Modeling of Convolutive Audio Mixtures Applied to Source Separation

Ngoc Q.K. Duong

N° d'ordre : 4437 ANNÉE 2011

UNIVERSITÉ DE RENNES 1

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du Signal et Télécommunications

Ecole doctorale Matisse

présentée par

Quang-Khanh-Ngoc DUONG

préparée à l'unité de recherche l'INRIA
Institut National de Recherche en Informatique et Automatique

---

Full-rank Gaussian Modeling of Convolutive Audio Mixtures Applied to Source Separation

Thèse soutenue à Rennes
le 15 novembre 2011

devant le jury composé de :

Jean-Jacques Fuchs
Professeur, Université de Rennes 1 / président

Gaël Richard
Professeur, Télécom ParisTech / rapporteur

Christian Jutten
Professeur, Université Joseph Fourier / rapporteur

Shoji Makino
Professor, University of Tsukuba / examinateur

Emmanuel Vincent
Chargé de Recherche, INRIA Rennes / examinateur

Rémi Gribonval
Directeur de Recherche, INRIA Rennes / directeur de thèse

Acknowledgements

I would like to acknowledge many people who have helped me along the way to this milestone. Foremost, I sincerely thank my advisors Emmanuel Vincent and Rémi Gribonval for continuous support to my research, and for patient correction of my work. I would not have been able to complete this thesis without their guidance. Besides, I would like to thank Gaël Richard, Christian Jutten, Jean-Jacques Fuchs, and Shoji Makino for accepting to be part of my thesis committee and for their insightful comments on my pre-defense.

I would like to express my most sincere gratitude to Frédéric Bimbot, Stephanie Lemaile and all other member of METISS research team for a great working atmosphere with lots of support they have shared daily with me. I have learned a lot from their motivation, enthusiasm, and immense knowledge. Many thanks also to Alexey Ozerov, Valentin Emiya, Prasad Sudhakar, and Simon Arberet for the numerous interesting discussions they provided.

I would also like to thank Shigeki Sagayama, Nobutaka Ono, Nobutaka Ito, Hideyuki Tachibana, and many other members of the Sagayama/Ono Lab at the University of Tokyo for warmly welcoming me into their lab and the helpful research collaboration they offered. I am grateful to my master supervisor Seung-Hyon Nam, who kindly supported my study in Korea when I deeply learned about signal processing and started to look into the source separation problem.

Grateful thanks also go to many Vietnamese friends in Rennes, who have been with me in the "Union des Étudiants Vietnamiens à Rennes" to share the good and bad times. Especially, I would like to express my profound appreciation to my girlfriend Anh-Thanh-Huyen Do for taking good care of me, and for her patience even during my hard times.

Last but not the least, I would like to express my love to my parents and sisters who have always been there for me, encouraging me and helping me to be who I am. This thesis would not have been possible without the love and affection that they have provided.

Abstract

We consider the task of under-determined and determined reverberant audio source separation, that is the extraction of the signal of each source from a multichannel audio mixture.

We propose a general Gaussian modeling framework whereby the contribution of each source to all mixture channels in the time-frequency domain is modeled as a zero-mean Gaussian random variable whose covariance encodes both the spatial and the spectral characteristics of the source. In order to better account for the reverberant mixing process, we relax the conventional narrowband assumption resulting in rank-1 spatial covariance and compute the upper bound on the separation performance achievable with full-rank spatial covariance. Experimental results indicate an improvement of up to 6 dB Signal-to-Distortion Ratio (SDR) in moderate to high reverberant conditions which supports this generalization. We also consider the use of quadratic time-frequency representations and that of the auditory-motivated equivalent rectangular bandwidth (ERB) frequency scale to increase the amount of exploitable information and decrease the overlap between the sources in the input representation.

After this theoretical validation of the proposed framework, we focus on estimating the model parameters from a given mixture signal in a practical blind source separation scenario. We derive a family of Expectation-Maximization (EM) algorithms to estimate the parameters either in the maximum likelihood (ML) sense or in the maximum a posteriori (MAP) sense. We propose a family of spatial location priors inspired by the theory of room acoustics as well as a spatial continuity prior and investigate the use of two spectral priors previously used in a single-channel or rank-1 multichannel context, namely spectral continuity and Nonnegative Matrix Factorization (NMF). The source separation results given by the proposed approach are compared with several baseline and state-of-the-art algorithms on both simulated mixtures and real-world recordings in various scenarios.

Titre: Modélisation gaussienne de rang plein des mélanges audio convolutifs appliquée à la séparation de sources.

Résumé

Nous considérons le problème de la séparation de mélanges audio réverbérants déterminés et sous-déterminés, c'est-à-dire l'extraction du signal de chaque source dans un mélange multicanal.

Nous proposons un cadre général de modélisation gaussienne où la contribution de chaque source aux canaux du mélange dans le domaine temps-fréquence est modélisée par un vecteur aléatoire gaussien de moyenne nulle dont la covariance encode à la fois les caractéristiques spatiales et spectrales de la source. Afin de mieux modéliser la réverbération, nous nous affranchissons de l'hypothèse classique de bande étroite menant à une covariance spatiale de rang 1 et nous calculons la borne théorique de performance atteignable avec une covariance spatiale de rang plein. Les résultats expérimentaux indiquent une augmentation du rapport Signal-à-Distorsion (SDR) de 6 dB dans un environnement faiblement à très réverbérant, ce qui valide cette généralisation. Nous considérons aussi l'utilisation de représentations temps-fréquence quadratiques et de l'échelle fréquentielle auditive ERB (*equivalent rectangular bandwidth*) pour accroître la quantité d'information exploitable et décroître le recouvrement entre les sources dans la représentation temps-fréquence.

Après cette validation théorique du cadre proposé, nous nous focalisons sur l'estimation des paramètres du modèle à partir d'un signal de mélange donné dans un scénario pratique de séparation aveugle de sources. Nous proposons une famille d'algorithmes *Expectation-Maximization* (EM) pour estimer les paramètres au sens du maximum de vraisemblance (ML) ou du maximum *a posteriori* (MAP). Nous proposons une famille d'*a priori* de position spatiale inspirée par la théorie de l'acoustique des salles ainsi qu'un *a priori* de continuité spatiale. Nous étudions aussi l'utilisation de deux *a priori* spectraux précédemment utilisés dans un contexte monocanal ou multicanal de rang 1: un *a priori* de continuité spatiale et un modèle de factorisation matricielle positive (NMF). Les résultats de séparation de sources obtenus par l'approche proposée sont comparés à plusieurs algorithmes de base et de l'état de l'art sur des mélanges simulés et sur des enregistrements réels dans des scénarios variés.

Contents

List of Figures

List of Tables

LIST OF TABLES

Notations and glossary

Standard mathematical symbols

$\mathbb{Z}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\Re$	Real part of a complex number
$\Im$	Imaginary part of a complex number
$\mathcal{N}_c$	Complex Gaussian distribution
$\mathbb{E}$	Expectation of a random variable

Vectors and matrices

a	Scalar
$\mathbf{a}$	Vector
$\mathbf{A}$	Matrix
$\mathbf{A}^T$	Matrix transpose
$\mathbf{A}^H$	Matrix conjugate transpose
$\mathbf{A}^{-1}$	Matrix inverse
$\mathbf{I}_I$	Identity matrix of size $I \times I$
$\text{diag}(\mathbf{a})$	Diagonal matrix with $\mathbf{a}$ as its diagonal
$\text{tr}(\mathbf{A})$	Matrix trace
$\det(\mathbf{A})$	Matrix determinant

Indices

t	Time sample index
f	Frequency index
n	Time frame index
i	Channel index
j	Source index

Sizes

I	Number of channels
J	Number of sources
N	Number of time frames
L	STFT filter length

General parameters

$\mathbf{x}(t) \in \mathbb{R}^I$	Time-domain mixture signal
$\mathbf{s}(t) \in \mathbb{R}^J$	Time-domain source signals
$\mathbf{c}_j(t) \in \mathbb{R}^I$	Time-domain j-th source image
$\mathbf{h}_j(\tau) \in \mathbb{R}^I$	Mixing filter from the j-th source to all microphones
$\mathbf{x}(n,f) \in \mathbb{C}^I$	Time-frequency domain mixture signals
$\mathbf{s}(n,f) \in \mathbb{C}^J$	Time-frequency domain source signals
$\mathbf{c}_j(n,f) \in \mathbb{R}^I$	Time-frequency domain j-th source image
$v_j(n,f) \in \mathbb{R}^+$	Variance of the j-th source
$\mathbf{R}_j(n,f) \in \mathbb{C}^{I\times I}$	Spatial covariance matrix of the j-th source
$\mathbf{\Sigma}_j(n,f) \in \mathbb{C}^{I\times I}$	Covariance matrix of the j-th source image
$\hat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f) \in \mathbb{C}^{I\times I}$	Empirical mixture covariance

Abbreviations

BSS	Blind Source Separation
DFT	Discrete Fourier transform
DoA	Direction of Arrival
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
ICA	Independent Component Analysis
i.i.d	independent and identically distributed
ML	Maximum Likelihood
MAP	Maximum a Posteriori
MMSE	Minimum Mean Square Error
PCA	Principal Component Analysis
STFT	Short-Time Fourier Transform
T-F	Time-Frequency

LIST OF TABLES

Part I

Introduction and state of the art

Chapter 1

Introduction

In this chapter, we first present the motivation for considering the problem of real-world audio source separation. We then provide an overview of this problem focusing on its basic formulation and general framework, followed by a summary of the thesis contributions. By the end of this chapter, the organization of the thesis is presented chapter by chapter.

1.1 Motivation

The human being is endowed with a multitude of senses. Among them, hearing is one of the most natural and fundamental sensing systems which provides key inputs for human perception. However, real-world sound scenes are usually mixtures of various sound sources and require humans great ability to locate, identify and differentiate sound sources which are heard simultaneously. Luckily, humans with normal hearing are generally very good at this task. For instance, one can turn round so as to face someone talking and follow a conversation in a noisy hall.

Contrary to humans, machines have been less successful at separating sound sources. Current audio source separation systems perform fairly well for studio recordings with a small number of sources. But their performance drops far below satisfaction in real-world recording environments, where there are many sound reflections on the walls and/or background noise [1].

Therefore, this thesis focuses on the challenging problem of audio source separation in real-world environments.

1.1.1 Audio source separation is essential

In daily life, recorded sound scenes often result from the superposition of several *sources* which is known as the *cocktail party* effect. A typical situation is in a cocktail party room where many people talk simultaneously in the presence of music and noise as shown in Fig. 1.1. Similar situations also happen for instance in outdoor recordings, where there is interference from a variety of environmental sounds, or in a music concert scenario, where the sound is generated by different instruments. In such situations, the resulting sound is a so-called *mixture*, which prevents people from well perceiving the target sound sources. Audio *source separation* aims to extract the signals of individual sound sources from the observed mixture with little prior information about the sources. This is formally known as *blind* source separation (BSS).

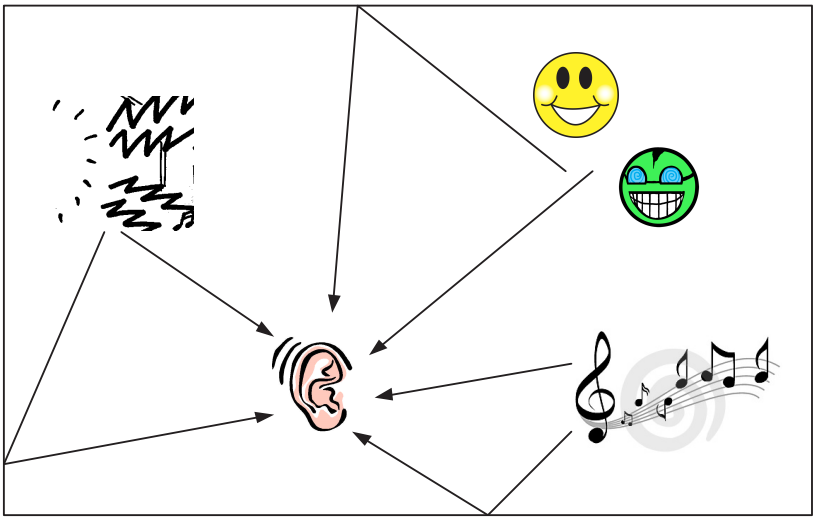


Figure 1.1: A typical cocktail party effect where the listened/recorded sound is the mixture of several sound sources such as speech, background music and background noise.

Audio source separation would benefit many practical applications. Humans would definitely better understand well-separated sound sources than mixtures with interfering sources [2, 3]. Hearing-impaired people would strive for a hearing aid equipment employing BSS to help them focus on a target source since they have lost the normal ability to separate it from the mixture [4, 5]. Moreover, the estimated sources may be further processed to extract features for automatic speech recognition (ASR) systems, automatic audio indexing systems, etc, [6, 7]. In a musical context, the extraction of individual musical components offered by BSS plays a key role in many music information

retrieval (MIR) tasks such as vocal extraction, instrument identification, *etc*, [8, 9, 10].

1.1.2 Target challenges: reverberation, diffuse sources and small movements

There are two typical types of real-world recording environments: studio or outdoor recording environments, known as *anechoic* environments, where the microphones capture only the direct sound propagation from a source, and indoor recording environments, known as *reverberant* environments, where the microphones capture not only the direct sound but also many sound reflections on the walls. State-of-the-art BSS algorithms perform quite well in noiseless anechoic conditions, *i.e.* result in the order of up to 12 dB Signal-to-Distortion Ratio (SDR) separation performance, but poorly and far from satisfaction in reverberant conditions, *i.e.* result in the order of 3 dB SDR. These numerical performance results are clearly shown in the recent community-based Signal Separation Evaluation Campaigns (SiSEC) [11, 12, 13, 1]. Therefore, addressing the separation of reverberant mixtures remains one of the key scientific challenges in the source separation community.

Besides, while most BSS algorithms are sensitive to background noise, all real-world recordings do suffer from nonstationary background noise which generally acts as an additional spatially *diffuse* source and remarkably decreases the separation performance [13]. Diffuse sources also arise for instance in music concerts where the sound of large musical instruments is generated from different spatial positions. Additionally, many other practical challenging situations such as source movements, short signal duration, etc, also need to be taken into account in order to radically address the BSS problem.

To sum up, audio source separation remains very difficult when considering real-world situations. **In this thesis, we focus on three main challenges: reverberant conditions, diffuse sources and small source movements, and leave aside other challenges such as large source movements and short signal duration. These three target challenges are related in the sense that they all result in spatial diffusion of the sound field generated by each source.**

1.2 Overview of the problem

In the previous section, we mentioned several terms like sources, mixture, source separation, reverberation, under-determined case. In this section, we define them more precisely while formulating the considered audio source separation problem.

1.2.1 Problem formulation

Let us consider a multichannel mixture signal recorded by an array of I microphones and denote it by $\mathbf{x}(t) = [x_1(t), ..., x_I(t)]^T$. This mixture signal can be expressed as [14]

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1.1}$$

where J denotes the number of sources, and $\mathbf{c}_j(t) = [c_{1j}(t), ..., c_{Ij}(t)]^T$ is the *spatial image* of the j-th source, that is the contribution of this source to all mixture channels. Note that background noise is also considered as a source in (1.1). In the case when the j-th source is a *point source*, *i.e.* it emits sound from a single position in space, $\mathbf{c}_j(t)$ is characterized as [15]

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t - \tau) \tag{1.2}$$

where $\mathbf{h}_j(\tau) = [h_{1j}(\tau), ..., h_{Ij}(\tau)]^T$ are linear *mixing filters* modeling the acoustic path from the j-th source to all I microphones and $s_j(t)$ is the emitted single-channel source signal.

Source separation consists in recovering either the J original source signals $s_j(t)$ or their spatial images $\mathbf{c}_j(t)$ given the I-channel mixture signal $\mathbf{x}(t)$. In the former case, the estimation of $s_j(t)$ usually results in a *scaling or filtering ambiguity* since both the sources and the filters can be arbitrarily scaled or filtered. However, this troublesome indeterminacy does not occur for the recovery of the source images. Only a *permutation ambiguity* remains, unless specific prior information is exploited about each source.

Industrial applications often require that the hardware resources be minimum. In other words, the number of mixture channels (*i.e.* microphones for recording) in the system should be as small as possible but make it possible to separate as many sources as possible. Many current BSS techniques have been designed for separating *determined* or *over-determined* mixtures where $I \geq J$: the number of microphones is equal or larger, respectively, than the number of sources. *Under-determined* mixtures where $I < J$ have remained much more difficult due in particular to the fact that the number of observable variables is less than the number of unknown variables, especially in reverberant conditions [12, 13, 1]. For instance, it was shown in [12] that the separation of determined mixtures of three sources in lab/office rooms resulted in a SDR of up to 6 dB while that of stereo mixtures only resulted in a SDR on the order of 3 dB.

In this thesis, we focus on the recovery of the source spatial images both in the determined case and in the under-determined case, assuming that the

number of sources J is known.

1.2.2 Mixing filters

The shape of the mixing filters h_{ij} varies between different physical environments. In *anechoic* environments where the microphones just capture the direct sound propagation path from a source, the mixing filters are scaled pure delay filters as shown in the left side of Fig. 1.2, that is $h_{ij} = a_{ij}\delta(t - t_{ij})$, where a_{ij} is the sound amplitude attenuation through the environment and t_{ij} the sound transmission time from the j-th source to the i-th microphone. A special and simpler form of the mixing filters occurs in *instantaneous* mixtures such as for instance certain music mixtures generated by amplitude panning. In this case, there is no time delay and $t_{ij} = 0$.

Indoor reverberant environments result in *convolutive* mixing with a more complex form of the mixing filters. In this case, h_{ij} is often modeled as a finite impulse response (FIR) filter [15] as shown in the right side of Fig. 1.2. The length L of h_{ij} varies depending on the acoustical properties of the room, which are often specified by the *reverberation time* T_{60}, that is the time required for the reflections of a direct sound to decay by 60 dB below the level of that direct sound.

1.2.3 General framework for source separation

In order to discriminate the sources from their mixture, source separation systems exploit a number of cues about the source spatial positions or about the source spectral structures [16, 15]. These two types of cues called *spatial cues* and *spectral cues*, respectively, are briefly discussed in the following and in more detail in Section 2.2 and 2.1, respectively. Source separation generally necessitates to exploit both types of cues. In particular, spatial cues alone do not suffice to discriminate sources coming from close directions while spectral cues alone may not enable one to discriminate sources with similar pitch range and timbre.

1.2.3.1 Spatial cues

Spatial cues encode any information related to the spatial position of a source over time. Many BSS algorithms exploit basic spatial cues derived from the phase and amplitude of the mixture channels, that is *interchannel time difference* (ITD) and *interchannel intensity difference* (IID). Fig. 1.3 illustrates a trivial example of use of IID in an anechoic mixing condition. At the source position p nearer to microphone 1, the recorded signal level x_1 is higher than x_2 and the corresponding IID when the source amplitude

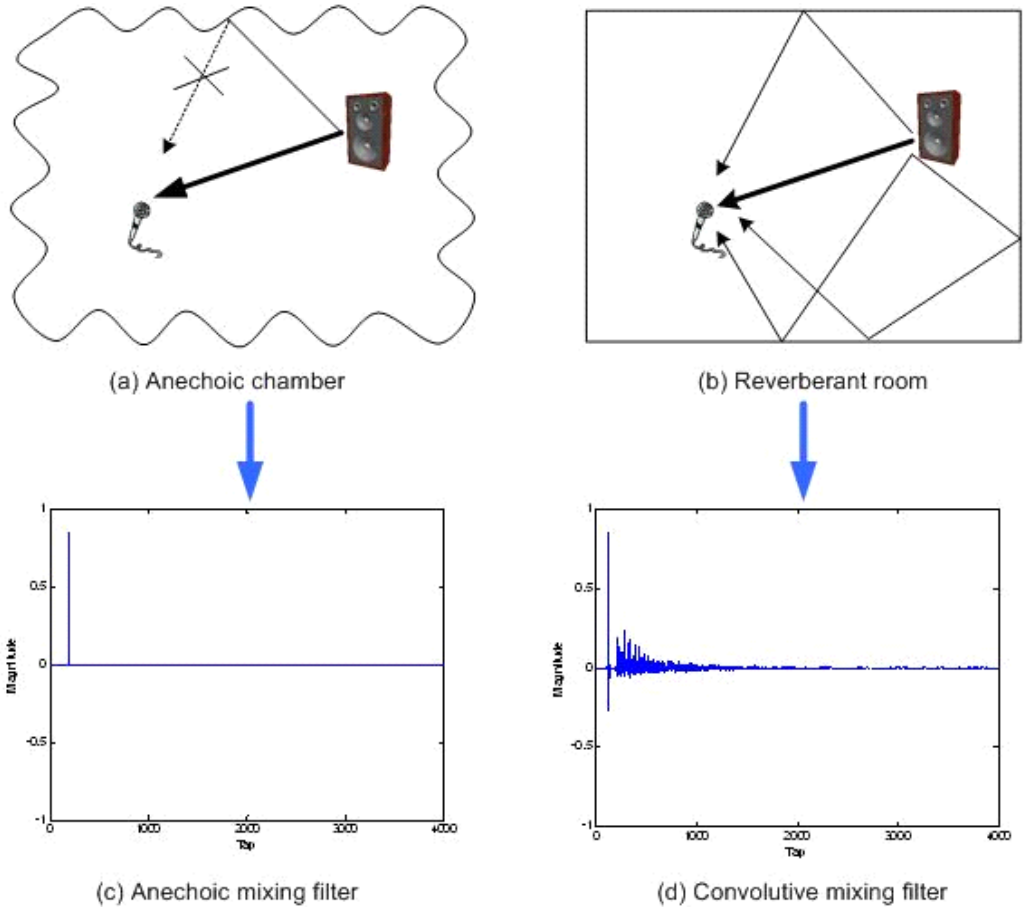


Figure 1.2: Types of mixing filters.

varies is modeled by the green plain line p in Fig. 1.3b. On the contrary, the source position q results in smaller x_1 than x_2, and the corresponding IID is represented by the red dotted line in Fig. 1.3b. The observed IID is therefore constant over time and directly related to the source *direction of arrival* (DoA). The idea of ITD and IID has been widely exploited in the history of both anechoic and convolutive source separation [17, 18, 19, 16, 20, 21]. In reverberant mixtures, the underlying assumption is that the ITD and IID deviate over frequency and time but remain close to their values as a function of the source DoAs in an anechoic mixture.

1.2.3.2 Spectral cues

Spectral cues encode the information related to the spectro-temporal structure of a source. This includes widely used features in audio signal processing in general as well

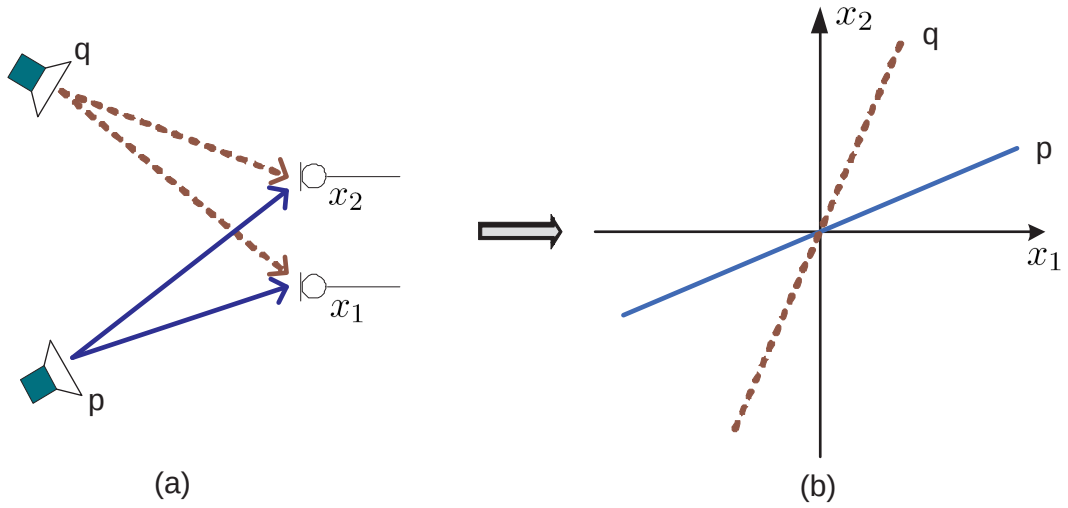


Figure 1.3: IID corresponding to two different source positions in an anechoic environment.

as in Computational Auditory Scene Analysis (CASA) [19] and BSS in particular due to the fact that different audio sources exhibit different recognizable spectro-temporal structures. For example, *sparsity* stands for an important feature of most audio signals in the time-frequency domain in the sense that many time-frequency coefficients are zero or close to zero [18, 16] while *harmonicity* and *smoothness* usually characterize the spectra of human voice and musical instruments [10]. Fig. 1.4 depicts an example of sparse and harmonic spectrum of a piano excerpt where the horizontal axis represents time, the vertical axis is frequency, and the color indicates the signal power in each time-frequency bin. Higher-level cues based on structured spectral models such as Spectral Gaussian Mixture Model (Spectral-GMM) [22, 23] or Nonnegative Matrix Factorization (NMF) [24, 25] have also been considered in the context of audio source separation.

1.2.3.3 General implementation under the narrowband assumption

Most BSS algorithms operate in the time-frequency (T-F) domain. The mixture signals are first transformed into the T-F domain via the short time Fourier transform (STFT). Denoting by F the number of frequency bins and by N the number of time frames, the mixing process (1.1) is hence rewritten as

$$\mathbf{x}(n,f) = \sum_{j=1}^{J} \mathbf{c}_j(n,f) \tag{1.3}$$

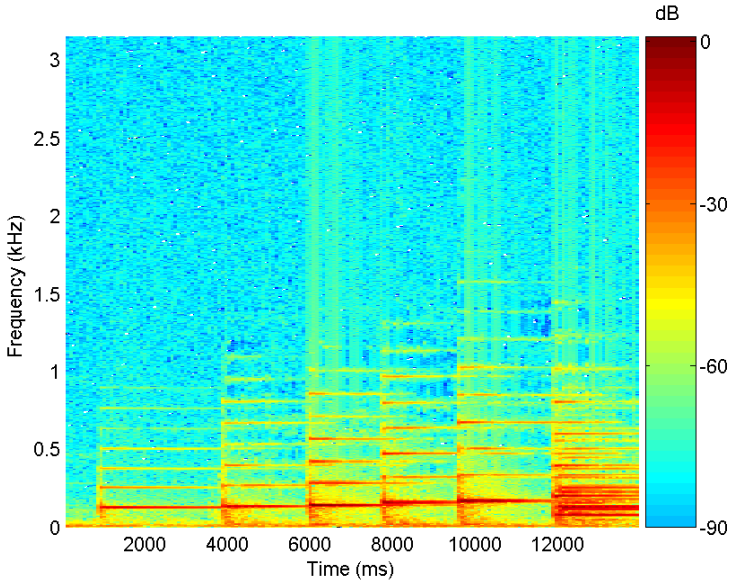


Figure 1.4: An example of sparse and harmonic spectrum.

where $1 \leq n \leq N$ denotes the time frame index and $1 \leq f \leq F$ the frequency bin index. The two major reasons for using a T-F representation are that it enables the use of spectral cues and that, under the so-called *narrowband assumption*, the linear convolution in (1.2) can be approximated by the simpler operation of complex-valued multiplication in each time-frequency bin (n, f)

$$\mathbf{c}_j(n, f) \approx \mathbf{h}_j(f) s_j(n, f) \tag{1.4}$$

where $\mathbf{c}_j(n, f)$ and $s_j(n, f)$ are the STFT coefficients of $\mathbf{c}_j(t)$ and $s_j(t)$, respectively, and $\mathbf{h}_j(f)$ is the Fourier transform of $\mathbf{h}_j(\tau)$. This equation can be seen as a *deterministic* parameterization of the mixing process where the mixture STFT coefficients can be perfectly predicted from the source coefficients given $\mathbf{h}_j(f)$. The amplitude and phase differences between the entries of $\mathbf{h}_j(f)$ are related to the IID and ITD at frequency f, respectively. Note that the narrowband assumption holds only when the length M of the mixing filters is short compared to the STFT window size [26]. In particular, it does not hold for reverberated or diffuse sources, whose sound comes from several directions. In such situations, the theory of room acoustics treats reverberation as a *probabilistic* phenomenon [27].

Given the mixing formulation in the T-F domain by (1.3) and (1.4), most existing approaches build a parametric model of the spatial cues, *i.e.* relating to $\mathbf{h}_j(f)$, and of

the spectral cues, *i.e.* relating to $|s_j(n,f)|$, and estimate the parameters in order to recover the original sources or source images. This general workflow is depicted in Fig. 1.5.

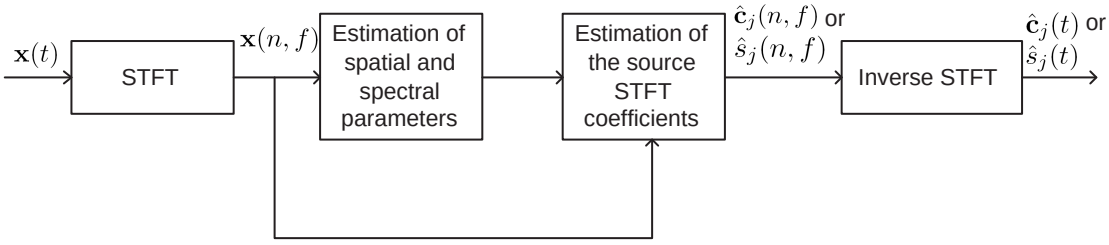


Figure 1.5: General source separation workflow under the narrowband assumption.

1.3 Contributions

Motivated by the fact that state-of-the-art source separation approaches have paid more attention to spectral cues than spatial cues, and that most existing work on spatial cues has relied on deterministic parameterization of the mixing process which is far from the actual characteristics of reverberation and diffuse sources, we focus in this thesis on the probabilistic modeling and parameterization of the mixing process for reverberated or diffuse sources. The achieved results have been described in our publications [28, 29, 30, 31, 32, 33, 34]. Our major contributions are to:

- propose a novel probabilistic modeling framework for spatial cues targeting reverberated or diffuse sources and overcoming the narrowband approximation to a certain extent [28, 30],
- design a general architecture for model parameter estimation and source separation [29, 30, 31],
- propose probabilistic spatial and spectral priors and demonstrate the potential of the proposed framework compared to state-of-the-art algorithms in various settings considering both simulated synthetic data and real-world recorded data [32, 33, 34].

One of our additional contributions which is not presented in this thesis is the participation in the organizing committee of the 2$^{\text{nd}}$ community-based Signal Separation

Evaluation Campaign (SiSEC 2010)[1]. The considered datasets and the results of state-of-the-art source separation systems are described in [13, 1].

Although we have not investigated the extension of the proposed modeling framework to other microphone array applications so far, this framework and the general parameter estimation algorithm described in this thesis could also largely be reused in different areas of audio signal processing such as diffuse noise suppression or spatial audio processing where accurate modeling of diffuse or reverberated sources is crucial. As a matter of fact, several authors have recently built up upon our work: Togami et al. applied our proposed probabilistic parameterization and modeling of spatial cues in the context of acoustic echo reduction [35] and online speech separation [36]; Ozerov et al. combined our proposed spatial cues model with advanced spectral cue models for robust audio source separation [37]; Araki et al. designed a *hybrid* source separation algorithm combining their existing time-frequency masking method as the first step with our proposed approach as the second step [38].

1.4 Outline of the thesis

This section describes the structure of the rest of the thesis which includes nine chapters organized in four parts:

- Part 1: Introduction and state of the art (Chapters 1 and 2)
- Part 2: Proposed probabilistic modeling framework (Chapters 3 and 4)
- Part 3: Estimation of the model parameters and source separation (Chapters 5 to 8)
- Part 4: Conclusion and perspectives (Chapter 9)

Chapter 2 presents a literature review of the considered audio source separation problem and summarizes existing work on spectral cues and spatial cues. It also describes two baseline approaches known as binary masking and ℓ_1-norm minimization used as benchmarks in subsequent experiments.

Chapter 3 presents the general probabilistic modeling framework where the STFT coefficients of the source images $\mathbf{c}_j(n,f)$ are modeled by a multivariate zero-mean Gaussian distribution whose unknown covariance is factorized into the product of a scalar

[1] http://sisec2010.wiki.irisa.fr/

source variance $v_j(n,f)$ and a *spatial covariance matrix* $\mathbf{R}_j(n,f)$. Several parameterizations of the spatial covariance matrix are then provided and followed by the general source separation architecture.

Chapter 4 is devoted to the investigation of upper bounds on the source separation performance when all or part of the model parameters are estimated from known source image signals. At this point, the potential of the proposed approach compared to state-of-the-art algorithms is shown in several experimental settings.

Chapter 5 focuses on the estimation of the model parameters $\{v_j(n,f), \mathbf{R}_j(n,f)\}_{j,n,f}$ from the observed mixture signals in a *blind* source separation context. We first present the general estimation architecture for both maximum likelihood (ML) and maximum a posteriori (MAP) criteria. Then the details of the proposed ML estimation algorithm are described including three steps: parameter initialization, parameter estimation by the expectation-maximization (EM) algorithm, and permutation alignment.

Chapter 6 derives and investigates several MAP estimation algorithms in order to improve the separation performance in certain scenarios where prior knowledge about the spatial cues is available. Two types of spatial priors are considered: a family of spatial location priors motivated by the theory of statistical room acoustics, and a spatial continuity prior. The algorithms for learning the hyper-parameters of these priors from training data are also presented.

Chapter 7 concentrates on the possibility to incorporate prior knowledge about the spectral cues via deterministic constraints or probabilistic priors into the proposed modeling framework. Two types of spectral priors are considered: a nonnegative matrix factorization (NMF) spectral model and a spectral continuity prior. The corresponding ML and MAP parameter estimation algorithms are then presented.

Chapter 8 evaluates the resulting separation performance in various scenarios. First, common data and evaluation metrics are described. Then the separation performance given by the proposed ML estimation algorithm is analyzed and compared with state-of-the-art and baseline algorithms in both simulated synthetic mixtures and real-world recorded mixtures. The potential of the MAP estimation algorithms presented in Chapters 6 and 7 is also addressed later in this chapter.

Chapter 9 ends the thesis by summarizing of the conclusions and presenting perspectives of future research directions, including fully blind source separation based on the estimation all acoustical parameters (*e.g.* reverberation time), and the use of the proposed framework in real-world information retrieval scenarios.

