propose to express the low-pass character of $a(n)$ by taking $a(n)$ to be piecewise constant (so that $a^2(n)$ is also piecewise constant) over a given time frames.The length of theses frames T_a should be selected a multiple of the signal period T. T_a can

be also time-varying, to accommodate for the time-varying speed of variation of the amplitude (attack versus decay).
Finally, $a(n)$ gets estimated using:

$$\widehat{a}(n) = \sqrt{\frac{1}{\overline{\overline{\theta^2}}} \left\langle y^2(n) - (y(n) - \widehat{s}(n))^2 \right\rangle_{T_a}} \tag{2.24}$$

where $\langle\ .\ \rangle_{T_a}$ denotes temporal averaging over the piecewise interval containing T_a; $\widehat{s}(n) = \left[\widehat{\mathbf{A}}\widehat{\mathbf{F}}\widehat{\boldsymbol{\theta}}\right](n)$ denotes the latest estimate of the signal of interest.

Instantaneous frequency estimation

As for the instantaneous amplitude, the instantaneous frequency gets estimated on a frame-by-frame basis. The length of these time frames T_f can differ from T_a and is typically longer since the frequency varies more slowly than the amplitude. In each frame, the instantaneous frequency is optimized using (2.20):

$$\begin{cases} \min\limits_{f} \left\| \mathbf{y} - \widehat{\mathbf{A}}\widehat{\mathbf{F}}(f)\widehat{\boldsymbol{\theta}} \right\| \\ \frac{\Delta f}{f_0} \leq \alpha_{max} \end{cases} \tag{2.25}$$

where Δf denotes the maximum relative frequency variation in the current frame compared to the previous frame, reflecting an assumed limited frequency variation rate. The optimal instantaneous frequency value for the current frame gets determined from a finite set of discrete values within the thus limited range.

2.4 QPSE with global amplitude and phase modulation

In the previous section, we have assumed that the global modulating phase signal (in (3)) is piecewise linear, i.e. $\exists T_{wl}$

$$\varphi_{wl}(n) = n\,(f_p - f_0) + \Phi_p \ \ \forall n \in [pT_{wl} \ \ (p+1)T_{wl}]$$

where $f_{wl}(n) + f_0 = f_p + f_0$ is the instantaneous frequency assumed to be piece-wise constant. In such a case, the global phase modulation can be interpreted in term of time warping. In this section, we relax further our assumptions on the global phase modulation signal:

$$\varphi(n) = nf_0 + \varphi_{wl}(n) + \widetilde{\varphi}(n) \tag{2.26}$$

where $\widetilde{\varphi}(n)$ is assumed to be slowly time-varying and with small magnitude $|2\pi\widetilde{\varphi}(n)| \ll 1$.
Thus, the audio signal can be modeled as:

$$s(n) = a(n) \sum_k a_k \cos\left(2\pi k \left(nf_0 + \varphi_{wl}(n)\right) + 2\pi k \widetilde{\varphi}(n) + \Phi_k\right)$$

A first-order approximation of the phase dependence produces an additive term involving the derivative of the periodic signal multiplied by a phase variation function

$$\begin{aligned} s(n) &\approx a(n) \sum_k a_k \cos\left(2\pi k \left(nf_0 + \varphi_{wl}(n)\right) + \Phi_k\right) \\ &\quad - a(n) \sum_k a_k (2\pi k \widetilde{\varphi}(n)) \sin\left(2\pi k \left(nf_0 + \varphi_{wl}(n)\right) + \Phi_k\right) \\ &= a(n)\theta\left(nf_0 + \varphi_{wl}(n)\right) + a(n) \frac{\widetilde{\varphi}(n)}{f_0 + \varphi^{'}_{wl}(n)} \theta^{'}\left(nf_0 + \varphi_{wl}(n)\right) \\ &= a(n)\theta\left(n + \tfrac{\varphi_{wl}(n)}{f_0}\right) + a(n) \underbrace{\frac{\widetilde{\varphi}(n)}{f_0 + f_{wl}(n)}}_{b(n)} \theta^{'}\left(n + \tfrac{\varphi_{wl}(n)}{f_0}\right) \end{aligned}$$

where $\theta(n) = \sum_k a_k \cos\left(2\pi k n f_0 + \Phi_k\right)$ is the signal defined in (2.15) ($T = \frac{1}{f_0}$ not necessarily an integer).
The derivative $\theta^{'}(n)$ denotes a sampled version of the continuous-time signal of which $\theta(n)$ is the sampled version. If the sampling satisfies Nyquist's criterion, then $\theta^{'}(n)$ can be obtained from $\theta(n)$ by filtering with the transfer function $j2\pi f$, $f \in (-\frac{1}{2}, \frac{1}{2})$ which we shall approximate with an FIR filter that is optimized as follows:

$$H(z) = \sum_{n=-P}^{P} h_n z^{-n} \tag{2.27}$$

$$\min_{h_n} \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f) \left| j2\pi f - H\left(e^{j2\pi f}\right)\right|^2 df \tag{2.28}$$

where P is the order of the FIR derivative filter approximation, and $S_{yy}(f)$ denotes the power spectrum of the signal $y(n)$ (see appendix 2.A for details). In summary, the audio signal can be written as

$$\mathbf{y} = \mathbf{A}\ \mathbf{F}\boldsymbol{\theta} + \underbrace{\mathbf{A}\ \mathbf{B}}_{\mathbf{C}}\ \mathbf{H}\mathbf{F}\boldsymbol{\theta}\ +\ \mathbf{v} \tag{2.29}$$

where :
- $\mathbf{y} = [y(1) \cdots y(N)]^T$, represents the observation vector
- $\mathbf{v} = [v(1) \cdots v(N)]^T$, denotes the noise vector
- $\boldsymbol{\theta} = [\theta(1) \cdots \theta(\lceil T \rceil)]$, characterizes the harmonic signature over essentially one period
- $\mathbf{A} = diag\,\{A(1) \cdots A(N)\}$, represents the global amplitude modulation
- $\mathbf{B} = diag\left\{\frac{\widetilde{\varphi}(1)}{f_0+f_{wl}(1)} \cdots \frac{\widetilde{\varphi}(N)}{f_0+f_{wl}(N)}\right\}$, characterizes the global phase modulation
- $\mathbf{F}$ is an $(N+2P) \times \lceil T \rceil$ is an interpolation matrix characterizing the time-warping
- $\mathbf{H}$ is an $N \times (N+2P)$ banded matrix that characterizes the derivative filter

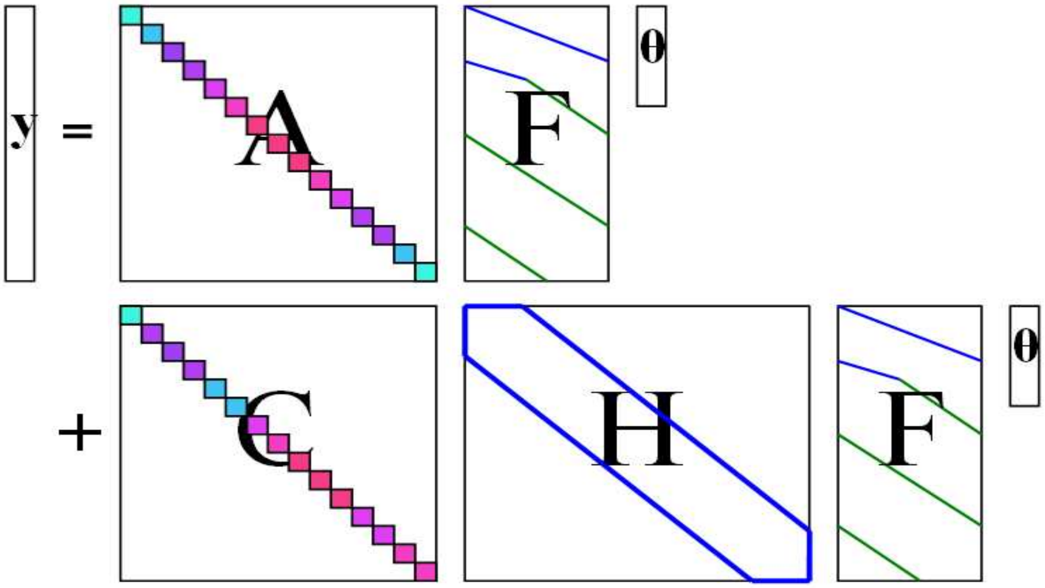

Figure 2.7: QPSE with global amplitude and phase modulation.

2.4.1 Periodic signal extraction procedure

As previous, the signal model in (2.29) is linear in $\boldsymbol{\theta}$, $\mathbf{A}$, and $\mathbf{C}$ (separately). Trying to estimate all of them at the same time is a difficult nonlinear problem. However, the estimation can easily be done iteratively.

Harmonic signature estimation

If we assume that the matrices $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{F}}$ are given, harmonic signature θ can be isolated as

$$\mathbf{y} = \widehat{\mathbf{A}} \left(\mathbf{I} + \widehat{\mathbf{B}}\mathbf{H}\right) \widehat{\mathbf{F}}\, \boldsymbol{\theta} + \mathbf{v} = \breve{\mathbf{G}}_\theta\, \boldsymbol{\theta} + \mathbf{v} \tag{2.30}$$

As the noise is assumed to be a white Gaussian signal, the ML approach leads to a least-squares problem. Then

$$\widehat{\theta} = \left(\breve{\mathbf{G}}_\theta^T \breve{\mathbf{G}}_\theta\right)^{-1} \breve{\mathbf{G}}_\theta^T Y \tag{2.31}$$

Instantaneous frequency estimation

As previously, the instantaneous frequency gets estimated on a frame-by-frame basis. In each frame, the instantaneous frequency is optimized using (2.25).

Instantaneous global amplitude and phase estimation

Using the current estimates of $(\widehat{F}, \widehat{\theta})$, the audio source signal can be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\widehat{\mathbf{F}}\widehat{\boldsymbol{\theta}} + \mathbf{C}\mathbf{H}\widehat{\mathbf{F}}\widehat{\boldsymbol{\theta}} + \mathbf{v} \\ &= \breve{\mathbf{G}}_{\mathbf{a}}\, \mathbf{A} \ + \ \breve{\mathbf{G}}_{\mathbf{c}}\, \mathbf{C} + \mathbf{v} \end{aligned}$$

where $\breve{\mathbf{G}}_{\mathbf{a}} = \operatorname{diag}\left\{\widehat{\mathbf{F}}\widehat{\boldsymbol{\theta}}\right\}$, and $\breve{\mathbf{G}}_{\mathbf{c}} = \operatorname{diag}\left\{\mathbf{H}\widehat{\mathbf{F}}\widehat{\boldsymbol{\theta}}\right\}$ are two $N \times N$ diagonal matrices.

On the other hand, the global modulating amplitude (respectively phase) signals are supposed to be lowpass band. Then, $a(n)$ (resp. $c(n)$) can be down-sampled. The remaining samples can be estimated using linear interpolation. Linear interpolation can be formalized as a linear transformation:

$$\begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ \vdots \\ a(N) \end{bmatrix} = \begin{bmatrix} 1 \quad 0 \cdots 0 \\ P_{21} P_{22} \cdots 0 \\ P_{31} P_{32} \cdots 0 \\ 0 \quad 1 \cdots 0 \\ \vdots \\ 0 \cdots \cdots 1 \end{bmatrix} \begin{bmatrix} a(1) \\ a(1+T_a) \\ \vdots \\ a(N) \end{bmatrix} = \mathbf{P}_a \, \mathbf{a}_{\downarrow}$$

where $\mathbf{a}_{\downarrow}$ (resp. $\mathbf{c}_{\downarrow}$) contains the degrees of freedom of our model (characterizing the amplitude (resp. phase) modulating signal), and $\mathbf{P}_a$ (resp. $\mathbf{P}_c$) represents the interpolation matrix of the global modulating amplitude (resp. phase). For the design of the interpolation matrices P_a and P_c, we suggest using Hamming window for linear interpolation. In fact, the linear interpolation can be interpreted as a linear filtering operation of the upsampled signal with a low-pass filter (see appendix 2.A for details). By using a smooth window with energy concentrated essentially in the principal lobe, the interpolation error gets amplified less. Thus, we do better amplitude and phase estimation.

We can also vary the interpolation window length with time. In fact, in the transient state the instantaneous amplitude is much more large-band than in harmonic steady-state. Therefore, one should use windows in the transient portions (to best model the audio signal attack and decay), and longer windows in the harmonic steady-state (to better remove the additive noise).

In sum, the estimation problem can be formalized as follows

$$\mathbf{y} = \left[\breve{\mathbf{G}}_{\mathbf{a}} \mathbf{P}_a \quad \breve{\mathbf{G}}_{\mathbf{c}} \mathbf{P}_c \right] \begin{bmatrix} \mathbf{a}_{\downarrow} \\ \mathbf{c}_{\downarrow} \end{bmatrix} + \mathbf{v}$$

and, $\widehat{\mathbf{A}}$, and $\widehat{\mathbf{C}}$ gets estimated using the least-squares technique (via the estimation of $\begin{bmatrix} \hat{\mathbf{a}}_{\downarrow} \\ \hat{\mathbf{c}}_{\downarrow} \end{bmatrix}$).

As expected, simulation shows that applied to clean music signal, the periodic signal extraction based on global amplitude and phase modulation performs better than the technique based on global time warping. However using noisy received signal, one can show that the proposed technique is not robust to initialization (in the cyclic optimization problem). In fact, the estimation of the parameters $\mathbf{A}$ and $\mathbf{C}$ strongly relies on the quality of estimation of the periodic signal $\boldsymbol{\theta}$ and the time-warping matrix $\mathbf{F}$. Therefore, we suggest using the first version (modeling the audio signal using a global amplitude modulation and global time-warping) for the algorithm initialization. Then, we refine the estimation taking into consideration small low-pass phase fluctuation.

2.5 Audio Modeling with Global Frequency-Selective Modulation and Global Time-Warping

In previous, we have presented the quasi-periodic signal models with global (flat) amplitude and frequency modulation. Such a model allows for no spectral variation throughout the note duration, only for amplitude and (synchronized) frequency modulation. The global amplitude modulation implies that all harmonic amplitudes evolve proportionally in time; whereas the global time-warping emphasizes the signal harmonicity. However, the ratio of the different harmonics (modeled through the basic waveshape $\boldsymbol{\theta}$) is assumed to be constant throughout the whole note duration.

The problem with such a model though is that in reality, periodic signals produced by musical instruments, e.g. string instruments, have harmonic components that decay at different speeds. Typically higher harmonics decay faster than lower harmonics. This means that the global amplitude modulation assumption is not satisfied.

The assumptions of global amplitude and frequency modulation were introduced to have a parsimonious signal representation. Indeed, the higher the number of parameters per second describing the signal, the noisier the parameter estimates, and consequently the reconstructed signal estimate. Introducing an amplitude modulating signal per harmonic would allow significant degrees of freedom in describing the signal, but would lead to a high parameter rate (the average number of parameters that appear in the description of one second of the signal). An intermediate parameter rate can be obtained by filtering the periodic signal with a short FIR filter that can introduce frequency-selective attenuation, and this in a time-varying fashion to reflect the time-varying amplitude.

In summary, we model the audio signal as a superposition of harmonic components with global frequency-selective amplitude modulation and global time-warping, i.e.,

$$y(n) = a_n(q)\ \theta\left(n + \frac{\varphi(n)}{f_0}\right)\ +\ v(n) \tag{2.32}$$

where $a_n(q) = a_{n,L}q^L + \cdots + a_{n,0} + \cdots + a_{n,L}q^{-L}$ is a symmetric linear-phase FIR filter, $2 * L + 1$ is the amplitude modulating filter length, and q^{-1} is the time delay operator.

Using matrix notations, the audio signal gets expressed as in (2.19), where the diagonal matrix $\mathbf{A}$ (characterizing the global amplitude modulation) is replaced by an $L+1$ symmetric band matrix (figure 2.8).

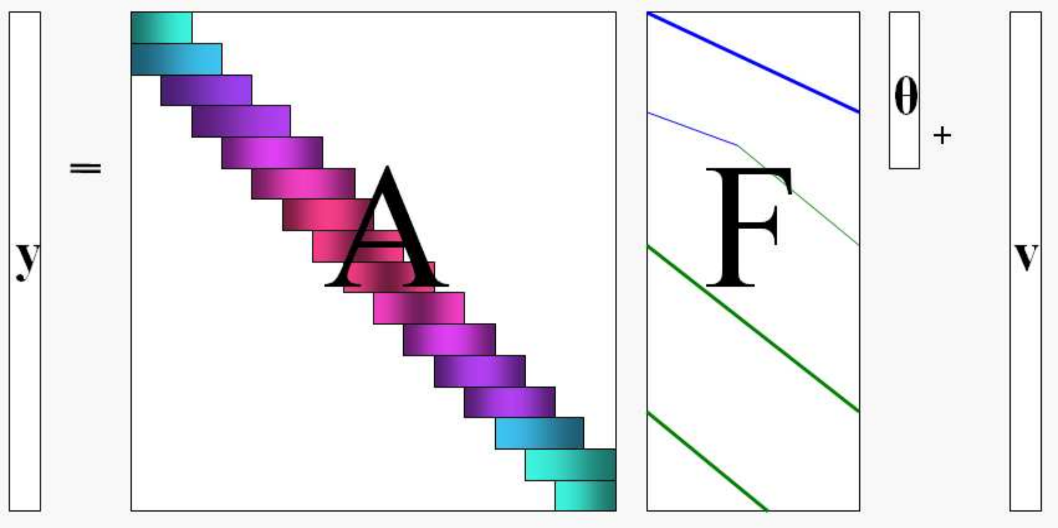

Figure 2.8: QPSE with global frequency-selective amplitude modulation and global timewarping.

The rows of $\mathbf{A}$ contain the coefficients of the FIR modulating amplitude ($a_n(q)$). These filters model the evolution of the note power as well as the relative decay of the different harmonics. Typically, as high frequencies decay faster than low frequencies, the modulating filters become more and more low-pass.

2.5.1 Instantaneous frequency selective attenuation estimation

As in previous, the freedom degrees of the model get estimated in an iterative (cyclic) fashion. Assuming a white Gaussian noise, the ML approach leads to simples least-squares problems. The estimation of the harmonic signature and the instantaneous frequency are performed as in the section 2.3.
If we assume that the normalized waveshape $\theta(n)$ and the time-warping function $\varphi(n)$ (via $\mathbf{F}(f)$) are given, the received signal $y(n)$ is linear with respect to the amplitude modulating filter coefficients, i.e.,

$$\begin{aligned} y(n) &= a_{n,0}\breve{\theta}(n) + \sum_{i=1}^{L} a_{n,i}\left(\breve{\theta}(n-i) + \breve{\theta}(n+i)\right) + v(n) \\ &= \left[\breve{\theta}(n) \cdots \breve{\theta}(n{-}L){+}\breve{\theta}(n{+}L)\right] \begin{bmatrix} a_{n,0} \\ \vdots \\ a_{n,L} \end{bmatrix} + v(n) \\ &= \qquad\quad \breve{\boldsymbol{\theta}}(n) \qquad\qquad\quad \mathbf{a}(n) \;\; + \;\; v(n) \end{aligned}$$

where $\breve{\theta}(n) = \theta\left(n + \frac{\varphi(n)}{f_0}\right)$ is the warped normalized waveshape. Thus, using the current estimates of $(\widehat{\mathbf{F}}, \widehat{\boldsymbol{\theta}})$, the observation vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{G}_a \; \mathbf{a} + \mathbf{v}$$

where $\mathbf{G}_a$ is a $N{\times}(N(L{+}1))$ block diagonal matrix, and $\mathbf{a} = [\mathbf{a}(1)^H \cdots \mathbf{a}(N)^H]^H$ is a $(N(L+1)) \times 1$ vector characterizing the amplitude modulation.
On the other hand, the coefficients of the frequency-selective modulating filter signals are assumed to be lowpass band. Therefore,
$\{a_{n,i}\}_{i=0:L}$ can be down-sampled. The remaining samples can be estimated using linear interpolation, i.e.,

$$\mathbf{a}_i = \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ \vdots \\ \vdots \\ a_{N,i} \end{bmatrix} = \begin{bmatrix} 1 \;\; 0 \cdots 0 \\ P_{21} P_{22} \cdots 0 \\ P_{31} P_{32} \cdots 0 \\ 0 \;\; 1 \cdots 0 \\ \vdots \\ 0 \;\cdots\;\cdots\; 1 \end{bmatrix} \begin{bmatrix} a_{1,i} \\ a_{T_a+1,i} \\ \vdots \\ a_{N,i} \end{bmatrix} = \;\; \mathbf{P}_a \; \mathbf{a}_{\downarrow i}$$

where $\mathbf{a}_{\downarrow i}$ contains the i^{th} coefficients of the frequency selective modulating filter $a_n(q)$, downsampled by the factor T_a. $\mathbf{P}_a$ represents the interpolation matrix used to reconstruct $\mathbf{a}_i$ from its downsampled version $\mathbf{a}_{\downarrow i}$ (see appendix 2.B for further discussion on the design of $\mathbf{P}_a$). In summary, the estimation problem can be formalized as:

$$\mathbf{y} = \underbrace{\mathbf{G}_a \left(\mathbf{P}_a \otimes \mathbf{I}_{L+1}\right)}_{\mathbf{G}_{\downarrow a}} \ \mathbf{a}_{\downarrow} + \mathbf{v} \tag{2.33}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{a}_{\downarrow} = \left[\mathbf{a}^H(1)\ \mathbf{a}^H(T_a+1) \cdots \mathbf{a}^H(N)\right]^H$ represents the freedom degrees of our model. Thus, the elements of $\widehat{\mathbf{A}}$ are estimated using the least-squares technique (via the estimation of $\mathbf{a}_{\downarrow}$).

Experimental results [180] validate that the global frequency-selective modulation based model fits real audio signals better than flat amplitude modulation based approach. However, simulations shows that only short FIR filter is required to model the diverse mode variations (no considerable gain are noticed by taking long filters).
On the other hand, the frequency-selective modulation induces additional degrees of freedom. Such a model leads to a parsimonious signal representation (decreasing modeling error). However, in noisy environment, the higher the number of parameters describing the signal, the noisier the parameter estimates and consequently the reconstructed signal estimate. That is why at high SNR, the frequency-selective modulation outperforms the flat modulation (the estimation error is neglected). However, at low SNR, flat amplitude modulation produces comparable enhancement accuracy.

2.6 Implementation issues and experimental results

2.6.1 Implementation and complexity issues

We comment the implementation of the Quasi-Periodic Signal Extraction (QPSE) algorithm based on global amplitude and phase modulation. The complexity of the scheme is investigated at the end of this section. Despite the model parameters being estimated using iterative (cyclic) LS approaches, the QPSE scheme can be complemented in an efficient way: by exploiting the sparsity and the structure of the interpolation matrices $\mathbf{F}$, $\mathbf{P}_a$, and $\mathbf{P}_c$,

one can considerably reduce the required memory and the computation complexity.
In fact as linear interpolation is used to upsample the frequency, amplitude, and phase modeling signals, each row of the matrices $\mathbf{F}$, $\mathbf{P}_a$, and $\mathbf{P}_c$ contains at most two non-zero elements. In addition, for two non-adjacent columns, the sets of non-zero elements do not overlap. So that for a $N \times M$ interpolation matrix $\mathbf{P}$ ($\mathbf{P} = F,\ \mathbf{P}_a,\ \text{or}\ \mathbf{P}_c$), the matrix $\mathbf{P}^H\mathbf{P}$ is a tri-diagonal matrix. In addition, taking into account the matrix structure, the computation complexity of such operation is $4N$ (instead of MN^2). Moreover, for a given $N \times 1$ vector $\mathbf{y}$, $\mathbf{P}^H\mathbf{y}$ can be computed using $2N$ multiplications (instead of MN).
Next, one can show that for a given K-band matrix $\mathbf{G}$, $\breve{\mathbf{G}} = \mathbf{P}^H\ \mathbf{G}\ \mathbf{P}$ is a $(K+2)$-band matrix. Thus, to solve the linear system $\breve{\mathbf{G}}\mathbf{x} = \mathbf{b}$, one should consider the LDU decomposition rather than the inversion of the matrix $\breve{\mathbf{G}}$. In such a case, the upper diagonal matrix L in the LDU decomposition is also a band matrix (figure 2.9).

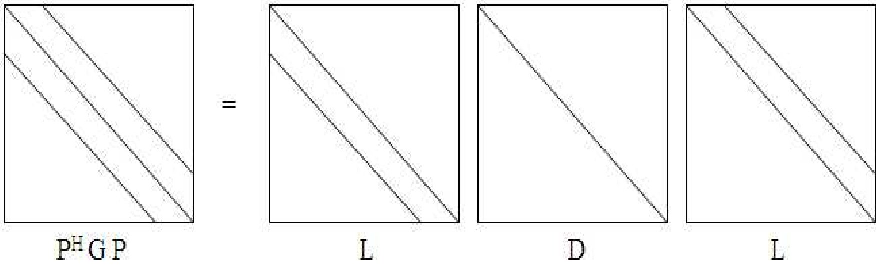

Figure 2.9: LDU decomposition of band matrix.

One the LDU decomposition is performed, the linear system comes back to a simple forward, instantaneous, and backward triangular systems. Thus, the computation complexity of the linear system inversion is $O\left(K^2M\right)$ (instead of $O\left(M^3\right)$).

2.6.2 Application to music signal enhancement

In this section, we will consider the application of the QPSE algorithm to music signal enhancement. First, the use of the phase and amplitude global modulation model is motivated by analyzing the stringed instruments sound production and validated through audio example. Then, the enhancement accuracy of the proposed algorithm is investigated and compared to the classic MP-based approaches.

The physics of how the sound is produced by the bowed stringed instruments is highly complicated and still not entirely understood. Hence, in the following, we try to provide a basic explanation of how these instruments work. Basically, stringed instruments produce sounds by bowing or plucking the strings. The vibrations of the strings have the form of standing waves which produce the combination of a fundamental frequency and a mixture of partials (almost multiple integers of the fundamental frequency):

- The fundamental frequency f_0 produced by a given string depends on the mass per unit length μ of the string, the length L and tension T of that string. This relationship is given by the formula:

$$f_0 = \frac{1}{2L}\sqrt{\frac{T}{\mu}}, \tag{2.34}$$

- The harmonics make the sound timbre fuller and richer than the fundamental alone. The energy and the frequencies of the harmonics depend upon the tension, mass and length of the string. They depend also upon the method of excitation of the string (bowing, plucking...).

In addition to the string properties and the method of excitation of strings, the sound timbre is significantly affected by resonances in the body of the instrument itself (figure 2.10). For instance, in the case of the violin, the resonances have been thoroughly studied and to them are attributed the excellent qualities of the legendary Stradivarius and others.
In sum, the three basic features of a musical sound are [95]:

- Pitch: related to the perception of the fundamental frequency of the sound.

- Intensity: related to the amplitude, and thus to the energy, of the vibration.

- Timbre: related to resonance in the body of the instrument, and the method of excitation of the string.

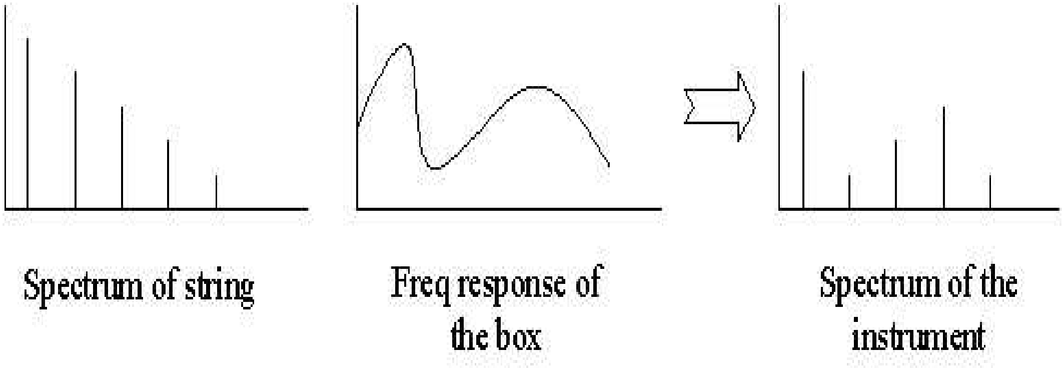

Figure 2.10: Stringed instrument response spectrum.

Thus, periodic signal modelling with global amplitude and phase modulation seems to be convenient for musical signals. In fact, the tree basic features of the musical sound are modeled by three distinct quantities: the instantaneous phase (frequency) models the pitch variations,the instantaneous amplitude models the intensity variation, whereas the basic periodic signal θ models the timber characteristics.

Next, we use real musical record to validate the use of the previous model for music signal representation. Figures 2.11, and 2.12 plot respectively the spectrogram, and the evolution of the harmonics energy in time domain for a guitar and a piano signals.

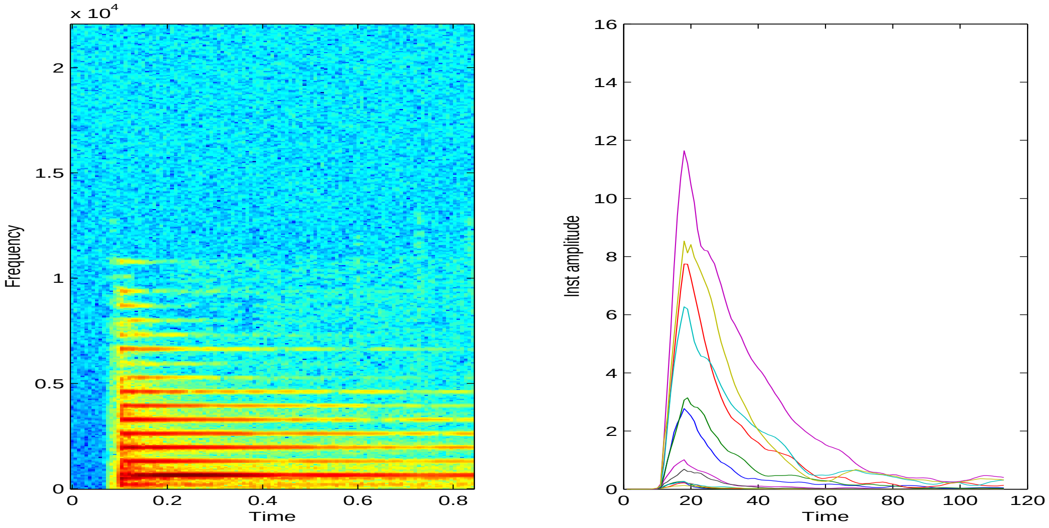

Figure 2.11: The spectrogram, and the time evolution of the harmonics energy for a guitar signal.

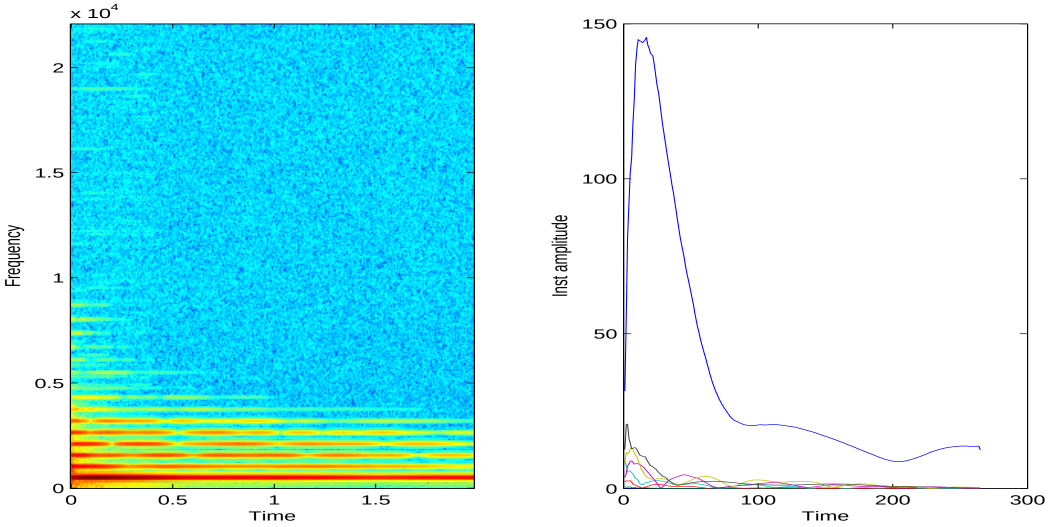

Figure 2.12: The spectrogram, and the time evolution of the harmonics energy for a piano signal.

The spectrogram shows that the signals are quite harmonic, and that the different harmonics evolve proportionally with time. Thus, the model assumptions seem to be hold for string instrument audio signal. Similar results are obtained by analyzing wind instrument audio signal (see figures 2.13, and 2.14)

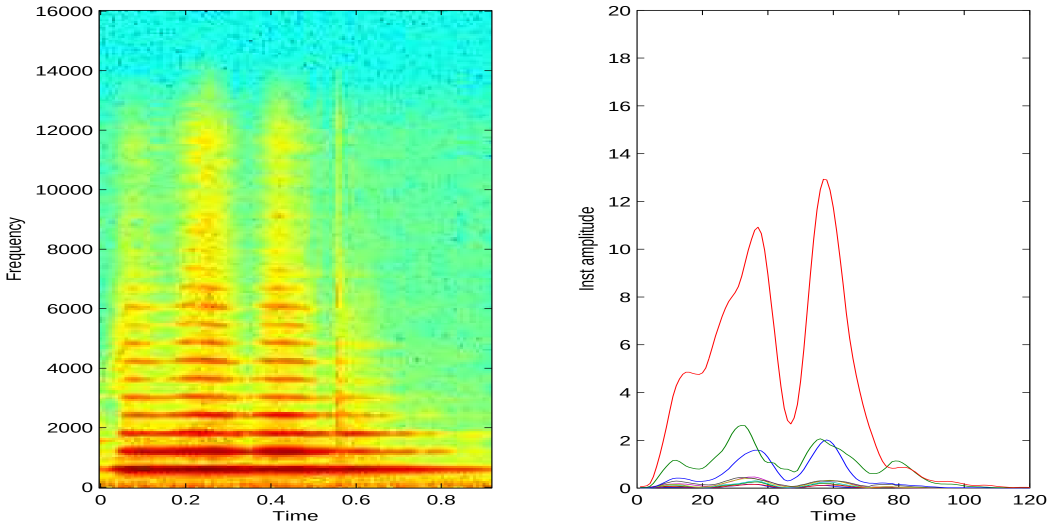

Figure 2.13: The spectrogram, and the time evolution of the harmonics energy for a flute signal.

As a result, the QPSE algorithm can be applied to music signal enhancement. The noise reduction is performed by extracting the audio components from the received signal. If the noise variance (σ_v^2) is available, this prior information can be used to enhance the estimation of the instantaneous amplitude $\left(\widehat{a}(n) = \sqrt{\frac{1}{\theta^2}\left\langle y^2(n) - \sigma_v^2 \right\rangle_n}\right)$.

On the other hand, MP-based approaches are also proposed to solve the noise reduction problem [22]. Once the received signal is decomposed into atoms, noise reduction is performed by classifying the atoms into "noise" vs. "signal"; then resynthesizing the enhanced audio signal.

To compare the enhancement accuracy of the two approaches, we have experimented with a real music signal. The proposed signal represents a single note (pitch = 84 Hz) played by an acoustic guitar. The record has a duration of 1s and is sampled at 22.050 Khz (see figure 2.15). A synthetic Gaussian white noise is added to the audio signal. Furthermore, we consider

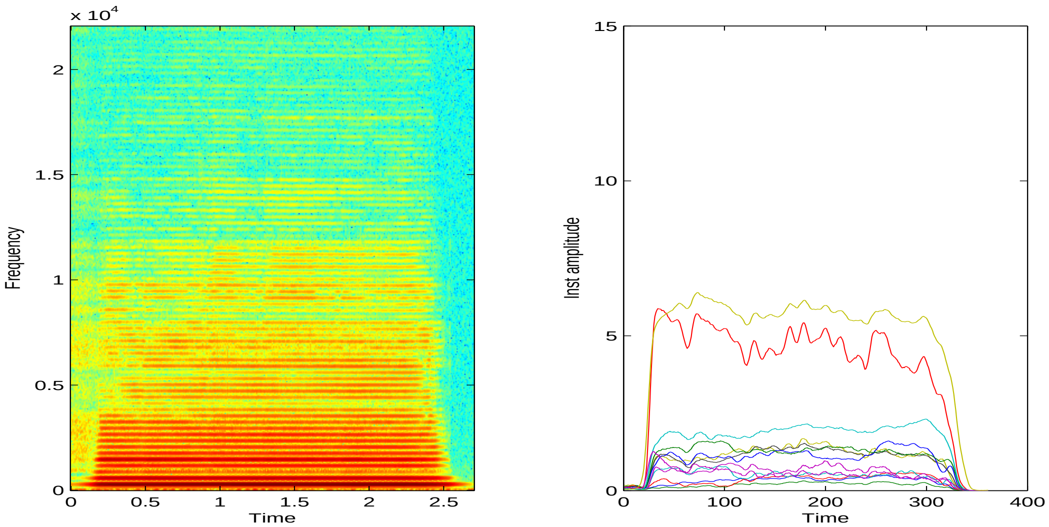

Figure 2.14: The spectrogram, and the time evolution of the harmonics energy for a saxophone signal.

the global signal-to-noise ratio (SNR_{out}) (possibly limited to the steady-state portion) as an objective evaluation criterion

$$SNR_{out} = 10 \log \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} (s(n) - \widehat{s}(n))^2}$$

which is consistent with previous enhancement studies [74, 155]. Fig. 2.16 plots curves of the averaged output SNR (evaluated by Monte-Carlo techniques) of the Matching Pursuit and Quasi-Periodic Signal Extraction techniques.

We observe that the QPSE and MP approaches have comparable enhancement performance. However, the QPSE approach outperforms the MP in the steady-state region (where the quasi-periodic model allows a better fit of the audio signal). The MP is better in the transition region, where the structure of QPSE is too constrained. We also remark that knowing the noise variance does not significantly increase the enhancement performance for the QPSE approach.

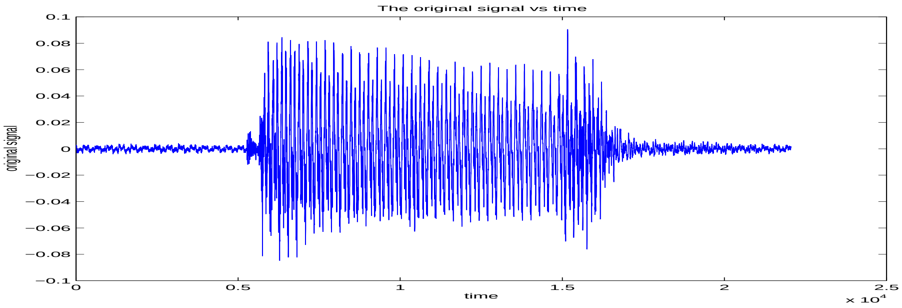

Figure 2.15: Original guitar signal.

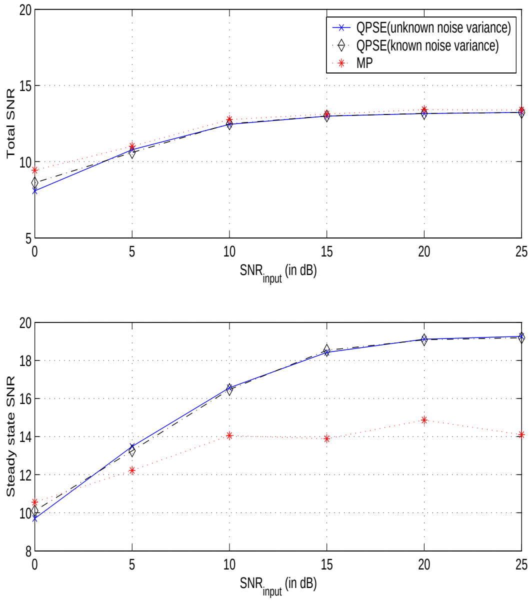

Figure 2.16: Noise Reduction (MP on solid line, QPSE on dotted line).

2.6.3 Application to speech signal enhancement

Speech enhancement can be described as the processing of speech signals to improve one or more perceptual aspects of speech, such as overall quality, intelligibility for human or machine recognizers, or degree of listener fatigue. The need for enhancing speech signals arises in many situations in which the speech either originates from some noisy location or is affected by noise over the channel or at the receiving end. In the presence of background noise, the human auditory system is capable of employing effective mechanisms to reduce the effect of noise on speech perception. Although such mechanisms are not well understood at the present state of knowledge to allow the design of speech enhancement systems based on auditory principles, several practical methods for speech enhancement have already been developed. Several reviews can be found in the literature [106, 129, 46].

On the other hand, the nature of the human speech dictates that not every short segment can be treated in the same fashion. In fact, speech segments can be classified in terms of the sounds they produce [111]. Basically, there are two sound categories: i) Unvoiced sounds, such as the /s/ in 'soft', are created by air passing through the vocal tract without the vocal cords vibrating. They exhibit low signal energy, no pitch, and a frequency spectrum biased towards the higher frequencies of the audio band, ii) Voiced sounds, such as /AH/ in 'and', are created by air passing through the glottis causing it to vibrate. And contrarily to unvoiced speech, voiced speech has greater signal energy, a pitch, and a spectrum biased towards the lower frequencies. In order to take advantage of the voicing in the glottal source signal, we propose modelling voiced sounds as a periodic signal with a global amplitude and phase modulation; and to take into account this structure to denoise the voiced segment.

Note that the global modulated signal model can be interpreted in terms of long-term prediction. Long-term prediction is typically used for voiced-speech coding. The most basic long-term predictor is the one tap filter given by

$$s_p(n) \; = \; a \; s(n-T) \tag{2.35}$$

where $s(n)$ is the input signal, $s_p(n)$ is the predicted signal, T is an integer value characterizing the speaker pitch, and a is a positive gain. In [112],

the authors propose a long-term prediction scheme enabling fractional delay. They show that this technique enables a more accurate representation of the voiced speech and achieves an improvement of synthetic quality for female speakers. Our model generalizes the previous approach by allowing tracking (slow) variations of gain and fractional delay (global amplitude and frequency modulation variations). Such an approach enables, not only a good tracking of the signal of interest, but also the rejection of signals having a different structure (white noise, PC noise, car noise, and human voice...), especially if the spectrum of this colored noise is concentrated in different frequency regions than the voiced speech.
Remark also that the described extraction technique models, and takes advantage of the correlation between the different partials. And contrary to classical sinusoidal modeling based techniques, it does not make any assumption on the value of P (in (2.11)). Implicitly, P is the maximum integer such that $f_0 P < \frac{1}{2}$ (the sampling frequency satisfy the Nyquist-Shannon sampling theorem).

Speech enhancement technique

The proposed enhancement algorithm (figure 2.17) is based on a different treatment of the voiced and unvoiced speech components. The processing steps are discussed in the following sections.

i) Enhancement Stage

i-a) Voiced speech extraction: As the voiced speech signal is assumed to be quasi-periodic (following (2.29)), it can be written as

$$\mathbf{s} \approx \mathbf{A}\ \mathbf{F}\boldsymbol{\theta}$$

The previous model is linear in $\boldsymbol{\theta}$, $\mathbf{A}$, or $\mathbf{F}$ (separately), $\mathbf{F}$ being parameterized nonlinearly. As the noise is assumed to be a white Gaussian signal, the Maximum Likelihood (ML) approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|\mathbf{y} - \mathbf{A}\,\mathbf{F}\,\boldsymbol{\theta}\|^2 \tag{2.36}$$

where $\mathbf{A}$ and $\mathbf{F}$ are parameterized in terms of subsamples. Trying to estimate all factors jointly is a difficult nonlinear problem. However, the estimation can easily be performed iteratively (as in [178, 181]).

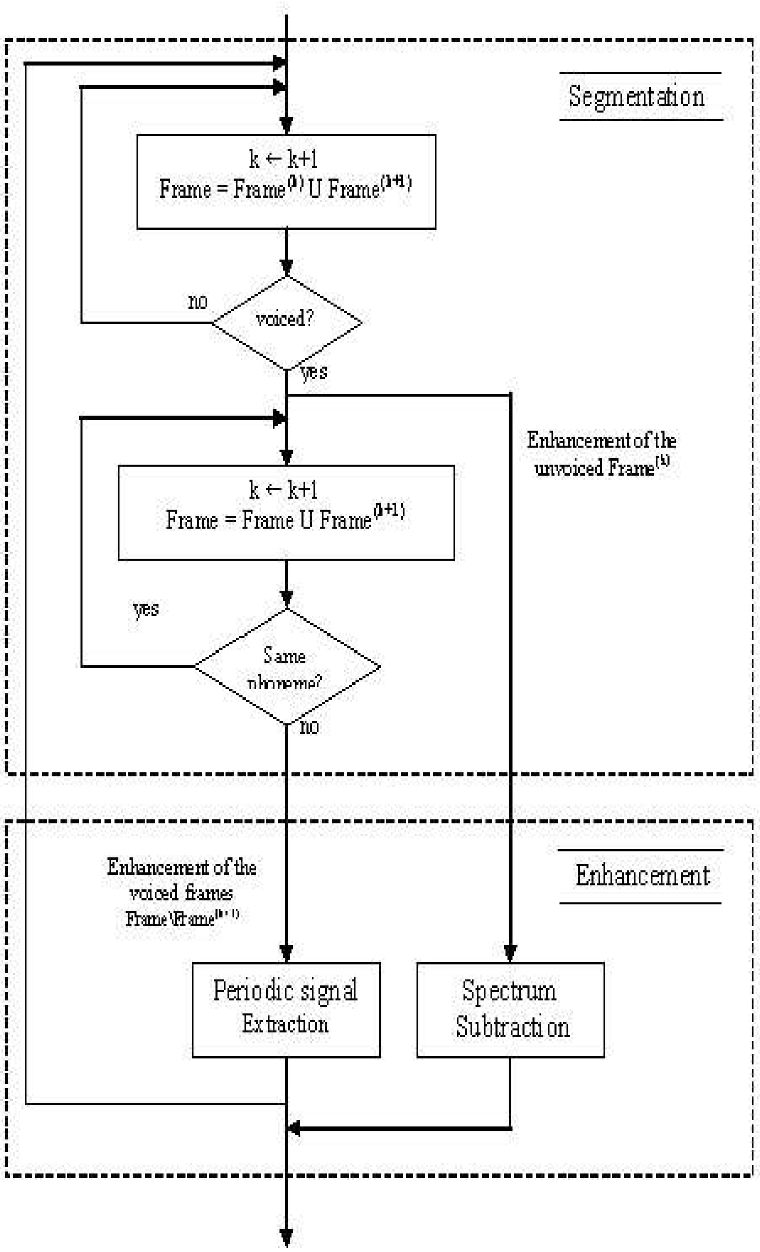
Segmentation
k ← k+1
Frame = Frame(l) U Frame(l+1)
no
voiced?
yes
Enhancement of the unvoiced Frame(k)
k ← k+1
Frame = Frame U Frame(l+1)
yes
Same phoneme?
no
Enhancement
Enhancement of the voiced frames Frame\Frame(l+1)
Periodic signal Extraction
Spectrum Subtraction

Figure 2.17: Speech Enhancement Technique.

i-b) Unvoiced speech extraction: In our preliminary experiments, the well-known spectral subtraction is employed to the unvoiced speech segments, for simplicity [128, 155]. In this conventional method, a frequency-domain Wiener filter is constructed from the speech and noise spectral estimates at each time frame, which is then used to obtain a clean speech estimate. The noisy signal power spectral density (S_{yy}) is estimated (by a Periodogram technique) using the observed signal of the current frame; whereas the estimate of the noise spectrum (S_{vv}) is updated during periods of non-speech activity. The tracking of the noise spectrum can be performed, also, on voiced frames (using the noise estimate $\widehat{v}(n) = y(n) - \widehat{s}(n)$). Finally, enhanced speech is reconstructed by Wiener filtering in the frequency domain:

$$\widehat{S}(w) = H(w)Y(w) \tag{2.37}$$

where $H(w) = \left(\frac{|\widehat{S}_{yy} - \widehat{S}_{vv}|}{\widehat{S}_{yy}}\right)^{\frac{1}{2}}$ denotes the estimated square root of the Wiener filter.

ii) Segmentation stage

The segmentation of the speech signal, i.e. classification of speech into voiced/unvoiced frames, is a crucial issue to ensure the performance of the Enhancement stage. In fact, the estimation accuracy of the quasi-periodic signal, as well as the spectrum of the noisy speech, depends on the speech frame length. On the other hand, the time resolution of these parameters is only as fine as the window length, itself. Since a speech signal is strongly non-stationary, it is not always possible to find a constant frame length giving a good tradeoff between estimation and localization accuracy.

There is a vast literature on speech segmentation with applications to speech analysis, synthesis, and coding [171, 173]. In some speech applications, the digital signal processing techniques are augmented by linguistic constraints or may be "supervised" by a human operator. However, manual phonetic segmentation is very costly and requires much time and effort. Automatic segmentation methods utilize from energy and zero crossings for silence and/or endpoint detection, to much more sophisticated spectral analysis methods for detecting changes in the speech spectrum. Each of these methods monitors one or more indicators, such as energy, number of zero crossings, pitch period, prediction error energy, or a spectral distortion measure, to detect significant changes.

Note that here the segmentation stage is not designed for recognition or classification applications. Its purpose is just to identify frames having similar spectrum characteristics (essentially spectrum envelope, and periodicity); so that they can be treated together. This motivates the choice of a distance criterion based on the energies of the extracted signal and the noise,

$$D = \max_{T} \frac{\sigma^2_{\widehat{s}_T} + \sigma^2_v}{\sigma^2_y} \tag{2.38}$$

where:
- $\widehat{s}_T$ is the quasi-periodic signal with a period T extracted as described is sections 2.3 or 2.4.
- $\sigma^2_{\widehat{s}_T}$, σ^2_v, and σ^2_y represent, respectively, the power of the extracted quasi-periodic signal, the noise and the received signal.

As we have seen in section (i-a), for a given period T, the proposed extraction algorithm approximates the projection of the noisy signal onto the subspace spanned by the set of T-periodic signals with low-pass amplitude and phase modulations. Thus, if the received signal corresponds to a unique voiced phoneme, $\exists T \ / \ \sigma^2_{\widehat{s}_T} + \sigma^2_v \approx \sigma^2_y$, then $D \approx 1$. However, if the received signal corresponds to an unvoiced phoneme ($\forall T \quad \sigma^2_{\widehat{s}_T} \approx 0$), or if it contains more than one phoneme ($\exists T_1 \neq T_2 \ / \ \sigma^2_{\widehat{s}_{T_1}} \neq 0, \ \sigma^2_{\widehat{s}_{T_2}} \neq 0$), we have $1 > D \rightarrow \frac{\sigma^2_v}{\sigma^2_y}$.
Consequently, the distance D seems to be suitable for our application.

The proposed segmentation procedure is described in figure 2.17. The main idea is to split speech signal into 10 ms frames; then make use of the distance D to group together frames belonging to the same voiced phonemes.

Experimental results

We now introduce some tests to evaluate the performance of the proposed speech enhancement scheme. The sampling rate is 8 kHz. A synthetic Gaussian white noise is added to speech signal. We first see the performance of the proposed scheme on a speech signal with relatively high SNR (SNR = 20 dB) in figure 2.18. In the figure 2.18.(b), we superpose curves of the extracted voiced signal, and the envelope of the original (noise free) signal. Obviously, the quasi-periodic model holds (with a good accuracy) for the voiced speech segments.

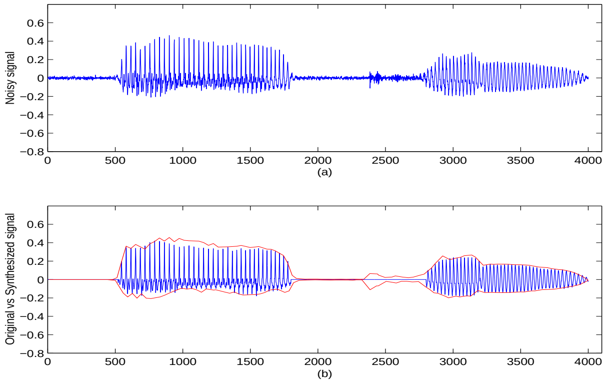

Figure 2.18: Noisy speech, extracted voiced speech, and noisefree signal envelope (SNR=20dB).

We then test the proposed scheme in a very noisy environment (SNR = 0 dB) (figure 2.19). In this second set of simulations, we treat only voiced frames (as spectral subtraction gives poor results); unvoiced frames are set to zero. Remark that in a noisy environment, the speakers have a tendency to stretch voiced phonemes (Lombard effect). We observe that the quasi-periodic characteristic is robust to the additional noise, and allows speech enhancement in a very noisy environment.

Furthermore, we consider a global measure of signal-to-noise ratio (SNR_{out}) as an objective evaluation criterion through this work

$$SNR_{out} = 10 \log \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} \left(s(n) - \widehat{s}(n)\right)^2}$$

which is consistent with previous enhancement studies [74, 155]. Figure 2.20 plots curves of the averaged output SNR (evaluated by Monte-Carlo techniques) for our proposed scheme and the classical spectral subtraction technique [128, 155].

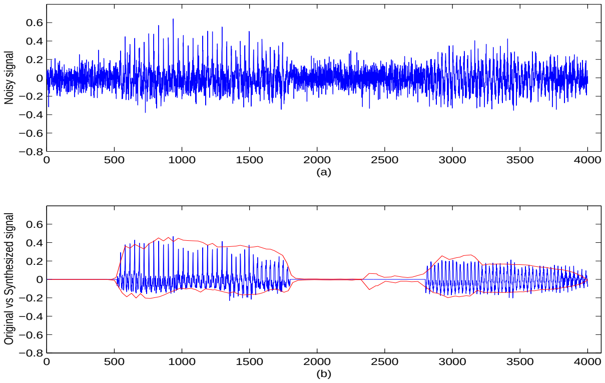

Figure 2.19: Noisy speech, extracted voiced speech, and noisefree signal envelope (SNR=0dB).

The output SNR has straightforward interpretation; and it can provide indications of the perceived audio quality in some cases [199]. Unfortunately, the output SNR shows a limited correlation with perceived speech quality. Therefore, some speech quality assessment algorithms try to include explicit models of the human auditory perception system. The ITU P.862 PESQ (Perceptual Evaluation of Speech Quality [151, 84]) is one of the most recently introduced methods, that is found implemented in many commercially available testing devices and monitoring systems [36].
Figure 2.21 plots curves of the averaged PESQ criterion (evaluated by Monte-Carlo techniques) for our proposed scheme and the classical spectral subtraction technique.

As can be observed in the previous graphs, the proposed scheme outperforms the spectral subtraction in low to high SNR regions. However, at very high SNR, the achievable output SNR of the proposed method is saturated due to approximation error in the periodicity model.
Remark that in our simulations, the noise spectrum is assumed to be known. It could be estimated during silence periods. It can be noted that the knowledge of the noise spectrum is required for spectral subtraction but not for the modulated periodic signal extraction. Nevertheless, the performance of this last technique is affected by the color of the noise. In this respect, a white noise will tend to lead to worse results than a colored noise (PC noise, car noise, human voice), especially if the spectrum of this colored noise is

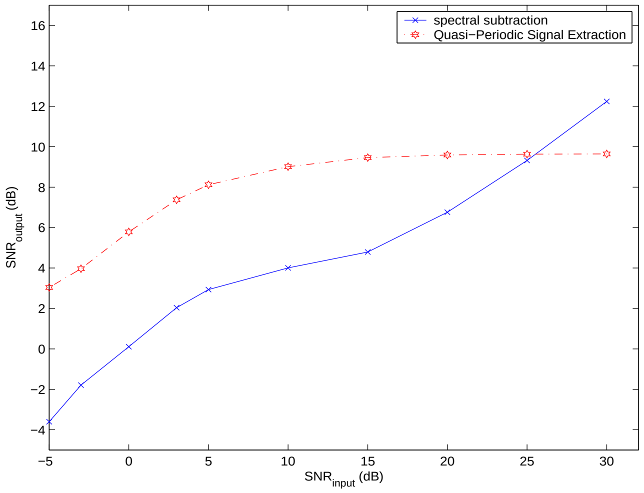

Figure 2.20: Comparison of our proposed scheme and the spectral subtraction technique for white noise corrupted speech signal.

concentrated in different frequency regions than the voiced speech.

2.7 Conclusion

In this chapter, we have investigated signal enhancement techniques exploiting the harmonic structure of the audio signal. We have modeled an audio signal as a periodic signal with (slow) global variation of amplitude (characterizing the evolution of the signal power) and phase (emphasizing the harmonic structure). The global phase variation is decomposed into a piecelinear part (interpreted in terms of global time-warping), and an excess part (assumed to have small magnitude). The bandlimited variation of global amplitude and phase gets expressed through a subsampled representation and parametrization of the corresponding signals.

Assuming additive white Gaussian noise and small time warping variation, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems.

Simulations show that the extraction technique is suitable for the analysis of isolated musical notes, and produces good auditive synthetic results. We

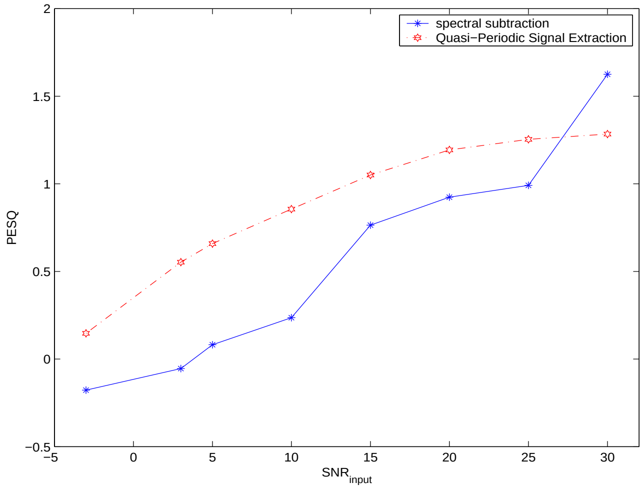

Figure 2.21: Comparison of our proposed scheme and the spectral subtraction technique for white noise corrupted speech signal.

have also considered application to the speech enhancement. The harmonic structure was exploited to identify and enhance voiced frames. Simulations show that the enhancement technique achieves quite good performance (especially in very noisy environments).

2.A Derivative filter determination

2.A.1 Problem statement

We consider to approximation of the analogical derivation of a sampled signal $\{y(n)\}_n$ at the frequency f_s .
It is common to use difference operator as an approximation of derivation. In fact, derivation is by definition

$$\acute{y}(t) = lim_{h \to 0} \frac{y(t+h) - y(t)}{h}$$

If $h = \frac{1}{f_s}$ is sufficiently small (i.e. y vary sufficiently slowly between $\frac{n-1}{f_s}$ and $\frac{n}{f_s}$), then $\acute{y}_n \approx f_s(y_n - y_{n-1})$.
This result can be interpreted in a different way. In fact, in Laplace domain, the derivation operator is p (which correspond to $j2\pi f$ in frequency domain). The difference operator can be seen as an FIR filter with transfer function $H(z) = f_s(1 - z^{-1})$. In frequency domain

$$H(f) = f_s \left(1 - e^{-j2\pi \frac{f}{f_s}}\right)$$

If y vary sufficiently slowly with time (i.e., if $y(f)$ is sufficiently low frequency), then

$$H(f) \approx f_s \left(1 - (1 - j2\pi \frac{f}{f_s})\right) \approx \; j2\pi f$$

Thus, with low frequency signals, $H(f)$ well approximates derivative operator. If the signal contains high frequencies, the previous approximation is not valid; and $H(z) = f_s(1 - z^{-1})$ can not be considered as an approximation of the derivation operator. Motivated by the previous observations, we propose to perform an approximation of the derivation operator taking into account signal spectral information.

2.A.2 Derivative filter determination

We propose to estimate the derivation operator as an FIR filter that approximates as well as possible $j2\pi f$. The problem can be formalized as follow:

$$H(z) = \sum_{n=-p}^{p} h_n z^{-n}$$
$$\min_{h_n} \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f) \left| j2\pi f_s f - H\left(e^{j2\pi f}\right) \right|^2 df$$

Where $S_{yy}(f)$ denotes the spectral density function of the signal $\{y(n)\}_n$. let us denote by $\mathbf{f} = [e^{j2\pi fp} \cdots 1 \cdots e^{-j2\pi fp}]^T$, and by $\mathbf{h} = [h_{-p} \cdots h_0 \cdots h_p]^T$. The cost function can be written as

$$\begin{aligned} C(\mathbf{h}) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f) \left| j2\pi f_s f - \mathbf{f}^T \mathbf{h} \right|^2 df \\ &= cst - 2\mathbf{p}^T \mathbf{h} + \mathbf{h}^T \mathbf{R} \mathbf{h} \end{aligned}$$

where :

- $cst = 4\pi f_s^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} f^2 S_{yy}(f) df$
- $p_k = 2\pi f_s \int_{-\frac{1}{2}}^{\frac{1}{2}} f S_{yy}(f) sin(2\pi f(p-k+1))\ df$
- $R_{kl} = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f) e^{2\pi f(k-l)}\ df = r_{yy}(k-l)$, which represents the covariance of the stationary signal $y(n)$ at the time-lag $k-l$

The optimum filter coefficients is given by:

$$\widehat{\mathbf{h}} = \mathbf{R}^{-1}\mathbf{p} \tag{2.39}$$

To know whether this extremum is a local minimum, we apply the derivation a second time to obtain

$$\frac{\partial^2 C}{\partial^2 \mathbf{h}} = 2\mathbf{R} \geq 0 \tag{2.40}$$

Hence the extremum at $H_{opt} = \mathbf{R}^{-1}\mathbf{p}$ is a local minimum, and it is furthermore the global minimum since it is the unique local extremum.
The minimum cost is given by

$$\begin{aligned} C_{min} &= cst - \mathbf{h}^T\mathbf{p} \\ &= cst - \mathbf{p}^T\mathbf{R}^{-1}\mathbf{p} \end{aligned}$$

2.B Linear Interpolation for Signal Reconstruction

Mathematical interpolation focuses on the estimation of an unknown value of a function f, defined on a regular grid N. If we restrict our consideration to a linear case, the desired solution will take the following general form:

$$f(x) = \sum_{n \in N} w(x,n) f(n) \tag{2.41}$$

where $f(x)$ is the unknown value , and $w(x,n)$ is a given linear weight function.
The linear weighting function must verify two properties:

- The interpolation of a constant function $f(n)$ remains constant (i.e., $\sum_{n \in N} w(x,n) = 1$)
- The interpolation at a given point n does not change the value $f(n)$ (i.e., $w(n,n) = 1$)

In addition, one can verify that mathematical interpolation is equivalent to *filtering* an impulse train carrying the signal sample with a continuous-time filter:

$$f(x) = \sum_{n \in \mathbb{N}} w(x-n) f(n) \tag{2.42}$$

where $w(.)$ characterizes the filter impulse response.
In fact, the nearest-neighbor interpolation can be achieved by filtering the signal using a rectangular window

$$w(t) = \begin{cases} 1 & \text{for } |t| < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{2.43}$$

The linear interpolation can, also, be performed by filtering the sampled signal with a continuous-time filter having a triangular impulse response

$$w(t) = \begin{cases} 1 - |t| & \text{for } |t| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.44}$$

We also can use smoother windows to perform interpolation, such as Hanning window (we can easily verify that the resulting interpolating weight function satisfy the two previous properties).

$$w(t) = \begin{cases} \frac{1}{2} + \frac{1}{2} cos(\pi k) & \text{for } |t| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.45}$$

The use of a smooth window (with energy concentrated essentially on the principle lobe), the interpolation error is less amplified.

Chapter 3

Underdetermined Audio Source Separation

In this chapter, we exploit the temporal and harmonic structure of audio signals to perform underdetermined source separation. We model an audio signal as a periodic signal with slow global variation of amplitude (reflecting the temporal evolution of the signal power) and frequency (limited time warping). Also voiced speech admits such a representation. The periodic signals are assumed to have distinct periodicity (sparse time-frequency mixture), and/or to arrive at a set of sensors with different amplitudes and delays (spatially distributed sources). Assuming additive white Gaussian noise, a maximum likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Simulation results reveal that the proposed approach allows extracting such signals accurately from an underdetermined mixture of several musical notes, using iterated successive interference cancelation.

