Modèles de mélange pour la séparation multicanale de sources sonores en milieu réverbérant

Simon Leglaive

▶ **To cite this version:**

Simon Leglaive. Modèles de mélange pour la séparation multicanale de sources sonores en milieu réverbérant. Traitement du signal et de l'image [eess.SP]. Télécom ParisTech, 2017. Français. NNT : . tel-01721933v2

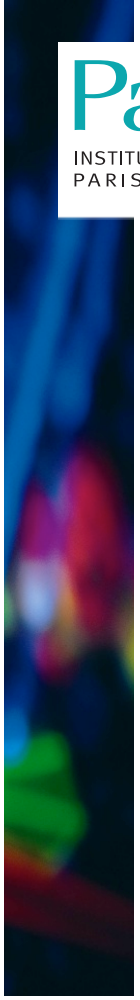
HAL Id: tel-01721933

https://theses.hal.science/tel-01721933v2

Submitted on 25 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUT OF TECHNOLOGY

TELECOM ParisTech

2017-ENST-0068

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Simon LEGLAIVE

le 12 décembre 2017

Modèles de mélange pour la séparation multicanale de sources sonores en milieu réverbérant

Directeurs de thèse :

Roland BADEAU et **Gaël RICHARD**

Jury composé de :

M. Sharon GANNOT, Professeur, Bar-Ilan University, Israël	Président du jury
M. Cédric FÉVOTTE, Directeur de Recherche, CNRS - IRIT, France	Rapporteur
M. Laurent GIRIN, Professeur, Grenoble INP, France	Rapporteur
Mme. Nancy BERTIN, Chargée de Recherche, CNRS - IRISA, France	Examinatrice
M. Matthieu KOWALSKI, Maître de Conférences, Université Paris-Sud, France	Examinateur
M. Roland BADEAU, Maître de Conférences, Télécom ParisTech, France	Directeur de thèse
M. Gaël RICHARD, Professeur, Télécom ParisTech, France	Directeur de thèse

Télécom ParisTech
École de l'Institut Mines-Télécom - Membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Je vous ai ouï malgré moi, sans que j'aie écouté [...], mais je n'ai pas compris ce que j'entendais.

Pierre Schaeffer
Traité des objets musicaux

Résumé

Cette thèse traite du problème de séparation de sources sonores pour les mélanges multicanaux enregistrés en milieu réverbérant. Nous focalisons nos travaux sur le cas sous-déterminé, c'est-à-dire lorsque le nombre de sources à séparer est supérieur au nombre de canaux du mélange. Afin d'aborder un tel problème, il est souvent utile de développer un modèle paramétrique permettant d'expliquer les données observées, c'est-à-dire le mélange. Nous adoptons dans cette thèse une approche probabiliste et hiérarchique où l'on distingue la modélisation des signaux sources monophoniques de celle du processus de mélange.

Les sources sont caractérisées dans un domaine temps-fréquence afin d'obtenir une représentation parcimonieuse, propice au développement d'un modèle car mettant en évidence la structure spécifique des signaux audio et plus particulièrement musicaux. Nous mettons en œuvre une modélisation probabiliste des sources où leurs coefficients temps-fréquence sont représentés comme des variables aléatoires latentes. Définir le modèle de source revient alors à définir la distribution jointe *a priori* de ces coefficients. Les modèles employés dans cette thèse se basent principalement sur les distributions gaussienne et t de Student. Nous utiliserons de plus des approches par factorisation en matrices non-négatives. L'intérêt de cette technique de réduction de rang réside notamment dans le caractère sous-déterminé du problème, elle permet en effet de réduire le nombre de paramètres à estimer.

Les principales contributions de cette thèse concernent la modélisation du mélange en présence de réverbération. Celui-ci est naturellement représenté dans le domaine temporel par la convolution des signaux sources avec les réponses impulsionnelles de salle qui caractérisent le chemin acoustique entre chaque source et chaque microphone. Ces réponses sont appelées *filtres de mélange* dans le contexte de la séparation de sources. Ces derniers sont généralement traités dans la littérature comme des paramètres déterministes estimés uniquement à partir des données observées. On sait cependant qu'ils correspondent à des réponses de salle, ils ont par conséquent une structure bien précise qu'il serait possible d'exploiter afin de guider leur estimation.

Dans une première partie nous considérons une approximation fréquente dans la littérature, qui consiste à approcher la convolution temporelle par une simple multiplication dans le domaine de la transformée de Fourier à court-terme, sous une hypothèse de filtres de mélange à réponse impulsionnelle courte. Le mélange est alors caractérisé par la réponse en fréquence des filtres. A partir de concepts d'acoustique géométrique des salles nous modélisons le trajet direct et les premiers échos de la réponse de salle par un processus autorégressif en fréquence. Suivant des résultats d'acoustique statistique des salles, la réverbération tardive est modélisée comme un processus gaussien en fréquence. Nous exploitons la décroissance exponentielle de la réverbération tardive dans le domaine temporel pour obtenir des expressions théoriques de la fonction d'autocovariance et de la densité spectrale de puissance de ce processus. Nous proposons également une paramétrisation autorégressive à moyenne ajustée de ces quantités. Nous développons finalement une méthode de séparation de sources basée sur un algorithme espérance-maximisation et permettant d'exploiter ces modèles par l'intermédiaire d'*a priori* sur les filtres de mélange, dans le cadre d'une estimation au sens du maximum *a posteriori*.

Dans une seconde partie nous souhaitons relâcher l'hypothèse de filtres de mélange courts car celle-ci limite fondamentalement les performances de séparation pour des mélanges fortement réverbérants. Nous proposons alors une méthode d'inférence variationnelle des coefficients temps-fréquence des sources à partir des observations temporelles du mélange. Cette approche permet de représenter de façon exacte le processus de mélange convolutif. Des résultats préliminaires obtenus en supposant la connaissance des filtres de mélange permettent de montrer la robustesse de cette approche en présence de forte réverbération. Nous développons ensuite un modèle de réponse impulsionnelle de salle basé sur la distribution t de Student. Celle-ci permet de prendre en compte le trajet direct et les premiers échos qui d'un point de vue statistique correspondent à des valeurs aberrantes par rapport au modèle de réverbération gaussien à amplitude exponentiellement décroissante. Nous développons finalement une méthode de séparation de sources basée sur une technique d'inférence variationnelle où les filtres de mélange sont considérés comme des variables aléatoires latentes dans le domaine temporel. Nous montrons également que cette approche permet d'avoir une représentation temps-fréquence adaptée à chaque source composant le mélange, notamment en terme de résolution.

Abstract

In this thesis we address the problem of audio source separation for multichannel mixtures recorded in a reverberant environment. Our work focuses on the under-determined case, that is, when the number of sources to be separated is greater than the number of channels in the mixture. In order to tackle such a problem, it is often useful to develop a parametric model that explains the observed data. In this thesis we adopt a probabilistic and hierarchical approach in which the modeling of the monophonic source signals is distinguished from that of the mixing process.

The sources are characterized in a time-frequency domain in order to obtain a sparse representation, suitable for the development of a model because highlighting the specific structure of audio signals and particularly musical ones. We rely on a probabilistic modeling of the sources where their time-frequency coefficients are represented as latent random variables. Defining the source model then amounts to defining the prior joint distribution of these coefficients. The source models in this thesis are mainly based on the Gaussian and the Student's t distributions. We will also use non-negative matrix factorization approaches. One advantage of this rank reduction technique is that the number of parameters to be estimated is reduced.

The main contributions of this thesis concern the modeling of the mixture in the presence of reverberation. Such a mixture is naturally represented in the time domain by the convolution of the source signals with the room impulse responses which characterize the acoustic path between each source and each microphone. These responses are called *mixing filters* in the context of source separation. The latter are generally treated in the literature as deterministic parameters, that are only estimated from the observed data. It is known, however, that they correspond to room responses, so they have a very specific structure that could be used to guide their estimation.

In a first part we consider a common approximation in the literature, which consists in approaching the temporal convolution by a simple multiplication in the short-time Fourier transform domain, under the hypothesis that the impulse response of the mixing filters is short. The mixture is then characterized by the frequency response of the filters. Based on geometrical room acoustics concepts, we model the direct path and the first echoes of the room response by an autoregressive process in the frequency domain. According to statistical room acoustics results, late reverberation is modeled as a Gaussian random process also in the frequency domain. We exploit the exponential temporal decay of late reverberation to obtain theoretical expressions of the autocovariance function and power spectral density of this process. We also propose an autoregressive moving average parametrization of these two quantities. Finally, we develop a source separation method based on an expectation-maximization algorithm which exploits priors on the mixing filters in order to perform maximum *a posteriori* estimation.

In a second part, we wish to relax the short mixing filters assumption because it fundamentally limits the separation performance for highly reverberant mixtures. We propose to infer the time-frequency source coefficients from the time-domain mixture observations, using a variational method. This approach makes it possible to exactly represent the convolutive mixing process, in the time domain. Preliminary results obtained by assuming that the mixing filters are known show the robustness of this approach in the presence of high reverberation. We then develop a room im-

pulse response model based on the Student's t distribution. This distribution allows us to take into account the direct path and the first echoes which, from a statistical point of view, correspond to outliers with respect to the Gaussian reverberation model with exponentially decaying amplitude. Finally, we develop a source separation method based on a variational inference technique where the mixing filters are considered as latent random variables in the time domain. We also show that this approach allows us to adapt the time-frequency representation to each individual source in the mixture, especially in terms of resolution.

Remerciements

Je tiens en tout premier lieu à remercier mes deux directeurs de thèse, Roland Badeau et Gaël Richard. Cela a été un réel plaisir de travailler avec vous pendant ces trois années, tant sur le plan personnel que scientifique. Vous avez toujours su me guider, car vous avez toujours pris le temps pour cela. Vos conseils et idées lors de nos nombreuses réunions m'ont à chaque fois fait avancer, et il est certain que sans cela les aboutissements de cette thèse n'auraient pas été les mêmes. Grâce à vous j'ai énormément appris, et c'est sûrement ce qui est le plus satisfaisant au sortir de cette thèse.

Ce travail de thèse a été évalué par un jury composé de chercheur·euse·s dont les travaux en traitement du signal audio ont eu une grande importance dans mon travail : Le président du jury Sharon Gannot, les rapporteurs Cédric Févotte et Laurent Girin, et les examinateurs Nancy Bertin et Matthieu Kowalski. Je souhaite ici les remercier pour avoir accepté de faire partie de ce jury de thèse et pour leurs remarques. Merci particulièrement à Laurent Girin, qui s'est intéressé à ce travail de thèse depuis le début, et dont les remarques ont toujours été source de réflexions personnelles enrichissantes, qui m'ont souvent fait voir les choses sous un autre angle.

Merci à l'École Doctorale Informatique, Télécommunication et Électronique (EDITE) de Paris de m'avoir permis par son financement d'effectuer cette thèse. J'ai également durant ces trois ans été impliqué dans le projet EDiSon3D (Edition et diffusion sonore spatialisée en 3 dimensions) financé par l'Agence Nationale de la Recherche, que je souhaite également remercier. Ce projet a été une opportunité unique de collaborer avec les partenaires de Radio France et de l'Université de Bretagne Occidentale, afin de comprendre les attentes et besoins des ingénieurs du son vis-à-vis de la séparation de sources, et également d'évaluer les possibilités de remixage après séparation de sources.

Merci aux permanents de l'anciennement nommé groupe AAO, Bertrand David, Yves Grenier, Slim Essid, Alexandre Gramfort et une nouvelle fois Roland Badeau et Gaël Richard. J'ai eu la chance de vous avoir comme enseignants avant cette thèse, vos cours et votre pédagogie ont eu un rôle essentiel dans mon choix de poursuivre dans le domaine du traitement du signal audio.

Un grand merci à Umut Şimşekli et Antoine Liutkus pour avoir initié notre collaboration et pour avoir partagé vos idées. Merci également pour certains moments mémorables, notamment en conférence.

Merci à toute l'équipe de la licence électronique, énérgie électrique et automatique (EEA) de l'Université Pierre et Marie Curie (UPMC), notamment Farouk Vallette, Mohamed Chetouani et Julien Denoulet. Je gardais d'excellents souvenirs de cette licence en tant qu'étudiant, cela a été un vrai plaisir pour moi d'y revenir en tant que chargé de mission d'enseignement. Merci à vous de m'avoir offert cette opportunité.

Je souhaite également remercier le Groupement de Recherche en Traitement du Signal et des Images (GRETSI), pour l'école d'été de Peyresq à laquelle j'ai eu la chance de participer en 2016. Le contenu de cette semaine de cours autour des modèles probabilistes et de l'inférence en signal et image était aussi exceptionnel que le cadre qui nous entourait. Merci également pour l'organisation biennale du colloque GRETSI, notamment pour la session plénière dédiée aux 50 ans du colloque en 2017, qui a été une opportunité unique de découvrir l'histoire de la discipline du traitement du signal.

Merci à tous les doctorants et post-doctorants que j'ai pu croiser à Télécom ParisTech et en dehors et qui ont pu contribuer à faire de ces trois années une excellente expérience.

Un merci spécial à Victor Bisot avec qui j'ai eu le plaisir de passer ces 6 ans successivement en tant qu'étudiant à Télécom ParisTech, en Master ATIAM et en tant que doctorant. Les projets étudiants, les soirées à la Maisel, les organisations de concerts à LaScène, nos projets musicaux aussi aboutis soient-ils, la pizza de la Nouvelle-Orléans, et tout ce que j'oublie ici resteront d'excellents souvenirs.

Un merci spécial également à Paul Magron, pour ta bonne humeur, tes délires, tes équations au tableau à base de *curly theta* et autres, bref, tout ce qui a fait de toi un co-bureau plus qu'au top.

Merci également à Laurent Benaroya, ça a été un plaisir de travailler avec toi dans le cadre du projet EdiSon3D. J'aurais aimé pouvoir continuer jusqu'à la fin du projet.

Merci à l'ensemble de mes amis, parce que par votre présence vous avez contribué à créer un équilibre qui a fait que ces trois ans se sont déroulés au mieux. Une mention spéciale à mes deux futurs témoins, Paul A. avec qui tout a commencé lors de notre TPE sur la distorsion du son au lycée, et Clément G. parce que si j'en suis là c'est aussi grâce à tout ce qu'on a vécu ensemble, sur 25 ans la liste serait beaucoup trop longue pour être détaillée ici.

Sur un plan plus personnel, je tiens à profondément remercier mes parents, pour votre affection, votre fierté et votre soutien. Merci également à ma sœur pour tout ce que tu m'as apporté, y compris l'immense joie de voir grandir mes trois magnifiques neveux et nièces.

Enfin, je remercie ma future épouse Melody, bien sûr pour ton soutien durant ces trois ans, pour m'avoir écouté répéter mes présentations, pour tes avis sur mes exemples audio même s'ils semblaient parfois avoir été «enregistrés dans une salle de bain», mais surtout pour ton amour qui dure déjà depuis bientôt 10 ans.

Table des matières

Résumé **v**

Abstract **vii**

Remerciements **ix**

Table des matières **xi**

Liste des figures **xv**

Abréviations **xvii**

Notations **xix**

I Introduction et état de l'art **1**

1 Introduction **3**
1.1 Contexte général 4
1.2 Formulation du problème 6
1.3 Structure du manuscrit et contributions 15
1.4 Publications associées à cette thèse 17

2 État de l'art **19**
2.1 Séparation de sources aveugle 20
2.2 Modèles probabilistes de source non stationnaire dans le domaine temps-fréquence 25
2.3 Factorisation en matrices non-négatives 30
2.4 Modèles pour les mélanges fortement réverbérants 36
2.5 Estimation et inférence statistique 39
2.6 Évaluation de la qualité de séparation 45
2.7 Bases de données 47

II Modélisation du mélange dans le domaine fréquentiel **49**

3 Modèles de réponse en fréquence de salle **51**
3.1 Introduction 52
3.2 Réponses impulsionnelle et fréquentielle de salle 53
3.3 Modèle de contributions précoces 53
3.4 Modèle de réverbération tardive 55

3.5 Conclusion 63

4 Séparation de sources avec a priori sur la réponse en fréquence des filtres de mélange 65
4.1 Modèles et estimation des filtres au sens du maximum de vraisemblance 66
4.2 A priori sur les filtres de mélange 69
4.3 Estimation des filtres au sens du maximum a posteriori 71
4.4 Résultats expérimentaux 74
4.5 Conclusion 79

III Modélisation du mélange dans le domaine temporel 81

5 Filtres de mélange déterministes 83
5.1 Représentation temporelle du mélange 85
5.2 Représentation temps-fréquence des sources 85
5.3 Modèle de source gaussien 88
5.4 Modèle de source *t* de Student 98
5.5 Conclusion 105

6 Modèle *t* de Student pour les filtres de mélange 107
6.1 Modèle 108
6.2 Inférence variationnelle 112
6.3 Résultats expérimentaux 120
6.4 Conclusion 127

IV Conclusion et perspectives 129

V Annexes 135

A Distributions de probabilité univariées 137
A.1 Distribution gaussienne 137
A.2 Distribution *t* de Student 137
A.3 Distribution inverse-gamma 138

B Preuve de l'équation (3.19) 139

C Preuve de l'équation (3.16) 141

D Méthode du gradient conjugué 143
D.1 Gradient conjugué 143
D.2 Gradient conjugué avec préconditionnement 144

E Éléments de démonstration des équations (5.22) à (5.26) 145

F Formulation alternative de la méthode du gradient conjugué pour l'étape E au chapitre 5, section 5.3 149

G Détails de calcul pour l'algorithme VEM du chapitre 6 **151**
G.1 Log-vraisemblance des données complètes 151
G.2 Étape E 151
G.3 Énergie variationnelle libre 156

Bibliographie **159**

Liste des figures

1.1 Formes d'onde et spectrogrammes de puissance en décibels calculés à partir de la transformée de Fourier à court-terme pour différents instruments de musique. . . 8
1.2 Illustration du processus d'analyse/synthèse par transformée à fenêtre glissante. . 10
1.3 Réponse impulsionnelle de salle. 12
1.4 Illustration du problème de séparation multicanale de sources sonores en milieu réverbérant. 14

2.1 Densité de probabilité de la distribution $\mathcal{S}\alpha\mathcal{S}(1/\sqrt{\alpha})$ (en échelle logarithmique) pour différentes valeurs de α. 28
2.2 Densité de probabilité de la distribution $\mathcal{T}_\alpha(0,1)$ (en échelle logarithmique) pour différentes valeurs de α. 29
2.3 Exemple de factorisation en matrices non-négatives d'un spectrogramme de puissance calculé à partir de la TFCT. Le signal se compose de deux notes de piano. Elles sont tout d'abord jouées séparément puis simultanément tel qu'indiqué par les activations temporelles. 31

3.1 Réponse impulsionnelle de salle provenant de la base de donnée RWCP [Nakamura et al., 2000]. La réponse impulsionnelle a été mesurée dans une salle avec un temps de réverbération d'environ 660 ms. 52
3.2 En haut : RIR provenant de la base RWCP. La distance source-microphone est d'environ 2 m et le temps de réverbération de 0.75 s. En bas : Relation de module et phase illustrant les équations (3.8)-(3.9). Les points bleus représentent les données tandis que les droites rouges correspondent aux droites d'équation $y = x$. $A_e(f)$ est calculé à partir de la partie précoce $a_e(t)$ représentée sur la figure du haut en bleu. 56
3.3 Fonctions d'autocovariance empirique et théorique calculées à partir de RIRs simulées. 60
3.4 Fonctions d'autocovariance empirique et théorique calculées à partir de RIRs mesurées. 60
3.5 Illustration de l'estimation des paramètres d'un modèle ARMA sur un exemple synthétique. Les observations ont été générées à partir d'un modèle ARMA(5,2). 62
3.6 Paramétrisation ARMA(7,2) de la DSP et de l'ACVF. 63
3.7 Réverbération tardive synthétisée à partir du filtrage ARMA(7,2) dans le domaine fréquentiel d'un bruit blanc gaussien complexe propre. 64

4.1 Illustration de l'approche suivie dans ce chapitre. 66
4.2 Approximation du modèle ARMA(7,2) inverse de fonction de transfert $\Phi(L)/\Theta(L)$ par un modèle MA d'ordre $N_\psi = 2048$ de fonction de transfert $\Psi(L)$. 74

4.3 SDR moyen (en dB) pour un des mélanges de la base de données, calculé en fonction des valeurs des variances des a priori σ_ϵ^2 et σ_κ^2. 78
4.4 SDR (en dB) des 29 sources de la base de données obtenu avec l'estimation MAP des filtres de mélange (en ordonnées) en fonction de l'estimation MV (en abscisses). Le point de SDR proche de -20 dB est associé aux balais de batterie présents dans un des mélanges. Cette source n'est pas bien modélisée par une NMF. 79

5.1 Erreur de modélisation due à l'hypothèse de filtres de mélange courts par rapport à la longueur de la fenêtre d'analyse utilisée dans le calcul de la TFCT, fixée ici à 128 ms. 84
5.2 SDR moyen en fonction du paramètre de forme du modèle de source *t* de Student. La courbe bleue correspond au modèle par NMF et celle en rouge au modèle basé sur la parcimonie. 104

6.1 Amplitude (figure du haut), énergie (figure du milieu) et énergie normalisée (figure du bas) d'une RIR de la base de données MIRD [Hadad et al., 2014]. Le temps de réverbération est de 610 ms et la distance source-microphone d'environ 2 m. . . . 110
6.2 Densité de probabilité empirique (trait plein noir) et densités de probabilité des distributions gaussienne (trait bleu en pointillé) et *t* de Student (trait rouge tireté) estimées à partir de 624 RIRs normalisées. 111
6.3 Réseau bayésien correspondant au modèle proposé. Les variables aléatoires latentes sont représentées par des cercles vides, les observations par des cercles grisés, et les paramètres du modèle par des points. Chaque sous-graphe contenu dans un rectangle est répété suivant les indices indiqués. Chaque arête traversant le côté d'un rectangle est également répliquée. 113
6.4 Évolution de la variance du bruit σ_i^2 au cours des itérations de l'algorithme VEM. 122
6.5 SDR moyen en fonction des paramètres de forme α_v et α_u (figure de gauche) et en fonction de la taille de la fenêtre de MDCT (en ms) pour la batterie et les autres sources (figure de droite). Le temps de réverbération des mélanges est de 360 ms. 123
6.6 SDR (en dB) et OPS (en %) des 29 sources de la base de données. Résultats obtenus avec la méthode de référence «SCM rang 1» en fonction de ceux obtenus avec la méthode proposée dans le cas d'une fenêtre MDCT adaptée en fonction de la source. 126
6.7 Évolution temporelle du paramètre de forme maximisant la vraisemblance calculée à partir de 624 RIRs (avec temps de réverbération de 610 ms), découpées en blocs non recouvrants d'une durée de 10 ms. La zone en niveau de gris représente la log-vraisemblance normalisée sur chaque bloc. Les carrés blancs indiquent la valeur optimale du paramètre de forme sur chaque bloc. 128

Abréviations

TF	Temps-fréquence
TFTC	Transformée de Fourier à temps continu
TFD	Transformée de Fourier discrète
TFCT	Transformée de Fourier à court-terme
MDCT	Transformée en cosinus discrète modifiée (de l'anglais *modified discrete cosine transform*)
RIR	Réponse impulsionnelle de salle (de l'anglais *room impulse response*)
RFR	Réponse en fréquence de salle (de l'anglais *room frequency response*)
NMF	Factorisation en matrices non-négatives (de l'anglais *non-negative matrix factorization*)
AR	Autorégressif
ARMA	Autorégressif à moyenne ajustée
DSP	Densité spectrale de puissance
ACVF	Fonction d'autocovariance (de l'anglais *autocovariance function*)
KL	Kullback-Leibler
IS	Itakura-Saito
EM	Espérance-maximisation
VEM	Espérance-maximisation variationnel (de l'anglais *variational expectation-maximization*)
ICA	Analyse en composantes indépendantes (de l'anglais *independent component analysis*)
SCA	Analyse en composantes parcimonieuses (de l'anglais *sparse component analysis*)
i.i.d	Indépendants et identiquement distribués
MV	Maximum de vraisemblance
MAP	Maximum a posteriori
SDR	Rapport signal sur distorsion (de l'anglais *signal to distortion ratio*)
SIR	Rapport signal sur interférences (de l'anglais *signal to interference ratio*)
SAR	Rapport signal sur artéfacts (de l'anglais *signal to artifact ratio*)
ISR	Rapport signal sur distorsion spatiale (de l'anglais *source image to spatial distortion ratio*)
OPS	Score perceptif global (de l'anglais *overall perceptual score*)

Notations

a	Scalaire
$\imath$	Nombre imaginaire : $\imath = \sqrt{-1}$
$\Re(\cdot)$	Partie réelle
$\Im(\cdot)$	Partie imaginaire
a^*	Complexe conjugué de a : $a^* = \Re(a) - \imath\Im(a)$
$\mathbf{a}$	Vecteur colonne ou ensemble de coefficients
$\mathbf{A}$	Matrice
$\mathbf{A}^{-1}$	Matrice inverse de $\mathbf{A}$
$\mathbf{A}^\dagger$	Matrice pseudo-inverse de $\mathbf{A}$: $\mathbf{A}^\dagger = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$
$\mathbf{A}^\top$	Matrice transposée de $\mathbf{A}$
$\mathbf{A}^H$	Matrice transposée conjuguée (conjugué hermitien) de $\mathbf{A}$: $\mathbf{A}^H = (\mathbf{A}^\top)^*$
$[\cdot]_i$	Groupe les éléments dans une matrice dont les colonnes sont indicées par i
$[\cdot]_{i,j}$	Groupe les éléments dans une matrice dont les lignes et les colonnes sont indicées par i et j respectivement
$(\cdot)_i$	i-ème élément d'un vecteur
$(\cdot)_{i,j}$	élément d'indice (i,j) d'une matrice
$(\cdot)_{i,:}$	i-ème ligne d'une matrice
$(\cdot)_{:,i}$	i-ème colonne d'une matrice
$\mathbf{AB}$	Produit des matrices $\mathbf{A}$ et $\mathbf{B}$: $(\mathbf{AB})_{i,j} = \sum_{r=1}^{R}(\mathbf{A})_{i,r}(\mathbf{B})_{r,j}$
$\mathbf{A} \otimes \mathbf{B}$	Produit de Kronecker des matrices $\mathbf{A}$ et $\mathbf{B}$: $\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1J}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & \cdots & a_{IJ}\mathbf{B} \end{pmatrix}$ où $a_{ij} = (\mathbf{A})_{i,j}$.
$\mathbf{A} \odot \mathbf{B}$	Produit d'Hadamard (terme à terme) des matrices $\mathbf{A}$ et $\mathbf{B}$: $(\mathbf{A} \odot \mathbf{B})_{i,j} = (\mathbf{A})_{i,j}(\mathbf{B})_{i,j}$
$\frac{\mathbf{A}}{\mathbf{B}}$	Division terme à terme des matrices $\mathbf{A}$ et $\mathbf{B}$: $\left(\frac{\mathbf{A}}{\mathbf{B}}\right)_{i,j} = \frac{(\mathbf{A})_{i,j}}{(\mathbf{B})_{i,j}}$
$\mathbf{A}^{\odot\eta}$	Exponentiation terme à terme de la matrice $\mathbf{A}$: $(\mathbf{A}^{\odot\eta})_{i,j} = (\mathbf{A})_{i,j}^{\eta}$
$\mathrm{diag}(\mathbf{A})$	Matrice contenant la partie diagonale de $\mathbf{A}$
$\mathrm{diag}(\{a_i\}_i)$	Matrice diagonale formée à partir de la suite de coefficients $\{a_i\}_i$
$\det(\mathbf{A})$	Déterminant de la matrice $\mathbf{A}$
$\mathrm{trace}(\mathbf{A})$	Trace de la matrice $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	Concatène les colonnes de la matrice $\mathbf{A}$ en un vecteur colonne
$\parallel \mathbf{a} \parallel_p$	Norme p du vecteur $\mathbf{a}$: $\parallel \mathbf{a} \parallel_p = \left(\sum_{i=1}^{I} \lvert(\mathbf{a})_i\rvert^p\right)^{1/p}$
$\parallel \mathbf{A} \parallel_F$	Norme de Frobenius de la matrice $\mathbf{A}$: $\parallel \mathbf{A} \parallel_F = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} \lvert(\mathbf{A})_{i,j}\rvert^2}$

$\mathbf{I}_M$	Matrice identité de taille $M \times M$
$\mathbf{1}_M$	Vecteur colonne de taille M dont tous les coefficients sont égaux à 1

$\mathbb{R}$	Ensemble des réels
$\mathbb{R}_+$	Ensemble des réels positifs
$\mathbb{C}$	Ensemble des complexes
$\mathbb{Z}$	Ensemble des entiers relatifs
$\mathbb{N}$	Ensemble des entiers naturels

$\mathbb{E}[x], \langle x \rangle_p$	Espérance mathématique d'une variable aléatoire continue de densité de probabilité p : $\mathbb{E}[x] = \langle x \rangle_p = \int x\, p(x) dx$
$\sim$	Suit une loi de probabilité
$D_{\mathrm{KL}}(q(x)) \vert\vert p(x))$	Divergence de Kullback-Leibler entre deux distributions de probabilité de densités p et q : $D_{\mathrm{KL}}(q(x)) \vert\vert p(x)) = \langle \ln\left(q(x)/p(x)\right) \rangle_q$

$\mathbb{1}_{\mathcal{S}}(x)$	Fonction indicatrice de l'ensemble $\mathcal{S}$: $\mathbb{1}_{\mathcal{S}}(x) = \begin{cases} 1 & \text{si } x \in \mathcal{S} \\ 0 & \text{sinon} \end{cases}$
$[f(\cdot) \star g(\cdot)](t)$	Produit de convolution discret entre f et g : $[f(\cdot) \star g(\cdot)](t) = \sum_{\tau \in \mathbb{Z}} f(\tau) g(t-\tau) = \sum_{\tau \in \mathbb{Z}} g(\tau) f(t-\tau)$
$\ln$	Logarithme naturel
$\log$	Logarithme décimal : $\log(x) = \ln(x)/\ln(10)$
$d_{IS}(x; y)$	Divergence d'Itakura-Saito : $d_{IS}(x; y) = x/y - \ln(x/y) - 1$
$\lfloor x \rfloor$	Partie entière par défaut de $x \in \mathbb{R}$: $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \,\vert\, n \leq x\}$
$\overset{c}{=}$	Égalité à une constante additive près
$\propto$	Égalité à un facteur multiplicatif près

Première partie

Introduction et état de l'art

Chapitre 1

Introduction

Sommaire

1.1 Contexte général 4
1.1.1 La séparation de sources en traitement du signal sonore 4
1.1.2 Applications de la séparation de sources audio 5
1.1.3 Le son et l'espace acoustique 6
1.2 Formulation du problème 6
1.2.1 Transformation temps-fréquence 7
1.2.2 Mélange convolutif 12
1.2.3 Formulation du problème 14
1.3 Structure du manuscrit et contributions 15
1.3.1 Première partie : Introduction et état de l'art 15
1.3.2 Seconde partie : Modélisation du mélange dans le domaine fréquentiel. 15
1.3.3 Troisième partie : Modélisation du mélange dans le domaine temporel . 16
1.3.4 Quatrième partie : Conclusion et perspectives 17
1.4 Publications associées à cette thèse 17

1.1 Contexte général

1.1.1 La séparation de sources en traitement du signal sonore

De nombreuses techniques en traitement du signal sonore ont pour objectif de permettre à une machine (un système informatique tel qu'un ordinateur, un téléphone, un robot, etc.) d'effectuer des tâches que l'humain fait plus ou moins naturellement, voire de le surpasser dans l'exécution de celles-ci. C'est pourquoi il est intéressant dans un premier temps de définir les différentes intéractions que nous avons quotidiennement avec le son et qui ont pu motiver la recherche en traitement du signal sonore. Prenons comme point de départ les quatre modes de l'écoute introduits par Pierre Schaeffer dans son *Traité des objects musicaux* [Schaeffer, 1966] et résumés comme suit par Michel Chion dans le *Guide des objets sonores* [Chion, 1983, p. 25] :

> «*Ouïr*, c'est percevoir par l'oreille, c'est être frappé de sons, c'est le niveau le plus brut, le plus élémentaire de la perception ; on «oit» ainsi, passivement, beaucoup de choses qu'on ne cherche ni à écouter ni à comprendre.
>
> *Écouter*, c'est prêter l'oreille à quelqu'un, à quelque chose ; c'est, par l'intermédiaire du son, viser la source, l'événement, la cause, c'est traiter le son comme indice de cette source, de cet événement.
>
> *Entendre*, c'est, d'après l'étymologie, manifester une intention d'écoute, c'est sélectionner dans ce qu'on oit ce qui nous intéresse plus particulièrement, pour opérer une «qualification» de ce qu'on entend.
>
> *Comprendre*, c'est saisir un sens, des valeurs, en traitant le son comme un signe renvoyant à ce sens, en fonction d'un langage, d'un code.»

Ces quatre modes d'écoute sont le plus souvent mis en jeu simultanément dans la perception du son. On voit néanmoins qu'il est nécessaire d'avoir ouï pour impliquer les autres modes. On remarque d'ailleurs que c'est ce verbe qui est en premier lieu utilisé dans l'épigraphe de cette thèse, illustrant les quatre modes d'écoute.

Prenons l'exemple d'une situation que l'on rencontre souvent pour introduire le problème de séparation de sources, celle d'une discussion dans un environnement bruyant tel qu'une réception ou un *cocktail*. Un auditeur *oit* passivement l'ensemble sonore qui l'entoure. Un locuteur lui parle, il décide alors de l'*écouter*. Pour l'*entendre*, il est naturellement capable de focaliser son attention sur sa voix, dans le but final de *comprendre* le message qui lui est communiqué. Il met ainsi en œuvre naturellement un procédé de séparation de sources, car il arrive en quelque sorte à isoler la voix du locuteur du reste de l'environnement sonore. Il est cependant important de nuancer par le fait que cette séparation n'est pas totale. L'auditeur perçoit toujours les autres sons qui l'entourent, c'est pourquoi il pourra se détourner de cette conversation si par exemple quelqu'un prononce son nom.

Remplaçons dans la situation précédente l'auditeur par une machine ayant été programmée dans un but précis que nous allons préciser. Ce sont un ou plusieurs microphones qui permettent à la machine d'ouïr, de capter le signal de l'environnement sonore dans le but de le traiter. Cette machine dispose d'une méthode de détection de l'évènement sonore associé à la prise de parole du locuteur, qui lui permet ainsi de l'écouter. Les téléphones utilisent par exemple un mot clé pour déclencher cette action. La machine emploie ensuite une méthode de séparation de sources afin d'isoler le flux de parole du locuteur des autres sons ambiants, qui dans ce cas précis constituent un bruit. Enfin, des techniques de reconnaissance automatique de la parole permettent à la machine de «comprendre» le message encodé et d'effectuer l'action qui lui est demandée (recherche sur internet, envoi d'un message, appel vocal, etc.).

De façon plus générale, la séparation de sources est une technique de traitement du signal qui vise à retrouver l'ensemble des signaux sources composant un mélange enregistré avec un ou plusieurs capteurs. On parlera plutôt de débruitage quand il s'agit d'isoler un signal d'intérêt noyé dans du bruit, ou de réhaussement lorsqu'on souhaite simplement augmenter la contribution du signal d'intérêt par rapport au bruit dans le mélange. On comprend par cette définition que la séparation de sources vise à aller au delà de ce que nous sommes naturellement capables de faire ; bien que l'on puisse focaliser notre attention sur un son, nous ne pouvons l'isoler parfaitement du reste de l'environnement sonore.

1.1.2 Applications de la séparation de sources audio

La séparation de sources peut être utilisée comme pré-traitement pour des tâches de classification ou de reconnaissance automatique. Dans l'exemple précédent, la séparation de la voix n'est pas l'objectif final, elle est utilisée dans le but d'aider à la reconnaissance automatique de la parole.

Dans cette thèse nous nous intéressons plus particulièrement au traitement des signaux musicaux. Dans ce contexte, la séparation de sources peut être utile pour l'extraction automatique d'information dans la musique. Dans un travail antérieur à cette thèse, nous utilisions par exemple une technique de séparation en composantes harmoniques, percussives et vocales dans le but de détecter la voix chantée dans un morceau de musique [Leglaive et al., 2015c]. Cette approche consistait à extraire des descripteurs audio à court-terme à partir des signaux séparés, et à les fournir en entrée d'un réseau de neurones récurrent effectuant la classification suivant la présence ou l'absence de voix chantée. Des techniques de séparation de sources ont également été utilisées dans d'autres applications comme la reconnaissance automatique d'instruments [Heittola et al., 2009] et l'estimation de mélodie [Durrieu et al., 2011; Tachibana et al., 2010; Rigaud et Radenen, 2016].

Le problème de séparation est plus critique quand les sources isolées ont pour vocation à être écoutées par des humains, car l'aspect perceptif lié au son est alors mis en jeu pour juger de la qualité. Nous serons très sensibles à des interférences entre sources, présentes dans les signaux séparés, ou bien à des sons non naturels appelés artéfacts et introduits par les techniques de traitement du signal sonore employées.

Des méthodes de séparation de sources peuvent par exemple être utilisées pour réduire les interférences dans les signaux captés par des microphones de proximité pour l'enregistrement d'un morceau de musique [Carabias-Orti et al., 2013; Prätzlich et al., 2015]. Ces interférences limitent en effet les possibilités de mixage des ingénieurs du son.

Un autre objectif important de la séparation de sources musicales est de permettre le remixage des morceaux de musique. Au cours de cette thèse nous avons été impliqués dans le projet ANR EDiSon3D (Edition et diffusion sonore spatialisée en 3 dimensions). Ce projet s'inscrit dans le cadre de l'émergence du son dit 3D, ayant comme objectif l'amélioration du rendu de l'espace sonore pour la musique et l'audiovisuel. Les productions (documentaires, fictions, musique, etc.) en binaural ou au format 5.1 du récent label «nouvOson» [1] de Radio France illustrent par exemple ce nouveau courant. Le concept de son 3D est étroitement lié au développement d'un «format objet» pour décrire une scène sonore, indépendamment du système de reproduction. A l'inverse des formats multicanaux actuels (stéréophonique, 5.1, etc.) où le mixage est figé, le format objet permet à chaque source audio d'être accompagnée de «méta-données» encodant par exemple l'information de spatialisation. Ce n'est qu'au moment de la diffusion que les sources sont positionnées dans l'espace, selon la configuration du dispositif de restitution. Dans ce contexte, la séparation de sources est nécessaire pour adapter un contenu audio existant dans un format multicanal standard vers ce nouveau paradigme objet, à des fins de remixage.

1. `http://hyperradio.radiofrance.fr/son-3d/`

1.1.3 Le son et l'espace acoustique

a) Le son

Le terme de «son» désigne à la fois une cause (e.g. un son de guitare), une onde acoustique aux propriétés physiques mesurables (e.g. un son harmonique de fréquence fondamentale 110 Hz) et une expérience perceptive (e.g. un son riche et clair).

On s'intéressera dans cette thèse principalement à la deuxième définition ; une source sonore sera traitée comme un signal résultant de la mesure d'une onde acoustique par un capteur, en l'occurence un microphone. Nous serons amenés pour traiter le problème de séparation de sources à développer des modèles de signaux, c'est-à-dire à les caractériser par l'intermédiaire d'une représentation mathématique. On cherchera à exploiter au travers de ces modèles les spécificités propres aux signaux musicaux, qui pourront être mises en évidence grâce à une représentation à deux dimensions, suivant le temps et la fréquence.

Supposer la connaissance de la cause du son peut également être utile. Il paraît naturel par exemple d'élaborer un modèle différent pour une batterie et un piano devant être séparés. On peut également s'inspirer du procédé de production sonore de la source, c'est le cas par exemple du modèle source/filtre [Durrieu et al., 2010, 2011].

Prendre en compte un aspect perceptif peut également s'avérer utile. Dans une application de remixage par exemple, il est éventuellement moins gênant d'avoir des interférences entre les sources après séparation plutôt que des artéfacts.

b) L'espace acoustique

Il ne peut y avoir de son sans un milieu permettant la propagation des ondes acoustiques. On s'intéresse dans cette thèse au cas de sources sonores enregistrées dans un milieu réverbérant, c'est-à-dire un espace muni de parois sur lesquelles le son émis par une source se réfléchit.

Prenons l'exemple d'un instrument de musique émettant une onde dans une salle quelconque, le signal capté par le microphone ne caractérise pas uniquement le son de l'instrument, il correspond à l'image de la source sonore vue au travers du milieu d'enregistrement. Il est clair qu'un même morceau de musique sera perçu de façon tout à fait différente s'il a été enregistré dans une chambre, un studio d'enregistrement ou une cathédrale. En plus de la modélisation de la source sonore, il nous faudra donc également caractériser la façon dont le son se propage jusqu'au microphone. La problématique que nous souhaitons traiter dans cette thèse correspond précisément à ce point. Nous allons chercher à développer de nouveaux modèles de mélange prenant en compte le caractère réverbérant du milieu d'enregistrement.

1.2 Formulation du problème

On considère un mélange de J signaux sources $s_j(t) \in \mathbb{R}$, enregistré par I microphones. On note $\mathbf{x}(t) = [x_i(t)]_i^\top \in \mathbb{R}^I$ le mélange multicanal, où $\cdot^\top$ est l'opérateur de transposition et $[\cdot]_i$ groupe les éléments dans une matrice dont les colonnes sont indicées par i ; ces éléments étant ici scalaires, $\mathbf{x}(t)$ est un vecteur colonne. Le mélange peut être décomposé d'après le modèle additif suivant [Cardoso, 1998b] :

$$\forall t \in \mathbb{Z}, \qquad \mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{y}_j(t), \tag{1.1}$$

où $\mathbf{y}_j(t) = [y_{ij}(t)]_i^\top \in \mathbb{R}^I$ représente la j-ème source image, c'est-à-dire le vecteur contenant l'image du signal source monophonique $s_j(t)$ au niveau de chaque microphone. Pour illustrer

le concept de source image considérons une pièce avec une unique source vocale et deux microphones (cas stéréo $I = 2$). $s_1(t)$ correspond au signal monophonique de voix, tel que nous pourrions le mesurer à proximité directe de la bouche du locuteur. La source image $\mathbf{y}_1(t)$ est un signal stéréophonique qui caractérise le signal source de voix mais qui incorpore également diverses informations spatiales. En particulier $\mathbf{y}_1(t)$ permet de rendre compte de la position relative de la source par rapport aux microphones, grâce aux différences de temps d'arrivée et d'intensité entre les deux microphones. L'effet de salle, c'est-à-dire le caractère réverbérant du milieu, est également présent dans le signal $\mathbf{y}_1(t)$.

Le problème de séparation de sources tel que nous le considérerons en général dans cette thèse consiste à retrouver l'ensemble des sources images à partir de l'observation du mélange. Lorsque les signaux traités sont musicaux, il est commun d'observer un mélange sur plusieurs canaux ($I > 1$). En effet les enregistrements sont généralement effectués en utilisant plusieurs microphones. De plus, un morceau de musique est souvent produit dans un format stéréophonique (ou simplement stéréo), c'est-à-dire avec deux canaux.

Le problème de séparation de sources est habituellement défini en fonction du nombre relatif de sources et de microphones ainsi que de la nature du mélange. Dans cette thèse nous considérons le cas le plus difficile, celui d'un problème *sous-déterminé*, ce qui signifie que nous devons séparer plus de sources que le nombre de canaux dans le mélange ($J > I$). De plus nous nous concentrons sur la séparation des mélanges réverbérants, on dit dans ce cas que le processus de mélange est *convolutif*. En raison de la nature sous-déterminée du problème, nous adopterons une approche basée sur une modélisation des signaux des sources et du mélange. L'information fournie par les données observées n'étant pas suffisante pour résoudre le problème, nous devons incorporer des connaissances supplémentaires sur les sources et/ou le mélange par l'intermédiaire de modèles. Le développement d'un modèle se fera généralement en deux temps :

1. Modéliser les signaux sources monophoniques $s_j(t)$. Il s'agira le plus souvent de caractériser ces derniers dans un domaine temps-fréquence comme nous allons le voir par la suite.
2. Modéliser le processus de mélange. Dans le cas d'un mélange additif tel que représenté à l'équation (1.1), il s'agira de modéliser le processus qui permet de passer de la source monophonique $s_j(t)$ à la source image multicanale $\mathbf{y}_j(t)$ qui lui est associée.

1.2.1 Transformation temps-fréquence

Dans cette thèse nous ne travaillerons qu'avec des signaux à support fini et enregistrés par des capteurs dont la dynamique est limitée, ces signaux sont par conséquent sommables, bornés et d'énergie finie. Les signaux sources sont très souvent représentés dans un domaine temps-fréquence (TF) afin d'obtenir une représentation parcimonieuse mettant en évidence la structure particulière des signaux audio et plus spécifiquement musicaux [Vincent et al., 2014]. On voit en effet sur la figure 1.1 qu'il est plus facile de discriminer les instruments de musique à partir de leur représentation TF qu'à partir de la forme d'onde, c'est-à-dire du signal temporel. On remarque notamment plusieurs caractéristiques qui pourront être exploitées dans le cadre du développement d'un modèle de source :

- Le signal de saxophone possède une structure harmonique constituée d'une fréquence fondamentale et de multiples de celle-ci. Il est moins stable que les signaux associés aux autres instruments car il est produit par la vibration d'une colonne d'air contrôlée par l'action de l'instrumentiste, notamment sur l'anche et le bec. Cette action dépend de nombreux paramètres (souffle, pression des lèvres sur l'anche, ...) qu'il est difficile de garder fixes, d'où de plus grandes variations au niveau du contenu TF du signal.
- Le signal de basse comme son nom l'indique a son énergie concentrée en basse fréquence.

- Le signal de guitare possède de fortes attaques identifiables par des lignes verticales suivies de sons harmoniques particulièrement stables temporellement, que l'on peut identifier par des lignes horizontales. On peut par ailleurs remarquer que ce signal ne présente que quelques motifs spectraux se répétant au cours du temps. Il s'agit en effet de quelques accords de guitare. La redondance en musique est un aspect important qu'il pourra être intéressant d'exploiter.
- Le signal de batterie est essentiellement percussif car sa représentation TF comporte majoritairement des lignes verticales. On remarque quelques composantes fortement énergétiques en basse fréquence qui correspondent à une grosse caisse.

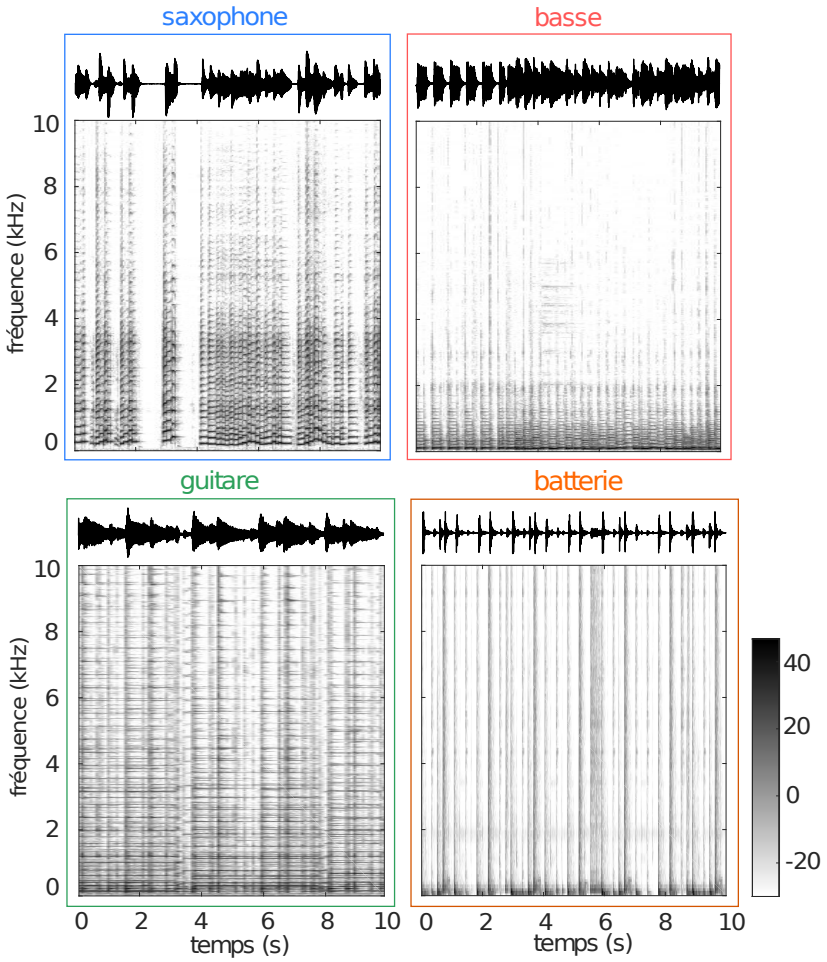


FIGURE 1.1 – Formes d'onde et spectrogrammes de puissance en décibels calculés à partir de la transformée de Fourier à court-terme pour différents instruments de musique.

De façon très générale, à partir d'un signal $s(t) \in \mathbb{R}$ défini dans le domaine temporel $t \in \mathbb{Z}$, l'opération d'*analyse* TF consiste à calculer une matrice $\mathbf{S} = [s_{fn}]_{f,n}$, $s_{fn} \in \mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, dans le domaine TF $(f, n) \in \{0, ..., F-1\} \times \mathbb{Z}$. On appellera spectrogramme d'amplitude la matrice $|\mathbf{S}|$ et spectrogramme de puissance la matrice $|\mathbf{S}|^{\odot 2}$ où $\odot$ représente une opération terme à terme sur les éléments de la matrice, ici il s'agit de prendre le carré de chacun des termes. Il existe plusieurs approches permettant de décrire le processus d'analyse/synthèse TF. Nous les présentons brièvement ci-dessous car plusieurs d'entre elles nous serviront dans la suite de ce manuscrit.

a) Transformée à fenêtre glissante

L'approche peut-être la plus intuitive pour calculer une transformée TF consiste à découper le signal en trames recouvrantes de courte durée et à calculer une transformation linéaire de chacune de celles-ci. Cette procédure est illustrée sur la figure 1.2. Soit pour tout $n \in \mathbb{Z}$ une trame de signal définie par :

$$s_n(t) = s(t + nH)w_a(t), \tag{1.2}$$

où $w_a(t)$ est une fenêtre d'analyse de support $\{0, ..., L_w - 1\}$ et H correspond à la taille de l'incrément d'analyse tel que $L_w/H \in \mathbb{N}$. Autrement dit le recouvrement entre deux trames successives est égal à $L_w - H$ échantillons. La matrice $\mathbf{S}$ est ensuite construite colonne par colonne telle que pour tout $n \in \mathbb{Z}$:

$$(\mathbf{S})_{:,n} = \mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w} \to \mathbb{K}^F}(\mathbf{s}_n), \tag{1.3}$$

où $\mathbf{s}_n = [s_n(t)]_t^\top \in \mathbb{R}^{L_w}$. $\mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w} \to \mathbb{K}^F}$ représente une transformation linéaire définie par une matrice $\mathbf{\Psi} \in \mathbb{K}^{F \times L_w}$. Pour tout vecteur $\mathbf{u} \in \mathbb{R}^{L_w}$, $\mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w} \to \mathbb{K}^F}(\mathbf{u}) = \mathbf{\Psi}\mathbf{u} \in \mathbb{K}^F$.

Soit $\mathcal{T}^{\tilde{\mathbf{\Psi}}}_{\mathbb{K}^F \to \mathbb{R}^{L_w}}$ une seconde transformation linéaire définie par une matrice $\tilde{\mathbf{\Psi}} \in \mathbb{K}^{L_w \times F}$. Pour tout vecteur $\mathbf{v} \in \mathbb{K}^F$, et en supposant qu'il existe $\mathbf{u} \in \mathbb{R}^{L_w}$ tel que $\mathbf{v} = \mathbf{\Psi}\mathbf{u}$, on a $\mathcal{T}^{\tilde{\mathbf{\Psi}}}_{\mathbb{K}^F \to \mathbb{R}^{L_w}}(\mathbf{v}) = \tilde{\mathbf{\Psi}}\mathbf{v} \in \mathbb{R}^{L_w}$. L'opération de synthèse procède également colonne par colonne. Elle consiste tout d'abord à calculer pour tout $n \in \mathbb{Z}$:

$$\hat{\mathbf{s}}_n = \mathcal{T}^{\tilde{\mathbf{\Psi}}}_{\mathbb{K}^F \to \mathbb{R}^{L_w}}\left((\mathbf{S})_{:,n}\right). \tag{1.4}$$

On construit ensuite le signal temporel $\hat{s}(t)$ pour tout $t \in \mathbb{Z}$ par addition-recouvrement :

$$\hat{s}(t) = \sum_{n \in \mathbb{Z}} w_s(t - nH)\hat{s}_n(t - nH), \tag{1.5}$$

où $w_s(t)$ est une fenêtre de synthèse de même support que $w_a(t)$ et $\hat{s}_n(t) = (\hat{\mathbf{s}}_n)_t$.

Dans cette thèse nous travaillerons uniquement avec des fenêtres d'analyse et de synthèse sinusoïdales identiques définies par :

$$w_a(t) = w_s(t) = \begin{cases} \sin\left(\dfrac{\pi}{L_w}\left(t + \dfrac{1}{2}\right)\right) & \text{si } 0 \leq t \leq L_w - 1 \\ 0 & \text{sinon} \end{cases}. \tag{1.6}$$

Il est important de mentionner que l'inversibilité de la transformation TF n'est pas forcément conditionnée au fait que $\mathcal{T}^{\tilde{\mathbf{\Psi}}}_{\mathbb{K}^F \to \mathbb{R}^{L_w}}$ soit la transformation inverse de $\mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w} \to \mathbb{K}^F}$. En effet, l'étape d'addition-recouvrement peut jouer un rôle très important dans le caractère inversible de la transformation TF. C'est le cas notamment des transformées basées sur le principe d'annulation du repliement dans le domaine temporel (TDAC d'après l'anglais *time-domain aliasing cancellation*). La transformée en cosinus discrète modifiée (MDCT d'après l'anglais *modified discrete cosine transform*) appartient par exemple à cette catégorie [Malvar, 1992]. Pour cette transformation nous avons $\mathbb{K} = \mathbb{R}$ et $H = F = L_w/2$. Comme $F < L_w$ la transformation $\mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w} \to \mathbb{R}^F}$ n'est pas inversible et l'opération (1.4) engendre un repliement temporel au niveau du signal $\hat{s}_n(t)$. Cependant grâce à l'opération d'addition-recouvrement (1.5) ce repliement s'annule et sous certaines conditions concernant les fenêtres d'analyse/synthèse $w_a(t)$ et $w_s(t)$ nous pouvons obtenir une reconstruction parfaite telle que $\hat{s}(t) = s(t)$. La fenêtre sinusoïdale permet d'obtenir une reconstruction parfaite dans le cas $H = L_w/2$.

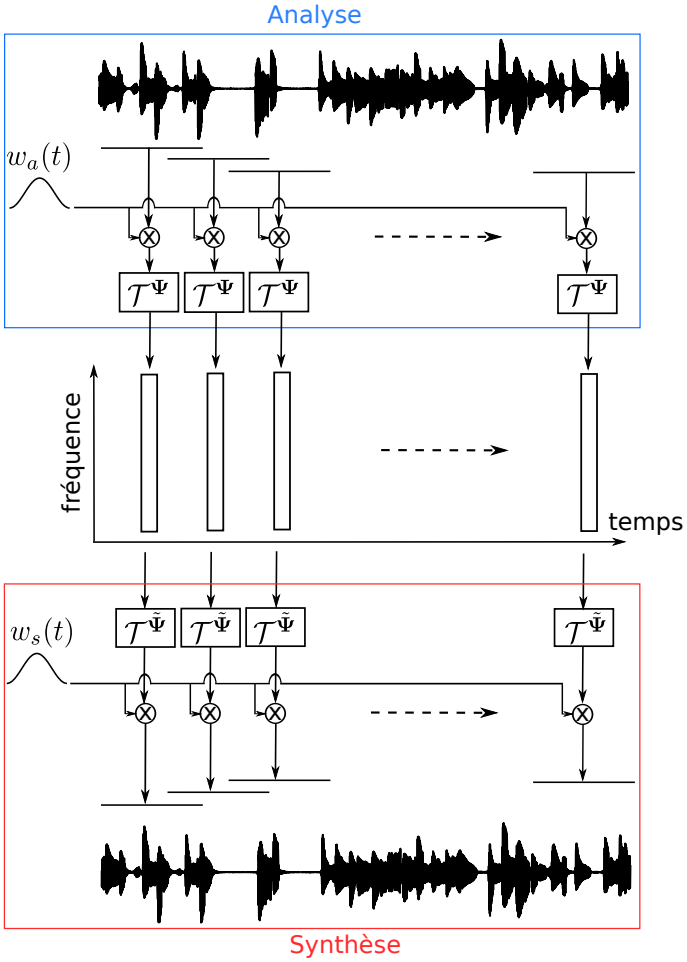


FIGURE 1.2 – Illustration du processus d'analyse/synthèse par transformée à fenêtre glissante.

b) Banc de filtres

L'approche par transformée à fenêtre glissante permet de comprendre une représentation TF comme le résultat d'une transformée fréquentielle appliquée à chaque trame n du signal. Cependant nous pouvons également interpréter une transformée TF comme fournissant pour chaque canal fréquentiel f un signal temporel sous-échantillonné d'un facteur H. Cette interprétation se formalise dans le cadre général des bancs de filtres. Les coefficients TF composant la matrice $\mathbf{S}$ sont alors calculés par filtrage et décimation :

$$\begin{aligned} s_{fn} &= [s(\cdot) \star \psi_f(\cdot)](nH) \\ &= \sum_{t \in \mathbb{Z}} s(t)\psi_f(nH - t) \end{aligned} \tag{1.7}$$

où $\psi_f(t)$ est un filtre d'analyse de support de longueur L_w et H est appelé facteur de décimation dans ce contexte. L'opération de synthèse à partir de la représentation TF du signal est définie par :

$$\hat{s}(t) = \sum_{f=0}^{F-1} \sum_{n \in \mathbb{Z}} s_{fn} \tilde{\psi}_f(t - nH), \tag{1.8}$$

où les filtres de synthèse $\tilde{\psi}_f(t)$ de support de longueur L_w doivent être choisis pour permettre la reconstruction parfaite du signal. Dans la littérature des bancs de filtres, le terme de reconstruction

parfaite signifie généralement que le signal de synthèse est égal au signal d'origine à un facteur d'échelle et un décalage près [Mertins, 1999]. Nous considérons ici une définition plus stricte : $\hat{s}(t) = s(t)$.

c) Décomposition sur un dictionnaire d'atomes

Finalement nous pouvons voir l'analyse TF d'un signal comme une décomposition sur un dictionnaire d'atomes TF d'analyse $\{\psi_{fn}(t)\}_{(f,n)\in\{0,\dots,F-1\}\times\mathbb{Z}}$:

$$s_{fn} = \sum_{t\in\mathbb{Z}} s(t)\psi_{fn}(t). \tag{1.9}$$

L'opération de synthèse consiste à reconstruire le signal temporel par combinaison linéaire d'atomes TF de synthèse $\{\tilde{\psi}_{fn}(t)\}_{(f,n)\in\{0,\dots,F-1\}\times\mathbb{Z}}$:

$$\hat{s}(t) = \sum_{f=0}^{F-1}\sum_{n\in\mathbb{Z}} s_{fn}\tilde{\psi}_{fn}(t). \tag{1.10}$$

Comme précédemment nous avons reconstruction parfaite si $\hat{s}(t) = s(t)$. En identifiant (1.7) avec (1.9) et (1.8) avec (1.10) il est clair que $\psi_{fn}(t) = \psi_f(nH - t)$ et $\tilde{\psi}_{fn}(t) = \tilde{\psi}_f(t - nH)$.

d) Exemple de la transformée de Fourier à court-terme

Nous allons illustrer ces trois approches dans le cas de la transformée de Fourier à court-terme (TFCT). Les coefficients TF sont alors à valeurs complexes ($\mathbb{K} = \mathbb{C}$) et $F = L_w$.

Transformée à fenêtre glissante Les deux transformations $\mathcal{T}^{\mathbf{\Psi}}_{\mathbb{R}^{L_w}\to\mathbb{C}^F}$ et $\mathcal{T}^{\tilde{\mathbf{\Psi}}}_{\mathbb{C}^F\to\mathbb{R}^{L_w}}$ correspondent respectivement aux transformées de Fourier discrète (TFD) directe et inverse, telles que pour tout $(f,t) \in \{0,\dots,L_w-1\}^2$:

$$(\mathbf{\Psi})_{f,t} = \frac{1}{\sqrt{L_w}} e^{-\imath 2\pi \frac{ft}{L_w}}; \tag{1.11}$$

$$(\tilde{\mathbf{\Psi}})_{t,f} = (\mathbf{\Psi})^*_{f,t}. \tag{1.12}$$

Banc de filtres Les filtres d'analyse et de synthèse sont définis pour tout $(f,t) \in \{0,\dots,L_w-1\}\times\mathbb{Z}$ par :

$$\psi_f(t) = \frac{1}{\sqrt{L_w}} w_a(-t) e^{+\imath 2\pi \frac{ft}{L_w}}; \tag{1.13}$$

$$\tilde{\psi}_f(t) = \frac{1}{\sqrt{L_w}} w_s(t) e^{+\imath 2\pi \frac{ft}{L_w}}. \tag{1.14}$$

Dictionnaire d'atomes TF Par conséquent dans le cas d'une décomposition sur un dictionnaire de TFCT nous avons pour tout $(f,t) \in \{0,\dots,L_w-1\}\times\mathbb{Z}$:

$$\psi_{fn}(t) = \psi_f(nH - t) = \frac{1}{\sqrt{L_w}} w_a(t - nH) e^{-\imath 2\pi \frac{f(t-nH)}{L_w}}; \tag{1.15}$$

$$\tilde{\psi}_{fn}(t) = \tilde{\psi}_f(t - nH) = \frac{1}{\sqrt{L_w}} w_s(t - nH) e^{+\imath 2\pi \frac{f(t-nH)}{L_w}}. \tag{1.16}$$

On peut finalement montrer qu'afin d'avoir reconstruction parfaite les fenêtres d'analyse et de synthèse doivent vérifier :

$$\sum_{n\in\mathbb{Z}} w_a(t-nH)w_s(t-nH) = 1. \tag{1.17}$$

Cette condition est notamment vérifiée pour des fenêtres d'analyse et de synthèse sinusoïdales avec $H = L_w/2$.

1.2.2 Mélange convolutif

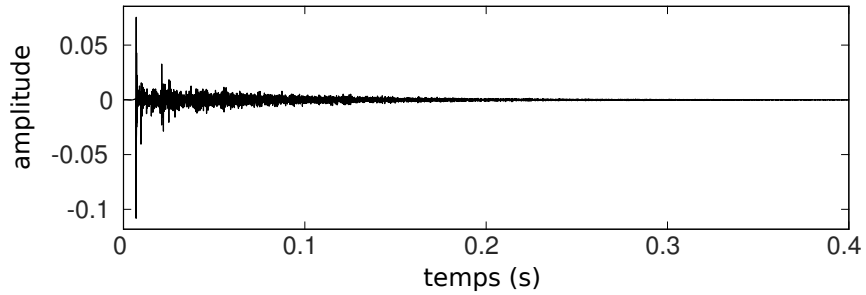


FIGURE 1.3 – Réponse impulsionnelle de salle.

Représentation temporelle du mélange Lorsqu'une source sonore ponctuelle est enregistrée dans un environnement réverbérant, le signal capté par le microphone correspond à une version filtrée du signal source. Le filtre impliqué dans cette opération est une réponse impulsionnelle de salle (RIR d'après l'anglais *room impulse response*) et représente la propagation du son dans la salle, entre la source et le microphone. Un exemple de RIR est représenté sur la figure 1.3. Dans le cadre de la séparation de sources audio on parle alors de mélange convolutif et chaque réponse de salle entre un microphone et une source est appelée *filtre de mélange*. Notons $a_{ij}(t)$ la réponse impulsionnelle du filtre de mélange représentant la propagation entre la source j et le microphone i. Ce filtre est généralement considéré comme étant à réponse impulsionnelle finie, nous notons son support $\{0, ..., L_a - 1\}$. Il nous est maintenant possible d'expliciter le lien entre le signal source $s_j(t)$ et la source image $y_{ij}(t)$ impliquée dans l'équation du mélange (1.1) :

$$\begin{aligned} y_{ij}(t) =& [s_j(\cdot) \star a_{ij}(\cdot)](t) \\ =& \sum_{\tau\in\mathbb{Z}} a_{ij}(\tau)s_j(t-\tau), \end{aligned} \tag{1.18}$$

où $\star$ représente le produit de convolution discret.

Approximation dans le domaine de la TFCT Comme indiqué précédemment, il est très fréquent de développer un modèle de source basé sur une représentation TF du signal $s_j(t)$. Un grand nombre de méthodes travaillent également à partir d'une représentation TF du mélange et donc des sources images. Généralement ces méthodes ont recours à la TFCT. Dans le cadre du formalisme par banc de filtres, les coefficients de la TFCT de la source image $y_{ij}(t)$ sont calculés par :

$$\begin{aligned} y_{ij,fn} =& [y_{ij}(\cdot) \star \psi_f(\cdot)](nH) \\ =& [s_j(\cdot) \star a_{ij}(\cdot) \star \psi_f(\cdot)](nH), \end{aligned} \tag{1.19}$$

où $\psi_f(t)$, $f = 0, ..., L_w - 1$, est défini à l'équation (1.13).

Soit $\Psi_f(\nu)$ la transformée de Fourier à temps discret (TFTD) de $\psi_f(t)$ définie pour $\nu \in \mathbb{R}$ par :

$$\Psi_f(\nu) = \sum_{t\in\mathbb{Z}} \psi_f(t) e^{-\imath 2\pi\nu t}. \tag{1.20}$$

On définit de la même façon la TFTD du filtre $a_{ij}(t)$ notée $A_{ij}(\nu)$. Par définition la TFTD est périodique de période 1, on peut donc restreindre ν à l'intervalle $[0, 1[$.

En utilisant l'équation (1.13), nous avons $\Psi_f(\nu) = (1/\sqrt{L_w})W_a^*(\nu - f/L_w)$ où $W_a(\nu)$ est la TFTD de la fenêtre d'analyse $w_a(t)$. Les fenêtres d'analyse employées pour le calcul d'une représentation TF correspondent à des réponses impulsionnelles de filtres passe-bas. En effet le spectre d'amplitude $|W_a(\nu)|$ présente un lobe principal centré sur la fréquence nulle ainsi que des lobes secondaires d'amplitude plus faible. Les caractéristiques du filtre passe-bas (notamment sa largeur de bande) dépendent de la largeur du lobe principal, de la hauteur du premier lobe secondaire et de la décroissance des lobes secondaires en fonction de la fréquence. Par conséquent $\Psi_f(\nu)$ correspond à la fonction de transfert d'un filtre passe-bande de fréquence centrale f/L_w. Si l'on fait l'hypothèse que $A_{ij}(\nu)$ varie lentement sur la bande passante de ce filtre on peut alors écrire :

$$A_{ij}(\nu)\Psi_f(\nu) \approx A_{ij}(f/L_w)\Psi_f(\nu), \tag{1.21}$$

ce qui rapporté au domaine temporel par TFTD inverse donne :

$$[a_{ij}(\cdot) \star \psi_f(\cdot)](t) \approx A_{ij}(f/L_w)\psi_f(t). \tag{1.22}$$

Sous cette hypothèse on obtient finalement l'expression suivante pour les coefficients de la TFCT de la source image $y_{ij}(t)$:

$$\begin{aligned} y_{ij,fn} =& A_{ij}(f/L_w)[s_j(\cdot) \star \psi_f(\cdot)](nH) \\ =& A_{ij}(f/L_w)s_{j,fn}, \end{aligned} \tag{1.23}$$

où $s_{j,fn} \in \mathbb{C}$ est la TFCT du signal $s_j(t)$ au point TF (f, n). Les valeurs f/L_w pour $f = 0, ..., L_w - 1$ correspondent à une discrétisation de l'intervalle $[0, 1[$; on obtient un nombre fini de L_w fréquences régulièrement espacées. Par définition $A_{ij}(f/L_w)$ correspond donc à la TFD sur L_w points de la réponse impulsionnelle $a_{ij}(t)$. Nous noterons par la suite $a_{ij,f} = A_{ij}(f/L_w)$ tel que :

$$y_{ij,fn} = a_{ij,f}s_{j,fn}. \tag{1.24}$$

Dans le cas limite où $L_w \to \infty$, le filtre de fonction de transfert $\Psi_f(\nu)$ devient parfaitement sélectif, c'est-à-dire que $\Psi_f(\nu)$ tend vers un dirac centré en f/L_w (c'est ainsi que nous comprenons le terme de «bande étroite» associé à cette approximation). L'équation (1.21) devient alors une égalité. C'est pourquoi il est généralement considéré dans la littérature qu'une convolution dans le domaine temporel peut s'écrire comme une simple multiplication dans le domaine TF si la réponse impulsionnelle du filtre de mélange est courte devant la longueur de la fenêtre d'analyse utilisée dans le calcul de la TFCT. La même conclusion peut être obtenue par un raisonnement dans le domaine temporel comme détaillé dans [Avargel et Cohen, 2007a].

Nous pouvons finalement mentionner que nous supposons ici des sources et microphones immobiles, c'est pourquoi la réponse en fréquence des filtres de mélange ne dépend pas de l'indice de trame n. Cependant ce formalisme peut être adapté au cas de filtres variant dans le temps. Une approche assez directe consiste à traiter le mélange par blocs de trames, où pour chaque bloc les filtres sont supposés constants. Une approche plus élaborée consisterait à modéliser l'évolution temporelle des filtres. C'est ce qui a été proposé dans [Kounades-Bastian et al., 2015] par l'intermédiaire d'un modèle de chaîne de Markov d'ordre 1. Ce modèle permet de prendre en compte une certaine régularité au niveau de l'évolution temporelle des filtres.

1.2.3 Formulation du problème

Nous venons de définir plusieurs concepts clés pour le problème de séparation de sources considéré dans cette thèse :

1. Les signaux sources sont généralement représentés dans un domaine **temps-fréquence** dans le but d'exploiter leur structure spécifique par l'intermédiaire d'un **modèle paramétrique**.

2. L'expression du **mélange convolutif** permet de prendre en compte le caractère réverbérant du milieu d'enregistrement par l'intermédiaire de paramètres appelés **filtres de mélange**.

3. Sous une **hypothèse de filtres de mélange à réponse impulsionnelle courte**, la convolution dans le domaine temporel peut être approchée par une simple multiplication dans le domaine de la TFCT. Les paramètres du modèle de mélange sont alors les **réponses en fréquence** des filtres de mélange.