Chapter 2

State of the art

This chapter presents a literature review of existing audio source separation approaches. Though the whole BSS picture built in more than two decades of research is very large and requires an intensive survey, *e.g.* see [16, 20, 21], we limit our survey to major approaches and classify them into two categories depending on whether the underlying assumptions relate mostly to spectral cues or to spatial cues. For that purpose, several established works exploiting statistical models for spectral cues such as nongaussian independent and identically distributed (i.i.d.) models, nonstationary Gaussian models, and higher-level cue models are reported in Section 2.1. Some existing works exploiting deterministic parameterization for spatial cues are presented in Section 2.2. Section 2.3 is devoted to two well-known baseline approaches, namely binary masking and ℓ_1-norm minimization, which are used as benchmarks in subsequent sections.

2.1 Much work on spectral cues

This section reviews the classical families of source spectral models. We start by simple sparsity models and move towards more advanced models.

2.1.1 Nongaussian i.i.d. models

Let us start by two state-of-the-art classes of algorithms based on either independent component analysis (ICA) or sparse component analysis (SCA). In both cases, the original sources are assumed to be nongaussian and i.i.d..

2.1.1.1 Independent component analysis

A breakthrough in source separation came along the advent of ICA where the sources are assumed to be *statistically independent* of each other, *i.e.* the joint probability distribution of the sources is the product of the marginal probability distributions of individual sources, and *nongaussian* i.i.d. distributed. ICA was first proposed for the separation of instantaneous mixtures in the determined case where it attempts to make the output signals as statistically independent as possible. More formally, source separation is typically achieved by minimizing some approximation of the mutual information between the estimated sources. Assuming a fixed prior distribution for the sources, this is equivalent to ML estimation [39].

Various ICA-based instantaneous source separation algorithms have been proposed. For instance, an Infomax algorithm that maximizes the mutual information between the input and output of a nonlinear neural network was proposed in [40]. Another famous algorithm known as FastICA [41] combines the information theoretic approach proposed by Comon [42] and the projection pursuit approach for fast optimization. Other popular ICA algorithms are based on higher-order statistics, *e.g.* the fourth-order approximation of mutual information proposed by Cardoso [43, 44]. All these algorithms estimate the sources by inverting the estimated mixing system $\mathbf{H}$ composed of the mixing filters $\mathbf{h}_j$.

When considering convolutive determined mixtures, most ICA approaches operate in the time-frequency domain where the convolution in the time domain (1.2) is approximated by the complex-valued multiplication in (1.4). Then, complex-valued instantaneous ICA can be applied to separate the source coefficients in each frequency bin independently. A range of frequency-domain ICA (FDICA) algorithms have been proposed, *e.g.* [45, 46, 47, 48, 49, 50, 51, 52]. However FDICA suffers from the well-known permutation ambiguity, *i.e.* the source indexes are arbitrarily ordered in each frequency bin, because the source coefficients are estimated independently in each bin. This issue has remained challenging in reverberant mixtures where the source DoAs, which help aligning the source orders, are poorly estimated [48]. A popular FDICA algorithm including four steps described in [51, 48] is shown in Fig. 2.1.

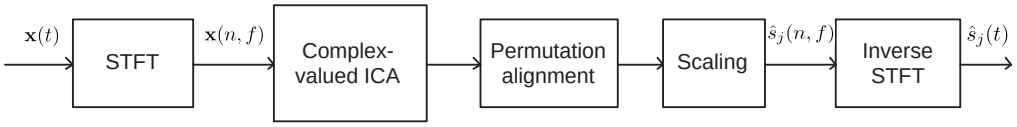


Figure 2.1: Typical structure of a frequency-domain ICA algorithm.

Despite the difficulty in solving the permutation problem, the primary limitation

of ICA-based approaches arises from the fact that they are not applicable to under-determined mixtures because the mixing system $\mathbf{H}$ is non-invertible.

2.1.1.2 Sparse component analysis

Similar nongaussian i.i.d assumptions in under-determined mixtures lead to sparse component analysis. SCA relies on the sparsity hypothesis, that is most source coefficients are zero or close to zero. This assumption generally holds for audio signals in the time-frequency domain as it can be seen in Fig. 1.4 where most time-frequency coefficients are close to zero. A closely related concept to sparsity is known as the disjointness of the supports of the source STFT coefficients, that is formulated by [53] as

$$s_{j_1}(n,f)s_{j_2}(n,f) \approx 0, \forall n, f, j_1 \neq j_2. \tag{2.1}$$

Most SCA-based source separation algorithms include two main steps [54]: (1) mixing parameter estimation, (2) source coefficient estimation. In the first step, the mixing vectors $\mathbf{h}_j(f)$ are estimated from the STFT coefficients of the mixture signal $\mathbf{x}(n,f)$, typically by a clustering algorithm [55, 56, 57]. In the second step, given the estimated mixing matrices, the vector of the source STFT coefficients are estimated either by *binary masking* [18]

$$\hat{s}_j(n,f) = \begin{cases} \frac{\mathbf{h}_j^H(f)\mathbf{x}(n,f)}{\|\mathbf{h}_j(f)\|^2} & \text{if source } j \text{ is predominant} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

or by solving the constrained minimization problem [58, 59]

$$\hat{\mathbf{s}}(n,f) = \arg \min_{\mathbf{s}(n,f)\in\mathbb{C}^J} \|\mathbf{s}(n,f)\|_p \text{ subject to } \sum_j \mathbf{h}_j(f)s_j(n,f) = \mathbf{x}(n,f) \tag{2.3}$$

where $\|\mathbf{s}\|_p$ is the ℓ_p norm of the vector $\mathbf{s}$ defined by

$$\|\mathbf{s}\|_p = \Big(\sum_{j=1}^{J} |s_j|^p\Big)^{\frac{1}{p}}. \tag{2.4}$$

Sparsity can be measured by the number of nonzero coefficients in a vector, that is its ℓ_0 norm. However the ℓ_0 norm is nonconvex and difficult to minimize directly. Hence, in practice relaxed versions of the ℓ_0 norm such as the ℓ_p norm are used. When $p \geq 1$, the ℓ_p norm is convex and the optimization problem is easier to solve. Note

that the optimization (2.3) is equivalent to MAP estimation when the source STFT coefficients are assumed to follow a generalized Gaussian prior with shape parameter p [58]. And it becomes a ℓ_1-*norm minimization problem* ($p = 1$) in the particular case when the STFT coefficients are Laplace distributed [55].

Various BSS algorithms exploiting sparsity in the T-F domain have been proposed to deal with under-determined mixtures. An early approach for anechoic mixing condition was presented by Aoki et al. [17]. Then a famous algorithm called Degenerate Unmixing Estimation Technique (DUET) was proposed [53, 18] where anechoic mixing matrices are first estimated, then only one dominant source in each time-frequency bin is extracted by binary masking. Other SCA-based algorithms have been designed for instantaneous and anechoic mixtures [60, 61, 62, 63, 58, 64] and convolutive mixtures [58, 65, 55, 66, 67]. ℓ_p-norm minimization approaches usually extract on the order of I sources per time-frequency [58]. Binary masking and ℓ_1-norm minimization will be presented in more details in Section 2.3.

2.1.2 Nonstationary Gaussian models

A distinct framework from nongaussian i.i.d. source modeling has emerged whereby the source STFT coefficients $s_j(n, f)$ are modeled as independent zero-mean Gaussian random variables with free variances $v_j(n, f)$ [21]:

$$p\big(s_j(n, f)\big) = \mathcal{N}_c\big(s_j(n, f)|0, v_j(n, f)\big). \tag{2.5}$$

Given this model, the mixture coefficients $\mathbf{x}(n, f)$ also follow a zero-mean Gaussian distribution with covariance matrix $\mathbf{\Sigma_x}(n, f)$. Under the narrowband assumption, $\mathbf{\Sigma_x}(n, f)$ is given by

$$\mathbf{\Sigma_x}(n, f) = \mathbf{H}(f)\mathbf{\Sigma_s}(n, f)\mathbf{H}^H(f) \tag{2.6}$$

where $\mathbf{\Sigma_s}(n, f)$ is the diagonal source covariance matrix whose diagonal entries are $v_j(n, f)$, and $\mathbf{H}(f)$ is the mixing matrix.

The model parameters $\theta = \{v_j(n, f), \mathbf{H}(f)\}_{j,n,f}$ are then estimated in the maximum likelihood (ML) sense by maximizing the log-likelihood

$$\log p(\mathbf{X}|\theta) = \sum_{n,f} -\mathrm{tr}(\mathbf{\Sigma_x}^{-1}(n, f)\widehat{\mathbf{\Sigma}}_\mathbf{x}(n, f)) - \log\det(\pi\mathbf{\Sigma_x}(n, f)) \tag{2.7}$$

where tr(.) and det(.) denote the trace and the determinant of a square matrix, respectively. The matrix $\hat{\mathbf{\Sigma}}_\mathbf{x}(n, f)$ is the empirical mixture covariance, which is computed

Cov1. either from the observed mixture signal in a single T-F bin as

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f) := \mathbf{x}(n,f)\mathbf{x}^H(n,f) \tag{2.8}$$

Cov2. or by locally averaging over the neighborhood of each T-F bin as [68]

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f) := \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^H(n',f') \tag{2.9}$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood such that $\sum_{n',f'} w_{nf}^2(n',f') = 1$. This model is known as Local Gaussian Model (LGM) because the source variances in different time-frequency bins are independent from each other. The time-varying source variances result in a *nonstationary* source model. The source STFT coefficients are finally obtained in the minimum mean square error (MMSE) sense by *Wiener filtering* as

$$\hat{\mathbf{s}}(n,f) = \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{2.10}$$

Note that the multichannel Wiener filter is equivalent to the combination of a spatial filter (minimum variance distortionless response (MVDR) beamformer) and a soft spectral mask (single-channel Wiener filter) [69]. Contrary to binary masking, which enables only one dominant source in each T-F bin, the source STFT coefficients estimated by multichannel Wiener filtering are all nonzero provided that $v_j(n,f) \neq 0, \forall j = 1,...,J$. This framework was introduced for the separation of determined instantaneous mixtures under the name of nonstationarity-based ICA [70], and was applied to the astronomical imaging problem in [71] where the separated components themselves naturally have Gaussian distributions. Then, it has been extended to the separation of underdetermined instantaneous audio mixtures in [72] and [68] and shown to provide better separation performance than ℓ_1-norm minimization.

Inspired by this Gaussian source model, in this thesis we will also consider a Gaussian modeling framework. But instead of modeling the source STFT coefficients $s_j(n,f)$, which encode only the source spectral information, we propose to model the source images $\mathbf{c}_j(n,f)$, which encode both spectral and spatial information of the source. The detail of our proposed modeling framework will be presented in Chapter 3.

2.1.3 Spectral Gaussian mixture model

Regardless of the mixing parameter estimation, the main limitation of nonstationarity-based or nongaussianarity-based ICA and SCA approaches is that they do not take into account the redundancy and structure of each audio source. This is because the T-F source coefficients are assumed to be independent of each other. In the following, we will review Gaussian model-based approaches where the redundancy and structure of each audio source can be exploited. In this subsection, we describe a more advanced source model, known as Spectral GMM, which incorporates a structured spectral model of each source [22, 73, 74, 23, 75].

The Spectral GMM approach models the short time Fourier spectrum of the j-th source $\mathbf{s}_j(n) = [s_j(n,f)]_f$, which is a column vector composed of all elements $s_j(n,f),\ f = 1,...,F$, as a multidimensional zero-mean complex-valued K-state Gaussian mixture with *probability density function* (pdf) given by

$$p\big(\mathbf{s}_j(n)\big) = \sum_{k=1}^{K} \pi_{jk} \mathcal{N}_c\big(\mathbf{s}_j(n)|\mathbf{0}, \mathbf{\Sigma}_{jk}\big) \tag{2.11}$$

where $\mathbf{0}$ denotes a vector of zeroes, π_{jk}, which satisfies $\sum_{k=1}^{K} \pi_{jk} = 1,\ \forall j$, and $\mathbf{\Sigma}_{jk} = \mathrm{diag}\big([v_{jk}(f)]_f\big)$ are the weight and the diagonal spectral covariance matrix of the k-th state of the j-th source, respectively, and

$$\mathcal{N}_c\big(\mathbf{s}_j(n)|\mathbf{0}, \mathbf{\Sigma}_{jk}\big) = \prod_f \frac{1}{\pi v_{jk}(f)} \exp\Big(- \frac{|s_j(n,f)|^2}{v_{jk}(f)} \Big). \tag{2.12}$$

This model can be viewed as a two-step generative process: at each time frame n of the j-th source, the first step is to pick one state $k(n)$ with probability $\pi_{jk(n)}$; given $k(n)$, the vector of STFT coefficients $\mathbf{s}_j(n)$ is generated from a zero-mean Gaussian distribution with covariance $\mathbf{\Sigma}_{jk(n)}$ in the second step. Source separation amounts to computing the posterior probability of all states in each time frame.

While the LGM presented in Section 2.1.2 assumes as many variance parameters $v_j(n,f)$ as T-F points, the Spectral GMM defines $K \times F$ free variances $v_{jk}(f)$ only and exploits the global structure of the sources to estimate them. However, since GMM does not explicitly model amplitude variation of sound sources, signals with different amplitude level but similar spectral shape may result in different estimated spectral variance templates $[v_{jk}(f)]_f$. To overcome this issue, another version of GMM called Spectral Gaussian Scaled Mixture Model (Spectral GSMM) was proposed in [76] where

a time-varying scaling parameter $g_{jk}(n)$ is incorporated in each Spectral-GMM. The pdf of the GSMM is then given by

$$p\big(\mathbf{s}_j(n)\big) = \sum_{k=1}^{K} \pi_{jk} \mathcal{N}_c\big(\mathbf{s}_j(n) | \mathbf{0}, g_{jk}(n)\mathbf{\Sigma}_{jk}\big). \tag{2.13}$$

Spectral GMM and Spectral GSMM have been applied to single-channel audio source separation in [22, 23, 76, 77], and stereo separation of moving sources in [78]. The GMM was also considered in multichannel instantaneous music mixtures [75] where the Spectral-GMMs are learnt from the mixture signals. Separation performance was then reported to improve by 5 dB SDR compared to binary masking in stereo mixtures of three sources.

2.1.4 Non-negative matrix factorization

NMF is an efficient decomposition technique which helps reducing data dimension [79]. It has been applied to various fields of machine learning and audio signal processing, *e.g.* music transcription [80, 24, 10]. In the following, we will present NMF as structured spectral source model known as Spectral NMF, applied to audio source separation.

In the Spectral NMF model, each source is the sum of K_j latent components as [25]

$$s_j(n,f) = \sum_{k=1}^{K_j} c_k(n,f) \quad \text{with} \quad c_k(n,f) \sim \mathcal{N}_c(0, h_{nk} w_{kf}) \tag{2.14}$$

where $h_{nk},\ w_{kf} \in \mathbb{R}^+$. The latent components are assumed to be mutually independent in each time-frequency bin, which implies that

$$p\big(s_j(n,f)\big) = \mathcal{N}_c\Big(s_j(n,f) | 0, \sum_{k=1}^{K_j} h_{nk} w_{kf}\Big). \tag{2.15}$$

Denoting by $\mathbf{S}_j$ the $N \times F$ matrix of STFT coefficients of the j-th source $[s_j(n,f)]_{nf}$, $\mathbf{H}_j = [h_{nk}]_{nk}$ with dimension $N \times K_j$ and $\mathbf{W}_j = [w_{kf}]_{kf}$ with dimension $K_j \times F$, it is shown in [24] that ML estimation of the latent variables $\mathbf{H}_j$ and $\mathbf{W}_j$ is equivalent to NMF of the power spectrogram $|\mathbf{S}_j|^2$ into $\mathbf{H}_j\mathbf{W}_j$ according to the Itakura-Saito (IS) divergence *i.e.*

$$-\log p\big(\mathbf{S}_j | \mathbf{H}_j, \mathbf{W}_j\big) \overset{c}{=} \sum_{nf} d_{IS}\big(|s_j(n,f)|^2 | [\mathbf{H}_j\mathbf{W}_j]_{nf}\big) \tag{2.16}$$

where $\overset{c}{=}$ denotes equality up to a constant and $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$. It now turns out into a form of *dimension reduction* in the sense that instead of estimating NF values of the power spectrogram $|\mathbf{S}_j|^2$, NMF offers the estimation of only NK_j values of $\mathbf{H}_j$ and K_jF values of $\mathbf{W}_j$ where $NK_j + K_jF \ll NF$.

Note that in the Spectral NMF model each source STFT coefficient is modeled as the sum of K elementary components (2.14) while in the Spectral GMM described in subsection 2.1.3 $s_j(n,f)$ is modeled as a process which can take only one of a number of states. Spectral NMF has been applied to single-channel speech separation [81, 82, 83] and multichannel speech/music separation [25] and shown to improve SDR by more than 1 dB compared to ℓ_1-norm minimization in under-determined mixtures [25]. In [84, 85], Spectral NMF is combined with Spectral GMM to form a more efficient, hybrid model (FS-HMM) compared to the Spectral NMF itself, which is applied to single-channel speech/music separation.

To sum up, various spectral models such as nongaussian i.i.d. models, LGM, GMM, GSMM, NMF have been proposed and deeply investigated in the literature of spectral cues for audio source separation scenarios. They can either be implemented separately into specific algorithms or incorporated as choices of spectral priors in a common general framework, *e.g.* in [37, 86].

2.2 Less work on spatial cues

In order to exploit spatial information about the sources for the purpose of separation, BSS approaches usually first parameterize the spatial cues and then estimate the spatial parameters. Contrary to much work on spectral cues exploiting probabilistic models as presented in Section 2.1, there has been less work on spatial cues. Most of it, which is reviewed in this section, employs either deterministic or probabilistic modeling of the IID and the ITD under the narrowband approximation. However the parameterization by IID/ITD itself remains *deterministic*.

2.2.1 Deterministic IID/ITD model

Most state-of-the-art spatial cue models rely on the *point source* assumption, by which sound from a source is assumed to come from a single position, and on the *narrowband approximation* in equation (1.4). The mixing process from the j-th source to the microphone array is hence characterized by the frequency dependent mixing vector $\mathbf{h}_j(f)$. For the sake of simplicity, let us consider in this section only the stereo case with unit-norm normalization and zero delay on the first channel. Remember that this is possible

without loss of generality due to the filtering indeterminacy about the original source signals. The mixing vector can now be deterministically written as

$$\mathbf{h}_j(f) = \begin{pmatrix} \cos\theta_j(f) \\ \sin\theta_j(f) e^{-2i\pi\phi_j(f)} \end{pmatrix} \tag{2.17}$$

where $\tan\theta_j(f)$ stands for the IID and $\phi_j(f)$ for interchannel phase difference (IPD), which relates to the ITD. Note that $\theta_j(f) = \theta_j$ is frequency-independent in anechoic mixtures and $\phi_j(f) = 0$ in instantaneous mixtures.

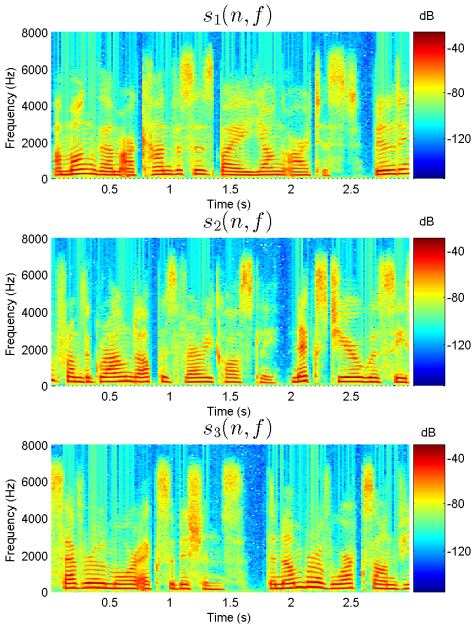
Most existing approaches that have been proposed to estimate the deterministic variables $\{\theta_j(f), \phi_j(f)\}_{j,f}$, and hence $\mathbf{h}_j(f)$, rely on a sparsity assumption [18, 87]. If the j-th source is predominant in a given time-frequency bin (n, f), then the ratio of the mixture signals provides an approximation of IID and ITD:

$$\frac{x_2(n,f)}{x_1(n,f)} \approx \tan\theta_j(f) e^{-2i\pi f\phi_j(f)}. \tag{2.18}$$

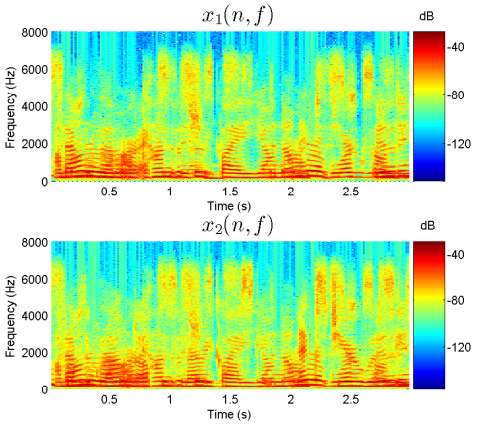
Therefore, if the sources have disjoint time-frequency supports, the data points $\{\mathbf{x}(n,f)\}_{nf}$ tend to form clusters along $\mathbf{h}_j(f)$. This phenomenon is shown by the example scatter plot in Fig. 2.2 for a stereo mixture of three sources. The geometrical structure of the scatter plot can be clustered, *e.g.* by clustering algorithms [88, 56], to obtain the estimates of $\mathbf{h}_j(f)$ in instantaneous, anechoic or mildly reverberant mixing conditions.

In line with this idea, many algorithms have been proposed to estimate the mixing parameters in under-determined instantaneous and anechoic mixtures [89, 90, 91, 92, 87]. Among them, the famous DUET algorithm was proposed under the above disjointness assumption in [53, 18]. The amplitude ratios and the phase differences of the STFT coefficients of stereo anechoic mixture signals are first computed in every time-frequency bin to build a two-dimensional histogram. Then the source directions are identified as the peaks of this IID-ITD histogram. Arberet et al. proposed another method [87, 57] known as Direction Estimation of Mixing MatrIX (DEMIX), which relies on a weaker sparsity assumption that, for each source, there exists one time-frequency region where it essentially contributes alone. DEMIX employs a sequential clustering algorithm over the data points weighted by a confidence measure for more robust mixing matrix estimation.

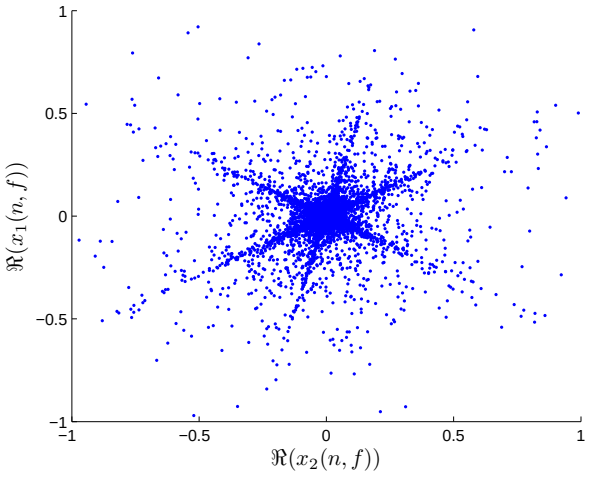
For diffuse or reverberated sources, the model in (2.17) and the approximation (2.18) do not hold and therefore do not provide accurate estimates of the mixing parameters due to the frequency dependence and variability of $\theta_j(f)$ and $\phi_j(t)$. However this deterministic IID/ITD model is still helpful to provide an initial estimate of $\mathbf{h}_j(f)$ [55],



(a) STFT magnitude of the three original sources



(b) STFT magnitude of the stereo mixture



(c) Scatter plot of the real part of the mixture STFT coefficients.

Figure 2.2: Scatter plot of the STFT coefficients of an instantaneous mixture of three sources.

and can also be exploited for solving the permutation problem in FDICA algorithms [49, 51, 93]. In the next subsection we review several existing approaches which exploit

probabilistic modeling of the IID and ITD for reverberated sources.

2.2.2 Probabilistic IID/ITD model

Probabilistic IID/ITD modeling has been considered for reverberated sources in a few source separation algorithms, *e.g.* [94, 95, 84, 66, 96, 97], where the IID and the ITD now vary with f. Mandel et al. [98, 95] proposed to model the IID and the IPD as Gaussian variables whose means and variances are estimated in the ML sense by the EM algorithm. This method treats the IID and the IPD as random variables. However it still relies on the narrowband approximation and the major assumption that the IID and the IPD are conditionally independent may not hold in practice since they both relate to the source DoAs.

The probabilistic spatial cue parameterization of diffuse sources has been considered in beamforming techniques [99, 100, 101] and source localization algorithms [102] to model background noise. However it has not yet been applied in the context of source separation to model reverberated sources. By probabilistic parameterization, we mean that the spatial image $\mathbf{c}_j(n,f)$ of each source cannot be deterministically computed from the source signal $s_j(n,f)$ and the mixing parameters anymore, but follows a certain probabilistic distribution. Izumi et al. consider a diffuse sound model in convolutive source separation [66] but only for the residual noise, while the transfer functions from the sources to the microphones are still deterministically parameterized by the ITD. One exception is the concurrent work of Pham et al. [96, 103, 97] where the observed log-ratio between the two channels in each time-frequency bin (n,f) $\log R(n,f) = \log[x_1(n,f)/x_2(n,f)]$ is modeled as a random variable whose real and imaginary parts are the log-IID and the IPD of the predominant source, respectively. This approach takes into account the correlation between the mixture channels which depends on the reverberation. Performance remains limited however, due to the sparsity assumption.

In summary, most state-of-the-art works on spatial cues, even when they follow probabilistic models, are based on deterministic parameterization by the IID and the ITD. This parameterization is relevant for instantaneous and anechoic mixtures but far from the actual characteristics of reverberation. This thesis therefore focuses on the probabilistic parameterization of spatial cues in order to deal with real-world diffuse and reverberated sources.

2.3 Baseline approaches

In this section, we detail the two baseline approaches used to benchmark the separation performance when considering the problem described in Fig. 2.3:

"*given the mixing matrix* $\mathbf{H}(f)$*, how to estimate the source or source image coefficients?*"

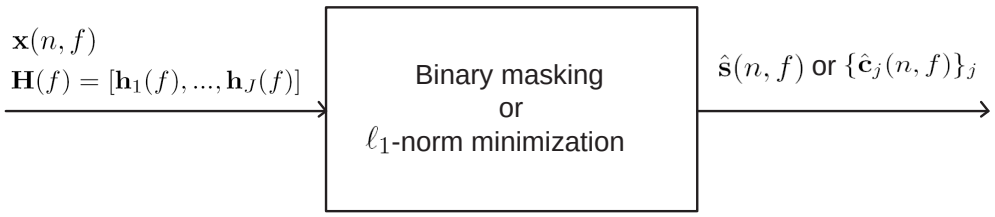


Figure 2.3: Baseline approaches for blind source STFT coefficient estimation

In the determined case where $\mathbf{H}(f)$ is square and invertible, the source coefficients can be easily computed by $\hat{\mathbf{s}}(n,f) = \mathbf{H}^{-1}(f)\mathbf{x}(n,f)$. However this computation is not applicable in the under-determined case since the mixing matrix is non-invertible. In this case, binary masking [18] or ℓ_1-norm minimization [55, 62] are typically applied. These approaches are also used as benchmarks in SiSEC. We use the reference software available for download [2].

2.3.1 Binary masking

Binary masking assumes that only one source is active in each time-frequency bin as in (2.1). In order to detect this active source, the mixture STFT coefficients $\mathbf{x}(n,f)$ are projected onto the subspace spanned by each mixing vector $\mathbf{h}_j(f)$, which is normalized to unit norm, and the source index j_0 yielding the largest projection is chosen:

$$j_0 = \arg\max_j \left| \frac{\mathbf{h}_j^H(f)}{\|\mathbf{h}_j(f)\|_2} \mathbf{x}(n,f) \right|. \tag{2.19}$$

[2] http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+speech+and+music+mixtures

The spatial image $\hat{\mathbf{c}}_{j_0}(n,f)$ of this source is then set to the projected mixture STFT coefficients, while these of the other sources are set to zero, that is

$$\hat{\mathbf{c}}_j(n,f) = \begin{cases} \frac{\mathbf{h}_j(f)\mathbf{h}_j^H(f)}{\|\mathbf{h}_j(f)\|^2}\mathbf{x}(n,f) & \text{if } j = j_0 \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{2.20}$$

Since binary masking extracts only one source per time-frequency bin, it usually produces strong musical noise artifacts due to discontinuities between neighboring time-frequency bins, especially in mixtures with many sources or strong reverberation where the disjointness assumption is more violated [21].

2.3.2 ℓ_1-norm minimization

Contrary to binary masking, ℓ_1-norm minimization estimates at least I active sources in each time-frequency bin by solving the constrained problem (2.3) with $p = 1$. This problem is equivalent to MAP estimation where the source STFT coefficients are assumed to follow the Laplace distribution [55]. In the real-valued case, it can be solved by linear programming [104] or by the shortest-path method [105, 63]. In [55], complex-valued ℓ_1-norm minimization is transformed into a second-order cone programming (SOCP) problem.

For simplicity, we adopt the shortest-path method [105, 63] which provides an approximate result but is several orders of magnitude faster than SOCP for the considered small number of sources and microphones. Let us denote by $\mathcal{K}_k^I \subset [1,2,...,J]$ the k-th subset of I sources among J sources, and $\tilde{\mathbf{H}}_k(f)$ the $I \times I$ matrix whose columns are $\mathbf{h}_j(f)$ where $j \in \mathcal{K}_k^I$. Assuming that the source indexes in $\mathcal{K}_k^I$ are predominant, their STFT coefficients can be computed as $\hat{\mathbf{s}}(n,f) = \tilde{\mathbf{H}}_k^{-1}(f)\mathbf{x}(n,f)$. This motivation leads to the detection of the I predominant sources in the ℓ_1-norm minimization sense by

$$k_0 = \arg\min_k \|\tilde{\mathbf{H}}_k^{-1}(f)\mathbf{x}(n,f)\|_1. \tag{2.21}$$

The source STFT coefficients are then estimated as

$$\hat{\mathbf{c}}_j(n,f) = \begin{cases} \mathbf{h}_j(f)[\tilde{\mathbf{H}}_{k_0}^{-1}(f)\mathbf{x}(n,f)]_{j_{k_0}} & \text{if } j \in \mathcal{K}_{k_0}^I \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{2.22}$$

where $[.]_j$ denotes the j-th entry of a vector, and j_{k_0} is the index of the j-th source in the subset $\mathcal{K}_{k_0}^I$.

ℓ_1-norm minimization usually produces fewer artifacts than binary masking due in

particular to the extraction of a larger number of sources in each time-frequency bin. However, the detection of the predominant sources via (2.21) often remains inaccurate since the mixture STFT coefficients do not suffice to discriminate predominant sources as shown in [21]. In this thesis, we present a *covariance* modeling framework in Chapter 3 which offers additional information about the correlation between mixture channels for better source discrimination.

Part II

Proposed probabilistic modeling framework

Chapter 3

General Gaussian framework

This chapter is devoted to the proposed source separation framework based on probabilistic modeling of spatial cues for reverberated or diffuse sources. For that purpose, we first describe the general Gaussian modeling framework where the STFT coefficients of the source images in each time-frequency bin are modeled by multivariate Gaussian random variables parameterized by their spectral variance and their spatial covariance. We then introduce four parameterizations of the spatial covariance depending on the level of constraint. Finally, a four-step general source separation architecture is presented.

3.1 Modeling framework

We start this section by describing the general probabilistic modeling framework adopted from now on. We then define four parameterizations with different degrees of flexibility resulting in rank-1 or full-rank spatial covariance matrices. We finally introduce the possibility to exploit spatial and spectral priors to enhance source separation performance in certain situations.

3.1.1 General Gaussian model

Our framework builds upon the nonstationary Gaussian modeling framework in Section 2.1.2 whereby the STFT coefficients of the source signals $s_j(n,f)$ are modeled by a Gaussian distribution whose variance is a function of (n,f). By contrast with [72, 68], we do not consider the source $s_j(n,f)$ but model the source image $\mathbf{c}_j(n,f)$ instead as a zero-mean Gaussian random vector with covariance matrix $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$.

Furthermore, we factor $\mathbf{\Sigma}_j(n,f)$ as

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(n,f) \tag{3.1}$$

where $v_j(n,f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(n,f)$ are $I \times I$ *spatial covariance matrices* encoding their spatial position and spatial spread. **This parameterization is probabilistic in the sense that $\mathbf{c}_j(n,f)$ can not be deterministically computed from the chosen parameters, but is randomly generated according to the considered Gaussian distribution. Note that this framework does not rely on the point source assumption nor on the narrowband assumption, hence it appears applicable to reverberated or diffuse sources**.

Under the classical assumption that the sources are uncorrelated, the vector $\mathbf{x}(n,f)$ of STFT coefficients of the mixture signal is also zero-mean Gaussian with covariance matrix

$$\mathbf{\Sigma}_\mathbf{x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(n,f). \tag{3.2}$$

The likelihood of the set of observed mixture STFT coefficients $\mathbf{x} = \{\mathbf{x}(n,f)\}_{n,f}$ given the set of variance parameters $v = \{v_j(n,f)\}_{j,n,f}$ and that of spatial covariance matrices $\mathbf{R} = \{\mathbf{R}_j(n,f)\}_{j,f}$ is given by

$$P(\mathbf{x}|v,\mathbf{R}) = \prod_{n,f} \frac{1}{\det\left(\pi\mathbf{\Sigma}_\mathbf{x}(n,f)\right)} e^{-\mathbf{x}^H(n,f)\mathbf{\Sigma}_\mathbf{x}^{-1}(n,f)\mathbf{x}(n,f)} \tag{3.3}$$

where $\mathbf{\Sigma}_\mathbf{x}(n,f)$ implicitly depends on $v_j(n,f)$ and $\mathbf{R}_j(n,f)$ according to (3.2). When the empirical mixture covariances are observed instead, the considered likelihood is expressed as

$$P(\widehat{\mathbf{\Sigma}}_\mathbf{x}|v,\mathbf{R}) = \prod_{n,f} \frac{1}{\det\left(\pi\mathbf{\Sigma}_\mathbf{x}(n,f)\right)} e^{-\mathrm{tr}\left(\mathbf{\Sigma}_\mathbf{x}^{-1}(n,f)\widehat{\mathbf{\Sigma}}_\mathbf{x}(n,f)\right)} \tag{3.4}$$

where $\widehat{\mathbf{\Sigma}}_\mathbf{x}(n,f)$ is the empirical mixture covariance computed either by (2.8) or (2.9).

Under this model, source separation can be achieved by first estimating the variance parameters v and the spatial parameters $\mathbf{R}$. Then the spatial images of all sources are obtained in the MMSE sense by multichannel Wiener filtering. This general source separation architecture will be presented in more detail in Section 3.2.