3.1 Introduction

A myriad of applications requires the extraction of a set of signals which are not directly accessible [121]. Instead, this extraction must be carried out from another set of measurements (called observations) which were generated as mixtures of the original signals. Typically, the observations are obtained at the output of a set of sensors. Since usually neither the original signals (called sources), nor the mixing transformation are known, this is certainly a challenging problem of multichannel blind estimation.

One of the most typical examples is the so-called ”cocktail party” problem. In this situation, any person attending the party can hear the speech of the speaker they want to listen to, together with surrounding sounds coming from other ”competing” speakers, music, background noise, etc. Everybody has experienced how the human brain is able to separate all these incoming sound signals and to switch to the desired one. Similar results can be achieved by adequately processing the output signal of an array of microphones, as long as the signals to be extracted fulfill certain conditions[175, 29, 210]. Wireless communication is another usual application. Blind separation of sources proves also useful in biomedical application. The separation of the maternal and the fetal electrocardiograms is one of them[15]. Other applications comprise many diverse areas such as radar and sonar, speech proceeding, semiconductor manufacturing, etc. With so many practical applications, it is no wonder that the problem of the Blind Source Separation (BSS) has aroused such enormous research interest among the signal processing community since the mid-eighties.

The majority of blind separation algorithms are based on the theory of Independent Component Analysis (ICA). The idea is to estimate the inverse mixing matrix using statistical independence of source signals. However, one area of research in Blind Source Separation, the Underdetermined BSS, is relatively untouched. It refers to the case when there are less mixtures than sources (figure 3.1). The underdetermined BSS poses a challenge because the mixing matrix is not invertible and the traditional ICA methods does not work. And, contrary to most blind separation algorithms, the source extraction itself requires additional assumptions on the source statistics or structure.
Several approaches are proposed for underdetermined BSS algorithms, which

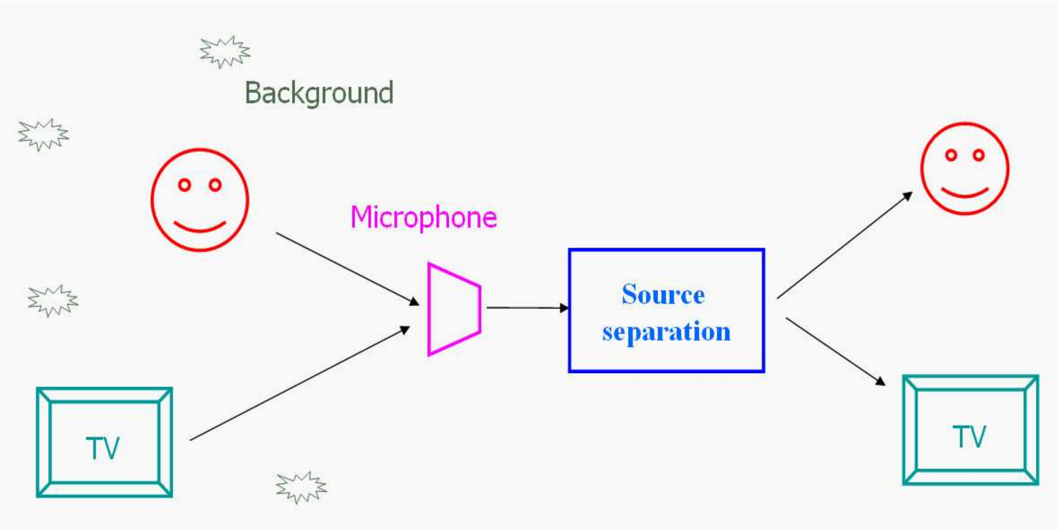

Figure 3.1: Underdetermined audio separation: problem statement.

are based on some sparse representation of the data[14, 122]. The key observation is that a good data representation often makes it possible to decompose a single underdetermined BSS problem into several (over)determined problems. In the one microphone setting, the underlying hypothesis is that at most one source is "active" in each component of the representation. However, majority of the proposed solutions performs separation, independently, on each time-frequency frame; and do not take advantage of signal correlation in different frames.
In this chapter, we assume that the received signals are a mixture of harmonic sources (as in (2.15)). We assume that at each time instant, the harmonic signals present in the mixture have distinct periods (which leads to a sparse mixture in time-frequency domain). The global amplitude and phase modulation assumption allows taking into consideration the correlation between different partials and time-domain frames. For a multi-mixture propagation, we take into account some propagation parameters such as the time of arrival and the signal attenuation. The proposed procedure combines the structural signal and propagation information using a Successive Interference Cancelation (SIC) scheme.

This chapter is organized as follows. In the section 3.2, a overview of blind source separation problem is presented. The multi-channel multi-monopath separation scheme is presented in section 3.3, as well as its SIC implementation. Finally, application to underdetermined musical source separation and comparison with classic separation method is investigated in section 3.4.

3.2 Audio source separation: a brief overview

The main goal of the present chapter is to supply an insight into the blind source separation problem and its basic foundations as well as to present a number of methods for BSS. In the first place, the BSS problem is presented with a general mathematical formulation. We will focus then respectively on the real instantaneous and convolutive linear mixture. The underdetermined case is treated in the last section.

3.2.1 Problem statement

Suppose there are N_s audio sources in a room $\mathbf{s}(n) = [s_1(n), s_2(n) \ \cdots \ s_{N_s}(n)]^T$, and M microphones capturing the auditory scene, by recording the observation signals $\mathbf{y}(n) = [y_1(n), y_2(n) \ \cdots \ y_M(n)]^T$. In the general case, the measurements can be regarded as mixtures of transformed versions of the sources, contaminated by some additive noise $\mathbf{v}(n) = [v_1(n), v_2(n) \ \cdots \ v_M(n)]^T$:

$$y_i(n) = \sum_{j=1}^{N_s} A_{ij} \left\{ s_j(n) \right\} + v_i(n), \qquad i = 1, 2, \ldots, M \tag{3.1}$$

where $A_{ij} \left\{ . \right\}$ denotes the transformation carried out on the j^{th} source contributing to the i^{th} sensor signal.

One can model the recording environment and illustrate the relation between the observed signals and the original signals [115] . A first approximation can be that each microphone captures a portion of each source. Even though this seems to be a rather simplified model, if we refer to studio recording, where audio signals are mixed by a mixing desk, the mixed signals can be modeled as summed portions of the original source, i.e. instantaneous mixtures of sources. Therefore,

$$\mathbf{y}(n) = \mathbf{A}_(n) + \mathbf{v}(n) \tag{3.2}$$

where $\mathbf{A}$ is an $M \times N_s$ matrix, but not necessary to be full rank .

Unfortunately, the instantaneous mixtures are rather incomplete in the case of sources recorded in acoustic room environment. In fact, due to different propagation paths between the sources and sensors, relative delays between the source signals occur. The effects must be modeled with the matrix of filters instead of the matrix of scalars. Assume the case of one sound source and a microphone in a room. Previous showed that the signal captured by the microphone is well represented by a convolution of the source signal with a FIR filter, modeling the room acoustics between the source and the sensor. In the case of many sensors, the signal at each sensor can be modeled by the following equation:

$$y_i(n) = \sum_{j=1}^{M} \sum_{\tau=1}^{\mathcal{T}} A_{ij}(\tau) s_j(n-\tau) + v_i(n), \qquad i = 1, 2, \ldots, M \tag{3.3}$$

where $\mathcal{T}$ denotes the maximum delay in terms of discrete points. Observations represent, thus, the superposition of N sources (distorted with N filters of maximum length $\mathcal{T}$). These mixtures are referred to as convolutive mixtures.

3.2.2 Instantaneous mixtures

Problem formulation

The problem of blind source separation was traditionally approached by observing instantaneous mixtures of sources. As we have seen, the observation vector $\mathbf{y}(n)$ can be expressed as:

$$\begin{aligned} \mathbf{y}(n) &= \mathbf{A}\mathbf{s}(n) + \mathbf{v}(n) \\ &= \mathbf{x}(n) + \mathbf{v}(n) \end{aligned}$$

where $\mathbf{A}$ is an unknown matrix called the mixing matrix.
The objective is to recover the original signals $\{s_i(n)\}_{i=1:N_s}$ given only the vectors $\mathbf{y}(n)$ (we denote $\mathbf{u} = \mathbf{B}\mathbf{y}$ the recovered signal). In this section, we assume that $\mathbf{A}$ is a full column rank and $M \geq N$ (there are fewer sources than sensors).

Blind source separation consists in identifying $\mathbf{A}$ and/or retrieving the source signals without resorting to any prior information about the mixing matrix $\mathbf{A}$; it exploits only the information carried by the received signals themselves, hence, the term *blind.* Of course, the lack of information on the structure of $\mathbf{A}$ must be compensated by some additional assumptions on source signals. The blind-identification of source separation techniques rely on the mutual independence of the source signals received at a given time. We note that the assumption of independence between sources is a statistically strong hypothesis but very plausible in practice for physically separated emitters. We assume also that:

- There is at the most one Gaussian source.
- The columns of $\mathbf{A}$ are linearly independent but otherwise arbitrary.
- The additive noise is spatially white $\mathbf{R}_v = \sigma_v^2 \mathbf{I}_M$, with unknown variance σ_v^2.
- The additive noise is independent of the sources.

Blind identifiability

In the blind context, a full identification of the mixing matrix is impossible because the exchange of a fixed scalar factor between a given source signal and the corresponding column does not affect the observations, as is shown by the following relation:

$$\mathbf{y}(n) = \mathbf{A}.\mathbf{s}(n) + \mathbf{v}(n) = \sum_{j=1}^{M} \frac{\mathbf{a}_j}{\alpha_j} \, \alpha_j s_j(n) + \mathbf{v}(n)$$

where $\{\alpha_j\}_j$ are arbitrary complex factors, and $\mathbf{a}_j$ denotes the j^{th} column of $\mathbf{A}$.

Also, one can note that the numbering of the sources is a pure notational convention but otherwise immaterial. These simple remarks show that without additional a priori information, the matrix $\mathbf{A}$ (and the sources) can be at best identified up to permutation and scaling factors [29, 30]. Thus, we can assume, without any loss of generality, that the source signals have unit

variance. This normalization convention does not affect the performance results of BSS algorithms.
Since the sources are assumed to be uncorrelated, we have:

$$\mathbf{R}_s(0) = E\left\{\mathbf{s}(n)\mathbf{s}(n)^T\right\} = \mathbf{I}_{N_s}$$

A square matrix $\mathbf{C}$ is said to be nonmixing if it has one and only one nonzero entry in each row and each column[30]. If $\mathbf{C}$ is nonmixing then $\widehat{\mathbf{s}}(n) = \mathbf{C}\mathbf{s}(n)$ is a copy of $\mathbf{s}(n)$, i.e., its entries are identical to those of $\mathbf{s}(n)$ up to permutations and changes of scales. Source separation is achieved if such a copy is obtained.
When the distribution of $\mathbf{s}(n)$ is unknown, we cannot expect to do better than signal copy. But the situation is a bit different if some prior information about the distribution of sources is available. For example, if the sources have distinct distributions, a possible permutation can be detected; or if the scale of a given source is known, the amplitude of the corresponding column of $\mathbf{A}$ can be estimated, etc.

Second-order identification

We consider exploiting second-order information to separate signals. The statistical independence leads to imposing decorrelation constraints. This is done by whitening the signal part $\mathbf{x}(n)$ of the observation; and can be achieved by applying to $\mathbf{x}(n)$ a whitening matrix $\mathbf{W}$, i.e., an $M \times N_s$ matrix such that $\mathbf{W}\mathbf{x}(n)$ is spatially white. The whiteness condition is :

$$E\left\{\mathbf{W}\mathbf{x}(n)\mathbf{x}(n)^T\mathbf{W}^T\right\} = \mathbf{W}\mathbf{R}_x(0)\mathbf{W}^T = \mathbf{W}\mathbf{A}\mathbf{A}^T\mathbf{W}^T = \mathbf{I}_{N_s} \tag{3.4}$$

where $\mathbf{R}_x(0) = E\left\{\mathbf{x}(n)\mathbf{x}(n)^T\right\}$ is the covariance matrix of $\mathbf{x}(n)$ at the time-lag $\tau = 0$. Equation (3.4) shows that if $\mathbf{W}$ is a whitening matrix, then $\mathbf{WA}$ is a $N_s \times N_s$ unitary matrix. It follows that for any whitening matrix $\mathbf{W}$, there exists a $N_s \times N_s$ unitary matrix such that $\mathbf{WA} = \mathbf{U}$. As a consequence, matrix $\mathbf{A}$ can be factored as:

$$\mathbf{A} = \mathbf{W}^\sharp\mathbf{U}$$

where $(.)^\sharp$ denotes the Moore-Penrose pseudoinverse[21].
We notice that the whitening procedure reduces the determination of the

mixture matrix $\mathbf{A}$ to that of a $N_s \times N_s$ unitary matrix $\mathbf{U}$. The whitened process $\mathbf{z}(n) = \mathbf{Ws}(n)$ still obeys a linear model:

$$\mathbf{z}(n) = \mathbf{Wy}(n) = \mathbf{Us}(n) + \mathbf{Wv}(n)$$

We can write then:

$$\mathbf{W}\left(\mathbf{R}_y(0) - \sigma_v^2 \mathbf{I}_M\right)\mathbf{W}^T = \mathbf{I}_{N_s}$$

Therefore, the whitening matrix can be determined from the array output covariance, provided the noise covariance matrix is known or can be estimated. A whitening matrix may also be determined from a linear combination of a set of covariance matrices taken at nonzero time lags, as suggested in[174]. In any case, finding a whitening matrix still leaves undetermined a unitary factor in $\mathbf{A}$. On the other hand, we notice that all the information contained in the covariance is "exhausted" after the whitening. In fact, changing the matrix $\mathbf{U}$ to any unitary matrix leaves the covariance of $\mathbf{z}$ unchanged. This "missing factor" can be determined using other procedures.

Unitary factor determination

As we have seen, the separation process may be decomposed in two steps (see fig. 3.2):

- the first step of the separation is the search for a whitening matrix $\mathbf{W}$ that transforms the original set of M signal into a reduced base of N_s orthogonal and normalized signals (named the whitened data set).

- The second step is the identification of the rotation $\mathbf{U}$.

The unknown rotation can be found either by exploiting the time dependence structure of signals or by minimizing some cost function. In the following section, we give a brief overview of those approaches.

i) Blind source separation using second-order statistics
This approach is based on matrix correlation functions, defined, for any $x(t)$, by

$$\mathbf{R}_y(\tau) = E[\mathbf{y}(n)\mathbf{y}^T(n-\tau)]$$

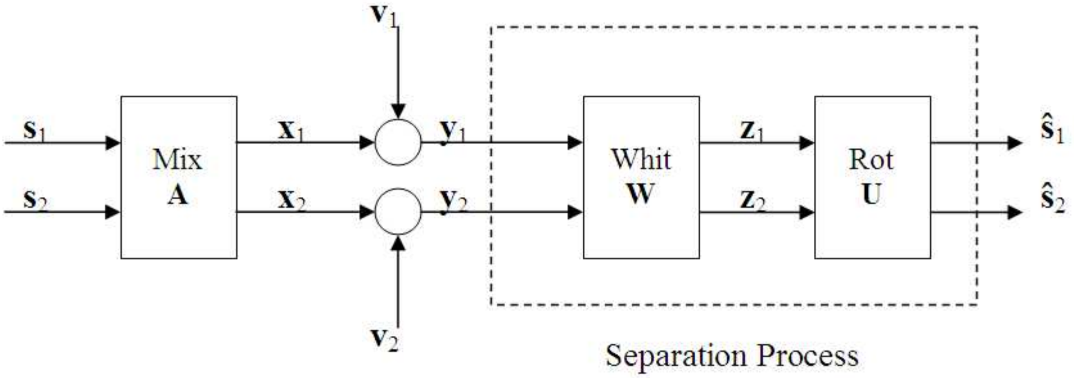

Figure 3.2: Bloc diagram for BSS for N=2 sources and observations.

The rotation can be found from any covariance matrix of the whitened observations at non-zero lag. In fact, as the additive noise is white, we have:

$$\mathbf{R}_z(\tau) = \mathbf{U}\mathbf{R}_s(\tau)\,\mathbf{U}^T \qquad \forall\tau \neq 0 \tag{3.5}$$

As $\mathbf{R}_z(\tau)$ is a covariance matrix of $\mathbf{z}(n)$ at time-lag τ, it is Hermitian and can be diagonalised by a unitary matrix $\mathbf{V}_\tau$. Moreover, if the eigenvalues of $\mathbf{R}_z(\tau)$ are distinct, the matrix $\mathbf{V}_\tau$ will be essentially unique, i.e. up to scaling and permutation of columns[21]. So,

$$\mathbf{V}_\tau = \mathbf{PDU}$$

Where $\mathbf{P}$ is a permutation matrix,and $\mathbf{D}$ is a non degenerated diagonal matrix.

In theory, one covariance matrix at non-zero lag is sufficient to estimate the rotation. In practice, however, it is useful to use a set of matrices as this would enhance the statistical efficiency of the algorithm and prevent an unfortunate choice of lags [124]. Generally, only a small number of these matrices is computed (in [124], the authors propose using 4 non-zero time-lag covariance matrices). $\mathbf{U}$ is estimated as the matrix that jointly diagonalizes the set of $\mathbf{R}_z(\tau)$ matrices.

The advantage of this approach is that the second part of the process relies solely on the manipulation of $\mathbf{R}_z(\tau \neq 0)$ matrices, that are not biased by the noise as stated by (3.5). This approach has the advantage of being a simple method as it only requires second-order information. However, from (3.5), it should be clear that the sources must have some temporal dependencies [94]. If the sources were white, $\mathbf{R}_s$ (and consequently $\mathbf{R}_z$) would be null matrices and no additional information would be available to estimate the rotation.

ii) Blind source separation using contrast functions

Source separation can be obtained by optimizing a contrast function. Contrast functions for source separation are generically denoted $\Phi[\mathbf{u}]$. They are a scalar measure of some distributional property of the output $\mathbf{u} = \mathbf{B}\mathbf{y}$. They must be designed in such a way that source separation is achieved when they reach their minimum value. In other words, a contrast function should satisfy:

- $\Phi[\mathbf{Cs}] \geq \Phi[\mathbf{s}] \quad \forall C$
- $\Phi[\mathbf{Cs}] = \Phi[\mathbf{s}]$ if $\mathbf{Cs}$ is a copy of $\mathbf{s}$

Contrast functions are based on many measures, such as entropy, mutual independence, high-order decorrelations, etc[177].

iii) Maximizing entropy (ME)

The idea of ME originated from the neural networks [208]. Let us transform $u_i = \sum_j B_{ij} y_j$ by a sigmoid function g_i to $\breve{u}_i = g_i(u_i)$; which is regarded as the output from an analog neuron. Let $\breve{\mathbf{u}} = [g_1(u_1), \ldots, g_N(u_N)]$ be the transformed output vector by sigmoid functions g_i $, i = 1, \ldots, N$. It is expected that the entropy of the output $\breve{\mathbf{u}}$ is maximized when the component $\breve{u}_i$ of $\breve{\mathbf{u}}$ are mutually independent.

iv) Minimizing mutual information (MMI)

The basic idea of MMI is to choose $\mathbf{U}$ that minimizes the dependence among the component of $\mathbf{u}$. The dependence is measured by the Kullback-Leibler divergence:

$$\Phi_{MMI}(\mathbf{u}) = \int_u p(\mathbf{u}) \ \log\left(\frac{p(\mathbf{u})}{\prod p(u_i)}\right)$$

where $p(\mathbf{u})$ and $p(u_i)$ denote respectively the probability and the marginal density functions of the random vector $\mathbf{u}(n)$. This function measures the distance between a distribution and the closest distribution with independent entries[30]. Under the whiteness constraint, minimizing the mutual information between the entries of $\mathbf{u}$ is equivalent to minimizing the sum of the entropies of $\mathbf{u}$. The contrast function becomes then:

$$\Phi^{\circ}_{MMI}\left(\mathbf{u}\right) = \sum_{i} H\left[u_i\right]$$

Thus, minimizing the mutual information requires entropy of the output components, which are not available, but can be estimated[114].

v) High-order approximation

High-order statistics can be used to define contrast functions which are simple approximations of those delivered from ME and MMI approaches. Several contrast functions have been proposed in the literature, in which high-order information is expressed mainly using cumulants[30, 208, 35].

3.2.3 Convolutive mixtures

Problem formulation

In the previous section, we have seen that there are many ICA methods that can perform separation of linearly instantaneous mixtures. However, if we try to apply these techniques on observation signals acquired from a microphone in a real environment, they will fail to separate audio sources. This is mainly because the previous model does not account for the room acoustics. In fact, the acoustic environment imposes a different impulse response between each source and microphone pair. Moreover, microphones may have different characteristics, or at least their frequency responses may differ for sources in different directions. This scenario can be described as a finite response (FIR) convolutive mixture:

$$\mathbf{y}\left(n\right) = \sum_{\tau=1}^{\mathcal{T}} \mathbf{A}\left(\tau\right)\mathbf{s}\left(n-\tau\right) + \underbrace{\sum_{\tau_1=1}^{\mathcal{T}_1} \mathbf{A}_1\left(\tau_1\right) v\left(n-\tau_1\right)}_{v(n)} \tag{3.6}$$

where $\mathcal{T}$ and $\mathcal{T}_1$ denote characterizing respectively to the source and noise propagation, and $\mathbf{A}(\tau)$ characterizes the impulse response between the source and the microphone positions.

This situation is considerably more complicated than in the previous section as we have now a matrix of filter rather than a matrix of scalars mixing. And even once the channel has been identified, inverting it is a more difficult task as the inverse can be an instable infinite impulse response (IIR) filter. Alternatively, one may formulate an FIR inverse $\mathbf{W}(\tau)$:

$$\mathbf{u}\left(n\right)=\sum_{\tau=1}^{Q}\mathbf{W}\left(\tau\right)\mathbf{y}\left(n-\tau\right)$$

where Q denote the length of the FIR inverse filter $\mathbf{W}$. $\mathbf{W}(\tau)$ should be estimated so as the output of the separation bloc $\mathbf{u}\left((n)=\left[u_1\left(n\right)\right)\cdots u_N\left(n\right)\right)]$ are statistically independent.

Now, there are two avenues to take. In the first one, every thing including the actual separation could be done in frequency domain. The second avenue is that actual separation is not done in the frequency domain but only one or some aspects of the separation algorithm.

Frequency domain methods

If we look at equation (3.6), one can rewrite it as follows using convolution:

$$\mathbf{y}\left(n\right)=\mathbf{A}*\mathbf{s}\left(n\right)+\mathbf{v}\left(n\right) \tag{3.7}$$

By applying z-transform on equation (3.7), we will have:

$$Y(z)=A(z)\ S(z)+V(z)$$

Where $A\left(z\right)=\sum_{\tau}\mathbf{A}\left(\tau\right)z^{-\tau}$ represents the matrix of z-transforms of the FIR filter $\mathbf{A}\left(\tau\right)$; $Y(z),S(z)$ and $V(z)$ denote respectively the z-transforms of $\mathbf{y}\left(n\right),\mathbf{s}\left(n\right)$ and $\mathbf{v}\left(n\right)$.

For practical purposes, we have to restrict ourselves to a limited number of sampling points of z. Naturally, we will take N equidistant samples on the unit circle in such a way that we can use the discrete Fourier transform (DFT). For periodic signal, the DFT allows us to express circular convolutions as products. However, in (3.7) we assume linear convolution[136]. A linear convolution can be approximated by a circular convolution if $\mathcal{T} << N$ (N denotes the frame size). Then we can write approximately

$$Y(f) \approx A(f) \cdot S(f) + V(f) \quad for \;\; \mathcal{T} << N_f \tag{3.8}$$

Thus, we can easily see that by using a Fourier transform, we have transformed a convolutional problem into N linear problems. In other words, the whole separation problem is broken into N instantaneous source separation problems. Hence, we can use the established theory behind instantaneous mixtures separation and solve this problem.

Another advantage of the frequency domain methods arises from the statistic properties of audio signals. In fact, if we examine the statistical properties of an audio signal over shorter quasi-stationary periods, the signal is better modeled as super-Gaussian in the frequency domain than in the time domain [40, 115]. And since performance of ICA methods depends on the non-gaussianity of sources, moving to the frequency domain provides a better achievable performance.

However, this case is not as simple as the separation of instantaneous mixtures. This is due to the following reasons:

- This problem involves complex valued signals, which is quite different compared to the real number case. For example, the common sigmoid functions do not have the same properties, as in the real case, where are applied to complex variables. Therefore, an important point is to choose a proper sigmoid function that fulfills all complex activation function requirement and performs complex domain separation.

- The ICA algorithms shown in the previous section and used to separate the frequency bins are invariant to scaling and permutation. The scaling invariance means that the scaling of every frequency bin can be different, which will result in spectral deformation of the original sounds.

- The permutation invariance is a more difficult problem. In fact, the algorithm may produce different permutations of separated sources along the frequency axis; and therefore sources remain mixed.

The distortion created by last two factors can be reduced by constraining the nonmixing filter update [168, 169] and/or order [137, 146], tacking into account the cross-frequency correlation between the separated signals [148], and exploiting the spatial information [2, 83, 138, 156, 157].

Time domain methods

Moving to the frequency domain is not the only way to approach convolved mixtures. There are many techniques that work exclusively in time domain or both in frequency and time domain. In fact, the first efforts on the separation of convolved mixtures were made in time domain. The first robust solutions to this problem are introduced by Torkolla [176]. The proposed scheme estimates the nonmixing filter coefficients by using an information maximization approach. The optimization leads to a gradient algorithm similar to the one proposed by Bell and Sejnowski for the separation of instantaneous mixtures. Sattar et al [158] notice that the Torkkola's algorithm works only if a stable causal inverse of direct channel filter ($\{A_{ii}(z)\}_{i=1:N_s}$) exist. However, this is not always guaranteed in audio signal separation problems. So, they propose to adapt the Torkkola method by using a non causal filter (even if $A_{ii}(z)$ does not have a stable causal inverse, it still has a stable non causal inverse). The major drawback of the scheme is its computational complexity. In fact, given that acoustical channels need to have a considerable length in real propagation environment, the algorithm seems to be very expensive.

Different approaches perform source separation by exploiting some of the source signal properties. In the literature, the demixing matrix $W(q)$ is blindly estimated utilizing one of the following properties:

- Non-whiteness : exploited by simultaneous minimization of output correlation matrices over multiple time-lags.
- Non-stationarity : exploited by simultaneous minimization of output correlation matrices at different time-instants.
- Non-Gaussianity : exploited using higher order statistics for independent component analysis.

The basic idea is that mixtures are typically more stationary, Gaussian, and whiter than sources; and by emphasizing one or more properties, sources can be recovered. Although it is commonly believed that each one of these properties is sufficient for separation, it has demonstrated that the combination of these criteria can lead to improved performance[117, 24]. In [110], the authors show that combining non-stationarity and non-whiteness leads to a significant improvement in performance, and exhibits a quasi-performance (its behavior is almost independent of the mixing matrix). A generic approach exploiting simultaneously the three properties was introduced in [24], called TRINICON (TRIple-N ICA for CONvolutive mixtures). It was shown that the proposed approach has links to a variety of popular algorithms, and several novel approaches [25, 26, 27].

3.2.4 Underdetermined mixtures

Blind source separation is a problem that arises when one or several sensor(s) record data to which can contribute several generating physical processes. It consists in recovering N_s unknown sources from M instantaneous mixtures. Classically, the idea is to estimate the inverse mixing matrix using statistical independence of source signals. However, one area of research in blind source separation, the Underdetermined BSS, is relatively unexplored. It refers to the case when there are less mixtures than sources. The underdetermined BSS poses a challenge because the mixing matrix is not invertible and the traditional ICA methods do not work. And, contrary to most blind separation algorithms, the source extraction itself requires additional assumptions on the source statistics or structure.

Several approaches are proposed for underdetermined BSS algorithms, which are based on some sparse representation of the data. The key observation is that a good data representation often makes it possible to decompose a single underdetermined BSS problem into several (over)determined problems. In the one microphone setting, the underlying hypothesis is that at most one source is "active" in each component of the representation.
On the other hand, audio sources have been shown to have sparse decomposition in a variety of time-frequency dictionaries. Sparse decomposition is typically performed using a short-time Fourier transform (STFT), wavelet transform (WT), or matching pursuit (MP).

Sparsity based separation methods

Sparsity assumption can be sufficient to solve the separation problem, without any additional information. Aissa-El-Bey et al. consider instantaneous mixtures. The proposed scheme minimizes a contrast function based on an l_p norm. The algorithm makes solution as sparse as possible (the norm l_p is a sparsity measure). Simulations show that, applied to overdetermined mixtures, the scheme provides good performance compared to other separation techniques [3]. Despite there is no-constraints preventing applying the scheme to underdetermined mixtures, the approach is not yet tested in such configuration.

Olson and Hansen exploits PARAFAC formulation to solve the convolutive separation problem [125]. First, the mixing matrices $\{\mathbf{A}(\tau)\}_\tau$ are estimated through k-means clustering (assuming source sparseness). Then, the sources are estimated by a Maximum A Posteriori (MAP) approach (as opposed to the usual binary masking reconstruction). Using the framework of PARAFAC model, one can show that the problem is identifiable if

$$\frac{1}{2}M\left(M+1\right) \geq N_s \tag{3.9}$$

Hence, the algorithm is convenient for underdetermined BSS only if the number of source and microphones satisfy the previous relation.

Masking based separation methods

Masking based source separation methods relays on the source sparseness assumption . If signals are sufficiently sparse, i.e. most of the samples of each signal are almost zero, we can assume that the source rarely overlaps. Then, each source signal can be extracted by applying a given mask to the observed mixture.

The design of the mask is a crucial issue, and generally depends on prior information on signals and/or propagation environment:

- In the case of a multi-microphone reception, if we assume a mono-path propagation environment (reverberation is negligible), direction of arrival can be exploited for the mask design. The direction of arrival are typically estimated based on observation vector clustering [14, 89, 118, 12].

- If the different signals can be decomposed into distinct dictionaries (harmonic vs transient, musical notes with different pitches...), the masking can be performed taking into account this structural diversity [152].

Binary masks leads to a very intuitive scheme: separation is performed by solving a simple classification problem. However, the use of binary masks leads to too much discontinuous zero-padding to extract signal, and therefore, they tend to contain loud musical noise, which is undesirable for audio applications. Several approaches are proposed to solve the problem by designing smooth masks [13], or combining masking and beamforming techniques [32].

Signal structure based separation methods

Additionally to the sparsity assumption, if prior knowledge about the sources is available, it can be exploited to increase the source separation accuracy. In the literature, some prior information was considered such as spectral, statistical, and temporal structures of mixed sources.

In [17], the authors propose a separation scheme where the sources are modelled AutoRegressive processes. It has been shown that one can reconstruct the processes and estimate the sources from their degenerate mixture using only second order statistics.

A second point of view tries to take advantage of the statistical prior on the audio signal. The main idea is to exploit statistical structure of the sound sources by learning either the signal structure [18], or the masking functions [152]. Simulations show a slide improvement comparing to the binary masking techniques.

The time structure of sound sources was also exploited to increase the separation performance. The idea is to learn a priori sets of basis filter (in time domain) that encode the source in an efficient manner (generalized exponential [88], damped sinusoid [64, 4], harmonic atoms [69]).

3.3 Multi-channel mono-path periodic signal extraction

As an approximation of the propagation environment, we use the delay-mixing model. In this model, only direct path signal components are considered. Signal components from a source arrives at a set of microphones with a given relative attenuations and a fractional delays between the time of arrivals at the different microphones. By fractional delays, we mean that delays between the received signal are not generally integer multiples of the sampling period. The signal attenuation and delay depend on the position of the source with respect to the microphone array orientation and geometry. Under the previous propagation assumptions, without any loss of generality, observations can be written as:

$$\begin{aligned} y_1(n) &= \sum_{i=1}^{N_s} s_i(n) + v_1(n) \\ y_k(n) &= \sum_{i=1}^{N_s} \beta_{ki} s_i(n - \tau_{ki}) + v_k(n) \quad k = 2 : M \end{aligned}$$

where $\{s_i(n)\}_{i=1:N_s}$ represent N_s distinct audio source signals following (2.15)(with distinct periodicities); $\{v_k(n)\}_{k=1:M}$ a spatially and temporally white gaussian noise signals; β_{ki} the relative attenuation of the i^{th} source at the k^{th} sensor; and τ_{ki} the propagation delay (function of the direction of arrival ϕ_i, and the microphone array geometry (that we suppose fix but unknown)).

As in [178], the time delay operation can be expressed using an interpolation matrix $\mathbf{H}_\tau$ (as it can be interpreted as a particular time warping):

$$\mathbf{s}_{i,\tau} = \mathbf{H}_\tau \mathbf{s}_i$$

where $\mathbf{s}_i = [s_i(1) \cdots s_i(N)]^T$, $\mathbf{s}_{i,tau} = [s_i(1-\tau) \cdots s_i(N-\tau)]^T$, and H_τ is an $N \times N$ toeplitz, band matrix characterizing the time delay operation.

Thus, the total observation vector can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{v} \tag{3.10}$$

where
- $\mathbf{y} = [\mathbf{y}_1^T \cdots \mathbf{y}_M^T]^T$, is a $MN \times 1$ vector representing the observation vector
- $\mathbf{s} = [\mathbf{s}_1^T \cdots \mathbf{s}_{N_s}^T]^T$, is a $NN_s \times 1$ vector representing the signals of interest.
- $\mathbf{v} = [\mathbf{v}_1^T \cdots \mathbf{v}_M^T]^T$, is a $MN \times 1$ vector denoting the noise vector
- $\mathbf{H} = \begin{bmatrix} I_N & \cdots & \mathbf{I}_N \\ \beta_{2,1}\mathbf{H}_{\tau_{2,1}} & \cdots & \beta_{2,L}\mathbf{H}_{\tau_{2,N_s}} \\ \vdots & & \vdots \\ \beta_{M,1}\mathbf{H}_{\tau_{M,1}} & \cdots & \beta_{M,N_s}\mathbf{H}_{\tau_{M,N_s}} \end{bmatrix}$ is an $NM \times NN_s$ interpolation matrix characterizing the propagation environment.

The previous model is linear in $\mathbf{H}$ and $\mathbf{s}$ (separately); $\mathbf{H}$ and $\mathbf{s}$ being parameterized nonlinearly. Trying to estimate all factors jointly is a difficult nonlinear problem. Indeed, as the noise is assumed to be a white Gaussian signal, the ML approach leads to the following least-squares problem:

$$\min_{H,s} \|\mathbf{y} - \mathbf{Hs}\|^2 \tag{3.11}$$

Once again, such optimization can easily be performed iteratively though (figure 3.3).

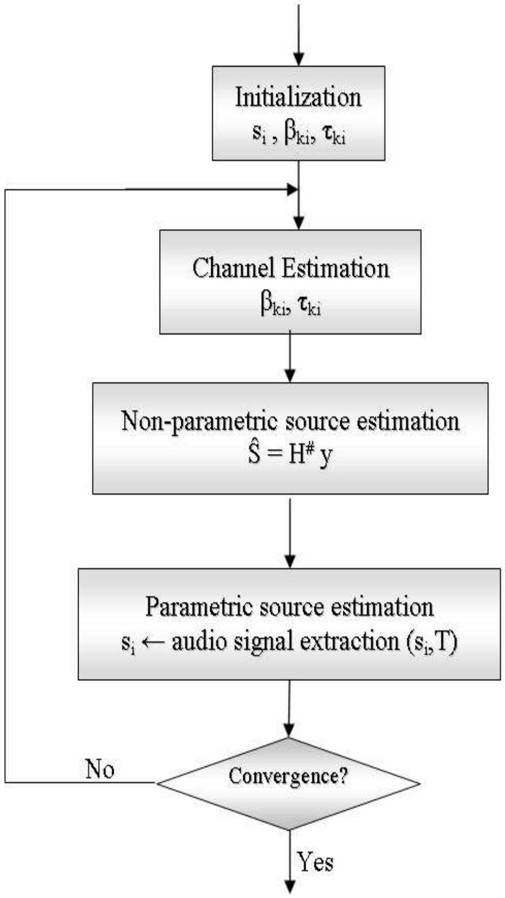
Initialization
s_i , β_{ki}, τ_{ki}
Channel Estimation
β_{ki}, τ_{ki}
Non-parametric source estimation
$\hat{S} = H^{\#}\,y$
Parametric source estimation
$s_i \leftarrow$ audio signal extraction (s_i,T)
No
Convergence?
Yes

Figure 3.3: Multi-channel mono-path separation scheme.

Channel estimation

Under the current estimate of the source signal $\widehat{S}$, the Channel coefficients are optimized using

$$\min_{\tau_{ki},\beta_{ki}} \|\mathbf{y} - \mathbf{H}\widehat{\mathbf{s}}\| \tag{3.12}$$

On the other hand,

$$\|\mathbf{y} - \mathbf{H}\widehat{\mathbf{s}}\|^2 = \sum_{k=1}^{M} \left\| \mathbf{y}_k - \sum_{i=1}^{N_s} \beta_{ki} \mathbf{H}_{\tau_{ki}} \widehat{\mathbf{s}}_i \right\|^2$$

and,

$$\begin{aligned} \left\| \mathbf{y}_k - \sum_{i=1}^{N_s} \beta_{ki} \mathbf{H}_{\tau_{ki}} \widehat{\mathbf{s}}_i \right\|^2 &= \|\mathbf{y}_k\|^2 + \sum_{i=1}^{N_s} \beta_{ki}^2 \|\widehat{\mathbf{s}}_{\tau_{ki}}\|^2 \\ &- \sum_{i=1}^{N_s} \beta_{ki} \widehat{\mathbf{R}}(\mathbf{y}_k, \mathbf{s}_{\tau_{ki}}) - \sum_{i \neq j} \beta_{ki}\beta_{kj} \widehat{\mathbf{R}}(\mathbf{s}_{\tau_{ki}}, \mathbf{s}_{\tau_{kj}}) \end{aligned}$$

where $\widehat{\mathbf{s}}_{\tau_{ki}} = \mathbf{H}_{\tau_{ki}} \widehat{\mathbf{s}}_i$ denotes the estimate of the i^{th} source delayed by τ_{ki}; and $\widehat{R}(x, y) = \frac{1}{N} \sum_{j=1}^{N} x(j)y(j)$ represents the estimate of the correlation between signals x and y.

Note that the quantities $\widehat{R}\left(s_{\tau_{ki}}, s_{\tau_{kj}}\right) \quad i \neq j$ can be neglected, as source signals are assumed to be independent, and have distinct periodicities. Then, the optimization problem in (3.12) is separable; and can be solved, independently, for each channel parameter.

The optimization over a given time-lag τ_{ki} can be interpreted in terms of maximizing the correlation $\widehat{R}\left(y_k, s_{\tau_{kj}}\right)$ between the observed signal on the sensor and the i^{th} source signal delayed by τ_{ki} .

$$\tau_{ki} = \arg\max_{\tau_{ki}} \widehat{R}\left(y_k, s_{\tau_{kj}}\right) \tag{3.13}$$

Ones the different time lags are estimated, the optimal attenuation coefficients are computed using:

$$\widehat{\beta}_{ki} = \frac{\widehat{R}(y_k, s_{\tau_{ki}})}{\left\|\widehat{S}_{\tau_{ki}}\right\|^2} \tag{3.14}$$

Source signal estimation

If we assume that the channel parameters known, the ML source estimation is given by:

$$\widehat{\widehat{\mathbf{s}}} = \mathbf{H}^{\#}\mathbf{y}$$

where $\mathbf{H}^{\#} = \left(\mathbf{H}^T\mathbf{H}\right)\mathbf{H}^T$ denotes the pseudoinverse of $\mathbf{H}$. This leads to an optimal beamforming processing.

Beamforming exploits the spatial information to focus in the source direction and decrease the interference due to other sources. This interference is further suppressed by exploiting the spectral structure of the mixtures (time-frequency sparseness of the sources). In fact, each audio source signal is assumed to have harmonic structure(as in (2.15)). Thus, it can be written as

$$\widehat{\widehat{\mathbf{s}}}_i \;=\; \widehat{\mathbf{A}}_i\widehat{\mathbf{F}}_i\widehat{\boldsymbol{\theta}}_i + \mathbf{v}_i \;=\; \widehat{\mathbf{s}}_i + \mathbf{v}_i \qquad i = 1 : N_s$$

where $\widehat{\mathbf{A}}_i$, $\widehat{\mathbf{F}}_i$, and $\widehat{\boldsymbol{\theta}}_i$ are estimated in an iterative (cyclic) fashion (as in previous chapter) from $\widehat{\widehat{\mathbf{s}}}_i$.

3.3.1 An ISIC Implementation for the multi-channel mono-path periodic signal separation scheme

In previous, we have proposed an audio separation scheme tacking into account simultaneously the source signal structure and the propagation environment model. The inherent complexity, however, is cubic on MN (as the technique requires the inversion of the non Toeplitz matrix $\mathbf{H}$). For practical implementation, Iterated Successive Interference Cancelation (ISIC) approach can be used to implement the previous technique.

Iterated successive interference cancelation is a nonlinear parameters estimation scheme in which parameters are estimated successively. The approach successively cancels concurrent parameters using their current estimate. The ISIC audio separation algorithm appears in the table 3.1.

Note that the non parametric source estimation (computed using a simple matched filter) can be interpreted as a delay and sum beamformer. It constitutes the second interference cancelation stage (the first is performed by removing the contribution of other (interfering) sources using the sources current estimate).

3.4 Experimental results

Using the proposed approach, we perform underdetermined separation using a single musical record. The proposed signal represents a synthesized mixture of three notes played by an acoustic guitar. The record has a duration of 1s and is sampled at 22.050 kHz (see figure 1). Their pitch frequencies are respectively 82 Hz, 92 Hz, and 116 Hz. The SNR of the input signal is 26 dB.

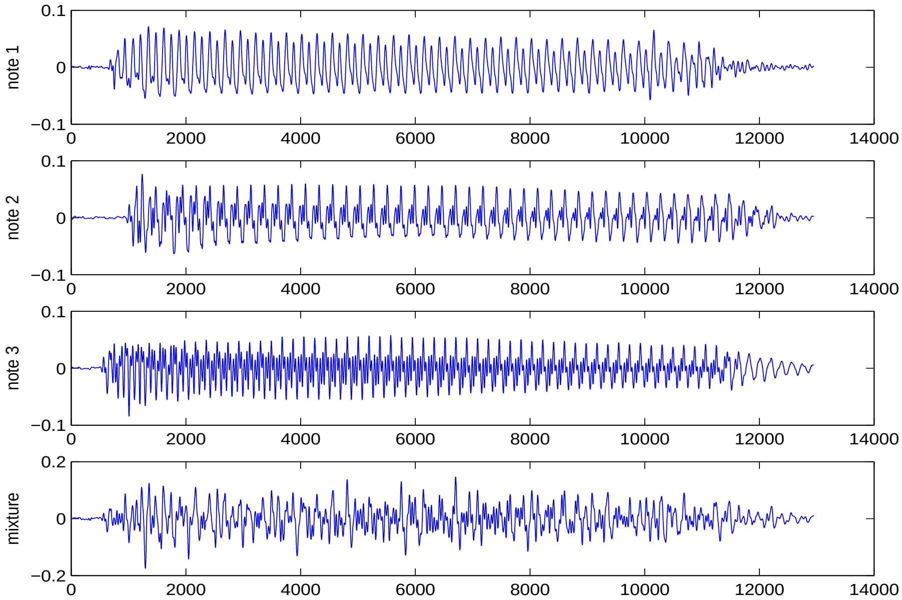

Figure 3.4: Original guitar signals.

As an evaluation criterion, we consider the signal to (measurement plus approximation) noise ratio for the estimated model (computed on the total note duration, and on the steady state region), i.e.

$$SNR_i = \frac{\sum_n s_i^2(n)}{\sum_n \left(s_i(n) - \hat{s}_i(n)\right)^2} \tag{3.15}$$

In figure 3.5, we plot the extraction SNR based respectively on global amplitude modulation and global time-warping model, and global amplitude and phase modulation (as described in the chapter 1).

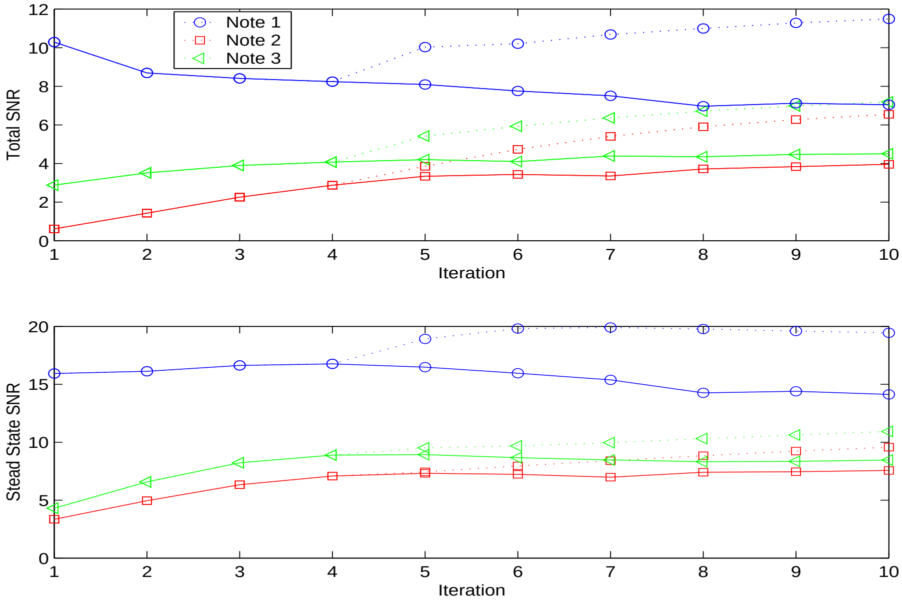

Figure 3.5: Estimation SNR for mono-mixture audio source separation (global time-warping model on solid line, and global amplitude and phase modulation on dotted line).

We observe that the second version achieves better performance, not only on the transient region but also on the steady state region).
We subplot also the different algorithm outputs (concerning the note 1) on figure 3.6. We note that although the notes are not synchronous (do not begin and vanish in the same time), the algorithm was able to detect the begin and the end of the musical note. This can be critical for some applications such as music transcription.

Next, we compare the separation accuracy of the proposed Iterated SIC with the sparse representation based approaches. Indeed, several authors have proposed underdetermined audio separation algorithms that are based on sparse time-frequency/time-scale representation of the data followed by binary masking [70]. One of the well known sparse decomposition techniques

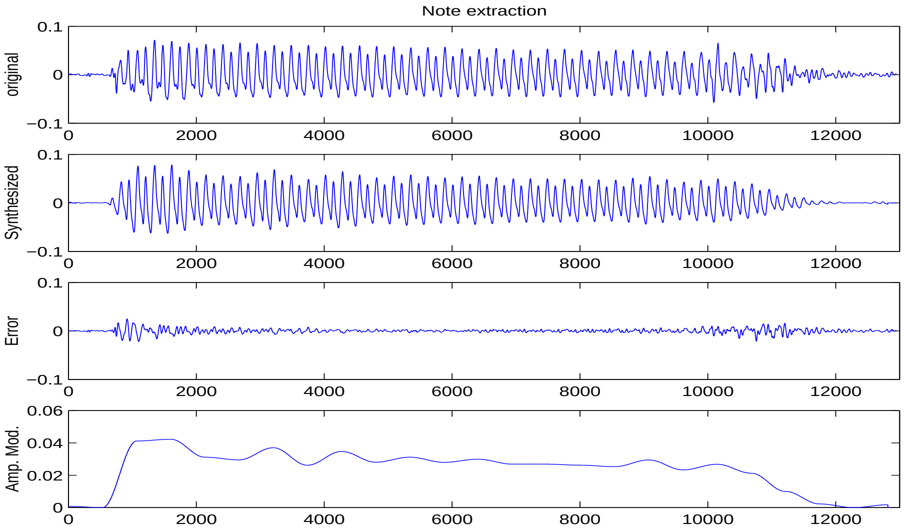

Figure 3.6: "Note 1" Extracted parameters.

is the Matching Pursuit (MP) algorithm [108]. The MP is a greedy strategy to decompose a signal into a linear combination of atoms chosen among a given dictionary. In each step, the element which "closely" matches the residual signal is selected. And, its contribution gets subtracted. Gribonval and Bacry propose a variant of the MP algorithm for audio applications (called Harmonic Matching Pursuit (HMP))[69]. They introduce the harmonic dictionary which extends the Gabor dictionary and better fits the harmonic structure of audio signals. At each step, an atom and all its (approximately) harmonically related atoms get selected.
The key observation for blind separation is that a good data representation often makes it possible to decompose a single underdetermined BSS problem into several (over)determined problems. In the one microphone setting, the underlying hypothesis is that at most one source is "active" in each component of the representation. The basic separation principle is simply to:

- decompose the observations into "components" (atoms).
- perform separation on each atom (which comes back to a classification problem).

The comparison between the Iterated SIC, MP, and HMP is summarized in the tables below.

We remark that matching pursuit fails to recover the note 3 (in figure 3.4); and that taking into account the harmonic structure of the audio signal (in the Iterated SIC, and HMP) increases the separation performance (especially in the steady state region). We see also that the QPSE-based approach outperforms the MP and the HMP approaches, and produces even much better auditive results.

We consider now the Multi-Input Multi-Output (MIMO) problem (figure 3.7). In our simulations, the audio source signals are captured by two microphones (spaced by $d = 0.2m$). The angles of arrival of the three source signals are respectively $\phi_1 = -\frac{\pi}{3}$, $\phi_2 = 0$, and $\phi_3 = +\frac{\pi}{3}$. The relative attenuations at the second microphones are respectively $\beta_{21} = 0.9$, $\beta_{22} = 1$, and $\beta_{23} = 1.1$.

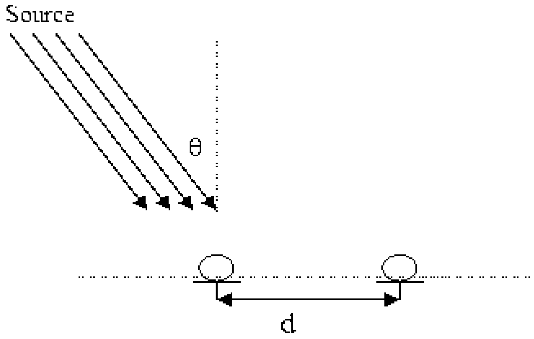

Figure 3.7: Multi-Input Multi-Output propagation scenario.

Figure 3.8 shows curves of the estimation SNR (for the total note duration) for MISO (slide line) and MIMO (dotted line) scenarios. As it was expected, we observe that, ones relative delays and attenuations are well estimated, using multiple output enable algorithm to achieve better performances.

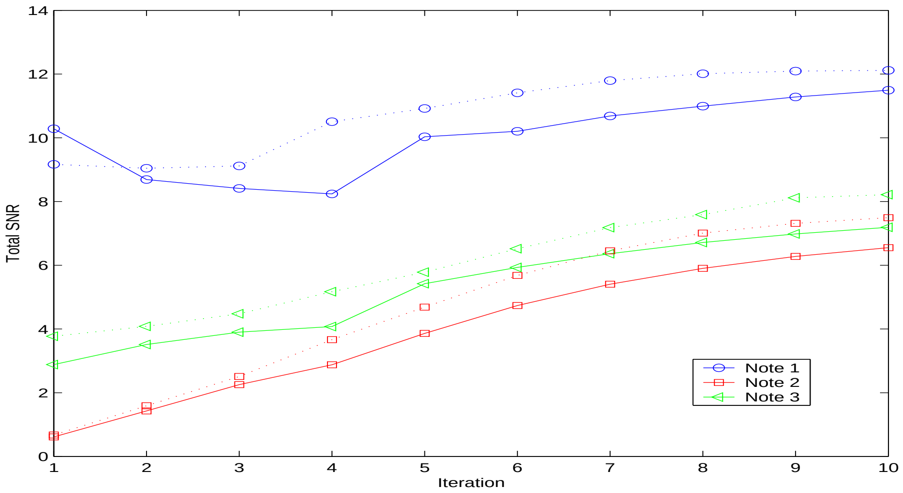

Figure 3.8: Estimation SNR for MISO (solid line), and MIMO (dotted line) audio source separation.

3.5 Conclusion

In this chapter, we have investigated the underdetermined convolutive source separation of audio mixtures. We have considered the periodic signal model with a slow global amplitude and phase variation. The global amplitude and phase modulation assumption allows taking into consideration the correlation between the different partials and time-domain frames. We have proposed a separation technique that takes into account simultaneously the source signal structure and the propagation environment parameters (ToA, signal attenuation). Experimental results reveal that the proposed approach allows extracting several musical notes accurately from an underdetermined mixture, and produces good auditive synthetic results. Simulations show also that the proposed scheme outperforms the classic separation schemes in terms of separation accuracy and robustness.

Iterated SIC Multichannel Audio Source Separation

Computation

```
Initialization
for  $i = 1 : N_s$  do
    $s_i(n) \leftarrow$ Periodic Source Extraction$(y_1(n), T_i)$
    for  $k = 1 : M$  do
        $\tau_{ki} = \arg\max_{\tau_{ki}} \widehat{R}\left(y_k, s_{\tau_{ki}}\right)$
        $\beta_{ki} = \frac{\widehat{R}\left(y_k, s_{\tau_{ki}}\right)}{\left\|s_{\tau_{ki}}\right\|^2}$
    end for
end for

Iteration
for  $i = 1 : N_s$  do
    Interference Cancellation
    for  $k = 1 : M$  do
        $y_k(n) \leftarrow y_k(n) - \sum_{p \neq i} \beta_{kp} \; s_p(n - \tau_{kp})$
    end for
    Channel Estimation
    for  $k = 1 : M$  do
        $\tau_{ki} = \arg\max_{\tau_{ki}} \widehat{R}\left(y_k, s_{\tau_{ki}}\right)$
        $\beta_{ki} = \frac{\widehat{R}\left(y_k, s_{\tau_{ki}}\right)}{\left\|s_{\tau_{ki}}\right\|^2}$
    end for
    Non parametric source estimation
    $s_i(n) \leftarrow \frac{1}{\sum_k \beta_{ki}^2} \sum_k \beta_{ki} y_k(n - \tau_{ki})$
    parametric source estimation
    $s_i(n) \leftarrow$ Periodic Source Extraction$(s_i(n), T_i)$
end for
```

Table 3.1: Iterated SIC Multichannel Audio Source Separation

	I-SIC	MP	HMP
Note 1	11.8	10.81	8.6
Note 2	7.11	4.57	6.1
Note 3	7.6	1.3	6.76

Table 3.2: Separation SNR (in dB) for the Iterated SIC, MP, and HMP (computed on the total note duration)

	I-SIC	MP	HMP
Note 1	19	14	12.57
Note 2	10.4	4.3	9.6
Note 3	11.61	0.21	11.09

Table 3.3: Separation SNR (in dB) for the Iterated SIC, MP, and HMP (computed on the steady state note region)

Chapter 4

Blind Source Dereverberation

In this chapter, we consider the blind multichannel dereverberation problem for a single source. The multichannel reverberation impulse response is assumed to be stationary enough to allow estimation of the correlations it induces from the received signals. It is well-known that a single-input multi-output (SIMO) filter can be equalized blindly by applying multichannel linear prediction (LP) to its output when the input is white. When the input is colored, the multichannel linear prediction will both equalize the reverberation filter and whiten the source. We exploit the channel spatiotemporal diversity and the speech signal non-stationarity to estimate the source correlation structure, which can hence be used to determine a source whitening filter. Multichannel linear prediction is then applied to the sensor signals filtered by the source whitening filter, to obtain source dereverberation. Particular attention is paid to the blind estimation of the source color (via the optimization of the AR coefficients and order). We also investigate the robustness of the scheme to the presence of additive noise.

4.1 Introduction

The quality of speech captured in real-world environments is invariably degraded by acoustic interference. This interference can be broadly classified into two distinct categories: additive and convolutive. The convolutive interference (commonly referred to as reverberation) is due to sound wave reflections from surrounding walls and objects. It leads to a modification of the speech signal characteristics. Therefore, it constitutes a major problem in speech recognition, speaker verification, and general auditive comfort in "hands-free" telephony applications. Blind dereverberation is the process of removing the effect of reverberation from an observed reverberant signal.
Let us consider a clean speech signal $s(n)$ produced in a reverberant room. The reverberant signal $\mathbf{y}(n)$ received on M microphones can be modeled as (see figure 4.1):

$$\mathbf{y}(n) = \sum_{i=0}^{\infty} \mathbf{h}(n,i)s(n-i) + \mathbf{v}(n) \tag{4.1}$$

where $\mathbf{v}(n)$ represents additive interference, and $\{\mathbf{h}(n,i)\}_i$ characterizes time-varying convolutional interference.

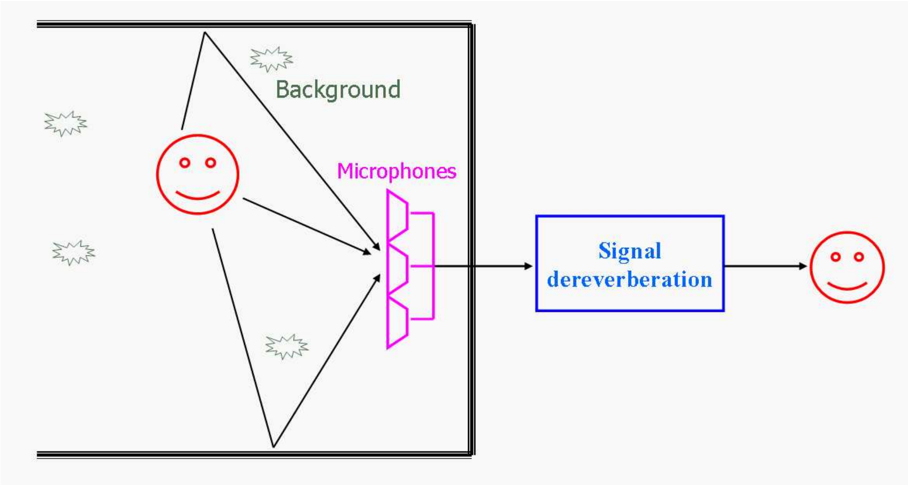

Figure 4.1: Speech dereverberation: problem statement.

As we have seen in chapter I, additive noise can be significantly reduced using

audio signal structure. On the other hand, removing reverberation appears easy: first, estimate the speaker-to-receiver filter $\{\mathbf{h}(n,i)\}_{k,i}$; then design an inverse filter $\{\mathbf{f}(n,i)\}_{n,i}$ to undo the reverberation effect. Although this sounds simple, in practice, reducing the distortion caused by reverberation is a difficult blind deconvolution problem. Speech enhancement for dereverberation and noise reduction in reverberant environments has been addressed extensively; but no adequate solution has yet been established [76, 57].
Intuitively, comparing to the classic speech enhancement problem, the speech dereverberation is trickier because the source signal itself belongs to the interference. Moreover, speech dereverberation is a challenging problem because:

- Both source signal statistics and the room reverberation are time-varying.
- Extra-information is required to make the problem identifiable. In fact, by filtering respectively the source signal and the Acoustic Impulse Response (AIR) using a given time-varying filter and its inverse, we obtain an acceptable solution to the blind deconvolution problem..
- Source signal correlations leads to a poor equalization performance.

In the literature, several schemes are proposed to solve the dereverberation problem exploiting essentially:

- Spatial diversity : due to the multichannel aspect (see section 4.3).
- Temporal diversity : the room impulse responses are slowly varying with time, whereas speech signal statistics change quickly.
- Spectral diversity: Dereverberation can be performed in the cepstral domain (where the room reverberation and the speech signal are better separated), or in the spectral domain (exploiting the harmonic structure of the speech signal).

In this chapter, the multichannel reverberation impulse response is assumed to be stationary enough to allow estimation of the correlations it induces from the received signals. The single-input multi-output (SIMO) channel is equalized blindly by applying multichannel linear prediction (LP) to its output when the input is white. To encounter for the input source color, we exploit the spatiotemporal channel diversity, and we estimate the source spectrum by averaging the received correlations (instead of averaging the received signals). We propose a tree-stage dereverberation procedure (figure 4.3):

- First, the colored non-stationary speech signal is transformed into an iid-like signal: exploiting the channel spatiotemporal diversity and the speech non-stationarity, we estimate an autoregressive model based on received correlations (averaged over the subchannels).

- Then, a blind channel predictor is computed based on pre-processed reverberant speech.

- Finally, speech signal dereverberation is performed using a zero-forcing equalizer based on the predictor computed in the previous step.

This chapter is organized as follows. After a quick overview of the dereverberation state of the art (section 4.2), the multichannel spatiotemporal diversity is examined in section 4.3. The speech dereverberation procedure, and the prewhitening order optimization will then be derived respectively in sections 4.4 and 4.5. We next investigate the robustness of the proposed scheme in presence of additive noise (section 4.6). Finally, simulation results are provided in section 4.7.

4.2 Speech dereverberation: a brief overview

4.2.1 Dereverberation based on spatial processing

As discussed in the previous paragraph, speech signals captured by a microphone located away from the user can be significantly corrupted by additive noise and reverberation. One method of reducing the signal distortion and improving the quality of the signal is the use of spatial processing. Spatial processing refers to the joint processing of signals captured by multiple spatially-separated sensors. Spatial filtering aims to discriminate between signals based on the physical location of the signal source. Spatial processing is relatively mature field, developed initially to process narrowband signals for radar and sonar applications, and later applied to broadband signals such as speech. As different signal replica come from different directions, spatial filtering seems to be appropriate for speech dereverberation.

Generally, spatial filtering exploits the fact that the source and the interferers are spatially distributed. The signals observed on the M distinct microphones are combined in order to focus on the direction of the audio source (see figure 4.2). The concept of algorithmically focusing in a desired direction is

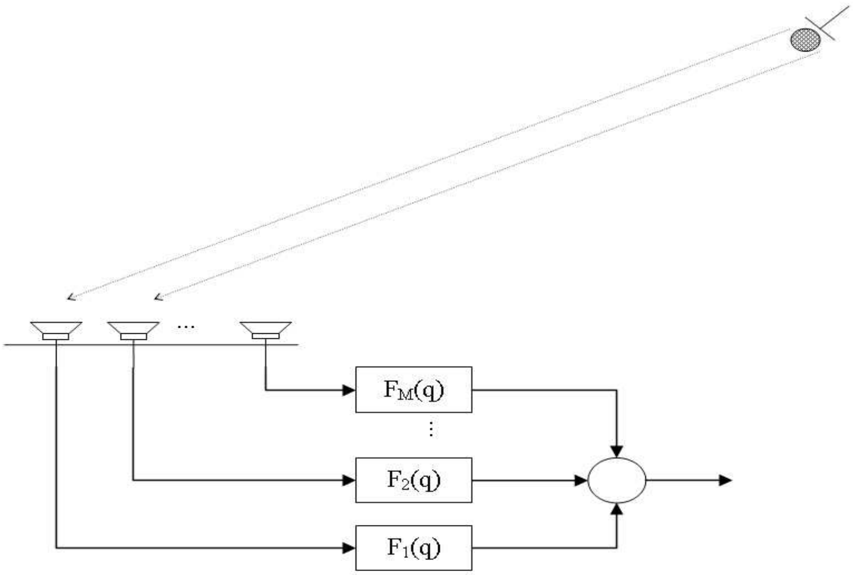

Figure 4.2: Spatial filtering: problem statement.

called Beamforming.
A simple (and most commonly used) spatial filtering system is the Delay-and-Sum (D-&-S) beamformer. In order to steer an array of arbitrary configuration and number of sensors, the received signals are first delayed to compensate for the path length differences from the source to the various microphones; and then the signals are combined together, i.e.,

$$\widehat{s}(n) = \sum_{i=1}^{M} f_i y_i(n - \tau_i) \tag{4.2}$$

where f_i and τ_i represent respectively the weight and the delay applied to the signal $y_i(n)$ received on the i^{th} microphone. There are several methods for choosing the weights $f_1 \cdots f_M$. The simplest and most common method is to set them all equal to $1/M$. Thus, beamforming is performed by a simple averaging over the sensor outputs, delayed to focus in the direction of the desired speaker. The process of finding the delays is known as time-delay estimation (TDE) and is closely related to the problem of source localization. Many methods exist in the literature, and most are based on cross correlations (see section 4.6.1 for further details).