Le problème de séparation de sources tel qu'il est fréquemment formulé dans la littérature consiste finalement à estimer les paramètres des modèles de source et de mélange, à partir desquels les coefficients TF des sources peuvent ensuite être estimés. Finalement, les signaux temporels des sources sont reconstruits par transformation TF inverse. Cette formulation du problème de séparation multicanale de sources sonores en milieu réverbérant est illustrée sur la figure 1.4.

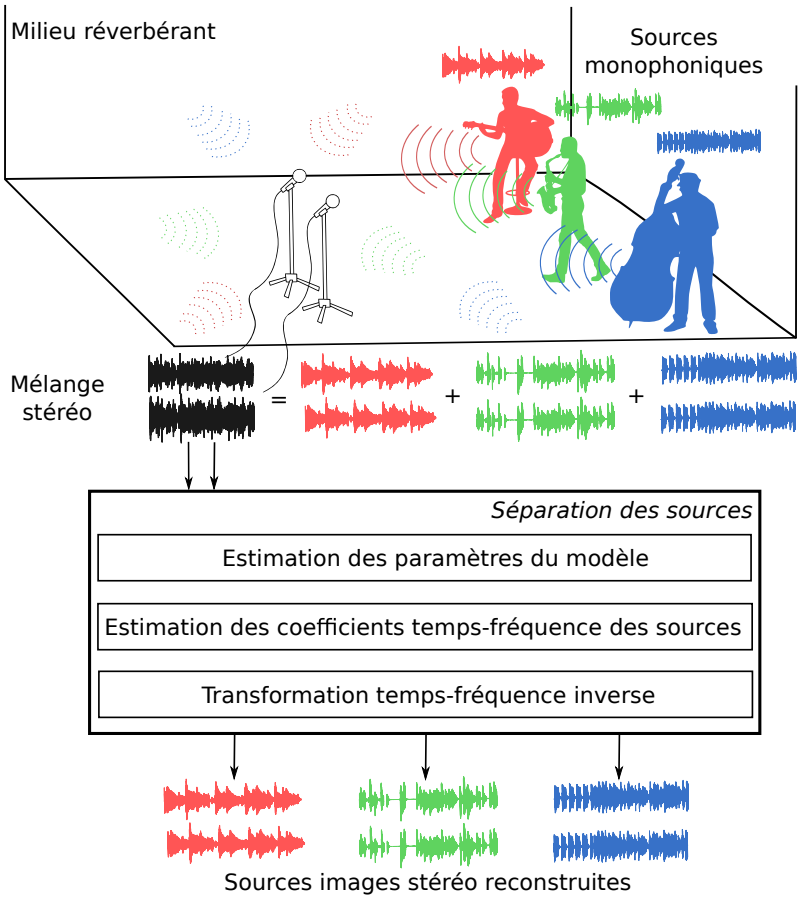


FIGURE 1.4 – Illustration du problème de séparation multicanale de sources sonores en milieu réverbérant.

1.3 Structure du manuscrit et contributions

Nous présentons ci-dessous la structure du manuscrit ainsi que les principales contributions liées à cette thèse.

1.3.1 Première partie : Introduction et état de l'art

Nous introduisons dans cette partie le problème de séparation de sources ainsi qu'un ensemble d'approches de la littérature qui lui ont été consacré.

Chapitre 2 - Etat de l'art Dans le Chapitre 2 nous dressons un état de l'art du domaine axé sur les approches probabilistes pour la séparation de sources.

Nous commençons par introduire les méthodes par analyse en composantes indépendantes et celles basées sur l'analyse en composantes parcimonieuses.

Nous présentons ensuite un certain nombre de modèles supposant une distribution de probabilité non stationnaire des coefficients TF des sources. Nous introduisons les techniques de factorisation en matrices non-négatives fréquemment utilisées dans le contexte des modèles de source probabilistes.

Nous avons montré au chapitre 1 que la convolution dans le domaine temporel pouvait être approchée par une simple multiplication dans le domaine de la TFCT sous une hypothèse de filtres de mélange courts. Cependant cette approximation n'est pas adaptée pour la séparation de mélanges fortement réverbérants car la réponse impulsionnelle des filtres est dans ce cas longue. Nous verrons que d'autres modèles ont été proposés dans la littérature pour la séparation des mélanges fortement réverbérants.

Nous présenterons ensuite les techniques d'inférence statistique qui seront utilisées dans le développement des méthodes de séparation de sources proposées dans cette thèse.

Finalement nous conclurons ce chapitre par les mesures de qualité de séparation communément employées dans le domaine et les bases de données qui seront utilisées dans cette thèse.

1.3.2 Seconde partie : Modélisation du mélange dans le domaine fréquentiel.

Les travaux développés dans cette partie reposent sur l'approximation du mélange convolutif dans le domaine de la TFCT, sous l'hypothèse de filtres de mélange courts.

Chapitre 3 - Modèles de réponse en fréquence de salle L'objectif du chapitre 3 est d'introduire de nouveaux modèles de réponse en fréquence de salle. Nous distinguons la modélisation des premières contributions (trajet direct et premiers échos) de la réverbération tardive.

Inspiré par le modèle de réponse de salle par sources images nous représentons le trajet direct et les premiers échos de la réponse impulsionnelle de salle comme des impulsions auxquelles sont associés un retard et une atténuation. Cette représentation nous amène à une modélisation autorégressive de la réponse en fréquence de salle associée aux premières contributions. Nous avons utilisé ce modèle pour la séparation des mélanges convolutifs multicanaux et réverbérants dans [Leglaive et al., 2015a,b]. Dans ces deux travaux préliminaires nous ne modélisions pas la réverbération tardive.

D'après la théorie de l'acoustique statistique des salles, la réponse en fréquence associée à la réverbération tardive peut être modélisée comme un processus aléatoire gaussien complexe centré, propre et stationnaire au sens large. Ce processus est donc totalement caractérisé par sa densité spectrale de puissance ou sa fonction d'autocovariance. En utilisant le fait que l'énergie de la réverbération tardive décroît exponentiellement au cours du temps, on montre que l'on peut obtenir

des expressions théoriques de ces deux quantités qui dépendent de certains paramètres de salle. Nous vérifions expérimentalement par simulation de Monte-Carlo la validité des expressions théoriques proposées. Nous montrons finalement que la densité spectrale de puissance et la fonction d'autocovariance peuvent être paramétrés de façon précise par un modèle autorégressif à moyenne ajustée. Ce modèle fréquentiel de réverbération tardive a été publié dans [Leglaive et al., 2016a].

Chapitre 4 - Séparation de sources avec a priori sur la réponse en fréquence des filtres de mélange Dans ce chapitre nous utilisons les modèles de réponse en fréquence de salle introduits au chapitre précédent afin de développer deux a priori probabilistes distincts pour les parties précoce et tardive des réponses en fréquence des filtres de mélange.

Dans un premier temps, nous présentons la méthode de séparation de sources proposée dans [Ozerov et Févotte, 2010] qui s'appuie sur un modèle de source gaussien basé sur la factorisation en matrices non-négatives. Dans cette article, la convolution dans le domaine temporel est approchée par une simple multiplication dans le domaine de la TFCT. La réponse en fréquence des filtres est alors estimée uniquement à partir des données observées, grâce à un algorithme espérance-maximisation.

Nous introduisons ensuite la procédure d'estimation proposée permettant de prendre en compte les a priori sur les filtres de mélange. Nous adaptons pour cela l'algorithme espérance-maximisation proposé dans [Ozerov et Févotte, 2010] afin d'estimer les filtres au sens du maximum a posteriori.

Finalement nous montrons expérimentalement l'intérêt de cette méthode, qui a fait l'objet d'une publication dans un article de revue [Leglaive et al., 2016b].

1.3.3 Troisième partie : Modélisation du mélange dans le domaine temporel

Nous explorons dans cette troisième partie de nouvelles approches pour la séparation de sources permettant de relâcher l'hypothèse de filtres de mélange courts et donc mieux adaptées pour la séparation de mélanges enregistrés en présence de forte réverbération. Nous proposons dans un cadre probabiliste d'inférer les coefficients TF des sources à partir des observations temporelles du mélange. Nous utilisons pour cela des techniques d'inférence variationnelle.

Chapitre 5 - Filtres de mélange déterministes Dans ce chapitre, les filtres de mélange dans le domaine temporel sont traités comme des paramètres déterministes uniquement estimés à partir des données observées.

Nous présentons tout d'abord une approche basée sur un modèle de source gaussien. Au chapitre précédent il était nécessaire de travailler dans le domaine de la TFCT afin d'approcher la convolution temporelle par une simple multiplication. Nous n'avons désormais plus de telle contrainte et pouvons choisir d'autres transformations TF. Nous considérons ici l'utilisation de la MDCT et de la TFCT à fréquence impaire.

Dans le cas de la MDCT, les résultats expérimentaux oracles [2] permettent de montrer le potentiel de cette approche pour la séparation de mélanges fortement réverbérants. Ces résultats ont été publiés dans [Leglaive et al., 2017a]. Nous montrons ensuite dans un contexte semi-aveugle, en supposant uniquement la connaissance des filtres de mélange, que la MDCT permet d'obtenir une qualité de séparation similaire à la TFCT à fréquence impaire. Cependant, la MDCT étant à échantillonnage critique, elle permet de réduire fortement le temps de calcul. Ces résultats ont été publiés dans [Leglaive et al., 2017c].

Finalement nous adaptons la technique d'inférence variationnelle proposée au cas d'une modélisation des sources reposant sur la distribution *t* de Student. Nous étudions dans un même cadre

2. Le terme «oracle» signifie que les paramètres de l'algorithme sont initialisés à partir de la vérité terrain, c'est-à-dire de la connaissance des vrais signaux sources et filtres de mélange.

un modèle basé sur une hypothèse de parcimonie des coefficients MDCT des sources et un second exploitant une paramétrisation par factorisation en matrices non-négatives. Cette approche nous permet de montrer l'intérêt d'utiliser à la fois une représentation temporelle du mélange convolutif et un modèle exploitant la dynamique TF des sources. Ces résultats ont été publiés dans [Leglaive et al., 2017b].

Chapitre 6 - Modèle *t* de Student pour les filtres de mélange Ce chapitre présente un cadre bayésien pour la séparation de sources audio où les filtres de mélange sont également traités comme des variables aléatoires latentes, dans le domaine temporel. Il se base sur l'article de revue [Leglaive et al., 2017d] qui a été soumis et est en phase de relecture au moment de la rédaction de cette thèse.

Comme au chapitre précédent, les coefficients TF des sources sont représentés par des variables latentes *t* de Student dont les paramètres d'échelle sont structurés par un modèle de factorisation en matrices non-négatives. Nous exploitons la décroissance exponentielle des réponses impulsionnelles de salle pour guider l'estimation des filtres de mélange grâce à un a priori également basé sur la distribution *t* de Student. La lourdeur de la queue de cette loi de probabilité permet de prendre indirectement en compte le trajet direct et les premiers échos des filtres. Nous montrons également que cette approche permet d'avoir une représentation temps-fréquence adaptée à chaque source composant le mélange, notamment en terme de résolution.

Nous développons à partir de ce modèle un algorithme d'inférence variationnel. Les résultats expérimentaux montrent le potentiel de cette méthode pour la séparation de mélanges enregistrés en présence de forte réverbération, sans supposer la connaissance des filtres de mélange.

1.3.4 Quatrième partie : Conclusion et perspectives

Cette dernière partie a pour objectif de conclure sur les méthodes de séparation de sources développées dans cette thèse. Nous présentons plusieurs perspectives pour de futurs travaux, essentiellement reliées à l'approche détaillée au chapitre 6.

1.4 Publications associées à cette thèse

a) Articles dans des revues avec comité de lecture

1. S. Leglaive, R. Badeau, G. Richard, «Multichannel audio source separation with probabilistic reverberation priors», *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, 2016.
2. S. Leglaive, R. Badeau, G. Richard, «Student's *t* source and mixing models for multichannel audio source separation», *IEEE Transactions on Audio, Speech and Language Processing*, 2017 (**article soumis**).

b) Articles dans des conférences avec comité de lecture

1. S. Leglaive, R. Badeau, G. Richard, «A priori probabiliste anéchoïque pour la séparation sous-déterminée de sources sonores en milieu réverbérant», *dans les actes du XXVe Colloque GRETSI*, Lyon, France, 2015.
2. S. Leglaive, R. Badeau, G. Richard, «Multichannel audio source separation with probabilistic reverberation modeling», *dans les actes de : IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, États-Unis d'Amérique, 2015.

3. S. Leglaive, R. Badeau, G. Richard, «Autoregressive moving average modeling of late reverberation in the frequency domain», *dans les actes de : European Signal Processing Conference (EUSIPCO)*, Budapest, Hongrie, 2016.

4. S. Leglaive, R. Badeau, G. Richard, «Multichannel audio source separation : variational inference of time-frequency sources from time-domain observations», *dans les actes de : IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Nouvelle-Orléans, LA, États-Unis d'Amérique, 2017.

5. S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, G. Richard, «Alpha-stable multichannel audio source separation», *dans les actes de : IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Nouvelle-Orléans, LA, États-Unis d'Amérique, 2017.

6. S. Leglaive, R. Badeau, G. Richard, «Semi-blind Student's t source separation for multichannel audio convolutive mixtures», *dans les actes de : European Signal Processing Conference (EUSIPCO)*, Ile de Kos, Grèce, 2017.

7. S. Leglaive, R. Badeau, G. Richard, «Séparation de sources audio en milieu réverbérant : Factorisation en matrices non-négatives et représentation temporelle du mélange convolutif», *dans les actes du XXVIe Colloque GRETSI*, Juan-Les-Pins, France, 2017.

8. S. Leglaive, R. Badeau, G. Richard, «Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization», *dans les actes de : IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, États-Unis d'Amérique, 2017.

Chapitre 2

État de l'art

Sommaire

2.1 Séparation de sources aveugle 20
2.1.1 Mélanges instantanés 20
2.1.2 Mélanges convolutifs 23
2.2 Modèles probabilistes de source non stationnaire dans le domaine temps-fréquence 25
2.2.1 Hypothèse d'indépendance des points temps-fréquence 26
2.2.2 Modèle Gaussien 26
2.2.3 Modèles avec distribution à queue lourde 27
2.3 Factorisation en matrices non-négatives 30
2.3.1 Modèle de factorisation en matrices non-négatives 30
2.3.2 Estimation par minimisation d'une divergence 31
2.3.3 Méthode de la fonction auxiliaire 32
2.3.4 La NMF dans les modèles de source probabilistes 33
2.4 Modèles pour les mélanges fortement réverbérants 36
2.4.1 Modèle convolutif diffus 36
2.4.2 Matrice de covariance spatiale 37
2.4.3 Convolution dans le domaine temps-fréquence 38
2.4.4 Représentation exacte dans le domaine temporel 39
2.5 Estimation et inférence statistique 39
2.5.1 Modèle probabiliste 39
2.5.2 Modèle à variables latentes 40
2.5.3 Inférence statistique 41
2.6 Évaluation de la qualité de séparation 45
2.6.1 Évaluation en matière de rapports d'énergie 46
2.6.2 Évaluation en matière de scores perceptifs 47
2.7 Bases de données 47

Ce chapitre a pour objectif de présenter un état de l'art du domaine. Ne pouvant être exhaustif compte tenu de la richesse de la littérature en séparation de sources, nous avons souhaité orienter cette revue du domaine vers les approches probabilistes et les modèles de mélanges réverbérants. Nous ne présenterons par exemple pas les méthodes basées sur le filtrage spatial multicanal, qui permettent de séparer les sources suivant leur direction d'arrivée. Le lecteur intéressé pourra se référer à la revue récente [Gannot et al., 2017]. Nous ne traiterons pas non plus des méthodes d'analyse de scène auditive computationnelle, qui cherchent à effectuer la séparation par regroupement des composantes du mélange selon des critères psychoacoustiques [Wang et Brown, 2006].

La section 2.1 introduit la définition historique du problème de séparation de sources aveugle, notamment grâce aux approches par analyse en composantes indépendantes. Nous présentons leur application pour la séparation des mélanges instantanés et convolutifs. Nous introduisons également les méthodes par analyse en composantes parcimonieuses.

Dans la section 2.2 nous présentons plus en profondeur certains modèles de source basés sur une hypothèse de distribution de probabilité non stationnaire dans le plan TF. Nous distinguons le modèle gaussien de ceux basés sur des distributions à queue lourde. Ces modèles de source seront utilisés dans le développement des méthodes proposées dans cette thèse.

Un modèle probabiliste non stationnaire dans le plan TF met en jeu un grand nombre de paramètres pour modéliser les sources, typiquement autant que le nombre de points TF. C'est pourquoi, dans un contexte de séparation sous-déterminée, il est fréquent d'utiliser des méthodes de réduction de rang telles que la factorisation en matrices non-négatives, conjointement avec les modèles probabilistes non stationnaires. L'objectif de la section 2.3 est d'introduire ces méthodes.

Dans la section 2.4 nous présentons différentes approches de la littérature qui ont cherché à aller au delà du modèle de mélange «bande étroite» introduit au chapitre 1, section 1.2.2, afin de modéliser des mélanges fortement réverbérants.

La formulation du problème telle qu'introduite au chapitre précédent faisait référence à une étape d'estimation des paramètres du modèle pour la séparation de sources. Dans le cadre d'approches probabilistes nous parlons d'estimation et d'inférence statistique. La section 2.5 a pour objectif d'introduire certaines de ces techniques qui seront employées dans cette thèse et qui sont de plus fréquemment utilisées dans la littérature de séparation de sources.

Finalement nous conclurons ce chapitre avec les sections 2.6 et 2.7 qui présentent respectivement les mesures de la qualité de séparation et les bases de données utilisées dans cette thèse.

2.1 Séparation de sources aveugle

2.1.1 Mélanges instantanés

a) Première définition du problème

Le problème de séparation de sources aveugle est né dans les années 1980. Dans un article fondateur publié dans les actes du Xème colloque du Groupement de Recherche en Traitement du Signal et des Images (GRETSI) de 1985, Hérault, Jutten et Ans définissent le problème de la façon suivante [Hérault et al., 1985] :

> «Supposons que, pour l'analyse d'un système complexe, nous disposions d'un ensemble de n voies de mesure. Chaque voie, issue d'un capteur spécifique, fournit à chaque instant t un signal image d'une certaine combinaison d'un nombre limité de variables internes du système. En supposant que ces variables soient des primitives indépendantes, ou simplement non-corrélées, est-il possible de les isoler par une méthode d'analyse statistique, c'est-à-dire sans connaissance a priori sur ces variables ni leur poids relatif dans chaque voie de mesure ?»

Le problème de séparation de sources aveugle tel que défini à l'origine dans cet article est associé à l'observation d'un mélange linéaire instantané sur plusieurs canaux («voies de mesure») [1]. L'équation du mélange est donnée par :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{2.1}$$

où $\mathbf{x}(t) = [x_i(t)]_i^\top \in \mathbb{R}^I$, $\mathbf{s}(t) = [s_j(t)]_j^\top \in \mathbb{R}^J$ et $\mathbf{A} \in \mathbb{R}^{I\times J}$ est la matrice de mélange représentant une transformation linéaire supposée régulière dans [Hérault et al., 1985] de sorte que sa transformation inverse représentée par $\mathbf{A}^{-1}$ existe. Cette matrice de mélange rend compte de la contribution de chaque source (les «variables internes») dans chaque canal du mélange, elle encode leur «poids relatif».

La caractéristique *aveugle* du problème de séparation de sources signifie qu'on ne suppose aucune connaissance a priori des signaux sources et de la matrice de mélange. La seule hypothèse formulée ici sur les sources est d'ordre statistique. Il s'agit de considérer leur indépendance ou de façon moins forte leur non corrélation. L'objectif de la séparation de sources aveugle dans ce contexte est d'estimer une matrice de séparation $\mathbf{B} \in \mathbb{R}^{J\times I}$ à appliquer aux observations $\mathbf{x}(t)$ afin de retrouver les sources :

$$\begin{aligned}\hat{\mathbf{s}}(t) &= \mathbf{B}\mathbf{x}(t)\\ &= \mathbf{B}\mathbf{A}\mathbf{s}(t).\end{aligned} \tag{2.2}$$

On considère que le problème de séparation est résolu si la matrice $\mathbf{C} = \mathbf{B}\mathbf{A}$ est non mélangeante [Cardoso, 1998a], c'est-à-dire que chaque ligne et chaque colonne possède un unique coefficient non nul. On comprend donc qu'on ne peut retrouver les sources qu'à une permutation et un facteur multiplicatif près.

Nous pouvons finalement mentionner que dans le cas sous-déterminé $I < J$, même si la matrice de mélange $\mathbf{A}$ était connue il serait impossible de retrouver les signaux sources, car celle-ci n'admet pas d'inverse. Si nous avions cette connaissance dans le cas déterminé $I = J$, nous choisirions simplement $\mathbf{B} = \mathbf{A}^{-1}$. Dans le cas sur-déterminé $I > J$ nous prendrions $\mathbf{B} = \mathbf{A}^\dagger$ où la matrice pseudo-inverse $\mathbf{A}^\dagger = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$ est telle que $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$. Certaines méthodes cherchent explicitement à identifier la matrice de mélange pour effectuer la séparation. C'est par exemple le cas des approches géométriques qui utilisent le fait que les échantillons du mélange se regroupent suivant les directions associées aux colonnes de la matrice de mélange [Puntonet et Prieto, 1995; Theis et al., 2003].

b) Analyse en composantes indépendantes

Il est intéressant de remarquer que dans cette première définition du problème de séparation de sources aveugle, les auteurs font explicitement référence à une hypothèse d'indépendance ou de non corrélation des sources. En effet, pendant plusieurs années les termes de séparation de

1. Nous noterons toujours I et non n le nombre de canaux.

sources aveugle et d'analyse en composantes indépendantes (ICA d'après l'anglais *independent component analysis*) vont être utilisés indifféremment pour désigner un seul et même problème [Comon, 1994]. C'est en 1998 que Cardoso propose de redéfinir le problème par l'intermédiaire d'un modèle additif tel que nous l'avons introduit à l'équation (1.1), page 6 [Cardoso, 1998b].

Comme nous l'avons vu avec l'équation (2.2), la séparation de sources aveugle consiste à retrouver les signaux sources à une permutation et un facteur multiplicatif près tel que $\hat{\mathbf{s}}(t) = \mathbf{C}\mathbf{s}(t)$ avec $\mathbf{C}$ une matrice non mélangeante. L'ICA utilise comme hypothèse de départ l'indépendance des signaux sources. Le problème de séparation consiste alors à restaurer l'indépendance statistique au niveau de l'estimée $\hat{\mathbf{s}}(t)$. Cette approche est motivée par le théorème d'identifiabilité suivant [Comon, 1994] : Soient $\mathbf{s}(t)$ un vecteur d'entrées indépendantes dont au plus une est gaussienne et $\mathbf{C}$ une matrice inversible, alors $\hat{\mathbf{s}}(t) = \mathbf{C}\mathbf{s}(t)$ est un vecteur à composantes mutuellement indépendantes si et seulement si la matrice $\mathbf{C}$ est non mélangeante. Le problème de séparation de sources est donc résolu (dans le cas d'au plus une source gaussienne) si les sources estimées sont statistiquement indépendantes.

C'est pourquoi certaines méthodes d'ICA cherchent à estimer une matrice de séparation $\mathbf{B}$ par minimisation d'une *fonction de contraste* $\phi[p(\hat{\mathbf{s}}(t))]$, avec $p(\cdot)$ une densité de probabilité. C'est une fonction scalaire à valeurs réelles atteignant son minimum lorsque les composantes du vecteur $\hat{\mathbf{s}}(t)$ sont indépendantes [Comon, 1994]. Afin d'alléger les notations nous notons simplement $\phi[\hat{\mathbf{s}}(t)]$ ce critère par la suite. Une fonction de contraste doit donc satisfaire $\phi[\hat{\mathbf{s}}(t) = \mathbf{C}\mathbf{s}(t)] \geq \phi[\mathbf{s}(t)]$ pour toute matrice $\mathbf{C}$ avec égalité si et seulement si $\mathbf{C}$ est non mélangeante. La séparation des sources est effectuée en minimisant cette fonction objectif par rapport à la matrice de séparation $\mathbf{B}$. Nous pouvons mentionner qu'une matrice de séparation peut être factorisée comme le produit d'une matrice de blanchiment $\mathbf{W}$ définie positive et d'une matrice de rotation $\mathbf{U}$ [Cardoso, 1998a]. La matrice $\mathbf{W}$ permet de décorréler les observations entre les différents canaux et peut être calculée à partir de la matrice de covariance empirique des données. Dans ce cas la minimisation de la fonction de constraste se fait par rapport à la matrice de rotation $\mathbf{U}$. De nombreuses fonctions de contraste ont été proposées dans la littérature, citons comme exemple l'information mutuelle entre les composantes de $\hat{\mathbf{s}}(t)$, calculée par l'intermédiaire d'une approximation utilisant les cumulants d'ordre inférieur ou égal à 4 [Comon, 1994].

Un résultat de Darmois [Darmois, 1953] permet de montrer que la séparation des sources est possible dans les deux cas suivants :

1. les sources sont indépendantes et identiquement distribuées (i.i.d) et non gaussiennes ;
2. les sources sont gaussiennes et non i.i.d.

Ce résultat a motivé de nombreuses méthodes en ICA. Deux courants se sont créés suivant que les méthodes qui leur étaient associées se basaient sur l'une ou l'autre de ces deux hypothèses sur les sources. La méthode *FastICA* [Bingham et Hyvärinen, 2000] appartient par exemple à la première catégorie. Elle cherche à maximiser la non-gaussianité des sources par l'intermédiaire d'une mesure basée sur le kurtosis. Le lien entre non-gaussianité et indépendance, qui n'est de premier abord pas évident, est particulièrement bien introduit dans [Hyvärinen et Oja, 2000, section 4.1]. A l'inverse, les méthodes supposant la gaussianité des sources exploitent leur structure temporelle par l'intermédiaire de statistiques d'ordre deux. L'approche introduite dans [Belouchrani et al., 1997] suppose par exemple des sources stationnaires au second ordre, exploitant ainsi leur dynamique spectrale. Cette méthode procède par diagonalisation des matrices de covariance $\mathbb{E}[\mathbf{x}(t+\tau)\mathbf{x}(t)^\top]$ conjointement pour plusieurs valeurs de τ afin d'identifier la matrice de séparation $\mathbf{B}$. Il s'agit de l'algorithme *second-order blind identification (SOBI)*. L'approche proposée dans [Pham et Cardoso, 2001] procède également par diagonalisation jointe mais suppose la non-stationarité des sources au second ordre. Elle exploite donc leur dynamique temporelle.

2.1.2 Mélanges convolutifs

a) Analyse en composantes indépendantes

Approche générale Un mélange convolutif peut s'écrire dans le domaine de la transformée en z de la façon suivante :

$$\mathbf{x}(z) = \mathbf{A}(z)\mathbf{s}(z), \tag{2.3}$$

où les éléments de la matrice $\mathbf{A}(z)$ sont donnés par $(\mathbf{A}(z))_{i,j} = \sum_{t=0}^{L_a-1} a_{ij}(t)z^{-t}$. De la même façon que dans le cas instantané, l'objectif de la séparation de sources pour les mélanges convolutifs est d'estimer une matrice de filtres $\mathbf{B}(z)$ à appliquer aux observations afin de retrouver les sources :

$$\begin{aligned}\hat{\mathbf{s}}(z) &= \mathbf{B}(z)\mathbf{x}(z) \\ &= \mathbf{B}(z)\mathbf{A}(z)\mathbf{s}(z).\end{aligned} \tag{2.4}$$

On considère que le problème est résolu si la matrice $\mathbf{C}(z) = \mathbf{B}(z)\mathbf{A}(z)$ est non mélangeante, cela signifie ici qu'elle peut s'écrire comme $\mathbf{C}(z) = \mathbf{D}(z)\mathbf{P}$, où $\mathbf{P}$ est une matrice de permutation et $\mathbf{D}(z)$ est une matrice de filtres diagonale. Le problème de séparation consiste donc à retrouver les sources à une permutation et un filtre près.

Comme dans le cas instantané, les approches par ICA pour la séparation des mélanges convolutifs sont motivées par le théorème d'identifiabilité suivant [Yellin et Weinstein, 1994; Thi et Jutten, 1995] : Soient $\mathbf{s}(t)$ un vecteur d'entrées indépendantes dont au plus une est gaussienne et $\mathbf{C}(z)$ une matrice de filtres inversible, alors $\hat{\mathbf{s}}(t)$, défini comme la transformée en z inverse de $\hat{\mathbf{s}}(z) = \mathbf{C}(z)\mathbf{s}(z)$, est un vecteur à composantes mutuellement indépendantes si et seulement si la matrice $\mathbf{C}(z)$ est non mélangeante. Le problème de séparation de sources est donc résolu (dans le cas d'au plus une source gaussienne) si les sources estimées sont statistiquement indépendantes.

Séparation dans le domaine de la TFCT De nombreuses approches en ICA pour les mélanges convolutifs cherchent à effectuer la séparation à partir d'une représentation du mélange dans le domaine de la TFCT. Notons $a_{ij,f}$ la TFD du filtre de mélange $a_{ij}(t)$ calculée sur L_w points où L_w est la longueur de la fenêtre d'analyse utilisée pour calculer la TFCT [2]. D'après l'approximation bande-étroite introduite à la section 1.2.2 du chapitre 1, le mélange $\mathbf{x}_{fn} = [x_{i,fn}]_i \in \mathbb{C}^I$ peut s'écrire de façon matricielle dans le domaine de la TFCT de la façon suivante :

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} \tag{2.5}$$

où $\mathbf{s}_{fn} = [s_{j,fn}]_j^\top \in \mathbb{C}^J$ et $\mathbf{A}_f = [a_{ij,f}]_{i,j} \in \mathbb{C}^{I\times J}$ est la matrice de mélange qui cette fois dépend de la fréquence.

On voit à partir de l'équation (2.5) que le fait d'exprimer le filtrage temporel des sources par une simple multiplication dans le domaine de la TFCT grâce à l'approximation bande-étroite permet de représenter le mélange convolutif par un simple mélange instantané dans chaque bande de fréquence de la TFCT. Les méthodes de séparation développées dans le cas instantané peuvent alors être étendues au cas convolutif. Il s'agit en particulier d'estimer une matrice de séparation $\mathbf{B}_f$ qui cette fois dépend de la fréquence. La méthode développée dans [Smaragdis, 1998] consiste précisément à appliquer dans chaque bande de fréquence des techniques d'ICA proposées pour la séparation de mélanges instantanés. La méthode proposée dans [Parra et Spence, 2000] se base également sur le modèle de mélange (2.5). Elle exploite la non-stationarité des sources dans le domaine de la TFCT par l'intermédiaire de statistiques d'ordre deux. Ces statistiques ainsi que la

2. Notons que le filtre de mélange est potentiellement sous paramétré dans le domaine fréquentiel si la longueur de la fenêtre d'analyse est plus petite que la longueur de la réponse impulsionnelle du filtre de mélange.

matrice de séparation qui dépend donc de la fréquence sont estimées en résolvant un problème des moindres carrés sous contraintes, par un algorithme de descente de gradient.

On se souvient néanmoins que généralement, on considère le problème comme étant résolu si les sources sont obtenues à une permutation et un facteur d'échelle près. Dans le cas d'un mélange convolutif ces indéterminations vont dépendre de la fréquence, car le problème est résolu indépendamment dans chaque canal fréquentiel. L'indétermination d'échelle n'est en général pas traitée, on se contente de retrouver les sources à un facteur multiplicatif près qui dépend de la fréquence, ce qui correspond en quelque sorte à un filtrage. En revanche le problème de permutations est beaucoup plus critique et il est nécessaire de lever cette indétermination. Pour cela, la méthode proposée dans [Parra et Spence, 2000] prend en compte pour l'estimation de la matrice de séparation une contrainte de filtres à réponse impulsionnelle courte. Celle-ci se traduit dans le domaine fréquentiel par une hypothèse de filtres à réponse en fréquence «lisse». Cette structure imposée pour l'estimation de la matrice de séparation permet de résoudre le problème de permutations. Une hypothèse similaire est également utilisée dans [Smaragdis, 1998]. D'autres approches ont recours à un post-traitement, indépendant du problème de séparation. La méthode proposée dans [Sawada et al., 2006, 2007] exploite par exemple un modèle de propagation anéchoïque. La réponse en fréquence des filtres de mélange étant supposée proche de celle associée à une propagation en champ libre (dans un milieu sans parois). L'approche proposée dans [Weihua et Fenggang, 2008] se base quant à elle sur les corrélations entre les amplitudes des signaux à fréquences adjacentes.

b) Analyse en composantes parcimonieuses

Les méthodes basées sur l'ICA ne permettent de résoudre le problème que dans le cas (sur-) déterminé. D'autres techniques doivent être employées lorsque le nombre de sources à séparer est supérieur au nombre de microphones.

L'analyse en composantes parcimonieuses (SCA d'après l'anglais *sparse component analysis*) [Gribonval et Zibulevsky, 2010] constitue un courant important en séparation de sources sous-déterminée pour les mélanges convolutifs. Ces méthodes exploitent la parcimonie des signaux sources dans le domaine TF. Elles supposent par exemple que les sources ont des supports TF disjoints, c'est-à-dire qu'une seule source est supposée active par point TF [Rickard, 2007; Sawada et al., 2007]. L'objectif est alors d'estimer un motif d'activations pour chaque source, la séparation étant ensuite effectuée par masquage du mélange dans le plan TF. La méthode proposée dans [Aissa-El-Bey et al., 2007] repose sur une hypothèse moins forte (quand le nombre de microphones est supérieur à 2), correspondant à un nombre de sources actives par point TF strictement inférieur au nombre de microphones.

Les méthodes basées sur la SCA exploitent aussi généralement la diversité spatiale des sources, c'est-à-dire le fait que celles-ci se situent à des positions différentes dans l'espace. Pour cela il peut s'agir d'utiliser des modèles de propagation anéchoïque [Rickard, 2007; Sawada et al., 2007] ou des indices spatiaux tels que les différences interaurales d'intensité et de temps [Mandel et al., 2010; Deleforge et al., 2015].

D'autres méthodes en SCA s'appuient sur des pénalités promouvant la parcimonie des sources dans le domaine TF. La norme ℓ_1 a par exemple été utilisée dans [Winter et al., 2007] et [Kowalski et al., 2010]. Des distributions de probabilité super-gaussiennes ont également été considérées pour développer des modèles probabilistes de sources parcimonieuses [Gribonval et Zibulevsky, 2010]. La distribution gaussienne généralisée, qui est reliée à la norme ℓ_p, a par exemple été utilisée dans [Vincent, 2007]. La distribution t de Student a été également proposée dans [Févotte et Godsill, 2006; Cemgil et al., 2007]. Tous ces modèles ont pour particularité de supposer que les coefficients TF des sources sont i.i.d.

2.2 Modèles probabilistes de source non stationnaire dans le domaine temps-fréquence

Un autre courant plus récent en séparation de sources consiste à modéliser les coefficients TF des sources par l'intermédiaire d'une distribution de probabilité non stationnaire [Vincent et al., 2010, 2014]. Les coefficients TF ne sont donc plus supposés i.i.d, ils peuvent par exemple être représentés comme des variables aléatoires gaussiennes centrées dont la variance varie en temps et en fréquence.

Un morceau de musique est généralement composé d'évènements sonores qui se répètent au cours du temps, tels que des sons pourvus d'une hauteur (des notes) ou des sons percussifs. Ces redondances peuvent être mises en évidence par l'intermédiaire d'une représentation TF des signaux. En séparation de sources audio il est particulièrement pertinent de les prendre en compte, précisément du fait du caractère sous-déterminé du problème. En effet cela permet de réduire le nombre de paramètres impliqués dans la modélisation des sources. Une approche courante consiste à supposer que le spectre (d'amplitude ou de puissance) à court-terme des sources suit un modèle compositionnel [Virtanen et al., 2015]. Sous des hypothèses de gaussianité et de stationnarité locale (nous reviendrons plus tard sur ce point) [Liutkus et al., 2011], il s'agit de modéliser la densité spectrale de puissance (DSP) à court-terme des signaux sources comme étant composée d'atomes fréquentiels qui se répètent au cours du temps. Pour ce faire les techniques de factorisation en matrices non-négatives (NMF d'après l'anglais *non-negative matrix factorization*) sont souvent employées. La contrainte de positivité de ces factorisations permet généralement d'obtenir une décomposition facilement interprétable [Virtanen et al., 2015], offrant alors la possibilité de guider le processus de séparation par de l'information extérieure [Vincent et al., 2014] (une partition de musique [Ganseman et al., 2010; Hennequin et al., 2011], un signal de référence [Smaragdis et Mysore, 2009], un modèle de production physique du son [Durrieu et al., 2011], des annotations fournies par l'utilisateur [Ozerov et al., 2011], etc.).

Dans la littérature récente, un grand nombre de distributions de probabilité ont été proposées pour modéliser les coefficients TF des signaux audio tout en s'appuyant sur une décomposition NMF [Févotte et al., 2009; Virtanen et al., 2008; Liutkus et al., 2015; Şimşekli et al., 2015; Yoshii et al., 2016; Magron et al., 2017]. Nous allons détailler certains de ces modèles dans cette section.

On peut a priori penser que la distribution de probabilité doit être choisie afin de respecter les statistiques réelles du signal. Cependant comme tous les coefficients TF sont représentés comme des variables aléatoires suivant des distributions différentes, il nous est impossible d'accéder à plusieurs réalisations de ces dernières pour pouvoir par exemple tracer un histogramme et vérifier la conformité du modèle. Néanmoins l'exactitude du modèle par rapport aux statistiques réelles du signal n'est pas le seul point à prendre en compte. Reprenons l'exemple de l'ICA, même si l'hypothèse d'indépendance des sources paraît a priori forte (surtout pour des sources musicales), cette méthode n'en reste pas moins très intéressante d'un point de vue performance de séparation. Comme mentionné dans [Cardoso, 1998a], «une approche bien conçue peut en fait être étonnamment robuste, même à des erreurs grossières de modélisation de la distribution des sources» [3].

Le terme de robustesse évoqué ici est particulièrement intéressant. C'est en effet le critère qui est recherché par les modèles de source qui se basent sur des distributions à queue lourde, comme les distributions alpha-stable ou t de Student par exemple. Il faut bien sûr préciser vis à vis de quoi le modèle doit être robuste. Il peut s'agir de robustesse par rapport à un bruit impulsif dans le plan TF [Magron et al., 2017]. L'estimation des paramètres d'un modèle gaussien est en effet particulièrement sensible aux valeurs aberrantes ; celles-ci vont fortement faire dévier la moyenne

3. Dans sa version originale en anglais : «*However, well designed approaches are in fact surprisingly robust even to gross errors in modeling the source distributions*».

et la variance estimées au sens du maximum de vraisemblance par exemple. Au contraire les modèles basés sur des distributions à queue lourde vont être beaucoup plus robustes à ces valeurs extrêmes. Il peut également s'agir de robustesse vis à vis de l'initialisation d'algorithmes itératifs tels que ceux employés dans le cadre de la NMF [Yoshii et al., 2016]. Finalement, utiliser une distribution à queue lourde peut permettre de prendre en compte une forte incertitude vis à vis du modèle gaussien sous-jacent.

2.2.1 Hypothèse d'indépendance des points temps-fréquence

Définir un modèle probabiliste de source dans le domaine TF consiste à définir la distribution jointe des coefficients TF $\{s_{j,fn}\}_{f,n}$ pour chaque source $j = 1, ..., J$. Dans la suite de cette section, nous omettons l'indice j car on suppose que chaque source suit indépendamment le même modèle. Une hypothèse importante très souvent utilisée dans la littérature consiste à supposer que tous les points TF sont indépendants, de sorte que :

$$p(\{s_{fn}\}_{f,n}) = \prod_{f,n} p(s_{fn}). \tag{2.6}$$

La densité de probabilité $p(\cdot)$ dans l'équation précédente est paramétrée par un ensemble de paramètres déterministes que nous ne faisons pour l'instant pas apparaître pour des raisons de concision.

Cette hypothèse d'indépendance peut se justifier en supposant que toutes les trames issues d'une analyse à court-terme d'un signal sont indépendantes et que les signaux sources sont localement stationnaires (i.e. sur le support de chaque trame d'analyse) [Liutkus, 2012]. Pour certaines catégories de processus, en particulier gaussiens et plus généralement harmonisables alpha-stables [Liutkus et Badeau, 2015], la stationnarité locale implique l'indépendance des points TF de la TFCT.

Cette hypothèse est utilisée pour des raisons d'ordre pratique. Cela permet en effet de simplifier fortement les modèles. Cependant elle n'est clairement pas consistante avec une représentation à court-terme des signaux : d'une part le recouvrement des trames d'analyse entraîne des corrélations entre trames successives, d'autre part des dépendances fréquentielles proviennent de l'effet de fuite spectrale induit par le fenêtrage temporel. Par ailleurs, cette hypothèse ne peut être vérifiée pour des signaux sinusoïdaux ou impulsifs par exemple, qui possèdent une structure lisse en temps ou en fréquence. Certains modèles de la littérature permettent cependant de prendre en compte explicitement ces dépendances [Badeau et Plumbley, 2014].

2.2.2 Modèle Gaussien

Le modèle de source gaussien consiste à représenter chaque coefficient $s_{fn} \in \mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$ comme une variable aléatoire suivant une loi gaussienne centrée de variance v_{fn} . Dans le cas $\mathbb{K} = \mathbb{C}$ il s'agit plus précisément de la distribution gaussienne complexe et à symétrie circulaire. Cela signifie que la distribution est invariante à un déphasage (une rotation dans le plan complexe) tel que s_{fn} et $s_{fn}e^{i\Psi}$ ont la même loi de probabilité pour tout $\Psi \in \mathbb{R}$. Cette distribution correspond à la gaussienne complexe propre [Adali et al., 2011] de moyenne nulle (cf. annexe A). Plus formellement nous avons :

$$s_{fn} \sim \mathcal{N}(0, v_{fn}). \tag{2.7}$$

La densité de probabilité de la loi gaussienne peut s'écrire de façon générale [4] :

$$N(x;\mu,\sigma^2) = \frac{1}{(\phi\pi\sigma^2)^{\frac{1}{\phi}}} \exp\left(-\frac{|x-\mu|^2}{\phi\sigma^2}\right), \tag{2.8}$$

où $\phi = 1$ si $x \in \mathbb{K} = \mathbb{C}$ et $\phi = 2$ si $x \in \mathbb{K} = \mathbb{R}$.

La variance v_{fn} représente la puissance «espérée» de la source dans le domaine TF. En effet, supposons que l'on observe un ensemble de réalisations $\{\tilde{s}_{fn}\}_{f,n}$ correspondant aux coefficients d'une représentation TF d'un signal audio quelconque. L'estimateur au sens du maximum de vraisemblance de la variance v_{fn} est alors donné par $|\tilde{s}_{fn}|^2$, c'est-à-dire le spectrogramme de puissance de la source.

2.2.3 Modèles avec distribution à queue lourde

Comme nous l'avons mentionné précédemment, le modèle gaussien peut présenter certaines limites dans le sens où la distribution gaussienne n'est pas suffisamment «flexible». Elle ne permet par exemple pas de prendre en compte une grande incertitude par rapport à un modèle de variance, ou encore l'estimation de ses paramètres sera fortement influencée par des valeurs aberrantes dans les observations. Les distributions à queue lourde sont en revanche considérées comme étant plus robustes aux valeurs aberrantes [West, 1984].

Une famille de distributions à queue lourde très utilisée dans la littérature correspond aux mélanges d'échelle de gaussiennes (SMoG d'après l'anglais *scale mixture of Gaussians*). La définition de cette famille fournie dans [Andrews et Mallows, 1974; West, 1987] est la suivante : Supposons que $z \sim \mathcal{N}(\mu,\sigma^2)$ et $u \in \mathbb{R}_+$ est une variable aléatoire de densité de probabilité $q(u)$. Alors $x = z\sqrt{u}$ est un SMoG de densité de mélange $q(u)$. La densité de probabilité de x peut alors s'écrire sous la forme suivante :

$$p(x) = \int_0^{+\infty} N(x;\mu,u\sigma^2)q(u)du. \tag{2.9}$$

Dans une terminologie bayésienne, $q(u)$ définit la distribution a priori sur le paramètre u, telle que conditionnellement à cette variable, x est gaussienne : $x|u \sim \mathcal{N}(0,u\sigma^2)$. L'équation (2.9) correspond à la densité de probabilité marginale de x.

On peut montrer que tout SMoG est associé à une distribution super-gaussienne [Palmer et al., 2006], ou autrement dit à queue lourde. Les SMoG généralisent un grand nombre de distributions, notamment les lois de Laplace, gaussienne généralisée, symmétrique alpha-stable ou encore *t* de Student. Plusieurs articles ont récemment proposé d'utiliser de telles distributions pour modéliser les signaux audio dans le domaine TF, comme par exemple les distributions de Cauchy [Liutkus et al., 2015], symétrique alpha-stable [Liutkus et Badeau, 2015; Şimşekli et al., 2015], ou encore *t* de Student [Yoshii et al., 2016].

a) Modèle symétrique alpha-stable

La distribution alpha-stable notée $\mathcal{S}_\alpha(\beta,\sigma,\mu)$ est caractérisée par quatre paramètres [Samorodnitsky et Taqqu, 1994] : $\alpha \in]0,2]$ est appelé exposant caractéristique, c'est un paramètre de forme qui détermine la lourdeur de la queue de la distribution, $\beta \in [-1,1]$ est le paramètre d'asymétrie, $\sigma \in \mathbb{R}_+$ est le paramètre d'échelle ou de dispersion et $\mu \in \mathbb{K}$ est le paramètre de position. Si $\alpha > 1$ ce dernier correspond à la moyenne (sinon la moyenne n'est pas définie).

4. Tout au long de ce manuscrit nous utiliserons des notations en lettres droites similaires à l'équation (2.8) pour désigner les densités des lois de probabilités que nous noterons en lettres rondes. $N(x;\mu,\sigma^2)$ représente ainsi la densité de probabilité de la loi gaussienne notée $\mathcal{N}(\mu,\sigma^2)$, cette densité de probabilité doit être comprise comme une fonction de la variable x paramétrée par μ et σ^2.

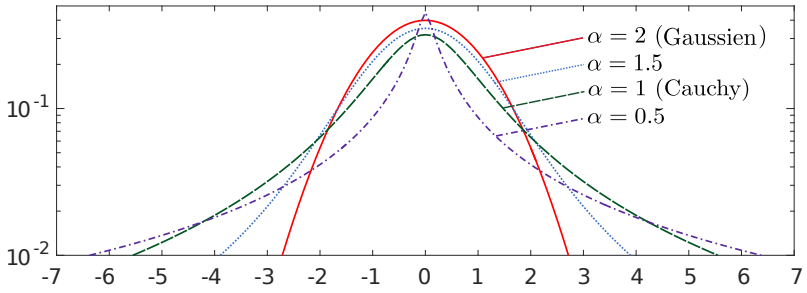


FIGURE 2.1 – Densité de probabilité de la distribution $\mathcal{S}\alpha\mathcal{S}(1/\sqrt{\alpha})$ (en échelle logarithmique) pour différentes valeurs de α.

Le cas particulier $\mathcal{S}_\alpha(0, \sigma, 0)$ correspond à la distribution symétrique alpha-stable que l'on notera $\mathcal{S}\alpha\mathcal{S}(\sigma)$. Sa densité de probabilité est représentée sur la figure 2.1 pour $\sigma = 1/\sqrt{\alpha}$. Pour une variable aléatoire à valeur complexe ($\mathbb{K} = \mathbb{C}$), cette distribution est à symétrie circulaire. La densité de probabilité de cette loi existe mais n'admet pas de forme analytique, sauf pour les cas particuliers $\alpha = 1$ et $\alpha = 2$ qui correspondent respectivement aux lois de Cauchy et gaussienne. On utilise alors la fonction caractéristique pour définir la loi symétrique alpha-stable. Soit $x \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$, sa fonction caractéristique s'écrit de façon générale (pour $x \in \mathbb{K} = \mathbb{C}$ ou $\mathbb{R}$) :

$$\forall \theta \in \mathbb{K}, \qquad \mathbb{E}\left[e^{\imath \Re(x\theta^*)}\right] = e^{-\sigma^\alpha |\theta|^\alpha}. \tag{2.10}$$

La distribution symétrique alpha-stable admet la représentation sous forme de SMoG suivante :

$$x \sim \mathcal{S}\alpha\mathcal{S}(\sigma) \Leftrightarrow \begin{cases} u & \sim \mathcal{S}_{\alpha/2}\left(1, 2\cos(\frac{\pi\alpha}{4})^{2/\alpha}, 0\right); \\ x|u & \sim \mathcal{N}(0, u\sigma^2). \end{cases} \tag{2.11}$$

La distribution de u dans l'équation précédente est appelée distribution positive alpha-stable dans [Magron et al., 2017], c'est une distribution à queue lourde de support $\mathbb{R}_+$.

Cette distribution a été proposée pour modéliser les coefficients TF des signaux audio dans [Liutkus et Badeau, 2015] puis conjointement avec un modèle NMF dans [Şimşekli et al., 2015]. Chaque coefficient $s_{fn} \in \mathbb{K}$ est modélisé de la façon suivante :

$$s_{fn} \sim \mathcal{S}\alpha\mathcal{S}\left(v_{fn}^{1/\alpha}\right). \tag{2.12}$$

Nous avons également proposé dans [Leglaive et al., 2017e] un modèle alpha-stable multivarié pour la séparation informée des mélanges multicanaux.

La distribution symétrique alpha-stable possède une propriété qui la rend très intéressante dans le contexte de la séparation de sources : Si s_1 et s_2 sont deux variables aléatoires indépendantes telles que $s_1 \sim \mathcal{S}\alpha\mathcal{S}\left(v_1^{1/\alpha}\right)$ et $s_2 \sim \mathcal{S}\alpha\mathcal{S}\left(v_2^{1/\alpha}\right)$ alors $x = s_1 + s_2 \sim \mathcal{S}\alpha\mathcal{S}\left(0, (v_1 + v_2)^{1/\alpha}\right)$. Cela se montre immédiatement à partir de l'équation (2.10) en utilisant le fait que la fonction caractéristique de la somme de deux variables aléatoires indépendantes est égale au produit de leurs fonctions caractéristiques. Dans le contexte de la séparation de sources, on comprend alors que si des sources indépendantes sont symétriques alpha-stables de même exposant caractéristique, le mélange est également symétrique alpha-stable. Dans le cas d'un mélange de sources monocanal, il est possible grâce à cette propriété d'obtenir un estimateur des sources par filtrage de Wiener généralisé [Liutkus et Badeau, 2015], de façon similaire au filtrage de Wiener classique dans le cas gaussien [Liutkus et al., 2011].

En revanche, du fait de l'absence d'expression analytique pour la densité de probabilité de cette distribution, il n'est pas possible d'avoir recours aux approches d'estimation statistique usuelles

telles que le maximum de vraisemblance. Une méthode d'inférence approchée basée sur une technique d'échantillonage est par exemple employée dans [Şimşekli et al., 2015]. Une approche basée sur la méthode des moments a été proposée dans [Fontaine et al., 2017b,a] dans le cadre d'une application de localisation de sources utilisant un modèle symétrique alpha-stable.

b) Modèle *t* de Student

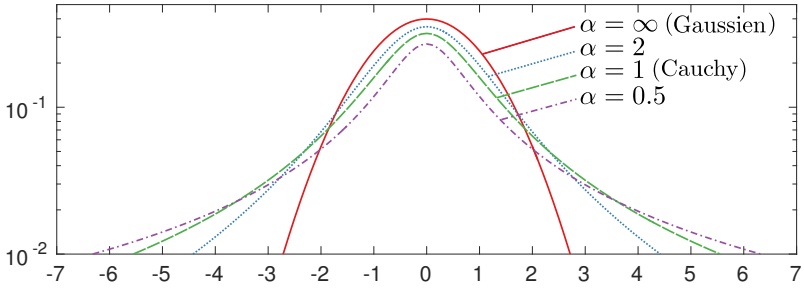


FIGURE 2.2 – Densité de probabilité de la distribution $\mathcal{T}_\alpha(0,1)$ (en échelle logarithmique) pour différentes valeurs de α.

La distribution *t* de Student (non standardisée) notée $\mathcal{T}_\alpha(\mu,\sigma)$ est caractérisée par trois paramètres : $\alpha \in \mathbb{R}_+$ est le paramètre de forme, également appelé degrés de liberté, qui tout comme pour la distribution alpha-stable caractérise la lourdeur de la queue de la distribution, $\mu \in \mathbb{K}$ est le paramètre de position qui correspond également à la moyenne si $\alpha > 1$ (sinon la moyenne n'est pas définie), et $\sigma \in \mathbb{R}_+$ est le paramètre d'échelle. La variance est égale à $\sigma^2 \dfrac{\alpha}{\alpha-2}$ si $\alpha > 2$, sinon celle-ci n'est pas définie. La densité de probabilité de cette distribution est représentée sur la figure 2.2 pour $\mu = 0$ et $\sigma = 1$. Pour une variable aléatoire à valeur complexe et lorsque $\mu = 0$, cette distribution est à symétrie circulaire. La distribution *t* de Student généralise également les lois de Cauchy et gaussienne qui correspondent respectivement aux cas $\alpha = 1$ et $\alpha \to \infty$. Sa densité de probabilité est définie par :

$$T_\alpha(x;\mu,\sigma) = \frac{2}{\phi}\frac{1}{(\alpha\pi\sigma^2)^{1/\phi}}\frac{\Gamma(\alpha/2+1/\phi)}{\Gamma(\alpha/2)}\left(1+\frac{2}{\phi\alpha}\frac{|x-\mu|^2}{\sigma^2}\right)^{-(\alpha/2+1/\phi)}, \tag{2.13}$$

où $\Gamma(\cdot)$ est la fonction gamma et on rappelle que $\phi = 1$ si $x \in \mathbb{K} = \mathbb{C}$ et $\phi = 2$ si $x \in \mathbb{K} = \mathbb{R}$.

En introduisant une variable aléatoire suivant une distribution inverse-gamma notée $\mathcal{IG}$, la loi *t* de Student admet la représentation suivante sous forme de SMoG :

$$x \sim \mathcal{T}_\alpha(\mu,\sigma) \Leftrightarrow \begin{cases} u & \sim \mathcal{IG}\left(\frac{\alpha}{2},\frac{\alpha}{2}\right); \\ x|u & \sim \mathcal{N}(\mu,u\sigma^2). \end{cases} \tag{2.14}$$

La densité de probabilité de la distribution inverse-gamma est définie par :

$$IG(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}\exp\left(\frac{-\beta}{x}\right). \tag{2.15}$$

Un modèle basé sur la distribution *t* de Student a récemment été proposé dans [Yoshii et al., 2016] pour une application de séparation de sources monocanale. Dans cet article, les coefficients de la TFCT ($\mathbb{K} = \mathbb{C}$) d'une source sont modélisés de la façon suivante :

$$s_{fn} \sim \mathcal{T}_\alpha\left(0, v_{fn}^{1/2}\right). \tag{2.16}$$

Comme nous le verrons à la section 2.3.4b ci-après, les paramètres d'échelle v_{fn} sont de plus structurés par un modèle NMF.

Nous pouvons également mentionner qu'au même moment, un modèle de source hiérarchique similaire à la représentation SMoG (2.14) et également basé sur la NMF a été proposé dans [Kounades-Bastian et al., 2016]. Bien qu'il n'y soit pas explicitement fait référence, ce modèle correspond bien à une distribution marginale pour les coefficients TF des sources qui soit *t* de Student.

2.3 Factorisation en matrices non-négatives

La popularité du modèle gaussien local présenté précédemment vient notamment du fait qu'il permet d'introduire une modélisation paramétrique de la puissance des sources dans le plan TF, au travers donc du terme de variance v_{fn}. Comme déjà mentionné, l'importance d'une telle paramétrisation réside en partie dans le caractère sous-déterminé du problème de séparation de sources. En effet chaque source est caractérisée par $F \times N$ paramètres $\{v_{fn}\}_{f,n}$ qui doivent être estimés. L'idée est alors de représenter la puissance de chaque source par un ensemble de paramètres de cardinal inférieur. Dans cette section nous présentons un modèle paramétrique basé sur la technique de réduction de rang par NMF. Avant d'introduire la NMF dans le cadre des modèles de source probabilistes présentés ci-dessus, nous revenons sur les origines de la NMF, comme technique de réduction de rang pour l'analyse de données à valeurs positives.

2.3.1 Modèle de factorisation en matrices non-négatives

Le modèle de NMF a été introduit dans [Lee et Seung, 1999]. Soit $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ une matrice de données positives observées, la NMF a pour objectif d'estimer deux matrices $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ et $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ telles que :

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\mathbf{H}. \tag{2.17}$$

K correspond au rang de la factorisation et est généralement choisi tel que $K(F+N) \ll FN$. Contrairement à d'autres techniques de réduction de dimensions comme l'analyse en composantes principales, la contrainte de positivité de la NMF permet d'obtenir une décomposition dont les facteurs sont naturellement interprétables. En effet les données sont expliquées par des combinaisons additives et non soustractives des colonnes de la matrice $\mathbf{W}$, parfois appelée dictionnaire. La NMF correspond donc à un *modèle compositionnel* [Virtanen et al., 2015] où les données sont représentés par une somme de K composantes de rang 1, elles-mêmes à coefficients positifs :

$$\hat{\mathbf{V}} = \sum_{k=1}^{K} \hat{\mathbf{V}}_k, \qquad \hat{\mathbf{V}}_k = (\mathbf{W})_{:,k}(\mathbf{H})_{k,:}\,. \tag{2.18}$$

En audio, la matrice de données correspond généralement au spectrogramme d'amplitude ou de puissance du signal à analyser. La matrice $\mathbf{W}$ coïncide alors avec un dictionnaire d'atomes spectraux et $\mathbf{H}$ contient les activations de ces atomes au cours du temps. Par exemple, le signal à partir duquel la NMF de la figure 2.3 est calculée contient deux notes de piano jouées individuellement puis simultanément. Nous voyons que les atomes fréquentiels correspondent aux spectres harmoniques de ces deux notes. Les activations temporelles indiquent quant à elles l'évolution de l'amplitude de ces notes au cours du temps.

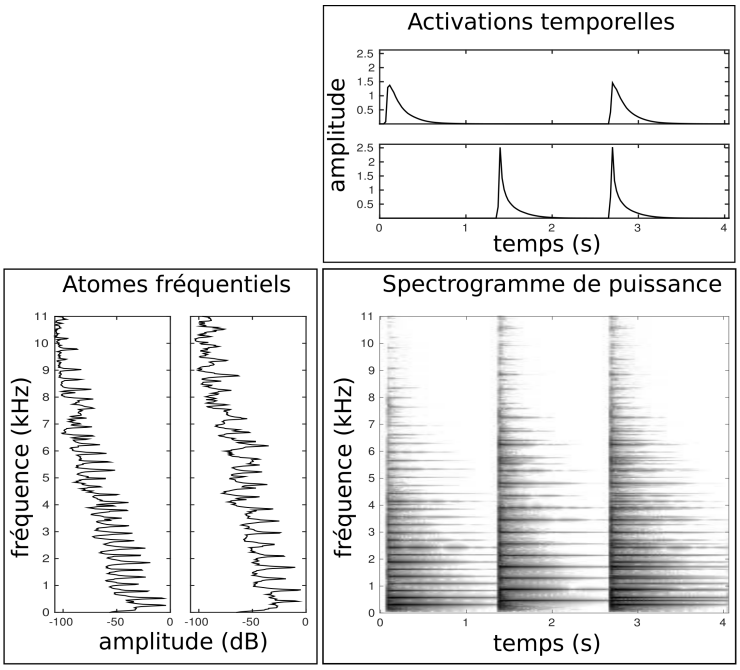


FIGURE 2.3 – Exemple de factorisation en matrices non-négatives d'un spectrogramme de puissance calculé à partir de la TFCT. Le signal se compose de deux notes de piano. Elles sont tout d'abord jouées séparément puis simultanément tel qu'indiqué par les activations temporelles.

2.3.2 Estimation par minimisation d'une divergence

Les paramètres du modèle NMF sont le plus souvent estimés en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}, \mathbf{WH}), \tag{2.19}$$

où $\mathbf{W}, \mathbf{H} \geq 0$ représente la contrainte de positivité des coefficients des matrices $\mathbf{W}$ et $\mathbf{H}$, et $D(\mathbf{V}, \mathbf{WH})$ est une mesure de dissimilarité séparable en chaque point de la matrice de données :

$$D(\mathbf{V}, \mathbf{WH}) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} d((\mathbf{V})_{f,n}, (\mathbf{WH})_{f,n}). \tag{2.20}$$

$d(x, y)$ est généralement choisi comme étant une divergence, c'est-à-dire une fonction à valeurs positives telle que $d(x, y) = 0$ si et seulement si $x = y$. La famille des β-divergences [Cichocki et Amari, 2010] est particulièrement populaire dans la littérature de NMF, celles-ci sont définies par :

$$d_\beta(x, y) = \begin{cases} \dfrac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta\, x\, y^{\beta-1}) & \beta \in \mathbb{R} \setminus \{0, 1\} \\ \dfrac{x}{y} - \ln\left(\dfrac{x}{y}\right) - 1 & \beta = 0 \\ x \ln\left(\dfrac{x}{y}\right) + y - x & \beta = 1 \end{cases} . \tag{2.21}$$

La famille des β-divergences admet comme cas particuliers la divergence d'Itakura-Saito ($\beta = 0$) que l'on notera par la suite $d_{IS}(\cdot, \cdot)$, la divergence de Kullback-Leibler généralisée ($\beta = 1$) et la distance euclidienne ($\beta = 2$).

Pour toute valeur de β et tout facteur $\gamma \in \mathbb{R}_+$, la β-divergence satisfait la propriété suivante [Févotte et al., 2009] :

$$d_\beta(\gamma x, \gamma y) = \gamma^\beta d_\beta(x, y). \tag{2.22}$$

Cette propriété indique que pour $\beta > 0$ (respectivement $\beta < 0$), les coefficients de la matrice de données ayant une grande (respectivement faible) amplitude auront un poids amplifié dans la résolution du problème d'optimisation défini par les équations (2.19) et (2.20). A l'inverse, dans le cas $\beta = 0$, la divergence est invariante à un changement d'échelle, ce qui peut être une propriété désirable. Par exemple, l'énergie dans les sons naturels a tendance à décroître plus rapidement dans les hautes fréquences que dans les basses fréquences. Cependant il peut être important de ne pas négliger ces composantes moins énergétiques dans l'estimation du modèle NMF car d'un point de vue perceptif elles ont un rôle important. Dans ce cas on pourra préférer utiliser la divergence d'Itakura-Saito.

Une approche heuristique pour estimer les paramètres du modèle NMF consiste à mettre à jour de façon itérative chaque paramètre scalaire θ (une entrée des matrices $\mathbf{W}$ ou $\mathbf{H}$) en multipliant sa valeur courante par le rapport des parties négative et positive de la dérivée du critère par rapport à ce paramètre notée $\nabla\mathcal{C}(\theta) = \nabla^+\mathcal{C}(\theta) - \nabla^-\mathcal{C}(\theta)$. La mise à jour s'écrit : $\theta \leftarrow \theta \times \nabla^-\mathcal{C}(\theta)/\nabla^+\mathcal{C}(\theta)$. Cette approche conduit aux mises à jours multiplicatives suivantes lorsque le critère correspond aux β-divergences [Févotte et al., 2009] :

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^\top \left[(\mathbf{WH})^{\odot\beta-2} \odot \mathbf{V}\right]}{\mathbf{W}^\top (\mathbf{WH})^{\odot\beta-1}}; \tag{2.23}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left[(\mathbf{WH})^{\odot\beta-2} \odot \mathbf{V}\right] \mathbf{H}^\top}{(\mathbf{WH})^{\odot\beta-1} \mathbf{H}^\top}, \tag{2.24}$$

où l'on rappelle que $\odot$ représente une opération élément par élément (ici la multiplication et la mise à la puissance) et la division de matrices doit également être comprise élément par élément. La contrainte de positivité des coefficients des matrices $\mathbf{W}$ et $\mathbf{H}$ est satisfaite si ces derniers sont initialisés avec des valeurs positives.

Les propriétés de convergence des algorithmes de mises à jour multiplicatives pour les β-divergences ont été analysées dans [Badeau et al., 2010], dans le cadre de la théorie de la stabilité au sens de Lyapunov.

Finalement, il a été démontré dans [Févotte et Idier, 2011] que pour $\beta \in [0, 2]$, les règles de mise à jour multiplicatives (2.23) et (2.24) garantissent la décroissance du critère. Les auteurs ont pour cela utilisé la méthode de la fonction auxiliaire, en décomposant $d_\beta(x, y)$ comme la somme de fonctions convexe, concave et constante par rapport à y. Une majorante de la partie convexe est obtenue grâce à l'inégalité de Jensen tandis que la partie concave est majorée localement par sa tangente. Le problème d'optimisation de départ est alors remplacé par une minimisation de la fonction auxiliaire permettant de garantir la décroissance du critère original.

2.3.3 Méthode de la fonction auxiliaire

Comme par la suite nous serons amenés à utiliser la méthode de la fonction auxiliaire, nous la présentons ici (voir par exemple [Kameoka et al., 2008]). Soit $\mathcal{C}(\theta)$ une fonction de coût que l'on souhaite minimiser par rapport à θ. La fonction $\mathcal{G}(\theta, \tilde{\theta})$ est une fonction auxiliaire de $\mathcal{C}(\theta)$ si elle satisfait :

$$\mathcal{C}(\theta) = \min_{\tilde{\theta}} \mathcal{G}(\theta, \tilde{\theta}), \tag{2.25}$$

où $\tilde{\theta}$ est une variable auxiliaire. Autrement dit, $\mathcal{G}(\theta, \tilde{\theta})$ est une majorante de $\mathcal{C}(\theta)$; pour tout $\tilde{\theta}$, $\mathcal{C}(\theta) \leq \mathcal{G}(\theta, \tilde{\theta})$. Le problème de minimisation de $\mathcal{C}(\theta)$ peut alors être remplacé par une minimisation alternée de la majorante suivant θ et $\tilde{\theta}$. En pratique, pour toute mise à jour $\theta^{\mathtt{it}+1}$ telle que

$$\mathcal{G}(\theta^{\mathtt{it}+1}, \tilde{\theta}^{\mathtt{it}}) \leq \mathcal{G}(\theta^{\mathtt{it}}, \tilde{\theta}^{\mathtt{it}}), \tag{2.26}$$

avec

$$\tilde{\theta}^{\mathtt{it}} = \arg\min_{\tilde{\theta}} \mathcal{G}(\theta^{\mathtt{it}}, \tilde{\theta}), \tag{2.27}$$

nous avons $\mathcal{C}(\theta^{\mathtt{it}+1}) \leq \mathcal{C}(\theta^{\mathtt{it}})$ car

$$\mathcal{C}(\theta^{\mathtt{it}+1}) \leq \mathcal{G}(\theta^{\mathtt{it}+1}, \tilde{\theta}^{\mathtt{it}}) \leq \mathcal{G}(\theta^{\mathtt{it}}, \tilde{\theta}^{\mathtt{it}}) = \min_{\tilde{\theta}} \mathcal{G}(\theta^{\mathtt{it}}, \tilde{\theta}) = \mathcal{C}(\theta^{\mathtt{it}}). \tag{2.28}$$

On voit en particulier d'après l'équation (2.27) que la variable auxiliaire $\tilde{\theta}$ à l'itération `it` dépend de la valeur courante du paramètre θ à l'itération `it`.

L'algorithme majoration-minimisation [Hunter et Lange, 2004] consiste à choisir la mise à jour suivante pour $\theta^{\mathtt{it}+1}$:

$$\theta^{\mathtt{it}+1} = \arg\min_{\theta} \mathcal{G}(\theta, \tilde{\theta}^{\mathtt{it}}). \tag{2.29}$$

Cependant, toute mise à jour satisfaisant (2.26) conduit à une décroissance monotone de $\mathcal{C}(\theta)$.

2.3.4 La NMF dans les modèles de source probabilistes

Nous avons présenté ci-dessus la NMF dans un cadre déterministe. Il est également possible de formaliser la NMF dans un cadre probabiliste, comme un problème d'estimation de paramètres d'un modèle d'observation probabiliste. Deux approches sont possibles :

1. Modéliser la matrice de données à coefficients positifs $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ (par exemple un spectrogramme d'amplitude ou de puissance). Un modèle de bruit additif gaussien est par exemple utilisé dans [Schmidt et Laurberg, 2008], tandis qu'un modèle de bruit multiplicatif de loi gamma est introduit dans [Févotte et al., 2009]. L'analyse probabiliste en composantes latentes correspond à un modèle de comptage où la matrice $\mathbf{V}$ est vue comme une distribution de probabilité [Shashanka et al., 2008]. Le modèle proposé dans [Virtanen et al., 2008] suppose que la matrice $\mathbf{V}$ est générée suivant une somme de composantes latentes non-négatives distribuées selon une loi de Poisson.

2. Modéliser la matrice $\mathbf{S} = [s_{fn}]_{f,n} \in \mathbb{K}^{F \times N}$ (à coefficients réels ou complexes) contenant les coefficients TF du signal. Un modèle génératif gaussien a par exemple été proposé dans [Févotte et al., 2009] ; chaque coefficient TF s_{fn} est modélisé comme une somme de composantes latentes gaussiennes dont la variance s'exprime par un modèle NMF. Un modèle génératif reposant sur la distribution de Cauchy a été introduit dans [Liutkus et al., 2015] et généralisé au cas des distributions symétriques alpha-stables dans [Şimşekli et al., 2015]. Enfin, un modèle d'observation basé sur la distribution *t* de Student a récemment été proposé dans [Yoshii et al., 2016].

Nous allons détailler ci-après l'utilisation de la NMF pour les modèles de source gaussien (introduit à la section 2.2.2) et *t* de Student (introduit à la section 2.2.3b). Dans ces deux cas, la NMF est utilisée pour structurer les paramètres d'échelle de la distribution des coefficients TF des sources.

a) Modèle gaussien structuré en variance

A notre connaissance, un modèle génératif gaussien TF exploitant une décomposition non-négative de la puissance à court-terme des sources a été pour la première fois proposé dans [Benaroya et al., 2003], dans le cadre d'une application de séparation de sources monocanale. Le lien entre ce modèle et la NMF a ensuite été explicitement établi dans [Févotte et al., 2009], où notamment il a été prouvé l'équivalence entre une estimation des paramètres au sens du maximum de vraisemblance et une NMF basée sur la divergence d'Itakura-Saito. Ce modèle génératif gaussien a ensuite été exploité pour la séparation de sources multicanale dans [Ozerov et Févotte, 2010].