Note that, in the general situation involving possibly moving sources, the spatial

covariance matrices are time-varying as described by $\mathbf{R}_j(n,f)$. However, in most of the following, we will assume that the source positions are fixed and the reverberation is moderate such that the spatial covariance matrices are *time-invariant*: $\mathbf{R}_j(n,f) = \mathbf{R}_j(f)$. This assumption is also made by most state-of-the-art approaches where the mixing process is modeled under the narrowband approximation by the time-invariant mixing vector $\mathbf{h}_j(f)$.

3.1.2 Spatial covariance parameterization

We present four spatial covariance parameterizations resulting in either rank-1 or full-rank matrices. The first two rank-1 parameterizations correspond to existing techniques. The third parameterization called full-rank direct+diffuse is considered in [28] for the first time in the context of source separation though it has already been introduced in the context of source localization. The fourth parameterization called full-rank unconstrained is our proposed parameterization, which offers the greatest flexibility to model the convolutive mixing process compared to the three other parameterizations.

3.1.2.1 Rank-1 convolutive parameterization

Most existing approaches to audio source separation rely on narrowband approximation of the convolutive mixing process (1.2) by the complex-valued multiplication (1.4). The covariance matrix of $\mathbf{c}_j(n,f)$ is then given by (3.1) where $v_j(n,f)$ is the variance of $s_j(n,f)$ and $\mathbf{R}_j(f)$ is equal to the rank-1 matrix

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f) \tag{3.5}$$

with $\mathbf{h}_j(f)$ denoting the Fourier transform of the mixing filters $\mathbf{h}_j(\tau)$. This *rank-1 convolutive parameterization* of the spatial covariance matrices has recently been exploited in [106, 25] together with an NMF model of the source variances.

3.1.2.2 Rank-1 anechoic parameterization

For omni-directional microphones in an anechoic recording environment without reverberation, each mixing filter boils down to the combination of a delay τ_{ij} and a gain κ_{ij} specified by the distance r_{ij} from the j-th source to the i-th microphone [102]

$$\tau_{ij} = \frac{r_{ij}}{c} \quad \text{and} \quad \kappa_{ij} = \frac{1}{\sqrt{4\pi} r_{ij}} \tag{3.6}$$

where c is the sound velocity. The spatial covariance matrix of the j-th source is hence given by the *rank-1 anechoic parameterization*

$$\mathbf{R}_j(f) = \mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}\big)^H(f) \tag{3.7}$$

where the Fourier transform $\mathbf{h}_j^{\text{ane}}(f) \in \mathbb{C}^{I\times 1}$ of the mixing filters is now parameterized as

$$\mathbf{h}_j^{\text{ane}}(f) = \begin{pmatrix} \kappa_{1j} e^{-2i\pi f \tau_{1j}} \\ \vdots \\ \kappa_{Ij} e^{-2i\pi f \tau_{Ij}} \end{pmatrix}. \tag{3.8}$$

Note that this parameterization is consistent with that of (2.17) but here the mixing vectors are not normalized.

3.1.2.3 Full-rank direct+diffuse parameterization

One possible interpretation of the narrowband approximation is that the sound of each source as recorded on the microphones comes from a single spatial position at each frequency f, as specified by $\mathbf{h}_j(f)$ or $\mathbf{h}_j^{\text{ane}}(f)$. This approximation is not valid in a reverberant environment, since reverberation induces some spatial spread of each source, due to echoes at many different positions on the walls of the recording room. This spread translates into full-rank spatial covariance matrices.

The theory of statistical room acoustics assumes that the spatial image of each source is composed of two uncorrelated parts: a direct part, which is modeled by $\mathbf{h}_j^{\text{ane}}(f)$ in (3.8) for omni-directional microphones, and a reverberant part [27]. The spatial covariance $\mathbf{R}_j(f)$ of each source is then a full-rank matrix defined as the sum of the covariance of its direct part and the covariance of its reverberant part

$$\mathbf{R}_j(f) = \mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}\big)^H(f) + \sigma^2_{\text{rev}}\mathbf{\Omega}(f) \tag{3.9}$$

where σ^2_{rev} is the variance of the reverberant part and $\Omega_{il}(f)$ is a function of the microphone directivity pattern and the distance d_{il} between the i-th and the l-th microphone such that $\Omega_{ii}(f) = 1$. This *full-rank direct+diffuse* model assumes that the reverberation recorded at all microphones has the same power but is correlated as characterized by $\Omega_{il}(f)$. This model has been employed for single source localization in [102] and is currently considered for multiple source localization in [107] but not for source separation yet.

Assuming that the reverberant part is diffuse, *i.e.* its intensity is uniformly dis-

tributed over all possible directions, for omni-directional microphones its normalized cross-correlation can be shown to be real-valued and equal to [27] (see appendix B for the detailed derivation)

$$\Omega_{il}^{\text{diff}}(f) = \frac{\sin(2\pi f d_{il}/c)}{2\pi f d_{il}/c}. \tag{3.10}$$

The assumption that the reverberant part is diffuse is valid if some practical conditions are fulfilled [27, 102]:

a) the dimensions of the room are large compared to the wavelength (this condition is usually satisfied for the considered audio frequencies, *e.g.* $f > 200$ Hz),

b) the frequencies of interest should be above Schroeder's frequency $f_S = 2000\sqrt{T_{60}/V}$, where T_{60} and V are the reverberation time, *i.e.* the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound, and the total volume of the room, respectively,

c) the sources and the microphones are located in the interior of the room, *i.e.* at least half a wavelength away from the walls.

Moreover, the power of the reverberant part within a parallelepipedic room with dimensions L_x, L_y, L_z is given by

$$\sigma_{\text{rev}}^2 = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{3.11}$$

where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time T_{60} via Eyring's formula [102]

$$\beta = \exp\left\{ -\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}} \right\}. \tag{3.12}$$

Note that the covariance matrix $\mathbf{\Omega}(f)$ is usually employed for the modeling of diffuse background noise [66, 100, 101]. For instance, the source separation algorithm in [66] assumes that the sources follow an anechoic model and represents the non-direct part of all sources by a shared diffuse noise component with covariance $\mathbf{\Omega}(f)$ and constant variance. Hence this algorithm does not account for the correlation between the variances of the direct part and the non-direct part. On the contrary, the proposed direct+diffuse parameterization scales the direct and non-direct part of $\mathbf{R}_j(f)$ by the same variance $v_j(n,f)$, which is more consistent with the physics of sound.

3.1.2.4 Full-rank unconstrained parameterization

In practice, the assumption that the reverberant part is diffuse is rarely satisfied in realistically reverberant environments. Indeed, early echoes accounting for most of its energy are not uniformly distributed on the boundaries of the recording room. When performing some experiments in a simulated rectangular room, we observed that (3.10) is valid on average when considering a large number of sources distributed at different positions in a room, but generally not valid for each individual source. As an example, we computed the normalized cross-correlation $\Omega_{il}(f)$ between the reverberant part of the spatial image of one source at 2 microphones spaced by 20 cm in a room with reverberation time $T_{60} = 250$ ms. Fig. 3.1a shows that the observed cross-correlation is very different from $\Omega_{il}^{\text{diff}}(f)$. We then repeated this experiment for 400 source and microphone positions and Fig. 3.1b depicts that the mean of the observed correlation is accurately modeled by $\Omega_{il}^{\text{diff}}(f)$ and the standard deviation equals for most frequency bins.

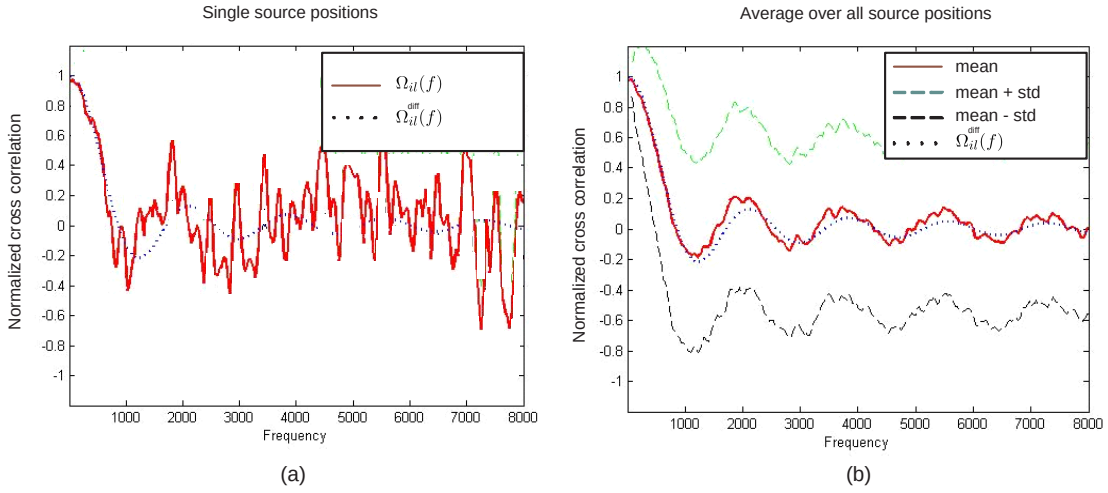


Figure 3.1: Normalized cross correlation of two microphone signals at $T_{60} = 250$ ms as a function of frequency

Therefore, we propose to parameterize the spatial information of each source via a *full-rank unconstrained* Hermitian positive semi-definite spatial covariance matrix $\mathbf{R}_j(f)$ whose coefficients are not deterministically related a priori. This unconstrained parameterization is the most general possible parameterization for a covariance matrix. It generalizes the above three parameterizations in the sense that any matrix taking the form of (3.5), (3.7) or (3.9) can also be considered as a particular form of unconstrained matrix. Fig. 3.2 depicts an example of a 2×2 rank-1 and a more general full-rank

unconstrained matrix. Because of this increased flexibility, this unconstrained parameterization better fits the data as measured by the likelihood. In particular, it improves the poor fit between the model and the data observed for rank-1 parameterizations due to the fact that the narrowband approximation underlying these models does not hold for reverberant mixtures. **In that sense, the proposed parameterization circumvents the narrowband approximation to a certain extent**.

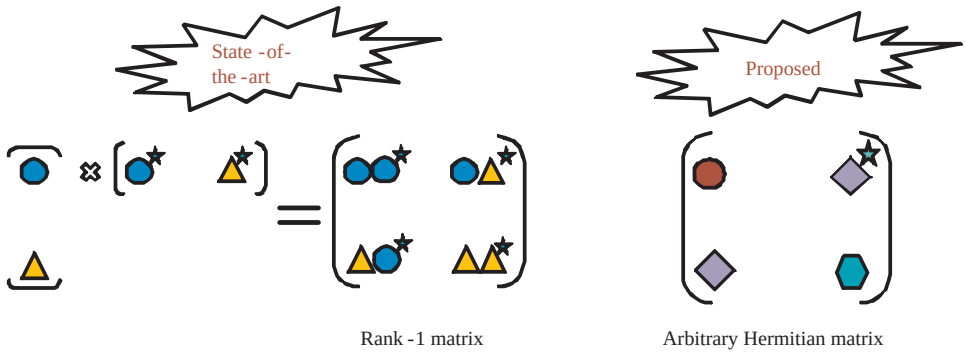


Figure 3.2: Example of a rank-1 and a full-rank spatial covariance parameterization.

The entries of $\mathbf{R}_j(f)$ are not directly interpretable in terms of simple geometrical quantities. The principal component of the matrix can be interpreted as a beamformer [108] pointing towards the direction of maximum sound power, while the ratio between its largest eigenvalue and its trace is equal to the ratio between the output and input power of that beamformer. In moderate reverberation conditions, the former is expected to be close to the source DoA while the latter is related to the ratio between the power of direct sound and that of reverberation. However, the strength of this parameterization is precisely that it remains valid to a certain extent in more reverberant environments, since it is the most general possible parameterization for a covariance matrix.

3.1.3 Number of parameters and priors on the parameters

Let us first discuss an example of the number of spatial parameters involved in each spatial covariance parameterization. Table 3.1 shows the number of real-valued parameters to be estimated for each spatial covariance parameterization after discounting redundant parameters for a stereo mixture of 3 sources with a FFT length of 1024. The rank-1 anechoic parameterization involves 6 distances r_{ij} from the sources to the microphones. The full-rank direct+diffuse parameterization involves two more parameters: the microphone spacing d and the variance of the reverberation part σ^2_{rev}. The rank-1 convolutive parameterization involves 2 real-valued parameters for each mixing

vector $\mathbf{h}_j(f)$, $f = 2, ..., 512$, since one of the two entries can be fixed arbitrary as in (2.17), and 1 parameter for $\mathbf{h}_j(1)$ and $\mathbf{h}_j(513)$. Hence it results in a total of 3072 parameters. Finally the full-rank unconstrained parameterization involves 4611 real-valued parameters: 3 parameters for each $\mathbf{R}_j(f)$, $f = 2, ..., 512$ because the scale can be fixed arbitrarily, and 2 parameters for $\mathbf{R}_j(1)$ and $\mathbf{R}_j(513)$. The two baseline approaches involves the same number of real-valued parameters as the rank-1 convolutive parameterization.

Rank-1 anechoic	6
Rank-1 convolutive	3072
Full-rank direct+diffuse	8
Full-rank unconstrained	4611
Binary masking	3072
ℓ_1-norm minimization	3072

Table 3.1: Number of real-valued parameters of the considered spatial models for a stereo mixture of 3 sources with a FFT length of 1024.

Let us now briefly discuss the possibility to incorporate priors on v_j and $\mathbf{R}_j$. The factorization of the covariance matrix in (3.1) offers a flexible treatment of spatial parameters $\mathbf{R}_j(f)$ and spectral parameters $v_j(n, f)$ in this Gaussian modeling framework. In the simplest case, the source variances are assumed to be unconstrained and the spatial covariance matrices follow one of the parameterizations presented in subsection 3.1.2, so that the model parameters $\theta = \{v_j(n, f), \mathbf{R}_j(f)\}_{j,n,f}$ can be estimated in the ML sense. This ML algorithm will be presented in Chapters 4 and 5.

However, in many practical situations either the spectral structure of the sources or the spatial source positions are known or can be estimated in advance. In these cases, deterministic constraints probabilistic prior distributions can be exploited to help enhance source separation performance. The incorporation of probabilistic spectral priors on v_j and probabilistic spatial priors on $\mathbf{R}_j$ is easy in the considered modeling framework where the parameters are subsequently estimated in the MAP sense. The general ML and MAP algorithms are introduced in section 5.1 and several spatial and spectral priors will be investigated in Chapters 6 and 7.

3.2 General source separation architecture

Under the *covariance modeling* framework presented in Section 3.1, source separation can be achieved using the following four-step general architecture. The overall structure

is shown in Fig. 3.3, and each step will be described by one subsection in the following.

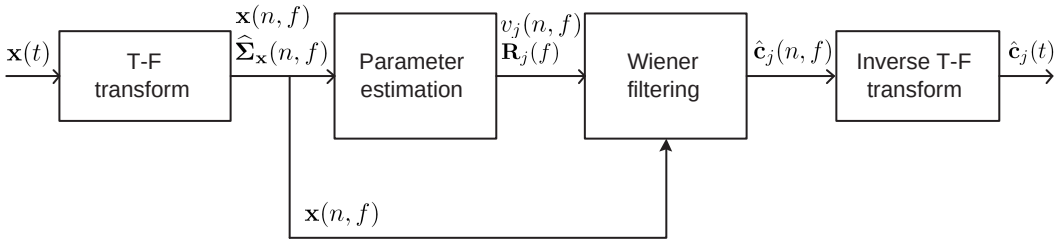


Figure 3.3: General source separation architecture.

3.2.1 Time-frequency transform

Most existing approaches transform the time-domain input signal $\mathbf{x}(t)$ into the time-frequency domain by the STFT which results in a *linear* T-F representation by a vector of complex-valued STFT coefficients $\mathbf{x}(n,f)$ [48, 21, 86]. In our proposed framework, we consider more generally a *quadratic* time-frequency representation where $\mathbf{x}(t)$ is represented via its empirical covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ in the neighborhood of each T-F bin. As a result, instead of estimating the model parameters from $\mathbf{x}(n,f)$, we infer them from $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$.

3.2.1.1 Linear time-frequency transform

We consider two linear T-F representations given by either the STFT or the auditory-motivated equivalent rectangular bandwidth (ERB) transform as follows.

L1. The STFT is the most popular T-F representation used for audio signals. The T-F mixture coefficients are given by

$$\mathbf{x}(n,f) = \sum_{\tau} H_f(\tau)\mathbf{x}(nM-\tau) \tag{3.13}$$

where $H_f(\tau) = w(\tau)e^{-i2\pi f_{\text{Hz}}\tau/f_s}$, w is a window function, f_s is the sampling frequency, M is the STFT frame shift, $f_{\text{Hz}} = (f-1)fs/L$ denotes frequency in Hz, and $1 \le n \le N$ and $1 \le f \le F$ denote time frame and frequency index, respectively. The mixing process is then written in the STFT domain by (1.3).

L2. In the ERB representation, the frequency scale is defined by [109]

$$f_{\text{Hz}} = \frac{e^{f_{\text{ERB}}/9.26} - 1}{0.00437} \tag{3.14}$$

where $f_{\text{ERB}} = \frac{f-1}{F-1} f_{\text{ERB}}^{\max}$ and $f_{\text{ERB}}^{\max} = 9.26 \log(0.00437 fs/2 + 1)$. The T-F mixture coefficients are obtained as in (3.13) except that $H_f(\tau) = w_f(\tau) e^{-i2\pi f_{\text{Hz}} \tau / f_s}$ where the length of the window $w_f(\tau)$ now depends on f. More precisely, its length is defined such that the width of its main lope equals four times the difference between the central frequencies of adjacent filters [84].

3.2.1.2 Quadratic time-frequency transform

We propose to exploit a quadratic T-F representation where the mixture signal is described by its $I \times I$ local covariance matrix in each T-F bin

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n, f) = \mathbb{E}[\mathbf{x}(n, f) \mathbf{x}^H(n, f)] \tag{3.15}$$

where $\mathbb{E}[.]$ denotes empirical expectation. This representation aims to improve the robustness of parameter estimation for two reasons. Firstly, it locally exploits the observed data in several T-F bins instead of a single one as given by the linear representation in (2.8). Secondly, besides the interchannel phase and intensity differences encoded by $\mathbf{x}(n, f)$ in the linear T-F representation, it offers additional information about the correlation between the mixture channels which decreases when the number of active sources or the angle between these sources increase. This additional information results in improved separation of instantaneous mixtures [110, 68]. In the thesis, we show that this approach also improves separation performance on reverberant mixtures, despite the fact that interchannel correlation is intrinsically lower in this context.

We consider two different quadratic representations corresponding to the two distinct frequency scales above:

Q1. $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n, f)$ is computed by local averaging over the neighborhood of each T-F bin of the linear STFT transform as in (2.9). Fig. 3.4 gives an example of local averaging over 5 neighboring T-F bins where w_{nf} has a size of 3×3. This computation was introduced in [68] in the context of instantaneous audio source separation as a sliding window variant of the T-F patch-based model in [72], but has not yet been exploited in the context of convolutive source separation.

Q2. $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n, f)$ is computed by local averaging over the neighborhood of each T-F bin of the linear ERB transform as in (2.9) again.

ERB-scale representations have been used in a few studies [84, 111]. Provided that the number of bands is large enough, they provide finer spectral resolution than the STFT at low frequencies, hence decrease the overlap between sources in this

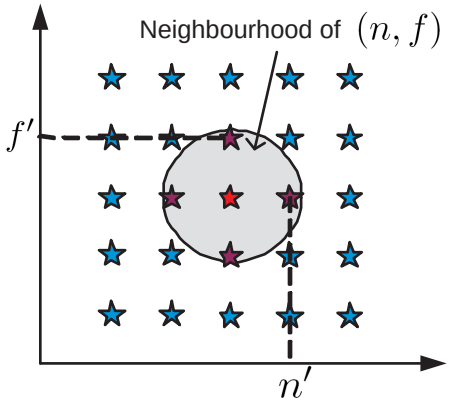


Figure 3.4: An example of quadratic representation taking into account 5 neighboring T-F bins.

crucial frequency range where most sound energy lies, while keeping good time resolution at higher frequencies. These representations were shown to improve separation performance based on ℓ_1-norm minimization for instantaneous mixtures [111] or single-channel Wiener filtering for convolutive mixtures [84]. Yet, they failed so far to improve separation performance based on multichannel Wiener filtering techniques for reverberant mixtures, due to the fact that the narrowband approximation does not hold at high frequencies because of the coarse spectral resolution at these frequencies [84]. We will show that ERB-scale representations are also beneficial for multichannel Wiener filtering-based separation provided that the proposed full-rank covariance parameterization is used.

3.2.2 Parameter estimation

Given the Gaussian modeling framework and the spatial covariance parameterization presented in Section 3.1.1, the core remaining problem is then **how to estimate the model parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ from the input signal.** Let us postpone the details of this core problem to Part 3 of the thesis. At this point we would like to point out that we have defined a family of EM algorithms for this purpose after suitable parameter initialization. The proposed general model parameter estimation architecture, which will be presented in Chapter 5, enables both ML and MAP estimation where spatial and spectral priors can be taken into account in the MAP case. Also, it accommodates both linear and quadratic input T-F representations thanks to the likelihood defined in (3.4) and the equivalence relation in (2.8).

3.2.3 Wiener filtering

Once the model parameters θ have been estimated, the source image T-F coefficients can be derived in the MMSE sense by multichannel Wiener filtering. This filter is applied to the mixture STFT coefficients as [71, 72]

$$\hat{\mathbf{c}}_j(n,f) := v_j(n,f)\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{3.16}$$

3.2.4 Inverse time-frequency transform

The final step is to reconstruct the time-domain source signals from their T-F representation. In the STFT domain, $\hat{\mathbf{c}}_j(t)$ is obtained from $\hat{\mathbf{c}}_j(n,f)$ via the inverse STFT with the overlap-add method. For the quadratic ERB representation, fullband signals $\hat{\mathbf{c}}_j(t)$ are recovered from subband signals $\hat{\mathbf{c}}_j(n,f)$ by upsampling followed by inverse ERB transform [84].

Chapter 4

Upper bound on source separation performance

This chapter focuses on evaluating the potential source separation performance achievable via the proposed framework. For that purpose, we first consider the simplest setting where all model parameters are known. This *oracle* setting provides an *upper bound* on the source separation performance achievable in a blind context. We then move toward a *semi-blind* setting where the spatial covariance matrices are known, but the source variances are blindly estimated from the observed mixture signal. Experimental results in both parameter estimation settings are reported by the end of the chapter and show the potential of the proposed approach compared to state-of-the-art and baseline approaches.

4.1 Oracle parameter estimation

We first present the estimation of the "true" spatial covariance matrices $\mathbf{R}_j(f)$ and the "true" source variance $v_j(n,f)$, considering the four parameterizations in Section 3.1.2, from the source image coefficients $\mathbf{c}_j(n,f)$ and the mixing filter coefficients $\mathbf{h}_j(f)$ which are assumed to be known. These estimates result in the oracle source separation algorithms shown in Fig. 4.1, which offer the highest separation performance achievable via each spatial covariance parameterization in the considered framework.

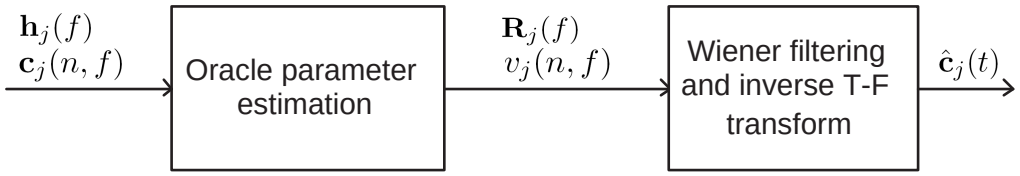


Figure 4.1: Oracle source separation workflow.

4.1.1 Oracle parameter estimation for the full-rank parameterizations

For the full-rank unconstrained parameterization, $\mathbf{R}_j(f)$ and $v_j(n,f)$ are estimated in the ML sense by maximizing the log-likelihood

$$\begin{aligned}\log \mathcal{L}_j &= \log \prod_{n,f} p\big(\mathbf{c}_j(n,f)|v_j,\mathbf{R}_j(f)\big) \\ &= \sum_{n,f} -\mathrm{tr}\Big(\frac{1}{v_j(n,f)}\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)\Big) - \log\det\big(\pi v_j(n,f)\mathbf{R}_j(f)\big)\end{aligned} \tag{4.1}$$

where $\widehat{\mathbf{\Sigma}}_j(n,f)$ is the empirical covariance of $\mathbf{c}_j(n,f)$, which is computed similarly to 2.8 or (2.9). By computing the derivatives of this expression with respect to $v_j(n,f)$ and each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain iterative updates for $v_j(n,f)$ and $\mathbf{R}_j(f)$ as (see Appendix A.2 for the detailed derivation)

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \tag{4.2}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f). \tag{4.3}$$

The parameters are initialized by $\mathbf{R}_j(f) = \frac{1}{N}\sum_{t=1}^{N}\widehat{\mathbf{\Sigma}}_j(n,f)$, $v_j(n,f) = 1$ for all n,f and convergence, as we observed in the experiments, is typically achieved in two or three iterations.

For the full-rank direct+diffuse parameterization, the spatial covariance $\mathbf{R}_j(f)$ is first computed by (3.9) instead assuming that the room characteristics and the geometry setting are known. Then $v_j(n,f)$ is computed by (4.2) given $\mathbf{R}_j(f)$.

4.1.2 Oracle parameter estimation for the rank-1 parameterizations

For the rank-1 spatial covariance parameterizations, the log-likelihood in (4.1) is undefined since $\mathbf{R}_j(f)$ is not invertible. The parameters are therefore computed directly from the mixing filter coefficients and the source image signal coefficients.

For the rank-1 convolutive parameterization, $\mathbf{R}_j(f)$ is computed by (3.5), and the source variances are derived by

$$v_j(n,f) = \frac{|\mathbf{h}_j^H(f)\mathbf{c}_j(n,f)|^2}{\|\mathbf{h}_j(f)\|_2^4}. \tag{4.4}$$

For the rank-1 anechoic spatial covariance parameterization, $\mathbf{R}_j(f)$ is computed by (3.7), and the source variances are derived by

$$v_j(n,f) = \frac{|\mathbf{h}_j^{\mathrm{ane}H}(f)\mathbf{c}_j(n,f)|^2}{\|\mathbf{h}_j^{\mathrm{ane}}(f)\|_2^4}. \tag{4.5}$$

4.1.3 Oracle parameter estimation for the baseline approaches

For binary masking and ℓ_1-norm minimization, the estimated source image T-F coefficients $\hat{\mathbf{c}}_j(n,f)$ are derived directly from the mixture T-F coefficients $\mathbf{x}(n,f)$, the mixing matrix $\mathbf{H}(f)$, and the true source images $\mathbf{c}_j(n,f)$ as in [112]. For each possible set of active source indexes j or k, we compute $\hat{\mathbf{c}}_j(n,f)$ by (2.20) for binary masking or by (2.22) for ℓ_1-norm minimization. We then choose the set of active source indexes that minimizes $\sum_j \|\hat{\mathbf{c}}_j(n,f) - \mathbf{c}_j(n,f)\|^2$ in each T-F bin. Matlab implementations of these two oracle estimators are available[4].

4.2 Semi-blind parameter estimation

In a BSS context, we need to estimate the model parameters from the mixture signal only. Recent evaluations of state-of-the-art algorithms [13, 1] have shown that the estimation of the spatial parameters remains difficult for real-world reverberant mixtures, due in particular to the existence of multiple local maxima in the ML criterion and to the source permutation problem arising when the model parameters at different frequencies are assumed to be independent. Therefore, we also investigate the potential separation performance achievable via the proposed framework in a *semi-blind* context, where the spatial covariance matrices $\mathbf{R}_j(f)$ are known but the source variances $v_j(n,f)$ are blindly estimated from the observed mixture as shown in Fig. 4.2.

The oracle computation of the spatial covariance matrices $\mathbf{R}_j(f)$ is presented in Section 4.1. We are left with the estimation of the source variances $v_j(n,f)$ given the mixture signal coefficients $\mathbf{x}(n,f)$ and $\mathbf{R}_j(f)$.

[4] http://bass-db.gforge.inria.fr/bss_oracle/

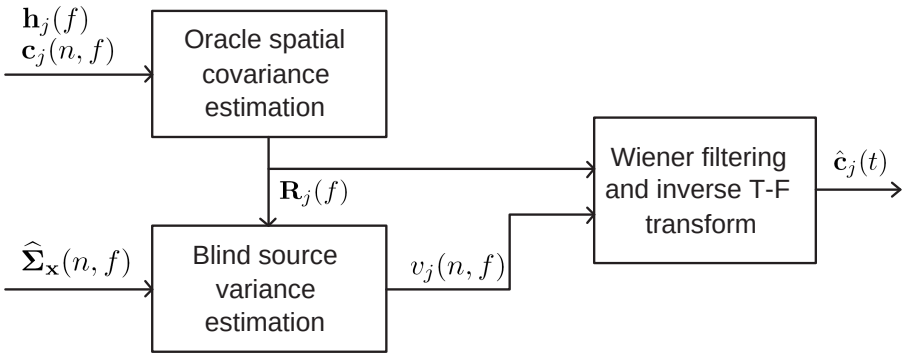


Figure 4.2: Semi-blind source separation workflow.

In our preliminary paper [28], we used a quasi-Newton algorithm for the estimation of $v_j(n,f)$ that converged in a very small number of iterations. However, due to the complexity of each iteration, we later found out that the EM algorithm, which is a popular choice for Gaussian models [71, 75, 66], provided a computational speedup despite a larger number of iterations. In order to be consistent with the blind parameter estimation in Part 3 where the EM algorithm is used, we use the same algorithm for the estimation of $v_j(n,f)$ in this semi-blind context. The detailed derivation of the EM algorithm will be presented in Chapter 5, where only the updates (5.19), (5.20) (in the E-step), and (4.2) (in the M-step) are used here.

Given the mixture T-F coefficients $\mathbf{x}(n,f)$ and the mixing matrix $\mathbf{H}(f)$, semi-blind source image estimations by binary masking and ℓ_1-norm minimization are presented in Section 2.3.1 and 2.3.2, respectively.

4.3 Experimental results

In order to evaluate the potential source separation performance achievable with each spatial covariance parameterization in an *oracle* context presented in Section 4.1, and in a *semi-blind* context presented in Section 4.2, we use the *simulated speech mixtures* dataset where the geometric setting and the room characteristics can be controlled. This dataset consists of 24 stereo mixtures of three speech sources generated with four different reverberation times $T_{60} = 50,\ 130,\ 250,\ 500$ ms and three different source-to-microphone distances $r = 50,\ 100,\ 150$ cm as described in detail in Section 8.1.1. We also compare the source separation performance with two baseline approaches, i.e. binary masking and ℓ_1-norm minimization described in Section 2.3, where the mixing matrix is known. The common parameter settings for the experiments are summarized

in Table 4.1 and the number of real-valued parameters to be estimated for each spatial covariance parameterization and baseline approaches is shown in Table 3.1.

Signal duration	10 s
Sampling rate	16 kHz
Number of channels	$I = 2$
STFT window type	sine window
STFT frame size	1024
STFT frame shift	512

Table 4.1: Experimental parameter settings for oracle and semi-blind separation.

We measure the source separation performance by the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), the signal-to-artifact ratio (SAR), and the source image-to-spatial distortion ratio (ISR), expressed in decibels (dB), detailed in Section 8.2.1, averaged over all male and female sources and all source-to-microphone distances. These energy ratio criteria account respectively for the overall distortion of the target source, residual crosstalk from other sources, musical noise, and spatial or filtering distortion of the target [11].

4.3.1 Oracle source separation performance

The oracle source separation performance achievable with the four spatial covariance parameterizations and the baseline approaches as function of the reverberation time is depicted in Fig. 4.3. As expected, the rank-1 convolutive parameterization performs best in terms of SDR and SAR for very low reverberation $T_{60} = 50$ ms where the narrowband assumption is satisfied and the mixing process is better modeled by the frequency dependent mixing vectors $\mathbf{h}_j(f)$ with fewer parameters than the full-rank $\mathbf{R}_j(f)$. In this case, the larger number of parameters in the full-rank unconstrained parameterization leads to overfitting.

When the reverberation time increases, the performance of the rank-1 convolutive parameterization drops below that of binary masking and ℓ_1-norm minimization. For $T_{60} \geq 130$ ms, the full-rank unconstrained spatial covariance parameterization outperforms all other parameterizations and binary masking according to all criteria and only results in slightly smaller SIR than ℓ_1-norm minimization. For instance, at $T_{60} = 250$ ms it improves the SDR by 6.8, 8.7, 3.3, 2.1, and 4.3 dB when compared to the rank-1 convolutive, rank-1 anechoic, full-rank direct+diffuse spatial covariance parameterizations, binary masking, and ℓ_1-norm minimization, respectively. This means that the full-rank unconstrained spatial covariances better fit the real-world reverberant mixing process.