Delay-and-sum (D-&-S) beamformer was proposed to perform speech dereverberation [48, 57]. By focusing on the source direction, the spatial filtering suppresses the signal replica coming from other directions. In addition to its low computational complexity, the D-&-S is robust to the presence of additive noise and errors in the time-delay estimations. However, one can remark that D-&-S exploits only partial spatial information (relative delays), and ignores the input signal characteristics; which leads to a poor dereverberation performance in strongly reverberant rooms and/or small microphone array.

4.2.2 Dereverberation based on speech signal features

The major objective of the multichannel dereverberation research has focused on improving the spatial filtering capability of a system in its operating environment. The previous section has addressed how spatial information can be exploited to dereverberate the speech signal. The goal of this section is to present an alternative and complementary strategy that emphasizes the incorporation of explicit speech modelling into the microphone array processing.

Dereverberation based on source statistical characteristics

A first class of speech dereverberation techniques suggests exploiting the statistical and spectral models of the speech signal to improve the enhancement accuracy. In [56], Gillespie et al. process the microphone signals by a subband adaptive filtering structure. The subband filters are adapted to maximize the kurtosis of the linear prediction residual of the reconstructed signal. The blind deconvolution filter is then designed to make the LP residual as non-Gaussian as possible. In this way, it exploits the a priori knowledge that the signal to be recovered (speech) is sub-Gaussian. It has been shown that a kurtosis is effective in measuring reverberation, and that the proposed technique achieves significant improvement in performance over the delay-and-sum beamformer.
A generic approach is proposed in [25, 27] exploiting simultaneously the non-Gaussianity, non-whiteness, and non-stationarity of the speech signal. In the previous references, the authors show that combining three properties leads

to a significant improvement in performances. It was also shown that the proposed approach has links to a variety of popular algorithms, and several novel approaches [25, 27].

Dereverberation based on source spectral characteristics

Some researchers have also proposed dereverberation methodologies that exploit the spectral structure of speech signals in a more direct manner.

Homomorphic filtering techniques was proposed for single-microphone dereverberation. These approaches are motivated by the fact that the convolutive interference becomes additive in the cepstral domain. In such an approach, it is assumed that the room impulse response and the original speech signal occupy separate regions in the cepstral domain. If the reverberation is produced by echoes equally spaced in time, the complex cepstral of the reverberation presents an impulsive structure. Then, a cepstral filtering procedure (using a comb filter) can be considered for reducing (even eliminating) the reverberation effect [126]. A second assumption (also discussed in [126]) associates the original speech signal with low quefrency components, and the room impulse response with high quefrency components. In such a way, dereverberation is performed by lowpass filtering the cepstrum of the reverberant signal.

In [20], the authors reveal some concerns about the applicability of cepstral processing to reverberant speech enhancement. Indeed, the accuracy of the computed cepstrum is critically dependent upon the segmentation error in time domain. This side effect can be alleviated by an appropriate choice of the segmentation window. Then, cepestral averaging is used to identify of the reverberation impulse response. Finally, the signal is dereverberated by inverting the estimated AIR. Satisfactory results related to the AIR estimation are obtained for minimum-phase or mixed-phase responses which have a few zeros outside the unit circle in the z-plane [20].

In all the previous research works, the major drawback is the assumption that the original speech signal and room impulse response must occupy non-overlapping regions. In general, such an assumption is valid for minimum-phase impulse responses. However, for mixed-phase responses there are contributions from the room response in the low quefrency region; and specifically, acoustic room responses have mixed-phase characteristic [170], restraining the use of dereverberation techniques based on cepstral analysis.

Nakatami and Miyoshi exploit differently the audio signal spectral prior. In [123], the authors suggest a frequency domain approach exploiting the harmonic structure of the speech signal to reduce late reverberation. The basic idea consists in estimating the direct sound by focusing on the local harmonic structure. The direct sound includes the direct path and some early reflections. Based on the received signal and the direct sound estimate, the acoustic impulse response is estimated in each frame. Assuming that the AIR are stationary enough, this estimate is enhanced by averaging over time frames. Finally, the speech signal is dereverberated by inverting the acoustic impulse response. The authors show that the technique is effective especially in the case of severe reverberation (reverberation time excides 0.5 seconds).

Dereverberation based on source-production techniques

An alternative solution is to explicitly incorporate the excitation speech model into the beamforming process. The source model describes speech signal in terms of an excitation sequence exciting a time-varying all-pole filter. These methods are motivated by the observation that in reverberant environments, the linear prediction residual signal contains the original impulse responses followed by several other peaks due to multipath propagation. Dereverberation is achieved by attenuating these peaks in the excitation sequence then synthesizing the enhanced speech using the enhanced LP residual and the all-pole filter (estimated from the reverberant speech).
Various methods for enhancing the LP residual exist. Griebel and Brandstein use coarse estimates of the room impulse response for each channel and apply a matched filter type operation to obtain weighting functions for the reverberant LP residuals [66]. Yegnanarayana et al. use Hilbert envelopes to represent the strength of the peaks in the LP residuals [209]. The time-aligned Hilbert envelopes from the individual channels are summed and used as a weight vector which is applied to the LP residual of one of the channels. In all these schemes, it is clear that an important assumption is made; that the speech LP coefficients are unaffected by reverberation. In [50], the authors show that spatial averaging of the LP coefficients (estimated on each microphone) is required to improve the accuracy of statement. They also demonstrate in [52] that LP coefficients obtained from spatially averaged multichannel speech signals achieves equally satisfactory results.

4.2.3 Dereverberation based on channel deconvolution

Another way to address the problem is the use of explicit model for the room reverberation; and contrary to spatial filtering the whole acoustic impulse response is toked into account. Depending if we try to exploit the statistical prior on the AIR, or if we consider the realization of the AIR, Bayesian and deterministic channel deconvolution techniques can be applied.

Dereverberation based on Bayesian channel deconvolution

Several methods are proposed to address the dereverberation problem in a Bayesian framework. In [39], the Gaussian model is used for modelling the input speech signal and the acoustic impulse response. MAP estimator is derived for the clean speech estimation (and marginalizing out the unknown channel parameter).

In [60], in order to avoid the channel source identification ambiguity, the speech source is modelled by an autoregressive process, whereas the channel is modelled by a MovingAverage (MA) process. The model parameters are estimated by maximizing the posterior probability. The maximization is performed using Markov Chain Monte Carlo (MCMC) method (to avoid the integration and the maximization of complicated posterior probabilities).

An alternative statistical model was proposed in [78, 79]. The non-stationary speech source is modelled by a block stationary AR process whereas each sub-channel by a stationary all-pole filter. In such a way, we avoid the channel source identification ambiguity. Using the Bayesian framework, the acoustic channel is estimated (source parameter are considered as nuisance parameters). Finally, the original signal is obtained by inverse filtering the observed reverberant signal.

Dereverberation based on deterministic channel deconvolution

Another general approach is based upon attempting to undo the effect of multipath propagation by considering the realization of acoustic impulse response. The acoustic impulse responses are in general not minimum phase and are not thus invertible. By beamforming to the direct path and the major images, it is possible to use the multipath propagation constructively to increase the SNR well beyond those achieved by the delay-&-sum beamformer. The result is a Matching Filtering (MF) process [85, 86, 87] which is shown to be effective to enhance the quality of reverberant speech, and attenuate the additive noise. Unfortunately, this technique has a number of practical limitations due to the blind channel estimation, the remaining late reflection, and the large equalization delay. In fact, the non-stationarity of the acoustic channels and the color of the input signal lead to a poor blind channel identification accuracy. On the other hand, matching filtering increases the propagation delay-spread and leads to an additional late reverberation component. Despite their low energy, the late reverberation components are very annoying for human perception and speech recognition applications. Finally, MF equalization introduces a large equalization delay (of about the AIR length), and produces a pre-echo that is also annoying for human perception and speech recognition applications.

On the other hand, a SIMO channel can be perfectly equalized using multiple Finite Impulse Response (FIR) filters (transverse filters) [116]. Let us consider a clean speech signal $s(n)$ produced in a reverberant room. The reverberant speech signal observed on M distinct microphones can be written as:

$$\mathbf{y}(n) = \mathbf{H}(q)s(n) \tag{4.3}$$

where $\mathbf{y}(n) = [y_1(n) \cdots y_M(n)]^T$ is the reverberant speech signal, $\mathbf{H}(q) = [H_1(q) \cdots H_M(q)]^T = \sum_{i=0}^{L_h-1} \mathbf{h}_i q^{-i}$ is the SIMO channel transfer function, $\{\mathbf{h}\}_i$ are the impulse response coefficients, and L_h is the channel length. $(.)^T$ denotes the transpose operator, and q^{-1} is the one sample time delay operator. According to the Bézout identity, if the channels $H_1(q) \cdots H_M(q)$ do not have common zeros, then $\exists \mathbf{F}(q) = [F_1(q) \cdots F_M(q)]^T$ (FIR) such that:

$$\mathbf{F}^T(q)\mathbf{H}(q) \;=\; \sum_{m=1}^{M} F_m(q)H_m(q) \;=\; 1 \tag{4.4}$$

If $\mathbf{H}(q)$ is known (or can be estimated), the coefficients of the FIR filters $F_m(q)$ can be computed by the well-known rules of matrix algebra. The blind AIR estimation should deal with the channel/speech identifiability problem. In fact, for any scalar filter $\alpha(q)$, $(H(q)/\alpha(q), \alpha(q)s(n))$ is also an acceptable solution for (4.3).

In [65], the authors compute the multi-channel FIR equalizer based on subspace methods. The identifiability problem is solved by using accurate information of the AIR length. The validity of the technique hinges critically

on the true channel impulse response being of strictly finite duration, and its successful identification requires knowledge of (at least a tight upper bound on) the channel length [197]. For the acoustic case, the true channel impulse response length is generally unknown, or/and not defined.
In [80], Huang et al. focus on the single-source two-microphone system. The authors notice that the AIR can be estimated by minimizing the mean squared value of the signal

$$e(n) = \widehat{H}_2(q)y_1(n) - \widehat{H}_1(q)y_2(n) \tag{4.5}$$

A solution of the optimization problem can be obtained through the eigenvalue decomposition of the autocorrelation matrix of the observed signal. The generalization to an arbitrary channel number is introduced in [81]. If the channel length is known, the solution estimates the right AIR. However, if the channel length is overestimated (let us denote by $\overline{L}_h$ the overestimated length), for any scalar filter $\alpha(q)$such that $order(\alpha(q)) < (\overline{L}_h - L_h)$, $\alpha(q)\mathbf{H}(q)$ is a solution of (4.5)

$$\begin{aligned} e(n) &= \alpha(q)H_2(q)y_1(n) - \alpha(q)H_1(q)y_2(n) \\ &= \alpha(q)\left(H_2(q)y_1(n) - H_1(q)y_2(n)\right) = 0 \end{aligned}$$

Hikichi et al. propose to solve the identification ambiguities by post-processing the estimated channel in order to estimate and compensate the common factor $\alpha(q)$ [76]. The common factor is extracted as the characteristic polynomial of the two-channel linear prediction matrix.
Another way to deal with identification ambiguities is the use of prior information on the source spectrum. In fact, if the source is white, the channel can be perfectly equalized using multichannel linear prediction. If the source spectrum is known, perfect equalization is still possible (after source prewhitening). For speech dereverberation, the source spectrum is unknown and should be estimated blindly. In [56], the authors propose a subband equalization structure. The equalization is done assuming a flat source spectrum in each subband. The spectral structure of the speech signal can also be exploited by assuming an AutoRegressive (AR) model. Cichocki and Amari estimate the AR coefficients based on the output correlations of a spatially filtered received signal [34].

In this chapter, we exploit differently the spatiotemporal channel diversity, and we propose estimating the AR source prewhitening filter by averaging the

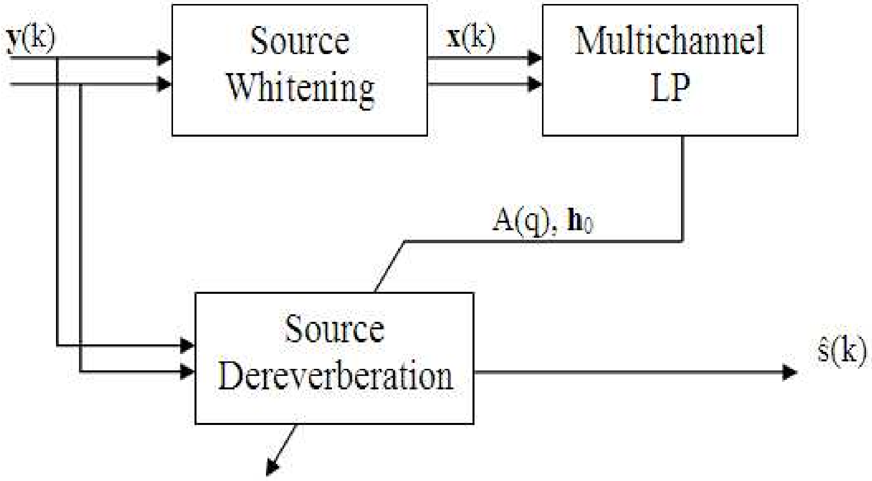

Figure 4.3: The dereverberation procedure.

received correlations (instead of averaging the received signals). We propose a tree-stage dereverberation procedure (figure 4.3):

- First, the colored non-stationary speech signal is transformed into an iid-like signal. Exploiting the channel spatiotemporal diversity and the speech non-stationarity, we estimate an AutoRegressive model based on received correlations (averaged over the subchannels).

- Then, a blind channel predictor is computed based on pre-processed reverberant speech.

- Finally, speech signal dereverberation is performed using a zero-forcing equalizer based on the predictor computed in the previous step.

4.3 Multichannel spatiotemporal diversity

4.3.1 Statistical room reverberation model

In an empty rectangular room, the room impulse response $h(t)$ can be computed by solving the wave propagation equations. At higher frequency, the complexity (in terms of the number of modes) of the deterministic wave equation modeling increases to a point where exact analysis is no longer feasible. To model $h(t)$, one could apply the theory of random (or diffuse) sound fields [160]. The signal captured by the microphone is the sum of contributions of a large number of modes. Consequently, the complex frequency response can be considered as a space-dependent Gaussian process. The two-dimensional Gaussian density arises from the central limit theorem assuming independence between modes.

In this section, we introduce the room reverberation model, which is built on some well-known results from statistical room acoustics. This theory closely describes the room acoustic behavior if the following conditions are met [51]:

A1) The dimension of the room are large relative to the wavelength of the source signal $s(t)$. For the frequencies of interest in speech processing, this condition is easily satisfied in almost all rooms.

A2) The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V (in m^3), and reverberation time T_{60} (in seconds), this condition is fulfilled for frequencies that exceed the “Schroeder large room frequency":

$$f_{sch} = 2000\sqrt{T_{60}/V} \tag{4.6}$$

A3) The source and the microphones are located in the interior of the room, at least a half-wavelength away from the walls.

Under the above conditions, the frequency response $H(f)$ (the Fourier transform of $h(t)$) can be treated as a random function of the source and microphone positions. These statistical properties are independent of the time-instant of the observation. They are determined by the room characteristics (volume, reverberation time, average wall absorption coefficient...). We write the transfer function $H(f)$ as

$$H(f) = H^r(f) + jH^i(f) \tag{4.7}$$

where $H^r(f)$ and $H^i(f)$ are real and imaginary parts of $H(f)$ respectively, and $j = \sqrt{(-1)}$ is the unitary imaginary number. We next cite a couple of useful results derived using the Statistical Room Acoustics (SRA) theory[51, 160, 161, 162]. Assuming the assumption (A1-A3) to be fulfilled:

- $H^r(f)$, and $H^i(f)$ are independent, zero-mean, Gaussian process.
- $\left\langle |H(f)|^2 \right\rangle = \left\langle |H^r(f)|^2 + |H^i(f)|^2 \right\rangle = \dfrac{1-\beta}{\pi A \beta}$.

where $\langle . \rangle$ denotes the spatial expectation (estimated by averaging over all possible source and microphone positions), β is the average wall absorption coefficients, and A is the total wall surface area.
By denoting $r_h(t) = \sum_\tau h(\tau)h(\tau - t)d\tau = \text{IFFT}\left(|H(f)|^2\right)$ the autocorrelation of the room impulse response, one can show that:

$$\begin{cases} \langle r_h(t) \rangle = \dfrac{1-\beta}{\pi A \beta}\ \delta(t) \\ \left\langle r_h^2(t) \right\rangle = \text{cst}\ \ \exp\left(-\left|t\right|/\tau_0\right) \\ \langle r_h(t_1) r_h(t_2) \rangle = 0 \ \ \forall t_1 \neq t_2 \end{cases} \tag{4.8}$$

where $\delta(t)$ denotes the Dirac delta function and τ_0 is the time for which the sound energy in the room decays to $1/e$ of its initial value after impulsive excitation ($\tau_0 = T_{60}/13.8$).

We also assume that the room impulse response between a source and M microphones (and the corresponding autocorrelations) are i.i.d. Thus, by averaging the correlation of the different subchannels ($r_M(t) = \frac{1}{M}\sum_{m=1}^{M} r_{h_m}(t)$), we have

$$\begin{cases} \langle r_M(t) \rangle = \langle r_{h_1}(t) \rangle = \dfrac{1-\beta}{\pi A \beta}\ \delta(t) \\ \text{var}\left(r_M(t)\right) = \dfrac{1}{M}\text{var}\left(r_{h_1}(t)\right) \end{cases} \tag{4.9}$$

where $\text{var}\,(x) = \langle x^2 \rangle - \langle x \rangle^2$ denotes the spatial variance of a spatially distributed variable x.

In [141], Polack developed a time-domain model extending and complementing the Schroeder frequency domain model . In this model, a room impulse response is described as one realization of a non-stationary stochastic process:

$$h(t) = b(t)e^{-\tau_0 t} \quad t \geq 0 \tag{4.10}$$

where $b(t)$ is centered stationary Gaussian noise, and τ_0 is defined as previously. The random noise is characterized by its Power Spectral Density (PSD) $P(f)$. $P(f)$ and τ_0 do not depend neither on the source nor on the microphone positions, and characterize the room acoustic.

Assuming the previous model, one can show:

$$\left\langle \left| H\left(f\right) \right|^2 \right\rangle = \frac{P(f).T_{60}}{6 \ln 10}. \tag{4.11}$$

The time-model based model coincides with Schroeder theory for a flat power spectral density $P(f)$. By considering a non flat PSD and/or frequency dependent reverberation time, Polack's model takes into consideration the frequency dependence of the reflection coefficients of walls and other objects, as well as the frequency dependence of the air absorption coefficient.

4.3.2 Spatiotemporal diversity of SIMO acoustic channels

To investigate the spatiotemporal diversity of SIMO acoustic channels, we consider a measured multichannel impulse response (from the MARDY database). The documentation of the MARDY database can be found in [206]. The sampling frequency is 48 kHz.
First, we fix the source at distance $d = 2m$ from the microphone array (in a central position). Figure 4.4(a) superposes the magnitudes of channel transfer functions $\left| H_m\left(f\right) \right|^2 \quad m = 1 : M$ between the source and the $M = 8$ microphones. The transfer function magnitude of the multichannel reverberation filter $\sum_{m=1}^{M} \left| H_m\left(f\right) \right|^2$ is plotted in figure 4.4 (b). We move the source (3m from the microphone array, in a position to the right), and we plot the same quantities in figure 4.5.
We verify that the averaged spectrum is independent from the source position; but it is not flat (as the room characteristics are frequency dependent). On the other hand, for speech processing we are mainly interested on the band 50-7000 Hz. If we consider this frequency band, we can see that the averaged spectrum is almost constant (figures 4.6 and 4.7). The frequency-dependence on the room characteristic can be neglected.
By summing the spectra of the received signals $S_{y_m y_m}(f)$, we get:

$$\sum_{m=1}^{M} S_{y_m y_m}(f) = \sum_{m=1}^{M} \left| H_m\left(f\right) \right|^2 S_{ss}(f) \approx c S_{ss}(f) \tag{4.12}$$

Then, due the multichannel spatiotemporal diversity, the superposition of the spectra of the received signals can estimate (up to a multiplicative constant c) the source spectrum $S_{ss}(f)$ (figure 4.8).

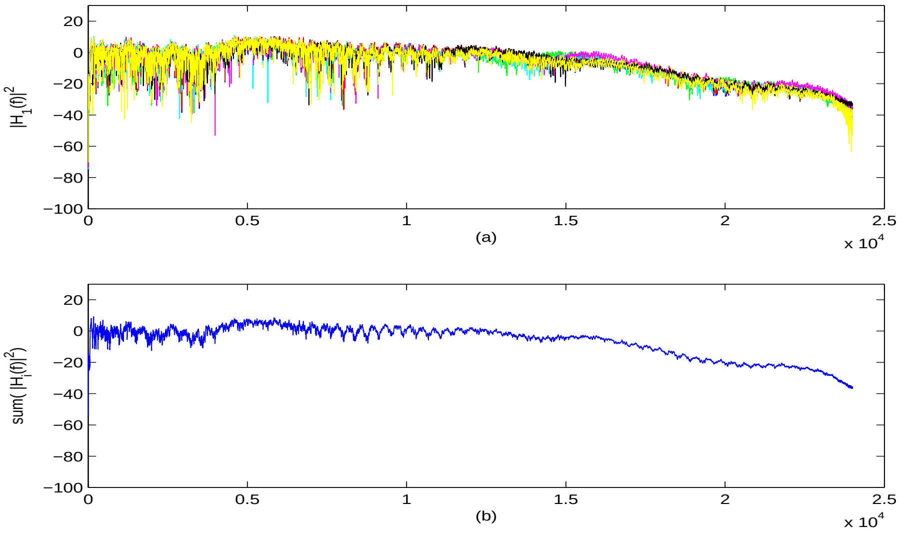

Figure 4.4: Mono and multichannel transfer function magnitudes (M=8, d=2, pos=C).

4.4 Speech dereverberation procedure

Motivated by the previous observation, we propose in this contribution a processing scheme that works with a cascade of three stages:

- Source whitening stage: removes correlation due to the speech signal.
- Multichannel prediction stage: computes a blind multichannel predictor (using pre-processed reverberant speech).
- Dereverberation stage : equalizes the channel impulse response (using a zero-forcing equalizer based on the predictor computed in the previous step).

In the following, we describe further the three stages.

4.4.1 Source whitening stage

The blind dereverberation should deal with the channel/source identifiability. In fact, for any scalar filter $\alpha(q)$, $(H(q)/\alpha(q), \alpha(q)s(n))$ is also an acceptable

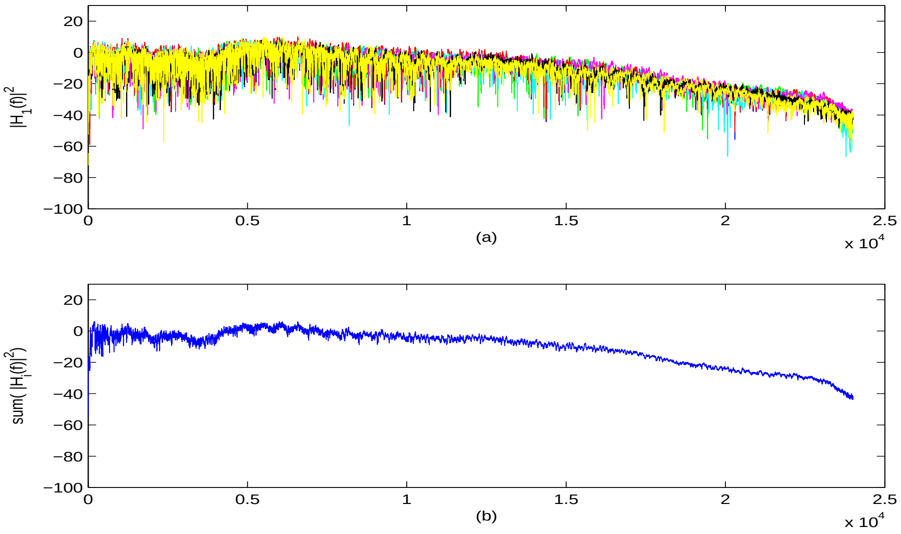

Figure 4.5: Mono and multichannel transfer function magnitudes (M=8, d=3, pos=R).

solution. One way to solve this problem is to assume a prior knowledge on the source color (and whiten the source by pre-processing the received signal).

As we have seen previously, due to the multichannel spatiotemporal diversity, the superposition of the spectra of the received signals estimates (up to a multiplicative factor) the source spectrum. This motivates us to remove correlation due to the source speech signal by compensating the common part in the multichannel impulse response. As this common part is due to the anechoic speech signal, it can be modeled as an AR process. The common AR coefficients can be estimated as those that minimize the sum of the prediction errors, averaged over the microphones:

$$\begin{aligned} e &= \sum_{m=1}^{M} \sum_{n=0}^{\infty} e_m^2(n) \\ &= \sum_{m=1}^{M} \sum_{n=0}^{\infty} \left[y_m(n) - \sum_{j=1}^{l} a_j y_m(n-j) \right]^2 \end{aligned} \tag{4.13}$$

This cost function was also considered in [73] for the estimation of the com-

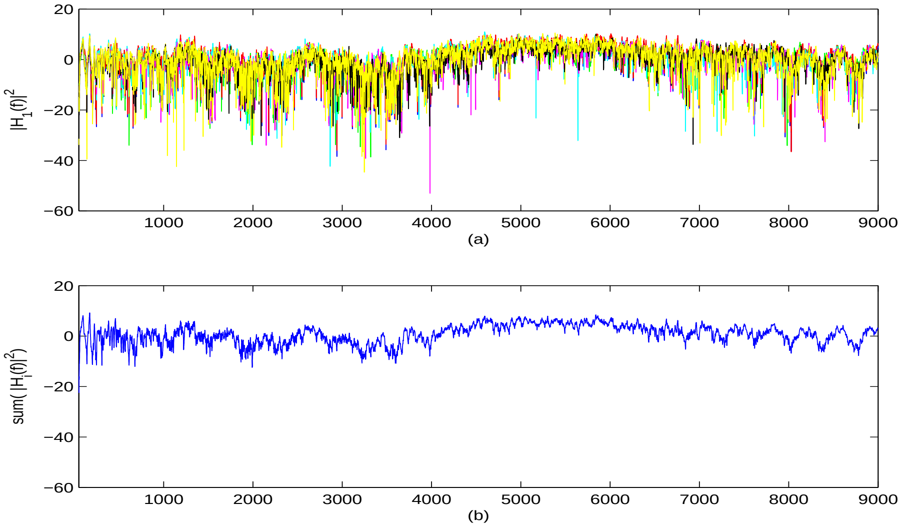

Figure 4.6: Mono and multichannel transfer function magnitudes (M=8, d=2, pos=C), restricted to 9 kHz.

mon acoustical poles of room transfer functions. The order of the AR process (whitening order) l is optimized in section 4.5.
In figure 4.9 we superpose the anechoic speech signal periodogram, and the AR spectral models estimated using either the source signal directly, or the sum of the correlation sequences of the M reverberant signals.
It can be seen that the AR spectrum estimated using reverberant signals gives a good estimation (up to a scalar) of the clean speech spectrum. Thus, it can be used to pre-process the reverberant speech in order to prewhiten the colored source speech signal.

A periodic input signal (which is perfectly predictible) may lead to identifiability problem for the SIMO channel: the predictor will have tendency to kill the signal rather than to whiten it. To alleviate this problem, we propose taking advantage from the signal non-stationarity (that can be interpreted as a form of temporal diversity). We suggest computing the AR whitening coefficients based on a long frame (where the channel is assumed to be constant, but the speech signal is not necessarily stationary). In such a way, the AR spectrum estimates the averaged speech spectrum (over the considered frame). It is important to emphasize that non-stationarity of the source

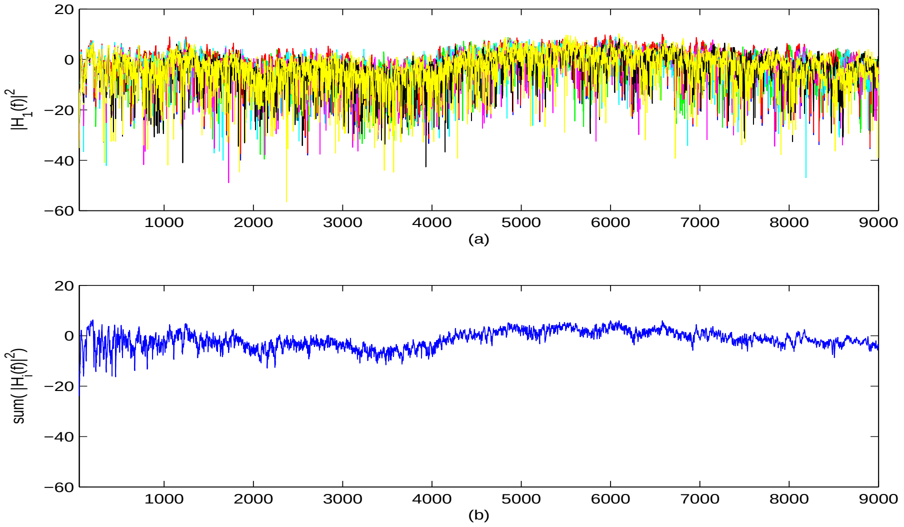

Figure 4.7: Mono and multichannel transfer function magnitudes (M=8, d=3, pos=R), restricted to 9 kHz.

is irrelevant as long as the source correlations are estimated with the same temporal averaging as for the multichannel linear prediction. The temporal diversity becomes a byproduct of this requirement.

4.4.2 Multichannel prediction stage

Blind multichannel identification and equalization exploiting the channel diversity introduced by sensor arrays has attracted a lot of interest in the communication and signal processing societies. Basically, the multichannel diversity introduces a useful 'signal overdetermination' which can be exploited in terms of signal/noise subspace decompositions. The multichannel linear prediction based technique (proposed and refined by Slock et al. [165, 167]) proved to be consistent in the presence of channel order error. This makes the LP-based equalization one of the more attractive solutions to blind speech dereverberation.

The source whitened reverberant signal observed on M distinct micro-

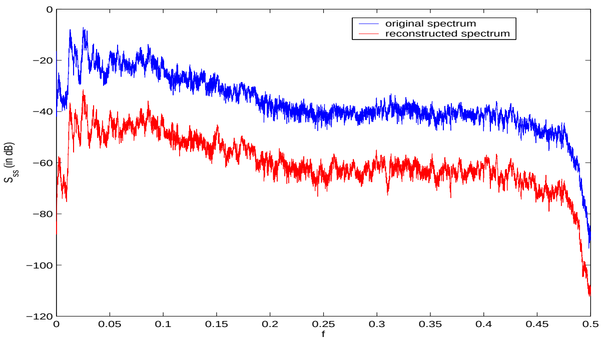

Figure 4.8: Welch periodogram of the original clean signal (on top), and the reconstructed one (by superposing the Welch spectra of the received signals).

phones can be written as:

$$\mathbf{x}(n) = \widehat{a}_{s,M}(q)\mathbf{y}(n) = \mathbf{H}(q)\widetilde{s}(n) \tag{4.14}$$

where $\mathbf{x}(n) = [x_1(n)\cdots x_M(n)]^T$, $\widehat{a}_{s,M}(q) = 1 + \sum_{j=1}^{l} \widehat{a}_{s,M}(j)q^{-j}$ is the prewhitening filter (performed in the previous stage), $\widetilde{s}(n) = \widehat{a}_{s,M}(q)s(n)$ is the prewhitened clean source signal.

Consider now the problem of predicting $\mathbf{x}(n)$ from the L_A latest observations $\mathbf{x}_{L_A}(n-1) = [\mathbf{x}^T(n-1)\cdots\mathbf{x}^T(n-L_A)]^T$. The prediction error is given by:

$$\widetilde{\mathbf{x}}(n) = \mathbf{x}(n) + \sum_{i=1}^{L_A} \mathbf{A}_{x,i}\mathbf{x}(n-i) = \mathbf{A}_x\mathbf{x}_{L_A+1}(n) \tag{4.15}$$

where $\mathbf{A}_x = [\mathbf{I}_M\ \mathbf{A}_{x,1}\ \cdots\ \mathbf{A}_{x,L_A}]$, $\{\mathbf{A}_{x,i}\}_i$ represents $M \times M$ matrices of the linear prediction coefficients, $\mathbf{I}_M$ is the identity matrix of size M, and L_A denotes the prediction order. The linear prediction matrices $\{\mathbf{A}_{x,i}\}_i$ are computed by minimizing the mean squared value of $\widetilde{\mathbf{x}}(n)$, which leads to normal equations (appendix 4.B).

According to (4.4), $\exists \mathbf{F}(q) = [f_1(q)\cdots f_M(q)]^T$ such that:

$$\mathbf{F}^T(q)\mathbf{x}(n) \ = \ \mathbf{F}^T(q)\mathbf{H}(q)\widetilde{s}(n) \ = \ \widetilde{s}(n) \tag{4.16}$$

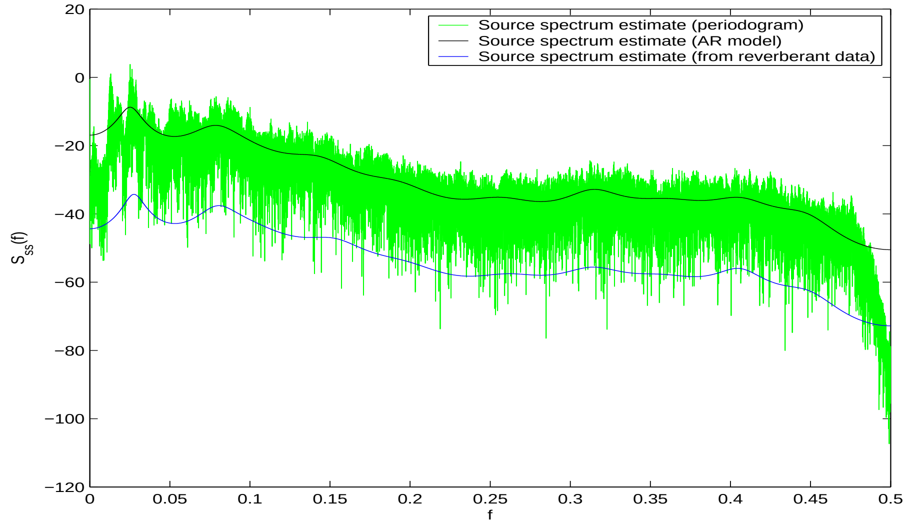

Figure 4.9: Source periodogram, spectrums of AR processes estimated using the clean and the reverberant signals ($l = 20, M = 8$).

Form (4.16), we see that $\mathbf{x}(n)$ is an autoregressive process.
As $\mathbf{x}(n) = \sum_{i=0}^{L_h-1} \mathbf{h}_i \widetilde{s}(n-i)$, the innovation process (the part of $\mathbf{x}(n)$ that can not be predicted from previous samples) is

$$\mathbf{x}_{inov}(n) \approx \mathbf{h}_0 \widetilde{s}(n) \tag{4.17}$$

where $\mathbf{h}_0 = \mathbf{H}(+\infty)$ represents the first vector coefficient of the SIMO channel filter (called the precursor coefficient). The approximation in (4.17) is due to the fact that $\widetilde{s}(n)$ is not perfectly white. Thus, using a long enough multichannel LP filter $\left(L_A \geq \frac{L_h - 1}{M-1}\right)$, solving the well-known normal equations leads to [133]

$$\widetilde{\mathbf{x}}(\mathbf{n}) = A_x(q)\mathbf{x}(n) = \mathbf{x}_{inov}(n) \approx \mathbf{h}_0 \widetilde{s}(n) \tag{4.18}$$

The minimum prediction error covariance is

$$\mathbf{\Sigma}_{\widetilde{\mathbf{x}}}(z) = A_x(z) S_{\mathbf{xx}}(z) A_x^{\dagger}(z) \approx \mathbf{h}_0 \sigma_{\widetilde{s}}^2 \mathbf{h}_0^T \tag{4.19}$$

where $\sigma_{\widetilde{s}}^2$ is the energy of the scalar process $\widetilde{s}(n)$, and $A_x^{\dagger}(z) = \sum_{i=0}^{L_A} \mathbf{A}_{x,i}^H z^i$ is the matched filter associated to $A_x(z)$. Therefore, the prediction error covariance has rank 1. Moreover, (4.19) allows estimating $\mathbf{h}_0$ (up to a scalar) as the eigenvector corresponding to the maximum eigenvalue of the prediction error covariance $\mathbf{\Sigma}_{\widetilde{\mathbf{x}}}$.

Note that the estimated LP filter $\widehat{A}_x(q)$ and precursor coefficient $\widehat{\mathbf{h}}_0$ depends only on the second order statistics of the reverberant signal. Thus, the proposed approach can be easily extended to the presence of an additive white noise, since the white noise variance can be identified and compensated for in the reverberant signal covariance matrix [132]. Further investigations are considered in the section 4.6.2.

4.4.3 Dereverberation stage

As we have shown previously, if $\widetilde{s}(n)$ is white and the channel satisfy the no-common zero condition, using a long enough multichannel linear predictor $\left(L_A \geq \dfrac{L_h - 1}{M - 1}\right)$ removes the multi-propagation effect, i.e.,

$$\widetilde{\mathbf{x}}(\mathbf{n}) = A_x(q)\mathbf{x}(n) = \mathbf{h}_0 \widetilde{s}(n)$$

A zero forcing equalizer can be defined by combining the multichannel LP outputs:

$$\mathbf{F}_{\mathbf{D\&P}}^{\mathbf{T}}(q) = \widehat{\mathbf{h}}_0^H \widehat{A}_x(q) \tag{4.20}$$

The proposed equalizer is called Delay-and-Predict (D-&-P). The choice of the name is justified in the section 4.6.1.
Note that $\widehat{\mathbf{h}}_0^H$ is the optimal gain combiner (the constant norm vector that maximizes the energy of the desired signal output).
Finally, the dereverberated speech signal can be computed as:

$$\widehat{s}(n) = \mathbf{F}_{\mathbf{D\&P}}^{\mathbf{T}}(q)\mathbf{y}(n) = \widehat{\mathbf{h}}_0^H \widehat{A}_x(q)\mathbf{y}(n) \tag{4.21}$$

4.5 Whitening order optimization

A key parameter in our dereverberation scheme is the order of the whitening filter. In fact, if the correlation matrix of the pre-processed speech signal

$\widetilde{s}(n) = a_s(q)s(n)$ is spherical and if we take a long enough multichannel LP filter $\left(L_A \geq \frac{L_h - 1}{M-1}\right)$, Delay&Predict equalizes perfectly the channel. To investigate the choice of this parameter, we consider a rectangular room with dimensions $L_x = 8m$ $L_y = 10m$ and $L_z = 4m$, and with wall reflection coefficients $\rho_x = \rho_y = \rho_z = 0.9$ ($T_{60} \approx 500ms$). A speech signal with duration of 8.8s, and sampled at 8 kHz is used as the original source signal (figure 4.10). The reverberant speech signal is observed on 8-elements microphone array.

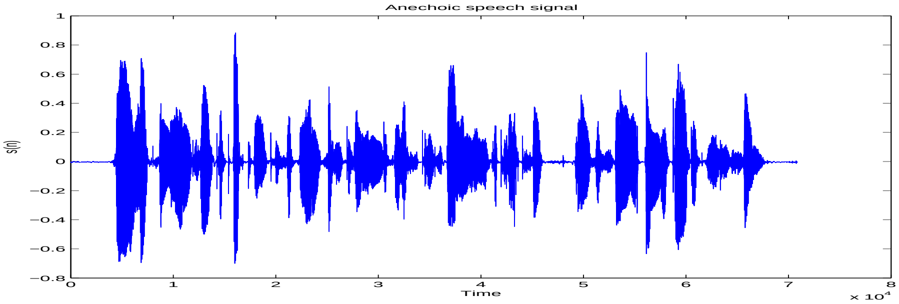

Figure 4.10: Anechoic speech signal.

A computer implementation (graciously provided by Geert Rombouts from K.U. Leuven) of the image method as described in [7, 139] is used to generate synthetic room impulse response between the source and the microphones. Figure 4.11 (a) plots the equalized channel ($f_{D\&P} * h$) impulse response and spectrum, and the spectrum of the whitened source speech signal (preprocessed using a 20-order linear predictor). We remark that due to the fact that the speech signal is a bandpass signal (observe values on very high and low frequencies), the Delay-and-Predict equalizer has a tendency to amplify the missing frequency components (as it is a zero-forcing equalizer); the fact that degrade the dereverberation performance. However, if we keep increasing the value of the order of the whitening LP filter and specially if it exceed the pitch period, this side effect is reduced (see figure 4.11 (b)). In such a case, the whitening LP is able to remove both short-terms and long-terms correlations, and the signal $\widetilde{s}$ fits better the whiteness assumption.

Next, we consider the Direct to Reverberant energy Ratio (DRR) as an evaluation criterion for the dereverberation accuracy:

$$DRR = 10 \log_{10} \left\{ \frac{\sum_{t=0}^{\tau-1} \widetilde{h}^2(t)}{\sum_{t=\tau}^{L-1} \widetilde{h}^2(t)} \right\} \quad dB \qquad (4.22)$$

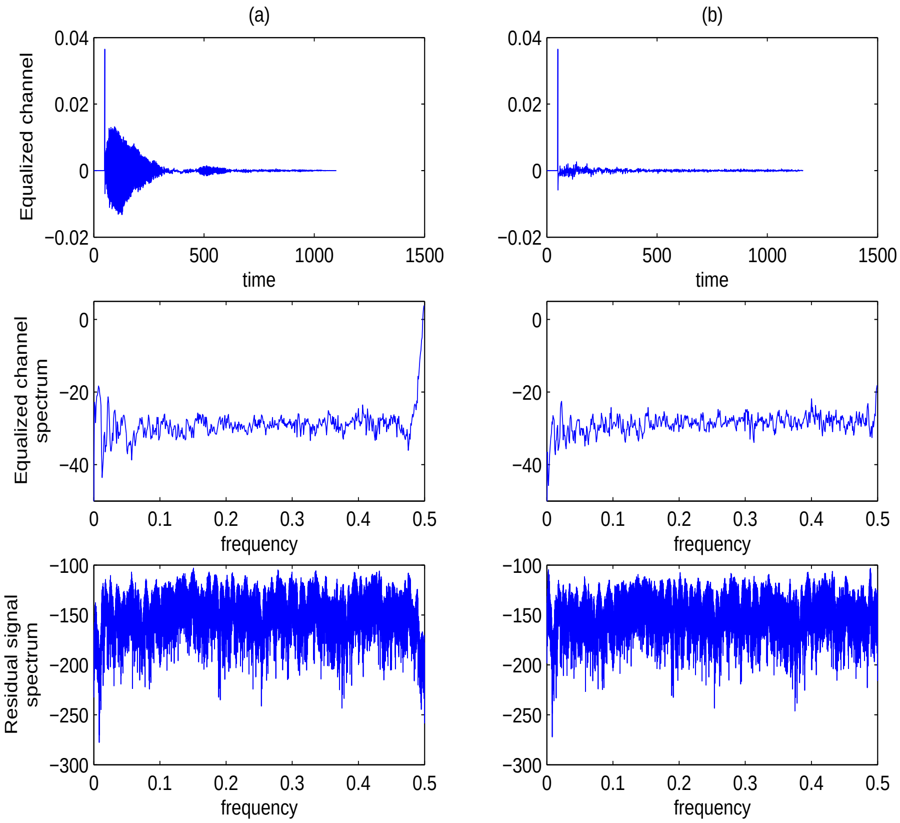

Figure 4.11: Equalized channel impulse response and spectrum, and the source preprocessed speech signal. (a) $l = 20$. (b) $l = 100$.

where $\widetilde{h}^2(t) = h * f(t)$ denotes the equalized channel, τ is the number of samples to include as the direct component, and $L = T_{60}f_s$ is the length of the impulse response (T_{60} is the reverberation time, and f_s is the sampling frequency).

Figure 4.12 shows the curves of the output DRR (function of the whitening filter order) using 2, 4, and 8 microphone array setup ($\tau = 10ms$). The order l of the AR process $a_s(q)$ (whitening order)is plotted in logarithmic scale $20\log_{10}(l)$.

We observe two distinct behaviors (depending on the size of our microphone array):

Figure 4.12: The output DRR (function of the whitening filter order), using 2, 4, and 8 microphones.

- If we have only 2-microphones, the two-channel filter cannot be assumed to be all-pass (spatiotemporal diversity is not enough). Then, by increasing the order of the whitening filter ($l > 100$) we are capturing details belonging either to the clean speech, and/or the channel. The whitening in the first stage will also remove some channel correlation before the multichannel equalization. The fact that affects the overall dereverberation accuracy.
- However, for the 8 microphone array setup, the all-pass multi-channel assumption is better matched. Then by increasing the whitening LP order, we remove essentially more source correlation. And the whiteness assumption of $\widetilde{s}$ is better fitted.

Remark that this problem is quite different from the classic AR order selection problem, where the estimation of the source correlations is troubled by the finite number of the available observations [23]. In our problem, we assume having enough observations to have an accurate estimation of the received signals correlations. The disturbance is due to the blind estimation of the source color: the channels are not flat and the number of microphones is not infinite. The whitening order should optimize the tradeoff between the modeling error (limited source whitening) and the estimation error (due to the blind estimation of the source correlations). In this section we propose, using a statistical room reverberation model, a design to optimize the whitening order (function of the room characteristics, and the number of subchannels).

4.5.1 Speech source whitening

As it is reported in the section 4.3.1, using the SRA theory one can show that for frequencies $f > f_{sch}$, the average reverberation spectrum is flat , i.e.,

$$\left\langle |H\,(f)|^2 \right\rangle = \frac{1-\beta}{\pi A \beta}$$

Then, the superposition of the spectra of the received signals can estimate (up to a multiplicative factor) the source spectrum. As this common part is due to the anechoic speech signal, it can be modeled as an AR process, i.e.,

$$s(n) = \frac{1}{a_s(q)} u_s(n) \tag{4.23}$$

where $u_s(n)$ is a zero-mean white process. The common AR coefficients can be estimated as those that minimize the sum of the squared prediction error signal, averaged over the M microphones:

$$e = \sum_{m=1}^{M} \sum_{n=0}^{\infty} e_m^2(n) = \sum_{m=1}^{M} \sum_{k=0}^{\infty} \left[y_m(n) - \sum_{j=1}^{l} a_j y_m(n-j) \right]^2 \tag{4.24}$$

The previous optimization problem leads to the normal equations:

$$\underbrace{\begin{bmatrix} r_{y,M}(0) & r_{y,M}(1) & \cdots & r_{y,M}(l-1) \\ r_{y,M}(1) & r_{y,M}(0) & \cdots & r_{y,M}(l-2) \\ \vdots & & \ddots & \vdots \\ r_{y,M}(l-1) \cdots & r_{y,M}(1) & & r_{y,M}(0) \end{bmatrix}}_{R_{y,M}} \underbrace{\begin{bmatrix} \widehat{a}_{s,M}(1) \\ \widehat{a}_{s,M}(2) \\ \vdots \\ \widehat{a}_{s,M}(l) \end{bmatrix}}_{\widehat{\mathbf{a}}_{s,M}} = - \underbrace{\begin{bmatrix} r_{y,M}(1) \\ r_{y,M}(2) \\ \vdots \\ r_{y,M}(l) \end{bmatrix}}_{\mathbf{P}_{y,M}}$$

where - $r_{y,M}(j) = \dfrac{1}{M} \displaystyle\sum_{m=1}^{M} r_{y_m y_m}(j)$ is the averaged correlation of the received signals at time-lag j.

- $r_{y_m y_m}(j)$ represents the correlation at the time-lag j of the received signal at the m^{th} microphone.

- $\{\widehat{a}_{s,M}(j)\}_j$ are the common AR parameters estimate (computed by solving the previous normal equations).

If the whitening filter is estimated using the source correlations, $\widetilde{s}(n) = a_{s,M}(q)s(n) = \frac{\widehat{a}_{s,M}(q)}{a_s(q)}u_s(n)$ will be perfectly white if the AR order goes to infinity. However, as we use a noisy correlation, infinite order is no longer optimal. The optimal whitening order should be choosing as that minimizing the mean of the prediction error variance $\sigma^2_{\widetilde{s}} = E\left\{\widetilde{s}(n)^2\right\}$, i.e.,

$$\hat{l} \; = \; \arg\min_l \sigma^2_{\widetilde{s}}(l) \tag{4.25}$$

On the other hand, the averaged received correlations $r_{y,M}(t)$ can be written as a function of the source correlations $r_s(t)$ and the averaged channel correlations $r_{h,M}(t) = \frac{1}{M}\sum_{m=1}^{M} r_{h_m h_m}(t)$, i.e.,

$$r_{y,M}(t) = r_s(t) * r_{h,M}(t) \tag{4.26}$$

By decomposing the averaged channel correlation into a deterministic and a zero-mean random processes, we have:

$$r_{y,M}(t) = c_0 \left(r_s(t) + \frac{c_1}{\sqrt{M}} \underbrace{r_s * r_{\tilde{h},M}(t)}_{r_{e,M}(t)} \right) \tag{4.27}$$

where $c_0 = \frac{1-\beta}{\pi A \beta}$, and $r_{\tilde{h},M}(t)$ is a zero-mean random process.

If we assume that $\frac{c_1}{\sqrt{(M)}} \ll 1$, using second-order approximation one can show that

$$\begin{aligned} \widehat{\mathbf{a}}_{s,M} \approx \; & \mathbf{R}_s^{-1}\mathbf{p}_s + \frac{c_1}{\sqrt{M}}\left(\mathbf{R}_s^{-1}\mathbf{p}_{e,M} - \mathbf{R}_s^{-1}\mathbf{R}_{e,M}\mathbf{R}_s^{-1}\mathbf{p}_s\right) \\ & + \frac{c_1^2}{M}\left(\mathbf{R}_s^{-1}\mathbf{R}_{e,M}\mathbf{R}_s^{-1}\mathbf{R}_{e,M}\mathbf{R}_s^{-1}\mathbf{p}_s - \mathbf{R}_s^{-1}\mathbf{R}_{e,M}\mathbf{R}_s^{-1}\mathbf{p}_{e,M}\right) \end{aligned} \tag{4.28}$$

where $\mathbf{R}_s$, and $\mathbf{R}_{e,M}$ (resp. $\mathbf{p}_s$, $\mathbf{p}_{e,M}$) have the same structure as $\mathbf{R}_{y,M}$ (resp. $\mathbf{p}_{y,M}$), in which $r_{y,M}(t)$ is replaced by $r_s(t)$ and $r_{e,M}(t)$. Then, we use the predictor $\widehat{\mathbf{a}}_{s,M}$ (performed using the noisy source correlation $r_{y,M}(t)$) to whiten the speech source. The prediction error variance is given by:

$$\begin{aligned}\sigma_{\tilde{s}}^2(l) = & \ \sigma_s^2 - \widehat{\mathbf{a}}_{s,M}^H \mathbf{p}_s - \mathbf{p}_s^H \widehat{\mathbf{a}}_{s,M} + \widehat{\mathbf{a}}_{s,M}^H \mathbf{R}_s \widehat{\mathbf{a}}_{s,M} \\ = & \ \sigma_s^2 - \mathbf{p}_s^H \mathbf{R}_s^{-1} \mathbf{p}_s \\ & + \frac{c_1^2}{M} \left(\mathbf{p}_{e,M} - \mathbf{R}_{e,M} \mathbf{R}_s^{-1} \mathbf{p}_s\right)^H \mathbf{R}_s^{-1} \left(\mathbf{p}_{e,M} - \mathbf{R}_{e,M} \mathbf{R}_s^{-1} \mathbf{p}_s\right)\end{aligned} \tag{4.29}$$

We observe that the prediction error variance can be decomposed into two terms:

- A deterministic term $\sigma_s^2 - \mathbf{p}_s^H \mathbf{R}_s^{-1} \mathbf{p}_s$ representing the error due to the use of finite order filter predictor.
- A stochastic term

  $$\frac{c_1^2}{M} \left(\mathbf{R}_{e,M}^{-1} \mathbf{p}_{e,M} - \mathbf{R}_s^{-1} \mathbf{p}_s\right)^H \mathbf{R}_{e,M} \mathbf{R}_s^{-1} \mathbf{R}_{e,M} \left(\mathbf{R}_{e,M}^{-1} \mathbf{p}_{e,M} - \mathbf{R}_s^{-1} \mathbf{p}_s\right)$$

  representing the error due to the use of noisy correlations $r_{y,M}(t)$ (instead of the source correlations $r_s(t)$) to estimate to source color. Note that this term increases with the AR order, and is inversely proportional to the number of microphones.

The whitening order should be optimized to give the best tradeoff between these two terms.

4.5.2 Whitening order determination

Stochastic whitening order estimation

As we can see from (4.29), $\sigma_{\tilde{s}}^2$ depends on the channel realization (via $\mathbf{p}_{e,M}$ and $\mathbf{R}_{e,M}$). These information are not available (our goal is to perform blind equalization). Thus, we propose relaxing the cost function in (4.25), and computing the prediction order that minimize the spatially averaged prediction error variance, i.e.,

$$\hat{l} = \arg\min_{l} \left\langle \sigma_{\tilde{s}}^2(l) \right\rangle \tag{4.30}$$

In such a way, we select a whitening order optimal in the average (over source and microphones positions), but not necessarily for the given channel realization. Note also that the (4.30) depends on the room statistics (function of the reverberation time, room volume...), but no-longer on the channel realization. Knowing the source correlations and the statistics of the room impulse response, one can have an analytical expression of $\langle \sigma^2_{\tilde{s}}(l) \rangle$. However, this analytical expression is very complex to derive and to implement (even using second order approximations). So that, we propose computing the spatial averaging using a Monte-Carlo approach:

1. We generate Gaussian random channels $\mathbf{h}(t)$ using (4.10) (having the same statistics as the room impulse responses)

2. We compute $\sigma^2_{\tilde{s}}(l)$ using the random channels $\{r_h(t)\}_{t=1:l}$.

3. We average $\langle \sigma^2_{\tilde{s}}(l) \rangle$ over the random channel realizations.

Remark that $\langle \sigma^2_{\tilde{s}}(l) \rangle$ still dependent on the unknown source correlations (averaged over a given period of time). However, the correlation details are not relevant, only the shape of the speech correlations is important. So that, we propose compute (4.30) using a priori speech correlation estimate $\overline{r}_s(t)$ (averaged over a long period of time, speakers ...). Fig. 4.13 subplots the curves of the averaged prediction error $\langle \sigma^2_{\tilde{s}}(l) \rangle$ function of the whitening order for 2 and 4 microphones. As it was expected from (4.29), the optimal whitening order for 4 microphones is higher than the one for 2 microphones. We also remark that the optimization results are coherent with the dereverberation results (Fig. 4.12).

Deterministic whitening order estimation

The order selected in the previous section is optimal in the average (over all possible channel realizations), but not necessarily for the given source and microphones position. In this section, we reconsider the *blind* AR order selection for a given channel realization (solving (4.25)). To solve this problem, we propose looking to the AR modeling problem from a different point of view: the source correlations are considered as noisy version of the received signal correlations (corrupted by the channel correlation inverse). On the other hand, for large enough whitening order (such that the covariance matrices $\mathbf{R}_s$ and $\mathbf{R}_{\widetilde{h},M}$ are almost band), we have:

$$\begin{aligned} \mathbf{R}_{y,M} &\approx \mathbf{R}_{\widetilde{h},M}\mathbf{R}_s \\ \mathbf{p}_{y,M} &\approx \mathbf{R}_{\widetilde{h},M}\mathbf{p}_s \end{aligned} \tag{4.31}$$

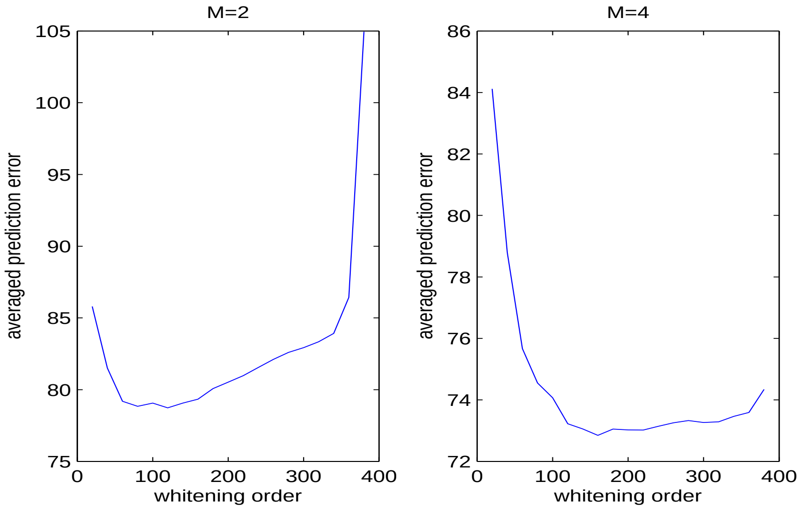

Figure 4.13: The averaged prediction error variance $\langle \sigma_{\tilde{s}}^2 \rangle$ function of the whitening order for 2 and 4 microphones.

Using these approximations,the prediction error variance becomes

$$\sigma_{\tilde{s}}^2(l) = \left[\, 1 \ \widehat{\mathbf{a}}_M^H \,\right] \mathbf{R}_s \begin{bmatrix} 1 \\ \widehat{\mathbf{a}}_M \end{bmatrix} \approx \left[\, 1 \ \widehat{\mathbf{a}}_M^H \,\right] \mathbf{R}_{h,M}^{-1} \mathbf{R}_{y,M} \begin{bmatrix} 1 \\ \widehat{\mathbf{a}}_M \end{bmatrix} \tag{4.32}$$

where $\mathbf{R}_s$, $\mathbf{R}_{y,M}$, and $\mathbf{R}_{h,M}$ are $(l+1)\times(l+1)$ matrices defined as in previous. Finally, the unknown matrix $\mathbf{R}_{h,M}^{-1}$ is replaced by its spatial average $\langle \mathbf{R}_{h,M}^{-1} \rangle$:

$$\sigma_{\tilde{s}}^2(l) = \left[\ 1 \ \widehat{\mathbf{a}}_M^H \ \right] \langle \mathbf{R}_{h,M}^{-1} \rangle \, \mathbf{R}_{y,M} \begin{bmatrix} 1 \\ \widehat{\mathbf{a}}_M \end{bmatrix} \tag{4.33}$$

Once again, the expectation $\langle \mathbf{R}_{h,M}^{-1} \rangle$ is computed using Monte-Carlo method. Fig. 4.14 subplots the curves of the prediction error $\sigma_{\tilde{s}}^2$ computed by (4.32) (using the source covariance matrix) or "blindly" by (4.33). We remark that the minima in the two curves match well; and that (4.33) can be used to select the whitening order. However, the approximation in (4.31) is valid only for large order ($l \geq 100$). Thus, it can happen that one sees some minima for $l < 100$. Those minima should be ignored. Another drawback of this approach is due to local minima. To alleviate this problem, we propose using stochastic whitening order selection to situate approximately the optimal

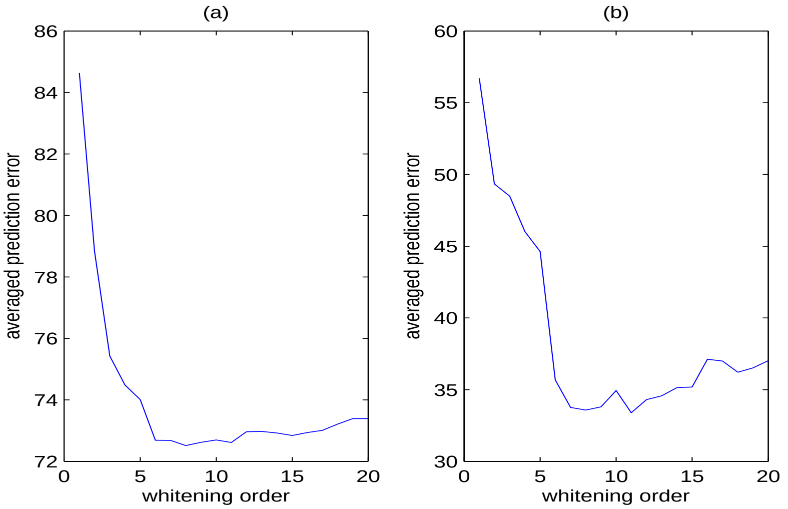

Figure 4.14: The prediction error variance $\sigma_{\tilde{s}}^2$ computed using the source covariance matrix (a), or using (4.33) (b).

order. Then deterministic whitening order selection is computed to select the AR order for the given channel realization.

4.6 Speech dereverberation in noisy environment

The dereverberation algorithms are generally introduced in a noiseless environment (the problem is still very difficult even in this ideal case). However for practical applications, the robustness of these algorithms to the presence of additive noise is required. As we have seen in section (4.4), the multivariable linear prediction estimates blindly a zero-forcing equalizer ($\mathbf{F}_{D\&P}$) for SIMO channels. The equalizer depends only on the reverberant signal second order statistics. Thus, the proposed approach can be easily extended in the presence of an additive white noise, since the white noise variance can be easily identified and compensated for in the reverberant signal covariance matrix. However, the presence of the additive noise has so far not been considered for the design of the ZF equalizer, and the resulting equalizer is no-longer optimal.

In this section, we present two issues in the design of the LP-based equalizer in order to increase the robustness of the scheme in the presence of additive white noise:

- First, we investigate the effect of relative subchannel delay compensation on the output SNR. We show that such relative delay can reduce considerably the output SNR.
- Then, we optimize the transformation of the multivariate prediction filter to a longer equalizer using the SNR criterion. The optimization corresponds to MMSE-ZF design, and the filter length increase allows for the introduction of some equalization delay, that can also be optimized.

This section does not focus the effect of the speech signal correlations, only equalization accuracy is considered. Thus within this section (except indication of the contrary), the input signal $s(n)$ refers to a white process.

4.6.1 Time delay compensation for LP equalization

Time delay compensation for SIMO dereverberation

Several authors point the lack of robustness of the LP equalizer in presence of additive noise. In particular, the algorithm overall performance rely on the particular realization of the multichannel precursor coefficient $\mathbf{h}_0$, yielding a prediction error signal with uncontrollable symbol-to-noise ratio [59]. In [107], Li et al. remark that some problems may arise when $\mathbf{h}_0$ have small entries. In [59], Gesbert and Duhamel use several multistep linear prediction to triangularize the multichannel system. In such a way, the proposed prediction scheme exploits the full channel structure. Thus, it provides more statistical efficiency in channel identification. In this section, we suggest alleviating this side effect by aligning the received signals on the various microphones (delay compensation for direct path). We demonstrate that it leads not only to an increase in the signal part energy $\left(\sigma_s^2 \left\|\mathbf{h}_0\right\|^2\right)$, but also to a decrease on the output MSE $= \left(\sigma_v^2 \text{ tr } \left\{A_x A_x^T\right\}\right)$.

Theorem 1: For a noisy SIMO dereverberation problem, the output SNR increases by relative subchannel delay compensation.

Proof:

see appendix 4.A.