Rappelons que dans le modèle de source gaussien présenté précédemment, chaque coefficient TF d'une source $s_{fn} \in \mathbb{K}$ est représenté comme une variable aléatoire suivant une loi gaussienne $\mathcal{N}(0, v_{fn})$. Les points TF sont de plus supposés indépendants. Le modèle proposé dans [Févotte et al., 2009] consiste à structurer la matrice $\mathbf{V} = [v_{fn}]_{f,n} \in \mathbb{R}_+^{F \times N}$ par l'intermédiaire d'un modèle NMF :

$$\mathbf{V} = \mathbf{W}\mathbf{H}, \tag{2.30}$$

où $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, $\mathbf{H} \in \mathbb{R}_+^{K \times N}$.

Soit $\mathbf{s} = \{s_{fn}\}_{f,n}$. D'après l'expression (2.8) de la densité de probabilité de loi gaussienne, la log-vraisemblance de ce modèle s'écrit de la façon suivante :

$$\begin{aligned} \ln p\left(\mathbf{s}; \mathbf{W}, \mathbf{H}\right) &= \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \ln p(s_{fn}; \mathbf{W}, \mathbf{H}) \\ &= -\frac{FN}{\phi} \ln(\phi\pi) - \frac{1}{\phi} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \left[\ln\left[(\mathbf{W}\mathbf{H})_{f,n}\right] + \frac{|s_{fn}|^2}{(\mathbf{W}\mathbf{H})_{f,n}} \right], \\ &\stackrel{c}{=} -\frac{1}{\phi} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} d_{IS}(|s_{fn}|^2, (\mathbf{W}\mathbf{H})_{f,n}), \end{aligned} \tag{2.31}$$

où la divergence d'Itakura-Saito est définie à l'équation (2.21) et $\stackrel{c}{=}$ traduit l'égalité à une constante additive près, ici par rapport aux paramètres de NMF $\mathbf{W}$ et $\mathbf{H}$.

Comme cela a été montré dans [Févotte et al., 2009], on voit que maximiser la log-vraisemblance par rapport aux paramètres de NMF sous une contrainte de positivité est équivalent à minimiser la divergence d'Itakura-Saito entre le spectrogramme de puissance de la source et la factorisation NMF, sous la même contrainte. Nous pouvons alors utiliser les règles de mises à jour multiplicatives (2.23) et (2.24) dans le cas $\beta = 0$.

b) Modèle *t* de Student

Rappelons le modèle de source *t* de Student introduit à la section 2.2.3b : $s_{fn} \sim \mathcal{T}_\alpha(0, v_{fn}^{1/2})$. Le modèle proposé dans [Yoshii et al., 2016] consiste à structurer les facteurs d'échelle v_{fn} par l'intermédiaire d'un modèle NMF de la même façon que pour le modèle Gaussien à l'équation (2.30). Ci-dessous nous montrons comment les paramètres NMF de ce modèle de source *t* de Student peuvent être estimés au sens du maximum de vraisemblance. De façon très similaire à l'approche développée dans [Yoshii et al., 2016], nous allons utiliser la méthode de la fonction auxiliaire. Cependant nous traitons conjointement les cas de transformations TF à valeurs réelles et complexes. De plus nous montrons que la fonction auxiliaire obtenue fait intervenir la divergence d'Itakura-Saito, ce qui nous permet d'utiliser les règles de mise à jour multiplicatives qui ont été données précédemment. Dans l'article [Yoshii et al., 2016], la divergence d'Itakura-Saito

n'intervient pas explicitement dans le problème d'estimation des paramètres du modèle NMF *t* de Student.

D'après l'expression (2.13) de la densité de probabilité de loi *t* de Student, la log-vraisemblance s'écrit de la façon suivante :

$$\begin{aligned}\ln p\left(\mathbf{s};\mathbf{W},\mathbf{H}\right) &= \sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\ln p(s_{fn};\mathbf{W},\mathbf{H}) \\ &= FN\left[\ln\left(\frac{2}{\phi}\right)-\frac{1}{\phi}\ln(\alpha\pi)+\ln\Gamma\left(\frac{\alpha}{2}+\frac{1}{\phi}\right)-\ln\Gamma\left(\frac{\alpha}{2}\right)\right] \\ &\quad -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\frac{1}{\phi}\ln\left[(\mathbf{WH})_{f,n}\right]+\left(\frac{\alpha}{2}+\frac{1}{\phi}\right)\ln\left(1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}\right)\right] \\ &\overset{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\frac{1}{\phi}\ln\left[(\mathbf{WH})_{f,n}\right]+\left(\frac{\alpha}{2}+\frac{1}{\phi}\right)\ln\left(1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}\right)\right]. \quad (2.32)\end{aligned}$$

Nous allons utiliser la méthode de la fonction auxiliaire afin de minimiser l'opposée de la log-vraisemblance par rapport aux paramètres de NMF. Soit $g:\mathbb{R}\mapsto\mathbb{R}$ une fonction continûment dérivable et strictement concave. Alors pour tout réel u_0, nous pouvons majorer g par sa tangente en ce point :

$$g(u)\leq g(u_0)+g'(u_0)(u-u_0), \quad (2.33)$$

avec égalité si et seulement si $u=u_0$. Comme $u\mapsto\ln(u)$ est une fonction strictement concave définie sur $]0,+\infty[$, pour tout ensemble de variables auxiliaires $\mathbf{c}=\{c_{fn}\in]0,+\infty[\}_{f,n}$ nous pouvons écrire l'inégalité suivante :

$$\ln\left(1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}\right)\leq\ln(c_{fn})+\frac{1}{c_{fn}}\left(1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}-c_{fn}\right), \quad (2.34)$$

avec égalité si et seulement si

$$c_{fn}=1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}. \quad (2.35)$$

En appliquant cette inégalité à l'équation (2.32) on obtient la fonction auxiliaire suivante qui majore l'opposée de la log-vraisemblance :

$$\begin{aligned}\mathcal{G}\left(\mathbf{W},\mathbf{H},\mathbf{c}\right) &= \sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\frac{1}{\phi}\ln\left[(\mathbf{WH})_{f,n}\right]\right. \\ &\qquad \left.+\left(\frac{\alpha}{2}+\frac{1}{\phi}\right)\left(\ln(c_{fn})+\frac{1}{c_{fn}}\left(1+\frac{2}{\phi\alpha}\frac{|s_{fn}|^2}{(\mathbf{WH})_{f,n}}-c_{fn}\right)\right)\right] \\ &\overset{c}{=}\frac{1}{\phi}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\ln\left[(\mathbf{WH})_{f,n}\right]+\frac{\phi\alpha+2}{\phi\alpha}\frac{|s_{fn}|^2}{c_{fn}}\frac{1}{(\mathbf{WH})_{f,n}}\right] \\ &\overset{c}{=}\frac{1}{\phi}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}d_{IS}\left(p_{fn},(\mathbf{WH})_{f,n}\right), \quad (2.36)\end{aligned}$$

avec $p_{fn}=\dfrac{\phi\alpha+2}{\phi\alpha}\dfrac{|s_{fn}|^2}{c_{fn}}$.

De la même façon que pour le modèle gaussien nous faisons apparaître la divergence d'Itakura-Saito. Cependant elle intervient ici non directement dans l'expression de la log-vraisemblance mais dans celle de la fonction auxiliaire $\mathcal{G}(\mathbf{W}, \mathbf{H}, \mathbf{c})$. La méthode de la fonction auxiliaire consiste finalement à itérer les deux étapes suivantes :

1. Minimiser (ou seulement faire décroître) $\mathcal{G}(\mathbf{W}, \mathbf{H}, \mathbf{c})$ par rapport aux paramètres de NMF sous une contrainte de positivité, et à variables auxiliaires fixées. Une fois de plus nous pouvons avoir recours aux règles multiplicatives (2.23) et (2.24), en utilisant les valeurs courantes des variables auxiliaires ;
2. Minimiser $\mathcal{G}(\mathbf{W}, \mathbf{H}, \mathbf{c})$ par rapport aux variables auxiliaires, à paramètres NMF fixés. Il est immédiat de vérifier à partir de l'équation (2.34) que cela correspond à mettre à jour les variables auxiliaires suivant l'équation (2.35), en utilisant les valeurs courantes des paramètres NMF.

Comme indiqué à la section 2.3.3, cette procédure garantie la décroissance du critère.

Par ailleurs, en injectant (2.35) dans l'expression de p_{fn} on obtient :

$$p_{fn} = \left(\frac{\phi\alpha |s_{fn}|^{-2} + 2(\mathbf{WH})_{f,n}^{-1}}{\phi\alpha + 2} \right)^{-1}. \tag{2.37}$$

Il est intéressant de remarquer que p_{fn} correspond à une moyenne harmonique pondérée entre le spectrogramme de puissance et la factorisation NMF au point TF (f, n), en utilisant la valeur courante des paramètres. A chaque itération de l'algorithme on cherche donc à mettre à jour les paramètres en calculant une NMF sur la matrice $\mathbf{P} = [p_{fn}]_{f,n} \in \mathbb{R}_+^{F \times N}$ formée par ces moyennes harmoniques. De plus, lorsque $\alpha \to \infty$, la distribution t de Student approche la gaussienne et $p_{fn} \to |s_{fn}|^2$. On se retrouve alors à résoudre le même problème d'optimisation que dans le cas gaussien, en utilisant les mêmes mises à jour. Finalement, pour des valeurs suffisamment faibles de α, il a été montré expérimentalement dans [Yoshii et al., 2016] que le modèle de NMF t de Student est très robuste à l'initialisation des paramètres. Il s'agissait dans cet article de décomposer individuellement des signaux de piano, guitare ou clarinette (comprenant au maximum 3 notes) en composantes de rang 1.

2.4 Modèles pour les mélanges fortement réverbérants

Nous présentons maintenant plusieurs modèles de mélange convolutif ayant été proposés dans la littérature afin de remédier aux limitations de l'hypothèse de filtres de mélange à réponse impulsionnelle courte, celle-ci n'étant pas valide lorsque le mélange est fortement réverbérant.

2.4.1 Modèle convolutif diffus

Le modèle convolutif sous l'hypothèse bande étroite introduit à la section 1.2.2 du chapitre 1 nous permet de faire le lien suivant entre les coefficients de la TFCT d'un signal source monophonique et la source image multicanale qui lui est associée :

$$\mathbf{y}_{j,fn} = \mathbf{a}_{j,f} s_{j,fn}, \tag{2.38}$$

avec $\mathbf{a}_{j,f} = [a_{ij,f}]_i^\top \in \mathbb{C}^I$.

Ce modèle, que nous appellerons «modèle convolutif ponctuel» par la suite, permet d'écrire le mélange comme à l'équation (2.5), par l'intermédiaire d'une matrice qui dépend de la fréquence. Il peut être généralisé au cas de sources étendues spatialement ou diffuses (émettant dans plusieurs

directions) [Ozerov et al., 2012; Duong et al., 2013]. Supposons que chaque source image dans le domaine de la TFCT ne soit plus associée à une unique source, mais corresponde à la somme de R_j sous-sources indépendantes et identiquement distribuées, «filtrées» par R_j filtres de mélange différents. On peut alors écrire :

$$\mathbf{y}_{j,fn} = \sum_{r=1}^{R_j} \mathbf{a}_{jr,f} s_{jr,fn}, \tag{2.39}$$

où pour chaque source et point TF, les variables aléatoires $s_{jr,fn}$, $r = 1, ..., R_j$, sont i.i.d. L'équation (2.39) peut se réécrire de façon matricielle :

$$\mathbf{y}_{j,fn} = \mathbf{A}_{j,f} \mathbf{s}_{j,fn}, \tag{2.40}$$

où $\mathbf{A}_{j,f} = [\mathbf{a}_{jr,f}]_r \in \mathbb{C}^{I \times R_j}$ et $\mathbf{s}_{j,fn} = [s_{jr,fn}]_r^\top \in \mathbb{C}^{R_j}$. Finalement, le mélange s'écrit de la façon suivante pour chaque point TF :

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn}, \tag{2.41}$$

où $\mathbf{A}_f = [\mathbf{A}_{j,f}]_j \in \mathbb{C}^{I \times R}$ est la matrice de mélange et $\mathbf{s}_{fn} = [\mathbf{s}_{j,fn}^\top]_j^\top \in \mathbb{C}^R$ est le vecteur contenant les coefficients des TFCTs des $R = \sum_{j=1}^J R_j$ sous-sources. Le modèle de mélange pour des sources diffuses est donc identique au cas ponctuel, il s'agit uniquement de considérer que chaque source j est associée à R_j réalisations d'une même variable aléatoire. Si pour tout j, $R_j = 1$, on retrouve le modèle convolutif ponctuel sous l'approximation bande étroite.

2.4.2 Matrice de covariance spatiale

Les modèles convolutifs diffus et ponctuels peuvent être regroupés sous un même formalisme plus général introduit dans [Vincent et al., 2010] et largement utilisé depuis [Duong et al., 2010; Arberet et al., 2010; Ozerov et al., 2012; Duong et al., 2013]. Le mélange est uniquement exprimé comme une somme de sources images modélisées comme des vecteurs aléatoires de $\mathbb{C}^I$. Ce modèle fait intervenir une *matrice de covariance spatiale*. Nous l'écrivons ci-dessous dans sa formulation originale c'est-à-dire dans le cas gaussien :

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{j,fn}, \qquad \mathbf{y}_{j,fn} \sim \mathcal{N}(0, v_{j,fn} \mathbf{R}_{j,f}). \tag{2.42}$$

Le paramètre $v_{j,fn} \in \mathbb{R}_+$ modélise la puissance de la source dans le domaine TF. La matrice $\mathbf{R}_{j,f} \in \mathbb{C}^{I \times I}$, appelée matrice de covariance spatiale, représente les caractéristiques spatiales de la source. Elle encode notamment les corrélations entre les différents canaux.

Matrice de covariance spatiale structurée Les modèles convolutifs présentés précédemment correspondent à des instances particulières de l'équation du mélange (2.42). Dans le cas particulier du mélange convolutif ponctuel (2.38), la matrice de covariance spatiale est de rang 1 et admet la représentation suivante :

$$\mathbf{R}_{j,f} = \mathbf{a}_{j,f} \mathbf{a}_{j,f}^H, \qquad \mathbf{a}_{j,f} \in \mathbb{C}^I. \tag{2.43}$$

Dans le cas du mélange convolutif diffus (2.40), la matrice de covariance spatiale est de rang $0 < R_j \leq I$ et s'écrit :

$$\mathbf{R}_{j,f} = \mathbf{A}_{j,f} \mathbf{A}_{j,f}^H, \qquad \mathbf{A}_{j,f} \in \mathbb{C}^{I \times R_j}. \tag{2.44}$$

Cette formulation du problème de séparation par sources images est très générale. En effet il s'agit uniquement de supposer des sources images gaussiennes (ou plus généralement suivant toute

distribution multivariée elliptique également définie par une «matrice d'échelle») dont la matrice de covariance se factorise en le produit de deux termes : la puissance spectrale $v_{j,fn}$ et la matrice de covariance spatiale $\mathbf{R}_{j,f}$. Le modèle de source image par matrice de covariance spatiale de rang plein a été introduit dans [Duong et al., 2010] pour modéliser des sources spatialement diffuses. Néanmoins, il a été montré expérimentalement dans ce même article que la flexibilité induite par une matrice de covariance spatiale de rang plein permet en présence de forte réverbération d'améliorer les résultats de séparation par rapport au modèle convolutif ponctuel. Il est cependant peut-être moins direct d'interpréter le rôle physique de la matrice de covariance spatiale et donc d'introduire des contraintes sur cette dernière. Par ailleurs, dans le cas du modèle structuré (2.44), $\mathbf{A}_{j,f}$ ne peut être interprétée comme une matrice de réponses en fréquence de salle que sous l'hypothèse de filtres de mélange courts par rapport à la fenêtre d'analyse de la TFCT. Cette hypothèse est une des principales limitations du modèle convolutif ponctuel et nous voyons ici que d'un point de vue théorique, elle n'est pas levée par l'introduction des matrices de covariance spatiale. Rappelons cependant encore une fois que d'un point de vue pratique, le modèle de matrice de covariance spatiale de rang plein s'avère généralement plus robuste (en terme de performances) à des mélanges fortement réverbérants, par rapport au modèle convolutif ponctuel [Duong et al., 2010].

2.4.3 Convolution dans le domaine temps-fréquence

Il a été démontré dans [Badeau et Plumbley, 2014] que pour tout banc de filtres TF à reconstruction parfaite, la convolution dans le domaine temporel peut être représentée de façon exacte par une convolution à deux dimensions dans ce domaine TF. Cette représentation TF de la convolution temporelle est appelée *cross-band filtering* (CBF) dans [Avargel et Cohen, 2007b] lorsque le banc de filtres correspond à la TFCT.

Indépendamment, il a été proposé dans [Attias, 2003] d'approcher la convolution dans le domaine temporel par un filtrage dans chaque sous-bande de la TFCT du signal source. Cette approche revient à négliger les relations entre bandes de fréquence dans la représentation par CBF. Cela correspond au modèle *convolutive transfer function* (CTF) récemment utilisé dans les méthodes de séparation de sources [Li et al., 2017a] et [Li et al., 2017b]. Dans ces travaux, la TFCT d'une source image est reliée à celle du signal source monophonique par l'équation suivante :

$$y_{ij,fn} = [a_{ij,f\cdot} \star s_{f\cdot}](n), \tag{2.45}$$

où $a_{ij,fn} = [a_{ij}(\cdot) \star \xi_f(\cdot)](nH)$ avec :

$$\xi_f(t) = e^{\imath 2\pi \frac{ft}{L_w}} \sum_{m=0}^{L_w-1} w_a(m) w_s(t+m), \tag{2.46}$$

où $w_a(t)$ et $w_s(t)$ sont respectivement les fenêtres d'analyse et de synthèse de support $\{0, ..., L_w - 1\}$. Le signal $\xi_f(t)$ étant défini sur le support $\{-(L_w-1), ..., L_w-1\}$, le filtre $a_{ij,fn}$ impliqué dans l'équation (2.45) est en général non-causal. D'après l'équation (2.46), $a_{ij,fn}$ peut se réécrire de la façon suivante :

$$a_{ij,fn} = \sum_{\tau=0}^{L_a-1} a_{ij}(\tau) e^{-\imath 2\pi \frac{f\tau}{L_w}} e^{\imath 2\pi \frac{fnH}{L_w}} \epsilon(nH-\tau), \tag{2.47}$$

où $\epsilon(t) = \sum_{m=0}^{L_w-1} w_a(m) w_s(t+m)$ est un signal de support $\{-(L_w-1), ..., L_w-1\}$.

Dans [Li et al., 2017a], le problème de séparation de sources est formulé sous la forme d'un problème d'optimisation où le terme d'attache aux données fait intervenir le modèle CTF tandis

qu'une pénalisation de la norme ℓ_1 des coefficients TF des sources permet de promouvoir leur parcimonie. Dans cet article les filtres de mélange sont supposés connus, à la fois pour l'algorithme développé et pour les expériences. Les mêmes auteurs ont ensuite proposé dans [Li et al., 2017b] d'exploiter le modèle de mélange CTF dans le cadre d'une approche probabiliste basée sur un modèle de source gaussien dans le domaine TF. Des règles de mises à jour pour les paramètres du modèle CTF sont fournies, mais comme précédemment les expériences sont conduites en supposant la connaissance des filtres de mélange. Les résultats présentés dans ces deux articles ont permis de montrer la grande robustesse du modèle CTF pour représenter des mélanges fortement réverbérants.

2.4.4 Représentation exacte dans le domaine temporel

Finalement, il a été proposé dans [Kowalski et al., 2010] de représenter le mélange convolutif de façon exacte en restant dans le domaine temporel, tout en exploitant cependant un modèle de source défini dans le domaine TF. Cette méthode repose également sur la formulation d'un problème d'optimisation où le terme d'attache aux données fait intervenir les signaux temporels captés par les microphones, tandis qu'un modèle de source parcimonieux est utilisé par l'intermédiaire d'une pénalisation ℓ_1 des coefficients TF des sources. Cette approche a ensuite été étendue dans [Arberet et Vandergheynst, 2014] en introduisant une contrainte de rang faible sur le spectrogramme des sources.

Un des avantages lié à l'utilisation d'une représentation temporelle du mélange est que cela permet plus de flexibilité au niveau de la modélisation des sources. Il est par exemple possible d'utiliser des dictionnaires TF différents pour chaque source, ou bien des dictionnaires hybrides permettant une décomposition des sources en composantes tonales et percussives [Feng et Kowalski, 2014].

2.5 Estimation et inférence statistique

L'ensemble des méthodes de séparation de sources développées dans cette thèse s'appuiera sur des modèles probabilistes. C'est pourquoi il nous semble important de définir dans cette première partie le principe et les techniques d'estimation et d'inférence statistique qui seront utilisées par la suite.

2.5.1 Modèle probabiliste

Soit $\mathbf{x}$ un ensemble de données observées et $\boldsymbol{\theta}$ l'ensemble des paramètres inconnus ayant généré les observations. Une des approches les plus populaires en statistique consiste à estimer les paramètres du modèle au sens du maximum de vraisemblance (MV) :

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}). \tag{2.48}$$

La vraisemblance $p(\mathbf{x};\boldsymbol{\theta})$ caractérise la façon dont les données ont été générées à partir des paramètres. La définition d'un modèle probabiliste consiste précisément à expliciter cette vraisemblance. Sous certaines hypothèses, l'estimateur MV est notamment convergent, c'est-à-dire qu'il converge en probabilité vers le vrai paramètre, et asymptotiquement efficace, c'est-à-dire qu'aucun autre estimateur convergent n'a de variance asymptotique plus faible [Wasserman, 2013].

Nous pouvons également être amenés à définir des a priori sur les paramètres du modèle, afin d'exprimer une certaine connaissance sur ces derniers. Dans ce cas les paramètres ne sont plus

déterministes mais sont considérés comme des variables aléatoires. On espère alors que l'information ajoutée par l'intermédiaire de l'a priori permettra d'obtenir une meilleure estimation des paramètres. De plus, cet ajout d'information peut être particulièrement important dans le contexte d'un problème mal posé où une infinité de solutions permettent d'expliquer les mêmes données observées. Dans ce cas l'a priori permet de régulariser le problème en imposant certaines contraintes aux variables d'intérêt devant être estimées.

Une approche répandue pour prendre en compte un a priori sur les paramètres consiste à estimer ces derniers au sens du maximum a posteriori (MAP) :

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}), \tag{2.49}$$

où $p(\boldsymbol{\theta}|\mathbf{x})$ caractérise la distribution a posteriori des paramètres (une fois les données observées). En utilisant la règle de Bayes, la densité de probabilité de cette distribution se réécrit $p(\boldsymbol{\theta}|\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{x})$. L'estimation MAP consiste alors à calculer :

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.50}$$

En général on cherche plutôt à maximiser le logarithme de la vraisemblance, $\ln p(\mathbf{x};\boldsymbol{\theta})$, appelé log-vraisemblance, ou dans le cas d'une estimation MAP le logarithme de la densité de probabilité a posteriori, $\ln p(\boldsymbol{\theta}|\mathbf{x}) \stackrel{c}{=} \ln p(\mathbf{x}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$.

2.5.2 Modèle à variables latentes

Comme nous l'avons vu précédemment, définir un modèle probabiliste consiste à expliquer le processus de génération des données au travers de la vraisemblance. Il peut être parfois utile d'introduire un ensemble de *variables latentes* (ou cachées), noté $\mathbf{z}$, afin de simplifier la définition du modèle. Ces variables permettent de faire le lien entre les paramètres et les observations, de telle sorte qu'il est en général plus facile de définir la distribution conditionnelle $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$ que la vraisemblance.

Dans ce cadre probabiliste à variables latentes, définir un modèle consiste à définir la distribution jointe des données cachées et observées. Sa densité de probabilité, également appelée *vraisemblance des données complètes* (comprendre les données observées et cachées), peut se décomposer sous la forme suivante :

$$p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})p(\mathbf{z};\boldsymbol{\theta}), \tag{2.51}$$

où $p(\mathbf{z};\boldsymbol{\theta})$ caractérise la distribution a priori sur les variables latentes et $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$ caractérise le processus de génération des données observées à partir des variables latentes. Dans un cadre bayésien, $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$ est parfois appelée vraisemblance et $p(\mathbf{x};\boldsymbol{\theta})$ vraisemblance marginale car obtenue en marginalisant $p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ par rapport aux variables latentes :

$$p(\mathbf{x};\boldsymbol{\theta}) = \int p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})d\mathbf{z}. \tag{2.52}$$

Le modèle probabiliste défini et la vraisemblance marginale calculée, nous pouvons finalement avoir accès à la distribution a posteriori des variables cachées grâce à la règle de Bayes :

$$p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) = \frac{p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})p(\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta})}. \tag{2.53}$$

Le calcul de cette distribution a posteriori est l'objectif principal de l'inférence bayésienne. De plus, il est souvent nécessaire d'obtenir une estimation ponctuelle des variables latentes à partir de cette distribution. On peut par exemple calculer la moyenne a posteriori :

$$\mathbf{z}^\star = \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^\star}[\mathbf{z}], \tag{2.54}$$

où $\boldsymbol{\theta}^\star$ désigne une estimation des paramètres, par exemple au sens du maximum de vraisemblance ou du maximum a posteriori.

On parlera d'inférence exacte lorsque le modèle probabiliste permet de calculer la distribution a posteriori des variables cachées. Cependant, pour des modèles complexes, il peut être impossible d'avoir accès à cette distribution car la vraisemblance (marginale) à l'équation (2.53) ne peut être calculée. La distribution a posteriori est donc connue a une constante de normalisation près qu'il n'est pas possible de calculer. Dans ce cas nous devons avoir recours à des techniques d'inférence approchée. Celles-ci se regroupent principalement en deux catégories :

1. Les méthodes d'échantillonnage telles que les approches de Monte-Carlo par chaînes de Markov [Bishop, 2006, chapitre 11]. Ces méthodes cherchent à approcher de façon numérique une intégrale n'ayant pas de forme analytique. Dans le cas du calcul de la moyenne a posteriori, ces méthodes consistent à tirer un ensemble d'échantillons i.i.d suivant la distribution a posteriori et à calculer une moyenne empirique comme estimateur de la moyenne a posteriori. Ces méthodes ont pour avantage de présenter des résultats théoriques assurant que sous certaines hypothèses concernant la chaîne de Markov, on échantillonne la vraie loi a posteriori. On peut de plus quantifier le biais et la variance de l'estimateur. Cependant en pratique elles sont coûteuses en temps de calcul et nécessitent de contrôler la convergence de la chaîne de Markov afin de s'assurer d'obtenir des échantillons i.i.d.
2. Les approches variationnelles, qui cherchent à calculer une approximation de la distribution a posteriori minimisant la divergence de Kullback-Leibler et vérifiant certaines contraintes [Bishop, 2006, chapitre 10]. Ces méthodes ne permettent en général pas d'obtenir la distribution a posteriori réelle, ni théoriquement, ni en pratique, cependant elles sont souvent plus rapides que les méthodes d'échantillonnage.

2.5.3 Inférence statistique

Dans un premier temps nous allons traiter le problème d'inférence exacte, lorsque la distribution a posteriori des variables latentes est connue. Nous allons pour cela introduire l'algorithme espérance-maximisation (EM). Celui-ci a été proposé à l'origine dans [Dempster et al., 1977], nous allons cependant le présenter en suivant l'approche utilisée dans [Bishop, 2006, chapitre 9], car cela nous sera utile pour introduire ensuite le concept d'inférence variationnelle.

Soit $\mathcal{F}$ un ensemble de densités de probabilités sur les variables latentes appelé famille variationnelle. Pour toute densité de probabilité $q \in \mathcal{F}$ et pour toute fonction $f(\mathbf{z})$, on note l'espérance mathématique $\langle f(\mathbf{z})\rangle_q = \int f(\mathbf{z})q(\mathbf{z})d\mathbf{z}$. On appellera *distribution variationnelle* la distribution de probabilité définie par la densité q.

Pour toute densité de probabilité $q \in \mathcal{F}$ et tout ensemble de paramètres $\boldsymbol{\theta}$, on peut montrer que la log-vraisemblance se décompose de la façon suivante :

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}(q;\boldsymbol{\theta}) + D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})), \tag{2.55}$$

avec $\mathcal{L}(q;\boldsymbol{\theta})$ *l'énergie variationnelle libre* définie par :

$$\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})}\right)\right\rangle_q, \tag{2.56}$$

et $D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}))$ la divergence de Kullback-Leibler entre la distribution variationnelle et la distribution a posteriori définie par :

$$D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \left\langle \ln\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}\right)\right\rangle_q. \tag{2.57}$$

L'énergie variationnelle libre se décompose en la somme d'un terme d'énergie $E(q;\boldsymbol{\theta})$ et de l'entropie différentielle de la distribution variationnelle notée $H(q)$:

$$\mathcal{L}(q;\boldsymbol{\theta}) = E(q;\boldsymbol{\theta}) + H(q), \tag{2.58}$$

avec

$$E(q;\boldsymbol{\theta}) = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_q; \tag{2.59}$$

$$H(q) = -\langle \ln q(\mathbf{z})\rangle_q. \tag{2.60}$$

La divergence de Kullback-Leibler satisfait $D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) \geq 0$, avec égalité si et seulement si $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$. On a donc d'après l'équation (2.55) $\ln p(\mathbf{x};\boldsymbol{\theta}) \geq \mathcal{L}(q;\boldsymbol{\theta})$; l'énergie variationnelle libre est une minorante de la log-vraisemblance.

a) Inférence exacte

L'algorithme EM peut être compris comme une maximisation alternée de l'énergie variationnelle libre par rapport à la densité de probabilité q et aux paramètres $\boldsymbol{\theta}$ [Bishop, 2006, chapitre 9]. Il s'agit en fait d'un algorithme de type minoration-maximisation, qui consiste à alterner deux étapes jusqu'à convergence :

- Étape E : $q^\star = \underset{q\in\mathcal{F}}{\arg\max}\, \mathcal{L}(q;\boldsymbol{\theta}')$;
- Étape M : $\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{L}(q';\boldsymbol{\theta})$,

où $\boldsymbol{\theta}'$ et q' désignent respectivement les estimées courantes des paramètres et de la distribution variationnelle.

A l'étape E on cherche donc à maximiser $\mathcal{L}(q;\boldsymbol{\theta}')$ par rapport à q à paramètres fixés. D'après l'équation (2.55) et en remarquant que $\ln p(\mathbf{x};\boldsymbol{\theta})$ ne dépend pas de $q(\mathbf{z})$, on voit que si la famille variationnelle $\mathcal{F}$ est non contrainte, alors $\mathcal{L}(q;\boldsymbol{\theta}')$ est maximale lorsque la divergence de Kullback-Leibler est nulle, c'est-à-dire que $q^\star(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')$. Dans ce cas, l'énergie variationnelle libre $\mathcal{L}(q^\star;\boldsymbol{\theta}')$ est égale à la log-vraisemblance $\ln p(\mathbf{x};\boldsymbol{\theta}')$; on a égalité entre la log-vraisemblance et sa minorante. L'estimée courante $q'(\mathbf{z})$ utilisée ensuite pour l'étape M est donc donnée par $q^\star(\mathbf{z})$.

Si maintenant on substitue $q(\mathbf{z})$ dans l'équation (2.56) par cette expression de l'estimée courante $q'(\mathbf{z})$, on obtient :

$$\begin{aligned}\mathcal{L}(q',\boldsymbol{\theta}) &= \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')} - \langle \ln p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')\rangle_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')} \\ &= Q(\boldsymbol{\theta};\boldsymbol{\theta}') + \text{constante},\end{aligned} \tag{2.61}$$

où $Q(\boldsymbol{\theta};\boldsymbol{\theta}') = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')}$ est l'espérance conditionnelle de la log-vraisemblance des données complètes et la constante par rapport aux paramètres $\boldsymbol{\theta}$ est égale à l'entropie différentielle de $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')$. Comme proposé à l'origine dans [Dempster et al., 1977], l'étape E de l'algorithme EM peut être vue comme consistant à calculer cette quantité $Q(\boldsymbol{\theta};\boldsymbol{\theta}')$.

A l'étape M de l'algorithme, on cherche à maximiser $\mathcal{L}(q',\boldsymbol{\theta})$ par rapport aux paramètres $\boldsymbol{\theta}$. Cela entraînera un accroissement de l'énergie variationnelle libre et donc nécessairement de la log-vraisemblance, car à la fin de l'étape E nous avions égalité entre les deux quantités. D'après l'équation (2.61), on voit finalement que l'étape M revient à maximiser $Q(\boldsymbol{\theta};\boldsymbol{\theta}')$ par rapport à $\boldsymbol{\theta}$.

Nous avons vu précédemment que l'estimation MAP des paramètres consiste à maximiser par rapport à $\boldsymbol{\theta}$ la quantité suivante :

$$\begin{aligned}\ln p(\boldsymbol{\theta}|\mathbf{x}) &\stackrel{c}{=} \ln p(\mathbf{x}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\ &\stackrel{c}{=} \mathcal{L}(q;\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})),\end{aligned} \tag{2.62}$$

où nous avons utilisé l'équation (2.55) pour décomposer la log-vraisemblance. Comme précédemment, on peut alors montrer que l'étape M dans le cas d'une estimation MAP des paramètres consiste à maximiser $Q(\boldsymbol{\theta};\boldsymbol{\theta}') + \ln p(\boldsymbol{\theta})$.

L'algorithme EM nécessite de connaître la distribution a posteriori des données cachées $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$, ou au moins l'espérance conditionnelle $\langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')}$. Dans certains cas de figure, nous n'avons pas cette connaissance et devons recourir à l'approche variationnelle, où la distribution variationnelle correspondra seulement à une approximation de la distribution a posteriori satisfaisant certaines contraintes.

b) Inférence variationnelle sous l'approximation de champ moyen

L'inférence variationnelle consiste à chercher une distribution variationnelle sur les variables latentes, de densité de probabilité $q \in \mathcal{F}$, qui minimise la divergence de Kullback-Leibler $D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}))$. Pour résoudre ce problème nous devons en général formuler certaines hypothèses sur la famille variationnelle $\mathcal{F}$. L'approximation de champ moyen est pour cela très répandue, elle consiste à supposer que la famille variationnelle correspond à l'ensemble des densités de probabilités qui se factorisent sous la forme suivante :

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(z_i), \tag{2.63}$$

où l'ensemble des variables latentes est supposé pouvoir se factoriser en M partitions z_i, $i = 1, ..., M$.

L'étape E de l'algorithme EM variationnel (VEM), sous l'approximation de champ moyen, consiste finalement à chercher la distribution $q(\mathbf{z})$ sous cette forme factorisée telle que l'énergie variationnelle libre soit maximale, à paramètres $\boldsymbol{\theta}$ fixés. En effet, d'après l'équation (2.55) on voit que minimiser la divergence de Kullback-Leibler est équivalent à maximiser l'énergie variationnelle libre. En utilisant l'équation (2.63) on peut réécrire l'énergie libre d'après l'équation (2.56) [Bishop, 2006, chapitre 10] :

$$
\begin{aligned}
\mathcal{L}(q;\boldsymbol{\theta}) &= \int \prod_{i=1}^{M} q_i(z_i) \ln \left(\frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{\prod_{i=1}^{M} q_i(z_i)} \right) d\mathbf{z} \\
&= \int \prod_{i=1}^{M} q_i(z_i) \Big[\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \sum_{i=1}^{M} \ln q_i(z_i) \Big] d\mathbf{z} \\
&= \int \prod_{i=1}^{M} q_i(z_i) \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} - \sum_{i=1}^{M} \int \prod_{k=1}^{M} q_k(z_k) \ln q_i(z_i) d\mathbf{z} \\
&= \int \prod_{i=1}^{M} q_i(z_i) \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} - \sum_{i=1}^{M} \int q_i(z_i) \ln q_i(z_i) dz_i \qquad \text{car } \int q_k(z_k) dz_k = 1 \\
&= \int q_j(z_j) \left[\int \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \prod_{i \neq j} q_i(z_i) dz_i \right] dz_j - \int q_j(z_j) \ln q_j(z_j) dz_j \\
&\qquad - \sum_{i \neq j} \int q_i(z_i) \ln q_i(z_i) dz_i. \qquad (2.64)
\end{aligned}
$$

On pose :

$$
\ln \tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta}) = \int \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \prod_{i \neq j} q_i(z_i) dz_i = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \rangle_{\prod_{i \neq j} q_i}. \qquad (2.65)
$$

Alors,

$$
\begin{aligned}
\mathcal{L}(q;\boldsymbol{\theta}) &= \int q_j(z_j) \ln \tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta}) dz_j - \int q_j(z_j) \ln q_j(z_j) dz_j - \sum_{i \neq j} \int q_i(z_i) \ln q_i(z_i) dz_i \\
&= \int q_j(z_j) \ln \frac{\tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta})}{q_j(z_j)} dz_j - \sum_{i \neq j} \int q_i(z_i) \ln q_i(z_i) dz_i \\
&= -D_{\mathrm{KL}}(q_j(z_j) || \tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta})) - \sum_{i \neq j} \int q_i(z_i) \ln q_i(z_i) dz_i. \qquad (2.66)
\end{aligned}
$$

On suppose maintenant que les $\{q_i\}_{i \neq j}$ sont fixes et on cherche à maximiser $\mathcal{L}(q;\boldsymbol{\theta})$ par rapport à $q_j(z_j)$. D'après l'équation (2.66) on voit que cela est équivalent à minimiser la divergence de Kullback-Leibler $D_{\mathrm{KL}}(q_j(z_j)||\tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta}))$, qui est minimale quand $q_j(z_j) = \tilde{p}(\mathbf{x}, z_j;\boldsymbol{\theta})$. La distribution optimale $q_j^\star(z_j)$ doit donc satisfaire :

$$
\ln q_j^\star(z_j) = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \rangle_{\prod_{i \neq j} q_i} + \text{constante}, \qquad (2.67)
$$

où la constante peut être déterminée par normalisation ou en reconnaissant une certaine forme de loi de probabilité.

Les équations (2.67) pour $j = 1, ..., M$ correspondent à un ensemble de conditions de consistance que doivent satisfaire les distributions $q_j^\star(z_j)$ qui maximisent l'énergie libre. Elles définissent un ensemble de solutions couplées car $q_j^\star(z_j)$ dépend des autres facteurs $q_i(z_i)$ pour $i \neq j$. Une solution consistante est alors trouvée en cyclant sur ces facteurs $q_j^\star(z_j)$ pour $j = 1, ..., M$ et en utilisant à chaque cycle les estimées précédentes.

L'algorithme VEM sous l'approximation de champ moyen peut finalement être résumé par deux étapes devant être alternées jusqu'à convergence :

- Étape E : Evaluer $q^\star(\mathbf{z})$ qui maximise $\mathcal{L}(q, \boldsymbol{\theta}')$ sous l'approximation de champ moyen (2.63) à partir du système d'équations (2.67) ;
- Étape M : Calculer $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q', \boldsymbol{\theta})$.

L'algorithme VEM garantit la monotonicité de l'énergie variationnelle libre, ce qui en pratique est très utile pour vérifier le bon comportement d'une implémentation de l'algorithme. Cependant, comme à la fin de l'étape E l'énergie variationnelle libre n'est plus forcément égale à la log-vraisemblance (du fait d'une divergence de Kullback-Leibler non nulle à l'équation (2.55)), il n'est pas garanti que l'algorithme VEM fasse croître la vraisemblance de façon monotone.

2.6 Évaluation de la qualité de séparation

Les méthodes de séparation de sources dans le cas d'un mélange multicanal et réverbérant évaluent généralement la qualité de séparation en matière de sources images reconstruites. En effet, reconstruire les signaux sources monophoniques sans connaître les filtres de mélange implique non seulement de séparer les sources images mais également de les déconvoluer de façon aveugle, ce qui est un autre problème difficile en soi. Par ailleurs, le problème de séparation de sources est fondamentalement mal posé, une infinité de solutions pour les filtres de mélange et les sources permettent d'expliquer les mêmes données observées. Le fait d'évaluer la séparation en matière de sources images permet d'ignorer ces indéterminations.

Dans cette section nous allons détailler différentes mesures objectives de la qualité de séparation. Le calcul de celles-ci nécessite les sources images estimées et originales. Dans un cadre général d'application des méthodes de séparation de sources, cette vérité terrain n'est bien évidemment pas connue, nous n'avons accès qu'au mélange des sources et non aux sources individuelles. Cependant dans un cadre de développement des méthodes de séparation, nous avons besoin d'avoir accès aux sources originales afin de calculer des indicateurs objectifs des performances de séparation. Cela est nécessaire pour le développement à proprement parler de la méthode, certains paramètres devant être ajustés en fonction des performances de séparation, et également pour comparer objectivement les résultats de la méthode proposée avec d'autres de la littérature.

Cependant ces mesures quantitatives (notamment les rapports d'énergie présentés ci-après) ne doivent pas être prises comme indicateur absolu de la qualité de séparation. Tout d'abord parce que le son n'est pas seulement un phénomène physique, un signal capté, c'est également l'origine d'une sensation perceptive. La meilleure façon de se rendre compte des performances en séparation de sources est encore d'écouter les sources estimées. C'est pourquoi nous nous sommes efforcés durant cette thèse de toujours accompagner nos résultats d'exemples audio. Ensuite, il nous est arrivé plusieurs fois d'observer de très bons scores (notamment durant les premières itérations d'algorithmes de type EM) obtenus par ces méthodes d'évaluation objectives, alors que les sources estimées étaient très éloignées des sources originales (privées de la partie haute fréquence de leur spectre par exemple). Ces mesures ne sont donc pas infaillibles. Comme nous allons le voir par la suite, certains travaux ont néanmoins cherché à développer des mesures objectives prenant en compte un aspect perceptif.

Finalement, il arrive souvent que la tâche de séparation de sources ne soit pas une finalité en soi mais corresponde plutôt à une étape de pré-traitement. C'est le cas par exemple en rehaussement de la parole où l'objectif est de séparer la parole du bruit. Cette étape de rehaussement est souvent effectuée comme pré-traitement avant une étape de reconnaissance automatique de la parole. Dans ce cas il peut être intéressant d'évaluer la qualité de séparation par l'intermédiaire du gain obtenu sur les performances de reconnaissance de la parole.

2.6.1 Évaluation en matière de rapports d'énergie

Nous présentons dans un premier temps les mesures de performance proposées dans [Vincent et al., 2006] pour évaluer la qualité des sources monophoniques estimées, puis étendues dans [Vincent et al., 2007] pour l'évaluation des sources images. Ces mesures se basent sur la décomposition suivante d'une source image estimée :

$$\hat{y}_{ij}(t) = y_{ij}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t), \tag{2.68}$$

où $y_{ij}(t)$ est la vraie source image et $e_{ij}^{\text{spat}}(t), e_{ij}^{\text{interf}}(t)$ et $e_{ij}^{\text{artif}}(t)$ correspondent à des composantes d'erreur représentant respectivement les distorsions spatiales, les interférences, c'est-à-dire des composantes issues des autres sources $j' \neq j$, et les artéfacts qui correspondent à des sons non présents dans le mélange d'origine. Le bruit musical est un exemple d'artéfact, il est la conséquence de points d'énergie résiduels localisés dans le plan TF qui du fait d'une certaine structure sont perçus comme un son présentant quasiment une hauteur. Ces trois erreurs sont définies de la façon suivante :

- Distorsion spatiale :
$$e_{ij}^{\text{spat}}(t) = \mathcal{P}_j^L\{\hat{y}_{ij}(t)\} - y_{ij}(t), \tag{2.69}$$
où $\mathcal{P}_j^L\{\cdot\}$ correspond à l'opérateur de projection orthogonale sur le sous-espace engendré par $y_{i'j}(t-\tau)$, $i' \in \{1, ..., I\}$, $\tau \in \{0, ..., L-1\}$. Ce terme d'erreur intègre donc les composantes provenant des autres canaux dans la source estimée. L'introduction du délai τ permet de ne pas pénaliser l'estimation d'une version filtrée de la source originale [Vincent et al., 2006]. L correspond à la longueur maximale autorisée pour la réponse impulsionnelle de ce filtre (fixée à 32 ms dans les références précédentes pour des signaux audio).
- Interférences :
$$e_{ij}^{\text{interf}}(t) = \mathcal{P}_{j=1:J}^L\{\hat{y}_{ij}(t)\} - \mathcal{P}_j^L\{\hat{y}_{ij}(t)\}, \tag{2.70}$$
où $\mathcal{P}_{j=1:J}^L\{\cdot\}$ correspond à l'opérateur de projection orthogonale sur le sous-espace engendré par $y_{i'j'}(t-\tau)$, $i' \in \{1, ..., I\}$, $j' \in \{1, ..., J\}$ et $\tau \in \{0, ..., L-1\}$.
- Artéfacts :
$$e_{ij}^{\text{artif}}(t) = \hat{y}_{ij}(t) - \mathcal{P}_{j=1:J}^L\{\hat{y}_{ij}(t)\}. \tag{2.71}$$

A partir de cette décomposition on calcule pour chaque source j un ensemble de mesures sous forme de rapports d'énergie exprimés en décibels (dB) et définis par :

- ISR (*source image to spatial distortion ratio*) :
$$\text{ISR}_j = 10 \log \frac{\sum_{i=1}^{I} \sum_{t=0}^{T-1} y_{ij}(t)^2}{\sum_{i=1}^{I} \sum_{t=0}^{T-1} e_{ij}^{\text{spat}}(t)^2}. \tag{2.72}$$
- SIR (*signal to interference ratio*) :
$$\text{SIR}_j = 10 \log \frac{\sum_{i=1}^{I} \sum_{t=0}^{T-1} \left[y_{ij}(t) + e_{ij}^{\text{spat}}(t) \right]^2}{\sum_{i=1}^{I} \sum_{t=0}^{T-1} e_{ij}^{\text{interf}}(t)^2}. \tag{2.73}$$

- SAR (*signal to artifact ratio*) :

$$\mathrm{SAR}_j = 10 \log \frac{\sum_{i=1}^{I} \sum_{t=0}^{T-1} \left[y_{ij}(t) + e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) \right]^2}{\sum_{i=1}^{I} \sum_{t=0}^{T-1} e_{ij}^{\text{artif}}(t)^2}. \tag{2.74}$$

Finalement, la qualité globale de séparation est mesurée par le SDR (*signal to distortion ratio*) défini par :

$$\mathrm{SDR}_j = 10 \log \frac{\sum_{i=1}^{I} \sum_{t=0}^{T-1} y_{ij}(t)^2}{\sum_{i=1}^{I} \sum_{t=0}^{T-1} \left[e_{ij}^{\text{spat}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t) \right]^2}. \tag{2.75}$$

2.6.2 Évaluation en matière de scores perceptifs

La méthode d'évaluation initialement introduite dans [Emiya et al., 2011] puis améliorée dans [Vincent, 2012] a pour objectif d'intégrer un aspect perceptif dans le calcul des mesures objectives.

Cette méthode propose tout d'abord de calculer les composantes d'erreur de la décomposition (2.68) grâce à un processus d'analyse/synthèse par banc de filtres gammatone de la source estimée $\hat{y}_{ij}(t)$, ceci afin d'approcher le traitement effectué par le système auditif.

Ces composantes sont ensuite utilisées pour calculer de nouvelles mesures objectives intégrant un aspect perceptif. L'importance des distorsions individuelles (spatiale, interférences et artéfacts) est évaluée grâce à une mesure de similarité perceptive entre le signal source estimé $\hat{y}_{ij}(t)$ et ce même signal après lui avoir soustrait l'erreur considérée : $\hat{y}_{ij}(t) - e_{ij}^{\text{spat/interf/artif}}(t)$. Une quatrième mesure de distorsion globale est calculée en évaluant la similarité perceptive entre la source estimée $\hat{y}_{ij}(t)$ et la vraie source $y_{ij}(t)$. On obtient ainsi un vecteur de descripteurs contenant quatre mesures de similarité.

Ces descripteurs sont ensuite utilisés comme vecteur d'entrée de quatre réseaux de neurones indépendants. Chaque réseau fournit en sortie un score exprimé en pourcent permettant d'évaluer les performances de séparation suivant la qualité globale, les distorsions spatiales, les interférences et les artéfacts ; respectivement l'OPS (*overall perceptual score*), le TPS (*target perceptual score*), l'IPS (*interference perceptual score*) et l'APS (*artifact perceptual score*). Les réseaux de neurones sont entraînés afin de minimiser l'erreur quadratique moyenne entre le score prédit et celui obtenu par un ensemble d'individus dans le cadre d'une évaluation subjective.

2.7 Bases de données

Nous présentons dans cette section les bases de données qui seront utilisées pour évaluer les méthodes de séparation de sources développées dans cette thèse. Nous utilisons les signaux sources de la base de données MASS du MTG [Vinyes, 2008]. Les sources sélectionnées sont celles pour lesquelles aucun effet n'a été appliqué d'après les informations fournies avec la base de données. Chaque source étant proposée dans un format stéréophonique, nous les convertissons tout d'abord au format monophonique et les sous-échantillonnons à 16 kHz. Ces sources sont ensuite convoluées avec des filtres de mélanges pour donner des sources images stéréophoniques qui sont finalement sommées pour créer un mélange stéréophonique. On obtient ainsi 8 mélanges décrits dans le tableau 2.1.

Dans cette thèse nous avons utilisé des filtres de mélange correspondant à des réponses de salle synthétiques ou mesurées. Les bases de données créées sont les suivantes :

1. «MASS-Synthétique» :
   - ▷ Réponses impulsionnelles de salle simulées par la méthode des sources images grâce à la boîte à outils «Roomsimove» [Vincent et Campbell, 2008].
   - ▷ Temps de réverbération : Ajustable.
   - ▷ Distance entre les sources et le centre de la paire de microphones : 1 m.
2. «MASS-RWCP» :
   - ▷ Réponses impulsionnelles de salle mesurées et fournies avec [Nakamura et al., 2000].
   - ▷ Temps de réverbération : 470 ms.
   - ▷ Distance entre les sources et le centre de la paire de microphones : 2 m.
3. «MASS-MIRD» :
   - ▷ Réponses impulsionnelles de salle mesurées et fournies avec [Hadad et al., 2014].
   - ▷ Temps de réverbération : 160, 360 ou 610 ms.
   - ▷ Distance entre les sources et le centre de la paire de microphones : 1 m.

Il est important de mentionner que toutes les sources sont disjointes spatialement dans les mélanges que nous utilisons.

mélange	durée	source 1	source 2	source 3	source 4	source 5
1	28 s	piano	balais	basse	-	-
2	14 s	batterie	voix	piano	basse	-
3	24 s	batterie	guitare	basse	-	-
4	28 s	batterie	voix	guitare	guitare	-
5	18 s	basse	guitare	batterie	-	-
6	15 s	batterie	voix	guitare	basse	-
7	25 s	batterie	voix	guitare	guitare	basse
8	12 s	batterie	voix	basse	-	-

TABLEAU 2.1 – Description des mélanges des bases de données.

Deuxième partie

Modélisation du mélange dans le domaine fréquentiel

Chapitre 3

Modèles de réponse en fréquence de salle

Sommaire

3.1 Introduction 52
3.2 Réponses impulsionnelle et fréquentielle de salle 53
3.3 Modèle de contributions précoces 53
3.3.1 Modèle autorégressif 54
3.3.2 Choix de l'ordre du modèle 55
3.4 Modèle de réverbération tardive 55
3.4.1 Autocovariance et densité spectrale de puissance théoriques 57
3.4.2 Paramétrisation autorégressive à moyenne ajustée 59
3.5 Conclusion 63

3.1 Introduction

Considérons un espace clos dans lequel se trouve une source ponctuelle et un point d'observation distinct identifié par la présence d'un microphone. Comme introduit au chapitre 1, lorsque la source émet un son, le signal capté par le microphone est égal à la convolution du signal source avec une réponse impulsionnelle de salle (RIR d'après l'anglais *room impulse response*). La RIR décrit la propagation entre les deux points de la salle, cette dernière pouvant être considérée comme un système linéaire invariant dans le temps (si l'on néglige les changements de température, de pression, etc.). Un tel système est en effet totalement caractérisé par sa réponse impulsionnelle, c'est-à-dire le signal en sortie lorsque l'entrée est une impulsion de dirac, qui est l'élément neutre du produit de convolution. En toute rigueur il nous faudrait également prendre en compte l'influence du microphone sur la mesure. Cependant nous négligerons cet aspect par la suite.

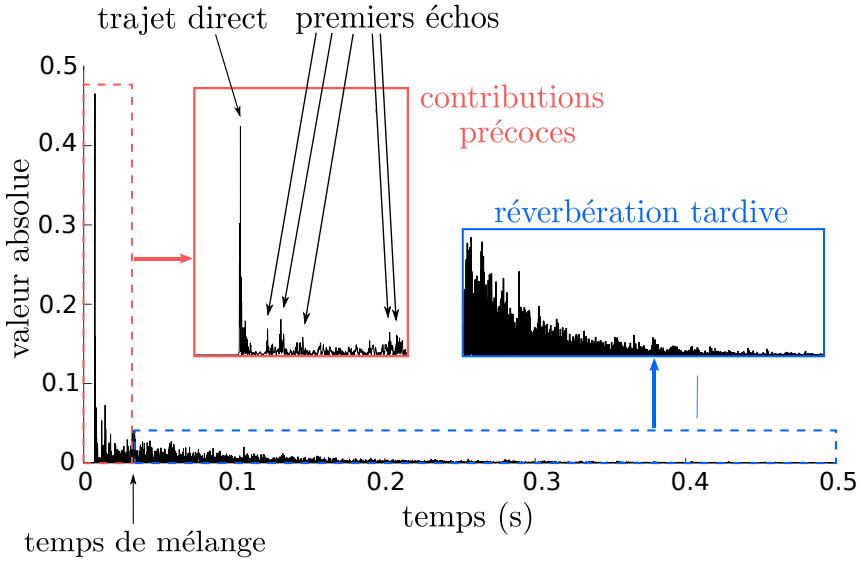


FIGURE 3.1 – Réponse impulsionnelle de salle provenant de la base de donnée RWCP [Nakamura et al., 2000]. La réponse impulsionnelle a été mesurée dans une salle avec un temps de réverbération d'environ 660 ms.

Une RIR possède une structure bien particulière. Nous pouvons en effet distinguer deux régions comme illustré sur la figure 3.1. La première contient ce que nous appellerons par la suite les *contributions précoces*. Il s'agit du trajet direct entre la source et le microphone ainsi que les premiers échos associés aux premières réflexions sur les parois et objets de la salle. La seconde région correspond à la *réverbération tardive*.

Le trajet direct est uniquement caractérisé par la distance entre la source et le microphone et la vitesse de propagation du son (que l'on considérera constante et égale à 340 m/s dans cette thèse).

Après le trajet direct arrivent les premières réflexions sur les parois de la salle. L'ensemble des premières contributions varie fortement avec la position de la source et du microphone dans la salle. On considère généralement que les premiers échos complètent l'information de localisation de la source portée par le trajet direct. Ils ont donc une grande importance dans la perception de la scène sonore spatialisée.

La réverbération tardive se met en place après un certain *temps de mélange*. Cette partie de la RIR est généralement associée au champ sonore diffus, c'est-à-dire que l'on considère avoir atteint un stade de la propagation dans la salle où l'énergie sonore est uniformément répartie dans l'espace et suivant toutes les directions [Schultz, 1971].

L'objectif de ce chapitre est d'introduire un modèle de réponse en fréquence de salle (RFR

d'après l'anglais *room frequency response*). La RFR est définie comme la transformée de Fourier de la RIR. Ces modèles seront ensuite utilisés au chapitre 4 dans le cadre d'une application de séparation de sources où le mélange est représenté dans le domaine fréquentiel. Bien que notre objectif final soit de développer un modèle fréquentiel, notre point de départ concernera toujours la structure bien particulière d'une RIR dans le domaine temporel. Nous voyons très clairement à partir de la figure 3.1 qu'il nous faudra considérer deux approches différentes pour modéliser les contributions précoces et la réverbération tardive. En effet ces deux parties sont structurellement très différentes ; tandis qu'il est envisageable de caractériser chaque première contribution individuellement, nous devrons employer des méthodes statistiques pour modéliser la réverbération tardive du fait de la forte densité d'échos qui lui est associée.

Dans la section 3.2 nous introduisons les notations et définissons le lien entre RIR et RFR. A partir de concepts d'acoustique géométrique des salles nous développerons dans la section 3.3 le modèle de contributions précoces. Celui-ci a été publié dans [Leglaive et al., 2015a,b], dans le cadre d'une application en séparation de sources. Dans la section 3.4 nous emploierons des résultats d'acoustique statistique des salles pour modéliser la réverbération tardive. Le modèle obtenu a été publié dans [Leglaive et al., 2016a]. Nous aboutirons ainsi à un modèle complet de réponse de salle, défini dans le domaine fréquentiel, et qui sera utilisé dans le chapitre 4 pour la séparation de sources.

3.2 Réponses impulsionnelle et fréquentielle de salle

Soit $a(t) = a_e(t) + a_l(t)$, $t = 0, ..., T-1$, une RIR de longueur T décomposée en la somme des contributions précoces $a_e(t)$ et de la réverbération tardive $a_l(t)$. Comme représenté sur la figure 3.1, ces deux parties ont des supports temporels disjoints : $a_e(t) \neq 0$ si $t < t_0$ et $a_l(t) \neq 0$ si $t \geq t_0$. L'instant t_0 est appelé temps de mélange, il est généralement défini dans la littérature à partir du volume V de la pièce en m^3 [Jot et al., 1997; Lindau et al., 2010] :

$$t_0 = \lfloor C_0 \sqrt{V} f_s \rfloor \text{ échantillons}, \tag{3.1}$$

où $C_0 = 2 \times 10^{-3}$ est une constante de normalisation et f_s est la fréquence d'échantillonnage en Hz. Nous pouvons mentionner qu'il est également parfois fixé arbitrairement entre 40 et 80 ms [Naylor et Gaubitch, 2010]. En effet la connaissance du volume de la salle n'est pas toujours disponible.

La RFR est définie de façon similaire par $A(f) = A_e(f) + A_l(f)$ où pour $f = 0, ..., T-1$:

$$A_{(\cdot)}(f) = \mathcal{F}_T\{a_{(\cdot)}(t)\} = \sum_{t=0}^{T-1} a_{(\cdot)}(t) e^{-\imath 2\pi f t/T}. \tag{3.2}$$

L'équation (3.2) définit la transformée de Fourier discrète (TFD) sur T points du signal $a_{(\cdot)}(t)$. On définit également la TFD inverse par :

$$a_{(\cdot)}(t) = \mathcal{F}_T^{-1}\{A_{(\cdot)}(f)\} = \frac{1}{T} \sum_{f=0}^{T-1} A_{(\cdot)}(f) e^{\imath 2\pi f t/T}. \tag{3.3}$$

3.3 Modèle de contributions précoces

En acoustique géométrique des salles, on représente la réflexion d'une onde acoustique sur une surface plane de la même façon qu'une réflexion optique sur un miroir [Allen et Berkley, 1979].

Cette approche débouche sur le concept de *source image* [1]. On considère que le son réfléchi sur une paroi se propage comme s'il avait été émis par une source virtuelle, symétrique de la source réelle par rapport au plan de la paroi. Il s'agit ici des sources images du premier ordre. De façon plus générale, chaque source image peut ensuite engendrer d'autres réflexions, caractérisées par de nouvelles sources images dites d'ordre supérieur. C'est donc un procédé itératif qui revient à déplier infiniment la salle par symétrie. On peut ensuite trouver l'ensemble des sources images d'ordre $n \geq 1$ comme les symétriques des sources images d'ordre $n-1$ où la source d'ordre 0 correspond à la source réelle.

Dans ce modèle de sources images, chaque réflexion est dite spéculaire. C'est-à-dire que chaque «rayon» incident donne naissance à un unique «rayon» réfléchi. Ce modèle ne permet donc pas de représenter des réflexions diffuses, où du fait des irrégularités des parois de la salle le son serait réfléchi dans plusieurs directions. Ce modèle simple caractérise donc chaque contribution précoce (trajet direct ou réflexion) par une impulsion à laquelle sont associés une amplitude (ou plutôt une atténuation) et un retard. Pour le trajet direct, le temps d'arrivée est égal à la distance source-microphone divisée par la vitesse du son tandis que l'atténuation est inversement proportionnelle à la distance source-microphone. Pour chaque réflexion, le temps d'arrivée dépend de la distance entre la source image qui lui est associée et le microphone. L'atténuation quant à elle tient compte non seulement de la distance parcourue mais également du coefficient d'absorption de la paroi.

3.3.1 Modèle autorégressif

Formellement, nous considérons que la partie précoce de la RIR est constituée de R contributions, chacune étant représentée par un dirac auquel est associé un terme d'atténuation ρ_k et un retard τ_k, $k = 0, ..., R-1$. La RFR associée $A_e(f)$ est ainsi approchée par $G(f)$, $f = 0, ..., T-1$, avec :

$$G(f) = \sum_{k=0}^{R-1} \rho_k \delta_k^f, \qquad \delta_k = e^{-\imath 2\pi \tau_k / T}. \tag{3.4}$$

Dans l'équation ci-dessus nous avons utilisé le fait qu'un retard dans le domaine temporel correspond à un déphasage (une multiplication par une exponentielle complexe) dans le domaine fréquentiel. D'après l'équation (3.4) il est possible de montrer que $\{G(f)\}_{f=R,...,T-1}$ satisfait une équation récurrente de la forme suivante :

$$\sum_{r=0}^{R} \varphi_r^e G(f-r) = 0, \tag{3.5}$$

tel que $\{\varphi_r^e\}_{r=0}^R$ et $\{\delta_k\}_{k=0}^{R-1}$ sont respectivement les coefficients et racines d'un même polynôme de degré R. Ce résultat nous vient de la littérature liée à l'estimation des paramètres des modèles sinusoïdaux à modulation d'amplitude exponentielle (*exponential sinusoidal models* (ESM) en anglais) [Kumaresan, 1983]. En effet l'équation (3.4) est un cas particulier d'un tel modèle ESM. Par ailleurs, il est important de remarquer que l'on représente ici un signal dans le domaine fréquentiel alors qu'en général un modèle ESM est défini dans le domaine temporel, pour caractériser l'amortissement exponentiel des systèmes vibratoires libres par exemple.

Sans perte de généralité nous considérons que le premier coefficient $\varphi_0^e = 1$; l'équation (3.5) étant toujours valide si l'on multiplie les deux membres par une constante. Nous supposons finalement que $A_e(f)$ satisfait l'équation récurrente (3.5) à un terme de déviation près $\kappa(f)$ modélisé

1. Dans ce chapitre, le terme de source image désigne une source virtuelle et non la convolution du signal source avec un filtre de mélange comme c'était le cas aux chapitres précédents.

par un bruit blanc gaussien complexe propre :

$$\sum_{r=0}^{R} \varphi_r^e A_e(f-r) = \kappa(f), \qquad \kappa(f) \sim \mathcal{N}_c(0, \sigma_\kappa^2). \tag{3.6}$$

Nous voyons alors à partir de cette dernière équation que la RFR $A_e(f)$ associée aux contributions précoces est représentée par un modèle autorégressif (AR) d'ordre R, noté AR(R).

Nous avons proposé ce modèle dans [Leglaive et al., 2015a,b] pour guider l'estimation des filtres de mélange dans une application de séparation de sources. Nous supposions dans ces travaux préliminaires que $A(f) = A_e(f)$, c'est-à-dire que nous négligions la réverbération tardive.

3.3.2 Choix de l'ordre du modèle

L'ordre de ce modèle AR doit théoriquement être égal au nombre de contributions précoces. Par exemple, la RIR représentée en haut de la figure 3.2 semble contenir environ 5 premières réflexions significatives qui ajoutées au trajet direct donnent $R \approx 6$ contributions précoces. Cependant si l'on considère que le trajet direct domine les premiers échos, il peut être suffisant de choisir $R = 1$. Dans ce cas, d'après les équations (3.4)-(3.6) nous avons :

$$A_e(f) = \delta_0 A_e(f-1) + \kappa(f), \tag{3.7}$$

où $\delta_0 = e^{-\imath 2\pi\tau_0/T}$ et $\tau_0 = \lfloor (r_0/c) f_s \rfloor$ avec r_0 la distance source-microphone en mètres et c la vitesse du son en mètres par secondes. Ce modèle AR(1) peut également être interprété comme exprimant le fait que la réponse en fréquence associée aux contributions précoces varie lentement en terme de module et phase (pour σ_κ^2 suffisamment faible), car la réponse impulsionnelle associée est concentrée aux instants proches de l'origine. Nous pouvons en effet écrire les relations suivantes pour le module et la phase de $A_e(f)$ d'après l'équation (3.7) :

$$|\, A_e(f) \,| \approx |\, A_e(f-1) \,| \tag{3.8}$$

$$\arg(A_e(f)) \approx \arg(A_e(f-1)) - 2\pi\tau_0/T. \tag{3.9}$$

Connaissant la distance source-microphone, nous pouvons vérifier ces relations à partir d'une RIR réelle provenant de la base de donnée RWCP [Nakamura et al., 2000]. Celle-ci est représentée en haut de la figure 3.2 où la partie précoce utilisée pour calculer $A_e(f)$ par TFD est en bleu. Les relations de module et phase sont représentées en bas de cette même figure. Nous observons que les points bleus associés aux coefficients de la RFR sont concentrés sur les droites rouges d'équation $y = x$, validant ainsi les relations (3.8)-(3.9). Par conséquent, considérer un modèle AR d'ordre 1 pour $A_e(f)$ semble être un choix raisonnable lorsque les premiers échos sont effectivement d'amplitude négligeable par rapport au trajet direct.

3.4 Modèle de réverbération tardive

Passons maintenant au développement d'un modèle de réverbération tardive. Nous rappelons qu'il s'agit de la partie de la RIR où un grand nombre de réflexions se superposent dans le domaine temporel comme illustré sur la figure 3.1. Contrairement à la partie précoce où la structure relativement parcimonieuse nous permettait de caractériser individuellement chaque réflexion, il nous faut maintenant employer des méthodes statistiques pour décrire la réverbération tardive. Le modèle présenté dans cette section a été publié dans [Leglaive et al., 2016a].

La théorie de l'acoustique statistique des salles a tout d'abord été initiée dans le domaine fréquentiel par Schroeder [Schroeder et Kuttruff, 1962; Schroeder, 1962, 1987]. C'est ensuite

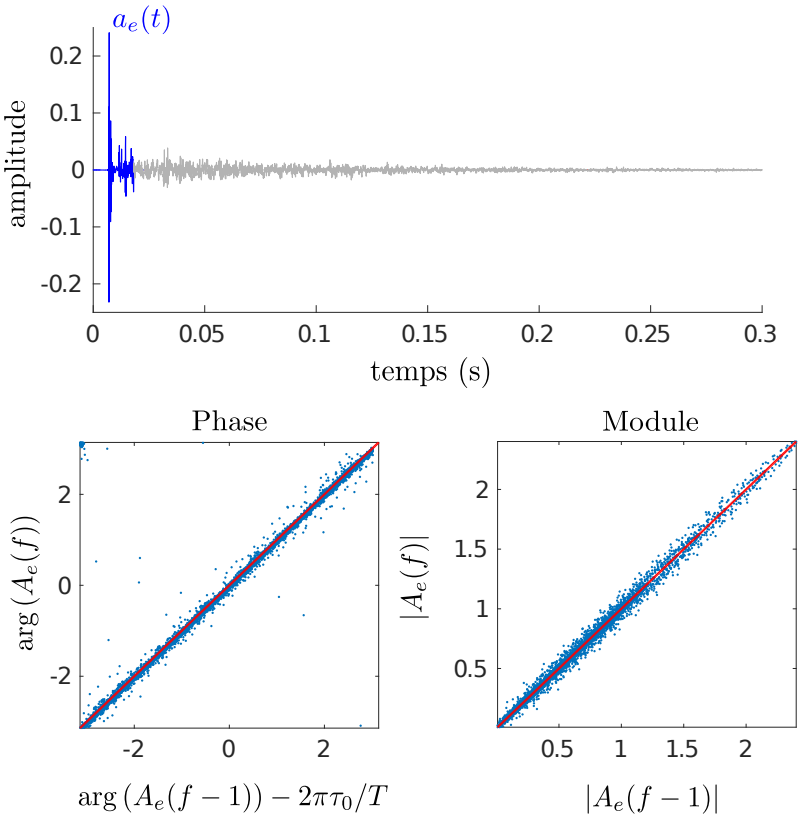


FIGURE 3.2 – En haut : RIR provenant de la base RWCP. La distance source-microphone est d'environ 2 m et le temps de réverbération de 0.75 s. En bas : Relation de module et phase illustrant les équations (3.8)-(3.9). Les points bleus représentent les données tandis que les droites rouges correspondent aux droites d'équation $y = x$. $A_e(f)$ est calculé à partir de la partie précoce $a_e(t)$ représentée sur la figure du haut en bleu.

Polack qui l'a complétée dans le domaine temporel [Polack, 1988, 1993]. En acoustique, une salle est caractérisée par ses modes propres, ce sont les solutions de l'équation de Helmoltz vérifiant certaines conditions aux limites [Kuttruff, 2009, ch. 3]. La réponse de la salle à une excitation est alors constituée de la superposition des réponses de chacun des modes (il y en a une infinité). Chaque mode propre peut être vu comme un résonateur caractérisé par un gain et une fréquence de résonance. La RFR (qui est également appelée fonction de Green en acoustique) correspond alors à une somme infinie des fonctions de transfert de ces résonateurs. Il est possible de montrer que la densité de modes, c'est-à-dire le nombre moyen de modes (identifiés par leur fréquence propre) par Hertz, croît avec le carré de la fréquence. Par conséquent, au dessus d'une certaine fréquence limite, une source sonore même sinusoïdale excitera un très grand nombre de modes. Ce résultat constitue la base de l'acoustique statistique des salles.

La fréquence à partir de laquelle un nombre «suffisant» de modes se superposent pour permettre un traitement statistique de la RFR est appelée fréquence de Schroeder et est définie par [Schroeder et Kuttruff, 1962] :

$$f_{sch} = C_1 \sqrt{\frac{T_{60}}{V}} \text{ Hz}, \tag{3.10}$$

où C_1 est une constante de normalisation qui vaut environ 2000 et T_{60} est le temps de réverbération en secondes, défini comme le temps au bout duquel l'énergie sonore a diminué de 60 dB après extinction de la source. Cette fréquence correspond à la limite au dessus de laquelle l'espacement moyen entre les fréquences de résonance est inférieur à un tiers de la largeur de bande

moyenne des modes. En supposant des modes dont les amplitudes et phases sont individuellement indépendantes et identiquement distribuées, le théorème central limite indique que la RFR peut être considérée comme un processus aléatoire dont les parties réelle et imaginaire sont des processus gaussiens indépendants de même variance. Nous considérons de plus que ce processus est centré et stationnaire au sens large (SSL). Nous pouvons alors définir sa fonction d'autocovariance (ACVF d'après l'abréviation anglaise de *autocovariance function*) $\gamma(m)$ et sa densité spectrale de puissance (DSP) $\phi(t)$ par :

$$\gamma(m) = \mathbb{E}[A_l(f)A_l(f-m)^*]; \tag{3.11}$$

$$\phi(t) = \frac{1}{T}\mathbb{E}[|\mathcal{F}_T\{A_l(f)\}|^2]. \tag{3.12}$$

En tant que processus gaussien complexe centré, propre et SSL, la RFR est totalement caractérisée par son ACVF ou sa DSP. Nous devons mentionner que comme nous travaillons à temps et fréquence discrets, tous les signaux sont T-périodiques. La RFR est donc précisément un processus SSL T-périodique. De plus, $\phi(t)$ doit être vue comme une fonction de DSP discrétisée. On pourra se référer à [Gu et al., 2007] pour une revue de certaines propriétés des processus aléatoires périodiques, appliqués à la modélisation fréquentielle des canaux de transmission sans fil.