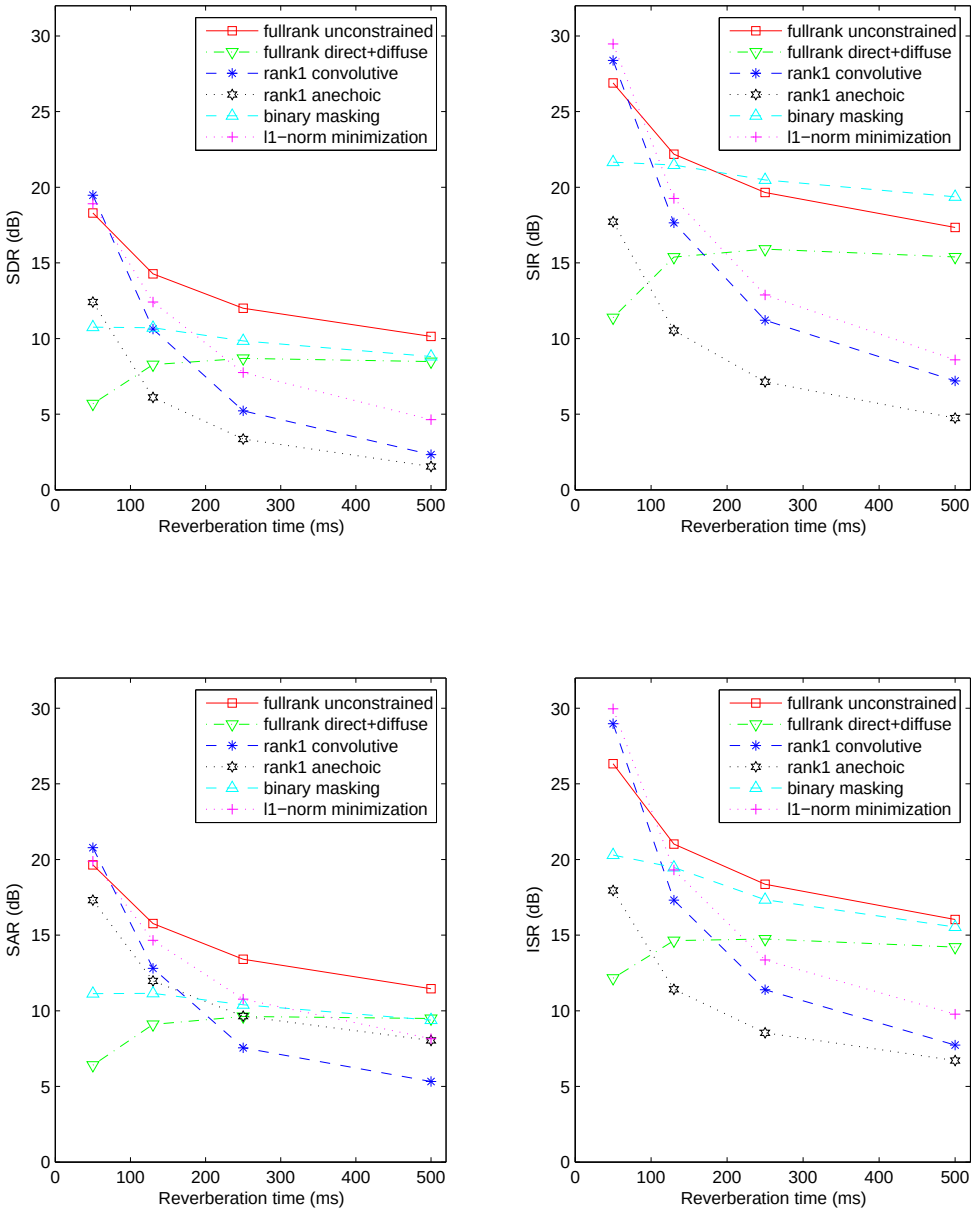


Figure 4.3: **Oracle** source separation performance achievable by the four spatial covariance parameterizations and the baseline approaches.

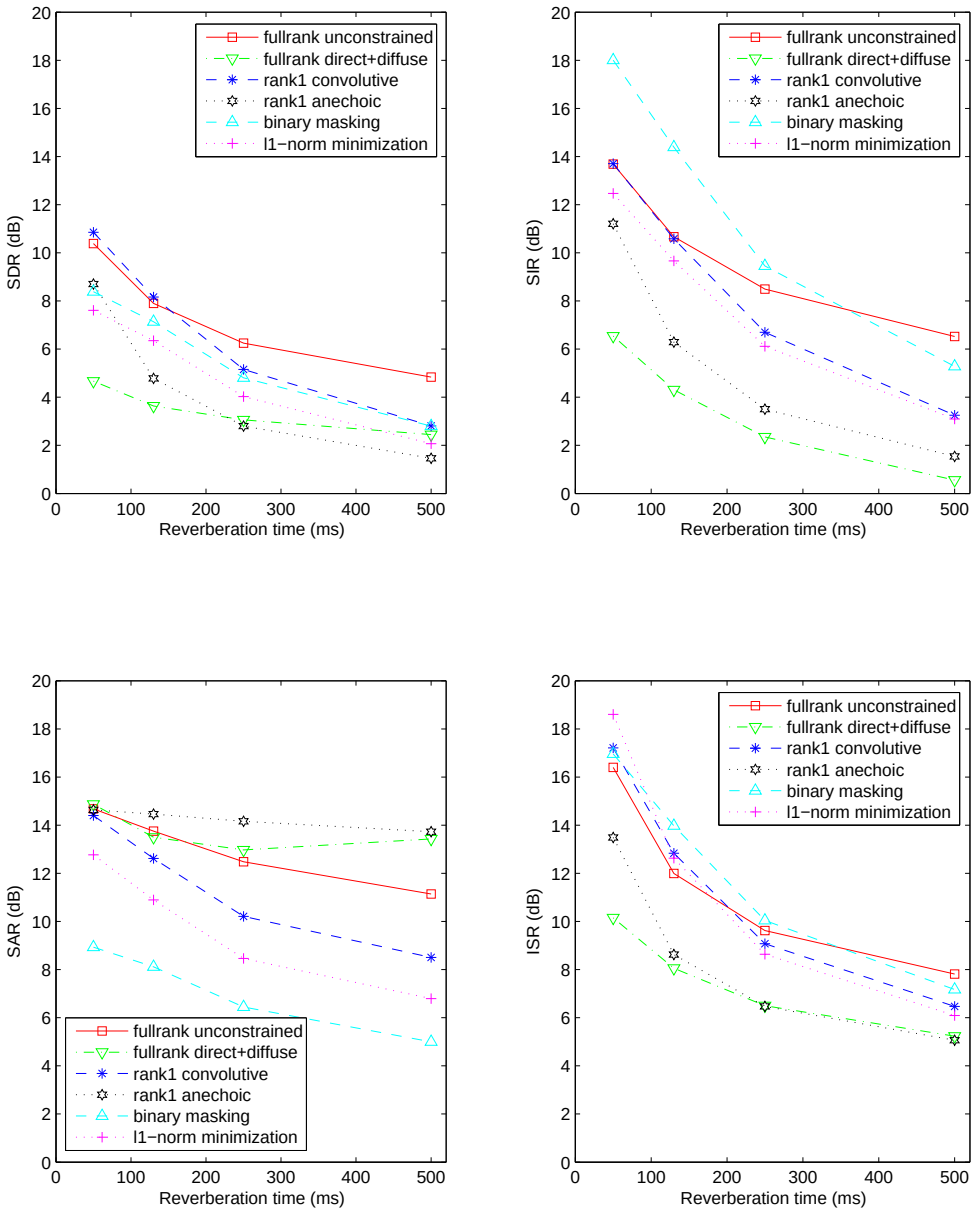


Figure 4.4: **Semi-blind** source separation performance achievable by the four spatial covariance parameterizations and the baseline approaches.

Also, the two constrained parameterizations perform more poorly in general. For moderate reverberation conditions $T_{60} \geq 130$ ms, the rank-1 anechoic parameterization results in lower SDR, SIR, and ISR than the three other spatial covariance parameterizations because it only accounts for the direct path. The full-rank direct+diffuse parameterization has lowest performance at $T_{60} = 50$ ms because the reverberation part is not diffuse so that the parameterization does not fit the mixing condition. It improves with increasing reverberation time but remains poorer than binary masking.

4.3.2 Semi-blind source separation performance

The semi-blind source separation performance achieved by the four spatial covariance parameterizations and the baseline approaches is shown in Fig. 4.4 as a function of the reverberation time. Obviously, performance has dropped compared to that obtained in the oracle context due to the blind estimation of $v_j(n, f)$.

The rank-1 anechoic parameterization offers very high SAR but low SDR and SIR. Similarly binary masking results in very high SIR but low SAR because it extracts only one dominant source per time-frequency bin, and results in strong musical noise. The SDR and the SIR given by the full-rank direct+diffuse spatial covariance parameterization are now below those obtained by the rank-1 convolutive parameterization for all reverberation conditions. Nevertheless, the full-rank unconstrained spatial covariance parameterization still offers the best SDR, which measures the overall distortion of the system, compared to all other approaches in real-world reverberation conditions when $T_{60} \geq 250$ ms. For instance, at $T_{60} = 500$ ms it improves the SDR by 2.0, 3.4, 2.3, 2.0, and 2.8 dB when compared to the rank-1 convolutive, rank-1 anechoic, full-rank direct+diffuse spatial covariance parameterizations, binary masking, and ℓ_1-norm minimization, respectively.

4.3.3 Robustness to small source movement

In a similar semi-blind source separation setting, we also investigated the robustness of the full-rank unconstrained spatial covariance parameterization to *small source movements* using real-world stereo recorded mixtures of three speech sources in a meeting room with 250 ms reverberation time. For the detail of this experiment and result, please refer to our paper [30]. As we observed, the separation performance of the rank-1 convolutive spatial covariance parameterization degraded more than that of the full-rank unconstrained parameterization when one of the source was slightly moved by 5° or 10° while its spatial covariance matrices were computed from its original posi-

tion. This result can be explained when considering the fact that the full-rank spatial covariance parameterization accounts for the spatial spread of each source as well as its spatial direction. Therefore, small source movements remaining in the range of the spatial spread do not affect much separation performance. This result indicates that the proposed parameterization could also offer a promising approach to the separation of moving sources due to its greater robustness to parameter estimation errors.

4.3.4 Source separation performance with different input time-frequency representations

As an additional experiment, we investigate the effect of the choice of the input T-F representation presented in Section 3.2.1 on the source separation performance given by the proposed full-rank unconstrained spatial covariance parameterization. We consider the linear T-F representation by the STFT (denoted linear STFT) where the input empirical mixture covariance $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ is computed by (2.8), the quadratic T-F representation by the STFT (denoted quadratic STFT) where $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ is computed by (2.9) with a window w_{nf} of size 3×3, and the quadratic T-F representation by the ERB transform (denoted quadratic ERB) where $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ is computed by (2.9) with 250 frequency bins. The ERB frame size is 1 and the ERB window is Hanning. Only the SDR achieved by both oracle and semi-blind parameter estimation is drawn in Fig. 4.5 since the SIR, the SAR and the ISR follow similar curves.

As expected, the quadratic ERB offers the highest source separation performance in both oracle and semi-blind parameter estimation context because it provides finer spectral resolution than the STFT at low frequencies where most sound energy lies. However, the quadratic STFT results in lower source separation performance than the linear STFT. This can be explained by the fact that the spatial covariance matrices are estimated from known source images in this experiment so that the estimates computed from a single T-F point are more accurate. We shall see in [31] and Section 8.3.1 that the quadratic STFT becomes better than the linear STFT in a blind estimation context.

4.4 Summary of Part 2

In this part, we first presented the proposed general Gaussian modeling framework for audio source separation targeting real-world reverberated or diffuse sources where the spatial position of the sources is encoded by frequency-dependent spatial covariance matrices and the time-frequency power of the sources is encoded by time-varying spectral

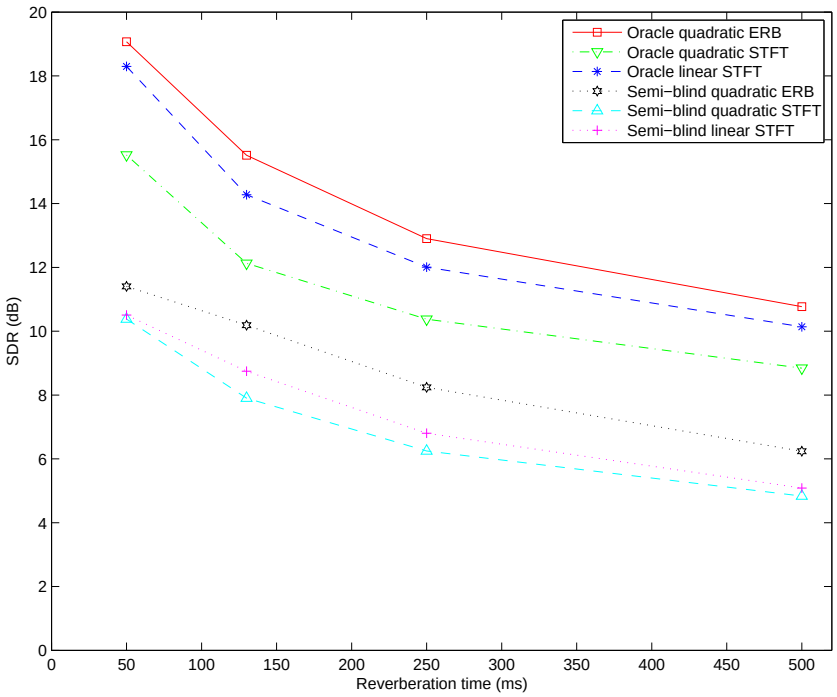


Figure 4.5: SDR achieved by the full-rank unconstrained parameterization in the oracle and the semi-blind context with different input time-frequency representations.

variances. The reasons for which we consider the Gaussian modeling framework are that in physical acoustics the reverberation is generally modeled as random Gaussian and that it is easy to handle computation as shown in the parameter estimation algorithm. We then introduced four spatial covariance parameterizations with different degrees of flexibility resulting in either rank-1 or full-rank matrices. Given the modeling framework and the parameterizations, we later designed a general source separation architecture which includes four major steps as presented in Section 3.2 and accepts either linear or quadratic time-frequency input representations.

In order to demonstrate the potential of the proposed framework, we computed the upper bound on the source separation performance achievable by the four spatial covariance parameterizations and compared the results with that given by baseline approaches. The overall results presented in Section 4.3 suggested that the rank-1 anechoic and the full-rank direct+diffuse spatial covariance parameterization provide poorer separation than the state-of-the-art rank-1 convolutive and the newly proposed full-rank

unconstrained spatial covariance parameterization. Hence, for the rest of the thesis, we will focus on designing source separation algorithms where the spatial covariance matrices are full-rank and unconstrained.

Part III

Estimation of model parameters and source separation

Chapter 5

General estimation architecture and ML algorithm

Chapter 4 has revealed that the proposed modeling framework and parameterization result in a higher *upper bound* on the source separation performance than both baseline approaches and state-of-the-art spatial cue parameterization. However, in order to perform actual *blind* source separation, the model parameters must be estimated from the mixture signals only. Hence, this chapter will present a general parameter estimation architecture which allows both ML and MAP estimation by the EM algorithm as briefly discussed in Section 3.2.2. For that purpose, we will first describe the general estimation workflow followed by ML and MAP criteria. We then provide details of the ML algorithm including two major steps: parameter initialization by hierarchical clustering and permutation alignment, and parameter estimation by EM updates. Note that the parameter estimation architecture and the EM algorithms derived in this chapter are applicable to either input representation, *i.e.* to the linear T-F representation or to the quadratic T-F representation presented in Section 3.2.1.

5.1 General parameter estimation architecture

Given the modeling framework and a choice of parameterization, the *blind* estimation of the model parameters from the mixture signal plays a central role for realistic BSS application. We propose a general and flexible parameter estimation architecture described in Fig. 5.1:

- this architecture mainly relies on iterative parameter updates via the EM algorithm, which is well-known as an appropriate choice for Gaussian models [113, 71],

- as any iterative optimization algorithm, EM is sensitive to the initialization [25] so that a suitable parameter initialization scheme is necessary,

- when some additional information about either the original source signals or the mixing condition is known in advance, priors can be designed and incorporated in the probabilistic model so as to exploit this knowledge and help improve the source separation performance by estimating the parameters in the MAP sense.

The general model parameter estimation criteria are:

- the ML criterion
$$\hat{\theta} = \arg\max_{\theta} \prod_{n,f} p\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)|\theta\big) \tag{5.1}$$

- the MAP criterion
$$\hat{\theta} = \arg\max_{\theta} \prod_{n,f} p\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)|\theta\big)p(\theta) \tag{5.2}$$

Though the general parameter estimation architecture is applicable to both ML and MAP approaches, in the rest of this chapter we consider ML parameter estimation only and postpone MAP estimation to Chapters 6 and 7. Besides, we will only consider the ML parameter estimation for the proposed full-rank unconstrained and the state-of-the-art rank-1 convolutive spatial covariance parameterizations since they were confirmed to provide a higher upper bound on the source separation performance compared to the other parameterizations in Chapter 4. The initialization scheme will be presented in Section 5.2 and the derivation of EM algorithm for the rank-1 convolutive and the full-rank unconstrained parameterizations will be presented in Section 5.3.

5.2 Parameter initialization

Preliminary experiments showed that the initialization of the model parameters greatly affects the separation performance resulting from the EM algorithm. Therefore, we carefully initialize the spatial parameters $\mathbf{h}_j^{\text{init}}(f)$ and $\mathbf{R}_j^{\text{init}}(f)$ via two steps: hierarchical clustering and permutation alignment as depicted in Fig. 5.2 and detailed below. In order to ensure a fair comparison of both input T-F representations independently of the parameter initialization procedure, the initial parameters in the quadratic ERB domain are derived from those in the STFT domain within a third step also detailed below.

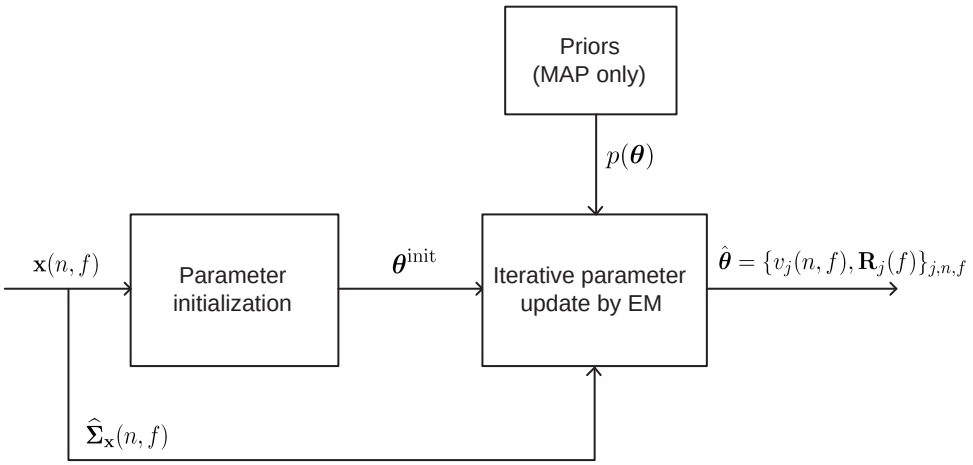


Figure 5.1: General blind parameter estimation workflow.

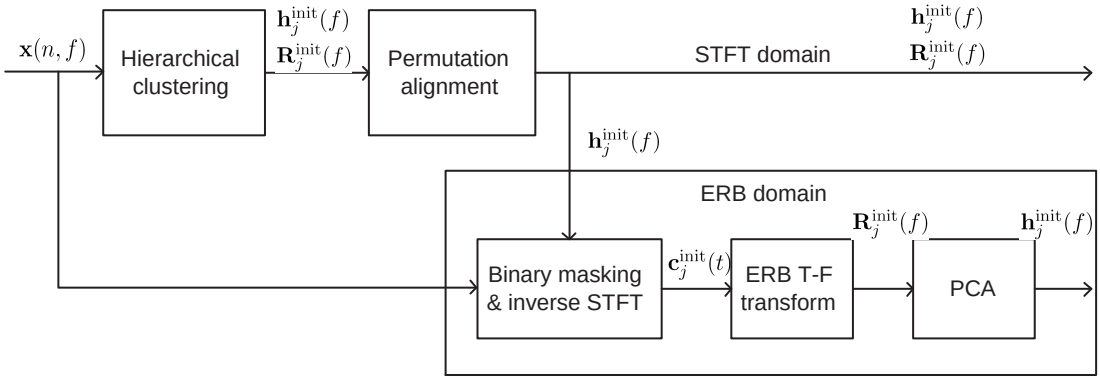


Figure 5.2: Parameter initialization workflow.

5.2.1 Hierarchical clustering

The parameter initialization schemes previously proposed for rank-1 Gaussian models are either restricted to instantaneous mixtures [75] or require some prior knowledge about the sources [28, 25]. By contrast, a number of clustering algorithms have been proposed for blind estimation of the mixing vectors in the context of sparsity-based convolutive source separation. In the following, we use up to minor improvements the hierarchical clustering-based algorithm in [55] for the initialization of rank-1 spatial covariance matrices and introduce a modified version of this algorithm for the initialization of full-rank unconstrained spatial covariance matrices.

The algorithm in [55] relies on the assumptions that at each frequency f the sounds of all sources come from disjoint regions of space and that a single source predominates

in most time-frequency bins. The vectors $\mathbf{x}(n,f)$ of mixture STFT coefficients then cluster around the direction of the associated mixing vector $\mathbf{h}_j(f)$ in the time frames n where the jth source is predominant. It is well known that the validity of this sparsity assumption decreases with increasing reverberation. Nevertheless, this algorithm was explicitly developed for reverberant mixtures.

In order to estimate these clusters, the vectors of mixture STFT coefficients are first normalized as

$$\bar{\mathbf{x}}(n,f) \leftarrow \frac{e^{-i\arg(x_1(n,f))}}{\|\mathbf{x}(n,f)\|_2}\mathbf{x}(n,f) \tag{5.3}$$

where arg(.) denotes the phase of a complex number. We then define the distance between two clusters C_1 and C_2 as the average distance between the associated normalized mixture STFT coefficients

$$d(C_1,C_2) = \frac{1}{|C_1||C_2|}\sum_{\bar{\mathbf{x}}_{c_1}\in C_1}\sum_{\bar{\mathbf{x}}_{c_2}\in C_2}\|\bar{\mathbf{x}}_{c_1}-\bar{\mathbf{x}}_{c_2}\|_2. \tag{5.4}$$

In a given frequency bin f, each normalized vector of mixture STFT coefficients $\bar{\mathbf{x}}(n,f)$ at a time frame n is first considered as a cluster containing a single item. The distance between each pair of clusters is computed and the two clusters with the smallest distance are merged. This bottom-up process called linking is repeated until the number of clusters is smaller than a predetermined threshold K. This threshold is usually much larger than the number of sources J, so as to eliminate outliers [55]. We finally choose the J clusters with the largest number of samples and compute the initial mixing vector and spatial covariance matrix, up to the permutation ambiguity, for each source as

$$\mathbf{h}_j^{\text{init}}(f) = \frac{1}{|C_j|}\sum_{\bar{\mathbf{x}}(n,f)\in C_j}\tilde{\mathbf{x}}(n,f) \tag{5.5}$$

$$\mathbf{R}_j^{\text{init}}(f) = \frac{1}{|C_j|}\sum_{\bar{\mathbf{x}}(n,f)\in C_j}\tilde{\mathbf{x}}(n,f)\tilde{\mathbf{x}}(n,f)^H \tag{5.6}$$

where $\tilde{\mathbf{x}}(n,f) = e^{-i\arg(x_1(n,f))}\mathbf{x}(n,f)$, and $|C_j|$ denotes the total number of samples in cluster C_j, which depends on the considered frequency bin f.

Note that, compared to the algorithm in [55], we propose some minor modifications:

- the distance between clusters is defined as the average distance between the normalized mixture STFT coefficients instead of the minimum distance between them, and
- the mixing vector $\mathbf{h}_j^{\text{init}}(f)$ is computed from the phase-normalized mixture STFT

coefficients $\tilde{\mathbf{x}}(n,f)$ instead of the phase- and amplitude-normalized coefficients $\bar{\mathbf{x}}(n,f)$. This increases the weight of time-frequency bins of large amplitude where the modeled source is more likely to be prominent, in a way similar to [18].

These modifications were found to provide better initial approximation of the mixing parameters in our experiments. We also tested random initialization and DOA-based initialization, *i.e.* where the initial mixing vectors $\mathbf{h}_j^{\text{init}}(f)$ are equal to $\mathbf{h}_j^{\text{ane}}(f)$ assuming known source and microphone positions. Both schemes were found to result in slower convergence and poorer separation performance than the chosen scheme.

5.2.2 Permutation alignment

Since the spatial parameters of the rank-1 convolutive and the full-rank unconstrained parameterization, *i.e.* $\mathbf{h}_j^{\text{init}}(f)$ and $\mathbf{R}_j^{\text{init}}(f)$, are estimated independently in each frequency bin f, they should be ordered so as to correspond to the same source across all frequency bins. This so-called *permutation problem* has been widely studied in the context of sparsity-based source separation. In the following, we apply the DoA -based algorithm in [51] to the estimated mixing vector $\mathbf{h}_j^{\text{init}}(f)$ and align $\mathbf{R}_j^{\text{init}}(f)$ in the same order.

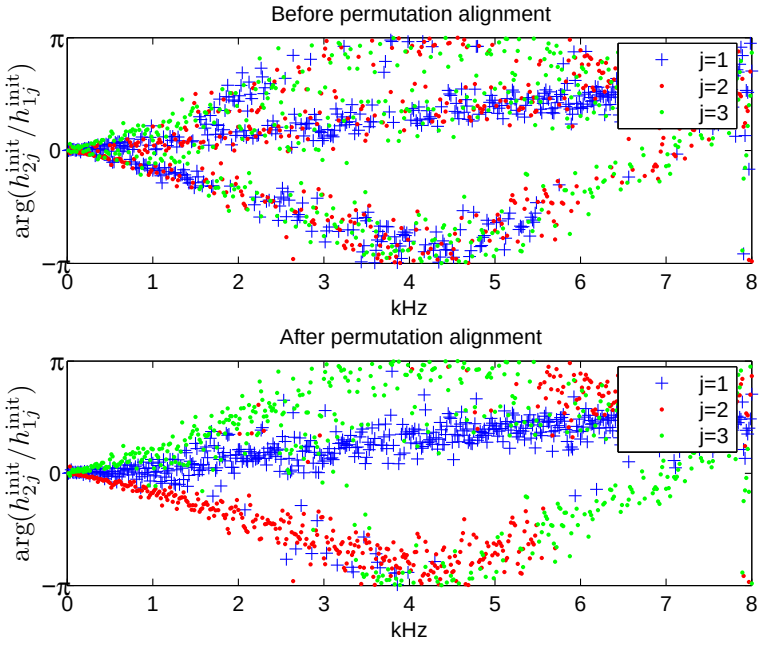


Figure 5.3: Argument of $h_{2j}^{\text{init}}(f)/h_{1j}^{\text{init}}(f)$ before and after permutation alignment for a synthetic stereo mixture of three sources with $T_{60} = 130$ ms.

The principle of this algorithm is as follows. Given the geometry of the microphone array, a critical frequency is determined above which spatial aliasing may occur. The mixing vectors $\mathbf{h}_j^{\text{init}}(f)$ are each unambiguously related to a certain DoA below that frequency, while phase wrapping may occur at higher frequencies. The algorithm first estimates the source DoAs and the permutations at low frequencies by clustering the mixing vectors after suitable normalization assuming no phase wrapping and then re-estimates them at all frequencies by taking phase wrapping into account. Then the order of the spatial covariance matrices $\mathbf{R}_j^{\text{init}}(f)$ in each frequency bin is permuted identically to that of the mixing vectors $\mathbf{h}_j^{\text{init}}(f)$. As a result, the mixing vectors and the spatial covariance matrices are aligned and initialized to the same source order after solving the permutation problem. Fig. 5.3 depicts an example of the phase difference between the two entries of the mixing vectors before and after solving the permutation for a stereo synthetic mixture of three speech sources with room reverberation time $T_{60} = 130$ ms. The critical frequency below which phase is unambiguously related to the source DoAs is here equal to 5 kHz [51]. The source order appears globally aligned for most frequency bins after solving the permutation.

5.2.3 Parameter computation in ERB T-F representation

Further computation is required for the ERB T-F representation. While a similar two-step procedure, *i.e.* hierarchical clustering and permutation alignment, could be applied to the first principal components of the empirical mixture covariances $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n, f)$, we found that this did not result in good performance in the high frequency range. Indeed, due to the broadness of high frequency subbands, the reduction of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n, f)$ to their first principal components results in some information loss.

Given the estimated mixing vectors $\mathbf{h}_j^{\text{init}}(f)$, initial estimates of the source images in the STFT domain $\mathbf{c}_j^{\text{init}}(n, f)$ are obtained by binary masking as described in Section 2.3.1. The empirical source covariances $\widehat{\mathbf{\Sigma}}_j^{\text{init}}(n, f)$ are then derived by STFT inversion of $\mathbf{c}_j^{\text{init}}(n, f)$ followed by ERB-scale representation. Finally the spatial covariance matrices $\mathbf{R}_j^{\text{init}}(f)$ are initialized by averaging $\widehat{\mathbf{\Sigma}}_j^{\text{init}}(n, f)$ over all time frames n, while the initial mixing vectors $\mathbf{h}_j^{\text{init}}(f)$ are taken as the first principal component of $\mathbf{R}_j^{\text{init}}(f)$.

5.2.4 Variance initialization

Given this suitable initialization of the spatial parameters, the initialization of the source variances does not significantly affect the convergence of the EM algorithm and the final separation performance. In our experiments, the source variances are simply initialized

to $v_j^{\text{init}}(n,f) = 1$. We have found that this basic initialization scheme provided similar performance to the more advanced but slower scheme consisting of finding the $v_j^{\text{init}}(n,f)$ most consistent with $\mathbf{h}_j^{\text{init}}(f)$ and $\mathbf{R}_j^{\text{init}}(f)$ by running EM without updating the mixing vectors or the spatial covariance matrices.

5.3 ML estimation by the EM algorithm

We first present the derivation of EM algorithm for the rank-1 convolutive parameterization. This algorithm is inspired from the study in [25], which relies on the same parameterization of the spatial covariance matrices but on a distinct model of the source variances and was restricted to a linear input T-F transform. We here express the updates in terms of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ instead of $\mathbf{x}(n,f)$, so that they become applicable to quadratic T-F transforms, and extent them to the proposed full-rank unconstrained parameterization.

5.3.1 EM updates for the rank-1 convolutive parameterization

Similarly to [25], EM cannot be directly applied to the noiseless mixture model, *i.e.* $\mathbf{x}(n,f) = \mathbf{H}(f)\mathbf{s}(n,f)$, since the estimated mixing vectors would remain fixed to their initial value under the resulting updates (see [25] for a theoretical explanation). This issue can be addressed by considering the noisy mixture model

$$\mathbf{x}(n,f) = \mathbf{H}(f)\mathbf{s}(n,f) + \mathbf{b}(n,f) \tag{5.7}$$

where $\mathbf{b}(n,f)$ denotes some additive zero-mean Gaussian noise. Following [25], we assume that $\mathbf{b}(n,f)$ is stationary and spatially uncorrelated and denote by $\mathbf{\Sigma}_{\mathbf{b}}(f) = \sigma_b^2\mathbf{I}$ its $I \times I$ time-invariant diagonal covariance matrix. This matrix is initialized to a small value related to the average empirical channel variance as discussed in [25].

EM is separately derived for each frequency bin f for the *complete data* $\{\mathbf{x}(n,f), \mathbf{s}(n,f)\}_n$ that is the set of observed mixture T-F coefficients and hidden source T-F coefficients of all time frames (see Appendix A.1 for the detailed description). The resulting parameter updates are summarized in Algorithm 5.1. In the E-step, the Wiener filter $\mathbf{W}(n,f)$ and the expectation of the natural statistics $\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)$ of the sources are computed. Then, in the M-step, the source variances, the mixing matrix and the noise covariance are updated.

Algorithm 5.1 EM algorithm for the rank-1 convolutive parameterization

E-step:

$$\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) = \text{diag}(v_1(n,f),...,v_J(n,f)) \tag{5.8}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \mathbf{H}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f) + \boldsymbol{\Sigma}_{\mathbf{b}}(f) \tag{5.9}$$

$$\mathbf{W}(n,f) = \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{5.10}$$

$$\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f) = \mathbf{W}(n,f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) + (\mathbf{I} - \mathbf{W}(n,f)\mathbf{H}(f))\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) \tag{5.11}$$

M-step:

$$v_j(n,f) = \widehat{\mathbf{R}}_{\mathbf{ss}\,jj}(n,f) \tag{5.12}$$

$$\mathbf{H}(f) = \widehat{\mathbf{R}}_{\mathbf{xs}}(f)\widehat{\mathbf{R}}_{\mathbf{ss}}^{-1}(f) \tag{5.13}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{b}}(f) =& \text{Diag}(\widehat{\mathbf{R}}_{\mathbf{xx}}(f) - \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{xs}}^H(f) - \widehat{\mathbf{R}}_{\mathbf{xs}}(f)\mathbf{H}^H(f) \\ &+ \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{ss}}(f)\mathbf{H}^H(f))\end{aligned} \tag{5.14}$$

where Diag(.) projects a matrix onto its diagonal, and

$$\widehat{\mathbf{R}}_{\mathbf{ss}}(f) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f) \tag{5.15}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs}}(f) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) \tag{5.16}$$

$$\widehat{\mathbf{R}}_{\mathbf{xx}}(f) = \frac{1}{N}\sum_{n=1}^{N}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f). \tag{5.17}$$

5.3.2 EM updates for the full-rank unconstrained parameterization

The derivation of EM for the full-rank unconstrained parameterization is easier than for the rank-1 convolutive parameterization since we can stick to the exact mixture model (1.1). This can be seen as an advantage of full-rank *vs.* rank-1 parameterization. EM is again separately derived for each frequency bin f. Since the mixture can be recovered from the spatial images of all sources, the *complete data* reduces to $\{\mathbf{c}_j(n,f)\}_{n,f}$, that is the set of hidden T-F coefficients of the spatial images of all sources on all time frames.

The resulting parameter updates are summarized in Algorithm 5.2, while the details of the derivation are presented in Appendix A.2. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the covariance $\widehat{\mathbf{\Sigma}}_j(n,f)$ of the spatial image of the jth source are computed similarly as for the rank-1 convolutive parameterization. Then in the M-step, $v_j(n,f)$ and $\mathbf{R}_j(f)$ are updated as in (4.2) and (4.3).

Algorithm 5.2 EM algorithm for the full-rank unconstrained parameterization

E step:

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f) \tag{5.18}$$

$$\mathbf{W}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{5.19}$$

$$\widehat{\mathbf{\Sigma}}_j(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{\Sigma}_j(n,f) \tag{5.20}$$

M step:

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \tag{5.21}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f). \tag{5.22}$$

5.3.3 Computational complexity

Each iteration of the EM update for the full-rank unconstrained parameterization mostly involves the computation of $(N+1)FJ$ inversions and $5NFJ$ multiplications of $I \times I$ matrices. The overall computational complexity of one iteration is therefore $O(6NFJI^3)$. It is linear as a function of the number of sources and the duration of the signal, and cubic as a function of the number of channels.

Chapter 6

MAP algorithms exploiting spatial priors

Following the general source separation architecture presented in Section 5.1 where the model parameters can be estimated in either the ML or the MAP sense, this chapter is devoted to MAP parameter estimation when some extra knowledge about the spatial source position is available in order to enhance the source separation performance. We first consider a family of spatial *location* priors in Section 6.1 where the theory of statistical room acoustics is exploited in particular scenarios where the geometric setting, *i.e.* the microphone spacing and the source-to-microphone distances, and the room reverberation time are known. **Let us also emphasize that this resulting MAP algorithms do not suffer from the troublesome permutation problem thanks to the prior location information of the sources.** We then consider a spatial *continuity* prior in Section 6.2 where the spatial covariance matrices are time-varying but assumed to vary smoothly over time. This continuity prior could be potentially employed to model *moving sources* which goes beyond the scope of this thesis. For each spatial prior, we derive the MAP spatial parameter estimation updates by EM where the E-step is similar to that of the ML algorithm and the M-step is modified.