To illustrate the effect of the data alignment on the SNR of the LP output, we consider the reverberation scenario described in section 4.5. A white noise is used as the source signal (sampled at 8 Khz).
The Matched Filter Bound (MFB) is defined as

$$\text{MFB} = \frac{\sigma_s^2}{\sigma_v^2} \|H\|^2 \tag{4.34}$$

The MFB is be also called "channel SNR" [55]. The MFB can be interpreted as the SNR of the maximum likelihood estimation of the input $s(n)$ assuming that all other inputs $s(n) \quad k \neq n$ are known [166]. It is clear that the MFB constitute an upper bound on the output SNR. Furthermore, we consider the evaluation criterion:

$$\frac{MFB}{SNR_{out}} \geq 1 \tag{4.35}$$

Note that for any zero-forcing equalizer (particularly D-&-P), this criterion do not depend on $\frac{\sigma_s^2}{\sigma_v^2}$.
Figure 4.15 compares the performance of the LP algorithms with and without relative time-delay compensation (averaged over 100 Monte Carlo runs). One can remark that the alignment of the received signals increases the robustness of the algorithm to additive noise, specially when the number of subchannels increases.
Taking into consideration the relative subchannel time delay compensation, the spatiotemporal zero-forcing equalizer becomes

$$\mathbf{F_{D\&P}}(q) = \mathbf{h}_0^H A_{L_p}(q) D(q) \tag{4.36}$$

where $D(q)$ is a diagonal matrix of delays aligning the direct path contributions in the M reverberant signal.
We called the dereverberation scheme “Delay-and-Predict (D-&-P) equalizer” as opposed to “Delay-and-Sum (D-&-S) beamformer”. Remark that the D-&-S beamformer is a special case of the D-&-P equalizer (where the multichannel linear prediction order $L_A = 0$).

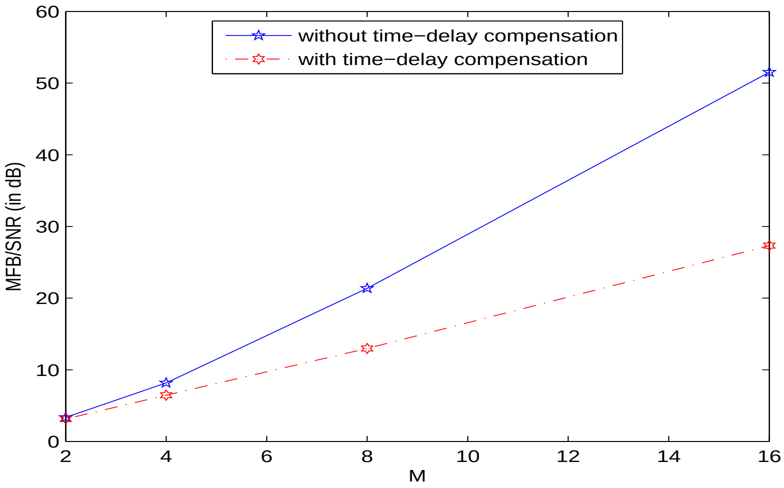

Figure 4.15: $\frac{MFB}{SNR_{out}}$with and without relative time-delay compensation.

Interpretation in terms of Generalized Sidelobe Canceller (GSC)

Multivariable linear prediction based equalizer can be interpreted as a particular generalized sidelobe canceller (figure 4.16). Assuming the precursor coefficient $\mathbf{h}_0$ is known, the desired response signal $d(n)$ is performed by a spatial matched filtering. Taking into consideration the input whiteness, the noise reference signal is computed using time-delay operation.
One can easily show that the noise reduction causal Winner filter can be expressed as

$$W_{D\&P}(q) = \mathbf{h}_0^H(\mathbf{I}_M - A_x(q)) \tag{4.37}$$

where $A_x(q)$ is a multivariable linear predictor assumed to be long enough to equalize the channel (i.e. $A_x^{-1}H(q) = \mathbf{h}_0$).
Thus, the multivariable linear prediction based equalizer and the previously described generalized sidelobe canceller coincide:

$$\begin{aligned} \widehat{s}(n) &= \mathbf{h}_0^H\mathbf{x}(n) - W_{D\&P}(q)\mathbf{x}(n-1) \\ &= \mathbf{h}_0^H A_x(q)\mathbf{x}(n) \\ &= F_{D\&P}(q)\mathbf{x}(n) \end{aligned} \tag{4.38}$$

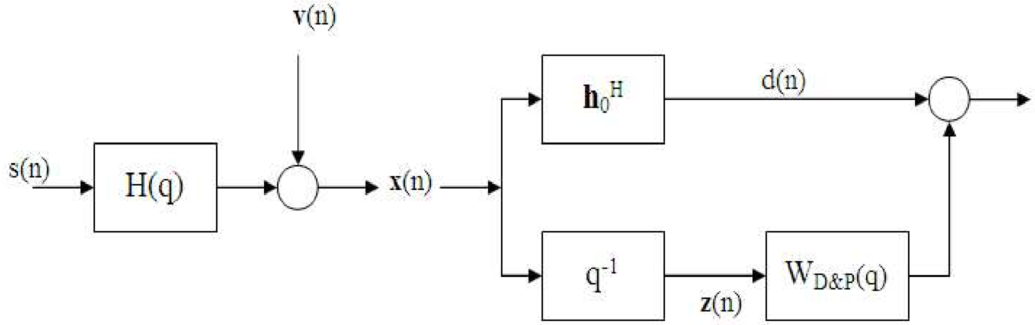

Figure 4.16: GSC interpretation of the LP based equalization.

Interpreting the D&P as a GSC provides some intuitions on the effect of the relative delay compensation on the dereverberation accuracy. In fact, using relative time delay compensation leads to an optimally weighted spatial filtering, which leads to better noise reduction (in the direct branch), and better enhancement accuracy of the overall scheme.

Next, we comment the difference between the multivariable linear prediction based equalization and the classic generalized sidelobe cancellation. The classic GSC scheme exploits spatial prior information to compute the noise reference $\mathbf{z}(n)$ instead of using statistical prior (input whiteness). The blocking channel $\mathbf{h}_0^{\perp H}$ removes the contribution of $s(n)$ and yield to the noise reference signal $\mathbf{z}(n)$. Then, a causal noise canceller $W_{GSC}(q)$ is applied to eliminate the stationary noise that leaks through the sidelobes of the fixed beamformer $\mathbf{h}_0^H$
The major drawback of such scheme is that it leads to a reduction of the spatial dimension of the SIMO problem . For instance, in two microphone array configuration, we show, in appendix 4.C, that if $\mathbf{h}_0^{\perp H} qH(q)$ is minimum phase, the noise canceller $W_{GSC}(q)$ has an infinite length, and can be expressed as

$$W_{GSC}(q) = \frac{\mathbf{h}_0^H \left(A_x^{-1}(q) - \mathbf{I}\right) \mathbf{h}_0}{\mathbf{h}_0^{\perp H} A_x^{-1}(q) \mathbf{h}_0} = \frac{\mathbf{h}_0^H H(q) - \|\mathbf{h}_0\|^2}{\mathbf{h}_0^{\perp H} H(q)} \tag{4.39}$$

We show also that, if $\mathbf{h}_0^{\perp H} qH(q)$ is non-minimum phase, perfect dereverberation is no-longer possible using a GSC scheme. And even when perfect dere-

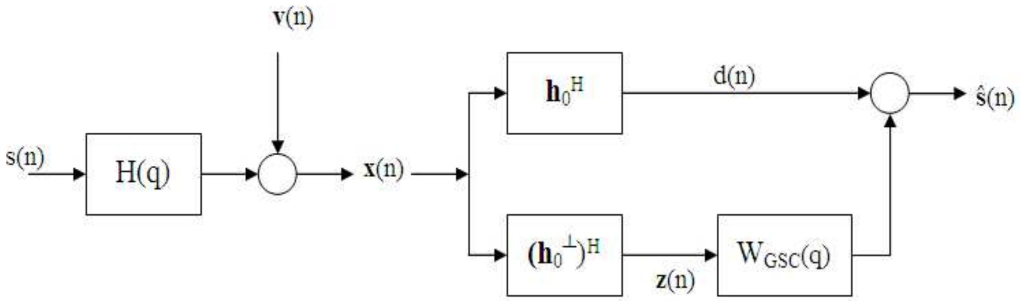

Figure 4.17: Generalized Sidelobe Cancellation scheme.

verberation is possible, it can not be performed using FIR filters. Contrary to GSC scheme, in the LP based approach, the multi-channel delay-spread diversity enables FIR perfect dereverberation.

Time delay estimation in multipath propagation environment

Time Delay Estimation (TDE) is a classic signal processing problem. In its simplest form, a signal is emitted from a source, and arrives with additive noise at two (or several) spatially separated sensors with different delays and attenuations, i.e.

$$\begin{array}{l} x_1(n) = s(n) + v_1(n) \\ x_2(n) = \alpha s(n - \tau^o) + v_2(n) \end{array} \tag{4.40}$$

In spite of its simple structure, several approaches based on quite different points of view have been proposed and studied to solve the problem [207]. The classical methods for TDE are based on cross-correlation (CC) and generalized cross-correlation (GCC) functions [96].

Assuming the signal $s(n)$ and noises $(v_1(n), v_2(n))$ are mutually independent processes, the cross correlation function between the received signals is given by:

$$\begin{aligned} R_{12}(\tau) &= E[x_1(n)x_2(n+\tau)] \\ &= \int_f X_1(f)X_2^H(f)e^{j2\pi f\tau}df \\ &= \alpha R_{ss}\left(\tau - \tau^o\right) \end{aligned} \tag{4.41}$$

where $X_1(f)$ and $X_2(f)$ denotes respectively the Fourier transform of the processes $x_1(n)$, and $x_2(n)$; and $(.)^H$ represents the transpose conjugation. It is clear that the delay τ^o can be estimated by locating the peak of $R_{12}(\tau)$. Typically, a parabolic fit is performed about the peak in $R_{12}(\tau)$ to achieve sub-sample resolution.

In reality, multi-path propagation can cause significant time delay estimator bias and ambiguities which can not be solved by the temporal CC method alone.

The generalized cross-correlation method extends the previous technique by introducing a weighting function, $W(f)$:

$$R_{12}(\tau) = \int_f W(f) X_1(f) X_2^H(f) e^{j2\pi f\tau} df \tag{4.42}$$

There exist many publications investigating the design and the effect of this weighting function, but still insufficient to solve the bias introduced by multi-path propagation [99]. Moreover, in all cases cross-correlation based techniques align the most powerful delays. However, due to the multi-path propagation, it does not correspond generally to the first path alignment. To face this problem, we reduce the multi-path propagation effect using multistep multichannel linear prediction. Next, we apply the cross-correlation techniques on the multi-steps LP residual signals (rather on the received signal).

Consider now the problem of predicting $\mathbf{x}(n)$ from the L_B observations $[\mathbf{x}^T(n-\lambda) \cdots \mathbf{x}^T(n-\lambda+L_B)]^T$ ($\lambda \geq 1$ is the prediction step). The prediction error signal is given by:

$$\tilde{\mathbf{x}}(n) = \mathbf{x}(n) + \sum_{i=0}^{L_B-1} \mathbf{B}_{x,i} \mathbf{x}(n-\lambda-i) \tag{4.43}$$

where $\{B_{x,i}\}_i$ are the matrices of the multistep multichannel LP coefficients, and L_B denotes the order of the multi-steps linear prediction.

Using the same reasoning as in the section 4.4.2, one can show that, using a long enough multichannel LP filter $\left(L_B \geq \frac{L_h-\lambda}{M-1}\right)$, the prediction error becomes

$$\tilde{\mathbf{x}}(n) = \mathbf{x}_{inov}(n) \approx \sum_{i=1}^{\lambda-1} \mathbf{h}_i \widetilde{s}(n-i) \tag{4.44}$$

Thus, the multi-step forward linear prediction "shorten" the impulse response $\{\mathbf{h}_i\}_i$, and removes all correlations at time-lag $\tau > \lambda$.

Thus, if there exists a time-lag λ_0 such that the direct paths on all channels are situated before and all reflections after (figure 4.18), the multi-path propagation problem can be alleviated by considering the multistep multichannel linear prediction with delay $\lambda = \lambda_0$. In such a case, the residual signal contains all/only the contributions of the first paths. Then, applying cross-correlation on the LP residual signal allows the alignment of the received signals.

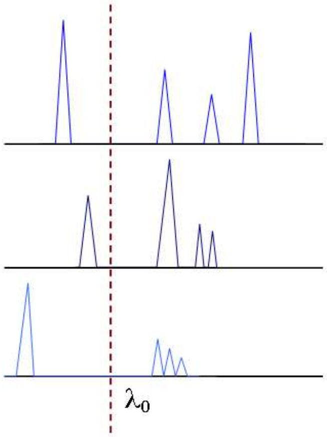

Figure 4.18: A case where direct paths and reverberation are separable.

However, if the microphones are not too close, some early reflections can arrive on some channels before direct paths on some other channels. In such a case, it will be impossible to find λ_0. Therefore, if λ_0 does not exist or if a prior information is not available, we propose an iterative scheme to align the received signals:

1. perform multichannel LP (multi-step LP with a time-lag $\lambda = 1$)

2. compute $\widehat{\mathbf{h}}_0$, which can be estimated as the eigenvector corresponding to the maximum eigenvalue of the LP residual correlation matrix

3. detect the positions of non-zero coefficients in $\widehat{\mathbf{h}}_0$, and delay the corresponding received signals by 1

4. repeat, until all received signals are aligned.

To illustrate the proposed scheme, we will take a simple example. Let us consider the channel impulse response:

$$\mathbf{h} = \begin{bmatrix} 0\cdots 0 \; 0.9 \; 1.1 \; 0\cdots 0 \\ 0\cdots 0 \;\; 0 \;\; 1 \;\; 0\cdots 0 \end{bmatrix}$$

From the cross-correlation point of view, the two subchannels are well aligned. However, direct paths are not. On the other hand, the eigenvector corresponding to the maximum eigenvalue of the LP residual correlation matrix estimates (up a scalar) $\widehat{\mathbf{h}}_0 \propto \begin{bmatrix} 0.9 \\ 0 \end{bmatrix}$.

According to $\widehat{\mathbf{h}}_0$, we should delay the first subchannel:

$$\begin{aligned} \begin{bmatrix} q^{-1}x_1(n) \\ x_2(n) \end{bmatrix} &= \begin{bmatrix} q^{-1}bfh_1(n) \\ bfh_2(n) \end{bmatrix} * \widetilde{s}(n) \\ &= \begin{bmatrix} 0\cdots 0 \; 0 \; 0.9 \; 1.1 \; 0\cdots 0 \\ 0\cdots 0 \; 0 \;\; 1 \;\; 0 \;\; 0\cdots 0 \end{bmatrix} * \widetilde{s}(n) \end{aligned}$$

Finally, by re-estimating $\widehat{\mathbf{h}}_0 \propto \begin{bmatrix} 0.9 \\ 1 \end{bmatrix}$, we observe that the two channels are aligned.

The major drawback of the proposed scheme is that the alignment resolution is equal to the sampling period. To increase the delay resolution, this procedure can be followed by a CC based refinement step, possibly using multichannel LP residuals.

4.6.2 MMSE-ZF LP postfiltering for blind multichannel equalization

The output of the multichannel linear predictor is

$$\mathbf{x}(n) = \mathbf{h_0}s(n) + A_x(q)\mathbf{v}(n) \tag{4.45}$$

In original LP equalizer, the columns of the predictor $A_x(q) = \mathbf{I}+\sum_{i=1}\mathbf{A}_{x,i}q^{-i}$ are combined using the weighing vector $\mathbf{h}_0^H$, i.e.,

$$F_{LP}(q) = \mathbf{h}_0^H A_x(q) \tag{4.46}$$

This choice maximizes the power of the signal part but not necessarily the output SNR. In[55], Gazzah computes the weighing vector by maximizing the output SNR, i.e.,

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\sigma_s^2}{\sigma_v^2} \frac{\|\mathbf{w}\|^2}{\mathbf{w}\mathbf{A}_x\mathbf{A}_x^H\mathbf{w}^H} \tag{4.47}$$

The proposed equalizer is:

$$F_{MLP}(q) = \mathbf{h}_0^H(\mathbf{A}_x\mathbf{A}_x^H)^{-1}A_x(q) \tag{4.48}$$

where $\mathbf{A}_x = [\mathbf{I}_M\mathbf{A}_{x,1}\cdots\mathbf{A}_{x,L_A}]$. The author shows that the proposed equalizer output not only outperforms the original LP equalizer, but also attains the lowest achievable (by any no-delay ZF equalizer) MSE.

In the following, we generalize the previous approach by considering a weighting filters to combines the columns of the $A_x(q)$. This will allow the design of non-zero-delay ZF equalizer. For a given length filter L_w, and an equalization delay $d \leq (L_w - 1)$ The weighting filter are optimized by maximizing the output SNR, under the d-delay zero-forcing constraint, .i.e.

$$\begin{cases} \mathbf{w} = \arg\max_{\mathbf{w}} \dfrac{\sigma_s^2}{\sigma_v^2} \dfrac{1}{\dfrac{1}{2\pi j}\displaystyle\oint \mathbf{w}(q)A_x(q)A_x^{\dagger}(q)\mathbf{w}^{\dagger}(q)\dfrac{dz}{z}} \\ \mathbf{w}(q).\mathbf{h_0} = q^{-d} \end{cases} \tag{4.49}$$

where $A_x^{\dagger}(q) = \sum_{i=1}^{L_A} A_{x,i}q^i$ denotes the $A_x(q)$ matched filter.
To solve the optimization problem, it is easier to first form the $(L_w.M) \times ((L_w{+}L{-}1).M)$ and $(L_w.M) \times L_w$ block Toeplitz matrices

$$\mathcal{A} = \begin{bmatrix} \mathbf{I}_M & \mathbf{A}_{x,1} & \cdots & \mathbf{A}_{x,L_A} & 0 & \cdots & 0 \\ 0 & \mathbf{I}_M & \mathbf{A}_{x,1} & \cdots & \mathbf{A}_{x,L_A} & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & & \\ \vdots & & \ddots & \ddots & & \ddots & \\ 0 & \cdots & 0 & \mathbf{I}_M & \mathbf{A}_{x,1} & \cdots & \mathbf{A}_{x,L_A} \end{bmatrix}$$

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{h_0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{h_0} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{h_0} \end{bmatrix}$$

The optimization in (4.49) becomes

$$\begin{cases} \mathbf{W}_{L_w,d} = \arg\min_{W} \mathbf{W}\mathbf{R}_w\mathbf{W}^H \\ \mathbf{W}_{L_w,d}\mathbf{H}_0 = \mathbf{e}_d \end{cases} \tag{4.50}$$

where $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_{L_w}]$ is a $(M.L_w)$ vector characterizing the weighting filter coefficients $\left(\mathbf{w}(q) = \sum_{i=1}^{L_w} \mathbf{w}_i q^{-i}\right)$, $\mathbf{R}_w = \sigma_v^2 \mathcal{A}\mathcal{A}^H$ represents the output noise covariance matrix, and $\mathbf{e}_d = [0...0\ 1\ 0...0]$ is the $(d+1)^{th}$ vector of the $I\!R^{L_w}$ canonical basis. Using Lagrange optimization, on can show that the optimal weighting filter is given by

$$\mathbf{W}_{L_w,d} = \mathbf{e}_d \left(\mathbf{H}_0^H \mathbf{R}_w^{-1} \mathbf{H}_0\right)^{-1} \mathbf{H}_0^H \mathbf{R}_w^{-1} \tag{4.51}$$

The achieved output MSE is

$$\mathrm{MSE} = \sigma_v^2 \mathbf{e}_d \left(\mathbf{H}_0^H \mathbf{R}_w^{-1} \mathbf{H}_0\right)^{-1} \mathbf{e}_d^H \tag{4.52}$$

Note that the delay $d \geq 0$ can be easily optimized by minimizing the output MSE. The optimal delay correspond to the largest diagonal element of $\mathbf{e}_d \left(\mathbf{H}_0^H \mathbf{R}_w^{-1} \mathbf{H}_0\right)^{-1}$.

Special cases:

- For $L_w = 1,\ d = 0$, we recover the solution proposed in [55], i.e.,

$$W_{1,0} \propto \mathbf{h_0} \left(\mathbf{A}\mathbf{A}^H\right)^{-1} \tag{4.53}$$

- If $L_w \to \infty$, and for an appropriate choice of the delay d_∞, one can show that

$$\mathbf{w}_{\infty,d_\infty}(q) = \left(\mathbf{h}_0^H A^{-\dagger}(q) A^{-1}(q) \mathbf{h}_0\right)^{-1} \mathbf{h}_0^H A^{-\dagger}(q) A^{-1}(q) \tag{4.54}$$

Exploiting the fact that $A^{-1}(q)\mathbf{h}_0 = H(q)$, one can show that the obtained ZF equalizer corresponds to the MMSE-ZF equalizer:

$$\begin{aligned} F_{\infty,d_\infty}(q) &= \mathbf{w}_{\infty,d_\infty}(q) A(q) \\ &= \left(H^\dagger(q) H(q)\right)^{-1} H^\dagger(q) \end{aligned} \tag{4.55}$$

Next, we illustrate the behavior of the proposed scheme, and we provide a comparison with the scheme proposed in[55] and the classic MMSE-ZF equalizer. Monte-Carlo simulations are constructed using the reverberation scenario described in section 4.5. A white noise is used as the source signal.

Figure 4.19 compares the performance of the different ZF equalizer (averaged over 10 Monte Carlo runs). We verify that if we consider zero delay equalization, increasing the order of weighting filter do not increase the performance; which is coherent with the results reported in[55]. On the other hand, despite achieving the MMSE-ZF equalization performance requires long filters and large delays (due to the acoustic channel length); considerable gains can be achieved by allowing even small delays (7.5 dB using 9 taps weighting filters (M=8)).

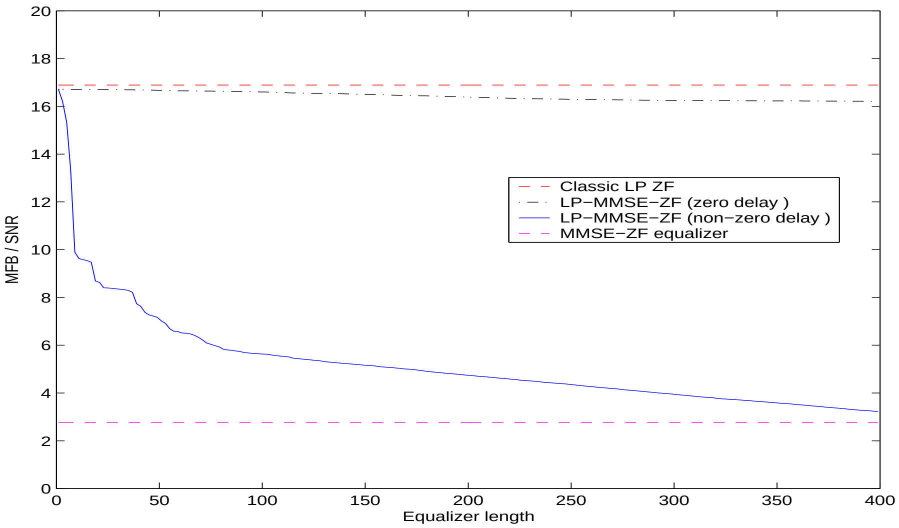

Figure 4.19: $\dfrac{MFB}{SNR_{out}}$ for different ZF equalization scheme(averaged over 20 Monte Carlo runs).

Then, we investigate the performance of the proposed scheme function of the number of sub-channels M (figure 4.20). Curves show that the gain, due to the use of non-zero delay equalization, increases with M. The reason is: the more sub-channel we have, the more freedom degrees (in the weighting filters) we can optimize, and the better output SNR we achieve.

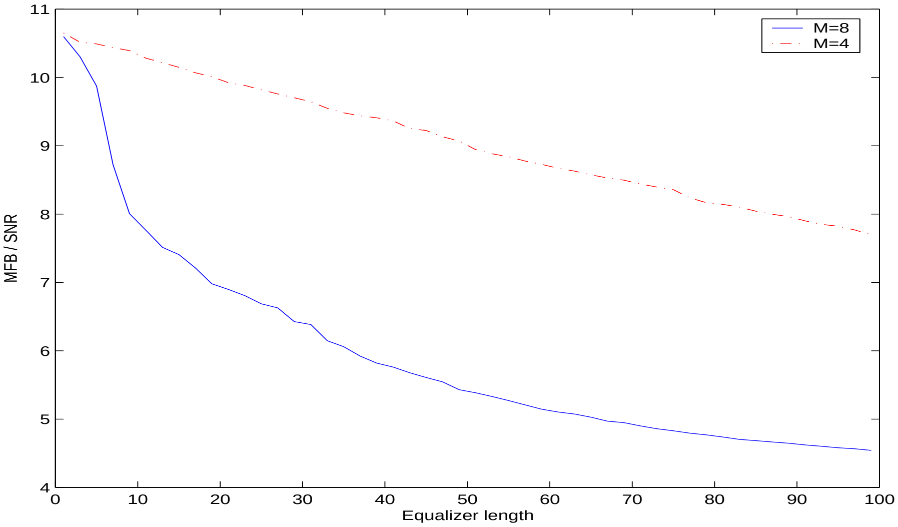

Figure 4.20: $\dfrac{MFB}{SNR_{out}}$ for non-zero delay ZF equalization function of the number of sub-channels.

We next illustrate the behaviour of the proposed scheme applied to speech dereverberation, and we provide a comparison with the classic Delay-&-Predict equalizer. A speech signal with duration of 8.8s, and sampled at 8 kHz is used as the original source signal. The reverberant speech signal is observed on 2 distinct microphones. The post-filter length (then the equalization delay) is constrained to be $L_w \leq 100$. Figure 4.21 plots the Signal-to-Echo+Noise Ratio ($SENR = \dfrac{\sum_k s(k)^2}{\sum_k (s(k) - \widehat{s}(k))^2}$) as function of the input Signal-to-Noise Ratio ($SNR = \dfrac{\sum_k (y(k) - v(k))^2}{\sum_k v(k)^2}$). Curves show that, in all regions, the Robust D-&-P performs better than both the classic D-&-P and D-&-S. Particularly in noisy environment, the post-filtering becomes essential in order to have acceptable enhancement accuracy. On the other hand, one can also remark that the post-processing still has a positive effect even in absence of ambient noise (SNR=60 dB). The reason is that the post-filtering compensates also for the errors due to the estimation of the clean spectrum (the estimation is done by averaging only two observation spectra).

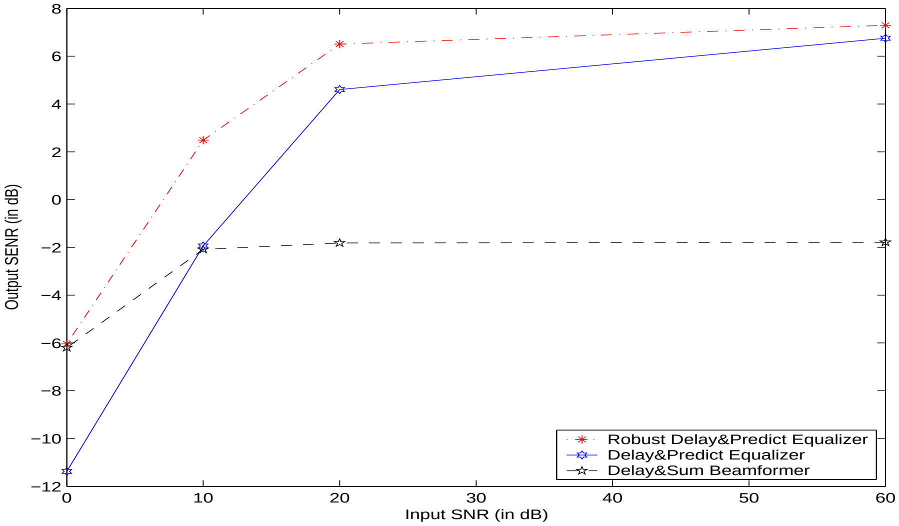

Figure 4.21: The SENR function of the input SNR.

Robust Delay-&-Predict Equalization under channel length underestimation

Ambient noise is not the unique origin of the additive noise. In fact, acoustic reverberation is theoretically infinite. As we assume that the channel has a finite length L_h, the late reverberation will be considered as additive noise, i.e.,

$$\mathbf{y}(k) = \sum_{i=0}^{L_h-1} \mathbf{h}_i s(k-i) + \underbrace{\sum_{i=L_h}^{\infty} \mathbf{h}_i s(k-i)}_{\mathbf{v}(k)} . \tag{4.56}$$

Classically the channel length is chosen long enough such that the energy of the additive noise is negligible (typically $L_h \geq T_{60} f_s$). With such choice, the acoustic channels may have considerable lengths in real propagation environments. Hence, the algorithm may be computationally very expensive. In this section, we investigate the effect on the dereverberation performance of the underestimation of the reverberation response.

We model the late reverberation as a spherically diffuse noise [101] (although strictly specking this additive noise (late reverberation) is neither white nor independent from the reverberated signal). Then, we apply the post-filtering designed in the previous section to reduce the noise effect. We consider the Direct to Reverberant energy Ratio (DRR) as an evaluation criterion for the dereverberation accuracy:

$$DRR = 10\log_{10}\left\{\frac{\sum_{t=0}^{\tau-1}\widetilde{h}^2(t)}{\sum_{t=\tau}^{L-1}\widetilde{h}^2(t)}\right\} \quad dB \tag{4.57}$$

where $\widetilde{h}^2(t) = \mathbf{h} * \mathbf{f}(t) = \sum_i \mathbf{h}_i \mathbf{f}_{t-i}$ denotes the equalized channel (with a given equalizer $\mathbf{f}(q)$), and τ is the number of samples to include as the direct component. The choice of the parameter τ depends on the application (how much early and late reverberation is annoying in the given application). By increasing the value of τ, we give more weight to the degradation due to the late reverberation. If τ is small ($\tau \leq 1\ ms$), the DRR criterion will be correlated with the dereverberation SENR (equal if the input is white). Figures 4.22 and 4.23 plots the curves of the output DRR of the classic, robust Delay-&-Predict equalizers, and the Delay-&-Sum beamformer (function of the assumed channel length), respectively using 2 and 4 microphone array setup (for $\tau = 10\ ms$ and $\tau = 1\ ms$).

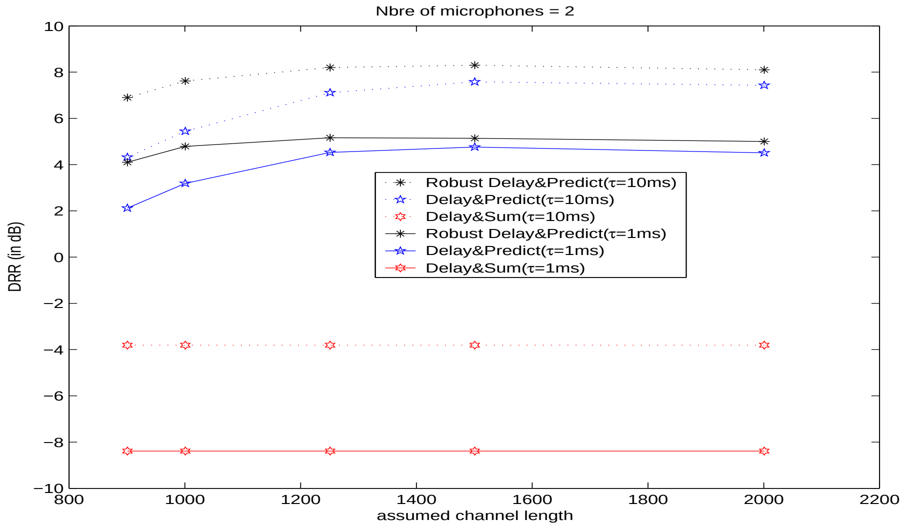

Figure 4.22: The output DRR function of the assumed channel length, using 2 microphone array setup ($\tau = 10\ ms$ and $\tau = 1\ ms$)).

In these simulations, the channel length is finite ($L_h = 2000$). One can remark that the robust D-&-P outperforms the classic D-&-P in terms of dereverberation accuracy. Then, it is more robust to the channel length underestimation. In all cases, the two schemes (classic and robust D-&-P) outperform the D-&-S beamformer. Remark also that even when the channel

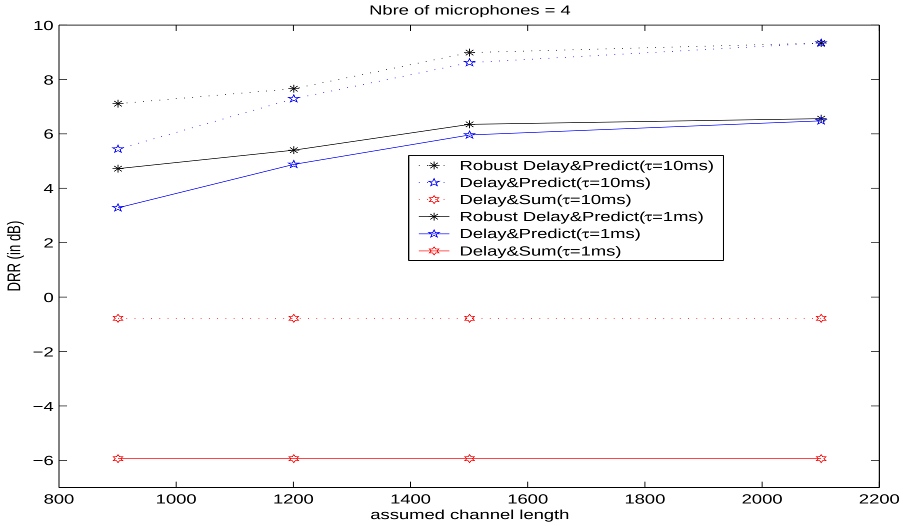

Figure 4.23: The output DRR function of the assumed channel length, using 4 microphone array setup ($\tau = 10\ ms$ and $\tau = 1\ ms$).

length is over-estimated, the robust D-&-P still performs better than the classic scheme, especially when only few microphones are available. As stated in previous, this is due to the fact that the robust D-&-P can compensate for the errors due to the estimation of the source correlations. These errors are more considerable as the number of microphones decreases.

4.7 Experimental results

In this section, we compare the accuracy of the Delay-&-Predict equalization, and the Delay-&-Sum beamforming using impulse responses measured in real environment, and real speech signal. the channel impulse responses are taking from the MARDY database, the convolved with a real speech signal sampled at 8kHz, and having a duration of 8.8s (figure 4.10). The distance between the source and the microphones is $d = 3m$. As an evaluation criterion, we consider the DRR ($\tau = 1ms$). Figure 4.24 shows that the D-&-P outperforms the D-&-S (in terms of DRR). For instance, the 2-microphones D-&-P performs better than the 8-microphones D-&-S. The D-&-P equalizer is particularly efficient if only few microphones are available. This is due to the fact that multichannel linear prediction performs well even using only two microphones; whereas the beamforming technique becomes an equalizer only as the number of microphones increases.

Figure 4.24: The D-&-P, D-&-S, and average reverberant channel DRR (for 2, 4, and 8 microphones).

The output DRR has straightforward interpretation; and it can provide indications of the perceived audio quality in some cases [199]. Unfortunately, the output DRR shows a limited correlation with perceived speech quality. Figure 4.25 compares of the PESQ of our proposed scheme and D-&-S beamforming, and the average PESQ of the reverberant signals.
Again, we see that the Delay-&-Sum beamformer gives poor results using a few number of microphones . However, the Delay-&-Predict scheme enhance the speech signal even using only 2 microphones.

4.8 Conclusion

In this chapter, a linear prediction based dereverberation technique was proposed. The multichannel reverberation impulse response is assumed stationary enough to allow estimation of the correlations it induces in the received signals. Spatial, temporal, and spectral diversities are exploited to transform the source speech signal into a whiter signal. An equalizer is then computed based on a multichannel linear prediction technique. Simulations

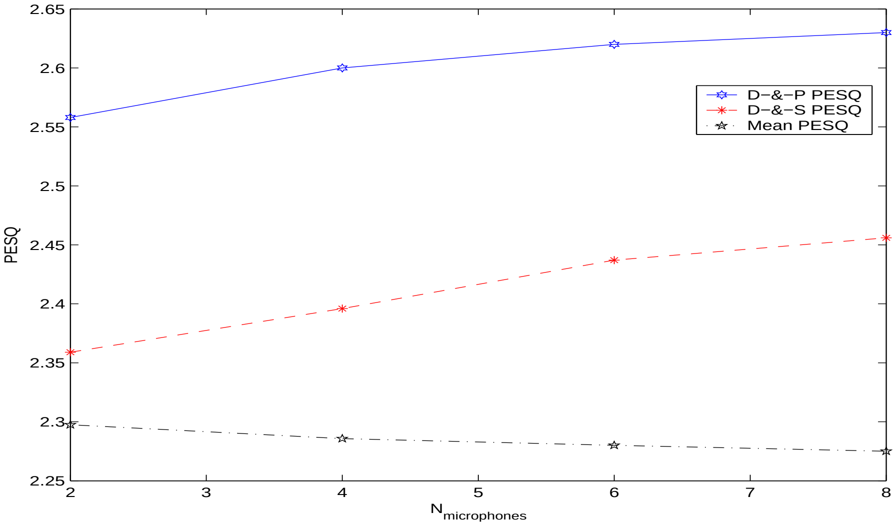

Figure 4.25: The D-&-P, D-&-S, and averaged reverberant channel PESQ (for 2, 4, and 8 microphones).

show that the Delay-and-Predict equalizer performs better than the delay-and-Sum beamformer, specially if only few microphones are available. We have also considered two robustness issues in the design of the LP-based equalizer in the presence of additive white noise. First, we have investigated the effect of relative subchannel delay compensation on the output SNR. We show that such relative delay compensation can increase considerably the output SNR. Then, we have optimized the transformation of the multivariate prediction filter to a longer equalizer filter using the SNR criterion. The optimization corresponds to MMSE-ZF design, and the filter length increase allows for the introduction of some equalization delay, that can also be optimized. Simulations show that considerable gains can be achieved by allowing even small equalization delays.

4.A Proof of Theorem 1

We consider the noisy SIMO system with M outputs:

$$\mathbf{x}(n) = \sum_{i=0}^{L_h-1} \mathbf{h}_i s(n-i) + \mathbf{v}(n) \tag{4.58}$$

where the channel noise $\mathbf{v}$ is assumed zero mean white process. The input signal and the noise covariances are denoted respectively $E\left\{s^2(n)\right\} = \sigma_s^2$, and $E\left\{\mathbf{v}(n)\mathbf{v}^T(n)\right\} = \sigma_v^2 I_M$. If σ_v^2 is known, one can compensate for the noise covariance in the reverberant signal covariance matrix. And noise-free multichannel LP can be computed (using the cleaned covariance matrix). This linear predictor is then applied to the noisy signal $\mathbf{x}(n)$.
To describe the blind linear predictive algorithm, it is easier to first form the $(L_A M) \times (L_A{+}L_h{-}1)$ block Toeplitz matrix [43]

$$\mathcal{H} = \begin{bmatrix} \mathbf{h}_0\ \mathbf{h}_1 & \cdots & \mathbf{h}_{L_h-1} & 0 & \cdots & 0 \\ 0 & \mathbf{h}_0\ \mathbf{h}_1 & \cdots & \mathbf{h}_{L_h-1} & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{h}_0\ \mathbf{h}_1 & \cdots & \mathbf{h}_{L_h-1} \end{bmatrix}$$

Equation (4.58) can be now written :

$$\mathbf{x}_{L_A}(n) = \mathcal{H}\,\mathbf{s}_{L_A}(n) + \mathbf{v}_{L_A}(n) \tag{4.59}$$

where: $\mathbf{x}_{L_A}(n) = \left[\ \mathbf{x}^T(n)\ \cdots\ \mathbf{x}^T(n{-}L_A{+}1)\ \right]^T$, $\mathbf{s}_{L_A}(n) = \left[\ s(n)\ \cdots\ s(n{-}L_A{-}L_h{+}2)\ \right]^T$, and $\mathbf{v}_{L_A}(n) = \left[\ \mathbf{v}^T(n)\ \cdots\ \mathbf{v}^T(n{-}L_A{+}1)\ \right]^T$.
With this notation, one can show that if $\mathbf{H}(q) = \sum_{i=0}^{L_h-1} \mathbf{h}_i q^{-i}$ has no zeros, the matrix $\mathcal{H}$ has full column rank. Then, the pseudoinverse $\mathcal{H}^{\#}$ exists, and the multichannel LP coefficients are given by [43]

$$\left[\ A_{L_A,1}\ \cdots\ A_{L_A,L_A}\ \right] = -\left[\ \mathbf{h}_1 \cdots \mathbf{h}_{L_h-1}\ 0 \cdots 0\ \right]\mathcal{H}^{\#}$$

By applying the linear prediction to the noisy observation, the residual signal becomes:

$$\begin{aligned} \widetilde{\mathbf{x}}(n) &= \mathbf{x}(n) + \sum_{i=1}^{L_A} A_{L_A,i}\mathbf{x}(n-i) \\ &= \mathbf{h}_0 s(n) + \sum_{i=0}^{L_A} A_{L_A,i}\mathbf{v}(n-i) \end{aligned} \tag{4.60}$$

The output MSE is given by

$$\begin{aligned}\text{MSE} &= \sigma_v^2 \text{tr}\left\{ I_M + \sum_{i=1}^{L_A} A_{L_A,i} A_{L_A,i}^T \right\} \\ &= \sigma_v^2 \text{tr}\left\{ I_M + \left[\mathbf{h}_1 \cdots \mathbf{h}_{L_h-1} 0 \cdots 0 \right] \left(\mathcal{H}^T \mathcal{H} \right)^{-1} \begin{bmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_{L_h-1}^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\}\end{aligned}$$

where $tr\{.\}$ denotes the trace operator.
Note that the MSE depends only on the $(L_h-1) \times (L_h-1)$ upper block of the matrix $\left(\mathcal{H}^T\mathcal{H}\right)^{-1}$. On the other hand, taking into consideration the whiteness of reverberation and its decaying energy (statistical channel model in (4.10)), one can show that this $(L_h-1) \times (L_h-1)$ upper block is almost diagonal, and that the MSE is given by

$$\begin{aligned} MSE &= \sigma_v^2 \left(M + \frac{||\mathbf{h}_1||^2}{||\mathbf{h}_0||^2} + \frac{||\mathbf{h}_2||^2}{||\mathbf{h}_0||^2 + ||\mathbf{h}_1||^2} + \cdots \right. \\ &+ \left. \cdots + \frac{||\mathbf{h}_{L_h-1}||^2}{||\mathbf{h}_0||^2 + \cdots + ||\mathbf{h}_{L_h-2}||^2} \right) \end{aligned} \tag{4.61}$$

The above equation shows how critical the energy of $\mathbf{h}_0$ is. In fact if $\|\mathbf{h}_0\|^2 \rightarrow 0$, not only the desired signal energy $(\sigma_s^2 \|\mathbf{h}_0\|^2) \rightarrow 0$, but also the MSE $\rightarrow \infty$. On the other hand, one can show that

$$\frac{\partial \text{MSE}}{\partial ||\mathbf{h}_0||^2} < 0. \tag{4.62}$$

Equation (4.62) is not sufficient to prove that the relative compensation decreases the output MSE. In fact, by aligning the received data, we are not increasing the energy of $\mathbf{h}_0$ independently of $\mathbf{h}_i$ $i \neq 0$. We denote $\{\delta_i\}_{i \geq 0}$ the difference between the energy of the i^{th} channel tap before and after relative delay compensation. After time aligning, the output MSE becomes:

$$\begin{aligned} MSE &= \sigma_v^2 \left(M + \frac{||\mathbf{h}_1||^2 + \delta_1}{||\mathbf{h}_0||^2 + \delta_0} + \frac{||\mathbf{h}_2||^2 + \delta_2}{||\mathbf{h}_0||^2 + ||\mathbf{h}_1||^2 + \delta_0 + \delta_1} \right. \\ &+ \left. \cdots + \frac{||\mathbf{h}_{L_h-1}||^2 + \delta_{L_h-1}}{||\mathbf{h}_0||^2 + \cdots + ||\mathbf{h}_{L_h-2}||^2 + \delta_0 + \cdots + \delta_{L_h-2}} \right) \end{aligned}$$

At first, we consider the delay compensation of only one subchannel. We denote by $\tau_d \neq 0$ the relative delay of this subchannel. Energy conservation leads to:

$$\sum_{s=0}^{\infty} \delta_{k+s\tau_d} = 0 \quad k \in [0, (\tau_d - 1)]$$

If we assume the channel energy to be decreasing with time lag and/or the relative delay τ_d is large enough, we have (figure 4.26):

$$\begin{aligned} \delta_i > 0 \ , \quad \forall i < \tau_d \\ \delta_i < 0 \ , \quad \forall i \geq \tau_d \end{aligned}$$

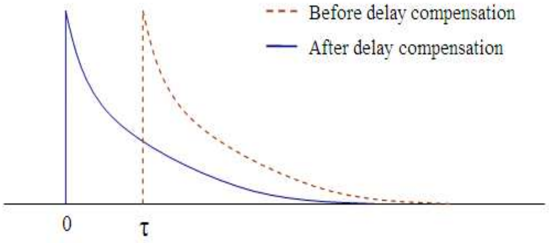

Figure 4.26: Subchannel impulse response before and after relative delay compensation .

Now, we consider the relative time compensation of the whole multichannel. $\tau_d \neq 0$ will denote the minimum non-zero relative delay on different subchannels. If M increases, and if the subchannel impulse responses follow (4.10), we have:

$$\begin{aligned} \frac{\delta_i}{\delta_0 + \cdots + \delta_{i-1}} &\approx \frac{||\mathbf{h}_i||^2}{||\mathbf{h}_0||^2 + \cdots + ||\mathbf{h}_{i-1}||^2} \ \forall i < \tau_d \\ \frac{\delta_i}{\delta_0 + \cdots + \delta_{i-1}} &< \frac{||\mathbf{h}_i||^2}{||\mathbf{h}_0||^2 + \cdots + ||\mathbf{h}_{i-1}||^2} \quad \forall i \geq \tau_d \end{aligned}$$

Therefore, the output MSE decreases after relative delay compensation. On the other hand, the desired signal energy $(\sigma_s^2 \left\|\mathbf{h}_0\right\|^2)$ increases.

Conclusion: By aligning the received data, the output SNR $= \dfrac{\sigma_s^2 \left\|\mathbf{h}_0\right\|^2}{MSE}$ increases.

4.B Multichannel LP computation and adaptation

Consider the problem of predicting $\mathbf{x}(n)$ from the L_A latest observations $\mathbf{x}_{L_A}(n{-}1) = [\mathbf{x}^T(n{-}1) \cdots \mathbf{x}^T(n{-}L_A)]^T$. The prediction error is given by:

$$\widetilde{\mathbf{x}}(n) = \mathbf{x}(n) + \sum_{i=1}^{L_A} \mathbf{A}_{x,i}\mathbf{x}(n-i) = \mathbf{A}_x \mathbf{x}_{L_A+1}(n) \tag{4.63}$$

where $\mathbf{A}_x = [\mathbf{I}_M \ \mathbf{A}_{x,1} \ \cdots \ \mathbf{A}_{x,L_A}]$, $\{\mathbf{A}_{x,i}\}_i$ represents $M \times M$ matrices of the linear prediction coefficients, $\mathbf{I}_M$ is the identity matrix of size M, and L_A denotes the prediction order. The spatio-temporal prediction filter is adapted by considering an RLS problem. In fact, the linear prediction matrices $\{\mathbf{A}_{x,i}\}_i$ are computed by minimizing the mean squared value of $\widetilde{\mathbf{x}}(n)$, i.e.,

$$\begin{cases} \min\limits_{A_i} \ J(n) = \sum\limits_{k=0}^{N-1} \|\widetilde{\mathbf{x}}(n-k)\|^2 \\ \mathbf{A}_0 = \mathbf{I}_M \end{cases} \tag{4.64}$$

where N is the length of the frame in which the channel is assumed to be stationary. $J(n)$ is a recursive-in-time measure which changes at each point to reflect the arrival of a new data sample. The objective of the RLS algorithm is to maintain a solution which is optimal with respect to (4.64) at each iteration.

Differentiating (4.64) leads:

$$\begin{aligned} \frac{\partial J(n)}{\partial A_{x,i_0}} &= 2\sum_{k=0}^{N-1} \widetilde{\mathbf{x}}(n-k)\mathbf{x}^T(n-k-i) \\ &= 2\sum_{k=0}^{N-1} \sum_{i=0}^{L_A} \mathbf{A}_{x,i} \left(\mathbf{x}(n-k-i)\mathbf{x}^T(n-k-i_0)\right) \\ &= 2\sum_{i=0}^{L_A} \mathbf{A}_{x,i}\mathbf{R}^T_{i_0,i}(n) \end{aligned}$$

with $\mathbf{R}_{i,j}(n) = \sum\limits_{k=0}^{N-1} \mathbf{x}(n-k-i)\mathbf{x}^T(n-k-j)$.

Transposing, and equating to zero leads to a set of a block normal equations

$$\mathbf{R}_x(n)\ \mathbf{A}(n) \ = \ \mathbf{P}_x(n) \tag{4.65}$$

where - $\mathbf{R}_x(n) = \begin{bmatrix} \mathbf{R}_{1,1}(n) & \cdots & \mathbf{R}_{1,L_A}(n) \\ \vdots & & \vdots \\ \mathbf{R}_{L_A,1}(n) & \cdots & \mathbf{R}_{L_A,L_A}(n) \end{bmatrix}$

- $\mathbf{A}(n) = \begin{bmatrix} \mathbf{A}^T_{x,1}(n) \\ \vdots \\ \mathbf{A}^T_{x,L_A}(n) \end{bmatrix}$

- $\mathbf{P}_x(n) = -\begin{bmatrix} \mathbf{R}_{1,0}(n) \\ \vdots \\ \mathbf{R}_{L_A,0}(n) \end{bmatrix}$

Note that $\mathbf{R}_x(n)$ can be computed recursively using

$$\mathbf{R}_x(n) = \mathbf{R}_x(n-1) + \mathbf{x}_{L_A}(n)\mathbf{x}_{L_A}(n)^T - \mathbf{x}_{L_A}(n-N)\mathbf{x}_{L_A}(n-N)^T \tag{4.66}$$

Thus, we can solve the problem recursively using

$$\begin{aligned}
\mathbf{R}_x^{-1}(n-\frac{1}{2}) &= \mathbf{R}_x^{-1}(n-1) - \frac{\mathbf{R}_x^{-1}(n-1)\mathbf{x}_{L_A}(n)\mathbf{x}^T_{L_A}(n)\mathbf{R}_x^{-1}(n-1)}{1+\mathbf{x}^T_{L_A}(n)\mathbf{R}_x^{-1}(n-1)\mathbf{x}_{L_A}(n)} \\
\mathbf{R}_x^{-1}(n) &= \mathbf{R}_x^{-1}(n-\frac{1}{2}) + \frac{\mathbf{R}_x^{-1}(n-\frac{1}{2}\mathbf{x}_{L_A}(n-N)\mathbf{x}^T_{L_A}(n-N)\mathbf{R}_x^{-1}(n-\frac{1}{2})}{1-\mathbf{x}^T_{L_A}(n-N)\mathbf{R}_x^{-1}(n-\frac{1}{2})\mathbf{x}_{L_A}(n-N)} \qquad (4.67) \\
\mathbf{P}_x(n) &= \mathbf{P}_x(n-1) - \mathbf{x}_{L_A}(n-1)\mathbf{x}^T(n) + \mathbf{x}_{L_A}(n-N-1)\mathbf{x}^T(n-N)
\end{aligned}$$

In summary, the multichannel linear prediction coefficient are updated as:

$$\begin{aligned}
\mathbf{C}(n) &= \mathbf{R}_x^{-1}(n)\mathbf{x}_{L_A}(n) \\
\gamma_k &= 1 + \mathbf{x}_{L_A}^T(n)\mathbf{C}(n) \\
\mathbf{R}_x^{-1}(n - \frac{1}{2}) &= \mathbf{R}_x^{-1}(n) - \mathbf{C}(n)\gamma^{-1}\mathbf{C}^T(n) \\
\mathbf{P}_x(n + \frac{1}{2}) &= \mathbf{P}_x(n) - \mathbf{x}_{L_A}(n)\mathbf{x}^T(n+1) \\
\\
\mathbf{D}(n) &= \mathbf{R}_x^{-1}(n - \frac{1}{2})\mathbf{x}_{L_A}(n - N + 1) \\
\delta_k &= 1 - \mathbf{x}_{L_A}(n - N + 1)^T\mathbf{D}(n) \\
\mathbf{R}_x^{-1}(n+1) &= \mathbf{R}^{-1}(n - \frac{1}{2}) + \mathbf{D}(n)\delta^{-1}\mathbf{D}^T(n) \\
\mathbf{P}_x(n+1) &= \mathbf{P}_x(n + \frac{1}{2}) + \mathbf{x}_{L_A}(n - N)\mathbf{x}_{L_A}^T(n - N + 1) \\
\\
\mathbf{A}(n+1) &= \mathbf{R}_x^{-1}(n+1)\mathbf{P}_x(n+1)
\end{aligned}$$

The spatial covariance matrix update

The sampled spatial covariance matrix is defined as

$$\mathbf{\Sigma}_{\tilde{x}}(n) = \sum_{k=0}^{N-1} \widetilde{\mathbf{x}}(n-k)\widetilde{\mathbf{x}}^T(n-k) \tag{4.68}$$

Using the same approach, we compute recursively the previous quantity using

$$\begin{aligned}
\Sigma_{\tilde{x}}^{-1}(n + \frac{1}{2}) &= \Sigma_{\tilde{x}}^{-1}(n-1) - \frac{\Sigma_{\tilde{x}}^{-1}(n-1)\widetilde{\mathbf{x}}(n)\widetilde{\mathbf{x}}^T(n)\Sigma_{\tilde{x}}^{-1}(n-1)}{1 + \widetilde{\mathbf{x}}^T(n)\mathbf{R}_x^{-1}(n)\widetilde{\mathbf{x}}(n)} \\
\Sigma_{\tilde{x}}^{-1}(n+1) &= \Sigma_{\tilde{x}}^{-1}(n + \frac{1}{2}) + \frac{\Sigma_{\tilde{x}}^{-1}(n + \frac{1}{2})\widetilde{\mathbf{x}}(n-N)\widetilde{\mathbf{x}}^T(n-N)\Sigma_{\tilde{x}}^{-1}(n + \frac{1}{2})}{1 - \widetilde{\mathbf{x}}^T(n-N)\mathbf{R}_x^{-1}(n + \frac{1}{2})\widetilde{\mathbf{x}}^T(n-N)}
\end{aligned}$$

4.C GSC for speech dereverberation

Let us consider the configuration where the reverberant speech signal is observed on two distinct microphones

$$\mathbf{x}(n) = H(q)s(n) = \sum_{i=0}^{L_h} \mathbf{h}_i s(n-i) \tag{4.69}$$

we denotes by $h_d(q) = \mathbf{h}_0^H H(q)$ and $h_z(q) = \mathbf{h}_0^{\perp^H} H(q)$ the two scalar filters characterizing the reverberation in the desired and noise reference signals respectively. We assume that $qh_z(q)$ is minimum phase (remark that $qh_z(q)$ is causal as the zero-lag component of $h_z(q)$ is equal to zero).
The noise canceller $W_{GSC}(q)$ is designed to perform the causal LMMSE estimation of $d(n)$ from the noise reference $\{z(n), z(n-1), \cdots\}$. Function of the (cross) second order statistics of the desired signal and the noise reference, the causal-Wiener filter can be expressed as:

$$W_{GSC}(q) = \frac{1}{S_{zz}^{+}(q)} \left\{ \frac{S_{dz}(q)}{S_{zz}^{+}(q^{-1})} \right\}_{+} \tag{4.70}$$

where

- $\{G(q)\}_+$: takes the causal part of $G(q)$.
- $S_{zz}(q) = S_{zz}^{+}(q) S_{zz}^{+}(q^{-1})$: is the spectral factorization. Subject to certain conditions, a power spectral density function (PSDF) can be factored into its causal minimum-phase factor $S_{zz}^{+}(q)$ and its anti-causal maximum phase counterpart $S_{zz}^{+}(q^{-1})$.

Taking into consideration the SIMO propagation structure, the PSDF of the noise reference can be written as:

$$\begin{aligned} S_{zz}(q) &= \mathbf{h}_0^{\perp^H} \mathbf{S}_{yy}(q) \mathbf{h}_0^{\perp} \\ &= \mathbf{h}_0^{\perp^H} A_x^{-1}(q) \mathbf{h}_0 \sigma_s^2 \mathbf{h}^H A_x^{-\dagger}(q) \mathbf{h}_0^{\perp} \end{aligned}$$

On the other hand, as the linear predictor is chosen long enough (to equalize perfectly the channel), i.e.,

$$A_x^{-1}(q)\mathbf{h}_0 = H(q) \tag{4.71}$$

The PSDF of the noise reference becomes

$$S_{zz}(q) \;=\; \underbrace{\mathbf{h}_0^{\perp^H} A_x^{-1}(q)\mathbf{h}_0}_{h_z(q)}\ \sigma_s^2\ \underbrace{\mathbf{h}^H A_x^{-\dagger}(q)\mathbf{h}_0^{\perp}}_{h_z(q^{-1})} \tag{4.72}$$

As $qh_z(q)$ is assumed minimum phase, the spectral factorization of S_{zz} can be expressed as

$$S_{zz}(q) = \underbrace{\left(q\sigma_s h_z(q)\right)}_{S_{zz}^{+}(q)}\ \cdot\ \underbrace{\left(q^{-1}\sigma_s h_z(q^{-1})\right)}_{S_{zz}^{+}(q^{-1})} \tag{4.73}$$

Finally the noise canceller is given by:

$$\begin{aligned} W_{GSC}(q) &= \frac{1}{q\sigma_s h_z(q)}\left\{\frac{\sigma_s^2 h_d(q)h_z(q))}{q^{-1}\sigma_s h_z(q^{-1})}\right\}_+ \\ &= \frac{1}{qh_z(q)}\left\{qh_d(q)\right\}_+ \\ &= \frac{h_d(q)-\|\mathbf{h}_0\|^2}{h_z(q)} = \frac{\mathbf{h}_0^H\left(A_x^{-1}(q)-\mathbf{I}\right)\mathbf{h}_0}{\mathbf{h}_0^{\perp^H} A_x^{-1}(q)\mathbf{h}_0} \end{aligned} \tag{4.74}$$

In sum, if $qh_z(q)$ is minimum phase, the generalized sidelobe cancellation enables perfect dereverberation:

$$\begin{aligned} \widehat{s}(n) &= d(n) - W_{GSC}(q)z(n) \\ &= h_d(q)s(n) - \frac{h_d(q)-\|\mathbf{h}_0\|^2}{h_z(q)}h_z(q)s(n) \\ &\propto s(n) \end{aligned}$$

If $qh_z(q)$ is non-minimum phase, $\frac{h_d(q)-\|\mathbf{h}_0\|^2}{h_z(q)}$ is no-longer stable and perfect dereverberation is no-longer possible using GSC scheme. On the other hand, It has be shown that the energy of minimum phase system is most concentrated in the beginning, i.e., the energy of minimum phase systems is delayed the least of all systems having the same magnitude response function [127]. Taking this remark into consideration, relative time delay compensation is also beneficial for a GSC scheme: the channel alignment concentrate the energy of $H(q)$ (then of $qh_z(q)$) around $n = 0$.

Part II

CWCU Estimation and Application to Mobile Localization

Chapter 5

CWCU Bayesian Parameter Estimation

Bayesian parameter estimation techniques such as Linear Minimum Mean Squared Error (LMMSE) often lead to useful MSE reduction, but they also introduce a bias (annoying in several applications). In this chapter, we introduce the concept of Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation, in which unbiasedness is forced for one parameter at a time. The more general introduction of the CWCU concept is motivated by LMMSE channel estimation, for which the implications of the concept are illustrated in various ways, including the effect on angle of arrival estimation, repercussion for trained channel estimation etc. Motivated by the channel tracking application, we also introduce CWCU Kalman filtering.

5.1 Introduction

In most applications, estimator designs are subject to a tradeoff between bias and variance. Bias is due to 'mismatch' between the average value of the estimator and the true parameter (conditional bias); whereas variance arises from fluctuations in the estimator due to statistical sampling.

If prior information on the parameter statistics is available, Bayesian estimation theory shows that under the Bayesian unbiasedness constraint, the MSE is bounded below by the Bayesian Cramer-Rao Bound (B-CRB). Moreover, the MMSE estimator minimizes $\mathbf{R}_{\widetilde{\theta}\widetilde{\theta}}$, the parameter estimation error correlation matrix, and not only the MSE (which is the trace of $\mathbf{R}_{\widetilde{\theta}\widetilde{\theta}}$). Nevertheless, Bayesian unbiasedness for random parameters corresponds to unbiasedness on the average, which is a very weak requirement. In particular the MMSE estimator is unbiased, and the MMSE estimator minimizes $\mathbf{R}_{\widehat{\theta}\widehat{\theta}}$ and the MSE, regardless of whether the Bayesian unbiasedness constraint is imposed or not. Thus, the Bayesian estimation leads then to a (conditionally) biased estimation. This bias is detrimental for a number of applications: MultiUser Signal Detection (MLSD) in a SISO system using the Viterbi algorithm (the bias is as detrimental as in biased LMMSE symbol receivers), fitting a parametric (pathwise) model to the channel impulse response, or using the channel estimate for the design of the receiver or the transmitter.

On the other hand, requiring that all parameter components to be jointly unbiased (which corresponds to zero-forcing when the parameters are multiple symbols) prevents the exploitation of prior statistical information. Hence, it leads to a significant reduction in estimation MSE.
This motivates us to introduce the Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation. Instead of constraining the estimator to be globally unbiased, we impose conditional unbiasedness on one parameter component at a time. In such a way, every parameter in turn is treated as deterministic while the others are being treated as Bayesian. If the parameters are transmitted symbols, the CWCU approach corresponds to unbiased symbol detection whereas joint deterministic unbiasedness leads to a zero-forcing approach.

In this chapter, we show that the CWCU estimation (and the Bayesian estimation in general) is particularly interesting if the parameter components are correlated and/or convolved trough a colored signal. This is typically the case in audio applications. Nevertheless, the Bayesian estimation is rarely used for audio processing because conditional bias is often annoying for such applications. Although the application of the CWCU concept to audio processing seems natural, it was not be considered in the context of this work. In this thesis, we consider some applications to digital communication such as supervised channel and direction of arrival estimation (sections 5.7 and 5.6), and application to mobile terminal positioning (chapter 6).

This chapter is organized as follows. In section 5.3, we investigate lower bounds for the CWCU estimation. The CWCU-LMMSE estimation, and CWCU linear filtering are derived respectively in sections 5.4, and 5.5. The interplay between block-size, joint bias and prior covariance rank is investigated in section 5.6. Application of the concept to channel estimation for mobile localization is presented in section 5.7.