Le terme de densité «*spectrale*» de puissance est ici trompeur. En effet, l'étude des séries temporelles, comme son nom l'indique, considère généralement des processus aléatoires indexés par le temps. L'ACVF d'un tel processus est alors une fonction d'un décalage temporel et la DSP une fonction de la fréquence. Dans notre cas, nous travaillons sur un processus aléatoire $A_l(f)$ indexé par les fréquences. L'indice m dans la définition de l'ACVF à l'équation (3.11) correspond ainsi à un décalage fréquentiel et la DSP définie à l'équation (3.12) est une fonction du temps.

Schroeder a fourni dans [Schroeder, 1962] une expression théorique de l'ACVF en considérant la décroissance exponentielle de l'énergie de la RIR au cours du temps. Cette décroissance exponentielle n'est cependant valide que pour un champ sonore diffus, donc uniquement pour la partie tardive d'une réponse de salle comme on peut l'observer sur la figure 3.1. Schroeder a par conséquent négligé dans ce travail l'influence du trajet direct et des premiers échos. Ceci explique le décalage entre l'ACVF théorique et les résultats expérimentaux observés dans [Gustafsson et al., 2003]. C'est pourquoi nous allons ici raffiner les résultats de Schroeder en prenant en compte le fait que la décroissance exponentielle n'est valide que pour la partie tardive de la RIR.

3.4.1 Autocovariance et densité spectrale de puissance théoriques

Sachant que pour un champ sonore diffus l'énergie de la RIR décroît exponentiellement au cours du temps, on définit le Profil Temporel d'Énergie (PTE) pour $t = 0, ..., T-1$ par :

$$\bar{a}_l(t) = \mathbb{E}[a_l^2(t)] = P_0^2 e^{-2t/\tau}\mathbb{1}_{\{t_0,...,T-1\}}(t), \tag{3.13}$$

où $\mathbb{1}_{\mathcal{T}}(t)$ est la fonction indicatrice qui vaut 1 si $t \in \mathcal{T}$, 0 sinon, P_0^2 est une constante liée à l'énergie totale de la réverbération tardive et τ est reliée au temps de réverbération T_{60} par :

$$\tau = \frac{T_{60} f_s}{3\ln(10)} \text{ échantillons.} \tag{3.14}$$

La fonction indicatrice à l'équation (3.13) permet de prendre en compte le fait que la décroissance exponentielle n'est valide que pour la réverbération tardive, là où le champ sonore est considéré comme diffus. Il est par ailleurs important de mentionner que plusieurs réalisations d'une RIR peuvent être interprétées comme plusieurs observations pour différentes positions de la source et du microphone dans la salle. L'espérance mathématique doit donc être comprise comme une moyenne spatiale. Rappelons également que sous une hypothèse de champ sonore diffus, généralement acceptée pour la partie tardive d'une réponse de salle, l'expression (3.13) ne dépend pas

de la position de la source et du microphone car l'énergie sonore est considérée comme étant uniformément répartie dans la salle et suivant toutes les directions. Certaines conditions doivent néanmoins être vérifiées pour que la théorie de l'acoustique statistique des salles dans le domaine fréquentiel et pour un champ sonore diffus s'applique [Schroeder, 1962] : (1) les dimensions de la salle sont grandes par rapport à la longueur d'onde du signal source ; (2) l'espacement moyen entre les fréquences de résonance des modes est inférieur à un tiers de leur largeur de bande moyenne (ce qui est vérifié au dessus de la fréquence de Schroeder par définition) ; (3) la source et le microphone sont situés au moins à une demi-longueur d'onde des parois de la salle.

Nous montrons en annexe C que la DSP du processus $\{A_l(f)\}_f$ s'exprime en fonction du PTE défini à l'équation (3.13) de la façon suivante :

$$\phi(t) = T\bar{a}_l(T-t). \tag{3.15}$$

En appliquant le théorème de Wiener-Khinchin à partir de cette expression de la DSP on obtient l'expression théorique de l'ACVF :

$$\gamma(m) = \mathcal{F}_T^{-1}\{\phi(t)\} = P_0^2 e^{-2T/\tau}\frac{1-e^{(\imath 2\pi m/T+2/\tau)(T-t_0+1)}}{1-e^{\imath 2\pi m/T+2/\tau}}. \tag{3.16}$$

P_0^2 est relié à la variance $\sigma_{rev}^2 = \gamma(0)$ de la RFR par :

$$P_0^2 = \sigma_{rev}^2 e^{2T/\tau}\frac{1-e^{2/\tau}}{1-e^{2(T-t_0+1)/\tau}}. \tag{3.17}$$

On peut de plus montrer que (voir annexe B) :

$$\sigma_{rev}^2 = \frac{1-\alpha}{\pi\alpha\mathcal{S}}, \tag{3.18}$$

où α est le coefficient d'absorption moyen (sans dimension) et $\mathcal{S}$ l'aire total des parois de la salle en m^2. Le coefficient d'absorption moyen peut être calculé à partir de la formule de Norris-Eyring [Naylor et Gaubitch, 2010, p. 24] :

$$\alpha = 1 - e^{-24\ln(10)V/(c\mathcal{S}T_{60})}, \tag{3.19}$$

avec c la vitesse du son en m.s^{-1}.

Validation expérimentale Nous souhaitons maintenant valider expérimentallement l'expression théorique de l'ACVF à l'équation (3.16) à partir de RIRs simulées et réelles. Pour un ensemble de RIRs donné, correspondant à plusieurs positions dans une salle, nous commençons par retirer les contributions précoces en forçant à zéro les échantillons pour $t < t_0$. Pour chaque réalisation $\tilde{a}_l^r(t)$ (une RIR pour une position particulière de la source et du microphone), on calcule la RFR $\tilde{A}_l^r(f)$ et on ne garde que les K échantillons dans l'intervalle $[f_{sch},\ f_s/2]$ Hz. L'ACVF empirique pour cette réalisation est alors obtenue à partir de l'estimateur suivant, pour $m \in \{-K+1, ..., K-1\}$:

$$\tilde{\gamma}_r(m) = \begin{cases} \dfrac{1}{K}\displaystyle\sum_{k=k_0}^{k_0+K-1-m} \tilde{A}_l^r(k)^*\tilde{A}_l^r(k+m), & m \geq 0 \\ \tilde{\gamma}_r(-m)^*, & m < 0 \end{cases}, \tag{3.20}$$

où k_0 est l'indice associé à la fréquence de Schroeder. D'après cette simulation de Monte Carlo, l'estimation finale de l'ACVF est obtenue en moyennant sur l'ensemble complet de réalisations :

$$\hat{\gamma}(m) = \frac{1}{N_r}\sum_{r=1}^{N_r}\tilde{\gamma}_r(m). \tag{3.21}$$

Durant les expériences nous avons tout d'abord observé une différence entre la variance empirique calculée sur les données et l'expression théorique à l'équation (3.18). Bien que très répandue dans la littérature, cette seule expression de la variance n'a jamais été validée expérimentalement à notre connaissance. Comme mentionné dans [Bies et Hansen, 2009, ch. 7], l'expression de la pression quadratique moyenne du champ réverbérant qui permet d'obtenir l'équation (3.18) (voir annexe B) peut être imprécise selon les dimensions de la salle et la distance source/microphone. De plus, cette formule est obtenue en considérant que le champ réverbérant est établi après la première réflexion du trajet direct sur une paroi de la salle, ce qui ne correspond pas à la définition de la réverbération tardive que nous avons donnée précédemment, en fonction du temps de mélange t_0. Cela peut donc expliquer la différence entre la théorie et les expériences pour la valeur de la variance σ^2_{rev}. Cependant, une correction empirique qui conduit à de bons résultats consiste à multiplier σ^2_{rev} telle que définie à l'équation (3.18) par une constante $C_{rev} = 75$. Cette correction sera donc toujours appliquée par la suite. Il pourrait être intéressant de chercher à comprendre plus précisément d'où vient ce décalage entre la théorie et les expériences. Cependant cela dépasse le cadre de cette thèse et ne sera pas traité ici.

Nous avons simulé 196 RIRs à partir de la méthode des sources images en utilisant la boîte à outils Roomsimove [Vincent et Campbell, 2008]. La salle est rectangulaire et de dimensions $10 \times 6.6 \times 3$ m avec un temps de réverbération $T_{60} = 250$ ms. La position de la source reste fixe et seule celle du microphone varie. On observe sur la figure 3.3 que l'ACVF empirique calculée à partir des RIRs simulées correspond bien à l'expression théorique (3.16).

Nous effectuons la même expérience à partir de 130 réponses de salle réelles provenant de la base de données C4DM (*Center for Digital Music*) [Stewart et Sandler, 2010]. Les RIRs sont mesurées dans une salle de dimensions $7.5 \times 9 \times 3.5$ m avec un temps de réverbération $T_{60} = 1.8$ s. La source est également figée alors que la position du microphone varie dans la salle. Pour cet ensemble de données on observe à nouveau une bonne correspondance entre théorie et pratique sur la figure 3.4.

3.4.2 Paramétrisation autorégressive à moyenne ajustée

Nous avons vu dans la sous-section précédente que les propriétés statistiques de la réverbération tardive sont totalement résumées par l'ACVF définie à l'équation (3.16). Cette fonction dépend uniquement du temps de réverbération, du volume et de l'aire totale des parois de la salle. Si dans le cadre des processus gaussiens la connaissance de la fonction d'autocovariance est totalement suffisante pour définir la distribution d'un vecteur d'observations, il peut être utile en pratique de résumer cette information par l'intermédiaire d'un modèle paramétrique impliquant un faible de nombre de paramètres. Cela peut par exemple nous éviter d'inverser une matrice de covariance de grande dimension.

Dans cette partie nous allons montrer qu'un modèle autorégressif à moyenne ajustée (ARMA) en fréquence permet de paramétrer de façon précise la DSP et l'ACVF théoriques définies aux équations (3.15) et (3.16) respectivement. On représente la partie tardive de la RFR $\{A_l(f)\}_f$ par le modèle ARMA(P, Q) suivant :

$$\Phi(L)A_l(f) = \Theta(L)\epsilon(f), \qquad \epsilon(f) \sim \mathcal{N}_c(0, \sigma^2_\epsilon), \tag{3.22}$$

où $\Phi(L) = \sum_{p=0}^{P} \varphi_p^l L^p$, $\Theta(L) = \sum_{q=0}^{Q} \vartheta_q L^q$ avec $\varphi_0 = \vartheta_0 = 1$ et L l'opérateur de décalage, i.e. $LA_l(f) = A_l(f-1)$. $\epsilon(f)$ est un bruit blanc gaussien complexe propre de variance σ^2_ϵ pour $f \in [0, ..., T-1]$ et est étendu par T-périodicité en dehors de cet intervalle. $A_l(f)$ correspond donc à la sortie d'un filtre de fonction de transfert $\Psi(z^{-1}) = \Theta(z^{-1})/\Phi(z^{-1})$ dont l'entrée est $\epsilon(f)$. On suppose que les zéros des polynômes $\Theta(z^{-1})$ et $\Phi(z^{-1})$ se trouvent à l'intérieur du cercle unité tel que $\Psi(z^{-1})$ soit la fonction de transfert d'un filtre causal stable et inversible. A partir de

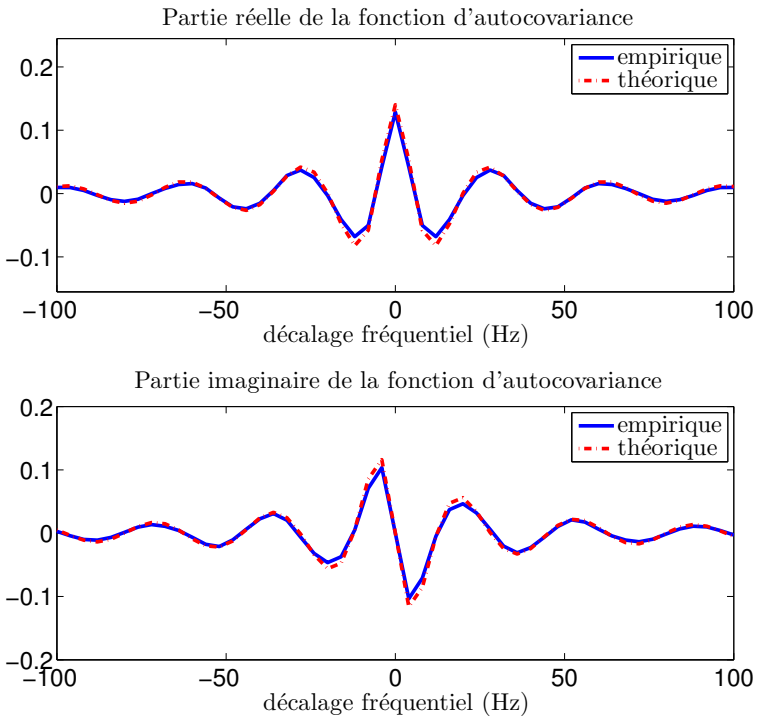


FIGURE 3.3 – Fonctions d'autocovariance empirique et théorique calculées à partir de RIRs simulées.

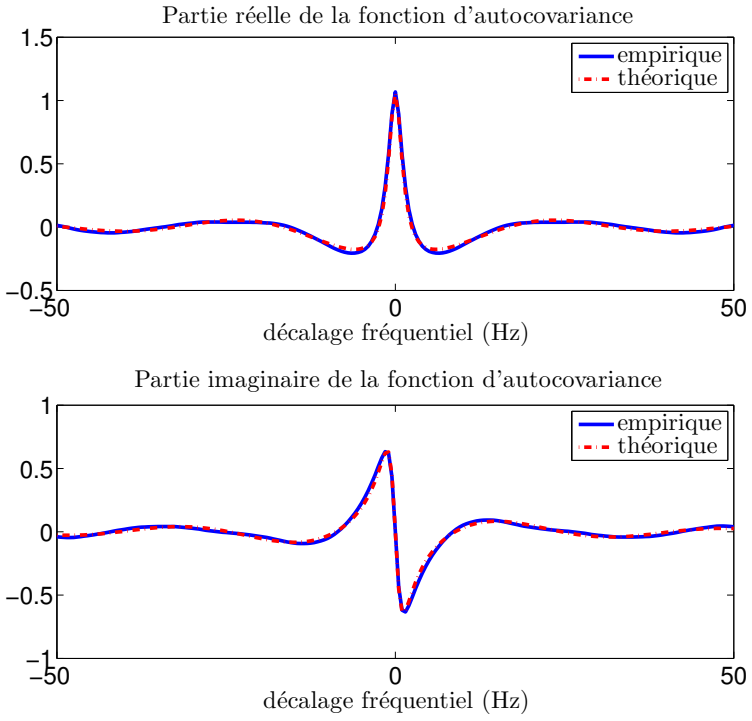


FIGURE 3.4 – Fonctions d'autocovariance empirique et théorique calculées à partir de RIRs mesurées.

cette représentation ARMA nous pouvons écrire les paramétrisations suivantes de la DSP et de l'ACVF :

$$\phi(t) = \sigma_\epsilon^2 \frac{\left|\mathcal{F}_T\left\{\{\vartheta_q\}_{q=0,\dots,Q}\right\}\right|^2}{\left|\mathcal{F}_T\left\{\{\varphi_p^l\}_{p=0,\dots,P}\right\}\right|^2}, \tag{3.23}$$

$$\sum_{p=0}^{P} \varphi_p^l \gamma(m-p) = \begin{cases} \sigma_\epsilon^2 \sum\limits_{q=\underline{m}}^{Q} \vartheta_q \psi_{q-\underline{m}}^* & \text{if } 0 \leq \underline{m} \leq Q \\ 0 & \text{if } \underline{m} > Q, \end{cases} \tag{3.24}$$

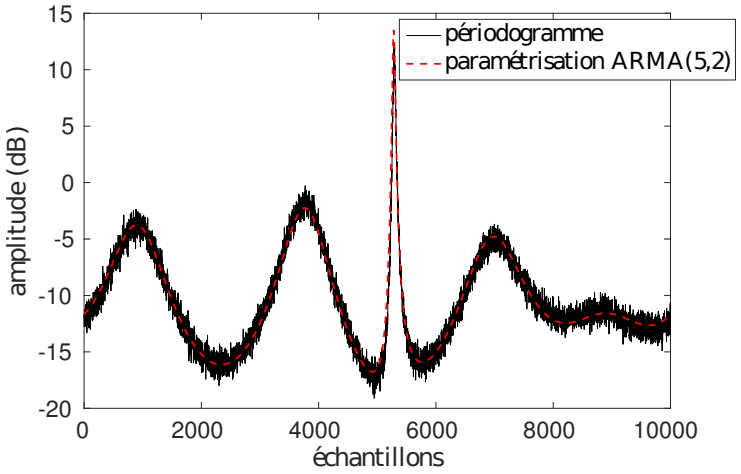
où $\Psi(z^{-1}) = \sum_{k=0}^{+\infty} \psi_k z^{-k}$ et $\underline{m} = m \,(\mathrm{mod}\, T)$ tel que $\underline{m} \in [0, ..., T-1]$. Pour que l'équation (3.24) soit valide nous devons supposer que le support de l'ACVF sur une période est limité afin qu'il n'y ait pas de recouvrement, i.e. $\psi_k = 0$ for $k > \lfloor T/2 \rfloor$.

a) Estimation des paramètres ARMA

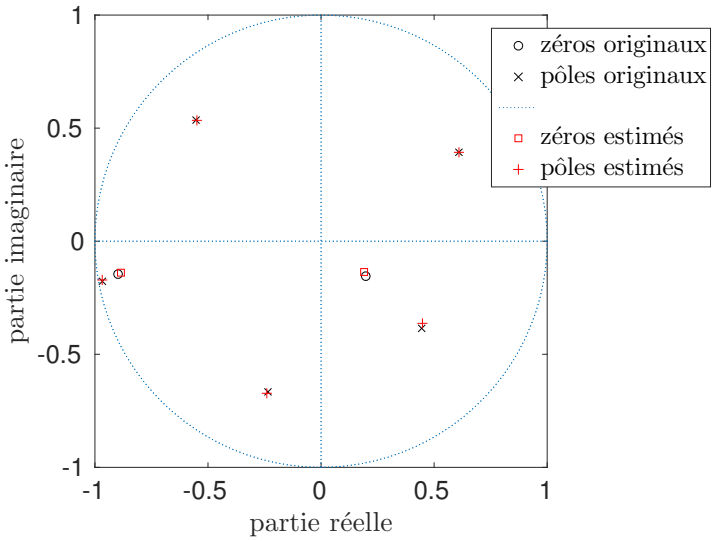
Les paramètres du modèle ARMA peuvent ensuite être estimés à partir de la seule connaissance de l'ACVF théorique donnée par l'équation (3.16). On voit ici l'avantage d'avoir une expression théorique, nous n'avons besoin d'aucune donnée pour estimer les paramètres du modèle ARMA. La procédure d'estimation est décrite dans [Kay, 1988, ch. 2], nous la rappelons brièvement ici : Les paramètres AR $\{\varphi_p^l\}_{p=1}^P$ sont tout d'abord estimés en résolvant les équations de Yule et Walker modifiées, définies à partir de l'équation de récurrence (3.24) pour $m = Q+1, ..., Q+P$. A partir des paramètres AR estimés on définit l'ACVF $\gamma'(m) = \Phi^*(L^{-1})\Phi(L)\gamma(m)$ où $\Phi^*(L^{-1}) = \sum_{p=0}^{P} (\varphi_p^l)^* L^{-p}$. Il s'agit de l'ACVF d'un processus MA dont les paramètres correspondent à ceux du processus ARMA de départ défini à l'équation (3.22), c'est-à-dire les coefficients MA $\{\theta_q\}_{q=1}^Q$ et la variance du bruit σ_ϵ^2. Nous employons la méthode de Durbin pour estimer ces paramètres MA [Durbin, 1959]. Cette méthode se base sur le fait que tout processus MA admet une représentation sous forme de processus AR d'ordre infini. En pratique nous approchons le processus MA(Q) par un processus AR d'ordre $L = 10Q$ dont nous notons les paramètres $\{\varphi'_p\}_{p=1}^L$. A partir de $\gamma'(m)$ nous calculons les coefficients AR de ce processus en résolvant les équations de Yule et Walker. Nous obtenons ainsi une estimation de la variance σ_ϵ^2 et des coefficients $\{\varphi'_p\}_{p=1}^L$. Finalement, nous estimons un modèle AR(Q) à partir de la séquence de coefficients $\{1, \varphi'_1, ..., \varphi'_L\}$ (vus comme des observations issues de ce processus). Les paramètres ainsi obtenus par résolution des équations de Yule et Walker correspondent à une estimation des coefficients $\{\theta_q\}_{q=1}^Q$. Nous vérifions sur la figure 3.5 la validité de cette procédure à partir d'un exemple synthétique.

b) Validation expérimentale

Nous allons maintenant vérifier la validité de cette paramétrisation ARMA de l'ACVF et de la DSP. Nous considérons les mêmes paramètres de salle et le même temps de réverbération (250 ms) que pour les RIRs simulées dans la section précédente. On calcule à partir de ces paramètres l'ACVF théorique donnée à l'équation (3.16) et nous estimons à partir de celle-ci les paramètres d'un modèle ARMA$(7, 2)$. L'ordre de la partie AR est choisi arbitrairement tandis que l'ordre de la partie MA est fixé en fonction du support de l'ACVF $\gamma'(m)$ définie au paragraphe précédent. En effet il s'agit de l'ACVF d'un processus MA, on sait donc que $\gamma'(m) = 0$ pour $m > Q$. On trace sur la figure 3.6 la DSP et l'ACVF calculées à partir de cette paramétrisation. On observe que le modèle ARMA permet de très bien représenter les statistiques de la partie tardive de la RFR. La DSP est tracée en décibels afin de mettre en évidence le fait que la décroissance exponentielle est bien respectée.



(a) Densité spectrale de puissance empirique (périodogramme) et paramétrisation ARMA.



(b) Carte des pôles et zéros de la fonction de transfert utilisée pour générer le processus ARMA(5,2) et leur estimation.

FIGURE 3.5 – Illustration de l'estimation des paramètres d'un modèle ARMA sur un exemple synthétique. Les observations ont été générées à partir d'un modèle ARMA(5,2).

Finalement, afin de vérifier la validité du modèle complet (ACVF, DSP théoriques et paramétrisation ARMA), nous synthétisons la partie tardive d'une réponse de salle à partir du modèle ARMA. Il s'agit pour cela d'effectuer un filtrage ARMA dans le domaine fréquentiel d'un bruit blanc gaussien complexe propre. Les paramètres du filtre correspondent à ceux estimés à partir de l'ACVF théorique. La RIR synthétique est ensuite reconstruite par TFD inverse en utilisant la symétrie hermitienne de la RFR. Celle-ci est représentée sur la figure 3.7. On observe clairement la décroissance exponentielle de la réverbération tardive, commençant à partir du temps de mélange tel que décrit par le PTE à l'équation (3.13). Le temps de réverbération estimé grace à la méthode de Schroeder [Schroeder, 1965] est de 280 ms, ce qui est proche de la valeur cible de 250 ms.

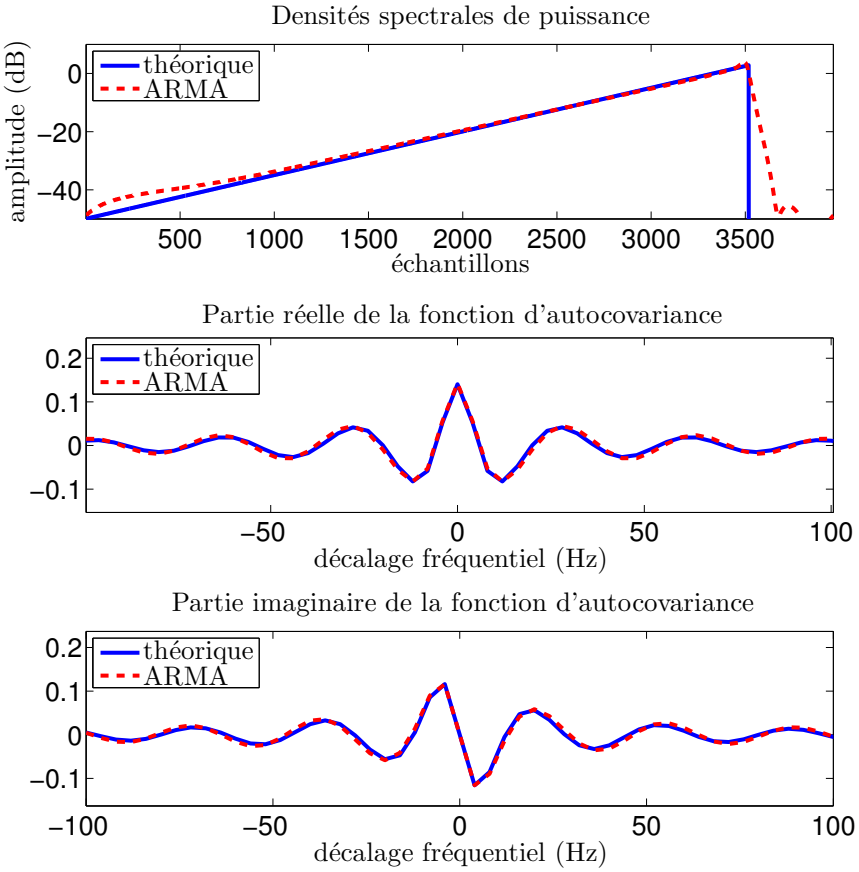


FIGURE 3.6 – Paramétrisation ARMA(7,2) de la DSP et de l'ACVF.

3.5 Conclusion

Dans ce chapitre nous avons proposé deux nouveaux modèles distincts pour les parties précoce et tardive des réponses en fréquence de salle, l'idée principale étant de chercher à transcrire la dynamique temporelle des réponses de salle sous forme de corrélations fréquentielles.

A partir de résultats d'acoustique géométrique des salles, les premières contributions ont été représentées dans le domaine temporel par des impulsions auxquelles correspondent une atténuation et un retard. Cette approche nous a amené à modéliser la réponse en fréquence associée aux premières contributions comme un processus AR. En supposant que le trajet direct domine les premiers échos, l'ordre de ce processus AR peut être fixé à 1.

D'après des résultats d'acoustique statistique des salles nous représentons la réverbération tardive en fréquence par un processus gaussien complexe, centré, propre et stationnaire au sens large. Nous avons montré qu'il est possible d'exploiter la décroissance exponentielle de la réverbération tardive dans le domaine temporel afin d'obtenir des expressions théoriques des fonctions de densité spectrale de puissance et d'autocovariance de ce processus. Nous avons finalement proposé de paramétrer ces deux quantités par un modèle ARMA.

Ces modèles fréquentiels de réponse de salle vont nous permettre dans le chapitre suivant de guider l'estimation des filtres de mélange dans un contexte de séparation de sources.

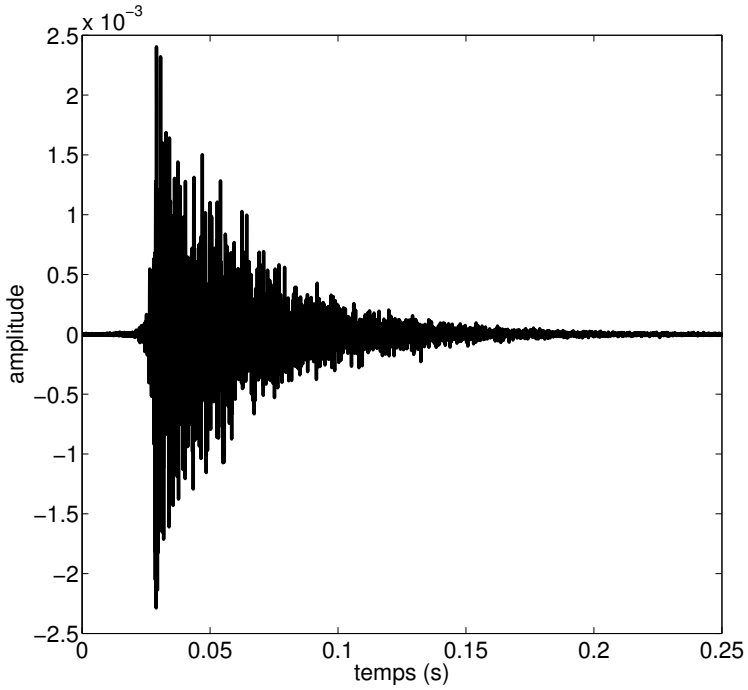


FIGURE 3.7 – Réverbération tardive synthétisée à partir du filtrage ARMA(7,2) dans le domaine fréquentiel d'un bruit blanc gaussien complexe propre.

Chapitre 4

Séparation de sources avec a priori sur la réponse en fréquence des filtres de mélange

Sommaire

4.1 Modèles et estimation des filtres au sens du maximum de vraisemblance . . 66
4.1.1 Modèle 66
4.1.2 Inférence par l'algorithme EM 67
4.1.3 Reconstruction des sources 69
4.2 A priori sur les filtres de mélange 69
4.2.1 A priori pour les contributions précoces 70
4.2.2 A priori pour la réverbération tardive 70
4.3 Estimation des filtres au sens du maximum a posteriori 71
4.3.1 Étape M pour la partie précoce des filtres de mélange 71
4.3.2 Étape M pour la partie tardive des filtres de mélange 72
4.4 Résultats expérimentaux 74
4.4.1 Expérience préliminaire concernant l'hypothèse de filtres de mélange courts 75
4.4.2 Initialisation des paramètres du modèle 76
4.4.3 Résultats de séparation : Comparaison des approches MV et MAP . . . 77
4.5 Conclusion 79

La méthode de séparation de sources développée dans ce chapitre se base sur l'approche proposée dans [Ozerov et Févotte, 2010; Ozerov et al., 2011]. Dans cette référence les sources sont représentées comme des variables aléatoires latentes dans le domaine de la TFCT. Les filtres de mélange sont quant à eux considérés comme des paramètres déterministes dans le domaine fréquentiel, estimés au sens du maximum de vraisemblance, c'est-à-dire à partir des données observées uniquement. L'inférence des sources et l'estimation des paramètres sont effectuées grâce à un algorithme EM. Nous commençons par présenter cette méthode à la section 4.1.

Nous proposons ensuite à la section 4.2 d'utiliser les modèles de réponse en fréquence de salle développés au chapitre précédent afin de définir un a priori probabiliste sur la réponse en fréquence des filtres de mélange. L'approche suivie dans ce chapitre pour définir le modèle est illustrée par la figure 4.1.

Le modèle de filtres de mélange est pris en compte à la section 4.3 afin de modifier l'étape M de l'algorithme EM précédent, dans le but d'estimer les filtres au sens du maximum a posteriori.

Enfin, dans la section 4.4 nous montrons expérimentalement que la prise en compte de l'a priori permet d'améliorer les résultats de séparation de sources. Cette méthode a fait l'objet d'une publication dans un article de revue [Leglaive et al., 2016b].

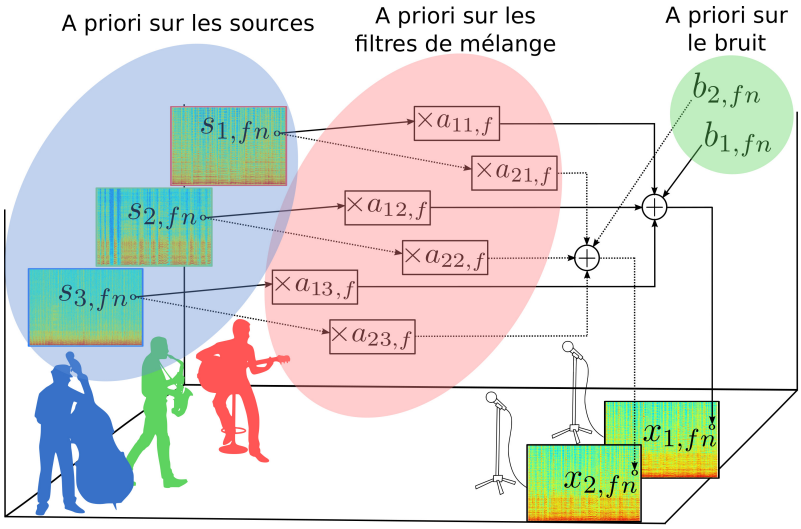


FIGURE 4.1 – Illustration de l'approche suivie dans ce chapitre.

4.1 Modèles et estimation des filtres au sens du maximum de vraisemblance

Dans cette section nous présentons le modèle et l'algorithme d'inférence développés dans [Ozerov et Févotte, 2010; Ozerov et al., 2011].

4.1.1 Modèle

a) Modèle de mélange

On considère un mélange bruité de J sources sur I canaux exprimé dans le domaine de la TFCT tel que pour tout $(f, n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$:

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{4.1}$$

où $\mathbf{x}_{fn} = [x_{i,fn}]_i^\top \in \mathbb{C}^I$, $\mathbf{s}_{fn} = [s_{j,fn}]_j^\top \in \mathbb{C}^J$, $\mathbf{A}_f = [a_{ij,f}]_{i,j} \in \mathbb{C}^{I\times J}$ est la matrice de mélange formée à partir des réponses en fréquence de salle et $\mathbf{b}_{fn} = [b_{i,fn}]_i^\top \in \mathbb{C}^I$ est un bruit blanc gaussien complexe stationnaire en temps et isotrope spatialement :

$$\mathbf{b}_{fn} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{b},f}), \qquad \mathbf{\Sigma}_{\mathbf{b},f} = \sigma_{b,f}^2 \mathbf{I}_I, \tag{4.2}$$

où $\sigma_{b,f}^2 > 0$ et $\mathbf{I}_I$ est la matrice identité de taille I. $\mathcal{N}_c(\boldsymbol{\mu}, \mathbf{\Sigma})$ représente la loi gaussienne complexe multivariée propre de densité de probabilité :

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\det(\pi\mathbf{\Sigma})} \exp[-(\mathbf{x}-\boldsymbol{\mu})^H \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})], \tag{4.3}$$

avec $\det(\mathbf{M})$ le déterminant de la matrice $\mathbf{M}$. Le terme «propre» signifie que la matrice de pseudo-covariance $\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]$ est nulle [Adali et al., 2011]. De plus, si le vecteur moyenne $\boldsymbol{\mu}$ est également nul alors la distribution est à symétrie circulaire.

Il est important de mentionner que ce terme $\mathbf{b}_{fn}$ ne correspond en général pas à un bruit réel dans le mélange. Il permet dans le cadre d'un modèle probabiliste d'indiquer que conditionnellement aux variables de sources, le mélange est gaussien.

b) Modèle de source

Les sources sont supposées mutuellement et individuellement indépendantes pour chaque point TF tel que :

$$\mathbf{s}_{fn} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{s},fn}), \qquad \mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left(\{v_{j,fn}\}_j\right), \tag{4.4}$$

où $\mathrm{diag}(\{c_m\}_m)$ est la matrice diagonale construite à partir des coefficients c_m pour $m = 1, ..., M$.

Les variances dans le modèle de source (4.4) sont de plus structurées par l'intermédiaire d'un modèle NMF dont nous rappelons l'expression :

$$v_{j,fn} = (\mathbf{W}_j \mathbf{H}_j)_{f,n}, \tag{4.5}$$

où $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$ et $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$ avec K_j le rang de la factorisation pour la source j.

4.1.2 Inférence par l'algorithme EM

Dans la méthode originellement proposée dans [Ozerov et Févotte, 2010], chaque source $s_{j,fn}$ est décomposée comme la somme de K_j composantes gaussiennes $c_{k,fn}$, représentées par une NMF de rang 1. Ces composantes sont alors considérées comme les variables latentes pour développer un algorithme EM. Il a ensuite été proposé, notamment par les mêmes auteurs, de considérer directement les sources $s_{j,fn}$ comme variables latentes [Ozerov et al., 2011]. Cette seconde approche permet d'accélérer la convergence de l'algorithme EM.

Nous notons donc $\mathbf{s} = \{\mathbf{s}_{fn}\}_{f,n}$ l'ensemble des variables latentes et $\mathbf{x} = \{\mathbf{x}_{fn}\}_{f,n}$ l'ensemble des variables observées. $\boldsymbol{\theta} = \left\{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{\mathbf{A}_f, \sigma_{b,f}^2\}_f\right\}$ correspond à l'ensemble des paramètres du modèle à estimer.

Vraisemblance D'après le modèle présenté à la section 4.1.1, la distribution du mélange est donnée par :

$$\mathbf{x}_{fn}; \boldsymbol{\theta} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{x},fn}), \qquad \mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f}. \tag{4.6}$$

Log-vraisemblance des données complètes A l'étape E de l'algorithme EM nous devons calculer l'espérance conditionnelle de la log-vraisemblance des données complètes $Q_{\text{MV}}(\boldsymbol{\theta};\boldsymbol{\theta}') = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}'}[\ln p(\mathbf{x},\mathbf{s}|\boldsymbol{\theta})]$ où $\boldsymbol{\theta}'$ représente la valeur courante des paramètres. D'après les équations (4.1), (4.2) et (4.4) la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned}\ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) &= -FN(I+J)\ln(\pi) \\ &\quad - \sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\Big[\ln\det(\boldsymbol{\Sigma}_{\mathbf{b},f}) + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})\Big] \\ &\quad - \sum_{j=1}^{J}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\Big[\ln[(\mathbf{W}_j\mathbf{H}_j)_{f,n}] + \frac{|s_{j,fn}|^2}{(\mathbf{W}_j\mathbf{H}_j)_{f,n}}\Big]. \qquad (4.7)\end{aligned}$$

Distribution a posteriori des sources On sait que $\mathbf{s}_{fn}$ et $\mathbf{b}_{fn}$ sont deux vecteurs aléatoires gaussiens indépendants, donc $[\mathbf{s}_{fn}^\top,\mathbf{b}_{fn}^\top]^\top$ est également un vecteur gaussien. Par linéarité de la loi normale, $\begin{bmatrix}\mathbf{s}_{fn}\\ \mathbf{x}_{fn}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_J & \mathbf{0}\\ \mathbf{A}_f & \mathbf{I}_I\end{bmatrix}\begin{bmatrix}\mathbf{s}_{fn}\\ \mathbf{b}_{fn}\end{bmatrix}$ est également un vecteur aléatoire gaussien. On peut alors montrer d'après les propriétés de la loi gaussienne multivariée que la distribution a posteriori des variables cachées s'écrit :

$$p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}) = \prod_{f,n} p(\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}) \qquad \text{où} \qquad \mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta} \sim \mathcal{N}_c(\hat{\mathbf{s}}_{fn},\boldsymbol{\Sigma}_{\mathbf{s},fn}^{post}), \qquad (4.8)$$

avec

- $\hat{\mathbf{s}}_{fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}$;
- $\boldsymbol{\Sigma}_{\mathbf{s},fn}^{post} = \boldsymbol{\Sigma}_{\mathbf{s},fn} - \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}$.

Nous remarquons que pour calculer l'estimateur des sources donné par la moyenne a posteriori $\hat{\mathbf{s}}_{fn}$, il est nécessaire d'inverser la matrice de covariance du mélange $\boldsymbol{\Sigma}_{\mathbf{x},fn}$ définie à l'équation (4.6). La matrice de covariance du bruit dans cette équation permet de prévenir l'inversion d'une matrice mal conditionnée. En effet si dans le cas sous-déterminé la matrice $\boldsymbol{\Sigma}_{\mathbf{s},fn}$ a plus de $(J-I)$ termes diagonaux nuls, alors le rang de $\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H$ est inférieur à I. Ce cas de figure peut arriver dans certaines régions TF où plusieurs sources sont inactives.

Étape E A partir des équations (4.7) et (4.8), on peut calculer l'espérance conditionnelle de la log-vraisemblance des données complètes. La fonction à maximiser à l'étape M s'écrit :

$$\begin{aligned}Q_{\text{MV}}(\boldsymbol{\theta};\boldsymbol{\theta}') &\overset{c}{=} -N\sum_{f=0}^{F-1}\Big[\ln\det(\boldsymbol{\Sigma}_{\mathbf{b},f}) \\ &\quad + \operatorname{trace}\Big(\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{xs},f}^H - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\Big)\Big] \\ &\quad - \sum_{j=1}^{J}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\Big[\ln[(\mathbf{W}_j\mathbf{H}_j)_{f,n}] + \frac{\hat{p}_{j,fn}}{(\mathbf{W}_j\mathbf{H}_j)_{f,n}}\Big], \qquad (4.9)\end{aligned}$$

où

- $\hat{\mathbf{R}}_{\cdot\cdot,f} = \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1}\hat{\mathbf{R}}_{\cdot\cdot,fn}$;
- $\hat{\mathbf{R}}_{\mathbf{xx},fn} = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}'}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H] = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$;

▷ $\hat{\mathbf{R}}_{\mathbf{xs},fn} = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}'}[\mathbf{x}_{fn}\mathbf{s}_{fn}^H] = \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H$;

▷ $\hat{\mathbf{R}}_{\mathbf{ss},fn} = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}'}[\mathbf{s}_{fn}\mathbf{s}_{fn}^H] = \boldsymbol{\Sigma}_{\mathbf{s},fn}^{post} + \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H$;

▷ $\hat{p}_{j,fn} = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}'}[|s_{j,fn}|^2] = \left(\hat{\mathbf{R}}_{\mathbf{ss},fn}\right)_{j,j}$.

Étape M - matrice de mélange En annulant le gradient de $Q_{\text{MV}}(\boldsymbol{\theta};\boldsymbol{\theta}')$ par rapport à la matrice de mélange $\mathbf{A}_f$ on obtient la règle de mise à jour suivante :

$$\mathbf{A}_f = \hat{\mathbf{R}}_{\mathbf{xs},f}\hat{\mathbf{R}}_{\mathbf{ss},f}^{-1}. \tag{4.10}$$

Étape M - covariance du bruit De la même façon on obtient la règle de mise à jour de la matrice de covariance du bruit $\boldsymbol{\Sigma}_{\mathbf{b},f}$:

$$\boldsymbol{\Sigma}_{\mathbf{b},f} = \text{trace}(\hat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{xs},f}^H - \hat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H)\frac{\mathbf{I}_I}{I}. \tag{4.11}$$

Étape M - paramètres NMF On reconnaît dans l'expression de $Q_{MV}(\boldsymbol{\theta};\boldsymbol{\theta}')$ la divergence d'Itakura-Saito définie à l'équation (2.21), page 31. Les matrices $\mathbf{W}_j$ et $\mathbf{H}_j$ peuvent alors être mises à jour en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{W}_j,\mathbf{H}_j\geq 0} \sum_{f=0}^{F-1}\sum_{n=0}^{N-1} d_{IS}\big(\hat{p}_{j,fn}, (\mathbf{W}_j\mathbf{H}_j)_{f,n}\big). \tag{4.12}$$

Ce problème peut être résolu grâce aux règles multiplicatives présentées dans [Févotte et al., 2009] et rappelées page 32 aux équations (2.23) et (2.24) (dans le cas $\beta = 0$).

4.1.3 Reconstruction des sources

Les sources et les filtres sont estimés à un facteur d'échelle près dépendant de la fréquence. C'est pourquoi nous fournirons en sortie de cet algorithme les sources images estimées :

$$\hat{\mathbf{y}}_{j,fn} = \hat{\mathbf{a}}_{j,fn}\hat{s}_{j,fn}, \tag{4.13}$$

où $\hat{s}_{j,fn} = (\hat{\mathbf{s}}_{fn})_j$ est la moyenne a posteriori de la source j au point TF (f,n) et $\hat{\mathbf{a}}_{j,f}$ correspond à la j-ème colonne de la matrice de mélange estimée. Les signaux temporels peuvent ensuite être reconstruits par TFCT inverse.

Comme indiqué dans [Ozerov et Févotte, 2010], à l'inverse des approches basées sur l'ICA par sous-bandes de fréquences (voir section 2.1.2a, page 23), la méthode que nous venons de présenter ne souffre pas du problème de permutation des canaux fréquentiels. Cette propriété provient du couplage entre bandes de fréquences induit par le modèle de source NMF et également de la procédure d'estimation conjointe des paramètres de source et de la matrice de mélange.

4.2 A priori sur les filtres de mélange

Nous souhaitons désormais utiliser les modèles développés au chapitre 3 pour définir des a priori sur les filtres de mélange. On décompose pour cela la matrice de mélange $\mathbf{A}_f$ sous la forme suivante :

$$\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}, \tag{4.14}$$

où $\mathbf{A}_{e,f} = [a_{ij,f}^e]_{i,j} \in \mathbb{C}^{I\times J}$ et $\mathbf{A}_{l,f} = [a_{ij,f}^l]_{i,j} \in \mathbb{C}^{I\times J}$ correspondent respectivement aux parties précoce et tardive des filtres de mélange.

4.2.1 A priori pour les contributions précoces

On considère que chaque filtre $\{a^e_{ij,f}\}_f$, $i = 1, ..., I$, $j = 1, ..., J$, suit le modèle AR(1) défini à l'équation (3.7), page 55 :

$$a^e_{ij,f} = \delta_{ij} a^e_{ij,f-1} + \kappa(f), \qquad \kappa(f) \sim \mathcal{N}_c(0, \sigma^2_\kappa), \tag{4.15}$$

où $\delta_{ij} = e^{-\imath 2\pi\tau_{ij}/L_a}$ avec L_a la longueur de la réponse impulsionnelle des filtres de mélange et $\tau_{ij} = \lfloor (r_{ij}c)/f_s \rfloor$ avec r_{ij} la distance entre la source j et le microphone i exprimée en mètres. On considère de plus que la longueur des filtres L_a est égale à la longueur de la fenêtre d'analyse de la TFCT. On suppose donc que le temps de réverbération est inférieur ou égal à la longueur de la fenêtre d'analyse.

On peut alors écrire la distribution jointe des coefficients $\{a^e_{ij,f}\}_f$:

$$p(\{a^e_{ij,f}\}_f) = p(a^e_{ij,0}) \prod_{f=1}^{F-1} p(a^e_{ij,f} | a^e_{ij,f-1}) = p(a^e_{ij,0}) \prod_{f=1}^{F-1} N(a^e_{ij,f};\, \delta_{ij} a^e_{ij,f-1}, \sigma^2_\kappa). \tag{4.16}$$

On suppose de plus que pour tout couple d'indices i et j, $a^e_{ij,0}$ suit un a priori «plat» [1]. Le logarithme de la densité de probabilité a priori sur l'ensemble des coefficients $\{a^e_{ij,f}\}_{i,j,f}$ peut alors s'écrire sous la forme matricielle suivante :

$$\ln p(\{\mathbf{A}_{e,f}\}_f) \stackrel{c}{=} -IJ(F-1)\ln(\sigma^2_\kappa) - \frac{1}{\sigma^2_\kappa} \sum_{f=1}^{F-1} \Big|\Big| \mathbf{A}_{e,f} - \mathbf{\Delta} \odot \mathbf{A}_{e,f-1} \Big|\Big|^2_F, \tag{4.17}$$

où $\mathbf{\Delta} = [\delta_{ij}]_{i,j} \in \mathbb{C}^{I \times J}$ et $||\cdot||_F$ est la norme de Frobenius.

Cet a priori possède comme hyperparamètres les coefficients AR $\{\delta_{ij}\}_{i,j}$ et la variance σ^2_κ. Les coefficients AR peuvent être estimés au sein de l'algorithme EM ou fixés si l'on connaît la distance r_{ij} entre chaque source j et microphone i. Dans un cadre bayésien, la variance σ^2_κ est fixée et exprime la confiance que nous avons dans l'a priori, c'est-à-dire sur le fait que $a^e_{ij,f}$ est proche de $\delta_{ij} a_{ij,f-1}$.

4.2.2 A priori pour la réverbération tardive

On considère que chaque filtre $\{a^l_{ij,f}\}_f$, $i = 1, ..., I$, $j = 1, ..., J$, suit le modèle ARMA causal et inversible défini à l'équation (3.22), page 59 :

$$\Phi(L) a^l_{ij,f} = \Theta(L)\epsilon(f), \qquad \epsilon(f) \sim \mathcal{N}(0, \sigma^2_\epsilon). \tag{4.18}$$

On remarque que les polynômes $\Phi(L)$ et $\Theta(L)$ ne dépendent pas des indices de source et de microphone. Cela provient en effet du caractère diffus du champ sonore associé à la réverbération tardive. Pour cette partie de la réponse de salle, l'énergie sonore est uniformément répartie dans la salle et suivant toutes les directions. Le modèle (4.18) peut de façon équivalente s'écrire en fonction du vecteur $\mathbf{a}^l_{j,f} = [a^l_{ij,f}]^\top_i \in \mathbb{C}^I$:

$$\Phi(L) \mathbf{a}^l_{j,f} = \Theta(L) \boldsymbol{\epsilon}_f, \qquad \boldsymbol{\epsilon}_f \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\boldsymbol{\epsilon},f} = \sigma^2_\epsilon \mathbf{I}_I). \tag{4.19}$$

1. Soit θ un paramètre réel ou complexe à estimer à partir d'observations $\mathbf{x}$. On appelle a priori «plat» une distribution de probabilité de densité (potentiellement impropre) $p(\theta)$ constante par rapport à θ. Ainsi l'estimateur MAP de θ correspond à l'estimateur MV, i.e. $\theta^\star = \arg\max_\theta\, p(\theta|\mathbf{x}) = \arg\max_\theta\, p(\mathbf{x}|\theta)p(\theta) = \arg\max_\theta\, p(\mathbf{x}|\theta)$. Cet a priori caractérise précisément l'absence d'information a priori sur le paramètre θ.

La matrice de covariance du bruit $\mathbf{\Sigma}_{\boldsymbol{\epsilon},f}$ étant proportionnelle à l'identité, nous ne considérons aucune dépendance spatiale entre la partie tardive des filtres de mélange associés à la source j. Nous avons alors la relation suivante caractérisant la distribution jointe des coefficients $\{a^l_{ij,f}\}_{i,f}$ exprimée en fonction des vecteurs $\{\mathbf{a}^l_{j,f}\}_f$:

$$p(\{\mathbf{a}^l_{j,f}\}_f) \propto \prod_{f=0}^{F-1} N_c\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{a}^l_{j,f}; 0, \mathbf{\Sigma}_{\boldsymbol{\epsilon},f}\right). \tag{4.20}$$

Il s'agit uniquement d'une relation de proportionnalité car nous avons omis un terme lié au jacobien de la transformation inverse du filtrage ARMA qui ne dépend pas des coefficients $\{a^l_{ij,f}\}_{i,f}$ du fait de la linéarité de la transformation.

Finalement, en supposant de plus l'indépendance des filtres suivant l'indice j, le logarithme de la densité de probabilité a priori sur l'ensemble des coefficients $\{a^l_{ij,f}\}_{i,j,f}$ peut s'écrire sous la forme suivante faisant intervenir la matrice de mélange $\mathbf{A}_{l,f}$:

$$\ln p(\{\mathbf{A}_{l,f}\}_f) \overset{c}{=} \sum_{f=0}^{F-1} -J \ln\det(\pi\mathbf{\Sigma}_{\boldsymbol{\epsilon},f}) - \operatorname{trace}\left[\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)^H \mathbf{\Sigma}^{-1}_{\boldsymbol{\epsilon},f}\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)\right]. \tag{4.21}$$

Les hyperparamètres de cet a priori sont les coefficients ARMA $\{\varphi_p\}_p$, $\{\vartheta_q\}_q$ et la variance du bruit σ^2_ϵ. Comme au chapitre 3, nous utiliserons un modèle ARMA(7,2) dont les coefficients sont appris à partir de l'ACVF théorique définie à l'équation (3.16), page 58. Cette fonction dépend du temps de réverbération, du volume de la pièce et de la surface totale des parois. Ces paramètres qui seront supposés connus permettent de définir le facteur de décroissance exponentielle τ (équation (3.14), page 57), le temps de mélange t_0 (équation (3.1), page 53), la variance de la réverbération tardive σ^2_{rev} (équation (3.18), page 58) et le coefficient d'absorption moyen de la salle α (équation (3.19), page 58). Comme précédemment, la variance du bruit du modèle ARMA σ^2_ϵ est supposée fixée et exprime la confiance que nous avons dans l'a priori.

4.3 Estimation des filtres au sens du maximum a posteriori

Nous reprenons l'algorithme EM présenté à la section 4.1 dans le cas de l'estimation MV des paramètres et notamment des filtres de mélange. Afin de prendre en compte les a priori que nous venons de développer, seule l'estimation des filtres de mélange à l'étape M de cet algorithme est modifiée. On cherche à minimiser la quantité suivante :

$$\mathcal{L}(\{\mathbf{A}_{e,f}\}_f, \{\mathbf{A}_{l,f}\}_f) = -Q_{\text{MV}}(\boldsymbol{\theta};\boldsymbol{\theta}') - \ln p(\{\mathbf{A}_{e,f}\}_f) - \ln p(\{\mathbf{A}_{l,f}\}_f), \tag{4.22}$$

où $Q_{\text{MV}}(\boldsymbol{\theta};\boldsymbol{\theta}')$ est donné à l'équation (4.9) avec $\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}$, $\ln p(\{\mathbf{A}^e_f\}_f)$ est défini à l'équation (4.17) et $\ln p(\{\mathbf{A}^l_f\}_f)$ à l'équation (4.21).

4.3.1 Étape M pour la partie précoce des filtres de mélange

En annulant le gradient de $\mathcal{L}(\{\mathbf{A}_{e,f}\}_f, \{\mathbf{A}_{l,f}\}_f)$ par rapport à $\mathbf{A}_{e,f}$ on obtient comme mise à jour pour $f = 0$:

$$\begin{aligned}\operatorname{vec}(\mathbf{A}_{e,f}) =& \Big[N\hat{\mathbf{R}}^\top_{\mathbf{ss},f} \otimes \mathbf{I}_I + \frac{1}{\sigma^2_\kappa}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f})\Big]^{-1} \\ &\times \operatorname{vec}\Big[N\hat{\mathbf{R}}_{\mathbf{xs},f} - N\mathbf{A}_{l,f}\hat{\mathbf{R}}_{\mathbf{ss},f} + \frac{1}{\sigma^2_\kappa}\mathbf{\Sigma}_{\mathbf{b},f}\big(\mathbf{\Delta}^* \odot \mathbf{A}_{e,f+1}\big)\Big];\end{aligned} \tag{4.23}$$

pour $1 \leq f \leq F-2$:

$$\begin{aligned}\mathrm{vec}(\mathbf{A}_{e,f}) = &\Big[N\hat{\mathbf{R}}_{\mathbf{ss},f}^{\top} \otimes \mathbf{I}_I + \frac{2}{\sigma_\kappa^2}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f})\Big]^{-1} \\ &\times \mathrm{vec}\left[N\hat{\mathbf{R}}_{\mathbf{xs},f} - N\mathbf{A}_{l,f}\hat{\mathbf{R}}_{\mathbf{ss},f} + \frac{1}{\sigma_\kappa^2}\mathbf{\Sigma}_{\mathbf{b},f}\big(\mathbf{\Delta} \odot \mathbf{A}_{e,f-1} + \mathbf{\Delta}^* \odot \mathbf{A}_{e,f+1}\big)\right];\end{aligned} \tag{4.24}$$

pour $f = F-1$:

$$\begin{aligned}\mathrm{vec}(\mathbf{A}_{e,f}) = &\Big[N\hat{\mathbf{R}}_{\mathbf{ss},f}^{\top} \otimes \mathbf{I}_I + \frac{1}{\sigma_\kappa^2}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f})\Big]^{-1} \\ &\times \mathrm{vec}\left[N\hat{\mathbf{R}}_{\mathbf{xs},f} - N\mathbf{A}_{l,f}\hat{\mathbf{R}}_{\mathbf{ss},f} + \frac{1}{\sigma_\kappa^2}\mathbf{\Sigma}_{\mathbf{b},f}\big(\mathbf{\Delta} \odot \mathbf{A}_{e,f-1}\big)\right];\end{aligned} \tag{4.25}$$

où $\mathrm{vec}(\cdot)$ concatène les colonnes d'une matrice en un seul vecteur colonne et $\otimes$ est le produit de Kronecker.

La mise à jour de la matrice $\mathbf{A}_{e,f}$ pour une fréquence donnée dépend des valeurs de cette matrice aux fréquences adjacentes. Il s'agit donc d'une méthode de relaxation (*coordinate descent* en anglais) et en théorie les mises à jour (4.23)-(4.25) doivent être répétées plusieurs fois au sein de chaque étape M. En pratique nous n'effectuons que deux itérations.

Les coefficients AR $\{\delta_{ij}\}_{i,j}$ qui correspondent aux paramètres de l'a priori (4.17) peuvent être fixés si la distance relative entre chaque source et chaque microphone est connue, ou bien ils peuvent être estimés au sein de l'étape M de l'algorithme EM, à partir de l'estimation courante de la partie précoce des filtres de mélange. Nous choisissons la seconde option car il peut être difficile en pratique de connaître les distances sources/microphones.

Nous cherchons donc à estimer les coefficients AR qui minimisent le critère (4.22) sous la contrainte $|\delta_{ij}| = 1$ imposée par le modèle (4.15). Ceci est équivalent à maximiser le logarithme de la densité de probabilité a priori $\ln p(\{\mathbf{A}_{e,f}\}_f)$ défini à l'équation (4.17) sous la même contrainte. Ce terme peut être compris ici comme une log-vraisemblance, car on suppose les filtres de mélange observés, à partir de leur estimation courante. On obtient finalement la mise à jour des coefficients AR par la méthode des multiplicateurs de Lagrange :

$$\delta_{ij} = \frac{\sum_{f=1}^{F-1} a_{ij,f}^e (a_{ij,f-1}^e)^*}{\Big|\sum_{f=1}^{F-1} a_{ij,f}^e (a_{ij,f-1}^e)^*\Big|}, \tag{4.26}$$

4.3.2 Étape M pour la partie tardive des filtres de mélange

Pour l'estimation de la partie tardive des filtres de mélange, nous ne pouvons malheureusement pas adopter la même approche que pour la partie précoce car il n'est pas possible d'annuler facilement le gradient $\mathcal{L}(\{\mathbf{A}_{e,f}\}_f, \{\mathbf{A}_{l,f}\}_f)$ par rapport à $\mathbf{A}_{l,f}$ à cause de la partie MA du modèle ARMA. Nous allons alors nous orienter vers une méthode de descente de gradient pour minimiser ce critère par rapport à $\{\mathbf{A}_{l,f}\}_f$. Nous avons étudié différentes approches dont la méthode de descente de gradient à pas optimal, l'approche de Barzilai et Borwein [Barzilai et Borwein, 1988], et la méthode du gradient conjugué avec ou sans préconditionnement. Nous avons obtenu les meilleurs résultats en terme de vitesse de convergence avec la dernière approche (avec précondionnement). Nous détaillons cette méthode ci-dessous.

Notre objectif est de minimiser le critère $\mathcal{L}(\{\mathbf{A}_{e,f}\}_f, \{\mathbf{A}_{l,f}\}_f)$ par rapport au vecteur de paramètres de taille IJF suivant :

$$\mathbf{a}_l = \left[\text{vec}\left(\mathbf{A}_{l,0}\right)^\top, \text{vec}\left(\mathbf{A}_{l,1}\right)^\top, ..., \text{vec}\left(\mathbf{A}_{l,F-1}\right)^\top \right]^\top. \tag{4.27}$$

Nous décomposons le gradient $\mathbf{g}$ (vecteur colonne de taille IJF) comme la concaténation de F gradients :

$$\mathbf{g} = [\mathbf{g}_0^\top, \mathbf{g}_1^\top, ..., \mathbf{g}_{F-1}^\top]^\top, \tag{4.28}$$

avec $\mathbf{g}_f = \text{vec}\left(\frac{1}{2}\nabla_{\mathbf{A}_{l,f}} \mathcal{L}(\{\mathbf{A}_{e,f}\}_f, \{\mathbf{A}_{l,f}\}_f)\right)$. Nous pouvons montrer à partir de l'équation (4.22) que :

$$\begin{aligned}\mathbf{g}_f =& N \,\text{vec}\left(\mathbf{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{A}_{e,f}\hat{\mathbf{R}}_{\mathbf{ss},f} - \hat{\mathbf{R}}_{\mathbf{xs},f})\right) + N\left(\hat{\mathbf{R}}_{\mathbf{ss},f}^\top \otimes \mathbf{\Sigma}_{\mathbf{b},f}^{-1}\right) \text{vec}\left(\mathbf{A}_{l,f}\right) \\ &+ \frac{\Phi^*(L^{-1})}{\Theta^*(L^{-1})}\left((\mathbf{I}_J \otimes \mathbf{\Sigma}_{\boldsymbol{\epsilon},f}^{-1})\frac{\Phi(L)}{\Theta(L)} \text{vec}\left(\mathbf{A}_{l,f}\right)\right),\end{aligned} \tag{4.29}$$

où on rappelle que $\Phi^*(L^{-1}) = \sum_{p=0}^{P} \varphi_p^* L^{-p}$ et $\Theta^*(L^{-1}) = \sum_{q=0}^{Q} \vartheta_q^* L^{-q}$. Nous rappelons également que dans cette équation, les ratios de polynômes en L et L^{-1} représentent des opérateurs de filtrage.

Nous voulons résoudre $\mathbf{g} = 0$ ce qui est ici équivalent à résoudre un système d'équations linéaires. De façon classique, nous utilisons la méthode du gradient conjugué avec préconditionnement [Golub et Van Loan, 1996] pour résoudre ce problème, celle-ci est présentée dans un cadre général à l'annexe D. A partir de l'équation (4.29) et en utilisant le fait que la matrice $\mathbf{\Sigma}_{\boldsymbol{\epsilon},f}$ est diagonale, nous définissons la matrice de préconditionnement diagonale suivante :

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_0 & 0 & \cdots & 0 \\ 0 & \mathbf{D}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{D}_{F-1} \end{pmatrix}, \tag{4.30}$$

où la matrice $\mathbf{D}_f$ est définie par :

$$\mathbf{D}_f = N \,\text{diag}(\hat{\mathbf{R}}_{\mathbf{ss},f}^\top \otimes \mathbf{\Sigma}_{\mathbf{b},f}^{-1}) + \frac{1}{\sigma_\epsilon^2}\mathbf{I}_{IJ} \sum_{s=0}^{N_\psi} |\psi_s|^2. \tag{4.31}$$

Les variables $\{\psi_s\}_{s=0}^{N_\psi}$ correspondent aux coefficients d'un polynôme $\Psi(L)$ caractérisant la fonction de transfert d'un modèle MA d'ordre N_Ψ. Nous cherchons ici à représenter le modèle ARMA inverse de fonction de transfert $\Phi(L)/\Theta(L)$ par ce modèle MA. Il s'agit d'une représentation exacte si l'ordre N_Ψ est infini. En pratique nous obtenons une approximation très précise en choisissant un ordre $N_\psi = 2048$ comme l'indique la densité spectrale de puissance représentée sur la figure 4.2.

La méthode du gradient conjugué avec préconditionnement pour la mise à jour de la partie tardive des filtres de mélange $\{\mathbf{A}_{l,f}\}_f$ est finalement résumée dans l'algorithme 1. Le critère d'arrêt pour cet algorithme itératif correspond à un nombre d'itérations égal à 20.

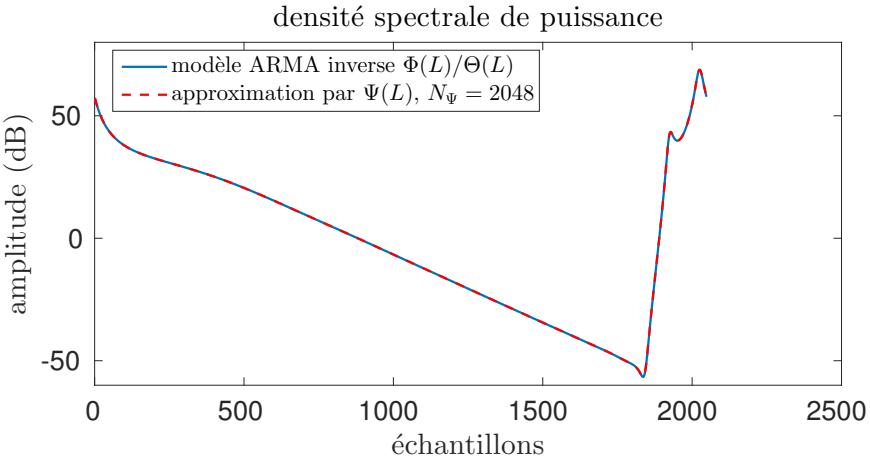


FIGURE 4.2 – Approximation du modèle ARMA(7,2) inverse de fonction de transfert $\Phi(L)/\Theta(L)$ par un modèle MA d'ordre $N_\psi = 2048$ de fonction de transfert $\Psi(L)$.

Algorithme 1 Mise à jour de $\{\mathbf{A}_{l,f}\}_f$ à l'étape M par la méthode du gradient conjugué avec préconditionnement

Initialiser $\mathbf{g}$ d'après (4.28) et (4.29)
Initialiser $\boldsymbol{\omega} = [\boldsymbol{\omega}_0^\top, \boldsymbol{\omega}_1^\top, ..., \boldsymbol{\omega}_{F-1}^\top]^\top$ où $\boldsymbol{\omega}_f = \mathbf{D}_f^{-1}\mathbf{g}_f$ avec $\mathbf{D}_f$ défini à l'équation (4.31)
Tant que critère d'arrêt non rencontré
    $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_0^\top, \boldsymbol{\gamma}_1^\top, ..., \boldsymbol{\gamma}_{F-1}^\top]^\top$ avec $\boldsymbol{\gamma}_f = N(\hat{\mathbf{R}}_{\mathbf{ss},f}^\top \otimes \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1})\boldsymbol{\omega}_f + \frac{\Phi^*(L^{-1})}{\Theta^*(L^{-1})}(\mathbf{I}_J \otimes \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},f}^{-1})\frac{\Phi(L)}{\Theta(L)}\boldsymbol{\omega}_f$
    $\mu = (\boldsymbol{\omega}^H\mathbf{g})/(\boldsymbol{\omega}^H\boldsymbol{\gamma})$
    $\mathbf{a}_l \leftarrow \mathbf{a}_l - \mu\boldsymbol{\omega}$
    Calculer $\mathbf{g}$ d'après (4.28) et (4.29)
    $\mathbf{g}_p = \mathbf{D}^{-1}\mathbf{g}$ avec $\mathbf{D}$ défini par les équations (4.30) et (4.31)
    $\alpha = -(\boldsymbol{\gamma}^H\mathbf{g}_p)/(\boldsymbol{\omega}^H\boldsymbol{\gamma})$
    $\boldsymbol{\omega} \leftarrow \mathbf{g}_p + \alpha\boldsymbol{\omega}$
Fin Tant que

4.4 Résultats expérimentaux

Dans cette section nous détaillons les expériences conduites à partir de la base de données «MASS-Synthétique» présentée à la section 2.7 du chapitre 2, pour laquelle le temps de réverbération est fixé à 128 ms.

Nous commençons par une expérience préliminaire ayant pour objectif de discuter l'hypothèse de filtres de mélange courts. Nous détaillons ensuite la procédure d'initialisation des paramètres du modèle. Finalement, après avoir étudié l'influence des hyper-paramètres des a priori, nous comparons les résultats de séparation obtenus par estimation MV des filtres de mélange [Ozerov et Févotte, 2010] [2] et ceux obtenus avec l'estimation MAP proposée. Des exemples audio sont disponibles en ligne [3].

2. Comme indiqué à la section 4.1.2, page 67, nous utilisons en fait l'algorithme proposé dans [Ozerov et al., 2011], qui reste néanmoins très proche de celui introduit dans [Ozerov et Févotte, 2010]. Pour des raisons de concision nous ne ferons donc référence qu'à la version originale de la méthode [Ozerov et Févotte, 2010].

3. `https://perso.telecom-paristech.fr/leglaive/demoMASSwPRP.html`

4.4.1 Expérience préliminaire concernant l'hypothèse de filtres de mélange courts

L'objectif de cette section est de discuter l'hypothèse de filtres de mélange courts utilisée pour exprimer le mélange convolutif dans le domaine de la TFCT. Nous supposons que la longueur des filtres de mélange est égale au temps de réverbération.

D'une part, pour écrire la convolution dans le domaine temporel comme une simple multiplication dans le domaine de la TFCT, nous devons théoriquement utiliser une fenêtre d'analyse plus longue que le temps de réverbération. D'autre part, nous ne pouvons pas choisir une fenêtre d'analyse trop longue en raison de l'hypothèse de stationnarité locale des signaux sources. Par ailleurs, les vrais filtres de mélange peuvent toujours être retrouvés en théorie dans le cas limite où la longueur de la fenêtre d'analyse est égale au temps de réverbération. Par exemple, en considérant le jeu de données utilisé ici, nous pouvons choisir une fenêtre de longueur égale à celle des filtres de mélange (128 ms) et interpréter la matrice de mélange $\mathbf{A}_f$ comme étant construite à partir des réponses en fréquence des filtres. En revanche si la fenêtre d'analyse est plus courte, les filtres sont sous-paramétrés et cette interprétation n'est plus valide.

Même si cette approche est théoriquement discutable, elle peut s'avérer efficace en termes de résultats de séparation de sources oracles. Cela peut expliquer pourquoi ce modèle de mélange est encore largement utilisé dans la littérature. En effet, même s'il ne permet pas de représenter le véritable processus de mélange (convolution dans le domaine temporel), sa simplicité est utile pour la séparation aveugle de sources audio et il permet d'atteindre des performances oracles (à paramètres du modèle connus) satisfaisantes.

Pour valider cet argument, nous avons mené une expérience préliminaire de séparation de sources en utilisant la méthode originale proposée dans [Ozerov et Févotte, 2010], c'est-à-dire avec estimation MV des paramètres. La longueur de la fenêtre d'analyse pour la TFCT est égale au temps de réverbération (128 ms). Nous considérons trois configurations :

- $\mathbf{A}_f$ est estimé de façon aveugle à l'étape M selon l'équation (4.10).
- La partie précoce des filtres $\mathbf{A}_{e,f}$ est fixée aux valeurs oracles en utilisant les filtres de mélange originaux tronqués au temps de mélange, tandis que la partie tardive des filtres $\mathbf{A}_{l,f}$ est fixée à zéro.
- La matrice de mélange complète $\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}$ est fixée de façon oracle grâce aux vrais filtres de mélange.