6.1 MAP algorithms exploiting spatial location priors

While BSS requires to recover the source signals from a given mixture without any other knowledge, in many practical situations the geometric setting and the room acoustical characteristics can be known in advance. Such situations can happen, for instance, in a

formal meeting where the position of each delegate is fixed or in a car where the positions of the driver and the passengers are fixed. Therefore, in this section we propose and investigate two different acoustically-motivated spatial priors which exploit the known geometric setting. These priors all rely on the theory of statistical room acoustics in order to express the mean and/or the variance of the prior as a function of the geometric setting and the room characteristics. The resulting MAP algorithms offer an acoustically principled solution to the estimation of the model parameters and to the permutation problem. Most importantly, they provide a proof of concept of the benefit of the proposed priors towards their future use in a BSS context.

The choices of two prior distributions in the following sections, *i.e* Wishart and inverse-Wishart, are due in particular to the *engineering* constraints such that they can:

- applies to Hermitian matrices
- has a closed-form expression for the mean
- results in close-form updates.

6.1.1 MAP algorithm exploiting a Wishart spatial location prior

6.1.1.1 Wishart spatial location prior

According to the theory of statistical room acoustics [27, 102], for a given microphone spacing and source position relative to the microphones, the mean spatial covariance matrix over all possible microphone positions is given by (3.9). However, our preliminary experiments have confirmed that the actual value of $\mathbf{R}_j(f)$ varies depending on the microphone positions and cannot be set to the fixed value (3.9). Therefore, we first investigate the modeling of $\mathbf{R}_j(f)$ as

$$p(\mathbf{R}_j(f)) = \mathcal{W}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \tag{6.1}$$

where

$$\mathcal{W}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^{-m}|\mathbf{R}|^{(m-I)}e^{-\mathrm{tr}(\mathbf{\Psi}^{-1}\mathbf{R})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \tag{6.2}$$

is the Wishart density over a Hermitian positive definite matrix $\mathbf{R}$ with positive definite scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $m\mathbf{\Psi}$ [114]. This distribution, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. We define

the scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = \frac{\mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}(f)\big)^H + \sigma_{\text{rev}}^2 \mathbf{\Omega}(f)}{m} \tag{6.3}$$

so that the mean of $\mathbf{R}_j(f)$ is coherent with (3.9). This model extends the full-rank direct+diffuse spatial covariance parameterization by allowing deviations of the spatial covariance matrices around their mean controlled by the number of degrees of freedom m. Such deviations occur since the full-rank dirrect+diffuse model is only valid on average when considering a large number of sources as shown in Fig. Fig. 3.1, or for instance when the source or the microphones are close to the walls, resulting in a strong directional early echo.

6.1.1.2 Learning the hyper-parameter

In order to obtain the best fit between the prior distribution and the actual spatial covariance matrices, we learn the number of degrees of freedom m of the Wishart prior (6.9) from training data. For a given geometric setting and reverberation time, several training signals are generated by convolving training source signals with mixing filters generated for many source and microphone positions p. The corresponding spatial covariance matrices $\mathbf{R}_p(f)$ are then estimated via the oracle estimator in Section 4.1.1. Since $\mathbf{R}_p(f)$ can be measured only up to an arbitrary scaling factor α, the number of degrees of freedom m may be estimated in the ML sense by maximizing

$$\mathcal{L}_W = \prod_p \prod_f \int_0^\infty p\big(\mathbf{R}_p(f)|\alpha, \mathbf{\Psi}_p(f), m\big) p(\alpha) d\alpha \tag{6.4}$$

where $p\big(\mathbf{R}_p(f)|\alpha, \mathbf{\Psi}_p(f), m\big) = J_\alpha \mathcal{W}\big(\alpha \mathbf{R}_p(f)|\mathbf{\Psi}_p(f), m\big)$, $J_\alpha = \alpha^{I^2}$ is the Jacobian of the scaling transform, $p(\alpha) = 1/\alpha$ is the non-informative Jeffreys prior, and $\mathbf{\Psi}_p(f)$ was computed by (6.3) for each geometry setting p. By computing the integral and discarding constants, the log-likelihood to be maximized is

$$\begin{aligned} \log \mathcal{L}_W \stackrel{\text{c}}{=} \sum_{p,f} & -m \log |\mathbf{\Psi}_p(f)| + (m - I) \log |\mathbf{R}_p(f)| - I\ m\ \text{tr}\big(\mathbf{\Psi}_p^{-1}(f) \mathbf{R}_p(f)\big) \\ & + \log \Gamma(I\ m) - \sum_{i=1}^{I} \log \Gamma(m - i + 1). \end{aligned} \tag{6.5}$$

Given $\mathbf{\Psi}_p(f)$ and $\mathbf{R}_p(f)$ for all p, f, $\log \mathcal{L}_W$ is then maximized using Matlab's fmincon Newton-based optimizer. As a result, the optimal value of m is found, which increases with the reverberation time as will be shown in Table 8.5. However, we observed in our experiment that the value of m does not greatly affect the separation result.

6.1.1.3 MAP spatial parameter update

Given the prior hyper-parameters $\mathbf{\Psi}_j(f)$ and m, we now estimate the model parameters θ in the MAP sense. We consider an EM algorithm where the *complete data* is chosen as in the ML algorithm as $\{\mathbf{c}_j(n,f)\ \forall j,n,f\}$. In the E-step, the Wiener filter and the expected covariance matrices $\widehat{\mathbf{\Sigma}}_j(n,f)$ are computed as in the ML algorithm (see Algorithm 5.2).

In the M-step, the expectation of the log-posterior of the complete data

$$Q_{\mathcal{W}}(\theta|\theta^{\text{old}}) = \sum_{j,f}\Big(\sum_{n}\log p\big(\mathbf{c}_j(n,f)|\mathbf{0},\mathbf{\Sigma}_j(n,f)\big) + \gamma \log \mathcal{W}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f),m\big)\Big) \quad (6.6)$$

is maximized with respect to the parameters, where γ is a tradeoff hyper-parameter determining the strength of the prior. By computing the partial derivatives of $Q_{\mathcal{W}}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_j(n,f)$ and equating them to zero, we obtain a quadratic matric equation. By solving this equation (see the detailed description in Appendix A.3), we obtain the spatial covariance update as

$$\mathbf{R}_j(f) = \frac{1}{2}\mathbf{A}^{-1/2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}\big)\mathbf{A}^{-1/2} \quad (6.7)$$

where $(.)^{1/2}$ denotes the square root of a Hermitian matrix, and

$$\begin{aligned}\mathbf{A} &= \gamma\mathbf{\Psi}_j^{-1}(f)\\ b &= -\gamma(m-I)+N\\ \mathbf{C} &= \sum_{n=1}^{N} -\frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)}.\end{aligned} \quad (6.8)$$

The update of $v_j(n,f)$ is not affected by the prior.

6.1.2 MAP algorithm exploiting an inverse-Wishart spatial location prior

6.1.2.1 Inverse-Wishart spatial location prior

We consider, as an alternative to the Wishart prior distribution, an inverse-Wishart distribution over each spatial covariance matrix $\mathbf{R}_j(f)$ since it is the *conjugate prior* for the likelihood of the considered Gaussian observation model, so that it results in simple closed-form parameter updates as shown below. We propose to model $\mathbf{R}_j(f)$ as

$$p(\mathbf{R}_j(f)) = \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \tag{6.9}$$

where

$$\mathcal{IW}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^m |\mathbf{R}|^{-(m+I)} e^{-\mathrm{tr}(\mathbf{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(m-i+1)} \tag{6.10}$$

is the inverse Wishart density [114] over a Hermitian positive definite matrix $\mathbf{R}$ with positive definite inverse scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $\mathbf{\Psi}/(m-I)$ [114], with Γ the gamma function. This distribution, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. Similarly to the Wishart prior, we define the inverse scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = (m-I)\Big(\mathbf{h}_j^{\mathrm{ane}}(f)\big(\mathbf{h}_j^{\mathrm{ane}}(f)\big)^H + \sigma_{\mathrm{rev}}^2 \mathbf{\Omega}(f)\Big) \tag{6.11}$$

so that the mean of $\mathbf{R}_j(f)$ is coherent with (3.9).

6.1.2.2 Learning the hyper-parameter

Similarly to the Wishart prior, the number of degrees of freedom m, which determines the deviation of $\mathbf{R}_j(f)$ from its mean, is learned from training data by maximizing the log-likelihood (6.4), where $p\big(\mathbf{R}_p(f)|\alpha, \mathbf{\Psi}_p(f), m\big) = J_\alpha \mathcal{IW}\big(\alpha\mathbf{R}_p(f)|\mathbf{\Psi}_p(f), m\big)$. By computing the integral and discarding contants, the log-likelihood to be maximized is

$$\begin{aligned}\log \mathcal{L}_{IW} \stackrel{\mathrm{c}}{=} &\sum_{p,f} m \log|\mathbf{\Psi}_p(f)| - (m+I)\log|\mathbf{R}_p(f)| - I\ m\ \mathrm{tr}\big(\mathbf{\Psi}_p(f)\mathbf{R}_p^{-1}(f)\big) \\ &+ \log\Gamma(I\ m) - \sum_{i=1}^{I} \log\Gamma(m-i+1).\end{aligned} \tag{6.12}$$

Given $\mathbf{\Psi}_p(f)$ and $\mathbf{R}_p(f)$ for all p, f, $\log \mathcal{L}_{IW}$ is then maximized using Matlab's fmincon Newton-based optimizer. As a result, the optimal value of m is found, which increases with the reverberation time. It can be shown that this expression is equal to (6.4) in the particular case when $I = 2$. Therefore, in this case, the optimal value of m is equal for both priors.

6.1.2.3 MAP spatial parameter update

Given the hyper-parameters $\mathbf{\Psi}_j(f)$ and m, we estimate the model parameters in the MAP sense using the EM algorithm where the complete data and the E-step are the same as for the Wishart prior.

In the M-step, the expectation of the log-posterior of the complete data

$$Q_{\mathcal{IW}}(\theta|\theta^{\text{old}}) = \sum_{j,f} \Big(\sum_{n} \log p\big(\mathbf{c}_j(n,f)|\mathbf{0}, \mathbf{\Sigma}_j(n,f)\big) + \gamma \log \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \Big) \tag{6.13}$$

is maximized with respect to the parameters, where γ is a tradeoff hyper-parameter determining the strength of the prior, and $\mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big)$ is defined in (6.10). By computing the partial derivatives of $Q_{\mathcal{IW}}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain the update rule for $\mathbf{R}_j(f)$ by

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N} \Big(\gamma \mathbf{\Psi}_j(f) + \sum_{n=1}^{N} \frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)} \Big). \tag{6.14}$$

Note that when $\gamma = 0$, *i.e.* the contribution of the prior is excluded, (6.14) becomes equal to the ML update (5.22). $v_j(n,f)$ is updated by (5.21), which means that the spatial prior does not affect the update of source variances. This MAP parameter update does not significantly increase the computational time compared to the ML parameter updates.

6.2 MAP algorithm exploiting a spatial continuity prior

6.2.1 Motivation for a smoothly time-varying spatial covariance

Our proposed Gaussian modeling framework offers the most general parameterization of the source covariance in (3.1). However, we have only considered a particular case where the spatial covariance matrices are time-invariant so far. Indeed, when each spatial image $\mathbf{c}_j(t)$ corresponds to a point source with fixed spatial position, the spatial

covariance matrices are time-invariant and may be modeled by $\mathbf{R}_j(f)$. In this section we target and investigate a more difficult and general setting of slow source movement. Such situations occur, for instance, in real human recordings where people often move their head while talking or in general music mixtures where the considered spatial images may consist of several instruments such as *e.g* the drums source including bass drum, snare drum, hi-hat, etc. In these cases, the spatial covariance matrices of each source are actually time-varying as parameterized by $\mathbf{R}_j(n,f)$ but can be assumed to *vary smoothly over time*. Indeed the speed of head movements is limited and in music mixtures one instrument usually predominates in a given time-frequency neighborhood. In the following we present a Markov chain inverse-Wishart distribution over $\mathbf{R}_j(n,f)$ to enforce this temporal continuity structure. **Again this prior is chosen because it is conjugate to the likelihood**.

6.2.2 Markov chain inverse Wishart prior

We write the Markov chain inverse-Wishart prior for $\mathbf{R}_j(n,f)$ when $n>1$ as

$$p\big(\mathbf{R}_j(n,f)\big) = \mathcal{IW}\big(\mathbf{R}_j(n,f)|(m_j-I)\mathbf{R}_j(n-1,f), m_j\big) \tag{6.15}$$

where $\mathcal{IW}\big(\mathbf{R}|\mathbf{\Psi},m\big)$ is defined in (6.10). This prior distribution enforces the mean values of the spatial covariance matrices at the current time frame to be equal to those at the previous time frame, *i.e.* $\mathbb{E}\{\mathbf{R}_j(n,f)\} = \mathbf{R}_j(n-1,f)$. The initial distributions $p\big(\mathbf{R}_j(1,f)\big)$ for the first time frame index are chosen as uniform for all f.

6.2.3 MAP spatial parameter update

Similarly to the MAP parameter estimation exploiting spatial location priors, $\mathbf{R}_j(n,f)$ is iteratively updated by an EM algorithm. The computations of the Wiener filter and the expected covariance matrices $\widehat{\mathbf{\Sigma}}_j(n,f)$ in the E-step are unchanged compared to algorithm 5.2.

In the M-step, the expectation of the log-posterior of the complete data

$$Q_{\mathcal{MIW}}(\theta|\theta^{\text{old}}) = \sum_{j,n,f} \log p\big(\mathbf{c}_j(n,f)|\mathbf{0},\mathbf{\Sigma}_j(n,f)\big) + \gamma \log \mathcal{IW}\big(\mathbf{R}_j(n,f)|(m_j-I)\mathbf{R}_j(n-1,f), m_j\big) \tag{6.16}$$

is maximized with respect to $\mathbf{R}_j(n,f)$, where γ is a tradeoff hyper-parameter determining the strength of the continuity prior. By computing the partial derivatives of $Q_{\mathcal{MIW}}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_j(n,f)$ and equating them to zero, we ob-

tain a quadratic matrix equation. By solving this equation (see the detailed description in Appendix A.4), we obtain the spatial covariance update as

$$\mathbf{R}_j(n,f) = \frac{1}{2}\mathbf{A}^{-1/2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}\big)\mathbf{A}^{-1/2} \tag{6.17}$$

where

$$\begin{aligned} \mathbf{A} &= \gamma(m_j - I)\mathbf{R}_j^{-1}(n+1,f) \\ b &= \gamma I + 1 \\ \mathbf{C} &= -\widehat{\mathbf{\Sigma}}_j(n,f)/v_j(n,f) - \gamma(m_j - I)\mathbf{R}_j(n-1,f) \end{aligned} \tag{6.18}$$

Note that in (A.18), due to the choice of initial distribution, $\mathbf{R}_j(n-1,f)$ are zero for all f when $n = 1$. Again, the update of $v_j(n,f)$ is not affected by the prior. This resulting MAP update requires remarkably more computational time compared to the ML update, as observed experimentally, since the M-step involves the computation of square root of matrices.

6.3 Summary

In this chapter, we have introduced two families of prior distributions over the spatial covariance to enhance the source separation performance in certain contexts. The Wishart and inverse-Wishart spatial location priors, which exploit knowledge from the theory of statistical room acoustics, are applicable when the geometric setting is known. They both result in closed-form MAP parameter updates and, more interestingly, the MAP algorithms do not suffer from the permutation problem thanks to this prior information about the source location. The spatial continuity prior enables the modeling of time-varying spatial covariances constrained by a Markov chain structure. This new model is essential to deal with slowly moving sources, which is beyond the scope of this thesis.

Chapter 7

Parameter estimation algorithms exploiting spectral priors

As a complement to the investigation of spatial priors in Chapter 6, this chapter will focus on investigating the spectral structures of the sources and the resulting parameter estimation algorithms to enhance the source separation performance in certain contexts. For that purpose, we first consider a NMF spectral model in Section 7.1 where $v_j(n,f)$ is the sum of several latent components. We then address spectral continuity priors in Section 7.2 with the application to music component separation discussed in Chapter 8. For both models, we estimate the spectral parameters in the MAP sense by EM updates where only the M-step is modified compared to the ML algorithm. **Note that, when considering some spectral priors, the troublesome permutation problem is reduced thanks to the structured coupling between the spectral parameters across all frequency bins and the joint estimation of $v_j(n,f)$ and $\mathbf{R}_j(f)$.** This can be seen as an additional advantage when incorporating a spectral source model in our source separation framework.

7.1 ML algorithm exploiting a NMF spectral model

In this section we briefly show the possibility of incorporating a spectral source model in the proposed Gaussian modeling framework in order to enhance the source separation performance. Among many well-known spectral models like GMM, GSMM, and NMF briefly reviewed in Section 2.1, we consider the NMF spectral model. This model was recently combined with the rank-1 spatial covariance parameterization in the same context of convolutive audio source separation in [24] and shown to provide better

separation than sparsity based approaches like binary masking or ℓ_p-norm minimization. We extend here the formulation of [24] to full-rank covariances and quadratic T-F representation. Examples of integration of the full-rank spatial covariance model with other spectral models (hierarchical NMF, GMM, GSMM) are given in [37].

7.1.1 NMF model

In the spectral NMF model (2.14), the source variances are constrained as

$$v_j(n,f) = \sum_{k=1}^{K} h_{j,nk} w_{j,kf} \tag{7.1}$$

where $h_{j,nk},\ w_{j,kf} \in \mathbb{R}^+$.

7.1.2 ML spectral parameter update

The derivation of the EM algorithm, which is similar to that of Algorithm 5.2, is based on the complete data $\{\mathbf{c}_j(n,f)\}_{n,f}$, that is the set of hidden T-F coefficients of the spatial images of all components on all time frames. The E-step is the same as Algorithm 5.2. Then in the M step, unconstrained estimates $\xi_j(n,f)$ of the source variances are first obtained as in (5.21), then $h_{j,nk}$ and $w_{j,kf}$ are iteratively updated via multiplicative update (MU) rules as [24, 37]

$$\xi_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \tag{7.2}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \odot \frac{\mathbf{W}^T((\mathbf{W}\mathbf{H})^{\odot -2} \odot \mathbf{\Xi}_j)}{\mathbf{W}^T(\mathbf{W}\mathbf{H})^{\odot -1}} \tag{7.3}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \odot \frac{((\mathbf{W}\mathbf{H})^{\odot -2} \odot \mathbf{\Xi}_j)\mathbf{H}^T}{(\mathbf{W}\mathbf{H})^{\odot -1}\mathbf{H}^T} \tag{7.4}$$

where $\mathbf{\Xi}_j = [\xi_j(n,f)]_{n,f}$, $\mathbf{H}_j = [h_{j,nk}]_{n,k}$, and $\mathbf{W}_j = [w_{j,kf}]_{k,f}$ are $N \times F$, $N \times K$, and $K \times F$ matrices, respectively, and $\odot$ denotes Hadamard entrywise product.

Note that, similarly to the MAP algorithms presented in Chapter 6 where the spectral parameter update was not affected by the spatial priors, in this algorithm the update of $\mathbf{R}_j(f)$ is the same as in Algorithm 5.2, which is not affected by the spectral model.

7.2 MAP algorithm exploiting a spectral continuity prior

When observing time-frequency representations of real-world sound sources, one can see that the spectra of certain sources are *smooth in the time direction*, while those of some other sources are piecewise *smooth in the frequency direction* [115]. The first class of sources typically includes harmonic sources such as vocals, piano, or violin, while the second incluides percussive sources such as drums. When considering the separation of these sources, we can exploit this spectral smoothness structure to enhance the separation performance. Therefore, in this section, we propose spectral continuity priors and investigate the resulting MAP parameter estimation algorithm in the local Gaussian modeling framework.

7.2.1 Markov chain inverse-gamma prior

Since the spectrum of harmonic sources is usually *smooth* over the time axis, we consider the following Markov chain prior for $v_j(n,f)$, when j is an index of a harmonic source, with $n > 1$:

$$p\big(v_j(n,f)\big) = \mathcal{IG}\big(v_j(n,f)|\alpha_j, (\alpha_j - 1)v_j(n-1,f)\big) \tag{7.5}$$

where $\mathcal{IG}(v|\alpha,\beta)$ denotes the inverse-gamma density with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$

$$\mathcal{IG}(v|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} e^{-\beta/v} \tag{7.6}$$

whose mean is $\beta/(\alpha-1)$.

Similarly, since the spectrum of percussive sources is usually *smooth* over the frequency axis, we consider the following Markov chain prior for $v_j(n,f)$, when j is an index of a percussive source, with $f > 1$:

$$p\big(v_j(n,f)\big) = \mathcal{IG}\big(v_j(n,f)|\alpha_j, (\alpha_j - 1)v_j(n,f-1)\big). \tag{7.7}$$

Similarly to the spatial continuity prior, these prior distributions enforce $\mathbb{E}\{v_j(n,f)\} = v_j(n-1,f)$ for harmonic sources, and $\mathbb{E}\{v_j(n,f)\} = v_j(n,f-1)$ for percussive sources.

Similarly to the choice of initial distribution, *i.e.* when $n = 1$ or $f = 1$, for the spatial covariances in Section 6.2.2, $p\big(v_j(1,f)\big)\ \forall f$ (for harmonic sources) and $p\big(v_j(n,1)\big)\ \forall n$ (for percussive sources) are chosen as uniform. **The choice of an inverse-gamma prior, which is the conjugate prior for the considered likelihood, brings not only simpler computation compared to the Gaussian prior investigated in [115] but also better separation performance as shown in our experiments**

in Section 8.5. This prior distribution was also used to model temporal continuity in [24] in the different context of multipitch estimation.

7.2.2 MAP spectral parameter update

The resulting EM updates for $v_j(n,f)$ have a similar form as those of $\mathbf{R}_j(n,f)$ under a spatial continuity prior. In the E-step, the Wiener filters and the expected covariance matrices $\widehat{\mathbf{\Sigma}}_j(n,f)$ are computed as in Algorithm 5.2.

In the M-step, the expectation of the log-posterior of the complete data

$$Q_{\mathrm{JG}}(\theta|\theta^{\mathrm{old}}) = \sum_{j,n,f} \log p\big(\mathbf{c}_j(n,f)|\mathbf{0},\mathbf{\Sigma}_j(n,f)\big) + \gamma \log p(v_j(n,f)) \tag{7.8}$$

is maximized with respect to $v_j(n,f)$, where $p(v_j(n,f))$ is given by either (7.5) (for harmonic sources) or (7.7) (for percussive sources), and γ is a tradeoff hyper-parameters determining the strength of the continuity priors. By computing the partial derivative of $Q_{\mathrm{JG}}(\theta|\theta^{\mathrm{old}})$ with respect to $v_j(n,f)$ and equating it to zero, we obtain a second order polynomial equation (see Appendix A.5) with a single positive solution

$$v_j(n,f) = (-b + \sqrt{b^2 - 4ac})/(2a) \tag{7.9}$$

where

$$\begin{aligned} a &= \begin{cases} \gamma(\alpha-1)/v_j(n+1,f) & \text{for harmonic sources} \\ \gamma(\alpha-1)/v_j(n,f+1) & \text{for percussive sources} \end{cases} \\ b &= I \\ c &= \begin{cases} -\mathrm{tr}\big(\mathbf{R}_j^{-1}(n,f)\widehat{\mathbf{\Sigma}}_j(n,f)\big) - \gamma(\alpha-1)v_j(n-1,f) & \text{for harmonic sources} \\ -\mathrm{tr}\big(\mathbf{R}_j^{-1}(n,f)\widehat{\mathbf{\Sigma}}_j(n,f)\big) - \gamma(\alpha-1)v_j(n,f-1) & \text{for percussive sources.} \end{cases} \end{aligned} \tag{7.10}$$

7.3 Summary

In this chapter, we have investigated the possibility of incorporating spectral source models in our proposed Gaussian modeling framework. We considered two spectral structures of the sources: an NMF model and spectral continuity priors. Both models result in closed-form parameter updates and do not require significantly more computational time, as observed experimentally, compared to the ML parameter update. They both help enhance the source separation performance in certain contexts, as will be shown in the experiments in Chapter 8.

Chapter 8

Evaluation of the separation performance in various scenarios

This chapter presents a complete evaluation of the source separation performance achieved by the proposed framework in different blind and non-blind scenarios. We first describe the test data consisting of several simulated and real-world datasets corresponding to different mixing conditions. We then present two popular families of evaluation metrics used in the source separation community to quantify the separation performance. Finally the results of various experiments are reported to compare the proposed algorithms to state-of-the-art and baseline approaches.

8.1 Test data

We consider four datasets ranging from simulated speech mixtures to more sophisticated real-world speech recordings, without and with background noise, and to real-world music mixtures. These datasets will be used to evaluate the source separation performance in Sections 8.3, 8.4, and 8.5.

8.1.1 Simulated speech mixtures

Simulated mixtures allow the generation of a wide range of recording configurations while keeping precise control over the configuration parameters, including the geometric setting and the room characteristics. More precisely, we generate simulated speech mixtures for the following purposes:

- in Chapter 4 we investigate the potential source separation performance achievable

by the proposed approach in the *oracle* and the *semi-blind* context where all or part of the model parameters are known,

- in Section 8.3.1 we investigate the *blind* source separation performance as a function of the reverberation time and the distance from the sources to the microphones,
- in Section 8.4 we investigate the source separation using spatial location priors where the geometric setting and the room characteristics need to be known.

For those purposes, we generated room impulse responses via the *image method* [116] from three source positions to a microphone pair position using the Roomsim toolbox[5]. The room dimensions were $4.45 \times 3.55 \times 2.5$ m, that are the dimensions used in the SiSEC campaign [12], and the microphone spacing was fixed to 5 cm. The positions of the sources and the microphones are illustrated in Fig. 8.1. Four different reverberation times were considered: $T_{60} = 50$, 130, 250 and 500 ms, and, for each T_{60}, three source-to-microphone distances were considered: $r = 50$, 100, 150 cm. The source images were computed by convolving 10 s speech signals with the simulated impulse responses. Two sets of speech signals were considered: male and female speech. This resulted in two mixtures for each T_{60} and r and 24 mixtures in total.

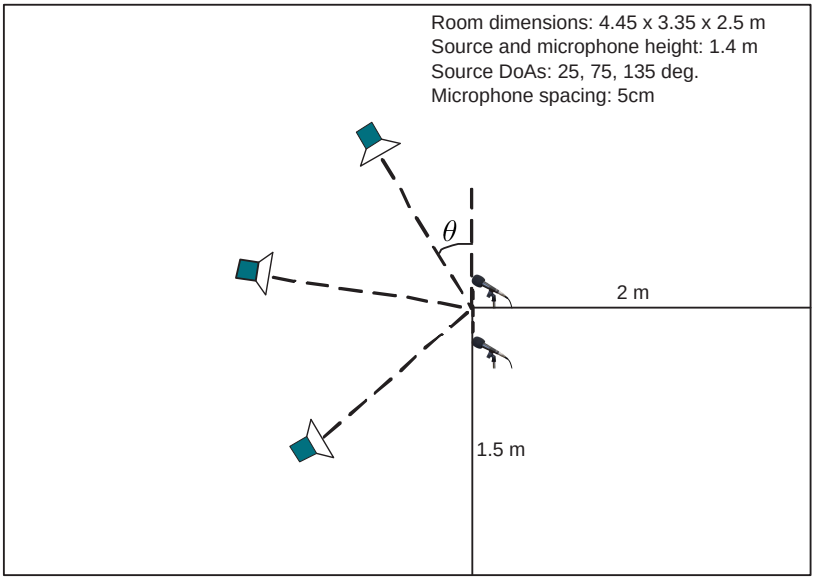


Figure 8.1: Room geometry setting for simulated speech mixtures.

[5] http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip

8.1.2 Real-world speech recordings

We also used a real-world speech dataset for the following reasons:

- in Section 8.3.2 we investigate the source separation performance of the proposed approach in real-world mixing conditions in both determined and under-determined cases,
- we also compare the separation performance of the proposed approach with state-of-the-art algorithms submitted for evaluation to SiSEC 2008 and SiSEC 2010.

8.1.2.1 Under-determined mixtures

These mixtures were taken from the *live recording test data* of the underdetermined-speech and music mixtures task[6] of SiSEC 2010, formerly used in SiSEC 2008 [12]. They consist of 8 real-world mixtures of 3 or 4 speech sources sampled at 16 kHz. The room reverberation time was either 130 ms or 250 ms and the microphone spacing 5 cm. Two mixtures were recorded for each given number of sources and each reverberation time, using either male or female speech signals. The source directions of arrival vary between -60 and +60 degrees with a minimal spacing of 15 degrees and the distances between the sources and the center of the microphone pair vary between 80 cm and 1.20 m.

8.1.2.2 Determined mixtures

These mixtures were taken from part of the *loesch* dataset of the determined and over-determined speech and music mixtures task[7] of SiSEC 2010 [13]. They consist of 2 stereo mixtures of 2 speech sources recorded in an office room whose reverberation time is about 450-500 ms. The sampling rate is 16 kHz and the microphone spacing is about 10 cm. Sources were played back using small loudspeakers placed approximately 120-140 cm far from the center of the microphones.

8.1.3 Real-world speech in background noise recordings

Since the ultimate goal of sound source separation techniques is to separate sources in the real world, we collected a dataset for the evaluation of source separation in the presence of real-world background noise. The separation results will be reported

[6] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Underdetermined-+speech+and+music+mixtures

[7] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Determined+and+over-determined+speech+and+music+mixtures

in Section 8.3. We originally proposed this dataset for the *Source separation in the presence of real-world background noise*[8] task of the SiSEC 2010 campaign [13] which attracted 6 entries. This dataset involved both stereo and 4-channel mixtures. In this thesis, for simplicity and consistency with the other datasets, we restricted ourself to stereo mixtures of one speech source and one of several kinds of real-world background noise. All mixtures are exactly 10 s long and sampled at 16 kHz. The microphones are omnidirectional and the microphone spacing is 8.6 cm.

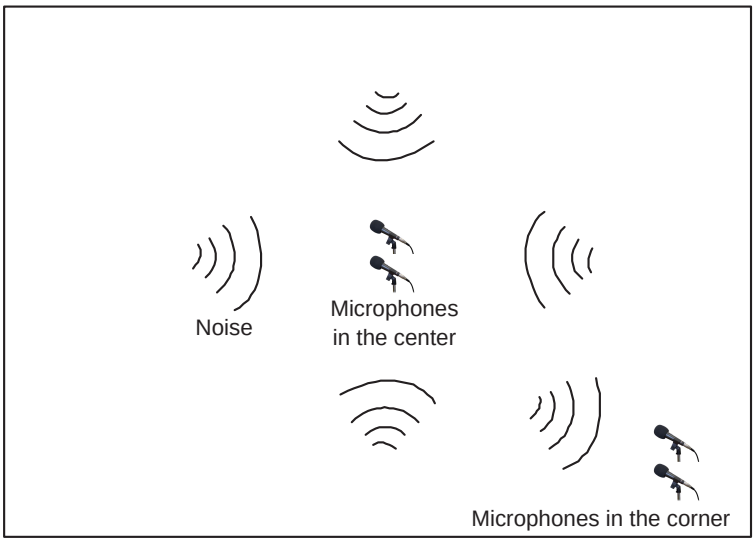


Figure 8.2: Room geometry setting for real-world noise recordings.

Noise recording: We recorded noise signals in three different real-world environments: subway car (Su), cafeteria (Ca), and square (Sq). In the (Ca) and the (Sq) environments, the noise signals were recorded at two different positions within the environment as shown in Fig. 8.2: center (Ce), where the noise is more isotropic, and corner (Co), where the noise may not be very isotropic. However in the (Su) environment, noise signals were recorded only in the center.

Mixture generation: The speech signals were separately acquired and subsequently added to the noise signals. This is because we need to have access to their spatial images in order to evaluate the separation performance [12]. For the (Su) and the (Ca) environments where there is reverberation, we recorded the source signals in a room with a reverberation time of approximately 250 ms. For the (Sq) environment where there is little or no reverberation, we simulated the source contributions to the

[8] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Source+separation+in+the+presence+of+real-world+background+noise

microphones under the assumption of anechoic spherical wave propagation. For each mixing condition, we considered two speech sources: male voice and female voice. This results in a total of 10 mixtures in the dataset.

8.1.4 Real-world music

In addition to the above three speech datasets, we consider a real-world music dataset. Sound mixing techniques vary depending on the music genre: for certain genres, instruments are placed at different spatial positions while for other genres most instruments are mixed to the center. This dataset consists of 8 stereo music mixtures of harmonic and percussive sources corresponding to two different mixing conditions. These mixtures are part of the Quaero project database, which was used for the *Professionally produced music recordings* task of SiSEC 2010[9][13]. The first 4 mixtures were originally mixed by a sound engineer where most instruments are panned close to the center with artificial reverb, and the total number of harmonic and percussive instruments in each mixture varies from four to eight. In order to investigate the contribution of spatial information, in the second set of mixtures (named Pan+), we moved each instrument to a random position by amplitude panning from 0 to 3 dB but keeping the same reverb. All the mixtures are sampled at 44.1 kHz and have 10 s duration. The separation performance given by the proposed harmonic and percussive source separation algorithm and state-of-the-art systems will be presented in Section 8.5.

8.2 Evaluation criteria

Several studies have been devoted to the evaluation of the source separation performance in the literature, both in terms of *subjective* quality [117, 118, 119] and *objective* quality [120, 121, 122]. We focus here on two popular families of *objective* evaluation criteria, which can be applied to any audio mixture and any algorithm and do not require the knowledge of the unmixing parameters or filters. These criteria, namely energy ratio criteria and perceptually-motivated criteria, have been widely used in recent evaluation campaigns [11, 12, 1].