5.2 Bias vs. MSE in parameter estimation: a brief overview

In most applications, estimator designs are subject to a tradeoff between bias and variance. Ideally, we would like to minimize the Mean Squared Error (MSE) over all possible estimators and hence over all bias vectors $\mathbf{b}(\boldsymbol{\theta})$. Unfortunately, if no limitations are imposed on $\widehat{\boldsymbol{\theta}}$, an estimator can always be found that makes both the bias and variance zero at a given point $\boldsymbol{\theta}$. Thus, instead of attempting to minimize the MSE over all possible estimators, we may restrict attention to estimator with a bias vector that lies in a suitable class. Then, the bias / variance tradeoff is fixed by minimizing the MSE under some constraints on the bias. A second problem is that such minimization is generally difficult to solve. One way to fix the tradeoff is to develop bounds on the best achievable performance in estimating parameters of interest, as well as to determine estimators that achieve these bounds. Using the Biased Cramer-Rao Bound B-CRB (which bounds the total MSE), we can bound the MSE of any estimator $\hat{\boldsymbol{\theta}}$ with a given bias vector $\mathbf{b}(\boldsymbol{\theta})$ by

$$\|\mathbf{b}(\boldsymbol{\theta})\|^2 + \operatorname{tr}\left\{(\mathbf{I}+\mathbf{D}(\boldsymbol{\theta}))\,\mathbf{J}^{-1}(\boldsymbol{\theta})\,(\mathbf{I}+\mathbf{D}(\boldsymbol{\theta}))^T\right\} \tag{5.1}$$

where $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher Information Matrix, and $\mathbf{D}(\boldsymbol{\theta}) = \dfrac{\partial \mathbf{b}^T(\boldsymbol{\theta})}{\partial \theta}$ is the bias gradient matrix. tr $\{.\}$ and $(.)^T$ denote respectively the trace and the transpose operators. $\mathbf{I}$ represents the identity matrix.
Traditionally, we consider the class of unbiased estimators. The MSE is bounded by the CRB. It can also be shown that for a Gaussian linear model, the Maximum Likelihood (ML) estimator achieves the CRB; and that is asymptotically unbiased for independent identically distributed (i.i.d.) measurements (under suitable regularity assumption).
For biased estimators, given a specified bias, the B-CRB serves as a bound on the smallest attainable variance. It turns out that the B-CRB does not depend directly on the bias but only on the bias gradient matrix. However, it may not be obvious how to choose a particular bias gradient. Hero et al. propose the Uniform CRB (U-CRB), which is a bound on the smallest attainable variance that can be achieved using any scalar estimator with a bounded bias gradient norm [75]. Reference [44] extends the U-CRB for vector parameter, and develops a class of estimators which asymptotically achieve the bound when estimating an unknown vector from i.i.d. vector measurements. However, Shahtalebi and Gasor show that for a linear model, the U-CRB is achievable by a class of linear estimators [53]. All estimators in this class have the same variance and the same gradient matrix. However, their performances (in terms of achievable MSE) are not the same. They conclude that the B-CRB is not a sufficient criterion to design optimal estimators.
Eldar considers the minimization of the MSE bound under a linear biased constraint[45]. In fact, bias vectors are allowed to be linear in $\boldsymbol{\theta}$, so that $\mathbf{b}(\theta) = \mathbf{M}\boldsymbol{\theta}$ for some matrix $\mathbf{M}$ (which includes unbiased estimation as a special case). An advantage of this class of estimators is that we can use results on unbiased estimation theory to find estimators which achieve the corresponding MSE bound. In fact, if $\widehat{\boldsymbol{\theta}}_0$ is an efficient estimator, i.e., an unbiased estimator that achieves the CRB, the $\widehat{\boldsymbol{\theta}} = (\mathbf{I} + \mathbf{M})\widehat{\boldsymbol{\theta}}_0$ achieves the MSE bound for estimators whose bias is equal to $\mathbf{b}(\boldsymbol{\theta}) = \mathbf{M}\boldsymbol{\theta}$. The problem is that minimization cannot be solved in the general case. However, the author shows that there often exists linear biased vectors that result in an MSE bound that dominate the CRB. The dominating bound can be obtained by solving a certain minimax optimization problem.

If prior information on the parameter statistics is available, Bayesian estimation theory shows that under Bayesian unbiasedness constraint, the MSE is bounded by the Bayesian CRB. Bayesian unbiasedness for random parameters corresponds to unbiasedness on the average, which is a very weak requirement. So that, we can achieve better bias vs. variance tradeoff. In recent years, the Bayesian formulation of channel estimation has become popular, as it allows for instance the exploitation of the power delay profile. This allows to reduce the number of parameters to be estimated from an a priori delay spread range to the effective delay spread of the power delay profile. For SIMO, MISO or MIMO channels, the Bayesian formulation allows to exploit correlation between antennas and to reduce the number of parameters from the physical number of antennas to an effective number of uncorrelated antennas. When the channel is fading in time, the Doppler spectrum and hence correlation in time can be exploited via Wiener or Kalman filtering to further reduce the MSE.

In all these cases, Bayesian estimation leads to biased channel estimates. This bias is detrimental for a number of applications: MLSD in a SISO system using the Viterbi algorithm (the bias is as detrimental as in biased LMMSE symbol receivers), fitting a parametric (pathwise) model to the channel impulse response, or using the channel estimate for the design of the receiver or the transmitter. The type of unbiasedness that is required here is conditional unbiasedness (where unbiasedness for Bayesian estimation corresponds to unbiasedness on the average, which is very weak requirement). However, conditional unbiasedness for vectors of parameters is usually introduced globally, requiring all parameter components to be jointly unbiased. However, such a stringent requirement, which corresponds to zero-forcing when the parameters are multiple symbols, prevents the exploitation of correlations between the parameters, and hence leads to a significant reduction in the benefits brought about by the Bayesian framework, the prior knowledge.

This motivates us to introduce the Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation. Instead of constraining the estimator to be globally unbiased, i.e., $E_{/\theta}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=0$, we impose conditional unbiasedness on one parameter component at a time, i.e.,

$$E_{Y|\theta_k}\left(\widehat{\theta}_k-\theta_k\right)=0 \qquad k=1:K \tag{5.2}$$

where $E_{Y|x}\left[Z(Y,X)\right]=E_Y\left[Z(Y,X)|x\right]=\displaystyle\int Z(Y,x)f_{Y|x}(y|x)dY$ denotes the

expectation of $Z(X,Y)$ on Y conditional to $X=x$; and $\theta=[\theta_1\cdots\theta_K]^T$ is the parameter vector to be estimated.
In such a way, the parameter of interest is constrained to be conditionally unbiased. Other parameters are treated as nuisance parameters. Note that the component-wise concept can be defined at different levels. For example, if we consider multichannel impulse response estimation; the component-wise concept can be defined at scalar level (by considering conditional unbiasedness separately for different channels and time lags). It can also be defined at a block level (by considering conditional unbiasedness jointly for different channels, and separately for different time lags).

5.3 Lower bounds for CWCU-MMSE estimation

We consider the estimation of a random parameter vector $\boldsymbol{\theta}=[\theta_1\cdots\theta_K]^T$ given a set of measurements collected in $\mathbf{y}$. The MMSE parameter estimation under the component-wise conditionally unbiasedness constraint can be formulated as:

$$\begin{cases} \min\limits_{\widehat{\theta}} E\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|^2=\sum\limits_k E\left\|\widehat{\theta}_k-\theta_k\right\|^2 \\ E_{\mathbf{y}|\theta_k}\left(\widehat{\theta}_k-\theta_k\right)=0 \quad k=1:K \end{cases}$$

It is easy to see that the minimization problem is separable, and $\widehat{\theta}_k$ is a solution of

$$\begin{cases} \min\limits_{\widehat{\theta}_k} E\left\|\widehat{\theta}_k-\theta_k\right\|^2 \\ E_{\mathbf{y}|\theta_k}\left(\widehat{\theta}_k-\theta_k\right)=0 \end{cases}$$

Without loss of generality, we can assume that $\boldsymbol{\theta}$ can be decomposed as $\boldsymbol{\theta}^T=[\theta_k^T \ \ \bar{\boldsymbol{\theta}}_k^T]$. A Bayesian lower bound on the error variance is given by,

$$E\left\|\widehat{\theta}_k-\theta_k\right\|^2\geq E_{\theta_k}\left(\mathbf{J}_{CRB,k}^{-1}\right) \tag{5.3}$$

where $\mathbf{J}_{CRB,k}=E_{\mathbf{y}|\theta_k}\left(\frac{\partial\ln f\left(\mathbf{y}|\theta_k\right)}{\partial\theta_k}\right)\left(\frac{\partial\ln f\left(\mathbf{y}|\theta_k\right)}{\partial\theta_k}\right)^T$ is the Fisher Information Matrix (FIM) where θ_k is considered as deterministic, and $\bar{\boldsymbol{\theta}}_k$ as nuisance parameters. To evaluate the above expression, we should evaluate the conditional pdf with respect to θ_k:

$$f\left(\mathbf{y}|\theta_k\right) = \int f\left(\mathbf{y}, \bar{\boldsymbol{\theta}}_k|\theta_k\right) d\bar{\boldsymbol{\theta}}_k = \int f\left(\mathbf{y}|\boldsymbol{\theta}\right) f\left(\bar{\boldsymbol{\theta}}_k|\theta_k\right) d\bar{\boldsymbol{\theta}}_k$$

Usually, the above bound is difficult to compute because either the above integration is not solvable, or the resulting expectation is not analytically tractable. This difficulty motivates the use of the modified Cramer-Rao bound (MCRB) (introduced by D'Andrea et al. in [11]),

$$E\left\|\widehat{\theta}_k - \theta_k\right\|^2 \geq E_{\theta_k}\left(\mathbf{J}_{MCRB,k}^{-1}\right) \tag{5.4}$$

where $\mathbf{J}_{MCRB,k} = E_{\mathbf{y},\bar{\theta}_k|\theta_k}\left(\frac{\partial \ln f\left(\mathbf{y}|\boldsymbol{\theta}\right)}{\partial \theta_k}\right)\left(\frac{\partial \ln f\left(\mathbf{y}|\boldsymbol{\theta}\right)}{\partial \theta_k}\right)^T$.
With respect to the classical CRB, The MCRB is much easier to compute, but generally lower. The problem is that the MCRB can be not tight enough for use in practical applications.

Another approach to facilitate the calculation of the CRB is to resort the CRB of the joint estimation of the desired parameter together with the nuisance terms [150]. Under the CWCU constraints, the estimation problem becomes

$$\begin{cases} \min\limits_{\widehat{\theta}} E\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\ E_{\mathbf{y},\bar{\theta}_k|\theta_k}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = 0 \end{cases}$$

Note that only $\widehat{\theta}_k$ is of interest; $\widehat{\bar{\boldsymbol{\theta}}}_k$ is only estimated to reduce the interference. In such way, the component of interest gets treated as deterministic whereas the other (correlated) parameter components continue to be treated as Bayesian. So that, other components are estimated better (taking into account prior information); as well as the component of interest (due to the coupling through prior and/or data).
As in the Bayesian and deterministic case, a performance bound on CWCU estimation can be defined based on the Fisher Information Matrix (FIM). The FIM for component-wise conditional estimation problem with respect to the parameter θ_k can be defined as:

$$\mathbf{J}_{/k}(\theta_k){=}E_{\mathbf{y},\bar{\theta}_k|\theta_k}\left(\frac{\partial \ln f\left(\mathbf{y}, \bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial \theta}\right)\left(\frac{\partial \ln f\left(\mathbf{y}, \bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial \theta}\right)^T \tag{5.5}$$

The Hessian of $\ln f\left(\boldsymbol{\theta}|\mathbf{y}\right)$ can be formulated as:

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(\frac{\partial \ln f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial\theta}\right)^T =& \frac{1}{f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}\frac{\partial}{\partial\theta}\left(\frac{\partial f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial\theta}\right)^T \\ &- \left(\frac{\partial \ln f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial\theta}\right)\left(\frac{\partial \ln f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial\theta}\right)^T \end{aligned} \tag{5.6}$$

For the expectation of the first term, we get

$$E_{\mathbf{y},\bar{\theta}_k|\theta_k}\frac{1}{f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}\frac{\partial}{\partial\theta}\left(\frac{\partial f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial\theta}\right)^T = 0 \tag{5.7}$$

Thus, using Bayes' rule, (5.6), and (5.7), the CWCU-FIM can be decomposed onto:

$$\begin{aligned}\mathbf{J}_{/k}(\theta_k) &= \frac{\partial}{\partial\theta}\left(\frac{\partial \ln f_{\theta_k}(\theta_k)}{\partial\theta}\right)^T - E_{\theta|\theta_k}\frac{\partial}{\partial\theta}\left(\frac{\partial \ln f_{\theta}(\boldsymbol{\theta})}{\partial\theta}\right)^T - E_{\mathbf{y},\theta|\theta_k}\frac{\partial}{\partial\theta}\left(\frac{\partial \ln f_{\mathbf{y}|\theta}(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta}\right)^T \\ &= \qquad -\mathbf{J}^{cw}_{/k} \qquad + \qquad \mathbf{J}^{prior}_{/k} \qquad + \qquad \mathbf{J}^{data}_{/k}\end{aligned} \tag{5.8}$$

As expected, we see that $J^{prior}_{/k} + J^{data}_{/k} \geq J_{/k} \geq J^{data}_{/k}$, since the CWCU estimation exploits the correlation between the parameters, and imposes an unbiasedness constraint for the parameter of interest.
Using the Schur components lemma, and the block matrix inversion formula, one can show that the error variance is bounded by:

$$E\left\|\widehat{\theta}_k - \theta_k\right\|^2 \geq E_{\theta_k}\left(J^{-1}_{HCRB,k}\right) \tag{5.9}$$

where $\mathbf{J}_{HCRB,k} = \mathbf{J}_{/k}(\theta_k,\theta_k) - \mathbf{J}^T_{/k}(\theta_k,\bar{\theta}_k)J^{-1}_{/k}(\bar{\theta}_k,\bar{\theta}_k)\mathbf{J}_{/k}(\theta_k,\bar{\theta}_k)$, and $\mathbf{J}_{/k}(x,z) = E_{\mathbf{y},\theta|\theta_k}\left(\frac{\partial \ln f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial x}\right)\left(\frac{\partial \ln f\left(\mathbf{y},\bar{\boldsymbol{\theta}}_k|\theta_k\right)}{\partial z}\right)^T$.

Remark that if θ_k and $\bar{\boldsymbol{\theta}}_k$ are independent $\mathbf{J}_{/k}(\theta_k,\theta_k) = \mathbf{J}_{MCRB,k}$. $\mathbf{J}_{HCRB,k}$, and $\mathbf{J}_{MCRB,k}$ overlap if and only if there is no coupling between the different parameters (through prior nor data).

Now, we consider a linear Gaussian model in (5.10). One can show that,

$$\begin{aligned}
\mathbf{J}_{CRB,k} &= \mathbf{C}^{-1}_{\theta_k\theta_k}\mathbf{C}_{\theta_k\theta}\mathbf{H}^T\ \left(\mathbf{C}_{vv}\mathbf{H}\left(\mathbf{C}_{\theta\theta}-\mathbf{C}_{\theta\theta_k}\mathbf{C}^{-1}_{\theta_k\theta_k}\mathbf{C}_{\theta_k\theta}\right)\mathbf{H}^T\right)^{-1}\mathbf{C}_{\theta\theta_k}\mathbf{C}^{-1}_{\theta_k\theta_k}\\
\mathbf{J}_{MCRB,k} &= \mathbf{h}_k^T\mathbf{C}_{vv}^{-1}\mathbf{h}_k\\
\mathbf{J}_{/k} &= \mathbf{C}_{\theta\theta}^{-1}-\mathbf{C}^{-1}_{\theta_k\theta_k}\mathbf{e}_k^T\mathbf{e}_k+\mathbf{H}^T\mathbf{C}_{vv}^{-1}\mathbf{H}
\end{aligned}$$

where $\mathbf{e}_k=[0\cdots 0\ 1\ 0\cdots 0]^T$ is the k^{th} element of the standard R^K basis $\left(\mathbf{e}_k^T\theta=\theta_k\right)$; and $\mathbf{h}_k=\mathbf{He}_k$ is the k^{th} column of $\mathbf{H}$. Using Monte Carlo simulations with the linear model, we obtain:

$$\mathbf{J}^{-1}_{CRB,k}\geq\mathbf{J}^{-1}_{HCRB,k}\geq\mathbf{J}^{-1}_{MCRB,k}$$

5.4 CWCU-LMMSE estimation for linear gaussian model

We consider a linear Gaussian model:

$$\mathbf{y}=\mathbf{H}\boldsymbol{\theta}+\mathbf{v} \tag{5.10}$$

where $\mathbf{y}$ is $N\times 1$ vector containing the received signal, $\boldsymbol{\theta}\sim N\left(0,\mathbf{C}_{\theta\theta}\right)$ is a $K\times 1$ vector containing the parameters to be estimated, and $\mathbf{v}\sim N\left(0,\sigma_v^2\mathbf{I}_N\right)$ is an $N\times 1$ additive white Gaussian noise independent from θ. $\mathbf{I}_N$ represents the identity matrix of size N.

As $\boldsymbol{\theta}$ and $\mathbf{y}$ are jointly Gaussian, minimizing the MSE leads to the LMMSE estimator:

$$\begin{aligned}
\widehat{\theta}_{lmmse} &= \arg\min_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|^2\\
&= \arg\min_{\widehat{\theta}=Fy}\operatorname{tr}\left\{\left(\mathbf{FH}-\mathbf{I}_K\right)\mathbf{C}_{\theta\theta}\left(\mathbf{FH}-\mathbf{I}_K\right)^H\right\}+\sigma_v^2\operatorname{tr}\left\{\mathbf{FF}^H\right\}\\
&= \mathbf{C}_{\theta\theta}\mathbf{H}^H\left(\mathbf{HC}_{\theta\theta}\mathbf{H}^H+\sigma_v^2\mathbf{I}_K\right)^{-1}\mathbf{y}
\end{aligned} \tag{5.11}$$

Under the joint unbiasedness constraint, minimizing the MSE leads to

$$\widehat{\boldsymbol{\theta}}=\left\{\begin{array}{l}\arg\min\limits_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|^2\\ E_{Y|\theta}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=0\end{array}\right. = \left\{\begin{array}{l}\arg\min\limits_{F}\operatorname{tr}\left\{\mathbf{FF}^H\right\}\\ \mathbf{FH}=\mathbf{I}_K\end{array}\right.$$

Then, joint unbiasedness prevents the exploitation of correlations between the parameters, and leads to a significant reduction in the benefits brought about by the Bayesian framework: the prior knowledge. In such a case, the MMSE estimator corresponds to the BLUE, i.e.,

$$\widehat{\boldsymbol{\theta}}_{blue} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{y} \tag{5.12}$$

LMMSE and BLUE estimators are related by (see appendix 5.A)

$$\widehat{\boldsymbol{\theta}}_{lmmse} = \underbrace{\mathbf{C}_{\theta\theta}\left(\left(\mathbf{H}^H\mathbf{H}\right)\mathbf{C}_{\theta\theta} + \sigma_v^2\mathbf{I}_K\right)^{-1}\left(\mathbf{H}^H\mathbf{H}\right)}_{\mathbf{B}_{lmmse}}\widehat{\boldsymbol{\theta}}_{blue} \tag{5.13}$$

where $\mathbf{B}_{lmmse} = \mathbf{C}_{\theta\theta}\left(\left(\mathbf{H}^H\mathbf{H}\right)\mathbf{C}_{\theta\theta} + \sigma_v^2\mathbf{I}_K\right)^{-1}\left(\mathbf{H}^H\mathbf{H}\right)$ represents the bias of the LMMSE estimation.

Imposing the CWCU constraints leads to the optimization problem

$$\widehat{\boldsymbol{\theta}}_{cwculmmse} = \begin{cases} \arg\min\limits_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\ E_{\mathbf{y}|\theta_k}\left(\widehat{\theta}_k - \theta_k\right) = 0 \quad k = 1:K \end{cases}$$

As $\boldsymbol{\theta}$ is assumed to be Gaussian, we have

Then, the CWCU-LMMSE is computed by optimizing

$$\begin{cases} \arg\min\limits_{\widehat{\theta}=Fy} \operatorname{tr}\left\{\left(\mathbf{FH} - \mathbf{I}_K\right)\mathbf{C}_{\theta\theta}\left(\mathbf{FH} - \mathbf{I}_K\right)^H\right\} + \sigma_v^2\operatorname{tr}\left\{\mathbf{FF}^H\right\} \\ \mathbf{e}_k^H\ \mathbf{FHC}_{\theta\theta}\ \mathbf{e}_k = \mathbf{e}_k^H\mathbf{C}_{\theta\theta}\mathbf{e}_k \qquad k = 1:K \end{cases}$$

If $\{\theta_k\}_k$ are decorrelated ($\mathbf{C}_{\theta\theta}$ is block-diagonal). The CWCU constraint becomes $\mathbf{e}_k^H\mathbf{FHe}_k = 1$. In this case, the component of interest θ_k is treated as deterministic, whereas the other (correlated) parameter components $\{\theta_p\}_{p\neq k}$ continue to be treated as Bayesian.
Using Lagrange optimization, one can show that the CWCU-LMMSE is given by (see appendix 5.B):

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{cwculmmse} &= \mathbf{D}_{cw}\widehat{\boldsymbol{\theta}}_{lmmse} \\ &= \mathbf{D}_{cw}\mathbf{B}_{lmmse}\widehat{\boldsymbol{\theta}}_{blue} \end{aligned} \tag{5.14}$$

where $\mathbf{D}_{cw} = \left(\operatorname{diag}\left(\mathbf{C}_{\theta\theta}\right)\right)\left(\operatorname{diag}\left(\mathbf{B}_{lmmse}\mathbf{C}_{\theta\theta}\right)\right)^{-1}$ is a diagonal matrix that ensures the component-wise unbiasedness constraint; and $\operatorname{diag}(\mathbf{B}) =$

$\sum_{k=1}^{K} \mathbf{e}_k \left(\mathbf{e}_k^H \mathbf{B} \mathbf{e}_k\right) \mathbf{e}_k^H$ is a $K \times K$ diagonal matrix formed by the diagonal elements of $\mathbf{B}$.

Special cases:

- If the parameters θ_k are decorrelated ($\mathbf{C}_{\theta\theta}$ is diagonal), $\mathbf{D}_{cw}$ can be simplified as:

$$\mathbf{D}_{cw} = \left(\operatorname{diag}\left(\mathbf{B}_{lmmse}\right)\right)^{-1} \tag{5.15}$$

  Thus, the CWCU-LMMSE corresponds to the classic Unbiased-LMMSE.

- If there is no-coupling between the parameters θ_k neither through prior nor through data ($\mathbf{B}_{lmmse}$ is diagonal), the BCWCU-LMMSE corresponds to the BLUE estimation. The CWCU-LMMSE estimation is of interest if there is a coupling through prior ($\mathbf{C}_{\theta\theta}$ is not diagonal), and/or data ($\left(\mathbf{H}^H\mathbf{H}\right)$ is not diagonal).

Reciprocally, one can show that

- If $\mathbf{D}_{cw}$ is diagonal, then $\{\theta_k\}_k$ are decorrelated.
- If $\mathbf{D}_{cw} = \mathbf{B}_{lmmse}^{-1}$, then $\{\theta_k\}_k$ are decoupled.

Remark that from linear multi-user detection ($\mathbf{C}_{\theta\theta}$ is diagonal), the CWCU-LMMSE estimation corresponds to the Unbiased LMMSE; whereas, the (jointly) conditionally unbiased estimator (BLUE) corresponds the MMSE-ZF.
If we suppose that $\boldsymbol{\theta} = [\theta_1^H \cdots \theta_L^H]^H$ can be decomposed on L sub-sets. $\{\theta_l\}_{l=1:L}$ can be either a scalar or vector parameter. The concept of CWCU-LMMSE can be easily generalized to the Block-CWCU-LMMSE. The estimator is computed simply by replacing the diagonal matrices (diag (.)) in the expression of $\mathbf{D}_{cw}$ by block-diagonal matrices (bdiag(.)), and the inverse by the pseudo-inverse [190].

5.5 CWCU linear filtering

Consider two stochastic processes $\{\mathbf{x}_k\}_{k\in Z}$ and $\{\mathbf{y}_k\}_{k\in Z}$ that are correlated. We observe the process $\mathbf{y}_k$ but we are interested in the process $\mathbf{x}_k$ that we cannot observe. The linear estimation of $\mathbf{x}_k$ form $\mathbf{y}_k$ can be formulated as a filtering operation, i.e.,

$$\widehat{\mathbf{x}}_k = \mathbf{F}(q)\mathbf{y}_k = \sum_i \mathbf{F}_j \mathbf{y}_{k-i} \tag{5.16}$$

where $\mathbf{F}(q) = \sum_i \mathbf{F}_i q^{-i}$ is a given linear filter; and q^{-1} is the one sample time delay operator.
We assume that

$$E_{|\mathbf{x}_k} \widehat{\mathbf{x}}_k = \sum_i \mathbf{F}_i \; E_{|\mathbf{x}_k} \mathbf{y}_k = \mathbf{B}\mathbf{x}_k \tag{5.17}$$

where $\mathbf{B}$ represents the filtering Bias. Note that if $\{\mathbf{y}_k, \mathbf{x}_k\}_k$ are jointly Gaussian, or if $\mathbf{y}_k$ is a linear mixture of independent parameters $\mathbf{x}_k$, the previous assumption is valid. Assuming (5.17), one can show that the bias $\mathbf{B}$ is given by

$$\begin{aligned} \mathbf{B} &= E\left[\widehat{\mathbf{x}}_k \mathbf{x}_k^H\right] \; \left(E\left[\mathbf{x}_k \mathbf{x}_k^H\right]\right)^{-1} \\ &= \sum_i \mathbf{F}_i E\left[\mathbf{y}_{i-k} \mathbf{x}_k^H\right] \mathbf{R}_{xx}^{-1} = \oint \mathbf{F}(z) \mathbf{S}_{yx}(z) \frac{dz}{z} \mathbf{R}_{xx}^{-1} \end{aligned} \tag{5.18}$$

where $\mathbf{S}_{yx}(z)$ is the z-domain cross-Power Spectral Density Function (cross-PSDF). Then, one can define the associated CWCU linear filtering by compensating the filtering bias, i.e.,

$$\mathbf{F}_{cw}(q) = \mathbf{D}_{cw} \mathbf{F}(q) \tag{5.19}$$

If $\mathbf{x}_k$ is a stochastic vector process, the notion of "component-wise" can be defined on different levels:
- per vector sample (removing bias using $\mathbf{D}_{cw} = \mathbf{B}^{-1}$).
- per scalar sample (removing bias using $\mathbf{D}_{cw} = (\text{ diag } (\mathbf{R}_{xx}))\,(\text{ diag } (\mathbf{B}\mathbf{R}_{xx}))^{-1}$).
In the following, we will derive the bias update for the Kalman, and Wiener filtering.

5.5.1 CWCU Kalman filtering

Consider the signal process model

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{G}_k \mathbf{u}_k \\ \quad \mathbf{y}_k \quad = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \end{cases} \tag{5.20}$$

where $\mathbf{F}_k$,$\mathbf{G}_k$, and $\mathbf{H}_k$ are given matrices. The initial state $\mathbf{x}_0$, the driving disturbance $\mathbf{u}_k$, and the measurement disturbance $\mathbf{v}_k$ are unknown complex vectors. The output $\mathbf{y}_k$ is assumed to be known for all k. We assume also that $E\left[\mathbf{x}_0\mathbf{x}_0^H\right] = \mathbf{R}_{x,0}$, $E\left[\mathbf{u}_k\mathbf{u}_l^H\right] = \mathbf{Q}\delta_{kl}$, $E\left[\mathbf{v}_k\mathbf{v}_l^H\right] = \mathbf{R}\delta_{kl}$, and $E\left[\mathbf{u}_k\mathbf{v}_l^H\right] = \mathbf{0}$.

The Kalman filter estimates the process $\mathbf{x}_{k+1}$ by using a form of feedback control: the filter estimates the process state at some time ($\widehat{\mathbf{x}}_{k+1/k}$) and then obtains feedback in the form of (noisy) measurements ($\mathbf{y}_{k+1}$). As such, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations. The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the a priori estimates for the next time step. The measurement update equations are responsible for the feedback, i.e., for incorporating a new measurement into the a priori estimate to obtain an improved a posteriori estimate. The time update equations can also be thought of as predictor equations, while the measurement update equations can be thought of as corrector equations. Indeed the final estimation algorithm resembles that of a predictor-corrector algorithm for solving numerical problems (see figure 5.1).

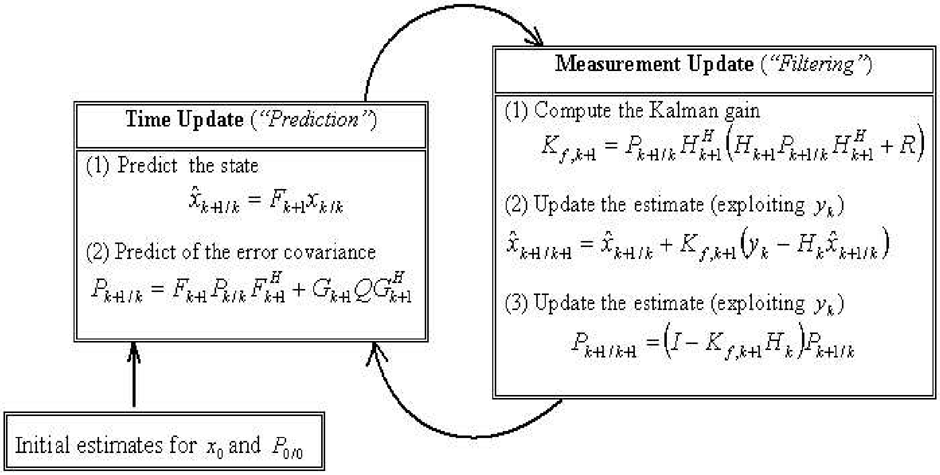
Time Update ("Prediction")
(1) Predict the state
$\hat{x}_{k+1/k} = F_{k+1}x_{k/k}$
(2) Predict of the error covariance
$P_{k+1/k} = F_{k+1}P_{k/k}F_{k+1}^H + G_{k+1}QG_{k+1}^H$
Measurement Update ("Filtering")
(1) Compute the Kalman gain
$K_{f,k+1} = P_{k+1/k}H_{k+1}^H\left(H_{k+1}P_{k+1/k}H_{k+1}^H + R\right)$
(2) Update the estimate (exploiting y_k)
$\hat{x}_{k+1/k+1} = \hat{x}_{k+1/k} + K_{f,k+1}\left(y_k - H_k\hat{x}_{k+1/k}\right)$
(3) Update the estimate (exploiting y_k)
$P_{k+1/k+1} = \left(I - K_{f,k+1}H_k\right)P_{k+1/k}$
Initial estimates for x_0 and $P_{0/0}$

Figure 5.1: Kalman filtering: time and measurement updates.

The Kalman filtering computes the MMSE estimation of the signal process

$\mathbf{x}_k$ given the noisy observations $\{\mathbf{y}_1 \cdots \mathbf{y}_N)$. The basic idea is using the prediction error signal $\{\widetilde{\mathbf{y}}_k\}_k$ (of $\{\mathbf{y}_k\}_k$) to update the MMSE estimate. As $\{\widetilde{\mathbf{y}}_k\}_k$ and $\{\mathbf{y}_k\}_k$ are related by a linear invertible transformation,

$$\widehat{\mathbf{x}}_{k+1/k+1} = E\left\{\mathbf{x}_{k+1}|\mathbf{y}_1 \cdots \mathbf{y}_{k+1}\right\} = E\left\{\mathbf{x}_{k+1}|\widetilde{\mathbf{y}}_1 \cdots \widetilde{\mathbf{y}}_{k+1}\right\} \tag{5.21}$$

And as $\{\widetilde{\mathbf{y}}_k\}_k$ forms an orthogonal family,

$$\begin{aligned} \widehat{\mathbf{x}}_{k+1/k+1} &= \sum_{i=0}^{k+1} E\left\{\mathbf{x}_{k+1}\widetilde{\mathbf{y}}_i^H\right\}\left(E\left\{\widetilde{\mathbf{y}}_i\widetilde{\mathbf{y}}_i^H\right\}\right)^{-1}\widetilde{\mathbf{y}}_i \\ &= \widehat{\mathbf{x}}_{k+1/k} + \underbrace{E\left\{\mathbf{x}_{k+1}\widetilde{\mathbf{y}}_{k+1}^H\right\}\left(E\left\{\widetilde{\mathbf{y}}_{k+1}\widetilde{\mathbf{y}}_{k+1}^H\right\}\right)^{-1}}_{K_{f,k+1}}\widetilde{\mathbf{y}}_{k+1} \end{aligned} \tag{5.22}$$

where $\widehat{\mathbf{x}}_{k+1/k} = E\left\{\mathbf{x}_{k+1}|\mathbf{y}_1 \cdots \mathbf{y}_k\right\}$ is the predicted value of $\mathbf{x}_{k+1}$ given the observations $\{\mathbf{y}_1 \cdots \mathbf{y}_k\}$. One can also show that the Kalman gain can be updated using [10]:

$$\begin{aligned} \mathbf{K}_{f,k+1} &= \mathbf{P}_{k+1/k}\mathbf{H}_k^H\left(\mathbf{H}_k\mathbf{P}_{k+1/k}\mathbf{H}_k^H\right)^{-1} \\ \mathbf{P}_{k+1/k} &= \mathbf{F}_{k+1}\mathbf{P}_{k/k}\mathbf{F}_{k+1}^H + \mathbf{G}_{k+1}\mathbf{Q}\mathbf{G}_{k+1}^H \\ \mathbf{P}_{k/k} &= \left(\mathbf{I} - \mathbf{K}_{f,k}\mathbf{H}_k\right)\mathbf{P}_{k/k-1} \end{aligned} \tag{5.23}$$

where $\mathbf{P}_{k+1/k} = E\left\{\left(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1/k}\right)\left(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1/k}\right)^H\right\}$ and $\mathbf{P}_{k+1/k+1} = E\left\{\left(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1/k+1}\right)\left(\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1/k+1}\right)^H\right\}$ represent respectively the a priori and the a posteriori estimate error covariance.

Using Kalman update equations (5.23), one can show that the Kalman Bias can be computed and updated using:

$$\begin{aligned} \mathbf{R}_{x,k+1} &= \mathbf{F}_k\mathbf{R}_{x,k}\mathbf{F}_k^H + \mathbf{G}_k\mathbf{Q}\mathbf{G}_k^H \\ \mathbf{B}_{k+1}^{pred} &= \left(\mathbf{F}_k\mathbf{B}_k^{filt}\mathbf{R}_{x,k}\mathbf{F}_k^H\right)\mathbf{R}_{x,k+1}^{-1} \\ \mathbf{B}_{k+1}^{filt} &= \left(\mathbf{I} - \mathbf{K}_{f,k+1}\mathbf{H}_{k+1}\right)\mathbf{B}_{k+1}^{pred} + \mathbf{K}_{f,k+1}\mathbf{H}_{k+1} \end{aligned} \tag{5.24}$$

where $\mathbf{R}_{x,k} = E\left(\mathbf{x}_k\mathbf{x}_k^H\right)$ is the correlation matrix of $\mathbf{x}_k$, $\mathbf{B}_{k+1}^{pred}, \mathbf{B}_{k+1}^{filt}$ represent respectively the time and the measurement update of the Kalman bias matrices.

5.5.2 CWCU Wiener filtering

Consider the signal process model

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k \tag{5.25}$$

where $\mathbf{x}_k$ is the desired signal to be estimated, and $\mathbf{v}_k$ represents an additive noise (assumed to have zero mean, and to be uncorrelated with the signal of interest $\mathbf{x}_k$).
The problem is performing the MMSE estimation of $\mathbf{x}_k$ given the observations $\{\mathbf{y}_1 \cdots \mathbf{y}_k\}$. One can show that if the z-domain PSDF of $\mathbf{y}_k$ is non-singular on the unit circle, it can be decomposed onto:

$$\mathbf{S}_{yy}(z) = \mathbf{A}^{-1}(z)\boldsymbol{\Sigma}\mathbf{A}^{-\dagger}(z) \tag{5.26}$$

where $\mathbf{S}_{yy}(z)$ represents the PSDF of the observed signal, $\mathbf{A}\left(z\right)$ denotes the optimal prediction filter for the observed signal y_k, and $\boldsymbol{\Sigma}$ is the associate prediction error variance.
The causal and non-causal Wiener filters are then given by [10]:

$$\begin{aligned}
\mathbf{F}_{wiener}(z) &= \mathbf{S}_{xy}(z)\mathbf{S}_{yy}^{-1}(z)\\
\mathbf{F}_{wiener}^{causal}(z) &= \mathbf{I} - \mathbf{S}_{vv}(z)\boldsymbol{\Sigma}\mathbf{A}(z)
\end{aligned}$$

where $S_{vv}(z)$ is the PSDF of the noise signal. Using (5.18), one can show that the bias of the Wiener, and the causal Wiener filters are given by

$$\begin{aligned}
\mathbf{B}_{wiener} &= \left(\oint \mathbf{S}_{xx}(z)\mathbf{S}_{yy}^{-1}(z)\mathbf{S}_{xx}(z)\frac{dz}{z}\right)\mathbf{R}_x^{-1}\\
B_{wiener}^{causal} &= \left(\oint \left(\mathbf{I} - \mathbf{S}_{vv}(z)\boldsymbol{\Sigma}^{-1}\mathbf{A}(z)\right)\mathbf{S}_{xx}(z)\frac{dz}{z}\right)\mathbf{R}_x^{-1}
\end{aligned}$$

If we assume the observed process follow the model (5.20). If we assume the $\mathbf{F}_k$, $\mathbf{G}_k$, and $\mathbf{H}_k$ are time invariant, one can show that in steady state we have

$$\begin{aligned}
\mathbf{S}_{xx}(z) &= H\left(z\mathbf{I} - \mathbf{F}\right)^{-1}\mathbf{G}\mathbf{Q}\mathbf{G}^H\left(z\mathbf{I} - \mathbf{F}^H\right)^{-1}\mathbf{H}^H\\
\mathbf{S}_{vv}(z) &= \mathbf{R}\\
\mathbf{S}_{xx}(z) &= \mathbf{H}\mathbf{S}_{xx}(z)\mathbf{H}^H + \mathbf{R}
\end{aligned} \tag{5.27}$$

Anderson and Moore show that the causal Wiener solution of the prediction problem is [10]

$$\mathbf{H}_{wiener}^{causal}(z) = \left(z\mathbf{I} - \left(\mathbf{F} - \mathbf{K}_p\mathbf{H}\right)\right)^{-1}\mathbf{K}_p\mathbf{H} \tag{5.28}$$

where $\mathbf{K}_p = \mathbf{F}\mathbf{K}_f$ denotes the Kalman prediction gain. The bias matrix can be then extended as:

$$\mathbf{B}^{pred}\mathbf{R}_x = \sum_{k=0}^{\infty} \left(\mathbf{F} - \mathbf{K}_p\mathbf{H}\right)^k \mathbf{K}_p\mathbf{H}\left(\mathbf{F}^H\right)^{k+1} \tag{5.29}$$

which is consistent with the results derived using (5.24), corresponding to the steady state Kalman Bias.

5.6 Interplay between global bias and prior covariance rank for BCWCU-LMMSE

As we have seen previously, imposing a joint conditionally unbiasedness constraint on a vector of parameters reduces Bayesian estimation to deterministic parameter estimation. However, in Block-CWCU Bayesian parameter estimation, every (set of) parameter in turn is treated as deterministic while the others are being treated as Bayesian. This leads to an intermediate approach between the classic deterministic and Bayesian approaches.
In this section, we consider the case where the prior covariance matrix has a limited rank. We investigate the interplay between block-size, joint bias and prior covariance rank. And, we show that B-CWCU-LMMSE, with appropriate block sizes, reduces the estimation noise, while guaranteeing joint unbiasedness. The result will be illustrated through a concrete example.

Consider a Base Station (BS) using M-element antenna array. The received signal over single-path propagation is an $M \times 1$ vector given by:

$$\mathbf{y}(k) = x(k)\mathbf{H}\mathbf{A}\left(\theta\right) + \mathbf{v}(k) \tag{5.30}$$

where $x(k)$ is a known scalar training sequence transmitted by the user, $\mathbf{H}$ in a $M \times M$ known matrix describing the coupling between the antenna elements, $\mathbf{v}(k)$ is an additive white Gaussian noise, i.e.,$\sim \mathcal{N}(0, \sigma_v^2\mathbf{I}_M)$, and $\mathbf{A}\left(\theta\right)$ denotes the array response (function of the array geometry, and the direction of arrival θ).

The Direction of Arrival (DoA) θ is generally estimated using a two step approach:

1. Estimate the array response vector $\mathbf{A}(\theta)$.

2. Compute the DoA based on the array manifold $\mathbf{A}(.)$.

In the literature, the Least-Squares (LS) technique (which corresponds to BLUE in this problem) is proposed for the estimation of the array response vector [130, 131, 172]

$$\widehat{\mathbf{A}}_{blue} = \sigma_x^{-2} \left(\mathbf{H}^H \mathbf{H}\right)^{-1} \mathbf{H}^H \sum_k x^*(k)\mathbf{y}(k) \tag{5.31}$$

where $\sigma_x^2 = \sum_k |x(k)|^2$ represents the energy of the training sequence $\{x(k)\}_k$. As we have seen in the previous section, BLUE provides an unbiased, but noisy estimate, i.e.

$$E_{|A}\left\{\widehat{\mathbf{A}}_{blue}\right\} \;=\; \mathbf{A}(\theta) \tag{5.32}$$

On the other hand, if prior information is available, it can be used to enhance the estimation SNR. In the following, we will investigate the effect of the use of a Bayesian prior on the estimation bias. We assume that the direction of arrival is varying around an unknown nominal DoA θ_0, i.e.,

$$\theta \;=\; \theta_0 \;+\; \delta\theta \tag{5.33}$$

And, we will have to estimate multiple instances of θ. Using a first order approximation, we have

$$\mathbf{A}\left(\theta\right) \;\approx\; \mathbf{A}\left(\theta_0\right) \;+\; \delta\theta \;\mathbf{A}'\left(\theta_0\right) \tag{5.34}$$

where $\mathbf{A}'\left(\theta_0\right) = \left.\frac{\partial \mathbf{A}(\theta)}{\partial \theta}\right]_{|\theta=\theta_0}$ denotes the gradient of $\mathbf{A}(\theta)$ at $\theta = \theta_0$. $\mathbf{A}(\theta)$ is random due to $\delta\theta$. Assuming $\delta\theta$ to have zero mean and variance σ_δ^2, the covariance matrix of $\mathbf{A}(\theta)$ becomes:

$$\begin{aligned}\mathbf{C}_A &= E\left\{\mathbf{A}(\theta)\mathbf{A}^H(\theta)\right\}\\ &= \mathbf{A}(\theta_0)\mathbf{A}^H(\theta_0) + \sigma_\delta^2 \mathbf{A}'(\theta_0)\mathbf{A}'^H(\theta_0).\end{aligned}$$

As $\text{rank}(\mathbf{C}_A) = 2$, using the eigen decomposition, the prior covariance matrix can be written as:

$$\mathbf{C}_A = \mathbf{U}\left[\begin{array}{cc}\mathbf{\Lambda}_C & 0\\ 0 & 0\end{array}\right]\mathbf{U}^H \tag{5.35}$$

where $\mathbf{\Lambda}_c$ is a 2×2 diagonal matrix, and $\mathbf{U}$ is a unitary matrix.
By introducing $\mathbf{U}_A = \mathbf{U} \left[\begin{array}{cccc} \mathbf{I}_2 & 0 & \cdots & 0 \end{array} \right]^H$, $\mathbf{C}_A$ can be simplified to

$$\mathbf{C}_A = \mathbf{U}_A \; \mathbf{\Lambda}_C \mathbf{U}_A^H \tag{5.36}$$

Note that $\mathbf{A}(\theta_0)$, $\mathbf{A}'(\theta_0)$ are unknown. Only the covariance $\mathbf{C}_A$ (and $\mathbf{U}_A$) is known. Remark also that $\mathbf{A}(\theta)$ lives in the subspace spanned by $\mathbf{U}_A$. Then, we can introduce zero-mean random variables η, and γ such that

$$\mathbf{A}(\theta) = \mathbf{U}_A \left[\begin{array}{c} \eta \\ \gamma \end{array} \right]. \tag{5.37}$$

From (5.13), one can show that

$$\mathbf{B}_{lmmse}\mathbf{U}_A = \mathbf{U}_A \underbrace{\left(\mathbf{\Lambda}_C \mathbf{\Lambda}_H + \sigma_v^2 \mathbf{I}_2 \right)^{-1} \mathbf{\Lambda}_C \mathbf{\Lambda}_H}_{\mathbf{\Lambda}_B} \tag{5.38}$$

where $\mathbf{\Lambda}_H = \mathbf{U}_A \left(\mathbf{H}^H \mathbf{H} \right) \mathbf{U}_A^H$ is a 2×2 matrix, not necessarily diagonal. Then, the expected value of the LMMSE estimate is

$$\begin{aligned} E_{|A} \left\{ \widehat{\mathbf{A}}_{lmmse} \right\} &= \mathbf{B}_{lmmse} \mathbf{A}(\theta) \\ &= \mathbf{U}_A \mathbf{\Lambda}_B \left[\begin{array}{c} \eta \\ \gamma \end{array} \right] = \mathbf{U}_A \left[\begin{array}{c} \eta' \\ \gamma' \end{array} \right]. \end{aligned} \tag{5.39}$$

In summary, the LMMSE estimates the array response in the right subspace, but with biased weighting. If $\mathbf{\Lambda}_B$ is not a multiple of identity, the LMMSE estimate of $\mathbf{A}(\theta)$ leads to erroneous DoA estimation.

If we impose Block-CWCU constraints, under some regularity assumptions, one can show that using a block-size 2 ($L_k \geq 2 \;\; \forall k$) (see appendix 5.C):

$$\mathbf{D}_{bcw} \mathbf{B}_{lmmse} \mathbf{U}_A = \mathbf{D}_{bcw} \mathbf{U}_A \mathbf{\Lambda}_B = \mathbf{U}_A$$

Thus, the BCWCU-LMMSE (with a bloc-size 2), guarantees **joint** unbiasedness, i.e., the expected value of the BCWCU-LMMSE estimate is

$$\begin{aligned} E_{|A} \left\{ \widehat{\mathbf{A}}_{bcwculmmse} \right\} &= \mathbf{D}_{cw} \mathbf{B}_{lmmse} \mathbf{A}(\theta) \\ &= \mathbf{U}_A \left[\begin{array}{c} \eta \\ \gamma \end{array} \right] = \mathbf{A}(\theta) \end{aligned} \tag{5.40}$$

In figure 5.2, we plot the estimation MSE $=$ tr $\left\{\mathbf{C}_{\tilde{A}\tilde{A}}\right\}$ of the BLUE, LMMSE, and BCWCU-LMMSE estimates. The MSE is averaged over 500 Monte Carlo runs. The matrices $\mathbf{C}_{\theta\theta}$ (having rank 2), and $\mathbf{H}$ are generated randomly. M was chosen equal to 10.

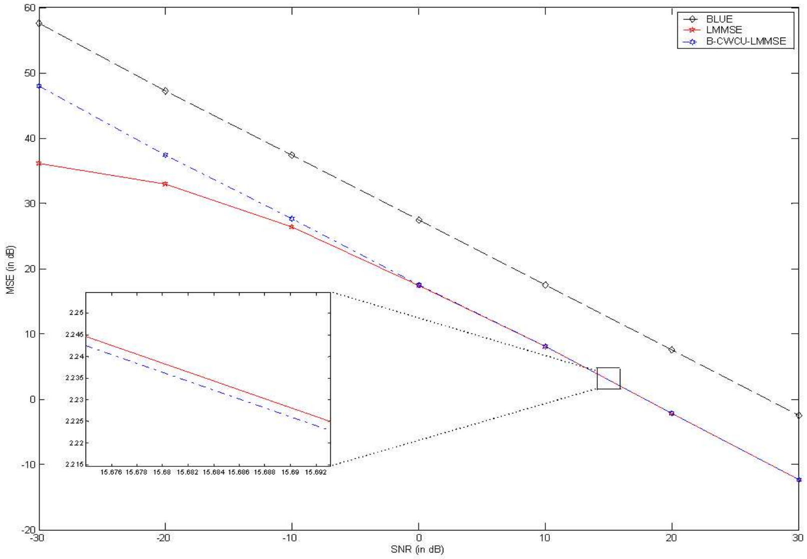

Figure 5.2: Estimation MSE of the BLUE, LMMSE, and BCWCU-LMMSE estimators as a function of SNR.

Thus, for the limited rank prior covariance matrix case, BCWCU-LMMSE reduces the estimation noise, while guaranteeing joint unbiasedness.
The result can be easily generalized to an arbitrary prior covariance rank. This leads to the following theorem.

Theorem: Let m denote the rank of the prior covariance matrix $\mathbf{C}_{\theta\theta}$. Then BCWCU-LMMSE , with block sizes of at least m, guarantees the joint unbiasedness.

5.7 CWCU-LMMSE: application to multiple channel estimation

In this section, we will focus on one particular problem setting, in which the channels from different Base Stations (BSs) to a Mobile Station (MS) need to be estimated jointly. The estimation of the transmission channel plays a crucial role in communication systems (for mobile positioning applications, multi-user detection...).

5.7.1 IPDL method for multiple channel estimation

In Code Division Multiple Access (CDMA) scheme, the fundamental problem is that because of the near-far problem it is difficult to hear multiple BSs. In downlink transmission, the received signal strength, when coming from a distant BS can be quite weak, especially when the mobile terminal is close to the serving BS. This situation is usually referred to as the hearability problem. In order to improve the hearability of neighboring BSs, the serving BS provides idle periods in continuous or burst mode. This technique is known as Idle Period-Down Link transmission (IPDL). The idle periods are short and arranged in a pseudo random way made known to all MSs in advance. The pseudo randomness assures that the effect of simultaneous idle periods in adjacent BSs is minimized. The length of the idle periods is a parameter, which the operator can change to trade off positioning response time and accuracy against capacity loss in the DL. With longer idle periods, the achievable accuracy would be better because of longer integration time at the MS, but the system capacity would be reduced and some assumptions about the channel model can't take the way. During these periods the serving BS completely ceases its transmission and the MS is scheduled to make the needed measurements from the neighbor BSs now hearable. By supporting the IPDL, the localization performance in MS will improve, as there will be less interference present during idle periods.

An example of IPDL method has been shown in figure 5.3. When BS#1 entered in idle period i, the MSs in BS#1 could detect other BSs (i.e. from BS#2 to BS#n, where n is an uncertain number and the number of neighbor BSs) signal. In that time the other BSs did not anything else but just transmit their CPICH (Common PIlot Channel) and other downlink channels.

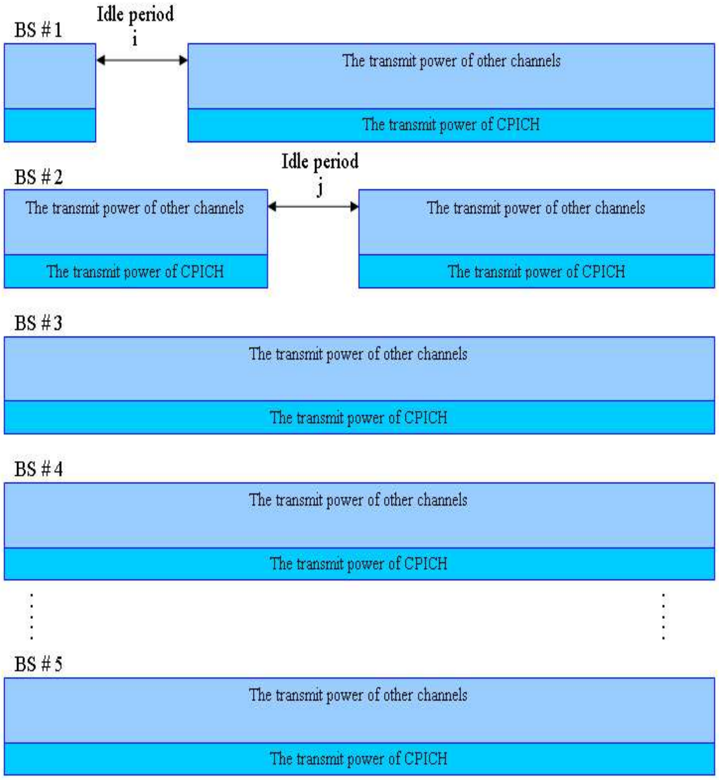
BS # 1
Idle period
i
The transmit power of other channels
The transmit power of CPICH
BS # 2
Idle period
j
The transmit power of other channels
The transmit power of CPICH
The transmit power of other channels
The transmit power of CPICH
BS # 3
The transmit power of other channels
The transmit power of CPICH
BS # 4
The transmit power of other channels
The transmit power of CPICH
BS # 5
The transmit power of other channels
The transmit power of CPICH

Figure 5.3: IPDL method for multiple channel estimation.

Generally, the idle period leads to a tradeoff between "the capacity loss" and "the estimation noise". In fact, the length of the idle period should be as short as possible to ensure that the capacity loss is minimized, but enough to allow acceptable channel estimation accuracy. The use of prior power delay profile statistical information can be advantageous; and leads to a better "capacity loss" vs. "estimation noise" tradeoff.

5.7.2 Block-CWCU-LMMSE to multiple channel estimation

In this section, we will focus on one particular problem setting, in which the channels from different Base Stations (BSs) to a Mobile Station (MS) need to be estimated jointly. Channel parameters are observed indirectly by the received data : convolved with a known training sequence and embedded in a (white Gaussian) noise.

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_k \mathbf{h}_k + \mathbf{v} \tag{5.41}$$

where

- $\mathbf{y} = \left[y_1^H \cdots y_N^H\right]^H$ denotes received data. N is the data length.
- $\mathbf{v} = \left[v_1^H \cdots v_N^H\right]^H$ represents the additive white gaussian noise.
- K is the number of base stations.
- $\mathbf{h}_k = \left[h_{k,1}^H \cdots h_{k,L_k}^H\right]^H$ denotes the Channel Impulse Response (CIR) between the MS and the k^{th} BS. L_k is k^{th} CIR length.
- $\mathbf{X}_k = \begin{bmatrix} x_1 & \cdots & x_L \\ \vdots & & \vdots \\ x_N & \cdots & x_{N+L-1} \end{bmatrix}$ is an $N \times L_k$ Hankel matrix characterizing the training sequence of the k^{th} BS.

Using a compact notation, the received data can be written as:

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{v} \tag{5.42}$$

where $\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_K]$, and $\mathbf{h} = \left[\mathbf{h}_1^H \cdots \mathbf{h}_K^H\right]^H$.

Note that the problem has a special structure. In fact, the channel impulse responses and their individual coefficients are decorrelated ($\mathbf{C}_{hh}$ is diagonal). On the other hand, the data covariance matrix $\mathbf{X}^H\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^H\mathbf{X}_1 & \cdots & \mathbf{X}_1^H\mathbf{X}_K \\ \vdots & & \vdots \\ \mathbf{X}_K^H\mathbf{X}_1 & \cdots & \mathbf{X}_K^H\mathbf{X}_K \end{bmatrix}$ can not be assumed to be block-diagonal due to:

- The limited length of the training sequence (the channel estimation is done only in the Idle Period Down-Link (IPDL)). Thus, despite the input being white, the training sequence is not long enough to lead to a spherical estimate of the input covariance matrix $\mathbf{X}^H\mathbf{X}$.

- The range of the CIR powers. In fact, despite the quantities $\mathbf{X}_k^H\mathbf{X}_k$ being approximately white ($\approx \sigma_k^2\mathbf{I}_{L_k}$), e.g. $\mathbf{X}_1^H\mathbf{X}_K$ can not be neglected with respect to $\mathbf{X}_K^H\mathbf{X}_K$.

Whereas the direct use of a Bayesian channel estimate for an interfering signal allows to better suppress the interference, its use for the user of interest may lead to a bias problem. This bias is detrimental for a number of applications. For example, the estimation bias is undesirable for mobile localization applications (e.g. Time of Arrival (ToA) is estimated by fitting a parametric model to the channel impulse response) [188]. That is why the Bayesian prior is rarely taking into account for such applications. Channel estimation is done typically based on Least-Squares (LS) or Matching Pursuit (MP) approaches [130, 172, 92, 38]. On the other hand, imposing joint unbiasedness between the CIRs coming from different base-stations is not required: we can allow for interference (contribution of other base-stations), if this can be motivated by a noise reduction.

As we have seen previously, even if the prior covariance matrix is diagonal, the channel impulses are coupled through the data covariance matrix $\mathbf{X}^H\mathbf{X}$ and then, the Block-CWCU-LMMSE is of interest.
The LMMSE estimate is given by:

$$\begin{aligned}\widehat{\mathbf{h}}_{lmmse} &= \underbrace{\left(\sigma_v^2\left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{C}_{hh}^{-1}+\mathbf{I}\right)^{-1}}_{\mathbf{B}_{lmmse}}\left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{y}\\ &= \mathbf{B}_{lmmse}\;\widehat{\mathbf{h}}_{blue}\end{aligned}$$

The B-CWCU constraint is formulated as

$$E\left[\widehat{\mathbf{h}}_k|\mathbf{h}_k\right] = \mathbf{h}_k \qquad k=1:K \tag{5.43}$$

As the prior covariance matrix is diagonal, minimizing the MSE (under B-CWCU constraints) leads to

$$\widehat{\mathbf{h}}_{bcwculmmse} = \left(\mathrm{bdiag}\left(\mathbf{B}_{lmmse}\right)\right)^{-1}\mathbf{B}_{lmmse}\;\widehat{\mathbf{h}}_{blue} \tag{5.44}$$

5.7.3 SIC implementation of the BCWCU-LMMSE estimator

The inherent complexity of the B-CWCU-LMMSE scheme is cubic in $L = \sum_k L_k$ (the same as for the LMMSE and the BLUE estimators). For practical implementation, the Successive Interference Cancellation (SIC) approach can be used to approximate the BCWCU-LMMSE estimator, with a complexity linear in L.

Successive interference cancellation multi-channel estimation is a scheme in which CIR's are estimated successively. The approach successively cancels the interference from the next strongest channel. Assume that channels have been ordered in order of decreasing $SNR_i = \frac{\sigma_x^2 \left\|\mathbf{h}_i\right\|^2}{\sigma_v^2} = \frac{\sigma_i^2}{\sigma_v^2}$ at the channel estimator input. First, we compute an unbiased estimate of the first (strongest) CIR (the BLUE is proportional to the matched filter). The contribution of weaker CIRs is ignored, i.e.,

$$\widehat{\mathbf{h}}_1 = \frac{1}{N\sigma_x^2}\mathbf{X}_1^H\mathbf{y} \tag{5.45}$$

Then, the LMMSE estimator is derived, the interfering signal is recreated at the receiver, and subtracted from the received waveform.

$$\begin{aligned}
&\widehat{\mathbf{h}}_1^{SIC} = \left(\sigma_v^2\mathbf{C}_{hh,1}^{-1} \;+\; \mathbf{X}_1^H\mathbf{X}_1\right)^{-1}\left(N\sigma_x^2\right)\widehat{\mathbf{h}}_1\\
&\widehat{\mathbf{y}}_1 = \mathbf{X}_1\widehat{\mathbf{h}}_1^{SIC}\\
&\mathbf{y} \leftarrow \mathbf{y} - \widehat{\mathbf{y}}_1
\end{aligned}$$

Remark that even if $\mathbf{X}^H\mathbf{X}$ is not approximately diagonal, the non-diagonal elements of $\mathbf{X}_1^H\mathbf{X}_1$ can be neglected (as the number of unknowns is M times less). One recursion of the SIC implementation of the BCWCU-LMMSE is described in the table below:

In this manner successive BS CIRs does not have to encounter interference caused by initial BS CIRs. SIC leads to good performance for all channel estimates: initial CIR estimates improve because the later channels have less power which means less interference for the initial channels, and later CIR estimates improve because early BS's interference has been cancelled out. Figure 5.4 shows that the SIC well approximates the B-CWCULMMSE estimator (specially for low SNR).

5.7.4 Modified SIC implementation of the BCWCU-LMMSE estimator

In the linear SIC approach above there are two sources of error:

SIC implementation of the BCWCU-LMMSE		
#	**Computation**	**cost**
channel estimation		
1	$\widehat{\mathbf{h}}_k = \frac{1}{N\sigma_k^2}\mathbf{X}_k^H\mathbf{y}$	$O(NL_k)$
Interference cancellation		
2	$\widehat{\mathbf{h}}_k^{lmmse} = \left(\frac{\sigma_v^2}{N\sigma_x^2}\mathbf{C}_{hh,k}^{-1} + \mathbf{I}_{L_k}\right)^{-1}\widehat{\mathbf{h}}_k$	$O(L_k)$
3	$\widehat{\mathbf{y}}_k = \mathbf{X}_k\widehat{\mathbf{h}}_k^{lmmse}$	$O(NL_k)$
4	$\mathbf{y} = \mathbf{y} - \widehat{\mathbf{y}}_k$	

Table 5.1: SIC implementation of the BCWCU-LMMSE

- Ignoring the contribution of channels with lower powers.

- Non-perfect cancellation of estimated channels.

In the following, we will try to alleviate the propagation of the estimation error (due to non-perfect interference cancellation). We suggest taking, at each step k, the estimate $\widehat{\mathbf{h}}_k$ computed from the joint LMMSE estimation of $\mathbf{h}^{(k)} = \left[\mathbf{h}_1^H \cdots \mathbf{h}_k^H\right]^H$. As in the classic SIC approach, the contribution of channels with lower powers $\bar{\mathbf{h}}^{(k)} = \left[\mathbf{h}_{k+1}^H \cdots \mathbf{h}_K^H\right]^H$ is ignored.

The LMMSE solutionis given by

$$\widehat{\mathbf{h}}^{LMMSE,(k)} = \underbrace{\left(\mathbf{C}_{hh}^{(k)^{-1}} + \frac{1}{\sigma_v^2}\mathbf{X}^{(k)^H}\mathbf{X}^{(k)}\right)^{-1}}_{\mathbf{B}_k^{-1}} \underbrace{\frac{1}{\sigma_v^2}\mathbf{X}^{(k)^H}\mathbf{y}}_{\mathbf{y}_{MF}^{(k)}}$$

where $\mathbf{C}_{hh}^{(k)} = E\left\{\mathbf{h}^{(k)}.\ \mathbf{h}^{(k)H}\right\}$, and $\mathbf{X}^{(k)} = [\mathbf{X}_1 \cdots \mathbf{X}_k]$.

By denoting $\mathbf{b}_k = \left(\mathbf{C}_{hh,k}^{-1} + \frac{1}{\sigma_v^2}\mathbf{X}_k^H\mathbf{X}_k\right)$, and

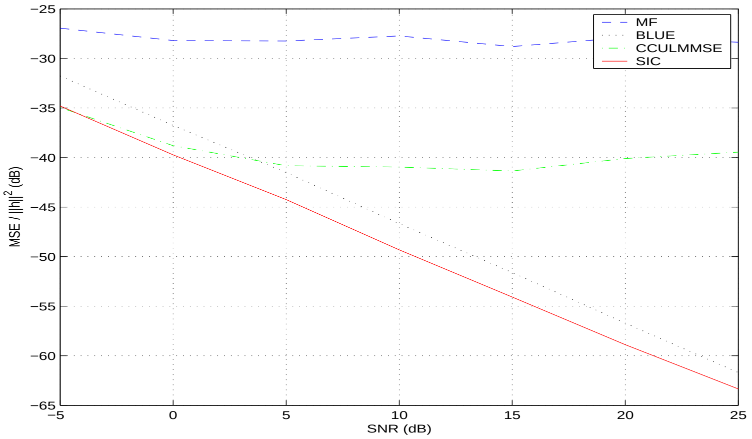

Figure 5.4: CIR estimation accuracy using MF, BLUE, CWCULMMSE, and SIC estimators.

$\mathbf{R}_{k-1,k} = \frac{1}{\sigma_v^2}\mathbf{X}^{(k-1)^H}\mathbf{X}_k$, $\mathbf{B}_{k+1}$ can be decomposed as

$$\begin{aligned}\mathbf{B}_{k+1} &= \begin{bmatrix} \mathbf{B}_k & \mathbf{R}_{k,k+1} \\ \mathbf{R}_{k,k+1}^H & \mathbf{b}_{k+1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{R}_{k,k+1}^H\mathbf{B}_k^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{B}_k & 0 \\ 0 & \mathbf{b}_{k+1} - \mathbf{R}_{k,k+1}^H\mathbf{B}_k^{-1}\mathbf{R}_{k,k+1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{B}_k^{-1}\mathbf{R}_{k,k+1} \\ 0 & \mathbf{I} \end{bmatrix}\end{aligned} \tag{5.46}$$

Then, the component of interest is given by:

$$\begin{aligned}\widehat{\mathbf{h}}_{k+1}^{mod-SIC} &= [0\cdots 0\ \mathbf{I}]\ \ \mathbf{B}_{k+1}^{-1}\ \ \mathbf{y}_{MF}^{(k+1)} \\ &= \frac{1}{\sigma_v^2}\left(\mathbf{b}_{k+1} - \mathbf{R}_{k,k+1}^H\mathbf{B}_k^{-1}\mathbf{R}_{k,k+1}\right)^{-1}\mathbf{X}_{k+1}^H\left(\mathbf{y} - \mathbf{X}^{(k)}\mathbf{B}_k^{-1}\mathbf{y}_{MF}^{(k)}\right)\end{aligned}$$

We recognize the same structure as in the classic SIC. The modified SIC algorithm is described in the table below

Remark that the complexity of the scheme is $O(L^3)$ (as BCWCU-LMMSE). From this point of view, it presents no advantage. However, the performance of the proposed scheme can be interpreted as a bound on the performance of the SIC approach in section 4.1 (there is no propagation of the estimation error).