Tous les autres paramètres du modèle sont estimés de façon aveugle.

Les résultats moyens de séparation dans ces trois cas sont présentés dans le tableau 4.1. Nous observons tout d'abord que même si la longueur de la fenêtre d'analyse de la TFCT est égale à la longueur des filtres de mélange, nous obtenons de bons résultats de séparation lorsque la matrice $\mathbf{A}_f$ est fixée à sa valeur oracle. Nous observons également que lorsque nous négligeons la réverbération tardive en fixant $\mathbf{A}_{l,f}$ à zéro, les performances de séparation diminuent fortement par rapport au cas précédent. Ceci montre l'importance d'obtenir une bonne estimation pour les parties précoce et tardive des filtres de mélange.

Grâce à cette expérience préliminaire, nous pouvons conclure que l'on peut espérer de bonnes performances de séparation même si la longueur des filtres de mélange est égale à la longueur de la fenêtre d'analyse de la TFCT. En effet nous avons vu que lorsque les filtres sont connus nous obtenons une qualité de séparation satisfaisante. Le but de notre approche est précisément d'intégrer des contraintes sur les filtres de mélange afin de guider leur estimation dans le but de se rapprocher des vrais filtres de mélange, qui correspondent à des réponses de salle.

Mesures	Configuration	Résultats moyens
SDR	$\mathbf{A}_f$ est estimé de façon aveugle	0.0
	$\mathbf{A}_{e,f}$ est fixé de façon oracle et $\mathbf{A}_{l,f}$ est fixé à zéro	0.9
	$\mathbf{A}_f$ est fixé de façon oracle	**7.9**
ISR	$\mathbf{A}_f$ est estimé de façon aveugle	5.3
	$\mathbf{A}_{e,f}$ est fixé de façon oracle et $\mathbf{A}_{l,f}$ est fixé à zéro	7.6
	$\mathbf{A}_f$ est fixé de façon oracle	**15.6**
SIR	$\mathbf{A}_f$ est estimé de façon aveugle	2.1
	$\mathbf{A}_{e,f}$ est fixé de façon oracle et $\mathbf{A}_{l,f}$ est fixé à zéro	3.9
	$\mathbf{A}_f$ est fixé de façon oracle	**13.5**
SAR	$\mathbf{A}_f$ est estimé de façon aveugle	6.4
	$\mathbf{A}_{e,f}$ est fixé de façon oracle et $\mathbf{A}_{l,f}$ est fixé à zéro	10.8
	$\mathbf{A}_f$ est fixé de façon oracle	**12.0**

TABLEAU 4.1 – Résultats préliminaires de séparation de sources (en dB) moyennés sur les 29 sources de la base de données.

4.4.2 Initialisation des paramètres du modèle

La solution obtenue par un algorithme EM est très sensible aux valeurs initiales des paramètres du modèle. Nous décrivons ci-après la procédure d'intialisation utilisée, elle s'inspire de celle proposée dans [Ozerov et Févotte, 2010, section IV.H].

1. Superposer les TFCTs des canaux gauche et droite du mélange afin de créer une matrice à valeurs complexes de taille $2F \times N$.
2. Calculer une NMF de rang K_{init} avec la divergence d'Itakura-Saito à partir du module au carré de cette matrice. Séparer les atomes spectraux obtenus suivant leur association au canal gauche ou droite.
3. Reconstruire par filtrage de Wiener (voir par exemple [Févotte et al., 2009]) K_{init} composantes $c_{ik,fn}$ pour les canaux gauche ($i = 1$) et droit ($i = 2$). En considérant un modèle de propagation anéchoïque nous pouvons écrire $c_{ik,fn} = \rho_{ik} e^{-\imath 2\pi f \tau_{ik}/L_a} |c_{k,fn}| e^{\imath \arg(c_{k,fn})}$.
4. Calculer $\tilde{c}_{ik,fn} = c_{ik,fn}/e^{\imath \arg(c_{1k,fn})}$. Ces variables vérifient :
$$\tilde{c}_{1k,fn} = \rho_{1k}|c_{k,fn}|;$$
$$\tilde{c}_{2k,fn} = \rho_{2k}|c_{k,fn}|e^{-\imath 2\pi f(\tau_{2k}-\tau_{1k})/L_a}.$$
5. Calculer $\bar{c}_{ik,f} = \dfrac{1}{N}\sum_n \tilde{c}_{ik,fn}$.
6. Calculer et dérouler $\xi_{k,f} = \arg\left(\dfrac{\bar{c}_{2k,f}}{\bar{c}_{1k,f}}\right)$. Ces variables vérifient :
$$\xi_{k,f} = -2\pi f(\tau_{2k} - \tau_{1k})/L_a.$$
7. En supposant des sources ponctuelles et spatialement disjointes, les ensembles $\{\xi_{k,f}\}_f$, $k = 1, ..., K_{\text{init}}$, forment J partitions car de multiples composantes sont associées spatialement à la même source et possèdent donc la même différence de temps d'arrivée $(\tau_{2k} - \tau_{1k})$. Utiliser l'algorithme des «K-moyennes» pour partitionner les ensembles $\{\xi_{k,f}\}_f$ suivant k.

8. En notant $\mathcal{K}_j$ l'ensemble des indices k associés à la source j, initialiser les filtres de mélange suivant $a_{ij,f} = \dfrac{1}{\#\mathcal{K}_j} \sum_{k \in \mathcal{K}_j} \bar{c}_{ik,f}$.
9. Calculer une pré-séparation par masquage TF en utilisant les filtres de mélange ainsi obtenus. Nous utilisons pour cela le code fournit dans le cadre du challenge SiSEC 2008 [4].
10. Initialiser les paramètres NMF à partir des sources ainsi séparées en utilisant la divergence de Kullback-Leibler généralisée (afin d'être plus robuste aux artéfacts créés par le masquage TF).
11. Initialiser $\sigma^2_{b,f} = 10^{-2} \sum_{i,n} |x_{i,fn}|^2/(IN)$.

La méthode d'estimation MAP requiert l'initialisation individuelle des parties précoce et tardive des filtres de mélange. Pour cela nous séparons temporellement les filtres obtenus par la procédure précédente en accord avec le temps de mélange t_0 défini à l'équation (3.1), page 53.

4.4.3 Résultats de séparation : Comparaison des approches MV et MAP

Comme mentionné précédemment, nous utilisons une fenêtre d'analyse/synthèse pour la TFCT d'une longueur de 128 ms ou 2048 points à 16 kHz. Le rang du modèle de NMF est arbitrairement fixé à 10 pour toutes les sources. Les algorithmes EM dans le cas des estimations MV et MAP sont lancés pendant 500 itérations à partir de la même initialisation.

Nous rappelons que les coefficients AR du modèle pour la partie précoce des filtres de mélange sont estimés à l'étape M de l'algorithme EM. Les coefficients du modèle ARMA(7,2) pour la réverbération tardive sont en revanche estimés à partir de l'ACVF théorique définie à l'équation (3.16), page 58, connaissant certains paramètres de salle.

Les variances des deux a priori correspondent à des paramètres critiques pour les résultats de séparation comme l'indique la figure 4.3. Nous observons que selon le choix de ces deux hyper-paramètres, le SDR moyen pour un des mélanges de la base de données varie entre 0.2 et 4.2 dB. Ces variances contrôlent l'influence des a priori pour l'estimation des filtres de mélange, plus leur valeur est faible, plus les a priori ont d'importance et inversement. Si leur valeur est trop élevée, les a priori deviennent «plats» et n'ont pas d'effet sur les résultats par rapport à l'estimation MV des filtres. Dans ce cas, c'est la vraisemblance qui domine dans la fonction de coût définie à l'équation (4.22). Du fait de cette dépendance des résultats aux variances des a priori, nous évaluons les performances de séparation avec l'approche MAP sur une grille de valeurs pour ces hyper-paramètres. La grille utilisée correspond aux graduations des axes de la figure 4.3.

a) Variances optimales pour chaque mélange

Pour chaque mélange nous calculons les résultats de séparation pour l'ensemble des valeurs des variances appartenant à la grille de recherche. Nous sélectionnons ensuite individuellement pour chaque mélange les variances associées aux meilleurs performances. Les résultats de séparation suivant cette stratégie sont présentés dans la colonne MAP* du tableau 4.2. Nous les comparons aux résultats de séparation sans a priori sur les filtres de mélange reportés dans la colonne MV.

Nous observons que pour tous les mélanges, l'amélioration en terme de SDR grâce à l'estimation MAP varie de 0.1 à 3.6 dB. En moyenne nous améliorons les résultats de 1.2 dB en terme de SDR, 0.9 dB pour l'ISR, 1.3 dB pour le SIR et 1.1 dB en SAR.

4. `http://sisec2008.wiki.irisa.fr`

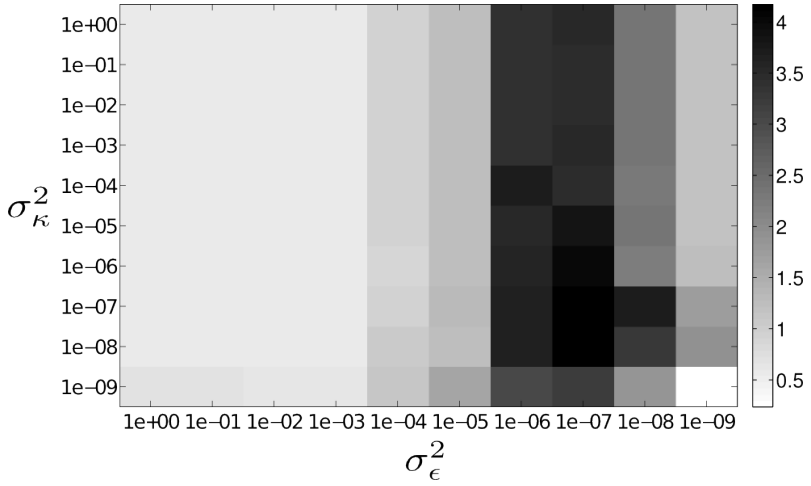


FIGURE 4.3 – SDR moyen (en dB) pour un des mélanges de la base de données, calculé en fonction des valeurs des variances des a priori σ^2_ϵ et σ^2_κ.

Cette première expérience permet de montrer que pour chaque mélange il existe des valeurs pour les variances des a priori qui permettent d'améliorer les résultats de séparation par rapport à une estimation MV des filtres de mélange.

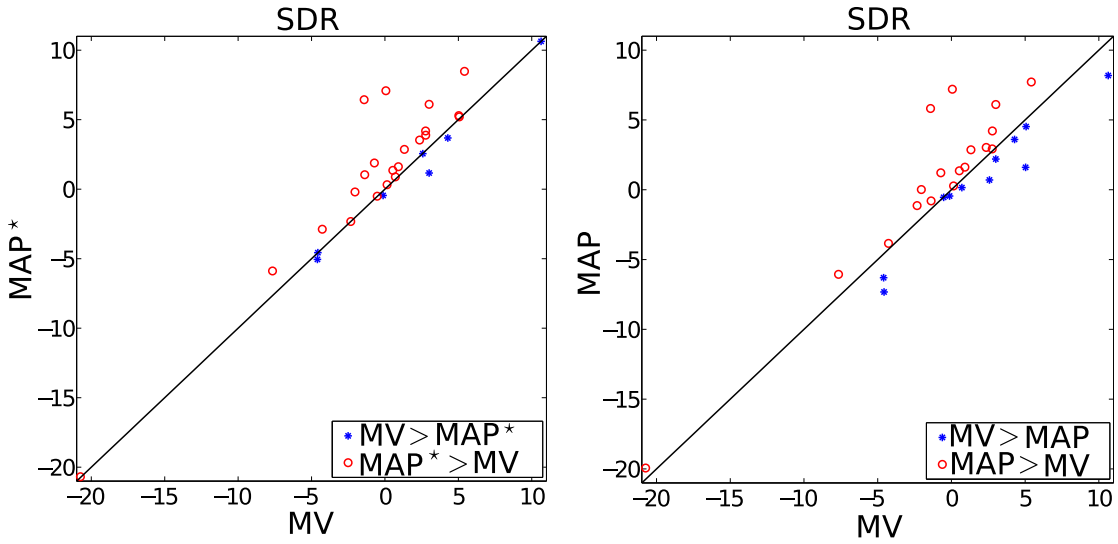
b) Variances choisies par validation croisée

Nous présentons maintenant les résultats de séparation obtenus lorsque les variances sont choisies par validation croisée. 7 mélanges parmi les 8 sont choisis comme échantillons d'entraînement afin de sélectionner les valeurs optimales des variances (maximisant le SDR moyenné sur l'ensemble des sources). La qualité de séparation est évaluée sur le dernier mélange qui sert d'échantillon test. Nous répétons cette procédure 8 fois afin de traiter successivement chaque mélange comme échantillon test. De façon intéressante, nous obtenons les variances optimales suivantes $(\sigma^2_\kappa, \sigma^2_\epsilon) = (10^{-6}, 10^{-7})$ pour chacune des 8 configurations entraînement/test. Les résultats obtenus avec cette technique de validation croisée sont présentés dans la colonne MAP du tableau 4.2. Par rapport à l'approche de référence MV nous voyons que le SDR décroît de 1.5 et 1.3 dB pour les mélanges 1 et 8 respectivement. Cependant, pour tous les autres mélanges, nous améliorons les performances en terme de SDR grâce à la prise en compte des a priori sur les filtres de mélange. L'amélioration varie entre 0.1 et 3.4 dB. En moyenne, l'approche MAP avec validation croisée permet d'améliorer le SDR de 0.7 dB, l'ISR de 0.7 dB, le SIR de 0.6 dB et le SAR de 0.2 dB.

Finalement, afin d'évaluer la significativité des résultats moyennés obtenus, nous traçons sur la figure 4.4 un nuage de points représentant le SDR (en dB) des 29 sources de la base de données obtenu avec l'estimation MAP des filtres de mélange (en ordonnées) en fonction de l'estimation MV (en abscisses). Les points rouges au dessus de la droite d'équation $y = x$ indiquent une meilleure séparation avec l'approche proposée. Sur la figure 4.4a nous présentons les résultats dans le cas où les variances pour les a priori sont choisies pour maximiser le SDR moyen individuellement pour chaque mélange (cas MAP*). Nous obtenons une majorité de points rouges indiquant une amélioration significative des performances de séparation grâce à la prise en compte des a priori. Sur la figure 4.4b nous traçons les résultats dans le cas où les variances sont choisies par validation croisée (cas MAP). Nous observons plus de points bleus que précédemment, néanmoins les points rouges restent majoritaires.

	SDR			ISR			SIR			SAR		
Mélange	MV	MAP⋆	MAP	MV	MAP⋆	MAP	MV	MAP⋆	MAP	MV	MAP⋆	MAP
1	-4.4	**-4.3**	-5.9	3.9	**4.1**	3.5	**0.4**	**0.4**	-3.4	**10.4**	9.8	7.3
2	-1.3	**0.0**	-1.0	4.1	4.0	**4.5**	-0.7	**1.2**	0.5	7.4	**8.5**	7.5
3	2.6	**3.8**	3.7	6.1	8.1	**8.4**	5.6	6.5	**6.7**	6.6	**9.4**	6.9
4	0.6	**4.2**	4.0	5.8	**8.4**	8.2	2.3	**6.8**	6.7	7.9	**9.6**	9.5
5	1.1	**2.0**	1.2	6.5	**7.6**	6.3	**2.0**	1.6	1.0	6.2	**6.3**	5.1
6	1.3	**2.5**	**2.5**	4.4	**5.9**	**5.9**	2.9	**5.7**	**5.7**	4.6	**6.1**	**6.1**
7	-0.7	**0.1**	0.0	5.0	**5.1**	**5.1**	0.5	**0.8**	0.7	3.4	**4.6**	4.3
8	**1.2**	**1.2**	-0.1	**7.5**	**7.5**	6.9	**5.2**	**5.2**	3.6	6.7	6.7	**6.8**
Moyenne	0.0	**1.2**	0.7	5.3	**6.2**	6.0	2.1	**3.4**	2.7	6.4	**7.5**	6.6

TABLEAU 4.2 – Résultats de séparation de sources en dB : sans a priori (MV) [Ozerov et Févotte, 2010; Ozerov et al., 2011], avec a priori et variances optimales pour chaque mélange (MAP⋆), avec a priori et variances choisies par validation croisée (MAP). La moyenne est calculée sur les 29 sources de la base de données.



(a) Cas où les variances des a priori sont optimales pour chaque mélange.

(b) Cas où les variances des a priori sont choisies par validation croisée.

FIGURE 4.4 – SDR (en dB) des 29 sources de la base de données obtenu avec l'estimation MAP des filtres de mélange (en ordonnées) en fonction de l'estimation MV (en abscisses). Le point de SDR proche de -20 dB est associé aux balais de batterie présents dans un des mélanges. Cette source n'est pas bien modélisée par une NMF.

4.5 Conclusion

Nous avons introduit au chapitre 3 de nouveaux modèles fréquentiels pour la partie précoce des réponses de salle et pour la réverbération tardive. Ces modèles visent à transcrire la dynamique temporelle des réponses de salle sous forme de corrélations dans le domaine fréquentiel. Nous avons ensuite utilisé ces modèles dans le chapitre 4 pour définir des a priori probabilistes sur les filtres de mélange dans le cadre d'un problème de séparation de sources sous-déterminé pour des mélanges multicanaux et réverbérants. Ces a priori ont été pris en compte dans une approche de type maximum a posteriori pour guider l'estimation des filtres de mélange. Nous avons montré que cette technique permet d'améliorer les performances de séparation par rapport à une estimation au

sens du maximum de vraisemblance.

Cependant, les résultats obtenus dépendent fortement des valeurs des variances des distributions a priori, ces dernières permettant de contrôler l'influence respective des données et des a priori dans l'estimation des filtres. Dans de futurs travaux, nous pourrions utiliser un a priori sur ces variances afin que leurs valeurs soient contraintes à être proches de celles obtenues par validation croisée, mais non exactement égales.

Dans ce travail nous nous sommes placés dans un cas semi-informé où le temps de réverbération, le volume et la surface des parois de la pièce étaient supposés connus, ceci afin de pouvoir calculer les paramètres du modèle ARMA à l'origine de l'a priori sur la réverbération tardive. Nous pourrions par la suite essayer d'estimer de façon aveugle certains de ces paramètres comme par exemple le temps de réverbération, en s'appuyant sur des travaux de la littérature tels que [Ratnam et al., 2003]. Par ailleurs, il pourrait être intéressant d'analyser la sensibilité de la méthode par rapport à des erreurs concernant ces paramètres de salle.

Nous pourrions également étendre l'approche proposée à un modèle de sources non-ponctuelles par l'intermédiaire de matrices de covariance spatiale. Ces matrices peuvent en effet être factorisées comme le produit extérieur de matrices de mélange associées à des sous-sources (voir section 2.4.2, page 37) pour lesquelles nous pourrions utiliser les mêmes a priori que ceux présentés précédemment.

Enfin, nous pourrions considérer non seulement la modélisation des corrélations fréquentielles des filtres de mélange, mais aussi des corrélations spatiales. Comme cela a été utilisé dans [Duong et al., 2010] et [Duong et al., 2013] pour la séparation de sources audio stéréo, la fonction de corrélation spatiale est théoriquement définie en fonction de la distance entre les deux microphones. Afin de prendre en compte ces corrélations spatiales, nous pourrions envisager d'utiliser une matrice de covariance du bruit à l'équation (4.19) qui soit pleine, ou bien un modèle ARMA vectoriel pour représenter la réverbération tardive dans le domaine fréquentiel.

Troisième partie

Modélisation du mélange dans le domaine temporel

Chapitre 5

Filtres de mélange déterministes

Sommaire

5.1 Représentation temporelle du mélange 85
5.2 Représentation temps-fréquence des sources 85
5.2.1 Transformée en cosinus discrète modifiée 85
5.2.2 Transformée de Fourier à court-terme 86
5.2.3 Transformée de Fourier à court-terme à fréquence impaire 87
5.2.4 Formulation générale 87
5.3 Modèle de source gaussien 88
5.3.1 Modèle de source 88
5.3.2 Inférence variationnelle 88
5.3.3 Résultats expérimentaux oracles pour la MDCT 95
5.3.4 Résultats expérimentaux semi-aveugles 96
5.3.5 Conclusion 98
5.4 Modèle de source *t* de Student 98
5.4.1 Modèle de source 98
5.4.2 Inférence variationnelle 99
5.4.3 Résultats expérimentaux semi-aveugles 103
5.5 Conclusion 105

L'hypothèse de filtres de mélange courts par rapport à la longueur de la fenêtre d'analyse de la TFCT limite fondamentalement les performances de séparation dans le cas de mélanges enregistrés en présence de forte réverbération. Cette observation a déjà été faite dans plusieurs articles de la littérature [Duong et al., 2010; Kowalski et al., 2010; Li et al., 2017a,b] et c'est ce qui a motivé le développement d'autres modèles de mélange tels que ceux introduits au chapitre 2, section 2.4.

Nous rappelons que la longueur des filtres est en général considérée comme égale au temps de réverbération. Dans un environnement domestique ou de travail, celui-ci varie souvent entre 0.1 et 0.8 s, alors qu'il peut monter jusqu'à 2 s dans des salles de concert [Bies et Hansen, 2009]. A l'inverse, les fenêtres d'analyse utilisées pour le calcul de la TFCT sont souvent courtes (de l'ordre de la centaine voire de la dizaine de millisecondes) afin de satisfaire l'hypothèse de stationnarité locale des signaux sources. Par conséquent, plus le temps de réverbération est grand, plus l'approximation de la convolution temporelle comme une simple multiplication dans le domaine de la TFCT devient fausse. Afin d'illustrer cela nous représentons sur la figure 5.1 l'erreur quadratique relative moyenne entre la TFCT du mélange convolutif et son approximation par multiplication de la TFCT des sources avec la réponse en fréquence des filtres de mélange :

$$\Delta x = \frac{1}{IFN} \sum_{i=1}^{I} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \frac{|x_{i,fn} - \hat{x}_{i,fn}|^2}{|x_{i,fn}|^2}, \tag{5.1}$$

où $x_{i,fn}$ est le point TF (f, n) de la TFCT du mélange au niveau du microphone i et $\hat{x}_{i,fn}$ son approximation. Il est important de mentionner que pour calculer $\hat{x}_{i,fn}$ nous avons dû tronquer les filtres de mélange dans le domaine temporel lorsque ces derniers étaient plus longs que la fenêtre d'analyse. Cette procédure arbitraire limite l'interprétation des résultats fournis par la figure 5.1, car il est éventuellement envisageable de trouver d'autres sous-paramétrisations des filtres induisant une erreur plus faible.

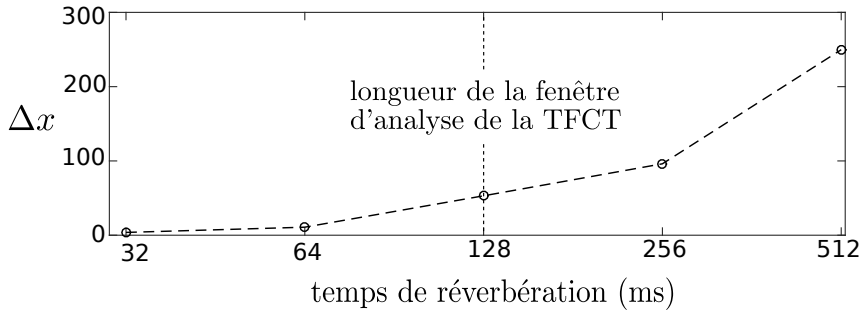


FIGURE 5.1 – Erreur de modélisation due à l'hypothèse de filtres de mélange courts par rapport à la longueur de la fenêtre d'analyse utilisée dans le calcul de la TFCT, fixée ici à 128 ms.

Nous souhaitons explorer dans ce chapitre de nouvelles approches pour la séparation de sources permettant de s'affranchir de cette approximation. Dans la lignée des travaux initiés dans [Kowalski et al., 2010], nous allons directement travailler à partir des signaux temporels du mélange, tout en considérant un modèle TF pour les sources. Nous proposons une approche probabiliste permettant d'inférer les variables aléatoires latentes TF des sources à partir des observations temporelles du mélange. Comme nous le verrons au chapitre suivant, cela permet non seulement de représenter le processus de mélange convolutif de façon exacte, mais aussi de prendre en compte des a priori simples sur les filtres afin d'exploiter leur structure temporelle.

Nous commençons par rappeler dans la section 5.1 la représentation temporelle du mélange convolutif. Nous présentons ensuite à la section 5.2 les transformations TF qui seront utilisées pour

modéliser les signaux sources. Nous développons dans la section 5.3 une technique d'inférence variationnelle pour la séparation de sources dans le cadre d'un modèle gaussien. Cette méthode a été publiée dans [Leglaive et al., 2017a,c]. Nous discutons notamment le choix de la représentation TF (à valeurs réelles ou complexes, redondante ou à échantillonnage critique, etc.). Finalement, dans la section 5.4 nous adaptons cette technique au cas d'un modèle de source *t* de Student. Cette approche a été publiée dans [Leglaive et al., 2017b].

5.1 Représentation temporelle du mélange

Il nous semble utile de rappeler ici quelques éléments déjà introduits au chapitre 1, concernant l'expression du mélange convolutif. Le signal $x_i(t) \in \mathbb{R}$, capté par le microphone $i \in \{1, ..., I\}$ est représenté comme la somme de sources images $y_{ij}(t) \in \mathbb{R}$, $j \in \{1, ..., J\}$, et d'un bruit blanc gaussien $b_i(t) \in \mathbb{R}$ tel que pour $t \in \{0, ..., T-1\}$:

$$x_i(t) = \sum_{j=1}^{J} y_{ij}(t) + b_i(t), \qquad b_i(t) \sim \mathcal{N}_{\mathbb{R}}(0, \sigma_i^2), \tag{5.2}$$

où $\mathcal{N}_{\mathbb{R}}$ représente dans ce chapitre la loi gaussienne pour une variable aléatoire à valeur réelle, sa densité de probabilité est définie à l'annexe A. Chaque source image s'exprime comme le produit de convolution du signal source $s_j(t) \in \mathbb{R}$, $t \in \{0, ..., L_s-1\}$, avec le filtre de mélange $a_{ij}(t) \in \mathbb{R}$, $t \in \{0, ..., L_a - 1\}$. On a donc pour tout $t \in \{0, ..., T-1\}$ avec $T = L_s + L_a - 1$:

$$y_{ij}(t) = [a_{ij}(\cdot) \star s_j(\cdot)](t). \tag{5.3}$$

5.2 Représentation temps-fréquence des sources

Inspirés par le modèle génératif temporel proposé dans [Févotte et Kowalski, 2014], nous représentons chaque source $s_j(t)$ par un ensemble de coefficients TF de synthèse $\{s_{j,fn} \in \mathbb{K}\}_{f,n}$, où $\mathbb{K}$ correspond à l'ensemble des réels ou des complexes. Ce sont les coefficients associés à une représentation du signal source par combinaison linéaire d'atomes TF, comme cela a été introduit au chapitre 1, section 1.2.1. Dans la suite de ce manuscrit nous notons $\{\psi_{fn}(t) \in \mathbb{K}\}_{(f,n)\in\{0,...,F-1\}\times\{0,...,N-1\}}$ l'ensemble des atomes TF de synthèse ; nous omettons le tilde qui était précédemment utilisé à l'équation (1.10), page 11.

La représentation TF utilisée est définie par l'ensemble $\mathbb{K}$ auquel appartiennent les coefficients de synthèse et par le choix du dictionnaire TF. Nous introduisons ci-dessous les transformées qui seront utilisées dans cette partie de la thèse.

5.2.1 Transformée en cosinus discrète modifiée

La transformée en cosinus discrète modifiée (MDCT) est à valeurs réelles ; $\mathbb{K} = \mathbb{R}$. Les atomes TF de synthèse sont définis pour tout $t \in \{0, .., L_s - 1\}$, $n \in \{0, ..., N-1\}$ et $f \in \{0, ..., F-1\}$ par [Malvar, 1992] :

$$\psi_{fn}(t) = \sqrt{\frac{4}{L_w}} w(t - nH) \cos\left(\frac{2\pi}{L_w}\left(f + \frac{1}{2}\right)\left(t - nH + \frac{1}{2} + \frac{L_w}{4}\right)\right), \tag{5.4}$$

où $w(t)$ est une fenêtre sinusoïdale de longueur L_w définie à l'équation (1.6), page 9, $H = L_w/2$ est le facteur d'incrément, N le nombre de trames de signal et $F = L_w/2$ le nombre de points en fréquence. Dans toute la suite nous supposons que L_w est un entier pair.

Un des avantages de la MDCT correspond au fait que c'est une transformation à échantillonnage critique, c'est-à-dire que le nombre de coefficients TF est égal au nombre d'échantillons dans le domaine temporel. Cette propriété permet de limiter le temps de calcul d'une méthode de séparation de sources où l'objectif est d'estimer l'ensemble des coefficients TF. En revanche, un potentiel inconvénient de la MDCT est que l'information temporelle portée par la phase est directement encodée dans les amplitudes des coefficients TF. Un spectrogramme calculé à partir d'une MDCT ne correspond donc pas une représentation invariante à un décalage temporel.

5.2.2 Transformée de Fourier à court-terme

La TFCT est une transformation à valeurs complexes, c'est-à-dire que $\mathbb{K} = \mathbb{C}$. Les atomes TF de synthèse ont été définis à l'équation (1.16), page 11, où le nombre de points en fréquence F est égal à la longueur de la fenêtre d'analyse/synthèse L_w. La TFCT est donc une transformation redondante où le nombre de coefficients TF est supérieur au nombre d'échantillons dans le domaine temporel. Le facteur de redondance est contrôlé par la taille de l'incrément H, ou de façon équivalente par le taux de recouvrement entre trames successives. Un spectrogramme d'amplitude ou de puissance calculé à partir d'une TFCT est une représentation invariante par décalage temporel car la phase est contenue dans l'argument des coefficients TF complexes.

Il est bien connu que pour un signal temporel à valeur réelle, la TFCT étant construite à partir de la TFD, elle possède la propriété de symétrie hermitienne ; les coefficients TF vérifient $s_{j,f'n} = s^*_{j,fn}$ avec $f' = F - f$. On voit en particulier qu'à la fréquence de Nyquist $f = L_w/2$, les coefficients sont réels car $s_{j,\frac{L_w}{2}n} = s^*_{j,\frac{L_w}{2}n}$. Par ailleurs, la discrétisation de l'axe fréquentiel implique une périodisation des trames d'analyse tel que $s_{j,f'n} = s_{j,fn}$ où $f' = f + F$. Cette périodicité combinée avec la propriété de symétrie hermitienne implique que les coefficients associés à la fréquence nulle $f = 0$ sont également réels.

On se rappelle que dans un contexte probabiliste, développer un modèle de source consiste à définir la distribution jointe de l'ensemble des coefficients $\{s_{j,fn}\}_{(f,n)\in\{0,\dots,F-1\}\times\{0,\dots,N-1\}}$. La propriété de symétrie hermitienne de la TFCT est dans ce cas problématique car elle crée un lien déterministe entre certains coefficients, de telle sorte que cette distribution jointe est dégénérée. Ce problème est résolu dans [Févotte et Kowalski, 2014] en supposant des signaux temporels à valeur complexe. Nous souhaitons ici prendre explicitement en compte le fait que les signaux audio sont réels. Pour cela nous pouvons exploiter la symétrie hermitienne afin d'introduire une nouvelle équation de synthèse n'impliquant que la partie non redondante du spectre :

$$s_j(t) = \sum_{n=0}^{N-1} \left[\underbrace{s_{j,0n}\,\psi_{0n}(t)}_{\text{fréquence nulle}} + \underbrace{s_{j,\frac{L_w}{2}n}\,\psi_{\frac{L_w}{2}n}(t)}_{\text{fréquence de Nyquist}} + 2\Re\left(\sum_{f=1}^{L_w/2-1} s_{j,fn}\psi_{fn}(t) \right) \right], \tag{5.5}$$

où $\Re(\cdot)$ est la partie réelle.

Grâce à cette reformulation et en posant $F = L_w/2$, la distribution jointe des coefficients $\{s_{j,fn}\}_{(f,n)\in\{0,\dots,F-1\}\times\{0,\dots,N-1\}}$ n'est plus dégénérée. Cependant un autre problème se pose : Les coefficients aux fréquences nulle et de Nyquist sont réels tandis que tous les autres sont complexes. Il nous est donc impossible de définir de façon rigoureuse un modèle probabiliste unique pour traiter l'ensemble des fréquences.

Du fait de ces limitations liées à l'utilisation de la TFCT, nous proposons ci-dessous d'avoir recours à la TFCT à fréquence impaire, notée TFCTFI.

5.2.3 Transformée de Fourier à court-terme à fréquence impaire

a) Transformée de Fourier discrète à fréquence impaire

De la même façon que pour la TFCT (voir chapitre 1 section 1.2.1), la TFCTFI est construite à partir de la TFD à fréquence impaire (TFDFI) que nous introduisons ici.

La TFDFI d'un signal $u(t)$, $t = 0, ..., L_u - 1$, est définie pour $f = 0, ..., L_u - 1$ par

$$u(f) = \frac{1}{\sqrt{L_u}} \sum_{t=0}^{L_u-1} u(t) \exp\left(-\imath \frac{2\pi}{L_u}\left(f + \frac{1}{2}\right) t\right). \tag{5.6}$$

Il s'agit d'un cas particulier de la TFD généralisée [Bongiovanni et al., 1976]. Par rapport à la TFD, nous voyons grâce au terme $\left(f + \frac{1}{2}\right)$ qu'il s'agit uniquement d'une rotation au niveau de l'échantillonage du cercle unité en fréquences discrètes, de sorte qu'aucun coefficient ne se trouve sur l'axe des réels.

Pour un signal réel $u(t)$, la TFDFI possède la propriété de symétrie suivante : $u(L_u - f - 1) = u(f)^*$. Contrairement à la TFD tous les coefficients sont complexes, la TFDFI est donc plus appropriée lorsque les coefficients fréquentiels sont modélisés comme des variables aléatoires à valeur complexe, comme c'est souvent le cas en séparation de sources audio. De plus, cette propriété de symétrie nous permet d'écrire une expression simple de la TFDFI inverse n'impliquant que la partie non-redondante du spectre :

$$u(t) = \frac{2}{\sqrt{L_u}} \Re\left(\sum_{f=0}^{L_u/2-1} u(f) \exp\left(\imath \frac{2\pi}{L_u}\left(f + \frac{1}{2}\right) t\right)\right). \tag{5.7}$$

Finalement, la TFCTFI est définie de la même façon que la TFCT mais en utilisant la TFDFI au lieu de la TFD. Nous pouvons également mentionner qu'il a été montré dans [Badeau, 2016] que la TFCTFI tout comme la MDCT est plus appropriée que la TFCT pour supposer l'indépendance des points TF (voir section 2.2.1), qui est une hypothèse fréquente en séparation de sources audio.

b) Transformation à court-terme

L'équation de synthèse dans le cas de la TFCTFI s'écrit alors :

$$s_j(t) = 2\Re\left(\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t)\right), \tag{5.8}$$

où les atomes TF de synthèse sont définis pour tout $t \in \{0, .., L_s - 1\}$, $n \in \{0, ..., N - 1\}$ et $f \in \{0, ..., F - 1\}$ avec $F = L_w/2$ par :

$$\psi_{fn}(t) = \sqrt{\frac{1}{L_w}} w(t - nH) \exp\left(\imath \frac{2\pi}{L_w}\left(f + \frac{1}{2}\right)(t - nH)\right). \tag{5.9}$$

5.2.4 Formulation générale

Nous souhaitons finalement regrouper dans un même formalisme les représentations par MDCT et TFCTFI. Nous définissons pour cela l'équation de synthèse générale suivante à partir des coefficients TF $\{s_{j,fn} \in \mathbb{K} = \mathbb{C} \text{ ou } \mathbb{R}\}_{(f,n)\in\mathcal{B}}$ où $F = L_w/2$ et $\mathcal{B} = \{0, ..., F - 1\} \times \{0, ..., N - 1\}$:

$$s_j(t) = \frac{2}{\phi} \Re\left(\sum_{(f,n)\in\mathcal{B}} s_{j,fn} \psi_{fn}(t)\right), \tag{5.10}$$

avec $\phi = 1$ et $\psi_{fn}(t)$ défini à l'équation (5.9) si $\mathbb{K} = \mathbb{C}$ (cas de la TFCTFI) ou $\phi = 2$ et $\psi_{fn}(t)$ défini à l'équation (5.4) si $\mathbb{K} = \mathbb{R}$ (cas de la MDCT).

A partir des équations (5.3) et (5.10), une source image peut finalement se réécrire :

$$y_{ij}(t) = \frac{2}{\phi} \Re \left(\sum_{(f,n) \in \mathcal{B}} s_{j,fn} g_{ij,fn}(t) \right), \tag{5.11}$$

où $g_{ij,fn}(t) = [a_{ij}(\cdot) \star \psi_{fn}(\cdot)](t)$.

5.3 Modèle de source gaussien

Dans cette section nous développons une technique d'inférence variationnelle des coefficients TF des sources en utilisant un modèle gaussien. Nous avons tout d'abord introduit cette approche pour le cas d'une transformation TF réelle basée sur la MDCT dans [Leglaive et al., 2017a]. Nous l'avons ensuite généralisée au cas de la TFCTFI dans [Leglaive et al., 2017c]. Afin d'éviter d'introduire des redondances dans ce manuscrit, nous présentons le cas le plus général, basé donc sur l'équation de synthèse (5.10) qui inclut à la fois le cas de la MDCT et de la TFCTFI. Néanmoins il est important de mentionner que la méthode d'inférence développée ci-dessous est beaucoup plus simple dans le cas de la MDCT, précisément car nous n'avons pas à traiter de dépendances a posteriori entre les parties réelle et imaginaire des coefficients TF des sources. Pour s'en convaincre, le lecteur intéressé pourra remarquer que les équations de l'article [Leglaive et al., 2017a] correspondent à une simplification des équations développées ci-après.

5.3.1 Modèle de source

Nous considérons à nouveau le modèle de source gaussien introduit au chapitre 2 section 2.2.2. Les coefficients TF des sources sont donc modélisés comme des variables aléatoires indépendantes, gaussiennes, centrées, et à symétrie circulaire dans le cas complexe :

$$s_{j,fn} \sim \begin{cases} \mathcal{N}_{\mathbb{R}}(0, \lambda_{j,fn}^2) & \text{si } \mathbb{K} = \mathbb{R} \\ \mathcal{N}_{\mathbb{C}}^{p}(0, \lambda_{j,fn}^2) & \text{si } \mathbb{K} = \mathbb{C} \end{cases}, \tag{5.12}$$

où $\mathcal{N}_{\mathbb{C}}^{p}$ représente la loi gaussienne complexe propre dont la densité de probabilité est définie à l'annexe A. Dans la suite de cette section, l'égalité $\mathbb{K} = \mathbb{R}$ doit être comprise comme équivalente à «la transformation TF utilisée correspond à la MDCT». Il en est de même pour $\mathbb{K} = \mathbb{C}$ avec la TFCTFI.

Les variances $\lambda_{j,fn}^2 \in \mathbb{R}_+$ de ces variables aléatoires sont de plus structurées par l'intermédiaire d'un modèle NMF que nous rappelons ici :

$$\lambda_{j,fn}^2 = (\mathbf{W}_j \mathbf{H}_j)_{f,n}, \tag{5.13}$$

avec $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$, $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$.

5.3.2 Inférence variationnelle

Soit $\mathbf{x} = \{x_i(t)\}_{i,t}$ l'ensemble des données observées, $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$ l'ensemble des variables latentes et $\boldsymbol{\theta} = \{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\}$ l'ensemble des paramètres du modèle.

Nous pouvons montrer d'après le modèle de mélange (5.2) et le modèle de source (5.12) que la distribution (multivariée) a posteriori des variables latentes, $p(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta})$, est gaussienne. Dans le cas

où la transformation TF utilisée est la TFCTFI, nous pouvons de plus montrer que cette distribution n'est pas «propre» [Adali et al., 2011], ce qui signifie qu'en plus d'être paramétrée par un vecteur moyenne $\mathbb{E}[\mathbf{s}]$ et une matrice de covariance $\mathbb{E}\left[(\mathbf{s}-\mathbb{E}[\mathbf{s}])\,(\mathbf{s}-\mathbb{E}[\mathbf{s}])^{H}\right]$, sa matrice de pseudo-covariance $\mathbb{E}\left[(\mathbf{s}-\mathbb{E}[\mathbf{s}])\,(\mathbf{s}-\mathbb{E}[\mathbf{s}])^{\top}\right]$ est non-nulle. Le problème principal que l'on rencontre ici est alors que ces matrices de (pseudo-)covariances sont pleines et de si grande dimension qu'il n'est pas envisageable de les stocker en mémoire sur un ordinateur.

Cela peut s'expliquer intuitivement en comprenant le processus de génération du mélange à partir des variables latentes : A chaque instant $t \in \{0, ..., T-1\}$, l'ensemble des JFN variables latentes ($2JFN$ si l'on distingue les parties réelle et imaginaire dans le cas complexe) sont mélangées au travers des équations (5.11) puis (5.2), pour donner un échantillon du mélange $x_i(t)$. Il est donc normal qu'a posteriori, une fois le mélange observé, toutes les variables latentes soient dépendantes, y compris leurs parties réelle et imaginaire lorsque l'on est dans le cas complexe. C'est pourquoi, bien qu'une technique d'inférence exacte basée sur un algorithme EM soit théoriquement possible, nous allons avoir recours à une approche variationnelle qui sous une hypothèse de champ moyen va permettre de briser certaines de ces dépendances a posteriori. On s'appuiera pour cela sur les résultats développés lors de l'introduction aux méthodes variationnelles dans le chapitre 2, section 2.5.3b.

a) Approximation de champ moyen

Nous cherchons la densité de probabilité $q \in \mathcal{F}$ sur les variables latentes $\mathbf{s}$ qui maximise l'énergie variationelle libre :

$$\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right)\right\rangle_{q}. \tag{5.14}$$

Nous formulons l'hypothèse de champ moyen qui consiste à supposer que la famille variationnelle $\mathcal{F}$ correspond à l'ensemble des densités de probabilité qui se factorisent sous la forme suivante :

$$q(\mathbf{s}) = \prod_{j=1}^{J}\prod_{f=0}^{F-1}\prod_{n=0}^{N-1} q_{jfn}(s_{j,fn}). \tag{5.15}$$

Cette hypothèse signifie précisément que l'on néglige toutes les dépendances a posteriori entre variables latentes, mais pas entre leurs parties réelle et imaginaire dans le cas complexe.

Sous cette hypothèse, l'estimation du point TF (f,n) de la source j est donnée par :

$$\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_{q}. \tag{5.16}$$

Les signaux temporels $\hat{s}_j(t)$ sont ensuite reconstruits par transformation TF inverse :

$$\hat{s}_j(t) = \frac{2}{\phi}\Re\left(\sum_{(f,n)\in\mathcal{B}} \hat{s}_{j,fn}\psi_{fn}(t)\right), \tag{5.17}$$

et la source image $\hat{y}_{ij}(t)$ est obtenue par convolution avec le filtre de mélange correspondant :

$$\hat{y}_{ij}(t) = [a_{ij}(\cdot)\star\hat{s}_j(\cdot)](t) = \frac{2}{\phi}\Re\left(\sum_{(f,n)\in\mathcal{B}} \hat{s}_{j,fn}g_{ij,fn}(t)\right). \tag{5.18}$$

b) Vraisemblance des données complètes

D'après le modèle présenté ci-dessus, la log-vraisemblance des données complètes $\ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) + \ln p(\mathbf{s};\boldsymbol{\theta})$ s'écrit :

$$\ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta}) \overset{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] - \frac{1}{\phi}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln(\lambda_{j,fn}^2) + \frac{|s_{j,fn}|^2}{\lambda_{j,fn}^2}\right]. \tag{5.19}$$

c) Étape E

Sous l'approximation de champ moyen, on peut montrer (cf. chapitre 2, section 2.5.3b) que les densités de probabilité $q_{jfn}(s_{j,fn})$ qui maximisent l'énergie variationnelle libre satisfont :

$$\begin{aligned}
\ln q_{jfn}^{\star}(s_{j,fn}) &= \langle \ln p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})\rangle_{q(\mathbf{s}\setminus s_{j,fn})} \\
&\overset{c}{=} \quad \Re(s_{j,fn})^2\left[-\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))^2 - \frac{1}{\phi\lambda_{j,fn}^2}\right] \\
&\quad + \Im(s_{j,fn})^2\left[-\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))^2 - \frac{1}{\phi\lambda_{j,fn}^2}\right] \\
&\quad - \Re(s_{j,fn})\left[-\frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\left(x_i(t) - \frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)\right] \\
&\quad - \Im(s_{j,fn})\left[\frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\left(x_i(t) - \frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)\right] \\
&\quad + \frac{4}{\phi^2}\Re(s_{j,fn})\Im(s_{j,fn})\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\Im(g_{ij,fn}(t)),
\end{aligned} \tag{5.20}$$

où $\Im(\cdot)$ représente la partie imaginaire et $\mathbf{s}\setminus s_{j,fn}$ correspond à l'ensemble $\mathbf{s}$ des variables latentes privé de $s_{j,fn}$.

On reconnaît à l'équation (5.20) le logarithme de la densité de probabilité d'une loi gaussienne (cf. annexe A) tel que

$$q_{jfn}^{\star}(s_{j,fn}) = \begin{cases} N_{\mathbb{C}}(\rho_{j,fn}, \hat{s}_{j,fn}^r, \hat{s}_{j,fn}^\imath, \gamma_{j,fn}^r, \gamma_{j,fn}^\imath) & \text{si } \mathbb{K}=\mathbb{C} \\ N_{\mathbb{R}}(\hat{s}_{j,fn}^r, \gamma_{j,fn}^r) & \text{si } \mathbb{K}=\mathbb{R} \end{cases}, \tag{5.21}$$

où

- $\hat{s}_{j,fn}^r = \langle \Re(s_{j,fn})\rangle_{q^\star}$;
- $\hat{s}_{j,fn}^\imath = \langle \Im(s_{j,fn})\rangle_{q^\star}$;
- $\gamma_{j,fn}^r = \langle (\Re(s_{j,fn}) - \hat{s}_{j,fn}^r)^2\rangle_{q^\star}$;
- $\gamma_{j,fn}^\imath = \langle (\Im(s_{j,fn}) - \hat{s}_{j,fn}^\imath)^2\rangle_{q^\star}$;

▷ $\rho_{j,fn} = \dfrac{\langle(\Re(s_{j,fn}) - \hat{s}^r_{j,fn})(\Im(s_{j,fn}) - \hat{s}^\imath_{j,fn})\rangle_{q^\star}}{\sqrt{\gamma^r_{j,fn}\gamma^\imath_{j,fn}}} \in [-1, 1]$.

Dans le cas complexe, on voit que $q^\star_{jfn}(s_{j,fn})$ correspond à la densité de probabilité de la distribution gaussienne complexe non propre. Cela signifie que les parties réelle et imaginaire sont corrélées a posteriori ($\rho_{j,fn}$ encode leur corrélation) et qu'elles ont des variances différentes. Dans le cas réel, comme $\Im(s_{j,fn}) = 0$, nous obtenons une distribution gaussienne réelle.

Les expressions des paramètres de cette distribution variationnelle s'obtiennent par identification de l'équation (5.20) avec le logarithme de la densité de probabilité de la loi gaussienne complexe définie à l'équation (A.2). Les calculs sont détaillés en annexe E. Nous obtenons les expressions suivantes :

$$\rho_{j,fn}=\frac{\dfrac{2}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\Im(g_{ij,fn}(t))}{\left(\dfrac{2}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))^2+\frac{1}{\phi\lambda^2_{j,fn}}\right)^{0.5}\left(\dfrac{2}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))^2+\frac{1}{\phi\lambda^2_{j,fn}}\right)^{0.5}}; \tag{5.22}$$

$$\gamma^r_{j,fn} = \left[2(1-\rho^2_{j,fn})\left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))^2 + \frac{1}{\phi\lambda^2_{j,fn}}\right)\right]^{-1}; \tag{5.23}$$

$$\gamma^\imath_{j,fn} = \left[2(1-\rho^2_{j,fn})\left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))^2 + \frac{1}{\phi\lambda^2_{j,fn}}\right)\right]^{-1}; \tag{5.24}$$

$$\hat{s}^r_{j,fn} = \hat{s}^r_{j,fn} - \gamma^r_{j,fn}(1-\rho^2_{j,fn})d^r_{j,fn}; \tag{5.25}$$

$$\hat{s}^\imath_{j,fn} = \hat{s}^\imath_{j,fn} - \gamma^\imath_{j,fn}(1-\rho^2_{j,fn})d^\imath_{j,fn}; \tag{5.26}$$

avec

$$d^r_{j,fn} = \frac{2}{\phi}\left[\frac{\hat{s}^r_{j,fn}}{\lambda^2_{j,fn}} - \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\left(x_i(t) - \sum_{j'=1}^{J}\hat{y}_{ij'}(t)\right)\right]; \tag{5.27}$$

$$d^\imath_{j,fn} = \frac{2}{\phi}\left[\frac{\hat{s}^\imath_{j,fn}}{\lambda^2_{j,fn}} + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\left(x_i(t) - \sum_{j'=1}^{J}\hat{y}_{ij'}(t)\right)\right]. \tag{5.28}$$

Il est important de mentionner que les mises à jour (5.25) et (5.26) ne sont valables que si elles sont effectuées séquentiellement suivant les indices j, f et n. Par ailleurs, on peut mentionner que si l'on injecte (5.23) et (5.27) dans (5.25), les coefficients $\hat{s}^r_{j,fn}$ disparaissent du membre de droite de cette dernière équation. Il en est de même pour les coefficients $\hat{s}^\imath_{j,fn}$ si l'on injecte (5.24) et (5.28) dans (5.26). Les équations (5.25) et (5.26) doivent donc être comprises comme des règles de mise à jour.

Finalement, les moments du premier et du second ordre de cette distribution variationnelle sont donnés par :

▷ Moyenne : $\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_q = \hat{s}^r_{j,fn} + \imath\hat{s}^\imath_{j,fn}$;

▷ Variance : $\gamma_{j,fn} = \langle|s_{j,fn} - \hat{s}_{j,fn}|^2\rangle_q = \gamma^r_{j,fn} + \gamma^\imath_{j,fn}$;

▷ Pseudo-variance : $\tilde{\gamma}_{j,fn} = \langle(s_{j,fn} - \hat{s}_{j,fn})^2\rangle_q = \gamma^r_{j,fn} - \gamma^\imath_{j,fn} + 2\imath\rho_{j,fn}\sqrt{\gamma^r_{j,fn}\gamma^\imath_{j,fn}}$.

d) Énergie variationnelle libre

A partir de la définition (2.56) de l'énergie variationnelle libre, de l'expression de la log-vraisemblance des données complètes à l'équation (5.19) et des résultats de l'étape E, nous avons :

$$\mathcal{L}(q^\star;\boldsymbol{\theta}) \stackrel{c}{=} -\frac{T}{2}\sum_{i=1}^{I}\ln(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\bar{e}_i - \frac{1}{\phi}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln\left(\lambda_{j,fn}^2\right) + \frac{|\hat{s}_{j,fn}|^2+\gamma_{j,fn}}{\lambda_{j,fn}^2}\right] + \frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\begin{cases}\ln\left(\gamma_{j,fn}^r\right) & \text{si } \mathbb{K}=\mathbb{R};\\ \ln\left(\gamma_{j,fn}^r\gamma_{j,fn}^\imath(1-\rho_{j,fn}^2)\right) & \text{si } \mathbb{K}=\mathbb{C},\end{cases} \tag{5.29}$$

où $\bar{e}_i = \left\langle \left\| \mathbf{x}_i - \sum_{j=1}^{J}\mathbf{y}_{ij}\right\|_2^2 \right\rangle_{q^\star}$ avec $\mathbf{x}_i = [x_i(t)]_t^\top \in \mathbb{R}^T$ et $\mathbf{y}_{ij} = [y_{ij}(t)]_t^\top \in \mathbb{R}^T$.

Soient $\mathbf{T}_{fn} \in \mathbb{C}^{T\times L_a}$ et $\hat{\mathbf{S}}_j \in \mathbb{R}^{T\times L_a}$ les matrices de Toeplitz [Golub et Van Loan, 1996] permettant d'implémenter le produit de convolution de $\psi_{fn}(t)$ et $\hat{s}_j(t)$ respectivement, $t = 0, ..., L_s$, avec un signal de longueur L_a. On définit également $\mathbf{a}_{ij} = [a_{ij}(t)]_t^\top \in \mathbb{R}^{L_a}$ tel que $\hat{\mathbf{y}}_{ij} = \hat{\mathbf{S}}_j\mathbf{a}_{ij} \in \mathbb{R}^T$. En utilisant l'approximation de champ moyen (5.15) et les résultats de l'étape E, $\bar{e}_i$ se développe de la façon suivante :

$$\bar{e}_i = \left\|\mathbf{x}_i - \sum_{j=1}^{J}\hat{\mathbf{y}}_{ij}\right\|_2^2 + \sum_{j=1}^{J}\mathbf{a}_{ij}^\top\left[\frac{2}{\phi^2}\sum_{(f,n)\in\mathcal{B}}\left(\gamma_{j,fn}\Re\left(\mathbf{T}_{fn}^H\mathbf{T}_{fn}\right) + \Re\left(\tilde{\gamma}_{j,fn}\mathbf{T}_{fn}^\top\mathbf{T}_{fn}\right)\right)\right]\mathbf{a}_{ij}. \tag{5.30}$$

e) Méthode du gradient conjugué avec préconditionnement pour l'étape E

Nous allons maintenant proposer une alternative aux règles de mise à jour séquentielles (5.25) et (5.26). Il s'agira d'avoir recours à une technique de descente par blocs de variables afin d'accélérer l'étape E de l'algorithme.

Nous pouvons montrer à partir de l'expression de l'énergie variationnelle libre (5.29) que $d_{j,fn}^{(\cdot)}$ défini à l'équation (5.27) ou (5.28) vérifie :

$$d_{j,fn}^{(\cdot)} = \frac{\partial - \mathcal{L}(q^\star;\boldsymbol{\theta})}{\partial \hat{s}_{j,fn}^{(\cdot)}}. \tag{5.31}$$

On voit ainsi que les équations (5.25) et (5.26) correspondent à une mise à jour des variables $\hat{s}_{j,fn}^{(\cdot)}$ dans la direction de l'opposée de la dérivée partielle $d_{j,fn}^{(\cdot)}$, avec un pas égal à $\gamma_{j,fn}^{(\cdot)}(1-\rho_{j,fn}^2)$. Quand la dérivée est nulle, il est clair que l'on a atteint un point stationnaire de l'algorithme car la mise à jour devient : $\hat{s}_{j,fn}^{(\cdot)} \leftarrow \hat{s}_{j,fn}^{(\cdot)}$.

Cette technique de descente par coordonnées peut s'avérer très lente quand le nombre de variables sur lesquelles on doit boucler est important, ce qui est le cas ici. Une approche possible pour accélérer l'algorithme consiste à mettre à jour simultanément un ensemble de variables en utilisant une méthode d'optimisation basée sur le calcul du gradient de la fonction de coût. Du fait de la nature gaussienne des distributions variationnelles, nous devons résoudre un problème d'optimisation convexe quadratique pour mettre à jour les variables $\{\hat{s}_{j,fn}^r, \hat{s}_{j,fn}^\imath\}$. Nous proposons alors d'utiliser la méthode du gradient conjugué avec préconditionnement [Golub et Van Loan, 1996] pour résoudre ce problème. Cette méthode est décrite à l'annexe D.

Pour des raisons de concision, nous choisissons de travailler directement avec des vecteurs à valeurs complexes, cependant une formulation alternative basée sur des matrices à valeurs réelles

est décrite à l'annexe F. Nous utilisons la dérivée généralisée aux fonctions de variables complexes définie par [Adali et al., 2011; Petersen et al., 2008] :

$$\frac{1}{2}d_{j,fn} = \frac{\partial - \mathcal{L}(q^\star;\boldsymbol{\theta})}{\partial \hat{s}^*_{j,fn}} = \frac{1}{2}\left(d^r_{j,fn} + \imath d^\imath_{j,fn}\right). \tag{5.32}$$

D'après les équations (5.27) et (5.28) on a :

$$d_{j,fn} = \frac{2}{\phi}\left[\frac{\hat{s}_{j,fn}}{\lambda^2_{j,fn}} - \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1} g^*_{ij,fn}(t)\Big(x_i(t) - \sum_{j'=1}^{J}\hat{y}_{ij'}(t)\Big)\right]. \tag{5.33}$$

Nous introduisons les définitions suivantes :

- ▷ $\hat{\mathbf{s}} = (\hat{\mathbf{s}}^r + \imath\hat{\mathbf{s}}^\imath) \in \mathbb{C}^{JFN}$ avec $\hat{\mathbf{s}}^{(\cdot)} \in \mathbb{R}^{JFN}$ le vecteur colonne d'entrées $\hat{s}^{(\cdot)}_{j,fn}$;
- ▷ $\mathbf{d} = (\mathbf{d}^r + \imath\mathbf{d}^\imath) \in \mathbb{C}^{JFN}$ avec $\mathbf{d}^{(\cdot)} \in \mathbb{R}^{JFN}$ le vecteur colonne d'entrées $d^{(\cdot)}_{j,fn}$;
- ▷ $\mathbf{g}_i(t) \in \mathbb{C}^{JFN}$ le vecteur colonne d'entrées $g_{ij,fn}(t)$;
- ▷ $\mathbf{D}$ la matrice de préconditionnement diagonale de taille $JFN \times JFN$ et d'entrées :

$$\frac{(\gamma^r_{j,fn})^{-1} + (\gamma^\imath_{j,fn})^{-1}}{1-\rho^2_{j,fn}} = \left(\frac{4}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}|g_{ij,fn}(t)|^2 + \frac{2}{\phi\lambda^2_{j,fn}}\right).$$

L'ordre des coefficients indicés par j, f et n pour construire ces vecteurs et cette matrice diagonale n'a pas d'importance, tant qu'il est gardé identique. La méthode du gradient conjugué avec préconditionnement pour la mise à jour des coefficients $\{\hat{s}_{j,fn}\}_{j,f,n}$ à l'étape E est détaillée dans l'algorithme 2.

Algorithme 2 Mise à jour de $\{\hat{s}_{j,fn}\}_{j,f,n}$ à l'étape E par la méthode du gradient conjugué avec préconditionnement

1: Initialiser $\mathbf{d}$ à partir de l'équation (5.33)
2: Initialiser $\boldsymbol{\omega} = \mathbf{D}^{-1}\mathbf{d}$
3: **Tant que** critère d'arrêt non rencontré
4: Calculer $\boldsymbol{\kappa}$ le vecteur colonne de taille JFN et d'entrées
$$\kappa_{j,fn} = \frac{2}{\phi}\left[\frac{\omega_{j,fn}}{\lambda^2_{j,fn}} + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1} g^*_{ij,fn}(t)\sum_{j'=1}^{J}\frac{2}{\phi}\Re\left(\sum_{(f',n')\in\mathcal{B}}\omega_{j',f',n'}g_{ij',f'n'}(t)\right)\right]$$
5: $\mu = (\boldsymbol{\omega}^H\mathbf{d})/(\boldsymbol{\omega}^H\boldsymbol{\kappa})$
6: $\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} - \mu\boldsymbol{\omega}$
7: Calculer $\mathbf{d}$ à partir de l'équation (5.33)
8: $\mathbf{d}_p = \mathbf{D}^{-1}\mathbf{d}$
9: $\beta = -(\boldsymbol{\kappa}^H\mathbf{d}_p)/(\boldsymbol{\omega}^H\boldsymbol{\kappa})$
10: $\boldsymbol{\omega} \leftarrow \mathbf{d}_p + \beta\boldsymbol{\omega}$
11: **Fin Tant que**

f) Étape M

L'étape M consiste à maximiser (ou seulement augmenter) $\mathcal{L}(q^\star;\boldsymbol{\theta})$ définie à l'équation (5.29) par rapport aux paramètres du modèle $\boldsymbol{\theta}$.

Variance du bruit Annuler la dérivée du critère par rapport à la variance σ_i^2 conduit à la mise à jour suivante :

$$\sigma_i^2 = \frac{1}{T}\bar{e}_i, \tag{5.34}$$

où $\bar{e}_i$ est donné à l'équation (5.30).

Paramètres NMF Comme précédemment, on reconnaît dans l'expression de l'énergie variationnelle libre la divergence d'Itakura-Saito entre la moyenne a posteriori (sous l'approximation variationnelle) du spectrogramme de puissance des sources $\langle |s_{j,fn}|^2 \rangle_{q^\star} = |\hat{s}_{j,fn}|^2 + \gamma_{j,fn}$ et $\lambda_{j,fn}^2 = (\mathbf{W}_j\mathbf{H}_j)_{f,n}$. Les matrices $\mathbf{W}_j$ et $\mathbf{H}_j$ peuvent alors être mises à jour en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{W}_j,\mathbf{H}_j \geq 0} \sum_{(f,n)\in\mathcal{B}} d_{IS}\big(|\hat{s}_{j,fn}|^2 + \gamma_{j,fn}, (\mathbf{W}_j\mathbf{H}_j)_{f,n}\big). \tag{5.35}$$

On utilise pour cela quelques itérations des règles multiplicatives (2.23) et (2.24) (dans le cas $\beta = 0$) définies à la page 32.

Filtres de mélange A partir de l'équation (5.29) on peut montrer que maximiser l'énergie variationnelle libre par rapport à $\mathbf{a}_{ij}$ est équivalent à résoudre le système d'équations suivant :

$$\left[\hat{\mathbf{S}}_j^\top\hat{\mathbf{S}}_j + \frac{2}{\phi^2}\sum_{(f,n)\in\mathcal{B}}\left(\gamma_{j,fn}\Re\left(\mathbf{T}_{fn}^H\mathbf{T}_{fn}\right) + \Re\left(\tilde{\gamma}_{j,fn}\mathbf{T}_{fn}^\top\mathbf{T}_{fn}\right)\right)\right]\mathbf{a}_{ij} = \hat{\mathbf{S}}_j^\top\left(\mathbf{x}_i - \sum_{j'\neq j}\hat{\mathbf{S}}_{j'}\mathbf{a}_{ij'}\right). \tag{5.36}$$

Soient $\epsilon_{ij}(t) = x_i(t) - \sum_{j'\neq j}\hat{y}_{ij'}(t)$, $t \in \{0, ..., T-1\}$, et $T_M\{\tau(k)\}$ une matrice de Toeplitz symétrique de taille $M \times M$ formée à partir de la suite de coefficients $\{\tau(k)\}_{k=0}^{M-1}$. On peut montrer que l'équation (5.36) se réécrit :

$$\left[T_{L_a}\left\{\hat{r}_j^{ss}(k) + \frac{2}{\phi^2}\sum_{(f,n)\in\mathcal{B}}\left(\gamma_{j,fn}\Re\left(\hat{r}_{fn}^{\psi\psi}(k)\right) + \Re\left[\tilde{\gamma}_{j,fn}\left(\hat{r}_{fn}^{\psi\psi}(k)\right)^*\right]\right)\right\}\right]\mathbf{a}_{ij} = \hat{\mathbf{r}}_{ij}^{s\epsilon}, \tag{5.37}$$

où

- $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k}\hat{s}_j(t)\hat{s}_j(t+k)$;
- $\hat{r}_{fn}^{\psi\psi}(k) = \sum_{t=0}^{L_s-1-k}\psi_{fn}^*(t)\psi_{fn}(t+k)$;
- $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1}\hat{s}_j(t)\epsilon_{ij}(t+k)$;
- $\hat{\mathbf{r}}_{ij}^{s\epsilon} = \left[\hat{r}_{ij}^{s\epsilon}(k)\right]_k^\top \in \mathbb{R}^{L_a}$.

Ce système d'équations peut être résolu par inversion matricielle. Cependant lorsque les filtres de mélange sont longs et pour des raisons de stabilité on préférera utiliser à nouveau la méthode du gradient conjugué avec préconditionnement [Golub et Van Loan, 1996]. Nous pouvons finalement mentionner que comme le montre l'équation (5.36), la mise à jour de $\mathbf{a}_{ij}$ dépend de $\{\mathbf{a}_{ij'}\}_{j'\neq j}$, il est donc nécessaire de procéder séquentiellement. En pratique une seule mise à jour de chaque filtre $\mathbf{a}_{ij}$ est effectuée par étape M, et l'ordre de mise à jour suivant les indices i et j est choisi arbitrairement et gardé fixe tout au long de l'algorithme VEM.

5.3.3 Résultats expérimentaux oracles pour la MDCT

Dans un premier temps nous souhaitons étudier la robustesse de l'approche proposée vis à vis du temps de réverbération. Nous considérons pour cela uniquement le cas d'une transformation TF par MDCT. Les expériences sont réalisées à partir de la base de données «MASS-Synthétique» (voir section 2.7 du chapitre 2) pour différents temps de réverbération : 32, 64, 128, 256 et 512 ms.

Nous comparons l'approche proposée à celle présentée dans [Ozerov et Févotte, 2010] [1], que nous avons détaillée au chapitre 4, section 4.1 (voir page 66). Les deux méthodes ont pour objectif la séparation de mélanges convolutifs multicanaux avec un modèle de source gaussien basé sur la NMF. La différence principale se situe au niveau de la représentation du processus de mélange convolutif. Dans [Ozerov et Févotte, 2010], les variables latentes TF des sources sont inférées à partir de l'observation de la TFCT du mélange. Par conséquent la convolution temporelle est approchée par une simple multiplication dans le domaine de la TFCT, sous une hypothèse de filtres de mélange à réponse impulsionnelle courte. Dans notre cas, le processus de mélange convolutif est représenté de façon exacte.

Pour les deux méthodes, nous utilisons une fenêtre d'analyse/synthèse TF sinusoïdale d'une longueur de 128 ms (2048 points à 16kHz). L'ordre de la NMF est arbitrairement fixé à 10 pour toutes les sources. Les deux algorithmes sont exécutés à partir d'une initialisation oracle des paramètres ; les paramètres de NMF sont calculés sur les sources monophoniques originales et les filtres de mélange sont initialisés à leur vraie valeur. Pour la méthode [Ozerov et Févotte, 2010], lorsque cela est nécessaire, nous tronquons les filtres dans le domaine temporel à la longueur de la fenêtre d'analyse, avant de calculer leur réponse en fréquence par TFD. Nous considérons dans un premier temps un cas oracle car nous souhaitons comparer les meilleurs performances que les deux méthodes peuvent atteindre. De plus, vérifier le bon comportement d'une méthode dans un tel scénario idéal est toujours intéressant. En effet il n'est pas certain que même initialisé de façon oracle, l'algorithme converge vers un point de l'espace des paramètres associé à de bonnes performances de séparation. Cela permet également de vérifier que l'approximation variationnelle de la loi a posteriori des variables latentes est raisonnable. Nous effectuons 100 itérations de ces algorithmes (V)EM à partir de l'initialisation oracle. Les performances de séparation sont en effet stables après 100 itérations.

Les résultats moyens de séparation évalués en terme de sources images reconstruites sont présentés dans le tableau 5.1. Nous utilisons les rapports d'énergie introduits au chapitre 2, section 2.6. Il est intéressant de remarquer que pour un temps de réverbération de 32 ms, la méthode de référence [Ozerov et Févotte, 2010] conduit à des résultats légèrement meilleurs que pour la méthode proposée. L'hypothèse de filtres de mélange est en effet ici vérifiée. Nous attribuons ce comportement au fait que la méthode de référence utilise des filtres de longueur égale à celle de la fenêtre d'analyse de la TFCT. Les filtres sont donc dans ce cas précis sur-paramétrés, ce qui peut être favorable dans le cadre d'une initialisation oracle de l'ensemble des paramètres du modèle. On voit néanmoins que pour tous les autres temps de réverbération, la méthode proposée obtient de meilleurs résultats. De plus, plus la réverbération est forte, plus l'écart de performance entre les deux méthodes se creuse ; pour $T_{60} = 64$ ms, nous améliorons le SDR de 0.7 dB tandis que pour $T_{60} = 512$ ms nous obtenons un gain de 5.5 dB. Ces résultats sont confirmés par l'écoute des exemples audio disponibles en ligne [2]. Alors que pour la méthode de référence la réverbération semble répartie dans les sources estimées, entraînant de fortes interférences pour des temps de

1. Comme indiqué au chapitre 4, section 4.1.2 (voir page 67), nous utilisons en fait l'algorithme proposé dans [Ozerov et al., 2011], qui reste néanmoins très proche de celui introduit dans [Ozerov et Févotte, 2010]. De la même façon qu'au chapitre précédent, pour des raisons de concision, nous ne ferons référence qu'à la version originale de la méthode [Ozerov et Févotte, 2010].

2. `https://perso.telecom-paristech.fr/leglaive/demo-icassp17.html`

	SDR		ISR		SIR		SAR	
Mélange convolutif	TFCT	temporel	TFCT	temporel	TFCT	temporel	TFCT	temporel
$T_{60} = 32$ ms	**16.7**	16.0	**24.3**	23.0	**24.4**	23.0	**18.6**	**18.6**
$T_{60} = 64$ ms	14.9	**15.6**	21.5	**22.6**	22.4	**22.6**	17.1	**18.5**
$T_{60} = 128$ ms	11.8	**15.3**	17.6	**22.3**	18.8	**22.6**	14.6	**18.2**
$T_{60} = 256$ ms	8.5	**13.8**	13.7	**20.5**	14.7	**21.3**	12.0	**16.7**
$T_{60} = 512$ ms	6.3	**11.8**	10.9	**18.1**	11.8	**19.2**	10.1	**14.6**

TABLEAU 5.1 – Résultats de séparation moyens en dB en fonction du temps de réverbération T_{60} pour la méthode [Ozerov et Févotte, 2010], où le processus de mélange convolutif est approché dans le domaine de la *TFCT*, et pour la méthode proposée, où celui-ci est représenté de façon exacte dans le domaine *temporel*.

réverbération élevés, la qualité de séparation obtenue avec la méthode proposée est plus constante avec l'augmentation du temps de réverbération.

Nous devons cependant mentionner que du fait de l'ajout de la dimension temporelle dans la formulation du problème de séparation de sources, notre méthode est significativement plus coûteuse en temps de calcul, celui-ci augmentant avec la longueur des filtres de mélange. Par exemple, pour un mélange de 3 sources d'une durée de 28 secondes, une itération de l'algorithme VEM dure environ 33, 37, 47, 69 et 111 secondes pour les temps de réverbération indiqués au tableau 5.1 pris en ordre croissant, alors qu'une itération de l'algorithme EM pour la méthode de référence dure environ 1 seconde. Ces données sont obtenues avec un processeur cadencé à 3.70 GHz.