[9] http://sisec.wiki.irisa.fr/

8.2.1 Energy ratio criteria

Both families of criteria derive from the perceptual decomposition of each estimated source image $\hat{c}_{ij}(t)$ as [11]

$$\hat{c}_{ij}(t) = c_{ij}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{inter}}(t) + e_{ij}^{\text{artif}}(t) \tag{8.1}$$

where $c_{ij}(t)$ is the true spatial image of the j-th source at the i-th microphone, and e_{ij}^{spat}, e_{ij}^{inter}, and e_{ij}^{artif} are different error components representing spatial (or filtering) distortion, interference from the other sources, and artifacts (or musical noise). In the first family of criteria, these error components are computed as [11]:

$$e_{ij}^{\text{spat}}(t) = P_j^L(\hat{c}_{ij})(t) - c_{ij}(t) \tag{8.2}$$

$$e_{ij}^{\text{inter}}(t) = P_{\text{all}}^L(\hat{c}_{ij})(t) - P_j^L(\hat{c}_{ij})(t) \tag{8.3}$$

$$e_{ij}^{\text{artif}}(t) = \hat{c}_{ij}(t) - P_{\text{all}}^L(\hat{c}_{ij})(t) \tag{8.4}$$

where P_j^L is the least-squares projector onto the subspace spanned by $c_{kj}(t-\tau)$, $1 \leq k \leq I$, $0 \leq \tau \leq L-1$, P_{all}^L is the least-squares projector onto the subspace spanned by $c_{kl}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $0 \leq \tau \leq L-1$, and L is the filter length which is set to 32 ms.

Given this decomposition, four energy ratios in decibels (dB) measuring overall distortion, artifacts, interference, and spatial distortion called *Signal to Distortion Ratio* (SDR), *Signal to Artifacts Ratio* (SAR), *Signal to Interference Ratio* (SIR), and *source Image to Spatial distortion Ratio* (ISR), respectively, are computed as [122, 11]

$$\text{SDR} = 10\log_{10} \frac{\sum_{i=1}^{I}\sum_t c_{ij}(t)^2}{\sum_{i=1}^{I}\sum_t \left(e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{inter}}(t) + e_{ij}^{\text{artif}}(t)\right)^2} \tag{8.5}$$

$$\text{SAR} = 10\log_{10} \frac{\sum_{i=1}^{I}\sum_t \left(c_{ij}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{inter}}(t)\right)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\text{artif}}(t)^2} \tag{8.6}$$

$$\text{SIR} = 10\log_{10} \frac{\sum_{i=1}^{I}\sum_t \left(c_{ij}(t) + e_{ij}^{\text{spat}}(t)\right)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\text{inter}}(t)^2} \tag{8.7}$$

$$\text{ISR} = 10\log_{10} \frac{\sum_{i=1}^{I}\sum_t c_{ij}(t)^2}{\sum_{i=1}^{I}\sum_t e_{ij}^{\text{spat}}(t)}. \tag{8.8}$$

When the source images are estimated up to an arbitrary order, all possible orders are evaluated and the order leading to the largest average SIR is selected. These criteria

were implemented in Matlab and distributed for public use[10].

8.2.2 Perceptually-motivated criteria

In addition to these energy ratio criteria, we consider the recently proposed perceptually-motivated objective criteria in [119] to assess the quality of the estimated source image signals. Four performance criteria akin to SDR, SAR, SIR, and ISR were defined and termed *Overall Perceptual Score* (OPS), *Artifacts-related Perceptual Score* (APS), *Interference-related Perceptual Score* (IPS), and *Target-related Perceptual Score* (TPS), respectively. These criteria score from 0 to 100 where higher values indicate better performance. They were derived based on the decomposition of the estimated source image signals into several distortion components similarly to (8.2), (8.3), and (8.4) and on the use of the PEMO-Q perceptual salience measure [123]. They were shown to improve the correlation with subjective scores compared to the energy ratio criteria [119] and were recently used in SiSEC 2010. The source code of these perceptually-motivated criteria is also available[11].

8.3 ML source separation performance in blind scenarios

In this section we investigate the *blind* source separation performance achieved with the proposed full-rank unconstrained spatial covariance parameterization where the model parameters are estimated from the observed mixture signals only, *i.e.* without any prior knowledge about either the source locations or the source spectral structures, in the ML sense as presented in Chapter 5. The mixing scenarios vary from simulated to real-world recorded mixtures and from noiseless to noisy mixtures. The detail ML parameter estimation algorithm including parameter initialization and EM parameter update is presented in Chapter 5.

8.3.1 BSS performance on simulated speech mixtures

The first experiment is devoted to the BSS performance of the proposed full-rank unconstrained spatial covariance parameterization with different sound scene parameters and with either the quadratic STFT representation (named fullrank unconstrained STFT) or the quadratic ERB T-F representation (named fullrank unconstrained ERB), compared to the state-of-the-art rank-1 convolutive parameterization and to the baseline

[10] http://bass-db.gforge.inria.fr/bss_eval/
[11] http://bass-db.gforge.inria.fr/peass/

approaches. Note that we consider the quadratic T-F representations here since they were confirmed to provide better separation performance than the linear STFT in Section 4.3.4 and in [31]. For that purpose, we consider the simulated speech mixtures dataset described in Section 8.1.1 and the parameter settings in Table 8.1.

Window type	sine window
STFT frame size	1024
STFT frame shift	512
Number of ERB bins	250
ERB frame shift	512
Propagation velocity	343 m/s
Number of EM iterations	10
Cluster threshold	$K = 30$

Table 8.1: Experimental parameter settings for source separation in blind scenarios.

Fig. 8.3 and Fig. 8.4 show the BSS performance, as evaluated by the energy ratio criteria, as a function of the reverberation time and the distance from the sources to the microphones, respectively. The results are averaged over all sources, all male and female speech mixtures, and either all distances from the sources to the microphones (for Fig. 8.3) or all reverberation times (for Fig. 8.4). The average computational time for each mixtures using Matlab 7.0 in an Intel core 2.4 GHz computer is about 3 minutes. As can be seen, the separation performance of all algorithms have dropped compared to the oracle and semi-blind results presented in Chapter 4. Also, as predicted, the BSS performance of all algorithms decreases with higher reverberation time and further distance from the sources to the microphones. However the SDR and SIR of the rank-1 convolutive approaches decrease faster compared to the fullrank unconstrained approaches. Binary masking offers very similar SIR to the fullrank unconstrained ERB and higher SIR than the other approaches, however it results in the poorest SAR meaning very strong musical noise. Overall, we observe that the proposed full-rank unconstrained STFT and full-rank unconstrained ERB approaches always provide higher SDR than the rank-1 convolutive STFT, the rank-1 convolutive ERB, and baseline approaches for all reverberation times and all distances from the sources to the microphones. For instance, with $T_{60} = 250$ ms, the SDR achieved via the full-rank unconstrained STFT approach is 1.7 dB, 0.9 dB and 2.3 dB larger than that of the rank-1 convolutive approach, binary masking, and ℓ_0-norm minimization respectively. These results confirm the effectiveness of our proposed model parameter estimation scheme and show in addition that full-rank spatial covariance matrices better approximate the mixing process

in a reverberant room.

Regarding the input T-F representation, the rank-1 convolutive ERB results in poorer SDR and SIR than and very similar SAR and ISR to the rank-1 convolutive STFT. This can be explained by the fact that the narrowband assumption, for which the rank-1 model relies on, does not hold at high frequencies in the ERB scale. On the contrary, the fullrank unconstrained ERB approach results in very similar SDR, when averaged over all T_{60} but slightly larger SAR and SIR than the fullrank unconstrained STFT approach. This suggests that full-rank models are needed from now to achieve further improvement in ERB scale-based source separation, since they partially overcomes the narrowband assumption and do not suffer from the large bandwidth at high frequencies.

8.3.2 BSS performance on real-world speech mixtures

In the second and third experiment, we compare the proposed full-rank unconstrained spatial covariance parameterization, with or without a NMF spectral model, with state-of-the-art BSS algorithms submitted for evaluation to SiSEC 2008 and SiSEC 2010 over the real-world speech recording dataset described in Section 8.1.2. The parameter settings for our proposed algorithm are listed in Table 8.1 where the input T-F representation is given by the quadratic STFT.

8.3.2.1 BSS performance on under-determined speech mixtures

Table 8.2 shows the average SDR and OPS achieved by each algorithm on the under-determined mixtures dataset described in Section 8.1.2.1. We consider SDR and OPS for comparison since they both measure the overall distortion of the system. Note that only the proposed approach and three other algorithms in SiSEC 2010 [124, 32, 125] were evaluated in term of OPS since this criterion was not available in 2008. The SDR and OPS figures of all algorithms besides the proposed algorithm were taken from the websites of SiSEC 2008[12] and SiSEC 2010[13] except for Izumi's algorithm [66] whose results were provided by its author.

[12] http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+speech+and+music+mixtures
[13] http://www.irisa.fr/metiss/SiSEC10/underdetermined/underdetermined-test-all.html

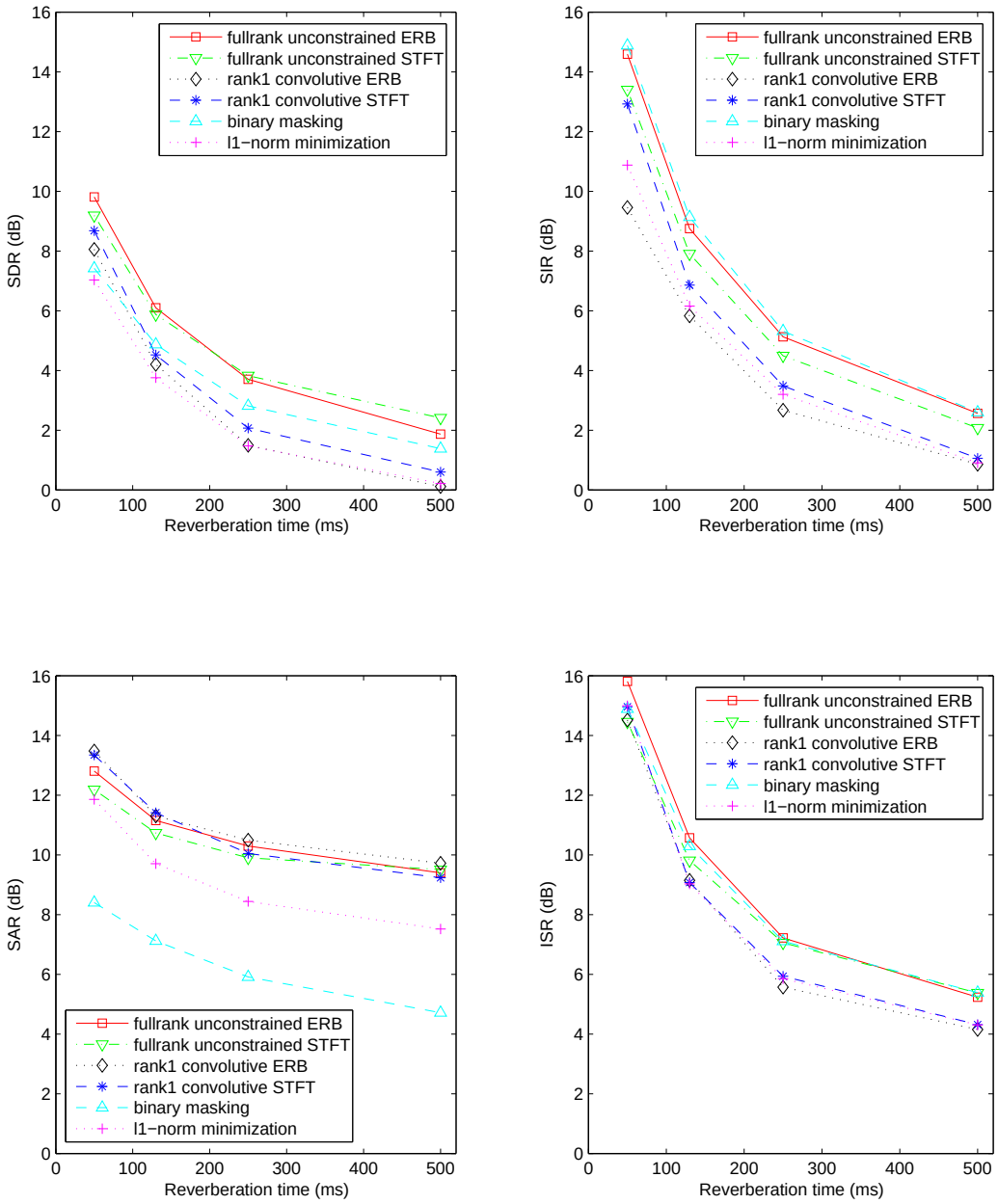


Figure 8.3: BSS performance on simulated speech mixtures **as a function of the reverberation time**.

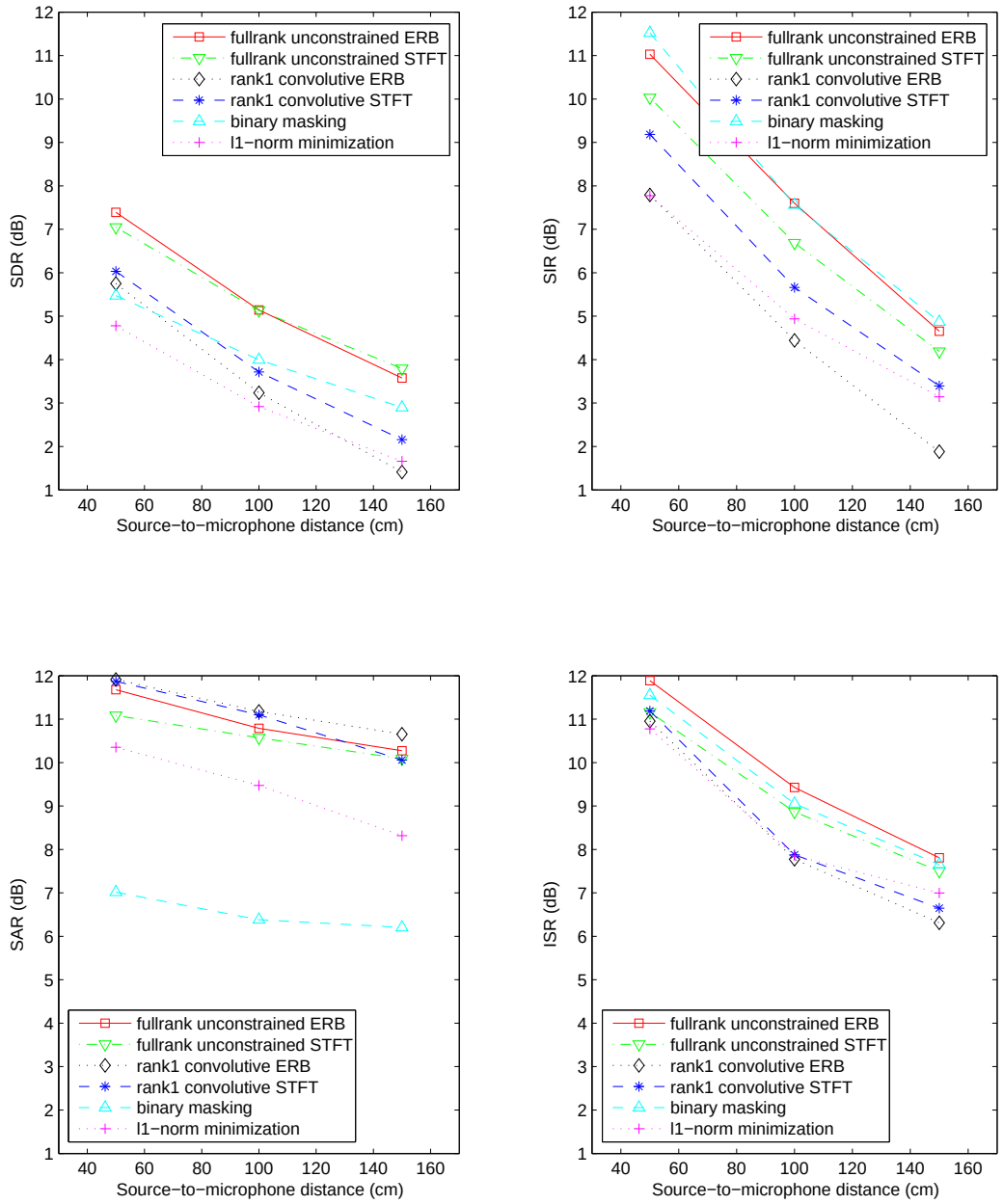


Figure 8.4: BSS performance on simulated speech mixtures **as a function of the distance from the sources to the microphones**.

T_{60}	Algorithm	SiSEC	3 sources		4 sources	
			SDR	OPS	SDR	OPS
130 ms	M. Cobos [126]	2008	2.3	-	2.1	-
	M. Mandel [98]		0.1	-	-3.7	-
	R. Weiss [127]		2.9	-	2.3	-
	S. Araki [128]		2.9	-	-	-
	Z. El Chami [129]		2.3	-	2.1	-
	Full-rank		**3.3**	35	**2.8**	32
	Ozerov [124]	2010	3.4	**55**	2.4	29
	Full-rank+NMF		3.8	46	3.1	**33**
	H. Sawada [125]		**7.2**	40	**3.4**	19
250 ms	M. Cobos [126]	2008	2.2	-	1.0	-
	M. Mandel [98]		0.8	-	1.0	-
	R. Weiss [127]		2.3	-	1.5	-
	S. Araki [128]		3.7	-	-	-
	Z. El Chami [129]		3.1	-	1.4	-
	Full-rank		**3.8**	40	**2.0**	25
	Ozerov [124]	2010	4.5	**47**	1.0	**30**
	Full-rank+NMF		4.4	40	2.2	25
	H. Sawada [125]		**6.1**	38	**3.7**	21

Table 8.2: Average SDR in dB and OPS over the real-world test data of SiSEC 2008 and SiSEC 2010 with 5 cm microphone spacing.

Let us first discuss the comparison with the algorithms in SiSEC 2008 [126, 98, 127, 128, 129] when this thesis work was started. For three-source mixtures, the proposed algorithm provides 0.4 dB and 0.1 dB SDR improvement compared to the best results given by Araki's algorithm [128] with $T_{60} = 130$ ms and $T_{60} = 250$ ms, respectively. For four-source mixtures, it provides even higher SDR improvements of 0.5 dB and 0.4 dB respectively compared to the best results given by Weiss's [127] and Izumi's algorithms [66]. More detailed comparison (not shown in the Table) indicates that the proposed algorithm also outperforms most others in terms of SIR, SAR and ISR. For instance, it achieves higher SIR than all other algorithms on average except Weiss's. Compared to Weiss's, it achieves the same average SIR but a higher SAR.

Let us now compare to the three new algorithms [124, 32, 125] submitted to SiSEC 2010. The full-rank approach performs better than Ozerov's [124] with the four-source mixtures in terms of SDR and OPS (for $T_{60} = 130$ ms), but worse with the three-source mixtures. The full-rank+NMF algorithm [32] whose authors we are part of, results in better SDR than the full-rank with both reverberation times and number of sources but similar OPS with $T_{60} = 250$ ms. Sawada's method [125] offers the best SDR overall, but

poorer average OPS than all three other algorithms, while Ozerov's [124] offers the best OPS overall. However, let us clarify that the algorithms [124, 32] were both built upon our proposed Gaussian modeling framework with the full-rank unconstrained spatial covariance parameterization. The difference lies in the use of spectral source models, namely harmonic NMF in [124] and NMF in [32]. This confirms our expectation that by incorporating a spectral source model in the proposed Gaussian modeling framework, we can obtain better separation performance.

Besides, the different results achieved by our proposed approach, [124], and [32], which all used different initialization schemes, also suggest that the parameter initialization could highly affect the separation performance in the full-rank spatial covariance framework. Sawada's method [125] uses bin-wise probabilistic clustering of the mixture coefficient $\mathbf{x}(n, f)$ in the STFT domain followed by a new probabilistic method for solving the permutation problem. It would be easily feasible to use this method, as the initialization step in our framework. We leave this potential improvement for future work.

8.3.2.2 BSS performance on determined speech mixtures

Table 8.2 shows the average energy ratio criteria and perceptually motivated criteria achieved by each algorithm on the determined mixtures dataset described in Section 8.1.2.2. The result of all algorithms besides the proposed algorithm were taken from the websites of SiSEC 2010[14].

Algorithm	Energy ratio criteria (dB)				Perceptually-motivated criteria			
	SDR	SIR	SAR	ISR	OPS	IPS	APS	TPS
Proposed full-rank	4.6	6.6	11.9	8.7	22.9	31.2	70.7	45.0
H. Sawada [130]	**8.8**	13.7	11.9	14.2	**37.6**	63.9	53.1	55.9
N. Ono [131]	3.0	5.2	7.7	7.5	29.0	55.7	45.8	49.7
N. Ono2 [131]	5.5	9.3	9.2	10.6	25.4	47.2	42.2	46.1

Table 8.3: Average BSS performance over the real-world stereo mixtures of two sources with 10 cm microphone spacing.

Our algorithm results in the best SAR and APS, *i.e.* less musical noise, but poorer SDR and SIR than other algorithms [130, 131], which used more advanced techniques for solving the permutation problem. Let us clarify that all other algorithms were based on ICA and, on contrary to ours, can not handle the separation of under-determined mixtures.

[14] http://www.irisa.fr/metiss/SiSEC10/determined/determined_all.html

8.3.3 BSS performance on real-world speech and noise mixtures

The fourth experiment aims to examine the source separation performance of the proposed approach in a challenging scenario where speech signals are mixed with real-world background noise recorded in different noisy environments. For that purpose, we applied our algorithm to the real-world speech in background noise dataset described in Section 8.1.3 and compared the results with the four other state-of-the-art approaches which participated in SiSEC 2010[15]. The average performance in terms of both energy ratio criteria and perceptually-motivated criteria of all algorithms is summarized in Table 8.4.

Algorithm	Energy ratio criteria (dB)				Perceptually-motivated criteria			
	SDR	SIR	SAR	ISR	OPS	IPS	APS	TPS
Proposed full-rank	**2.7**	4.4	11.9	16.1	36.3	45.2	88.3	69.9
J. Even [133]	-9.3	14.1	8.0	-7.2	33.5	55.5	53.8	49.4
M. Go [134]	-12.2	6.2	8.8	-7.7	19.3	34.8	40.6	35.2
N. Hirata [135]	-8.6	7.5	5.4	-3.2	16.3	35.2	28.3	39.4
H. Sawada [130]	2.5	4.0	12.8	16.0	**38.4**	51.2	78.4	73.4

Table 8.4: Average BSS performance in the presence of real-world background noise

We can see that the algorithms in [133, 134, 135] performed poorly with very low SDR and OPS. Generally, our proposed algorithm performed as well as Sawada's [130] where we obtain larger SDR, SIR, ISR, and APS, but smaller SAR, OPS, IPS, and TPS. This evaluation confirms the appropriateness of our proposed approach to real-world background noise where the full-rank unconstrained spatial covariance parameterization accounts well for diffuse noise sources.

8.4 Source separation using spatial location priors

Our fifth experiment is devoted to the investigation of the separation performance achieved by the MAP algorithms exploiting spatial location priors presented in Section 6.1. For this purpose, we consider the simulated speech mixtures dataset described in Section 8.1.1 where the geometric settings and the room characteristics are known. We compare the separation results obtained by the MAP algorithm exploiting the inverse-Wishart spatial location prior presented in Section 6.1.2, named MAP inverse-Wishart, and the MAP algorithm exploiting the Wishart spatial location prior presented in Sec-

[15] One approach [132] estimated the target source only, so that it cannot be compared with others which estimated both the target source and the noise

tion 6.1.1, named MAP Wishart, with the ML algorithms where the spatial parameters were either blindly initialized as presented in Chapter 5, named ML blind init, or initialized from the known geometric setting by (3.9), named ML geom. init. We used the quadratic STFT representation where STFT was computed with a sine window of length 1024, w_{nf} of size 3×3, and 10 EM iterations for all algorithms. The trade-off parameter γ, which determines the strength of the priors, does not significantly affect the result but we observed that $\gamma = 100$ is globally a good choice. The optimal number of degrees of freedom m of the priors was learned from training data and is shown in Table 8.5 together with the mean power σ^2_{rev} of echoes and reverberation computed by (3.11), which both depend on the reverberation time.

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.8	4.2	6.4
σ^2_{rev}	0.011	0.057	0.131	0.287

Table 8.5: Learned value of m and predicted value of σ^2_{rev} for the Wishart and inverse-Wishart prior.

As expected, σ^2_{rev} strongly increases with reverberation, such that the direct-to-reverberant energy ratio is 14 dB lower when $T_{60} = 500$ ms than when $T_{60} = 50$ ms. The variance of the inverse-Wishart prior, which is inversely related to m [114], decreases with reverberation time while that of the Wishart prior, on the contrary, increases with reverberation time.

The separation performance of all algorithms, as evaluated by the energy ratio criteria and averaged for all sources and all distances from the source to the microphone, as a function of the reverberation time is shown in Fig. 8.5. The ML algorithm initialized from known geometry settings results in larger SDR, SAR and ISR than the blindly initialized ML algorithm at $T_{60} \geq 130$ ms. The MAP Wishart algorithm offers the best SAR but the worst SIR and ISR compared to all other algorithms, and overall it results in very similar SDR to the ML geom. init algorithm. Overall, the MAP inverse-Wishart algorithm outperforms all other algorithms for all considered reverberation times in terms of SDR, SIR, and ISR. For instance, at $T_{60} = 250$ ms, it improves the SDR by 0.8 dB, 0.4 dB and 0.3 dB compared to the ML blind init, the ML geom. init and the MAP Wishart algorithms, respectively. This confirms the benefit of the proposed inverse-Wishart spatial location prior and the associated MAP algorithm.

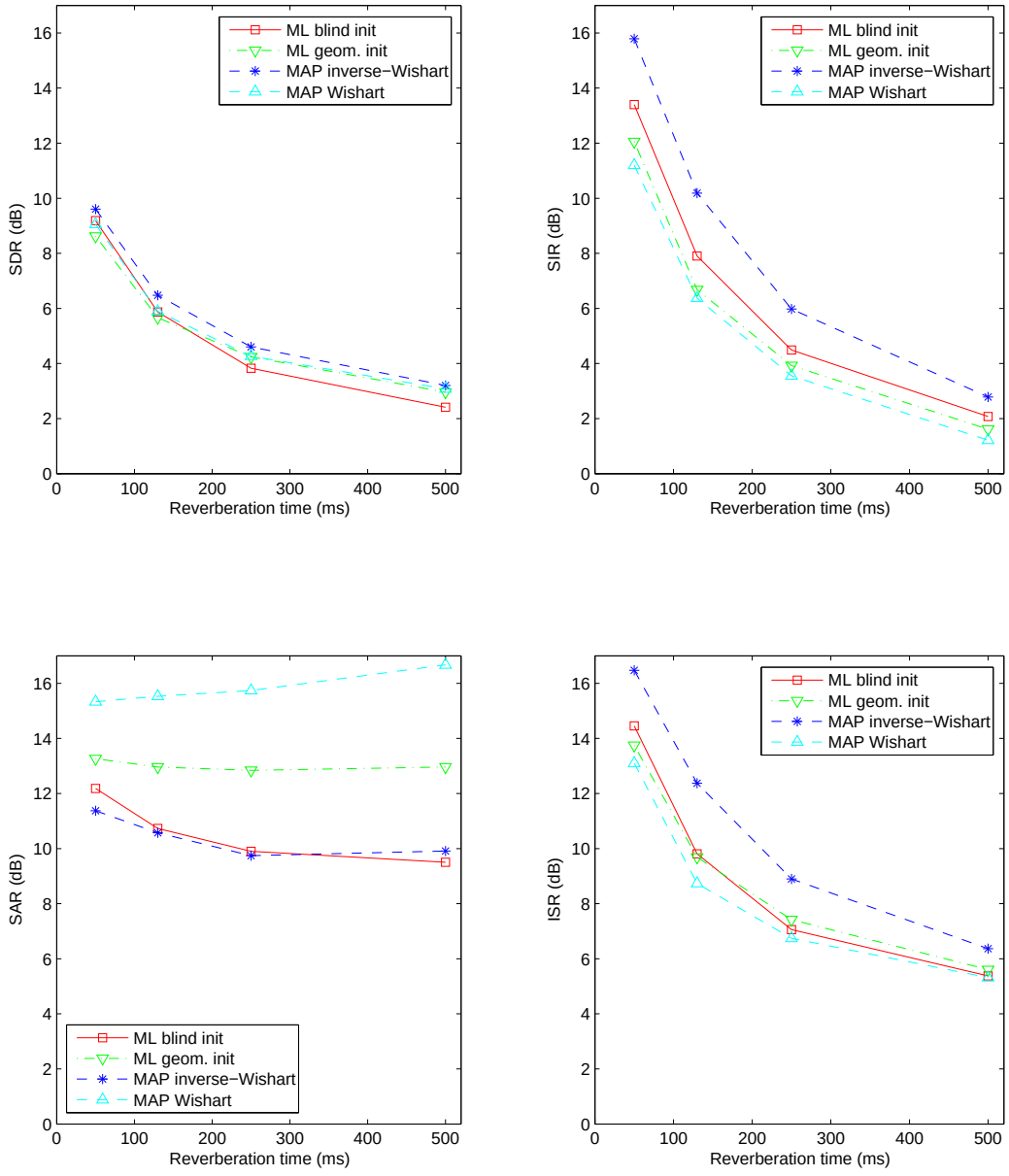


Figure 8.5: ML and MAP results on simulated speech mixtures as a function of the reverberation time.

8.5 Harmonic and percussive musical component separation using spatial and spectral continuity priors

Real-world musical mixtures can be modeled as the sum of two source images: a harmonic source and a percussive source. These two sources typically consist of several instruments, *i.e.* the harmonic source may involve vocals, violin, piano, etc. The separation of these sources from a mixture is useful for remixing [115] and also for various music information retrieval tasks [8, 136, 9]. Hence, our last experiment is devoted to the application of the MAP parameter estimation algorithm exploiting spatial and spectral continuity priors presented in Sections 6.2 and 7.2, respectively, to the separation of these two sources in music mixtures. We jointly exploit the Markov chain inverse-Wishart spatial continuity prior presented in Section 6.2.2 and the Markov chain inverse-gamma spectral continuity priors presented in Section 7.2.1. The resulting spatial and spectral parameter updates are derived in Section 6.2.3 and 7.2.2, respectively. The parameter settings are summarized in Table 8.6 where the hyper-parameters m and α were heuristically fixed depending on the desired shape of the priors, which determines the degree of smoothness.

STFT frame size	4096
STFT frame shift	2048
Number of EM iterations	5
$\mathcal{IW}$ degrees of freedom	$m = 5$
Trade-off parameter for spatial prior	$\gamma = 0.5$
$\mathcal{IG}$ shape parameters	$\alpha = 10$
Trade-off parameter for spectral priors	$\gamma = 1$

Table 8.6: Experimental parameter settings for MAP parameter estimation exploiting spatial and spectral continuity priors

We compared the separation performance of the proposed multichannel harmonic and percussive sound separation algorithm (M-HPSS), where both spatial and spectral continuity priors are exploited, with that given by the single channel HPSS algorithm (HPSS$_{\mathcal{IG}}$), where only the inverse-gamma spectral continuity prior is exploited, and with that achieved by the state-of-the-art algorithm using I-divergence and Gaussian continuity priors (HPSS) [115]. We consider single channel HPSS because, to the best of our knowledge, there was not an algorithm exploiting stereo channel for harmonic and percussive source separation so far. The resulting energy ratio criteria and and perceptually-motivated criteria were averaged over all mixtures of the real-world music

dataset described in Section 8.1.4 and are shown in Table 8.7.

Dataset	Algorithm	Energy ratio criteria (dB)				Perceptually-motivated criteria			
		SDR	SIR	SAR	ISR	OPS	IPS	APS	TPS
Original	HPSS	3.8	5.2	7.6	8.7	27	36	77	45
	HPSS$_{\mathcal{IG}}$	4.8	7.9	8.0	10.7	**30**	42	73	50
	M-HPSS	**5.0**	7.2	8.6	10.1	**30**	39	79	51
Pan+	HPSS	3.8	5.1	7.5	8.6	27	36	77	44
	HPSS$_{\mathcal{IG}}$	4.7	7.7	8.2	10.4	29	41	73	51
	M-HPSS	**5.3**	7.4	8.8	10.3	**31**	40	80	53

Table 8.7: Average harmonic/percussive component separation performance.

The numerical results show the significant separation improvement of HPSS$_{\mathcal{IG}}$ compared to the original HPSS in terms of all criteria but APS over both the original and the Pan+ datasets. This means that the inverse-gamma prior better models the spectral continuity of harmonic and percussive components than the Gaussian prior introduced in [115]. The separation performance given by HPSS and HPSS$_{\mathcal{IG}}$ over the Pan+ dataset is very similar to that over the original dataset due to the fact that panning does not affect the spectral structure of the sources. But the performance achieved by M-HPSS has noticeably increased in the panned dataset, i.e. the SDR is 0.6 dB larger than that given by HPSS$_{\mathcal{IG}}$, showing the benefit of exploiting spatial information. This performance improvements are confirmed by informal listening tests and by the perceptually-motivated criteria.

8.6 Summary of Part 3

Following the general Gaussian modeling framework and the source separation architecture presented in Part 2, we addressed a core problem in this part of the thesis, that is how to estimate the model parameters from a given multichannel mixture signal so that BSS can be performed. We started by designing a general parameter estimation architecture, which accepts both linear and quadratic input T-F representations as well as both ML and MAP approaches. This architecture relies on the EM algorithm, which is appropriate for Gaussian models.

We then derived a ML parameter estimation algorithm consisting of two major steps: parameter initialization and EM parameter updates in Chapter 5. In order to enhance the source separation performance in certain situations where prior information about either the source spectra or the mixing process is available, we introduced several

spatial and spectral priors and derived the corresponding ML/MAP parameter updates in Chapter 6 and Chapter 7.

Finally, in Chapter 8, we performed a number of experiments with different datasets created in both simulated and real-world scenarios to quantify the source separation performance of the proposed algorithms and compare it with state-of-the-art algorithms. The experimental results demonstrate that the proposed algorithms are at or above the level of the state of the art.

Part IV

Conclusion and perspectives

Chapter 9

Conclusion

In this thesis, we presented a radically novel reverberant audio source separation framework grounded on the statistical modeling of spatial cues. The following sections summarize our contributions and discuss the potential extensions of the thesis.

9.1 Conclusions

We targeted the challenging problem of determined and *under-determined* source separation for *reverberated or diffuse* sources. Initially motivated by the Gaussian modeling of late reverberation in physical acoustics, we focused on the probabilistic modeling and parameterization of the reverberant mixing process by opposition to the deterministic parameterization adopted by most state-of-the-art source separation approaches. For that general purpose, we introduce in Chapter 3 a general probabilistic modeling framework whereby the T-F coefficients of the source images in each T-F bin are modeled by multivariate Gaussian random variables parameterized by their *spectral variance* and their *spatial covariance*. We then discussed four specific spatial covariance parameterizations, including the state-of-the-art rank-1 matrix parameterizations, which rely on the narrowband approximation, the full-rank direct+diffuse parameterization, which exploits the theory of statistical room acoustics and was considered in the context of source separation for the first time, and the proposed *full-rank unconstrained* parameterization, which generalizes the former three parameterizationss. The latter parameterization offers the greatest flexibility to model the convolutive mixing process and *overcomes the narrowband approximation, which does not hold for reverberant mixtures, to a certain extent*. Experimental results on simulated speech mixtures in Chapter 4 showed that the full-rank unconstrained spatial covariance parameterization offers the highest upper

bound on the source separation performance at moderate or large reverberation times compared to the other parameterizations and to the baseline approaches. Hence we can conclude that the proposed full-rank unconstrained parameterization better models the reverberant mixing process.