Motivated by the fact that channels are ordered by decreasing power and the observation that $\mathbf{X}_k^H\mathbf{X}_k$ is approximately proportional to the identity

Modified SIC implementation of the BCWCU-LMMSE		
#	**Computation**	**cost**
channel estimation		
1	$\widehat{\mathbf{h}}_k = \frac{1}{N\sigma_x^2}\mathbf{X}_k^H\mathbf{y}^{(k)}$	$O(NL_k)$
Interference cancellation		
2	$\mathbf{P}_k = \left(\mathbf{b}_k - \mathbf{R}_{k-1,k}^H\mathbf{B}_{k-1}^{-1}\mathbf{R}_{k-1,k}\right)$	$O(L_k(\sum_{p=1}^{k-1}L_p)^2)$
3	$\widehat{\mathbf{h}}_k^{mod-SIC} = \frac{N\sigma_x^2}{\sigma_v^2}\mathbf{P}_k^{-1}\widehat{\mathbf{h}}_k$	$O(L_k^3)$
4	$\widehat{\mathbf{y}}_k = \mathbf{X}_k\widehat{\mathbf{h}}_k^{mod-SIC}$	$O(NL_k)$
5	$\mathbf{y} = \mathbf{y} - \widehat{\mathbf{y}}_k$	
6	$\mathbf{B}_k^{-1}$ (updated using MIL on (24))	$O(L_k(\sum_{p=1}^{k-1}L_p)^2)$

Table 5.2: Modified SIC implementation of the BCWCU-LMMSE

matrix, we approximate $\mathbf{B}_k$ and $\mathbf{b}_{k+1}$ in (5.46) by diagonal matrices. We call the resulting scheme Modified & Simplified SIC, It has a computational complexity of $O(L^2)$.

We analyze the performance of the proposed algorithms by comparing their estimation MSE (computed by Monte Carlo simulations). The received signal is assumed to be the superposition of the contribution of 5 base stations, and embedded in a white Gaussian noise. The relative received signal powers are respectively 0, -5, -10, -15, -20 dB. The power delay profile is generated according to the channel model "Vehicular B". Figure 5.5 plots the curves of the estimation MSE of the 5^{th} (the weakest) BS. The curves show that the SIC implementations well approximate the BCWCU-LMMSE estimator at low SNR. We remark also that the simplifications introduced to the modified scheme do not affect the estimation accuracy, and that the modified SIC outperforms the classic one (at the expense of additional complexity).

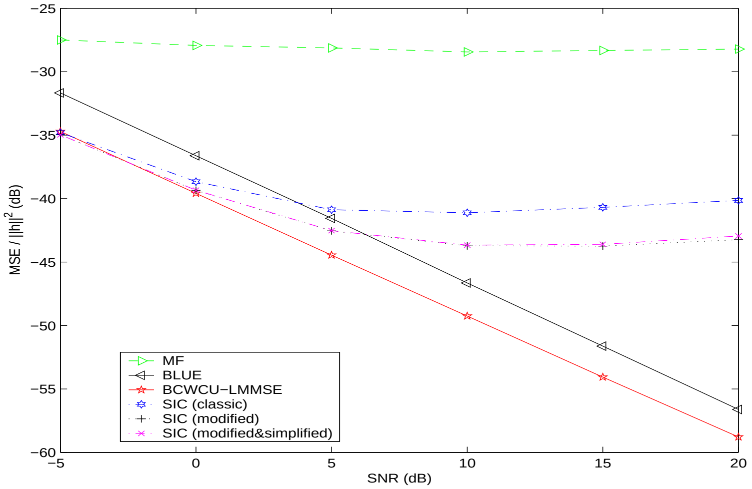

Figure 5.5: CIR estimation accuracy using MF, BLUE, BCWCU-LMMSE, SIC, Modified SIC estimators.

5.7.5 Power delay profile adaptation using the EM algorithm

The estimation of the statistical parameter of the transmission channel plays a crucial role in the CIR estimation procedure. Unlike the channel impulse response, the Power Delay Profile changes very slowly; and can be estimated with a good accuracy from the received data. In this section, the identification of the Power Delay Profile model is based on the concept of expectation maximization (EM) and an iterative optimization algorithm to produce maximum-likelihood (ML) estimates under certain conditions. The EM procedure is divided into two steps:

- The expectation step (E-step), computes the conditional expectation of unobserved sufficient information (complete data), under given observed insufficient information (incomplete data) and the current estimation of the parameters.

- The maximization step (M-step), provides the new estimate of parameters by maximizing the conditional expectation over unknown parameters.

As in [54], we propose using an adaptive EM Algorithm to jointly update the LMMSE channel estimates and the power delay profile parameters. The resulting algorithm (using the first SIC B-CWCU-LMMSE implementation) is listed in the table below (k denotes the idle frame index).

One can show that the power delay profile $\widehat{\mathbf{C}}_{hh,k}$ are updated such that:

$$\mathbf{C}_{hh,k} \;=\; \sum_{j=0} \lambda^j \left(\mathbf{C}_{\widetilde{h}\widetilde{h},k} + \widehat{\mathbf{h}}_{LMMSE,k}\widehat{\mathbf{h}}_{LMMSE,k}^H\right)$$

where in the uncorrelated channel coefficients case $\mathbf{C}_{hh,k}$ converges to a diagonal matrix.

PDP Estimation via Adaptive EM Algorithm	
#	**Computation**
Initialization	
1	$\widehat{\mathbf{h}}_{0/0} = 0, \mathbf{P}_{0/0} = \mathbf{I}, \mathbf{C}_{hh}^{(0)} = \mathbf{I}, \mathbf{M}^{(0)} = 0, \gamma^{(0)} = 0$
LMMSE estimation	
2	$\mathbf{C}_{\widetilde{h}\widetilde{h},k} = \left(\mathbf{C}_{hh,k-1}^{-1} \;+\; \frac{1}{\sigma_v^2}\,\mathbf{X}_k^H\mathbf{X}_k\right)^{-1}$
3	$\widehat{\mathbf{h}}_{LMMSE,k} = \mathbf{C}_{\widetilde{h}\widetilde{h},k}\,\frac{1}{\sigma_v^2}\,\mathbf{X}^H\mathbf{y}$
Adaptive EM parameter estimation	
4	$\mathbf{M}^{(k)} = \lambda\mathbf{M}^{(k-1)} + \mathbf{C}_{\widetilde{h}\widetilde{h},k} + \widehat{\mathbf{h}}_{LMMSE,k}\widehat{\mathbf{h}}_{LMMSE,k}^H$
5	$\gamma^{(k)} = \lambda\gamma^{(k-1)} + 1$
6	$\mathbf{C}_{hh,k} = \frac{1}{\gamma^{(k)}}\mathbf{M}^{(k)}$

Table 5.3: PDP Estimation via Adaptive EM Algorithm

5.8 Conclusion

Bayesian parameter estimation techniques such as LMMSE often lead to useful MSE reduction, but they also introduce a bias. On the other hand, imposing a joint conditionally unbiasedness constraint on a vector of parameters reduces Bayesian estimation to deterministic parameter estimation.

In this chapter, we introduce the concept of Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation, in which unbiasedness is forced for one parameter at a time. In the CWCU parameter estimation, every parameter in turn is treated as deterministic while the others are being treated as Bayesian. If the parameters are transmitted symbols, the CWCU approach corresponds to unbiased symbol detection whereas joint deterministic unbiasedness leads to a zero-forcing approach. Moreover, if the prior covariance matrix has a limited rank, we show that the block-CWCU estimation (with an appropriate block size) reduces the estimation noise, while guaranteeing the joint unbiasedness.

The more general introduction of the CWCU concept is motivated by LMMSE channel estimation, for which the implications of the concept are illustrated in various ways, including the effect on angle of arrival estimation, repercussion for blind channel estimation etc. Motivated by the channel tracking application, we also introduce CWCU Kalman filtering.

5.A Bias for LMMSE estimation

We consider a linear Gaussian model in (5.10). The LMMSE estimation is given by:

$$\widehat{\boldsymbol{\theta}}_{lmmse} = \mathbf{C}_{\theta\theta}\mathbf{H}^H \left(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}_K\right)^{-1}\mathbf{y}$$

We denote by m the rank of the prior covariance matrix $\mathbf{C}_{\theta\theta}$ ($m \leq L$). Using eigenvalue decomposition, $\mathbf{C}_{\theta\theta}$ can be written as

$$\mathbf{C}_{\theta\theta} = \mathbf{U}_C\boldsymbol{\Lambda}_C\mathbf{U}_C^H \tag{5.47}$$

where $\boldsymbol{\Lambda}_C$ is a $m \times m$ diagonal matrix containing non-zero eigenvalue of $\mathbf{C}_{\theta\theta}$, and $\mathbf{U}_C$ is a $L \times m$ matrix containing the corresponding eigenvectors $\left(\mathbf{U}_C^H\mathbf{U}_C = \mathbf{I}_m\right)$.

By decomposing $\mathbf{C}_{\theta\theta}$ as in (5.47), and applying twice the Matrix Inversion Lemma (MIL), one can show that

$$\begin{aligned}\left(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}_K\right)^{-1} &= \sigma_v^{-2}\mathbf{I}_L - \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1} \\ &\quad \times \left(\mathbf{C} + \sigma_v^2\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right)^{-1}\mathbf{C}_{\theta\theta}\mathbf{H}^H\sigma_v^{-2}\end{aligned}$$

Then,

$$\mathbf{H}^H\left(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}_K\right)^{-1} = \left(\mathbf{C} + \sigma_v^2\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right)^{-1}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H$$

Finally, we show the relation (5.13).

5.B CWCU-LMMSE for linear gaussian model

In the appendix, we will compute the Block CWCU-LMMSE estimator for a linear Gaussian model (as in (5.10)). We assume that $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^H \cdots \boldsymbol{\theta}_L^H]^H$ can be decomposed on L sub-sets. $\{\boldsymbol{\theta}_l\}_{l=1:L}$ can be either a scalar or vector parameter. We denote by L_k the size of $\boldsymbol{\theta}_k$ ($\sum_k L_k = L$).

The BCWCU-LMMSE is computed by minimizing the MSE under the BCWCU constraints, i.e.,

$$\widehat{\boldsymbol{\theta}}_{cwculmmse} = \begin{cases} \arg\min\limits_{\widehat{\theta}=Fy} \ \mathrm{tr}\left\{(\mathbf{FH}-\mathbf{I}_K)\,\mathbf{C}_{\theta\theta}\,(\mathbf{FH}-\mathbf{I}_K)^H\right\} + \sigma_v^2 \mathrm{tr}\left\{\mathbf{FF}^H\right\} \\ \mathbf{E}_k^H\left(\mathbf{FHC}_{\theta\theta}\right)\mathbf{E}_k = \mathbf{E}_k^H \mathbf{C}_{\theta\theta}\mathbf{E}_k \qquad k = 1:K \end{cases}$$

where $\mathbf{E}_k = [\ 0\ 0\ \mathbf{I}_{L_k}\ \ 0\ 0\ ]^H$ is the $L \times L_k$ matrix such that $\mathbf{E}_k^H \boldsymbol{\theta} = \boldsymbol{\theta}_k$. The Lagrangian of the constrained optimization problem is defined as

$$\begin{aligned} \mathcal{L}(\mathbf{F}, \boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_K) = {} & \mathrm{tr}\left\{(\mathbf{FH}-\mathbf{I}_K)\,\mathbf{C}_{\theta\theta}(\mathbf{FH}-\mathbf{I}_K)^H + \sigma_v^2 \mathbf{FF}^H\right\} \\ & + 2\sum_{k=1}^{K} \mathrm{tr}\ \left\{\boldsymbol{\Lambda}_k \mathbf{E}_k^H\left(\mathbf{FHC}_{\theta\theta} - \mathbf{C}_{\theta\theta}\right)\mathbf{E}_k\right\} \end{aligned} \tag{5.48}$$

Taking the gradient with respect to F gives

$$\frac{\partial \mathcal{L}}{\partial F} = 2\mathbf{F}\left(\mathbf{HC}_{\theta\theta}\mathbf{H}^H + \sigma_v^2 \mathbf{I}\right) - 2\mathbf{C}_{\theta\theta}\mathbf{H}^H + 2\left(\sum_{k=1}^{K} \mathbf{E}_k \boldsymbol{\Lambda}_k \mathbf{E}_k^H\right)\mathbf{C}_{\theta\theta}\mathbf{H}^H$$

The Lagrangian is minimum for

$$\mathbf{F}_{cw} = \left(\mathbf{I}_K - \sum_{k=1}^{K} \mathbf{E}_k^H \Lambda_k \mathbf{E}_k\right)\mathbf{C}_{\theta\theta}\mathbf{H}^H\left(\mathbf{HC}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}\right)^{-1}$$

Then,

$$\mathbf{F}_{cw}\mathbf{H} = \left(\mathbf{I}_K - \sum_{k=1}^{K} \mathbf{E}_k^H \boldsymbol{\Lambda}_k \mathbf{E}_k\right)\underbrace{\mathbf{C}_{\theta\theta}\mathbf{H}^H\left(\mathbf{HC}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}\right)^{-1}\mathbf{H}}_{\mathbf{B}_{lmmse}}$$

By considering the k^{th} constraint, one can show that the k^{th} Lagrange multiplier is

$$\boldsymbol{\Lambda}_k = \mathbf{I}_{L_k} - \left(\mathbf{E}_k^H \mathbf{C}_{\theta\theta}\mathbf{E}_k\right)\left(\mathbf{E}_k^H \mathbf{B}_{lmmse}\mathbf{C}_{\theta\theta}\mathbf{E}_k\right)^{\#} \tag{5.49}$$

where $(.)^{\#}$ denotes the pseudo-inverse operator. Thus

$$\begin{aligned}\sum_{k=1}^{K}\mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^H &= \mathrm{bdiag}\left[\mathbf{\Lambda}_1\ \cdots\mathbf{\Lambda}_K\right]\\ &= \mathbf{I}_L - \left(\mathrm{bdiag}\left(\mathbf{C}_{\theta\theta}\right)\right)\left(\mathrm{bdiag}\left(\mathbf{B}_{lmmse}\mathbf{C}_{\theta\theta}\right)\right)^{\#}\end{aligned}$$

Finally,

$$\mathbf{F}_{cw} = \mathbf{D}_{cw}\mathbf{C}_{\theta\theta}\mathbf{H}^H\left(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}\right)^{-1}$$

where $\mathbf{D}_{cw} = \left(\mathrm{bdiag}\left(\mathbf{C}_{\theta\theta}\right)\right)\left(\mathrm{bdiag}\left(\mathbf{B}_{lmmse}\mathbf{C}_{\theta\theta}\right)\right)^{\#}$ is a block diagonal matrix that ensures the component-wise unbiasedness constraints.

5.C B-CWCU-LMMSE bias for limited prior covariance rank

We consider the case when the prior covariance matrix $\mathbf{C}_A$ has a limited rank $m = 2$. All the notations correspond to those of section 5.6.
Using the eigen decomposition, $\mathbf{C}_A$ can be written as

$$\mathbf{C}_A = \mathbf{U}_A \mathbf{\Lambda}_C \mathbf{U}_A^H \tag{5.50}$$

where $\mathbf{\Lambda}_C$ is a $m \times m$ non-degenerated diagonal matrix, and the m columns of $\mathbf{U}_A$ form an orthonormal family. We have shown in (5.39) that LMMSE estimate is biased. In this appendix, we investigate the bias of the Block-CWCU-LMMSE estimate.

As we have seen in (5.14), the B-CWCU-LMMSE can be deduced from the LMMSE by compensating the component-wise bias, i.e.,

$$\widehat{\boldsymbol{\theta}}_{bcwculmmse} = \mathbf{D}_{bcw} \widehat{\boldsymbol{\theta}}_{lmmse} \tag{5.51}$$

where $\mathbf{D}_{bcw} = (\mathrm{bdiag}\,(\mathbf{C}_A))\,(\mathrm{bdiag}\,(\mathbf{B}_{lmmse}\mathbf{C}_A))^{\#}$ is a block-diagonal matrix that ensures the component-wise unbiasedness constraint. The k^{th} diagonal block of $\mathbf{D}_{bcw}$ is

$$\mathbf{E}_k^T \mathbf{D}_{bcw} \mathbf{E}_k = \left(\mathbf{U}_{A,k} \mathbf{\Lambda}_C \mathbf{U}_{A,k}^H\right) \left(\mathbf{U}_{A,k} \mathbf{\Lambda}_B \mathbf{\Lambda}_C \mathbf{U}_{A,k}^H\right)^{\#} \tag{5.52}$$

where $\mathbf{U}_{A,k} = \mathbf{E}_k^T \mathbf{U}_A$ is the k^{th} block of the matrix $\mathbf{U}_A$.
Assuming that $\left(\mathbf{U}_{A,k}^H \mathbf{U}_{A,k}\right)$ is invertible, and $L_k \geq m$, developing the pseudo-inverse leads to

$$\begin{aligned}
\mathbf{E}_k^T \mathbf{D}_{bcw} \mathbf{E}_k &= \left(\mathbf{U}_{A,k} \mathbf{\Lambda}_C \mathbf{U}_{A,k}^H\right) \left(\mathbf{U}_{A,k} \left(\mathbf{U}_{A,k}^H \mathbf{U}_{A,k}\right)^{-1} \mathbf{\Lambda}_C^{-1} \mathbf{\Lambda}_B^{-1} \left(\mathbf{U}_{A,k}^H \mathbf{U}_{A,k}\right)^{-1} \mathbf{U}_{A,k}^H\right) \\
&= \mathbf{U}_{A,k} \mathbf{\Lambda}_B^{-1} \left(\mathbf{U}_{A,k}^H \mathbf{U}_{A,k}\right)^{-1} \mathbf{U}_{A,k}^H
\end{aligned}$$

Finally, one can derive the bias of the B-CWCU-LMMSE:

$$\begin{aligned}
E_{|A}\left\{\mathbf{A}_{bcwculmmse}\right\} &= \mathbf{D}_{bcw}\; E_{|A}\left\{\mathbf{A}_{lmmse}\right\} \\
&= \mathbf{D}_{bcw}\mathbf{U}_A\mathbf{\Lambda}_B\begin{bmatrix}\eta\\ \gamma\end{bmatrix} \\
&= \sum_k \mathbf{E}_k\left(\mathbf{E}_k^T\mathbf{D}_{bcw}\mathbf{E}_k\right)\mathbf{E}_k^T\mathbf{U}_A\mathbf{\Lambda}_B\begin{bmatrix}\eta\\ \gamma\end{bmatrix} \\
&= \sum_k \mathbf{E}_k\mathbf{U}_{A,k}\begin{bmatrix}\eta\\ \gamma\end{bmatrix} \\
&= \mathbf{A}(\theta)
\end{aligned}$$

Thus, we show that the block-CWCU-LMMSE (with block-size at least $m = 2$) guarantees joint unbiasedness.

Chapter 6

Mobile Terminal Positioning via PDP Fingerprinting

Non-Line-of-Sight and multipath propagation conditions pose significant problems for most mobile terminal positioning approaches. In contrast, Power Delay Profile Fingerprinting (PDP-F) thrives on multipath propagation. This multipath extension of T(D)oA is based on matching an estimated power delay profile from one or several base stations (BSs) (or other transmitters (broadcast, ...)) with a memorized power delay profile map for a given cell. It is obvious that the overall localization accuracy depends strongly of the quality of the PDP estimation. In this chapter, we propose exploiting the prior structural and statistical information to enhance the PDP estimation and increase the localization accuracy.

6.1 Introduction

Mobile positioning systems have received significant attention in both research and industry over the past few years. Conventional localization techniques are organized in a two step procedure. The first step involves the measurement of one or more given physical parameters of the transmitted signal (typically Time or Time-Difference of Arrival (ToA, TDoA), path delays...). The later step combines multiple measurements to estimate the mobile position.
Classically, the delay parameters (ToA, TDoA...) are estimated by analyzing the local maxima of the channel impulse response. One way to enhance the estimation accuracy and the scheme robustness is using a fitting between a parametric model and the estimated channel impulse response. In this context, the estimation of transmission channel becomes crucial. In fact, although the Bayesian formulation of channel estimation allows better noise suppression, the introduced bias is very annoying for the delay estimation (as it leads to a modification of the pulse-shape structure). That is why the Bayesian prior is rarely exploited in such applications. Channel estimation is done typically based on Least-Squares, which leads to unbiased but noisier channel estimate. This fact was one major (and our initial) motivation for the introduction of the CWCU parameter estimation. Joint unbiasedness is not necessary for such application; only block component-wise unbiasedness is required. For instance, only channel taps forming the principal-lobe of a given pulse-shape need to be jointly unbiased (see figure 6.1). One can allow for interference (contribution of other paths), if this can be motivated by a noise reduction.

Non-Line-of-Sight (NLoS) and multipath propagation conditions pose significant problems for most geometric and satellite assisted mobile terminal positioning approaches. In contrast, power delay profile fingerprinting (PDP-F) thrives on multipath propagation. This multipath extension of T(D)oA is based on matching an estimated power delay profile from one or several base stations (BSs) (or other transmitters (broadcast, ...)) with a memorized power delay profile map for a given cell. It becomes obvious that the overall localization accuracy depends strongly of the quality of the PDP estimation. We propose exploiting the prior knowledge on the received signal structure to enhance the PDP estimation and increase the localization accuracy. Moreover, since the received signal comes from a MS located at an

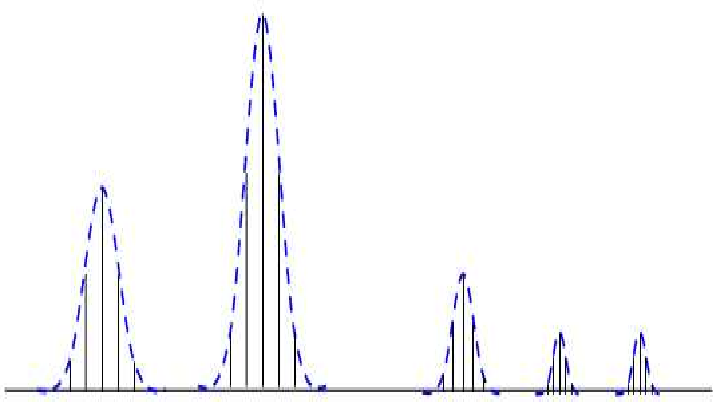

Figure 6.1: Parametric channel model.

unknown (but existing) position, the PDP parameters can not have arbitrary values. Classic PDP-F ignores these constraints, and it is thus sub-optimal. Using a Bayesian framework, we introduce a one step localization approach taking the localization constraints into consideration. In the case of a multi-antenna reception, we propose an extension of the PDP-F (PSDP-F) taking into account the spatial information. PSDP-F can be considered as a multipath extension of a combination of T(D)oA and AoA methods, that does not require an explicit requirement for antenna array calibration.
We also propose a validation of PDP-F via simulations that can easily be reproduced. In these simulations, the multicellular environment consists of a big box in which multipath arises by reflection off the six sides. The resulting PDP depends on the positions of BS and terminal, the attenuation mechanism and the reflection coefficients of the six sides.

This chapter is organized as follows. After a brief overview of the state of the art in the mobile terminal positioning, the PDP Fingerprinting approach is described (section 6.3), and evaluated using a ray-tracing multipath simulation environment (section 6.4). In sections 6.5 and 6.6, we investigate the enhancement of the PDP using parametric deterministic and Bayesian models, and the corresponding PDP-fingerprinting schemes respectively. Next, we introduce an extension of the PDP-F approach for the multi-antenna reception scenario (section 6.7). Finally, a discussion and some concluding remarks are also provided in section 6.8.

6.2 Mobile terminal positioning: a brief overview

Mobile positioning systems have received significant attention in both research and industry over the past few years [97, 28]. Indeed, the localization of the mobile phone has become one of the most important features of communication systems due its various potential applications (effective intra and inter-system handoff, localization of emergency caller...). The basic function of localization system is to collect information about position-dependent parameters of a Mobile Station (MS) signal and to process that information to get a location estimate .

Conventional localization techniques aiming at higher accuracy than simple cell identification are organized in a two steps procedure [201]. The first step involves the measurement of certain physical parameters of the received signal (e.g. time, or time-difference, of arrival (ToA, TDoA), angle of arrival (AoA), signal strength...). The signal is assumed to be received under Line of Sight (LoS) conditions, in which case the parameters of multiple MS-BS links are required to have position identifiability. The second step combines multiple measurements from the link to a convenient number of Base Stations (BSs) to estimate the mobile position.
Weiss et al. underline the sub-optimality of the two-step approach [8, 205]. In fact, the signal parameters are estimated separately and independently for each MS-BS link, ignoring the constraint that all measurements must correspond to the same source. Weiss et al. introduce the “Direct Position Determination (DPD) approach": the estimated channel impulse responses for each MS-BS link are processed jointly and the MS position is computed as the best match to all data simultaneously. Monte Carlo simulations demonstrate that the DPD method provides better localization accuracy (especially in the presence of multipath propagation/fading [9]), and allows to work in an extended (lower) SNR range.

The main cause of inaccuracies observed in conventional localization systems is the realistic propagation conditions imposed by the wireless channel: multipath propagation and often Non Line-of-Sight (NLoS) conditions. In fact, the conventional methods rely on the line-of-Sight path between a base

station and the Mobile station. However, in an urban environment, a LoS condition (i.e. the LoS path being present) is rarely satisfied for three BSs at the same time. This fact degrades the localization performance (identifiability and accuracy) of conventional techniques and creates the need to develop more accurate techniques suited for these propagation.
To alleviate this problem, Porretta et al. suggest tracking the MS position to obtain more reliable position estimates [142, 143]. The combination of a ToA and AoA measurement allows localization identifiability from just one MS-BS link. Based on the ToA and AoA measurements, the MS location is estimated by following two alternative procedures. When the MS is in the LoS condition, the location is determined through the parameters relevant to the first path received at the BS (AoA and ToA). On the other hand, under the NLoS condition, the MS position is determined by minimizing a given cost function (taking into account the ToA, the AoA, and the coordinates of the obstacles found along the AoA for the first N paths). An alternative approach is proposed by Nájar et al. [119, 120]. In LoS condition, the estimation of the ToA of the LoS and a NLoS path allows the determination of an offset (bias) between the two ToAs. This bias is then subtracted to the ToA of the NLoS path during NLoS conditions to provide an estimate of the LoS ToA. In general, the position estimate accuracy and its identifiability can always be improved by adding a Kalman filtering stage to track the location trajectory, on the basis of brute position estimate. The use of the Kalman filter allows the tracking, not only of the position and the velocity of the mobile, but also of the ToA bias caused by multipaths, and NLoS conditions.

While previous techniques try to reduce the multipath and the NLoS effects, those cannot be eliminated, and the errors they induce are difficult to predict. For that reason, some new localization methods (e.g. Received Signal Strength and Location Fingerprinting) have been designed to obtain optimal performance in urban environment. Those techniques not only overcome the problems related to the propagation environment, but also take advantage from the temporal diversity of the wireless channel. The idea is to use a previously collected or predicted signal database (location dependent parameters) from the coverage area. The terminal measures the same parameters, and sends it to the localization server in the network. The position is then determined by a correlation algorithm, which compares the measured signal parameters with the information stored in the database.

The Enhanced Signal Strength (ESS) method is based on this principle, and has allowed the deployment of personal locator systems in PHS service areas in Japan. The position of the mobile is determined using the signal strength of preferably three to five base stations. From this input plus information from the base station database, the system can calculate the position of the MS [97]. The database is built by simulating the signal propagation characteristics of every wireless transmitting antenna in the area of interest. Heikki et al. propose building the signal strength database through measurements instead of computation [102].
Instead of exploiting signal strength, the Location Fingerprinting (LF) (introduced by U.S. Wireless Corp. of San Ramon, Calif.) relies on signal structure characteristics [97, 202, 203, 204]. By combining multipath pattern with other characteristics, the LF creates a signature unique to a given location. The position of the mobile is determined by matching the transmitter's signal characteristics to an entry of the database. For LF, multipoint signal reception is not required: the system can use data for only a single point to determine location. Ahonen and Eskelinen suggest using the measured Power Delay Profiles (PDPs) in the database [5, 6]. Thus, the location estimation is possible by using only one BS due to the additional information provided by the PDP, i.e., the amplitudes and the delays of the multipath components.

Due to channel reciprocity the knowledge of the Channel Impulse Response (CIR) and its parameters can be exploited either in the downlink at the MS or in the uplink at the BS. In case the CIRs to multiple BSs are used, they can be either be exploited in the downlink at the MS, or in the uplink at a switching center connecting several BSs.

6.3 PDP fingerprinting, with and without time reference

6.3.1 Fingerprinting for localization

Location Fingerprinting is a general localization method that can be applied to any cellular or WLAN (Wireless Local Area Network) network. The key idea is to store signal structure information, from the whole coverage area of the localization system, in a database. The database should contain collected or predicted position dependent signal information (a position signature called fingerprints) with a resolution comparable to the accuracy that can be achieved with the method. The MS measures the same parameters, and sends it to the localization server in the network. The position is then determined by a correlation algorithm, which compares the measured signal parameters with the information stored in the database.

The major effort in applying location fingerprinting is the creation and maintenance of the database. The signal fingerprints for the database can be collected either by:

- measurements: a vehicle drives through the coverage area collecting the signal fingerprints at each position.
- or by a computational network planning tool: by simulating the signal propagation characteristics of every transmitting antenna in the area of interest.

Measurements are more laborious but produce more accurate fingerprint data. Also a combination of measured and computed fingerprints can be used.

6.3.2 Basic synchronous PDF fingerprinting method

An important consideration in the location fingerprinting technique is the choice of the signal fingerprints. Any location-dependent signal information that can be measured by the MS or the BSs is useful for the location fingerprinting technique. The signal fingerprints could include signal strength, signal time delay, or even channel impulse response. Ahonen and Eskelinen suggest using the measured PDPs as a signal fingerprints for UMTS systems. The power delay profile shows the power and the arrival times of the different ray-paths between the selected transmitter and the selected receiver (see figure 6.2). In the case of synchronous network, the first peak of the measured PDP determines the time of arrival (ToA) of the received signal, which is used for the ToA algorithm. In addition, the PDP fingerprinting (with time reference) takes advantage of the entire measured PDP (the whole temporal diversity). Therefore, the system can use data from only a single point to determine location; multipoint signal reception is not required, although it is highly desirable.

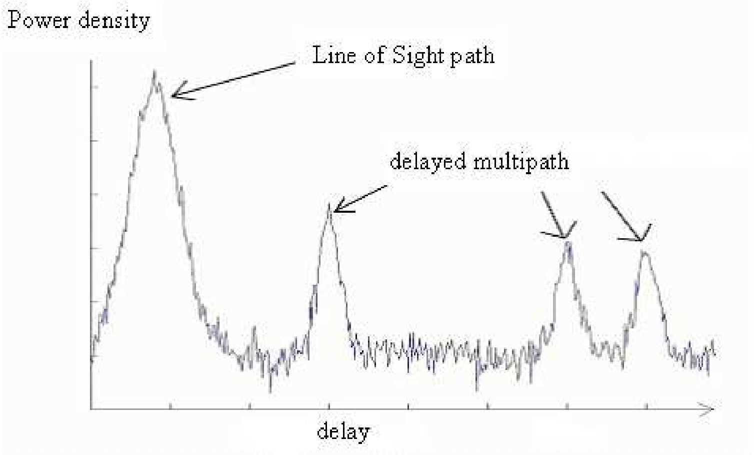

Figure 6.2: Example of Power Delay profile with a LoS path.

In the case of an asynchronous network, absolute ToA information is not available (only relative ToA, or TDoA between paths). The measured PDP could match with any delayed version of the PDPs stored in the database. We call this variant "PDP fingerprinting without time reference". In theory, if the number of multiple paths increases, multipoint signal reception is not required. In practice, multipoint signal reception for PDP-F (without time reference) is desirable to have a good positioning accuracy (see figure 6.7). PDP-F with and without time reference can be interpreted as an extension of the ToA and TDoA classic approaches.

6.3.3 Synchronous and asynchronous matching score function

A second important consideration in the location fingerprinting technique is the choice of the matching score [212]. In fact, for each location (x, y) a matching score can be computed from the measured $\widehat{PDP}$ and the stored $PDP_{(x,y)}$. The MS position is determined by minimizing the distance between the two quantities (called matching score).

For one-BS signal reception scenario, typically we use the Least Square (LS) cost function:

$$C(x,y) = \left\| \widehat{PDP} - PDP_{(x,y)} \right\|^2 \tag{6.1}$$

For multi-BS signal reception scenario, one can use the following LS cost function (or a weighted version):

$$C(x,y) = \sum_{k=1}^{K} \left\| \widehat{PDP}^{(k)} - PDP^{(k)}_{(x,y)} \right\|^2 \tag{6.2}$$

where K is the number of BSs, and $\widehat{PDP}^{(k)}$ denotes the observed PDP at the k^{th} BS.
In the asynchronous (absence of time reference) case, the criterion (6.1) would becomes

$$C(x,y) = \min_{\Delta\tau} \left\| \widehat{PDP}_{\Delta\tau} - PDP_{(x,y)} \right\|^2 \tag{6.3}$$

where $\widehat{PDP}_{\Delta\tau}$ is $\widehat{PDP}$ slighted over a synchronization delay $\Delta\tau$. In the single path case, this becomes the Received Signal Strength (RSS) fingerprinting method. So, the criterion in (6.3) can be considered to be an extension of the RSS method to the frequency selective channel.

6.4 Ray tracing multipath in a box

6.4.1 Localization validation using ray tracing multipath in a box

Generally, we assume an approximate signal/propagation model to yield a practical implementation of the proposed localization techniques. In this subsection, the proposed scheme will be validated using more accurate propagation and/or received signal models. The simulation environment is a crucial issue for the validation of localization algorithms. In the literature, two main strategies are adopted to validate the proposed algorithms:

- Evaluation in real or like-real scenario: the evaluation is done via an experimental localization trial or by simulating the propagation environment using a network planning tool featuring a three-dimensional ray-launching method.

- Evaluation on a fixed CIR: the delays and the gains of the different channel taps are fixed according to a given channel model (Vehicular A, B, indoor to outdoor A, B...).

There is no doubt that the most accurate evaluation technique will rely to real scenario evaluation. However, those techniques are often expensive, time-consuming, labor intensive, and not easy to reproduce. On the other hand, using a fixed CIR seems to be insufficient and does not allow position tracking scenarios.
In this section, we propose a validation of the PDP fingerprinting using a ray-tracing multi-path in a box. The proposed validation technique produces a new tradeoff between complexity and evaluation accuracy; and it can easily be reproduced. The multicellular environment consists of a big box in which multipath arises by reflection from the six sides (figure 6.3).

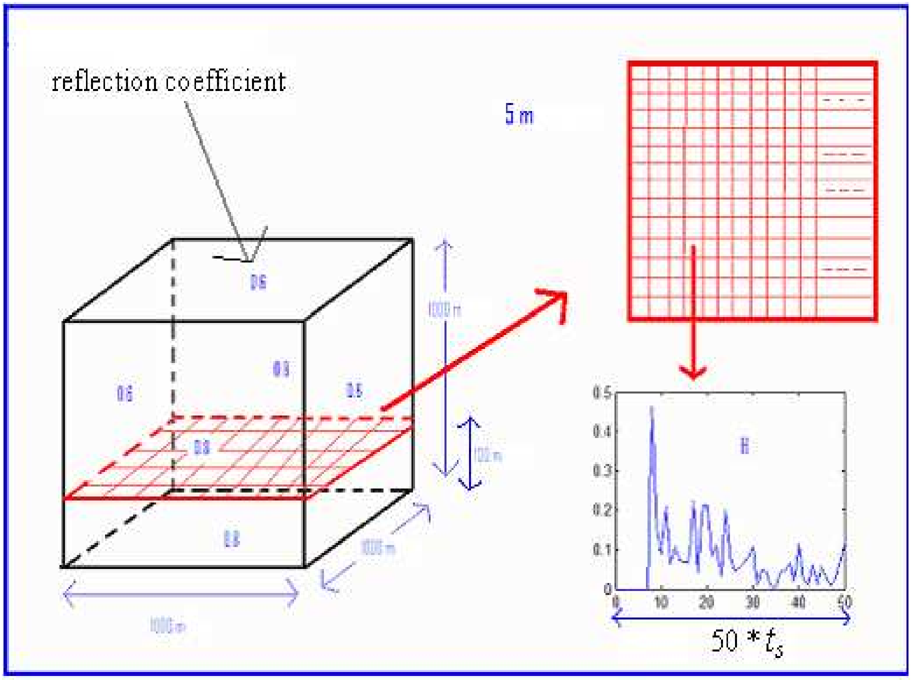

Figure 6.3: The multicellular simulation environment.

Note that even if this propagation environment may be a far cry from realistic wireless environments (except for certain indoor or street scenarios), it allows to generate realistic power delay profiles, which is the key ingredient of the method considered here.

The ray tracing multipath environment is taken from the acoustics word (figure 6.4). The method used for the simulation of the impulse response between the MS and the multiple BSs is similar to the image method [139].

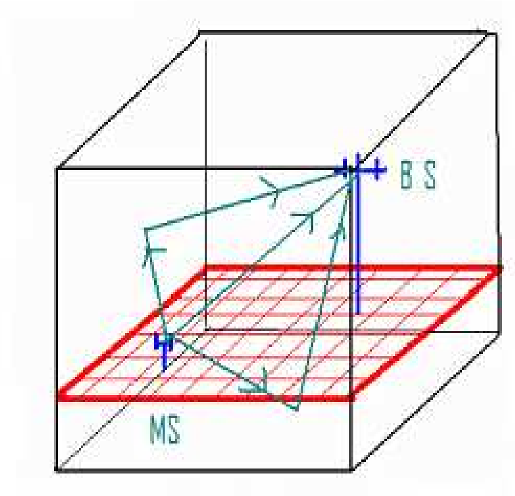

Figure 6.4: The ray tracing simulator.

The image method originates in geometrical acoustics. It states that only specular reflections of sound are important. The real scene is complemented with additional images of original space mirrored by walls that are to reflect the sound. The intensity of the new sound sources is decreased according to the absorption of walls and air. Only direct propagation of sound from the original source and from new sources (resulting from mirroring) is then taken into account [7].

However, as we consider electromagnetic propagation at a certain carrier frequency, several modifications should be taking into account:

- Due to the narrow-band transmission at a certain carrier frequency, we should consider complex channel impulse response, and take into account the phase modification caused by the electromagnetic waves reflections.

- We include a wave attenuation model (with respect to the distance).
- We perform a more precise sampling operation via a pulse shape (instead of a simple delays ceiling).

Note that if we multiply the sampling frequency and divide the box dimensions by the same factor, we recover the same sampled impulse response as previously (up to a multiplicative factor). Thus, the sampling frequency and the box dimension do not have, independently, intrinsic interpretations. The box can be sized according to the cell volume, the distribution of the major obstacles...

Due to the symmetry of our simulated propagation environment, another artifact is being introduced. In fact, if the reflection coefficients of the different faces are equal and if the BS is located at the center of the box, the environment will have four mirror symmetries. Thus, the same channel impulse response will be received at eight distinct positions (see figure 6.5). A real propagation environment is too complex to be symmetric. To alleviate this artifact, one can use distinct reflection coefficients on the different box faces, avoid positioning the BS at a box symmetry plan, and/or force the MS to move on a limited area (figure 6.5).

6.4.2 PDP fingerprinting validation using ray tracing multipath method

In this section, we propose a validation of the PDP fingerprinting via the ray tracing multipath environment proposed above. We consider a cubic box with dimensions $1000 \times 1000 \times 1000$. To simulate the channel impulse response we refer to the CDMA2000 standard. The CDMA chip-rate in the simulation is 3.8 MHz, with an up-sampling factor equal to 4. We also consider a raised cosine pulse-shape to perform more precise sampling operation.
Given an MS position, the channel impulse response is simulated using the ray-tracing multipath routine. White Gaussian noise is added. The PDP estimate is computed by taking the magnitude of the noisy CIR. Finally, the MS position is determined by matching the estimated PDP and the pre-stored PDP database. Figure 6.6 plots the Root Mean Square positioning

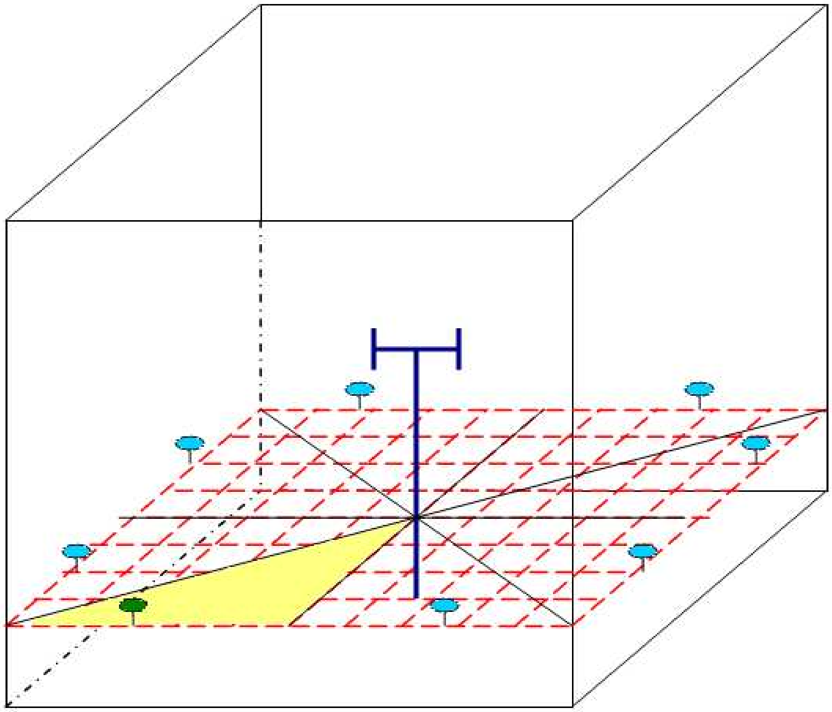

Figure 6.5: Symmetry artifact in the "ray tracing multipath in a box" simulation environment.

Error (RMSE) for the PDP-Fingerprinting function of the spatial resolution for $SNR = +\infty$, and $SNR = 10dB$.

We see that the sensibility of positioning error to discretization depends on the SNR of the channel estimation. Figure 6.7 compares the PDP-F using 1 and 2 base stations with and without time reference. For the case of 2 BSs, the noise power on the 2 BSs is assumed to be the same. The SNR on the x-axis corresponds the SNR at the strongest BS signal. We see that for the one point signal reception scenario, the synchronization between the MS and BSs increases significantly the positioning accuracy. However, the effect of synchronization is not too spectacular for the multipoint signal reception scenario (TDoA-like information is sufficient to give satisfying precision).

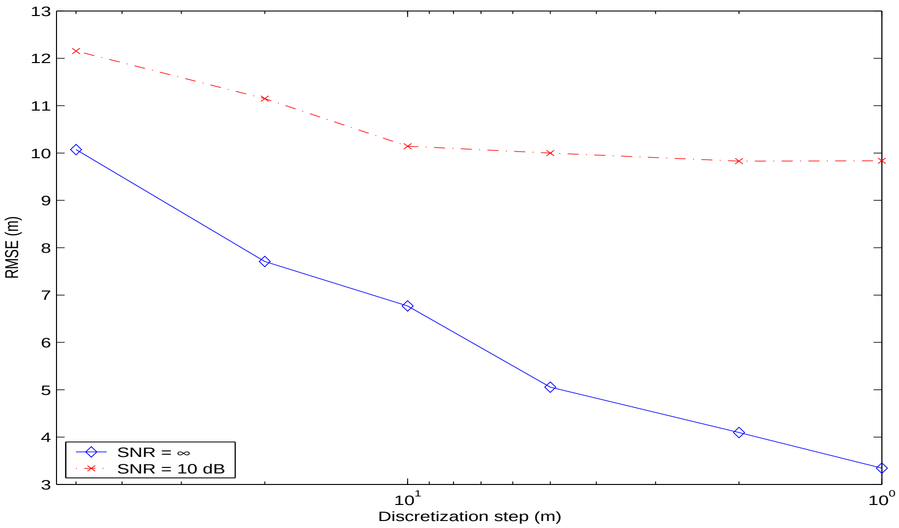

Figure 6.6: RMSE vs. Discretization step for $SNR = +\infty$, and $SNR = 10dB$.

6.5 Parametric power delay profile estimation

In a multipath propagation environment, the received impulse response between a MS and a BS antenna can be written as:

$$h(t,\tau) = \sum_{l=1}^{L} A_l(t)\; e^{j\varphi_l(t)}\; p(\tau - \tau_l(t)) \tag{6.4}$$

where L denotes the number of paths, $p(t)$ is the convolution of the transmit and receive filters, $\tau_l(t)$, $A_l(t) \geq 0$ and $\varphi_l(t)$ are respectively the delay, the fading amplitude and phase of the l^{th} path. The path delay and fading amplitude vary slowly with the position; whereas the fading phase varies rapidly (with 2π over one wavelength). If the MS moves slowly, one can assume delays and fading amplitudes to be constant over T channel observations, but the fading phases are certainly not.
The (filtered) PDP is obtained by averaging the squared CIR magnitude over the path phases:

$$PDP(\tau) = E_{\varphi}\left|h(t,\tau)\right|^2 = \sum_{l=1}^{L} A_l^2 p^2(\tau - \tau_l). \tag{6.5}$$

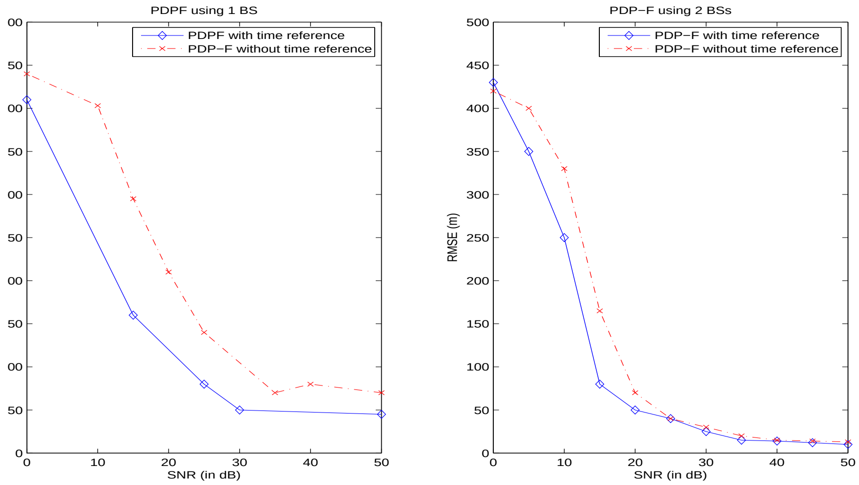

Figure 6.7: Positioning accuracy for PDP-F with and without time reference using 1 BS (left) or 2 BSs (right).

The channel parameters are observed indirectly through the received data. Usually the channel estimation is based on a known sequence of symbols (training or pilot sequence), which is unique for a certain transmitter and which appears in every transmission burst. Thus, the channel estimator is able to estimate the CIR for each burst separately by exploiting the known transmitted symbols and the corresponding received samples. In the majority of mobile positioning techniques, CIR estimation are performed using a simple matched filtering [93], or the Least-Squares estimation [49]. If the training sequence is long enough, the estimated CIR can be written as:

$$\widehat{h}(t,\tau) = \sum_{l=1}^{L} A_l \; e^{j\varphi_l(t)} \; p(\tau - \tau_l) + v(t,\tau) \qquad t = 1 : T \tag{6.6}$$

where $v(t,\tau)$ denotes the additive white Gaussian estimation error with a variance σ_v^2.

Then, the PDP is commonly estimated by averaging the squared magnitude of the estimated taps, i.e.,

$$\widehat{PDP}(\tau) = \frac{1}{T} \sum_{t=1}^{T} \left| \widehat{h}(t,\tau) \right|^2 \tag{6.7}$$

However, if the mobile moves rapidly and/or some paths are not resolvable (due to the limited bandwidth of the pulse-shape $p(t)$, path contributions may strongly overlap in delay), the averaging may provide a poor PDP estimation, and hence poor localization accuracy. In the following, structural and statistical prior information about the channel are exploited to enhance the PDP estimation.

6.5.1 Deterministic PDP estimation

In this section, a structural priors for the wireless channel is considered such as the multipath propagation model (in (6.6)), and the prior knowledge of the pulse-shape. The parameters τ_l, A_l, and $\varphi_l(t)$ are considered unknown deterministic parameters. The exploitation of this structural information leads to the following two-step procedure:

- First, estimate the model parameters by optimizing the Maximum Likelihood criterion

$$\widehat{\tau}_l, \widehat{A}_l, \{\widehat{\varphi}_l(t)\}_{t=1:T} = \arg \min_{\tau_l, A_l, \varphi_l(t)} \sum_{t=1}^{T} \sum_{\tau=1}^{N} \left\| \widehat{h}(t,\tau) - \sum_{l=1}^{L} A_l \; e^{j\varphi_l(t)} \; p(\tau - \tau_l) \right\|^2 \tag{6.8}$$

- Then, construct the PDP estimate as

$$\widehat{PDP}(\tau) \;=\; \sum_{l=1}^{L} \widehat{A}_l^2 \; p^2(\tau - \widehat{\tau}_l) \tag{6.9}$$

Minimizing (6.8) leads to a difficult non-linear optimization problem. Although the least-squares problem (6.8) is separable in the complex path amplitude $A_l e^{j\varphi_l(t)}$ a difficulty arises from imposing that A_l does not depend on t. To have a tractable solution, we propose a two step optimization scheme. First, we estimate the paths delays τ_l and the complex fading coefficients $b_l(t) = A_l e^{j\varphi_l(t)}$. Then, the constant fading amplitudes and the varying phases are extracted from the varying complex coefficients $b_l(t)$ using an LS based technique.

For the clarity of the algorithm description, we shall consider matrix notation. Equation (6.6) becomes

$$\widehat{\mathbf{h}}(t) \;=\; \underbrace{[\mathbf{p}_{\tau_1}\cdots\mathbf{p}_{\tau_L}]}_{\mathbf{P}_{\boldsymbol{\tau}}}\underbrace{\begin{bmatrix} A_1\; e^{j\varphi_1(t)} \\ \vdots \\ A_L\; e^{j\varphi_L(t)} \end{bmatrix}}_{\mathbf{b}(t)} + \mathbf{v}(t) \qquad t = 1:T$$

where $\widehat{\mathbf{h}}(t) = \left[\widehat{h}(t,t_0)\cdots\widehat{h}(t,t_0+(N-1)t_s)\right]^T$, and similarly for $\mathbf{v}(t)$, $\boldsymbol{\tau} = [\tau_1\cdots\tau_L]^T$, and $\mathbf{p}_\tau = [p(t_0-\tau)\cdots p(t_0+(N-1)t_s-\tau)]^T$. N is the channel impulse length, and t_0 is the sampling period. Note that $\mathbf{p}_\tau^T = \mathbf{p}_\tau^H$
The paths delays $\boldsymbol{\tau}$ and fading coefficients $\mathbf{b}(t)$ should minimize:

$$\widehat{\boldsymbol{\tau}}, \widehat{\mathbf{b}}(t) = \arg\min_{\boldsymbol{\tau},\mathbf{b}(t)} \sum_{t=1}^{T} \left\|\widehat{\mathbf{h}}(t) - \mathbf{P}_{\boldsymbol{\tau}}\mathbf{b}(t)\right\|^2 \tag{6.10}$$

The problem is quadratic in $\mathbf{b}(t)$, leading to the estimates $\widehat{\mathbf{b}}(t) = \left\{\mathbf{P}_\tau^T\mathbf{P}_\tau\right\}^{-1}\mathbf{P}_\tau^T\widehat{\mathbf{h}}(t)$ for a given $\boldsymbol{\tau}$. The resulting problem for $\boldsymbol{\tau}$ is non-linear:

$$\widehat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \sum_{t=1}^{T} \widehat{h}^H(t)\mathcal{P}^{\perp}_{P_\tau}\widehat{h}^H(t) \tag{6.11}$$

where $\mathcal{P}_{P_\tau} = \mathbf{P}_\tau\left(\mathbf{P}_\tau^T\mathbf{P}_\tau\right)^{-1}\mathbf{P}_\tau^T$, $\mathcal{P}^{\perp}_{P_\tau} = \mathbf{I} - \mathcal{P}_{P_\tau}$ represent the projection on the column space of $\mathbf{P}_\tau$, and its orthogonal subspace.
We propose estimating these parameters by exploiting the sparse nature of the CIR through the use a Matching Pursuit (MP) algorithm. The MP has been used in a variety of applications [149], and particular to derive accurate channel estimates [37, 92, 38, 104]. Using the standard form of the MP algorithm, we first find the delay τ_1, such as $\mathbf{p}_{\tau_1}$ is that best aligned with the different channel realizations $\mathbf{h}^{(0)}(t) = \widehat{\mathbf{h}}(t)$. Then, for each channel realization, the projection of $\mathbf{h}^{(0)}(t)$ along $\mathbf{p}_{\tau_1}$ is removed from $\mathbf{h}^{(0)}(t)$ and the residual $\mathbf{h}^{(1)}(t)$ is found. Now, the delay τ_2 which best aligns $\mathbf{p}_{\tau_2}$ and $\mathbf{h}^{(1)}(t)$ is computed and a new residual $\mathbf{h}^{(2)}(t)$ is formed. The algorithm proceeds by sequentially choosing the column that best matches the residual until some termination criterion is met. The l^{th} iteration is described in the following paragraph.
For a given delay τ_0, we denote the projection onto the vector $\mathbf{p}_\tau$ as $\mathcal{P}_\tau =$

$\mathcal{P}_{\mathbf{P}_\tau} = \mathbf{p}_\tau \mathbf{p}_\tau^H / \|\mathbf{p}_\tau\|^2$. The delay $\widehat{\tau}_l$ is selected such that $\mathbf{p}_{\tau_l}$ is best aligned with the residual $\mathbf{h}^{(l-1)}(t)\ \ t = 1 : T$, i.e.,

$$\widehat{\tau}_l = \arg\max_{\tau_l} \sum_{t=1}^{T} \left\| \mathcal{P}_{\tau_l} \mathbf{h}^{(l-1)}(t) \right\|^2 = \arg\max_{\tau_l} \sum_{t=1}^{T} \left| \mathbf{p}_{\tau_l}^H \ \mathbf{h}^{(l-1)}(t) \right|^2 \tag{6.12}$$

where the second equality exploit the fact the $\|\mathbf{p}_\tau\|^2$ is essentially constant over most of the range of τ.
The previous maximization can be carried out using a two step procedure. First, a brute-force exhaustive search is performed on a quantized delay grid (to avoid local minima). Then the optimization is refined using e.g. the golden section algorithm.
Once the path delay is estimated, the complex fading coefficients are deduced:

$$\widehat{b}_l(t) = \frac{\left(\mathbf{p}_{\widehat{\tau}_l}^T \ \mathbf{h}^{(l-1)}(t)\right)}{\|\mathbf{p}_{\widehat{\tau}_l}\|^2} \qquad t = 1 : T \tag{6.13}$$

Finally, the new residual vectors are computed:

$$\mathbf{h}^{(l)}(t) = \mathbf{h}^{(l-1)}(t) - \widehat{b}_l(t)\mathbf{p}_{\widehat{\tau}_l} \qquad t = 1 : T \tag{6.14}$$

The recursions are repeated until a specified number of taps L_{max} been selected or the residual becomes sufficiently small, i.e., $\frac{\sum_t \left\|\mathbf{h}^{(l)}(t)\right\|^2}{\sum_t \left\|\widehat{h}(t)\right\|^2} < \epsilon$, where ϵ can be chosen as a function of σ_v^2.
In a somewhat more coupled version of the MP approach, a re-estimation of all complex path amplitudes can be performed with each newly added path, i.e., $\widehat{\mathbf{b}}_{1:l}(t) = \left(\mathbf{P}_{\widehat{\tau}_{1:l}}^T \mathbf{P}_{\widehat{\tau}_{1:l}}\right)^{-1} \mathbf{P}_{\widehat{\tau}_{1:l}}^T \widehat{\mathbf{h}}(t)$ would replace (6.13), and $\mathbf{h}^{(l)}(t) = \widehat{\mathbf{h}}(t) - \mathbf{P}_{\widehat{\tau}_{1:l}} \widehat{\mathbf{b}}_{1:l}(t)$ which would replace (6.14).

In the previous step, we have ignored the constraint that $\forall l \ \ b_l(t), \ \ t = 1 : T$, have the same magnitude. In fact, the estimated complex fading coefficient (corresponding to the l^{th} path) can be written as:

$$\widehat{b}_l(t) = A_l e^{j\varphi_l(t)} + \widetilde{b}_l(t) \quad t = 1 : T \tag{6.15}$$

where $\widetilde{b}_l(t)$ are the fading estimation errors. If the paths are resolvable (path contributions do not overlap much), the estimation errors can be assumed to be white Gaussian. In this case, a Maximum Likelihood formulation leads to the following LS problem:

$$\begin{aligned}\widehat{A}_l, \widehat{\varphi}_l(t) &= \arg\min_{A_l,\varphi_l(t)} \sum_{t=1}^{T} \left|\widehat{b}_l(t) - A_l e^{j\varphi_l(t)}\right|^2 \\ &= \arg\min_{A_l,\varphi_l(t)} \sum_{t=1}^{T} \left|\widehat{b}_l(t) e^{-j\varphi_l(t)} - A_l\right|^2 . \qquad (6.16)\end{aligned}$$

Thus, the fading amplitudes and phases are estimated as:

$$\begin{cases} \widehat{A}_l = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \left|\widehat{b}_l(t)\right| \\ \widehat{\varphi}_l(t) = \text{angle}\left(\widehat{b}_l(t)\right) \quad / \quad e^{j\widehat{\varphi}_l(t)} = \widehat{b}_l(t)/\left|\widehat{b}_l(t)\right| \end{cases} \qquad (6.17)$$

Finally, the refined PDP estimate is computed as in (6.9).

6.5.2 Bayesian PDP estimation

If the propagation paths are resolvable (in delay), the deterministic approach in (6.8) is appropriate. However, if this is not the case and the channel taps are the superpositions of different paths arriving at almost the same delay, i.e.,

$$A_l(t) e^{j\varphi_l(t)} = \sum_{k=1}^{K_l} A_{l,k} e^{j\varphi_{l,k}(t)} \qquad (6.18)$$

Therefore, modeling the fading amplitudes as deterministic quantities is no longer appropriate and a Bayesian modeling is warranted. In this section, we assume that the complex fading vector $\mathbf{b}(t)$, and the additive noise $\mathbf{v}(t)$ are independent i.i.d. zero-mean Gaussian vector processes, i.e.,

$$\begin{matrix} \mathbf{b}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\mathbf{b}) \\ \mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathbf{I}_N) \end{matrix} , \quad t = 1 \cdots T \qquad (6.19)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{C})$ denotes the zero-mean complex normal distribution with covariance matrix $\mathbf{C}$, $\mathbf{C}_b = \text{diag}\left(\sigma_{b,1}^2 \cdots \sigma_{b,L}^2\right)$ is a diagonal matrix characterizing the covariance of the random complex fading amplitudes.

The statistical model (6.19) implies that the $\widehat{\mathbf{h}}(t)$ are modelled as i.i.d. complex Gaussian vectors with $\widehat{\mathbf{h}}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{h}})$, $\mathbf{C}_{\mathbf{h}} = \mathbf{P}_\tau \mathbf{C}_b \mathbf{P}_\tau^T + \sigma_v^2 \mathbf{I}_N$. Thus, whereas in the deterministic case the channel is parameterized by path delays and amplitudes, the Bayesian model parameterizes the channel with path delay and power. The considered approach is Bayesian for $\mathbf{h}(t)$, but Maximum Likelihood for the parameters $\boldsymbol{\tau}$ and $\mathbf{C}_b$. To distinguish from the deterministic ML approach in the previous section, the ML approach considered here will called Rayleigh ML.

Taking into account the statistical model, the likelihood of the channel parameters is given by:

$$L(\boldsymbol{\tau}, \mathbf{C}_b) \propto -T \ln(\det \mathbf{C}_h) - \sum_{t=1}^{T} \widehat{\mathbf{h}}^H(t) \mathbf{C}_h^{-1} \widehat{\mathbf{h}}(t) \tag{6.20}$$

Maximizing (6.20) (with respect to $\boldsymbol{\tau}$, $\mathbf{C}_b$) is again a difficult non-linear problem. In this section, we will not elaborate on the global maximization of the Rayleigh likelihood. We will restrict our interest to the local identifiability of the Bayesian localization approach. Application to mobile localization is considered in the section 6.6.2.

To investigate the local identifiability of Bayesian PDP-Fingerprinting, we assume that paths are well separated, which implies

$$\mathbf{p}_{\tau_i}^T \mathbf{p}_{\tau_j} \approx \sigma_p^2 \delta_{i,j} \tag{6.21}$$

where σ_p^2 is the energy of the pulse-shape, and $\delta_{i,j}$ is the Kronecker delta function. Under this assumption, $\frac{1}{\sigma_p}\mathbf{P}_\tau$ becomes an orthogonal matrix and the determinant of $\mathbf{C}_h$ does not depend on the path delays:

$$\det(\mathbf{C}_h) = \prod_{l=1}^{L} \left(\sigma_p^2 \sigma_{b,l}^2 + \sigma_v^2\right) \tag{6.22}$$

On the other hand, using the matrix inversion lemma (ShermanŰMorrisonŰ-Woodbury formula) [61], one can also show that

$$\mathbf{C}_h^{-1} = \sigma_v^{-2} \mathbf{I}_N - \sigma_v^{-2} \sum_{l=1}^{L} \frac{\sigma_{b,l}^2}{\sigma_{b,l}^2 \sigma_p^2 + \sigma_v^2} \mathbf{p}_{\tau_l} \mathbf{p}_{\tau_l}^H \tag{6.23}$$

Using (6.22) and (6.23), the Fisher Information Matrix (FIM) can be derived for the parameters τ_l, $\sigma^2_{b,l}$, $l = 1 : l$. On the other hand, differentials in these channel parameters can be coupled to differentials in the position (dx, dy). By assuming a certain scenario for obstacle positions and path attenuation exponent, this leads to a FIM for the estimation of (dx, dy). One can show that this FIM is non-singular (with a probability one over a random scenario distribution) for $L \geq 2$. Thus using one BS, the mobile localization is locally identifiable if we consider at least two paths, which is consistent with the identifiability results derived in the framework of classic geometric localization (in the LoS conditions, at least 2 BSs are needed for local identifiability).

6.6 Signature vs. direct position fingerprinting

Location fingerprinting, as any localization, can be implemented using two philosophies:

- Signature based fingerprinting, in which the localization is decomposed on two steps. First, a signal fingerprint (PDP in our case) is computed. Then, the MS location is determined by matching the measured and the stored signal signature.

- Direct position fingerprinting, in which signal signature is not explicitly computed. Using signal structure information previously collected from the whole coverage area, one can select the position from which the measured signal is likely coming from.

6.6.1 Signature based fingerprinting

Signature based fingerprinting performs separately the fingerprint estimation and the matching:

- Estimation stage: computes the signal signature (fingerprint). No-constraints are imposed on the estimate. The estimation scheme ignores the fact that the observed signal comes from an MS located at a given position.