5.3.4 Résultats expérimentaux semi-aveugles

Les expériences que nous décrivons ci-après sont réalisées à partir de la base de données «MASS-RWCP». Les mélanges sont donc créés à partir de réponses de salle mesurées et non plus simulées, pour un temps de réverbération de 470 ms. Nous considérons un scénario semi-aveugle en supposant la connaissance des filtres de mélange. Ces derniers sont alors fixés, c'est-à-dire qu'ils ne sont pas mis à jour au sein des algorithmes de séparation. Tous les autres paramètres du modèle sont en revanche estimés de façon aveugle. Nous utilisons la même fenêtre d'analyse (permettant une reconstruction parfaite pour tous les taux de recouvrement étudiés ci-après) et le même rang de NMF que précédemment. Les filtres de mélange étant fixés, nous évaluerons la séparation en terme de sources monophoniques reconstruites.

a) Étude de l'influence de la transformation temps-fréquence

Nous souhaitons dans un premier temps étudier l'influence du choix de la représentation TF sur les résultats de séparation, dans le cadre de l'approche proposée. Nous comparons pour cela les résultats obtenus en utilisant la MDCT, qui rappelons-le est une transformation à échantillonnage critique, avec ceux obtenus grâce à la TFCTFI qui est une transformation redondante. Plusieurs facteurs de redondance sont étudiés en faisant varier le taux de recouvrement entre trames de signal successives. Par exemple, un recouvrement de 50% conduit à un nombre de coefficients TF (un complexe équivaut à deux réels) double par rapport au nombre d'échantillons temporels. Plus le recouvrement est élevé, plus la redondance de la TFCTFI est importante. Nous effectuons 200 itérations de l'algorithme VEM. Les résultats de séparation moyens sont présentés dans la première partie du tableau 5.2 (lignes 2 à 5), on observe tout d'abord que d'après les SDR, SIR et SAR, la qualité de séparation augmente avec la redondance de la transformation TF. Le SDR dans le cas de la TFCTFI avec un recouvrement de 75% est en effet environ le double de celui

	SDR	SIR	SAR	OPS	temps de calcul
MDCT	4.8	10.9	8.2	**38.9**	132.5
TFCTFI - recouvrement 25%	6.5	12.9	9.5	35.7	375.4
TFCTFI - recouvrement 50%	7.6	14.9	10.4	35.1	538.8
TFCTFI - recouvrement 75%	**9.7**	**17.9**	**12.1**	34.1	1032.0
[Ozerov et Févotte, 2010]	-2.4	4.3	2.1	22.5	1.4
[Kowalski et al., 2010]	7.5	14.1	10.1	29.1	149.8

TABLEAU 5.2 – Résultats de séparation moyennés sur l'ensemble des sources de la base de données. SDR, SIR et SAR sont exprimés en dB, l'OPS en pourcentage et le temps de calcul en minutes.

obtenu dans le cas de la MDCT. Cependant nous estimons à l'écoute des résultats de séparation que ces conclusions ne concordent pas avec la qualité perçue (le lecteur est invité à se référer aux exemples audio disponibles en ligne [3]). C'est pourquoi nous avons également calculé l'OPS, qui est une mesure objective de la qualité globale de séparation incluant un aspect perceptif dans sa définition. Cette mesure a été présentée au chapitre 2, section 2.6. Comme on peut le voir dans le tableau 5.2, la qualité de séparation mesurée à partir de ce critère est beaucoup moins dépendante de la transformation TF utilisée. Nous obtenons même les meilleurs résultats avec la MDCT qui est à échantillonnage critique. Nous pensons que ces conclusions concordent davantage avec à la qualité de séparation perçue.

b) Comparaison avec les méthodes de référence

Nous comparons également l'approche proposée avec deux méthodes de la littérature (pour lesquelles les filtres de mélange sont également supposés connus). La première correspond de nouveau à l'approche introduite dans [Ozerov et Févotte, 2010], pour laquelle nous effectuons 200 itérations de l'algorithme EM. Nous voyons à la ligne 6 du tableau 5.2 que cette méthode obtient les moins bons résultats, du fait de l'hypothèse de filtres de mélange courts qui n'est pas respectée. Le temps de réverbération étant de 470 ms et la longueur de la fenêtre de 128 ms, les filtres sont fortement sous-paramétrés ce qui limite les performances de la méthode. De plus, cette approche est pénalisée par le fait que nous évaluions les performances à partir des sources monophoniques reconstruites. En effet nous pouvons entendre à partir des exemples audio disponibles en ligne que les sources monophoniques estimées restent réverbérées. Si nous évaluons les performances en terme de sources images reconstruites, cette méthode obtient des meilleurs résultats, ils restent cependant inférieurs à toutes les autres approches comparées.

La seconde méthode de la littérature que nous considérons dans cette expérience a été introduite dans [Kowalski et al., 2010]. Celle-ci repose également sur une représentation exacte du processus de mélange convolutif dans le domaine temporel, cependant elle utilise un modèle de source parcimonieux dans le domaine TF basé sur une régularisation ℓ_1 des coefficients de la TFCT des sources. Cette approche aboutit à un problème de type Lasso (*least absolute shrinkage and selection operator*) qui est résolu par l'algorithme FISTA (*fast iterative shrinkage-thresholding algorithm*). Comme proposé dans l'article nous effectuons 20000 itérations de cet algorithme. Les résultats obtenus par cette méthode sont présentés à la dernière ligne du tableau 5.2. D'après les SDR, SIR et SAR, nous obtenons de meilleures performances uniquement dans le cas d'une représentation TF des sources par TFCTFI avec un taux de recouvrement de 75%. Cependant, selon l'OPS, la qualité de séparation est également améliorée pour les autres représentations TF,

3. `https://perso.telecom-paristech.fr/leglaive/demo-waspaa17.html`

notamment pour la MDCT.

Pour conclure ces expériences, nous donnons dans la dernière colonne du tableau 5.2 le temps de calcul (en minutes) requis par les méthodes pour séparer un mélange de 3 sources, d'une durée de 12 secondes et échantillonné à 16 kHz. Ces données sont obtenues avec un processeur cadencé à 3.70 GHz. Il est clair que plus la représentation TF est redondante, plus le nombre de coefficients à estimer est grand et le temps de calcul élevé. Dans le cas d'une représentation par MDCT, ce dernier est du même ordre que pour la méthode [Kowalski et al., 2010]. L'approche proposée dans [Ozerov et Févotte, 2010] est de loin la plus rapide car elle ne repose pas sur une représentation temporelle du mélange.

5.3.5 Conclusion

Les expériences conduites dans cette section nous permettent de conclure qu'il est préférable d'utiliser la MDCT plutôt que la TFCTFI. La qualité de séparation perçue est en effet similaire et le temps de calcul avec la MDCT est beaucoup plus faible car c'est une transformation à échantillonnage critique. Dans la suite de cette thèse nous ne travaillerons donc plus qu'avec la MDCT. Pour de futurs travaux, il restera néanmoins à comprendre d'où provient la contradiction entre les performances indiquées par le SDR et l'OPS dans le cas de la méthode basée sur la MDCT. Une piste à explorer réside dans la différence d'encodage de la phase entre la MDCT et la TFCTFI. En effet, dans le premier cas celle-ci est encodée directement dans les amplitudes des coefficients TF, qui par l'intermédiaire du modèle de source sont représentées par une NMF.

Comparer notre méthode avec [Ozerov et Févotte, 2010] a permis de montrer l'intérêt d'utiliser une représentation exacte du processus de mélange convolutif dans le domaine temporel, malgré une surcharge en matière de temps de calcul. La comparaison avec [Kowalski et al., 2010] à quant à elle permis de montrer l'intérêt du modèle de source par NMF, par rapport à une approche basée sur la parcimonie des sources dans le plan TF, qui de fait ne permet pas de prendre en compte la dynamique des sources dans ce domaine.

5.4 Modèle de source *t* de Student

Nous présentons dans cette section une méthode de séparation similaire à celle développée dans la section précédente, dans le cas d'une représentation des sources dans le domaine de la MDCT. Nous utilisons cependant cette fois un modèle de source basé sur la distribution t de Student. Chaque source $j \in \{1, ..., J\}$ est donc représentée par un ensemble de coefficients TF de synthèse réels $\{s_{j,fn} \in \mathbb{R}\}_{f,n}$. L'approche développée ci-après a été publiée dans [Leglaive et al., 2017b].

5.4.1 Modèle de source

Chaque coefficient TF est modélisé comme une variable aléatoire suivant une loi t de Student, de paramètre de position nul et de paramètres de forme α et d'échelle $\lambda_{j,fn}$:

$$s_{j,fn} \sim \mathcal{T}_\alpha(0, \lambda_{j,fn}). \tag{5.38}$$

Les coefficients TF des sources sont toujours supposés indépendants. D'après ce modèle, nous avons $\mathbb{E}[s_{j,fn}^2] = \lambda_{j,fn}^2 \dfrac{\alpha}{\alpha - 2}$ pour $\alpha > 2$, sinon cette espérance n'est pas définie. Ce modèle t de Student admet une représentation sous forme de SMoG grâce à l'introduction d'une variable

aléatoire inverse-gamma $v_{j,fn} \in \mathbb{R}_+$:

$$\begin{cases} s_{j,fn}|v_{j,fn} & \sim \mathcal{N}_{\mathbb{R}}(0, v_{j,fn}) \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\lambda_{j,fn}^2\right) \end{cases}. \tag{5.39}$$

Dans la suite nous allons considérer deux instances particulières de ce modèle :

1. **Modèle parcimonieux** : Les coefficients TF pour une source donnée sont supposés i.i.d. de sorte que pour tous f et n :

$$\lambda_{j,fn}^2 = \lambda_j^2. \tag{5.40}$$

   Pour un paramètre de forme α petit, c'est-à-dire pour une distribution *t* de Student à queue suffisamment lourde, cette approche suppose la parcimonie des coefficients MDCT des sources. Ce modèle a déjà été utilisé dans [Févotte et Godsill, 2006; Cemgil et al., 2007] pour la séparation de mélanges instantanés et non convolutifs. La flexibilité de la distribution *t* de Student due à la lourdeur de sa queue permet ici de réduire considérablement le nombre de paramètres impliqués dans la modélisation des sources (un seul paramètre d'échelle par source), ce qui peut être intéressant dans un contexte sous-déterminé pour la séparation de sources.

2. **Modèle NMF** : On considère une distribution non-stationnaire des points TF où les paramètres d'échelle sont structurés par l'intermédiaire d'un modèle NMF :

$$\lambda_{j,fn}^2 = (\mathbf{W}_j\mathbf{H}_j)_{f,n}, \tag{5.41}$$

   avec $\mathbf{W}_j = [w_{j,fk}]_{f,k} \in \mathbb{R}_+^{F\times K_j}$ and $\mathbf{H}_j = [h_{j,kn}]_{k,n} \in \mathbb{R}_+^{K_j\times N}$. Quand α tend vers l'infini, la distribution *t* de Student tend vers la gaussienne et nous retrouvons le modèle de source utilisé à la section précédente (dans le cas de la MDCT).

   Ce modèle NMF basé sur la distribution *t* de Student a été récemment introduit dans [Yoshii et al., 2016] pour la séparation monocanale de sources de rang 1, sans avoir recours cependant à une formulation par SMoG (voir chapitre 2, section 2.3.4b). Au même moment, un modèle de source hiérarchique également basé les distributions gaussienne et inverse-gamma a été proposé dans [Kounades-Bastian et al., 2016]. Dans cet article, chaque source est représentée dans le domaine de la TFCT comme la somme de composantes latentes gaussiennes. Les variances TF de chacune de ces composantes sont représentées comme des variables aléatoires latentes inverse-gamma dont les paramètres d'échelle sont contraints par une NMF de rang 1. Bien que cela ne soit pas mentionné explicitement, ce modèle correspond bien à une distribution marginale pour les composantes de rang 1 qui soit *t* de Student.

5.4.2 Inférence variationnelle

Comme précédemment $\mathbf{x} = \{x_i(t)\}_{i,t}$ représente l'ensemble des observations temporelles du mélange. Nous notons $\mathbf{z} = \{\mathbf{s} = \{s_{j,fn}\}_{j,f,n}, \mathbf{v} = \{v_{j,fn}\}_{j,f,n}\}$ l'ensemble des variables latentes de ce modèle et $\boldsymbol{\theta} = \{\{\lambda_{j,fn}^2\}_{j,f,n}, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\}$ l'ensemble des paramètres. Comme auparavant nous considérons un cas semi-aveugle en supposant la connaissance des filtres de mélange.

Dans le cas du modèle gaussien introduit à la section précédente, nous avions recours à une méthode d'inférence variationnelle pour des raisons pratiques (voir Section 5.3.2, page 88) ; afin de forcer la matrice de covariance a posteriori des coefficients MDCT des sources à être diagonale. Dans le cas du modèle *t* de Student considéré ici, la distribution a posteriori des variables latentes ne peut être calculée explicitement, c'est la raison pour laquelle nous considérons également une approche d'inférence variationnelle.

a) Approximation de champ moyen

Nous cherchons une densité de probabilité $q \in \mathcal{F}$ sur les variables latentes $\mathbf{z}$ qui maximise l'énergie variationelle libre. A nouveau nous formulons l'hypothèse de champ moyen qui consiste à supposer que la famille variationnelle $\mathcal{F}$ correspond à l'ensemble des densités de probabilité qui se factorisent sous la forme suivante :

$$q(\mathbf{s}) = \prod_{j=1}^{J} \prod_{f=0}^{F-1} \prod_{n=0}^{N-1} q^s_{jfn}(s_{j,fn}) q^v_{jfn}(v_{j,fn}). \tag{5.42}$$

Afin de simplifier les notations on note simplement $q(\cdot)$ les distributions $q^{s/v}_{jfn}(\cdot)$. L'estimateur des coefficients TF des sources sous cette approximation est le même que dans le cas gaussien à l'équation (5.16), il s'agit de la moyenne a posteriori sous l'hypothèse de champ moyen $\hat{s}_{j,fn} = \langle s_{j,fn} \rangle_q$. La source monophonique $\hat{s}_j(t)$ est comme précédemment reconstruite par MDCT inverse et la source image $\hat{y}_{ij}(t)$ est obtenue par convolution avec le filtre de mélange associé $a_{ij}(t)$.

b) Vraisemblance des données complètes

A partir des équations (5.2) et (5.39), la log-vraisemblance des données complètes $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta}) + \ln p(\mathbf{s}|\mathbf{v}) + \ln p(\mathbf{v}; \boldsymbol{\theta})$ s'écrit :

$$\begin{aligned}
\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \stackrel{c}{=} & -\frac{1}{2} \sum_{i=1}^{I} \sum_{t=0}^{T-1} \left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2} \left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t) \right)^2 \right] \\
& - \sum_{j=1}^{J} \sum_{(f,n) \in \mathcal{B}} \left[\ln(v_{j,fn}) \left(\frac{\alpha}{2} + 1 + \frac{1}{2} \right) + \ln \Gamma \left(\frac{\alpha}{2} \right) \right. \\
& \left. + \frac{1}{v_{j,fn}} \left(\frac{\alpha}{2} \lambda_{j,fn}^2 + \frac{s_{j,fn}^2}{2} \right) + \frac{\alpha}{2} \ln \left(\frac{2}{\alpha \lambda_{j,fn}^2} \right) \right].
\end{aligned} \tag{5.43}$$

c) Étape E

Étape E-V D'après l'approximation de champ-moyen et l'équation (5.43), on peut montrer que les densités de probabilité $q(v_{j,fn})$ qui maximisent l'énergie variationnelle libre satisfont :

$$\begin{aligned}
\ln q^\star(v_{j,fn}) & \stackrel{c}{=} \langle \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \rangle_{q(\mathbf{z} \setminus v_{j,fn})} \\
& \stackrel{c}{=} -\ln(v_{j,fn}) \left(\frac{\alpha + 1}{2} + 1 \right) - \frac{1}{v_{j,fn}} \left(\frac{\alpha}{2} \lambda_{j,fn}^2 + \frac{\langle s_{j,fn}^2 \rangle_{q(s_{j,fn})}}{2} \right).
\end{aligned} \tag{5.44}$$

On reconnaît alors $q^\star(v_{j,fn}) \propto IG(\delta, \beta_{j,fn})$ avec

$$\delta = \frac{\alpha + 1}{2}; \tag{5.45}$$

$$\beta_{j,fn} = \frac{\alpha}{2} \lambda_{j,fn}^2 + \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{2}, \tag{5.46}$$

où $\gamma_{j,fn} = \langle (s_{j,fn} - \hat{s}_{j,fn})^2 \rangle_{q(s_{j,fn})}$.

Étape E-S En utilisant le fait que $\left\langle v_{j,fn}^{-1} \right\rangle_{q(v_{j,fn})} = \dfrac{\delta}{\beta_{j,fn}}$, on trouve de la même façon :

$$\begin{aligned}
\ln q^\star(s_{j,fn}) &\stackrel{c}{=} \langle \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \rangle_{q(\mathbf{z} \setminus s_{j,fn})} \\
&\stackrel{c}{=} -\frac{1}{2} \Bigg(\sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}^2(t) + \frac{\delta}{\beta_{j,fn}} \Bigg) \\
&\times \left[s_{j,fn} - \frac{\sum\limits_{i=1}^{I} \frac{1}{\sigma_i^2} \sum\limits_{t=0}^{T-1} g_{ij,fn}(t) \Bigg(x_i(t) - \sum\limits_{(j',f',n') \neq (j,f,n)} \hat{s}_{j',f'n'} g_{ij',f'n'}(t) \Bigg)}{\sum\limits_{i=1}^{I} \frac{1}{\sigma_i^2} \sum\limits_{t=0}^{T-1} g_{ij,fn}^2(t) + \frac{\delta}{\beta_{j,fn}}} \right]^2 .
\end{aligned} \tag{5.47}$$

On reconnaît $q_{jfn}^\star(s_{j,fn}) = N(\hat{s}_{j,fn}, \gamma_{j,fn})$ avec :

$$\gamma_{j,fn} = \Bigg(\sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}^2(t) + \frac{\delta}{\beta_{j,fn}} \Bigg)^{-1}, \tag{5.48}$$

et après réécriture du terme de moyenne [4] :

$$\hat{s}_{j,fn} = \hat{s}_{j,fn} - \gamma_{j,fn} d_{j,fn}, \tag{5.49}$$

où

$$d_{j,fn} = \hat{s}_{j,fn} \frac{\delta}{\beta_{j,fn}} - \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}(t) \Bigg(x_i(t) - \sum_{j'=1}^{J} \hat{y}_{ij'}(t) \Bigg). \tag{5.50}$$

De la même façon qu'à la section 5.3, on peut montrer que $d_{j,fn} = \partial(-\mathcal{L}(q^\star; \boldsymbol{\theta}))/(\partial \hat{s}_{j,fn})$ où $\mathcal{L}(q^\star; \boldsymbol{\theta})$ est défini ci-dessous à l'équation (5.51). La mise à jour (5.49) permet donc de minimiser l'opposée de l'énergie variationnelle libre suivant une technique de descente par coordonnées. Afin d'accélérer l'étape E-S, on choisit comme précédemment d'utiliser la méthode du gradient conjugué avec préconditionnement pour résoudre ce problème d'optimisation.

d) Énergie variationnelle libre

L'énergie variationnelle libre s'écrit :

$$\begin{aligned}
\mathcal{L}(q^\star; \boldsymbol{\theta}) \stackrel{c}{=} &- JFN \left[\ln \Gamma \left(\frac{\alpha}{2} \right) - \left(\frac{\alpha + 1}{2} - \delta \right) \mathrm{di}\Gamma(\delta) - \delta - \ln \Gamma(\delta) + \frac{\alpha}{2} \ln \left(\frac{2}{\alpha} \right) \right] \\
&- \frac{T}{2} \sum_{i=1}^{I} \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \bar{e}_i \\
&- \sum_{j=1}^{J} \sum_{(f,n) \in \mathcal{B}} \left[\frac{1}{2} \ln \left(\frac{\beta_{j,fn}}{\gamma_{j,fn}} \right) + \frac{\alpha}{2} \ln \left(\frac{\beta_{j,fn}}{\lambda_{j,fn}^2} \right) + \frac{\alpha \delta}{2} \frac{\lambda_{j,fn}^2}{\beta_{j,fn}} + \frac{\delta}{2} \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\beta_{j,fn}} \right],
\end{aligned} \tag{5.51}$$

4. Comme précédemment l'équation (5.49) correspond à une règle de mise à jour, on peut en effet montrer en injectant (5.48) et (5.50) dans cette équation que la variable $\hat{s}_{j,fn}$ disparaît du membre de droite.

où comme à l'équation (5.30) :

$$\bar{e}_i = \left\langle \left\| \mathbf{x}_i - \sum\nolimits_{j=1}^{J} \mathbf{y}_{ij} \right\|_2^2 \right\rangle_{q^\star} = \left\| \mathbf{x}_i - \sum\nolimits_{j=1}^{J} \hat{\mathbf{y}}_{ij} \right\|_2^2 + \sum_{j=1}^{J} \sum_{(f,n)\in\mathcal{B}} \gamma_{j,fn} \left\| \mathbf{g}_{ij,fn} \right\|_2^2, \quad (5.52)$$

avec $\mathbf{g}_{ij,fn} = [g_{ij,fn}(t)]_t^\top \in \mathbb{R}^T$. Cette équation correspond à la simplification de (5.30) lorsque les variables latentes de source sont réelles (cas de la MDCT).

e) Étape M

Variance du bruit Annuler la dérivée de l'énergie variationnelle libre par rapport à la variance σ_i^2 conduit à la même mise à jour que pour le modèle gaussien donnée à l'équation (5.34).

Paramètres du modèle de source parcimonieux Dans le cas du modèle de source parcimonieux défini à l'équation (5.40), on obtient la règle de mise à jour suivante par annulation de la dérivée partielle du critère :

$$\lambda_j^2 = \left(\frac{1}{FN} \sum_{(f,n)\in\mathcal{B}} \left(\frac{\beta_{j,fn}}{\delta} \right)^{-1} \right)^{-1}. \quad (5.53)$$

Il est intéressant de remarquer qu'il s'agit d'une moyenne harmonique des coefficients $\{\beta_{j,fn}/\delta\}_{f,n}$.

Paramètres du modèle de source NMF Lorsque les paramètres d'échelle $\lambda_{j,fn}$ sont structurés par un modèle NMF comme à l'équation (5.41), on peut montrer que maximiser $\mathcal{L}(q^\star;\boldsymbol{\theta})$ par rapport à $\mathbf{W}_j, \mathbf{H}_j$ sous une contrainte de positivité des coefficients de ces matrices est équivalent à minimiser la fonction de coût suivante sous la même contrainte :

$$\begin{aligned} \mathcal{C}(\mathbf{W}_j, \mathbf{H}_j) &= \sum_{(f,n)\in\mathcal{B}} \left[\frac{(\mathbf{W}_j\mathbf{H}_j)_{f,n}}{\beta_{j,fn}/\delta} - \ln\left((\mathbf{W}_j\mathbf{H}_j)_{f,n} \right) \right] \\ &\overset{c}{=} \sum_{(f,n)\in\mathcal{B}} d_{IS}\left((\mathbf{W}_j\mathbf{H}_j)_{f,n}, \frac{\beta_{j,fn}}{\delta} \right). \end{aligned} \quad (5.54)$$

On fait de nouveau apparaître la divergence d'Itakura-Saito dans le problème d'estimation des paramètres NMF. Cependant ces derniers apparaissent cette fois dans le premier argument de cette divergence qui n'est pas symétrique. On ne peut donc avoir recours aux règles multiplicatives standard. On peut toutefois utiliser la méthode de la fonction auxiliaire (voir chapitre 2, section 2.3.3) pour résoudre ce problème d'optimisation.

Comme $u \mapsto -\ln(u)$ est une fonction convexe pour $u \in \mathbb{R}_+$, on utilise l'inégalité de Jensen pour obtenir une majorante de $\mathcal{C}(\mathbf{W}_j, \mathbf{H}_j)$. Pour tout ensemble de coefficients $\{c_{jk,fn} \in [0,1]\}_{k=1}^{K_j}$ tels que $\sum_{k=1}^{K_j} c_{jk,fn} = 1$:

$$-\ln\left((\mathbf{W}_j\mathbf{H}_j)_{f,n} \right) \leq -\sum_{k=1}^{K_j} c_{jk,fn} \ln\left(\frac{w_{j,fk}h_{j,kn}}{c_{jk,fn}} \right), \quad (5.55)$$

avec égalité si et seulement si

$$c_{jk,fn} = \frac{w_{j,fk}h_{j,kn}}{(\mathbf{W}_j\mathbf{H}_j)_{f,n}}. \quad (5.56)$$

En injectant (5.55) dans (5.54) on obtient une majorante du critère :

$$\mathcal{C}(\mathbf{W}_j, \mathbf{H}_j) \leq \sum_{(f,n)\in\mathcal{B}} \left[\frac{(\mathbf{W}_j\mathbf{H}_j)_{f,n}}{\beta_{j,fn}/\delta} - \sum_{k=1}^{K_j} c_{jk,fn} \ln\left(\frac{w_{j,fk}h_{j,kn}}{c_{jk,fn}}\right) \right]. \tag{5.57}$$

En annulant les dérivées partielles de cette majorante par rapport à $w_{j,fk}$ et $h_{j,kn}$, et en remplaçant $c_{jk,fn}$ par l'expression (5.56) qui minimise la majorante, on obtient les règles de mise à jour multiplicatives suivantes :

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \odot \frac{(\mathbf{W}_j\mathbf{H}_j)^{\odot -1}\,\mathbf{H}_j^\top}{(\mathbf{B}_j/\delta)^{\odot -1}\,\mathbf{H}_j^\top}; \tag{5.58}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \odot \frac{\mathbf{W}_j^\top\,(\mathbf{W}_j\mathbf{H}_j)^{\odot -1}}{\mathbf{W}_j^\top\,(\mathbf{B}_j/\delta)^{\odot -1}}, \tag{5.59}$$

où $\mathbf{B}_j = [\beta_{j,fn}]_{f,n} \in \mathbb{R}_+^{F\times N}$. La contrainte de positivité est satisfaite si les paramètres NMF sont initialisés avec des valeurs positives. On peut finalement effectuer au sein de l'étape M quelques itérations de ces règles multiplicatives.

Filtre de mélange Comme pour le modèle gaussien (voir équation (5.37), page 94), on peut montrer que maximiser l'énergie variationnelle libre par rapport aux filtres de mélange est équivalent à résoudre le système d'équations suivant :

$$\left[T_{L_a}\{\hat{r}_j^{ss}(k)\} + T_{L_a}\left\{ \sum_{(f,n)\in\mathcal{B}} \gamma_{j,fn}\hat{r}_{fn}^{\psi\psi}(k) \right\} \right] \mathbf{a}_{ij} = \hat{\mathbf{r}}_{ij}^{s\epsilon}. \tag{5.60}$$

Cette équation correspond à la simplification de (5.37) lorsque les variables latentes de source sont réelles. On rappelle que la mise à jour de $\mathbf{a}_{ij}$ dépend de $\{\mathbf{a}_{ij'}\}_{j'\neq j}$ au travers ici du terme $\hat{\mathbf{r}}_{ij}^{s\epsilon}$. Il est donc nécessaire de procéder séquentiellement. En pratique, comme précédemment, une seule mise à jour de chaque filtre $\mathbf{a}_{ij}$ est effectuée par étape M, et l'ordre de mise à jour suivant les indices i et j est choisi arbitrairement et gardé fixe tout au long de l'algorithme VEM.

5.4.3 Résultats expérimentaux semi-aveugles

Les expériences sont réalisées à partir de la base de données «MASS-Synthétique» pour un temps de réverbération de 256 ms. On utilise toujours une fenêtre d'analyse/synthèse TF sinusoïdale d'une longueur de 128 ms. Le rang de factorisation de la NMF est arbitrairement fixé à 10 pour toutes les sources. On considère un scénario semi-aveugle où les filtres de mélange sont supposés connus et fixés à leur valeur oracle tandis que tous les autres paramètres sont estimés de façon aveugle. Les performances de séparation sont évaluées sur les sources monophoniques reconstruites. Des exemples audio sont disponibles en ligne [5].

a) Influence du modèle de source

Nous comparons tout d'abord les résultats de séparation selon le choix du modèle de source, c'est-à-dire basé sur la parcimonie ou sur la NMF. Le SDR moyenné sur l'ensemble des sources de la base de données est représenté sur la figure 5.2 en fonction du paramètre de forme de la distribution *t* de Student.

5. `https://perso.telecom-paristech.fr/leglaive/demo-eusipco17.html`

On voit premièrement que le modèle de source parcimonieux requiert un paramètre de forme significativement plus petit que pour le modèle NMF. Les résultats chutent fortement lorsque la valeur de ce paramètre augmente, ce qui montre la nécessité d'utiliser une distribution à queue lourde si l'on suppose des coefficients TF i.i.d.

Deuxièmement, il est clair que le modèle NMF obtient des performances supérieures au modèle parcimonieux. Cela montre encore une fois l'avantage d'exploiter la dynamique des sources dans le plan TF et non simplement de supposer leur parcimonie.

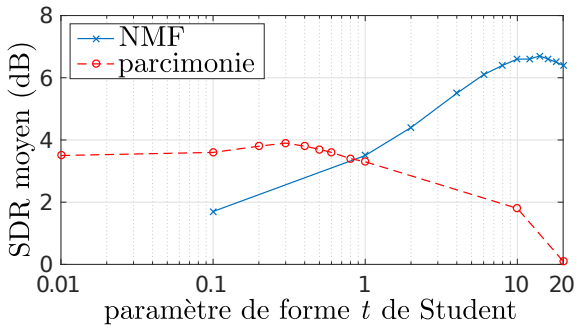


FIGURE 5.2 – SDR moyen en fonction du paramètre de forme du modèle de source t de Student. La courbe bleue correspond au modèle par NMF et celle en rouge au modèle basé sur la parcimonie.

b) Comparaison avec les méthodes de référence

On compare maintenant les résultats de séparation obtenus avec la méthode proposée et avec les méthodes de référence [Ozerov et Févotte, 2010; Kowalski et al., 2010] déjà présentées dans le cadre de l'évaluation du modèle gaussien à la section 5.3. Rappelons que ces méthodes travaillent dans le domaine de la TFCT. Nous incluons également dans cette comparaison la méthode [Leglaive et al., 2017a] qui correspond au modèle gaussien introduit en section 5.3, dans le cas de la MDCT. Les paramètres de forme de la distribution t de Student pour la méthode proposée sont choisis comme ceux maximisant le SDR moyen à la figure 5.2.

Les résultats sont présentés dans le tableau 5.3. Encore une fois, les méthodes basées sur une représentation exacte du processus de mélange convolutif dans le domaine temporel obtiennent les meilleures performances. Le SIR indique que les deux approches basées sur la parcimonie des coefficients TF des sources ([Kowalski et al., 2010] utilisant une pénalisation ℓ_1, équivalent d'un a priori de Laplace, et celle proposée utilisant la distribution t de Student) conduisent à des performances similaires en terme d'interférences. En revanche la méthode proposée dans [Kowalski et al., 2010] induit moins d'artéfacts, c'est pourquoi le SAR et le SDR moyens sont supérieurs. Finalement, le modèle de source t de Student proposé, avec paramétrisation NMF, et son équivalent gaussien [Leglaive et al., 2017a], obtiennent les meilleurs résultats en terme de SDR. Cependant, d'après l'OPS l'approche basée sur la distribution gaussienne permet d'obtenir une meilleure qualité de séparation.

	SDR	SIR	SAR	OPS
[Ozerov et Févotte, 2010]	1.7	8.5	4.9	25.1
[Kowalski et al., 2010]	5.5	11.7	8.8	29.4
[Leglaive et al., 2017a]	**6.7**	12.5	9.5	**42.0**
Modèle *t* de Student basé parcimonie ($\alpha = 0.4$)	3.8	11.8	5.9	28.7
Modèle *t* de Student basé NMF ($\alpha = 14$)	**6.7**	**12.7**	**10.0**	33.9

TABLEAU 5.3 – Résultats de séparation moyennés sur l'ensemble des sources de la base de données. SDR, SIR et SAR sont exprimés en dB et l'OPS en pourcentage.

5.5 Conclusion

Dans ce chapitre nous avons introduit un nouveau cadre probabiliste pour la séparation des mélanges convolutifs, basé sur l'inférence des coefficients TF des sources à partir des observations temporelles du mélange.

Nous avons tout d'abord proposé un modèle de source gaussien basé sur la NMF. Les expériences réalisées dans un contexte oracle et semi-aveugle ont permis de montrer la robustesse de cette approche pour la séparation des mélanges en présence de forte réverbération.

Les résultats expérimentaux ont également permis de montrer que la MDCT et la TFCTFI conduisent à des performances de séparation similaires d'après des mesures objectives prenant en compte un aspect perceptif. Cependant, la MDCT étant à échantillonnage critique elle induit un temps de calcul significativement plus faible que la TFCTFI.

Nous avons finalement adapté la méthode proposée à deux autres modèles de source basés sur la distribution *t* de Student. Le premier repose sur une hypothèse de parcimonie des coefficients TF des sources tandis que le second fait intervenir une paramétrisation NMF. Nous avons montré expérimentalement que l'approche par NMF permet d'obtenir de meilleures performances. Néanmoins, les résultats fournis par l'OPS semblent indiquer que la distribution gaussienne pour le modèle de source basé sur la NMF est mieux adaptée que la distribution *t* de Student.

Dans ce chapitre, les expériences ont été effectuées dans un contexte oracle ou semi-aveugle (à filtres de mélange connus). Nous allons voir au chapitre suivant qu'il est également possible d'estimer les filtres, mais cela nécessite la prise en compte d'un a priori probabiliste. Nous verrons que sans contraintes, il n'est pas possible d'obtenir une estimation des filtres de mélange qui conduise à des performances de séparation satisfaisantes.

Chapitre 6

Modèle *t* de Student pour les filtres de mélange

Sommaire

6.1 Modèle 108
6.1.1 Modèle de source 108
6.1.2 Modèle de réponse impulsionnelle de salle 109
6.2 Inférence variationnelle 112
6.2.1 Approximation de champ moyen 112
6.2.2 Modèle conjugué-exponentiel 113
6.2.3 Énergie variationnelle libre 114
6.2.4 Étape E 115
6.2.5 Étape M 119
6.3 Résultats expérimentaux 120
6.3.1 Méthodes de référence 120
6.3.2 Cadre expérimental 120
6.3.3 Initialisation des paramètres 121
6.3.4 Hyperparamètres du modèle 122
6.3.5 Comparaison avec les méthodes de référence 123
6.4 Conclusion 127

Utiliser une représentation temporelle du mélange pour la séparation de sources permet non seulement de représenter de façon exacte le processus de mélange convolutif mais également de développer des a priori simples pour les filtres de mélange. En effet ces derniers correspondent à des réponses impulsionnelles de salle, ils ont donc une structure temporelle bien précise qu'il est possible d'exploiter afin de guider leur estimation. Dans un contexte sous-déterminé, être en mesure d'exploiter de l'information supplémentaire sur les variables d'intérêt est un aspect très important que l'on propose de considérer ici dans un cadre bayésien. Nous pouvons mentionner que certains travaux en estimation simultanée de multiples RIRs ont déjà montré l'intérêt d'exploiter cette structure temporelle, dans un contexte non-aveugle (connaissant les signaux sources) et sous-déterminé (plus de coefficients pour les RIRs que d'échantillons pour le mélange) [Benichoux et al., 2014] [Giri, 2016, Ch. 5]. Dans la partie II du manuscrit, nous cherchions à transcrire la dynamique temporelle des filtres sous forme de corrélations fréquentielles. Nous allons voir dans cette partie qu'il est plus aisé de travailler directement avec la réponse impulsionnelle, et non plus fréquentielle.

Les méthodes de la littérature basées sur une représentation du mélange convolutif dans le domaine temporel telles que [Kowalski et al., 2010; Arberet et Vandergheynst, 2014] et [Feng et Kowalski, 2014], ainsi que celles développées au chapitre précédent, supposent la connaissance des vrais filtres de mélange. Cette hypothèse est fortement restrictive car dans une application réelle, il est difficile d'avoir cette connaissance. Cela suppose en effet de connaître l'environnement acoustique où a eu lieu l'enregistrement ainsi que d'avoir mesuré les réponses de salle entre chaque source et chaque microphone.

La contribution principale de ce chapitre est de proposer un nouveau cadre bayésien où les filtres de mélange sont traités comme des variables aléatoires latentes dans le domaine temporel, suivant une distribution *t* de Student. Le modèle proposé généralise le modèle gaussien [Polack, 1988], basé sur la décroissance exponentielle de la réverbération tardive et largement utilisé en acoustique statistique des salles. Nous formulons également le modèle de source afin de pouvoir adapter la représentation TF à la nature de chaque source présente dans le mélange, notamment en matière de résolution temps/fréquence. Enfin, les expériences sont réalisées dans un scénario plus réaliste, car nous ne supposons pas les filtres de mélange connus.

Le modèle complet est présenté à la section 6.1. Nous détaillons la méthode d'inférence variationnelle basée sur un algorithme VEM à la section 6.2. Enfin nous présentons les résultats expérimentaux dans la section 6.3 et concluons à la section 6.4. L'article de revue [Leglaive et al., 2017d] en cours de révision au moment de l'écriture de ce manuscrit est basé sur les résultats présentés dans ce chapitre.

6.1 Modèle

Nous considérons toujours le modèle de mélange convolutif bruité introduit à l'équation (5.2), page 85.

6.1.1 Modèle de source

a) Représentation temps-fréquence

Comme précédemment, les signaux sources sont représentés par un ensemble de coefficients TF de synthèse $\{s_{j,fn} \in \mathbb{R}\}_{(f,n)\in\mathcal{B}_j}$ avec $\mathcal{B}_j = \{0, ..., F_j - 1\} \times \{0, ..., N_j - 1\}$:

$$s_j(t) = \sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\psi_{j,fn}(t). \tag{6.1}$$

Nous choisissons de travailler dans le domaine de la MDCT, l'atome TF de synthèse est donc défini par :

$$\psi_{j,fn}(t) = \sqrt{\frac{2}{F_j}} w_j(t - nH_j) \cos\left(\frac{2\pi}{M_j}\left(t - nH_j + \frac{1}{2} + \frac{M_j}{4}\right)\left(f + \frac{1}{2}\right)\right), \tag{6.2}$$

où $w_j(t)$ est une fenêtre sinusoïdale de longueur M_j, $H_j = M_j/2$ est la taille de l'incrément et $F_j = M_j/2$. Par rapport à l'équation (5.4), page 85, on remarque cette fois que l'atome MDCT dépend de l'indice de source j au travers du paramètre de longueur M_j de la fenêtre. Cette spécificité permet d'adapter la résolution TF pour chaque source composant le mélange. On peut par exemple choisir une fenêtre courte pour des signaux percussifs afin d'avoir une bonne résolution temporelle, et une fenêtre plus longue pour des sources tonales (présentant une hauteur) afin d'avoir une bonne résolution fréquentielle.

On rappelle que d'après cette décomposition, une source image se réécrit :

$$y_{ij}(t) = [a_{ij}(\cdot) \star s_j(\cdot)](t) = \sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\, g_{ij,fn}(t), \tag{6.3}$$

avec $g_{ij,fn}(t) = [a_{ij}(\cdot) \star \psi_{j,fn}(\cdot)](t)$.

b) A priori sur les sources

Comme au chapitre 5, section 5.4, on suppose que chaque coefficient $s_{j,fn}$ suit une loi *t* de Student :

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}(0, \lambda_{j,fn}). \tag{6.4}$$

Nous utilisons cependant ici une formulation sous forme de SMoG différente, bien qu'équivalente :

$$\begin{cases} s_{j,fn}|v_{j,fn} & \sim \mathcal{N}(0, v_{j,fn}\lambda_{j,fn}^2) \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha_v}{2}, \frac{\alpha_v}{2}\right) \end{cases}. \tag{6.5}$$

On suppose de plus que la variable $\lambda_{j,fn}^2$ suit un modèle NMF comme à l'équation (5.41), page 99 : $\lambda_{j,fn}^2 = (\mathbf{W}_j\mathbf{H}_j)_{f,n}$.

Contrairement au modèle hiérarchique de l'équation (5.39), page 99, la factorisation NMF se trouve désormais reliée à la variance de la distribution gaussienne sur $s_{j,fn}$ et non plus au paramètre d'échelle de la distribution inverse-gamma sur $v_{j,fn}$. Nous comprendrons l'intérêt de cette reformulation à l'étape M de l'algorithme VEM développé dans ce chapitre. Il est important néanmoins de mentionner que ces deux modèles sont équivalents du fait de la propriété suivante de la distribution inverse-gamma : Si $z \sim \mathcal{IG}(\alpha, \beta)$, alors pour tout facteur $C \in \mathbb{R}_+$, $Cz \sim \mathcal{IG}(\alpha, C\beta)$.

6.1.2 Modèle de réponse impulsionnelle de salle

a) Modèle de RIR

On représente en haut de la figure 6.1 une RIR provenant de la base de données MIRD [Hadad et al., 2014]. Celle-ci a été enregistrée dans une salle avec un temps de réverbération de 610 ms et avec une distance source-microphone d'environ 2 m. L'énergie de cette RIR est représentée en dB sur la figure du milieu. On voit que celle-ci décroît exponentiellement au cours du temps (linéairement sur une échelle en dB). En acoustique statistique des salles, une RIR est généralement représentée par un processus aléatoire centré $a(t) \in \mathbb{R}$ indexé par le temps $t \in \mathbb{N}$. Comme

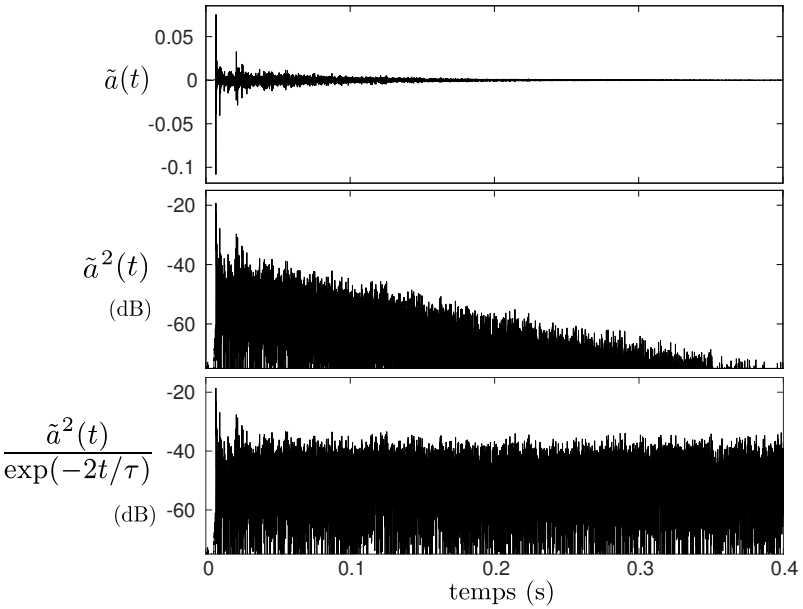


FIGURE 6.1 – Amplitude (figure du haut), énergie (figure du milieu) et énergie normalisée (figure du bas) d'une RIR de la base de données MIRD [Hadad et al., 2014]. Le temps de réverbération est de 610 ms et la distance source-microphone d'environ 2 m.

précédemment on suppose une RIR de longueur finie L_a. $\tilde{a}(t)$ à la figure 6.1 correspond à une réalisation de ce processus. La propriété de décroissance exponentielle de l'énergie est alors liée à l'expression de la variance $\mathbb{E}[a^2(t)]$, en supposant que celle-ci soit définie [Schroeder, 1987, 1962]. On rappelle que différentes réalisations de ce processus correspondent à différentes positions de la source et du microphone, l'espérance mathématique peut donc être comprise comme une moyenne spatiale.

Modèle gaussien avec décroissance exponentielle C'est Moorer qui pour la première fois évoque la ressemblance entre une RIR de salle de concert et un bruit blanc gaussien modulé par une enveloppe à décroissance exponentielle [Moorer, 1979]. Dans [Polack, 1988], Polack introduit un modèle formalisant cette observation et qui a depuis été largement utilisé en acoustique statistique des salles. Une RIR est représentée comme un processus gaussien centré non-stationnaire :

$$a(t) \sim \mathcal{N}(0, r^2(t)) \tag{6.6}$$

où

$$r^2(t) = \sigma_r^2 \exp(-2t/\tau), \qquad \tau = \frac{T_{60} f_s}{3\ln(10)}. \tag{6.7}$$

σ_r^2 est un facteur d'échelle global lié à l'énergie totale de la réverbération, τ est le paramètre de décroissance exponentiel déjà introduit au chapitre 3, équation (3.14), page 57. Sous une hypothèse de champ diffus, σ_r^2 ne dépend pas de la position de la source et du microphone dans la salle car l'énergie sonore est supposée identiquement distribuée dans toute la salle, suivant toutes les directions. Le choix de la distribution gaussienne à l'équation (6.6) provient naturellement du modèle de RFR introduit par Schroeder [Schroeder et Kuttruff, 1962; Schroeder, 1962] et de la linéarité de la transformée de Fourier. En effet dans ce modèle, la RFR est représentée par un processus aléatoire dont les parties réelle et imaginaire sont des processus gaussiens centrés, indépendants et de même variance.

Expérience préliminaire L'étude proposée par Polack dans [Polack, 1993] conclut qu'une RIR ne peut être représentée par un processus stochastique qu'après la frontière définie par le temps de mélange, c'est-à-dire quand la densité temporelle d'échos est suffisamment grande. Le modèle gaussien (6.6) n'est par conséquent valide que pour la partie diffuse de la RIR, c'est-à-dire pour la réverbération tardive.

On représente en bas de la figure 6.1 l'énergie normalisée $\tilde{a}^2(t)/\exp(-2t/\tau)$ de la RIR afin de compenser la décroissance exponentielle. Le facteur de décroissance τ est fixé d'après l'équation (6.7), en utilisant la connaissance du temps de réverbération T_{60}. On observe effectivement sur cette figure que l'énergie normalisée est presque constante au cours du temps. Il y a cependant de fortes déviations au début de la RIR qui correspondent au trajet direct et aux premiers échos.

D'après le modèle gaussien (6.6), les coefficients de la RIR normalisée $a(t)/\exp(-t/\tau)$ sont censés être i.i.d. suivant la distribution $\mathcal{N}(0, \sigma_r^2)$. Nous proposons ici de vérifier cette hypothèse en utilisant les RIRs de la base de données MIRD [Hadad et al., 2014]. Celles-ci sont enregistrées dans une salle avec un temps de réverbération de 610 ms en utilisant 3 réseaux différents de 8 microphones pour 26 positions de la source. On obtient ainsi un ensemble de 624 RIRs à partir desquelles on peut calculer la densité de probabilité empirique (un histogramme normalisé) représentée en trait plein noir sur la figure 6.2. On estime également à partir de ces données la distribution gaussienne la plus proche au sens du maximum de vraisemblance. La densité de probabilité associée est représentée sur la même figure en bleu ($\sigma_r = 0.006$). On voit clairement que la distribution empirique présente une queue plus lourde que la gaussienne. Cela est dû au trajet direct et aux premiers échos des RIR, qui d'un point de vue statistique correspondent à des valeurs aberrantes par rapport au modèle gaussien (6.6).

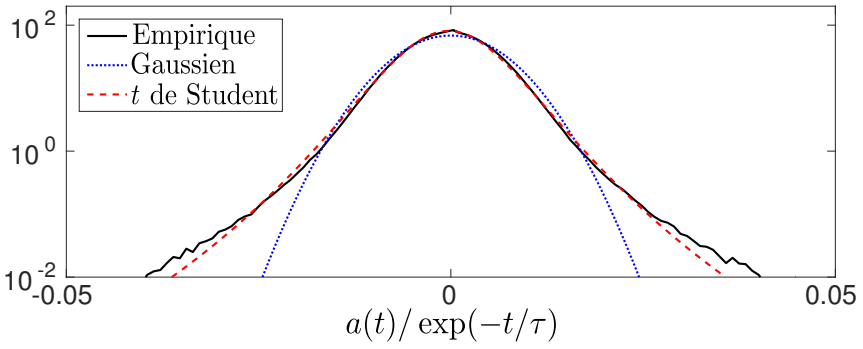


FIGURE 6.2 – Densité de probabilité empirique (trait plein noir) et densités de probabilité des distributions gaussienne (trait bleu en pointillé) et *t* de Student (trait rouge tireté) estimées à partir de 624 RIRs normalisées.

Modèle *t* de Student avec décroissance exponentielle C'est précisément pour cette raison que nous proposons d'utiliser la distribution *t* de Student. Sa robustesse vis à vis des valeurs aberrantes permet de prendre en compte indirectement le trajet direct et les premiers échos dans le modèle. Afin d'illustrer cela, nous représentons également en trait rouge sur la figure 6.2 la densité de probabilité associée à la distribution *t* de Student estimée à partir des 624 RIR normalisées. L'estimation au sens du maximum de vraisemblance conduit à $\sigma_r = 0.005$ et $\alpha_u = 7.2$. Il est clair que la distribution *t* de Student constitue un meilleur choix pour modéliser les RIRs. Elle permet de compenser l'hypothèse de champ diffus formulée pour écrire le modèle (6.6), qui n'est en général pas vérifiée.

A priori sur les filtres de mélange Suivant les résultats de cette expérience préliminaire, on suppose que chaque coefficient $a_{ij}(t)$ du filtre de mélange associé à la source j et au microphone i suit indépendamment une loi *t* de Student :

$$a_{ij}(t) \sim \mathcal{T}_{\alpha_u}(0, r(t)), \tag{6.8}$$

où $r(t)$ représente la décroissance exponentielle comme introduit à l'équation (6.7). Ce modèle admet celui de Polack à l'équation (6.6) comme cas particulier lorsque α_u tend vers l'infini. De la même façon que pour le modèle de source, il peut être écrit sous forme de SMoG par l'intermédiaire d'une variable aléatoire inverse-gamma $u_{ij}(t)$:

$$\begin{cases} a_{ij}(t)|u_{ij}(t) & \sim \mathcal{N}\left(0, u_{ij}(t) r^2(t)\right) ; \\ u_{ij}(t) & \sim \mathcal{IG}\left(\frac{\alpha_u}{2}, \frac{\alpha_u}{2}\right). \end{cases} \tag{6.9}$$

Ce modèle hiérarchique peut s'interpréter de la façon suivante : Une RIR $a_{ij}(t)$ est composée d'un champ sonore diffus (uniformément réparti dans la salle) dont l'énergie est définie par σ_r^2. Au niveau du trajet direct et des premiers échos, la RIR peut dévier fortement de ce comportement global. Ce sont précisément les variables inverse-gamma $u_{ij}(t)$ qui permettent de prendre en compte ces fortes déviations, qui elles dépendent de l'indice de source et de microphone, car étant reliées à leurs positions respectives.

Le modèle *t* de Student (6.8) permet grâce à l'introduction d'un unique paramètre α_u de prendre en compte implicitement le trajet direct et les premiers échos. Bien sûr nous supposons ici que les premières contributions et la réverbération tardive sont identiquement distribuées, ce qui est discutable. Néanmoins, comme nous allons le voir ce modèle simple permet d'avoir de bons résultats de séparation de sources, et comme nous l'évoquerons en conclusion de ce chapitre, considérer un modèle plus complexe, où l'on distingue la modélisation des premiers échos et celle de la réverbération tardive, ne nous a pas permis d'améliorer les performances par rapport à l'utilisation du modèle (6.8).

6.2 Inférence variationnelle

On définit les ensembles de variables latentes suivants : $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$; $\mathbf{v} = \{v_{j,fn}\}_{j,f,n}$; $\mathbf{a} = \{a_{ij}(t)\}_{i,j,t}$; $\mathbf{u} = \{u_{ij}(t)\}_{i,j,t}$ et $\mathbf{z} = \{\mathbf{s}, \mathbf{v}, \mathbf{a}, \mathbf{u}\}$. Soit $\mathbf{x} = \{x_i(t)\}_{i,t}$ l'ensemble des variables observées et $\boldsymbol{\theta} = \{\boldsymbol{\lambda} = \{\lambda^2_{j,fn}\}_{j,f,n}, \boldsymbol{\sigma} = \{\sigma_i^2\}_i\}$ l'ensemble des paramètres déterministes du modèle. Nous omettons les autres paramètres que l'on suppose fixés. Le modèle complet est représenté sous forme de réseau bayésien à la figure 6.3.

L'inférence exacte de la distribution a posteriori des variables latentes est ici analytiquement impossible. C'est pourquoi nous avons de nouveau recours à une approche variationnelle, que nous détaillons dans cette section.

6.2.1 Approximation de champ moyen

Nous utilisons à nouveau l'approximation de champ moyen qui consiste à supposer que la densité de probabilité q se factorise de la façon suivante :

$$q(\mathbf{z}) = \prod_{j=1}^{J} \left[\prod_{(f,n)\in\mathcal{B}_j} q^s_{jfn}(s_{j,fn}) q^v_{jfn}(v_{j,fn}) \right] \left[\prod_{i=1}^{I} \prod_{t=0}^{L_a-1} q^a_{ijt}(a_{ij}(t)) q^u_{ijt}(u_{ij}(t)) \right]. \tag{6.10}$$

Afin de simplifier les notations nous noterons simplement $q(\cdot)$ les densités introduites à l'équation (6.10), sans indices ni exposants.

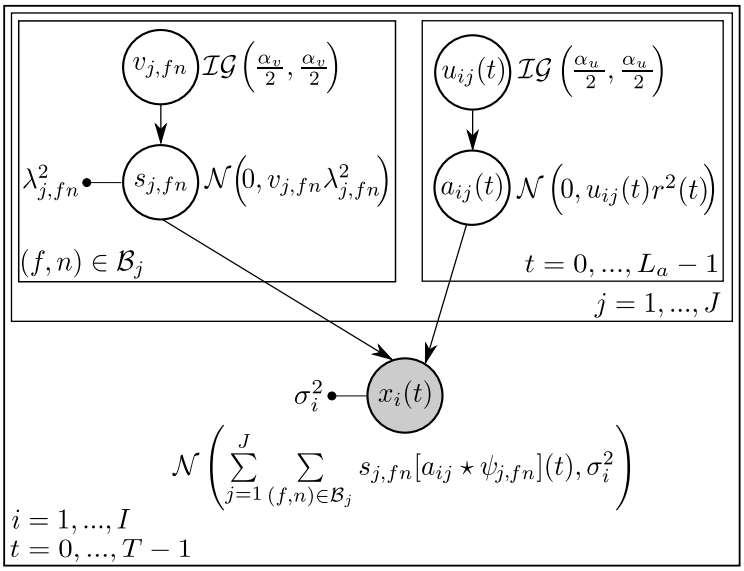


FIGURE 6.3 – Réseau bayésien correspondant au modèle proposé. Les variables aléatoires latentes sont représentées par des cercles vides, les observations par des cercles grisés, et les paramètres du modèle par des points. Chaque sous-graphe contenu dans un rectangle est répété suivant les indices indiqués. Chaque arête traversant le côté d'un rectangle est également répliquée.

On rappelle que sous cette approximation, l'estimation de la j-ème source au point TF (f, n) est donnée par $\hat{s}_{j,fn} = \langle s_{j,fn} \rangle_{q(s_{j,fn})}$. Le signal temporel $\hat{s}_j(t)$ associé est reconstruit par MDCT inverse :

$$\hat{s}_j(t) = \sum_{(f,n)\in\mathcal{B}_j} \hat{s}_{j,fn} \psi_{j,fn}(t). \tag{6.11}$$

De la même façon, l'estimation du filtre de mélange est donnée par la moyenne a posteriori sous l'approximation variationnelle : $\hat{a}_{ij}(t) = \langle a_{ij}(t) \rangle_{q(a_{ij}(t))}$. On définit également la variable $\hat{g}_{ij,fn}(t) = [\hat{a}_{ij} \star \psi_{j,fn}](t)$. Enfin, l'estimée de la j-ème source image au niveau du microphone i est donnée par :

$$\begin{aligned} \hat{y}_{ij}(t) =& [\hat{a}_{ij}(\cdot) \star \hat{s}_j(\cdot)](t) \\ =& \sum_{(f,n)\in\mathcal{B}_j} \hat{s}_{j,fn} \hat{g}_{ij,fn}(t). \end{aligned} \tag{6.12}$$

6.2.2 Modèle conjugué-exponentiel

Le modèle défini précédemment et représenté sous forme de réseau bayésien à la figure 6.3 est dit «conjugué-exponentiel» [Winn et Bishop, 2005]. En effet la distribution de chaque variable latente conditionnellement à ses parents appartient à la famille exponentielle, et les distributions de ses parents sont conjuguées par rapport à elle. Par conséquent on peut montrer que sous l'approximation de champ moyen, la distribution optimale $q^\star(z), z \in \mathbf{z}$, qui maximise l'énergie variationnelle libre sera de la même forme que la distribution a priori de z, conditionnellement à ses parents [Winn et Bishop, 2005]. Dans notre cas cela correspond aux distributions variationnelles suivantes :

$$q^\star(v_{j,fn}) = IG(v_{j,fn};\, \nu_v, \beta_{j,fn}); \tag{6.13}$$

$$q^\star(s_{j,fn}) = N(s_{j,fn};\, \hat{s}_{j,fn}, \gamma_{j,fn}); \tag{6.14}$$

$$q^\star(u_{ij}(t)) = IG(u_{ij}(t);\, \nu_u, d_{ij}(t)); \tag{6.15}$$

$$q^\star(a_{ij}(t)) = N\left(a_{ij}(t);\, \hat{a}_{ij}(t), \rho_{ij}(t)\right). \tag{6.16}$$

Il est intéressant de remarquer que le fait de travailler avec un modèle «conjugué-exponentiel» nous permet directement d'écrire l'énergie variationnelle libre, sans même connaître les expressions des paramètres de ces distributions.

6.2.3 Énergie variationnelle libre

L'énergie variationnelle libre peut être décomposée de la façon suivante d'après sa définition :

$$\begin{aligned}
\mathcal{L}(q^\star, \boldsymbol{\theta}) &= \left\langle \ln\left(\frac{p(\mathbf{x}, \mathbf{s}, \mathbf{v}, \mathbf{a}, \mathbf{u}; \boldsymbol{\theta})}{q^\star(\mathbf{s})q^\star(\mathbf{v})q^\star(\mathbf{a})q^\star(\mathbf{u})}\right)\right\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{v})q^\star(\mathbf{a})q^\star(\mathbf{u})} \\
&= \langle \ln p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \boldsymbol{\sigma})\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{a})} \\
&\quad + \langle \ln p(\mathbf{s}|\mathbf{v}; \boldsymbol{\lambda}) - \ln q^\star(\mathbf{s})\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{v})} \\
&\quad + \langle \ln p(\mathbf{v}) - \ln q^\star(\mathbf{v})\rangle_{q^\star(\mathbf{v})} \\
&\quad + \langle \ln p(\mathbf{a}|\mathbf{u}) - \ln q^\star(\mathbf{a})\rangle_{q^\star(\mathbf{a})q^\star(\mathbf{u})} \\
&\quad + \langle \ln p(\mathbf{u}) - \ln q^\star(\mathbf{u})\rangle_{q^\star(\mathbf{u})},
\end{aligned} \tag{6.17}$$

où $\langle -\ln q(\cdot)\rangle$ est l'entropie différentielle de la distribution q. Dans ce chapitre, $p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \boldsymbol{\sigma})$ est appelé vraisemblance tandis que $p(\mathbf{x}; \boldsymbol{\theta})$ correspond à la vraisemblance marginale.

Nous allons détailler ci-dessous chacun des termes du membre de droite de l'équation (6.17), à partir du modèle présenté à la section 6.1 et des distributions variationnelles données par les équations (6.13) à (6.16).

Terme de vraisemblance

$$\langle \ln p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \boldsymbol{\sigma})\rangle = -\frac{IT}{2}\ln(2\pi) - \frac{T}{2}\sum_{i=1}^{I}\ln(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\bar{e}_i, \tag{6.18}$$

où $\bar{e}_i = \left\langle \left\| \mathbf{x}_i - \sum_{j=1}^{J}\mathbf{y}_{ij} \right\|_2^2 \right\rangle$ se développe de la façon suivante :

$$\bar{e}_i = \left\| \mathbf{x}_i - \sum_{j=1}^{J}\hat{\mathbf{y}}_{ij} \right\|_2^2 + \sum_{j=1}^{J}\left[\|\hat{\mathbf{s}}_j\|_2^2 \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau) \right] + \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\left[\|\hat{\mathbf{g}}_{ij,fn}\|_2^2 + \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau) \right], \tag{6.19}$$

avec $\hat{\mathbf{g}}_{ij,fn} = [\hat{g}_{ij,fn}(t)]_t^\top \in \mathbb{R}^T$, $\hat{\mathbf{s}}_j = [\hat{s}_j(t)]^\top \in \mathbb{R}^{L_s}$ et $\hat{\mathbf{y}}_{ij} = [\hat{y}_{ij}(t)]^\top \in \mathbb{R}^T$. Les variables $\hat{s}_j(t)$ et $\hat{y}_{ij}(t)$ sont liées aux paramètres variationnels $\{\hat{s}_{j,fn}\}_{j,f,n}$ et $\{\hat{a}_{ij}(t)\}_{i,j,t}$ par les équations (6.11) et (6.12) respectivement. Les détails de calcul de ce terme $\bar{e}_i$ sont fournis dans l'annexe G, section G.3.

Terme S

$$\begin{aligned}
\langle \ln p(\mathbf{s}|\mathbf{v}; \boldsymbol{\lambda}) - \ln q^\star(\mathbf{s})\rangle = -\frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\Bigg[&\ln\left(\frac{\beta_{j,fn}}{\gamma_{j,fn}}\right) \\
&- \mathrm{di}\Gamma(\nu_v) + \ln(\lambda_{j,fn}^2) + \frac{\nu_v}{\beta_{j,fn}}\frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\lambda_{j,fn}^2} - 1\Bigg].
\end{aligned} \tag{6.20}$$

Terme V

$$\langle \ln p(\mathbf{v}) - \ln q^\star(\mathbf{v}) \rangle = -\sum_{j=1}^{J} \sum_{(f,n)\in\mathcal{B}_j} \left[\frac{\alpha_v \nu_v}{2} \frac{1}{\beta_{j,fn}} + \frac{\alpha_v}{2} \ln(\beta_{j,fn}) + \ln\Gamma\left(\frac{\alpha_v}{2}\right) + \frac{\alpha_v}{2} \ln\left(\frac{2}{\alpha_v}\right) + \mathrm{di}\Gamma(\nu_v)\left(\nu_v - \frac{\alpha_v}{2}\right) - \nu_v - \ln\Gamma(\nu_v) \right]. \tag{6.21}$$

Terme A

$$\langle \ln p(\mathbf{a}|\mathbf{u}) - \ln q^\star(\mathbf{a}) \rangle = \frac{IJL_a}{2} - \frac{IJ}{2} \sum_{t=0}^{L_a-1} \left[\ln(r^2(t)) - \mathrm{di}\Gamma\left(\nu_u\right) \right] - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=0}^{L_a-1} \left[\ln\left(\frac{d_{ij}(t)}{\rho_{ij}(t)}\right) + \frac{\nu_u}{d_{ij}(t)} \frac{\hat{a}_{ij}^2(t) + \rho_{ij}(t)}{r^2(t)} \right]. \tag{6.22}$$

Terme U

$$\langle \ln p(\mathbf{u}) - \ln q^\star(\mathbf{u}) \rangle = -\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=0}^{L_a-1} \left[\frac{\alpha_u \nu_u}{2} \frac{1}{d_{ij}(t)} + \frac{\alpha_u}{2} \ln(d_{ij}(t)) + \ln\Gamma\left(\frac{\alpha_u}{2}\right) + \frac{\alpha_u}{2} \ln\left(\frac{2}{\alpha_u}\right) + \mathrm{di}\Gamma(\nu_u)\left(\nu_u - \frac{\alpha_u}{2}\right) - \nu_u - \ln\Gamma(\nu_u) \right]. \tag{6.23}$$

6.2.4 Étape E

L'étape E de l'algorithme VEM consiste alors à maximiser l'énergie variationnelle libre par rapport aux paramètres variationnels, c'est-à-dire aux paramètres des distributions (6.13) à (6.16). De nombreuses méthodes d'optimisation sont envisageables pour résoudre ce problème.

a) Approche basée sur l'équation (6.24)

Nous avons vu au chapitre précédent qu'il était notamment possible d'utiliser une approche de type descente par coordonnées, en cyclant sur chaque paramètre variationnel scalaire. Dans ce cas nous utilisions le fait que la distribution optimale $q^\star(z)$, $z \in \mathbf{z}$, doit vérifier :

$$\ln q^\star(z) \overset{c}{=} \langle \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \rangle_{q(\mathbf{z}\setminus z)}. \tag{6.24}$$

L'étape E consistait à développer le membre de droite de l'équation (6.24) pour chaque variable latente, afin d'identifier la forme de la distribution variationnelle optimale ainsi que ses paramètres. Nous formulions ensuite certaines des mises à jour pour faire apparaître explicitement une approche de type descente par coordonnées. Nous détaillons dans l'annexe G l'ensemble des calculs permettant de résoudre l'étape E suivant cette procédure pour le modèle étudié dans ce chapitre.

Cette technique fournie un ensemble de solutions couplées ; on peut en effet montrer que la distribution $q^\star(z)$ dépend des paramètres des distributions variationnelles associées aux variables de la couverture de Markov (*Markov blanket* en anglais) de z, c'est-à-dire ses parents, enfants et co-parents dans la représentation du modèle par réseau bayésien [Winn et Bishop, 2005]. Du fait de ces potentiels nombreux couplages, cette approche peut s'avérer coûteuse. Nous avons alors proposé précédemment d'avoir recours à une méthode d'optimisation de type descente (ou montée) de gradient afin de mettre à jour directement des groupes de variables.

b) Approche par maximisation directe de l'énergie variationnelle libre

Cependant il est également possible d'exploiter le fait que nous ayons un modèle «conjugué-exponentiel» et donc que nous disposions déjà de l'expression de l'énergie variationnelle libre [1] à l'équation (6.17). Dans ce cas les mises à jour des paramètres variationnels peuvent s'obtenir en maximisant directement cette fonction objectif par des techniques d'optimisation standard. Lorsque la forme du critère le permet, on peut par exemple chercher à annuler ses dérivées partielles. De plus, du fait de la nature gaussienne des distributions variationnelles (6.14) et (6.16), nous devons résoudre individuellement deux problèmes d'optimisation quadratiques pour mettre à jour les paramètres variationnels $\{\hat{s}_{j,fn}\}_{j,f,n}$ et $\{\hat{a}_{ij}(t)\}_{i,j,t}$ (voir l'expression de l'énergie variationnelle libre à l'équation (6.17)). C'est pourquoi à nouveau nous allons utiliser la méthode du gradient conjugué avec préconditionnement.

Cette approche par maximisation directe de l'énergie variationnelle libre nécessite certes que certaines hypothèses sur le modèle soient vérifiées, mais elle permet d'utiliser des méthodes d'optimisation performantes et adaptées à la résolution de problèmes en grande dimension. On peut par exemple citer [Hoffman et al., 2013] où une approche d'optimisation stochastique est utilisée, dans le sens où le gradient est estimé à partir d'un sous-ensemble des données. Cette approche est donc adaptée lorsque l'énergie variationnelle libre s'écrit comme une somme de termes car les données sont i.i.d. Le gradient naturel peut également être utilisé afin de tenir compte de la géométrie du problème d'optimisation [Honkela et al., 2008, 2010]. Il a également été proposé récemment [Kingma, 2017] de paramétrer la distribution variationnelle par l'intermédiaire d'un réseau de neurones. Cette approche permet de tirer profit des techniques de rétropropagation utilisées pour l'entraînement des réseaux de neurones afin de calculer de façon efficace les gradients. L'algorithme résultant est alors adapté à la résolution de problèmes d'inférence en grande dimension. Nous pouvons également mentionner que de nouvelles méthodes d'inférence variationnelle ont récemment été proposées pour résoudre efficacement des problèmes en grande dimension. Ces méthodes sont basées sur la résolution directe d'un problème d'optimisation fonctionnelle, par des méthodes de type descente de gradient [Fraysse et Rodet, 2014] (adaptées à la nature du problème), ou par l'utilisation de méthodes par sous-espaces [Zheng et al., 2015], le principe de ces dernières étant de chercher la direction de descente dans un sous-espaces plutôt que de la fixer de façon unique.

Concernant le problème d'inférence variationnelle posé ici, nous pensons que pour dériver l'étape E de l'algorithme VEM il est intéressant d'exploiter à la fois l'approche par coordonnées (basée sur l'équation (6.24)) et celle par optimisation directe de l'énergie variationnelle libre. Cela permet en effet de vérifier les calculs, en comparant les règles des mise à jour obtenues dans les deux cas. Les résultats présentés ci-après ont été obtenus grâce aux calculs détaillés dans l'annexe G. Cependant, les gradients du critère par rapport aux paramètres $\{\hat{s}_{j,fn}\}_{j,f,n}$ et $\{\hat{a}_{ij}(t)\}_{i,j,t}$ peuvent de façon équivalente être directement calculés à partir de l'expression de l'énergie variationnelle libre à l'équation (6.17).

Étape E-V En utilisant l'équation (6.24) et l'expression de la log-vraisemblance des données complètes fournie à l'équation (G.1), on peut montrer que les paramètres de la densité de probabilité $q^\star(v_{j,fn})$ à l'équation (6.13) ont pour expression :

$$\nu_v = \frac{\alpha_v + 1}{2}; \tag{6.25}$$

1. Nous aurions également pu utiliser cette approche précédemment dans la thèse car les modèles du chapitre 5 sont également «conjugués-exponentiels».

$$\beta_{j,fn} = \frac{\alpha_v}{2} + \frac{\langle s_{j,fn}^2 \rangle_{q(s_{j,fn})}}{2\lambda_{j,fn}^2}, \tag{6.26}$$

où $\langle s_{j,fn}^2 \rangle_{q(s_{j,fn})} = \hat{s}_{j,fn}^2 + \gamma_{j,fn}$.

Étape E-U De la même façon on obtient les paramètres de la densité de probabilité $q^\star(u_{ij}(t))$ à l'équation (6.15) :

$$\nu_u = \frac{\alpha_u + 1}{2}; \tag{6.27}$$

$$d_{ij}(t) = \frac{\alpha_u}{2} + \frac{\langle a_{ij}^2(t) \rangle_{q(a_{ij}(t))}}{2r^2(t)}, \tag{6.28}$$

où $\langle a_{ij}^2(t) \rangle_{q(a_{ij}(t))} = \hat{a}_{ij}^2(t) + \rho_{ij}(t)$. Nous pouvons mentionner que les équations (6.26) et (6.28) s'obtiennent également par annulation de la dérivée du critère. En revanche, du fait des fonctions gamma et digamma impliquées dans l'expression de l'énergie variationnelle libre, la seule façon simple (à notre connaissance) d'obtenir les expressions de ν_v et ν_u est de s'appuyer sur l'équation (6.24) (voir annexe G).