Given the covariance modeling framework and the parameterization, we designed a general source separation architecture including four steps: T-F transform, parameter estimation, Wiener filtering, and inverse T-F transform. We proposed to consider *quadratic* input T-F representations via the empirical mixture covariance $\widehat{\Sigma}_{\mathbf{x}}(n,f)$ to improve the accuracy of parameter estimation compared to the linear STFT representation considered in state-of-the-art approaches. We presented two different quadratic representations corresponding to either a linear frequency scale via the STFT or an auditory-motivated nonlinear frequency scale via the ERB transform. The experiments in Chapter 4 also confirmed the effectiveness of the quadratic T-F representations, which resulted in higher source separation performance than the linear T-F representation.

The latter part of the thesis focused on the investigation of the full-rank unconstrained spatial covariance parameterization for BSS purposes. In order to perform *blind* source separation, the model parameters need to be estimated from the observed mixture signal itself. Hence, Chapter 5 was devoted to a general parameter estimation architecture applicable to both ML and MAP. This architecture accepts both linear or quadratic input T-F representations and was grounded on the EM algorithm, which is well-known as an appropriate choice for Gaussian mixture models. Since EM is very sensitive to the initialization, we proposed suitable procedures adapted from the state of the art to initialize and to align the order of the parameters across all frequency bins. A complete ML parameter estimation algorithm consisting of the initialization step and the estimation by EM updates was then presented. Experimental results revealed that the algorithm converges in about 10 iterations.

Chapter 6 and Chapter 7 completed this probabilistic framework by defining prior distributions over the spatial and spectral parameters so as to improve the robustness of parameter estimation with small amounts of data and to address the initialization and the permutation problem in a probabilistically principled fashion. Two classes of spatial priors were presented in Chapter 6: the spatial location priors exploited knowledge from the theory of statistical room acoustics assuming a known geometric setting, and the spatial continuity prior opened a new way to deal with *moving sources* whose spatial location was modeled via time-varying spatial covariances. In Chapter 7, we have investigated the spectral structures of the sources, *i.e.* a NMF spectral model and spectral continuity priors, and the resulting ML/MAP parameter estimation

algorithms.

An extensive set of experiments was presented in Chapter 8 to provide a comprehensive comparison of the source separation performance of the proposed algorithms with state-of-the-art algorithms and baseline approaches in various mixing scenarios ranging from simulated speech mixtures to real-world speech and music datasets. Experimental results, evaluated by both the widely used energy ratio criteria and the recently proposed perceptually-motivated criteria, prove the effectiveness of the proposed framework. The full-rank unconstrained spatial covariance parameterization improves the separation performance compared to rank-1 models and most state-of-the-art algorithms in realistic reverberant environments. The potential benefit of the proposed prior distributions over spatial and spectral parameters was also confirmed. For instance, the resulting MAP algorithms exploiting the spatial location priors presented in Chapter 6 provide a proof of concept of the benefit of the proposed priors towards their future use in a blind source separation context, and the resulting MAP algorithm exploiting spatial and spectral continuity priors improves the separation performance of harmonic and percussive sources in music mixtures.

9.2 Perspectives

A number of directions built upon the proposed framework could be explored in future work. Let us list them in the following.

Efficient initialization and estimation for ML algorithm

Accurate initialization of the spatial parameters is crucial for the success of the ML parameter estimation algorithm presented in Chapter 5. Our preliminary experiments have shown that the spatial parameters initially estimated by hierarchical clustering vary depending on the heuristic choice of the number of clusters K. Also the performance of the chosen permutation alignment method decreases when the microphone spacing is large due in particular to phase wrapping. Therefore, a more efficient initialization scheme appears necessary. Experiments with the real-world speech mixtures in Section 8.3.2 suggest a potential use of Sawada's bin-wise probabilistic clustering method [125] for this purpose. Another possibility would be to first estimate the source DoAs, for example by the algorithm in [137], then to find a better way of exploiting this information for initialization rather than simple DoA-based binary masking in Section 5.2.1 which was found not to work.

In addition to the initialization, the estimation of the spatial covariance matrices

$\mathbf{R}_j(f)$ from the T-F bins involving a single active source could also be considered. Indeed, it has been shown for instantaneous mixtures that the detection of these T-F bins improves the accuracy of the estimated mixing vectors [57]. However, this extension is not trivial in the convolutive case and therefore offers an interesting research on how to detect these single source T-F bins.

Efficient online algorithm

In order to employ source separation for real-time systems, the model parameters must be estimated in an online manner and the computation burden should be decreased. Tomagi et al. have recently extended our batch ML algorithm to an online algorithm and demonstrated its potential [36]. However their approach strongly relied on the good parameter initialization and the spatial parameters were heuristically treated without updating them at each EM iteration, *i.e.* only the source variances were updated. Therefore, any research effort to develop an online algorithm from our proposed framework would potentially bring good results.

Extension toward the use of spatial priors

The investigation of the spatial location priors in Chapter 6 has provided a proof of concept of their benefit in a blind source separation context. However, at this stage, the geometric setting and the reverberation time must be known to specify the prior distribution. Future work can hence be devoted to the estimation of these parameters from the mixture signal itself so as to perform blind source separation.

We have confirmed that the proposed full-rank unconstrained spatial covariance parameterization better accounts for reverberation and small source movements. Though addressing moving sources goes beyond the scope of this thesis, the time-varying spatial covariance parameterization and the spatial continuity prior presented in Chapter 6 have opened a promising extension of the proposed framework toward addressing more difficult recording scenarios involving moving speakers.

Extension toward the use of spectral priors

Chapter 7 has investigated the combination of the spatial covariance with an NMF spectral source model. However, depending on different types of audio sources, other spectral models such as GMM, GSMM, or HMM possibly provide a good combination to improve source separation performance. Similarly, different spectral priors rather than the spectral continuity priors investigated in Chapter 7, *e.g.* prior enforcing the harmonic structure, could also be considered for specific applications in the future work.

Most of these priors has recently been considered in [37].

Extension toward a more advanced direct+early echoes+diffuse model

The direct+diffuse spatial covariance parameterization introduced in Section 3.1.2.3 offers a promising modeling of reverberated sources, especially when considering the small number of spatial parameters it involves. However, this model is generally not valid for each individual source, as shown in Fig. 3.1b, due in particular to the non-uniformly distributed early echoes. Therefore, a potential research could be devoted to a novel model distinguishing 3 parts: the direct part, early echoes, and the late reverberation part. This improved parameterization could require the design of a model for echoes, which does not exist today.

Extension to other fields

The thesis has addressed the general modeling of reverberated or diffuse sound sources. This is a core problem with applications in many areas of audio signal processing, including speech enhancement, audio information retrieval and spatial audio processing. As an example, in Chapter 8 we have demonstrated the benefit of the proposed framework for the separation of harmonic and percussive sound in music mixtures, which can be exploited as a pre-processing step for some MIR tasks. Togami et al. have also recently applied our full-rank covariance modeling framework in a multichannel acoustic echo cancelation system [35]. Therefore, in a long-term research direction we would like to apply the framework presented in this thesis to other audio processing fields where the modeling of reverberated or diffuse sound sources is necessary.

Part V

Appendices

Appendix A

Derivation of the EM algorithm

This appendix present the derivation of the EM algorithms for the estimation of the model parameters $\boldsymbol{\theta} = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n}$ given the empirical mixture covariances $\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\}_{n,f}$ in each frequency bin f of both the rank-1 convolutive spatial covariance parameterization presented in Section 5.3.1 and the full-rank unconstrained parameterization presented in Section 5.3.2. The detailed derivations for MAP parameter updates in Chapters 6 and 7 are also provided.

A.1 ML estimation for the rank-1 convolutive parameterization

A.1.1 E-step

The mixing model is described by equation (5.7). The source posterior distribution is given by

$$\begin{aligned}\log p(\mathbf{s}(n,f)|\mathbf{x}(n,f),\boldsymbol{\theta}) &\stackrel{\text{c}}{=} \log p\big(\mathbf{x}(n,f)|\mathbf{s}(n,f),\boldsymbol{\theta}\big) + \log p\big(\mathbf{s}(n,f)|\boldsymbol{\theta}\big) \\ &= \log \mathcal{N}_c\big(\mathbf{x}(n,f) - \mathbf{H}(f)\mathbf{s}(n,f)|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}(f)\big) + \log \mathcal{N}_c\big(\mathbf{s}(n,f)|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\big) \qquad \text{(A.1)} \\ &= \log \mathcal{N}_c\big(\mathbf{s}(n,f)|\hat{\mathbf{s}}(n,f), \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}(n,f)\big) \qquad \text{(A.2)}\end{aligned}$$

By computing the gradient of (A.1) with respect to $\mathbf{s}^H(n,f)$ and equating it to zero, we obtain $\hat{\mathbf{s}}(n,f)$ as

$$\hat{\mathbf{s}}(n,f) = \mathbf{W}(n,f)\mathbf{x}(n,f) \qquad \text{(A.3)}$$

where the Wiener filter $\mathbf{W}(n,f)$ is computed by (5.10).

By computing the second derivative $\frac{\partial^2}{\partial \mathbf{s}^H(n,f)\mathbf{s}(n,f)}$ of (A.1) and (A.2), we obtain

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}(n,f) &= \big(\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\mathbf{H}^H(f)\mathbf{H}(f) + \boldsymbol{\Sigma}_{\mathbf{s}}^{-1}(n,f)\big)^{-1} \\ &= \boldsymbol{\Sigma}_{\mathbf{s}}(n,f) - \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)\big(\mathbf{H}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f) + \boldsymbol{\Sigma}_{\mathbf{b}}\big)^{-1}\mathbf{H}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\end{aligned} \tag{A.4}$$

with the small noise variance ($\boldsymbol{\Sigma}_{\mathbf{b}} \to \mathbf{0}$), (A.4) can be rewritten as

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}(n,f) = \big(\mathbf{I} - \mathbf{W}(n,f)\mathbf{H}(f)\big)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) \tag{A.5}$$

The conditional expectation of the natural statistic $\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)$ is therefore given by equation (5.15) since $\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f) = \mathbb{E}\{\hat{\mathbf{s}}(n,f)\hat{\mathbf{s}}^H(n,f)\} + \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}(n,f)$.

A.1.2 M step

Log-likelihood of the *complete data* $C = \{\mathbf{x}(n,f), \mathbf{s}(n,f)\}_n$ is

$$\begin{aligned}\log \mathcal{L}_C &= \sum_n \log p\big(\mathbf{x}(n,f), \mathbf{s}(n,f)|\boldsymbol{\theta}\big) \\ &= \sum_n \log p\big(\mathbf{x}(n,f)|\mathbf{s}(n,f), \boldsymbol{\theta}\big) + \log p\big(\mathbf{s}(n,f)|\boldsymbol{\theta}\big) \\ &\overset{c}{=} \sum_n \log \mathcal{N}_c\big(\mathbf{x}(n,f) - \mathbf{H}(f)\mathbf{s}(n,f)|\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}}(f)\big) \\ &\overset{c}{=} -N\mathrm{tr}\Big[\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}(f)\Big(\mathbf{R}_{\mathbf{xx}}(f) - \mathbf{H}(f)\mathbf{R}_{\mathbf{xs}}^H(f) - \mathbf{R}_{\mathbf{xs}}(f)\mathbf{H}^H(f) + \mathbf{H}(f)\mathbf{R}_{\mathbf{ss}}(f)\mathbf{H}^H(f)\Big)\Big] \\ &\quad - N\log|\boldsymbol{\Sigma}_{\mathbf{b}}(f)|\end{aligned} \tag{A.6}$$

where the *natural statistics* are defined as

$$\begin{aligned}\mathbf{R}_{\mathbf{xx}}(f) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}(n,f)\mathbf{x}^H(n,f) \\ \mathbf{R}_{\mathbf{xs}}(f) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}(n,f)\mathbf{s}^H(n,f) \\ \mathbf{R}_{\mathbf{ss}}(f) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{s}(n,f)\mathbf{s}^H(n,f)\end{aligned} \tag{A.7}$$

By replacing the natural statistics by their conditional expectations with respected to the estimated parameters in the E-step and computing the gradient of this expression

with respect to $\mathbf{H}(f)$ and $\mathbf{\Sigma_b}(f)$ and equating them to zero, we obtain the parameter updates (5.13) and (5.14), where $\widehat{\mathbf{R}}_{\mathbf{ss}}(f)$, $\widehat{\mathbf{R}}_{\mathbf{xs}}(f)$, and $\widehat{\mathbf{R}}_{\mathbf{xx}}(f)$ are computed by (5.15), (5.16), and (5.17), respectively.

A.2 ML estimation for the full-rank spatial covariance parameterization

A.2.1 E-step

Let us denote by $\tilde{\mathbf{s}}(n,f) = [\mathbf{c}_1(n,f)^T, ..., \mathbf{c}_J(n,f)^T]^T$ and $\tilde{\mathbf{H}} = [\mathbf{I}, ..., \mathbf{I}]$ the $I \times IJ$ matrix. The mixing model (1.3) is then equivalent to $\mathbf{x}(n,f) = \tilde{\mathbf{H}}\tilde{\mathbf{s}}(n,f)$, which is similar to (5.7) when $\mathbf{b} \to \mathbf{0}$. Therefore, the computation of the Wiener filter by (5.19) and that of the covariance $\widehat{\mathbf{\Sigma}}_j(n,f)$ of the spatial image of the jth source by (5.20) in the E-step are directly inherited from the computation of the Wiener filter and the the expectation of the natural statistics $\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)$ for the rank-1 convolutive parameterization.

A.2.2 M-step

The M-step maximizes the log-likelihood of the complete data defined in (4.1) to obtain the iterative parameter update. The partial derivatives of this log-likelihood with respect to $v_j(n,f)$ and to each entry (k,l) of $\mathbf{R}_j(f)$ are (see Appendix D for the gradient computation)

$$\frac{\partial \log \mathcal{L}_j}{\partial v_j(n,f)} = \frac{\mathrm{tr}\big(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)\big)}{v_j^2(n,f)} - \frac{I}{v_j(n,f)} \tag{A.8}$$

$$\frac{\partial \log \mathcal{L}_j}{\partial \mathbf{R}_j(f)_{kl}} = \mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\mathbf{E}_{kl}\left(\mathbf{R}_j^{-1}(f)\sum_n \frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)} - N\right)\right) \tag{A.9}$$

where $\mathbf{E}_{kl} = \frac{\partial \mathbf{R}_j(f)}{\partial \mathbf{R}_j(f)_{kl}}$ is an $I \times I$ matrix whose (k,l)-th entry is 1 and other entries are zero. By solving the equations $\frac{\partial \log \mathcal{L}_j}{\partial v_j(n,f)} = 0$ and $\frac{\partial \log \mathcal{L}_j}{\partial \mathbf{R}_j(f)_{kl}} = 0$ for all k,l, we obtain iterative updates for $v_j(n,f)$ and $\mathbf{R}_j(f)$ as

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \tag{A.10}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f). \tag{A.11}$$

A.3 MAP estimation using a Wishart spatial location prior

By computing the partial derivatives of the expectation of the log-posterior of the complete data defined in (6.6) with respect to each entry of $\mathbf{R}_j(n,f)$ and equating them to zero, we obtain the quadratic matrix equation

$$\mathbf{R}_j(f)\mathbf{A}\mathbf{R}_j(f) + b\mathbf{R}_j(f) + \mathbf{C} = \mathbf{0} \tag{A.12}$$

where

$$\begin{aligned} \mathbf{A} &= \gamma\boldsymbol{\Psi}_j^{-1}(f) \\ b &= -\gamma(m-I)+N \\ \mathbf{C} &= \sum_{n=1}^{N} -\frac{\widehat{\boldsymbol{\Sigma}}_j(n,f)}{v_j(n,f)}. \end{aligned} \tag{A.13}$$

After the variable exchange $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{R}_j(f)\mathbf{A}^{1/2}$ (when $\gamma \neq 0$), (A.12) is equivalent to

$$\mathbf{X}^2 + b\mathbf{X} + \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{0} \tag{A.14}$$

The first two coefficients of quadratic matrix equation are scalar multiples of the identity matrix and the last one is positive semi-definite. Therefore this equation can be solved [138, p. 304]. The unique positive definite solution is given by

$$\mathbf{X} = \frac{1}{2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}\big) \tag{A.15}$$

Therefore

$$\mathbf{R}_j(f) = \frac{1}{2}\mathbf{A}^{-1/2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}\big)\mathbf{A}^{-1/2}. \tag{A.16}$$

A.4 MAP estimation using a Markov chain inverse-Wishart spatial continuity prior

By computing the partial derivatives of the expectation of the log-posterior of the complete data defined in (6.16) with respect to each entry of $\mathbf{R}_j(n,f)$ and equating them to zero, we obtain the quadratic matrix equation

$$\mathbf{R}_j(n,f)\mathbf{A}\mathbf{R}_j(n,f) + b\mathbf{R}_j(n,f) + \mathbf{C} = \mathbf{0} \tag{A.17}$$

where $\mathbf{A}$, b, and $\mathbf{C}$ are given by

$$\begin{aligned}\mathbf{A} &= \gamma(m_j - I)\mathbf{R}_j^{-1}(n+1, f)\\ b &= \gamma I + 1\\ \mathbf{C} &= -\widehat{\mathbf{\Sigma}}_j(n, f)/v_j(n, f) - \gamma(m_j - I)\mathbf{R}_j(n-1, f)\end{aligned} \tag{A.18}$$

By the variable change $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{R}_j(n, f)\mathbf{A}^{1/2}$ we obtain the equivalent quadratic matrix equation

$$\mathbf{X}^2 + b\mathbf{X} + \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{0} \tag{A.19}$$

Again, the first two coefficients are scalar multiples of the identity and the last one is positive semi-definite. Therefore this equation can be solved [138, p. 304]. The unique positive definite solution is given by

$$\mathbf{R}_j(n, f) = \frac{1}{2}\mathbf{A}^{-1/2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}\big)\mathbf{A}^{-1/2}. \tag{A.20}$$

A.5 MAP estimation using a Markov chain inverse-gamma spectral continuity prior

By computing the partial derivative of $Q_{\mathcal{JG}}(\theta|\theta^{\text{old}})$ defined in 7.8 with respect to $v_j(n, f)$ and equating it to zero, we obtain a second order polynomial equation

$$av_j^2(n, f) + bv_j(n, f) + c = 0 \tag{A.21}$$

where

$$\begin{aligned}a &= \begin{cases}\gamma(\alpha-1)/v_j(n+1, f) & \text{for harmonic sources}\\ \gamma(\alpha-1)/v_j(n, f+1) & \text{for percussive sources}\end{cases}\\ b &= I\\ c &= \begin{cases}-\text{tr}\big(\mathbf{R}_j^{-1}(n, f)\widehat{\mathbf{\Sigma}}_j(n, f)\big) - \gamma(\alpha-1)v_j(n-1, f) & \text{for harmonic sources}\\ -\text{tr}\big(\mathbf{R}_j^{-1}(n, f)\widehat{\mathbf{\Sigma}}_j(n, f)\big) - \gamma(\alpha-1)v_j(n, f-1) & \text{for percussive sources.}\end{cases}\end{aligned} \tag{A.22}$$

This equation has single positive solution, that is

$$v_j(n, f) = (-b + \sqrt{b^2 - 4ac})/(2a). \tag{A.23}$$

Appendix B

Derivation of the normalized cross-correlation function of diffuse signals

This appendix presents the derivation of the normalized cross-correlation function in (3.10) to validate the full-rank direct+diffuse spatial covariance paramerization in (3.9).

We assume that the intensities of incident sounds are uniformly distributed over all possible directions in the 3D space. Under a far-field assumption, the complex-valued signals at the microphone pair (i, l) coming from a given direction (see Fig. B.1) can be written as [27]

$$\begin{aligned} x_i(t) &= A_0 e^{j(2\pi f t - \phi)} \\ x_l(t) &= A_0 e^{j(2\pi f t - \phi - 2\pi f d_{il}\cos(\theta)/c)} \end{aligned} \tag{B.1}$$

where A_0, d_{il}, ϕ, θ denotes amplitude, microphone spacing, phase, and sound incident direction, respectively.

The normalized cross-correlation function of $x_1(t)$ and $x_2(t)$ is

$$\begin{aligned} \Omega_{il,\theta}(f) &= \frac{x_1(t)x_2^*(t)}{\sqrt{E\{|x_1(t)|^2\}E\{|x_2(t)|^2\}}} \\ &= e^{j2\pi f d\cos(\theta)/c} \end{aligned} \tag{B.2}$$

By averaging $\Omega_{il,\theta}(f)$ over all directions under the diffuse sound assumption (see Fig. B.2), we obtain

B. DERIVATION OF THE NORMALIZED CROSS-CORRELATION FUNCTION OF DIFFUSE SIGNALS

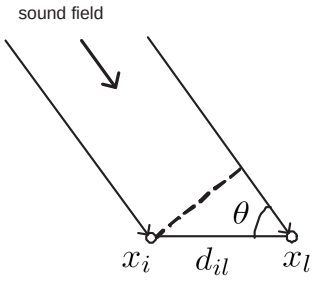


Figure B.1: Parameterization of the incident sound field

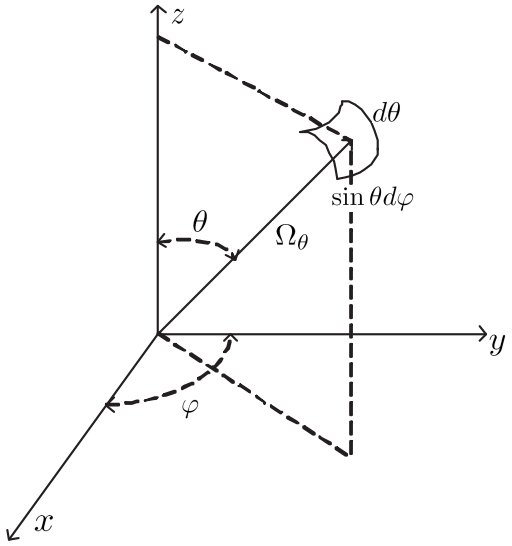


Figure B.2: 3D integration of the sound field covariance

$$\begin{aligned}\Omega_{il}^{\text{diff}}(f) &= \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\Omega_{il,\theta}(f)\sin(\theta)d\varphi d\theta \\ &= \frac{1}{4\pi}\int_0^{2\pi}d\varphi\int_0^{\pi}e^{j2\pi f d_{il}\cos(\theta)/c}\sin(\theta)d\theta \\ &= \frac{\sin(2\pi f d_{il}/c)}{2\pi f d_{il}/c}.\end{aligned} \tag{B.3}$$

Note that the normalized cross-correlation function is real-valued even with complex signals.

Appendix C

Formulas of some considered probability distributions

C.1 Multivariate complex Gaussian distribution

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\pi\boldsymbol{\Sigma})} e^{-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{C.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and the covariance matrix of $\mathbf{x}$, respectively.

C.2 Complex Wishart distribution

$$p(\mathbf{R}|\boldsymbol{\Psi}, m) = \frac{|\boldsymbol{\Psi}|^{-m}|\mathbf{R}|^{m-I} e^{-\mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{R})}}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(m-i+1)} \tag{C.2}$$

where $\boldsymbol{\Psi}$ is an $I \times I$ positive-definite Hermitian matrix and m is a positive scalar (number of degrees of freedom).

C.3 Complex inverse-Wishart distribution

$$p(\mathbf{R}|\boldsymbol{\Psi}, m) = \frac{|\boldsymbol{\Psi}|^{m}|\mathbf{R}|^{-(m+I)} e^{-\mathrm{tr}(\boldsymbol{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(m-i+1)} \tag{C.3}$$

where $\boldsymbol{\Psi}$ is an $I \times I$ positive-definite Hermitian matrix and m is a positive scalar (number of degrees of freedom).

C.4 Inverse gamma distribution

$$p(v|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} v^{-\alpha-1} e^{-\beta/v} \tag{C.4}$$

where α and β are shape parameter and scale parameter, respectively.

Appendix D

Differentiation formulas

Differentiation of a number of expressions with respect to a matrix $\mathbf{X}$ [139]:

$$\partial(\alpha\mathbf{X}) = \alpha\partial\mathbf{X} \tag{D.1}$$
$$\partial(\mathbf{XY}) = (\partial\mathbf{X})\mathbf{Y} + \mathbf{X}(\partial\mathbf{Y}) \tag{D.2}$$
$$\partial(\mathrm{tr}(\mathbf{X})) = \mathrm{tr}(\partial\mathbf{X}) \tag{D.3}$$
$$\partial(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1} \tag{D.4}$$
$$\partial(\det(\mathbf{X})) = \det(\mathbf{X})\mathrm{tr}(\mathbf{X}^{-1}\partial\mathbf{X}) \tag{D.5}$$
$$\partial(\log(\det(\mathbf{X}))) = \mathrm{tr}(\mathbf{X}^{-1}\partial\mathbf{X}) \tag{D.6}$$

Differentiation of a number of expressions with respect to the real and imaginary parts of a complex-valued vector $\mathbf{a}$

$$\frac{\partial\mathbf{b}^H\mathbf{a}}{\partial\Re\mathbf{a}} = \mathbf{b}^* \quad ; \quad \frac{\partial\mathbf{b}^H\mathbf{a}}{\partial\Im\mathbf{a}} = i\mathbf{b}^* \tag{D.7}$$
$$\frac{\partial\mathbf{a}^H\mathbf{b}}{\partial\Re\mathbf{a}} = \mathbf{b} \quad ; \quad \frac{\partial\mathbf{a}^H\mathbf{b}}{\partial\Im\mathbf{a}} = -i\mathbf{b} \tag{D.8}$$
$$\frac{\partial\mathbf{a}^H\mathbf{W}\mathbf{a}}{\partial\Re\mathbf{a}} = \mathbf{W}^T\mathbf{a}^* + \mathbf{W}\mathbf{a} \quad ; \quad \frac{\partial\mathbf{a}^H\mathbf{W}\mathbf{a}}{\partial\Im\mathbf{a}} = i\big(\mathbf{W}^T\mathbf{a}^* - \mathbf{W}\mathbf{a}\big) \tag{D.9}$$

Résumé étendu

Introduction

Dans la vie de tous les jours, les scènes sonores résultent souvent du mélange de plusieurs sources et requièrent la capacité de localiser, identifier et différencier les sources sonores présentes simultanément. Alors que les humains sont généralement très bons pour cette tâche, les systèmes numériques réussissent moins bien à séparer les sources sonores, particulièrement dans les environnements réverbérants présentant de nombreuses réflexions sur les murs et/ou du bruit de fond [1]. Par conséquent, cette thèse se concentre sur le problème épineux de la séparation aveugle de sources (SAS), qui vise à extraire les signaux sources d'un mélange observé avec peu d'information *a priori*, dans un environnement réverbérant réel.

Formulation du problème

Considérons un signal de mélange multicanal $\mathbf{x}(t) = [x_1(t), ..., x_I(t)]^T$ enregistré par une antenne de I microphones. Ce mélange peut être exprimé comme [14]

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t)$$

où J dénote le nombre de sources et $\mathbf{c}_j(t) = [c_{1j}(t), ..., c_{Ij}(t)]^T$ est l'*image spatiale* de la j-ième source, c'est-à-dire sa contribution à l'ensemble des canaux du mélange. Notons que le bruit de fond est aussi considéré comme une source dans l'équation ci-dessus. Dans le cas où la j-ième source est une *source ponctuelle*, c'est-à-dire qu'elle émet du son en un point unique de l'espace, $\mathbf{c}_j(t)$ se caractérise par [15]

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t - \tau)$$

où $\mathbf{h}_j(\tau) = [h_{1j}(\tau), ..., h_{Ij}(\tau)]^T$ est un *filtre de mélange* linéaire modélisant la propagation acoustique de la j-ième source à l'ensemble des I microphones et $s_j(t)$ est le signal source monophonique émis. La séparation de sources consiste à estimer soit les J signaux sources originaux $s_j(t)$ soit leurs images spatiales $\mathbf{c}_j(t)$ à partir des I canaux du signal de mélange $\mathbf{x}(t)$. Nous considérons ce deuxième problème dans cette thèse, en supposant que le nombre de sources J est connu.

État de l'art

Les systèmes de SAS exploitent généralement un ensemble d'indices pour discriminer les sources portant sur leur position spatiale ou sur leur structure spectrale [16, 15]. Ces deux types d'indices sont appelés respectivement *indices spatiaux*, tels que la différence de temps d'arrivée (ITD) et d'intensité (IID) entre les canaux, et *indices spectraux*, tels que la parcimonie, l'harmonicité, la continuité spectrale ou temporelle, *etc.* Afin de modéliser et d'exploiter ces indices, la plupart des approches de SAS opèrent dans le domaine temps-fréquence (T-F) par l'intermédiaire de la transformée de Fourier à court terme (TFCT), de sorte que le mélange peut se réécrire en chaque bande de fréquence f et chaque trame temporelle n comme

$$\mathbf{x}(n, f) = \sum_{j=1}^{J} \mathbf{c}_j(n, f).$$

Ces approches reposent de plus sur l'*approximation de bande étroite* de la convolution par une multiplication complexe en chaque point T-F

$$\mathbf{c}_j(n, f) \approx \mathbf{h}_j(f) s_j(n, f)$$

où le vecteur de mélange $\mathbf{h}_j(f)$ de taille $I \times 1$ est la transformée de Fourier de $\mathbf{h}_j(\tau)$, $s_j(n, f)$ sont les coefficients de TFCT des sources $s_j(t)$ et $\mathbf{c}_j(n, f) = [c_{1j}(n, f), ..., c_{Ij}(n, f)]^T$ les coefficients de TFCT de leurs images spatiales $\mathbf{c}_j(t)$.

Étant donnée cette formulation du mélange dans le domaine T-F, la plupart des approches existantes consistent à construire un modèle des indices spatiaux, c'est-à-dire de $\mathbf{h}_j(f)$, et des indices spectraux, c'est-à-dire de $|s_j(n, f)|$, et à estimer ses paramètres. Les approches connues incluent l'analyse en composantes indépendantes (ACI) [40, 41, 45, 46, 47, 48, 49, 50, 51, 52], l'analyse en composantes parcimonieuses [53, 18, 55, 56, 57], les modèles gaussiens non-stationnaires [71, 72, 21], *etc.* La performance de séparation atteignable par ces techniques reste cependant limitée pour des

sources diffuses ou réverbérées [12] car l'approximation de bande étroite n'est pas valable dans ce cas.

Contributions

Nous nous focalisons dans cette thèse sur la modélisation et la paramétrisation probabilistes du processus de mélange pour des sources diffuses ou réverbérées. Les résultats obtenus ont été décrits dans nos publications [28, 29, 30, 31, 32, 33, 34]. Nos contributions majeures sont de proposer un nouveau cadre de modélisation probabiliste pour les indices spatiaux qui contourne l'approximation de bande étroite dans une certaine mesure [28, 30], de concevoir une architecture générale pour l'estimation des paramètres et la séparation de sources [29, 30, 31], de proposer des *a priori* probabilistes sur les paramètres spatiaux et spectraux, et de démontrer le potentiel du cadre proposé par rapport aux algorithmes de l'état de l'art dans différents scénarios incluant des données synthétiques simulées comme des enregistrements réels [32, 33, 34].

La structure du reste de ce résumé est comme suit. Nous introduisons d'abord le cadre de modélisation général et la paramétrisation proposée. Ensuite, nous abordons l'estimation des paramètres du modèle au sens du maximum de vraisemblance (MV) ou du maximum *a posteriori* (MAP). Nous comparons la performance de séparation de sources obtenue à celle des algorithmes de l'état de l'art. Enfin, nous tirons les conclusions et discutons les pistes de recherche futures.

Cadre de modélisation général

Cadre de modélisation et architecture de séparation de sources

Notre cadre de modélisation s'appuie sur le cadre de modélisation gaussienne non-stationnaire [72, 68] où le vecteur de coefficients T-F $\mathbf{c}_j(n,f)$ de l'image spatiale de chaque source est modélisé comme un vecteur aléatoire gaussien de moyenne nulle et de matrice de covariance $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$. De plus, nous factorisons $\mathbf{\Sigma}_j(n,f)$ comme

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(n,f)$$

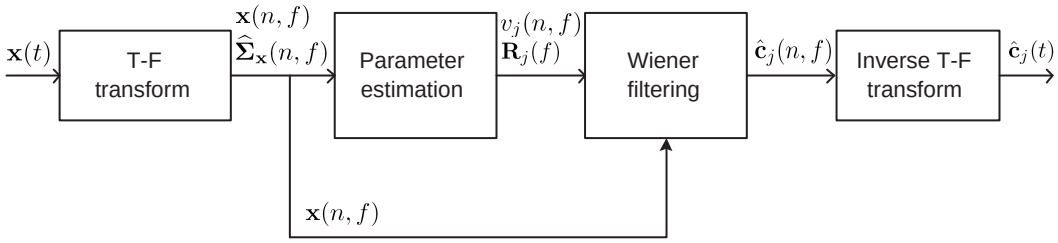
où $v_j(n,f)$ sont des *variances* non-stationnaires représentant le spectre de puissance à court terme de la source et $\mathbf{R}_j(n,f)$ est une *matrice de covariance spatiale* de taille $I \times I$ représentant sa position et sa largeur spatiale. **Cette paramétrisation est probabiliste dans le sens que $\mathbf{c}_j(n,f)$ ne peut pas être calculé de façon déterministe à partir des paramètres choisis, mais est généré de façon aléatoire**

selon la distribution gaussienne considérée. Ce cadre de modélisation ne repose ni sur l'hypothèse de sources ponctuelles ni sur l'approximation de bande étroite, de sorte qu'il apparaît applicable à des sources réverbérées ou diffuses.

Sous l'hypothèse classique que les sources sont décorrélées, le vecteur $\mathbf{x}(n,f)$ de coefficients de TFCT du mélange est aussi gaussien de moyenne nulle et sa matrice de covariance est égale à

$$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(n,f).$$

À partir de ce modèle, la séparation de sources peut être effectuée selon les quatre étapes décrites dans la figure ci-dessous.



General source separation architecture.

Le problème central consiste à estimer les paramètres du modèle $\theta = \{v_j(n,f), \mathbf{R}_j(n,f)\}_{j,n,f}$ en ajustant la covariance du modèle $\mathbf{\Sigma}_{\mathbf{x}}(n,f)$ à la covariance empirique observée $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ selon la fonction de vraisemblance

$$P(\widehat{\mathbf{\Sigma}}_{\mathbf{x}}|v,\mathbf{R}) = \prod_{n,f} \frac{1}{\det\left(\pi \mathbf{\Sigma}_{\mathbf{x}}(n,f)\right)} e^{-\mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)\right)}.$$

La covariance empirique $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ s'apparente à une représentation T-F *quadratique* préalablement calculée en moyennant localement au voisinage de chaque point T-F [68]

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f) := \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^H(n',f')$$

où w_{nf} est une fenêtre bi-dimensionnelle spécifiant la forme du voisinage telle que $\sum_{n',f'} w_{nf}^2(n',f') = 1$. Une fois les paramètres du modèle estimés, les images spatiales de toutes les sources sont obtenues au sens du minimum de l'erreur quadratique

moyenne (MMSE) par filtrage de Wiener multicanal [71, 72]

$$\hat{\mathbf{c}}_j(n,f) := v_j(n,f)\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f).$$

Paramétrisations de la covariance spatiale

Nous présentons quatre paramétrisations de la covariance spatiale correspondant à des matrices soit de rang 1 soit de rang plein. Les deux premières paramétrisations de rang 1 correspondent à des techniques existantes. La troisième appelée paramétrisation de rang plein directe+diffuse est considérée pour la première fois dans le contexte de la séparation de sources, bien qu'elle ait déjà été utilisée pour la localisation de sources. La quatrième paramétrisation appelée paramétrisation de rang plein non contrainte est la paramétrisation proposée, qui offre la plus grande flexibilité pour modéliser le processus de mélange par rapport aux trois autres paramétrisations.

Paramétrisation de rang 1 anéchoïque ou convolutive

Les paramétrisations de rang 1 résultent de l'approximation de bande étroite. La matrice de covariance spatiale prend alors la forme suivante pour des mélanges convolutifs:

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$$

où $\mathbf{h}_j^H(f)$ est le vecteur de mélange défini précédemment.

Pour des mélanges anéchoïques, le vecteur de mélange est dénoté par $\mathbf{h}_j^{\text{ane}}(f)$ et est une fonction directe de la distance entre les sources et les microphones [102].

Paramétrisation de rang plein directe+diffuse

La paramétrisation de rang plein directe+diffuse s'inspire de la théorie statistique de l'acoustique des salles dans laquelle l'image spatiale de chaque source est supposée être la somme de deux parties décorrélées: une partie directe modélisée par $\mathbf{h}_j^{\text{ane}}(f)$ et une partie diffuse provenant de toutes les directions de l'espace uniformément [27]. La matrice de covariance spatiale $\mathbf{R}_j(f)$ de chaque source est alors une matrice de rang plein définie comme la somme de la covariance de sa partie directe et de la covariance de sa partie diffuse

$$\mathbf{R}_j(f) = \mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}\big)^H(f) + \sigma_{\text{rev}}^2\mathbf{\Omega}(f)$$

où σ_{rev}^2 est la variance de la partie diffuse et $\Omega_{il}(f)$ est une fonction du diagramme de directivité des microphones et de la distance d_{il} entre le i-ième et le l-ième microphone.

À la fois σ^2_{rev} et $\Omega_{il}(f)$ sont calculables en fonction de la disposition spatiale des sources et des caractéristiques de la salle sous certaines hypothèses [27, 102]. Par exemple, dans le cas de microphones omnidirectionnels, $\Omega_{il}(f) = \frac{\sin(2\pi f d_{il}/c)}{2\pi f d_{il}/c}$.

Paramétrisation de rang plein non contrainte

Nous proposons une paramétrisation de rang plein non contrainte, où $\mathbf{R}_j(f)$ est une matrice hermitienne positive semi-définie de rang plein dont les coefficients ne sont pas contraints *a priori*. Cette paramétrisation non contrainte est la plus générale possible pour une matrice de covariance. Elle généralise les trois paramétrisations ci-dessus dans le sens où n'importe quelle matrice contrainte de rang 1 ou de rang plein peut être considérée comme une forme particulière de matrice non contrainte. En raison de sa flexibilité accrue, cette paramétrisation non contrainte épouse mieux les données en terme de vraisemblance. En particulier, elle améliore l'ajustement entre le modèle et les données sur des mélanges réverbérants par rapport aux modèles de rang 1 car l'approximation de bande étroite n'est pas valable dans ce cas. **Dans ce sens, la paramétrisation proposée contourne l'approximation de bande étroite dans une certaine mesure**.

Nous avons montré dans [28, 30] que cette paramétrisation offre la meilleure borne supérieure de performance de séparation de sources, en supposant $\mathbf{R}_j(f)$ connue, par rapport aux autres paramétrisations et à deux approches de base, le masquage binaire [18] et la minimisation de norme ℓ_1 [55]. Par la suite, nous nous focalisons sur l'estimation aveugle des paramètres dans un but de SAS.

Estimation des paramètres du modèle

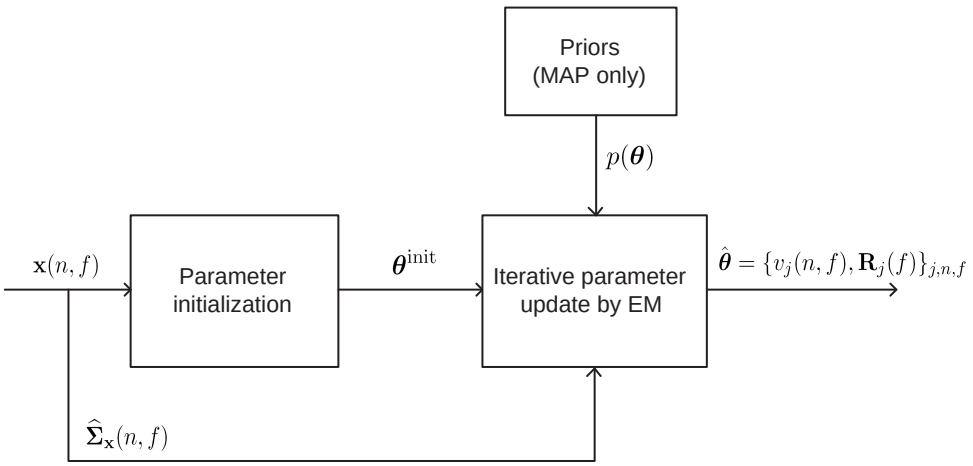
Nous présentons tout d'abord une architecture générale pour l'estimation de paramètres qui couvre aussi bien l'estimation au sens du MV qu'au sens du MAP, en s'appuyant sur l'algorithme espérance-maximisation (EM). Nous détaillons ensuite l'algorithme pour le cas MV et des exemples d'algorithmes pour le MAP avec des *a priori* spatiaux et spectraux.

Achitecture générale pour l'estimation de paramètres

La figure ci-dessous illustre l'architecture générale, flexible, que nous proposons pour l'estimation des paramètres du modèle. Cette architecture repose principalement sur la mise à jour itérative des paramètres via l'algorithme EM. Ce dernier est un choix appro-

prié bien connu pour les modèles gaussiens [113, 71]. EM étant sensible à l'initialisation [25], le choix du schéma d'initialisation des paramètres est important. Lorsque des informations additionnelles sont disponibles à l'avance, que ce soit sur les signaux sources originaux ou sur les conditions de mélange, des lois *a priori* peuvent être choisies et incorporées au modèle probabiliste pour exploiter cette connaissance et aider à améliorer la performance de séparation, en estimant les paramètres au sens du MAP. Les critères généraux d'estimation des paramètres sont:

- le critère MV
$$\hat{\theta} = \arg\max_{\theta} \prod_{n,f} p\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)|\theta\big)$$
- le critère MAP
$$\hat{\theta} = \arg\max_{\theta} \prod_{n,f} p\big(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)|\theta\big)p(\theta)$$



Architecture générale pour l'estimation aveugle des paramètres.

Estimation MV avec l'algorithme EM

L'estimation MV des paramètres se décompose en deux étapes: l'initialisation des paramètres puis leur mise à jour avec l'algorithme EM. Nous avons adapté une méthode existante de *clustering hiérarchique* [55] pour initialiser les paramètres spatiaux, et une technique d'estimation de directions d'arrivée (DoA) [51] pour aligner l'ordre des sources à toutes les fréquences à la première étape [29, 30]. L'explicitation de l'algorithme EM pour la paramétrisation de la matrice de covariance spatiale convolutive de rang 1, que

nous avons obtenue dans [30], s'inspire fortement de l'étude [25]. Pour la paramétrisation de rang plein non contrainte, l'algorithme EM est construit séparément pour chaque fréquence f pour les *données complètes* $\{\mathbf{c}_j(n,f)$ [30]. Il en résulte une procédure de mise à jour des paramètres résumée dans l'algorithme ci-dessous.

Algorithme EM pour la paramétrisation de rang plein non contrainte

---

Etape E :

$$\begin{aligned}\boldsymbol{\Sigma}_j(n,f) &= v_j(n,f)\mathbf{R}_j(f)\\ \mathbf{W}_j(n,f) &= \boldsymbol{\Sigma}_j(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)\\ \widehat{\boldsymbol{\Sigma}}_j(n,f) &= \mathbf{W}_j(n,f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I}-\mathbf{W}_j(n,f))\boldsymbol{\Sigma}_j(n,f)\end{aligned}$$

Etape M :

$$\begin{aligned}v_j(n,f) &= \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\boldsymbol{\Sigma}}_j(n,f))\\ \mathbf{R}_j(f) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\boldsymbol{\Sigma}}_j(n,f).\end{aligned}$$

Chaque itération de mise à jour EM pour la paramétrisation de rang plein non contrainte implique essentiellement le calcul de $(N+1)FJ$ inversions et de $5NFJ$ multiplications de matrices de taille $I \times I$. Le coût de calcul global d'une itération est donc $\mathcal{O}(6NFJI^3)$.

Algorithmes MAP exploitant des *a priori* spatiaux et spectraux

Pour améliorer la performance en séparation de sources dans certaines situations où de l'information *a priori* est disponible, que ce soit sur les spectres des sources ou sur le processus de mélange, nous avons introduit plusieurs lois *a priori* spatiale et spectrales, ainsi que les mises à jour MAP des paramètres par les algorithmes EM associés. L'étape E de ces algorithmes EM pour le MAP coïncide avec celle de l'Algorithme MV, et pour l'étape M les *a priori* spatiaux changent uniquement la mise à jour de $\mathbf{R}_j(f)$ tandis que les *a priori* spectraux affectent celle de $v_j(n,f)$.

A priori de position spatiale

Dans beaucoup de situations pratiques la disposition géométrique des sources et des micros, ainsi que les principales caractéristiques acoustiques de la salle, peuvent être

connues à l'avance. Pour ces raisons nous proposons de modéliser $\mathbf{R}_j(f)$ comme [33]

$$p(\mathbf{R}_j(f)) = \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big)$$

où

$$\mathcal{IW}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^m |\mathbf{R}|^{-(m+I)} e^{-\mathrm{tr}(\mathbf{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(m-i+1)}$$

est la densité de Wishart inverse [114] pour une matrice hermitienne définie positive $\mathbf{R}$, avec $\mathbf{\Psi}$ une matrice définie positive, m le nombre de degrés de liberté, Γ la fonction gamma. L'espérance selon cette distribution est $\mathbf{\Psi}/(m-I)$ [114]. On définit les matrices $\mathbf{\Psi}_j(f)$ comme

$$\mathbf{\Psi}_j(f) = (m-I)\Big(\mathbf{h}_j^{\mathrm{ane}}(f)\big(\mathbf{h}_j^{\mathrm{ane}}(f)\big)^H + \sigma^2_{\mathrm{rev}}\mathbf{\Omega}(f)\Big)$$

si bien que l'espérance de $\mathbf{R}_j(f)$ correspond aux calculs résultants de la théorie statistique de l'acoustique des salles [27, 102] et est cohérente avec la paramétrisation de rang plein directe+diffuse. Le nombre de degrés de liberté m, qui détermine de combien $\mathbf{R}_j(f)$ peut typiquement dévier de son espérance, est appris sur des données d'entraînement.

Cet *a priori conjugué* de la vraisemblance associé au modèle d'observation gaussien donne lieu à une forme explicite simple de la mise à jour des paramètres spatiaux dans l'étape M de l'algorithme EM [33]

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N}\Big(\gamma\mathbf{\Psi}_j(f) + \sum_{n=1}^{N}\frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)}\Big)$$

où γ détermine la contribution de l'*a priori*. Dans la thèse on considère aussi comme alternative un *a priori* Wishart sur $\mathbf{R}_j(f)$. Des résultats expérimentaux montrent que l'estimation MAP exploitant l'*a priori* spatial de type Wishart inverse améliore d'1 dB le rapport signal à distortion (SDR, mesure globale de la distortion en SAS) par rapport à l'approche MV, pour des mélanges stéréophoniques de trois sources avec un temps de réverbération de 250 ms [33]. Cela confirme l'efficacité de l'*a priori* proposé et de l'algorithme MAP associé.

A priori de continuité spatiale

En pratique, il arrive souvent que les sources sonores ne soient pas fixes mais se déplacent lentement. Il s'agit par exemple de légers déplacements de la tête lorsque l'on parle.

Cela se traduit par une variabilité au cours du temps des matrices de covariance spatiales paramétrisées par $\mathbf{R}_j(n,f)$, variabilité qui peut cependant être considérée comme *lente et fonction régulière du temps*. Pour cette raison nous proposons de modéliser $\mathbf{R}_j(n,f)$ pour $n>1$ par

$$p\big(\mathbf{R}_j(n,f)\big) = \mathcal{IW}\big(\mathbf{R}_j(n,f)|(m_j-I)\mathbf{R}_j(n-1,f), m_j\big).$$

Cette distribution *a priori* a pour effet d'imposer que les espérances des matrices de covariances spatiales pour une trame temporelle donnée soient égales à leurs valeurs à la trame précédente. Par ailleurs, il s'agit là encore d'un *a priori conjugué* vis-à-vis de la vraisemblance, ce qui induit une forme explicite de la mise à jour de $\mathbf{R}_j(n,f)$ à l'étape M de l'algorithme EM [34].

A priori de continuité spectrale / NMF

En complément de l'exploration d'*a priori* spatiaux, nous avons aussi cherché à exploiter les structures spectrales des sources et à construire des algorithmes d'estimation de paramètres en conséquence, le tout afin d'améliorer la performance de séparation de sources dans certains contextes. Pour cela, nous avons considéré dans un premier temps un modèle spectral NMF où $v_j(n,f)$ est modélisé comme la somme de plusieurs composantes latentes [32]

$$v_j(n,f) = \sum_{k=1}^{K} h_{j,nk} w_{j,kf}$$

où $h_{j,nk},\ w_{j,kf} \in \mathbb{R}^+$.

Nous avons aussi considéré des *a priori* de continuité spectrale, pour rendre compte du fait que le spectre à court terme de certaines sources (par exemple: la voix, le piano, le violon...) est "régulier" dans la direction du temps, tandis que celui d'autres sources (notamment les percussions) est régulier dans la direction de la fréquence [115]. La régularité dans le temps est favorisée sous la forme $\mathbb{E}\{v_j(n,f)\} = v_j(n-1,f)$, aussi nous avons proposé de modéliser $v_j(n,f)$ comme

$$p\big(v_j(n,f)\big) = \mathcal{IG}\big(v_j(n,f)|\alpha_j, (\alpha_j-1)v_j(n-1,f)\big)$$

où $\mathcal{IG}(v|\alpha,\beta)$ est la densité gamma inverse de paramètres de forme $\alpha>0$ et d'échelle $\beta>0$. De même, on peut favoriser la régularité en fréquence en modélisant $v_j(n,f)$ comme

$$p\big(v_j(n,f)\big) = \mathcal{IG}\big(v_j(n,f)|\alpha_j, (\alpha_j-1)v_j(n,f-1)\big).$$

Tous ces *a priori* spectraux mènent à des formes explicites des mises à jour de $v_j(n,f)$ dans l'étape M de l'algorithme EM, et nous avons montré qu'ils permettent d'améliorer la performance en séparation de sources dans certains contextes [32, 34].

Résultats expérimentaux

L'apport du cadre de modélisation gaussien et des algorithmes MV/MAP proposés dans cette thèse a été démontré par de nombreuses expériences, conduites dans des conditions expérimentales variées, et présentées dans plusieurs de nos publications [28, 29, 30, 32, 33, 34]. Pour prendre un exemple, nous comparons ci-dessous la paramétrisation de la matrice de covariance de rang plein non contrainte que nous avons proposée, avec ou sans modèle NMF spectral, aux algorithmes de SAS de l'état de l'art soumis aux campagnes d'évaluation SiSEC 2008 et 2010, pour les mélanges stéréophoniques enregistrés en conditions réelles. Le tableau ci-dessous indique pour chaque algorithme le SDR moyen et le score perceptuel global (OPS), qui mesurent tous les deux la distortion globale obtenue.

Algorithme	SiSEC	3 sources		4 sources	
		SDR	OPS	SDR	OPS
M. Cobos [126]	2008	2.3	-	2.1	-
M. Mandel [98]		0.1	-	-3.7	-
R. Weiss [127]		2.9	-	2.3	-
S. Araki [128]		2.9	-	-	-
Z. El Chami [129]		2.3	-	2.1	-
Full-rank		**3.3**	35	**2.8**	32
Ozerov [124]	2010	3.4	**55**	2.4	29
Full-rank+NMF		3.8	46	3.1	**33**
H. Sawada [125]		**7.2**	40	**3.4**	19

SDR moyen en dB et OPS pour les données de test acquises en conditions réelles de SiSEC 2008 et SiSEC 2010, avec des microphones espacés de 5 cm et $T_{60} = 130$ ms.

Les algorithmes de SiSEC 2008 [126, 98, 127, 128, 129] correspondent au début de ce travail de thèse, période où le critère OPS n'avait pas encore été proposé. Pour des mélanges de trois sources, l'algorithme que nous avons proposé améliore le SDR de 0.4 dB par rapport aux meilleurs résultats obtenus par l'algorithme d'Araki [128]. Pour des mélanges de quatre sources, l'amélioration du SDR est encore plus marquée, de l'ordre de 0.5 dB, par rapport aux meilleurs résultats obtenus par Weiss [127].

Comparons maintenant nos résultats avec ceux des trois nouveaux algorithmes

[124, 32, 125] soumis à SiSEC 2010. L'approche de rang plein apporte une amélioration par rapport à l'approche d'Ozerov [124] sur des mélanges de quatre sources, en termes de SDR et d'OPS. L'approche de rang plein avec NMF [32] améliore le SDR par rapport à l'approche de rang plein seule, pour trois ou quatre sources. La méthode de Sawada [125] aboutit au meilleur SDR global, mais au prix d'un OPS en retrait, par rapport aux trois autres algorithmes, tandis que celle d'Ozerov [124] aboutit au meilleur OPS. Il est intéressant de noter que ce dernier algorithme [124] a également été construit en exploitant le cadre de modélisation gaussienne que nous avons proposé, avec la paramétrisation de rang plein non contrainte que nous avons proposée, le tout combiné avec un modèle spectral NMF harmonique [124]. Cela confirme l'efficacité du cadre que nous avons proposé, et nous nous attendons à ce que l'incorporation d'un modèle spectral de source dans ce cadre puisse encore améliorer la performance.

Conclusion et perspectives

Dans cette thèse, nous avons présenté un cadre radicalement nouveau pour la séparation de sources sonores réverbérées, en s'appuyant sur la modélisation statistique d'indices spatiaux à travers la notion de matrice de covariance spatiale. Nous avons introduit quatre paramétrisations spécifiques des matrices de covariance spatiales, qui couvrent non seulement les paramétrisations de rang 1 de l'état de l'art basées sur l'approximation de bande étroite, mais également des paramétrisations de rang plein qui s'affranchissent de cette approximation dans une certaine mesure. Nous avons alors proposé une architecture générale pour l'estimation des paramètres de ces modèles s'appuyant sur la mise à jour des paramètres par EM. Pour cela nous avons proposé d'exploiter des représentations temps-fréquence quadratiques via des matrices de covariance empiriques du mélange $\widehat{\Sigma}_{\mathbf{x}}(n,f)$ afin d'améliorer la précision de l'estimation des paramètres par rapport aux représentations linéaires de type TFCT $\mathbf{x}(n,f)$ exploitées dans les approches de l'état de l'art. Nous avons exprimé un algorithme d'estimation de paramètre au sens MV et une famille d'algorithmes au sens MAP exploitant des *a priori* spatiaux et/ou spectraux pour améliorer la performance en séparation de sources dans certains contextes. Finalement, nous avons mené un ensemble conséquent d'expériences et de simulations pour comparer la performance en séparation de sources des algorithmes proposés à celle des algorithmes de l'état de l'art dans plusieurs scénarios de mélange. Tous les résultats expérimentaux confirment l'efficacité du cadre proposé.

Plusieurs directions s'appuyant sur le cadre que nous avons proposé pourront être explorées dans de futurs travaux. Tout d'abord, à court terme, il s'agira d'améliorer les

algorithmes d'estimation MV et MAP. Par exemple, un meilleur schéma d'initialisation est crucial pour un bon fonctionnement de l'algorithme MV puisque les paramètres spatiaux estimés initialements par le clustering hiérarchique peuvent varier fortement selon le choix heuristique du nombre maximum de clusters. En outre, nous pressentons que tout effort de recherche pour développer un algorithme "en-ligne" à partir du cadre que nous avons proposé a un potentiel immédiat. Par ailleurs, il serait certainement utile pour améliorer la performance en séparation de sources de s'intéresser à plusieurs extensions des *a priori* spatiaux et spectraux, comme par exemple l'estimation des hyper-paramètres des lois *a priori*. A moyen terme, il s'agira d'étudier un nouveau modèle, étendant la paramétrisation son direct + son diffus de la matrice de covariance, en distinguant trois contributions: son direct; premières réflections; réverbération tardive. Cette paramétrisation améliorée nécessitera vraisemblablement la conception d'un modèle des premières réflections qui n'existe pas à ce jour. Enfin, à plus long terme, le cadre développé de cette thèse pourra être déployé dans d'autres champs du traitement du signal audio où la modélisation de sources réverbérées ou diffuses s'avère utile.

Bibliography

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, to appear. 23, 25, 26, 32, 65, 103, 141

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, no. 6, pp. 4007–4018, 2006. 24

[3] T. Arai, H. Hodoshima, and K. Yasu, "Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1775–1780, 2010. 24

[4] D. L. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends amplif*, pp. 332–352, 2008. 24

[5] T. Lunner, M. Rudner, and J. Ronnberg, "Cognition and hearing aids," *Scandinavian journal of psychology*, vol. 50, pp. 395–403, 2009. 24

[6] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with mising feature theory," in *Proc. IEEE Int. Conf. on Robotic and Automation (ICRA)*, pp. 1517–1523, 2004. 24

[7] M. Mandel, S. Bressler, B. Shinn-Cunningham, and D. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1872–1883, 2010. 24

[8] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 169–172, 2008. 25, 115

[9] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signal based on temporal-variability of melodic source," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 425–428, 2010. 25, 115

[10] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010. 25, 29, 41

[11] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 552–559, 2007. 25, 67, 103, 104

[12] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 734–741, 2009. 25, 26, 100, 101, 102, 103, 143

[13] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010) - Audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 114–122, 2010. 25, 26, 32, 65, 101, 102, 103

[14] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1941–1944, 1998. 26, 141

[15] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007. 26, 27, 141, 142

[16] P. O'Grady, B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005. 27, 28, 29, 35, 142

[17] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22 (2001), no. 2, pp. 149–157, 2001. 28, 38

[18] O. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004. 28, 29, 37, 38, 43, 46, 81, 142, 146

[19] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006. 28, 29

[20] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, pp. 1–34, Springer, 2007. 28, 35

[21] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, IGI Global, 2010. 28, 35, 38, 47, 48, 59, 142

[22] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 613–616, 2003. 29, 40, 41

[23] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007. 29, 40, 41

[24] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009. 29, 41, 95, 96, 98

[25] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010. 29, 41, 42, 53, 78, 79, 83, 147, 148

[26] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000. 30

[27] H. Kuttruff, *Room Acoustics*. New York: Spon Press, 4th ed., 2000. 30, 54, 55, 88, 135, 145, 146, 149

[28] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 129–132, 2009. 31, 53, 66, 79, 143, 146, 151

[29] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 9–12, March 2010. 31, 143, 147, 151

[30] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010. 31, 70, 143, 146, 147, 148, 151

[31] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 73–80, Sep. 2010. 31, 71, 106, 143

[32] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1–4, 2010. 31, 107, 110, 111, 143, 150, 151, 152

[33] N. Q. K. Duong, E. Vincent, and R. Gribonval, "An acoustically-motivated spatial prior for under-determined reverberant source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 9–12, May 2011. 31, 143, 149, 151

[34] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Proc. IEEE Int. Conf. on*

Acoustics, Speech, and Signal Processing (ICASSP), pp. 205–208, May 2011. 31, 143, 150, 151

[35] M. Togami and K. Hori, "Multichannel semi-blind source separation via local Gaussian modeling for acoustic echo reduction," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2011. 32, 125

[36] M. Togami, "Online speech source separation based on maximum likelihood of local Gaussian modeling," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 213–216, May 2011. 32, 124

[37] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, to appear. 32, 42, 96, 125

[38] S. Araki and T. Nakatani, "Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel Wiever filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 225–228, 2011. 32

[39] J.-F. Cardoso, "Informax and maximum likelihood for blind source separation," *IEEE Signal Processing Letter*, no. 4, pp. 112–114, 1997. 36

[40] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995. 36, 142

[41] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999. 36, 142

[42] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994. 36

[43] J.-F. Cardoso, "Blind identification of independent components with higher-order statistics," *Proc. Workshop of higher-order spectral analysis*, pp. 157–160, 1989. 36

[44] J.-F. Cardoso, "Higher-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999. 36

[45] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998. 36, 142

[46] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 83683, 2006. 36, 142

[47] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004. 36, 142

[48] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, pp. 47–78, Springer, 2007. 36, 59, 142

[49] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing," in *Proc. 2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 77–80, 2006. 36, 44, 142

[50] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007. 36, 142

[51] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007. 36, 44, 81, 82, 142, 147

[52] F. Nesta, M. Omologo, and P. Svaizer, "Separating short signals in highly reverberant environment by a recursive frequency-domain BSS," in *Proc. of 2008 IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 232–235, 2008. 36, 142

[53] S. R. A. Jourjine and O. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2985–2988, June 2000. 37, 38, 43, 142

[54] F. Theis and E. Lang, "Formalization of the two-step approach to overcomplete BSS," in *Proc. Int. Conf. on Signal and Image Processing (SIP)*, pp. 207–212, 2002. 37

[55] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, article ID 24717, 2007. 37, 38, 43, 46, 47, 79, 80, 142, 146, 147

[56] S. Araki, H. Sawada, and S. Makino, "K-means based underdetermined blind speech separation," in *Blind Speech Separation*, pp. 243–270, Springer, 2007. 37, 43, 142

[57] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010. 37, 43, 124, 142

[58] E. Vincent, "Complex nonconvex ℓ_p norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 430–437, 2007. 37, 38

[59] R. Gribonval and M. Zibulevsky, "Sparse component analysis," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pp. 367–420, Academic Press, 2010. 37

[60] T.-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999. 38

[61] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001. 38

[62] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, no. 3-4, pp. 627–641, 2003. 38, 46

[63] M. Xiao, S. Xie, and Y. Fu, "A statistically sparse decomposition principle for underdetermined blind source separation," in *Proc. Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 165–168, 2005. 38, 47

[64] A. Nesbit, E. Vincent, and M. D. Plumbley, "Extension of sparse, adaptive signal decompositions to semi-blind audio source separation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 605–612, 2009. 38

[65] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 881–884, 2004. 38

[66] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 147–150, 2007. 38, 45, 55, 66, 107, 110

[67] P. Sudhakar, S. Arberet, and R. Gribonval, "Double sparsity: Towards blind estimation of multiple channels," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 571–578, Sep. 2010. 38

[68] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 775–782, 2009. 39, 51, 60, 143, 144

[69] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Analysis: Signal Processing Techniques and Applications*, ch. 3, pp. 39–60, Springer Verlag, 2001. 39

[70] D. Pham and J. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001. 39

[71] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. European Signal Processing Conference (EUSIPCO)*, vol. 1, pp. 561–564, 2002. 39, 62, 66, 77, 142, 145, 147

[72] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Work-*

shop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 78–81, 2005. 39, 51, 60, 62, 142, 143, 145

[73] H. Attias, "New EM algorithms for source separation and deconvolution with microphone array," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 297–300, 2003. 40

[74] H. Attias, "Gaussian mixture model for underdetermined source separation," in *Proc. IEEE Int. Conf. on Neural Networks and Brain*, pp. 1965–1969, 2005. 40

[75] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 751–758, 2009. 40, 41, 66, 79

[76] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006. 40, 41

[77] R. Blouet, G. Rapaport, and C. Févotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 37–40, 2008. 41

[78] J. Nix and V. Hohmann, "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 995–1008, 2007. 41

[79] D. D. Lee and H. S. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999. 41

[80] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180, 2003. 41

[81] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1833–1836, 2008. 42

[82] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007. 42

[83] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007. 42

[84] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006. 42, 45, 60, 61, 62

[85] A. Ozerov and C. Févotte, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 121–124, 2009. 42

[86] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval, "A tractable framework for estimating and combining spectral source models for audio source separation," *Signal Processing*, to appear. 42, 59

[87] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 745–748, 2007. 43

[88] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. on Neural Networks*, vol. 16, pp. 645–678, 2005. 43

[89] L. T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," in *Proc. IEEE Int. Sym. on Signal Processing and Its Applications (ISSPA)*, pp. 583–586, 2001. 43

[90] F. Abrard and Y. Deville, "Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach," in *Proc. IEEE Int. Sym. on Signal Processing and Its Applications (ISSPA)*, pp. 81–84, 2003. 43

[91] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 386–389, 2004. 43

[92] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 536–543, 2006. 43

[93] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," in *Proc. IEEE Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008. paper ID 9023. 44

[94] M. Mandel and D. Ellis, "A probability model for interaural phase difference," in *Proc. ISCA Workshop on Statistical and Perceptual Audio Processing (SAPA)*, pp. 1–6, 2006. 45

[95] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010. 45

[96] Z. El Chami, D. T. Pham, C. Servière, and A. Guerin, "A new model-based underdetermined speech separation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 147–150, 2008. 45

[97] Z. El Chami, D. T. Pham, C. Servière, and A. Guerin, "A new EM algorithm for underdetermined convolutive blind source separation," in *Proc. 17th European Signal Procesing Conference (EUSIPCO)*, pp. 1457–1461, 2009. 45

[98] M. Mandel and D. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 275–278, 2007. 45, 110, 151

[99] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2965–2968, 1999. 45

[100] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, pp. 61–85, Springer, 2001. 45, 55

[101] A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003. 45, 55

[102] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 791–803, 2003. 45, 53, 54, 55, 88, 145, 146, 149

[103] D. T. Pham, Z. El Chami, A. Guerin, and C. Servière, "Modeling the short time Fourier transform ratio and application to underdetermined audio source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 298–305, 2009. 45

[104] Y. Q. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, pp. 1193–1234, 2004. 47

[105] F. J. Theis, W. E. Lang, and C. G. Puntonet, "A geometric algorithm for overcomplete linear ICA," *Neurocomputing*, vol. 56, pp. 381–389, 2004. 47

[106] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation.," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3137–3140, 2009. 53

[107] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimentional TDOA estimation of multiple sources," *IEEE Trans. on Audio, Speech and Language Processing*, to appear. 54

[108] B. D. van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988. 57

[109] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the ASA*, vol. 114, pp. 2236–2252, October 2003. 59

[110] Y. Deville, "Temporal and time-frequency correlation-based blind source separation methods," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 1059–1064, 2003. 60

[111] J. Burred and T. Sikora, "Comparison of frequency-warped representations for source separation of stereo mixtures," in *Proc. 121st AES Convention*, October 2006. 60, 61

[112] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007. 65

[113] A. P. Dempster, N. M. Laird, and B. D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, pp. 1–38, 1977. 77, 147

[114] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices," *IEE Proceedings on Radar, Sonar and Navigation*, vol. 147, pp. 162–168, 2000. 88, 91, 113, 149

[115] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pp. 139–144, 2008. 97, 115, 116, 150

[116] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979. 100

[117] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guilines for subjective evaluation of audio source separation algorithms," in *Proc. UK ICA Research Network Workshop*, pp. 93–96, 2006. 103

[118] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 454–461, 2007. 103

[119] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*. to appear. 103, 105

[120] R. Lambert, "Difficulty measures and figures of merit for source separation," in *Proc. Int. Symp. on Independent Component Analysis and Blind Signal Separation(ICA)*, pp. 133–138, 1999. 103

[121] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pp. 763–768, 2003. 103

[122] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006. 103, 104

[123] R. Huber and B. Kollmeier, "PEMO-Q - A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 1902–1911, Nov. 2006. 105

[124] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 33–40, 2010. 107, 110, 111, 151, 152

[125] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011. 107, 110, 111, 123, 151, 152

[126] M. Cobos and J. López, "Blind separation of underdetermined speech mixtures based on DOA segmentation," *IEEE Trans. on Audio, Speech, and Language Processing*. submitted. 110, 151

[127] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech and Language*, vol. 24, pp. 16–20, Jan 2010. 110, 151

[128] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pp. 742–750, 2009. 110, 151

[129] Z. El Chami, D. T. Pham, C. Serviere, and A. Guerin, "A new model based underdetermined source separation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 147–150, 2008. 110, 151

[130] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 3247–3250, 2007. 111, 112

[131] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 165–172, Sep. 2010. 111

[132] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–244, 2010. 112

[133] J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Speech enhancement in presence of diffuse background noise: Why using blind signal extraction?," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4770–4773, 2010. 112

[134] R. Okamoto, Y. Takahashi, H. Saruwatari, and K. Shikano, "Mmse stsa estimator with nonstationary noise estimation based on ica for high-quality speech enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4778–4781, 2010. 112

[135] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009. 112

[136] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, pp. 529–540, 2008. 115

[137] C. Blandin, E. Vincent, and A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2616–2619, May 2011. 123

[138] N. J. Higham and H. M. Kim, "Solving a quadratic matrix equation by Newton's method with exact line searches," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, pp. 303–316, 2001. 132, 133

[139] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, 2008. Version 20081110. 139