- Matching stage: finds the best match between the computed and stored fingerprints. Prior information on the MS position can be useful to reduce the search area and avoid positioning ambiguities.

Obviously, the overall localization accuracy depends strongly on the fingerprint estimation quality. Particularly, the accuracy of the PDP estimation is affected by two major sources of impairment [31]:

- Additive noise: because the input signal is recorded in presence of noise, the estimated CIR (then the PDP) is always corrupted by a random fluctuation.

- Outlying noise: if the SNR at a given delay fall below a given threshold, the corresponding PDP component will contain almost no useful information on the source localization: these values must be interpreted as outlying components.

Using the simulation environment described in the section 6.4.1, we investigate the effect of the PDP estimation on the localization accuracy of the PDP-fingerprinting. Figure 6.8 compares the RMSE of the PDP-fingerprinting (function of the input SNR) where the PDP is estimated using non-parametric scheme (as in (6.7)) or parametric deterministic model (as in (6.9)).
We remark that the parametric PDP estimation outperforms the non-parametric scheme. In fact, exploiting the prior knowledge of the pulse-shape increases the robustness of the estimation scheme to additive noise and outlying components (by ignoring paths with low energy). This leads to more accurate PDP estimation, which in turn leads to better localization performance.

6.6.2 Direct position fingerprinting

The observed CIR comes from a MS located at an unknown (but possible) position. Thus, the PDP delays and fading amplitudes variances cannot have arbitrary values. Ignoring these constraints makes the previous localization scheme sub-optimal. The Bayesian modeling provides an appropriate framework to solve this problem.
Using the Bayesian structure, the PDP is parameterized by the time delay and the fading variance of the different paths. During the creation and the maintenance of the database, these parameters are estimated and stored at positions in the coverage area on some sampling grid.

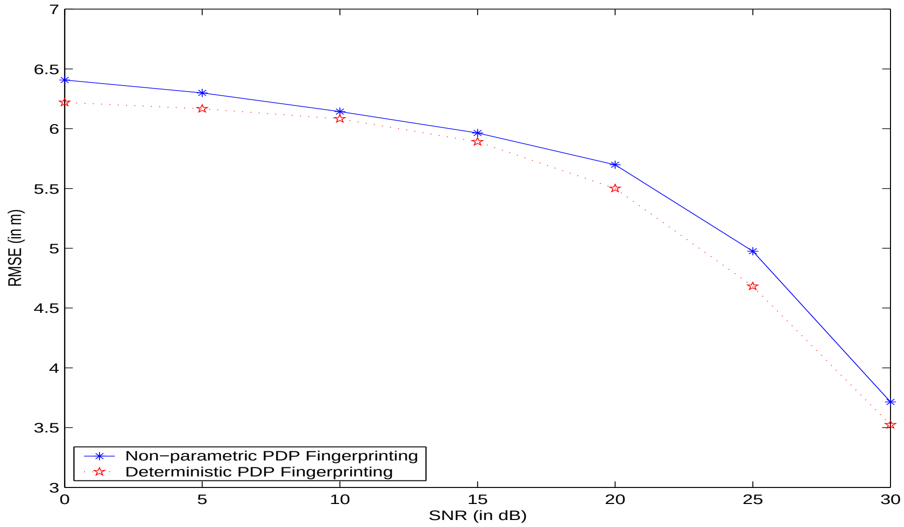

Figure 6.8: Positioning accuracy for PDP-F vs. SNR (using non-parametric and deterministic parametric PDP estimation schemes).

The likelihood that the received CIR $\widehat{\mathbf{h}}(t)$, $t = 1 : T$, comes from a MS located around the position corresponding to the p^{th} database entry is:

$$L\left(\widehat{\mathbf{h}}_1 \cdots \widehat{\mathbf{h}}_T | \boldsymbol{\tau}^{(p)}, \mathbf{C}_b^{(p)}\right) \sim -\ln\left(\det \mathbf{C}_h^{(p)}\right) - \operatorname{tr}\left\{\mathbf{C}_h^{-(p)} \widehat{\mathbf{C}}_h\right\} \qquad (6.24)$$

where $\mathbf{C}_h^{(p)}$ is the channel covariance matrix computed using $\tau^{(p)}$ and $\mathbf{C}_b^{(p)}$ (the time delay and amplitude covariance stored at the p^{th} database entry). $\widehat{\mathbf{C}}_h = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{h}}(t) \widehat{\mathbf{h}}^H(t)$ is the observed sample covariance matrix.

The MS position is the selected by maximizing this likelihood, i.e.,

$$\hat{p} = \arg\max_p L\left(\widehat{\mathbf{h}}_1, \cdots, \widehat{\mathbf{h}}_T | \boldsymbol{\tau}^{(p)}, \mathbf{C}_b^{(p)}\right) \qquad (6.25)$$

This leads to a one step localization approach taking into account the constraints imposed by the MS location. Prior information on the MS location (using, for example, a tracking scheme) can be exploited to reduce the optimization subspace and avoid potential positioning ambiguities.

A posteriori, the proposed approach can be interpreted as a kind of extension of the DPD localization scheme [8, 205] to multi-paths propagation environment.

Differentiating the likelihood leads to

$$\begin{aligned}
\partial L\left(\boldsymbol{\tau}, \mathbf{C}_b\right) &= \partial\Big(-\ln\left(\det \mathbf{C}_h\right) + \operatorname{tr}\left\{\mathbf{C}_h^{-1}\widehat{\mathbf{C}}_h\right\}\Big) \\
&= -\Big(\operatorname{tr}\left\{\mathbf{C}_h^{-1}\,\partial\mathbf{C}_h\right\} \;+\; \operatorname{tr}\left\{\mathbf{C}_h^{-1}\,\partial\mathbf{C}_h\;\mathbf{C}_h^{-1}\widehat{\mathbf{C}}_h\right\}\Big) \\
&= -\frac{1}{2}\;\partial\operatorname{tr}\left\{\mathbf{C}_h^{-1}\Big(\widehat{\mathbf{C}}_h-\mathbf{C}_h\Big)\mathbf{C}_h^{-1}\Big(\widehat{\mathbf{C}}_h-\mathbf{C}_h\Big)\right\} \\
&= -\frac{1}{2}\;\partial\left\|\mathbf{C}_h^{-\frac{1}{2}}\Big(\widehat{\mathbf{C}}_h-\mathbf{C}_h\Big)\mathbf{C}_h^{-\frac{H}{2}}\right\|_F^2 \qquad (6.26)
\end{aligned}$$

where $\left\|\mathbf{C}\right\|_F$ and $\mathbf{C}^{\frac{1}{2}}$ denote respectively the Frobenius norm and a square root of the matrix $\mathbf{C}$.

Thus, the Rayleigh maximum likelihood approach leads to the Optimally weighted Covariance Matching (OCM) method [58]. The parameters are selected in order to match the whole covariance matrix $\widehat{\mathbf{C}}_h$, and not only the PDP (the diagonal elements). It is also remarkable that if the channel impulse response is sufficiently sparse (pulse-shape supports do not overlap), the covariance matrix $\mathbf{C}_h$ is almost diagonal, and the deterministic and Bayesian estimation techniques coincide.

Moreover, prior information on the signal structure are available and can exploited to enhance the estimation of the observed covariance matrix. Different levels of structural information can be considered: subspace decomposition, and high resolutions methods can be used to emphasize the prior structure of the observed covariance matrix. Exploiting the prior structure improves the localization accuracy and resolution, and defines intermediate approaches between the classic geometric and mapping techniques.

6.7 Power space delay profile fingerprinting for mobile localization

In the MISO/MIMO (Multi-Input Single/Multi-Output) cases, we consider Base Stations equipped with an M-element antenna array. Traditional geometric techniques exploit the additional spatial information by estimating jointly the Angle and the Time of Arrival which leads to an increase in the

localization accuracy[172]. On the other hand, traditional mapping methods consider separately the different observations received at the antenna array elements, without exploiting the relation between these observations. In particular, the PDP removes all phase information. Thus the classic PDP fingerprinting techniques do not exploit the spatial information provided by the antenna array reception. Once again, we propose exploiting the prior information on the channel structural and statistical prior to enhance the localization accuracy.
The received impulse response from a MS on an antenna array :

$$\mathbf{h}(t,\tau) = \sum_{l=1}^{L} A_l(t) e^{j\varphi_l(t)}\ \mathbf{g}(\theta_l)\ p(\tau-\tau_l) \tag{6.27}$$

where $\mathbf{g}(\theta_l)$ is the vector response of the antenna array to the l^{th} path in direction θ_l. Under the narrow-band assumption, $\mathbf{g}(\theta_l)$ reflects mainly the phase-shifts that the carrier signal undergoes when imprinting on the consecutive antenna elements from direction θ_l.
We define the Power Spatial Delay Profile (PSDP) as

$$\begin{aligned}\underbrace{\mathbf{PSDP}(\tau)}_{M\times M} &= E_b\left\{\mathbf{h}(t,\tau)\mathbf{h}^H(t,\tau)\right\} \\ &= \sum_{l=1}^{L} \sigma_{b,l}^2\ \left|p(\tau-\tau_l)\right|^2\ \mathbf{g}(\theta_l)\mathbf{g}^H(\theta_l) \end{aligned} \tag{6.28}$$

where $E_b\{.\}$ denotes the expectation over the fading coefficients (assumed to be independent), which can be estimated by assuming local spatial or temporal ergodicity. Remark that the m^{th} diagonal element of the PSDP corresponds to the PDP of the received impulse response between the MS and the m^{th} BS antenna array element. And as the PDP varies slowly with position, those diagonal elements are almost equal. Notice also that the schemes proposed in the sections 6.5 and 6.6 can be easily generalized to the PSDP case. The path angle θ_l and delay τ_l can be jointly estimated using for example the JADE algorithm [194, 195, 196].

As for any fingerprinting based approach, the matching score design is a critical issue. Since the PSDP provides extra freedom degrees, it gives additional flexibility in the design of the matching score function. For instance, to extend the score function in (1), one can compute the matching score between the estimated CIR, and the stored $\mathbf{PSDP}_{x,y}$ as:

$$C(x,y) = \left\| \widehat{\mathbf{PSDP}}(\tau) - \mathbf{PSDP}_{x,y}(\tau) \right\|_F^2 \tag{6.29}$$

where $\widehat{\mathbf{PSDP}}(\tau) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}^H(t,\tau)\mathbf{h}(t,\tau)$ the sampling PSDP estimates. Another possible matching score function is:

$$C(x,y) = \sum_{t=1}^{T}\sum_{\tau} PDP^2_{x,y}(\tau)\left(\mathbf{h}^H(t,\tau)\mathbf{PSDP}^{-1}_{x,y}(\tau)\mathbf{h}(t,\tau) - M\right)^2 \tag{6.30}$$

Remark that for $M = 1$ (in which a case $\mathbf{PSDP} = PDP$), (6.29) and (6.30) lead to the PDP LS cost function (as in (1)).

We propose to investigate the PSDP fingerprinting via the ray tracing multipath environment. We consider $M = 2$, and the matching score (1) for the PDP-F and (6.29) for PSDP-F. We plot the Root Mean Square positioning Error (RMSE) as a function of CIR estimation SNR for 1 and 2 BSs (figure 6.9).

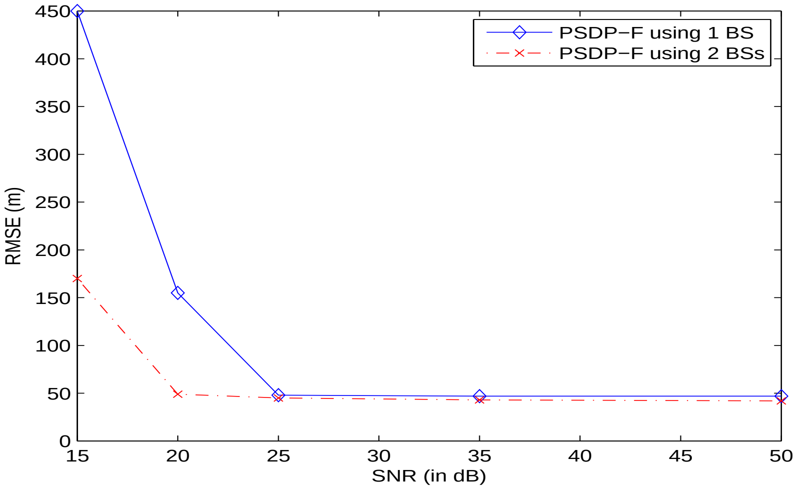

Figure 6.9: Positioning accuracy for PSDP-F vs. SNR using 1 and 2 BSs.

We remark that using multi-BSs reception is advantageous. As usual, performances saturate due to the discretization of the stored PSDP.

Next, we compare the PDP vs. PSDP based approaches (figure 6.10). We fix the spatial discretization for the PDP-F to twice that of the PSDP-F. Remark that with this resolution choice, we have a comparable number of score evaluations for the search algorithms for both PDP and PSDP. However, the PDP database construction and maintenance is much more expensive and time consuming.

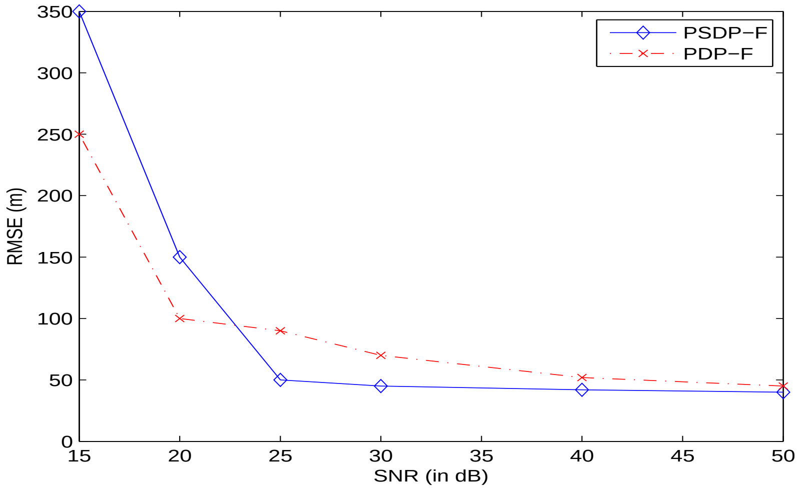

Figure 6.10: Positioning accuracy for PSDP-F vs. PDP-F using 1 BS.

We see that even when using lower resolution, the employment of PSDP is advantageous in the higher SNR region.

6.8 Conclusion

In this chapter, we have investigated the PDP fingerprinting localization technique on multipath propagation. This multipath extension of T(D)oA is based on matching an estimated power delay profile from one or several base stations (BSs)(or other transmitters (broadcast, ...)) with a memorized power delay profile map for a given cell. Not only is the PDP fingerprinting robust to the propagation conditions imposed by wireless communication, but also it exploits multipath propagation instead of combating it.
We have proposed a validation method for PDP-F via simulations that can

easily be reproduced. The multicellular environment consists of a big box in which multipath arises by reflection from the six sides. This ray tracing multipath environment is taken from the acoustics world, and adapted to electromagnetic propagation at a certain carrier frequency. The resulting PDP depends on the positions of BS and terminal, the attenuation mechanism and the reflection coefficients of the six sides.

We have also proposed a parametric deterministic and a Bayesian models to enhance the PDP estimation. The simulations show that the parametric estimation is more robust to the additive noise and outlying components, and leads to an enhancement of the PDP-F localization accuracy. On the other hand, the Bayesian framework seems to be an appropriate introduction to one step localization approaches, that takes into considerations the MS location constraints.

Finally, we have proposed an extension of the PDP-F taking into account spatial information available with a multi-antenna reception. PSDP-F can be considered as a multipath extension of the combined T(D)oA and AoA methods, without explicit requirement for antenna array calibration.

Chapter 7

General Conclusions

In this thesis, we have investigated the audio signal enhancement and Bayesian parameters estimation problems. Particularly, we have underlined the benefit of exploiting prior information on the spectral, spatial, and statistical signal structure. This work has been organized in two parts. The first part investigates several configurations of the acoustical signal enhancement and restoration problem. The second part focuses on Bayesian parameter estimation and applications to channel estimation and mobile localization.

First, we have investigated audio signal enhancement. We have exploited the time-frequency prior structure of audio signals. We have modeled an elementary audio signal as a periodic signal with slow global variation of amplitude (characterizing the temporal evolution of the signal power) and phase (emphasizing the harmonic structure). The bandlimited variation of global amplitude and phase gets expressed through a subsampled representation and parametrization of the corresponding signals. Assuming additive white Gaussian noise and small time warping variation, a Maximum Likelihood approach was proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems.
In chapter 2, we have applied the proposed structural decomposition to the classic noise reduction problem. Simulations show that the proposed scheme

is suitable for the analysis of musical notes, and produces good auditive synthetic results. We have also considered application to speech enhancement. The prior structure was exploited to identify and enhance voiced frames. Simulations show that the enhancement technique achieves quite good performance (especially in very noisy environments).
In chapter 3, we have investigated underdetermined convolutive audio source separation. We have proposed a separation technique that takes into account simultaneously the source signal structure and the propagation environment parameters (ToA, signal attenuation). Experimental results reveal that the proposed approach allows extracting several musical notes accurately from an underdetermined mixture, and produces good auditive synthetic results. Simulations show also that the proposed scheme outperforms the classic separation schemes in terms of accuracy and robustness.

For multi-microphone configurations, additional spatial information is available. We have showed that: despite the source position being unknown, a SIMO channel tends to become allpass as the number of sub-channels and/or the reverberation delay spread increases. We call such prior spatial information "spatiotemporal diversity". In chapter 4, we investigate the blind dereverberation of audio signals. We propose a multichannel linear prediction based equalizer, exploiting spatial, temporal, and spectral diversities. Simulations show that the proposed Delay-&-Predict Equalizer scheme performs better than the classic Delay-&-Sum Beamformer, especially if only few microphones are available.
We have also investigated two robustness issues in the design of the LP-based equalizer in the presence of additive white noise. First, we have examined the effect of relative subchannel delay compensation on the output SNR. We show that such relative delay compensation can increase considerably the output SNR. Then, we have optimized the transformation of the multivariate prediction filter to a longer equalizer filter using the SNR criterion. The optimization corresponds to MMSE-ZF design, and the post-filter length increase allows for the introduction of some equalization delay, that can also be optimized. Simulations show that considerable gains can be achieved by allowing even small equalization delays.

Part II focuses on Bayesian parameter estimation. Classic Bayesian approaches often lead to useful MSE reduction, but they also introduce a bias. On the other hand, imposing a joint conditionally unbiasedness constraint on a vector of parameters reduces Bayesian estimation to deterministic parameter estimation, throwing away all prior information. In chapter 5, we introduce the concept of Component-Wise Conditionally Unbiased (CWCU) Bayesian parameter estimation, in which unbiasedness is forced for one parameter at a time. In CWCU parameter estimation, every parameter in turn is treated as deterministic while the others are being treated as Bayesian. If the parameters are transmitted symbols, the CWCU approach corresponds to unbiased symbol detection whereas joint deterministic unbiasedness leads to a zero-forcing approach. Moreover, if the prior covariance matrix has a limited rank, we show that block-CWCU estimation (with an appropriate block size) reduces the estimation noise, while guaranteeing joint unbiasedness. The more general introduction of the CWCU concept is motivated by LMMSE channel estimation, for which the implications of the concept are illustrated in various ways, including the effect on angle of arrival estimation, repercussion for blind channel estimation etc. Application to mobile localization was also considered in more detail in chapter 6. In fact, we have investigated the PDP fingerprinting location technique on multipath propagation. This multipath extension of T(D)oA is based on matching an estimated power delay profile from one or several base stations with a memorized power delay profile map for a given cell. Not only is the PDP fingerprinting robust to the propagation conditions introduced by wireless communication, but also it exploits multipath instead of combating it. We have proposed a validation of PDP-F via simulations that can easily be reproduced. The multicellular environment consists of a big box in which multipath arises by reflection off the six sides. This ray tracing multipath environment is taken from the acoustics world, and adapted to electromagnetic propagation at a certain carrier frequency. The resulting PDP depends on the positions of BS and terminal, the attenuation mechanism and the reflection coefficients of the six sides. We have also proposed parametric deterministic and Bayesian channel models to enhance the PDP estimation. The simulations show that parametric estimation is more robust to additive noise and outlier components, and also leads to an enhancement of the PDP-F location accuracy. Finally, we have also proposed an extension of PDP-F, taking into account spatial information that becomes available with multi-antenna reception or transmission. The PSDP-F can be considered as a multipath extension of the combined T(D)oA and AoA methods, without explicit requirement for antenna array calibration.

7.1 Prespectives

This work has proposed different schemes and techniques, as well as the analysis of different practical and theoretical scenarios. On the other hand, our work opened new problems. We list hereafter some research directions arising from this thesis :

- In the speech enhancement scheme proposed in section 2.6.3, we have focused on denoising voiced frames (exploiting the prior structural/harmonic information). Extra investigations on unvoiced frame enhancement (possibly the combination of voiced and unvoiced denoising) are needed to increase the overall scheme enhancement accuracy.

- In the separation scheme introduced in chapter 3, the prior audio structure was exploited to enhance the sparse decomposition of the mixtures; whereas the signal classification exploits only the pitch information (as in the classic sparse decomposition methods). However, each of the structural decomposition outputs (power and pitch evolutions, spectral envelope) characterizes the audio source. Exploiting this information in the classification step should increase the total separation performance.

- The dereverberation scheme introduced in chapter 4 exploits spatial, temporal, and spectral diversity of the audio signal. On the other hand, it ignores the harmonic structure of such a signal. The harmonic prior was showed to be effective to remove late reverberation [123]. Post-processing the dereverberation output (taking into account local signal structure) seems to be an effective way to enhance the whole dereverberation performance.

- In chapter 4, we have considered mono-source blind speech dereverberation. Extending the Delay-&-Predict approach to multi-source dereverberation still an open problem.

- We have also introduced the general concept of the CWCU estimation. The concept was motivated and illustrated with concrete examples. However, the implications of the concept to several applications need to be investigated (in particular to audio processing and Maximum Likelihood symbol detection).

Résumé en Français

Chapter 8

Résumé en Français

8.1 Introduction

Le débruitage audio est un composant vital dont dépend la performance des systèmes de communication audio opérant dans des environnements bruyants. Typiquement, la qualité d'un signal audio (enregistré dans un environnement réel) est inévitablement dégradée par l'interférence acoustique (figure 8.1). En effet, un signal audio est soit produit dans un environnement bruité, soit distordu par le canal acoustique. Cette interférence peut être globalement classifiée en deux catégories: additive et convolutive.

- Le bruit ambiant provient des sources audio avoisinantes: bruit de fond, musique, etc. Compte tenue du principe de superposition, nous supposons que la contribution de bruit ambiant est additive. Nous assumons aussi que le bruit ambiant est indépendant de la source audio d'intérêt.

- L'interférence convolutive (généralement désignée sous le nom de la réverbération) est due aux réflexions des ondes sonores sur les murs et des objets avoisinants. Elle entraîne la modification des caractéristiques du signal de la parole. Par conséquent, elle constitue un problème majeur dans plusieurs applications.

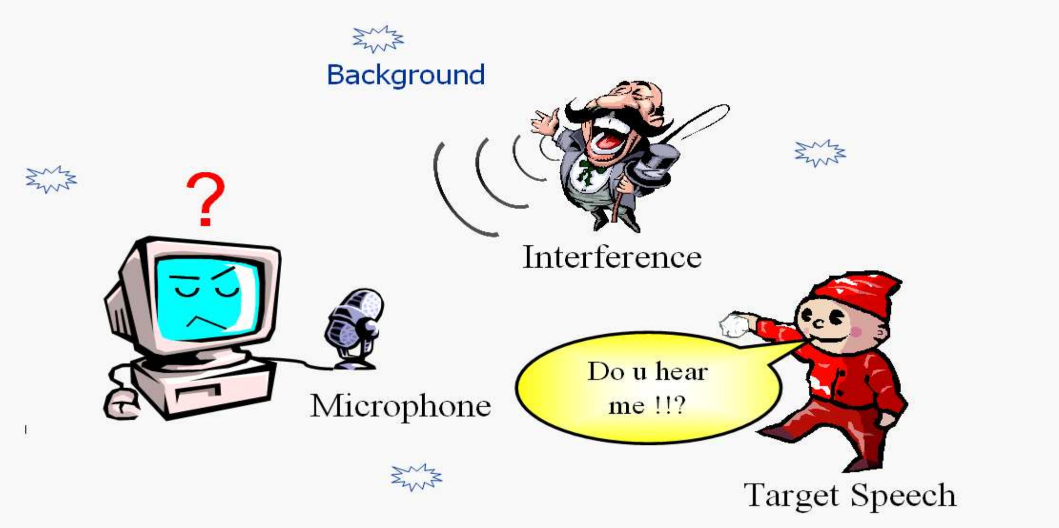

Figure 8.1: Audio signal captured in a real-world environment.

Le signal audio bruité est typiquement capté par un ensemble de microphones. En combinant les différentes observations et en utilisant des outils appropriés de traitement de signal, le débruitage audio vise restaurer au mieux le signal audio d'origine (figure 8.2)

Le débruitage des signaux audio est considéré comme problème difficile due à la nature aveugle du problème et aux variations rapides des caractéristiques des signaux de la parole et du bruit. Cependant, si des informations aprioris (sur la structure ou les statistiques du signal) sont disponibles, les performances du débruitage augmentent d'une manière significative en exploitant de tels aprioris. Dans cette thèse, nous étudions trois types d'apriori: spectrale, spatiale, et statistique; et nous considérons particulièrement des applications au débruitage audio et à la localisation des mobiles.

D'abord, nous étudions la représentation structurale du signal audio. Le modèle proposé exploite l'espacement et les corrélations tempofréquentielles du signal audio. Nous appliquons notre modèle au débruitage audio la séparation audio sous-déterminée. Les résultats expérimentaux montrent que l'approche proposée convient à l'analyse des signaux de musiques et de la parole, et produisent de bons résultats auditifs. Les simulations prouvent

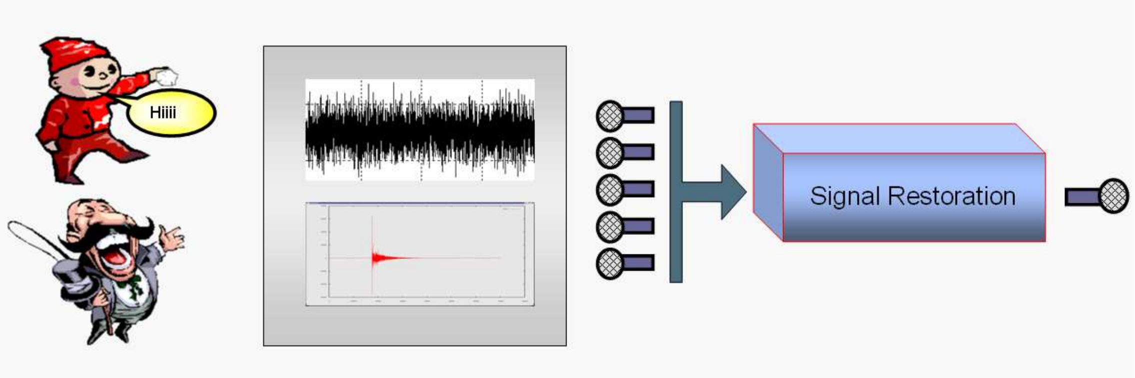

Figure 8.2: Acoustical signal enhancement and restoration.

également que le schéma proposé surpasse les schémas “matching pursuit” classiques en termes d’exactitude et de robustesse de séparation.

Ensuite, nous étudions le dereverberation aveugle des signaux audio. Nous proposons un égaliseur basé sur la prédiction linéaire multicanale, exploitant les diversités spatiales, temporelles, et spectrales. Les simulations prouvent que l’égaliseur proposé (Delay-&-Predict) surpasse le filtre spatial classique (Delay-&-Sum).

La dernière partie de la thèse se concentre sur l’estimation Bayésienne des paramètres. Les approches Bayésiennes classiques produisent une réduction utile du MSE, mais en dépit d’un biais non nul (souvent gênant dans plusieurs applications). Nous introduisons le concept d’estimation conditionnellement non-biaisé par morceau, pour laquelle la contrainte du biais concerne un paramètre à la fois. De cette manière, chaque paramètre est traité comme déterministe tandis que les autres paramètres sont traités comme Bayésiens. Une introduction plus générale du concept est motivée par l’estimation LMMSE des canaux, pour laquelle les implications du concept sont illustrées dans diverses manières. L’application à la localisation des mobiles est étudiée en détails.

8.2 Extraction de Signaux Périodiques avec Modulation Globale d'Amplitude et de Phase

8.2.1 Introduction

Dans le cadre de l'analyse/synthèse de signaux audio, le modèle sinusoïdal a reçu un intérêt considérable et s'est avéré efficace et utile dans plusieurs applications: compression, séparation de source, réduction de bruit...

Le modèle sinusoïdal représente le signal audio comme une somme discrète de sinusoïdes variantes dans le temps. Les paramètres du modèle sont typiquement estimés en utilisant une transformée de Fourier à court-terme (STFT). La taille et le recouvrement des fenêtres sont généralement fixés a priori. Les sinusoïdes sont identifiées et extraites dans chacune de ces fenêtres; les valeurs intermédiaires sont fixées par interpolation. Un problème fondamental des techniques basées sur les modèles sinusoïdaux classiques est que puisque le signal audio est fortement non-stationnaire, il n'est pas toujours possible de trouver un bon compromis entre la résolution temporelle et fréquentielle. Un autre inconvénient est que ces techniques ignorent la structure harmonique des signaux audio.

D'autre part, en traitant les signaux périodiques, l'état de l'art actuel se limite à l'estimation des signaux périodiques purs (avec une période égale à un nombre entier d'échantillons) [134]. Dans cette référence, les auteurs proposent une approche par maximum de vraisemblance pour analyser les signaux périodiques. Ils montrent que le schéma proposé peut être interprété comme une projection du signal sur des sous-espaces adaptés.

Dans le present travail, nous étendrons les résultats des références présentes et nous proposons de fusionner la modélisation sinusoïdal et les techniques d'analyse des signaux périodiques. Nous modélisons le signal audio comme un signal périodique de période pas nécessairement entière et une variation globale (lente) d'amplitude et de fréquence (time-warping).

Dans un second temps, nous introduisons plus de flexibilité sur la modélisation de la variation globale de phase. Nous décomposerons la phase instantanée en une composante linéaire par morceau (modélisant la variation lente de la fréquence instantanée), et une composante de faible amplitude (modélisant les fluctuations de la phase instantanée). Appliqués à des mélanges de signaux musicaux, ces nouveaux degrés de liberté permettent la modélisation de plusieurs phénomènes musicaux (vibrato, glissando...), et améliorent la performance de la séparation.

8.2.2 Modèle quasi-périodique avec modulation globale en fréquence et amplitude

La modélisation sinusoïdale représente le signal audio comme somme discrète de sinusoïdes variant dans le temps:

$$s(t) = \sum_{k=0}^{P} a_k(t) \cos\left(\theta_k(t)\right) \quad . \tag{8.1}$$

$A_k(t)$ et $\theta_k(t)$ représentent respectivement l'amplitude et la phase instantanée de la k^{eme} partielle. Compte tenu de l'harmonicité du signal audio, $\theta_k(t)$ se décompose en:

$$\theta_k(t) = 2\pi k t f_0 \; + \; 2\pi\varphi_k(t) \tag{8.2}$$

ou $\varphi_k(t)$ caractérise l'évolution de la phase instantanée autour de la k^{eme} harmonique; et varie lentement dans le temps.
L'hypothèse de modulation globale sous-entend que les amplitudes des différentes harmoniques évoluent proportionnellement dans le temps, et que les fréquences instantanées sont linéairement corrélées, i.e.,

$$\left\{ \begin{array}{l} a_k(n) = a_k \; a(n) \\ 2\pi\varphi_k(n) = 2\pi k \; \varphi(n) \; + \; \Phi_k \end{array} \right. \quad . \tag{8.3}$$

En résumé, nous modélisons un signal audio comme superposition de composantes harmoniques avec une modulation globale d'amplitude, et un time-warping (qui peut être interprété en termes de variations globale de phase):

$$\begin{array}{l} y(n) \; = \; s(n) \; + \; v(n) \\ = \; \sum_k a_k(n) \; \cos\left(2\pi k n f_0 + 2\pi\varphi_k(n)\right) + v(n) \\ = \; a(n) \underbrace{\sum_k a_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right)}_{\theta\left(n + \frac{\varphi(n)}{f_0}\right)} + v(n) \end{array}$$

où,

- v_n est un bruit blanc, additive, Gaussien.

- $a(n)$ représente le signal modulé en amplitude. Il caractérise/permet l'évolution de l'énergie du signal, reflétant l'attaque, le maintien, et l'atténuation du signal.

- $\varphi(n)$ représente le signal modulé en phase (qui peut être interprété en terme de time-warping).

- $\theta(n) = \sum_k a_k \cos(2\pi k f_0 n + \Phi_k)$ est un signal périodique de période $T = \frac{1}{f_0}$ (T pas nécessairement un nombre entier).

Dans [178], nous avons supposé que la fréquence instantané est constante par morceau; et nous avons exprimé le time-warping par le biai d'un opérateur d'interpolation agissant sur le signal périodique de base ($\theta(n)$). Nous avons également proposé un schéma d'extraction basé sur l'estimation cyclique des différents parametres du modèle.

8.2.3 Modèle quasi-périodique avec modulation globale en phase et amplitude

Dans la section précédente, nous supposons que le signal modulé en phase ($\varphi(n)$ en (8.1)) est linéaire par morceau, i.e., $\exists T$

$$\varphi_{wl}(n) = n\,(f_0 + f_p) + \Phi_p \quad \forall n \in [pT \quad (p+1)T]$$

ou $f_{wl}(n) + f_0 = f_p + f_0$ est la fréquence instantanée supposé constante par morceau. Dans ce cas, la modulation globale de phase peut être interprété en terme de time-warping.

Dans cette section, nous relaxons d'avantage nos hypothèses sur la modulation globale en phase, en supposant que:

$$\varphi(n) = nf_0 + \varphi_{wl}(n) + \widetilde{\varphi}(n) \tag{8.4}$$

$\widetilde{\varphi}(n)$ variant lentement dans le temps, et ayant une faible magnitude ($|2\pi\widetilde{\varphi}(n)| \ll 1$). Ainsi, le signal audio peut s'écrire comme:

$$s(n) = a(n) \sum_k a_k \cos(2\pi k\,(nf_0 + \varphi_{wl}(n)) + 2\pi k\widetilde{\varphi}(n) + \Phi_k)$$

L'approximation au premier ordre (par rapport à $\widetilde{\varphi}(n)$) se traduit par un terme additif (fonction de la dérivée du signal périodique $\theta(n)$), i.e.

$$\begin{aligned}
s(n) &\approx a(n)\sum_k a_k \cos\left(2\pi k\left(nf_0+\varphi_{wl}(n)\right)+\Phi_k\right)\\
&\quad -a(n)\sum_k a_k(2\pi k\widetilde{\varphi}(n))\sin\left(2\pi k\left(nf_0+\varphi_{wl}(n)\right)+\Phi_k\right)\\
&= a(n)\theta\left(nf_0+\varphi_{wl}(n)\right)+a(n)\frac{\widetilde{\varphi}(n)}{f_0+\varphi_{wl}^{'}(n)}\theta^{'}\left(nf_0+\varphi_{wl}(n)\right)\\
&= a(n)\theta\left(n+\tfrac{\varphi_{wl}(n)}{f_0}\right)+a(n)\underbrace{\frac{\widetilde{\varphi}(n)}{f_0+f_{wl}(n)}}_{B(n)}\theta^{'}\left(n+\tfrac{\varphi_{wl}(n)}{f_0}\right)
\end{aligned}$$

La dérivée $\theta^{'}(n)$ représente la version échantillonnée de la dérivée du signal en temps continu $\theta(t)$ (dont nous disposons uniquement de la version échantillonné $\theta(n)$). Si l'échantillonnage vérifie le critère de Nyquist, alors $\theta^{'}(n)$ peut être obtenue en filtrant $\theta(n)$ par la fonction de transfère $H^{o}(f)=j2\pi f$, $f\in(-\frac{1}{2},\frac{1}{2})$. Une approximation de la fonction de transfert précédente peut être obtenue en optimisant:

$$H(z)=\sum_{n=-P}^{P}h_n z^{-n} \tag{8.5}$$

$$\min_{h_n}\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{yy}(f)\left|j2\pi f-H\left(e^{j2\pi f}\right)\right|^2 df \tag{8.6}$$

où P est l'ordre du filtre á réponse impulsionelle finit $H(z)$ approximant le filtre dérivée $H^{o}(z)$, et $S_{yy}(f)$ représente le spectre du signal $y(n)$.
En résumé, le signal audio s'écrit comme:

$$Y=A\ F\theta+\underbrace{A\ B}_{C}\ HF\theta\ +\ V \tag{8.7}$$

où,

- $Y=[y(1)\cdots y(N)]^T$ représente le vecteur observation.
- $V=[v(1)\cdots v(N)]^T$ représente le vecteur bruit.
- $\theta=[\theta(1)\cdots\theta(\lceil T\rceil)]$ caractérise le signal périodique de base.

- $A = diag\,\{a(1) \cdots a(N)\}$ représente le signal modulé en amplitude.
- $B = diag\left\{\frac{\widetilde{\varphi}(1)}{f_0+f_{wl}(1)} \cdots \frac{\widetilde{\varphi}(N)}{f_0+f_{wl}(N)}\right\}$ caractérise le signal modulé en phase.
- F is an $(N+2P) \times \lceil T \rceil$ est une matrice d'interpolation caractérisant le time-warping.
- H is an $N \times (N+2P)$ est une matrice bande caractérisant le filtre dérivée.

8.2.4 Application à la séparation aveugle des mélanges sous-déterminés.

La majorité des algorithmes de séparation aveugle de sources se basent sur la théorie de l'Analyse en Composantes Indépendante. L'idée est d'estimer l'inverse de la matrice de mixage en utilisant l'indépendance statistique des sources. Cependant, un domaine de recherche, la séparation de mélanges sous-déterminés, reste relativement moins exploité. Il s'intéresse au cas ou il y a moins de mélanges que de sources. La séparation de mélanges sous-déterminés pose un défi parce que la matrice de mixage n'est pas inversible et les méthodes traditionnelles ne fonctionnent plus. Et, contrairement la plupart des algorithmes de séparation aveugle, l'extraction des sources elle-même nécessite des informations additionnelles sur les statistiques des sources ou de leurs structures.

Les modèles présentés ci-dessus peuvent être adaptés à la séparation de mélanges sous déterminés. Une approche d'annulation successive d'interférence (SIC) itérative (basée sur l'extraction de signaux Quasi-Périodiques) est dérivée pour la séparation de sources audio sous-déterminées. Nous utilisons le schéma proposé pour la séparation de sources audio à partir d'un mélange unique. Les observations représentent un mélange synthétique de trois notes (jouées par une guitare acoustique). L'enregistrement a une durée de 1 seconde et est échantillonné à 22.050 kHz. Les pitchs des différentes notes sont respectivement 82 Hz, 92 Hz, 116 Hz. Le Rapport Signal-sur-Bruit (SNR) d'entrée est de 26 dB.

Comme critère d'évaluation, nous utilisons le Rapport Signal sur Bruit (de mesure + modélisation) calculé sur la durée totale de la note ainsi que dans la région de convergence. Dans la figure 8.4, nous traçons le SNR basé sur les deux modèles proposés. Nous observons que la deuxième version atteint

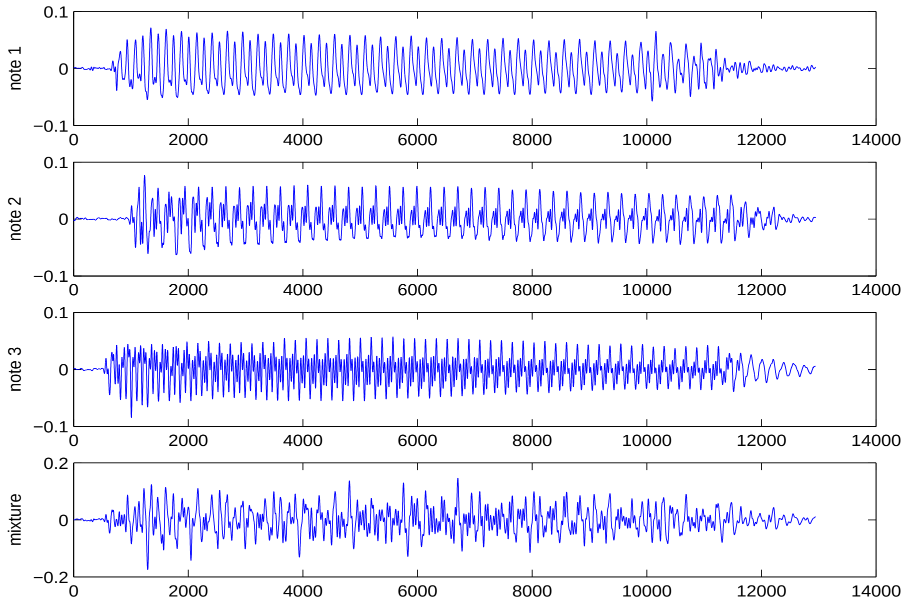

Figure 8.3: Mélangé audio sous-déterminé.

de meilleures performances, aussi bien dans les régions transitoires que dans la région de convergence. Cependant, les simulations montrent qu'elle est assez sensible à l'initialisation des paramètres. Ainsi, nous proposons utiliser le résultat du premier algorithme pour initialiser le second.

Ensuite, nous comparons des performances du schéma de séparation proposé avec les techniques classiques de décompositions parcimonieuses. En effet, plusieurs algorithmes de séparation de source sous-déterminée se basent sur une représentation parcimonieuse du signal audio, suivie par une opération de masquage permettant d'isoler la source d'intérêt. Un signal admet une décomposition parcimonieuse dans un dictionnaire $D = \{g_k(n)\}$ s'il peut être approximer par une combinaison linéaire d'un petit nombre d'atomes $g_k(n)$. L'algorithme Matching Pursuit (MP) selecte itérativement les atomes $g_k(n)$ et calcule leurs poids correspondant. Son principe de base est de sélectionner à chaque itération l'atome le plus proche (corrélé) avec le résidu, puis de mettre à jour le signal résidu en ôtant la contribution de l'atome sélectionnée. Les critères d'arrêt les plus courants sont basés sur le niveau absolu ou relatif de l'énergie du résidu et/ou sur un nombre fixe d'itération à effectuer.

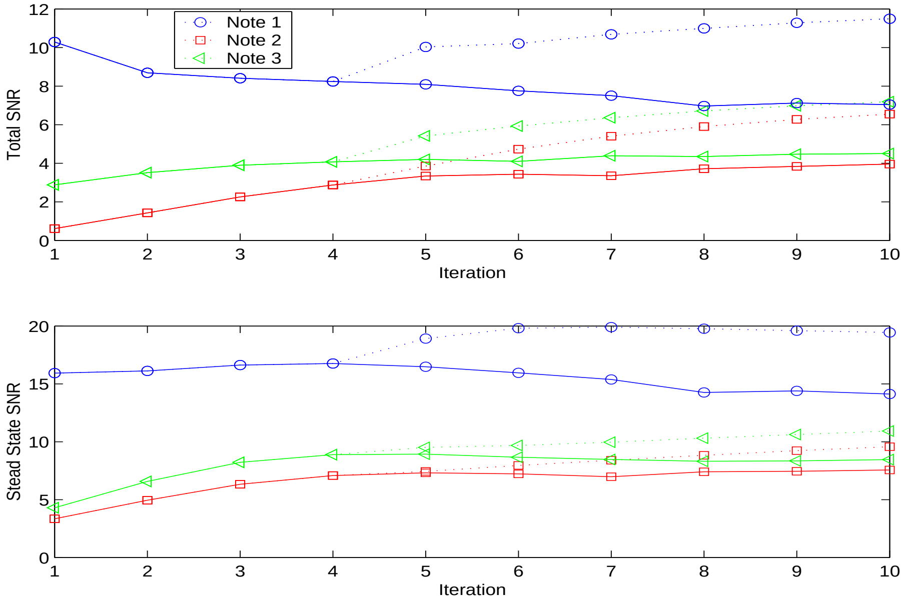

Figure 8.4: RSB de la séparation de source mono-canal (modèle avec modulation globale en amplitude et de fréquence en line continue, et modèle avec modulation globale en phase et amplitude en ligne pointillé).

Gribonval et Bacry proposent une variante de l'algorithme MP (Harmonic Matching Pursuit (HMP)) exploitant de l'harmonicité du signal audio pour structurer la sélection des atomes et accélérer la convergence.
La représentation parcimonieuse des signaux audio a été appliquée à la séparation de source sous-déterminée. Le principe de base est qu'une représentation "efficace" peut décomposer un problème de séparation sous-déterminé en plusieurs problèmes surdéterminés. Dans une configuration mono-microphone, le mélange est séparable si au plus une seule source est active dans n'importe quelle composante de la décomposition. La séparation est effectuée en deux étages:

- décomposer les mélanges en " composantes" (atomes).
- Effectuer la séparation pour chaque atome (ce qui revient a une opération de classification).

La comparaison entre les schémas SIC itérative, MP, et HMP est résumé dans les tableaux 8.1 et 8.2.

	I-SIC	MP	HMP
Note 1	11.8	10.81	8.6
Note 2	7.11	4.57	6.1
Note 3	7.6	1.3	6.76

Table 8.1: Rapport Signal sur Bruit (en dB) pour les schémas SIC itérative, MP, et HMP (calculé sur la durée totale de la note)

	I-SIC	MP	HMP
Note 1	19	14	12.57
Note 2	10.4	4.3	9.6
Note 3	11.61	0.21	11.09

Table 8.2: Rapport Signal sur Bruit (en dB) pour les schémas SIC itérative, MP, et HMP (calculé sur la région de convergence)

Nous remarquons que le Matching Pursuit n'arrive pas à reconstruire la 'note 3' (de la figure 8.3); et que l'exploitation de la structure harmonique du signal audio (dans les schémas SIC itérative et HMP) augmente les performances de séparation (spécialement dans la région de convergence). On note aussi que le schéma SIC itérative produit de meilleures résultats objectives (Rapport Signa-sur-Bruit) et subjective (résultats audibles).

Nous considérons également le cas multi-entrées multi-sorties (figure 8.5). Dans notre simulation, les sources audio est captée par deux microphones (séparés par d=0.2m). Les angles d'arrivés des trois sources sont respectivement $\phi_1 = -\frac{\pi}{3}$, $\phi_2 = 0$, et $\phi_3 = +\frac{\pi}{3}$. L'atténuation relative est respectivement$\beta_{21} = 0.9$, $\beta_{22} = 1$, et $\beta_{23} = 1.1$.
La figure 8.6 présente les courbes du rapport signal sur bruit (calculé sur la durée totale de la note) pour les scénarios mono-entrée (trait continu) et multi-entrée (trait interrompu). Nous observons que les délais et les atténuations relatives sont bien estimés, et que l'algorithme est capable d'exploiter la dimension supplémentaire du problème.

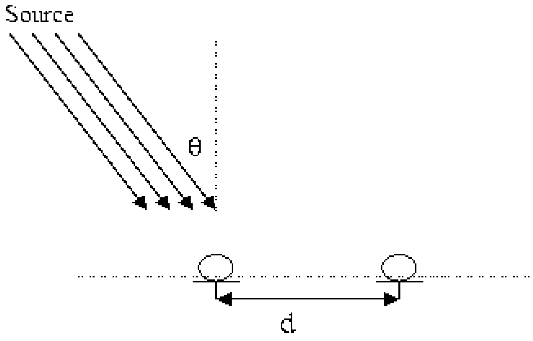

Figure 8.5: Scénario de propagation multi-entrées multi-sorties.

8.3 Égalisation Delay-and-Predict pour la déréverbération aveugle de la parole

8.3.1 Introduction

La qualité de la parole capturée dans les environnements réels est invariablement dégradée par l'interférence acoustique. Cette interférence peut être classifiée en deux catégories: additif et convolutive. L'interférence convolutive (généralement désignée sous le nom de la réverbération) est due aux réflexions des ondes sonores sur les murs et les objets avoisinants. Elle entraîne la modification des caractéristiques du signal de la parole. Par conséquence, elle constitue un problème majeur dans la reconnaissance de la parole, l'identification vocale, et le confort auditive général dans des applications de téléphonie "mains libres". La déréverbération aveugle est le processus d'enlever l'effet de la réverbération d'un signal réverbéré. La réduction de cette déformation est un problème difficile de déconvolution aveugle. Cette difficulté est due à la nature large bande du signal de la parole et à la longueur de la réponse impultionnelle acoustique. Ce problème a été adressé intensivement ces dernières années; mais sans grand succès.

Nous considérons le problème de déréverbération aveugle multicanaux.

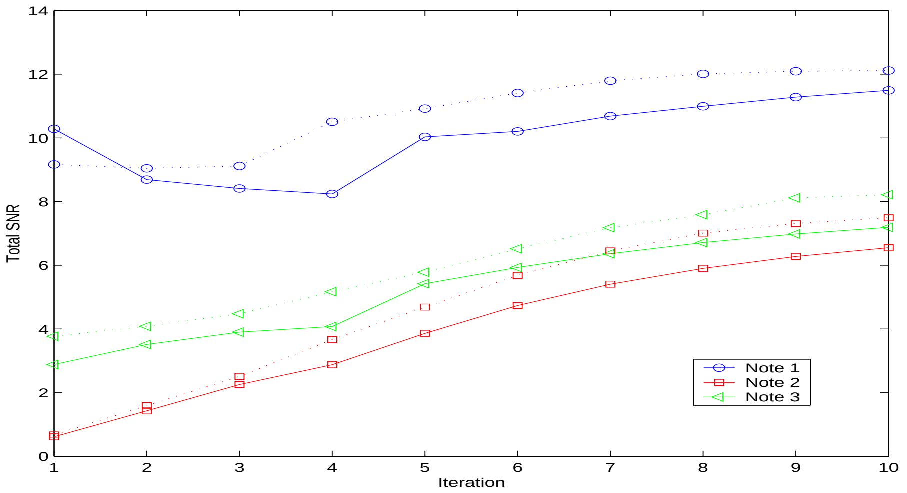

Figure 8.6: Rapport Signal sur Bruit (calculé sur la durée totale de la note) pour les scénarios mono-entrée (trait continu) et multi-entrée (trait interrompu).

Nous supposons que la réponse impultionnelle est suffisamment stationnaire pour permettre l'évaluation des corrélations des signaux reçus. Pour une entrée blanche,la prédiction linéaire multicanaux a été montré efficace pour l'égalisation aveugle des canaux SIMO. Cependant si l'entrée est colorée, la prédiction linéaire non seulement égalise le canal mais aussi blanchit la source. Ce qui produit un effet auditif désagréable.
Nous exploitons le fait qu'un filtre SIMO tend à devenir passe-tout avec l'augmentation du nombre des sous-canaux et/ou de l'étalement temporel, pour estimer la structure des corrélations de source. Cette structure est exploitée pour déterminer un filtre blanchisseur de l'entrée. Un égaliseur à forçage-à-zéro est ensuite estimé en appliquant la prédiction multicanaux aux signaux pré-blanchi. Il est important de souligner que la non-stationnarité de la source ne pose aucun problème tant que les corrélations de source, et la prédiction linéaire sont estimées sur la base des mêmes données.

Généralement, les schémas de déréverbération sont conçus en absence de bruit additif (vue que le problème reste assez difficile même dans ce cas idéal). Cependant pour une utilisation pratique, la robustesse des algorithmes proposés au bruit additive est indispensable. Dans le présent travail, nous proposons deux mesures pour améliorer la robustesse de notre schéma de déréverbération:

- L'alignement des signaux de prédiction qui entraîne l'amélioration des performances de prédiction ainsi que l'utilisation de prédicteurs plus court.
- L'optimisation de l'égaliseur qui entraîne une meilleur combinaison des sorties du prédicteur linéaire multi-variable (tenant compte du bruit additif)

8.3.2 Égalisation Delay-and-Predict pour la déréverbération aveugle de la parole

Le problème que pose l'utilisation de la prédiction linéaire pour l'égalisation aveugle des canaux SIMO est que, pour une entrée colorée, la prédiction linéaire égalise le canal; mais aussi blanchit la source. Cependant, si la couleur de l'entrée est connue (ou peut être estimé), ce problème peut être résolu par le biais d'un prétraitement du signal reçu.
Dans les références [183, 184], nous proposons un schéma de déréverbération en trois étages (voir figure 8.7):

- En premier lieu, nous exploitons les diversités spatiale et temporelle du signal parole afin d'estimer la couleur du signal d'entré; puis prétraiter le signal reçu pour omettre les corrélations due á la couleur de la source.
- Ensuite, un prédicteur aveugle multicanal est calculé (avec les corrélations du signal reçu pré-blanchi).
- Enfin, les colonnes du prédicteur linéaire sont combinées pour définir un égaliseur á forçage á zéro. Cet égaliseur sera utilisé pour la déréverbération du signal reçu.

i) Blanchiment de la source

D'après les résultats de la théorie statistique des salles acoustiques, on peut montrer que pour les fréquences $f > f_{sch} = 2000\sqrt{T_{60}/V}$, le spectre moyen de la réverbération est plat, i.e.,

$$\left\langle \left| H\left(\exp^{2j\pi f}\right) \right|^2 \right\rangle = \frac{1-\beta}{\pi A \beta} \tag{8.8}$$

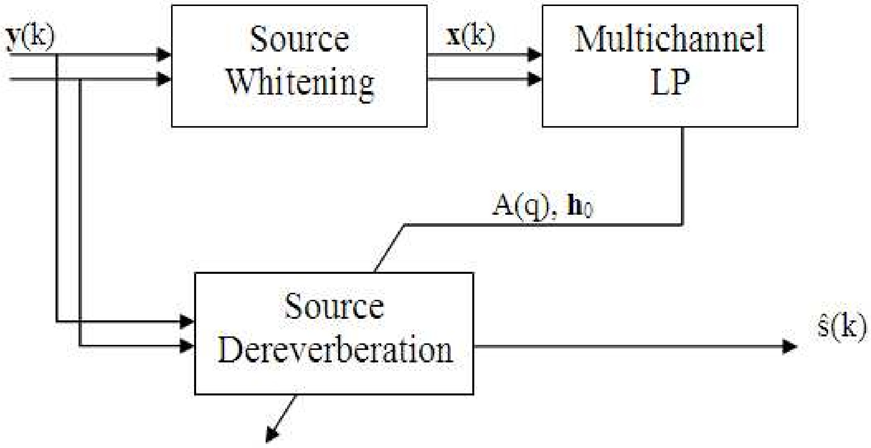

Figure 8.7: La procédure de déréverbération.

oú $\langle . \rangle$ est l'espérance spatiale, β est le coefficient moyen de l'absorption acoustique des murs, A est la surface totale des murs, f_{sch} est la fréquence de Schroeder, T_{60} est le temps de réverbération, et V est le volume de la salle. Dans [183], les simulations montrent que la superposition des spectres des sous canaux d'un SIMO devient plat si le nombre des sous canaux augmente. Par conséquence, la superposition des spectres des signaux reçus estime (á un facteur multiplicatif prés) le spectre de la source audio. Puisque ces corrélations communes sont celles d'un signal de parole, ils peuvent être modélisés par un processus AutoRégressif (AR). Les coefficients de ce processus AR sont calculés en minimisant la puissance de l'erreur de prédiction, moyenné sur les différents microphones, i.e.

$$e = \sum_{k=1}^{M} \sum_{n=0}^{\infty} e_k^2(n) = \sum_{k=1}^{M} \sum_{n=0}^{\infty} \left[y_k(n) - \sum_{j=1}^{l} a_j y_k(n-j) \right]^2 \tag{8.9}$$

Une fois la couleur de la source est estimée, le signal reçu est blanchi:

$$\mathbf{x}(k) = a(q)\mathbf{y}(k) \approx \mathbf{H}(q)\tilde{s}(k) \tag{8.10}$$

oú $\mathbf{x}(k) = [x_1(k) \cdots x_M(k)]^T$ est le signal reçu blanchi, $a(q) = 1 + \sum_{j=1}^{l} a_j q^{-j}$ est le filtre d'erreur de prédiction du signal source, $\tilde{s}(k)$ est le signal d'erreur de prédiction de la source (blanc).

ii) Prédiction multi-canal

Dans le paragraphe précédent, nous avons mis en évidence que la diversité spatio-temporel du canal peut être exploité pour estimer la structure des corrélations de la source. Le fait qui permet le blanchiment de l'observations.
Nous retrouvons ainsi le problème classique d'égalisation d'un canal SIMO (entrée "blanche"). Dans ce cas, le schéma de déconvolution classique (basé sur la prédiction linéaire multicanaux) peut être utilisé.
Le signal d'erreur du prédiction de $\mathbf{x}(k)$ (á partir des L_A dernières observations $\mathbf{X_{L_A}}(k-1) = [\mathbf{x}^T(k-1) \cdots \mathbf{x}^T(k-L_A)]^T$ s'écrit:

$$\tilde{\mathbf{x}}(k) = \mathbf{x}(k) + \sum_{i=1}^{L_A} A_{L_A,i}\mathbf{x}(k-i) = A_{L_A}\mathbf{X_{L_A+1}}(k) \tag{8.11}$$

oú $A_{L_A} = [I_m \; A_{L_A,1} \; \cdots \; A_{L_A,L_A}]$, $A_{L_A,i}$ sont les coefficients matricielle du filtre de prédiction linéaire. Ces coefficients sont déterminé en minimisant l'énergie de l'erreur de prédiction.
Á la sortie du prédicteur, le signal d'erreur du prédiction s'écrit:

$$\tilde{\mathbf{x}}(k) \approx \mathbf{h}_0\tilde{s}(k) \tag{8.12}$$

Ainsi $\mathbf{h_0}$ est colinéaire au vecteur propre maximal de la matrice de corrélation de l'erreur de prédiction. Un égaliseur á forçage-á-zéro peut être définit comme:

$$\mathbf{F_{D\&P}}(q) = \mathbf{h}_0^H A_{L_A}(q) \tag{8.13}$$

iii) Déréverberation du signal reçu

Finalement, l'égaliseur $F_{D\&P}$ est utilisé pour déréverbérer le signal de la parole:

$$\widehat{s}(k) = \mathbf{F^T_{D\&P}}(q)\mathbf{y}(k) = \mathbf{h}_0^T A_{L_A}(q)\mathbf{y}(k) \tag{8.14}$$

Les résultats de simulation montrent que l'égaliseur proposé performe beaucoup mieux de que schéma classique " Delay-and-Sum ", particulièrement dans le cas ou seulement un nombre limité de microphones est disponible (voir figure 8.8).

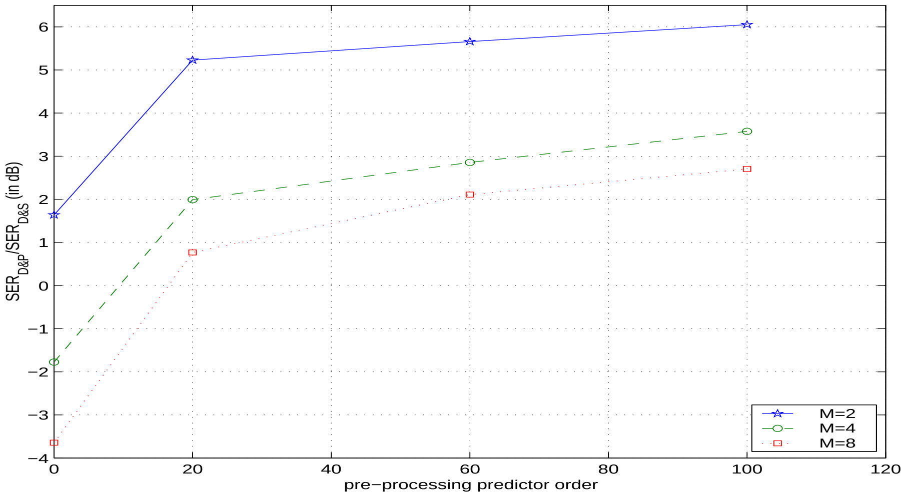

Figure 8.8: Le gain $G = \dfrac{SER_{D\&P}}{SER_{D\&S}}$ pour 2, 4, and 8 microphones.

8.3.3 Déréverbération aveugle de la parole en présence de bruit additif

Généralement, les schémas de déréverbération sont conçus en absence de bruit ambiant. Cependant pour une utilisation pratique la robustesse de ces algorithmes au bruit additive est indispensable. Dans le présent travail, nous proposons deux méthodes pour améliorer la robustesse de notre schéma de déréverbération: l'alignement des signaux de prédiction, et l'optimisation de l'égaliseur á forçage-á-zéro (via l'optimisation d'un post-filtrage).

i) Alignement des signaux de prédiction

Plusieurs auteurs soulignent le manque de robustesse de l'égaliseur basé sur la prédiction linéaire en présence de bruit additif. En particulier, la performance globale d'algorithme est fonction d'une réalisation particulière du coefficient $\mathbf{h}_0$, générant un signal d'erreur de prédiction avec un rapport signal-sur-bruit incontrôlable.

Nous suggérons atténuer cet effet en alignant les signaux reçus sur les divers microphones (compensation des délais des chemins direct). Nous démontrons que cette procédure mène non seulement à une augmentation de l'énergie

utile du signal $\left(\sigma_s^2 \|\mathbf{h}_0\|^2\right)$, mais également à la réduction de l'erreur quadratique moyenne á la sortie du prédicteur MSE $= \left(\sigma_v^2 \text{ trace}\left\{A_{L_A} A_{L_A}^T\right\}\right)$ (voir annexe 4.A).
La figure 8.9 compare les performances de la prédiction linéaire avec et sans compensation de délais (avec simulation Monte Carlo avec 100 exécutions). On remarque bien que l'alignement des signaux reçus augmente la robustesse de l'algorithme au bruit additif; et que ceci est d'autant plus crucial que le nombre de sous-canaux augmente.

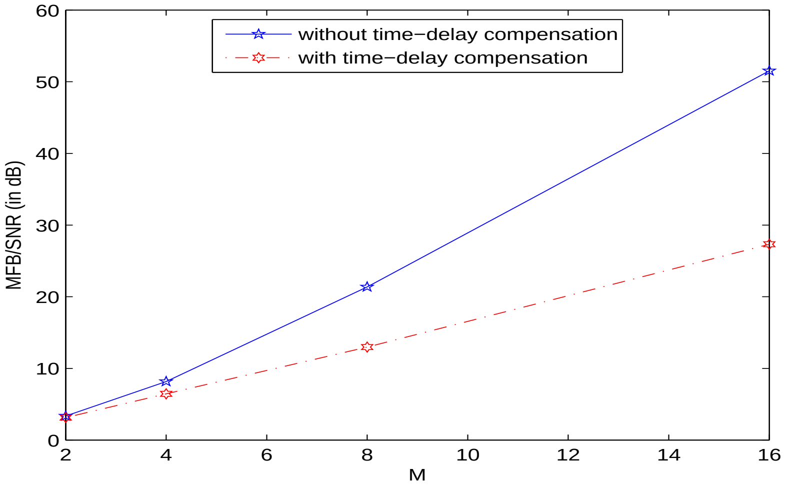

Figure 8.9: $\dfrac{MFB}{SNR_{out}}$ avec et sans compensation de délais.

ii) Optimisation de l'égaliseur á forçage-á-zéro

En présence de bruit additive $\mathbf{v}(k)$, la sortie du prédicteur multicanal est:

$$\mathbf{x}(k) = \mathbf{h_0} s(k) + A(q)\mathbf{v}(k) \tag{8.15}$$

Dans le schéma d'égalisation classique, les colonnes du prédicteur $A(q) = I + \sum_{i=1} A_i q^{-i}$ sont combiné par le vecteur pondérant $\mathbf{h}_0^H$, i.e.,

$$F_{LP}(q) = \mathbf{h}_0^H A(q) \tag{8.16}$$

Ce choix maximise la puissance du signal utile, mais pas nécessairement le rapport signal sur bruit de la sortie. Dans [55], Gazzah fixe le vecteur pondérant en maximisant le rapport signal sur bruit de la sortie, i.e.

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\sigma_s^2}{\sigma_v^2} \frac{\|\mathbf{w}\|^2}{\mathbf{w}AA^H\mathbf{w}^H} \tag{8.17}$$

Nous généralisons l'approche précédente en considérant les filtres pondérant (au lieu de simples facteurs scalaires). Le fait qui permet de définir des égaliseurs à forçage-à-zéro avec un délai d'égalisation non-nul. L'optimisation des filtres pondérant correspond à la conception d'un MMSE-ZF. L'ordre non-nul de ces filtres permet l'introduction d'un délai d'égalisation, qui peut être également optimisé. Les simulations prouvent que des gains considérables peuvent être réalisés (même pour un petit délai d'égalisation) (voir figure 8.10).

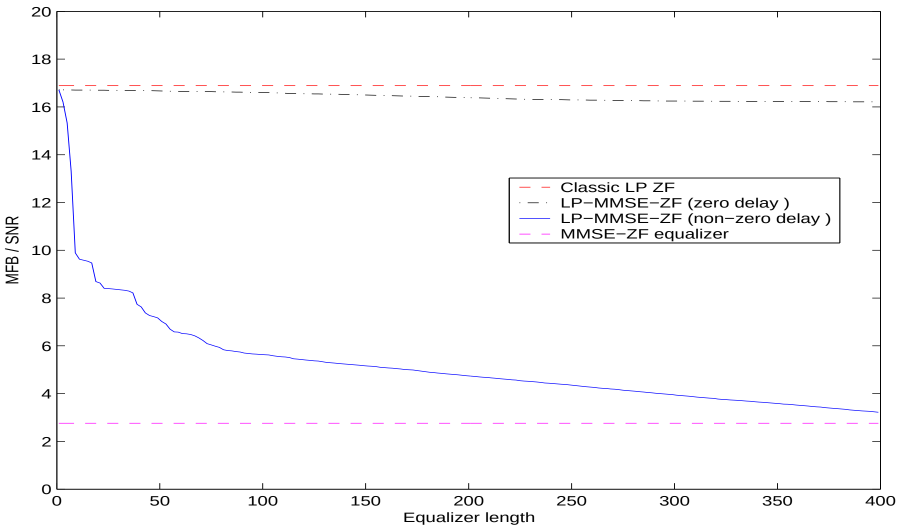

Figure 8.10: $\dfrac{MFB}{SNR_{out}}$ pour différents égaliseurs ZF.

Ensuite, nous illustrons le comportement du schéma proposé appliqué à la dereverberation de la parole, et nous comparons ses performance a celui du filtre spatiale Delay-&-Sum. Nous considérons le scénario de dereverberation précédant. Un signal de parole de durée 8.8 s, et échantillonné a 8 kHz est utilisé comme signal d'entrée. Le signal réverbéré est capté par deux microphones distant de $d = 0.2m$. L'ordre du post-filtrage est contraint á $L_w \leq 100$. La figure 8.11 trace le Rapport Signal sur Echo+Bruit ($SENR = \frac{\sum_k s(k)^2}{\sum_k (s(k) - \widehat{s}(k))^2}$) en fonction du Rapport Signal sur Bruit ($SNR = \frac{\sum_k (y(k) - v(k))^2}{\sum_k v(k)^2}$). Les courbes montrent que les performances du D-&-P sont meilleures que celles du D-&-S.

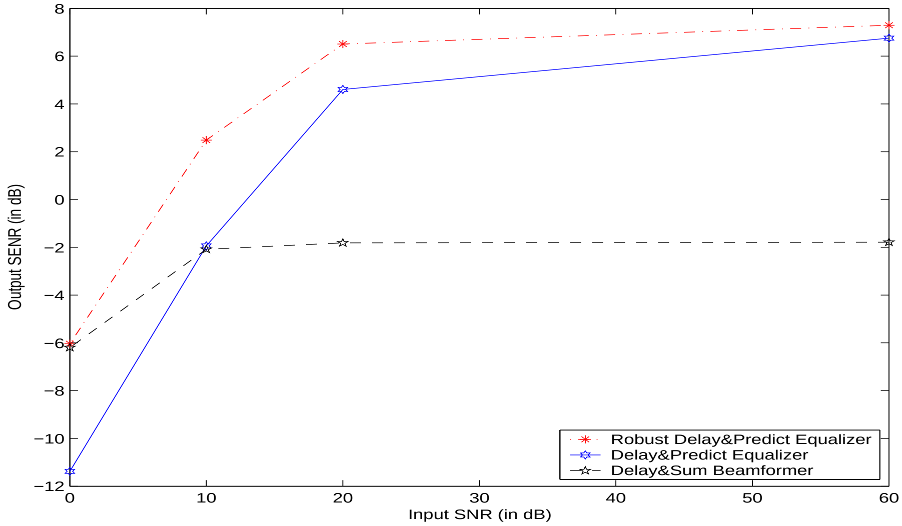

Figure 8.11: Rapport Signal sur Echo+Bruit (SENR) de sortie en fonction du Rapport Signal sur Bruit d'entrée (SNR).

Particulièrement, dans les régions bruitées, le post-filtrage devient indispensable. D'autre part, on peut aussi remarquer que ce filtrage a encore un effet positif même en absence de bruit (SNR=60 dB). Dans ce cas, il compense l'erreur d'estimation du spectre du signal d'origine.

Le bruit ambiant n'est pas l'unique source de bruit additif. En effet, la réverbération acoustique est théoriquement infinie. Et puisque nous imposons que la longueur de la réponse impultionnelle est finie (L_h), une partie de la réverbération (réverbération tardive) est considérée comme bruit additive, i.e.,

$$\mathbf{y}(k) = \sum_{i=0}^{L_h-1} \mathbf{h}_i s(k-i) + \underbrace{\sum_{i=L_h}^{\infty} \mathbf{h}_i s(k-i)}_{\mathbf{v}(k)} . \tag{8.18}$$

Classiquement, la réponse impultionnelle est choisie suffisamment longue pour que l'énergie de cette composante additive soit négligeable (typiquement $L_h \leq T_{60} f_s$, ou T_{60} est le temps de réverbération et f_s est la fréquence d'échantillonnage). Avec ce choix, le canal de propagation peut être excessivement long dans des conditions de propagation réelles. Ainsi, la complexité algorithmique de la déréverbération peut devenir très grande pour être pratique.

Dans cette section, nous étudions l'effet de la sous-estimation de la longueur de la réponse impultionnelle sur la performance de dereverberation. Nous modélisons la réverbération tardive comme un bruit diffus sphérique [101] (bien que, strictement dit, ce bruit additive est ni blanc ni indépendant de la source d'intérêt). Ensuit, nous appliquons le filtrage a posteriori (proposé dans la section précédente). Nous considérons le rapport d'énergie directe sur réverbération (DRR) comme critère d'évaluation:

$$DRR = 10 \log_{10} \left\{ \frac{\sum_{t=0}^{\tau-1} \widetilde{h}^2(t)}{\sum_{t=\tau}^{L-1} \widetilde{h}^2(t)} \right\} \quad dB \tag{8.19}$$

ou $\widetilde{h}^2(t) = \mathbf{h} * \mathbf{f}(t) = \sum_i \mathbf{h}_i \mathbf{f}_{t-i}$ est le canal égalisé (avec l'égaliseur $\mathbf{f}(q)$), et τ est le nombre d'échantillons considérés comme appartenant au chemin direct. Les figures 8.12 et 8.13 tracent les courbes du DRR du sortie des égaliseurs Delay-&-Predict classique et robuste, ainsi que du filtre spatial Delay-&-Sum, en utilisant respectivement 2 et 4 microphones (pour $\tau = 10\, ms$ et $\tau = 1\, ms$).

Dans ces simulation, la longueur du canal est finit ($L_h = 2000$). On peut remarquer que les performances du schéma robuste excèdent celles du schéma classique; et que les deux performent beaucoup mieux que le filtrage spatial classique. On note aussi que même quand la longueur de la réponse impultionnelle est surestimé, le filtrage a posteriori a encore un effet positif, spécialement quand uniquement deux microphones sont disponibles. Comme déjà mentionné, ceci est dû au fait que le post filtrage compense les erreurs d'estimation du spectre du signal d'origine. Ces erreurs sont plus importantes quand le nombre de microphones diminue.

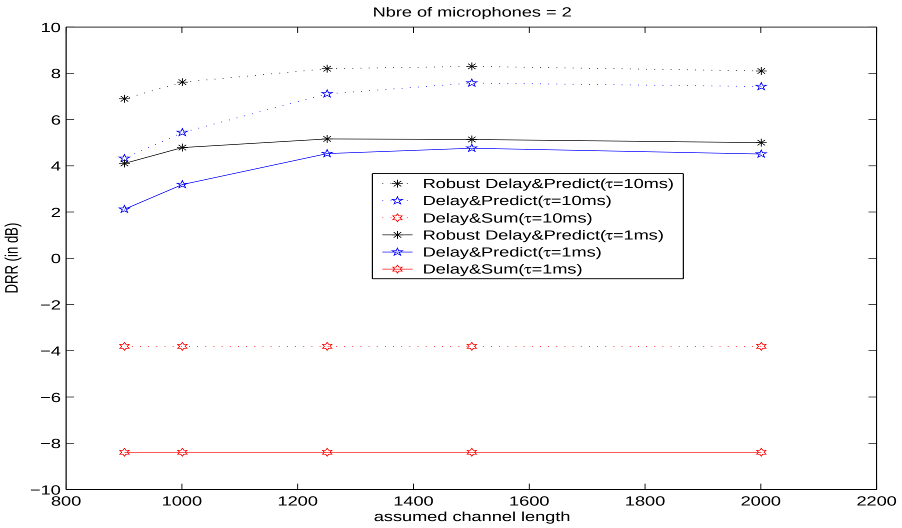

Figure 8.12: Le DRR de sortie fonction du présumée longueur du canal, configuration en 2 microphones ($\tau = 10\ ms$ et $\tau = 1\ ms$)).

8.4 Estimation Bayesienne Conditionnellement Non-Biaisé par Morceau

Généralement, les schémas d'estimation font l'objet d'un compromis entre le biais et la variance. Le biais est due à "l'écart" entre la valeur moyenne de l'estimateur et la véritable valeur du paramètre (biais conditionnel); alors que la variance est du aux fluctuations due à l'échantillonnage statistique.

Si des informations statistiques aprioris sont disponibles, la théorie de l'estimation Bayesienne montre que, sous la contrainte de non-biais Bayesien, l'erreur quadratique moyenne (MSE) est bornée par la Borne Cramer Rao Bayesienne (B-CRB). En outre, l'estimateur MMSE minimise $\mathbf{R}_{\widetilde{\theta}\widetilde{\theta}}$, la matrice de corrélation du signal d'erreur, et pas seulement l'erreur quadratique moyenne (qui correspond la trace de $\mathbf{R}_{\widetilde{\theta}\widetilde{\theta}}$). Néanmoins, le non-biais Bayesien d'un paramètre aléatoire correspond à un biais nul en moyenne, ce qui engendre une contrainte très faible. En particulier, l'estimateur MMSE est non-biaisé (au sens Bayesien), et l'estimateur MMSE minimise $\mathbf{R}_{\widehat{\theta\theta}}$ et le MSE, indépendamment du fait d'imposer ou non la contrainte du non-biais Bayesien. Ainsi, les estimations bayesiennes conduisent à une estimation (conditionnellement) biaisée. Ce biais est nuisible dans certain nombre d'applications:

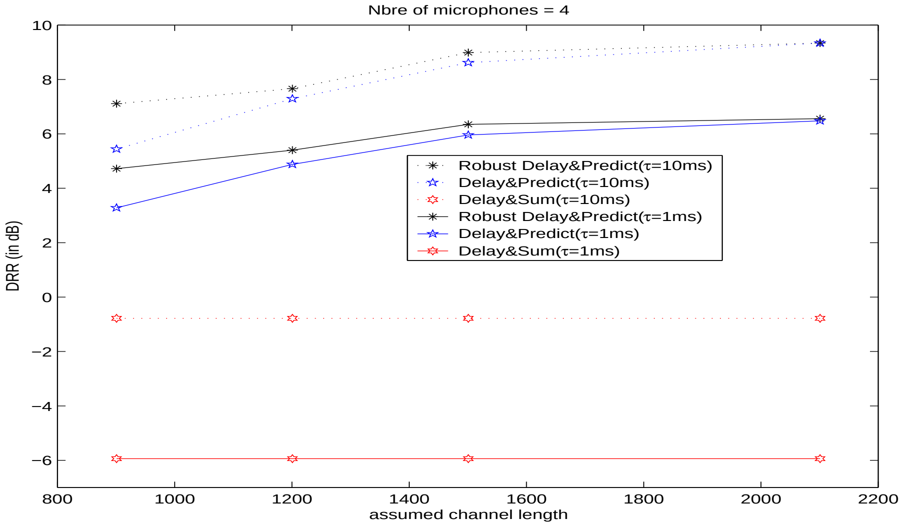

Figure 8.13: Le DRR de sortie fonction du présumée longueur du canal, configuration en 4 microphones ($\tau = 10\ ms$ et $\tau = 1\ ms$)).

détection multiutilisateurs (par le schéma de Viterbi ou LMMSE), estimation paramétrique des réponses impulsionnelles...

D'autre part, exiger que tous les composants du paramètre soient conjointement (conditionnellement) non-biaisé (ce qui correspond à un forçage a zéro, dans le cas d'une détection multiutilisateurs) empêche l'exploitation de l'information statistique apriori. Par conséquence, ça conduit à une réduction significative de l'erreur quadratique moyenne de l'estimation.
Ceci constitue une motivation pour introduire l'estimation Bayesienne conditionnellement non-biaisé par morceau (CWCU). Plutôt que de contraindre l'estimateur à être conjointement non-biais, nous imposons la contrainte du biais par composante. De cette manière, chaque paramètre est traité comme déterministe tandis que les autres paramètres sont traités comme Bayésiens. Dans le cas d'une détection multiutilisateurs, imposer le non-biais par morceau correspond à une détection Bayesienne non-baisé, tandis que le non-biais conjoint correspond à un forçage-á-zéro.
Une introduction plus générale du concept est motivée par l'estimation LMMSE des canaux de transmission, pour laquelle les implications du concept sont illustrées dans diverses manières. L'application à la localisation des mobiles est étudiée en détails.

8.4.1 Estimation Bayesienne Conditionnellement Non-Biaisé par Morceau pour un Modèle Linéaire Gaussien

Nous considérons un modèle linéaire Gaussien:

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{v} \tag{8.20}$$

ou $\mathbf{y}$ représente le signal reçu, $\boldsymbol{\theta} \sim N\left(0, \mathbf{C}_{\theta\theta}\right)$ symbolise les paramètres a estimer, et $\mathbf{v}$ représente un bruit blanc additif Gaussien.

Remarquons que tout estimateur linéaire est non biaisé en moyenne, i.e.,

$$E_{Y,\theta}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = FE_{Y,\theta}\left(\boldsymbol{\theta}\right) = \mathbf{0} \tag{8.21}$$

Ainsi, sous contrainte de non-biais (Bayesien), minimiser l'erreur quadratique moyenne détermine l'estimateur LMMSE:

$$\begin{aligned}\widehat{\theta}_{lmmse} &= \arg\min_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\ &= \arg\min_{\widehat{\theta}=Fy} \operatorname{tr}\left\{\left(\mathbf{FH} - \mathbf{I}_K\right)\mathbf{C}_{\theta\theta}\left(\mathbf{FH} - \mathbf{I}_K\right)^H\right\} + \sigma_v^2 \operatorname{tr}\left\{\mathbf{FF}^H\right\} \\ &= \mathbf{C}_{\theta\theta}\mathbf{H}^H\left(\mathbf{HC}_{\theta\theta}\mathbf{H}^H + \sigma_v^2\mathbf{I}_K\right)^{-1}\mathbf{y} \end{aligned} \tag{8.22}$$

D'autre part, sous contrainte de non-biais conditionnel conjoint, la minimisation de l'erreur quadratique moyenne donne:

$$\widehat{\boldsymbol{\theta}} = \left\{\begin{array}{l}\arg\min_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\ E_{Y|\theta}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = 0\end{array}\right. = \left\{\begin{array}{l}\arg\min_{F} \operatorname{tr}\left\{\mathbf{FF}^H\right\} \\ \mathbf{FH} = \mathbf{I}_K\end{array}\right.$$

Ainsi, imposer le non-biais conjoint empêche l'exploitation des corrélations aprioris entre les paramètres, et conduit à une réduction significative du bénéfice dû a l'apriori Bayesien. Dans ce cas, l'estimateur MMSE correspond à l'estimateur BLUE, i.e.,

$$\widehat{\boldsymbol{\theta}}_{blue} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{y} \tag{8.23}$$

Les estimateurs LMMSE et BLUE sont lié par (voir l'annexe 5.A)

$$\widehat{\boldsymbol{\theta}}_{lmmse} = \underbrace{\mathbf{C}_{\theta\theta}\left(\left(\mathbf{H}^H\mathbf{H}\right)\mathbf{C}_{\theta\theta} + \sigma_v^2\mathbf{I}_K\right)^{-1}\left(\mathbf{H}^H\mathbf{H}\right)}_{\mathbf{B}_{lmmse}}\widehat{\boldsymbol{\theta}}_{blue} \tag{8.24}$$

ou $\mathbf{B}_{lmmse} = \mathbf{C}_{\theta\theta}\left(\left(\mathbf{H}^H\mathbf{H}\right)\mathbf{C}_{\theta\theta} + \sigma_v^2\mathbf{I}_K\right)^{-1}\left(\mathbf{H}^H\mathbf{H}\right)$ représente le bias conditionnel de l'estimateur LMMSE.
Imposer la contrainte du non-bias conditionnel par morceau donne lieu au problème d'optimisation:

$$\widehat{\boldsymbol{\theta}}_{cwculmmse} = \begin{cases} \arg\min\limits_{\widehat{\theta}=Fy} E\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\ E_{\mathbf{y}|\theta_k}\left(\widehat{\theta}_k - \theta_k\right) = 0 \quad k = 1:K \end{cases}$$

Puisque $\boldsymbol{\theta}$ est supposé Gaussien,

$$\begin{aligned} E_{\mathbf{y}|\theta_k}\left\{\widehat{\theta}_k\right\} &= \mathbf{e}_k^H\ \mathbf{FH}\ \ E_{\mathbf{y}|\theta_k}\left\{\boldsymbol{\theta}\right\} \\ &= \left(\mathbf{e}_k^H\mathbf{FHC}_{\theta\theta}\mathbf{e}_k\right)\left(\mathbf{e}_k^H\mathbf{C}_{\theta\theta}\mathbf{e}_k\right)^{-1}\theta_k \end{aligned}$$

Par conséquence, l'estimateur CWCU-LMMSE est calculé en optimisant:

$$\begin{cases} \arg\min\limits_{\widehat{\theta}=Fy} \operatorname{tr}\left\{\left(\mathbf{FH} - \mathbf{I}_K\right)\mathbf{C}_{\theta\theta}\left(\mathbf{FH} - \mathbf{I}_K\right)^H\right\}\ +\ \sigma_v^2\operatorname{tr}\left\{\mathbf{FF}^H\right\} \\ \mathbf{e}_k^H\ \mathbf{FHC}_{\theta\theta}\ \mathbf{e}_k = \mathbf{e}_k^H\mathbf{C}_{\theta\theta}\mathbf{e}_k \qquad k = 1:K \end{cases}$$

En utilisant l'optimisation de Lagrange, on peut monter que le CWCU-LMMSE est donné par (voir l'annexe 5.B):

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{cwculmmse} &= \mathbf{D}_{cw}\widehat{\boldsymbol{\theta}}_{lmmse} \\ &= \mathbf{D}_{cw}\mathbf{B}_{lmmse}\widehat{\boldsymbol{\theta}}_{blue} \end{aligned} \tag{8.25}$$

ou $\mathbf{D}_{cw} = \left(\operatorname{diag}\left(\mathbf{C}_{\theta\theta}\right)\right)\left(\operatorname{diag}\left(\mathbf{B}_{lmmse}\mathbf{C}_{\theta\theta}\right)\right)^{-1}$ est une matrice diagonale garantissant le non-biais par morceau.

Cas particuliers:

- Si les composantes θ_k du paramètre $\boldsymbol{\theta}$ sont décorrélés ($\mathbf{C}_{\theta\theta}$ est diagonale), alors $\mathbf{D}_{cw}$ se simplifie:

  $$\mathbf{D}_{cw} = \left(\operatorname{diag}\left(\mathbf{B}_{lmmse}\right)\right)^{-1} \tag{8.26}$$

  Ainsi, le CWCU-LMMSE correspond a l'estimateur MMSE non-biaisé (U-MMSE).

- Si les composantes θ_k du paramètre $\boldsymbol{\theta}$ sont non-couplés ni a travers l'apriori Bayesian ni a travers les données ($\mathbf{B}_{lmmse}$ est diagonale), le CWCU-LMMSE correspond a l'estimateur BLUE. Ainsi, l'estimation CWCU-LMMSE définit un nouveau compromis biais-variance dans le cas ou les paramètre sont couplés a traves a travers l'apriori Bayesien ($\mathbf{C}_{\theta\theta}$ est non-diagonale) ou/et a travers les données ($\left(\mathbf{H}^H\mathbf{H}\right)$ est non-diagonale).

Remarquons que dans le cas d'une détection multiutilisateurs ($\mathbf{C}_{\theta\theta}$ est diagonale), le CWCU-LMMSE correspond à une détection Bayesienne non-baisé (ULMMSE), tandis que le BLUE correspond à un forçage-á-zéro (MMSE-ZF).

Dans cette thèse nous avons introduit le concept générale de l'estimation Bayesienne conditionnellement non-biaisé par morceau. Une attention particulaire a était accordé a l'estimation LMMSE des canaux de transmission, pour laquelle les implications du concept sont illustrées dans diverses manières. Particulièrement, nous avons

- Introduit les filtrages de Wiener et Kalman sous les contraintes CWCU.
- Etudié la relation entre le biais conjoint et le rang de la matrice de correlation apriori. Nous avons montré que si $\mathbf{C}_{\theta\theta}$ est de rang m, imposé la contrainte du biais conditionnelle par block (de taille $\geq m$), garantie le non-bias conjoint.
- Analysé l'estimation CWCU-LMMSE pour les canaux de transmission. Nous avons proposé deux schémas itératifs pour une implémentation efficace du schéma CWCU-LMMSE.

Par ailleurs, l'application du concept CWCU à la localisation des mobiles a était étudiée en détails, spécifiquement l'implication a la localisation des mobiles par signature. Nous avons proposé:

- Deux approches paramétriques (déterministe et Bayesienne) pour l'estimation du profil de puissance du canal de transmission.
- Deux schémas de localisation par signature (directe et indirecte).
- L'exploitation de l'information spatiale disponible lors d'une transmission multi-entrées ou/et multi-sorties.

- Une méthode de validation reproductible se basant sur une methode simple de “lancé de rayons”.

Bibliography

[1] E. Aboutanios,“A modified Dichotomous Search Frequency Estimator,” *Signal Processing Letters*, Vol.11, Issue 2, pp.186-188, Feb. 2004.

[2] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, “Time Domain Blind Source Separation of Non-Stationary Convolved Signals by Utilizing Geometric Beamforming,” *In Proceedings of IEEE Int. Conf. on Neural Networks for Signal Processing (NNSP)*, Sept. 2002.

[3] A. Aïssa-El-Bey, H. Bousbia-Salah, K. Abed-Meraim, and Y. Grenier, “Audio source separation using sparsity,” *In Proc. of the Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006.

[4] A. Aïssa-El-Bey, K. Abed-Meraim and Y. Grenier, “Underdetermined Blind Audio Source Separation Using Modal DecompositionŠ,” *EURASIP Journal on Audio, Speech & Music Processing*, Mar. 2007.

[5] S. Ahonen and P. Eskelinen, “Mobile Terminal Location for UMTS,” *IEEE Aerospace and Electronic Systems Magazine*, Vol.18, Issue 2, pp. 23-27, Feb. 2003.

[6] S. Ahonen and P. Eskelinen, “Performance Estimations of Mobile Terminal Location with Database Correlation in UMTS Networks,” *In Proc. of Int. Conf. on 3G Mobile Communication Technologies*, pp. 25-27, June 2003.

[7] J.B. Allen and D.A. Berkley, “Image Method for efficiently Simulating Small Room Acoustics,” *J. Acoust. Soc. Amer.*, Vol. 65, pp. 943Ű950, 1979.

[8] A. Amar and A.J. Weiss, “Direct Position Determination of Multiple Radio Signals,” *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.2, pp. 81-84, May 2004.

[9] A. Amar and A.J. Weiss, “Advances in Direct Position Determination,” *In Proc. of IEEE Work. on Sensor Array and Multichannel Signal Processing*, pp. 584-588, July 2004.

[10] B. Anderson and J. Moore, "Optimal Filtering," *Prentice Hall*, 1979.

[11] A.N. D’Andrea, U. Mengali, and R. Reggiannini, "The modified Cramèr-Rao bound and its application to synchronization problems," *IEEE Trans. on Communications*, Vol.42, pp.1391-1399, Feb./Mar./Apr. 1994.

[12] S. Araki, H. Sawada, R. Mukai, and S. Makino, “DOA estimation for multiple sparse sourceswith normalized observation vector clustering” *In Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp.969-972, May 2006.

[13] S. Araki, H. Sawada, R. Mukai, and S. Makino, “ Blind sparse source separation with spatially smoothed time-frequency masking,” *in Proc of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006.

[14] F. Abrard, Y. Deville, and P. R. White,“From Blind Source Separation to Blind Source Cancellation in the Underdetermined Case: a new Approach Based on Time-Frequency Analysis,” *In Proc. of Int. Conf. on Independent Component Analysis (ICA)*, San Diego, California, December 2001.

[15] E. Bacharakis, ”Separation of Maternal and Foetal ECG Using Blind Source Separation Methods,” *In Proc. of Europ. Signal Processing Conf. (EUSIPCO)*, pages 395-398, Trieste, Italy, September 1996.

[16] R. Badeau, B. David, and G. Richard, “High-Resolution Spectral Analysis of Mixtures of Complex Exponentials Modulated by Polynomials,” *IEEE Trans. on Signal Processing*, Vol. 54, No.4, pp.1341-1350, Apr.2006.

[17] R. Balan, , A. Jourjine, and J. Rosca, “ AR processes and sources can be reconstructed from degenerate mixtures",” *In Proc. of Int. Conf. on Independent Component Analysis (ICA)*, pp. 467-472, Jan. 2000.

[18] R. Balan and J. Rosca, “ Source Separation Using Sparse Discrete Prior Models,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp.1113-1116, May 2006.

[19] A.J. Barabell, “Improving the Resolution Performance of Eigenstructure-Based Direction-Finding Algorithms,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.8, pp.336-339, Apr.1983.

[20] D. Bees, M. Blostein, and P. Kabal, “Reverberant Speech Enhancement Using Cepstral Processing,” *In Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 977-980, Apr. 1991.

[21] A. Belouchrani, K. Abed-Meriam, J.F. Cardoso, and E. Moulines, ”A Blind Source Separation Technique Using Second-Order Statistics,” *IEEE Trans. On Signal Processing*, Vol.45, No.2, February 1997.

[22] L. Benaroya, F. Bimbot, G. Gravier, and R. Gribonval, “Audio source separation with one sensor for robust speech recognition,” *In Proc. of Int. Symposium on Computer Architecture*, May 2003.

[23] P.M.T. Broersen, “Finite Sample Criteria for Autoregressive Order Selection,” *IEEE Trans. on Signal Processing*, Vol.48, No.12, pp.3550-3558, Mec. 2000.

[24] H. Buchner, R. Aichner, and W. Kellermann, “Blind Source Separation for Convolutive Mixtures Exploiting Nongaussianity, Nonwhiteness, and Nonstationarity,” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, pp. 275-278, Sep. 2003 .

[25] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON: A Versatile Framework for Multichannel Blind Signal Processing,” *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp.889-892, May 2004.

[26] R. Aichner, H. Buchner, and W. Kellermann, “On the Causality Problem in Time-Domain Blind Source Separation and Deconvolution Algorithms ,” *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp.181-184, Mar. 2005.

[27] W. Kellermann, H. Buchner, and R. Aichner, “Separating Convolutive Mixtures with TRINICON ,” *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp.961-964, May 2006.

[28] J.J. Caffery and G.L. Stuber, “Overview of Radiolocation in CDMA Cellular Systems,” *IEEE Communications Magazine*, Vol.36, Issue 4, pp. 38-45, Apr. 1998.

[29] X.-R. Cao and R.-W. Liu, ”General Approach to Blind Source Separation,” *IEEE Transactions On Signal Processing*, Vol.44, No.3, March 1996.

[30] J.F. Cardoso, ”Blind Signal Separation: Statistical Principles,” *In Proc. of the IEEE*, Vol.86, No.10, October 1998.

[31] N. Castaneda, M. Charbit, and E. Moulines, “Source Localization from Quantized Time of Arrival Measurements,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp. 933-936, May 2006.

[32] J. Cermak, S. Araki, H. Sawada, and S. Makino, ”Blind speech separation by combining beamformers and a time frequency binary mask,” *in Proc of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006.

[33] S. Chen and D.L. Donoho, “Atomic decomposition by basis pursuit,” *Technical report*, Statistics Department,Stanford University, 1995.

[34] A. Cichocki and S. Amari ,“Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications,” *John Wiley & Sons*, 2002.

[35] P. Comon, “Independent Component Analysis, A New Concept?,” *Signal Processing*, Vol.36, No.3, pages 287-314, 1994 .

[36] A.E. Conway, “Output-based method of applying PESQ to measure the perceptual quality of framed speech signals,” *In Proc. of IEEE Wireless Communications and Networking Conf. (WCNC)*, Vol. 4, pp. 21-25, March 2004.

[37] S.F. Cotter and B.D. Rao, “Matching Pursuit Based Decision-Feedback Equalizers,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp. 2713-2716, Jun. 2000.

[38] S.F. Cotter and B.D. Rao, "Sparse Channel Estimation via Matching Pursuit with Application to Equalization," *IEEE Trans. on Communications*, Vol.50, Issues:3, pp. 374-377, March 2002.

[39] M.J. Daly and J.R. Reilly, “Blind Deconvolution Using Bayesian Methods with Application to the Dereverberation of Speech,” *In Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.2, pp.1009-1012, Apr.2004.

[40] N. Mitianoudis and M. Davis,“A Fixed Point Solution for Convolved Audio Source Separation,” *In Proc. of IEEE Work. on Applications of Signal Processing on Audio and Acoustics (WASPAA)*, Oct. 2001.

[41] L. Daudet, “Sparse and structured decompositions of audio signals in overcomplete spaces,” *In Proc. of Int. Conf. on Digital Audio Effects (DAFX)*, Oct. 2004.

[42] Yinong Ding and Xiaoshu Qian, “Estimating Sinusoidal Parameters of Musical Tones Based on Global Waveform Fitting,” *In Proc. of IEEE Work. on Multimedia Signal Processing (MMSP)*, pp.95-100, June 1997.

[43] Z. Ding, “Linear Predictive Algorithms for Blind Multichannel Identification,” *In Signal Processing Advances in Wireless and Mobile Communications, Vol. I: Trends in Channel Estimation and Equalization*, In G.B. Giannakis, Y. Hua, P. Stoica, and L. Tong (Editors), Prentice Hall, 2001.

[44] Y.C. Eldar, "Minimum Variance in Biased Estimation: Bounds and Asymptotically Optimal Estimators," *IEEE Trans. on Signal Processing*, Vol.52, pp.1915-1930, July 2004.

[45] Y.C. Eldar, "MSE Bounds Dominating the Cramèr-Rao Bound," *In Proc. of IEEE Statistical Signal Processing (SSP)*, July 2005.

[46] Y. Ephraim, “Statistical model based speech enhancement systems,” *In Proc. of the IEEE*, Vol. 80, No. 10, pp. 1526-1555, Oct. 1992.

[47] TSGR1#4(99)346. “Enhancing speech degraded by additive noise or interfering speakers,” *Ericsson*, Shin-Yokohama, Japan, TSG-RAN Working Group 1 meeting #4,1999.

[48] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” *J. Acoust. Soc. Amer.*, pp. 1508-1518, Nov. 1985.

[49] S. Fischer, H. Grubeck A. Kangas,H. Koorapaty, E. Larsson, and P. Lundqvist, “Time of arrival estimation of narrowband TDMA signals for mobile positioning,” *In Proc. of IEEE Int. Symp. on Personal Indoor and Mobile Radio Communications (PIMRC)*, Vol.1, pp.451-455, Sept. 1998.

[50] N. Gaubitch, P.A. Naylor, and D.B. Ward, “On the Use of Linear Prediction for Dereverberation of Speech,” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, pp. 99-102, Sept. 2003.

[51] N.D. Gaubitch and P.A. Naylor, “Analysis of the Dereverberation Performance of Microphone Arrays.” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2005.

[52] N. Gaubitch, P.A. Naylor, and D.B. Ward, “Multi-microphone speech dereverberation via spatio-temporal averaging,” *In Proc. of Europ. Signal Processing Conf. (EUSIPCO)*, pp.809-812, Sept. 2004.

[53] K. Shahtalebi and S. Gazor, "Adaptive Linear Estimators Using Biased Cramèr-Rao Bound," *In proc. of IEEE Statistical Signal Processing (SSP)*, July 2005.

[54] W. Gao, S. Tsai, and J. S. Lehnert, “Diversity combining for DS/SS systems with time-varying, correlated multipath fading channels,” *IEEE Trans. on Communications*, Vol. 51, No. 2, Pages: 284-295, February 2003.

[55] H. Gazzah, “Optimum Blind Multichannel Equalization Using the Linear Prediction Algorithm,” *IEEE Trans. on Signal Processing*, Vol. 54, pp.3242-3247, Aug. 2006.

[56] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, “Speech Dereverberation via Maximum-Kurtosis Subband Adaptive Filtering,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 6, pp. 3701-3704, May 2001.

[57] B.W. Gillespie and L.E. Atlas, "Acoustic Diversity for Improved Speech Recognition in Reverberant Environments," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*,Vol. 1, pp. 557-560, May 2002.

[58] G.B. Giannakis and S.D. Halford, "Asymptotically Optimal Blind Fractionally Spaced Channel Estimation and Performance Analysis," *IEEE Trans. on Signal Processing*, Vol.45, Issue 7, pp.1815-1830, July 1997.

[59] D. Gesbert and P. Duhamel, "Robust Blind Channel Identification and Equalization Based on Multi-Step Predictors," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 5, pp.3621-3624, Apr. 1997.

[60] S.J. Godsill and C. Andrieu, "Bayesian Separation And Recovery Of Convolutively Mixed Autoregressive Sources," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp.1733-1736, Mar. 1999.

[61] G.H. Golub and C.F. Van Loan "Matrix Computations," *John Hopkins University Press*, 1996.

[62] M. Goodwin and M. Vetterli, "Time-Trequency Signal Models for Music Analysis, Transformation, and Synthesis," *In Proc. of IEEE Int. Symposium on Time-Frequency and Time-Scale Analysis (TFTS)*, pp.133-136, June 1996.

[63] M. Goodwin and M. Vetterli, "Atomic Decompositions of Audio Signals," *In Proc. of IEEE Work. on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1997.

[64] M. Goodwin, "Matching Pursuit With Damped Sinusoids," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp.2037-2040, Apr. 1997.

[65] A. Aïssa-El-Bey, K. Abed-Meraim, and Y. Grenier, "Blind Separation of Audio Sources Convolutive Mixtures Using Parametric Decomposition," *In proc of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2005.

[66] S. Griebel and M. Brandstein, "Microphone Array Speech Dereverberation Using Coarse Channel Estimation," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.1, pp.201-204, May 2001.

[67] R. Gribonval, "Fast Matching Pursuit with a Multiscale Dictionary of Gaussian Chirps," *IEEE Trans. Signal Processing*, Vol.49, No.5, May 2001.

[68] R. Gribonval and P. Vandergheynst, "On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries," *Technical report*, IRISA, April 2004.

[69] R. Gribonval and E. Bacry, "Harmonic Decomposition of Audio Signals with Matching Pursuit," *IEEE Trans. on Signal Processing*, Vol.51, No.1, Jan. 2003.

[70] R. Gribonval, "Piecewise Linear Source Separation," *In Proc. of SPIE*, Vol.5207, pp. 297-310, Aug. 2003.

[71] R. Gribonval, Ph. Depalle, X. Rodet, E. Bacry, and S. Mallat, "Sound Signals Decomposition Using a High Resolution Matching Pursuit," *In Proc. of Int. Computer Music Conf. (ICMC)*, Aug. 1996.

[72] R. Gribonval, E. Bacry, S. Mallat, Ph. Depalle, and X. Rodet, "Analysis of Sound Signal with High Resolution Matching Pursuit," *In Proc. of IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis (TFTS)* , June 1996.

[73] Y. Haneda, S. Makino, and Y. Kaneda, "Multiple-Point Equalization of Room Transfer Functions by Using Common Acoustical Poles," *IEEE Trans. on Speech, and Audio Processing*, Vol. 5, Issue 4, pp. 325-333, July 1997.

[74] J.H.L. Hansen and L.M. Arslan, "Markov model-based phoneme class partitioning for improved constrained iterative speech enhancement," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 3, Issue 1, pp. 98-104, Jan. 1995.

[75] A.O. Hero, J.A. Fessler, and M. Usman, "Exploring Estimator Bias-Variance Tradeoffs Using the Uniform CR Bound," *IEEE Trans. on Signal Processing*, Vol.44, pp.2026-2041, August 1996.

[76] T. Hikichi, M. Delcroix, and M. Miyoshi, “Blind dereverberation based on estimates of signal transmission channels without precise information of channel order,” *In Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2005.

[77] K. Hofbauer, “Estimating frequency and amplitude of sinusoids in harmonic signalsŮa survey and the use of shifted fourier transforms,” *MasterŠs thesis, Graz University of Technology*, May 2004.

[78] J.R. Hopgood. “Bayesian Blind MIMO Deconvolution of Nonstationary Autoregressive Sources Mixed Through All-Pole Channels,” *In Proc. of IEEE Statistical Signal Processing (SSP)*, pp.422-425, Sep. 2003.

[79] J.R. Hopgood and P.J.W. Rayner, “Blind Single Channel Deconvolution using Nonstationary Signal Processing,” *IEEE Trans. on Speech and Audio Processing*, Vol.11, Issue 5, pp.476-488, Sep. 2003.

[80] Y. Huang, J. Benesty, and G.W. Elko, “Adaptive Eigenvalue Decomposition Algorithm for Real-Time Acoustic Source Localization System,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.2, pp.937-940, Mar. 1999.

[81] J. Chen, Y. Huang, and J. Benesty, “An Adaptive Blind SIMO Identification Approach to Joint Multichannel Time Delay Estimation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp.53-56, May 2004.

[82] Y. Huang, J. Benesty, and J. Chen (ed.), “Acoustic MIMO Signal Processing,” *Springer*, Oct. 2006.

[83] M.Z. Ikram and D.R. Morgan,“A Beamforming Approach to Permutation Alignment for Multichannel Frequency-Domain Blind Speech Separation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002.

[84] *ITU-T Recommendation P.862*, “Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone network and speech codecs,” 2001

[85] J.L. Flanagan, A.C. Surendran, and E.E. Jan, “Spatially selective sound capture for speech and audio processing,” *Speech Communication*, Vol.13, pp.207-222, Oct. 1993.

[86] Q. Lin, E.E. Jan, and J. Flanagan, “Microphone Arrays and Speaker Identification,” *IEEE Trans. on Speech and Audio Processing*, Oct. 1994.

[87] E.E. Jan and J. Flanagan, “Microphone Arrays for Speech Processing,” *In Proc. of IEEE Int. Symp. on Signals, Systems, and Electronics (ISSSE)*, Oct. 1995.

[88] G. J., Jang and T. W. Lee, “ A probabilistic approach to single channel source separation. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in neural information processing systems* , MIT Press, 2003.

[89] A. Jourjine, S. Rickard, and O. Yilmaz,“Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp. 2985-2988, June 2000.

[90] D.T.M. Slock, “Form Sinusoids in Noise to Blind Deconvolution in communications”. In Kailath, A. Paulraj, V. Roychowdhury, and C.D. Shaper, editors, *Communications, computation, control and signal processing*, Kluwer Academic Publishers, 1997.

[91] F. Keiler and S. Marchand,“Survey On Extraction Of Sinusoids in Stationary Sounds,” *In Proc. of Digital Audio Effects Conf. (DAFX)*, pp. 59-64, Sept. 2002.

[92] S. Kim and R.A. Iltis, "A Matching-Pursuit/GSIC-based Algorithm for DS-CDMA Sparse-Channel Estimation," *IEEE Signal Processing Letters*, Vol.11, Issues:1, pp. 12-15, January 2004.

[93] S. Kim, Y. Jeong, and C. Lee, “Recapitulation of the IPDL positioning method,” *IEEE Trans. on vehicular technology*, Vol.54, No.1, Jan. 2005.

[94] M. Klajman and A.G. Constantinides, “A Combined Statistics Cost Function for Blind and Semi-Blind Source Separation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dec. 2001.

[95] A. Klapuri and M. Davy (ed.),“Signal Processing Methods for Music Transcription,” *Springer*, New York, 2006.

[96] C. H. Knapp and C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 24, Issue 4, pp. 320-327, Aug. 1976.

[97] H. Koshima and J. Hoshen, “Personal Locator Services Emerge,” *In IEEE Spectrum*, Vol.37, Issue 2, pp. 41-48, Feb. 2000.

[98] S. Krstulovic, R. Gribonval, P. Leveau, and L. Daudet, “A comparison of two extensions of the matching pursuit algorithm for the harmonic decomposition of sounds,” *In Proc. of IEEE Work. on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005.

[99] J. Krolik, M. Eizenman and S. Pasupathy, “Time Delay Estimation via Generalized Correlation with Adaptive Spatial Prefiltering,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.11, pp. 893-1896, Apr. 1986.

[100] S. Y. Kung, K. S. Arun, and D. B. Rao, “State-Space and Singular Value Decomposition Based Approximation Methods for Harmonic Retrieval Problem,” *J. of Opt. Soc. of America* , Vol.73, pp.1799-1811, Dec. 1983.

[101] A. Koul, J.E. Greenberg, “Using Intermicrophone Correlation to Detect Speech in Spatially Separated Noise,” *EURASIP Journal on Applied Signal Processing*, Issue 12, 2006.

[102] H. Laitinen, J. Lahteenmaki, and T. Nordstrom, “Database Correlation Method for GSM Location,” *In Proc. of IEEE Vehicular Technology Conference (VTC)*, Vol.4, pp. 2504-2508, May 2001.

[103] T. -W. Lee, A. J. Bell, and R. Lambert (ed.), “Blind Separation of Delayed and Convolved Sources,” *Advances in Neural Information Processing Systems*, Vol.9, pages 758-764, MIT Press 1997.

[104] M. Lenardi and D.T.M. Slock, "Estimation of Time-Varying Wireless Channels and Application to the UMTS W-CDMA FDD Downlink," *In Proc. of European Wireless (EW)*, Feb. 2002.

[105] S.N. Levine, T.S. Verma, and J.O. Smith, "Multiresolution Sinusoidal Modeling for Wideband Audio with Modifications," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.6, pp. 3585-3588, May 1998.

[106] J.S. Lim (Ed.), "Speech Enhancement," *Englewood Cliffs*, NJ: Prentice-Hall, 1983.

[107] X. Li and H. Fan, "Linear Prediction Methods for Blind Fractionally Spaced Equalization," *IEEE Trans. on Signal Processing*, Vol. 48, June 2000.

[108] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, Vol.41, pp. 3397-3415, December 1993.

[109] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *Technical report*, Courant Institute of Mathematical Sciences, 1993.

[110] U. Manmontri and P.A. Naylor, "Blind Identification Using Second-Order Statistics: a Nonstationarity and Nonwhiteness Approach," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp.305-308S, March 2005.

[111] J.A. Marks, "Real time speech classification and pitch detection," *In Proc. of Southern African Conf. on Communication and Signal Processing (COMSIG)*, pp.1-6, June 1988.

[112] J.S. Marques, I.M. Trancoso, J.M. Tribolet, and L.B. Almeida, "Improved pitch prediction with fractional delays in CELP coding," *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp.665-668, April 1995.

[113] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 34, Pages: 744-754, August 1986.

[114] P. Meinicke and H. Ritter, "Independent Component Analysis with Quantizing Density Estimators," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dec. 2001.

[115] N. Mitianoudis and M. Davies,"Audio Source Separation of convolved mixtures," *IEEE Trans. on Speech and Audio Processing*, 2002.

[116] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.36, Feb. 1988.

[117] L. Molgedey and G. Schuster, "Separation of a Mixture of Independent Signals Using Time Delayed Correlations," *Physical Review Letters*, Vol.72, No.23, pp.3634-3637, June 1994.

[118] R. Mukai, H. Sawada, S. Araki, and S. Makino, " Blind Source Separation of Many Signals in the Frequency Domain," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp.969-972, May 2006.

[119] M. Najar and J. Vidal, "Kalman Tracking for Mobile Location in NLOS Situations," *In Proc. of IEEE Personal, Indoor and Mobile Radio Communications (PIMRC)*, Vol.3, pp. 2203-2207, Sept. 2003.

[120] M. Najar, J.M. Huerta, J. Vidal, and J.A. Castro, "Mobile Location with Bias Tracking in Non-Line-Of-Sight," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* , Vol.3, pp. 956-959, May 2004.

[121] A.K. Nandi, "Blind Estimation Using Higher-Order Statistics, " *Kluwer Acadamic Publishers*, Boston 1999.

[122] L-T Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using Time-Frequency Distributions," *In Proc. of Int. Symp. on Signal Processing and its Applications (ISSPA)*, Aug. 2001.

[123] T. Nakatani and M. Miyoshi, "Blind Dereverberation of Single Channel Speech Signal Based on Harmonic Structure," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.1, pp.92-95, Apr. 2003.

[124] D. Nuzillard and J.-M. Nuzillard, “Second-Order Blind Source Separation in the Fourier Space of Data,” *Signal Processing*, Vol.83, pp.627-631, 2003.

[125] R.K. Olsson and L.K. Hansen, “Blind Separation of More Sources than Sensors in Convolutive Mixtures?,” *In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.5, pp. 657-660, May 2006.

[126] A.V. Oppenheim, R.W. Schafer, and T.G. Stockham Jr., “Nonlinear Filtering of Multiplied and Convolved Signals,” *IEEE Trans. on Audio Electroacoust.*, Vol.AU-16, No.3, pp. 437-466, Sept. 1968.

[127] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, “Discrete-Time Signal Processing,” *Prentice-Hall*, 2nd ed., 1999.

[128] J. Ortega-Garcia and J. Gonzalez-Rodriguez, “Overview of speech enhancement techniques for automatic speaker recognition,” *In Proc. of Int. Conf. on Spoken Language Processing (ICSLP)*, Vol. 2, pp. 929-932, 1996.

[129] D. O’Shaughnessy, “Enhancing speech degraded by additive noise or interfering speakers,” *IEEE Communications Magazine*, Vol. 27, Issue 2, pp. 46-52, Feb. 1989.

[130] S. Ozen, M.D. Zoltowski, and M. Fimoff, "A Novel Channel Estimation Method: Blending Correlation and Least-Squares Based Approaches," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp. 2281-2284, May 2002.

[131] S. Ozen and M.D. Zoltowski, "Time-of-Arrival (TOA) Estimation Based Structured Sparse Channel Estimation Algorithm, with Applications to Digital TV Receivers," *In proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.4, pp. 481-484, April 2003.

[132] K. Paliwal, “Estimation of Noise Variance from the Noisy AR Signal And Its Application in Speech Enhancement,” *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. 36, No.2, Feb. 1988.

[133] C.B. Papadias and D.T.M. Slock, "Fractionally Spaced Equalization of Linear Polyphase Channels and Related Blind Techniques Based on Multichannel Linear Prediction," *IEEE Trans. on Signal Processing*, Vol.47, Issue 3, pp. 641-654, March 1999.

[134] D.D. Muresan and T.W. Parks, "Orthogonal, Exactly Periodic Supspace Decomposition," *IEEE Trans. on Signal Processing*, Vol. 51, No. 9, Sept. 2003.

[135] J.D. Wise, J.R. Caprio, and T.W. Parks, "Maximum Likelihood pitch estimation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 51, pp. 418-421, May 1976.

[136] L.C. parra, "An Introduction to Independent Component Analysis and Blind Source Separation" *Part of a course on Neural Network in the EE department in Princeton University*, November 1998.

[137] L. Parra and C. Spence, "Convolutive Blind Source Separation Based On Multiple Decorrelation," *technical report*, April 1998.

[138] L.C. Parra and C.V. Alvino,"Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming," *IEEE Transaction on Speech and Audio Processing*, Vol. 10, No. 6, September 2002.

[139] P.M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, pp. 1527-1529, Nov. 1986.

[140] V. F. Pisarenko, "The Retrieval of Harmonics from a Covariance Function," *Geophysical J. Royal Astron. Soc.*, Vol.33, pp. 347-366, May 1973.

[141] J.-D. Polack, "La Transmission de l'Energie Sonore dans les Salles," *PhD thesis, Universite du Maine*, Le Mans, France, 1988.

[142] M. Porretta, P. Nepa, G. Manara, F. Giannetti, M. Dohler, B. Allen, and A.H. Aghvami, "User Positioning Technique for Microcellular Wireless Networks," *IEEE Electronics Letters*, Vol.39, Issue 9, pp. 745-747, May 2003.

[143] M. Porretta, P. Nepa, G. Manara, F. Giannetti, M. Dohler, B. Allen, and A.H. Aghvami, “A Novel Single Base Station Location Technique for Microcellular Wireless Networks: Description and Validation by a Deterministic Propagation Model,” *IEEE Trans. on Vehicular Technology*, Vol.53, Issue 5, pp. 1502-1514, Sept. 2004.

[144] P. Prandom, M. Goodwin, and M. Vetterli, “Optimal Time Segmentation for Signal Modeling and Compression,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, Pages: 2029-2032, April 1997.

[145] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, “Minimization or Maximization of Functions,” *Numerical Recipes in C (The Art of Scientific Computing)*, chapter 10, pp. 402–405. Cambridge University Press, USA, 2nd edition, 1992.

[146] J. C. Principe, and H.-C. Wu “Blind Separation of Convolutive Mixtures,” *In Proc. of International Joint Conference on Neural Networks (IJCNN)*, Vol.2, pages 1054-1058, July 1999.

[147] G. M. Riche de Prony, “Essai Expérimental et Analytique: sur les Lois de la Dilatabilité de Fluides Élastiques et sur celles de la Force Expansive de la Vapeur de l’Eau et de la Vapeur de l’Alcool á Différentes Températures,” *In Journal de l’école polytechnique*, Vol.1, No.22, pp. 24-76, 1795.

[148] K. Rahbarr and J.P. Reilly,“Blind Source Separation Algorithm for MIMO Convolutive Mixtures,” *In Proc. of Int. Conf. on Independent Component Analysis (ICA)*, pp.224-229, Dec. 2001.

[149] B.D. Rao, “Signal Processing with the Sparseness Constraint,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.3, pp. 1861-1864, May 1998.

[150] I. Reuven and H. Messer, "A Barankin-type lower bound on the estimation error of a hybrid parameter vector," *IEEE Trans. on Information Theory*, Vol.43, pp.1084-1093, May 1997.

[151] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality

assessment of telephone networks and codecs," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 749-752, May 2001.

[152] S. T. Roweis, " One Microphone Source Separation," *In Proc. of Neural Information Processing Systems (NIPS)*, pp. 793-799, Nov. 2000.

[153] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT-A Subspace Rotation Approach to Estimation of Parameters of Cisoids in Noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 5, pp. 1340-1342, Oct. 1986.

[154] R. Roy and T. Kailath, "Total Least Squares ESPRIT," *In Proc. of the IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 297-301, Nov. 1987.

[155] H. Sameti, H. Sheikhzadeh, L. Deng, and R.L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. on Signal Processing*, Vol. 6, Issue 5, pp. 445-455, Sept. 1998.

[156] H. Saruwatari, T. Kawamura, K. Sawai, A. Kaminuma, and M. Sakata,"Blind Source Separation Based on Fast-Convergence Algorithm Using ICA and Beamforming for real Convolutive mixture ," *In Proc. of IEEE Work. on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2001.

[157] H. Saruwatari, T. Kawamura, and K. Shikano,"Fast-Convergence Algorithm for ICA-Based Blind Source Separation Using Array Signal Processing ," *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* , pp.3097-3100, May 2002.

[158] F. Sattar, M.Y. Siyal, L.C. Wee, and L.C. Yen, "Blind Source Separation of Audio Signals Using Improved ICA Method," *In Proc. of IEEE Signal Processing*, pages 452-455, 2001.

[159] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propagat.*, Vol.43, No.3, pp. 276-280, Mar. 1986.

[160] M.R. Schroeder, “Statistical Parameters of the Frequency Response Curves of Large Rooms,” *J. of Audio Eng. Soc.*, Vol.35, pp.299-305, 1987.

[161] M.R. Schroeder and H. Kuttruff, “On Frequency Response Curves in Rooms. Comparison of Experimental, Theoretical and Monte-Carlo Results for the Average Frequency Spacing Between Maxima,” *J. Acoust. Soc. Amer.*, Vol.34, pp.34-76, 1962.

[162] M.R. Schroeder, “Frequency Correlation Functions of Frequency Responses in Rooms,” *J. Acoust. Soc. Amer.*, Vol.34, pp.1819-1823, 1963.

[163] D.J. Shyy and B. Rohani, “Indoor Location Technique for 2G and 3G Cellular/PCS Networks,” *In Proc. of IEEE Conf. on Local Computer Networks (LNC)*, pp. 264-271, Nov. 2000.

[164] A. de Cheveigné and M. Slama, “Acoustic Scene Analysis based on Power Decomposition ,” *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, May 2006.

[165] D.T.M. Slock, “Blind Fractionally-Spaced Equalization, Perfect-Reconstruction Filter Banks and Multichannel Linear Prediction,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 4, pp.585-588, Apr. 1994.

[166] D.T.M Slock and E. DeCarvalho, “Matched Filter Bounds for Reduced-Order Multichannel Models,” *In Proc. of IEEE GlobalCommunications Conf. (GLOBECOM)*, pp.77-81, Nov. 1996.

[167] C.B. Papadias and D.T.M. Slock, “Fractionally Spaced Equalization of Linear Polyphase Channels and Related Blind Techniques Based on Multichannel Linear Prediction,” *IEEE Trans. on Signal Processing*, Vol. 47, pp.641-654, Mar. 1999.

[168] P. Smaragdis, “Efficient Blind Separation of Convolved Sound Mixtures” *In Proc. of IEEE Work. on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 1997.

[169] P. Smaragdis, “Blind Separation of convolved mixtures in the Frequency Domain” *In Proc. of Int. Work. on Independence and Artificial Neural Networks (I&ANN)*, Feb. 1998.

[170] S. Subramaniam, A.P. Petropulu, and C. Wendt, “Cepstrum-Based Deconvolution for Speech Dereverberation,” *IEEE Trans. on Speech, and Audio Processing*, Vol. 4, Issue 5, pp. 392-396, Sep. 1996.

[171] L. Ta-Hsin and J.D. Gibson, “Speech analysis and segmentation by parametric filtering,” *IEEE Trans. on Speech and Audio Processing*, Vol. 4, Issue 3, pp. 203-213, May 1996.

[172] A. Tarighat, N. Khajehnouri, and A.H. Sayed, "CDMA Location Using Multiple Antennas and Interference Cancellation," *In Proc. of IEEE Vehicular Technology Conf. (VTC)*, Vol.4, pp. 2711-2715, April 2003.

[173] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, “Automatic phonetic segmentation,” *IEEE Trans. on Speech, and Audio Processing*, Vol. 11, Issue 6, pp. 617-625, Nov. 2003.

[174] L. Tong, V.C. Soon, R. Liu, and Y.-F. Huang, “AMUSE: A New Blind Identification Algorithm,” *In Proc. of IEEE Int. Symp. on Circuits and Systems (ISCAS)*, May 1990.

[175] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang, “Indeterminacy and Identifiablity of Blind Identification,” *IEEE Trans. On Circuits And Systems*, Vol.38, No.5, May 1991.

[176] K. Torkkola,“Blind Separation of Convolved Sources Based on Information Maximisation,” *In Proc. of IEEE Work. on Neural Networks for Signal Processing (NNSP)*, Sept. 1996.

[177] K. Torkkola, “Blind Separation for Audio Signals - Are we There Yet?,” *In Proc. of Int. Conf. on Independent Component Analysis (ICA)*, Jan. 1999.

[178] Mahdi Triki and Dirk T.M. Slock, “Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition,” *In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp.233-236, March 2005.

[179] Mahdi Triki and Dirk T.M. Slock, “A Novel Voiced Speech Enhancement Approach Based on Modulated Periodic Signal Extraction,” *In Proc. of European Signal Processing Conf. (EUSIPCO)*, Sept. 2006.

[180] Ahmed Triki, Mahdi Triki and Dirk T.M. Slock, “Periodic Signal Extraction with Frequency-Selective Amplitude Modulation and Global Time-Warping for Music Signal Decomposition,” *Submitted to the IEEE Int. Conf Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2008.

[181] Mahdi Triki and Dirk T.M. Slock, “Multi-channel mono-path periodic signal extraction with global amplitude and phase modulation for music and speech signal analysis,” *In Proc. of IEEE Work. on Statistical Signal Processing (SSP)*, pp.77-82, July 2005.

[182] Mahdi Triki and Dirk T.M. Slock, “Music Source Separation via Sparsified Dictionaries vs. Parametric Models,” *In Proc. of Int. Sym. on Communications, Control, and Signal Processing (ISCCSP)*, March 2006.

[183] M. Triki and D.T.M. Slock, “Blind Dereverberation of Quasi-periodic Sources Based on Multichannel Linear Prediction,” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2005.

[184] M. Triki and D.T.M. Slock, “Delay and Predict Equalization For Blind Speech Dereverberation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

[185] M. Triki and D.T.M. Slock, “Iterated Delay and Predict Equalization for Blind Speech Dereverberation,” *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2006

[186] M. Triki and D.T.M. Slock, “Multivariate LP Based MMSE-ZF Equalizer Design Considerations and Application to MultiMicrophone Dereverberation,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2007.

[187] M. Triki and D.T.M. Slock, “AR Source Modeling Based on Spatiotemporally Diverse Multichannel Outputs and Application to Multimicrophone Dereverberation,” *In Proc. of Int. Conf. on Digital Signal Processing (DSP)*, July 2007.

[188] Mahdi Triki, Salah Abdellatif, and Dirk T.M. Slock, "Interference Cancellation with Bayesian Channel Models and Application to TDOA/IPDL," *In Proc. of Int. Symp. on Signal Processing and its Applications (ISSPA)*, August 2005.

[189] Mahdi Triki, and Dirk T.M. Slock, "Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation," *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2005.

[190] Mahdi Triki, and Dirk T.M. Slock, "Investigation of Some Bias and MSE Issues in Block-Component-Wise Conditionally Unbiased LMMSE," *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2006.

[191] Mahdi Triki, Dirk T.M. Slock, Vincent Rigal, and Pierrick François, "Mobile Terminal Positioning via Power Delay Profile Fingerprinting: Reproducible Validation Simulations," *In Proc. of IEEE Vehicular Technology Conf.*, Sept. 2006.

[192] Mahdi Triki and Dirk T.M. Slock, “The Instrumental Variable Multichannel FAP-RLS and FAP Algorithms,” *In Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, July 2004.

[193] Tayeb Sadiki, Mahdi Triki, and Dirk T.M. Slock, “Window Optimization Issues in Recursive Least-Squares Adaptive Filtering and Tracking,” *In Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov. 2004.

[194] M.C. Vanderveen, A.-J. Van der Veen, and A. Paulraj, “Estimation of Multipath Parameters in Wireless Communications,” *IEEE Trans. on Signal Processing*, Vol.46, No.3, Mar. 1998.

[195] M.C. Vanderveen, C.B. Papadias, and A. Paulraj, “Joint Angle and Delay Estimation (JADE) for Multipath Signals Arriving at an Antenna Array,” *IEEE Communication Letters*, Vol.1, No.1, Jan. 1997.

[196] A.-J. Van der Veen, M.C. Vanderveen, and A. Paulraj, “Joint Angle and Delay Estimation Using Shift-Invariance Techniques,” *IEEE Trans. on Signal Processing*, Vol.46, No.2, Feb. 1998.

[197] A.J. van der Veen, S. Talwar, and A. Paulraj, “A Subspace Approach to Blind Space-Time Signal Processing for Wireless Communication Systems,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.45, pp. 173-190, Jan. 1997.

[198] T. Virtanen and A. Klapuri,"Separation of Harmonic Sound Sources Using Sinusoidal Modeling," *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[199] S. Voran, "Objective estimation of perceived speech quality - part I: Development of the measuring normalizing block technique," *IEEE Trans. on Speech, and Audio Processing*, Vol. 7, Issue 4, pp. 371-382, July 1999.

[200] G. H.Watson and K. Gilholm, "Signal and image feature extraction from local maxima of generalized correlation," *Pattern Recogn.*, Vol.31, No.11, pages: 1733-1745, 1998.

[201] M. Wax and T. Kailath, "Decentralized Processing in Sensor Arrays," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.33, Issue 5, pp. 1123-1129, Oct. 1985.

[202] O. Hilsenrath and M. Wax, "Radio Transmitter Location Finding for Wireless Communication Network Service and Management," *US Patent*, 6 026 304, Feb. 2000.

[203] M. Wax, Y. Meng and O. Hilsenrath, "Subspace signature matching for location ambiguity resolution in wireless communication systems," *US Patent*, 6 064 339, May 2000.

[204] M. Wax, O. Hilsenrath and A. Bar, "Radio Transmitter Location Finding in CDMA Wireless Communication Systems," *US Patent*, 6 249 680, June 2001.

[205] A.J. Weiss, "Direct Position Determination of Narrowband Radio Frequency Transmitters," *IEEE Signal Processing Letters*, Vol.11, Issue 5, pp. 513-516, May 2004.

[206] J.Y.C. Wen, N.D. Gaubitch, E.A.P. Habets, T. Myatt, and P.A. Naylor. "Evaluation of Speech Dereverberation Algorithms Using the MARDY Database." *In Proc. of Int. Work. on Acoustic Echo and Noise Control (IWAENC)*, Sep. 2006.

[207] C.Y. Wuu and A. Pearson, "On time delay estimation involving received signals," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 32, Issue 4, pp. 828-835, Aug. 1984.

[208] H.H. Yang and S.-I. Amari, “Adaptive On-Line Learning Algorithms for Blind Separation - Maximum Entropy and Minimum Mutual Information,” *Neural Computation*, Vol.9, pages 1457-1482, 1997.

[209] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, “Speech Enhancement Using Excitation Source Information,” *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.1, pp.541-544, May 2002.

[210] D. Yelling and E. Weinstein, ”Criteria for Multichannel Signal Separation,” *IEEE Transactions On Signal Processing*, Vol.42, No.8, August 1994.

[211] Y.V. Zakharov and T.C. Tozer,“Frequency Estimator with Dichotomous Search of Periodogram Peak,” *Electronics Letters*, Vol.35, Issue 19, pp.1608-1609, Sep. 1999.

[212] D. Zimmermann, J. Baumann, A. Layh, F. Landstorfer, R. Hoppe, and G. Wolfle, “Database Correlation for Positioning of Mobile terminals in Cellular Networks Using Wave Propagation Models,” *In Proc. of IEEE Vehicular Technology Conf. (VTC)*, Vol.7, pp. 4682-4686, Sept. 2004.

[213] M. Zoltawski and D. Stavrinides, “Sensor Array Signal Processing via a Procrustes Rotations Based Eigen-Analysis of the ESPRIT Data Pencil,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 6, pp. 832-861, June 1989.