Étape E-S En annulant la dérivée partielle de $\mathcal{L}(q^\star, \boldsymbol{\theta})$ par rapport à la variance $\gamma_{j,fn}$ de la distribution variationnelle sur $s_{j,fn}$ on obtient :

$$\gamma_{j,fn} = \left[\frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \left(\sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) + \|\hat{\mathbf{g}}_{ij,fn}\|_2^2 \right) \right]^{-1}, \tag{6.29}$$

où l'on a utilisé le fait que $\left\langle v_{j,fn}^{-1} \right\rangle = \nu_v / \beta_{j,fn}$.

Passons maintenant à la mise à jour des paramètres de moyenne $\{\hat{s}_{j,fn}\}_{j,f,n}$. Soient $\hat{\mathbf{s}} \in \mathbb{R}^{JFN}$ le vecteur colonne d'entrées $\hat{s}_{j,fn}$ et $\Delta\hat{\mathbf{s}} \in \mathbb{R}^{JFN}$ le gradient de l'énergie variationnelle libre par rapport au vecteur $\hat{\mathbf{s}}$. Soit $\hat{\mathbf{G}}_i \in \mathbb{R}^{T \times JFN}$ la matrice formée par concaténation des vecteurs colonnes $\hat{\mathbf{g}}_{ij,fn} \in \mathbb{R}^T$. On peut montrer que le gradient s'écrit de la façon suivante :

$$\Delta\hat{\mathbf{s}} = \boldsymbol{\Lambda}_{\hat{\mathbf{s}}}\hat{\mathbf{s}} - \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \hat{\mathbf{G}}_i^\top \mathbf{x}_i, \tag{6.30}$$

où étant donnée une correspondance arbitraire [2] entre le triplet (j, f, n) et $b \in \{1, ..., B\}$ avec $B = \sum_{j=1}^{J} F_j N_j$, $\boldsymbol{\Lambda}_{\hat{\mathbf{s}}} \in \mathbb{R}^{JFN \times JFN}$ s'écrit :

$$\boldsymbol{\Lambda}_{\hat{\mathbf{s}}} = \operatorname{diag}\left(\left\{ \frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \right\}_b \right) + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \hat{\mathbf{G}}_i^\top \hat{\mathbf{G}}_i. \tag{6.31}$$

D'après l'équation (6.30) il est clair qu'annuler le gradient $\Delta\hat{\mathbf{s}}$ par rapport à $\hat{\mathbf{s}}$ revient à résoudre un système d'équations linéaires, on utilisera donc la méthode du gradient conjugué avec préconditionnement. La matrice de préconditionnement correspond à la partie diagonale de $\boldsymbol{\Lambda}_{\hat{\mathbf{s}}}$. On peut alors montrer que :

$$(\boldsymbol{\Lambda}_{\hat{\mathbf{s}}})_{b,b} = \gamma_{j,fn}^{-1}. \tag{6.32}$$

2. Cette correspondance doit cependant être cohérente avec la construction des vecteurs $\hat{\mathbf{s}}$ et $\Delta\hat{\mathbf{s}} \in \mathbb{R}^{JFN}$ et de la matrice $\hat{\mathbf{G}}_i \in \mathbb{R}^{T \times JFN}$.

Chaque entrée du gradient sera par conséquent multipliée par la variance $\gamma_{j,fn}$. Cette mise à l'échelle permet de prendre en compte le fait que modifier la moyenne d'une distribution gaussienne de faible variance a un bien plus fort impact (pouvant être mesuré en terme de divergence de Kullback-Leibler) que lorsque la variance est grande. Par ailleurs $\gamma_{j,fn}^{-1}$ est égal à l'information de Fisher $\mathcal{I}(\hat{s}_{j,fn})$ définie par :

$$\mathcal{I}(\hat{s}_{j,fn}) = \left\langle -\frac{\partial^2}{\partial \hat{s}_{j,fn}^2} \ln q^\star(s_{j,fn}) \right\rangle_{q^\star(s_{j,fn})} = \gamma_{j,fn}^{-1}. \tag{6.33}$$

L'information de Fisher est en effet utile pour caractériser la géométrie d'un problème d'optimisation lié à des distributions de probabilité [Honkela et al., 2008, 2010] et elle intervient dans le calcul du gradient naturel.

Étape E-A Cette étape est très similaire à l'étape E-S. En annulant la dérivée partielle de $\mathcal{L}(q^\star, \boldsymbol{\theta})$ par rapport à la variance $\rho_{ij}(t)$ de la distribution variationnelle sur $a_{ij}(t)$ on obtient :

$$\rho_{ij}(t) = \left[\frac{\nu_u}{d_{ij}(t) r^2(t)} + \frac{1}{\sigma_i^2} \left(\|\hat{\mathbf{s}}_j\|_2^2 + \sum_{(f,n)\in\mathcal{B}_j} \gamma_{j,fn} \right) \right]^{-1}, \tag{6.34}$$

où l'on a utilisé le fait que $\left\langle u_{ij}^{-1}(t) \right\rangle = \nu_u / d_{ij}(t)$.

Soient $\hat{\mathbf{a}}_{ij} = [\hat{a}_{ij}(t)]_t^\top \in \mathbb{R}^{L_a}$ et $\Delta\hat{\mathbf{a}}_{ij} \in \mathbb{R}^{L_a}$ le gradient de l'énergie variationnelle libre par rapport au vecteur $\hat{\mathbf{a}}_{ij}$. Nous rappelons également les définitions suivantes déjà introduites au chapitre 5 :

- ▷ $T_{L_a}\{c(k)\}$: Matrice de Toeplitz symétrique construite à partir de la suite $\{c(k)\}_{k=0}^{L_a-1}$;
- ▷ $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k} \hat{s}_j(t)\hat{s}_j(t+k)$;
- ▷ $\hat{r}_{j,fn}^{\psi\psi}(k) = \sum_{t=0}^{L_s-1-k} \psi_{j,fn}(t)\psi_{j,fn}(t+k)$;
- ▷ $\epsilon_{ij}(t) = x_i(t) - \sum_{j'\neq j} \hat{y}_{ij'}(t)$;
- ▷ $\hat{\mathbf{r}}_{ij}^{s\epsilon} = \left[\hat{r}_{ij}^{s\epsilon}(k)\right]_k^\top \in \mathbb{R}^{L_a}$ avec $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1} \hat{s}_j(t)\epsilon_{ij}(t+k)$.

On peut montrer que le gradient s'écrit :

$$\Delta\hat{\mathbf{a}}_{ij} = \boldsymbol{\Lambda}_{\hat{\mathbf{a}}_{ij}} \hat{\mathbf{a}}_{ij} - \frac{1}{\sigma_i^2} \hat{\mathbf{r}}_{ij}^{s\epsilon}, \tag{6.35}$$

où la matrice $\boldsymbol{\Lambda}_{\hat{\mathbf{a}}_{ij}} \in \mathbb{R}^{L_a \times L_a}$ est définie par :

$$\boldsymbol{\Lambda}_{\hat{\mathbf{a}}_{ij}} = \operatorname{diag}\left(\left\{ \frac{\nu_u}{d_{ij}(t) r^2(t)} \right\}_t \right) + \frac{1}{\sigma_i^2} T_{L_a} \left\{ \hat{r}_j^{ss}(k) + \sum_{(f,n)\in\mathcal{B}_j} \gamma_{j,fn} \hat{r}_{j,fn}^{\psi\psi}(k) \right\}. \tag{6.36}$$

Comme précédemment, annuler le gradient $\Delta\hat{\mathbf{a}}_{ij}$ par rapport à $\hat{\mathbf{a}}_{ij}$ correspond à résoudre un système d'équations linéaires. Nous utilisons pour cela la méthode du gradient conjugué avec préconditionnement où la matrice de préconditionnement correspond à la partie diagonale de $\boldsymbol{\Lambda}_{\hat{\mathbf{a}}_{ij}}$. Il est immédiat de vérifier que :

$$[\boldsymbol{\Lambda}_{\hat{\mathbf{a}}_{ij}}]_{t,t} = \rho_{ij}(t)^{-1}, \tag{6.37}$$

car $\hat{r}_{j,fn}^{\psi\psi}(0) = 1$, les atomes de MDCT étant orthonormés, et $\hat{r}_j^{ss}(0) = \|\hat{\mathbf{s}}_j\|_2^2$. Comme précédemment, la mise à jour de $\hat{\mathbf{a}}_{ij}$ dépend de $\{\hat{\mathbf{a}}_{ij'}\}_{j'\neq j}$, il est donc nécessaire de procéder séquentiellement.

6.2.5 Étape M

a) Variance du bruit

En pratique nous ferons décroître la variance du bruit σ_i^2 progressivement au cours des itérations de l'algorithme VEM. Ce paramètre permet d'ajuster les contributions relatives des a priori et de la vraisemblance dans l'énergie variationnelle libre. Une variance élevée permet de favoriser les a priori, ce qui peut être utile durant les premières itérations de l'algorithme. A l'inverse, faire décroître la variance progressivement permet d'accroître la contribution des données observées dans l'estimation des paramètres.

Néanmoins, une approche alternative pour mettre à jour ce paramètre consiste à annuler la dérivée de $\mathcal{L}(q^\star, \boldsymbol{\theta})$ par rapport à σ_i^2, on obtient ainsi comme auparavant :

$$\sigma_i^2 = \frac{1}{T}\bar{e}_i, \tag{6.38}$$

avec $\bar{e}_i$ défini à l'équation (6.19).

b) Paramètres de NMF

On rappelle que $\lambda_{j,fn}^2 = (\mathbf{W}_j\mathbf{H}_j)_{f,n}$. On peut montrer que maximiser $\mathcal{L}(q^\star, \boldsymbol{\theta})$ par rapport à $\mathbf{W}_j, \mathbf{H}_j$ sous la contrainte $\mathbf{W}_j, \mathbf{H}_j \geq 0$ est équivalent à minimiser la fonction de coût suivante sous la même contrainte :

$$\mathcal{C}(\mathbf{W}_j, \mathbf{H}_j) = \sum_{(f,n)\in\mathcal{B}_j} d_{IS}\left(\hat{p}_{j,fn}, (\mathbf{W}_j\mathbf{H}_j)_{f,n}\right), \tag{6.39}$$

avec

$$\hat{p}_{j,fn} = \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\beta_{j,fn}/\nu_v}. \tag{6.40}$$

Une fois de plus nous pouvons avoir recours aux règles de mise à jour multiplicatives (2.23) et (2.24), page 32, dans le cas de la divergence d'Itakura-Saito.

Il est intéressant d'utiliser les résultats de l'étape E-V pour développer le premier argument de cette divergence. On peut montrer que :

$$\hat{p}_{j,fn} = \left(\frac{\alpha_v(\hat{s}_{j,fn}^2 + \gamma_{j,fn})^{-1} + (\mathbf{W}_j\mathbf{H}_j)_{f,n}^{-1}}{\alpha_v + 1}\right)^{-1}. \tag{6.41}$$

Il s'agit d'une moyenne harmonique pondérée entre la moyenne a posteriori du spectrogramme de puissance de la source j, $\langle s_{j,fn}^2\rangle = \hat{s}_{j,fn}^2 + \gamma_{j,fn}$, et la paramétrisation NMF de cette même source, en utilisant les valeurs courantes des paramètres. En injectant (6.41) dans l'équation (6.39) nous retrouvons un problème d'optimisation très similaire à celui présenté au chapitre 2, section 2.3.4b, équation (2.36), associé à l'estimation MV des paramètres NMF avec un modèle d'observations *t* de Student. L'unique différence vient du fait que nous avons ici l'espérance du spectrogramme de puissance, car les sources ne sont pas directement observées.

On comprend maintenant pourquoi nous avons choisi de reformuler le modèle de source NMF à l'équation (6.5) par rapport à celui proposé au chapitre 5, section 5.4, équation (5.39), page 99. Alors que pour ce dernier nous devions développer une nouvelle méthode d'estimation des paramètres NMF (voir section 5.4.2e du chapitre précédent), nous pouvons nous appuyer ici sur des méthodes de la littérature.

Il est finalement intéressant de remarquer que lorsque α_v tend vers l'infini (la distribution *t* de Student approche la gaussienne), $\hat{p}_{j,fn}$ tend vers $\hat{s}_{j,fn}^2 + \gamma_{j,fn}$ et on retrouve le même problème d'optimisation qu'à l'étape M des algorithmes (V)EM pour les modèles gaussiens présentés précédemment dans cette thèse.

6.3 Résultats expérimentaux

6.3.1 Méthodes de référence

Nous présentons ici les méthodes de la littérature avec lesquelles nous allons comparer l'approche proposée.

a) Modèle de source gaussien et filtres de mélange déterministes non contraints

Nous considérons tout d'abord la méthode que nous avons proposée dans [Leglaive et al., 2017a]. Celle-ci a été décrite au chapitre 5, section 5.3. Elle correspond au cas d'une représentation TF des sources par MDCT. Pour la présentation des résultats, cette approche sera nommée «gaussien - filtres non contraints» car elle est basée sur un modèle de source gaussien et des filtres de mélange déterministes et non contraints dans le domaine temporel.

b) Modèle de matrice de covariance spatiale

La seconde méthode que l'on considère a été proposée dans [Ozerov et al., 2012]. Elle généralise dans un même cadre les modèles de mélange convolutifs ponctuel et diffus (voir chapitre 2, section 2.4) utilisés respectivement dans [Ozerov et Févotte, 2010; Ozerov et al., 2011] et [Duong et al., 2010; Arberet et Vandergheynst, 2014]. Nous rappelons que cette méthode se base sur un modèle de source image multivariée gaussien dans le domaine de la TFCT :

$$\mathbf{y}_{j,fn} \sim \mathcal{N}(0, \lambda_{j,fn}^2 \mathbf{R}_{j,f}), \quad \text{avec} \quad \mathbf{R}_{j,f} = \mathbf{A}_{j,f}\mathbf{A}_{j,f}^H, \tag{6.42}$$

où $\mathbf{y}_{j,fn} = [y_{ij,fn}]_i^\top \in \mathbb{C}^I$, $\lambda_{j,fn}^2 \in \mathbb{R}_+$ représente la puissance de la source j, paramétrée par un modèle NMF, et $\mathbf{R}_{j,f} \in \mathbb{C}^{I\times I}$ est la matrice de covariance spatiale (SCM d'après l'anglais *spatial covariance matrix*), structurée à l'aide d'une matrice $\mathbf{A}_{j,f} \in \mathbb{C}^{I\times R_j}$ de rang $0 < R_j \leq I$. Le modèle convolutif ponctuel correspond au cas $R_j = 1$ et le modèle diffus au cas d'une matrice de rang plein ($R_j = 2$ pour des mélanges stéréophoniques).

Cette approche repose sur un algorithme EM. Nous utilisons ici l'implémentation fournie par les auteurs et disponible en ligne [3]. Pour la présentation des résultats, cette approche sera nommée «SCM rang 1» ou «SCM rang 2» selon le choix du rang de la matrice de covariance spatiale.

Nous effectuons 200 itérations des algorithmes (V)EM pour l'ensemble des méthodes comparées dans cette évaluation expérimentale. Pour l'approche proposée, nous effectuons 10 itérations de la méthode du gradient conjugué aux étapes E-S et E-A. A l'étape M, on effectue 10 itérations des règles de mise à jour multiplicatives des matrices d'activations NMF.

6.3.2 Cadre expérimental

Base de données Nous utilisons la base de données «MASS-MIRD» présentée à la section 2.7 et correspondant à des mélanges créés à partir de RIRs mesurées pour trois temps de réverbération différents : 160, 360 et 610 ms.

Scénario semi-aveugle La contribution principale de ce chapitre est de proposer un nouveau cadre bayésien pour la séparation de sources audio, où les filtres de mélange sont traités comme des variables aléatoires latentes dans le domaine temporel. C'est pourquoi nous sommes principalement intéressés par l'évaluation du modèle de mélange. Nous supposons par conséquent une certaine connaissance a priori des signaux sources. Afin de garder un cadre expérimental réaliste,

3. `http://bass-db.gforge.inria.fr/fasst/`

nous choisissons d'apprendre uniquement les dictionnaires NMF (les matrices $\mathbf{W}_j$ avec un rang $K_j = 10$) à partir des signaux sources originaux. Ces dictionnaires pré-entraînés de façon oracle sont ensuite gardés fixes durant l'algorithme VEM, seules les matrices d'activations $\mathbf{H}_j$ sont mises à jour à l'étape M. Nous supposons également que le temps de réverbération est connu afin de pouvoir définir le profil de décroissance exponentielle $r^2(t)$ à l'équation (6.7). Ce cadre expérimental semi-aveugle est utilisé pour la méthode proposée et les méthodes de référence.

6.3.3 Initialisation des paramètres

a) Paramètres variationnels

Nous présentons tout d'abord l'initialisation des paramètres de la méthode proposée. Les paramètres de forme ν_u et ν_v sont fixés d'après les équations (6.25) et (6.27). Les autres paramètres sont initialisés comme suit :

- $\beta_{j,fn} = \alpha_v/2$;
- $\gamma_{j,fn} = \tilde{v}_{j,fn}\lambda_{j,fn}^2$ où $\tilde{v}_{j,fn}$ correspond à une réalisation de la loi $\mathcal{IG}(\nu_v, \beta_{j,fn})$;
- $\hat{s}_{j,fn}$ est initialisé comme réalisation de la loi $\mathcal{N}(0, \gamma_{j,fn})$;
- $d_{ij}(t) = \alpha_u/2$;
- $\rho_{ij}(t) = \tilde{u}_{ij}(t)r^2(t)$ où $\tilde{u}_{ij}(t)$ correspond à une réalisation de la loi $\mathcal{IG}(\nu_u, d_{ij}(t))$;
- $\hat{a}_{ij}(t)$ est initialisé comme réalisation de la loi $\mathcal{N}(0, \rho_{ij}(t))$

Il est important de préciser que nous n'avons pas «optimisé» cette procédure d'initialisation. Il s'agit simplement d'initialiser les distributions variationnelles (6.13) à (6.16) en s'inspirant des distributions a priori associées présentées à la section 6.1.

b) Paramètres spatiaux des méthodes de référence

Les valeurs initiales des paramètres spatiaux des méthodes de référence sont fixées à partir de l'initialisation de $\hat{a}_{ij}(t)$ (voir paragraphe précédent), afin de fournir la même «information» à l'ensemble des algorithmes. Pour notre méthode [Leglaive et al., 2017a], les filtres de mélange déterministes sont exactement initialisés à $\hat{a}_{ij}(t)$. Pour les approches de la littérature basées sur l'utilisation d'une matrice de covariance spatiale, on calcule tout d'abord $\hat{a}_{ij,f}$, la TFD de $\hat{a}_{ij}(t)$ calculée sur un nombre de points égal à la longueur de la fenêtre d'analyse de la TFCT. Lorsque les filtres de mélange sont plus longs que cette dernière, on les tronque au préalable dans le domaine temporel. Soit $\hat{\mathbf{a}}_{j,f} = [\hat{a}_{ij,f}]_i^\top \in \mathbb{C}^I$, la matrice $\mathbf{A}_{j,f}$ à l'équation (6.42) est initialisée telle que $(\mathbf{A}_{j,f})_{:,r} = \hat{\mathbf{a}}_{j,f}$ pour tout $r \in \{1, ..., R_j\}$.

c) Paramètres NMF

Comme mentionné précédemment, on se place dans un cadre semi-aveugle où les matrices $\mathbf{W}_j$ sont calculées à partir des coefficients TF des sources originales. Ces matrices sont estimées au sens du maximum de vraisemblance en suivant la procédure décrite au chapitre 2, section 2.3.4a pour un modèle de source gaussien ou section 2.3.4b pour un modèle de source *t* de Student. Dans les deux cas il s'agit de résoudre un problème d'optimisation en utilisant des mises à jour multiplicatives. Afin de comparer les méthodes de la façon la plus juste possible, nous utilisons la même graine aléatoire pour initialiser les paramètres NMF avant d'effectuer ces mises à jour. Enfin, une fois que le dictionnaire NMF est estimé, on «jette» la matrice d'activations $\mathbf{H}_j$ obtenue et on la réinitialise par une matrice remplie de 1.

6.3.4 Hyperparamètres du modèle

Les expériences préliminaires décrites dans cette sous-section pour choisir les hyperparamètres du modèle sont effectuées à partir des mélanges créés avec un temps de réverbération de 360 ms.

a) Variance du modèle de bruit

Comme expliqué à la section 6.2.5a, la variance du bruit σ_i^2 est fixée manuellement pour décroître au cours des itérations de l'algorithme VEM suivant la courbe représentée sur la figure 6.4. Les valeurs initiale et finale sont respectivement 10^{-2} et 10^{-6}.

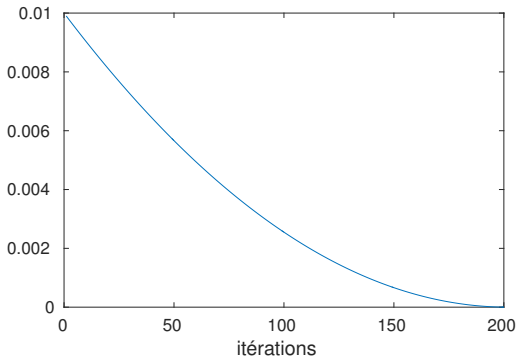


FIGURE 6.4 – Évolution de la variance du bruit σ_i^2 au cours des itérations de l'algorithme VEM.

b) Paramètre d'échelle du modèle de décroissance exponentielle

Du fait de l'indétermination d'échelle entre les filtres de mélange et les signaux sources, i.e. $[a_{ij}(\cdot) \star s_j(\cdot)](t) = [Ca_{ij}(\cdot) \star \frac{1}{C} s_j(\cdot)](t)$ pour tout facteur $C \in \mathbb{R}$, le paramètre d'échelle σ_r^2 à l'équation (6.7) peut être fixé arbitrairement. Nous choisissons $\sigma_r^2 = 10^{-2}$ afin que l'énergie totale des filtres de mélange ne soit pas trop élevée.

c) Paramètres de forme des distributions *t* de Student

Nous allons étudier l'influence des paramètres de forme des distributions *t* de Student utilisées pour modéliser les coefficients TF des sources et la réponse impulsionnelle des filtres de mélange. Pour cela nous considérons une grille de valeurs : $(\alpha_v, \alpha_u) \in \{0.1, 1, 10, 100, \infty\}^2$ (l'infini correspond à 4.5×10^{15}). On calcule pour chacun de ces couples de valeurs le SDR moyenné sur l'ensemble des sources de la base de données. Les résultats sont représentés à gauche sur la figure 6.5, les paramètres de forme qui maximisent le SDR sont $(\alpha_v, \alpha_u) = (100, 1)$.

On peut conclure que l'hypothèse gaussienne est raisonnable pour le modèle de source car la valeur optimale de α_v est particulièrement élevée. C'est un résultat qui est en accord avec les expériences déjà réalisées au chapitre 5, section 5.4. En revanche on remarque qu'il est important de choisir une valeur faible pour le paramètre de forme α_u du modèle de filtres de mélange. Cela confirme dans le cadre d'une application de séparation de sources que la distribution *t* de Student est mieux adaptée à la modélisation des RIRs que la gaussienne.

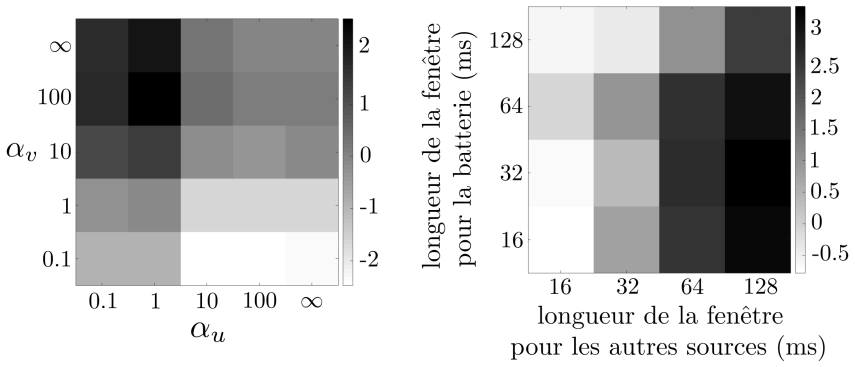


FIGURE 6.5 – SDR moyen en fonction des paramètres de forme α_v et α_u (figure de gauche) et en fonction de la taille de la fenêtre de MDCT (en ms) pour la batterie et les autres sources (figure de droite). Le temps de réverbération des mélanges est de 360 ms.

d) Résolution temps-fréquence

Dans l'expérience présentée au paragraphe précédent, la longueur de la fenêtre de MDCT était fixée à 64 ms pour toutes les sources. Cependant, un des avantages de travailler directement avec les signaux temporels captés par les microphones est que nous pouvons adapter la transformation TF à chaque source composant le mélange. Il est possible en particulier de choisir la longueur de la fenêtre de MDCT en fonction de la nature des sources, afin d'adapter les résolutions temporelle et fréquentielle. Nous considérons ici un cas d'application simple : On choisit une longueur de fenêtre différente pour la batterie et pour les autres sources du mélange. Comme précédemment nous avons recours à une grille de valeur qui est représentée à droite sur la figure 6.5. Les paramètres de forme des distributions *t* de Student sont fixés d'après les résultats obtenus au paragraphe précédent.

Comme nous pouvions nous y attendre, une courte fenêtre (32 ms) pour la batterie permet d'obtenir les meilleurs résultats. Cet instrument génère en effet des sons fortement percussifs d'où l'importance d'une bonne résolution temporelle. A l'inverse, pour tous les autres instruments (piano, basse, voix ou guitare) une fenêtre plus longue (128 ms) est mieux adaptée afin d'obtenir une bonne résolution fréquentielle.

Nous avons considéré ici une configuration extrêmement simple pour adapter le dictionnaire TF aux sources du mélange. Cela pourra faire l'objet d'une étude plus approfondie dans de futurs travaux. Il serait notamment intéressant d'utiliser un dictionnaire TF hybride comme cela a été proposé dans [Févotte et Kowalski, 2014; Feng et Kowalski, 2014].

6.3.5 Comparaison avec les méthodes de référence

On compare dans cette sous-section les performances de séparation obtenues avec l'approche proposée et avec les méthodes de la littérature présentées précédemment. Nous utilisons l'ensemble des mélanges, associés à trois temps de réverbération différents : 160, 360 et 610 ms. Les résultats présentés au paragraphe précédent et ceux qui vont être détaillés ici sont illustrés par des exemples audio disponibles en ligne [4].

En accord avec les expériences préliminaires, nous choisissons comme paramètres de forme des distributions *t* de Student $(\alpha_v, \alpha_u) = (100, 1)$. Afin de fournir une comparaison la plus juste possible, nous présenterons les résultats de la méthode proposée dans les deux configurations suivantes :

4. https://perso.telecom-paristech.fr/leglaive/demoSSMM.html

- ▷ «méthode proposée - sans fenêtre MDCT adaptée» : La longueur de la fenêtre de MDCT est la même pour toutes les sources. Elle est fixée comme pour les méthodes de référence à 64 ms.
- ▷ «méthode proposée - avec fenêtre MDCT adaptée» : La longueur de la fenêtre de MDCT est de 32 ms pour la batterie et 128 ms pour les autres sources.

Comme l'indiquent les résultats dans le tableau 6.1, l'approche proposée permet d'obtenir des performances robustes à la présence de forte réverbération. A l'inverse, on observe que les performances des méthodes de référence chutent avec l'augmentation du temps de réverbération. Cela montre que le modèle de matrice de covariance spatiale (de rang 1 ou 2) n'est pas suffisamment précis pour représenter un mélange fortement réverbérant, ce qui a également été observé dans [Li et al., 2017a] et [Li et al., 2017b]. Par ailleurs, on voit que les modèles de rang 1 et 2 obtiennent des performances très proches, ce qui est également confirmé à l'écoute des résultats.

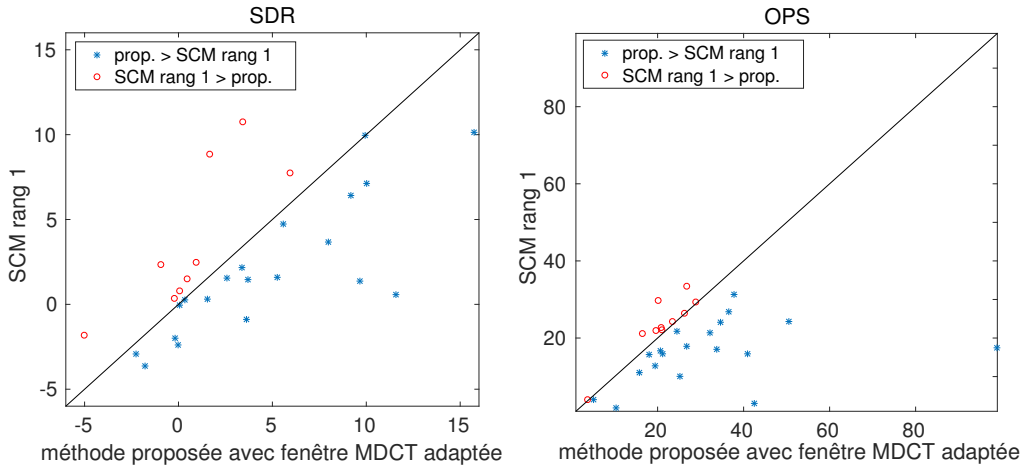
D'après le SDR, notre approche [Leglaive et al., 2017a] avec filtres de mélanges non contraints dans le domaine temporel obtient les moins bonnes performances, tandis que l'OPS semble indiquer qu'en moyenne cette méthode permet d'obtenir une qualité de séparation meilleure que dans le cas d'un modèle par matrice de covariance spatiale. En écoutant les résultats de séparation, nous pensons effectivement que la qualité n'est pas satisfaisante lorsque les filtres ne sont pas contraints, particulièrement dans le cas d'un mélange très réverbéré. Cela confirme la tendance du SDR à décroître avec l'augmentation du temps de réverbération. Bien que pour cette méthode le processus de mélange convolutif soit représenté de façon exacte, nous pensons que ces résultats peuvent s'expliquer par le fait que les filtres sont estimés sans contraintes, uniquement à partir des données observées. Nous avons en effet remarqué que pour certains morceaux, une partie des signaux sources pouvait se retrouver dans les filtres de mélange (cf. exemple audio en ligne). Nous supposons que cela provient des ambiguïtés qui existent entre les filtres et les sources dans l'expression du mélange convolutif. L'introduction d'a priori sur les filtres de mélange permet de résoudre ce problème.

On peut aussi remarquer que bien que les hyperparamètres du modèle (paramètres de forme des distributions t de Student et taille des fenêtres de MDCT) aient été optimisés en utilisant les mélanges associés à un temps de réverbération de 360 ms, les valeurs obtenues semblent se généraliser aux autres conditions de mélange.

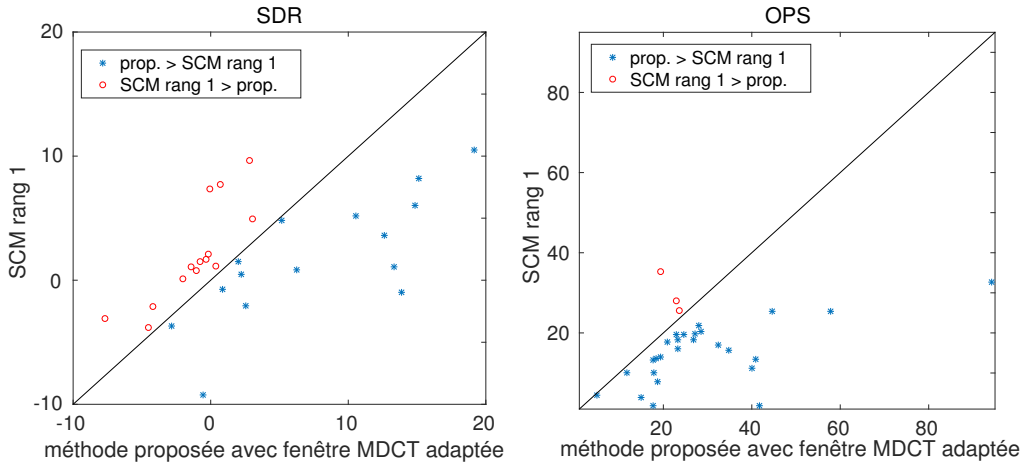
Enfin, pour évaluer la pertinence des résultats, nous traçons sur la figure 6.6 un nuage de points, pour chaque temps de réverbération, représentant le SDR ou l'OPS obtenu avec la meilleure des méthodes de référence («SCM rang 1», en ordonnées) et la meilleure des méthodes proposées (avec fenêtre MDCT adaptée, en abscisse). Chaque point correspond à une source de la base de données. D'après le SDR, un nombre non négligeable de sources est mieux séparé par la méthode de référence. Cependant lorsque l'approche proposée améliore les performances, le gain en SDR est en moyenne plus important. Si l'on regarde maintenant l'OPS, on voit que la méthode proposée permet d'améliorer les résultats pour quasiment l'ensemble des sources. Cela est d'autant plus vrai que le temps de réverbération est élevé. Tout comme au chapitre 5, nous pensons que l'OPS est plus révélateur de la qualité de séparation perçue. Cette différence de résultats entre le SDR et l'OPS pourrait comme précédemment provenir de l'utilisation de la MDCT, qui par rapport à la TFCT semble pénaliser le SDR.

Temps de réverbération : 160 ms					
	SDR	ISR	SIR	SAR	OPS
SCM rang 1 [Ozerov et al., 2012]	2.5	6.4	3.6	9.8	18.8
SCM rang 2 [Ozerov et al., 2012]	2.4	6.2	3.5	9.7	18.8
gaussien - filtres non contraints [Leglaive et al., 2017a]	1.2	6.2	3.2	8.7	20.8
méthode proposée - sans fenêtre MDCT adaptée	2.8	6.9	4.3	10.6	20.8
méthode proposée - avec fenêtre MDCT adaptée	**3.5**	**8.0**	**6.1**	**12.0**	**27.6**
Temps de réverbération : 360 ms					
	SDR	ISR	SIR	SAR	OPS
SCM rang 1 [Ozerov et al., 2012]	1.9	5.7	2.4	9.4	16.6
SCM rang 2 [Ozerov et al., 2012]	1.8	5.6	2.5	9.5	16.6
gaussien - filtres non contraints [Leglaive et al., 2017a]	0.6	6.3	2.6	6.9	20.6
méthode proposée - sans fenêtre MDCT adaptée	2.6	7.4	3.5	10.7	22.4
méthode proposée - avec fenêtre MDCT adaptée	**3.4**	**8.7**	**5.0**	**12.5**	**28.2**
Temps de réverbération : 610 ms					
	SDR	ISR	SIR	SAR	OPS
SCM rang 1 [Ozerov et al., 2012]	1.8	5.4	2.4	9.4	14.6
SCM rang 2 [Ozerov et al., 2012]	1.7	5.3	2.2	9.5	14.7
gaussien - filtres non contraints [Leglaive et al., 2017a]	0.3	5.8	1.8	5.3	19.0
méthode proposée - sans fenêtre MDCT adaptée	**3.2**	**8.2**	**4.9**	10.3	23.8
méthode proposée - avec fenêtre MDCT adaptée	3.1	8.1	4.1	**11.6**	**24.8**

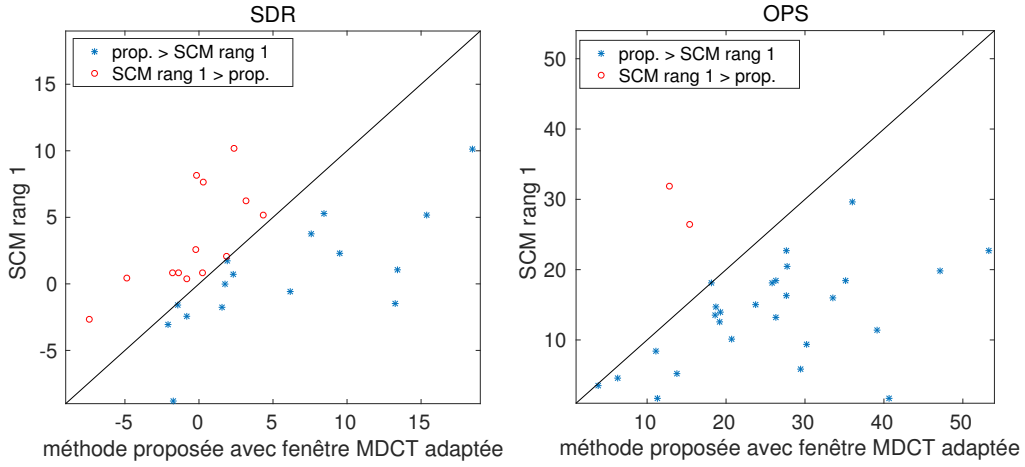
TABLEAU 6.1 – Résultats de séparation moyennés sur l'ensemble des sources de la base de données, pour différents temps de réverbération.



(a) Temps de réverbération : 160 ms



(b) Temps de réverbération : 360 ms



(c) Temps de réverbération : 610 ms

FIGURE 6.6 – SDR (en dB) et OPS (en %) des 29 sources de la base de données. Résultats obtenus avec la méthode de référence «SCM rang 1» en fonction de ceux obtenus avec la méthode proposée dans le cas d'une fenêtre MDCT adaptée en fonction de la source.

6.4 Conclusion

Dans ce chapitre nous avons introduit un nouveau cadre bayésien pour la séparation multicanale de sources sonores, basé sur une modélisation TF des sources et une modélisation temporelle des filtres de mélange.

Les expériences ont permis de montrer que l'hypothèse gaussienne pour le modèle de source semble raisonnable. A l'inverse, l'utilisation d'une distribution à queue lourde est d'une importance cruciale pour la modélisation de la réponse impulsionnelle des filtres de mélange. La robustesse de la distribution *t* de Student nous permet de prendre en compte indirectement le trajet direct et les premiers échos des RIRs, qui d'un point de vue statistique correspondent à des valeurs aberrantes par rapport au modèle gaussien à amplitude exponentiellement décroissante de l'équation (6.6), largement utilisé en acoustique statistique des salles.

On sait cependant que ce modèle gaussien est valide pour la réverbération tardive [Polack, 2015]. On peut de la même façon qu'à la figure 6.2 tracer un histogramme uniquement pour la réverbération tardive et vérifier qu'il s'apparente bien à une distribution gaussienne. C'est pourquoi il pourrait être naturel de considérer le modèle gaussien (6.6) uniquement pour la réverbération tardive, tandis qu'un modèle différent serait utilisé pour les contributions précoces. Cette approche peut se formaliser au travers du modèle de RIR suivant :

$$a(t) \sim \mathcal{T}_{\alpha_u(t)}(0, r(t)), \tag{6.43}$$

où $r(t)$ est défini à l'équation (6.7). Le paramètre de forme est maintenant dépendant du temps et peut être choisi tel que $\alpha_u(t) \to \infty$ pour les instants t associés à la réverbération tardive (t supérieur au temps de mélange). Dans ce cas la distribution *t* de Student correspond à la gaussienne.

D'après l'équation (6.43), les coefficients de la RIR normalisée $a(t)/\exp(-t/\tau)$ sont supposés suivre la distribution $\mathcal{T}_{\alpha_u(t)}(0, \sigma_r)$. On peut alors réutiliser les 624 RIRs de la base de données MIRD [Hadad et al., 2014] pour estimer le paramètre de forme optimal au sens du maximum de vraisemblance. On représente en niveau de gris sur la figure 6.7 la log-vraisemblance des données calculée sur des blocs de 10 ms non recouvrants. Pour chaque bloc on normalise la log-vraisemblance par sa valeur maximale. Le paramètre de forme qui maximise la log-vraisemblance normalisée est représenté par un carré blanc sur chaque bloc. On voit que le paramètre de forme optimal est petit ($\alpha_u(t) \approx 10$) pour les premières contributions (jusqu'à 60 ms environ), il augmente ensuite progressivement jusqu'à atteindre sa valeur maximale, illustrant ainsi la validité de l'hypothèse gaussienne pour la réverbération tardive. Nous voyons qu'à la toute fin de la RIR le paramètre de forme optimal diminue, nous pensons que ce comportement ne doit pas être pris en compte car la fin d'une RIR n'est que peu énergétique et donc potentiellement bruitée. Finalement, ces résultats sont à nuancer par le fait que la log-vraisemblance est très plate sur cette représentation à deux dimensions, comme l'indique l'échelle de couleur sur la figure 6.7. Ce critère semble donc peu impacté par le choix du paramètre de forme.

Nous avons essayé d'utiliser ce modèle de RIR avec paramètre de forme variant dans le temps pour la méthode de séparation de sources présentée précédemment. Nous avons testé plusieurs configurations pour définir son évolution temporelle (faible valeur pour les contributions précoces et grande valeur pour la réverbération tardive, évolution fixée suivant les résultats de la figure 6.7, etc.), cependant cela n'a pas permis d'améliorer les performances de séparation, c'est pourquoi nous n'avons pas présenté cette approche plus en détails. Ce résultat négatif peut éventuellement s'expliquer par le fait que même si utiliser un paramètre de forme dépendant du temps nous permet de mieux caractériser les statistiques réelles des filtres de mélange, cela n'implique pas nécessairement une meilleure séparation. Il est par exemple connu en déconvolution aveugle d'images que l'a priori optimal n'est pas celui qui caractérise le mieux les statistiques réelles des images naturelles, mais plutôt celui qui permet de discriminer au mieux l'image nette de sa version floutée

[Wipf et Zhang, 2014]. Il est possible que nous observions ici un phénomène similaire, ce point devra être approfondi dans de futurs travaux.

D'autres perspectives liées à la méthode présentée dans ce chapitre sont décrites en conclusion de cette thèse, dans la partie suivante.

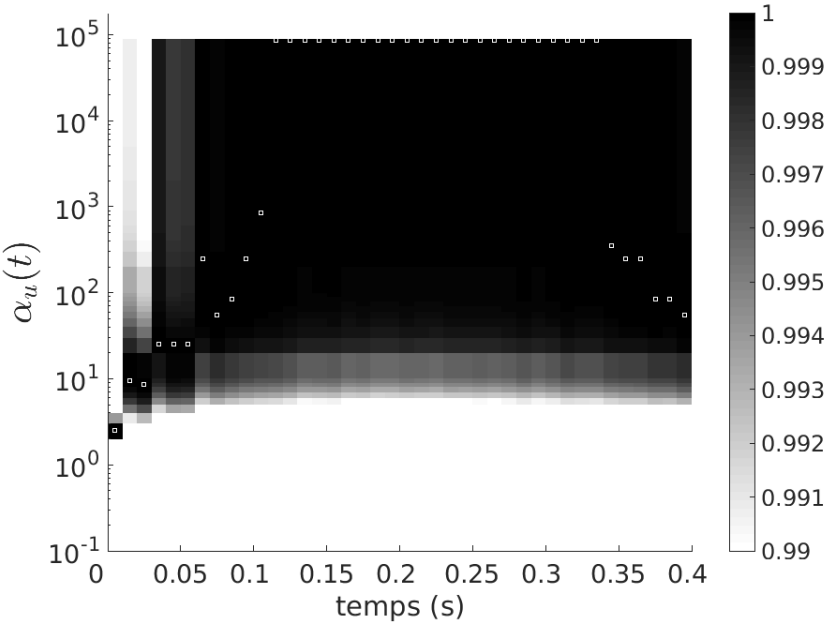


FIGURE 6.7 – Évolution temporelle du paramètre de forme maximisant la vraisemblance calculée à partir de 624 RIRs (avec temps de réverbération de 610 ms), découpées en blocs non recouvrants d'une durée de 10 ms. La zone en niveau de gris représente la log-vraisemblance normalisée sur chaque bloc. Les carrés blancs indiquent la valeur optimale du paramètre de forme sur chaque bloc.

Quatrième partie

Conclusion et perspectives

Conclusion

Nous avons cherché à montrer dans cette thèse l'intérêt qu'il pouvait y avoir pour la séparation de sources audio en milieu réverbérant à exploiter le fait que les filtres de mélange correspondent à des réponses de salle. Cette information nous a amené à développer des modèles probabilistes pour les filtres de mélange, que ce soit dans le domaine fréquentiel ou temporel, permettant de guider leur estimation dans un contexte de séparation de sources.

Approche fréquentielle

Dans un premier temps nous nous sommes appuyés sur une approche fréquemment employée qui consiste à formuler le processus de mélange convolutif temporel comme une simple multiplication dans le domaine de la TFCT. Ce modèle faisant intervenir la réponse en fréquence des filtres de mélange, nous avons cherché à transcrire leur dynamique temporelle sous forme de corrélations fréquentielles, par l'intermédiaire de modèles autorégressifs à moyenne ajustée. Ces modèles de réponse en fréquence de salle nous ont permis de développer des a priori probabilistes sur les filtres de mélange, pris en compte au travers d'une estimation au sens du maximum a posteriori dans le cadre d'un algorithme EM. Nous avons montré expérimentalement que cette approche permettait d'améliorer la qualité de séparation. Néanmoins, cette formulation du mélange reste une approximation qui devient de plus en plus fausse à mesure que le temps de réverbération augmente. Les performances de séparation sont donc fondamentalement limitées lorsque le mélange est enregistré en présence de forte réverbération.

Approche temporelle

C'est pourquoi dans un second temps nous avons souhaité développer des méthodes s'appuyant sur une représentation exacte du processus de mélange convolutif dans le domaine temporel, tout en gardant une modélisation probabiliste TF des sources basée sur des techniques de NMF. Nous avons proposé d'inférer les variables latentes TF des sources à partir des observations temporelles du mélange. En plus d'être appropriée pour la séparation de mélanges fortement réverbérants, cette approche nous a permis de développer un modèle simple de filtres de mélange. Nous avons représenté ces derniers comme des variables aléatoires latentes dans le domaine temporel, suivant un modèle t de Student exploitant la décroissance exponentielle de la réverbération tardive. La robustesse de cette distribution de probabilité nous a permis de prendre en compte implicitement le trajet direct et les premiers échos des filtres de mélange. Ces modèles ont été utilisés pour développer une méthode de séparation de sources basée sur une technique d'inférence variationnelle. Nous avons montré expérimentalement le potentiel de cette approche pour la séparation des mélanges fortement réverbérants, en supposant des dictionnaires NMF oracles.

Perspectives

Nous pensons que l'approche temporelle résumée au paragraphe précédent induit plusieurs perspectives qu'il serait intéressant d'étudier dans de futurs travaux, nous les présentons ci-dessous.

Approche supervisée pour la modélisation des sources

Pour la méthode développée au dernier chapitre, nous utilisions des dictionnaires NMF oracles, appris à partir des sources originales. Une extension directe de ce travail consisterait à apprendre des dictionnaires NMF spécifiques à chaque source en utilisant une base de données annexe, telle

que MedleyDB [Bittner et al., 2014]. Nous pourrions également utiliser des modèles compositionnels de sources plus sophistiqués, tels que ceux introduits dans [Ozerov et al., 2012]. Ces derniers permettent par exemple de prendre en compte la nature harmonique de certaines sources, ou bien d'exploiter la connaissance du système de production des sources, au travers d'un modèle source/filtre. Dans ce cas, seule l'étape M de l'algorithme VEM proposé au chapitre précédent devrait être modifiée.

Une autre perspective importante consiste à utiliser un réseau de neurones pour représenter les paramètres $\lambda_{j,fn}^2$ du modèle de source *t* de Student (6.4). Cette approche a récemment été introduite dans le cadre d'un modèle gaussien de sources images multicanales, basé notamment sur l'utilisation de matrices de covariance spatiales [Nugraha et al., 2016a,b]. Il s'agit d'utiliser au sein d'un algorithme EM un réseau de neurones pour mettre à jour les paramètres $\lambda_{j,fn}^2$ caractérisant la puissance des sources dans le domaine TF. Cette méthode a obtenu une des meilleures performances au challenge SiSEC (*signal separation evaluation campaign*) 2016 [Liutkus et al., 2017], pour la tâche de séparation de musique produite professionnellement. Les auteurs utilisent également un réseau de neurones pour initialiser les paramètres $\lambda_{j,fn}^2$. La qualité de cette initialisation permet d'effectuer seulement un faible nombre d'itérations de l'algorithme EM, le temps de calcul requis par la méthode est donc fortement limité.

Extension du modèle de mélange

Nous supposions au dernier chapitre la connaissance du temps de réverbération afin de pouvoir fixer le profil de décroissance exponentielle de la réverbération à l'équation (6.7). Il serait intéressant d'étudier la robustesse du modèle *t* de Student pour les filtres de mélange vis à vis d'une mauvaise estimation du temps de réverbération.

Nous pourrions également développer, voire apprendre, d'autres profils temporels $r(t)$ afin de représenter d'autres conditions de mélange. Par exemple, en production musicale, des effets comme des délais ou des échos sont souvent utilisés. Dans ce cas la réponse impulsionnelle d'un filtre de mélange est composée de plusieurs impulsions de dirac, chacune étant associée à un écho. Un tel filtre pourrait être représenté par le même modèle *t* de Student qu'à l'équation (6.8), mais en fixant $r(t)$ à une faible valeur pour tout t afin de représenter la parcimonie de la réponse impulsionnelle.

Techniques d'inférence

Dans cette thèse nous avons utilisé des techniques d'inférence approchée basées sur des approximations variationnelles. Il a été montré dans [Cemgil et al., 2007] l'intérêt d'utiliser des approches hybrides, alternant méthode de Monte Carlo par chaîne de Markov pour explorer rapidement l'espace des variables latentes et méthode variationnelle pour converger rapidement vers un mode de la distribution a posteriori. Il serait intéressant d'étudier cette approche dans le contexte de la méthode proposée au chapitre précédent. On pourrait pour cela utiliser des méthodes de Monte Carlo par chaîne de Markov dédiées à l'échantillonnage de distributions gaussiennes en grande dimension [Gilavert et al., 2015].

Il serait également intéressant d'étudier les nouvelles méthodes d'inférence bayésienne inspirées du domaine de l'apprentissage profond [Kingma, 2017] et dédiées à la résolution de problèmes en grande dimension ou faisant appel à de grandes bases de données.

Autres applications

Pour terminer cette partie concernant les perspectives liées à la méthode présentée au chapitre 6, nous présentons ci-dessous d'autres applications que la séparation de sources auxquelles

cette méthode pourrait s'appliquer et qu'il serait intéressant d'explorer.

Déconvolution Comme c'est souvent le cas en séparation de sources audio, nous avions comme objectif dans cette thèse de séparer les sources images multicanales et non les sources monophoniques. Cependant nous pourrions explorer dans de futurs travaux l'utilisation de notre méthode pour une application de déconvolution aveugle, correspondant au cas limite où le nombre de sources J est égal à 1. Il serait de plus intéressant de formuler le problème de déconvolution comme un problème d'apprentissage de dictionnaire, de la même façon que ce qui est proposé dans [Barchiesi et Plumbley, 2011]. Les éléments de ce dictionnaire correspondent à la convolution des atomes TF avec la réponse impulsionnelle du filtre de convolution. Nous chercherions alors à développer une méthode bayésienne exploitant des a priori sur les coefficients TF du signal d'intérêt et sur le dictionnaire.

Rehaussement et reconnaissance automatique de la parole Nous pourrions également utiliser notre approche afin de rehausser un signal de parole enregistré dans un environnement bruité et réverbérant, avec comme objectif final une application de reconnaissance automatique de la parole (ASR de l'anglais *automatic speech recognition*). En effet, la robustesse aux bruits interférant avec le signal de parole et aux conditions d'enregistrement réverbérantes est un critère important qui impacte de façon significative les performances en ASR. C'est de plus un problème toujours ouvert d'après une étude récente [Sivasankaran et al., 2017]. Les méthodes d'ASR se basent généralement sur des descripteurs audio calculés à partir du signal de parole rehaussé. Une telle approche pourrait bénéficier de la nature bayésienne du modèle présenté au chapitre précédent. En effet, comme nous cherchons à inférer la distribution a posteriori du signal à rehausser (sous une approximation variationnelle), nous disposons d'une information sur l'incertitude de l'estimateur, portée par la variance a posteriori des coefficients TF du signal. Cette incertitude peut alors être propagée au calcul des descripteurs audio. Il a été montré dans [Adiloğlu et Vincent, 2016] que ce type d'approche par propagation de l'incertitude permet d'améliorer les performances des méthodes d'ASR.

Localisation de sources Développer des descripteurs robustes à la réverbération et à des conditions d'enregistrement bruitées est un enjeu important pour le problème de localisation de sources [Li et al., 2016, 2017c]. Il serait intéressant d'étudier si l'information portée par les variables inverse-gamma $u_{ij}(t)$ du modèle de réponse de salle (6.9) peut être utilisée dans ce but. Ces variables sont supposées «encoder» la présence du trajet direct et des premiers échos dans les filtres de mélange. Nous pourrions alors chercher à extraire des descripteurs à partir de la distribution a posteriori de ces variables latentes et les utiliser pour développer une méthode supervisée de localisation de sources (voir par exemple [Deleforge et al., 2015]). Finalement, nous pourrions introduire dans le modèle bayésien représenté sur la figure 6.3 une dépendance des variables latentes $u_{ij}(t)$ à la position relative de la source j au microphone i.

Afin de développer ces méthodes de réhaussement et de localisation, nous pourrions nous appuyer sur une base de données récemment introduite dans [Bertin et al., 2016]. Il s'agit de signaux de parole enregistrés dans de multiples conditions réelles domestiques (salon, cuisine, chambre et salle de bain pour trois maisons différentes), en présence de bruit et de réverbération.

Cinquième partie

Annexes

Annexe A

Distributions de probabilité univariées

A.1 Distribution gaussienne

A.1.1 La distribution gaussienne réelle

Soit $\mathcal{N}_{\mathbb{R}}(\mu, \sigma^2)$ la distribution gaussienne définie pour une variable aléatoire (v.a.) à valeurs réelles x. Sa densité de probabilité est définie par [Bishop, 2006, annexe B] :

$$N_{\mathbb{R}}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{A.1}$$

A.1.2 La distribution gaussienne complexe

Soit $\mathcal{N}_{\mathbb{C}}(\rho, \mu_{x_r}, \mu_{x_\imath}, \sigma^2_{x_r}, \sigma^2_{x_\imath})$ la distribution gaussienne définie pour une v.a. à valeurs complexes $x = x_r + \imath x_\imath$. Sa densité de probabilité est définie par [Adali et al., 2011] :

$$\begin{aligned} N_{\mathbb{C}}(x; \rho, \mu_{x_r}, \mu_{x_\imath}, \sigma^2_{x_r}, \sigma^2_{x_\imath}) &= \frac{1}{2\pi\sigma_{x_r}\sigma_{x_\imath}\sqrt{1-\rho^2}} \\ &\times \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_r-\mu_{x_r})^2}{\sigma^2_{x_r}} + \frac{(x_\imath-\mu_{x_\imath})^2}{\sigma^2_{x_\imath}} - \frac{2\rho(x_r-\mu_{x_r})(x_\imath-\mu_{x_\imath})}{\sigma_{x_r}\sigma_{x_\imath}}\right)\right], \end{aligned} \tag{A.2}$$

où $\rho = \mathbb{E}[(x_r-\mu_{x_r})(x_\imath-\mu_{x_\imath})]/(\sigma_{x_r}\sigma_{x_\imath}) \in [-1, 1]$.

Le cas particulier $\mathcal{N}_{\mathbb{C}}(\rho = 0, \mu_{x_r}, \mu_{x_i}, \sigma^2_{x_r} = \sigma^2/2, \sigma^2_{x_i} = \sigma^2/2)$ correspond à la distribution gaussienne complexe *propre*. Nous la noterons également $\mathcal{N}_c(\mu, \sigma^2)$ ou bien $\mathcal{N}^p_{\mathbb{C}}(\mu, \sigma^2)$, avec $\mu = \mu_{x_r} + \imath\mu_{x_i}$ et $\sigma^2 = 2\sigma^2_{x_r} = 2\sigma^2_{x_i}$. Sa pdf est définie par :

$$N_c(x; \mu, \sigma^2) = N^p_{\mathbb{C}}(\mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right). \tag{A.3}$$

Finalement cette distribution gaussienne complexe est dite à *symétrie circulaire* si elle est propre et si de plus $\mu = 0$. Dans ce cas la distribution est invariante à un déphasage (une rotation dans le plan complexe) de sorte que x et $xe^{\imath\Psi}$ ont la même distribution pour tout $\Psi \in \mathbb{R}$.

A.2 Distribution *t* de Student

Soit $\mathcal{T}_\nu(\mu, \lambda)$ la distribution t de Student (non standardisée) définie pour une v.a. x à valeurs complexes si $\phi = 1$ et réelle si $\phi = 2$. Sa densité de probabilité est définie par [Kotz et Nadarajah,

2004] :

$$T_\nu(x;\mu,\lambda) = \frac{2}{\phi}\frac{1}{(\nu\pi\lambda^2)^{1/\phi}}\frac{\Gamma(\nu/2+1/\phi)}{\Gamma(\nu/2)}\left(1+\frac{2}{\phi\nu}\frac{|x-\mu|^2}{\lambda^2}\right)^{-(\nu/2+1/\phi)}, \tag{A.4}$$

où $\Gamma(\cdot)$ est la fonction gamma.

A.3 Distribution inverse-gamma

Soit $\mathcal{IG}(\alpha,\beta)$ la distribution inverse-gamma définie pour une v.a. $x \in \mathbb{R}_+$. Sa densité de probabilité est définie par [Carlin et Louis, 2008, annexe A] :

$$IG(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}\exp\left(\frac{-\beta}{x}\right). \tag{A.5}$$

La distribution inverse-gamma possède les propriétés suivantes :

$$\begin{aligned}\mathbb{E}[\ln(x)] &= \ln(\beta) - \text{di}\Gamma(\alpha);\\ \mathbb{E}[x^{-1}] &= \frac{\alpha}{\beta},\end{aligned}$$

où $\text{di}\Gamma(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ est la fonction digamma.

Annexe B

Preuve de l'équation (3.19)

Dans cette annexe nous démontrons l'expression théorique de la variance σ^2_{rev}, donnée à l'équation (3.18), page 58. Soit f un indice de fréquence en Hertz. La relation entre la RFR $A_l(f)$ et la pression sonore complexe $P_l(f)$ due à la réverbération à un point dans la salle pour une source ponctuelle est donnée par [Morse et Ingard, 1968, p. 311], [Radlovic et al., 2000] :

$$P_l(f) = -j2\pi f \rho_0 Q(f) A_l(f), \tag{B.1}$$

avec ρ_0 la densité volumique de l'air en kg.m^{-3} et $Q(f)$ la vitesse volumique de la source en m^3.s^{-1}. D'après l'égalité précédente on peut relier la variance de la pression et celle de la RFR :

$$\sigma^2_{rev} = \mathbb{E}[|A_l(f)|^2] = \frac{\mathbb{E}[|P_l(f)|^2]}{(2\pi f)^2 \rho_0^2 |Q(f)|^2}, \tag{B.2}$$

où $|Q(f)|^2$ s'exprime [Morse et Ingard, 1968, p. 311], [Radlovic et al., 2000] :

$$|Q(f)|^2 = \frac{W_s c}{\pi f^2 \rho_0}. \tag{B.3}$$

c est la vitesse du son en m.s^{-1} et W_s la puissance de la source en Watts. Nous devons maintenant exprimer la pression quadratique moyenne (PQM) due au champ réverbérant $\mathbb{E}[|P_l(f)|^2]$. On note α le coefficient moyen d'absorption de la salle (sans dimension) et $\mathcal{S}$ la surface totale des parois de la salle en m^2. Dans un champ réverbérant l'intensité est constante et est reliée à la PQM par [Morse et Ingard, 1968, p. 581] :

$$I_{rev} = \frac{\mathbb{E}[|P_l(f)|^2]}{4\rho_0 c}. \tag{B.4}$$

Pour un champ sonore diffus en conditions stationnaires, la puissance réverbérée provenant de la source doit être égale à la puissance absorbée par les murs [Bies et Hansen, 2009, ch. 7]. La puissance réverbérée peut être définie comme la puissance de la source présente après la première réflexion du trajet direct sur une paroi, $(1-\alpha)W_s$, et la puissance absorbée est $\alpha \mathcal{S} I_{rev}$. D'après cette condition on peut écrire :

$$(1-\alpha)W_s = \alpha \mathcal{S} \frac{\mathbb{E}[|P_l(f)|^2]}{4\rho_0 c}. \tag{B.5}$$

On obtient alors l'expression de la PQM due au champ réverbérant :

$$\mathbb{E}[|P_l(f)|^2] = \frac{4\rho_0 c(1-\alpha)W_s}{\alpha \mathcal{S}}. \tag{B.6}$$

En substituant les équations (B.6) et (B.3) dans (B.2) on obtient le résultat donné par l'équation (3.18).

Annexe C

Preuve de l'équation (3.16)

Dans cette annexe nous démontrons l'expression donnée à l'équation (3.15), page 58. A partir de la définition de la TFD on peut montrer que :

$$\mathcal{F}_T\{a_l^*(T-1-t)\} = e^{\imath 2\pi f/T} A_l^*(f). \tag{C.1}$$

En appliquant la TFD inverse, notée $\mathcal{F}_T^{-1}\{\cdot\}$, à l'équation (C.1) et d'après les théorèmes de la convolution circulaire et du retard :

$$a_l^*(T-1-t) = \delta(t+1) \circledast \mathcal{F}_T^{-1}\{A_l^*(f)\}, \tag{C.2}$$

où $\circledast$ représente l'opérateur de convolution circulaire, et $\delta(t)$ est l'impulsion de Dirac. En utilisant le fait que $\mathcal{F}_T^{-1}\{A_l^*(f)\} = \dfrac{1}{T}\mathcal{F}_T\{A_l(f)\}^*$ dans l'équation (C.2) on obtient :

$$\mathcal{F}_T\{A_l(f)\} = T a_l(T-t). \tag{C.3}$$

Finalement, en injectant (C.3) dans (3.12) on obtient le résultat de l'équation (3.15).

Annexe D

Méthode du gradient conjugué

La méthode du gradient conjugué a été introduite pour résoudre de façon itérative des systèmes d'équations linéaires impliquant une matrice symétrique définie positive [Hestenes et Stiefel, 1952]. Cela correspond de façon équivalente à minimiser une fonction quadratique, strictement convexe. Contrairement à une méthode de descente de gradient standard, la méthode du gradient conjugué tient compte d'une «mémoire» dans le choix à chaque itération de la nouvelle direction de descente, par l'intermédiaire de la notion de conjugaison. En effet, chaque nouvelle direction de descente est choisie comme étant orthogonale à la précédente, pour un certain produit scalaire.

D.1 Gradient conjugué

Soit $\mathbf{A} \in \mathbb{R}^{n \times n}$ une matrice hermitienne définie positive. On définit le produit scalaire par rapport à $\mathbf{A}$ entre deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ par $(\mathbf{u}, \mathbf{v})_{\mathbf{A}} = (\mathbf{A}\mathbf{u}, \mathbf{v}) = \mathbf{v}^H \mathbf{A}\mathbf{u}$. On dit que $\mathbf{u}$ et $\mathbf{v}$ sont conjugués s'ils sont orthogonaux pour ce produit scalaire. La méthode du gradient conjugué a pour objectif de résoudre $\mathbf{A}\mathbf{x} = \mathbf{b}$ en minimisant la forme quadratique $J(\mathbf{x}) = \frac{1}{2}\mathbf{x}^H \mathbf{A}\mathbf{x} - \mathbf{x}^H \mathbf{b}$. En effet, l'annulation du gradient de cette fonction revient à résoudre le système d'équations.

A l'itération k de l'algorithme, la nouvelle direction de descente $\mathbf{w}_{k+1}$ n'est pas directement égale au gradient $\mathbf{g}_{k+1}$, mais à une version corrigée de celui-ci, de telle sorte que la nouvelle direction de descente soit conjuguée à la précédente, i.e. $(\mathbf{w}_{k+1}, \mathbf{w}_k)_{\mathbf{A}} = 0$.

On peut montrer que la méthode du gradient conjugué converge vers la solution du problème de minimisation quadratique en au plus n itérations (la dimension du problème). C'est pourquoi il s'agit d'un algorithme largement utilisé pour résoudre ce type de problèmes. En pratique une solution satisfaisante peut souvent être obtenue en un nombre inférieur d'itérations. L'algorithme du gradient conjugué s'écrit finalement [Golub et Van Loan, 1996] :

- Initialisation ($k = 0$) : choix d'un vecteur $\mathbf{x}_0 \in \mathbb{C}^n$ et calcul de $\mathbf{g}_0 = \mathbf{A}\mathbf{x}_0 - \mathbf{b}$, $\mathbf{w}_0 = \mathbf{g}_0$.
- Itération k :
  - Si $\mathbf{g}_k = 0$ stop ;
  - Sinon :
    1. $\beta_k = (\mathbf{g}_k, \mathbf{w}_k)/(\mathbf{A}\mathbf{w}_k, \mathbf{w}_k)$
    2. $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{w}_k$
    3. $\mathbf{g}_{k+1} = \mathbf{A}x_{k+1} - \mathbf{b}$
    4. $\alpha_{k+1} = -(\mathbf{g}_{k+1}, \mathbf{A}\mathbf{w}_k)/(\mathbf{A}\mathbf{w}_k, \mathbf{w}_k)$
    5. $\mathbf{w}_{k+1} = \mathbf{g}_{k+1} + \alpha_{k+1}\mathbf{w}_k$
  - k = k + 1.

D.2 Gradient conjugué avec préconditionnement

Lorsque la matrice $\mathbf{A}$ est mal conditionnée, il est possible d'améliorer la convergence de l'algorithme en ayant recours à un préconditionnement. Nous présentons ci-dessous la méthode du gradient conjugué avec préconditionnement, où la matrice de préconditionnement $\mathbf{D}$ est choisie comme étant la partie diagonale de $\mathbf{A}$ (préconditionnement de Jacobi). Soient $\mathbf{A}' = \mathbf{D}^{-0.5}\mathbf{A}\mathbf{D}^{-0.5}$, $\mathbf{x}' = \mathbf{D}^{0.5}\mathbf{x}$ et $\mathbf{b}' = \mathbf{D}^{-0.5}\mathbf{b}$, alors résoudre $\mathbf{A}\mathbf{x} = \mathbf{b}$ est équivalent à résoudre $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$. On écrit la méthode du gradient conjugué pour résoudre ce système :

— Initialisation ($k = 0$) : choix d'un vecteur $\mathbf{x}'_0 \in \mathbb{C}^n$ et calcul de $\mathbf{g}'_0 = \mathbf{A}'\mathbf{x}'_0 - \mathbf{b}'$, $\mathbf{w}'_0 = \mathbf{g}'_0$.

— Itération k :

  — Si $\mathbf{g}'_k = 0$ stop ;

  — Sinon :

    1. $\beta_k = (\mathbf{g}'_k, \mathbf{w}'_k)/(\mathbf{A}'\mathbf{w}'_k, \mathbf{w}'_k)$
    2. $\mathbf{x}'_{k+1} = \mathbf{x}'_k - \beta_k \mathbf{w}'_k$
    3. $\mathbf{g}'_{k+1} = \mathbf{A}'\mathbf{x}'_{k+1} - \mathbf{b}'$
    4. $\alpha_{k+1} = -(\mathbf{g}'_{k+1}, \mathbf{A}'\mathbf{w}'_k)/(\mathbf{A}'\mathbf{w}'_k, \mathbf{w}'_k)$
    5. $\mathbf{w}'_{k+1} = \mathbf{g}'_{k+1} + \alpha_{k+1}\mathbf{w}'_k$

  — k = k + 1.

En utilisant $\mathbf{g}'_k = \mathbf{D}^{-0.5}\mathbf{g}_k$, $\mathbf{w}'_k = \mathbf{D}^{0.5}\mathbf{w}_k$ et $\mathbf{x}'_k = \mathbf{D}^{0.5}\mathbf{x}_k$, on peut réécrire l'algorithme de la méthode du gradient conjugué avec préconditionnement de la façon suivante :

— Initialisation ($k = 0$) : choix d'un vecteur $\mathbf{x}_0 \in \mathbb{C}^n$ et calcul de $\mathbf{g}_0 = \mathbf{A}\mathbf{x}_0 - \mathbf{b}$, $\mathbf{w}_0 = \mathbf{D}^{-1}\mathbf{g}_0$.

— Itération k :

  — Si $\mathbf{g}_k = 0$ stop ;

  — Sinon :

    1. $\beta_k = (\mathbf{g}_k, \mathbf{w}_k)/(\mathbf{A}\mathbf{w}_k, \mathbf{w}_k)$
    2. $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{w}_k$
    3. $\mathbf{g}_{k+1} = \mathbf{A}x_{k+1} - \mathbf{b}$
    4. $\alpha_{k+1} = -(\mathbf{D}^{-1}\mathbf{g}_{k+1}, \mathbf{A}\mathbf{w}_k)/(\mathbf{A}\mathbf{w}_k, \mathbf{w}_k)$
    5. $\mathbf{w}_{k+1} = \mathbf{D}^{-1}\mathbf{g}_{k+1} + \alpha_{k+1}\mathbf{w}_k$

  — k = k + 1.

Cette formulation de la méthode du gradient conjugué avec préconditionnement ne fait intervenir à chaque itération qu'une multiplication supplémentaire du gradient par l'inverse d'une matrice diagonale, ce qui est peu coûteux.

Annexe E

Éléments de démonstration des équations (5.22) à (5.26)

Dans cette annexe nous donnons quelques éléments de démonstration des équations (5.22) à (5.26), page 91. En utilisant les équations (5.20) et (A.2), identifier les paramètres de la distribution (5.21), page 90, consiste à résoudre le système d'équations suivant :

$$
\begin{cases}
\dfrac{1}{2(1-\rho_{j,fn}^2)\gamma_{j,fn}^r} = \dfrac{2}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))^2 + \frac{1}{\phi\lambda_{j,fn}^2}; \\
\dfrac{1}{2(1-\rho_{j,fn}^2)}\left(\dfrac{2}{\gamma_{j,fn}^r}\hat{s}_{j,fn}^r - \dfrac{2\rho_{j,fn}}{\sqrt{\gamma_{j,fn}^r\gamma_{j,fn}^\imath}}\hat{s}_{j,fn}^\imath\right) = \dfrac{2}{\phi}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t)) \\
\qquad\qquad \times\left(x_i(t) - \dfrac{2}{\phi}\displaystyle\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right); \\
\dfrac{\rho_{j,fn}}{(1-\rho_{j,fn}^2)}\dfrac{1}{\sqrt{\gamma_{j,fn}^r\gamma_{j,fn}^\imath}} = \dfrac{4}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\Im(g_{ij,fn}(t)); \\
\dfrac{1}{2(1-\rho_{j,fn}^2)\gamma_{j,fn}^\imath} = \dfrac{2}{\phi^2}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))^2 + \frac{1}{\phi\lambda_{j,fn}^2}; \\
-\dfrac{1}{2(1-\rho_{j,fn}^2)}\left(\dfrac{2}{\gamma_{j,fn}^\imath}\hat{s}_{j,fn}^\imath - \dfrac{2\rho_{j,fn}}{\sqrt{\gamma_{j,fn}^r\gamma_{j,fn}^\imath}}\hat{s}_{j,fn}^r\right) = \dfrac{2}{\phi}\displaystyle\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t)) \\
\qquad\qquad \times\left(x_i(t) - \dfrac{2}{\phi}\displaystyle\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right).
\end{cases}
\tag{E.1}
$$

Après résolution nous obtenons les équations (5.22), (5.23) et (5.24), page 91, définissant respectivement $\rho_{j,fn}$, $\gamma_{j,fn}^r$ et $\gamma_{j,fn}^\imath$, et nous obtenons les expressions suivantes des paramètres de

moyenne :

$$\hat{s}^r_{j,fn}=\frac{2}{\phi}\gamma^r_{j,fn}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)$$
$$-\frac{2}{\phi}\rho_{j,fn}\sqrt{\gamma^r_{j,fn}\gamma^i_{j,fn}}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right); \tag{E.2}$$

$$\hat{s}^i_{j,fn}=-\frac{2}{\phi}\gamma^i_{j,fn}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)$$
$$+\frac{2}{\phi}\rho_{j,fn}\sqrt{\gamma^r_{j,fn}\gamma^i_{j,fn}}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right). \tag{E.3}$$

Du fait de leur complexité, nous cherchons maintenant à simplifier les expressions de $\hat{s}^r_{j,fn}$ et $\hat{s}^i_{j,fn}$. On peut montrer d'après les équations (5.22), (5.23), (5.24), (E.2) et (E.3) que les trois égalités suivantes sont vérifiées :

$$\begin{aligned}\hat{s}^i_{j,fn}\rho_{j,fn}\sqrt{\frac{\gamma^r_{j,fn}}{\gamma^i_{j,fn}}} =& -\frac{2}{\phi}\rho_{j,fn}\sqrt{\gamma^r_{j,fn}\gamma^i_{j,fn}}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\\ &\times\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)\\ &+\frac{2}{\phi}\rho^2_{j,fn}\gamma^r_{j,fn}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\\ &\times\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right);\end{aligned} \tag{E.4}$$

$$\begin{aligned}\hat{s}^r_{j,fn}\rho_{j,fn}\sqrt{\frac{\gamma^i_{j,fn}}{\gamma^r_{j,fn}}} =& \frac{2}{\phi}\rho_{j,fn}\sqrt{\gamma^r_{j,fn}\gamma^i_{j,fn}}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\\ &\times\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right)\\ &-\frac{2}{\phi}\rho^2_{j,fn}\gamma^i_{j,fn}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\\ &\times\left(x_i(t)-\frac{2}{\phi}\sum_{j'f'n'\neq jfn}\Re\big(\hat{s}_{j',f'n'}g_{ij',f'n'}(t)\right);\end{aligned} \tag{E.5}$$

$$\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\Im(g_{ij,fn}(t)) = \frac{\rho_{j,fn}}{2\left(1-\rho_{j,fn}^2\right)\sqrt{\gamma_{j,fn}^r\gamma_{j,fn}^i}}. \tag{E.6}$$

En utilisant les équations (E.4) et (E.6) pour réécrire (E.2), et en reconnaissant $d_{j,fn}^r$ défini à l'équation (5.27), nous obtenons l'équation (5.25), page 91. De façon similaire, en utilisant les équations (E.5) et (E.6) pour réécrire (E.3), et en reconnaissant $d_{j,fn}^i$ défini à l'équation (5.28), nous obtenons l'équation (5.26), page 91.

Annexe F

Formulation alternative de la méthode du gradient conjugué pour l'étape E au chapitre 5, section 5.3

Dans cette annexe nous détaillons une formulation alternative de la méthode du gradient conjugué pour l'étape E présentée au chapitre 5, section 5.3 (voir plus précisément page 92). Nous travaillons ici avec des matrices à valeurs réelles.

On introduit tout d'abord les définitions suivantes :

- $\hat{\mathbf{s}} = [(\hat{\mathbf{s}}^r)^\top, (\hat{\mathbf{s}}^\imath)^\top]^\top \in \mathbb{R}^{2JFN}$ avec $\hat{\mathbf{s}}^{(\cdot)} \in \mathbb{R}^{JFN}$ le vecteur colonne d'entrées $\hat{s}^{(\cdot)}_{j,fn}$;
- $\mathbf{d} = [(\mathbf{d}^r)^\top, (\mathbf{d}^\imath)^\top]^\top \in \mathbb{R}^{2JFN}$ avec $\mathbf{d}^{(\cdot)} \in \mathbb{R}^{JFN}$ le vecteur colonne d'entrées $d^{(\cdot)}_{j,fn}$;
- $\mathbf{g}_i(t) \in \mathbb{C}^{JFN}$ le vecteur colonne d'entrées $g_{ij,fn}(t)$.
- $\mathbf{V}$ la matrice diagonale de taille $JFN \times JFN$ et d'entrées $\lambda^2_{j,fn}$.

L'ordre des coefficients indicés par j, f et n pour construire ces vecteurs et cette matrice diagonale n'a pas d'importance, tant qu'il est gardé identique.

On peut montrer d'après les équations (5.27) et (5.28), page 91, que :

$$\begin{aligned}\mathbf{d}^r =& \frac{2}{\phi}\left(\mathbf{V}^{-1} + \frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(\mathbf{g}_i(t))\Re(\mathbf{g}_i(t))^\top\right)\hat{\mathbf{s}}^r - \frac{4}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(\mathbf{g}_i(t))\Im(\mathbf{g}_i(t))^\top\hat{\mathbf{s}}^\imath \\ & - \frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(\mathbf{g}_i(t))x_i(t);\end{aligned} \tag{F.1}$$

$$\begin{aligned}\mathbf{d}^\imath =& \frac{2}{\phi}\left(\mathbf{V}^{-1} + \frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(\mathbf{g}_i(t))\Im(\mathbf{g}_i(t))^\top\right)\hat{\mathbf{s}}^\imath - \frac{4}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(\mathbf{g}_i(t))\Re(\mathbf{g}_i(t))^\top\hat{\mathbf{s}}^r \\ & + \frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(\mathbf{g}_i(t))x_i(t).\end{aligned} \tag{F.2}$$

On cherche le vecteur $\hat{\mathbf{s}}$ qui annule le gradient $\mathbf{d}$. Ceci est équivalent à résoudre le système d'équa-

tions $\mathbf{A}\hat{\mathbf{s}} + \mathbf{c} = \mathbf{0}$ où :

$$\mathbf{A}=\left(\begin{array}{c|c} \frac{2}{\phi}\left(\mathbf{V}^{-1}+\frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(\mathbf{g}_i(t))\Re(\mathbf{g}_i(t))^\top\right) & -\frac{4}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(\mathbf{g}_i(t))\Im(\mathbf{g}_i(t))^\top \\ -\frac{4}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(\mathbf{g}_i(t))\Re(\mathbf{g}_i(t))^\top & \frac{2}{\phi}\left(\mathbf{V}^{-1}+\frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(\mathbf{g}_i(t))\Im(\mathbf{g}_i(t))^\top\right) \end{array}\right); \tag{F.3}$$

$$\mathbf{c} = \frac{2}{\phi}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\begin{pmatrix}-\Re(\mathbf{g}_i(t)) \\ \Im(\mathbf{g}_i(t))\end{pmatrix} x_i(t). \tag{F.4}$$

On définit également la matrice de préconditionnement diagonale $\mathbf{D} = \mathrm{diag}(\mathbf{A})$ qui contient sur la première moitié de sa diagonale les coefficients $[\gamma_{j,fn}^r(1-\rho_{j,fn}^2)]^{-1}$ et sur la seconde moitié $[\gamma_{j,fn}^\imath(1-\rho_{j,fn}^2)]^{-1}$. La méthode du gradient conjugué avec préconditionnement est détaillée dans l'algorithme 3.

Algorithme 3 Mise à jour de $\{\hat{s}_{j,fn}^r, \hat{s}_{j,fn}^\imath\}_{j,f,n}$ à l'étape E par la méthode du gradient conjugué avec préconditionnement

Initialiser $\mathbf{d}$ à partir des équations (5.27) et (5.28) et $\boldsymbol{\omega} = \mathbf{D}^{-1}\mathbf{d}$
Tant que critère d'arrêt non rencontré
    $\boldsymbol{\kappa} = \mathbf{A}\boldsymbol{\omega}$
    $\mu = (\boldsymbol{\omega}^\top\mathbf{d})/(\boldsymbol{\omega}^\top\boldsymbol{\kappa})$
    $\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} - \mu\boldsymbol{\omega}$
    Calculer $\mathbf{d}$ à partir des équations (5.27) et (5.28)
    $\mathbf{d}_p = \mathbf{D}^{-1}\mathbf{d}$
    $\beta = -(\boldsymbol{\kappa}^\top\mathbf{d}_p)/(\boldsymbol{\omega}^\top\boldsymbol{\kappa})$
    $\boldsymbol{\omega} \leftarrow \mathbf{d}_p + \beta\boldsymbol{\omega}$
Fin Tant que

Annexe G

Détails de calcul pour l'algorithme VEM du chapitre 6

Nous détaillons dans cette annexe les calculs de l'algorithme VEM présenté au chapitre 6, section 6.2, à partir de la page 112.

G.1 Log-vraisemblance des données complètes

D'après le modèle défini au chapitre 6, section 6.1, la log-vraisemblance des données complètes s'écrit :

$$
\begin{aligned}
\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) &= \ln p(\mathbf{x}|\mathbf{z};\boldsymbol{\sigma}) + \ln p(\mathbf{s}|\mathbf{v};\boldsymbol{\lambda}) + \ln p(\mathbf{v}) + \ln p(\mathbf{a}|\mathbf{u}) + \ln p(\mathbf{u}) \\
&= -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] \\
&\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\left[\ln\left(2\pi v_{j,fn}\lambda_{j,fn}^2\right) + \frac{s_{j,fn}^2}{v_{j,fn}\lambda_{j,fn}^2}\right] \\
&\quad -\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\left[\ln\Gamma\left(\frac{\alpha_v}{2}\right) + \frac{\alpha_v}{2}\ln\left(\frac{2}{\alpha_v}\right) + \left(\frac{\alpha_v}{2}+1\right)\ln(v_{j,fn}) + \frac{\alpha_v}{2}\frac{1}{v_{j,fn}}\right] \\
&\quad -\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=0}^{L_a-1}\left[\ln\left(2\pi u_{ij}(t)r^2(t)\right) + \frac{a_{ij}^2(t)}{u_{ij}(t)r^2(t)}\right] \\
&\quad -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=0}^{L_a-1}\left[\ln\Gamma\left(\frac{\alpha_u}{2}\right) + \frac{\alpha_u}{2}\ln\left(\frac{2}{\alpha_u}\right) + \left(\frac{\alpha_u}{2}+1\right)\ln(u_{ij}(t)) + \frac{\alpha_u}{2}\frac{1}{u_{ij}(t)}\right].
\end{aligned}
\tag{G.1}
$$

G.2 Étape E

G.2.1 Étape E-V

On peut montrer en utilisant l'équation (G.1) que :

$$\ln q^{\star}(v_{j,fn}) \stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\setminus v_{j,fn})}$$

$$\stackrel{c}{=} -\ln(v_{j,fn})\left(\frac{\alpha_v+1}{2}+1\right)-\frac{1}{v_{j,fn}}\left(\frac{\alpha_v}{2}+\frac{\langle s_{j,fn}^2\rangle_{q(s_{j,fn})}}{2\lambda_{j,fn}^2}\right). \tag{G.2}$$

On reconnaît $q^\star(v_{j,fn}) = IG(\nu_v, \beta_{j,fn})$ avec :

$$\nu_v = \frac{\alpha_v+1}{2}; \tag{G.3}$$

$$\beta_{j,fn} = \frac{\alpha_v}{2}+\frac{\langle s_{j,fn}^2\rangle_{q(s_{j,fn})}}{2\lambda_{j,fn}^2}, \tag{G.4}$$

où $\langle s_{j,fn}^2\rangle_{q(s_{j,fn})} = \hat{s}_{j,fn}^2 + \gamma_{j,fn}$.

G.2.2 Étape E-U

De façon tout à fait similaire on trouve :

$$\begin{aligned}\ln q^\star(u_{ij}(t)) &\stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\setminus u_{ij}(t))}\\ &\stackrel{c}{=} -\ln(u_{ij}(t))\left(\frac{\alpha_u+1}{2}+1\right)-\frac{1}{u_{ij}(t)}\left(\frac{\alpha_u}{2}+\frac{\langle a_{ij}^2(t)\rangle_{q(a_{ij}(t))}}{2r^2(t)}\right).\end{aligned} \tag{G.5}$$

On reconnaît $q^\star(u_{ij}(t)) = IG(\nu_u, d_{ij}(t))$ avec :

$$\nu_u = \frac{\alpha_u+1}{2}; \tag{G.6}$$

$$d_{ij}(t) = \frac{\alpha_u}{2}+\frac{\langle a_{ij}^2(t)\rangle_{q(a_{ij}(t))}}{2r^2(t)}, \tag{G.7}$$

où $\langle a_{ij}^2(t)\rangle_{q(a_{ij}(t))} = \hat{a}_{ij}^2(t) + \rho_{ij}(t)$.

G.2.3 Étape E-S

On définit $\mathrm{Var}_q(\mathbf{z}) = \left\langle (\mathbf{z}-\langle\mathbf{z}\rangle_q)(\mathbf{z}-\langle\mathbf{z}\rangle_q)^\top\right\rangle_q = \left\langle \mathbf{z}\mathbf{z}^\top\right\rangle_q - \langle\mathbf{z}\rangle_q\langle\mathbf{z}\rangle_q^\top$ où $\mathbf{z}$ est un vecteur aléatoire réel.

Toujours d'après l'expression de la log-vraisemblance des données complètes à l'équation (G.1), on a :

$$\begin{aligned}\ln q^\star(s_{j,fn}) &\stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\setminus s_{j,fn})}\\ &\stackrel{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\left\langle\left\|\mathbf{x}_i-\sum_{j'=1}^{J}\sum_{(f',n')\in\mathcal{B}_{j'}} s_{j',f'n'}\mathbf{T}_{j',f'n'}\mathbf{a}_{ij'}\right\|_2^2\right\rangle_{q(\mathbf{z}\setminus s_{j,fn})}\\ &\quad -\frac{1}{2}\frac{\left\langle v_{j,fn}^{-1}\right\rangle_{q(v_{j,fn})}}{\lambda_{j,fn}^2}s_{j,fn}^2,\end{aligned} \tag{G.8}$$

où l'on rappelle que $\mathbf{x}_i = [x_i(t)]_t^\top \in \mathbb{R}^T$, $\mathbf{a}_{ij} = [a_{ij}(t)]_t^\top \in \mathbb{R}^{L_a}$ et $\mathbf{T}_{j,fn} \in \mathbb{R}^{T\times L_a}$ est la matrice de Toeplitz représentant le produit de convolution de $\psi_{j,fn}(t)$, $t = 0, ..., L_s-1$, avec un signal de longueur L_a. Dans la suite et lorsque cela ne porte pas à confusion, nous omettons de noter en indice la densité de probabilité par rapport à laquelle sont calculées les moyennes et variances.

A des constantes additives près par rapport à $s_{j,fn}$, on peut montrer que :

$$\begin{aligned}
&\left\langle \left\| \mathbf{x}_i - \sum_{j=1}^{J} \sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\mathbf{T}_{j,fn}\mathbf{a}_{ij} \right\|_2^2 \right\rangle \\
&\qquad \stackrel{c}{=} \left\| \mathbf{x}_i - \sum_{j=1}^{J} \sum_{(f,n)\in\mathcal{B}_j} \langle s_{j,fn}\mathbf{T}_{j,fn}\mathbf{a}_{ij} \rangle \right\|_2^2 + \operatorname{trace}\left[\operatorname{Var}\left(\sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\mathbf{T}_{j,fn}\mathbf{a}_{ij} \right) \right] \\
&\qquad \stackrel{c}{=} \left\| \mathbf{x}_i - s_{j,fn}\mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij} - \sum_{(j',f',n')\neq(j,f,n)} \hat{s}_{j',f'n'}\mathbf{T}_{j',f'n'}\hat{\mathbf{a}}_{ij'} \right\|_2^2 \\
&\qquad\qquad + s_{j,fn}^2 \operatorname{trace}\left[\operatorname{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn} \right] \\
&\qquad\qquad\qquad + 2 s_{j,fn} \sum_{(f',n')\neq(f,n)} \hat{s}_{j,f'n'} \operatorname{trace}\left[\operatorname{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,f'n'} \right], \qquad \text{(G.9)}
\end{aligned}$$

où $\hat{\mathbf{a}}_{ij} = \langle \mathbf{a}_{ij} \rangle$. En injectant l'équation (G.9) dans l'équation (G.8) et en développant on identifie $q^\star(s_{j,fn}) = N(\hat{s}_{j,fn}, \gamma_{j,fn})$ avec :

$$\gamma_{j,fn} = \left[\frac{\left\langle v_{j,fn}^{-1} \right\rangle}{\lambda_{j,fn}^2} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \operatorname{trace}\left[\left(\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \operatorname{Var}(\mathbf{a}_{ij}) \right) \mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn} \right] \right]^{-1}; \qquad \text{(G.10)}$$

$$\hat{s}_{j,fn} = \hat{s}_{j,fn} - \gamma_{j,fn}\Delta\hat{s}_{j,fn}; \qquad \text{(G.11)}$$

$$\Delta\hat{s}_{j,fn} = \hat{s}_{j,fn}\frac{\left\langle v_{j,fn}^{-1} \right\rangle}{\lambda_{j,fn}^2} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \left[\operatorname{trace}\left(\operatorname{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^\top \hat{\mathbf{S}}_j \right) - \hat{\mathbf{g}}_{ij,fn}^\top \left(\mathbf{x}_i - \sum_{j'=1}^{J} \hat{\mathbf{y}}_{ij'} \right) \right], \qquad \text{(G.12)}$$

où l'on rappelle que $\hat{\mathbf{g}}_{ij,fn} = \mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij} = [\hat{g}_{ij,fn}(t)]_t^\top \in \mathbb{R}^T$, $\hat{\mathbf{y}}_{ij} = [\hat{y}_{ij}(t)]_t^\top \in \mathbb{R}^T$ et $\hat{\mathbf{S}}_j = \langle \mathbf{S}_j \rangle$ avec $\mathbf{S}_j \in \mathbb{R}^{T\times L_a}$ la matrice de Toeplitz permettant de représenter le produit de convolution du signal source $s_j(t)$, $t = 0, ..., L_s$, avec un signal de longueur L_a.

a) Simplification du terme $\Delta\hat{s}_{j,fn}$

D'après l'approximation de champ moyen, $\operatorname{Var}(\mathbf{a}_{ij}) = \operatorname{diag}(\{\rho_{ij}(t)\}_{t=0,..,L_a-1})$. Nous pouvons alors simplifier le terme suivant intervenant dans l'expression de $\Delta\hat{s}_{j,fn}$:

$$\begin{aligned}
\operatorname{trace}\left(\operatorname{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^\top \hat{\mathbf{S}}_j \right) &= \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \left(\mathbf{T}_{j,fn}^\top \hat{\mathbf{S}}_j \right)_{\tau,\tau} \\
&= \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \sum_{(f',n')\in\mathcal{B}_j} \hat{s}_{j,f'n'} \left(\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,f'n'} \right)_{\tau,\tau} \\
&= \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \sum_{(f',n')\in\mathcal{B}_j} \hat{s}_{j,f'n'} \sum_{t=0}^{L_s-1} \psi_{j,fn}(t)\psi_{j,f'n'}(t) \\
&= \hat{s}_{j,fn} \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau). \qquad \text{(G.13)}
\end{aligned}$$

La dernière égalité s'obtient en utilisant le fait que les atomes MDCT sont orthonormés. $\Delta\hat{s}_{j,fn}$ se réécrit donc finalement de la façon suivante :

$$\Delta\hat{s}_{j,fn} = \hat{s}_{j,fn}\left[\frac{\left\langle v_{j,fn}^{-1}\right\rangle}{\lambda_{j,fn}^2} + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\right] - \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\hat{\mathbf{g}}_{ij,fn}^{\top}\left(\mathbf{x}_i - \sum_{j'=1}^{J}\hat{\mathbf{y}}_{ij'}\right). \quad \text{(G.14)}$$

b) Simplification du terme de variance $\gamma_{j,fn}$

De façon similaire nous pouvons montrer que $\mathrm{trace}\left[\mathrm{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^{\top}\mathbf{T}_{j,fn}\right] = \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)$. Nous avons également $\mathrm{trace}\left[\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^{\top}\mathbf{T}_{j,fn}^{\top}\mathbf{T}_{j,fn}\right] = \hat{\mathbf{g}}_{ij,fn}^{\top}\hat{\mathbf{g}}_{ij,fn} = \|\hat{\mathbf{g}}_{ij,fn}\|_2^2$. On peut donc réécrire $\gamma_{j,fn}$ de la façon suivante :

$$\gamma_{j,fn} = \left[\frac{\left\langle v_{j,fn}^{-1}\right\rangle}{\lambda_{j,fn}^2} + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\left(\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau) + \|\hat{\mathbf{g}}_{ij,fn}\|_2^2\right)\right]^{-1}. \quad \text{(G.15)}$$

Enfin, d'après l'étape E-V $\left\langle v_{j,fn}^{-1}\right\rangle = \nu_v/\beta_{j,fn}$.

c) Méthode du gradient conjugué avec préconditionnement

L'équation (G.11) définit un ensemble de solutions couplées pour les variables $\{\hat{s}_{j,fn}\}_{j,f,n}$. Ces dernières doivent donc être mises à jour séquentiellement. Cependant nous pouvons utiliser comme précédemment la méthode du gradient conjugué avec préconditionnement pour mettre à jour l'ensemble des variables $\{\hat{s}_{j,fn}\}_{j,f,n}$ et donc accélérer l'étape E-S. Il suffit pour cela de remarquer que $\Delta\hat{s}_{j,fn}$ est la dérivée partielle de l'opposé de l'énergie variationnelle libre (le critère que l'on optimise et défini au chapitre 6, section 6.2.3) par rapport à $\hat{s}_{j,fn}$.

G.2.4 Étape E-A

En utilisant à nouveau l'expression de la log-vraisemblance des données complètes nous avons :

$$\begin{aligned}\ln q^{\star}(a_{ij}(t)) &\stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\setminus a_{ij}(t))} \\ &\stackrel{c}{=} -\frac{1}{2}\frac{1}{\sigma_i^2}\left\langle\left\|\mathbf{x}_i - \sum_{j'=1}^{J}\mathbf{S}_{j'}\mathbf{a}_{ij'}\right\|_2^2\right\rangle_{q(\mathbf{z}\setminus a_{ij}(t))} - \frac{1}{2}\frac{\left\langle u_{ij}^{-1}(t)\right\rangle_{q(u_{ij}(t))}}{r^2(t)}a_{ij}^2(t). \quad \text{(G.16)}\end{aligned}$$

Soient $\mathbf{a}_{ij}(t), \hat{\tilde{\mathbf{a}}}_{ij}(t) \in \mathbb{R}^{L_a}$ deux vecteurs colonnes tels que :

$$(\mathbf{a}_{ij}(t))_{t'} = \begin{cases} a_{ij}(t) & \text{si } t' = t; \\ 0 & \text{sinon,}\end{cases} \quad \text{(G.17)}$$

$$(\hat{\tilde{\mathbf{a}}}_{ij}(t))_{t'} = \begin{cases} 0 & \text{si } t' = t; \\ [\hat{\mathbf{a}}_{ij}]_{t'} & \text{sinon.}\end{cases} \quad \text{(G.18)}$$

A des constantes additives près par rapport à $a_{ij}(t)$ on peut montrer que :

$$\begin{aligned}\left\langle \left\| \mathbf{x}_i - \sum_{j'=1}^{J} \mathbf{S}_{j'} \mathbf{a}_{ij'} \right\|_2^2 \right\rangle &\overset{c}{=} \left\| \mathbf{x}_i - \hat{\mathbf{S}}_j \left(\mathbf{a}_{ij}(t) + \hat{\tilde{\mathbf{a}}}_{ij}(t) \right) - \sum_{j' \neq j} \hat{\mathbf{y}}_{ij'} \right\|_2^2 + \operatorname{trace}\left[\operatorname{Var}\left(\mathbf{S}_j \mathbf{a}_{ij} \right) \right] \\ &\overset{c}{=} \left\| \mathbf{x}_i - \hat{\mathbf{S}}_j \left(\mathbf{a}_{ij}(t) + \hat{\tilde{\mathbf{a}}}_{ij}(t) \right) - \sum_{j' \neq j} \hat{\mathbf{y}}_{ij'} \right\|_2^2 \\ &\quad + \operatorname{trace}\left[\mathbf{a}_{ij}(t) \mathbf{a}_{ij}(t)^\top \operatorname{Var}\left(\mathbf{S}_j^\top \right) + 2 \mathbf{a}_{ij}(t) \hat{\tilde{\mathbf{a}}}_{ij}(t)^\top \operatorname{Var}\left(\mathbf{S}_j^\top \right) \right].\end{aligned} \tag{G.19}$$

En injectant l'équation (G.19) dans l'équation (G.16) et en développant on peut alors montrer que $q^\star(a_{ij}(t)) = N\left(\hat{a}_{ij}(t), \rho_{ij}(t)\right)$ avec :

$$\rho_{ij}(t) = \left[\frac{\left\langle u_{ij}^{-1}(t) \right\rangle}{r^2(t)} + \frac{1}{\sigma_i^2} \left(\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \operatorname{Var}\left(\mathbf{S}_j^\top \right) \right)_{t,t} \right]^{-1}; \tag{G.20}$$

$$\hat{a}_{ij}(t) = \hat{a}_{ij}(t) - \rho_{ij}(t) \Delta \hat{a}_{ij}(t); \tag{G.21}$$

$$\Delta \hat{a}_{ij}(t) = \hat{a}_{ij}(t) \frac{\left\langle u_{ij}^{-1}(t) \right\rangle}{r^2(t)} + \frac{1}{\sigma_i^2} \left(\left(\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \operatorname{Var}\left(\mathbf{S}_j^\top \right) \right) \hat{\mathbf{a}}_{ij} - \hat{\mathbf{S}}_j^\top \left(\mathbf{x}_i - \sum_{j' \neq j} \hat{\mathbf{y}}_{ij'} \right) \right)_t, \tag{G.22}$$

où d'après l'étape E-U $\left\langle u_{ij}^{-1}(t) \right\rangle = \nu_u / d_{ij}(t)$.

a) Réécriture du terme $\Delta \hat{a}_{ij}(t)$

D'après l'approximation de champ moyen nous pouvons écrire :

$$\operatorname{Var}\left(\mathbf{S}_j^\top \right) = \operatorname{Var}\left(\sum_{(f,n) \in \mathcal{B}_j} s_{j,fn} \mathbf{T}_{j,fn}^\top \right) = \sum_{(f,n) \in \mathcal{B}_j} \gamma_{j,fn} \mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}.$$

On remarque alors que $\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \operatorname{Var}\left(\mathbf{S}_j^\top \right)$ correspond à une matrice de Toeplitz symétrique donnée par :

$$\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \operatorname{Var}\left(\mathbf{S}_j^\top \right) = T_{L_a} \left\{ \hat{r}_j^{ss}(k) + \sum_{(f,n) \in \mathcal{B}_j} \gamma_{j,fn} \hat{r}_{j,fn}^{\psi\psi}(k) \right\}, \tag{G.23}$$

où l'on rappelle que $T_M\{\tau(k)\}$ représente une matrice de Toeplitz symétrique de taille $M \times M$ et formée à partir de la séquence $\{\tau(k)\}_{k=0}^{M-1}$, $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k} \hat{s}_j(t) \hat{s}_j(t+k)$ et $\hat{r}_{j,fn}^{\psi\psi}(k) = \sum_{t=0}^{L_s-1-k} \psi_{j,fn}(t) \psi_{j,fn}(t+k)$.

On rappelle de plus que :

$$\hat{\mathbf{S}}_j^\top \left(\mathbf{x}_i - \sum_{j' \neq j} \hat{\mathbf{y}}_{ij'} \right) = \hat{\mathbf{r}}_{ij}^{s\epsilon}, \tag{G.24}$$

où $\hat{\mathbf{r}}_{ij}^{s\epsilon} = \left[\hat{r}_{ij}^{s\epsilon}(k)\right]_k^\top \in \mathbb{R}^{L_a}$ avec $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1} \hat{s}_j(t)\epsilon_{ij}(t+k)$ et $\epsilon_{ij}(t) = x_i(t) - \sum_{j'\neq j} \hat{y}_{ij'}(t)$. Finalement, en injectant les équations (G.23) et (G.24) dans (G.22), $\Delta\hat{a}_{ij}(t)$ se réécrit de la façon suivante :

$$\Delta\hat{a}_{ij}(t) = \hat{a}_{ij}(t)\frac{\left\langle u_{ij}^{-1}(t)\right\rangle}{r^2(t)} + \frac{1}{\sigma_i^2}\left(T_{L_a}\left\{\hat{r}_j^{ss}(k) + \sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\hat{r}_{j,fn}^{\psi\psi}(k)\right\}\hat{\mathbf{a}}_{ij} - \hat{\mathbf{r}}_{ij}^{s\epsilon}\right)_t . \tag{G.25}$$

b) Réécriture du terme de variance $\rho_{ij}(t)$

D'après l'équation (G.23) et en utilisant le fait que $\hat{r}_{j,fn}^{\psi\psi}(0) = 1$ car les atomes MDCT sont orthonormés et que $\hat{r}_j^{ss}(0) = \|\hat{\mathbf{s}}_j\|_2^2$ avec $\hat{\mathbf{s}}_j = [\hat{s}_j(t)]_t \in \mathbb{R}^{L_s}$, le terme de variance $\rho_{ij}(t)$ à l'équation (G.20) se simplifie comme suit :

$$\rho_{ij}(t) = \left[\frac{\left\langle u_{ij}^{-1}(t)\right\rangle}{r^2(t)} + \frac{1}{\sigma_i^2}\left(\|\hat{\mathbf{s}}_j\|_2^2 + \sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\right)\right]^{-1} . \tag{G.26}$$

c) Méthode du gradient conjugué avec préconditionnement

Les équations résultant de l'étape E-A définissent un ensemble de solutions couplées pour les variables $\{\hat{a}_{ij}(t)\}_{i,j,t}$. Ces dernières doivent donc être mises à jour séquentiellement. Cependant, comme à l'étape E-S, nous pouvons utiliser la méthode du gradient conjugué avec préconditionnement. Il suffit pour cela de remarquer que $\Delta\hat{a}_{ij}(t)$ est la dérivée partielle de l'opposé de l'énergie variationnelle libre par rapport à $\hat{a}_{ij}(t)$.

G.3 Énergie variationnelle libre

L'énergie variationnelle libre est définie au chapitre 6, section 6.2.3 (voir page 114). Nous présentons dans cette section les détails de calcul du terme $\bar{e}_i$ impliqué dans l'expression de $\langle \ln p(\mathbf{x}|\mathbf{s},\mathbf{a};\boldsymbol{\sigma})\rangle$. D'après l'approximation de champ moyen, $\bar{e}_i$ s'écrit de la façon suivante :

$$\begin{aligned}
\bar{e}_i &= \left\langle \left\| \mathbf{x}_i - \sum_{j=1}^{J}\mathbf{y}_{ij}\right\|_2^2\right\rangle \\
&= \left\| \mathbf{x}_i - \left\langle \sum_{j=1}^{J}\mathbf{y}_{ij}\right\rangle\right\|_2^2 + \operatorname{trace}\left[\operatorname{Var}\left(\sum_{j=1}^{J}\mathbf{y}_{ij}\right)\right] \\
&= \left\| \mathbf{x}_i - \sum_{j=1}^{J}\hat{\mathbf{y}}_{ij}\right\|_2^2 + \sum_{j=1}^{J}\operatorname{trace}\left[\operatorname{Var}\left(\mathbf{y}_{ij}\right)\right] .
\end{aligned} \tag{G.27}$$

Or toujours d'après l'approximation de champ moyen,

$$\begin{aligned}
\operatorname{trace}\left[\operatorname{Var}\left(\mathbf{y}_{ij}\right)\right] &= \operatorname{trace}\left[\operatorname{Var}\left(\mathbf{S}_j\mathbf{a}_{ij}\right)\right] \\
&= \operatorname{trace}\left[\left\langle \mathbf{S}_j\mathbf{a}_{ij}\mathbf{a}_{ij}^\top\mathbf{S}_j^\top\right\rangle - \left\langle \mathbf{S}_j\mathbf{a}_{ij}\right\rangle\left\langle \mathbf{a}_{ij}^\top\mathbf{S}_j^\top\right\rangle\right] \\
&= \operatorname{trace}\left[\left\langle \mathbf{a}_{ij}\mathbf{a}_{ij}^\top\right\rangle\left\langle \mathbf{S}_j^\top\mathbf{S}_j\right\rangle - \hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top\hat{\mathbf{S}}_j^\top\hat{\mathbf{S}}_j\right] .
\end{aligned} \tag{G.28}$$

De plus,

$$\begin{aligned}\left\langle \mathbf{a}_{ij}\mathbf{a}_{ij}^\top \right\rangle &= \hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \mathrm{Var}(\mathbf{a}_{ij});\\ \left\langle \mathbf{S}_j^\top \mathbf{S}_j \right\rangle &= \hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \mathrm{Var}\left(\sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\mathbf{T}_{j,fn}^\top\right)\\ &= \hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \sum_{(f,n)\in\mathcal{B}_j} \gamma_{j,fn}\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}.\end{aligned}$$

On a donc :

$$\begin{aligned}&\mathrm{trace}\left[\mathrm{Var}\left(\mathbf{y}_{ij}\right)\right]\\ &\quad= \mathrm{trace}\left[\left(\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \mathrm{Var}(\mathbf{a}_{ij})\right)\left(\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j + \sum_{(f,n)\in\mathcal{B}_j} \gamma_{j,fn}\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right) - \hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top \hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j\right]\\ &\quad= \mathrm{trace}\left[\mathrm{Var}(\mathbf{a}_{ij})\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j\right] + \sum_{(f,n)\in\mathcal{B}_j} \gamma_{j,fn}\,\mathrm{trace}\left[\left(\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \mathrm{Var}(\mathbf{a}_{ij})\right)\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right].\end{aligned} \tag{G.29}$$

On obtient alors :

$$\begin{aligned}\bar{e}_i = \left\|\mathbf{x}_i - \sum_{j=1}^{J}\hat{\mathbf{y}}_{ij}\right\|_2^2 &+ \sum_{j=1}^{J}\mathrm{trace}\left[\mathrm{Var}(\mathbf{a}_{ij})\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j\right]\\ &+ \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\,\mathrm{trace}\left[\left(\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \mathrm{Var}(\mathbf{a}_{ij})\right)\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right].\end{aligned} \tag{G.30}$$

De plus, $\mathrm{Var}(\mathbf{a}_{ij}) = \mathrm{diag}(\{\rho_{ij}(t)\}_{t=0,..,L_a-1})$. Nous pouvons alors simplifier le terme suivant intervenant dans l'expression de $\bar{e}_i$:

$$\begin{aligned}\mathrm{trace}\left[\mathrm{Var}(\mathbf{a}_{ij})\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j\right] &= \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\left(\hat{\mathbf{S}}_j^\top \hat{\mathbf{S}}_j\right)_{\tau,\tau}\\ &= \|\hat{\mathbf{s}}_j\|_2^2\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau).\end{aligned} \tag{G.31}$$

Nous pouvons également montrer en utilisant le fait que les atomes de MDCT sont othonormés que :

$$\begin{aligned}&\mathrm{trace}\left[\left(\hat{\mathbf{a}}_{ij}\hat{\mathbf{a}}_{ij}^\top + \mathrm{Var}(\mathbf{a}_{ij})\right)\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right]\\ &\qquad= \mathrm{trace}\left[\mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij}(\mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij})^\top\right] + \mathrm{trace}\left[\mathrm{Var}(\mathbf{a}_{ij})\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right]\\ &\qquad= (\mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij})^\top(\mathbf{T}_{j,fn}\hat{\mathbf{a}}_{ij}) + \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\left(\mathbf{T}_{j,fn}^\top \mathbf{T}_{j,fn}\right)_{\tau,\tau}\\ &\qquad= \|\hat{\mathbf{g}}_{ij,fn}\|_2^2 + \sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau).\end{aligned} \tag{G.32}$$

En utilisant les équations (G.31) et (G.32) pour réécrire (G.30) on obtient l'expression finale de $\bar{e}_i$ donnée à l'équation (6.19).

Bibliographie

Adali, T., P. J. Schreier et L. L. Scharf. 2011, «Complex-valued signal processing : The proper way to deal with impropriety», *IEEE Transactions on Signal Processing*, vol. 59, n° 11, p. 5101–5125. (page 26, 67, 89, 93, 137)

Adiloğlu, K. et E. Vincent. 2016, «Variational Bayesian inference for source separation and robust feature extraction», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, n° 10, p. 1746–1758. (page 133)

Aissa-El-Bey, A., K. Abed-Meraim et Y. Grenier. 2007, «Blind separation of underdetermined convolutive mixtures using their time–frequency representation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 5, p. 1540–1550. (page 24)

Allen, J. B. et D. A. Berkley. 1979, «Image method for efficiently simulating small-room acoustics», *Journal of the Acoustical Society of America*, vol. 65, n° 4, p. 943–950. (page 53)

Andrews, D. F. et C. L. Mallows. 1974, «Scale mixtures of normal distributions», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, n° 1, p. 99–102. (page 27)

Arberet, S., A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot et P. Vandergheynst. 2010, «Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation», dans *Actes de International Conference on Information Science, Signal Processing and their Applications*, p. 1–4. (page 37)

Arberet, S. et P. Vandergheynst. 2014, «Reverberant audio source separation via sparse and low-rank modeling», *IEEE Signal Processing Letters*, vol. 21, n° 4, p. 404–408. (page 39, 108, 120)

Attias, H. 2003, «New EM algorithms for source separation and deconvolution with a microphone array», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. V – 297–300. (page 38)

Avargel, Y. et I. Cohen. 2007a, «On multiplicative transfer function approximation in the short-time Fourier transform domain», *IEEE Signal Processing Letters*, vol. 14, n° 5, p. 337–340. (page 13)

Avargel, Y. et I. Cohen. 2007b, «System identification in the short-time fourier transform domain with crossband filtering», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 4, p. 1305–1319. (page 38)

Badeau, R. 2016, «Preservation of whiteness in spectral and time-frequency transforms of second order processes», Rapport de recherche, Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI. (page 87)

Badeau, R., N. Bertin et E. Vincent. 2010, «Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization», *IEEE Transactions on Neural Networks*, vol. 21, n° 12, p. 1869–1881. (page 32)

Badeau, R. et M. D. Plumbley. 2014, «Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, n° 11, p. 1670–1680. (page 26, 38)

Barchiesi, D. et M. D. Plumbley. 2011, «Dictionary learning of convolved signals», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5812–5815. (page 133)

Barzilai, J. et J. M. Borwein. 1988, «Two-point step size gradient methods», *IMA Journal of Numerical Analalysis*, vol. 8, n° 1, p. 141–148. (page 72)

Belouchrani, A., K. Abed-Meraim, J.-F. Cardoso et E. Moulines. 1997, «A blind source separation technique using second-order statistics», *IEEE Transactions on signal processing*, vol. 45, n° 2, p. 434–444. (page 22)

Benaroya, L., L. Mcdonagh, F. Bimbot et R. Gribonval. 2003, «Non negative sparse representation for Wiener based source separation with a single sensor», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 613–616. (page 34)

Benichoux, A., L. S. R. Simon, E. Vincent et R. Gribonval. 2014, «Convex regularizations for the simultaneous recording of room impulse responses», *IEEE Transactions on Signal Processing*, vol. 62, n° 8, p. 1976–1986. (page 108)

Bertin, N., E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury et E. Jamet. 2016, «A French corpus for distant-microphone speech processing in real homes», dans *Interspeech 2016*. (page 133)

Bies, D. A. et C. H. Hansen. 2009, *Engineering noise control : theory and practice*, CRC press. (page 59, 84, 139)

Bingham, E. et A. Hyvärinen. 2000, «A fast fixed-point algorithm for independent component analysis of complex valued signals», *International journal of neural systems*, vol. 10, n° 01, p. 1–8. (page 22)

Bishop, C. M. 2006, *Pattern Recognition and Machine Learning*, Springer. (page 41, 42, 43, 137)

Bittner, R. M., J. Salamon, M. Tierney, M. Mauch, C. Cannam et J. P. Bello. 2014, «MedleyDB : A multitrack dataset for annotation-intensive MIR research.», dans *Actes de International Society for Music Information Retrieval (ISMIR) Conference*, p. 155–160. (page 132)

Bongiovanni, G., P. Corsini et G. Frosini. 1976, «One-dimensional and two-dimensional generalised discrete Fourier transforms», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, n° 1, p. 97–99. (page 87)

Carabias-Orti, J. J., M. Cobos, P. Vera-Candeas et F. J. Rodríguez-Serrano. 2013, «Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings», *EURASIP Journal on Advances in Signal Processing*, vol. 2013, n° 1, p. 184. (page 5)

Cardoso, J.-F. 1998a, «Blind signal separation : statistical principles», *Proceedings of the IEEE*, vol. 86, n° 10, p. 2009–2025. (page 21, 22, 25)

Cardoso, J.-F. 1998b, «Multidimensional independent component analysis», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, p. 1941–1944. (page 6, 22)

Carlin, B. P. et T. A. Louis. 2008, *Bayesian methods for data analysis*, CRC Press. (page 138)

Cemgil, A. T., C. Févotte et S. J. Godsill. 2007, «Variational and stochastic inference for Bayesian source separation», *Digital Signal Processing*, vol. 17, n° 5, p. 891–913. (page 24, 99, 132)

Chion, M. 1983, *Guide des objets sonores : Pierre Schaeffer et la recherche musicale*, Buchet/Chastel. (page 4)

Cichocki, A. et S.-I. Amari. 2010, «Families of alpha-beta-and gamma-divergences : Flexible and robust measures of similarities», *Entropy*, vol. 12, n° 6, p. 1532–1568. (page 31)

Comon, P. 1994, «Independent component analysis, a new concept ?», *Signal processing*, vol. 36, n° 3, p. 287–314. (page 22)

Şimşekli, U., A. Liutkus et A. T. Cemgil. 2015, «Alpha-stable matrix factorization», *IEEE Signal Processing Letters*, vol. 22, n° 12, p. 2289–2293. (page 25, 27, 28, 29, 33)

Darmois, G. 1953, «Analyse générale des liaisons stochastiques : etude particulière de l'analyse factorielle linéaire», *Revue de l'Institut international de statistique*, vol. 21, n° 1/2, p. 2–8. (page 22)

Deleforge, A., F. Forbes et R. Horaud. 2015, «Acoustic space learning for sound-source separation and localization on binaural manifolds», *International journal of neural systems*, vol. 25, n° 1, p. 1440 003. (page 24, 133)

Dempster, A. P., N. M. Laird et D. B. Rubin. 1977, «Maximum likelihood from incomplete data via the EM algorithm», *Journal of the royal statistical society. Series B (methodological)*, vol. 39, n° 1, p. 1–38. (page 41, 42)

Duong, N. Q., E. Vincent et R. Gribonval. 2010, «Under-determined reverberant audio source separation using a full-rank spatial covariance model», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 7, p. 1830–1840. (page 37, 38, 80, 84, 120)

Duong, N. Q., E. Vincent et R. Gribonval. 2013, «Spatial location priors for Gaussian model based reverberant audio source separation», *EURASIP Journal on Advances in Signal Processing*, vol. 2013, p. 149. (page 37, 80)

Durbin, J. 1959, «Efficient estimation of parameters in moving-average models», *Biometrika*, vol. 46, n° 3/4, p. 306–316. (page 61)

Durrieu, J.-L., B. David et G. Richard. 2011, «A musically motivated mid-level representation for pitch estimation and musical audio source separation», *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, n° 6, p. 1180–1191. (page 5, 6, 25)

Durrieu, J.-L., G. Richard, B. David et C. Févotte. 2010, «Source/filter model for unsupervised main melody extraction from polyphonic audio signals», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 3, p. 564–575. (page 6)

Emiya, V., E. Vincent, N. Harlander et V. Hohmann. 2011, «Subjective and objective quality assessment of audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n° 7, p. 2046–2057. (page 47)

Feng, F. et M. Kowalski. 2014, «Hybrid model and structured sparsity for under-determined convolutive audio source separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6682–6686. (page 39, 108, 123)

Févotte, C., N. Bertin et J.-L. Durrieu. 2009, «Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis», *Neural Computation*, vol. 21, n° 3, p. 793–830. (page 25, 32, 33, 34, 69, 76)

Févotte, C. et S. J. Godsill. 2006, «A bayesian approach for blind separation of sparse sources», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n° 6, p. 2174–2188. (page 24, 99)

Févotte, C. et J. Idier. 2011, «Algorithms for nonnegative matrix factorization with the β-divergence», *Neural computation*, vol. 23, n° 9, p. 2421–2456. (page 32)

Févotte, C. et M. Kowalski. 2014, «Low-rank time-frequency synthesis», dans *Actes de Advances in Neural Information Processing Systems (NIPS)*, p. 3563–3571. (page 85, 86, 123)

Fontaine, M., C. Vanwynsberghe, A. Liutkus et R. Badeau. 2017a, «Scalable Source Localization with Multichannel Alpha-Stable Distributions», dans *Actes de European Signal Processing Conference (EUSIPCO)*, Actes de European Signal Processing Conference (EUSIPCO), p. 11–15. (page 29)

Fontaine, M., C. Vanwynsberghe, A. Liutkus et R. Badeau. 2017b, «Sketching for nearfield acoustic imaging of heavy-tailed sources», dans *Actes de International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, p. 80–88. (page 29)

Fraysse, A. et T. Rodet. 2014, «A measure-theoretic variational Bayesian algorithm for large dimensional problems», *SIAM Journal on Imaging Sciences*, vol. 7, n° 4, p. 2591–2622. (page 116)

Gannot, S., E. Vincent, S. Markovich-Golan et A. Ozerov. 2017, «A consolidated perspective on multimicrophone speech enhancement and source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, n° 4, p. 692–730. (page 20)

Ganseman, J., P. Scheunders, G. J. Mysore et J. S. Abel. 2010, «Evaluation of a score-informed source separation system.», dans *Actes de International Society for Music Information Retrieval (ISMIR) Conference*, p. 219–224. (page 25)

Gilavert, C., S. Moussaoui et J. Idier. 2015, «Efficient Gaussian sampling for solving large-scale inverse problems using MCMC», *IEEE Transactions on Signal Processing*, vol. 63, n° 1, p. 70–80. (page 132)

Giri, R. 2016, *Bayesian sparse signal recovery using scale mixtures with applications to speech*, thèse de doctorat, UC San Diego. (page 108)

Golub, G. H. et C. F. Van Loan. 1996, *Matrix computations*, Johns Hopkins University Press. (page 73, 92, 94, 143)

Gribonval, R. et M. Zibulevsky. 2010, «Sparse component analysis», dans *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, édité par P. Comon et C. Jutten, Academic press, p. 367–420. (page 24)

Gu, G., X. Gao, J. He et M. Naraghi-Pour. 2007, «Parametric modeling of wideband and ultrawideband channels in frequency domain», *IEEE Transactions on Vehicular Technology*, vol. 56, n° 4, p. 1600–1612. (page 57)

Gustafsson, T., B. D. Rao et M. Trivedi. 2003, «Source localization in reverberant environments : Modeling and statistical analysis», *IEEE Transactions on Speech and Audio Processing*, vol. 11, n° 6, p. 791–803. (page 57)

Hadad, E., F. Heese, P. Vary et S. Gannot. 2014, «Multichannel audio database in various acoustic environments», dans *Actes de International Workshop on Acoustic Signal Enhancement (IWAENC)*, p. 313–317. (page xvi, 48, 109, 110, 111, 127)

Heittola, T., A. Klapuri et T. Virtanen. 2009, «Musical instrument recognition in polyphonic audio using source-filter model for sound separation.», dans *Actes de International Society for Music Information Retrieval (ISMIR) Conference*, p. 327–332. (page 5)

Hennequin, R., B. David et R. Badeau. 2011, «Score informed audio source separation using a parametric model of non-negative spectrogram», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 45–48. (page 25)

Hérault, J., C. Jutten et B. Ans. 1985, «Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé», dans *Colloque sur le traitement du signal et des images*, GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, p. 1017–1022. (page 20, 21)

Hestenes, M. R. et E. Stiefel. 1952, *Methods of conjugate gradients for solving linear systems*, vol. 49, NBS Washington, DC. (page 143)

Hoffman, M. D., D. M. Blei, C. Wang et J. Paisley. 2013, «Stochastic variational inference», *Journal of Machine Learning Research*, vol. 14, n° 1, p. 1303–1347. (page 116)

Honkela, A., T. Raiko, M. Kuusela, M. Tornio et J. Karhunen. 2010, «Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes», *Journal of Machine Learning Research*, vol. 11, n° Nov., p. 3235–3268. (page 116, 118)

Honkela, A., M. Tornio, T. Raiko et J. Karhunen. 2008, «Natural conjugate gradient in variational inference», dans *Actes de International Conference on Neural Information Processing*, p. 305–314. (page 116, 118)

Hunter, D. R. et K. Lange. 2004, «A tutorial on MM algorithms», *The American Statistician*, vol. 58, n° 1, p. 30–37. (page 33)

Hyvärinen, A. et E. Oja. 2000, «Independent component analysis : algorithms and applications», *Neural Networks*, vol. 13, n° 4, p. 411–430. (page 22)

Jot, J.-M., L. Cerveau et O. Warusfel. 1997, «Analysis and synthesis of room reverberation based on a statistical time-frequency model», dans *Audio Engineering Society Convention 103*. Papier no. 4629. (page 53)

Kameoka, H., N. Ono et S. Sagayama. 2008, «Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 29–32. (page 32)

Kay, S. 1988, «Spectral estimation», dans *Advanced Topics in Signal Processing*, édité par J. S. Lim et A. V. Oppenheim, Prentice-Hall, p. 58–122. (page 61)

Kingma, D. P. 2017, *Variational inference and deep learning : A new synthesis*, thèse de doctorat, University of Amsterdam. (page 116, 132)

Kotz, S. et S. Nadarajah. 2004, *Multivariate t-distributions and their applications*, Cambridge University Press. (page 137)

Kounades-Bastian, D., L. Girin, X. Alameda-Pineda, S. Gannot et R. Horaud. 2015, «A variational EM algorithm for the separation of moving sound sources», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5. (page 13)

Kounades-Bastian, D., L. Girin, X. Alameda-Pineda, S. Gannot et R. Horaud. 2016, «An inverse-gamma source variance prior with factorized parameterization for audio source separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 136–140. (page 30, 99)

Kowalski, M., E. Vincent et R. Gribonval. 2010, «Beyond the narrowband approximation : Wide-band convex methods for under-determined reverberant audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 7, p. 1818–1829. (page 24, 39, 84, 97, 98, 104, 105, 108)

Kumaresan, R. 1983, «On the zeros of the linear prediction-error filter for deterministic signals», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 31, n° 1, p. 217–220. (page 54)

Kuttruff, H. 2009, *Room acoustics*, CRC Press. (page 56)

Lee, D. D. et H. S. Seung. 1999, «Learning the parts of objects by non-negative matrix factorization», *Nature*, vol. 401, n° 6755, p. 788. (page 30)

Leglaive, S., R. Badeau et G. Richard. 2015a, «A priori probabiliste anéchoïque pour la séparation sous-déterminée de sources sonores en milieu réverbérant», dans *Colloque GRETSI*. Papier no. 127. (page 15, 53, 55)

Leglaive, S., R. Badeau et G. Richard. 2015b, «Multichannel audio source separation with probabilistic reverberation modeling», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5. (page 15, 53, 55)

Leglaive, S., R. Badeau et G. Richard. 2016a, «Autoregressive moving average modeling of late reverberation in the frequency domain», dans *Actes de European Signal Processing Conference (EUSIPCO)*, p. 1478–1482. (page 16, 53, 55)

Leglaive, S., R. Badeau et G. Richard. 2016b, «Multichannel audio source separation with probabilistic reverberation priors», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, n° 12, p. 2453–2465. (page 16, 66)

Leglaive, S., R. Badeau et G. Richard. 2017a, «Multichannel audio source separation : variational inference of time-frequency sources from time-domain observations», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 26–30. (page 16, 85, 88, 104, 105, 120, 121, 124, 125)

Leglaive, S., R. Badeau et G. Richard. 2017b, «Semi-blind Student's *t* source separation for multichannel audio convolutive mixtures», dans *Actes de European Signal Processing Conference (EUSIPCO)*, p. 2323–2327. (page 17, 85, 98)

Leglaive, S., R. Badeau et G. Richard. 2017c, «Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 264–268. (page 16, 85, 88)

Leglaive, S., R. Badeau et G. Richard. 2017d, «Student's t source and mixing models for multichannel audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*. Article en phase de relecture au moment de la rédaction de cette thèse. (page 17, 108)

Leglaive, S., U. Şimşekli, A. Liutkus, R. Badeau et G. Richard. 2017e, «Alpha-stable multichannel audio source separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 576–580. (page 28)

Leglaive, S., R. Hennequin et R. Badeau. 2015c, «Singing voice detection with deep recurrent neural networks», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 121–125. (page 5)

Li, X., L. Girin et R. Horaud. 2017a, «Audio source separation based on convolutive transfer function and frequency-domain lasso optimization», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 541–545. (page 38, 84, 124)

Li, X., L. Girin et R. Horaud. 2017b, «An EM algorithm for audio source separation based on the convolutive transfer function», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 56–60. (page 38, 39, 84, 124)

Li, X., L. Girin, R. Horaud et S. Gannot. 2016, «Estimation of the direct-path relative transfer function for supervised sound-source localization», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, n° 11, p. 2171–2186. (page 133)

Li, X., L. Girin, R. Horaud et S. Gannot. 2017c, «Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization», *IEEE Transactions on Audio, Speech, and Language Processing*. (page 133)

Lindau, A., L. Kosanke et S. Weinzierl. 2010, «Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses», dans *Audio Engineering Society Convention 128*. Papier no. 8089. (page 53)

Liutkus, A. 2012, *Processus gaussiens pour la séparation de sources et le codage informé*, thèse de doctorat, Télécom ParisTech. (page 26)

Liutkus, A. et R. Badeau. 2015, «Generalized Wiener filtering with fractional power spectrograms», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 266–270. (page 26, 27, 28)

Liutkus, A., R. Badeau et G. Richard. 2011, «Gaussian processes for underdetermined source separation», *IEEE Transactions on Signal Processing*, vol. 59, n° 7, p. 3155–3167. (page 25, 28)

Liutkus, A., D. Fitzgerald et R. Badeau. 2015, «Cauchy nonnegative matrix factorization», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5. (page 25, 27, 33)

Liutkus, A., F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono et J. Fontecave. 2017, «The 2016 Signal Separation Evaluation Campaign», dans *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)*, vol. 10169, édité par P. Tichavský, M. Babaie-Zadeh, O. J. Michel et N. Thirion-Moreau, Springer, Grenoble, France, p. 323 – 332. (page 132)

Magron, P., R. Badeau et A. Liutkus. 2017, «Lévy NMF for robust nonnegative source separation», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 259–263. (page 25, 28)

Malvar, H. S. 1992, *Signal Processing with Lapped Transforms*, Artech House, ISBN 0890064679. (page 9, 85)

Mandel, M. I., R. J. Weiss et D. P. Ellis. 2010, «Model-based expectation-maximization source separation and localization», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 2, p. 382–394. (page 24)

Mertins, A. 1999, *Signal Analysis : Wavelets, Filter Banks, Time-Frequency Transforms and Applications, English (revised edition)*, Wiley Online Library. (page 11)

Moorer, J. A. 1979, «About this reverberation business», *Computer music journal*, vol. 3, n° 2, p. 13–28. (page 110)

Morse, P. M. et K. U. Ingard. 1968, *Theoretical acoustics*, Princeton university press. (page 139)

Nakamura, S., K. Hiyane, F. Asano, T. Nishiura et T. Yamada. 2000, «Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition.», dans *Actes de International Conference on Language Resources and Evaluation (LREC)*, p. 965–968. (page xv, 48, 52, 55)

Naylor, P. A. et N. D. Gaubitch. 2010, *Speech dereverberation*, Springer Science & Business Media. (page 53, 58)

Nugraha, A., A. Liutkus et E. Vincent. 2016a, «Multichannel audio source separation with deep neural networks», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, n° 9, p. 1652–1664. (page 132)

Nugraha, A., A. Liutkus et E. Vincent. 2016b, «Multichannel music separation with deep neural networks», dans *Actes de European Signal Processing Conference (EUSIPCO)*, p. 1748–1752. (page 132)

Ozerov, A. et C. Févotte. 2010, «Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n° 3, p. 550–563. (page 16, 34, 66, 67, 69, 74, 75, 76, 79, 95, 96, 97, 98, 104, 105, 120)

Ozerov, A., C. Févotte, R. Blouet et J.-L. Durrieu. 2011, «Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 257–260. (page 25, 66, 67, 74, 79, 95, 120)

Ozerov, A., E. Vincent et F. Bimbot. 2012, «A general flexible framework for the handling of prior information in audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, n° 4, p. 1118–1133. (page 37, 120, 125, 132)

Palmer, J., K. Kreutz-Delgado, B. D. Rao et D. P. Wipf. 2006, «Variational EM algorithms for non-Gaussian latent variable models», dans *Advances in neural information processing systems*, p. 1059–1066. (page 27)

Parra, L. et C. Spence. 2000, «Convolutive blind separation of non-stationary sources», *IEEE Transactions on Speech and Audio Processing*, vol. 8, n° 3, p. 320–327. (page 23, 24)

Petersen, K. B., M. S. Pedersen et al.. 2008, «The matrix cookbook», *Technical University of Denmark*, vol. 7, p. 15. (page 93)

Pham, D.-T. et J.-F. Cardoso. 2001, «Blind separation of instantaneous mixtures of nonstationary sources», *IEEE Transactions on Signal Processing*, vol. 49, n° 9, p. 1837–1848. (page 22)

Polack, J.-D. 1988, *La transmission de l'énergie sonore dans les salles*, thèse de doctorat, Université du Maine. (page 56, 108, 110)

Polack, J.-D. 1993, «Playing billiards in the concert hall : The mathematical foundations of geometrical room acoustics», *Applied Acoustics*, vol. 38, n° 2-4, p. 235–244. (page 56, 111)

Polack, J.-D. 2015, «Are impulse responses Gaussian noises ?», dans *Acoustics, Information, and Communication*, édité par N. Xiang et G. M. Sessler, Springer, p. 77–91. (page 127)

Prätzlich, T., R. M. Bittner, A. Liutkus et M. Müller. 2015, «Kernel additive modeling for interference reduction in multi-channel music recordings», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 584–588. (page 5)

Puntonet, C. G. et A. Prieto. 1995, «An adaptive geometrical procedure for blind separation of sources», *Neural Processing Letters*, vol. 2, n° 5, p. 23–27. (page 21)

Radlovic, B. D., R. C. Williamson et R. A. Kennedy. 2000, «Equalization in an acoustic reverberant environment : Robustness results», *IEEE Transactions on Speech and Audio Processing*, vol. 8, n° 3, p. 311–319. (page 139)

Ratnam, R., D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing et A. S. Feng. 2003, «Blind estimation of reverberation time», *Journal of the Acoustical Society of America*, vol. 114, n° 5, p. 2877–2892. (page 80)

Rickard, S. 2007, «The duet blind source separation algorithm», dans *Blind Speech Separation*, Springer, p. 217–241. (page 24)

Rigaud, F. et M. Radenen. 2016, «Singing voice melody transcription using deep neural networks.», dans *Actes de International Society for Music Information Retrieval (ISMIR) Conference*, p. 737–743. (page 5)

Samorodnitsky, G. et M. S. Taqqu. 1994, *Stable non-Gaussian random processes : stochastic models with infinite variance*, vol. 1, CRC press. (page 27)

Sawada, H., S. Araki, R. Mukai et S. Makino. 2006, «Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 77–80. (page 24)

Sawada, H., S. Araki, R. Mukai et S. Makino. 2007, «Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 5, p. 1592–1604. (page 24)

Schaeffer, P. 1966, *Traité des objets musicaux*, Collection Pierres Vives, Éditions du Seuil. (page 4)

Schmidt, M. N. et H. Laurberg. 2008, «Nonnegative matrix factorization with Gaussian process priors», *Computational intelligence and neuroscience*, vol. 2008, p. 3. (page 33)

Schroeder, M. R. 1962, «Frequency-correlation functions of frequency responses in rooms», *Journal of the Acoustical Society of America*, vol. 34, n° 12, p. 1819–1823. (page 55, 57, 58, 110)

Schroeder, M. R. 1965, «New method of measuring reverberation time», *Journal of the Acoustical Society of America*, vol. 37, n° 3, p. 409–412. (page 62)

Schroeder, M. R. 1987, «Statistical parameters of the frequency response curves of large rooms», *Journal of the Audio Engineering Society*, vol. 35, n° 5, p. 299–306. (page 55, 110)

Schroeder, M. R. et K. Kuttruff. 1962, «On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima», *Journal of the Acoustical Society of America*, vol. 34, n° 1, p. 76–80. (page 55, 56, 110)

Schultz, T. 1971, «Diffusion in reverberation rooms», *Journal of Sound and Vibration*, vol. 16, n° 1, p. 17–28. (page 52)

Shashanka, M., B. Raj et P. Smaragdis. 2008, «Probabilistic latent variable models as nonnegative factorizations», *Computational intelligence and neuroscience*, vol. 2008. (page 33)

Sivasankaran, S., E. Vincent et I. Illina. 2017, «A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions», *Computer Speech and Language*, vol. 46, p. 444–460. (page 133)

Smaragdis, P. 1998, «Blind separation of convolved mixtures in the frequency domain», *Neurocomputing*, vol. 22, n° 1, p. 21–34. (page 23, 24)

Smaragdis, P. et G. J. Mysore. 2009, «Separation by "humming" : User-guided sound extraction from monophonic mixtures», dans *Actes de IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 69–72. (page 25)

Stewart, R. et M. B. Sandler. 2010, «Database of omnidirectional and b-format room impulse responses.», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 165–168. (page 59)

Tachibana, H., T. Ono, N. Ono et S. Sagayama. 2010, «Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 425–428. (page 5)

Theis, F. J., A. Jung, C. G. Puntonet et E. W. Lang. 2003, «Linear geometric ICA : Fundamentals and algorithms», *Neural Computation*, vol. 15, n° 2, p. 419–439. (page 21)

Thi, H.-L. N. et C. Jutten. 1995, «Blind source separation for convolutive mixtures», *Signal processing*, vol. 45, n° 2, p. 209–229. (page 23)

Vincent, E. 2007, «Complex nonconvex ℓ_p norm minimization for underdetermined source separation», dans *Actes de International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, p. 430–437. (page 24)

Vincent, E. 2012, «Improved perceptual metrics for the evaluation of audio source separation», dans *Actes de International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, p. 430–437. (page 47)

Vincent, E., N. Bertin, R. Gribonval et F. Bimbot. 2014, «From blind to guided audio source separation : How models and side information can improve the separation of sound», *IEEE Signal Processing Magazine*, vol. 31, n° 3, p. 107–115. (page 7, 25)

Vincent, E. et D. R. Campbell. 2008, «Roomsimove», `http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip`. (page 48, 59)

Vincent, E., R. Gribonval et C. Févotte. 2006, «Performance measurement in blind audio source separation», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n° 4, p. 1462–1469. (page 46)

Vincent, E., M. G. Jafari, S. A. Abdallah, M. D. Plumbley et M. E. Davies. 2010, «Probabilistic modeling paradigms for audio source separation», *Machine Audition : Principles, Algorithms and Systems*, p. 162–185. (page 25, 37)

Vincent, E., H. Sawada, P. Bofill, S. Makino et J. P. Rosca. 2007, «First stereo audio source separation evaluation campaign : data, algorithms and results», dans *Actes de International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, p. 552–559. (page 46)

Vinyes, M. 2008, «MTG MASS dataset», `http://mtg.upf.edu/download/datasets/mass`. (page 47)

Virtanen, T., A. T. Cemgil et S. Godsill. 2008, «Bayesian extensions to non-negative matrix factorisation for audio signal modelling», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1825–1828. (page 25, 33)

Virtanen, T., J. F. Gemmeke, B. Raj et P. Smaragdis. 2015, «Compositional models for audio processing : Uncovering the structure of sound mixtures», *IEEE Signal Processing Magazine*, vol. 32, n° 2, p. 125–144. (page 25, 30)

Wang, D. et G. J. Brown. 2006, *Computational auditory scene analysis : Principles, algorithms, and applications*, Wiley-IEEE press. (page 20)

Wasserman, L. 2013, *All of statistics : a concise course in statistical inference*, Springer Science & Business Media. (page 39)

Weihua, W. et H. Fenggang. 2008, «Improved method for solving permutation problem of frequency domain blind source separation», dans *Actes de IEEE International Conference on Industrial Informatics (INDIN)*, p. 703–706. (page 24)

West, M. 1984, «Outlier models and prior distributions in Bayesian linear regression», *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 431–439. (page 27)

West, M. 1987, «On scale mixtures of normal distributions», *Biometrika*, vol. 74, n° 3, p. 646–648. (page 27)

Winn, J. et C. M. Bishop. 2005, «Variational message passing», *Journal of Machine Learning Research*, vol. 6, n° Apr., p. 661–694. (page 113, 115)

Winter, S., W. Kellermann, H. Sawada et S. Makino. 2007, «MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization», *EURASIP Journal on Applied Signal Processing*, vol. 2007, n° 1, p. 024 717. (page 24)

Wipf, D. et H. Zhang. 2014, «Revisiting Bayesian blind deconvolution», *Journal of Machine Learning Research*, vol. 15, n° 1, p. 3595–3634. (page 128)

Yellin, D. et E. Weinstein. 1994, «Criteria for multichannel signal separation», *IEEE transactions on signal processing*, vol. 42, n° 8, p. 2158–2168. (page 23)

Yoshii, K., K. Itoyama et M. Goto. 2016, «Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation», dans *Actes de IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 51–55. (page 25, 26, 27, 29, 33, 34, 36, 99)

Zheng, Y., A. Fraysse et T. Rodet. 2015, «Efficient Variational Bayesian Approximation Method Based on Subspace optimization», *IEEE Transactions on Image Processing*, vol. 24, n° 2, p. 681–693. (page 116)

Modèles de mélange pour la séparation multicanale de sources sonores en milieu réverbérant

Simon LEGLAIVE

Résumé : Cette thèse porte sur la séparation sous-déterminée de sources sonores en milieu réverbérant. Nous adoptons une approche probabiliste où les signaux sources sont représentés comme des variables aléatoires latentes dans un domaine temps-fréquence. La structure spécifique des signaux musicaux dans ce domaine est exploitée par l'intermédiaire de modèles de factorisation en matrices non-négatives.

Les méthodes de la littérature traitent généralement les filtres de mélange comme des paramètres déterministes estimés uniquement à partir des données observées. Ces filtres correspondent cependant à des réponses de salle, ils ont donc une structure bien particulière qu'il est possible d'exploiter afin de guider leur estimation.

Dans une première partie, le processus de mélange convolutif temporel est approché dans le domaine de la transformée de Fourier à court-terme, sous une hypothèse de filtres de mélange à réponse impulsionnelle courte. Nous développons des modèles autorégressifs à moyenne ajustée ayant pour objectif de transcrire la dynamique temporelle des filtres sous forme de corrélations fréquentielles. Ces modèles sont ensuite utilisés dans une méthode de séparation de sources où les filtres sont estimés au sens du maximum *a posteriori*, par un algorithme espérance-maximisation.

Dans une seconde partie, nous proposons une méthode d'inférence variationnelle des coefficients temps-fréquence des sources à partir des observations temporelles du mélange. Le processus de mélange convolutif est donc cette fois représenté de façon exacte. En plus de convenir à la séparation de mélanges fortement réverbérants, cette approche nous permet de développer des *a priori* simples sur les filtres de mélange afin de guider leur estimation. Nous proposons un modèle basé sur la distribution *t* de Student et exploitant la décroissance exponentielle de la réverbération dans le domaine temporel.

Mots-clés : Séparation sous-déterminée de sources audio, mélanges multicanaux réverbérants, acoustique statistique des salles, factorisation en matrices non-négatives, modèles probabilistes, inférence statistique, inférence variationnelle.

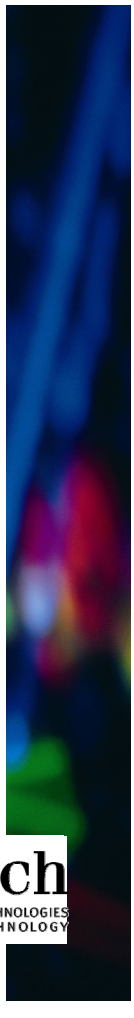
Abstract: This thesis addresses the problem of under-determined audio source separation for multichannel reverberant mixtures. We adopt a probabilistic approach where the source signals are represented as latent random variables in a time-frequency domain. The specific structure of musical signals in this domain is exploited by means of non-negative matrix factorization models.

In the literature, the mixing filters are generally treated as deterministic parameters, only estimated from the observed data. However, as these filters correspond to room responses, they exhibit a very particular structure that can be used to guide their estimation.

In a first part, the time-domain convolutive mixing process is approximated in the short-time Fourier transform domain, under the assumption that the impulse response of the mixing filters is short. We develop autoregressive moving average models that aim to transcribe the temporal dynamics of the filters into frequency-domain correlations. These models are then used in a source separation framework, for performing maximum *a posteriori* estimation of the mixing filters by means of an expectation-maximization algorithm.

In a second part, we propose to infer the time-frequency source coefficients from the time-domain mixture observations, using a variational approach. The convolutive mixing process is here exactly represented. In addition to being suitable for the separation of highly reverberant mixtures, this approach allows us to develop simple priors for the mixing filters in order to guide their estimation. We propose a model based on the Student's *t* distribution that exploits the exponential decay of reverberation in the time domain.

Key-words: Under-determined audio source separation, multichannel reverberant mixtures, statistical room acoustics, non-negative matrix factorization, probabilistic models, statistical inference, variational inference.



TELECOM ParisTech

EDITE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY