Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux

Emmanuel Vincent

UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE
ÉCOLE DOCTORALE EDITE
IRCAM – CENTRE POMPIDOU

THÈSE DE DOCTORAT

spécialité
ACOUSTIQUE, TRAITEMENT DU SIGNAL ET INFORMATIQUE
APPLIQUÉS À LA MUSIQUE

présentée par
EMMANUEL VINCENT

pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE

Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux

soutenue le 2 décembre 2004
devant le jury composé de

Xavier RODET	Université Paris VI	Directeur de thèse
Christian JUTTEN	INPG	Rapporteur
Bruno TORRÉSANI	Université de Provence	Rapporteur
Patrick FLANDRIN	École Normale Supérieure de Lyon	Examinateur
Jean-Dominique POLACK	Université Paris VI	Examinateur
Mark SANDLER	Queen Mary University of London	Examinateur

Tout est lié.
Lao Tseu

Résumé

Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux

Depuis une quinzaine d'années, l'étude des enregistrements de musique de chambre se focalise sous deux points de vue distincts : la séparation de sources et la transcription polyphonique. La séparation de sources cherche à extraire des enregistrements les signaux correspondant aux instruments présents. La transcription polyphonique vise à les décrire par un ensemble de paramètres : noms des instruments, hauteurs et volumes des notes jouées, *etc*. Les méthodes existantes, fondées sur l'analyse spatiale et spectro-temporelle des enregistrements, fournissent des résultats satisfaisants sur des cas simples. Mais généralement leur performance se dégrade vite au-delà d'un nombre d'instruments limite ou en présence de réverbération, d'instruments de même tessiture ou de notes à intervalle harmonique.

Notre hypothèse est que ces méthodes souffrent souvent de modèles de sources instrumentales trop génériques. Nous proposons d'y remédier par la création de modèles d'instruments spécifiques basés sur un apprentissage.

Dans ce travail, nous justifions cette hypothèse par l'étude des informations pertinentes présentes dans les enregistrements musicaux et de leur exploitation par les méthodes existantes. Nous construisons ensuite de nouveaux modèles probabilistes d'instruments inspirés de l'Analyse en Sous-espaces Indépendants (ASI) et nous donnons quelques exemples d'instruments appris. Enfin nous appliquons ces modèles à la séparation et la transcription d'enregistrements réalistes, parmi lesquels des pistes de CD et des mélanges synthétiques convolutifs ou sous-déterminés de ces pistes.

Mots-clefs : séparation de sources, transcription polyphonique, identification d'instruments, analyse de scènes sonores, modèles probabilistes de sources, Analyse en Sous-espaces Indépendants

Abstract

Instrument models for source separation and transcription of music recordings

For about fifteen years the study of chamber music recordings has focused on two distinct viewpoints : source separation and polyphonic transcription. Source separation tries to extract from a recording the signals corresponding to each musical instrument playing. Polyphonic transcription aims at describing a recording by a set of parameters : names of the instruments, pitch and loudness of the notes, *etc*. Existing methods, based on spatial and spectro-temporal analysis of the recordings, provide satisfying results in simple cases. But their performance generally degrades fast with more instruments than a fixed limit, under reverberant conditions, with instruments of similar playing ranges or with notes at harmonic intervals.

Our hypothesis is that these methods often suffer from too generic models of instrumental sources. We propose to address this by creating specific instrument models based on learning.

In this dissertation, we justify this hypothesis by studying the relevant information present in musical recordings and its use by existing methods. Then we describe new probabilistic instrument models inspired from Independent Subspace Analysis (ISA) and we give a few examples of learnt instruments. Finally we exploit these models to separate and transcribe realistic recordings, among which CD tracks and synthetic convolutive or underdetermined mixtures of these tracks.

Keywords : source separation, polyphonic transcription, instrument identification, auditory scene analysis, probabilistic source models, Independent Subspace Analysis

Remerciements

Je tiens avant tout à remercier Xavier Rodet pour m'avoir proposé de travailler sur ce sujet passionnant, à la fois proche des mathématiques et de la musique. Par son expérience exceptionnelle de la recherche appliquée en audio et par son ouverture d'esprit, il a su à la fois me conseiller et me laisser libre de creuser mes propres directions, ce dont je lui suis hautement reconnaissant.

Je remercie ensuite les fondateurs, enseignants et gestionnaires du DEA ATIAM. L'existence du DEA et la qualité de la formation transmise ont été décisives dans mon orientation. Merci de même à tous les enseignants de lycée et de conservatoire qui ont éveillé ma curiosité scientifique et musicale bien avant.

Merci beaucoup à tous les collègues qui m'ont éclairé par des discussions scientifiques de qualité et m'ont appris à programmer correctement en MATLAB, écrire un article ou présenter des *slides*. Merci en particulier à mes voisins de bureau, à mes camarades de l'Action Jeunes Chercheurs, aux membres de l'équipe Analyse-Synthèse, aux membres d'autres équipes de l'IRCAM, et aux nombreux "séparateurs de sources" et "transcripteurs polyphoniques" rencontrés dans diverses circonstances.

Merci également aux membres du jury de l'intérêt qu'ils ont bien voulu porter à mon travail.

Mes remerciements vont aussi à tous ceux qui m'ont permis d'effectuer ce travail dans de bonnes conditions et d'avoir plus de temps pour la recherche proprement dit. Merci entre autres au personnel administratif, à l'équipe système et à la production de l'IRCAM, aux développeurs des logiciels Prosper et Tkbibtex, aux développeurs des Hidden Markov Model toolbox, Time-Frequency toolbox, Fast ICA toolbox et Ear toolbox pour MATLAB, et aux créateurs de la base de données d'enregistrements musicaux RWC Database.

Merci au Ministère de l'Éducation Nationale et de la Recherche, qui a financé la plus grande partie de mes études supérieures, et au GdR ISIS, qui m'a permis par son financement de rencontrer des collègues très intéressants dans des lieux tout aussi intéressants. Merci aussi aux collègues enseignants de Jussieu avec qui j'ai collaboré durant mon monitorat.

J'exprime toute ma gratitude pour leur soutien et leur bonne humeur à ceux que j'ai côtoyé durant tout ce temps en dehors du boulot : mes parents, mes frère et soeurs, mes amis de l'association, mes potes de lycée, de prépa, de l'ENS, de Centrale Lille et d'ailleurs.

Merci aux artistes d'horizons divers qui m'ont donné de la bonne musique ou de la bonne peinture quand j'en avais besoin. Mention spéciale aux inventeurs méconnus de la coinche, de la pétanque, du roller, du thé vert japonais et de la bière belge qui ont aussi beaucoup contribué à mon équilibre.

Je conclus cette liste en ne remerciant pas tous ceux qui contribuent à déprécier les études supérieures et à ralentir l'avancée de la recherche, en particulier TF1 et Fun Radio par leur militantisme contre la réflexion et leur promotion de la culture bas de gamme, Mathworks et les maisons d'édition scientifique par la vente de licences MATLAB et de revues à des prix inabordables pour les labos pauvres de chez nous et des pays en voie de développement, ainsi que la SACEM et les maisons de disques par leur soutien aux mesures excessives du droit d'auteur au détriment des mesures plus défendables.

Table des matières

Liste des notations

Objets usuels

Nous adoptons les notations générales suivantes.

Les matrices, les vecteurs ou les signaux multicanal seront notés en caractères gras. Les scalaires ou les signaux monocanal seront notés en caractères standard. Les parenthèses désigneront des suites, les bornes des indices n'étant pas précisées lorsqu'elles sont définies auparavant. Les accolades désigneront des ensembles.

Les suites et les signaux seront souvent notés de façon abrégée sans indices ou variables temporelles. Par exemple la suite $(o_{tf})_{0 \leq f \leq F-1}$ sera notée comme un vecteur $\mathbf{o}_t$ et la suite $(\mathbf{o}_t)_{0 \leq t \leq T-1} = (o_{tf})_{0 \leq f \leq F-1, 0 \leq t \leq T-1}$ sera elle-même notée comme une matrice $\mathbf{o}$. De même le signal x désignera la séquence $(x(u))_{0 \leq u \leq U-1}$. Ces abbréviations seront généralement définies explicitement.

Fonctions usuelles

$\mathbf{a}^T$	Transposée du vecteur ou de la matrice $\mathbf{a}$
$\mathbf{a}^H$	Transposée-conjuguée du vecteur ou de la matrice $\mathbf{a}$
$\overline{a}$	Conjugué du nombre complexe a
$\Re a$	Partie réelle du nombre complexe a
$\lvert a \rvert$	Norme du nombre complexe a ou valeur absolue du nombre réel a
$a \mod 2\pi$	Reste (dans $]-\pi, \pi]$) de la division entière du nombre réel a par 2π
$\langle \mathbf{a}, \mathbf{b} \rangle$	Produit scalaire des vecteurs (ou des signaux) $\mathbf{a}$ et $\mathbf{b}$
$\lVert \mathbf{a} \rVert$	Norme euclidienne du vecteur $\mathbf{a}$
$\mathbf{a} \bullet \mathbf{b}$	Produit point-à-point des vecteurs $\mathbf{a}$ et $\mathbf{b}$
$a \star b$	Convolution des signaux a et b
$\angle \mathbf{a}$	Phase point-à-point (dans $]-\pi, \pi]$) du vecteur de nombres complexes $\mathbf{a}$
$\exp(\mathbf{a})$	Exponentielle point-à-point du vecteur $\mathbf{a}$
$\log(\mathbf{a})$	Logarithme point-à-point du vecteur $\mathbf{a}$
$P(a)$	Probabilité ou densité de probabilité de la variable a
$P(a\|b)$	Probabilité ou densité de probabilité conditionnelle de la variable a sachant l'événement b
$\mathcal{N}(a; \mu, \sigma)$	Densité de probabilité gaussienne de moyenne μ et d'écart-type σ évaluée pour la valeur a
$\#\mathcal{A}$	Cardinal de l'ensemble $\mathcal{A}$

Indices

f	Canal fréquentiel (de 0 à F-1)
h	Hauteur de note sur l'échelle MIDI (N_j hauteurs possibles de H_j à $H'j$ pour l'instrument j)
i	Canal du mélange (de 1 à m)
j	Source ou instrument présent (de 1 à n)
k	Ordre d'une composante de variation (de 1 à K_j)
t	Trame temporelle (de 0 à T-1)

Variables utilisées et paragraphe de définition

$\mathbf{x}$, $\mathbf{s}$, $\mathbf{n}$, $\mathbf{A}$, $s_{\mathrm{img}\,ij}$	§1.1
$\mathbf{x}_{\mathrm{rmx}}$	§1.3.2
s_j^f, s_j^{ft}, x_i^f, x_i^{ft}, g_f, o_{itf}	§2.2.2
$\mathbf{d}_{i'it}^{\mathrm{pha}}$, $\mathbf{d}_{i'it}^{\mathrm{vol}}$, $\mathbf{d}_{i'it}^{\mathrm{coh}}$	§2.2.3
$\mathbf{E}_j$, $\mathbf{p}_j$, $\mathbf{m}_j$	§3.1
$\mathbf{o}$, $\mathbf{\Theta}$	§3.2
$\mathcal{M}$, $\mathcal{O}$, $\mathcal{I}$, $\mathcal{M}_j$, P^{spat}, P^{spec}, P^{desc}, $P^{\mathrm{état}}$	§3.3
P^{comb}	§3.4.1
$\mathbf{m}'_j$, e_{jht}, $\mathbf{\Phi}'_{jht}$, $\mathbf{\Phi}_{jh}$, v_{jht}^k, $\mathbf{U}_{jh}^k$, $\boldsymbol{\alpha}_{jt}$, σ_j^α	§4.1
μ_{jh}^e, σ_{jh}^e, μ_{jhk}^v, σ_{jhk}^v, $\mathcal{A}_{jt}$	§4.2
Z_j, $\mathcal{D}_j^{\mathrm{note}}$, $\mathcal{D}_j^{\mathrm{segm}}$, $\mathcal{T}_j^{\mathrm{note}}$, $\mathcal{T}_j^{\mathrm{segm}}$, μ_j^n, σ_j^n, μ_j^s, σ_j^s	§4.3
w_{spat}, w_{spec}, w_{desc}, $w_{\mathrm{état}}$, w_{comb}	§4.4
$\mathbf{n}'$, $\boldsymbol{\epsilon}_t$, σ^ϵ, P_t^{trans}	§4.5.1
$\pi_{1,htf}$, π'_{tf}	§4.5.2
$harm$, $trans$	§4.5.3
$(\mathbf{\Phi}_{1,h}^{\mathrm{harm}})$, $(\mathbf{U}_{1,h}^{\mathrm{part}})$, $(\mathbf{U}_{1,h}^{\mathrm{fréq}})$, $(\mathbf{U}_{1,h}^{\mathrm{bruit}})$	§4.6.1

Abbréviations et paragraphe de définition

ACI	Analyse en composantes indépendantes	§2.4.1
ACP	Analyse en composantes principales	
ASA	Analyse de scènes auditives	§2.1.2
ASI	Analyse en sous-espaces indépendants	§2.3.4
EM	Algorithme espérance/maximisation	§3.5
MAP	Maximum *a posteriori*	§3.4
MG	Mélange de gaussiennes	§2.1.2
MIM	Maximum d'information mutuelle	§3.5
MMC	Modèle de Markov caché	§2.1.2
MV	Maximum de vraisemblance	§3.5
RSA	Rapport signal-à-artefacts	§1.3.1
RSD	Rapport signal-à-distortion	§1.3.1
RSI	Rapport signal-à-interférences	§1.3.1
RRA	Rapport remix-à-artefacts	§1.3.2
RRD	Rapport remix-à-distortion	§1.3.2
RRES	Rapport remix-à-erreur sur les sources	§1.3.2
SP	Soustraction de puissance	§6.2.1
TRF	Taux de reconnaissance de familles d'instruments	§1.2.1
TRI	Taux de reconnaissance d'instruments	§1.2.1

Liste des figures

Liste des tableaux

Introduction

De l'utilité supposée des modèles d'instruments

La majorité des signaux audio sont des *mélanges* auxquels contribuent plusieurs *sources*. Ainsi un dialogue de film réalisé à la table de mixage est un mélange *synthétique* entre les paroles successives de plusieurs locuteurs. Et un fond sonore de rue enregistré au micro est un mélange *réel* entre les bruits simultanés de plusieurs voitures.

L'étude des mélanges audio a de nombreuses applications dont la plus populaire est le problème du cocktail (*cocktail party problem*). Ce nom provient de la situation rencontrée lors d'une fête. Comment comprendre ce que dit votre jolie voisine parmi les autres voix, la musique, le bruit des verres et les autres bruits de fond ? L'audition humaine effectue dans ce cas une écoute sélective basée sur les *caractéristiques spatiales et spectro-temporelles* des sources sonores en présence. Elle prend également en compte des *informations a priori* telles que la position spatiale des sources fournie par la vision et les particularités apprises de la voix et du langage. Ce type de comportement peut être copié par un système artificiel visant la *séparation* de la source d'intérêt pour écoute ou bien la *transcription* du texte prononcé.

Dans ce travail, nous considérons plus particulièrement les enregistrements de *musique de chambre instrumentale*, c'est-à-dire la musique jouée par plusieurs instruments simultanément en effectif réduit sans chanteurs, récitants ou effets sonores. Ces enregistrements correspondent à des types de mélanges assez variés, des enregistrements mono aux CD audio stéréo et autres enregistrements multipistes.

Comme dans le cadre du problème du cocktail, nous étudions ces mélanges sous deux angles applicatifs différents : transcription et séparation. La transcription vise à décrire un mélange par un ensemble de paramètres perceptifs et musicaux utiles dans un but de classification : effectif instrumental et noms des instruments, notes jouées et instants d'attaque, nuances, *vibrato*, *etc*. La séparation cherche à extraire d'un mélange les signaux correspondant aux instruments présents pour les écouter directement ou après traitement et remixage.

Les premières méthodes pour résoudre ces questions sont apparues il y a une quinzaine d'années. Elles exploitent les informations spatiales et spectro-temporelles présentes dans les mélanges à l'aide de modèles de sources statistiques ou paramétriques. Les méthodes de transcription, proposées par la communauté musique/parole/psycho-acoustique, s'appliquent généralement aux mélanges *monocanal*. Les méthodes de séparation, étudiées par la communauté signal/télécoms/statistiques, se limitent souvent aux mélanges *multicanal*.

La plupart de ces méthodes utilisent des modèles de sources non spécifiques, applicables à tous les instruments voire à des sources non instrumentales. Leurs résultats sont globalement satisfaisants sur des enregistrements synthétiques générés à partir de fichiers MIDI ou par mélange panoramique, mais se dégradent sur des enregistrements réels. L'information disponible devient souvent insuffisante pour distinguer des instruments de même tessiture, transcrire et séparer des notes à intervalle harmonique ou séparer un enregistrement réverbérant. Deux directions sont alors possibles pour augmenter la quantité d'information disponible : ou bien ajouter des canaux supplémentaires au mélange ou bien prendre en

compte des caractéristiques supplémentaires des sources.

Le but essentiel de ce travail est de montrer que l'utilisation de modèles de sources spécifiques à chaque instrument peut aider à transcrire et séparer des mélanges réels habituellement considérés comme difficiles. Nous étudions en particulier une famille de modèles de sources *probabilistes* appelés *modèles d'instruments*, dont les paramètres sont *appris* sur des bases de données de sons instrumentaux.

Un autre but important est de montrer que l'utilisation de plusieurs canaux de mélange et de conditions de mélange bien calibrées facilite aussi la transcription et la séparation. Nous proposons en particulier des méthodes adaptées aux enregistrements stéréo instantanés ou convolutifs et aux enregistrements multipistes type "ingénieur du son".

Plan du document

À la suite de cette introduction, le document est découpé en sept chapitres, une conclusion et deux annexes qui se présentent comme suit.

Nous commençons dans le chapitre 1 par définir précisément les tâches de transcription et de séparation étudiées. Nous discutons comment évaluer la performance d'un algorithme et la difficulté d'un mélange.

Le chapitre 2 étudie les informations spatiales et spectro-temporelles contenues dans les mélanges et les informations sur les sources fournies par les modèles de sources usuels. Un état de l'art montre comment ces deux types d'informations sont combinés dans les méthodes existantes et quelles sont leurs limitations. Cela nous amène à sélectionner un certain nombre de mélanges considérés comme difficiles et pour lesquels l'utilisation de modèles d'instruments semble présenter un intérêt.

Nous proposons ensuite dans le chapitre 3 un cadre assez général pour construire des modèles probabilistes d'instruments représentant le spectre de puissance à court terme. Nous expliquons comment apprendre leurs paramètres et les utiliser pour la transcription et la séparation. Nous mettons ainsi en lumière des liens importants entre transcription et séparation.

Le chapitre 4 traite un modèle d'instrument particulier basé sur une Analyse en Sous-espaces Indépendants (ASI) non linéaire combinée à des modèles d'évolution temporelle. Nous construisons le modèle en plusieurs étapes, puis nous décrivons les algorithmes de transcription et d'apprentissage associés.

Nous présentons dans le chapitre 5 quelques exemples d'instruments appris. Nous testons également la performance des modèles pour l'identification d'instruments et de notes sur des enregistrements solo.

Le chapitre 6 explique comment combiner plusieurs modèles d'instruments en un modèle de musique de chambre monocanal. Nous évaluons la performance d'identification d'instruments sur des duos synthétiques. Puis nous donnons des exemples de transcription et de séparation de duos réels et synthétiques d'instruments de même tessiture ou jouant des notes à intervalle harmonique.

Le chapitre 7 étend ces résultats aux mélanges multicanal. Nous proposons des modèles de musique de chambre multicanal dans le cas de mélanges stéréo panoramiques, stéréo "AB étroits" ou multipistes. Nous donnons des exemples de transcription et de séparation pour chacun de ces types de mélange. Nous comparons les résultats à ceux des modèles correspondants en monocanal.

Nous concluons en proposant des pistes de recherche pour améliorer les modèles d'instruments construits et en construire d'autres semblables.

L'annexe A décrit les données d'apprentissage et de test utilisées.

L'annexe B présente le banc de filtres utilisé pour calculer les spectres à court terme des mélanges et extraire les sources par filtrage.

Chapitre 1

Présentation des tâches considérées

Ce premier chapitre est consacré à la définition précise des problèmes étudiés. Dans le paragraphe 1.1 nous décrivons les notions de mélange, de séparation et de transcription. Nous énumérons ensuite dans les paragraphes 1.2 et 1.3 quatre tâches typiques pour lesquelles nous proposons des critères numériques d'évaluation de performance des algorithmes. Enfin nous discutons brièvement la notion de difficulté d'un mélange dans le paragraphe 1.4.

La plupart des résultats de ce chapitre sont le résumé d'un travail original en collaboration avec l'équipe METISS de l'IRISA (Rennes) et l'équipe ADTS de l'IRCCyN (Nantes) au sein d'une Action Jeunes Chercheurs du GdR ISIS sur le sujet "Ressources pour la séparation de signaux audiophoniques". Cette action, qui s'est déroulée de mars 2002 à novembre 2003, a donné lieu à plusieurs articles [Vin03b, Vin04a, Gri03b, Vin03a] ainsi qu'à une base de données en ligne [BASS-dB] regroupant des signaux tests et des routines MATLAB d'évaluation de la performance pour les tâches de séparation. Nous adaptons ici ces résultats au cadre plus spécifique de l'audio musicale et nous rajoutons des critères d'évaluation de performance pour les tâches de transcription mentionnés dans la littérature.

1.1 Transcription et séparation : deux visions complémentaires d'un mélange

Un mélange est un signal audio où plusieurs sources audio distinctes sont en présence. Dans le cas de la musique de chambre instrumentale, c'est-à-dire des mélanges d'instruments musicaux, les mélanges se trouvent typiquement sous deux formes : les mélanges synthétiques provenant de CD audio stéréo et les mélanges réels réalisés par enregistrement multipistes. Bien qu'ils correspondent à des méthodes d'acquisition très différentes, tous peuvent s'exprimer grâce au même formalisme.

En appelant $(s_j)_{1\leq j\leq n}$ les signaux sources, $(x_i)_{1\leq i\leq m}$ les signaux captés, $(n_i)_{1\leq i\leq m}$ les bruits de fond et $(a_{ij})_{1\leq i\leq m,1\leq j\leq n}$ les filtres source-à-capteur (causals) variant dans le temps, un mélange $m \times n$ prend la forme d'une convolution

$$x_i(u) = \sum_{j=1}^{n} \sum_{\tau=0}^{+\infty} a_{ij}(u-\tau,\tau) s_j(u-\tau) + n_i(u). \tag{1.1}$$

En notant $(s_{\mathrm{img}\,ij})_{1\leq i\leq m,1\leq j\leq n}$ les images spatiales des sources sur les capteurs, définies par $s_{\mathrm{img}\,ij}(u) = \sum_{\tau=0}^{+\infty} a_{ij}(u-\tau,\tau) s_j(u-\tau)$, ce mélange prend aussi la forme d'une somme

$$x_i(u) = \sum_{j=1}^{n} s_{\mathrm{img}\,ij}(u) + n_i(u). \tag{1.2}$$

Enfin ce mélange s'exprime de façon plus concise grâce au formalisme des matrices de filtres comme

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{n}, \tag{1.3}$$

où $\mathbf{s}$ est le vecteur des signaux sources, $\mathbf{x}$ le vecteur des signaux observés, $\mathbf{n}$ le vecteur des signaux de bruit et $\mathbf{A}$ la matrice des filtres de mélange.

Chaque source correspond à un instrument présent et le bruit de fond regroupe toutes les autres sources audio non musicales. Plusieurs instruments peuvent avoir le même nom, par exemple dans le cas d'un duo de violons. Dans la suite, nous supposerons que le nombre de sources n vérifie $2 \leq n \leq 6$, suivant la définition usuelle du terme "musique de chambre". Nous n'étudierons pas les mélanges de type "musique d'ensemble" ($7 \leq n \leq 14$) ou "musique d'orchestre" ($n \geq 15$).

Les différentes situations de mélange sont généralement classées selon trois axes décrivant la structure des filtres de mélange $\mathbf{A}$. Un mélange peut être instantané (réglage du gain sur une table de mixage ou dans un logiciel d'édition sonore), instantané variant dans le temps (compression de dynamique), convolutif court (réglage de dynamique dans les sous-bandes aigu/médium/grave), convolutif court variant dans le temps, convolutif long (réverbération artificielle) ou convolutif long variant dans le temps (enregistrement au micro en salle avec des sources en mouvement). Un mélange peut être aussi monocanal (micro mono) ou multicanal (micro stéréo ou antenne). Un mélange peut être enfin sur-déterminé (plus de canaux que de sources), sous-déterminé (moins de canaux que de sources) ou déterminé (autant de canaux que de sources).

L'étude de la musique de chambre instrumentale peut se diviser en deux types d'applications. D'une part les applications de transcription visent à décrire un mélange par un ensemble de paramètres perceptifs et musicaux utiles dans un but de classification : effectif instrumental et noms des instruments, notes jouées et instants d'attaque, nuances, *vibrato*, *etc*. D'autre part les applications de séparation cherchent à extraire d'un mélange les signaux correspondant à chaque instrument présent pour les faire écouter directement ou après traitement et remixage.

Transcription et séparation sont deux visions complémentaires d'un mélange. D'une part bien transcrire ne permet pas forcément de bien séparer et *vice-versa*. D'autre part les résultats de transcription peuvent être traités automatiquement alors que les résultats de séparation sont destinés à un humain.

Il existe un grand nombre d'applications distinctes de transcription et de séparation et un grand nombre d'algorithmes utilisés pour les résoudre. Pour pouvoir évaluer notre travail et le comparer aux algorithmes existants, il est nécessaire de choisir un certain nombre de tâches typiques de transcription et de séparation, mais aussi de mesurer pour chaque tâche la performance des algorithmes et la difficulté des mélanges test choisis. Nous abordons ces questions dans le reste du chapitre.

1.2 Définition et évaluation des tâches de transcription

Nous décrivons pour commencer deux tâches de transcription que nous appelons identification d'instruments et identification de notes. Nous donnons quelques exemples d'applications, nous définissons les tâches précisément par la nature des entrées et sorties des algorithmes, et nous rappelons les critères numériques de performance usuels.

1.2.1 Identification d'instruments

L'identification d'instruments vise à retrouver l'effectif instrumental et le nom des instruments présents dans un mélange. Ses applications concernent principalement la segmentation, l'indexation et la recherche par similarité dans des bases de données musicales [Dow03].

Définition

Entrées : observations $\mathbf{x}$, structure de $\mathbf{A}$, liste de noms d'instruments potentiellement présents, base de données de sons instrumentaux.
Sorties : nombre d'instruments estimé $\widehat{n}$, ensemble des noms des instruments présents estimés $\widehat{\mathcal{O}}$.
Informations *a priori* facultatives : nombre d'instruments n.

Évaluation

La performance d'un algorithme d'identification d'instruments peut se mesurer par des critères de comparaison de chaînes [Rap02, God99, Kas95]. La mesure consiste à compter le nombre minimal de substitutions N_{subst}, d'insertions N_{inser} et de suppressions N_{suppr} nécessaires pour faire coïncider la liste d'instruments trouvée avec la vraie liste d'instruments à permutation près. Le Taux de Reconnaissance d'Instruments (TRI) est alors défini par

$$TRI = 1 - \frac{N_{\text{subst}} + N_{\text{inser}} + N_{\text{suppr}}}{N_{\text{total}}}, \tag{1.4}$$

où N_{total} est le vrai nombre d'instruments. Nous définissons de même le Taux de Reconnaissance de Familles d'instruments (TRF). Les quantités $N_{\text{subst}}/N_{\text{total}}$, $N_{\text{inser}}/N_{\text{total}}$ et $N_{\text{suppr}}/N_{\text{total}}$ sont appelées taux de substitution, taux d'insertion et taux de suppression.

Dans le cas où le nombre d'instruments est connu et égal à un, il est possible de préciser les erreurs (de substitution) en notant combien de fois chaque instrument est identifié parmi un ensemble d'extraits test d'un instrument donné. La représentation correspondante (instruments testés dans les lignes, instruments identifiés dans les colonnes) est appelée matrice de confusion. Elle permet d'observer quels instruments sont les plus sujets à confusion.

1.2.2 Identification de notes

Le but de l'identification de notes est de transcrire le son de tous les instruments dans un mélange en une suite d'accords. Chaque accord est décrit par un ensemble de notes attaquées simultanément et définies par leur hauteur sur l'échelle des demi-tons (ou échelle MIDI) et leur instrument associé. La durée des notes n'est pas prise en compte. L'effectif instrumental et le nom de chaque instrument sont supposés connus. Dans le cas où les instruments ont des tessitures disjointes, cette tâche peut se résoudre sans nécessairement utiliser de base de données de sons instrumentaux. Les applications de l'identification de notes sont semblables à celle de l'identification d'instruments : extraction automatique de partition pour l'étude des styles musicaux en musicologie ou recherche par similarité de mélodie pour identifier les éventuels droits d'auteur d'un enregistrement dans l'industrie musicale [Dow03].

Définition

Entrées : observations $\mathbf{x}$, structure de $\mathbf{A}$, liste des noms des instruments présents $\mathcal{O}$.
Sorties : liste d'accords estimés joués par tous les instruments.
Informations *a priori* facultatives : base de données de sons instrumentaux.

Évaluation

La qualité d'un résultat d'identification de notes peut aussi se mesurer par des critères de comparaison de chaînes [Rap02]. Au cours de ce travail, nous ne ferons qu'un nombre limité d'expériences d'identification de notes. Nous ne proposons donc pas de mesure numérique de performance. Nous nous contenterons d'une visualisation des notes substituées, insérées ou supprimées.

Notons que certaines applications acceptent des distorsions entre la mélodie trouvée et la mélodie réelle. Par exemple, dans le cadre de la recherche par sifflotement (*query by humming*) [Dow03], il peut être possible de reconnaître une mélodie même avec quelques notes insérées ou supprimées.

Notons aussi que la définition de la tâche que nous avons choisie est basée uniquement sur la hauteur des notes, leur succession temporelle et les instruments associés. Nous ne prenons pas en compte la description du rythme, des nuances (*piano* ou *forte*) et la présence ou non de *vibrato*. Ces informations peuvent être très utiles dans le cadre de systèmes d'écoute automatiques [Sch00], mais sont plus difficiles à évaluer.

1.3 Définition et évaluation des tâches de séparation

Après avoir discuté des tâches de transcription, nous considérons maintenant deux tâches de séparation : extraction de sources et modification de scène sonore. De même nous donnons quelques exemples d'applications, nous définissons les tâches par la nature des entrées et sorties des algorithmes et nous construisons des mesures de performance adaptées.

1.3.1 Extraction de sources

L'extraction de sources consiste à extraire d'un mélange le signal correspondant à un instrument particulier décrit par son nom, sa direction spatiale ou la tessiture de sa mélodie. Parfois cet instrument ne peut pas être spécifié de façon unique, par exemple lorsque plusieurs instruments ont le même nom ou jouent des mélodies semblables en monocanal. Dans ce cas, cette tâche se résout en extrayant plusieurs signaux sources puis en sélectionnant le signal voulu manuellement. L'application principale visée est l'utilisation de ce signal comme échantillon pour la création de musique électronique.

Définition

Entrées : observations $\mathbf{x}$, structure de $\mathbf{A}$, instrument voulu j.
Sorties : source estimée $\widehat{s_j}$ ou image estimée sur un capteur $\widehat{s_{\mathrm{img}\,ij}}$.
Informations *a priori* facultatives : liste des noms des instruments présents $\mathcal{O}$, base de données de sons instrumentaux, partition musicale, segmentation temporelle, directions spatiales des sources.

Évaluation : discussion générale

Évaluer la performance d'un algorithme d'extraction de sources consiste à calculer la distorsion perçue entre la source estimée $\widehat{s_j}$ et la source réelle correspondante s_j, ou bien entre l'image estimée sur un capteur $\widehat{s_{\mathrm{img}\,ij}}$ et l'image réelle $s_{\mathrm{img}\,ij}$. Comment effectuer ce calcul est un problème complexe, que nous n'abordons pas en détail ici. Pour les besoins de ce document, nous nous contentons de définir précisément les quatre mesures de performance utilisées. De nombreux détails, des justifications théoriques et expérimentales de nos affirmations et des résultats commentés à écouter sont disponibles dans [Vin04a].

La plupart des mesures de performance pour l'extraction de sources existant dans la littérature [Lam99, Sch99] ont plusieurs limitations théoriques et prennent mal en compte les propriétés de l'audition. Premièrement la plupart ne s'appliquent qu'à certains types de mélanges et de résultats. La mesure d'interférence de [Lam99] suppose que le mélange est (sur-)déterminé, et l'erreur quadratique normalisée de [Sch99] fournit des résultats grossiers lorsque la source est mal estimée. Deuxièmement ces mesures ne tiennent pas compte du caractère transparent pour l'oreille de certaines distorsions complexes du signal. Ainsi la performance estimée est maximale uniquement lorsque $\widehat{s_j}$ égale s_j à un gain près, et peut être très mauvaise lorsque $\widehat{s_j} = f(s_j)$ avec f un filtre passe-tout. Troisièmement ces mesures fournissent

un critère de performance unique qui peut avoir des valeurs similaires pour des résultats perceptifs très différents. De ce point de vue, il est utile de distinguer trois types d'erreur au sein d'une source estimée $\widehat{s_j}$: les interférences provenant des autres sources, les restes de bruit de fond et le *bruit musical* [Cap93]. En effet le bruit musical a un son caractéristique de "glou-glou" souvent perçu comme plus gênant que les interférences, qui sont elles-mêmes perçues comme plus gênantes que le bruit de fond. Il est particulièrement important de mesurer séparément ces quantités sur les sources extraites de mélanges sous-déterminés, qui contiennent souvent un bruit musical important.

Les mesures que nous proposons dans la suite de ce paragraphe essayent d'éviter ces limitations. Dans un premier temps nous discutons l'évaluation d'une source estimée $\widehat{s_j}$. Nous choisissons une famille de distorsions autorisées plus complexes que de simples gains et nous décomposons $\widehat{s_j}$ sous la forme

$$\widehat{s_j} = s_{\text{dist}} + e_{\text{interf}} + e_{\text{bruit}} + e_{\text{artef}}, \tag{1.5}$$

où $s_{\text{dist}} = f(s_j)$ est la vraie source s_j à distorsion près f, et où e_{interf}, e_{bruit} et e_{artef} sont les termes d'erreur d'interférences, de bruit et d'artefacts, ce dernier terme regroupant les distorsions non autorisées des sources et le bruit musical. Nous définissons quatre mesures de performance à l'aide de rapports d'énergie par trames entre s_{dist}, e_{interf}, e_{bruit} et e_{artef}. Dans un deuxième temps nous étendons ces mesures à l'évaluation de l'image estimée d'une source sur un capteur $\widehat{s_{\text{img}\,ij}}$. Nous expliquons alors quels paramètres de calcul choisir pour que ces mesures rendent compte en partie des jugements perceptifs de qualité.

Évaluation : décomposition d'une source estimée

Nous choisissons comme famille de distorsions autorisées l'ensemble des filtrages variant dans le temps. La plupart des distorsions courantes correspondent à des cas particuliers de filtrages variant dans le temps et il est possible de sélectionner une famille de distorsions plus réduite par un simple réglage de paramètres.

Nous définissons un filtrage variant dans le temps à F sous-bandes et T trames de la façon suivante. Soit w_F une fenêtre de taille F de support $[0, F-1]$. Nous considérons le banc de filtres $(H_f)_{0\leq f\leq F-1}$ défini par $H_f(u) = w_F(u)\exp(2i\pi fu/F)$ et nous découpons la source s_j en signaux à bande limitée $(s_j^f)_{0\leq f\leq F-1}$ définis par $s_j^f = H_f \star s_j$. Soit maintenant w_T une fenêtre rectangulaire de taille L. Nous découpons les signaux à bande limitée en signaux à bande limitée et à support fini $(s_j^{tf})_{0\leq f\leq F-1,\,0\leq t\leq T-1}$ définis par $s_j^{tf}(u) = w_T(u-tL)s_j^f(u)$. Toute distorsion $f(s_j)$ s'exprime alors comme une somme pondérée des (s_j^{tf}) dont les coefficients dépendent de f.

Nous proposons de décomposer de la source estimée $\widehat{s_j}$ comme dans l'équation 1.5 grâce à une série de projections orthogonales. Projeter $\widehat{s_j}$ sur le sous-espace engendré par les (s_j^{tf}) permet de trouver la distorsion f qui rend $s_{\text{dist}} = f(s_j)$ le plus proche possible de $\widehat{s_j}$ au sens de la distance euclidienne. D'autres projections orthogonales permettent ensuite de décomposer le résidu en plusieurs termes d'erreur.

Soient $(s_{j'}^{tf})_{1\leq j'\leq n,\,0\leq f\leq F-1,\,0\leq t\leq T-1}$ et $(n_i^{tf})_{1\leq i\leq m,\,0\leq f\leq F-1,\,0\leq t\leq T-1}$ les signaux à bande limitée et à support fini correspondant aux sources $(s_{j'})_{1\leq j'\leq n}$ et aux bruits $(n_i)_{1\leq i\leq m}$. Pour rendre la décomposition unique et éviter les effets de bord, nous fixons le support de ces signaux à $[0, U+F-2]$, où $[0, U-1]$ est le support original des sources et des bruits. Notons $\Pi\{y_1, \ldots, y_k\}$ le projecteur orthogonal sur le sous-espace engendré par les signaux $y_1, \ldots, y_k$. Nous considérons les trois projecteurs

$$\Pi_{s_j} = \Pi\{(s_j^{tf})_{0\leq f\leq F-1,\,0\leq t\leq T-1}\}, \tag{1.6}$$

$$\Pi_{\mathbf{s}} = \Pi\{(s_{j'}^{tf})_{1\leq j'\leq n,\,0\leq f\leq F-1,\,0\leq t\leq T-1}\}, \tag{1.7}$$

$$\Pi_{\mathbf{s},\mathbf{n}} = \Pi\{\{(s_{j'}^{tf})_{1\leq j'\leq n}, (n_i^{tf})_{1\leq i\leq m}\}_{0\leq f\leq F-1,\,0\leq t\leq T-1}\}. \tag{1.8}$$

Et nous décomposons $\widehat{s_j}$ comme la somme des quatre termes

$$s_{\text{dist}} = \Pi_{s_j} \widehat{s_j}, \tag{1.9}$$

$$e_{\text{interf}} = \Pi_{\mathbf{s}} \widehat{s_j} - \Pi_{s_j} \widehat{s_j}, \tag{1.10}$$

$$e_{\text{bruit}} = \Pi_{\mathbf{s},\mathbf{n}} \widehat{s_j} - \Pi_{\mathbf{s}} \widehat{s_j}, \tag{1.11}$$

$$e_{\text{artef}} = \widehat{s_j} - \Pi_{\mathbf{s},\mathbf{n}} \widehat{s_j}. \tag{1.12}$$

Le calcul des projections orthogonales est un problème de moindres carrés. Le vecteur de coefficients de projection $\mathbf{c}$ défini par $\Pi\{y_1, \ldots, y_k\}\widehat{s_j} = \sum_{l=1}^{k} \overline{c_l} y_l$ se calcule par $\mathbf{c} = \mathbf{R}_{\mathbf{yy}}^{-1}[\langle \widehat{s_j}, y_1 \rangle, \ldots, \langle \widehat{s_j}, y_k \rangle]^H$, où $\mathbf{R}_{\mathbf{yy}}$ est la matrice de Gram des signaux $y_1, \ldots, y_k$ définie par $(\mathbf{R}_{\mathbf{yy}})_{ll'} = \langle y_l, y_{l'} \rangle$. Dans notre cas, les matrices de Gram correspondant aux projecteurs Π_{s_j}, $\Pi_{\mathbf{s}}$ et $\Pi_{\mathbf{s},\mathbf{n}}$ ont une structure diagonale par blocs car le découpage des signaux en trames produit des signaux de supports disjoints. Pour calculer les signaux projetés, il suffit alors d'effectuer des projections séparées sur chaque trame, puis de sommer les résultats.

Évaluation : calcul de rapports d'énergie

Nous mesurons l'importance relative des signaux projetés s_{dist}, e_{interf}, e_{bruit} et e_{artef} à l'aide de quatre rapports d'énergie par trames exprimés en décibels (dB). L'utilisation de rapports par trames vise à évaluer l'évolution de la performance au cours du temps lorsque la puissance de ces signaux varie.

À partir d'une fenêtre rectangulaire w'_T de taille L', nous découpons s_{dist} en T' trames $(s^{t'}_{\text{dist}})_{0 \leq t' \leq T'-1}$ définies par $s^{t'}_{\text{dist}}(u) = w'_T(u - t'L')s_{\text{dist}}(u)$. Nous définissons de même $(e^{t'}_{\text{interf}})_{0 \leq t' \leq T'-1}$, $(e^{t'}_{\text{bruit}})_{0 \leq t' \leq T'-1}$ et $(e^{t'}_{\text{artef}})_{0 \leq t' \leq T'-1}$. Ce découpage est indépendant du découpage en trames des sous-bandes des sources et des bruits pour le calcul des projections (en particulier les tailles des fenêtres L et L' peuvent être différentes). Nous calculons alors le Rapport Source à Distorsion

$$\text{RSD}^{\text{t}'} = 10 \log_{10} \frac{\|s^{t'}_{\text{dist}}\|^2}{\|e^{t'}_{\text{interf}} + e^{t'}_{\text{bruit}} + e^{t'}_{\text{artef}}\|^2}, \tag{1.13}$$

le Rapport Source à Interférences

$$\text{RSI}^{\text{t}'} = 10 \log_{10} \frac{\|s^{t'}_{\text{dist}}\|^2}{\|e^{t'}_{\text{interf}}\|^2}, \tag{1.14}$$

le Rapport Sources à Bruit

$$\text{RSB}^{\text{t}'} = 10 \log_{10} \frac{\|s^{t'}_{\text{dist}} + e^{t'}_{\text{interf}}\|^2}{\|e^{t'}_{\text{bruit}}\|^2}, \tag{1.15}$$

et le Rapport Sources à Artefacts

$$\text{RSA}^{\text{t}'} = 10 \log_{10} \frac{\|s^{t'}_{\text{dist}} + e^{t'}_{\text{interf}} + e^{t'}_{\text{bruit}}\|^2}{\|e^{t'}_{\text{artef}}\|^2}. \tag{1.16}$$

Ces quatre mesures sont inspirées de la définition habituelle du RSB, avec quelques modifications destinées à rendre le RSB et le RSA plus proches de la quantité perçue de bruit et d'artefacts lorsque le RSI est faible.

La performance peut se visualiser en traçant chaque mesure en fonction du numéro de trame t', ou bien en résumant les variations temporelles sous forme d'histogramme cumulatif [Cap93]. Nous définissons également des valeurs globales du RSD, du RSI, du RSB et du RSA comme étant les médianes de leurs valeurs par trames.

Évaluation : extension des mesures à l'évaluation d'une image spatiale estimée

L'évaluation d'une image spatiale estimée utilise les mêmes mesures de SDR, SIR, SNR et SAR à deux changements près. Premièrement la décomposition de $\widehat{s_j}$ dans les équations 1.5 et 1.9 à 1.12 est remplacée par la décomposition de $\widehat{s_{\mathrm{img}\,ij}}$. Deuxièmement nous définissons les signaux $(s_{j'}^{tf})$ utilisés dans les équations 1.6 à 1.8 en appliquant un banc de filtres et un fenêtrage aux images spatiales $(s_{\mathrm{img}\,ij'})_{1\leq j'\leq n}$ au lieu des sources elles-mêmes $(s_{j'})_{1\leq j'\leq n}$. Dans le cas de mélanges instantanés, les sources sont liées à leurs images spatiales par de simples gains donc la définition des projecteurs reste inchangée. Par contre, dans le cas de mélanges convolutifs, cela permet de considérer la réverbération de s_j et des $(s_{j'})_{j'\neq j}$ comme une partie naturelle de s_{dist} et e_{interf}, et non pas comme du bruit musical intégré à e_{artef}.

Évaluation : choix des paramètres

Dans la suite, nous ne chercherons pas à extraire les sources elles-mêmes des mélanges convolutifs, mais uniquement leurs images spatiales, ce qui est un problème plus facile. Pour cette raison, nous choisissons les paramètres de calcul suivants : $F = 1$, $L = 200$ ms et $L' = 200$ ms. Ce choix vise à rapprocher si possible les quatre mesures définies des jugements perceptifs de performance, et se base sur notre propre expérience d'écoute de résultats d'extraction de sources. Pour étudier cette question de façon plus rigoureuse il faudrait bien sûr effectuer une série de tests d'écoute sur plusieurs personnes.

La valeur de F signifie que les distorsions de filtrage des images spatiales des sources ne sont pas autorisées, ce qui permet par exemple de comptabiliser comme des erreurs des partiels manquants dans certaines notes. Les algorithmes d'extraction de sources souffrent généralement d'une indétermination de filtrage sur les sources extraites. Avec ce choix de F, un filtrage résiduel sur une source estimée est considéré comme une erreur par les mesures de performance mais pas forcément par l'audition (par exemple un filtre passe-tout est transparent auditivement). Ce problème disparaît avec l'extraction d'images spatiales, qui est soumise à une simple indétermination de gain.

La valeur de L permet d'autoriser des distorsions des sources à support temporel limité, par exemple des notes manquantes dans une mélodie. Dans ce cas, l'erreur locale n'affecte pas la perception de la performance du reste de la mélodie.

La valeur de L' a moins d'influence, en particulier sur la valeur globale (médiane) des mesures.

Notons que malgré ce choix de paramètres les mesures définies ne peuvent pas expliquer toutes les propriétés de l'audition. Par exemple la décomposition par projections ne peut pas expliquer pourquoi deux bruits blancs ont le même son même lorsqu'ils sont orthogonaux. Les phénomènes de masquage fréquentiel et temporel et la notion de sonie [Cap93] ne sont pas non plus pris en compte.

1.3.2 Modification de scène sonore

La modification de scène sonore vise à obtenir un nouveau mélange appelé *remix* correspondant au mélange des sources et des bruits de fond à l'aide de nouveaux filtres de mélange. Les tâches de suppression de source et de débruitage, qui consistent à *remixer* en mettant une source ou les bruits de fond à zéro, en sont des cas particuliers. Sans information *a priori* sur les instruments présents, toutes les sources sont modifiées de la même manière, par exemple en diminuant ou en augmentant leur écart spatial ou en égalisant leurs volumes. Dans le cas contraire, chaque source peut être traitée de façon adaptée, par exemple en augmentant le volume d'un instrument particulier. La modification de scène sonore a beaucoup d'applications, comme le *remastering* d'un CD [Rad02], la diffusion sur plusieurs canaux d'enregistrements stéréo [Ave02, Dre00], le *karaoke* (suppression de la voix dans une chanson) [Abr01] et la restauration de vieux enregistrements [Cap93].

Définition

Entrées : observations $\mathbf{x}$, structure de $\mathbf{A}$, nouveaux filtres de mélange $\mathbf{B}$ (définis explicitement ou par une méthode permettant de les calculer en fonction des sources).
Sorties : remix estimé $\widehat{\mathbf{x}_{\mathrm{rmx}}} = \widehat{\mathbf{B} \star \mathbf{s}}$.
Informations *a priori* facultatives : liste des noms des instruments présents $\mathcal{O}$, base de données de sons instrumentaux, partition musicale, segmentation temporelle, directions spatiales des sources.

Évaluation

La qualité d'un remix estimé $\widehat{\mathbf{x}_{\mathrm{rmx}}}$ se mesure en le comparant auditivement au remix attendu $\mathbf{x}_{\mathrm{rmx}} = \mathbf{B} \star \mathbf{s}$. Par simplicité nous considérons uniquement le cas des mélanges monocanal. En effet il est difficile en multicanal d'évaluer de façon pertinente auditivement les distorsions spatiales possibles, c'est-à-dire les erreurs sur la direction d'une source ou la transformation d'une source ponctuelle en une source spatialement étendue.

Les mesures que nous proposons sont dérivées des mesures pour l'extraction de sources. La famille de distorsions autorisées et les paramètres F, L et L' sont inchangés. Nous décomposons $\widehat{x_{\mathrm{rmx}}}$ sous la forme

$$\widehat{x_{\mathrm{rmx}}} = x_{\mathrm{rmx\,dist}} + e_{\mathrm{sourc}} + e_{\mathrm{bruit}} + e_{\mathrm{artef}}, \tag{1.17}$$

où $x_{\mathrm{rmx\,dist}}$ est le vrai remix x_{rmx} à distorsion près f, et où e_{sourc}, e_{bruit} et e_{artef} sont les termes d'erreur sur les sources, de bruit et d'artefacts. Nous ne considérons ici que le cas où la scène remixée est assez proche de la scène originale. En cas de suppression de sources, il faudrait séparer le terme d'erreur sur les sources en deux termes e_{sourc} et e_{interf} pour distinguer les erreurs sur les sources présentes dans le remix, qui sont assez subtiles à percevoir, et les interférences provenant des sources mal supprimées, qui sont beaucoup plus gênantes.

La décomposition fait intervenir les projecteurs $\Pi_{\mathbf{s}}$ et $\Pi_{\mathbf{s},\mathbf{n}}$ introduits dans les équations 1.7 et 1.8, ainsi que le projecteur

$$\Pi_{x_{\mathrm{rmx}}} = \Pi\{(x^{tf}_{\mathrm{rmx}})_{0\leq f\leq F-1,\,0\leq t\leq T-1}\}, \tag{1.18}$$

où les signaux $(x^{tf}_{\mathrm{rmx}})_{0\leq f\leq F-1,\,0\leq t\leq T-1}$ sont calculés en appliquant à x_{rmx} le même banc de filtres et les mêmes fenêtrages que pour calculer $(s^{tf}_{j'})_{1\leq j'\leq n,\,0\leq f\leq F-1,\,0\leq t\leq T-1}$ et $(n^{tf}_{i})_{1\leq i\leq m,\,0\leq f\leq F-1,\,0\leq t\leq T-1}$. Nous définissons les termes de la décomposition par

$$x_{\mathrm{rmx\,dist}} = \Pi_{x_{\mathrm{rmx}}}\widehat{x_{\mathrm{rmx}}}, \tag{1.19}$$

$$e_{\mathrm{sourc}} = \Pi_{\mathbf{s}}\widehat{x_{\mathrm{rmx}}} - \Pi_{x_{\mathrm{rmx}}}\widehat{x_{\mathrm{rmx}}}, \tag{1.20}$$

$$e_{\mathrm{bruit}} = \Pi_{\mathbf{s},\mathbf{n}}\widehat{x_{\mathrm{rmx}}} - \Pi_{\mathbf{s}}\widehat{x_{\mathrm{rmx}}}, \tag{1.21}$$

$$e_{\mathrm{artef}} = \widehat{x_{\mathrm{rmx}}} - \Pi_{\mathbf{s},\mathbf{n}}\widehat{x_{\mathrm{rmx}}}. \tag{1.22}$$

À partir de ces quatre termes découpés en trames indexées par t' nous calculons le Rapport Remix à Distorsion

$$\mathrm{RRD}^{\mathrm{t}'} = 10\log_{10}\frac{\|x^{t'}_{\mathrm{rmx\,dist}}\|^2}{\|e^{t'}_{\mathrm{sourc}} + e^{t'}_{\mathrm{bruit}} + e^{t'}_{\mathrm{artef}}\|^2}, \tag{1.23}$$

le Rapport Remix à Erreur sur les Sources

$$\mathrm{RRES}^{\mathrm{t}'} = 10\log_{10}\frac{\|x^{t'}_{\mathrm{rmx\,dist}}\|^2}{\|e^{t'}_{\mathrm{sourc}}\|^2}, \tag{1.24}$$

le Rapport Remix à Bruit

$$\mathrm{RRB}^{\mathrm{t}'} = 10\log_{10}\frac{\|x^{t'}_{\mathrm{rmx\,dist}} + e^{t'}_{\mathrm{sourc}}\|^2}{\|e^{t'}_{\mathrm{bruit}}\|^2}, \tag{1.25}$$

et le Rapport Remix à Artefacts

$$\mathrm{RRA}^{\mathrm{t}'} = 10\log_{10}\frac{\|x^{t'}_{\mathrm{rmx\,dist}} + e^{t'}_{\mathrm{sourc}} + e^{t'}_{\mathrm{bruit}}\|^2}{\|e^{t'}_{\mathrm{artef}}\|^2}. \tag{1.26}$$

1.4 Notion de difficulté

Les mesures de performance décrites dans les paragraphes précédents permettent de classer plusieurs algorithmes résolvant la même tâche selon la performance de leurs résultats sur un seul mélange. Pour intégrer les résultats sur des mélanges différents, nous devons aussi être capables de classer les mélanges selon leur difficulté.

Une première famille de critères de difficulté est basée sur les filtres de mélange $\mathbf{A}$ [Sch99]. Dans les mélanges déterminés non bruités convolutifs courts, il existe des filtres de séparation invariants dans le temps $\mathbf{W} = \mathbf{A}^{-1}$ généralement faciles à estimer [Car98a]. Il est alors possible d'extraire les sources par $\mathbf{s} = \mathbf{W}\mathbf{x}$, et ensuite de les transcrire ou de les remixer. Dans les mélanges convolutifs longs ou variant dans le temps, ces filtres de séparation existent toujours en théorie. Mais leur taille élevée par rapport au nombre d'observations disponibles les rend difficiles à estimer car les algorithmes atteignent leurs limites pratiques (problèmes de maxima locaux des fonctions à optimiser par exemple) [Ser03]. Dans les mélanges sous-déterminés ou bruités, et en particulier dans les mélanges monocanal, ces filtres n'existent généralement pas et l'extraction repose sur un filtrage variant dans le temps [Gri03a]. Cela augmente encore le nombre de paramètres inconnus à estimer par rapport à la quantité d'information disponible.

Une deuxième famille de critères de difficulté consiste à décrire la quantité d'information *a priori* sur les sources et sur les filtres de mélange. L'utilisation d'une base de données de sons instrumentaux peut rendre les tâches considérées plus faciles en permettant l'apprentissage de modèles de sources spécifiques aux instruments potentiellement présents. Une partition musicale (liste d'accords et de rythmes joués par chaque instrument), une segmentation temporelle (instants d'entrée et de sortie de chaque instrument) ou une localisation des sources (azimuts par rapport aux capteurs) apportent aussi des informations sur un mélange précis qui facilitent la séparation. Dans certains cas intermédiaires, une seule source est modélisée de façon précise et la partition disponible est incomplète [Mer98].

Une troisième famille de critères de difficulté est liée à la ressemblance entre les sources au sein d'un mélange particulier. Par exemple l'extraction de sources est plus difficile lorsque les directions spatiales des sources sont proches [Sch99]. Et l'identification de notes et d'instruments est plus difficile lorsque plusieurs instruments ont le même nom, lorsqu'ils ont une tessiture ou un timbre proche, ou lorsqu'ils jouent des notes à intervalle harmonique créant des situations de masquage [Egg03].

Cette liste montre qu'il existe de nombreux critères de difficulté, mais pas de véritables mesures de difficulté actuellement.

Si possible les mesures de difficulté devraient être capables de réduire le nombre de critères. Ainsi il serait utile d'avoir une mesure unique correspondant à la taille des filtres et au nombre de sources.

Les mesures de difficulté devraient aussi pondérer les critères de façon différente selon la tâche considérée. Par exemple, les masquages entre sources sont gênants pour l'extraction de sources, mais parfois sans importance pour la modification de scène sonore car les erreurs d'estimation d'une source

peuvent être masquées au sein du remix par la présence des autres sources. De même l'utilisation de filtres de mélange à réponse non plate importe peu pour l'identification de notes car elle ne change pas leur caractère harmonique, mais elle peut influencer l'identification d'instruments en modifiant leur timbre.

Des mesures de difficulté vérifiant ces contraintes pourraient se définir en mesurant les limitations théoriques des algorithmes ou bien leurs performances expérimentales moyennes en fonction des critères ci-dessus.

Par la suite, nous ne fournirons pas de mesure de difficulté globale pour caractériser les différents mélanges testés.

Chapitre 2

État de l'art et objectifs

Ce deuxième chapitre étudie les méthodes existantes de transcription et de séparation sous le point de vue des deux types d'informations utilisées : celles provenant de l'analyse des mélanges et celles provenant des modèles de sources. Dans le paragraphe 2.1 nous décrivons la structure du son instrumental et sa traduction dans les modèles de sources les plus courants. Dans le paragraphe 2.2 nous rappelons les différents types d'enregistrements de musique de chambre et les analyses généralement effectuées pour en extraire des quantités pertinentes. Nous présentons ensuite un état de l'art des méthodes de transcription et de séparation pour les mélanges monocanal dans le paragraphe 2.3 et pour les mélanges multicanal dans le paragraphe 2.4. Enfin, dans le paragraphe 2.5, nous résumons les limitations de ces méthodes en terme de performance et de difficulté des mélanges étudiés, et nous formulons les objectifs de ce travail.

Ce chapitre ne contient pas de résultats originaux. Les limitations de certains algorithmes, constatées simultanément par d'autres auteurs, sont plus détaillées dans notre mémoire de DEA [Vin01] et dans un article [Vin04d].

2.1 Informations fournies par les modèles de sources

2.1.1 Structure du son instrumental

La notion fondamentale de la musique occidentale est celle de note. Un instrument de percussion peut jouer une seule note. La majorité des autres instruments peuvent jouer plusieurs notes, correspondant à un nombre fini de hauteurs différentes sur une échelle en demi-tons. La structure du son instrumental peut se décrire par trois types de structures : premièrement l'évolution des paramètres instantanés du son au sein d'une note, deuxièmement le regroupement de notes en phrases musicales et en accords, et troisièmement les contraintes liées au jeu en commun de plusieurs instruments. Nous abordons ces trois points successivement.

Description instantanée d'une note : hauteur, volume et timbre

Deux grandes familles d'instruments sont distinguées en musique et dans la littérature [Mar99b, Jen99], selon que l'oscillation produisant le son est libre ou entretenue. La première famille regroupe les percussions, les cordes pincées ou *pizzicato* et les cordes frappées. La deuxième est constituée des cordes frottées, des bois (à anche double, simple ou à air) et des cuivres (saxophones, autres cuivres bouchés ou non).

Une note d'un instrument à oscillations entretenues est constituée de trois parties successives : l'attaque, la partie entretenue et la relâche (*attack, sustain, release*). Durant la partie entretenue le son est constitué d'une somme de partiels sinusoïdaux presque harmoniques, dont les fréquences et les amplitudes varient lentement. Quelquefois une composante de souffle bruitée (comme dans la flûte) ou des

partiels supplémentaires non harmoniques s'ajoutent. Le transitoire d'attaque voit les partiels apparaître successivement, en commençant souvent par les plus graves, et peut contenir un bruit à bande large. Le transitoire de relâche voit les partiels et leur réverbération par la salle disparaître progressivement. La durée de l'attaque est de 5 à 500 ms selon l'instrument, le mode de jeu et la tessiture. La durée de la partie entretenue dépend du choix de l'instrumentiste et celle de la relâche du temps de réverbération de la salle.

Une note d'un instrument à oscillations libres est constituée des mêmes parties moins la partie entretenue. Le son contient seulement le transitoire d'attaque (qui peut ne pas contenir de partiels harmoniques dans le cas des percussions) et sa réverbération. L'instrumentiste peut limiter la durée de la note par étouffement.

À un instant donné de l'attaque, de la partie entretenue ou de la relâche, le son est décrit par trois variables instantanées : la hauteur, le volume et le timbre. La hauteur d'un son correspond à sa fréquence fondamentale (pour les instruments produisant des partiels harmoniques), le volume à sa puissance totale et le timbre à tous ses autres nombreux paramètres (inharmonicité, rapports de puissance entre partiels, *etc*). Le timbre dépend en partie des deux autres variables. Par exemple, pour la plupart des instruments, la puissance relative des partiels aigus diminue régulièrement plus la hauteur de la note est aiguë ou plus la nuance de jeu est *piano*. Cette diminution peut être localement moins régulière en présence de changements de registre ou de résonances étroites du corps de l'instrument.

Cette définition physique du timbre instantané ne doit pas être confondue avec celle du timbre d'une note ou d'un instrument, qui se définit comme une quantité perceptive et prend en compte non seulement les valeurs du timbre instantané mais aussi les variations de fréquence et de volume. Plus précisément des expériences de psycho-acoustique ont montré que le timbre d'une note est lié principalement au temps d'attaque, au centroïde spectral et à l'écartement spectral [Mar03].

La hauteur, le volume et le timbre instantanés évoluent de façon en partie prévisible selon l'instrument et peuvent aussi être modifiées par l'instrumentiste au cours de la partie entretenue. La hauteur est instable durant l'attaque. Durant la partie entretenue elle peut être stable, varier de façon contrôlée en jeu *vibrato* ou de façon aléatoire (*jitter*). Le volume augmente durant l'attaque, parfois par à-coups (*blips*) et diminue durant la relâche. Durant la partie entretenue il peut être stable, varier de façon contrôlée en jeu *crescendo*, *decrescendo* ou *tremolo* ou de façon aléatoire (*shimmer*). Enfin le timbre varie selon l'instrument, sa facture (unique), les conditions d'enregistrement et le mode de jeu.

Formation de phrases musicales et d'accords

Un instrument joue rarement des notes isolées, mais plutôt des phrases musicales comportant des transitions entre notes successives. Les caractéristiques de ces transitions dépendent du style de jeu *legato*, *staccato* ou *portamento*, mais aussi de l'intervalle entre les notes [Str85]. Cela se manifeste en particulier sur la profondeur et la durée de la baisse de volume entre deux notes successives. Des expériences de psycho-acoustique ont prouvé que la forme des transitions *legato* pouvait entrer en compte dans la perception du timbre d'un instrument [Dup99].

Certains instruments dits polyphoniques (par opposition à monophoniques) peuvent aussi jouer des accords formés de plusieurs notes simultanées. Le son d'un accord est constitué de la superposition du son de chaque note, à quelques non-linéarités près dues au couplage des oscillations sur la table d'harmonie.

La superposition de notes au sein d'un accord ou la succession de notes au sein d'une phrase sont réglées par les principes de l'harmonie et le style musical. La musique occidentale impose souvent des notes à intervalle harmonique dans un accord ou des intervalles particuliers entre notes successives dans une mélodie.

Contraintes de la musique de chambre

Lorsque plusieurs instrumentistes jouent ensemble, le jeu de chacun est à la fois dépendant et indépendant de celui des autres. Par exemple en musique d'ensemble ou d'orchestre la règle veut même que plusieurs instruments (de même nom) jouent la même partition simultanément, mais cela n'est pas vrai en musique de chambre. Plus généralement la musique occidentale favorise les notes à intervalle harmonique et les attaques les plus synchronisées possible, mais deux notes provenant d'instruments différents ont toujours des petites fluctuations de fréquence aléatoires et il est difficile pour les instrumentistes de produire des attaques parfaitement synchronisées.

2.1.2 Quelques modèles de sources usuels

Les modèles de sources utilisés dans la littérature sont à l'image de la structure du son instrumental. Ils comprennent généralement plusieurs niveaux d'hypothèses : sur le son à un instant donné, sur son évolution temporelle et sur les relations entre les sons d'instruments différents en musique de chambre. Dans la suite de ce paragraphe, nous décrivons trois modèles de sources parmi les plus courants, en regroupant des approches souvent présentées de façon différente. Les deux premiers modélisent directement les formes d'onde des sources, alors que le dernier modélise plus généralement des observations extraites des sources. Nous verrons que certaines méthodes de transcription et de séparation de musique de chambre utilisent plusieurs modèles à la fois.

Sinusoïdes et règles de groupement auditif

Un premier modèle de sources consiste à analyser une source comme la somme de plusieurs partiels sinusoïdaux et d'un bruit résiduel [Ser90, Dav02c, Ros02, Cem03]. En notant s_j la source analysée, $(e_{jh})_{h\in\mathbb{N}}$ les amplitudes instantanées des partiels, $(\omega_{jh})_{h\in\mathbb{N}}$ leurs fréquences instantanées, $(\delta_{jh})_{h\in\mathbb{N}}$ leurs phases et ϵ_j le bruit résiduel, cela s'écrit

$$s_j(u) = \sum_{h\in\mathbb{N}} e_{jh}(u)\cos(\omega_{jh}(u)u + \delta_{jh}) + \epsilon_j(u). \tag{2.1}$$

Certains modèles distinguent aussi les transitoires d'attaque et le bruit de fond stationnaire au sein du bruit résiduel [Ell96].

Les partiels d'une même source à un instant donné ou à des instants différents suivent les six règles de l'Analyse de Scènes Auditives (ASA) : similarité, proximité, continuité, variation commune, symétrie et connexité [Bre90]. Issues d'expériences de psycho-acoustique, ces règles expliquent comment l'oreille perçoit plusieurs sources sonores séparées dans un mélange. Par exemple des partiels d'une même note vérifient les propriétés suivantes : harmonicité et apparition simultanée [Ell96], variation commune de fréquence et d'amplitude [Sch00] et continuité de l'enveloppe spectrale [Vir03a]. Ces règles supposent implicitement que des sons venant de sources différentes ne vérifient pas toutes ces règles à la fois. Par exemple des partiels de notes différentes sont inharmoniques, ou bien apparaissent à des instants différents, ou bien subissent des variations de fréquence et d'amplitude sans rapport, ou bien forment une enveloppe spectrale chahutée. Les règles de l'ASA peuvent être complétées par des règles musicales voisines, régissant le rythme [Cem03], les intervalles typiques entre notes successives ou le fait que la basse et la mélodie se chevauchent rarement [Kas99]. Toutes ces règles sont modélisées soit comme des sources de connaissance à appliquer successivement [Ell96], soit comme des distributions de probabilité *a priori* [Dav02c, Ros02, Cem03].

Le son d'une source instrumentale donnée peut se décrire par plusieurs paramètres calculables à partir de son analyse en partiels : enveloppe spectrale, vitesse d'attaque et de relâche, inharmonicité, quantité de *jitter* et de *shimmer*, *etc*. Ces paramètres peuvent être appris sur des enregistrements solo [Jen99, Mar99b, Kit03].

Parcimonie et indépendance

Un deuxième modèle de sources consiste à décomposer une source comme la somme de plusieurs signaux appelés atomes, qui ne sont pas forcément des sinusoïdes, plus un résidu éventuel. Si s_j est la source modélisée, $\mathcal{D} = (\Phi_h)_{h\in\mathbb{N}}$ le dictionnaire contenant les atomes et ϵ_j le résidu, la décomposition s'écrit [Gri99, Mol03]

$$s_j = \sum_{h\in\mathbb{N}} e_{jh}\Phi_h + \epsilon_j. \tag{2.2}$$

Les coefficients de décomposition $(e_{jh})_{h\in\mathbb{N}}$ sont choisis de sorte à réaliser un compromis entre la qualité de représentation et le nombre (fini) d'atomes utilisés. Le résidu éventuel est généralement modélisé comme un bruit gaussien et les coefficients de décomposition comme des variables parcimonieuses (*sparse*), c'est-à-dire dont la densité de probabilité a un pic en zéro et des "queues lourdes". Des exemples de distributions parcimonieuses sont la distribution laplacienne [Lee99], les distributions de kurtosis positif [Car98a] et certaines distributions exponentielles généralisées [Pen01, Zib01]. Il est aussi possible d'en construire par mélange entre distributions de Dirac et distributions gaussiennes [Vie01], ou par mélange de distributions gaussiennes de variances différentes [Att99].

Les coefficients de décomposition sont le plus souvent supposés indépendants et identiquement distribués. Dans le cas où $\mathcal{D}$ contient des sinusoïdes à support fini, l'harmonicité et la continuité temporelle peuvent être favorisées en groupant les atomes de même support en sous-espaces constitués d'atomes harmoniques [Gri99] et en modélisant l'évolution temporelle des coefficients à chaque fréquence par une chaîne de Markov [Mol03]. D'autres structures de dépendance entre coefficients sont possibles [Pen01, Pen00, Hyv00, Car98b].

Des sources différentes sont modélisées en supposant que leurs coefficients de décomposition sur un même dictionnaire sont indépendants [Car98a] voire non nuls uniquement sur des sous-dictionnaires disjoints propres à chaque source [Ben01, Dub02, Car98b].

Les dictionnaires courants sont des dictionnaires fixés correspondant à des formes d'ondes simples : Diracs, ondelettes [Che98], atomes de Gabor [Gri99, Wol03], ondelettes et cosinus [Mol03]. Certains sont des bases (ils permettent de décomposer tout signal de façon unique sans résidu), d'autres sont redondants (ils contiennent une base). Ces dictionnaires représentent bien la majorité des sources instrumentales, dont l'énergie est concentrée sur les attaques et les partiels avec parfois des grandes zones de silence. Mais la faible structuration des coefficients de décomposition constitue un défaut par rapport au modèle de sources précédent, puisque le timbre des sources n'est pas modélisé. Ce défaut peut être compensé par l'utilisation d'un dictionnaire adapté à une source particulière, construit par exemple par translation de formes d'onde prototypes extraites d'enregistrements solo [McD03, Bel02] ou de bases de données de notes isolées [Kas99].

Mélange de Gaussiennes et indépendance

Un troisième modèle de sources consiste à représenter par un Mélange de Gaussiennes (MG) une suite d'observations extraites d'une source, par exemple son spectre de puissance à court terme [Ben03] ou son cepstre [Ero03]. Contrairement aux deux modèles précédents, celui-ci permet d'analyser une source mais pas toujours de la resynthétiser. Une suite de vecteurs observés $(\mathbf{m}_{jt})_{0\leq t\leq T-1}$ est modélisée par

$$\mathbf{m}_{jt} = \mathbf{\Phi}_{h_{jt}} + \boldsymbol{\epsilon}_{h_{jt}}, \tag{2.3}$$

où h_{jt} est un état caché à valeurs dans un espace d'états fini $[\,1, H_j]$ et à distribution multinômiale fixée, et $\boldsymbol{\epsilon}_{h_{jt}}$ est un bruit gaussien de moyenne nulle dont la covariance dépend de h_{jt}. Il est possible d'introduire des facteurs d'échelle dans le modèle de façon à découpler l'amplitude des observations et les observations normalisées [Ben03].

Les états cachés $(h_{jt})_{0\leq t\leq T-1}$ d'un MG sont supposés indépendants. Mais la structure temporelle d'une source peut être mieux représentée en les modélisant par une chaîne de Markov à valeurs dans l'espace d'états. Le modèle complet combinant probabilités de transition markoviennes et densités d'émission gaussiennes se nomme alors Modèle de Markov Caché (MMC) [Rab89, Gha01].

Des sources différentes sont généralement modélisées par un MMC factoriel [Gha97] où chaque source s_j correspond à une chaîne de Markov $(h_{jt})_{0\leq t\leq T-1}$ indépendante des autres. La chaîne de Markov factorielle $(h_{1,t},\ldots,h_{n,t})$ est à valeurs dans le produit cartésien des espaces d'états des sources $[1,H_1]\times\ldots\times[1,H_n]$.

Les paramètres du son d'une source donnée pour un MG ou un MMC sont le plus souvent appris sur une base de données de notes isolées [Ero03] ou sur un ensemble d'enregistrements solo [Bro01], éventuellement étiquetés par des partitions musicales [Rap02].

2.2 Informations fournies par l'analyse des mélanges

Comme pour l'information relative aux sources, il existe deux façons de modéliser l'information relative à un mélange : exploiter directement les formes d'onde des canaux du mélange, ou en faire une première analyse (*front end*) qui écarte une partie de l'information inutile. Ce paragraphe rappelle les types de mélanges rencontrés en musique de chambre et quelques analyses courantes. Nous distinguons les analyses spectro-temporelles en monocanal des analyses spatiales en multicanal.

2.2.1 Types de mélanges

Les enregistrements de musique de chambre se trouvent surtout sous deux formes : les enregistrements multipistes (huit ou seize pistes en pratique) et les CD audio stéréo.

Enregistrements multipistes

En musique jazz ou pop, les instrumentistes sont souvent enregistrés séparément en studio insonorisé ou en situation de concert avec des instruments électrifiés insensibles aux vibrations acoustiques des autres instruments. Les enregistrements multipistes correspondants ne sont pas vraiment des mélanges, puisque chaque canal contient une source séparée.

En musique classique ou contemporaine, les musiciens jouent ensemble dans une salle de concert et sont enregistrés avec une batterie de micros. La disposition des micros est généralement la suivante [Bar98] : un micro stéréo "lointain" placé à distance des musiciens et un micro mono "d'appoint" à proximité de chacun (éventuellement plusieurs pour les instruments spatialement étendus comme la batterie ou le piano). Les micros d'appoint sont des capsules directives. Le micro lointain peut être de plusieurs types, dont les plus courants sont : XY (deux capsules cardioïdes coïncidentes pointant vers l'instrumentiste le plus à gauche et celui le plus à droite), ORTF (deux capsules cardioïdes à une distance de 17 cm formant un angle de 110°), binaural (deux capsules omnidirectionnelles séparés par une tête artificielle) et AB étroit (deux capsules omnidirectionnelles à une distance de 40 cm). Le choix est effectué en fonction du style de diffusion voulu : un micro XY ou ORTF est préférable pour une diffusion sur hauts-parleurs, un micro binaural pour une diffusion sur écouteurs à un auditeur précis et un micro ORTF pour une diffusion sur écouteurs à un auditeur quelconque. Le micro ORTF constitue un bon compromis lorsque le style de diffusion n'est pas fixé.

Un enregistrement multipistes de ce type est un mélange sur-déterminé convolutif long variant dans le temps. Les filtres de mélange sont constitués de deux parties : une succession de pics distincts dus aux réflexions précoces localisées sur les murs de la salle, puis un signal plus continu dû aux réflexions tardives diffuses formant la réverbération [Ave02]. Généralement l'ensemble des réflexions précoces dure

50 ms environ et la réverbération une seconde ou plus. Les petits mouvements des musiciens (de l'ordre de quelques centimètres) affectent peu les réflexions précoces, qui peuvent donc être plus ou moins prédites connaissant la disposition des instruments et des micros et les caractéristiques de la salle. Par contre ces mouvements modifient complètement la réverbération, qui prend un caractère aléatoire. Les signaux des micros d'appoint sont caractérisés par une faible quantité relative d'interférences (phénomène de "repisse"), de réflexions précoces et de réverbération, contrairement au signal du micro lointain.

CD stéréo

Un CD stéréo de jazz ou de pop est un mélange en studio des instruments enregistrés séparément. Une technique de mélange courante consiste à faire un mélange panoramique (*panning*) à partir de gains positifs, puis à rajouter de la réverbération artificielle au tout [Ave02, Bar98]. Des gains variables ou des filtrages courts peuvent aussi être utilisés. Les mélanges résultants sont donc généralement sous-déterminés et convolutifs (courts ou longs) et varient parfois dans le temps.

Un CD stéréo de musique classique ou contemporaine est réalisé par remixage en studio d'un enregistrement multipistes du type décrit ci-dessus. Les signaux mono d'appoint sont utilisés pour corriger à des instants particuliers le volume ou le timbre des instruments sur le signal stéréo lointain. De cette façon la perception des caractéristiques de la salle d'enregistrement et de la disposition spatiale des musiciens est préservée [Bar98]. Les CD sont alors des mélanges sous-déterminés convolutifs longs variant dans le temps.

Enregistrements mono

Notons aussi l'existence d'enregistrements mono réalisés par ajout des deux canaux d'un mélange stéréo instantané (provenant lui-même d'un mélange en studio ou d'un enregistrement stéréo XY). Ces enregistrements sont rares en pratique. Mais ils sont très présents dans la littérature sur la transcription, qui prend peu en compte les informations spatiales en stéréo.

2.2.2 Analyse spectro-temporelle

Spectre à court terme

Une première analyse possible de ces différents types de mélanges est d'en extraire les propriétés spectro-temporelles en calculant leur spectre à court terme. Il existe plusieurs définitions non équivalentes de cette quantité. Nous présentons ici la définition par banc de filtres [Ell96].

Soit x_i un canal du mélange, $(H_f)_{0\leq f\leq F-1}$ un banc de filtres passe-bande et w_T une fenêtre. Le signal x_i est découpé en signaux à bande limitée $(x_i^f)_{0\leq f\leq F-1}$ définis par

$$x_i^f(u) = \sum_{\tau=-\infty}^{+\infty} H_f(\tau) x_i(u-\tau), \tag{2.4}$$

puis en signaux à bande limitée et à support fini $(x_i^{tf})_{0\leq f\leq F-1,\, 0\leq t\leq T-1}$ définis par

$$x_i^{tf}(u) = w_T(u - tL) x_i^f(u), \tag{2.5}$$

où L est le pas entre deux trames successives. Au point temps-fréquence (t, f) (dans la sous-bande f et la trame t), le spectre de log-puissance de x_i vaut alors $\log \|x_i^{tf}\|^2$, le spectre d'amplitude $\|x_i^{tf}\|$ et le spectre de puissance $\|x_i^{tf}\|^2$. Les filtres $(H_f)_{0\leq f\leq F-1}$ sont généralement espacés sur une échelle de fréquence linéaire, logarithmique ou ERB [Rom03], avec une largeur de bande proportionnelle à l'espacement entre filtres successifs.

Pour éviter une chute de la log-puissance vers $-\infty$ lorsque $\|x_i^{tf}\| \approx 0$, la valeur calculée peut-être mise à zéro en-dessous d'un certain seuil de silence. Lorsque ce seuil dépend de la fréquence, la log-puissance est alors définie relativement au seuil [Vir03b] par

$$o_{itf} = \log\left(\frac{\|x_i^{tf}\|^2}{g_f^2} + 1\right), \tag{2.6}$$

où $(g_f)_{0\leq f\leq F-1}$ est le seuil de silence en amplitude.

Cepstre

Le spectre de log-puissance peut être paramétré à transformation linéaire près. Par exemple la transformée de Fourier inverse du spectre de log-puissance est appelée cepstre. Le cepstre de x_i dans la trame t et à la "quéfrence" q vaut [Mar99a]

$$o'_{itq} = \frac{1}{F}\sum_{f=0}^{F-1} o_{itf} \exp(2i\pi qf/F). \tag{2.7}$$

Les premiers coefficients du cepstre ($q = 0, 1, 2\ldots$) décrivent l'enveloppe spectrale du son. D'autres changements de paramétrage linéaires peuvent être utiles [Cas02, Ero03].

Corrélogramme

Lorsque les filtres sont à bande large, les spectres à court terme sont trop grossiers pour distinguer les pics spectraux induits par des partiels d'un son instrumental. L'information de périodicité au sein d'un signal à bande limitée x_i^f peut alors se calculer grâce à son auto-corrélation locale définie pour le délai l dans la trame t par [Ell96]

$$c_{itfl} = \sum_u w_T^2(u - tL)x_i^f(u)\overline{x_i^f(u-l)}. \tag{2.8}$$

L'ensemble des (c_{itfl}) forme une représentation à trois dimensions appelée corrélogramme. Dans le cas d'un banc de filtres auditifs (espacement fréquentiel sur l'échelle ERB et largeur de bande égale à un ERB), le corrélogramme permet de rendre compte à la fois des effets de masquage auditif et de perception de la hauteur [Ell96].

2.2.3 Analyse spatiale

Différence de phase et de volume inter-canal

Une deuxième analyse possible d'un mélange est d'en extraire les propriétés spatiales en comparant ses canaux deux par deux. La comparaison se fait habituellement en calculant une différence de phase et de volume en chaque point temps-fréquence. La différence de phase entre les canaux i' et i au point (t, f) vaut

$$d_{i'itf}^{\text{pha}} = \angle\langle x_i^{tf}, x_{i'}^{tf}\rangle, \tag{2.9}$$

et la différence de volume [Rom03]

$$d_{i'itf}^{\text{vol}} = \log\|x_{i'}^{tf}\|^2 - \log\|x_i^{tf}\|^2. \tag{2.10}$$

Ce calcul est valable lorsque le banc de filtres $(H_f)_{0\leq f\leq F-1}$ est complexe, c'est-à-dire lorsque les filtres ont une réponse nulle en fréquence négative. Dans le cas d'un banc de filtres réel, $d_{i'itf}^{\text{pha}}$ est le retard pour lequel la corrélation entre x_i^{tf} et $x_{i'}^{tf}$ retardé est maximale [Rom03].

Lorsque les canaux i' et i correspondent aux deux capsules d'un micro stéréo, l'utilisation de ces quantités dépend du type de micro. Pour un micro XY seule la différence de volume est significative, pour un micro AB étroit seule la différence de phase l'est, et pour un micro binaural ou ORTF les deux quantités le sont. De plus il est possible de relier ces quantités à l'azimut observé en chaque point temps-fréquence (t, f). Par exemple pour un micro ORTF ou AB étroit la différence de phase inter-canal s'exprime en fonction de l'azimut θ_{tf} par la relation de formation de voies (*beamforming*) [Vis03]

$$d^{\mathrm{pha}}_{i'itf} = 2\pi \frac{fd}{c} \sin\theta_{tf} \mod 2\pi, \tag{2.11}$$

où f est la fréquence centrale de la sous-bande, d est la distance entre capsules, c la vitesse du son et $\theta_{tf} = \pm\pi/2$ correspond à l'axe des capsules. En dessous d'une fréquence critique c/d cette relation permet de déterminer $\sin\theta_{tf}$ en fonction de $d^{\mathrm{pha}}_{i'itf}$ de façon unique, mais au-dessus il peut exister une indétermination qui nécessite l'utilisation de $d^{\mathrm{vol}}_{i'itf}$ conjointement à $d^{\mathrm{pha}}_{i'itf}$ [Vis03]. De façon similaire la localisation auditive fait intervenir les informations de phase en-deçà de 1500 Hz environ et les informations de volume au-delà [Rom03].

Lorsque les canaux i' et i proviennent d'un mélange synthétique instantané, la différence de volume inter-canal est liée à l'azimut observé θ_{tf} par la loi de mélange panoramique [Pul01]

$$\tan\theta_{tf} = \tan\theta_0 \tanh\frac{d^{\mathrm{vol}}_{i'itf}}{4}, \tag{2.12}$$

où $\theta_0 \simeq 30^\circ$ est le demi-angle entre les hauts-parleurs de diffusion.

Cohérence inter-canal

Une autre quantité utile est la cohérence entre les canaux i' et i définie au point temps-fréquence (t, f) par [Ave02]

$$d^{\mathrm{coh}}_{i'itf} = \frac{|\langle x^{tf}_i, x^{tf}_{i'}\rangle|}{\|x^{tf}_i\|\|x^{tf}_{i'}\|} \tag{2.13}$$

lorsque le banc de filtres $(H_f)_{0\leq f\leq F-1}$ est complexe, ou par le maximum de la valeur absolue de la corrélation entre x^{tf}_i et $x^{tf}_{i'}$ retardé lorsque le banc de filtres est réel. Cette quantité est proche de 1 lorsque le son au point (t, f) est constitué d'une seule source et de ses réflexions précoces. Elle est inférieure lorsque le son est constitué de plusieurs sources ou de réverbération [Ave02].

2.3 Méthodes pour les enregistrements monocanal

Jusqu'à présent nous avons décrit l'information contenue dans les modèles de sources usuels et l'information extraite des mélanges. Dans ce paragraphe nous expliquons comment les algorithmes existants combinent ces deux types d'information pour la transcription et la séparation d'enregistrements monocanal. Nous portons une attention particulière aux algorithmes utilisant des modèles de sources appris, et nous mentionnons aussi quelques algorithmes intéressants appliquées à des enregistrements solo ou à la parole. Dans un souci de lisibilité nous séparons cet état de l'art en cinq catégories qui correspondent à peu près aux dénominations utilisées dans la littérature. Chacune est décrite par ses conditions d'application, ses hypothèses, quelques algorithmes, un commentaire global de leurs résultats et quelques références bibliographiques plus détaillées. Les performances numériques exactes des algorithmes ne sont pas fournies, car elles correspondent rarement au même ensemble test et aux mêmes critères de mesure.

2.3.1 Analyse sinusoïdale

L'analyse sinusoïdale consiste à décomposer le mélange sous forme d'une somme de partiels sinusoïdaux groupés en notes par harmonicité. La combinaison de notes la plus vraisemblable est sélectionnée selon un ensemble de distributions *a priori* fixées sur les amplitudes et les fréquences des partiels [Dav02c, Ros02] et éventuellement sur le rythme [Cem03]. La maximisation de vraisemblance fait appel à des techniques d'inférence bayésienne de type Monte Carlo par chaînes de Markov [Dav02c].

Les algorithmes existants choisissent des *a priori* identiques pour toutes les sources. Ils ne sont pas applicables à la transcription ou la séparation d'enregistrements réels car ils identifient les notes jouées sans les attribuer à un instrument particulier. Seuls des instruments de tessitures différentes peuvent être ainsi distingués.

Si l'instrument associé à chaque note est connu, l'analyse sinusoïdale peut séparer les sources en resynthétisant séparément les partiels de chacune. Cette resynthèse ne pose de problème particulier, sauf aux fréquences où des partiels de plusieurs sources se recouvrent. Dans ce cas un partiel d'amplitude faible peut être masqué par un autre, ce qui empêche d'estimer son amplitude à partir du signal observé. Il est nécessaire d'introduire un *a priori* sur l'enveloppe spectrale et temporelle de ce partiel pour estimer son amplitude et éviter une distorsion de timbre perceptible dans le signal resynthétisé. Virtanen [Vir03a] utilise un *a priori* de continuité spectrale et temporelle qui fournit des résultats satisfaisants, mais n'est pas toujours applicable. Par exemple les partiels impairs de clarinette sont plus puissants que les partiels pairs, et les partiels de violon sont renforcés à certaines fréquences par les résonances étroites de la table d'harmonie. Il est aussi possible d'apprendre au préalable la forme des enveloppes sur des extraits solo. Meron et Hirose [Mer98] suppriment le piano d'un mélange piano et voix en utilisant la partition du piano et en apprenant la durée de l'attaque et la vitesse de décroissance de chaque partiel de chaque note de piano.

2.3.2 Analyse de scènes auditives computationnelle

L'analyse de scènes auditives computationnelle (*computational auditory scene analysis*) est aussi basée sur la modélisation des sources en partiels harmoniques, mais ne décompose pas explicitement le mélange. Des hypothèses de partiels sont créées en étudiant les pics du spectre à court terme du mélange [Kas95] ou l'information de périodicité dans son corrélogramme [Bro94, Ell96, Got99, Wu03b]. Ces hypothèses engendrent à leur tour des hypothèses de groupement simultané (notes, accords) et séquentiel (phrases musicales). La vraisemblance de chaque hypothèse est calculée en appliquant les règles de groupement auditif de l'ASA, implémentées sous forme de tableau noir (*blackboard*) [Ell96] ou de réseau bayésien [Kas95], et la meilleure hypothèse est choisie comme explication.

Pour grouper les notes en différentes sources, la plupart des algorithmes existants utilisent uniquement l'information fournie par les intervalles entre notes successives. Cela ne permet pas de distinguer des instruments de même tessiture, sauf dans le cas contraint où les mélodies sont disjointes.

Il est possible d'augmenter la ségrégation entre instruments en ajoutant aux règles de base de l'ASA une règle de similarité de timbre. Godsmark et Brown [God99] associent à un groupe de notes une courbe de timbre (*timbre track*) qui relie son amplitude totale à son centroïde spectral, et mesurent la similarité de timbre entre deux groupes de notes par la distance entre leurs courbes de timbre. Cette distance (combinée à d'autres règles) est ensuite utilisée pour déterminer la vraisemblance associée au regroupement de ces notes au sein d'une même source.

La performance de leur algorithme reste assez faible en raison de trois limitations. Premièrement la hauteur des notes est souvent mal identifiée. En particulier deux notes à intervalle harmonique sont

souvent considérées comme une note unique (de même hauteur que la plus grave). L'information de corrélation d'amplitude entre les partiels devrait éviter cette erreur en théorie mais ne le permet pas toujours en pratique, que les attaques des notes soient synchronisées ou non. Par exemple il peut être difficile de distinguer sur l'enveloppe spectrale entre un *crescendo* rapide et le rajout d'une note d'attaque lente à intervalle harmonique. Deuxièmement, lorsque plusieurs sources comportent des partiels à la même fréquence, il est difficile de savoir quelle source masque les autres. En particulier, même lorsque les hauteurs des notes sont connues, il arrive que des partiels correspondant à deux sources différentes soient groupés au sein d'une même note dont le timbre nouveau est attribué à une source tierce. Troisièmement, l'amplitude des partiels masqués est comptée comme nulle, ce qui donne des valeurs inexactes de l'amplitude totale et du centroïde spectral pour les notes partiellement masquées. Cela peut mener à des erreurs même lorsque la source masquante est connue.

Une solution aux deux premières limitations est d'apprendre des modèles de timbre pour un certain nombre d'instruments et de les utiliser lors du calcul de vraisemblance des hypothèses pour vérifier qu'un groupe de partiels correspond bien à une note plausible d'un instrument donné. Par exemple l'algorithme OPTIMA [Kas95] utilise des modèles de timbre appris sur une base de données de notes isolées de cinq instruments pour transcrire des accords de deux ou trois notes. Ces modèles contiennent à la fois des données brutes comme les enveloppes d'amplitude de tous les partiels, et des quantités liées à la perception auditive du timbre comme la durée de l'attaque et la quantité de *vibrato*. Les résultats montrent que l'intégration de modèles d'instruments appris engendre une diminution significative des erreurs d'estimation de hauteur des notes sur des accords aléatoires.

Kinoshita, Sakai et Tanaka [Kin99] modifient OPTIMA pour résoudre la troisième limitation. En cas de masquage entre partiels de plusieurs notes, si l'instrument jouant une des notes est identifié, il soustraient la puissance apprise de ce partiel pour cet instrument à la puissance observée avant d'identifier les instruments jouant les autres notes. Cette adaptation des observations améliore encore les résultats.

Enfin, notons que l'analyse de scènes auditives computationnelle n'est pas conçue pour séparer les sources d'un mélange, mais pour expliquer ce mélange du point de vue perceptif [Sch00]. Il est possible d'extraire une source donnée en filtrant le mélange par un masque binaire égal à un dans les zones temps-fréquence où cette source masque les autres et à zéro ailleurs. Mais cela fournit de mauvais résultats dès que des partiels de plusieurs sources sont présents dans une même zone temps-fréquence. Cette situation est très courante, surtout lorsque le mélange est analysé à l'aide de filtres auditifs à bande large. Le principe d"'allocation exclusive" [Bre90], qui stipule que l'oreille associe une source unique à chaque zone temps-fréquence, n'est pas transposable à la séparation.

2.3.3 Décomposition parcimonieuse temporelle

Le principe de la décomposition parcimonieuse temporelle est de d'exprimer la forme d'onde du mélange sous forme d'une somme pondérée d'atomes d'un dictionnaire $\mathcal{D}$ plus un résidu éventuel. Souvent le dictionnaire est constitué d'un dictionnaire plus petit $\mathcal{D}' = (\Phi_h)_{h\in\mathbb{N}}$ et de versions translatées $\mathcal{D}'_t = (\Phi_{ht})_{h\in\mathbb{N},0\leq t\leq T-1}$ où $\Phi_{ht}(u) = \Phi_h(u - tL)$. Dans le cas où le support des atomes de $\mathcal{D}'$ est de même taille que le pas de translation L, il est équivalent de découper la forme d'onde du mélange en trames disjointes de taille L et de décomposer chaque trame sur $\mathcal{D}'$. La décomposition la plus vraisemblable est choisie connaissant la distribution de probabilité (parcimonieuse) des coefficients de décomposition et celle du résidu, à l'aide d'un algorithme de type Basis Pursuit [Ben01], Monte Carlo par chaînes de Markov[Wol03] ou gradient de Newton [Jan03].

Les dictionnaires standard ne permettent pas d'interpréter directement les coefficients de décomposition en terme de transcription et de séparation. Par exemple, la décomposition de deux notes simultanées à intervalle d'octave dans un dictionnaire de Gabor fait intervenir des atomes en commun. En modélisant la dépendance entre les coefficients de décomposition sur des atomes voisins, il est possible d'extraire

des partiels ou des transitoires [Mol03, Wol03]. En classant les atomes par sous-espaces il est aussi possible d'extraire des groupes de partiels harmoniques [Gri99]. Mais la modélisation de structures de dépendance plus complexes n'a pas été étudiée. L'approche habituelle considère plutôt les coefficients de décomposition comme indépendants et modélise la structure du son instrumental au niveau des atomes en choisissant un dictionnaire adapté.

La transcription d'un mélange par décomposition parcimonieuse s'obtient en construisant le dictionnaire $\mathcal{D}'$ par union de plusieurs sous-dictionnaires $(\mathcal{D}'_{jh})$, contenant chacun quelques formes d'onde typiques d'une note de hauteur h d'un instrument j, extraites par exemple d'une base de données de notes isolées. Les coefficients de décomposition du mélange non nuls correspondent alors aux notes jouées par chaque instrument. Cette approche souffre du fait que la forme d'onde d'une note n'est pas invariante par translation et varie beaucoup selon la nuance de jeu, la facture de l'instrument et les conditions d'enregistrement. L'utilisation d'un gros dictionnaire contenant toutes les formes d'onde translatées est assez lourde. Les méthodes existantes ajoutent plutôt une procédure sous-optimale d'alignement des atomes sur chaque trame [Kas99, Bel02]. De plus les atomes sont modifiés en fonction du mélange.

Kashino [Kas99] modifie les atomes localement sur chaque trame par un filtrage. Son algorithme est testé sur des enregistrements de trio, connaissant *a priori* la hauteur des notes et les instants d'attaque, dont l'identification n'est pas étudiée. Les résultats de groupement des notes par instrument sont décevants, avec par exemple des confusions entre violon et piano. Cependant, le taux d'erreur est réduit de moitié en prenant en compte le caractère monophonique des instruments et des informations d'ordre musical.

L'algorithme de transcription de piano solo de Bello, Daudet et Sandler [Bel02], basé sur un principe voisin de projection orthogonale, obtient des résultats satisfaisants en apprenant les atomes sur tout l'enregistrement par une estimation de la fréquence fondamentale prédominante à chaque instant suivie d'un filtrage. Les atomes correspondant à des notes non apprises sont estimés en fonction de ceux des notes voisines présentes. L'algorithme n'est pas testé sur des instruments à oscillations soutenues, dont les formes d'onde sont généralement plus variables, et la méthode d'adaptation des atomes n'est pas transposable à un mélange.

La décomposition parcimonieuse temporelle peut être appliquée à la séparation dans le cas plus général où le dictionnaire $\mathcal{D}'$ est l'union de plusieurs sous-dictionnaires $(\mathcal{D}'_j)$ correspondant chacun à une source. Chaque source peut alors être synthétisée en mettant à zéro les coefficients de décomposition du mélange correspondant aux autres sources. Les algorithmes existants font appel à la notion de base adaptée à un signal donné, qui se définit comme la base qui maximise la parcimonie des coefficients de la décomposition sans résidu de ce signal. Ces algorithmes sont limités à nouveau par le problème de non invariance par translation, qui empêche de bien représenter les caractéristiques du signal avec un nombre réduit d'atomes. L'utilisation de dictionnaires redondants ou de procédures d'alignement de phase dans cette optique n'a pas été étudiée.

Dubnov [Dub02] construit une base adaptée au mélange, puis regroupe les atomes en plusieurs sous-espaces disjoints par étude des dépendances entre coefficients de décomposition. Testé sur un mélange de deux percussions et un mélange de batterie et de voix, son algorithme donne des résultats moyens. En particulier les zones contenant des mélanges de voix et de batterie sont attribuées entièrement à la batterie. Cela est dû au fait que la batterie n'est jamais présente seule et que la voix n'est pas assez redondante. Il n'est alors pas possible d'estimer la voix aux instants où elle est masquée à partir des autres instants.

Benaroya [Ben01] apprend préalablement une base adaptée à chaque source sur des extraits solo, puis décompose le mélange sur la réunion de ces bases. Jang, Lee et Oh [Jan03] utilisent le même principe mais apprennent aussi la distribution des coefficients de décomposition. La qualité de séparation obtenue est moyenne sur des mélanges synthétiques de voix d'homme et de femme et de musique jazz et rock, et

mauvaise sur des mélanges synthétiques de piano et de violon. Les bases adaptées contiennent des atomes caractéristiques pour une voix et un style de musique donné, alors qu'elles ressemblent aux mêmes bases de cosinus pour des instruments différents.

2.3.4 Décomposition parcimonieuse spectrale

La décomposition parcimonieuse spectrale consiste à remplacer la décomposition de la forme d'onde du mélange par celle de son spectre d'amplitude à court terme, qui a de meilleures propriétés d'invariance par translation. Cela suppose que le spectre d'amplitude d'une somme de signaux est proche de la somme de leurs spectres d'amplitude. Généralement le spectre est calculé sur une échelle de fréquence linéaire [Abd01] ou logarithmique [Cas00], puis chaque trame est décomposée sur le même dictionnaire de spectres $\mathcal{D}'$. La décomposition du spectre de log-puissance et du cepstre a aussi été étudiée pour d'autres applications [Ero03, Cas02], et celle de formes temps-fréquence de plusieurs trames successives a été mentionnée [Abd01].

La plupart des approches supposent les coefficients de décomposition indépendants et s'attachent à trouver un dictionnaire adapté aux signaux et à l'application voulue.

Les algorithmes de transcription existants construisent le sous-dictionnaire $\mathcal{D}'$ à partir du spectre à court terme du mélange. La taille du dictionnaire est fixée à la main et ses atomes sont calculés pour maximiser la parcimonie des coefficients de décomposition des trames. Puis les atomes sont regroupés en sous-dictionnaires disjoints $(\mathcal{D}'_{jh})$, censés correspondre à des notes distinctes, par étude de leurs caractéristiques [Uhl03] ou de celles des coefficients de décomposition [Cas00, Uhl03]. Les notes présentes sont détectées par la non nullité des coefficients de décomposition [Abd01], ou par l'augmentation brusque de ces coefficients lors des attaques [Vir03b, Fit03b]. Chaque dictionnaire générant un sous-espace vectoriel de l'espace engendré par le dictionnaire complet, cette méthode prend aussi le nom d'Analyse en Sous-espaces Indépendants (ASI).

L'algorithme de Casey et Westner [Cas00], aussi utilisé par Uhle, Dittmar et Sporer [Uhl03], calcule une base adaptée aux trames après réduction de dimension linéaire par Analyse en Composantes Principales (ACP). Les résultats sur un mélange de trois percussions particulier sont satisfaisants, mais la performance baisse dès que les sources ont des volumes différents [Vin01]. Fitzgerald, Cowle et Lawlor [Fit03a, Fit03b] améliorent les résultats de transcription de percussions en remplaçant l'ACP par une réduction de dimension non linéaire ou par une projection orthogonale sur un sous-espace *a priori* appris sur des notes isolées.

L'algorithme d'Abdallah [Abd01] calcule un dictionnaire adapté directement sur les données sans réduction de dimension et réalise une bonne transcription de clavecin solo. L'algorithme similaire de Klingseisen et Plumbley [Kli00] a cependant des difficultés à distinguer des notes de hauteur différente de flûte et de clarinette dans un mélange de quatre instruments à vent. Virtanen [Vir03b] ajoute une contrainte de positivité des atomes et des coefficients de décomposition, modélise la continuité temporelle des coefficients et pondère l'énergie du résidu par la réponse fréquentielle de l'oreille externe et moyenne. Il produit des résultats satisfaisants pour l'identification des notes de grosse caisse dans des enregistrements de musique de chambre.

L'application de ces algorithmes à la transcription d'enregistrements réels n'a pas été étudiée en détail. Tous ces algorithmes sont souvent testés sur des mélanges synthétisés à partir de fichiers MIDI, de sorte que le son d'une note donnée ne varie pas d'une instance à l'autre. De plus lorsque les instruments ne sont pas des percussions les atomes trouvés ne correspondent pas toujours à une note unique [Abd01] et le regroupement de notes par instruments n'est pas considéré.

Comme précédemment, les résultats de décomposition peuvent être appliqués à la séparation en construisant $\mathcal{D}'$ par union de plusieurs sous-dictionnaires $(\mathcal{D}'_j)$ décrivant des sources distinctes.

L'algorithme de Casey et Westner [Cas00], mentionné ci-dessus, synthétise chaque source séparément à partir de la phase du mélange et de son spectre d'amplitude, estimé en mettant à zéro les coefficients de décomposition du mélange correspondant aux autres sources. Cela donne un timbre très artificiel aux sources séparées, car les petites variations aléatoires de l'enveloppe spectrale ne sont pas préservées (elles sont incluses dans le résidu de décomposition).

Benaroya [Ben03] utilise les spectres d'amplitude estimés pour réaliser un filtrage pseudo-Wiener du mélange (ou filtrage de Wiener variant dans le temps). Il fournit des résultats satisfaisants sur un mélange de piano et de batterie, avec apprentissage préalable des sous-dictionnaires sur des extraits solo des mêmes instruments enregistrés dans les mêmes conditions.

2.3.5 Combinaison de modèles

La combinaison de modèles décrit le spectre à court terme du mélange par un MG ou un MMC factoriel. Son utilisation nécessite de connaître les réponses fréquentielles des filtres de mélange [RG03]. En pratique le seul cas considéré est celui où le mélange est égal à la somme des sources. Le spectre de puissance du mélange est alors modélisé par la somme des spectres de puissance des sources. La densité d'émission d'un état du mélange $(h_1, \ldots, h_n)$ est calculée en fonction des densités d'émission gaussiennes des états des sources $h_1, \ldots, h_n$. Si ces gaussiennes modélisent des spectres de puissance, la combinaison est égale à une gaussienne dont les paramètres sont trouvés par addition des moyennes et des variances [Ben03, Gha97]. Si elles modélisent des spectres de log-puissance, la combinaison est approchée par une gaussienne dont les paramètres sont calculés à l'aide de formules complexes [Gal95], ou par maximum point-à-point des moyennes et utilisation d'une variance fixée [Row00]. La combinaison de gaussiennes représentant des cepstres a aussi été étudiée pour la reconnaissance de la parole bruitée [Gal95]. Les états cachés les plus probables sur chaque trame sont calculés en réduisant la taille de l'espace de recherche par des techniques heuristiques diverses [Pon03, Row00, Rap02, Ben03]. La suite d'états la plus probable est calculée par algorithme de Viterbi.

Bien qu'il soit possible d'apprendre un MMC factoriel directement sur un mélange [Gha97], les algorithmes existants supposent tous que les MG ou les MMC représentant chaque instrument sont appris sur des extraits solo. Ces modèles décrivent donc bien les caractéristiques timbrales de chaque instrument. Par contre il est difficile de modéliser correctement les accords jouables par ce type de modèle, y compris ceux créés par les instruments monophoniques par ajout de la réverbération d'une note aux notes suivantes. L'algorithme de transcription de tabla de Gillet et Richard [Gil04] apprend de nombreuses combinaisons possibles de deux percussions et de leurs réverbérations. Mais lorsque le nombre de notes jouables est plus élevé tous les accords jouables ne sont pas présents dans les extraits d'apprentissage et cette approche n'est pas applicable.

Dans son algorithme de transcription de piano solo, Raphael [Rap02] propose de modéliser explicitement les accords comme des états et de partager les paramètres des densités d'émission des accords faisant intervenir les mêmes notes. Il obtient de bons résultats après apprentissage sur des extraits enregistrés dans les mêmes conditions et étiquetés par leur partition.

Eggink et Brown [Egg03] identifient l'instrument associé à chaque note d'un accord séparément, sans combinaison de modèles. Ils supposent que la hauteur, la durée et les instants d'attaque des notes sont trouvés auparavant par un algorithme d'ASA computationnelle sans *a priori*. L'amplitude observée dans les zones temps-fréquence où des partiels sont masqués par d'autres partiels ou par le bruit de fond n'est pas prise en compte, ou bien est considérée comme une borne supérieure à l'amplitude des partiels [Jos99]. Cette approche donne de bons résultats d'association note-instrument sur des accords de deux notes. Cependant la performance est limitée par celle de l'algorithme d'ASA computationnelle et par l'utilisation insuffisante de l'information présente dans les zones de recouvrement de sources (par exemple en cas de notes à intervalle harmonique).

Roweis [Row00], repris par Pontoppidan et Dyrholm [Pon03], applique la combinaison de modèles à la séparation de sources à l'aide d'un masquage binaire spectral. Le spectre de log-puissance de chaque source sur chaque trame est estimé par la moyenne de la densité d'émission de son état le plus probable. Une source donnée est extraite en mettant à zéro les zones temps-fréquence du mélange où sa log-puissance est inférieure à celle des autres sources. Les résultats semblent satisfaisants sur des mélanges de deux sources de parole (homme et femme).

Dans le cas de sources de musique, qui sont moins disjointes dans le plan temps-fréquence, Benaroya [Ben03] utilise un filtrage pseudo-Wiener calculé à partir des moyennes des densités d'émission des états des sources. Tous les états d'une source à un instant donné sont pris en compte (pas seulement le plus probable) et pondérés par leur probabilité. Sur un mélange de piano et batterie, il note une forte amélioration de performance en utilisant des MG à facteurs d'échelle au lieu des MG standard. La performance reste cependant un peu inférieure à celle obtenue avec une décomposition parcimonieuse spectrale pour le même nombre d'états et d'atomes.

2.4 Méthodes pour les enregistrements multicanal

Nous abordons maintenant l'état de l'art de la transcription et de la séparation d'enregistrements multicanal. Nous séparons les algorithmes en cinq catégories présentes dans la littérature. Comme pour les enregistrements monocanal, nous les décrivons par les types de mélanges auxquels ils s'appliquent et un bref commentaire de leurs résultats.

2.4.1 Analyse en composantes indépendantes

L'Analyse en Composantes Indépendantes (ACI) s'applique aux mélanges (sur-)déterminés peu bruités invariants dans le temps. Son principe est d'estimer des filtres de démixage invariants dans le temps qui maximisent la vraisemblance des sources estimées. L'ACI généralise les méthodes de formation de voies, qui considèrent des filtres de démixage uniquement formés d'un gain et d'un délai. La maximisation de vraisemblance repose sur une méthode de Jacobi [Pha01], de gradient de Newton [Hyv01a] ou de gradient relatif [Ama97], les sources étant extraites successivement [Hyv01a] ou conjointement [Car98a].

L'ACI est conçue pour la séparation. Elle peut éventuellement être appliquée à la transcription en utilisant des algorithmes de transcription de solos sur chaque source estimée. Le lien entre performance de séparation et performance de transcription dans ce cas n'a pas été étudié.

La majorité des algorithmes pour les mélanges instantanés travaillent directement sur la forme d'onde du mélange. Ils modélisent les formes d'onde des sources comme des variables parcimonieuses indépendantes, dont les distributions ne sont pas fixées mais réestimées itérativement. La maximisation de vraisemblance des sources estimées est alors équivalente à la minimisation de leur information mutuelle [Car98a]. Les algorithmes JADE [Car98a], JD-BGL [Pha01] et FastICA [Hyv01a] estiment les sources à permutation et gain près de façon presque parfaite. Des algorithmes semblables donnent de bons résultats sur les mélanges convolutifs courts [Ama97], mais la performance se dégrade lorsque la taille des filtres de démixage dépasse une centaine d'échantillons [Ser03].

Les algorithmes pour les mélanges convolutifs longs décomposent le mélange en sous-bandes fréquentielles. Ils estiment les images spatiales des sources indépendamment dans chaque sous-bande à permutation près, en supposant qu'elles sont parcimonieuses et indépendantes. Le nombre de paramètres à estimer conjointement est diminué, mais les bonnes permutations doivent être trouvées pour regrouper les sous-bandes par source. Davies [Dav02a] modélise les sous-bandes d'une source par des variables laplaciennes de même amplitude et estime conjointement les permutations qui maximisent la vraisem-

blance de ce modèle. Parra et Alvino [Par02] posent des contraintes sur la réponse fréquentielle des filtres de démixage en fonction des azimuts des sources. D'autres algorithmes calculent les permutations successivement par des méthodes heuristiques basées sur la structure fréquentielle des sources [Ane00, Mur01, Saw03]. Les algorithmes les plus performants, comme celui de Servière [Ser03], utilisent des filtres de démixage de quelques milliers d'échantillons avec des résultats satisfaisants. Notons que tous ces algorithmes estiment les images spatiales des sources. L'estimation des sources elles-mêmes à gain près semble irréalisable en pratique.

Deux algorithmes étudient des modèles de sources plus élaborés appris sur des extraits solo. Mitianoudis et Davies [Mit02] séparent des mélanges instantanés déterminés de deux sources en modélisant le cepstre de chaque instrument par un MG. Le critère usuel de vraisemblance présente des maxima locaux pour certaines valeurs de mélange des sources. De meilleurs résultats sont obtenus en maximisant la différence entre la vraisemblance d'une source voulue et la vraisemblance de l'autre source (*cohort normalisation*). Reyes-Gomez, Raj et Ellis [RG03] séparent des mélanges convolutifs longs sur-déterminés de deux sources de parole en modélisant le spectre de log-puissance de chaque locuteur par un MMC. L'algorithme réestime alternativement les filtres de démixage, le spectre des sources et leur transcription en suites d'états. Pour éviter une mauvaise transcription d'une source estimée due à la présence d'interférences, toutes les sources sont transcrites simultanément sur chaque source estimée par combinaison de modèles (les réponses fréquentielles des filtres de mélange étant réapprises à chaque itération). Connaissant les phrases prononcées, cet algorithme semble plus performant que les algorithmes classiques.

Quelques études évaluent la performance de l'ACI sur des mélanges convolutifs longs indépendamment d'un algorithme particulier en calculant les filtres de démixage optimaux. Westner et Bove [Wes99] montrent que la qualité de séparation augmente en ajoutant des canaux au mélange. Balan, Rosca et Rickard [Bal01] prouvent que des filtres de démixage longs sont plus performants mais beaucoup moins robustes en cas de petits mouvements des sources. Le démixage par des filtres invariants dans le temps est donc inapplicable à des enregistrements réels.

Parra et Spence [Par00] séparent des enregistrements réels de deux sources de parole en calculant des filtres de démixage sur des trames de mélange. La performance est maximale lorsque le pas entre trames est minimal, mais elle reste limitée.

2.4.2 Sélection de zones temps-fréquence

Toujours dans le cas de mélanges déterminés, une façon d'améliorer l'ACI est de découper le mélange en sous-bandes et en trames (x_i^{tf}) et de prendre en compte seulement les points temps-fréquence (t, f) où une source unique est présente. Sur ces points la modélisation des formes d'onde des sources par des variables parcimonieuses indépendantes est valide, même si elles ne sont pas indépendantes globalement.

Dans le cas de mélanges stéréo instantanés, Abrard, Deville et White [Abr01] sélectionnent ces points par étude de la cohérence inter-canal. Ils utilisent ensuite la propriété suivante : si seule la source j est présente en un point temps-fréquence, alors l'azimut observé en ce point est égal à l'azimut de la source (lié au gain relatif de mélange $a_{2,j}/a_{1,j}$). Cette propriété permet de calculer les azimuts de toutes les sources à partir des différences de volume inter-canal. Cela fournit la matrice de mélange (à gain près sur chaque colonne) qui est inversée pour retrouver les sources.

Albouy et Deville [Alb03] étendent ces résultats aux mélanges stéréo convolutifs courts. Lorsque le spectre est une transformée de Fourier à court terme avec un nombre de sous-bandes supérieur à la taille des filtres, les différences de volume et de phase inter-canal permettent cette fois de retrouver les rapports $a_{2,j}(f)/a_{1,j}(f)$, où $a_{1,j}(f)$ et $a_{2,j}(f)$ sont les transformées de Fourier des filtres $a_{1,j}(\tau)$ et $a_{2,j}(\tau)$. Le problème de permutation entre sous-bandes est évité en considérant uniquement le cas de sources de parole qui présentent des silences sur toutes les sous-bandes simultanément.

Dans le cas de mélanges non stéréo, Févotte [Fév04b] utilise une méthode plus complexe de sélection de points temps-fréquence et applique une ACI unique aux signaux (x_i^{tf}) extraits des points sélectionnés.

Ces algorithmes sont généralement plus performants que l'ACI.

2.4.3 Maximisation de la parcimonie

Dans le cas de mélanges sous-déterminés, le mélange n'est plus séparable par inversion des filtres de mélange. Il faut estimer conjointement les filtres de mélange et les sources qui permettent de reconstruire le mélange, éventuellement avec une faible erreur résiduelle. Si les formes d'onde des sources sont modélisées par des décompositions parcimonieuses dans une base donnée, l'estimation consiste à réaliser un compromis entre la quantité d'erreur résiduelle et la parcimonie des coefficients de décomposition supposés indépendants. Les calculs sont effectués par des méthodes de type gradient naturel et gradient de Newton [Lee99], apprentissage bayésien variationnel [Att99, Mis01] ou Monte Carlo par chaînes de Markov [Fév04b].

Comme l'ACI, cette méthode s'applique avant tout à la séparation. La transcription des sources estimées par des algorithmes de transcription de musique soliste n'a pas été étudiée mais fournit probablement des résultats moyens. Lorsque des partiels de plusieurs sources se recouvrent, les sources estimées contiennent souvent des interférences et des artefacts qui modifient leur timbre et peuvent mener à des erreurs d'insertion ou de suppression de notes.

Les algorithmes existants ne considèrent que le cas de mélanges instantanés. Lee *et al.* [Lee99] et Theis, Puntonet et Lang [The03] séparent des mélanges de parole en modélisant la parcimonie dans la base de Dirac, c'est-à-dire la parcimonie des formes d'onde. Davies [Dav02b] et Févotte [Fév04a] utilisent la base de Transformée Cosinus Discrète Modifiée (TCDM) pour séparer des sources musicales.

Dans le cas particulier de mélanges stéréo, certains algorithmes proposent d'effectuer la séparation en deux étapes : estimation de la matrice de mélange, puis estimation des sources sachant la matrice de mélange [Zib01, The03, Gri03a]. La première étape ne repose pas sur le modèle probabiliste des sources. Elle consiste à décomposer le mélange dans la base où les sources sont parcimonieuses, puis à définir un azimut observé sur chaque atome (de façon similaire à la définition de l'azimut observé en un point temps-fréquence). Les azimuts des sources sont alors calculés en étudiant les pics [Zib01] ou les "champs réceptifs" [The03] de l'histogramme des azimuts, ce qui fournit la matrice de mélange (à gain près sur chaque colonne). La deuxième étape reste basée sur le modèle probabiliste, mais dans certains cas elle peut aussi se décrire à partir des informations spatiales. Par exemple, lorsque les coefficients de décomposition des sources sont supposés laplaciens de même variance, les coefficients estimés sur un atome donné sont non nuls pour deux sources et nuls pour les autres. Ces deux sources sont celles dont les azimuts sont les plus proches de l'azimut observé sur cet atome [The03].

La performance de ces algorithmes dépend du degré de parcimonie effectif des sources. Prenons l'exemple d'un mélange stéréo de sources de distribution laplacienne de même variance dans une base d'atomes temps-fréquence. Deux types d'erreur existent lorsque les sources ne sont pas parfaitement disjointes. Premièrement les sources sont mal séparées aux points temps-fréquence contenant trois sources ou plus car seules deux sources en sont extraites. Deuxièmement, la séparation est mauvaise aussi aux points contenant deux sources "périphériques" (provenant de la gauche et de la droite) car ils sont attribués aux sources "centrales" (provenant du milieu) dont l'azimut est plus proche de l'azimut observé [Vin04d]. D'autres distributions parcimonieuses diminuent ces problèmes sans les supprimer [Vie01]. En pratique la parcimonie limitée des sources musicales en temps-fréquence se manifeste par des artefacts sur les sources estimées [Fév04a] et par une dégradation de performance lorsque le nombre de sources augmente [Zib01].

Plusieurs approches sont possibles pour améliorer les résultats. Zibulevsky *et al.* [Zib01] augmentent la parcimonie des sources en utilisant un dictionnaire redondant au lieu d'une base. Vielva, Erdoğmuş et

Príncipe [Vie01] modélisent chaque source par une distribution de variance appropriée connue *a priori*.

2.4.4 Masquage temps-fréquence

Certains algorithmes, dits de masquage temps-fréquence, adaptent la méthode de séparation de mélanges stéréo en deux étapes décrite ci-dessus en remplaçant l'estimation probabiliste des sources par un filtrage plus heuristique. D'abord les azimuts observés en différents points temps-fréquence sont mis en commun au sein d'un histogramme, dont les pics fournissent les azimuts des sources [Vis03]. Puis l'image spatiale de chaque source est extraite en filtrant les canaux du mélange par un masque temps-fréquence, dont la valeur en un point est un gain fonction de l'azimut observé. Souvent un masque binaire est choisi : chaque point est attribué avec un gain unitaire à la source dont l'azimut est le plus proche de l'azimut observé et avec un gain nul aux autres sources [Yil02]. L'avantage de cette méthode tient à son application plus facile aux mélanges stéréo convolutifs. Mais en pratique l'azimut est difficile à calculer dans de tels mélanges. Certains algorithmes utilisent donc plusieurs quantités à la fois pour décrire les propriétés spatiales.

Avendano et Jot [Ave02] séparent et remixent sur cinq canaux des enregistrements stéréo synthétiques réalisés par mélange panoramique et ajout de réverbération artificielle. Les points temps-fréquence contenant de la réverbération sont mis de côté par étude de la cohérence inter-canal, et les azimuts sont calculés sur les points restants à partir de la différence de volume inter-canal.

Yilmaz, Radke et Rickard [Yil02, Rad02] étendent cet algorithme au cas où chaque filtre de mélange est formé d'un gain et d'un délai. Le caractère synthétique du mélange mène à une description spatiale bidimensionnelle formée de la différence de volume inter-canal et de l'azimut (fonction de la différence de phase inter-canal). Les délais de mélange sont supposés suffisamment courts pour calculer l'azimut observé de façon unique en haute fréquence.

Roman, Wang et Brown [Rom03] étudient des mélanges convolutifs réalisés à partir de réponses impulsionnelles binaurales d'une centaine d'échantillons enregistrées dans une chambre anéchoïque. Ces filtres correspondent à des délais de mélange plus longs et à des différences de volume inter-canal variables selon la fréquence. L'azimut des sources est calculé uniquement en fonction des différences de phase inter-canal, en prenant en compte tous les azimuts observés possibles en chaque point en haute fréquence. Ensuite les points temps-fréquence sont séparés en plusieurs groupes dans chaque canal en fonction des différences de volume et de phase inter-canal observées. Chaque groupe est alors attribué à la source dont l'azimut explique le mieux la différence de phase observée. Sur le même type de mélanges, Viste et Evangelista [Vis03] proposent d'exploiter la dépendance des informations spatiales pour définir un azimut unique en chaque point en haute fréquence. La différence de volume inter-canal théorique à chaque fréquence et pour chaque azimut est calculée sur une base de données de réponses impulsionnelles et utilisée pour choisir le bon azimut en chaque point parmi tous les azimuts compatibles avec la différence de phase inter-canal observée.

La performance de ces algorithmes est semblable à celle des algorithmes de maximisation de la parcimonie. En particulier les résultats sont satisfaisants sur des mélanges de parole [Yil02, Rom03] mais plus moyens sur des mélanges musicaux [Vis02]. Le déséquilibre en faveur des sources "centrales" subsiste [Rom03, Vis03, Vis02], et d'autres types d'erreurs apparaissent, parmi lesquels les "zéros abusifs" engendrés par le masquage binaire [Ave02] et la mauvaise estimation de phase due à la prise en compte de chaque canal du mélange séparément lors du filtrage [Gri03a].

À nouveau plusieurs approches ont été proposées pour améliorer les résultats. Viste et Evangelista [Vis02] détectent les points temps-fréquence contenant des partiels de plusieurs notes mal séparés. Ils recalculent les deux sources prépondérantes localement par un algorithme de type ACI minimisant l'erreur entre les enveloppes d'amplitude des partiels estimés et les enveloppes théoriques calculées en fonction des autres partiels des notes correspondantes. Gribonval [Gri03a] remplace le spectre à court terme classique par une décomposition parcimonieuse sur des atomes stéréo et teste la performance obtenue avec

plusieurs dictionnaires redondants sur des mélanges synthétiques instantanés.

2.4.5 Analyse de scènes auditives computationnelle

Une dernière méthode consiste à étendre l'analyse de scènes auditives computationnelle aux mélanges multicanal en ajoutant une règle de proximité spatiale aux règles de l'ASA monocanal. Cette règle peut être utilisée au sein d'une structure de tableau noir à la fois pour le groupement simultané (notes, accords) et le groupement séquentiel (phrases musicales). La manière la plus simple de définir la proximité spatiale est de calculer une distance entre azimuts observés. Cela suppose qu'un azimut unique puisse être défini en tout point temps-fréquence, par exemple en calculant tous les azimuts possibles en fonction de la différence de phase inter-canal et en sélectionnant le bon selon la différence de volume inter-canal.

Nakatani [Nak02] construit un système de détection de la fréquence fondamentale de chaque locuteur dans des mélanges binauraux de parole réalisés par convolution avec des filtres courts enregistrés en chambre anéchoïque. Il constate une nette amélioration de performance en ajoutant la règle de proximité spatiale aux règles monocanal de continuité et de proximité de fréquence fondamentale.

Sakuraba et Okuno [Sak03] étudient la transcription de musique de chambre stéréo. Les mélanges testés sont réalisés par enregistrement avec un micro AB large de quatre sources mono générées à partir de fichiers MIDI et diffusées chacune par un haut-parleur dans une chambre anéchoïque. En plus de la proximité spatiale, une règle de similarité de timbre est utilisée pour valider les hypothèses de groupement séquentiel (mais pas pour celles de groupement simultané). Les auteurs prouvent que l'ajout d'une de ces deux règles aux règles de l'ASA monocanal améliore à la fois les résultats de groupement simultané et de groupement séquentiel, et que l'ajout des deux règles à la fois améliore encore plus les résultats.

Notons que cette méthode permet aussi de séparer les sources par masquage binaire du mélange ou par resynthèse sinusoïdale [Nak02]. Les résultats sont alors comparables à ceux énoncés précédemment.

2.5 Résumé des méthodes et objectifs

Résumons maintenant les buts de ce travail au vu des tâches à accomplir et des résultats des algorithmes existants.

2.5.1 Utilisation de modèles d'instruments

Notre première constatation sur l'état de l'art est que la performance dépend de la quantité et de la structure de l'information contenue dans les modèles de sources. Le tableau 2.1, qui présente cinq catégories de modèles de sources, appelle en effet deux remarques générales. Premièrement, les modèles non spécifiques (identiques pour tous les instruments) semblent moins performants que les modèles de groupement par similarité, eux-mêmes moins performants que les modèles spécifiques appris. Deuxièmement, les modèles représentant le son instrumental globalement sans structuration intermédiaire par notes semblent limités aux tâches de séparation.

Nous proposons donc comme but essentiel de ce travail l'étude de l'apport de modèles de sources spécifiques appris et structurés, que nous appellerons modèles d'instruments, pour la transcription et la séparation. Nous ne nous limiterons pas à l'utilisation des modèles pour le regroupement simultané et séquentiel des notes par instruments. Nous explorerons d'autres utilisations possibles, par exemple pour l'estimation des hauteurs des notes et des azimuts des sources.

Des études semblables ont déjà été réalisées sur les mélanges monocanal, dont certaines contemporaines de ce travail. Les méthodes d'ASA computationnelle et de combinaison de modèles se sont

Structure	Type de modèle	Parag.	Performance
Modèle non spécifique	Dictionnaire fixé (Gabor, Dirac) ou modèle harmonique avec corrélation des amplitudes à diverses fréquences.	2.3.1 2.4.1 2.4.2 2.4.3 2.4.4	Inapplicable aux mélanges monocanal. Qualité limitée sur les mélanges sous-déterminés ou convolutifs longs.
Groupement par similarité	Dictionnaire adapté extrait du mélange et groupement des atomes par instruments.	2.3.3 2.3.4	Applicable uniquement aux mélanges assez redondants (avec notes répétées). Transcription satisfaisante de mélanges monocanal de percussions mais limitée en général (atomes représentant du bruit ou des accords).
Apprentissage non supervisé	Dictionnaire adapté ou MG appris sur des extraits solo sans partition.	2.3.3 2.3.4 2.3.5 2.4.1	Séparation satisfaisante de mélanges monocanal (sauf 2.3.3). Bonne association note-instrument mais nécessité de connaître la hauteur des notes préalablement. Difficulté à modéliser les accords et les notes absents de l'ensemble d'apprentissage.
Structuration par notes et groupement par similarité	Modèle harmonique et similarité de timbre.	2.3.2 2.4.5	Bonne transcription de mélanges monocanal et multicanal sous-déterminés. Risque de créer des sources avec un timbre ne correspondant à aucun instrument existant.
Structuration par notes et apprentissage supervisé	Dictionnaire adapté ou MG appris sur des extraits solo avec partition ou des notes isolées. Modèle harmonique avec caractéristiques timbrales apprises.	2.3.1 2.3.2 2.3.3 2.3.4 2.3.5 2.4.1	Bonne transcription de mélanges monocanal (sauf 2.3.3) avec diminution des erreurs sur la hauteur des notes.

TAB. 2.1 – Classement des méthodes de transcription et séparation selon les modèles de sources utilisés.

révélées performantes. Mais les résultats ont souvent été obtenus sur des mélanges particuliers (musique soliste, mélanges d'instruments MIDI, de percussions ou de parole) et parfois avec des connaissances *a priori* importantes (hauteur des notes, règles de composition musicale). Nous chercherons donc à confirmer et étendre ces résultats en construisant un nouveau modèle de sources inspiré de l'ASI et en effectuant quelques tests sur des enregistrements musicaux difficiles avec des connaissances *a priori* raisonnables. Nous étudierons en particulier des extraits de CD et des mélanges synthétiques d'extraits de CD contenant des instruments de même tessiture ou des notes à intervalle harmonique.

Notre modèle de sources se concentrera sur la modélisation du timbre des instruments et ne prendra pas en compte d'informations de type musical, comme les intervalles entre notes successives ou le rythme. Au mieux il ne pourra donc transcrire et extraire correctement que les notes partiellement masquées à chaque instant. Les notes dont l'attaque ou la relâche est entièrement masquée seront transcrites et extraites mais avec un début ou une fin imprécis. Les notes complètement masquées seront considérées comme absentes.

Structure	Type d'analyse	Parag.	Performance
Monocanal	Décomposition en sous-bandes et en trames ou sur un dictionnaire adapté.	2.3.1 2.3.2 2.3.3 2.3.4 2.3.5	Bonne transcription et séparation avec des modèles de sources appris et structurés par notes (sauf 2.3.3). Qualité limitée avec regroupement par similarité et autres modèles de sources.
Sous-déterminé convolutif court	Décomposition en sous-bandes et en trames.	2.4.4 2.4.5	Bonne transcription (surtout avec regroupement par similarité). Séparation limitée. Utilisation des propriétés binaurales.
Sous-déterminé instantané	Décomposition en sous-bandes et en trames ou sur un dictionnaire adapté.	2.4.3 2.4.4	Séparation limitée.
Sur-déterminé convolutif long	Décomposition en sous-bandes.	2.4.1	Séparation limitée. Utilisation des propriétés des micros AB pour les mélanges stéréo.
Sur-déterminé convolutif court ou instantané		2.4.1 2.4.2	Séparation parfaite.

TAB. 2.2 – Classement des méthodes de transcription et séparation selon les types de mélange abordés.

2.5.2 Utilisation de mélanges multicanal calibrés

Notre deuxième constatation sur l'état de l'art est que la performance dépend aussi du nombre de canaux et des contraintes sur le mélange. Le tableau 2.2, qui classe les résultats des méthodes existantes selon cinq types de mélanges, mène à nouveau à deux remarques générales. Premièrement, les mélanges contenant moins de capteurs semblent donner une performance moindre ou nécessiter des modèles de sources plus complexes à performance égale. Deuxièmement, les propriétés des micros utilisés apportent une information très utile dans le cas de mélanges stéréo convolutifs.

Nous proposons donc comme autre but de ce travail la mesure des changements de performance atteignables entre différents mélanges des mêmes sources.

Cette question a été abordée très récemment pour la transcription, mais uniquement dans le cadre de l'ASA computationnelle pour la comparaison de mélanges monocanal et convolutifs courts binauraux de certaines sources (musique générée à partir de MIDI, parole). Et elle a été laissée de côté pour la séparation. Nous tâcherons de l'étudier sur des mélanges synthétiques d'extraits de CD sous-déterminés, convolutifs longs ou multipistes type "ingénieur du son".

Nous n'étudierons pas les mélanges variant dans le temps, à cause de la difficulté à enregistrer des réponses impulsionnelles de salle correspondant à des sources en mouvement. Cependant notre approche de séparation ne sera pas basée sur l'estimation de filtres de démixage, et sera donc en théorie plus facilement applicable aux mélanges variant dans le temps. Nous n'étudierons pas non plus les mélanges comprenant du bruit de fond non stationnaire, par exemple des enregistrements de concerts avec des raclements de gorge et des sonneries de téléphones portables.

Chapitre 3

Cadre probabiliste de construction et d'utilisation de modèles d'instruments

Ce troisième chapitre décrit un cadre assez général de construction et d'utilisation de modèles probabilistes d'instruments. Nous commençons par proposer dans le paragraphe 3.1 une famille de modèles d'instruments basés sur un modèle génératif du spectre de puissance à court terme. Dans le paragraphe 3.2, nous combinons ces modèles d'instruments pour former des modèles de mélanges. Nous exprimons alors la distribution d'un mélange grâce à la loi de Bayes dans le paragraphe 3.3. Nous expliquons dans le paragraphe 3.4 comment utiliser ces modèles pour résoudre les quatre tâches de transcription et séparation considérées. Nous montrons enfin dans le paragraphe 3.5 comment apprendre les paramètres des modèles d'instruments sur des extraits solo et des notes isolées.

La plupart des résultats de ce chapitre sont nouveaux, mais s'inspirent de modèles génératifs existants et des méthodes d'inférence et d'apprentissage correspondantes, en particulier [Abd01, Row00, Pen00, Gal95]. Certains travaux contemporains abordent des questions similaires [Ben03, Gri03b, Vir03b]. Seules les grandes lignes du chapitre ont été présentées dans nos articles [Vin04d, Vin04c, Vin04b], avec quelques différences mineures.

3.1 Modèles probabilistes d'instruments

3.1.1 Modèle génératif à trois couches

Sur un enregistrement solo, les variables observées (forme d'onde, cepstre, *etc*) dépendent à la fois du timbre de l'instrument, des notes jouées et de l'interprétation de l'instrumentiste. Construire un modèle probabiliste d'un instrument consiste à modéliser ces variables par une distribution de probabilité générale capable de représenter toutes les notes et les interprétations possibles. La distribution doit alors décrire les diverses structures du son propres à l'instrument (évolution au sein d'une note, accords et phrases musicales jouables, styles de jeu) et à son répertoire (*tempi* usuels).

Souvent ces structures ne sont pas modélisées directement sur les variables observées mais sur une série de variables cachées discrètes (notes, accords, phrases) et continues (hauteur, inharmonicité, volume, centroïde spectral, *etc*). La distribution de probabilité est alors décomposée en deux "couches" : une distribution sur les variables cachées et une distribution conditionnelle des variables observées sachant les variables cachées. Un tel modèle est appelé modèle génératif à deux couches. La distribution sur les variables cachées peut elle-même se décomposer à l'aide d'une troisième couche, ce qui fournit un modèle génératif à trois couches, et ainsi de suite. Les modèles de sources décrits dans le paragraphe 2.1.2 sont des modèles génératifs à deux couches, où les variables cachées sont les paramètres sinusoïdaux, les coefficients de décomposition parcimonieuse ou les états du MG.

La construction de modèles d'instruments applicables à la transcription et la séparation nécessite quelques contraintes. Comme nous l'avons souligné dans le paragraphe 2.5.1, les variables cachées doivent reproduire la structure par notes du son instrumental pour que le modèle représente bien les accords et que la présence ou l'absence d'une note à un instant donné soit bien définie. Les variables observées doivent contenir toutes les informations nécessaires pour identifier des notes et des instruments.

Nous proposons d'utiliser une famille de modèles génératifs à trois couches, que nous appelons modèles d'instruments. Les trois couches sont définies de la façon suivante : la couche spectrale contient le spectre de puissance à court terme du son instrumental modélisé, la couche d'état contient l'état binaire de chaque note de l'échelle des demi-tons à chaque instant (présence ou absence) et la couche descriptive contient les descripteurs continus correspondants (volume, hauteur, timbre). Dans la suite du paragraphe, nous discutons les raisons du choix de cette famille de modèles et les différentes structures modélisées.

3.1.2 Modélisation du spectre de puissance à court terme

L'état de l'art montre que les modèles de sources représentant le spectre de puissance sont effectivement applicables à la transcription et la séparation.

En effet il est possible d'évaluer la vraisemblance du spectre d'une source sachant les notes jouées et la vraisemblance du mélange sachant les spectres des sources. La hauteur d'une note harmonique est liée aux fréquences des pics du spectre [Kas95, Abd01] et son timbre à leurs amplitudes [Kin99, Egg03]. Le spectre d'un son percussif est aussi lié à l'instrument correspondant [Fit03a]. Les spectres des sources permettent d'approcher le spectre d'un mélange monocanal (voir paragraphes 2.3.4 et 2.3.5), ainsi que la différence de volume et de phase inter-canal [Rom03].

À partir de ces deux vraisemblances, il est possible d'évaluer la vraisemblance d'une hypothèse sur les notes jouées ou sur les spectres des sources sachant le mélange. Les meilleures hypothèses de notes fournissent une transcription du mélange. Les hypothèses de spectres des sources combinées à un filtrage [Ben03] ou à une synthèse sinusoïdale [Vir03a, Mer98] permettent la séparation du mélange.

En particulier, des modèles d'instruments représentant le spectre de puissance aident à résoudre les problèmes classiques rencontrés lorsque plusieurs sources se chevauchent en un point temps-fréquence. Dans un mélange monocanal ils permettent d'attribuer l'amplitude observée à une source et de déterminer les amplitudes des sources masquées [Kin99]. Dans un mélange stéréo panoramique, ils permettent d'effectuer la distinction entre des sources "centrales" et des sources "périphériques" donnant la même différence de volume inter-canal [Vin04d].

Les modèles de sources utilisant d'autres variables observées sont plus limités ou contiennent des informations inutiles.

Les modèles représentant le cepstre ne sont pas applicables à l'identification de notes même en musique soliste et les modèles harmoniques ne sont pas applicables aux mélanges de percussions.

La représentation des différences de phase entre partiels harmoniques est inutile. Ces différences de phase peuvent aider à transcrire des enregistrements de piano lorsqu'elles sont apprises directement sur le mélange [Bel02]. Mais elles changent de façon peu prévisible selon les conditions d'enregistrement et les mouvements des sources. L'utilisation de descripteurs du timbre complexes semble aussi superflue dans l'état de l'art actuel. Les méthodes d'identification d'instruments basées uniquement sur les puissances instantanées des partiels donnent des résultats comparables aux autres [Mar99a, Bro01, Egg03].

3.1.3 Choix des structures modélisées

La modélisation des structures du son instrumental à l'aide de la famille de modèles génératifs proposée consiste à définir la distribution de probabilité conjointe des variables de toutes les couches. Nous

restreignons les structures modélisables en fixant des relations d'indépendance conditionnelle entre variables.

Dans la suite, nous appelons j l'indice de l'instrument modélisé ($1 \leq j \leq n$), h l'indice de hauteur de note sur l'échelle MIDI pour cet instrument ($H_j \leq h \leq H'_j$) et t l'indice de trame temporelle à court terme ($0 \leq t \leq T-1$). Nous notons $E_{jht} \in \{0,1\}$ l'état de la note h au temps t, $\mathbf{p}_{jht} \in \mathbb{R}^{K_j+1}$ le vecteur de descripteurs continus correspondants et $\mathbf{m}_{jt}$ le spectre de puissance de l'instrument au temps t. Nous définissons aussi $\mathbf{E}_{jt} = (E_{jht})_{H_j \leq h \leq H'_j}$ l'état de l'instrument au temps t et $\mathbf{p}_{jt} = (\mathbf{p}_{jht})_{H_j \leq h \leq H'_j}$ les descripteurs correspondants.

La figure 3.1 représente les dépendances entre ces variables sous forme de réseau bayésien dynamique [Gha01]. Dans ce schéma, chaque noeud correspond à une variable et les flèches entre noeuds permettent de définir les fils et les descendants, et les parents et les ancêtres d'une variable. Toute variable est alors indépendante de ses non-descendants sachant ses parents. Les dépendances respectent le principe de causalité : les flèches vont du passé vers le futur et de la couche haut niveau vers la couche bas niveau.

L'état de l'instrument $\mathbf{E}_{jt}$ dépend uniquement des états aux instants précédents $(\mathbf{E}_{jt'})_{0 \leq t' \leq t-1}$. Cela permet de modéliser la durée des notes [Ost96] et la dépendance entre les états des notes. Par exemple il est possible d'imposer une durée minimale pour chaque note, mais aussi une durée minimale entre deux notes successives sur un instrument monophonique ou la synchronie des notes d'un accord sur un instrument polyphonique.

Le vecteur de descripteurs d'une note $\mathbf{p}_{jht}$ dépend uniquement de son état à cet instant E_{jht}. Cela décrit le fait que le volume des notes absentes est nul et que les descripteurs des notes présentes sont indépendants. En effet les descripteurs d'une note et de la réverbération des notes précédentes et ceux des notes d'un accord d'un instrument à oscillations libres sont indépendants. Dans un souci de lisibilité, l'indépendance conditionnelle entre $\mathbf{p}_{jht}$ et les états et les descripteurs des autres notes $(E_{j'ht})_{j' \neq j}$ et $(\mathbf{p}_{j'ht})_{j' \neq j}$ n'est pas explicitée sur le schéma.

Le spectre de l'instrument $\mathbf{m}_{jt}$ dépend uniquement des descripteurs de toutes les notes à cet instant $(\mathbf{p}_{jh't})_{H_j \leq h' \leq H'_j}$.

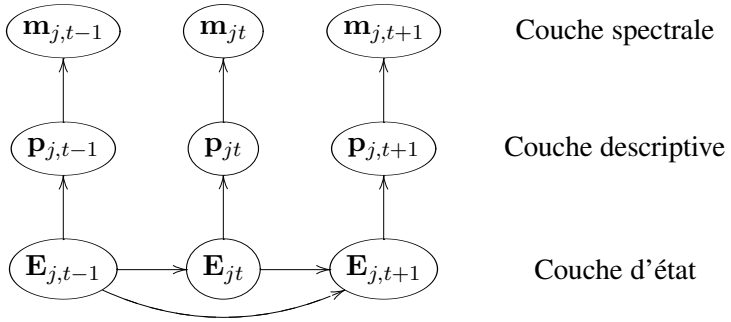

FIG. 3.1 – Représentation graphique simplifiée d'un modèle d'instrument

Le choix des variables du modèle et des relations d'indépendance conditionnelle est un compromis entre la performance et la quantité de calculs pour la transcription et la séparation. Pour construire des modèles plus proches de la réalité, nous aurions pu remplacer les états binaires des notes par les quatre états attaque, partie soutenue, relâche (et réverbération) et silence [Ero03]. Nous aurions pu aussi modéliser la continuité temporelle des descripteurs de chaque note [Vir03b], la dépendance entre les descripteurs des notes d'un accord sur instrument à oscillations soutenues, la continuité temporelle du spectre de l'instrument.

3.2 Modèles probabilistes de mélanges

Pour appliquer les modèles d'instruments à la transcription et la séparation, nous construisons maintenant des modèles de mélange. Rappelons que nous nous limitons aux mélanges réalisés par des filtres de mélange et du bruit fond stationnaires. La suite du chapitre reste cependant applicable aux mélanges variant dans le temps à quelques modifications près.

Notre hypothèse est que tous les instruments potentiellement présents dans un mélange sont modélisés par des modèles d'instruments construits comme au paragraphe précédent. Sous cette hypothèse, modéliser un mélange consiste à définir la distribution de probabilité conjointe du mélange et des variables des modèles d'instruments présents.

3.2.1 Choix des observations et des paramètres de mélange

Puisque les modèles d'instruments représentent les spectres à court terme des sources, il n'est pas utile de garder toutes les informations présentes dans la forme d'onde du mélange. Comme nous l'avons mentionné dans le paragraphe 3.1.2, il est suffisant d'en extraire une suite d'observations sur des trames à court terme $(\mathbf{o}_t)_{0\leq t\leq T-1}$, par exemple les spectres de puissance des canaux et les différences de volume et de phase inter-canal.

Nous supposons que $\mathbf{o}_t$ dépend uniquement des spectres des instruments à cet instant $(\mathbf{m}_{jt})_{1\leq j\leq n}$ et d'un vecteur de paramètres de mélange $\mathbf{\Theta}$ [RG03]. Dans un mélange convolutif long, cela ne correspond pas tout à fait à la réalité, car le mélange dépend aussi des sources sur les trames précédentes [Wu03a]. Cependant cette dépendance se manifeste surtout dans les zones de réverbération.

Pour un mélange instantané, $\mathbf{\Theta}$ regroupe les gains et les azimuts des sources et le spectre du bruit de fond stationnaire. Pour un mélange convolutif réverbérant, $\mathbf{\Theta}$ peut de plus contenir les réponses fréquentielles des filtres source-à-microphone, qui ne sont pas bien décrites par le seul azimut. Notons que $\mathbf{\Theta}$ ne contient pas nécessairement un paramétrage bijectif des filtres de mélange : les paramètres décrivant la réverbération tardive peuvent être omis.

3.2.2 Indépendance entre instruments

Nous supposons de plus que les variables des modèles d'instruments pour des sources différentes sont indépendantes. Cela rejoint notre choix de ne pas modéliser le style musical. Dans le cas contraire, nous aurions pu définir une couche d'état globale pour tous les instruments et modéliser par exemple la probabilité des différents accords ou la synchronie des attaques et des nuances.

3.3 Loi de Bayes

Suite à toutes nos hypothèses de modélisation, nous pouvons maintenant écrire la distribution conjointe des observations, des paramètres de mélange et des variables des modèles d'instruments.

Nous regroupons les variables à différents instants pour définir $\mathbf{o} = (\mathbf{o}_t)_{0\leq t\leq T-1}$, $\mathbf{m}_j = (\mathbf{m}_{jt})_{0\leq t\leq T-1}$, $\mathbf{p}_j = (\mathbf{p}_{jt})_{0\leq t\leq T-1}$ et $\mathbf{E}_j = (\mathbf{E}_{jt})_{0\leq t\leq T-1}$. Nous notons $\mathcal{M}_j$ le modèle de l'instrument j, c'est-à-dire l'ensemble des paramètres décrivant sa distribution, et $\mathcal{M}$ l'ensemble des modèles des instruments potentiellement présents. Nous appelons orchestre et nous notons $\mathcal{O}$ la liste des noms des instruments présents dans le mélange. La donnée de $\mathcal{M}$ et $\mathcal{O}$ est équivalente à celle de la liste des modèles des instruments présents $(\mathcal{M}_j)$. Enfin, nous notons $\mathcal{I}$ les informations *a priori* supplémentaires sur les observations qui influencent les valeurs de $\mathbf{\Theta}$, $\mathcal{O}$ et $(\mathbf{E}_j)$ (azimuts des sources, nombre d'instruments, partition musicale).

La loi de Bayes et les hypothèses d'indépendance conditionnelle mènent à

$$P(\mathbf{o}, \mathbf{\Theta}, (\mathbf{m}_j), (\mathbf{p}_j), (\mathbf{E}_j), \mathcal{O}, \mathcal{M}|\mathcal{I}) = P^{\text{spat}}\, P^{\text{spec}}\, P^{\text{desc}}\, P^{\text{état}}\, P(\mathbf{\Theta}|\mathcal{I})\, P(\mathcal{O}|\mathcal{I})\, P(\mathcal{M}) \qquad (3.1)$$

où

$$P^{\text{spat}} = P(\mathbf{o}|\mathbf{\Theta}, (\mathbf{m}_j)) = \prod_{t=0}^{T-1} P(\mathbf{o}_t|\mathbf{\Theta}, (\mathbf{m}_{jt})_{1\leq j\leq n}) \tag{3.2}$$

$$P^{\text{spec}} = \prod_{j=1}^{n} P(\mathbf{m}_j|\mathbf{p}_j, \mathcal{M}_j) = \prod_{t=0}^{T-1}\prod_{j=1}^{n} P(\mathbf{m}_{jt}|\mathbf{p}_{jt}, \mathcal{M}_j) \tag{3.3}$$

$$P^{\text{desc}} = \prod_{j=1}^{n} P(\mathbf{p}_j|\mathbf{E}_j, \mathcal{M}_j) = \prod_{t=0}^{T-1}\prod_{j=1}^{n}\prod_{h=H_j}^{H'_j} P(\mathbf{p}_{jht}|E_{jht}, \mathcal{M}_j) \tag{3.4}$$

$$P^{\text{état}} = \prod_{j=1}^{n} P(\mathbf{E}_j|\mathcal{M}_j) = \prod_{t=0}^{T-1}\prod_{j=1}^{n} P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1}, \mathcal{M}_j, \mathcal{I}) \tag{3.5}$$

Dans les chapitres suivants, nous proposerons des expressions paramétriques pour les distributions intervenant dans ces équations. Dans le chapitre 4, nous construirons les distributions $P(\mathbf{m}_{jt}|\mathbf{p}_{jt}, \mathcal{M}_j)$, $P(\mathbf{p}_{jht}|E_{jht}, \mathcal{M}_j)$ et $P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1}, \mathcal{M}_j, \mathcal{I})$ définissant les trois couches d'un modèle d'instrument $\mathcal{M}_j$. La distribution conditionnelle des observations $P(\mathbf{o}_t|\mathbf{\Theta}, (\mathbf{m}_{jt})_{1\leq j\leq n})$ ne sera pas définie explicitement. Dans les chapitres 4, 6 et 7 nous définirons plutôt la distribution $P(\mathbf{o}_t|\mathbf{\Theta}, (\mathbf{p}_{jt})_{1\leq j\leq n}, (\mathcal{M}_j))$ selon le type de mélange.

3.4 Cadre bayésien pour la transcription et la séparation

L'utilisation des modèles d'instruments s'effectue en deux phases : apprentissage des paramètres des modèles spécifiques à chaque instrument, puis intégration comme information *a priori* dans la transcription et la séparation. Nous expliquons dans ce paragraphe comment effectuer les quatre tâches définies dans le chapitre 1 une fois les modèles appris. Nous présentons uniquement les variables estimées et les critères d'estimation utilisés. Des algorithmes plus précis adaptés au cas particulier de modèle d'instrument choisi seront discutés dans le chapitre 4 et les suivants.

3.4.1 Identification de notes

L'identification de notes vise à trouver les listes d'accords joués par tous les instruments sachant les instruments présents. Dans les modèles d'instruments proposés, ces listes d'accords ne sont pas modélisées en tant que telles, mais en lien avec les instants de début et de fin des notes au sein des états $(\mathbf{E}_j)$. Il est donc naturel d'estimer plutôt les états $(\mathbf{E}_j)$ sachant les modèles des instruments présents $(\mathcal{M}_j)$. Notons que cela ne permet pas de distinguer deux notes de même hauteur jouées à la suite sans silence intermédiaire et une note unique de durée plus longue (il faudrait pour cela modéliser différemment les parties d'attaque et de relâche). Le critère d'estimation est le critère du Maximum *A Posteriori* (MAP) [Rab89]

$$\widehat{(\mathbf{E}_j)} = \arg\max_{(\mathbf{E}_j)} P((\mathbf{E}_j)|\,\mathbf{o}, (\mathcal{M}_j), \mathcal{I}), \tag{3.6}$$

qui consiste à intégrer la distribution conjointe de l'équation 3.1 par rapport à $\mathbf{\Theta}$, $(\mathbf{m}_j)$ et $(\mathbf{p}_j)$.

Pour certains modèles d'instruments et de mélange, l'intégrale par rapport à $(\mathbf{m}_j)$ est facilement approchable (voire calculable exactement) sous une forme paramétrique simple. L'intégrale du produit de distributions $P^{\text{spat}}\,P^{\text{spec}}$ est alors remplacée par

$$P^{\text{comb}} = P(\mathbf{o}|\mathbf{\Theta}, (\mathbf{p}_j), (\mathcal{M}_j)) = \prod_{t=0}^{T-1} P(\mathbf{o}_t|\mathbf{\Theta}, (\mathbf{p}_{jt})_{1\leq j\leq n}, (\mathcal{M}_j)). \tag{3.7}$$

Cette approche est utilisée par les algorithmes de décomposition parcimonieuse spectrale et de combinaison de modèles sur des mélanges monocanal (voir paragraphes 2.3.4 et 2.3.5). Nous serons en mesure de l'adopter aussi par la suite.

Par contre, l'intégrale par rapport à $\mathbf{\Theta}$ et $(\mathbf{p}_j)$ n'est généralement pas calculable sous une forme paramétrique simple. Les méthodes d'approximation classiques sont les méthodes variationnelles [Att99, Mis01, Gha01] et de Monte Carlo [Gha01, Dav02c, Wol03, Fév04b], assez lourdes en calculs, et la méthode de Laplace [Abd01]. Nous choisissons une approximation encore plus simple qui consiste à prendre en compte uniquement les valeurs les plus probables de $\mathbf{\Theta}$ et $(\mathbf{p}_j)$ sachant $\mathbf{o}$, $(\mathbf{E}_j)$, $(\mathcal{M}_j)$ et $\mathcal{I}$ [Abd01, Ben03].

Cela revient à considérer le nouveau critère

$$\widehat{(\mathbf{E}_j)} \approx \arg\max_{(\mathbf{E}_j)} \max_{\mathbf{\Theta},(\mathbf{p}_j)} P(\mathbf{\Theta}, (\mathbf{p}_j), (\mathbf{E}_j) | \, \mathbf{o}, (\mathcal{M}_j), \mathcal{I}). \tag{3.8}$$

Ce critère n'est pas optimal pour l'identification de notes, mais il trouve tout son intérêt pour des tâches de transcription plus avancées où $\mathbf{\Theta}$ et $(\mathbf{p}_j)$ doivent également être estimés. Dans ce cas il correspond au critère du MAP

$$\widehat{\mathbf{\Theta}}, \widehat{(\mathbf{p}_j)}, \widehat{(\mathbf{E}_j)} = \arg \max_{\mathbf{\Theta},(\mathbf{p}_j),(\mathbf{E}_j)} P(\mathbf{\Theta}, (\mathbf{p}_j), (\mathbf{E}_j) | \, \mathbf{o}, (\mathcal{M}_j), \mathcal{I}). \tag{3.9}$$

Les algorithmes d'identification de notes par décomposition parcimonieuse spectrale et par ASA computationnelle fonctionnent sur le même principe. Ils estiment les notes présentes en fonction des meilleurs coefficients de décomposition ou du meilleur regroupement des partiels (voir paragraphes 2.3.4, 2.3.2 et 2.4.5).

Remarquons que ce critère admet deux propriétés particulièrement intéressantes, pas toujours vérifiées par les algorithmes existants : le nombre de notes présentes à un instant donné n'est pas fixé à l'avance et les hauteurs et les instruments associés aux notes sont estimés conjointement.

3.4.2 Identification d'instruments

L'identification d'instruments consiste à estimer les instruments présents $\mathcal{O}$ sachant les modèles des instruments potentiellement présents $\mathcal{M}$. Cela se traduit par le critère du MAP

$$\widehat{\mathcal{O}} = \arg\max_{\mathcal{O}} P(\mathcal{O} | \, \mathbf{o}, \mathcal{M}, \mathcal{I}). \tag{3.10}$$

Son calcul implique l'intégration de la distribution conjointe de l'équation 3.1 par rapport à $\mathbf{\Theta}$, $(\mathbf{m}_j)$, $(\mathbf{p}_j)$ et $(\mathbf{E}_j)$. Comme pour l'identification de notes, nous choisissons d'approcher l'intégrale par rapport à $(\mathbf{m}_j)$ par une forme paramétrique simple et de prendre en compte uniquement les valeurs les plus probables de $\mathbf{\Theta}$, $(\mathbf{p}_j)$ et $(\mathbf{E}_j)$ sachant $\mathbf{o}$, $(\mathcal{M}_j)$ et $\mathcal{I}$.

Cela fournit le critère pratique

$$\widehat{\mathcal{O}} \approx \arg\max_{\mathcal{O}} P(\mathcal{O}|\mathcal{I}) \, P(\widehat{\mathbf{\Theta}}, \widehat{(\mathbf{p}_j)}, \widehat{(\mathbf{E}_j)} | \, \mathbf{o}, (\mathcal{M}_j), \mathcal{I}), \tag{3.11}$$

où $\widehat{\mathbf{\Theta}}$, $\widehat{(\mathbf{p}_j)}$ et $\widehat{(\mathbf{E}_j)}$ sont définis en fonction de $\mathcal{O}$ par l'équation 3.9. Cette expression fait apparaître un lien important entre identification d'instruments et identification de notes. En effet, l'estimation implique de calculer pour chaque orchestre possible la probabilité de la meilleure transcription au sens de l'équation 3.9.

Les algorithmes existants d'identification d'instruments estiment aussi simultanément les notes et les instruments présents [Kas95, Kin99], ou supposent les notes déjà identifiées auparavant [Egg03].

3.4.3 Extraction de sources

Le but de l'extraction de sources est d'estimer les sources $\mathbf{s}$ sachant les modèles des instruments présents $(\mathcal{M}_j)$. Cela correspond au critère du MAP

$$\widehat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}, (\mathcal{M}_j), \mathcal{I}). \tag{3.12}$$

Nous introduisons les variables des modèles d'instruments en développant $P(\mathbf{s}|\mathbf{x}, (\mathcal{M}_j), \mathcal{I})$ sous la forme

$$P(\mathbf{s}|\mathbf{x}, (\mathcal{M}_j), \mathcal{I}) = \int P(\mathbf{s}|\mathbf{x}, \mathbf{\Theta}, (\mathbf{m}_j))\, P(\mathbf{\Theta}, (\mathbf{m}_j)|\, \mathbf{o}, (\mathcal{M}_j), \mathcal{I})\, d\mathbf{\Theta}\, d(\mathbf{m}_j), \tag{3.13}$$

où $\mathbf{\Theta}$ et $(\mathbf{m}_j)$ sont supposés indépendants de $\mathbf{x}$ sachant $\mathbf{o}$, et $\mathbf{s}$ indépendant de $(\mathcal{M}_j)$ et $\mathcal{I}$ sachant $(\mathbf{m}_j)$. Le terme $P(\mathbf{\Theta}, (\mathbf{m}_j)|\, \mathbf{o}, (\mathcal{M}_j), \mathcal{I})$ correspond lui-même à l'intégrale de la distribution conjointe de l'équation 3.1 par rapport à $(\mathbf{p}_j)$ et $(\mathbf{E}_j)$. Nous proposons une approximation en deux étapes. Nous prenons en compte d'abord uniquement les valeurs les plus probables de $\mathbf{\Theta}$, $(\mathbf{p}_j)$ et $(\mathbf{E}_j)$ sachant $\mathbf{o}$, $(\mathcal{M}_j)$ et $\mathcal{I}$, puis uniquement la valeur la plus probable de $(\mathbf{m}_j)$ sachant $\mathbf{o}$, $\mathbf{\Theta}$, $(\mathbf{p}_j)$, $(\mathbf{E}_j)$ et $(\mathcal{M}_j)$.

Cela mène au nouveau critère

$$\widehat{\mathbf{s}} \approx \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}, \widehat{\mathbf{\Theta}}, \widehat{(\mathbf{m}_j)}), \tag{3.14}$$

où

$$\widehat{(\mathbf{m}_j)} = \arg\max_{(\mathbf{m}_j)} P((\mathbf{m}_j)|\, \mathbf{o}, \widehat{\mathbf{\Theta}}, \widehat{(\mathbf{p}_j)}, \widehat{(\mathbf{E}_j)}, (\mathcal{M}_j)) \tag{3.15}$$

et $\widehat{\mathbf{\Theta}}$, $\widehat{(\mathbf{p}_j)}$ et $\widehat{(\mathbf{E}_j)}$ sont définis par l'équation 3.9. L'application de ce critère met en lumière un lien important entre transcription et extraction de sources. En effet l'estimation des sources comprend trois étapes successives : estimation de la meilleure transcription au sens de l'équation 3.9, estimation du spectre de puissance des sources par l'équation 3.15 et filtrage du mélange selon l'équation 3.14. Le filtrage peut être invariant dans le temps ou non, et être éventuellement complété d'une resynthèse sinusoïdale des partiels cachés.

La plupart des algorithmes d'extraction de sources existants effectuent aussi des estimations approchées en plusieurs étapes. Le filtrage est réalisé en fonction du meilleur regroupement des partiels pour l'ASA computationnelle, en fonction des meilleurs coefficients de décomposition pour la décomposition parcimonieuse spectrale, en fonction du meilleur état à chaque instant pour la combinaison de modèles et en fonction des meilleurs paramètres de mélange pour la maximisation de parcimonie et le masquage temps-fréquence (voir paragraphes 2.3.2, 2.4.5, 2.3.4, 2.3.5, 2.4.3 et 2.4.4). Quelques algorithmes de maximisation de parcimonie estiment conjointement les sources et les paramètres de mélange [Lee99, Dav02b, Fév04a]. Mais souvent leur performance est identique à celle des algorithmes en deux étapes, qui peuvent estimer les paramètres de mélange avec une grande précision [Abr01, Fév04b]. L'algorithme de combinaison de modèles décrit dans [Ben03] réalise un filtrage pseudo-Wiener en intégrant les états cachés des MG, mais pas les facteurs d'échelle.

L'étape de filtrage dépend de la distribution $P(\mathbf{s}|\mathbf{x}, \mathbf{\Theta}, (\mathbf{m}_j))$. Cette distribution peut être remplacée par $P(\mathbf{s}_{\mathrm{img}\,i}|\mathbf{x}, \mathbf{\Theta}, (\mathbf{m}_j))$ dans le cas où on cherche à estimer l'image spatiale des sources sur le capteur i. Dans les chapitres 6 et 7 nous donnerons plusieurs méthodes de filtrage selon le type de mélange. Notons que dans le cas d'un mélange sur-déterminé convolutif court peu bruité, il est possible d'estimer des filtres de démixage invariants dans le temps uniquement en fonction des paramètres de mélange $\mathbf{\Theta}$. Dans ce cas, la deuxième étape est inutile. Notons aussi que la procédure d'estimation des sources ne change pas fondamentalement en remplaçant le critère MAP de l'équation 3.12 par le critère d'Espérance *A Posteriori* (EAP) [Ben03]. Seule l'équation 3.14 de filtrage par MAP est modifiée en un filtrage par EAP. Sous certaines hypothèses, les deux types de filtrage sont équivalents [Ben03].

3.4.4 Modification de scène sonore

La modification de scène sonore consiste à estimer le remix $\mathbf{x}_{\mathrm{rmx}}$ sachant les paramètres de remix $\mathbf{\Theta}_{\mathrm{rmx}}$ et les modèles des instruments présents $(\mathcal{M}_j)$. L'estimateur du MAP vaut

$$\widehat{\mathbf{x}_{\mathrm{rmx}}} = \arg\max_{\mathbf{x}_{\mathrm{rmx}}} P(\mathbf{x}_{\mathrm{rmx}}|\mathbf{x}, \mathbf{\Theta}_{\mathrm{rmx}}, (\mathcal{M}_j), \mathcal{I}). \tag{3.16}$$

Avec l'approximation utilisée pour l'extraction de sources, nous obtenons l'estimateur approché

$$\widehat{\mathbf{x}_{\mathrm{rmx}}} \approx \arg\max_{\mathbf{x}_{\mathrm{rmx}}} P(\mathbf{x}_{\mathrm{rmx}}|\mathbf{x}, \mathbf{\Theta}_{\mathrm{rmx}}, \widehat{\mathbf{\Theta}}, \widehat{(\mathbf{m}_j)}), \tag{3.17}$$

où $\widehat{\mathbf{\Theta}}$ et $\widehat{(\mathbf{m}_j)}$ sont définis par les équations 3.9 et 3.15. L'estimateur obtenu avec un critère EAP est semblable. Cela mène encore à un algorithme en trois étapes, similaire aux algorithmes existants [Ave02, Rad02]. Des méthodes de filtrage correspondant à la distribution $P(\mathbf{x}_{\mathrm{rmx}}|\mathbf{x}, \mathbf{\Theta}_{\mathrm{rmx}}, \mathbf{\Theta}, (\mathbf{m}_j))$ seront décrites dans les chapitres 6 et 7.

3.5 Cadre bayésien pour l'apprentissage des paramètres des modèles

Nous terminons ce chapitre en présentant le critère utilisé pour apprendre les paramètres des modèles d'instruments et en discutant le choix des données d'apprentissage. Un algorithme d'apprentissage adapté au modèle d'instrument choisi sera décrit dans le chapitre 4.

3.5.1 Critère d'estimation

L'apprentissage vise à estimer les paramètres de l'ensemble des modèles d'instruments $\mathcal{M}$ sur des observations $\mathbf{o}$, étiquetées par la liste des instruments présents $\mathcal{O}$ et éventuellement par d'autres informations $\mathcal{I}$. Le critère de MAP s'écrit

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{M}|\,\mathbf{o}, \mathcal{O}, \mathcal{I}). \tag{3.18}$$

Lorsque la distribution *a priori* $P(\mathcal{M})$ est uniforme, ce critère est appelé Maximum de Vraisemblance (MV) [Rab89]. Il est similaire au critère d'identification d'instruments, car il requiert une intégration de la distribution conjointe de l'équation 3.1 par rapport à $\mathbf{\Theta}$, $(\mathbf{m}_j)$, $(\mathbf{p}_j)$ et $(\mathbf{E}_j)$. Nous utilisons donc la même approximation, qui mène au critère pratique

$$\widehat{\mathcal{M}} \approx \arg\max_{\mathcal{M}} P(\mathcal{M})\, P(\widehat{\mathbf{\Theta}}, \widehat{(\mathbf{p}_j)}, \widehat{(\mathbf{E}_j)}|\,\mathbf{o}, (\mathcal{M}_j), \mathcal{I}), \tag{3.19}$$

où $\widehat{\mathbf{\Theta}}$, $\widehat{(\mathbf{p}_j)}$ et $\widehat{(\mathbf{E}_j)}$ sont définis en fonction de $\mathcal{M}$ par l'équation 3.9. Ce critère implique de calculer pour chaque ensemble de modèles possibles la probabilité de la meilleure transcription au sens de l'équation 3.9. En pratique la recherche exhaustive est remplacée par une réestimation alternative du meilleur ensemble de modèles d'instruments et de la meilleure transcription, qui converge lorsqu'un maximum local du critère est atteint. Cette procédure de réestimation est connue sous le nom d'algorithme Espérance/Maximisation (EM) approché [Ben03].

Cette approche est souvent utilisée pour l'apprentissage de dictionnaires de spectres [Abd01, Vir03b, Ben03]. Par contre, les MG ou MMC sont généralement appris par un algorithme EM exact car l'intégration par rapport aux états nécessite moins de calculs [Ben03, Rab89].

3.5.2 Choix des données d'apprentissage

L'algorithme EM est sensible à l'initialisation des modèles. Un mauvais modèle initial engendre une mauvaise transcription des données d'apprentissage, qui à son tour engendre un mauvais modèle réestimé, et ainsi de suite. Ce comportement est évitable en choisissant des données d'apprentissage simples et bien étiquetées. Plus les résultats de transcription sont contraints par l'étiquetage, plus les erreurs sont réduites. Malgré tout, un petit biais peut subsister sur certains paramètres avec le critère approché de l'équation 3.19 (voir paragraphe 4.6).

La majorité des algorithmes effectuent l'apprentissage sur des bases de données de notes isolées de hauteur connue [Kas95, Kin99, Kas99, Fit03b, RG03]. Dans ce cas seuls les descripteurs $(\mathbf{p}_{jht})_{0\leq t\leq T-1}$ pour une note h et un instrument j fixés sont réestimés à chaque transcription, ce qui exclut les erreurs grossières. Cependant la répartition des notes des bases de données selon la durée ou la nuance ne correspond pas forcément à la répartition observée sur des enregistrements solo.

D'autres algorithmes proposent d'apprendre chaque modèle sur des enregistrements solo à l'aide de partitions musicales [Rap02]. Chaque transcription réalise un alignement temporel entre les partitions et les observations pour réestimer $\mathbf{p}_j$ et $\mathbf{E}_j$. Cette méthode peut donner des erreurs lorsque l'interprétation diffère des partitions ou lorsque l'alignement est difficile, par exemple avec un *tempo* rapide, des notes successives à intervalle d'octave ascendante ou d'unison, ou des attaques peu marquées. De plus elle ne permet pas d'apprendre les paramètres des notes absentes des enregistrements ou masquées dans des accords.

Par la suite, nous considérerons uniquement des modèles appris sur une base de données de notes isolées car nous ne disposons pas d'extraits solo étiquetés en quantité suffisante.

3.5.3 À propos de l'apprentissage discriminant

Le critère d'apprentissage bayésien par MAP ne garantit pas que les modèles d'instruments appris sont optimaux en terme de performance de transcription et de séparation. Les critères d'apprentissage discriminant (non bayésien) par Maximum de l'Information Mutuelle (MIM) [Rab89] permettent en partie de corriger ce défaut. Par exemple, le critère

$$\widehat{\mathcal{M}} = \arg\max_{\mathcal{M}} P(\mathcal{O}|\,\mathbf{o}, \mathcal{M}, \mathcal{I}) \tag{3.20}$$

garantit que les modèles d'instruments appris sont optimaux pour l'identification d'instruments sur les données d'apprentissage. Un critère approché consiste à maximiser le rapport entre la probabilité de l'orchestre *a priori* $\mathcal{O}$ et celle du meilleur orchestre $\widehat{\mathcal{O}}$ défini par l'équation 3.11, ces probabilités étant elles-mêmes approchées comme dans l'équation 3.11. Cela donne encore un algorithme de réestimation alternée. Des critères semblables permettent d'apprendre des modèles optimaux pour les autres tâches.

La limitation des critères de MIM est que leur optimalité n'est prouvée que pour les données d'apprentissage et la tâche choisie. Un modèle optimal sur des notes isolées ne l'est pas forcément sur de la musique de chambre. L'apprentissage par MIM doit être réalisé directement sur des mélanges, ce qui compromet l'optimalité des critères à cause des limitations de l'algorithme EM. De plus un modèle optimal pour l'identification de notes ne l'est pas forcément pour l'extraction de sources.

3.5.4 Choix de la taille des modèles

La taille d'un modèle est définie comme le nombre de paramètres indépendants qu'il contient. Un modèle de grande taille modélise mieux les données d'apprentissage mais n'est pas forcément plus performant sur des données de test différentes. En effet il risque de prendre en compte des détails non reproduisibles [Rab89].

Pour éviter ce phénomène de sur-apprentissage, il est possible de définir la distribution *a priori* $P(\mathcal{M})$ de sorte à favoriser les modèles de petite taille dans le critère du MAP. Cette possibilité est difficile

à mettre en pratique. La méthode usuelle, appelée validation croisée, consiste à apprendre plusieurs modèles de taille différente selon le critère du MV, à tester leur performance sur des données de test différentes pour une tâche fixée et à conserver le plus performant.

Dans le chapitre suivant, la taille sera un critère important pour guider notre définition d'un modèle d'instrument particulier. Les paramètres du modèle seront choisis de sorte à être en nombre réduit et facilement partageables.

Chapitre 4

Présentation du modèle d'instrument choisi

Ce quatrième chapitre propose un modèle d'instrument particulier, faisant partie de la famille décrite au chapitre précédent, et basé sur une ASI non linéaire. Nous définissons ce modèle dans les paragraphes 4.1, 4.2, 4.3 et 4.4 en spécifiant la distribution paramétrique des variables de chaque couche et en exprimant leur distribution conjointe par une loi de Bayes pondérée. Nous proposons ensuite dans les paragraphes 4.5 et 4.6 des algorithmes associés au modèle pour la transcription et l'apprentissage sur des enregistrements solo. Nous discutons enfin dans le paragraphe 4.7 le choix de paramètres de calcul du spectre à court terme appropriés.

Le modèle d'instrument proposé et les algorithmes associés résultent d'un travail original, qui s'inspire des modèles existants cités dans le chapitre précédent. Nous fournirons dans les paragraphes 5.1 et 5.2 des exemples de modèles et une validation expérimentale de nos hypothèses. Par rapport à l'article [Vin04c], ce chapitre propose un nouveau modèle de couche d'état et de nouveaux algorithmes de transcription et d'apprentissage.

4.1 Couche spectrale

Définir la couche spectrale d'un modèle d'instrument c'est exprimer la distribution $P(\mathbf{m}_{jt}|\mathbf{p}_{jt}, \mathcal{M}_j)$ sous forme paramétrique. Au passage cette définition implique de définir aussi les descripteurs $\mathbf{p}_{jt}$.

4.1.1 Spectre d'un accord

Nous approchons le spectre à court terme $\mathbf{m}_{jt}$ de l'instrument j au temps t par un spectre modèle $\mathbf{m}'_{jt}$ défini par une combinaison linéaire des spectres de puissance des notes [Abd01]

$$\mathbf{m}'_{jt} = \sum_{h=H_j}^{H'_j} \exp(e_{jht}) \exp(\mathbf{\Phi'}_{jht}), \tag{4.1}$$

où $\mathbf{\Phi'}_{jht}$ est le spectre de log-puissance de la note h à cet instant et e_{jht} est sa log-puissance, liée à son volume instantané. D'autres modèles similaires utilisent une combinaison linéaire des spectres d'amplitude [Abd01] ou un maximum à chaque fréquence [Row00].

La décomposition d'un accord en notes permet d'éviter les procédures complexes de partage de paramètres nécessaires avec les modélisations par MG [Rap02]. L'utilisation de facteurs de puissance réduit

aussi le nombre de paramètres et augmente la performance du modèle [Ben03]. De plus, la modélisation des puissances (positives) par des exponentielles de log-puissances (réelles) permet de s'affranchir des contraintes de positivité explicites utilisées dans quelques algorithmes de décomposition parcimonieuse spectrale [Ben03, Vir03b].

4.1.2 Spectre d'une note

Le spectre de chaque note ${\mathbf{\Phi'}}_{jht}$ varie au cours du temps. Ces variations peuvent concerner la puissance des partiels, la fréquence fondamentale, ou bien l'énergie non harmonique. Elles sont dues à la fois au style de jeu (volume et hauteur contrôlés par l'instrumentiste), aux caractéristiques de l'instrument (apparition et disparition progressive des partiels à l'attaque et à la relâche) et aux conditions d'enregistrement (réponse de salle non plate). Nous approchons toutes ces variations par une combinaison linéaire de spectres de log-puissance

$$\mathbf{\Phi'}_{jht} = \mathbf{\Phi}_{jh} + \sum_{k=1}^{K_j} v_{jht}^k \mathbf{U}_{jh}^k, \tag{4.2}$$

où $\mathbf{\Phi}_{jh}$ est le spectre de log-puissance moyen de la note h, $\mathbf{U}_{jh}^k$ un "spectre de variation" décrivant les variations autour du spectre moyen et v_{jht}^k un "descripteur de variation" associé. Nous supposons que la puissance totale de $\mathbf{\Phi}_{jh}$ et la norme de $\mathbf{U}_{jh}^k$ sont unitaires. La donnée de e_{jht} et des $(v_{jht}^k)_{1 \leq k \leq K_j}$ définit les descripteurs $\mathbf{p}_{jht}$ de la note h au temps t. Ce modèle a été utilisé pour la reconnaissance d'instruments et la classification de genre musical, mais directement sur le spectre de log-puissance d'un enregistrement solo (sans passer par les notes) [Cas02]. Un modèle similaire de décomposition du cepstre a aussi été utilisé récemment pour la reconnaissance d'instruments sur des notes isolées [Ero03].

L'avantage d'une représentation du spectre par combinaison linéaire est de modéliser de façon unique toutes les variations possibles du spectre, et de façon plus compacte qu'un MG. En particulier le modèle s'applique aux percussions comme aux instruments harmoniques.

De plus, la modélisation par combinaison linéaire du spectre de log-puissance nécessite moins de composantes de variation que celle du spectre de puissance. Prenons l'exemple de la modélisation des variations de fréquence. Ces variations peuvent s'approcher avec une seule composante par note par un développement limité à l'ordre un. Chaque composante $\mathbf{U}_{jh}^k$ est alors définie grossièrement comme la dérivée de $(\mathbf{\Phi}_{jh})_{H_j \leq h \leq H'_j}$ par rapport à h (voir le paragraphe 4.6.1 pour une définition plus précise). Si la combinaison linéaire est réalisée en puissance, cette approximation est valable uniquement pour de très petits écarts de fréquence, et peut mener à des valeurs de puissance négatives pour des écarts de fréquence plus importants. Ceci est dû au caractère très pointu des pics représentant les partiels (la valeur absolue de la dérivée seconde est élevée). Si la combinaison linéaire est réalisée en log-puissance, les pics sont moins pointus et l'approximation est plus satisfaisante.

4.1.3 Distribution de l'erreur résiduelle

Pour définir la distribution de la couche spectrale, il reste à définir la distribution de l'erreur résiduelle, c'est-à-dire l'erreur entre le spectre réel $\mathbf{m}_{jt}$ et le spectre modèle $\mathbf{m}'_{jt}$.

Les algorithmes de décomposition parcimonieuse spectrale modélisent l'erreur résiduelle en puissance $\mathbf{m}_{jt} - \mathbf{m}'_{jt}$ comme un bruit gaussien [Abd01, Vir03b]. Appliqué à l'identification de notes, ce modèle tend à considérer les notes de faible volume comme absentes car il affecte peu d'importance aux erreurs de faible puissance [Vin01, Fit03a]. De plus, il discrimine mal les instruments car l'enveloppe spectrale de puissance des notes est semblable pour la plupart des instruments (la puissance du premier partiel est supérieure à celle des autres partiels) [Kli00].

Ces limitations semblent dues au fait que l'erreur en puissance n'est pas indépendante du spectre modèle : en pratique elle semble plutôt correspondre à un bruit multiplicatif qu'à un bruit additif [Abd01].

Cette observation correspond aux hypothèses des algorithmes de combinaison de modèles, qui modélisent généralement chaque instrument par un MG sur le spectre de log-puissance [Row00, Egg03]. Des MG appris sur le spectre de log-puissance donnent de meilleurs résultats d'extraction de sources que des MG appris sur le spectre de puissance [Ben03].

Nous définissons l'erreur résiduelle en log-puissance $\boldsymbol{\alpha}_{jt}$ par

$$\log(\mathbf{m}_{jt}) = \log(\mathbf{m}'_{jt}) + \boldsymbol{\alpha}_{jt}, \tag{4.3}$$

et nous supposons que $\boldsymbol{\alpha}_{jt}$ est un bruit gaussien à covariance diagonale isotrope d'écart-type σ_j^{α}. Le choix d'une covariance diagonale isotrope permet en effet de réduire le nombre de paramètres nécessaires pour représenter la distribution. La distribution conditionnelle de la couche spectrale est définie par

$$P(\mathbf{m}_{jt}|\mathbf{p}_{jt}, \mathcal{M}_j) = \prod_{f=0}^{F-1} \mathcal{N}(\alpha_{jtf}; 0, \sigma_j^{\alpha}). \tag{4.4}$$

Notons que cette hypothèse n'est pas complètement vérifiable expérimentalement, car le son d'un instrument est toujours observé en présence de bruit de fond. Le modèle résultant en présence de bruit de fond est présenté dans le paragraphe 4.5.1.

Cette modélisation de l'erreur résiduelle admet deux propriétés intéressantes. Premièrement elle est applicable à tous les types d'instruments, y compris aux percussions, aux instruments produisant des partiels non harmoniques ou des partiels harmoniques avec fondamentale absente. Deuxièmement, dans le cas d'instruments harmoniques, la hauteur des notes et les instruments associés correspondent à une erreur résiduelle unique. En particulier il n'est pas nécessaire que les partiels des notes présentes correspondent aux pics spectraux observés. Cette notion de pic pose habituellement problème car un partiel présent peut ne correspondre à aucun pic observé (si sa fréquence est proche de celle d'un autre partiel plus puissant) et un pic observé peut ne correspondre à aucun partiel présent (c'est le cas pour les pics du spectre du bruit de fond ou les pics correspondant aux lobes secondaires du spectre d'un partiel).

Le modèle global combinant les équations 4.1, 4.2 et 4.3 est un cas particulier d'ASI non linéaire, où chaque sous-espace représente une note et où les non-linéarités sont fixées (contrairement aux modèles d'ACI non linéaire plus généraux où les non-linéarités doivent être estimées [Jut03]).

4.1.4 Partage de paramètres

Le nombre de paramètres du modèle peut encore être réduit en pratique, car les valeurs des spectres (Φ_{jhf}) et (U^k_{jhf}) suivent des structures en partie prédictibles selon la hauteur h et la fréquence f.

Les valeurs de spectres de hauteurs différentes peuvent être partagées en identifiant la structure formantique de l'instrument [Dud02, O'L03], ou plus simplement en les représentant par un spectre unique transposé [Jen99] (ou par combinaison linéaire d'un nombre réduit de spectres transposés). Le partage doit rester limité à des notes de hauteur proche, car les variations de l'enveloppe spectrale selon la hauteur forment une caractéristique importante pour l'identification d'instruments [Kit03, Jen99, Mar99b].

Les valeurs d'un spectre dans différentes bandes de fréquence peuvent aussi être représentées par un nombre réduit de paramètres sous certaines contraintes. Par exemple, le spectre de puissance d'une note harmonique peut être représenté comme la somme des spectres de puissance de ses partiels, ou comme une somme de spectres d'enveloppe lisse représentant plusieurs partiels voisins [Vir03a]. Le spectre d'un partiel est défini par deux paramètres seulement : sa fréquence et sa puissance. D'autres types de contraintes sont proposées dans le paragraphe 4.6.1.

4.2 Couche descriptive

La couche descriptive est décrite par le produit des distributions conditionnelles $P(\mathbf{p}_{jht}|E_{jht}, \mathcal{M}_j)$.

Les log-puissances des notes absentes sont contraintes par définition à $e_{jht} = -\infty$ et les descripteurs de variation des notes absentes peuvent aussi être contraints à une valeur arbitraire $v^k_{jht} = 0$. Nous supposons que les descripteurs des notes présentes sont indépendants et suivent des lois gaussiennes de moyennes (μ^e_{jh}) et (μ^v_{jhk}) et d'écarts-type (σ^e_{jh}) et (σ^v_{jhk}).

La distribution résultante s'écrit

$$P(\mathbf{p}_{jt}|\mathbf{E}_{jt}, \mathcal{M}_j) = \prod_{h \in \mathcal{A}_{jt}} \Big(\mathcal{N}(e_{jht}; \mu^e_{jh}, \sigma^e_{jh}) \prod_{k=1}^{K_j} \mathcal{N}(v^k_{jht}; \mu^v_{jhk}, \sigma^v_{jhk}) \Big), \tag{4.5}$$

où $\mathcal{A}_{jt} = \{h / E_{jht} = 1\}$ est l'ensemble des notes présentes au temps t. Notons que cette distribution n'est pas à proprement parler la distribution des descripteurs, mais son intégrale par rapport aux descripteurs des notes absentes. La véritable distribution modélise les contraintes sur ces descripteurs à l'aide de distributions de Dirac, qui sont ensuite intégrées pour estimer le critère de transcription par MAP de l'équation 3.6 [Vie01].

Comme précédemment, le nombre de paramètres peut être réduit en partageant les moyennes et les écarts-types entre notes de hauteur voisine [Jen99].

4.3 Couche d'état

La couche d'état se définit par la distribution conditionnelle $P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0 \leq t' \leq t-1}, \mathcal{M}_j, \mathcal{I})$. Selon la tâche à accomplir et la présence ou non d'une partition musicale *a priori* $\mathcal{I}$, nous considérons deux modèles paramétriques différents pour cette distribution.

4.3.1 Modèle factoriel

Le modèle le plus simple, appelé modèle de Bernoulli factoriel ou plus simplement modèle factoriel, suppose que tous les états $(E_{jht})_{H_j \leq h \leq H'_j, 0 \leq t \leq T-1}$ sont indépendants et suivent une loi de Bernoulli de même paramètre Z_j.

L'indépendance des états implique l'indépendance des descripteurs, au lieu de leur simple indépendance conditionnelle, ce qui correspond à l'hypothèse habituelle de l'ASI (voir paragraphe 2.3.4). La probabilité Z_j est alors un facteur de parcimonie [Abd01] : plus elle est élevée, plus les transcriptions contenant un faible nombre de notes présentes sont favorisées.

La distribution s'exprime sous la forme

$$P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0 \leq t' \leq t-1}, \mathcal{M}_j) = (1 - Z_j)^{\#\mathcal{A}_{jt}} (Z_j)^{N_j - \#\mathcal{A}_{jt}}, \tag{4.6}$$

où $N_j = H'_j - H_j + 1$ est le nombre de notes jouables par l'instrument j.

4.3.2 Modèle segmental d'instrument monophonique

Ce modèle factoriel peut être amélioré en ajoutant à la propriété de parcimonie une propriété de continuité temporelle. Pour cela, le plus simple est de modéliser la suite d'états $(E_{jht})_{0 \leq t \leq T-1}$ pour chaque note h par une chaîne de Markov (d'ordre un) à valeurs dans $\{0, 1\}$. Le modèle global est alors une chaîne de Markov factorielle [Gha97] à valeurs dans $\{0, 1\}^{N_j}$. Appliqué à la transcription, ce modèle améliore un peu les résultats, mais mène toujours à un nombre élevé d'erreurs d'insertion [Vin04c, Rap02]. En effet il modélise la durée des notes par une distribution exponentielle, ce qui affecte une probabilité importante aux notes de courte durée et de très longue durée. Il est possible de concentrer la durée des notes

autour d'une valeur moyenne et d'imposer une valeur minimale en représentant la suite d'états pour chaque note par une chaîne de Markov avec plus de deux états [Bon96, Rap02]. Mais cela augmente encore la probabilité des notes de très longue durée. De plus, l'hypothèse d'indépendance des états de notes différentes ne permet pas de modéliser la durée minimale écoulée entre deux accords successifs d'une mélodie.

Pour ces raisons, nous n'utilisons pas le modèle de chaîne de Markov factorielle par la suite. Nous modélisons la durée des notes et leurs instants d'attaque à l'aide d'un modèle non factoriel plus complexe pouvant modéliser des distributions de durée arbitraires. Par simplicité, nous explicitons le modèle uniquement pour un instrument monophonique. Cependant, il reste valable pour un instrument polyphonique à quelques modifications près.

Description du modèle

Nous découpons la ligne temporelle en segments successifs, où le début de chaque segment correspond à l'attaque d'une nouvelle note (différente des notes présentes à l'instant précédent). La donnée des instants d'attaque des notes et des durées des segments est équivalente. Nous supposons que les durées des notes et des segments sont indépendantes et distribuées selon $\mathcal{D}_j^{\text{note}}$ et $\mathcal{D}_j^{\text{segm}}$. Nous supposons aussi que la hauteur de chaque nouvelle note est distribuée uniformément parmi les hauteurs des notes absentes à l'instant précédent. Nous ne modélisons pas le rythme ou les caractéristiques du jeu *legato* ou *staccato*. Il faudrait pour cela introduire une dépendance entre les durées de segments successifs [Cem03], ou entre la durée d'un segment et de la note correspondante.

Ce modèle généralise les modèles segmentaux utilisés en parole, où chaque segment contient un phonème unique [Ost96, Rab89]. Dans le cadre musical, chaque segment contient non seulement la note jouée au début de ce segment, mais aussi les parties réverbérées des notes précédentes. La prise en compte de ces parties réverbérées évite qu'elles soient attribuées à d'autres instruments que celui dont elles proviennent. Elle est donc essentielle pour la transcription et la séparation.

Les modèles segmentaux usuels contraignent le premier et le dernier segment à coïncider avec les limites temporelles d'observation $t = 0$ et $t = T - 1$. Nous supposons au contraire que le premier et le dernier segment et les notes présentes en $t = 0$ ou $t = T - 1$ peuvent se poursuivre au-delà. Cela diminue les erreurs de transcription au voisinage de ces limites et permet d'évaluer la probabilité de chemins partiels $(\mathbf{E}_{jt'})_{0 \leq t' \leq t}$ à chaque instant, et non seulement au début de chaque segment. Nous en verrons l'utilité dans le paragraphe 4.5 pour la transcription par algorithme de recherche en faisceaux.

En pratique, cette hypothèse mène à une expression assez complexe. Les durées effectives du dernier segment et des notes présentes en $t = T - 1$ sont indépendantes sachant leurs instants de début. La probabilité des durées observées (tronquées) est alors calculable par intégration et fait intervenir les fonctions de distribution cumulative $\mathcal{Q}_j^{\text{note}}$ et $\mathcal{Q}_j^{\text{segm}}$ définies par

$$\mathcal{Q}_j^{\text{note}}(d) = \sum_{d' \geq d} \mathcal{D}_j^{\text{note}}(d') \qquad \text{et} \qquad \mathcal{Q}_j^{\text{segm}}(d) = \sum_{d' \geq d} \mathcal{D}_j^{\text{segm}}(d'). \tag{4.7}$$

Par contre, les durées effectives du premier segment et des notes présentes en $t = 0$ ne sont pas indépendantes sachant leurs instants de fin. De façon approchée, nous supposons que les durées observées (tronquées) sont indépendantes et suivent des lois $\mathcal{D}'_j{}^{\text{note}}$ et $\mathcal{D}'_j{}^{\text{segm}}$ proportionnelles à $\mathcal{Q}_j^{\text{note}}$ et $\mathcal{Q}_j^{\text{segm}}$ (ces fonctions étant supposées sommables). La probabilité des états initiaux $P(\mathbf{E}_{j,0})$ est définie par l'équation 4.6.

Expression de la probabilité d'un chemin

Numérotons de 0 à R_j-1 les notes présentes en $t=0$, de R_j à R'_j-1 les notes entièrement observées, et de R'_j à R''_j-1 les notes présentes en $t=T-1$. Appelons $(t_r)_{R_j\leq r\leq R''_j-1}$ les instants d'attaque successifs observés et $(d_r)_{0\leq r\leq R''_j-1}$ les durées des notes observées. Nous définissons la probabilité du chemin global $\mathbf{E}_j$ par

$$\begin{aligned} P(\mathbf{E}_j|\mathcal{M}_j) =& {\mathcal{D}'}_j^{\text{segm}}(t_{R_j}) \prod_{r=R_j}^{R''_j-2} \mathcal{D}_j^{\text{segm}}(t_{r+1}-t_r)\mathcal{Q}_j^{\text{segm}}(T-t_{R''_j-1}) \\ & \prod_{r=0}^{R_j-1} {\mathcal{D}'}_j^{\text{note}}(d_r) \prod_{r=R_j}^{R'_j-1} \mathcal{D}_j^{\text{note}}(d_r) \prod_{r=R'_j}^{R''_j-1} \mathcal{Q}_j^{\text{note}}(d_r) \\ & P(\mathbf{E}_{j,0}) \prod_{r=R_j}^{R''_j-1} (N_j - \#\mathcal{A}_{j,t_r-1})^{-1}. \end{aligned} \tag{4.8}$$

Cette équation n'est parfaitement valable que lorsque la durée d'observation est assez longue (au moins une attaque est observée et les notes présentes en $t=0$ se terminent avant $t=T-1$). Sinon les distributions ${\mathcal{D}'}_j^{\text{note}}$ et ${\mathcal{D}'}_j^{\text{segm}}$ doivent être remplacées par les fonctions de distribution cumulative correspondantes.

Cette équation permet de calculer la probabilité conditionnelle $P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1},\mathcal{M}_j)$ à l'aide des probabilités des chemins partiels $P((\mathbf{E}_{jt'})_{0\leq t'\leq t},\mathcal{M}_j)$ et $P((\mathbf{E}_{jt'})_{0\leq t'\leq t-1},\mathcal{M}_j)$. Appelons d_{jt}^{segm} la durée écoulée au temps t depuis la dernière attaque (ou depuis $t=0$) et d_{jht}^{note} la durée écoulée au temps t depuis la dernière attaque de la note h (ou depuis $t=0$). Nous obtenons pour t assez grand

$$P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1},\mathcal{M}_j) =$$

$$\begin{cases} \mathcal{T}_j^{\text{segm}}(d_{jt}^{\text{segm}}) \displaystyle\prod_{h'\in\mathcal{A}_{j,t-1}\setminus\mathcal{A}_{jt}} (1-\mathcal{T}_j^{\text{note}}(d_{jh't}^{\text{note}})) \prod_{h'\in\mathcal{A}_{jt}} \mathcal{T}_j^{\text{note}}(d_{jh't}^{\text{note}}) & \text{si } \mathcal{A}_{jt}\subset\mathcal{A}_{j,t-1}, \\ \dfrac{1-\mathcal{T}_j^{\text{segm}}(d_{jt}^{\text{segm}})}{N_j-\#\mathcal{A}_{j,t-1}} \displaystyle\prod_{h'\in\mathcal{A}_{j,t-1}\setminus\mathcal{A}_{jt}} (1-\mathcal{T}_j^{\text{note}}(d_{jh't}^{\text{note}})) \prod_{h'\in\mathcal{A}_{j,t-1}\cap\mathcal{A}_{jt}} \mathcal{T}_j^{\text{note}}(d_{jh't}^{\text{note}}) & \text{si } \mathcal{A}_{jt}\not\subset\mathcal{A}_{j,t-1}, \end{cases} \tag{4.9}$$

où $\mathcal{T}_j^{\text{note}}$ et $\mathcal{T}_j^{\text{segm}}$ sont les probabilités de continuation définies par

$$\mathcal{T}_j^{\text{note}}(d) = \frac{\mathcal{Q}_j^{\text{note}}(d+1)}{\mathcal{Q}_j^{\text{note}}(d)} \qquad \text{et} \qquad \mathcal{T}_j^{\text{segm}}(d) = \frac{\mathcal{Q}_j^{\text{segm}}(d+1)}{\mathcal{Q}_j^{\text{segm}}(d)}. \tag{4.10}$$

Le premier cas de l'équation 4.9 décrit la probabilité de transition entre états à l'intérieur d'un segment, et le deuxième la probabilité de transition entre états à l'interface entre deux segments. L'équation reste valable pour les petites valeurs de t en remplaçant $\mathcal{T}_j^{\text{segm}}(d_{jt}^{\text{segm}})$ et $\mathcal{T}_j^{\text{note}}(d_{jh't}^{\text{note}})$ par ${\mathcal{T}'}_j^{\text{segm}}(d_{jt}^{\text{segm}})$ et ${\mathcal{T}'}_j^{\text{note}}(d_{jh't}^{\text{note}})$ lorsque $d_{jt}^{\text{segm}}=t+1$ ou $d_{jh't}^{\text{note}}=t+1$, où ${\mathcal{T}'}_j^{\text{note}}$ et ${\mathcal{T}'}_j^{\text{segm}}$ sont les probabilités de continuation définies à partir de ${\mathcal{D}'}_j^{\text{note}}$ et ${\mathcal{D}'}_j^{\text{segm}}$.

Choix des distributions de durée

La performance du modèle segmental dépend beaucoup des distributions de durée choisies. Dans la suite, nous utilisons des distributions de durée log-gaussiennes avec un seuil de durée minimale, définies

par

$$\mathcal{D}_j^{\text{note}}(d) = \begin{cases} \dfrac{\mathcal{N}(\log(d); \mu_j^n; \sigma_j^n)}{\sum_{d' \geq d_j^n} \mathcal{N}(\log(d'); \mu_j^n; \sigma_j^n)} & \text{si } d \geq d_j^n, \\ 0 & \text{sinon,} \end{cases} \tag{4.11}$$

$$\mathcal{D}_j^{\text{segm}}(d) = \begin{cases} \dfrac{\mathcal{N}(\log(d); \mu_j^s; \sigma_j^s)}{\sum_{d' \geq d_j^s} \mathcal{N}(\log(d'); \mu_j^s; \sigma_j^s)} & \text{si } d \geq d_j^s, \\ 0 & \text{sinon.} \end{cases} \tag{4.12}$$

Ces distributions sont proches des densités gamma souvent utilisées en traitement de la parole [Rab89].

Lorsqu'une partition $\mathcal{I}$ est disponible, la probabilité conditionnelle $P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1}, \mathcal{M}_j, \mathcal{I})$ se calcule de façon semblable à l'équation 4.9. Les probabilités de continuation $\mathcal{T}_j^{\text{note}}$ et $\mathcal{T}_j^{\text{segm}}$ sont calculées en fonction des durées des notes correspondantes dans la partition, et une seule hauteur de note donnée par la partition est considérée à chaque début de segment au lieu de $N_j - \#\mathcal{A}_{j,t-1}$ hauteurs possibles.

4.4 Loi de Bayes pondérée

Expérimentalement, les hypothèses d'indépendance conditionnelle définissant chaque couche ne sont pas parfaitement valides. Par exemple, la définition générale d'un modèle d'instrument au chapitre précédent suppose que les descripteurs $(\mathbf{p}_{jht})_{0\leq t\leq T-1}$ d'une note h sont indépendants sachant ses états $(\mathbf{E}_{jht})_{0\leq t\leq T-1}$ et que les spectres à court terme $(\mathbf{m}_{jt})_{0\leq t\leq T-1}$ de l'instrument sont indépendants sachant ses descripteurs $(\mathbf{p}_{jt})_{0\leq t\leq T-1}$. Mais en pratique il existe une dépendance entre descripteurs ou spectres à des instants voisins. De même dans ce chapitre la description de l'erreur résiduelle $\boldsymbol{\alpha}_{jt}$ par un bruit gaussien à covariance diagonale ne tient pas compte des dépendances entre les valeurs de l'erreur à des fréquences proches.

Généralement, ces dépendances sont faibles. Les modéliser explicitement conduit à des calculs supplémentaires, voire à des problèmes de sur-apprentissage. Les méthodes d'identification d'instruments sur des enregistrements solo modélisent généralement le cepstre au lieu du spectre, afin de décorréler les observations à des fréquences différentes [Bro01, Mar99a, Gal95]. Mais cela n'est pas applicable à des mélanges.

Nous approchons ces dépendances plus simplement dans l'expression de la probabilité conjointe des variables en remplaçant la loi de Bayes "naïve" de l'équation 3.1 par la loi de Bayes pondérée [Han01]

$$P(\mathbf{o}, \boldsymbol{\Theta}, (\mathbf{m}_j), (\mathbf{p}_j), (\mathbf{E}_j), \mathcal{O}, \mathcal{M}|\mathcal{I}) = (P^{\text{spat}})^{w_{\text{spat}}} (P^{\text{spec}})^{w_{\text{spec}}} (P^{\text{desc}})^{w_{\text{desc}}} (P^{\text{état}})^{w_{\text{état}}} \\ P(\boldsymbol{\Theta}|\mathcal{I})^{w_{\boldsymbol{\Theta}}} P(\mathcal{O}|\mathcal{I})^{w_{\mathcal{O}}} P(\mathcal{M})^{w_{\mathcal{M}}}. \tag{4.13}$$

Les poids w_{spat}, w_{spec}, w_{desc}, $w_{\text{état}}$, $w_{\boldsymbol{\Theta}}$, $w_{\mathcal{O}}$ et $w_{\mathcal{M}}$ sont compris entre 0 et 1, et d'autant plus proches de 1 que les variables correspondantes sont conditionnellement indépendantes. De même, nous associons un poids w_{comb} au terme P^{comb} de l'équation 3.7.

L'utilisation de ces poids est justifiée du point de vue probabiliste [Han01]. Par exemple, dans le cas de dépendance simple où les valeurs de l'erreur résiduelle vérifient $\alpha_{jt,2f} = \alpha_{jt,2f+1}$ pour tout f, la probabilité de l'erreur en tenant compte de cette dépendance vaut $\prod_{f=0}^{(F-1)/2} \mathcal{N}(\alpha_{jt,2f}; 0, \sigma_j^\alpha) = (\prod_{f=0}^{F-1} \mathcal{N}(\alpha_{jtf}; 0, \sigma_j^\alpha))^{1/2}$. Il est alors équivalent de calculer la probabilité exacte (tenant compte de la dépendance) ou de pondérer la probabilité approchée (n'en tenant pas compte) par un poids exponentiel égal à 1/2.

Le rapport de poids $w_{\text{desc}}/w_{\text{spec}}$ peut jouer aussi le rôle d'un facteur de parcimonie. Plus il est élevé, moins l'erreur résiduelle a d'importance, et plus les transcriptions avec un faible nombre de notes présentes sont favorisées.

D'autres méthodes de transcription et de séparation utilisent des poids similaires avec une justification non probabiliste. Les algorithmes d'ASA computationnelle construisent le score associé à chaque hypothèse par pondération des différentes sources de connaissance en fonction des règles de groupement auditif [Wu03b, God99, Ell96]. Les coefficients de pondération peuvent dépendre de la scène analysée et de la tâche effectuée : la détection du rythme peut se passer entièrement du spectre et ne conserver que les variations d'énergie, mais pas l'identification de notes. De même, quelques algorithmes de décomposition parcimonieuse spectrale pondèrent les probabilités des coefficients de décomposition et de l'erreur résiduelle, et interprètent le poids correspondant comme un facteur de compromis entre parcimonie et erreur résiduelle [Vir03b, Ben03].

4.5 Algorithmes de transcription d'enregistrements solo

Nous proposons maintenant plusieurs algorithmes de transcription d'enregistrements solo monocanal adaptés aux modèles d'instruments choisis. Ces algorithmes sont également utilisés pour l'apprentissage dans le paragraphe 4.6, et adaptés à la transcription et la séparation des mélanges dans les paragraphes 6.1.3 et 7.1.3. Dans tout ce paragraphe, nous supposons que seul l'instrument $j = 1$ est présent sur le capteur $i = 1$ avec pour filtre de mélange un gain unitaire.

4.5.1 Choix des observations

Ajout de bruit de fond

Le signal d'un enregistrement solo est la somme du signal de l'instrument et du bruit de fond stationnaire. Même les enregistrements de bonne qualité contiennent un certain bruit de fond dû à la quantisation de la forme d'onde. De plus, le seuil de silence utilisé dans l'équation 2.6 fixe une valeur minimale du spectre de log-puissance observé qui peut être assimilée à la log-puissance d'un bruit de fond stationnaire virtuel. La modélisation explicite du bruit de fond est nécessaire pour éviter que celui-ci ne soit appris comme une caractéristique de l'instrument présent, alors qu'il ne dépend que des conditions d'enregistrement. Nous approchons cette somme de signaux par la somme de spectres de puissance

$$\mathbf{o}_t = \log(\mathbf{m}_{1,t} + \mathbf{n}_t), \tag{4.14}$$

où les observations $(\mathbf{o}_t)$ sont le spectre de log-puissance observé et $\mathbf{n}_t$ est le spectre de puissance du bruit de fond. Nous modélisons le bruit de fond comme un instrument à spectre stationnaire par

$$\log(\mathbf{n}_t) = \log(\mathbf{n}') + \boldsymbol{\beta}_t, \tag{4.15}$$

où $\boldsymbol{\beta}_t$ est un bruit gaussien à covariance diagonale isotrope d'écart-type σ^β. Nous supposons que les écarts-type du modèle de bruit de fond et du modèle d'instrument σ^β et σ_1^α sont proches d'un écart-type moyen σ^ϵ, et nous approchons les observations par

$$\mathbf{o}_t \approx \log(\mathbf{m}'_{1,t} + \mathbf{n}') + \boldsymbol{\epsilon}_t, \tag{4.16}$$

où $\boldsymbol{\epsilon}_t$ est un bruit gaussien à covariance diagonale isotrope d'écart-type σ^ϵ. La probabilité des observations s'exprime alors directement en fonction des descripteurs par

$$P(\mathbf{o}_t|\boldsymbol{\Theta}, \mathbf{p}_{1,t}, \mathcal{M}_1) \approx \prod_{f=0}^{F-1} \mathcal{N}(\epsilon_{tf}; 0, \sigma^\epsilon), \tag{4.17}$$

où les paramètres de mélange $\boldsymbol{\Theta}$ contiennent le spectre modèle du bruit de fond $\mathbf{n}'$.

Expression de la loi de Bayes pondérée

La transcription des enregistrements solo est réalisée par le critère MAP de l'équation 3.9. Si la probabilité *a priori* du bruit de fond $P(\mathbf{\Theta})$ est uniforme, ce critère équivaut à maximiser $P^{\text{trans}} = P(\mathbf{o}, \mathbf{p}_1, \mathbf{E}_1 | \, \mathbf{\Theta}, \mathcal{M}_1, \mathcal{I})$ conjointement par rapport aux états $\mathbf{E}_1$, aux descripteurs $\mathbf{p}_1$ et au bruit de fond $\mathbf{\Theta}$. La loi de Bayes pondérée s'écrit $P^{\text{trans}} = \prod_{t=0}^{T-1} P_t^{\text{trans}}$ avec

$$\begin{aligned} \log P_t^{\text{trans}} = & \, w_{\text{comb}} \log P(\mathbf{o}_t | \mathbf{\Theta}, \mathbf{p}_{1,t}, \mathcal{M}_1) \\ & + w_{\text{desc}} \log P(\mathbf{p}_{1,t} | \mathbf{E}_{1,t}, \mathcal{M}_1) \\ & + w_{\text{état}} \log P(\mathbf{E}_{1,t} | (\mathbf{E}_{1,t'})_{0 \leq t' \leq t-1}, \mathcal{M}_1, \mathcal{I}). \end{aligned} \tag{4.18}$$

Le premier terme de la somme est développé dans l'équation 4.17, le deuxième dans l'équation 4.5 et le troisième dans l'équation 4.6 ou 4.9 selon le modèle de couche d'état choisi.

4.5.2 Estimation des descripteurs et du bruit de fond

Supposons pour commencer que les états sont connus et attachons-nous à l'estimation des descripteurs et du bruit de fond. Les descripteurs des notes absentes étant contraints à des valeurs fixées, seuls les descripteurs des notes présentes doivent être estimés.

Initialisation et réestimation des descripteurs

Le gradient de $\log P_t^{\text{trans}}$ par rapport aux descripteurs des notes présentes au temps t s'écrit

$$\frac{\partial \log P_t^{\text{trans}}}{\partial e_{1,ht}} = \frac{w_{\text{comb}}}{{\sigma^\epsilon}^2} \sum_{f=0}^{F-1} \epsilon_{tf} \, \pi_{1,htf} - \frac{w_{\text{desc}}}{{\sigma_{1,h}^e}^2} (e_{1,ht} - \mu_{1,h}^e), \tag{4.19}$$

$$\frac{\partial \log P_t^{\text{trans}}}{\partial v_{1,ht}^k} = \frac{w_{\text{comb}}}{{\sigma^\epsilon}^2} \sum_{f=0}^{F-1} \epsilon_{tf} \, U_{1,hf}^k \, \pi_{1,htf} - \frac{w_{\text{desc}}}{{\sigma_{1,hk}^v}^2} (v_{1,ht}^k - \mu_{1,hk}^v), \tag{4.20}$$

où

$$\pi_{1,htf} = \frac{\exp(\Phi'_{1,htf}) \exp(e_{1,ht})}{m'_{1,tf} + n'_f}. \tag{4.21}$$

La quantité $\pi_{1,htf}$, comprise entre 0 et 1, représente la proportion de puissance due à la note h dans le spectre modèle au temps t $\mathbf{m}'_{1,t} + \mathbf{n}'$ à la fréquence f. Cette quantité s'interprète comme une quantité de masquage : $\pi_{1,htf} \approx 1$ si la note h masque les autres notes et le bruit de fond au point (t, f) et $\pi_{1,htf} \approx 0$ si au contraire la note h est masquée par les autres notes ou par le bruit de fond. La valeur de l'observation o_{tf} en un point (t, f) n'intervient pas pour déterminer les descripteurs d'une note h lorsqu'elle est masquée en ce point. Cette propriété remarquable se retrouve dans les algorithmes d'inférence à données manquantes (*missing data inference*) [Egg03], mais pas dans les algorithmes d'ASI linéaire [Abd01].

Il n'existe pas de formule analytique exacte permettant l'annulation du gradient de $\log P_t^{\text{trans}}$ par rapport à tous les descripteurs simultanément. Nous proposons donc un algorithme de Newton au second ordre approché [Hyv01a]. Les termes diagonaux et non-diagonaux de la hessienne de $\log P_t^{\text{trans}}$ par rapport aux log-puissances des notes présentes valent respectivement

$$\frac{\partial^2 \log P_t^{\text{trans}}}{\partial {e_{1,ht}}^2} = -\frac{w_{\text{comb}}}{{\sigma^\epsilon}^2} \sum_{f=0}^{F-1} \left(\pi_{1,htf}^2 - \epsilon_{tf} \frac{\partial \pi_{1,htf}}{\partial e_{1,ht}} \right) - \frac{w_{\text{desc}}}{{\sigma_{1,h}^e}^2}, \tag{4.22}$$

$$\frac{\partial^2 \log P_t^{\text{trans}}}{\partial e_{1,ht} \partial e_{1,h't}} = -\frac{w_{\text{comb}}}{{\sigma^\epsilon}^2} \sum_{f=0}^{F-1} \left(\pi_{1,htf} \, \pi_{1,h'tf} - \epsilon_{tf} \frac{\partial \pi_{1,htf}}{\partial e_{1,h't}} \right) \qquad \text{si } h' \neq h \tag{4.23}$$

Les quantités de la seconde équation $\pi_{1,htf}\,\pi_{1,h'tf}$ et $\partial\pi_{1,htf}/\partial e_{1,h't}$ sont non négligeables uniquement lorsque les notes h et h' sont de puissance similaire au point (t, f) et masquent le bruit de fond. Généralement, la plupart des partiels de deux notes de hauteur différente ont des fréquences différentes et les quelques partiels à des fréquences communes sont de puissances différentes (souvent plus élevée pour les partiels de la note la plus aiguë). Les termes non diagonaux de la hessienne peuvent donc être négligés et l'inverse de la hessienne peut être approché par l'inverse de sa diagonale calculable plus rapidement.

Pour les mêmes raisons, la quantité $\partial\pi_{1,htf}/\partial e_{1,ht}$ peut être négligée dans la première équation. Enfin, nous remplaçons $\pi_{1,htf}^2$ par $\pi_{1,htf}$ dans la première équation. Cela garantit que la différence entre les valeurs de $e_{1,ht}$ avant et après réestimation de Newton est bornée par $\max_f |\epsilon_{tf}| + |e_{1,ht} - \mu_{1,h}^e|$. Sinon l'équation de réestimation est instable au temps t lorsque $\pi_{1,htf} \approx 0$ pour tout f.

Les mêmes approximations sont choisies pour la hessienne par rapport aux descripteurs de variation.

L'algorithme obtenu contient les étapes suivantes :

1. Initialiser les descripteurs des notes présentes au temps t par

$$e_{1,ht} \leftarrow 2\log\Bigg(\sum_{f=0}^{F-1}\exp\Big(\frac{o_{tf}}{2}\Big)\exp\Big(\frac{\Phi_{1,hf}}{2}\Big)\Bigg) \qquad \text{et} \qquad v_{1,ht}^k \leftarrow 0 \tag{4.24}$$

   et calculer la valeur initiale de $\log P_t^{\text{trans}}$.

2. Réestimer les descripteurs par

$$e_{1,ht} \leftarrow e_{1,ht} + \frac{\dfrac{w_{\text{comb}}}{\sigma^{\epsilon\,2}}\displaystyle\sum_{f=0}^{F-1}\epsilon_{tf}\,\pi_{1,htf} - \dfrac{w_{\text{desc}}}{\sigma_{1,h}^{e\,2}}(e_{1,ht} - \mu_{1,h}^e)}{\dfrac{w_{\text{comb}}}{\sigma^{\epsilon\,2}}\displaystyle\sum_{f=0}^{F-1}\pi_{1,htf} + \dfrac{w_{\text{desc}}}{\sigma_{1,h}^{e\,2}}}, \tag{4.25}$$

$$v_{1,ht}^k \leftarrow v_{1,ht}^k + \frac{\dfrac{w_{\text{comb}}}{\sigma^{\epsilon\,2}}\displaystyle\sum_{f=0}^{F-1}\epsilon_{tf}\,U_{1,hf}^k\,\pi_{1,htf} - \dfrac{w_{\text{desc}}}{\sigma_{1,hk}^{v\,2}}(v_{1,ht}^k - \mu_{1,hk}^v)}{\dfrac{w_{\text{comb}}}{\sigma^{\epsilon\,2}}\displaystyle\sum_{f=0}^{F-1}U_{1,hf}^{k\,2}\,\pi_{1,htf} + \dfrac{w_{\text{desc}}}{\sigma_{1,hk}^{v\,2}}}. \tag{4.26}$$

3. Recalculer $\log P_t^{\text{trans}}$. Si la variation relative de $\log P_t^{\text{trans}}$ est positive et inférieure à un seuil l_{augm}, proposer les descripteurs actuels comme résultat. Sinon retourner à l'étape 2.

Cet algorithme converge vers un maximum local de $\log P_t^{\text{trans}}$. Il est possible de vérifier que les éventuels maxima locaux correspondent à des valeurs proches des descripteurs lorsque les spectres des notes présentes possèdent un faible recouvrement spectral. Dans le cas contraire, l'algorithme peut être sensible à l'initialisation des descripteurs. L'initialisation choisie dans l'équation 4.24 signifie que l'amplitude initiale correspondant à $e_{1,ht}$ est égale au produit scalaire du spectre d'amplitude observé et du spectre d'amplitude moyen normalisé de la note h.

Dans la suite, le seuil de convergence est fixé à $l_{\text{augm}} = 1\%$.

Initialisation et réestimation du bruit de fond

Le gradient de $\log P^{\text{trans}}$ par rapport au spectre de log-puissance modèle du bruit de fond vaut

$$\frac{\partial \log P^{\text{trans}}}{\partial \log n'_f} = \frac{w_{\text{comb}}}{\sigma^{\epsilon\,2}}\sum_{t=0}^{T-1}\epsilon_{tf}\,\pi'_{tf}, \tag{4.27}$$

où

$$\pi'_{tf} = \frac{n'_f}{m'_{1,tf} + n'_f} \tag{4.28}$$

représente la proportion de puissance due au bruit de fond dans le spectre modèle au point (t, f).

À partir des mêmes approximations de la hessienne que précédemment, nous proposons l'algorithme suivant :

1. Initialiser le spectre de log-puissance du modèle de bruit de fond à chaque fréquence par

$$\log n'_f \leftarrow \min_{0 \leq t \leq T-1} o_{tf} \tag{4.29}$$

   et le lisser par convolution avec une fenêtre fixée w_n. Calculer la valeur initiale de $\log P^{\text{trans}}$.

2. Réestimer le spectre par

$$\log n'_f \leftarrow \log n'_f + \frac{\sum_{t=0}^{T-1} \epsilon_{tf}\, \pi'_{tf}}{\sum_{t=0}^{T-1} \pi'_{tf} + c_{\text{bruit}}} \tag{4.30}$$

   et le lisser par convolution avec la fenêtre w_n.

3. Recalculer $\log P^{\text{trans}}$. Si l'augmentation relative de $\log P^{\text{trans}}$ est comprise entre $-l_{\text{augm}}$ et $+l_{\text{augm}}$, proposer le spectre actuel comme résultat. Sinon retourner à l'étape 2.

La constante de régularisation c_{bruit} apparaissant dans l'équation 4.30 vise à obtenir un algorithme stable à la fréquence f lorsque $\pi'_{tf} \approx 0$ pour tout t. Le lissage de l'étape 2 ajoute une contrainte de continuité spectrale pour éviter le sur-apprentissage. En particulier cela empêche le modèle de bruit de fond de représenter les pics spectraux dus aux partiels dans les enregistrements de courte durée où une seule note est tenue. Il est possible de lisser le gradient du spectre au lieu du spectre lui-même pour un résultat similaire. Cette contrainte ne permet plus de garantir que $\log P^{\text{trans}}$ augmente à chaque itération, mais expérimentalement l'algorithme converge bien. Nous aurions pu aussi implémenter une contrainte de continuité sous forme d'une distribution *a priori* $P(\boldsymbol{\Theta})$ non uniforme, mais dans ce cas le réglage du poids approprié w_Θ est assez délicat.

Dans la suite, la fenêtre de lissage w_n est une fenêtre de Hanning (de somme unitaire) de taille 15 et la constante de régularisation est fixée à $c_{\text{bruit}} = 0,1 \times T$ où T est la durée totale des observations.

4.5.3 Recherche dans l'espace des états

Supposons maintenant que les états doivent aussi être estimés. Pour cela, nous imbriquons les deux algorithmes ci-dessus à une procédure de recherche dans l'espace des états. Nous décrivons deux procédures différentes selon le modèle de couche d'état choisi.

Généralement, la taille de l'espace d'état est trop importante pour tester tous les états possibles. Par exemple, dans le cas d'une distribution de Bernoulli factorielle sur la couche d'état, le nombre d'états possibles atteint sur chaque trame $2^{46} \simeq 7,0 \times 10^{13}$ pour un violon pouvant jouer 46 notes et $\sum_{a=0}^{4} \binom{46}{a} \simeq 1,8 \times 10^5$ si le nombre maximum de notes simultanées est limité à quatre (ce nombre étant souvent atteint avec des notes rapides et une réverbération élevée). Le point commun des deux procédures décrites est d'utiliser des critères heuristiques pour limiter le nombre d'états testés. Le compromis entre temps de calcul (et taille en mémoire) et qualité de la transcription est fixé manuellement.

Algorithme de saut

Nous proposons d'effectuer la transcription avec un modèle de Bernoulli factoriel par un algorithme de saut itératif. Son principe est de commencer par tester des états très simples (état de silence), puis à chaque itération d'ajouter ou de retrancher des notes au meilleur état déjà testé sur chaque trame pour créer de nouveaux états à tester. Les états testés sont évalués séparément sur chaque trame t par leur log-probabilité $\log P_t^{\mathrm{trans}}$.

En sus des variables usuelles, cet algorithme définit sur chaque trame t l'ensemble des notes possibles $\mathcal{P}_t$, ainsi que l'ensemble des états testés $\mathfrak{T}_t$ et l'ensemble des états à réestimer $\mathfrak{R}_t$. Son fonctionnement est le suivant :

1. Initialiser le spectre de log-puissance du bruit de fond par l'équation 4.29 et le lisser par convolution avec la fenêtre w_n.
2. Sur chaque trame t,
   (a) définir l'ensemble des notes possibles $\mathcal{P}_t$ comme les N_{poss} notes h pour lesquelles la valeur initiale de $e_{1,ht}$, définie par l'équation 4.24, est la plus élevée ;
   (b) initialiser l'ensemble des états testés $\mathfrak{T}_t$ et l'ensemble des états à réestimer $\mathfrak{R}_t$ avec un état unique de silence ;
   (c) calculer la valeur initiale de log-probabilité correspondant à cet état de silence ;
   (d) calculer la log-probabilité totale $\log P_{\mathrm{meil}}$ du chemin formé des états de silence.
3. Pour chaque trame t, proposer de nouveaux états à tester de la façon suivante :
   (a) sélectionner le premier état de $\mathfrak{T}_t$ comme "père" ;
   (b) énumérer tous les états "fils" obtenus en retranchant une note ou en ajoutant une note de $\mathcal{P}_t$ au père ;
   (c) ajouter à $\mathfrak{T}_t$ et $\mathfrak{R}_t$ les fils nouveaux (qui ne se trouvent pas déjà dans $\mathfrak{T}_t$) ;
   (d) pour chaque fils nouveau, initialiser les descripteurs correspondants par ceux du père pour les notes conservées et par l'équation 4.24 pour la note ajoutée éventuelle, et calculer la valeur initiale de log-probabilité correspondante.
4. Sur chaque trame t,
   (a) réestimer les descripteurs de chaque état de $\mathfrak{R}_t$ par les équations 4.25 et 4.26 et recalculer la log-probabilité correspondante ;
   (b) ôter de $\mathfrak{R}_t$ les états pour lesquels la variation relative de log-probabilité est positive et inférieure à l_{augm} ;
   (c) classer les états de $\mathfrak{T}_t$ par ordre décroissant de log-probabilité ;
   (d) ôter de $\mathfrak{R}_t$ les états dont la position de classement est supérieure à un entier N_{dern}, ou dont la log-probabilité est inférieure à celle du premier état de $\mathfrak{T}_t$ moins un seuil l_{rel}.
5. Réestimer le spectre de log-puissance du modèle de bruit de fond par l'équation 4.30, en utilisant les quantités ϵ_{tf} et π'_{tf} correspondant au premier état de $\mathfrak{T}_t$ sur chaque trame t, et le lisser par convolution avec la fenêtre w_n.
6. Recalculer la log-probabilité totale $\log P_{\mathrm{meil}}$ du chemin formé du premier état de $\mathfrak{T}_t$ sur chaque trame t. Si la variation relative de $\log P_{\mathrm{meil}}$ est comprise entre $-l_{\mathrm{augm}}$ et $+l_{\mathrm{augm}}$, proposer ce chemin et les descripteurs correspondants comme résultat. Sinon retourner à l'étape 3.

Notons que cet algorithme n'estime pas les descripteurs correspondant à tous les états avant de proposer de nouveaux états à tester. En effet l'algorithme de Newton approché est assez coûteux, et il est préférable de ne l'utiliser jusqu'à convergence que sur les états les plus probables, qui ne peuvent pas être connus lors des premières itérations. Cette idée est utilisée par l'algorithme d'ASI proposé dans [Abd01]. Pour la même raison, tous les états testés de faible probabilité sont conservés pour éviter de les tester à nouveau.

Dans la suite, les seuils sont fixés à $l_{\mathrm{rel}} = \log(10^{-3})$ et $N_{\mathrm{dern}} = 100$. Dans les expériences d'identification d'instruments, toutes les notes sont considérées comme possibles ($N_{\mathrm{poss}} = N_1$), et dans les autres expériences nous utilisons $N_{\mathrm{poss}} = 20$.

Recherche en faisceaux

La transcription avec un modèle segmental est réalisée par un algorithme de recherche en faisceaux (*beam search*) [Ort96]. Il consiste à prolonger itérativement des chemins partiels, en ne conservant à chaque itération que les chemins les plus probables (*acoustic pruning*, *histogram pruning*). Un chemin partiel de longueur $t+1$ est évalué à l'aide de sa log-probabilité $\log P_t^{\mathrm{chem}} = \sum_{t'=0}^{t} \log P_{t'}^{\mathrm{trans}}$.

Ce type d'algorithme est généralement utilisé avec des MMC, et ses caractéristiques sont proches de l'architecture de tableau noir souvent utilisée en ASA computationnelle [Ell96]. Nous l'adaptons facilement au modèle segmental grâce à notre définition de probabilités de transition permettant de calculer la probabilité de n'importe quel chemin partiel.

Partant du principe que le bruit de fond ne peut pas être estimé sur les premières trames des observations uniquement, nous effectuons une transcription préalable des observations complètes par l'algorithme de saut décrit ci-dessus. La transcription par recherche en faisceaux utilise alors le bruit de fond ainsi trouvé sans modification, et tient compte des états estimés pour limiter le nombre d'états testés.

Nous utilisons également une méthode grossière de détection d'attaques inspirée de [Hai03] et modifiée à l'aide de peignes harmoniques. Pour chaque note h formée de partiels harmoniques, nous notons $\mathcal{F}_h$ l'ensemble des bandes de fréquence contenant les partiels (chaque partiel est attribué à la bande de fréquence centrale la plus proche et chaque bande est comptée une seule fois). Nous définissons la fonction $harm(.,h)$ qui associe à un spectre de log-puissance quelconque $\mathbf{\Omega}$ le spectre de log-puissance harmonique de hauteur h $harm(\mathbf{\Omega}, h)$ tel que $harm(\mathbf{\Omega}, h)_f = \mathbf{\Omega}_f$ pour tout $f \in \mathcal{F}_h$. Le calcul de cette fonction est décrit dans l'annexe B.1.3. Nous définissons alors le peigne harmonique $\mathbf{\Psi}_h$ associé par

$$\Psi_{hf} = \frac{\exp(l_{\mathrm{peig}})}{\exp(l_{\mathrm{peig}}) + \exp(harm(\mathbf{0}, h)_f)}, \tag{4.31}$$

où l_{peig} est un seuil de log-puissance fixé et $\mathbf{0}$ est le spectre de log-puissance plat décrit par le vecteur nul. Pour les percussions, la notion de hauteur n'a pas de sens et le peigne peut être fixé à $\Psi_{hf} = 1$ pour tout f.

L'algorithme utilise deux nouveaux ensembles au temps t : l'ensemble des attaques de notes possibles $\mathcal{P}'_t$ et l'ensemble des chemins partiels testés et conservés $\mathfrak{C}_t$. Il est formé des étapes suivantes :

1. En utilisant l'algorithme de saut, estimer le bruit de fond $\widehat{\mathbf{n}'}$ et les états $\widehat{\mathbf{E}_1}$. Sur chaque trame t, définir l'ensemble des notes possibles $\mathcal{P}_t$ comme les notes h vérifiant $\widehat{E_{1,ht'}} = 1$ pour au moins une valeur de t' avec $t \leq t' \leq t + t_{\mathrm{att}}$.
2. Pour chaque note h,
   (a) calculer grossièrement la variation temporelle positive de log-puissance associée par

$$\delta_{ht} = \sum_{f=0}^{F-1} \Psi_{hf} \, \max(o_{t+1,f} - o_{tf}, 0) \tag{4.32}$$

   pour $0 \leq t \leq T-2$ et $\delta_{h,T-1} = 0$;

(b) lisser cette variation par convolution avec une fenêtre fixée w_{att}.

3. Sur chaque trame t, définir l'ensemble des attaques de notes possibles $\mathcal{P}'_t$ comme les notes présentes dans $\mathcal{P}_t$ pour lesquelles $(\delta_{ht})_{0 \leq t \leq T-1}$ admet un maximum local en t et δ_{ht} est supérieure à sa moyenne temporelle.
4. Sur la trame $t = 0$,
   (a) définir $\mathfrak{C}_0$ comme l'ensemble des états (ou chemins partiels de longueur 1) formés de silence ou d'une note de $\mathcal{P}_0$;
   (b) pour chaque état, initialiser les descripteurs par l'équation 4.24 et les réestimer par les équations 4.25 et 4.26 jusqu'à ce que la variation relative de $\log P_0^{\text{trans}}$ soit positive et inférieure à l_{augm}, puis conserver $\log P_0^{\text{chem}} = \log P_0^{\text{trans}}$;
5. Sur chaque trame $t \geq 1$,
   (a) énumérer tous les chemins partiels de longueur $t+1$ obtenus en prolongeant au temps t les chemins partiels de $\mathfrak{C}_{t-1}$ par un état identique à l'état au temps $t-1$, avec une note en moins par rapport à cet état ou avec une note de $\mathcal{P}'_t$ en plus par rapport à cet état ;
   (b) pour chaque chemin partiel, initialiser les descripteurs au temps t par l'équation 4.24 et les réestimer par les équations 4.25 et 4.26 jusqu'à ce que la variation relative de $\log P_t^{\text{trans}}$ soit positive et inférieure à l_{augm}, puis calculer $\log P_t^{\text{chem}} = \log P_{t-1}^{\text{chem}} + \log P_t^{\text{trans}}$;
   (c) lorsque plusieurs chemins partiels aboutissent sur le même état au temps t, conserver uniquement le meilleur chemin ;
   (d) classer les chemins partiels par ordre décroissant de log-probabilité ;
   (e) définir $\mathfrak{C}_t$ comme l'ensemble des chemins partiels dont la position de classement est inférieure à un entier N_{dern} et dont la log-probabilité est supérieure à celle du meilleur chemin partiel moins un seuil l_{rel}, ou bien dont la durée écoulée sur le segment en cours est inférieure strictement à un seuil $d_{\min}$.
6. Proposer le meilleur chemin global de $\mathfrak{C}_{T-1}$ et les descripteurs correspondants comme résultat.

L'étape 5a limite le nombre de transitions testées par rapport aux transitions possibles en pratique. Lorsque plusieurs chemins partiels de $\mathfrak{C}_t$ aboutissent sur le même état au temps t, l'étape 5b est effectuée une seule fois pour chaque état concerné. Avec un modèle de couche d'état par chaîne de Markov factorielle, la sélection à l'étape 5c du meilleur chemin partiel aboutissant sur un état donné correspondrait à l'algorithme de Viterbi [Rab89]. Avec un modèle segmental, cette sélection peut supprimer des hypothèses qui pourraient se révéler plus probables que les hypothèses conservées. Néanmoins, elle sélectionne uniquement les instants de début des segments, et pas les notes présentes dans les segments. L'utilisation d'un seuil de durée minimale dans l'étape 5e assure que chaque hypothèse de notes est conservée pendant un certain nombre de trames.

Dans la suite nous choisissons pour w_{att} une fenêtre de Hanning (de somme unitaire) de taille 5 et nous fixons $t_{\text{att}} = 10$ et $l_{\text{peig}} = -6$. La détection d'attaques dans l'étape 3 diminue alors le nombre d'attaques testées d'un facteur 10 environ, tout en surestimant le nombre d'attaques effectivement présentes pour éviter les erreurs de suppression de notes. Les paramètres de limitation de la taille de la recherche en faisceaux proprement dite sont fixés à $d_{\min} = 20$, $N_{\text{dern}} = 100$ et $l_{\text{rel}} = 10^{-6}$.

4.6 Algorithmes d'apprentissage des paramètres du modèle

Nous complétons dans ce paragraphe la liste d'algorithmes de transcription proposés par des algorithmes d'apprentissage des paramètres du modèle d'un instrument donné sur une base de donnés de notes isolées. À nouveau, nous supposons que seul l'instrument $j = 1$ est présent sur le capteur $i = 1$

avec pour filtre de mélange un gain unitaire. Par simplicité, nous décrivons la base de données par une suite d'observations unique $\mathbf{o}$ étiquetée par les états correspondant aux notes $\mathbf{E}_1$. Nous supposons qu'il existe des zones de silence dans les observations, de sorte que le spectre du modèle de bruit de fond $\mathbf{n}'$ est connu.

4.6.1 Apprentissage des paramètres spécifiques

Les paramètres de la couche spectrale et de la couche descriptive peuvent être appris avec le critère MV de l'équation 3.19, en supposant leur probabilité *a priori* $P(\mathcal{M})$ uniforme. Ce critère équivaut à maximiser la probabilité de transcription P^{trans}, définie par la loi de Bayes pondérée de l'équation 4.18, par rapport aux descripteurs et aux paramètres.

Réestimation

Le gradient de $\log P^{\text{trans}}$ par rapport aux paramètres vaut

$$\frac{\partial \log P^{\text{trans}}}{\partial \Phi_{1,hf}} = \frac{w_{\text{comb}}}{\sigma^{\epsilon\,2}} \sum_{t=0}^{T-1} \epsilon_{tf}\, \pi_{1,htf}, \tag{4.33}$$

$$\frac{\partial \log P^{\text{trans}}}{\partial U^k_{1,hf}} = \frac{w_{\text{comb}}}{\sigma^{\epsilon\,2}} \sum_{t=0}^{T-1} \epsilon_{tf}\, v^k_{1,ht}\, \pi_{1,htf}, \tag{4.34}$$

$$\frac{\partial \log P^{\text{trans}}}{\partial \mu^e_{1,h}} = \frac{w_{\text{desc}}}{\sigma^{e\,2}_{1,h}} \sum_{t=0}^{T-1} (e_{1,ht} - \mu^e_{1,h}), \tag{4.35}$$

$$\frac{\partial \log P^{\text{trans}}}{\partial \log \sigma^e_{1,h}} = \frac{w_{\text{desc}}}{\sigma^{e\,2}_{1,h}} \sum_{t=0}^{T-1} (e_{1,ht} - \mu^e_{1,h})^2 - w_{\text{desc}}\, T, \tag{4.36}$$

$$\frac{\partial \log P^{\text{trans}}}{\partial \mu^v_{1,hk}} = \frac{w_{\text{desc}}}{\sigma^{v\,2}_{1,hk}} \sum_{t=0}^{T-1} (v^k_{1,ht} - \mu^v_{1,hk}), \tag{4.37}$$

$$\frac{\partial \log P^{\text{trans}}}{\partial \log \sigma^v_{1,hk}} = \frac{w_{\text{desc}}}{\sigma^{v\,2}_{1,hk}} \sum_{t=0}^{T-1} (v^k_{1,ht} - \mu^v_{1,hk})^2 - w_{\text{desc}}\, T, \tag{4.38}$$

où $\pi_{1,htf}$ est défini par l'équation 4.21.

Nous adoptons à nouveau une maximisation par algorithme de Newton du second ordre approché, qui mène à l'algorithme EM approché suivant :

1. Fixer la tessiture de l'instrument (paramètres H_1 et H'_1) et le nombre de composantes de variation K_1. Pour chaque note h, initialiser $\mathbf{\Phi}_{1,h}$, $(\mathbf{U}^k_{1,h})$, $\mu^e_{1,h}$, $\sigma^e_{1,h}$, $(\mu^v_{1,hk})$ et $(\sigma^v_{1,hk})$.
2. Sur chaque trame t, initialiser les descripteurs par l'équation 4.24 et les réestimer par les équations 4.25 et 4.26 jusqu'à ce que la variation relative de $\log P^{\text{trans}}_t$ soit positive et inférieure à l_{augm}. Calculer $\log P^{\text{trans}}$.
3. Réestimer les paramètres de la couche spectrale de la façon suivante :
   (a) pour chaque note h, calculer les variations $\Delta\mathbf{\Phi}_{1,h}$ et $\Delta\mathbf{U}^k_{1,h}$ définies par

$$\Delta\Phi_{1,hf} \leftarrow \frac{\displaystyle\sum_{t=0}^{T-1} \epsilon_{tf}\, \pi_{1,htf}}{\displaystyle\sum_{t=0}^{T-1} \pi_{1,htf} + c_{\text{appr}}}, \tag{4.39}$$

$$\Delta U^k_{1,hf} \leftarrow \frac{\sum_{t=0}^{T-1} \epsilon_{tf}\, v^k_{1,ht}\, \pi_{1,htf}}{\sum_{t=0}^{T-1} v^{k\,2}_{1,ht}\, \pi_{1,htf} + c_{\mathrm{appr}}(\mu^{v\,2}_{1,hk} + \sigma^{v\,2}_{1,hk})}; \tag{4.40}$$

(b) lisser ces variations par $\Delta\mathbf{\Phi}_{1,h} \leftarrow \sum_{h'=H_1}^{H'_1} trans(\Delta\mathbf{\Phi}_{1,h'}, h-h')w_\Phi(h'-h)$ et $\Delta\mathbf{U}^k_{1,h} \leftarrow \sum_{h'=H_1}^{H'_1} trans(\Delta\mathbf{U}^k_{1,h'}, h-h')w_U(h'-h)$, où $trans(.,h)$ est la fonction (calculée par interpolation linéaire) qui transpose un spectre de h demi-tons et w_Φ et w_U sont des fenêtres telles que $\sum_{h'=H_1}^{H'_1} w_\Phi(h'-h) = 1$ et $\sum_{h'=H_1}^{H'_1} w_U(h'-h) = 1$;

(c) réestimer les spectres par $\mathbf{\Phi}_{1,h} \leftarrow \mathbf{\Phi}_{1,h} + \Delta\mathbf{\Phi}_{1,h}$ et $\mathbf{U}^k_{1,h} \leftarrow \mathbf{U}^k_{1,h} + \Delta\mathbf{U}^k_{1,h}$;

(d) pour chaque note h, normaliser les spectres par $\Phi_{1,hf} \leftarrow \Phi_{1,hf} - \log(\sum_{f=0}^{F-1} \exp(\Phi_{1,hf}))$ et $\mathbf{U}^k_{1,h} \leftarrow \mathbf{U}^k_{1,h}/\|\mathbf{U}^k_{1,h}\|$ pour tout k.

4. Réestimer les paramètres de la couche descriptive de la façon suivante :

(a) pour chaque note h, calculer

$$\mu^e_{1,h} \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} e_{1,ht}, \tag{4.41}$$

$$\sigma^e_{1,h} \leftarrow \left(\frac{1}{T}\sum_{t=0}^{T-1} e^2_{1,ht} - \left(\frac{1}{T}\sum_{t=0}^{T-1} e_{1,ht}\right)^2\right)^{1/2}, \tag{4.42}$$

$$\mu^v_{1,hk} \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} v^k_{1,ht}, \tag{4.43}$$

$$\sigma^v_{1,hk} \leftarrow \left(\frac{1}{T}\sum_{t=0}^{T-1} v^{k\,2}_{1,ht} - \left(\frac{1}{T}\sum_{t=0}^{T-1} v^k_{1,ht}\right)^2\right)^{1/2}; \tag{4.44}$$

(b) lisser $(\mu^e_{1,h})_{H_1\le h\le H'_1}$ et $(\log\sigma^e_{1,h})_{H_1\le h\le H'_1}$ par convolution avec une fenêtre $w_{\mu\sigma}$; lisser de même $(\mu^v_{1,hk})_{H_1\le h\le H'_1}$ et $(\log\sigma^v_{1,hk})_{H_1\le h\le H'_1}$ pour tout k par convolution avec $w_{\mu\sigma}$.

5. Recalculer $\log P^{\mathrm{trans}}$. Si l'augmentation relative de $\log P^{\mathrm{trans}}$ est comprise entre $-l_{\mathrm{augm}}$ et $+l_{\mathrm{augm}}$, proposer les paramètres actuels comme résultat. Sinon retourner à l'étape 2.

Les équations 4.41 à 4.44 sont les équations usuelles d'estimation de la moyenne et de l'écart-type d'une distribution gaussienne. Si nécessaire, les écarts-types peuvent être contraints à une valeur minimale strictement positive. La constante de régularisation c_{appr} stabilise la réestimation des spectres typiques de la note h à la fréquence f lorsque $\pi_{1,htf} \approx 0$ pour tout t. Ces spectres sont normalisés explicitement dans l'étape 3c, comme dans les algorithmes d'ACI usuels [Hyv01a, Abd01]. Le surapprentissage est limité par le lissage des paramètres de notes de hauteurs voisines dans les étapes 3b et 4b. La normalisation et le lissage empêchent de garantir que $\log P^{\mathrm{trans}}$ augmente à chaque itération, mais expérimentalement l'algorithme converge bien.

Dans la suite, nous avons utilisé comme constante de régularisation $c_{\mathrm{appr}} = 10$ et comme fenêtres de lissage des fenêtres de Hanning de taille 7 pour w_Φ, 15 pour w_U et 25 pour $w_{\mu\sigma}$.

Lorsque $K_1 > 0$ les paramètres ne sont pas estimés de façon unique. Ce serait le cas par exemple avec deux contraintes supplémentaires : fixer les moyennes des descripteurs de variation $(\mu^v_{1,hk})$ à zéro, et remplacer le critère MV lorsque $K_1 > 1$ par un critère de type ACP [Hyv01b] (sinon une permutation

de l'indice k ne change pas la probabilité du modèle). Cependant, cette non unicité n'a pas d'incidence sur la qualité du modèle et la non nullité des $(\mu^v_{1,hk})$ accélère l'apprentissage.

Notons que l'écart-type σ^ϵ est fixé durant tout l'apprentissage. Sa valeur n'est pas caractéristique de l'instrument, car elle est un compromis entre l'écart-type du modèle d'instrument σ^α_1 et celui du modèle de bruit σ^β. Dans le reste de cette étude, les valeurs des (σ^α_j) seront implicitement fixées à la même valeur pour tous les instruments j et ne seront pas utilisées.

Notons aussi que l'approximation du critère MV exact de l'équation 3.18 par le critère approché de l'équation 3.19 engendre un léger biais sur les valeurs des écarts-types $(\sigma^e_{1,h})$ et $(\sigma^v_{1,hk})$, qui tend à diminuer leur valeur. Dans le cas où $K_1 = 0$ et où la puissance du bruit de fond est nulle, il est possible de montrer que les valeurs apprises de $(\sigma^e_{1,h})$ sont identiques pour les deux critères en choisissant w_{comb} deux fois plus grand pour le critère approché ou bien $\sigma^{\epsilon\,2}$ deux fois plus faible.

Initialisation

Le résultat de l'algorithme EM approché dépend de l'initialisation des paramètres dans l'étape 1. Puisque l'algorithme utilise des procédures de lissage, il est nécessaire que les paramètres de notes de hauteurs voisines soient voisins dès l'initialisation.

D'une part, nous fixons des valeurs initiales identiques pour toutes les notes h aux paramètres de la couche descriptive : $\mu^e_{jh} = 15$, $\sigma^e_{jh} = 5$, $\mu^v_{jhk} = 0$ et $\sigma^v_{jhk} = 20$.

D'autre part, dans le cas d'instruments harmoniques, nous proposons d'initialiser les paramètres de la couche spectrale $(\mathbf{\Phi}_{1,h})$ et $(\mathbf{U}^k_{1,h})$ par des spectres harmoniques $(\mathbf{\Phi}^{\text{harm}}_{1,h})$ et par des spectres représentant la variation de log-puissance des partiels harmoniques $(\mathbf{U}^{\text{part}}_{1,h})$, la variation de fréquence fondamentale $(\mathbf{U}^{\text{fréq}}_{1,h})$, et la variation de log-puissance du spectre entre les partiels harmoniques $(\mathbf{U}^{\text{bruit}}_{1,h})$.

Ces spectres initiaux sont extraits des observations d'apprentissage en utilisant les fonctions $harm$ et $trans$ définies dans le paragraphe 4.5.3 et ci-dessus. Nous séparons les observations $\mathbf{o}$ en observations disjointes $(\mathbf{o}_h)_{H_1 \leq h \leq H'_1}$ correspondant aux différentes notes, et nous leur associons des masques binaires $(\xi_h)_{H_1 \leq h \leq H'_1}$ de masquage du bruit de fond, définis par $\xi_{htf} = 1$ si $o_{htf} > \log(2\, n'_f)$ et 0 sinon. L'algorithme d'initialisation fonctionne comme suit :

1. Calculer de nouvelles observations $\mathbf{o}^{\text{harm}}_h$ par normalisation à chaque instant : $o^{\text{harm}}_{htf} = o_{htf} - \log(\sum_{f=0}^{F-1} \exp(o_{htf}))$. Définir $\mathbf{\Phi}^{\text{harm}}_{1,h}$ comme la moyenne de ces observations hors zones masquées par $\Phi^{\text{harm}}_{1,hf} \leftarrow \sum_{t=0}^{T-1} o^{\text{harm}}_{htf} \xi_{htf} / \sum_{t=0}^{T-1} \xi_{htf}$. Contraindre les valeurs de $\Phi^{\text{harm}}_{1,hf}$ dans les sous-bandes ne contenant pas de partiels en posant

$$\mathbf{\Phi}^{\text{harm}}_{1,h} \leftarrow harm(\mathbf{\Phi}^{\text{harm}}_{1,h}, h). \tag{4.45}$$

2. Calculer de nouvelles observations $\mathbf{o}^{\text{part}}_h$ par centrage de $\mathbf{o}^{\text{harm}}_h$ et annulation des sous-bandes de ne contenant pas de partiels : $o^{\text{part}}_{htf} = o^{\text{harm}}_{htf} - \Phi^{\text{harm}}_{1,hf}$ si $f \in \mathcal{F}_h$ et $o^{\text{part}}_{htf} = 0$ sinon. Définir $\mathbf{U}^{\text{part}}_{1,h}$ comme la première composante de l'ACP des observations masquées $(o^{\text{part}}_{htf}\, \xi_{htf})_{0 \leq f \leq F-1,\, 0 \leq t \leq T-1}$. Contraindre les valeurs de $U^{\text{part}}_{1,hf}$ dans les sous-bandes ne contenant pas de partiels en posant

$$\mathbf{U}^{\text{part}}_{1,h} \leftarrow \frac{d\, harm(\mathbf{\Phi}_{1,h} + v\mathbf{U}^{\text{part}}_{1,h}, h)}{d\, v}. \tag{4.46}$$

   Si $h > H_1$, poser éventuellement $\mathbf{U}^{\text{part}}_{1,h} \leftarrow -\mathbf{U}^{\text{part}}_{1,h}$ de sorte que $\langle \mathbf{U}^{\text{part}}_{1,h}, \mathbf{U}^{\text{part}}_{1,h-1} \rangle > 0$.

3. Poser

$$\mathbf{U}^{\text{fréq}}_{1,h} \leftarrow \frac{d\, trans(\mathbf{\Phi}^{\text{harm}}_{1,h}, h)}{d\, h}. \tag{4.47}$$

4. Calculer de nouvelles observations $\mathbf{o}_h^{\text{bruit}}$ par soustraction de la partie harmonique à chaque instant et annulation des sous-bandes de fréquence inférieure à la fréquence fondamentale f_h : $o_{htf}^{\text{bruit}} = o_{htf} - harm(\mathbf{o}_{ht}, h)_f$ si $f > f_h$ et $o_{htf}^{\text{bruit}} = 0$ sinon. Définir $\mathbf{U}_{1,h}^{\text{bruit}}$ comme la première composante de l'ACP des observations masquées $(o_{htf}^{\text{bruit}}\, \xi_{htf})_{0 \leq f \leq F-1,\, 0 \leq t \leq T-1}$. Si $h > H_1$, poser éventuellement $\mathbf{U}_{1,h}^{\text{bruit}} \leftarrow -\mathbf{U}_{1,h}^{\text{bruit}}$ de sorte que $\langle \mathbf{U}_{1,h}^{\text{bruit}}, \mathbf{U}_{1,h-1}^{\text{bruit}} \rangle > 0$.
5. Normaliser et lisser $(\mathbf{\Phi}_{1,h}^{\text{harm}})$, $(\mathbf{U}_{1,h}^{\text{part}})$, $(\mathbf{U}_{1,h}^{\text{fréq}})$ et $(\mathbf{U}_{1,h}^{\text{bruit}})$ comme dans les étapes 3b et 3c de l'algorithme d'apprentissage.

Des exemples de spectres initiaux seront donnés dans le paragraphe 5.1. La meilleure initialisation à partir de ces spectres sera sélectionnée par validation croisée dans le paragraphe 5.3.1.

Apprentissage contraint

Il est intéressant d'essayer de conserver la structure de ces spectres initiaux durant l'apprentissage. D'une part, cela permet d'apprendre uniquement les caractéristiques des instruments utiles pour leur identification, qui ne sont pas forcément les caractéristiques qui maximisent la vraisemblance des données d'apprentissage. D'autre part, cela attribue un sens physique aux descripteurs de variation. Les descripteurs correspondant à $\mathbf{U}_{1,h}^{\text{fréq}}$ sont liés à la fréquence instantanée et peuvent permettre de localiser les notes jouées *vibrato*. De même, les descripteurs correspondant à $\mathbf{U}_{1,h}^{\text{part}}$ sont liés à la brillance du timbre et donnent des indications sur la nuance de jeu (*forte* ou *piano*) (en effet la log-puissance des notes dépend du volume d'enregistrement et ne détermine la nuance qu'à une constante près).

Malheureusement, les structures des spectres initiaux sont incompatibles avec notre définition de l'erreur résiduelle, et il n'est pas possible d'apprendre les paramètres de la couche spectrale au sens du critère MV en conservant ces structures. Par exemple, dans le cas où $K_1 = 0$ et où une part d'énergie non harmonique (due à la réverbération de l'instrument et pas au bruit de fond) est présente entre les partiels sur l'ensemble d'apprentissage, les spectres harmoniques $(\mathbf{\Phi}_{1,h})$ minimisant l'erreur résiduelle ne sont pas les mêmes que les spectres harmoniques initiaux $(\mathbf{\Phi}_{1,h}^{\text{harm}})$, car ils déterminent la puissance des partiels aussi en fonction des zones de bruit. Les spectres initiaux, qui minimisent approximativement l'erreur résiduelle dans les sous-bandes contenant des partiels uniquement, ont plus de sens physiquement et caractérisent mieux l'instrument.

Lorsque $K_1 = 0$, une solution partielle à ce problème consiste à introduire la contrainte de l'équation 4.45 entre les étapes 3a et 3b de l'algorithme EM approché, c'est-à-dire après les équations de réestimation non contraintes (au lieu d'introduire la contrainte dans les équations de réestimation). Cela permet d'apprendre les spectres harmoniques $(\mathbf{\Phi}_{1,h})$ uniquement en fonction des puissances des partiels. Cependant, cette solution pose des problèmes de convergence de l'algorithme et n'est plus valable pour $K_1 > 0$. La seule véritable solution consisterait à utiliser une autre définition de l'erreur résiduelle basée sur une distance entre pics spectraux [Car92].

Par simplicité, nous proposons de construire les modèles d'instruments contraints en conservant les spectres initiaux et en apprenant par l'algorithme EM approché uniquement les paramètres de la couche descriptive. La comparaison entre ces modèles et les modèles non contraints sera réalisée par validation croisée dans le paragraphe 5.3.1.

4.6.2 Choix des autres paramètres

Les paramètres de la couche d'état ne peuvent être appris comme ceux des autres couches. Dans une base de données de notes isolées, la durée écoulée entre deux notes successives n'a pas de sens et la durée des notes est généralement supérieure à la durée des notes dans un enregistrement réel pour les instruments à oscillations entretenues. De plus, ces durées ne sont pas caractéristiques des instruments. Nous

choisissons donc de fixer manuellement les paramètres des distributions de durée aux mêmes valeurs pour tous les instruments.

Les poids utilisés dans l'équation de Bayes pondérée ne peuvent être appris par le critère MV. De plus, le poids w_{comb} dépend de l'importance du bruit de fond au sein des observations. Il est généralement un peu plus élevé sur des extraits solo que sur des notes isolées, car le masquage entre notes rend les valeurs de spectre observées plus indépendantes conditionnellement aux descripteurs. Il est possible d'apprendre les poids optimaux sur un mélange donné par un critère MIM approché (voir paragraphe 3.5.3), mais un choix manuel se révèle aussi efficace.

4.7 Spectre à court terme adapté à la transcription et la séparation

La performance du modèle pour la transcription et la séparation dépend du type de spectre à court terme choisi. Le spectre est défini par les paramètres du banc de filtres (échelle fréquentielle, nombre de sous-bandes, largeur de bande) et du découpage temporel (longueur des fenêtres, pas entre fenêtres successives) et par le seuil de silence (voir paragraphe 2.2.2). Nous discutons dans ce paragraphe les arguments guidant le choix de ces paramètres. Les paramètres utilisés par la suite sont présentés dans l'annexe B.1.

4.7.1 Détection des hauteurs de notes et qualité de la modélisation

Dans le cas d'instruments harmoniques, la performance pour la transcription dépend avant tout de la possibilité d'identifier les hauteurs des notes présentes dans un mélange. Pour cela, il faut pouvoir détecter plusieurs partiels de chaque note et déterminer précisément leur fréquence à partir du spectre du mélange et des différences de volume et de phase inter-canal. La détection de tous les partiels n'est pas nécessaire, et elle n'est pas toujours possible. Les partiels aigus sont généralement de faible amplitude, et souvent masqués par des partiels d'autres notes dans des accords. Il faut donc s'assurer que la localisation des premiers partiels permet toujours de discriminer des notes de hauteurs semblables.

Pour cela, la largeur de bande doit être inférieure à un demi-ton dans la zone de moyenne et basse fréquence correspondant aux hauteurs de notes usuelles. Une largeur de bande trop faible engendre cependant une faible résolution temporelle, qui empêche de tenir compte des instants d'attaque des partiels pour les regrouper en notes.

De plus, il est souhaitable de limiter le nombre de sous-bandes pour éviter le sur-apprentissage des modèles ou l'utilisation de procédures de partage de paramètres complexes.

L'identification d'instruments harmoniques ne nécessite pas de connaître précisément les puissances des partiels aigus des notes. La puissance moyenne fournie par une enveloppe spectrale grossière semble suffire [Egg03, Bro01, Mar99b]. De même, l'identification de percussions est basée sur leurs formants situés en moyenne et basse fréquence. Il est donc possible de choisir une largeur de bande de l'ordre du demi-ton ou plus en haute fréquence et une largeur constante en très basse fréquence, ces deux zones contenant moins d'information utile. Une largeur de bande importante en haute fréquence permet de plus une meilleure représentation de la fréquence instantanée des notes par l'équation 4.2, puisque des petites variations de fréquence engendrent des petites variations du spectre.

Une fois la largeur de bande fixée à chaque fréquence, le nombre de sous-bandes dépend de l'écart de fréquence entre sous-bandes voisines. Généralement, le recouvrement entre sous-bandes (rapport entre largeur de bande et écart de fréquence) est choisi constant de l'ordre de 2 à 4. Un recouvrement trop faible ou trop élevé engendre une perte d'information.

Enfin, la longueur et le pas des fenêtres de découpage en trames doivent éviter la perte d'information tout en limitant la durée de calcul. Une longueur de quelques dizaines de millisecondes forme un bon

compromis.

4.7.2 Parcimonie des sources et inversibilité

L'utilisation du modèle pour la séparation comprend une étape de transcription et une étape de filtrage. La performance du modèle pour la séparation dépend donc de la performance de ces deux étapes.

Les algorithmes décrits dans les paragraphes 6.2 et 7.2 effectuent le filtrage par application d'un simple gain en chaque point temps-fréquence (sauf pour les mélanges stéréo panoramiques), puis inversion du banc de filtres. Dans ce cas, la séparation est d'autant meilleure que les sources sont disjointes dans le plan temps-fréquence. La séparation de partiels de même fréquence reste irréalisable sans resynthèse sinusoïdale des partiels cachés.

À nouveau, la largeur de bande doit être inférieure à un demi-ton en moyenne et basse fréquence pour séparer les premiers partiels de notes de hauteurs semblables. Par contre, il n'est pas indispensable de séparer exactement tous les partiels aigus. L'estimation des sources dans des bandes larges de deux tons suffit, car cela correspond à la résolution fréquentielle de l'audition en haute fréquence [Rom03]. De plus, une largeur de bande trop faible nuit à la qualité de l'inversion du banc de filtres. Un étalement temporel des transitoires d'attaque peut se produire lorsque la taille des filtres est supérieure à une cinquantaine de millisecondes. Des contraintes semblables existent dans le cadre du débruitage [Cap93].

Fixer le recouvrement entre sous-bandes à une valeur suffisamment élevée permet de compenser les erreurs de filtrage dans une sous-bande en moyennant avec les sous-bandes voisines lors de l'inversion du banc de filtres [Cap93].

Chapitre 5

Exemples d'instruments et transcription d'enregistrements solo

Le but de ce cinquième chapitre est de valider nos hypothèses de modélisation sur des enregistrements solo. Nous présentons brièvement les instruments appris dans le paragraphe 5.1. Puis nous comparons les hypothèses de modélisation à la distribution empirique des observations d'apprentissage dans le paragraphe 5.2. Enfin nous donnons des exemples de transcription d'enregistrements solo dans le paragraphe 5.3. L'expérience d'identification d'instruments nous conduit en particulier à sélectionner les modèles utilisés par la suite parmi tous les modèles appris.

Les résultats d'identification d'instruments sur des extraits solo ont été diffusés en partie dans l'article [Vin04b].

5.1 Présentation des instruments appris

Faute de disposer d'une quantité suffisante d'enregistrements de percussions, nous avons dû nous limiter à modéliser des instruments harmoniques. Pour comparer nos résultats aux méthodes existantes d'identification d'instruments, nous avons alors appris les cinq instruments utilisés dans [Egg03] : flûte (Fl), clarinette (Cl), hautbois (Ob), violon (Vn) et violoncelle (Vc). Les données d'apprentissage sont décrites dans l'annexe A.1.

Plusieurs initialisations différentes ont été effectuées en variant le nombre de composantes de variation K_j, et avec ou sans contrainte sur les spectres $(\mathbf{\Phi}_{jh})$ et $(\mathbf{U}_{jh}^k)$. Nous avons utilisé $\sigma^\epsilon = 1$, $w_{\mathrm{comb}} = 1$ et $w_{\mathrm{desc}} = 1$. Après apprentissage, nous avons contraint les paramètres des composantes modélisant les variations de fréquence à $\mu_{jhk}^v = 0$ et $\sigma_{jhk}^v = 0,005 \times \|d\,trans(\mathbf{\Phi}_{jh}^{\mathrm{harm}}, h)/d\,h\|$, de sorte que les fréquences fondamentales de notes de même hauteur d'instruments différents aient *a priori* la même moyenne et le même écart-type. Ceci évite que les diapasons utilisés dans l'ensemble d'apprentissage soient considérés comme caractéristiques des instruments.

Les figures 5.1 à 5.5 montrent les spectres initiaux $(\mathbf{\Phi}_{jh}^{\mathrm{harm}})$, $(\mathbf{U}_{jh}^{\mathrm{part}})$, $(\mathbf{U}_{jh}^{\mathrm{fréq}})$ et $(\mathbf{U}_{jh}^{\mathrm{bruit}})$ pour chaque note de chaque instrument. La figure 5.6 dessine le centroïde spectral $\overline{f}$ et l'écartement spectral $\widetilde{f}$ correspondant à chaque spectre $\mathbf{\Phi}_{jh}^{\mathrm{harm}}$, calculés par [Mar03]

$$\overline{f} = \frac{\sum_{f=0}^{F-1} f \exp(\Phi_{jhf})^{0,3}}{\sum_{f=0}^{F-1} \exp(\Phi_{jhf})^{0,3}}, \tag{5.1}$$

$$\widetilde{f} = \frac{\sum_{f=0}^{F-1} (f - \overline{f})^2 \exp(\Phi_{jhf})^{0,3}}{\sum_{f=0}^{F-1} \exp(\Phi_{jhf})^{0,3}}. \tag{5.2}$$

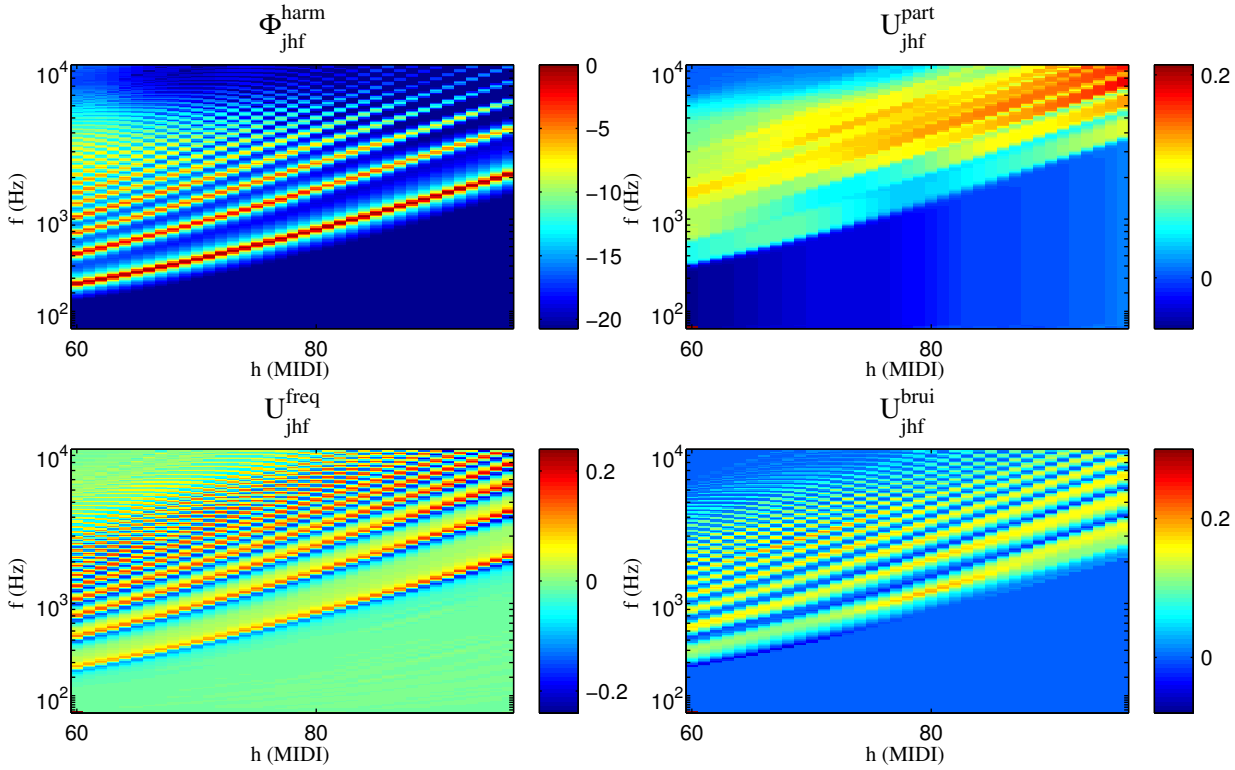

FIG. 5.1 – Spectres initiaux du modèle de flûte.

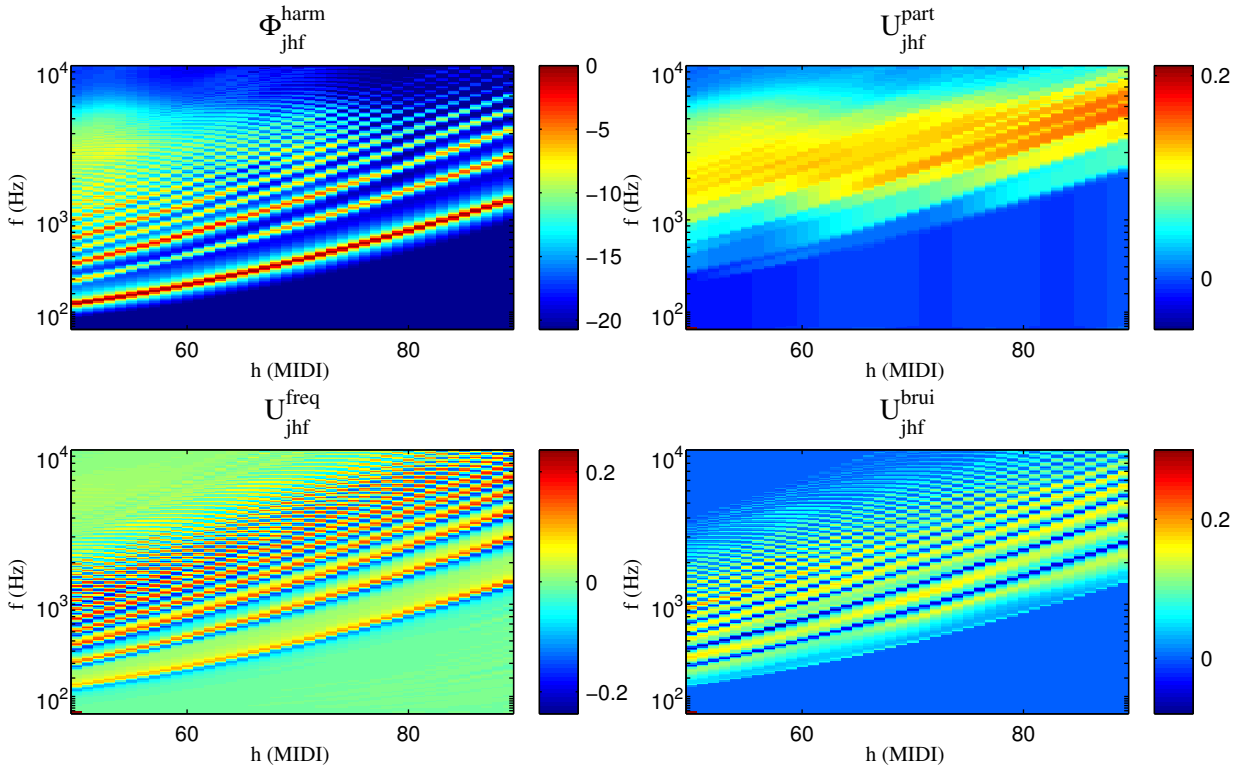

FIG. 5.2 – Spectres initiaux du modèle de clarinette.

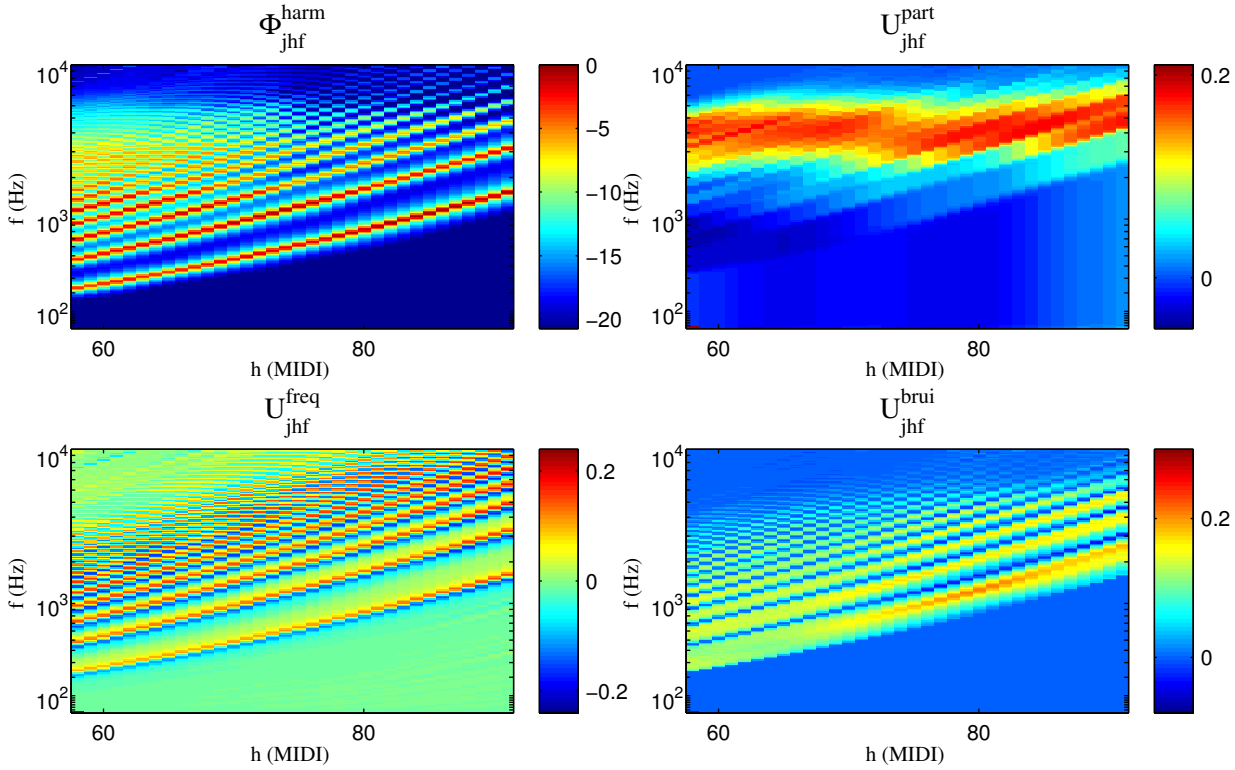

FIG. 5.3 – Spectres initiaux du modèle de hautbois.

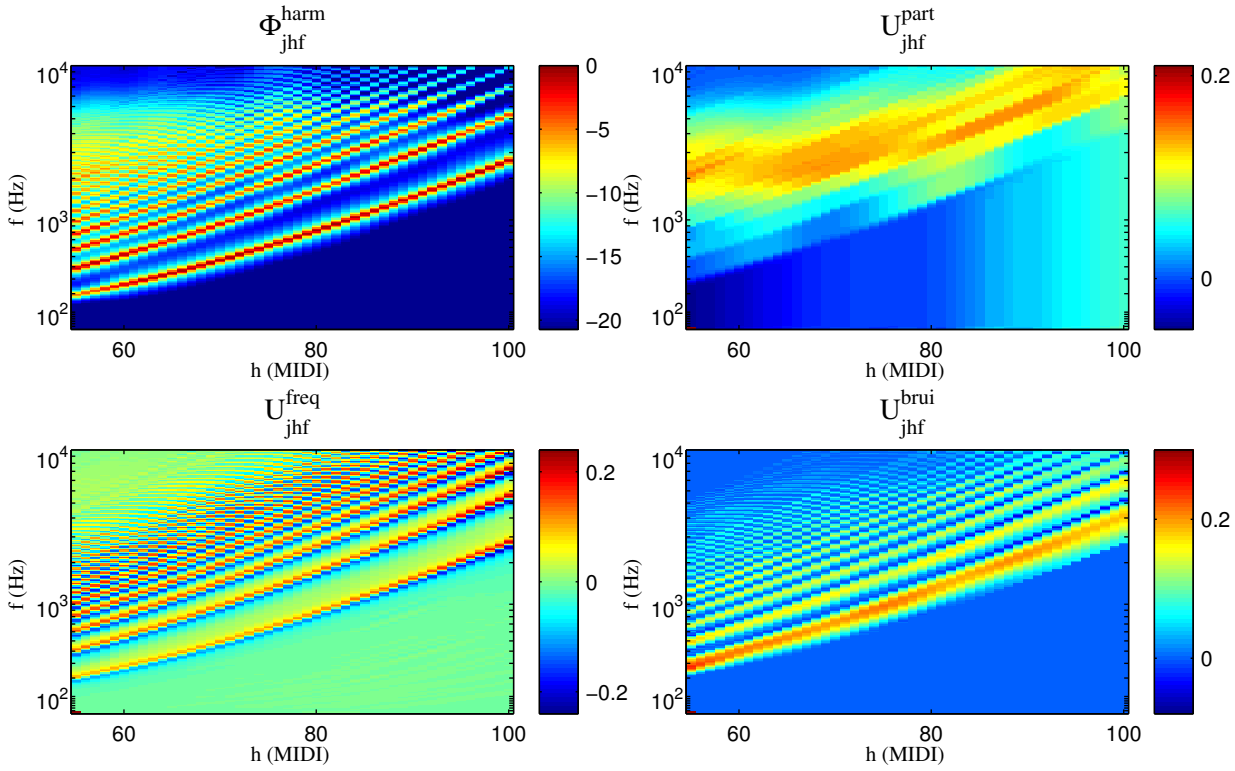

FIG. 5.4 – Spectres initiaux du modèle de violon.

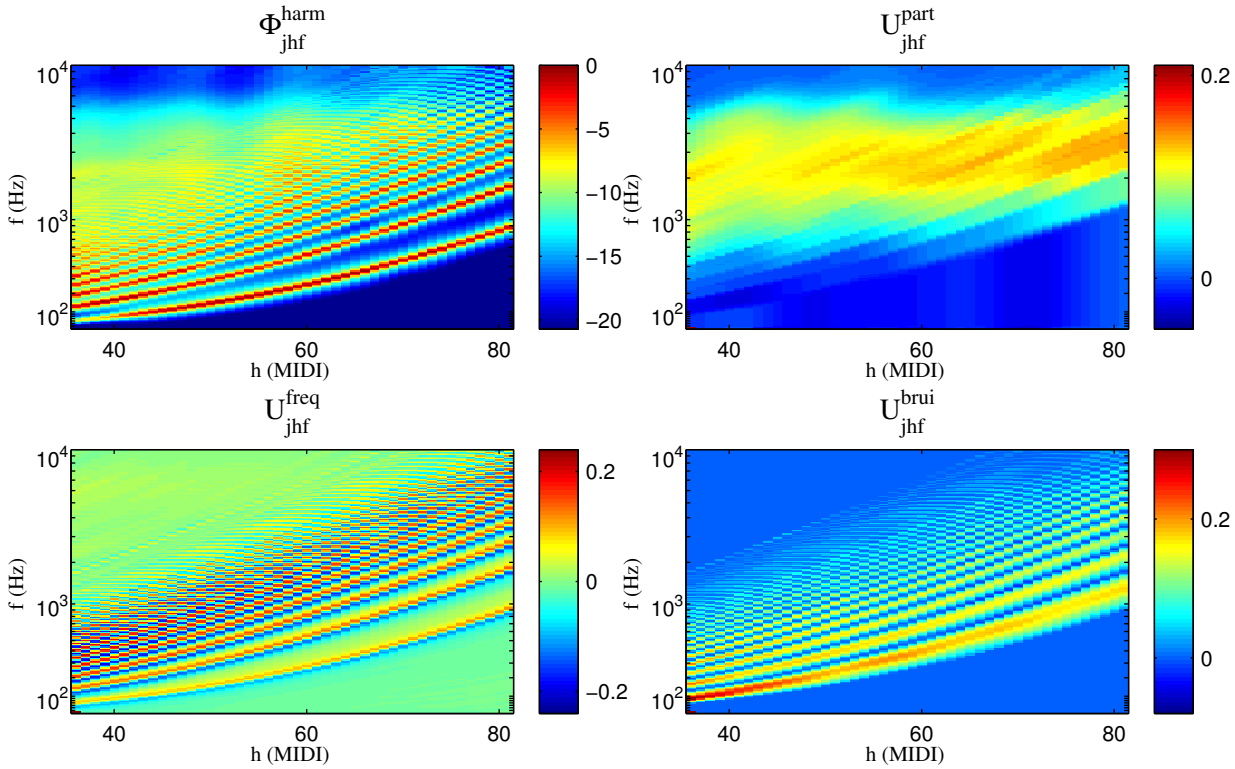

FIG. 5.5 – Spectres initiaux du modèle de violoncelle.

Nous constatons que les tessitures des instruments, ainsi que les centroïdes et écartements spectraux, peuvent permettre de discriminer des notes de certaines hauteurs d'instruments différents. Cependant les spectres complets ($\mathbf{\Phi}^{\mathrm{harm}}_{jh}$) apportent d'autres informations, par exemple les partiels d'ordre pair des notes graves de clarinette sont moins puissants que ceux d'ordre impair. De plus, nous remarquons que tous les spectres ($\mathbf{U}^{\mathrm{part}}_{jh}$) représentent un ajout de puissance sur les partiels d'ordre élevé qui contribue à modifier les centroïdes et écartements spectraux, en particulier en fonction de la nuance de jeu.

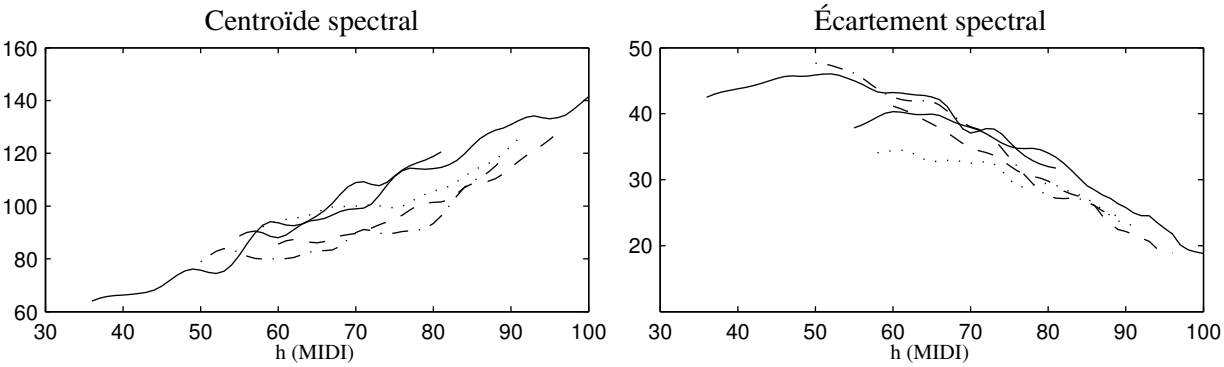

FIG. 5.6 – Centroïde spectral et écartement spectral des modèles d'instruments (tirets : flûte, tirets mixtes : clarinette, pointillés : hautbois, trait plein : violon et violoncelle).

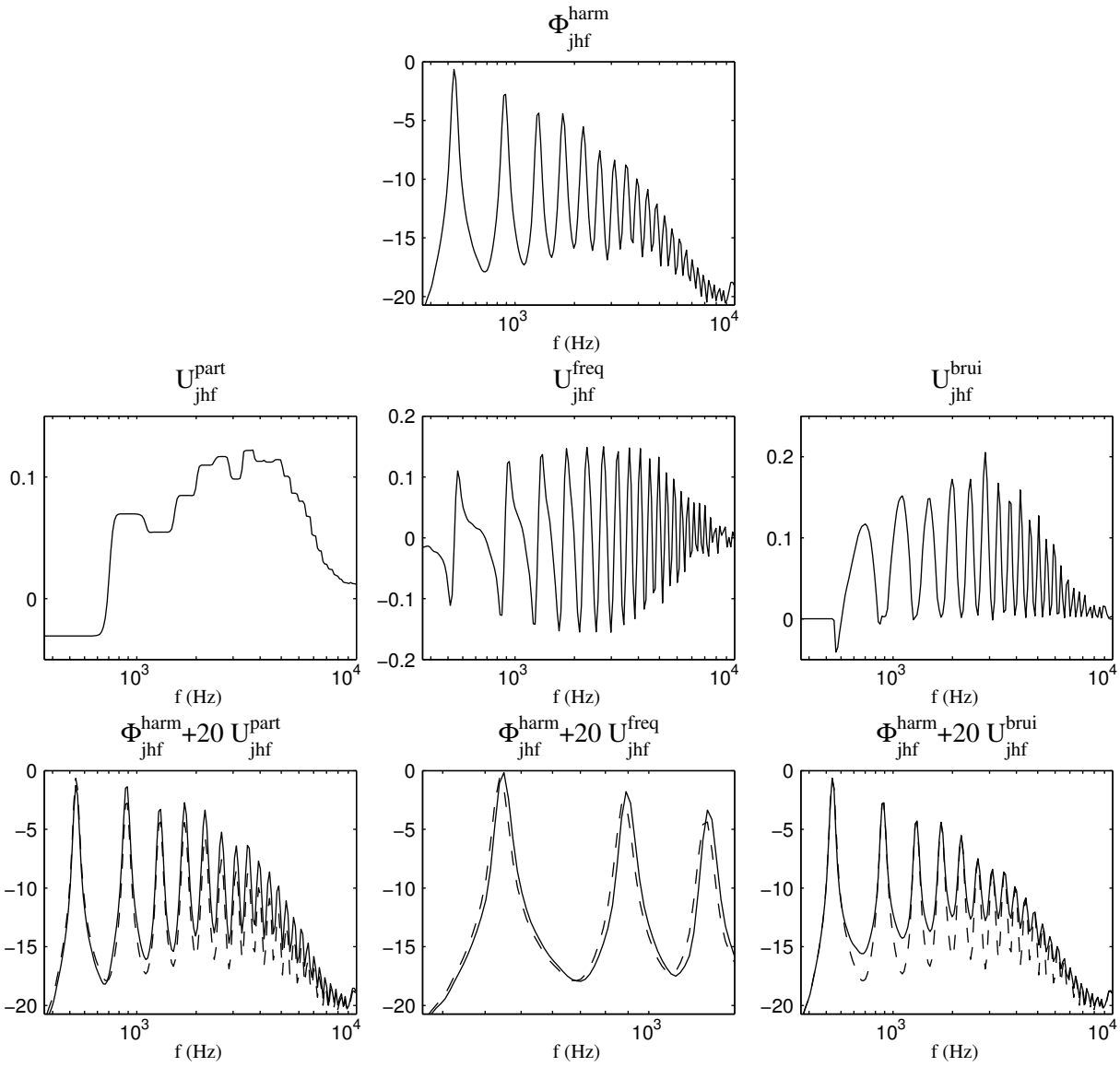

FIG. 5.7 – Exemples de spectres obtenus à l'aide du modèle de la note MIDI 69 de flûte.

5.2 Validation des hypothèses de modélisation

5.2.1 Spectres contraints, distribution de l'erreur résiduelle et des descripteurs

Afin de valider les hypothèses de modélisation effectuées, nous comparons la distribution des modèles à la distribution des données d'apprentissage dans le cas particulier de la note MIDI 69 (la4) du modèle de flûte. Le modèle est initialisé par $\mathbf{\Phi}_{jh} = \mathbf{\Phi}_{jh}^{\mathrm{harm}}$, $\mathbf{U}_{jh}^{1} = \mathbf{U}_{jh}^{\mathrm{part}}$ et $\mathbf{U}_{jh}^{2} = \mathbf{U}_{jh}^{\mathrm{fréq}}$ pour chaque note h, puis appris sur toutes les données avec contrainte sur les spectres. Ensuite nous sélectionnons les observations d'apprentissage correspondant à la note $h = 69$, et nous calculons les descripteurs optimaux.

La figure 5.7 montre des exemples de spectres obtenus en combinant le spectre moyen $\mathbf{\Phi}_{jh}^{\mathrm{harm}}$ aux spectres de variation $\mathbf{U}_{jh}^{\mathrm{part}}$, $\mathbf{U}_{jh}^{\mathrm{fréq}}$ et $\mathbf{U}_{jh}^{\mathrm{bruit}}$. Nous constatons que ces spectres initiaux représentent bien les trois types de variations du spectre voulues avec un nombre réduit de paramètres. Par exemple, les va-

riations de fréquence fondamentale sont modélisées correctement par un seul spectre de variation $\mathbf{U}_{jh}^{\text{fréq}}$ jusqu'à un quart de ton (situation représentée dans la figure).

La figure 5.8 compare les distributions empiriques de l'erreur résiduelle et des descripteurs (calculées par histogramme) à des distributions gaussiennes (calculées par les équations 4.41–4.44). Puisque l'instrument est observé en présence de bruit de fond, la distribution exacte de l'erreur résiduelle α_{jtf} n'est pas disponible. Nous l'approchons par l'erreur résiduelle combinée ϵ_{tf} dans les zones où l'instrument masque le bruit de fond, déterminées par la condition $\pi_{jhtf} > 1/2$. Nous remarquons que l'hypothèse de gaussianité des variables est assez bien vérifiée, particulièrement pour l'erreur résiduelle. Le caractère multimodal de la distribution empirique de e_{jht} est dû à la discrétisation des volumes de jeu et à la longue durée de la partie soutenue des notes dans la base de données : les deux modes correspondent au jeu *piano* et aux jeux *mezzo* et *forte*. Une distribution plus unimodale est obtenue en normalisant chaque forme d'onde de note isolée de la base de données indépendamment des autres.

Les coefficients de corrélation entre e_{jht} et v_{jht}^1, entre e_{jht} et v_{jht}^2 et entre v_{jht}^1 et v_{jht}^2 valent 0,56, -0,012 et -0,0038 respectivement. Cependant la corrélation entre e_{jht} et v_{jht}^1 devient proche de zéro en normalisant chaque forme d'onde de note isolée de la base de données indépendamment des autres. Cette normalisation est plus réaliste car il est peu probable de rencontrer de grosses différences de volume de jeu dans de courts extraits. Dans ce cas, l'hypothèse d'indépendance des descripteurs est vérifiée.

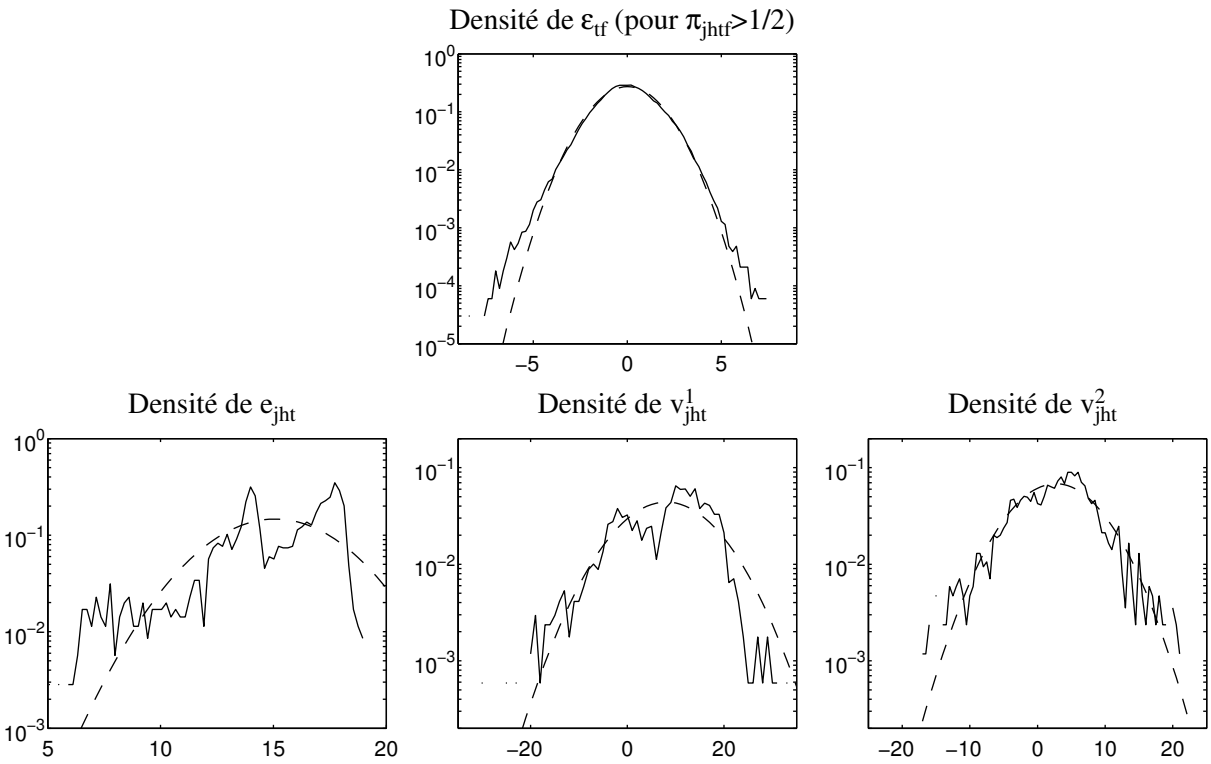

FIG. 5.8 – Densités empiriques de l'erreur résiduelle et des descripteurs sur les extraits d'apprentissage de la note MIDI 69 de flûte (traits pleins) comparées à des densités gaussiennes (tirets).

5.2.2 Distributions de durée des notes et des segments

Les hypothèses de distributions de durée des notes et des segments ne peuvent être validées sur la base de données de notes isolées. Faute de disposer d'une base de données de solos étiquetés, nous fixons les paramètres de distributions à $\mu_j^n = \log(50)$, $\sigma_j^n = 0,2$, $d_j^n = 20$, $\mu_j^s = \log(30)$, $\sigma_j^s = 0,2$ et $d_j^s = 20$.

Le temps de calcul dépend fortement des durées minimales d_j^n et d_j^s : des durées faibles augmentent le nombre d'hypothèses considérées et donc la complexité de l'algorithme de recherche en faisceaux.

La figure 5.9 compare ces distributions modèles aux distributions empiriques de durée des notes et des segments sur trois extraits solo de violoncelle, clarinette et violon utilisés par la suite (voir paragraphe 5.3.2). Cependant le nombre d'extraits est trop insuffisant pour en tirer des conclusions : chaque extrait a son propre *tempo* et un nombre limité de notes. En supposant que la proportion moyenne de blanches, de noires, de croches, *etc* est fixée, la distribution de durée des segments est proche d'une log-gaussienne dès lors que la distribution du *tempo* est log-gaussienne. De fait, les *tempi* autorisés par un métronome sont discrétisés sur une échelle logarithmique et les *tempi* moyens sont plus utilisés que les extrêmes, ce qui semble confirmer la pertinence de notre modélisation.

La figure 5.9 montre aussi la valeur de la probabilité de continuation associée à la distribution de durée des notes. Nous remarquons que cette probabilité est égale à un en-dessous de la durée minimale d_j^n, puis baisse à l'approche de la durée moyenne $\exp(\mu_j^n)$, puis remonte au-delà.

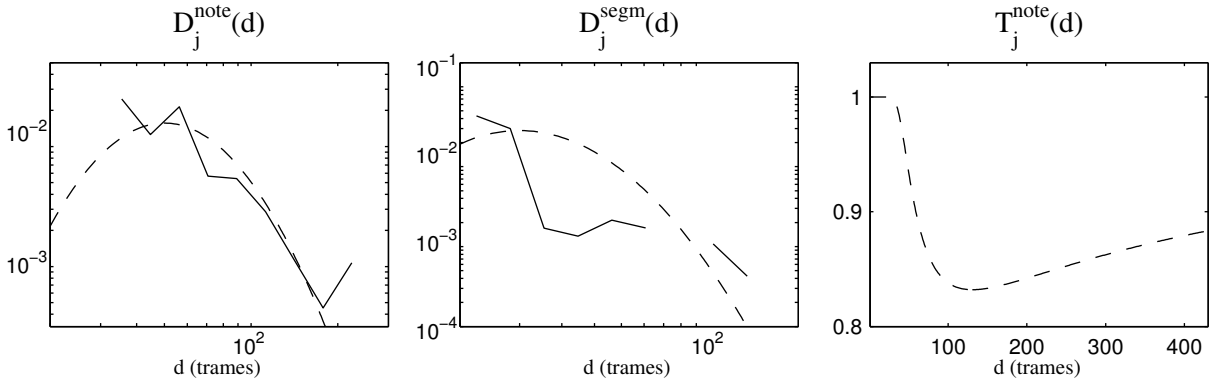

FIG. 5.9 – Probabilités empiriques de durée des notes et des segments (traits pleins) comparées à des probabilités log-gaussiennes (tirets).

5.3 Exemples

Nous testons maintenant la performance des modèles sur des tâches de transcription d'enregistrements solo afin de fixer une limite expérimentale à leur performance sur des enregistrements de musique de chambre, qui sont évidemment plus difficiles.

5.3.1 Identification d'instruments

Algorithme d'identification d'instruments

Notre première expérience consiste à identifier les instruments présents dans des courts enregistrements, sachant qu'il s'agit d'enregistrements solo. Nous considérons pour cela vingt extraits de cinq secondes par instrument, dont l'obtention est décrite dans l'annexe A.2.

Puisque l'identification précise des mélodies jouées n'est pas nécessaire, nous utilisons le modèle factoriel de couche d'état. L'algorithme d'identification d'instruments consiste simplement à estimer pour chaque instrument possible $\mathcal{O}$ la probabilité P^{trans} de la meilleure transcription par l'algorithme de saut du paragraphe 4.5.3, puis à proposer comme résultat l'instrument maximisant $P^{\text{trans}} P(\mathcal{O}|\mathcal{I})$.

Nous fixons le facteur de parcimonie Z_1 à la même valeur Z pour tous les instruments, et la distribution *a priori* des orchestres à

$$P(\mathcal{O}|\mathcal{I}) \propto \begin{cases} Z^{-T\,N_{j'}} & \text{si } \mathcal{O} \text{ contient un seul instrument } j', \\ 0 & \text{sinon.} \end{cases} \tag{5.3}$$

Ce choix corrige les différences de largeur de tessiture des instruments. Il attribue la même probabilité de couche d'état $P(\mathbf{E}_1|\mathcal{M}_1)P(\mathcal{O}|\mathcal{I})$ à tous les instruments pouvant produire les états $\mathbf{E}_1$, et par conséquent la même probabilité *a posteriori* $P^{\text{trans}}P(\mathcal{O}|\mathcal{I})$ à tous les instruments sur des extraits silencieux (où $E_{jht} = 0$ pour tout (h,t)). La dépendance de $P(\mathcal{O}|\mathcal{I})$ à la durée des extraits T peut être transférée à w_{comb}, w_{desc} et $w_{\text{état}}$ en divisant le logarithme de la loi de Bayes pondérée par T. Pour une application à l'indexation de bases de données de CD musicaux, $P(\mathcal{O}|\mathcal{I})$ pourrait être choisie en fonction du nombre d'enregistrements disponibles pour chaque instrument.

Résultats

Les hyper-paramètres sont fixés à $Z = 0,95$, $\sigma^\epsilon = 1,2$, $w_{\text{comb}} = 0,25$, $w_{\text{desc}} = 0,5$, $w_{\text{état}} = 1$ et $w_{\mathcal{O}} = 1$. Les résultats sont évalués par les critères de TRI et TRF moyennés sur tous les enregistrements de test, présentés dans le tableau 5.1. Les instruments sont séparés en deux familles : vents et cordes frottées.

Dans le meilleur des cas, sur la dernière ligne du tableau, le TRI et le TRF moyens sont égaux à 89% et 96%. Cette performance est satisfaisante vu le faible nombre de données d'apprentissage utilisé (tous les instruments sont enregistrés dans une salle unique, et sans les procédures de lissage des paramètres entre notes de hauteurs voisines le TRI n'est que de 82%). Le tableau 5.2 montre la matrice de confusion correspondante. La source d'erreur principale est l'attribution de certaines phrases de violoncelle ne contenant que des notes aiguës au violon. Cependant, les phrases de violoncelle contenant à la fois des notes aiguës et des notes graves sont correctement identifiées. Les informations ambigües apportées par certaines notes au sein d'une phrase sont compensées par les autres notes.

Pour déterminer l'importance relative des informations de tessiture et d'enveloppe spectrale, nous effectuons la même expérience avec $K_1 = 0$ et des spectres contraints ($\mathbf{\Phi}_{jh}^{\text{harm}}$) de même enveloppe spectrale de pente -12 dB par octave. Le TRI et le TRF moyens chutent à 33% et 60% respectivement, ce qui est proche d'un choix aléatoire (20% et 50%). Dans ce cas, seul le violoncelle garde un TRI satisfaisant de 80%.

Cela prouve d'une part que les modèles d'intrument capturent bien les caractéristiques spectrales des instruments utiles pour leur identification, et d'autre part qu'un TRI satisfaisant peut être obtenu même sans modéliser les caractéristiques temporelles du timbre instrumental comme la durée de l'attaque ou des transitions *legato*.

Le TRI moyen est comparable au TRI de 88% obtenu par Eggink et Brown [Egg03] pour les mêmes instruments avec un algorithme d'inférence à données manquantes basé sur un MG sur le spectre de log-puissance et utilisant une échelle de fréquence linéaire pour le calcul des spectres.

Par contre, il est supérieur au TRI obtenu par Brown, Houix et McAdams [Bro01] et Marques et Moreno [Mar99a] pour des instruments différents modélisés par MG sur le cepstre. L'amélioration est vraisemblablement due au fait que les coefficients cepstraux modélisent mal la réverbération et les différents niveaux de bruit de fond car ils représentent l'enveloppe de tout le spectre (y compris du bruit de fond) au lieu de l'enveloppe de chaque note. De plus ils ne tiennent pas compte de la dépendance entre la hauteur et l'enveloppe spectrale, qui est une caractéristique importante des instruments [Kit03].

Taille	Initialisation	Spectres contraints	Performance TRI	TRF
$K_1 = 0$	harm	NON	83%	95%
	harm	OUI	83%	94%
$K_1 = 1$	harm + part	NON	79%	94%
	harm + part	OUI	88%	95%
	harm + fréq	OUI	85%	94%
	harm + bruit	OUI	85%	93%
$K_1 = 2$	harm + part + fréq	OUI	89%	96%

TAB. 5.1 – Comparaison des performances des modèles appris pour l'identification d'instruments sur des enregistrements solo.

		Instrument identifié				
		Fl	Cl	Ob	Vn	Vc
Extrait testé	Fl	**95%**		5%		
	Cl	5%	**90%**		5%	
	Ob			**95%**	5%	
	Vn	5%			**95%**	
	Vc	5%			25%	**70%**

TAB. 5.2 – Matrice de confusion pour l'identification d'instruments sur des enregistrements solo avec les modèles les plus performants.

Validation croisée des modèles appris

Le tableau 5.1 montre que la performance varie de façon notable selon les conditions d'apprentissage. Trois remarques peuvent être effectuées à ce sujet.

Premièrement, le meilleur résultat obtenu augmente en fonction du nombre K_1 de composantes de variation jusqu'à $K_1 = 2$. La modélisation de chaque note par un spectre moyen semble donc être insuffisante.

Deuxièmement, la performance des modèles est généralement supérieure en contraignant les spectres $(\mathbf{\Phi}_{jh})$ et $(\mathbf{U}^k_{jh})$ à conserver leur valeur initiale au cours de l'apprentissage, plutôt qu'en apprenant les spectres optimaux. L'utilisation de spectres structurés semble donc importante pour l'identification d'instruments. L'apprentissage par critère MAP tend à supprimer la structure des spectres : les $(\mathbf{\Phi}_{jh})$ contiennent alors de l'énergie non harmonique entre les partiels, et les $(\mathbf{U}^k_{jh})$ modélisent à la fois les variations de cette énergie non harmonique et d'autres caractéristiques. Cette perte de structure peut se révéler nuisible. Par exemple, les modèles non contraints avec $K_1 = 1$ initialisés par $(\mathbf{\Phi}^{\mathrm{harm}}_{jh})$ et $(\mathbf{U}^{\mathrm{part}}_{jh})$ sont moins performants que ceux avec $K_1 = 0$ (bien qu'ils fournissent une meilleure représentation de l'ensemble d'apprentissage). De façon similaire, nous avons noté une faible perte de performance (de l'ordre de 1%) en absence de contrainte sur les lois gaussiennes des descripteurs modélisant les variations de fréquence.

Troisièmement, tous les types de structure n'ont pas la même importance pour l'identification d'instruments. La modélisation des variations de puissance des partiels semble la plus importante, suivie de celle des variations de fréquence fondamentale. La modélisation de l'énergie non harmonique entre les partiels semble inutile ou nuisible. Le modèle avec $K_1 = 3$ contenant les trois types de spectres de variation contraints donne un TRI moyen de 64%, nettement inférieur à celui du modèle avec $K_1 = 2$ sans composante d'énergie non harmonique.

Nous choisissons la tâche d'identification d'instruments solo comme critère de validation croisée des

modèles d'instruments. Dans le reste de ce chapitre et dans le chapitre 6, nous utiliserons uniquement les modèles d'instruments correspondant à la meilleure performance dans cette expérience, c'est-à-dire les modèles avec deux composantes de variation où $\mathbf{\Phi}_{jh} = \mathbf{\Phi}_{jh}^{\mathrm{harm}}$, $\mathbf{U}_{jh}^{1} = \mathbf{U}_{jh}^{\mathrm{part}}$ et $\mathbf{U}_{jh}^{2} = \mathbf{U}_{jh}^{\mathrm{fréq}}$ pour toute note h. Dans le chapitre 7, nous utiliserons les modèles sans composante de variation où $\mathbf{\Phi}_{jh} = \mathbf{\Phi}_{jh}^{\mathrm{harm}}$. En effet, nous verrons qu'en présence de plusieurs canaux de mélange, des modèles d'instruments simples suffisent et permettent de réduire le temps de calcul.

Robustesse au bruit de fond et à la réverbération

Nous avons testé la robustesse de ces modèles au niveau de bruit en ajoutant aux enregistrements testés du bruit blanc gaussien à divers Rapports Signal-à-Bruit (RSB). Le TRI moyen baisse à 83% pour un RSB de 20 dB et à 58% pour un RSB de 0 dB. Les modèles restent donc capables d'exploiter les informations utiles pour l'identification d'instruments à des niveaux de bruit élevés.

Nous avons obtenu ces résultats en appliquant l'algorithme de saut deux fois de suite sur chaque extrait et en utilisant l'estimation du bruit de fond à la première itération pour initialiser la deuxième. En effet les résultats de l'algorithme dépendent fortement de l'initialisation du bruit de fond lorsque le RSB est faible. Cette double application de l'algorithme ne modifie pas la performance sur les extraits non bruités, mais sans elle le TRI moyen baisse à 78% et 39% respectivement sur les extraits bruités.

Une autre possibilité d'améliorer la performance serait de choisir les hyper-paramètres σ^{ϵ}, w_{comb} et Z en fonction du niveau de bruit, avec par exemple une valeur plus faible pour σ^{ϵ} à fort niveau de bruit.

Nous avons aussi évalué la robustesse à la réverbération en convoluant les enregistrements testés avec une réponse impulsionnelle de salle de réponse fréquentielle en amplitude non plate, correspondant à un temps de réverbération perçu de 0,8 s environ (voir annexe A.3). Le TRI moyen décroît à 81%.

5.3.2 Identification de notes

Notre deuxième expérience consiste à identifier les notes à partir des deux modèles de couche d'état dans trois extraits solo réels de violoncelle, clarinette et violon. Nous changeons les hyper-paramètres à $Z = 0,96$ et $\sigma^{\epsilon} = 1,4$.

Les résultats sont présentés dans les figures 5.10, 5.11 et 5.12. Les transcriptions réalisées à partir du modèle segmental correspondent toutes aux véritables partitions des extraits. De plus les instants d'attaque des notes sont assez précis, et le nombre de notes présentes à chaque instant est correctement estimé, même en présence d'une réverbération élevée comme dans l'extrait de clarinette. Les transcriptions réalisées à partir du modèle factoriel sont aussi correctes visuellement, mais contiennent des états de très courte durée ou au contraire des "trous" au sein de certaines notes. Elles pourraient donc être utilisées pour de tâches de comparaison de mélodies, mais pas pour l'identification de notes proprement dit.

La figure 5.13 compare la densité empirique de l'erreur résiduelle sur l'extrait solo de violoncelle (calculée par histogramme après transcription par le modèle segmental) à la densité gaussienne définie par $\sigma^{\epsilon} = 1,4$ (différente de celle définie par la moyenne et l'écart-type empiriques de l'erreur). Nous constatons que la densité empirique contient un pic en zéro. En effet, l'écart-type du modèle de bruit de fond est légèrement plus faible que celui du modèle d'instrument en pratique. Le signal en très haute ou en très basse fréquence ne contient que le bruit de fond. L'erreur résiduelle dans ces zones est faible car les observations sont stationnaires et donc bien modélisées par le modèle de bruit de fond stationnaire.

Il est possible de modéliser cette différence d'écart-type en remplaçant l'équation 4.17 par la combinaison des densités des modèles d'instruments et de bruit de fond au sens de la combinaison de modèles [Gal95]. L'écart-type de l'erreur résiduelle en un point temps-fréquence donné devient alors proche de

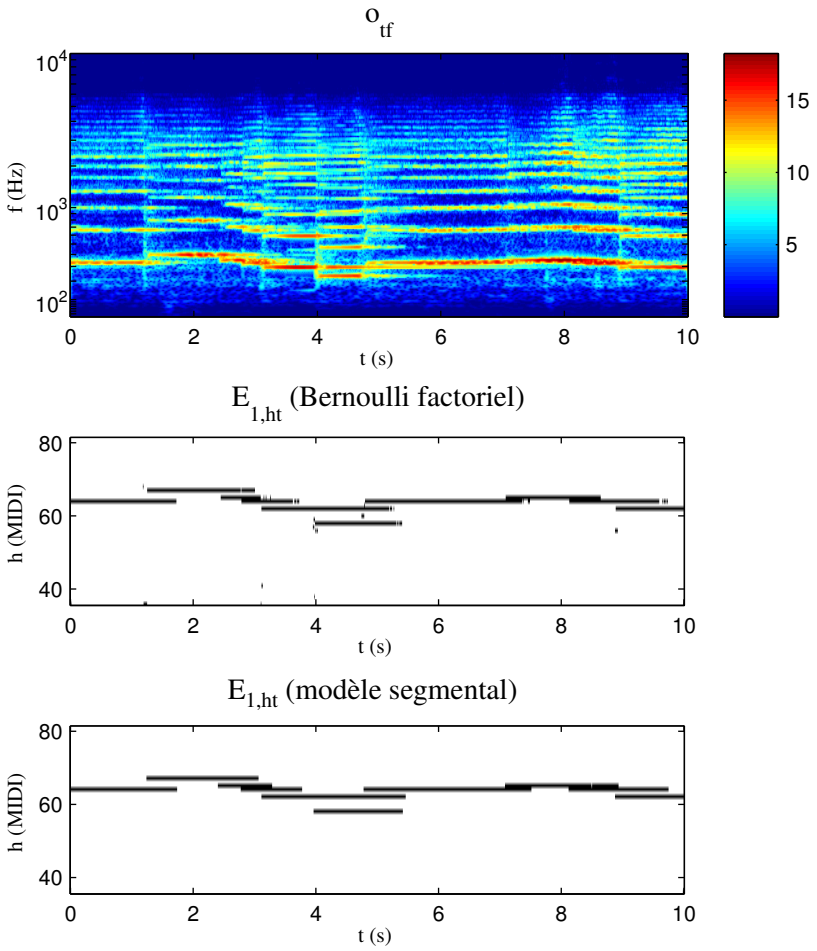

FIG. 5.10 – Notes identifiées sur l'extrait solo de violoncelle.

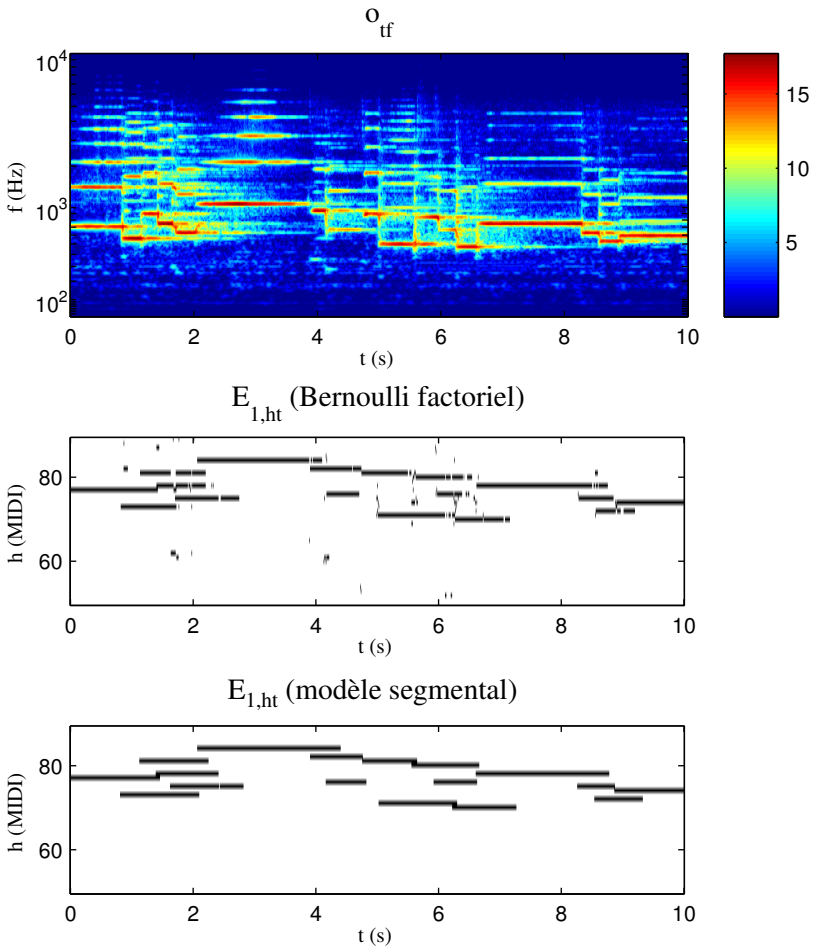

FIG. 5.11 – Notes identifiées sur l'extrait solo de clarinette.

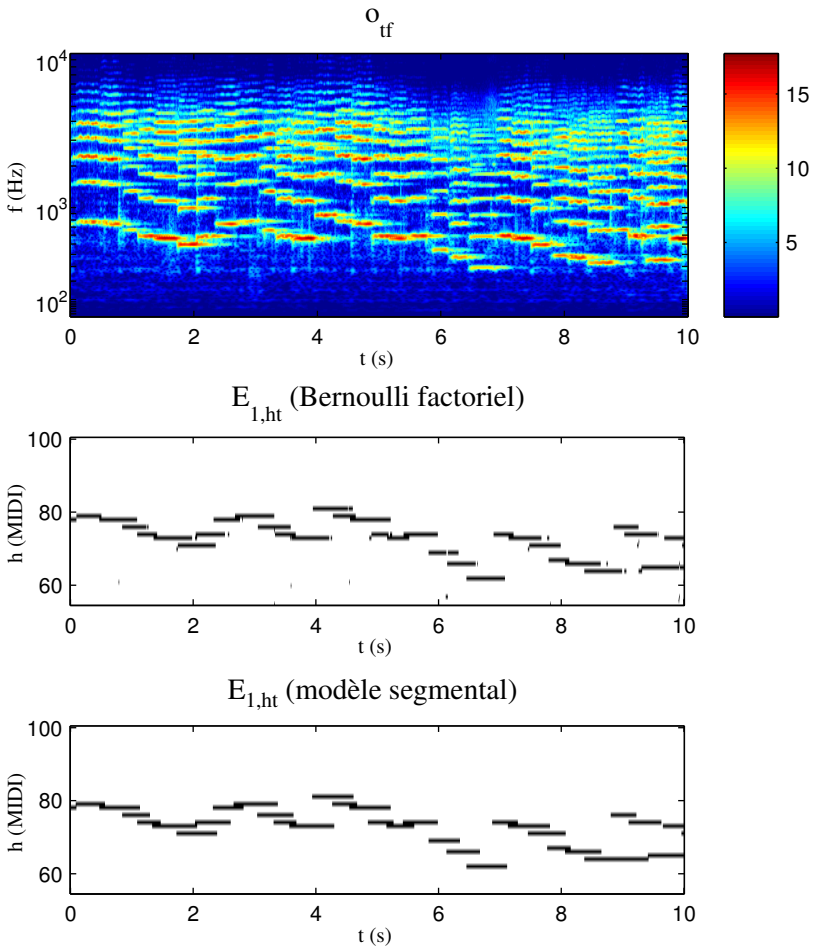

FIG. 5.12 – Notes identifiées sur l'extrait solo de violon.

celui du modèle d'instrument si l'instrument masque le bruit de fond en ce point et *vice-versa*. Nous avons testé ce modèle pour l'identification d'instruments et de notes sur les exemples de ce chapitre. Nous avons constaté une baisse de performance pour les deux tâches. En particulier, la sensibilité accrue aux variations du bruit de fond a tendance à rajouter des notes de très courte durée dans les transcriptions réalisées à l'aide du modèle factoriel. Et la probabilité de transcription, permettant de choisir le meilleur instrument, devient plus sensible à l'initialisation du bruit de fond. De plus, cela engendre un surplus de temps de calcul important.

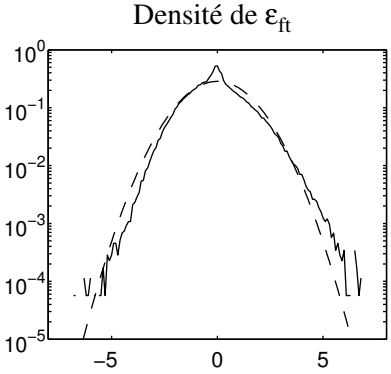

FIG. 5.13 – Densité empirique de l'erreur résiduelle sur l'extrait solo de violoncelle (traits pleins) comparée à une densité gaussienne (tirets).

5.3.3 Résumé des résultats

En résumé, les résultats de ce chapitre assurent que les modèles d'instruments proposés capturent bien certaines caractéristiques utiles du son instrumental. La performance d'identification d'instruments sur des extraits solo réels est comparable à celle d'un des meilleurs algorithmes existants. Et l'identification de notes donne un résultat parfait sur les trois extraits testés (avec un réglage manuel des hyperparamètres et des distributions de durée identique pour tous les extraits).

Les exemples de ce chapitre servent de limite expérimentale à la performance de transcription des extraits de musique de chambre étudiés dans les chapitres suivants, naturellement plus difficiles. En particulier, nous remarquons que la performance d'identification d'instruments sur des extraits solo n'est pas parfaite, particulièrement pour les instruments de timbre proche comme le violon et le violoncelle. Il est donc illusoire d'attendre une performance parfaite pour l'association note-instrument dans un extrait de musique de chambre sans utiliser d'autres informations que le timbre.

Chapitre 6

Transcription et séparation d'enregistrements monocanal

Ce sixième chapitre présente des applications des modèles d'instruments à la transcription et à la séparation d'enregistrements de musique de chambre monocanal. Nous expliquons dans le paragraphe 6.1 comment combiner les modèles d'instruments pour créer des modèles de mélanges monocanal, puis dans le paragraphe 6.2 comment extraire les sources d'un mélange par filtrage une fois sa transcription effectuée. Nous donnons ensuite quelques exemples de transcription et de séparation de mélanges réels et synthétiques dans le paragraphe 6.3.

Ce chapitre apporte peu de nouveautés théoriques par rapport aux chapitres précédents. Mais l'expérience d'identification d'instruments sur des mélanges synthétiques avec effectif instrumental inconnu n'a pas d'équivalent dans la littérature à notre connaissance. Faute de disposer d'une base de données de mélanges étiquetés, nous ne présentons les résultats d'identification de notes et de séparation que pour un nombre réduit de mélanges musicaux non percussifs. La performance de notre algorithme ne peut donc être comparée à celles des algorithmes existant pour le même type de mélange [Kin99, Egg03], eux aussi testés sur un nombre réduit de mélanges de courte durée. Nous reprenons quelques exemples de transcription sur des duos réels dans les articles [Vin04c, Vin04b].

6.1 Transcription

6.1.1 Approximation des spectres à court terme des sources

Après transcription d'un mélange, l'extraction de sources est effectuée en deux étapes : estimation des spectres à court terme des sources pour maximiser $P((\mathbf{m}_j)|\,\mathbf{o}, \mathbf{\Theta}, (\mathbf{p}_j), (\mathbf{E}_j), (\mathcal{M}_j))$, puis estimation de leurs formes d'ondes pour maximiser $P(\mathbf{s}|x, \mathbf{\Theta}, (\mathbf{m}_j))$. En pratique, nous constatons que les résultats de séparation varient peu en remplaçant dans la deuxième étape les spectres exacts des sources et du bruit de fond $(\mathbf{m}_j)$ et $\mathbf{n}$ par leurs spectres modèles $(\mathbf{m}'_j)$ et $\mathbf{n}'$. Dans le reste de ce chapitre et dans le suivant, nous adoptons cette approximation. Nous omettons donc la définition de la probabilité des observations en fonction des spectres exacts pour définir directement la probabilité en fonction des spectres modèles, autrement dit en fonction des descripteurs.

6.1.2 Choix des observations

Ajout des spectres de plusieurs sources

La combinaison de modèles d'instruments dans un mélange est similaire à la combinaison du modèle d'instrument et du bruit de fond dans un solo. Nous exprimons tout mélange de façon équivalente comme

un mélange instantané à gains unitaires (sources et images spatiales sur le canal unique sont confondues). La modélisation de filtres de mélange non triviaux n'est pas nécessaire dès lors que nous ne cherchons pas à déréverbérer les sources. Et l'hypothèse d'égalité des gains est valable lorsque les volumes des instruments dans le mélange sont proches de leurs volumes de jeu habituels, appris sur la base de données de notes isolées. Ceci est le cas pour les mélanges réels mais aussi généralement pour les mélanges synthétiques, puisque le mixage tend à égaliser les volumes des sources.

Nous choisissons comme observations $(\mathbf{o}_t)$ le spectre de log-puissance du mélange et nous supposons que

$$\mathbf{o}_t = \log\Big(\sum_{j=1}^{n} \mathbf{m}'_{jt} + \mathbf{n}'\Big) + \boldsymbol{\epsilon}_t, \tag{6.1}$$

où $\mathbf{n}'$ est le spectre de puissance du modèle de bruit de fond et $\boldsymbol{\epsilon}_t$ est un bruit gaussien à covariance diagonale isotrope d'écart-type σ^ϵ. Cette hypothèse est valable lorsque les écarts-types des erreurs résiduelles des modèles de tous les instruments et du bruit de fond sont proches de σ^ϵ. La probabilité des observations s'exprime en fonction des descripteurs par

$$P(\mathbf{o}_t|\boldsymbol{\Theta}, (\mathbf{p}_{jt}), (\mathcal{M}_j)) = \prod_{f=0}^{F-1} \mathcal{N}(\epsilon_{tf}; 0, \sigma^\epsilon), \tag{6.2}$$

où les paramètres de mélange $\boldsymbol{\Theta}$ contiennent le spectre modèle du bruit de fond $\mathbf{n}'$.

Expression de la loi de Bayes pondérée

Nous supposons que la probabilité *a priori* du bruit de fond $P(\boldsymbol{\Theta})$ est uniforme. Comme précédemment, le critère MAP de transcription de l'équation 3.9 est alors remplacé par la maximisation de $P^{\text{trans}} = P(\mathbf{o}, (\mathbf{p}_j), (\mathbf{E}_j)|\, \boldsymbol{\Theta}, (\mathcal{M}_j), \mathcal{I})$ conjointement par rapport aux états $(\mathbf{E}_j)$, aux descripteurs $(\mathbf{p}_j)$ et aux paramètres de mélange $\boldsymbol{\Theta}$. La loi de Bayes pondérée s'écrit $P^{\text{trans}} = \prod_{t=0}^{T-1} P_t^{\text{trans}}$ avec

$$\begin{aligned} \log P_t^{\text{trans}} = {} & w_{\text{comb}} \log P(\mathbf{o}_t|\boldsymbol{\Theta}, (\mathbf{p}_{jt}), (\mathcal{M}_j)) \\ & + w_{\text{desc}} \sum_{j=1}^{n} \log P(\mathbf{p}_{jt}|\mathbf{E}_{jt}, \mathcal{M}_j) \\ & + w_{\text{état}} \sum_{j=1}^{n} \log P(\mathbf{E}_{jt}|(\mathbf{E}_{jt'})_{0\leq t'\leq t-1}, \mathcal{M}_j, \mathcal{I}). \end{aligned} \tag{6.3}$$

Le premier terme de la somme est développé dans l'équation 6.2, le deuxième dans l'équation 4.5 et le troisième dans l'équation 4.6 ou 4.9 selon le modèle de couche d'état choisi.

6.1.3 Algorithmes de transcription

Les algorithmes de transcription de mélanges monocanal sont semblables aux algorithmes de transcription d'enregistrements solo. L'utilisation de techniques heuristiques de réduction de l'espace d'état est encore plus cruciale, car la taille de l'espace d'état augmente à peu près exponentiellement avec le nombre d'instruments.

Les descripteurs et le bruit de fond sont initialisés comme précédemment et réestimés par algorithme de Newton au second ordre approché. La quantité de masquage $\pi_{1,htf}$ est alors remplacée par des quantités équivalentes (π_{jhtf}) définies pour chaque instrument j par

$$\pi_{jhtf} = \frac{\exp(\Phi'_{jhtf}) \exp(e_{jht})}{\sum_{j=1}^{n} m'_{jtf} + n'_f}. \tag{6.4}$$

Dans l'algorithme de saut, à l'étape 2a l'ensemble des notes possibles $\mathcal{P}_t$ contient N_{poss} notes par instrument.

Dans l'algorithme de recherche en faisceaux, à l'étape 4a $\mathfrak{C}_0$ peut contenir des états formés de plusieurs notes de $\mathcal{P}_0$, à l'étape 5a le prolongement des chemins partiels de $\mathfrak{C}_{t-1}$ peut s'effectuer en rajoutant plusieurs notes de $\mathcal{P}'_t$ d'instruments différents simultanément. Dans l'étape 5e la durée écoulée sur le segment en cours est remplacée par le minimum des durées écoulées sur les segments en cours pour chaque instrument, et le nombre de chemins partiels pour lesquels le minimum des durées écoulées sur les segments en cours est inférieur à $d_{\min}$ est limité à N_{dern} (en ne gardant que les meilleurs).

6.2 Séparation

6.2.1 Filtrage pour l'extraction de sources

Nous estimons les sources par filtrage variant dans le temps du mélange. Le mélange x est découpé en sous-bandes $(x^f)_{0\leq f\leq F-1}$ puis en trames disjointes $(x^{tf})_{0\leq f\leq F-1,\,0\leq t\leq T-1}$. Nous supposons que le filtrage variant dans le temps est réduit à un gain $\Pi_{jtf} \in [0,1]$ en chaque point (t,f) pour chaque instrument j, de sorte que $\widehat{s_j^{tf}} = \Pi_{jtf}\, x^{tf}$. Chaque source $\widehat{s_j}$ est alors estimée à partir des $(\widehat{s_j^{tf}})_{0\leq f\leq F-1,\,0\leq t\leq T-1}$ par mise côte-à-côte des trames et inversion du banc de filtres. Le calcul et l'inversion du banc de filtres sont décrits dans l'annexe B.

Cette méthode attribue aux sources la phase locale du mélange. La phase locale des sources pourrait être mieux estimée en tenant compte des relations d'harmonicité entre partiels d'une même note. Nous n'abordons pas cette question par la suite.

Les méthodes de filtrage existantes expriment souvent le gain Π_{jtf} sous la forme

$$\Pi_{jtf} = \left(\frac{m'_{jtf}}{\sum_{j=1}^{n} m'_{jtf} + n'_f}\right)^{\gamma}, \tag{6.5}$$

où γ est un réel positif.

Si $\gamma = 1$, la méthode de filtrage est appelée pseudo-Wiener [Ben03]. Il s'agit d'une extension du filtrage de Wiener au cas où le signal non désiré n'est pas stationnaire. Cette méthode vérifie la contrainte $x = \sum_{j=1}^{n} \widehat{s_j}$ lorsque la reconstruction du banc de filtres est exacte. Lorsque les spectres des sources et du bruit de fond $(\mathbf{m}'_j)$ et $\mathbf{n}'$ sont bien estimés, le résultat a généralement un RSI faible mais un RSA élevé car les zones masquées se voient attribuer un gain très faible.

Si $\gamma = 1/2$, la méthode est appelée Soustraction de Puissance (SP) [Cap93]. Cette fois le résultat a un RSI plus élevé et peut avoir un RSA plus faible lorsque les spectres sont bien estimés. Cette méthode ne vérifie pas la contrainte $x = \sum_{j=1}^{n} \widehat{s_j}$.

Le choix de γ dépend du compromis voulu entre quantité d'interférences et d'artefacts. Dans le cas du débruitage, la soustraction de puissance est préférée car elle limite la quantité de bruit musical [Cap93]. Dans le cas de l'extraction de sources, ce n'est pas forcément le cas car les interférences sont plus gênantes que le bruit de fond résiduel.

D'autres méthodes pour limiter les artefacts d'extraction consistent à fixer une limite inférieure aux gains de filtrage suivant un modèle de masquage auditif [Wol00] ou bien à utiliser la continuité temporelle des sources. L'utilisation de modèles de sources structurés semble rendre ces méthodes moins nécessaires que dans le cas aveugle pour lequel elles ont été conçues [Ben03].

6.2.2 Filtrage pour la modification de scène sonore

La modification de scène sonore peut aussi être réalisée à l'aide de cette méthode simplement en extrayant les sources, puis en les mélangeant différemment. Dans ce cas, la contrainte $x = \sum_{j=1}^{n} \widehat{s}_j$ est naturelle, puisqu'elle garantit que le mélange est égal à son remix si les filtres de remix sont égaux aux filtres de mélange initiaux. Parmi les méthodes de filtrage définies ci-dessus, seul le filtrage de Wiener est donc utilisable.

Si les nouveaux filtres de mélange sont définis par un vecteur de gains positifs en chaque point temps-fréquence, le remix peut aussi s'exprimer par filtrage du mélange par un gain Π_{tf} en chaque point (t, f). Par exemple, dans le cas où la modification de scène consiste à multiplier l'amplitude de la source j par facteur a, le gain Π_{tf} est défini par

$$\Pi_{tf} = \frac{a\, m'_{jtf} + \sum_{j' \neq j} m'_{j'tf} + n'_f}{\sum_{j=1}^{n} m'_{jtf} + n'_f}. \tag{6.6}$$

6.3 Exemples

6.3.1 Identification d'instruments sur des duos monocanal synthétiques

Notre première expérience étend les résultats d'identification d'instruments à des duos monocanal synthétiques de courte durée. En additionnant les formes d'onde des extraits solo utilisés dans le paragraphe 5.3.1, nous construisons sept extraits de cinq secondes pour chaque paire d'instruments (soit trente-cinq duos d'instruments de même nom et soixante-dix duos d'instruments différents).

Nous proposons deux algorithmes différents pour l'identification d'instruments, selon que le nombre d'instruments présents est connu *a priori* ou non.

Effectif instrumental connu

Lorsque le nombre d'instruments est connu, l'algorithme d'identification le plus immédiat reprend l'algorithme utilisé sur des enregistrements solo en modifiant la probabilité *a priori* des orchestres en

$$P(\mathcal{O}|\mathcal{I}) \propto \begin{cases} Z^{-T\,(N_{j'}+N_{j''})} & \text{si } \mathcal{O} \text{ contient deux instruments } j' \text{ et } j'', \\ 0 & \text{sinon.} \end{cases} \tag{6.7}$$

La probabilité P^{trans} de la meilleure transcription est alors estimée pour chaque couple d'instruments $\mathcal{O}$ par algorithme de saut, et le couple maximisant $P^{\text{trans}} P(\mathcal{O}|\mathcal{I})$ est proposé comme résultat.

La probabilité *a priori* $P^{\text{trans}} P(\mathcal{O}|\mathcal{I})$ possède la même propriété intéressante que celle utilisée sur les enregistrements solo, à savoir que la probabilité de la couche d'état $P((\mathbf{E}_j)|(\mathcal{M}_j))P(\mathcal{O}|\mathcal{I})$ est identique pour tous les couples d'instruments pouvant jouer les états $(\mathbf{E}_j)$. Cependant elle ne permet pas d'identifier les bons instruments sur des duos d'instruments de même nom. En effet, quel que soit le type de mélange, la valeur de $P^{\text{trans}} P(\mathcal{O}|\mathcal{I})$ pour un couple de modèles d'instruments différents (j', j'') est toujours supérieure ou égale à sa valeur pour les couples de modèles d'instruments identiques (j', j') et (j'', j''). Il s'agit simplement de la constatation classique que plus la taille d'un modèle est élevée, mieux il approche les observations. Or un couple de modèles d'instruments identiques a une capacité de modélisation quasi équivalente à celle d'un seul modèle.

En théorie ce problème devrait diminuer en utilisant les modèles d'instruments segmentaux au lieu des modèles factoriels. Cependant la limitation du nombre de chemins partiels évalués par la recherche en faisceaux peut engendrer des transcriptions sous-optimales pour certains couples d'instruments testés mais pas pour d'autres, et donc mener à des erreurs d'identification d'instruments. Nous conservons donc la modélisation factorielle et nous testons uniquement les couples de modèles d'instruments différents

sur les mélanges d'instruments différents.

Nous fixons les hyper-paramètres à $Z = 0,95$, $\sigma^{\epsilon} = 1,4$, $w_{\text{comb}} = 0,5$, $w_{\text{desc}} = 0,5$ et $w_{\text{état}} = 1$. La matrice de confusion obtenue est présentée dans le tableau 6.1.

Le TRI moyen vaut 83%. Plus précisément 83% des extraits sont reconnus sans erreur (les deux instruments du couple sont bien identifiés), 17% des extraits avec une erreur (un seul instrument du couple est bien identifié) et aucun avec deux erreurs. La baisse de performance est assez faible par rapport aux enregistrements solo. La performance est supérieure à celle obtenue par Eggink et Brown [Egg03] avec les mêmes couples d'instruments et les hauteurs des notes connues *a priori* (taux d'exactitude de 74% correspondant à un TRI de 48%) et largement supérieure à celle d'un choix aléatoire (10% des extraits reconnus sans erreur et TRI moyen de -20%).

		Couple d'instruments identifié									
		Fl-Cl	Fl-Ob	Fl-Vn	Fl-Vc	Cl-Ob	Cl-Vn	Cl-Vc	Ob-Vn	Ob-Vc	Vn-Vc
Mélange testé	Fl-Cl	**57%**				29%	14%				
	Fl-Ob		**100%**								
	Fl-Vn	29%		**43%**					14%		14%
	Fl-Vc				**71%**			14%			14%
	Cl-Ob					**100%**					
	Cl-Vn	14%					**86%**				
	Cl-Vc	14%						**86%**			
	Ob-Vn								**100%**		
	Ob-Vc								14%	**86%**	
	Vn-Vc										**100%**

TAB. 6.1 – Matrice de confusion pour l'identification d'instruments sur des duos monocanal synthétiques.

Effectif instrumental inconnu

Lorsque le nombre d'instruments présents est inconnu, cet algorithme n'est plus applicable pour deux raisons. Premièrement, pour les raisons évoquées ci-dessus, une distribution *a priori* de type $P(\mathcal{O}|\mathcal{I}) \propto Z^{-T\sum_{j=1}^{n} N_j}$ engendre un biais systématique en faveur des orchestres constitués d'un nombre élevé d'instruments n. Deuxièmement, le coût d'évaluation de la probabilité de transcription pour chaque orchestre devient prohibitif. Une première possibilité serait d'utiliser une méthode semi-bayésienne basée sur un rapport de vraisemblance [Ros02]. Cependant l'algorithme de saut qui en découle nécessite encore beaucoup de calculs, et le choix du seuil de vraisemblance est délicat.

Nous proposons un algorithme non bayésien plus simple qui consiste à transcrire le mélange à l'aide de tous les modèles d'instruments simultanément (donc comme un quintette dans cette expérience), puis à classer comme présents uniquement les instruments pour lesquels le nombre total d'états estimés présents est important.

Cet algorithme se justifie particulièrement pour l'identification d'instruments sur de longs extraits. Dans ce cas, la tâche d'identification définie dans le paragraphe 1.2.1 n'est pas pertinente, car un instrument présent tout le long d'un morceau n'a pas la même importance qu'un instrument ne jouant qu'un court instant. Une tâche plus réaliste consiste à évaluer le temps de jeu de chaque instrument, ou bien à évaluer les instruments présents sur des trames temporelles de plus courte durée. L'algorithme que nous proposons est directement utilisable dans ce but.

Nous testons les résultats sous deux conditions différentes.

Dans une première phase, nous recherchons uniquement les noms des instruments présents, mais pas l'effectif de chacun (un duo d'instruments identiques est assimilé à un solo). Nous sélectionnons un nom d'instrument comme présent si son nombre total d'états estimés présents dépasse un certain pourcentage $N_{\min}$ du nombre de trames des observations T. Les résultats (obtenus avec les mêmes hyper-paramètres que ci-dessus) sont décrits dans la figure 6.1. Pour la valeur optimale de $N_{\min} = 76\%$, le TRI moyen vaut 56%. L'algorithme est assez robuste au choix de $N_{\min}$, puisque le TRI moyen reste supérieur à 50% pour des valeurs de $N_{\min}$ comprises entre 61% et 86%.

Dans une deuxième phase, nous recherchons à la fois les noms et les effectifs des instruments présents. Nous estimons l'effectif d'un nom d'instrument à n' si son nombre total d'états estimés présents est compris entre les pourcentages $(n'-1)N_{\min}$ et $n'N_{\min}$ du nombre de trames T. La valeur optimale de $N_{\min}$ ne change pas, mais le TRI moyen baisse à 38%. L'évaluation de ces résultats par le critère de TRI n'est cependant pas idéale car en pratique les erreurs d'effectif sont moins gênantes que les erreurs de noms d'instruments pour les applications potentielles.

Cet algorithme peut aussi être utilisé lorsque le nombre d'instruments différents présents est connu en sélectionnant les instruments dont le nombre d'états présents estimé est le plus élevé. Le TRI moyen sur les duos d'instruments différents vaut alors 79%, avec 21% des extraits reconnus avec une erreur et aucun avec deux erreurs. Cette performance est légèrement inférieure à celle du critère bayésien mais engendre un gain de temps de calcul d'un facteur quatre environ. Le gain pourrait être plus important encore en augmentant le nombre d'instruments à tester ou le nombre de sources présentes.

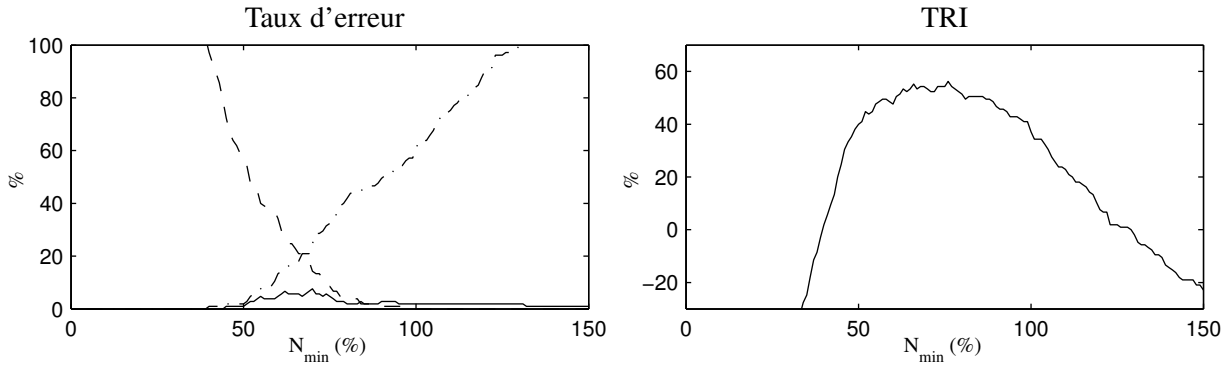

FIG. 6.1 – Performance d'identification d'instruments sur des duos monocanal synthétiques avec effectif instrumental inconnu (à gauche, trait plein : taux de substitution, tirets : taux d'insertion, tirets mixtes : taux de suppression).

6.3.2 Duo monocanal synthétique de clarinette et violon

Dans notre deuxième expérience, nous étudions un duo synthétique formé des extraits de clarinette (s_1) et de violon (s_2) utilisés dans le paragraphe 5.3.2. Le mélange est obtenu simplement par addition des formes d'onde des sources. Les sources ont un recouvrement limité en temps-fréquence, mais leurs mélodies s'entrecroisent dans la même zone de hauteur appartenant aux tessitures des deux instruments. Nous changeons l'hyper-paramètre Z à $0,96$.

La figure 6.2 présente les résultats de l'identification de notes pour les deux modèles de couche d'état. Ces résultats peuvent être comparés aux transcriptions exactes des sources dans les figures 5.11 et 5.12. Nous constatons que la plupart des notes ont leur hauteur et leur instrument bien identifiés, mais quelques erreurs subsistent.

Avec le modèle factoriel, de nombreuses notes de très courte durée sont présentes. De plus certaines notes sont reconnues comme faisant partie à la fois des deux instruments, par exemple la note $h = 84$ au temps $t \simeq 3,1$ s ou la note $h = 73$ au temps $t \simeq 7,4$ s, et d'autres sont essentiellement attribuées au mauvais instrument, par exemple la note $h = 76$ au temps $t \simeq 4,4$ s. Des erreurs de ce type sont inévitables car la performance d'identification de notes est limitée par la performance d'identification d'instruments sur des notes isolées, qui est naturellement inférieure à celle sur des extraits solo. En reprenant l'expérience d'identification d'instruments du chapitre précédent sur les notes isolées de la base de données *SOL* [SOL] de hauteur MIDI comprise entre 60 et 72 et de nuance *mezzo*, nous avons constaté une baisse de TRI pour tous les instruments, avec un TRI moyen de 63% seulement. Cette baisse de performance est comparable à celle obtenue par Eggink et Brown [Egg03].

Avec le modèle segmental, certaines erreurs du modèle factoriel sont corrigées, mais d'autres apparaissent. Les erreurs restantes sont une insertion et une suppression pour la clarinette, et une substitution et une suppression pour le violon (sur un total de cinquante notes). Les deux erreurs pour le violon, situés aux instants $t \simeq 1,7$ s et $t \simeq 9,3$ s, correspondent à des situations où les deux mélodies sont à l'unison. La note de clarinette de hauteur $h = 81$ supprimée au temps $t \simeq 5$ s est analysée comme la partie réverbérée d'une note de violon de même hauteur, la durée de cette partie étant manifestement trop longue par rapport à la durée de réverbération des autres notes de violon.

La performance du modèle segmental est donc satisfaisante, mais pourrait être encore améliorée en rajoutant des contraintes, par exemple sur la durée de la réverbération des notes, sur le rythme ou sur les intervalles entre notes successives. Ces contraintes permettraient aussi de réduire la taille de l'espace d'état et donc de rendre la recherche en faisceaux plus efficace pour le même temps de calcul.

Les résultats d'extraction de sources sont décrits dans le tableau 6.2. La transcription par oracle consiste à utiliser les vrais descripteurs estimés sur les sources solo.

Nous observons que la performance est globalement satisfaisante, et plutôt limitée par la méthode de filtrage que par la qualité des modèles d'instruments. Le RSD moyen obtenu avec les vrais spectres à court terme des sources n'est que 2 dB supérieur à celui obtenu avec les spectres modèles. Cela prouve que les spectres typiques contenus dans les modèles d'instruments permettent d'obtenir des coefficients de filtrage proches des coefficients théoriques. Les deux méthodes de filtrage donnent sur cet exemple des résultats similaires. Pour augmenter encore la performance, il faudrait modifier les méthodes de filtrage et prendre en compte les relations de phase entre les sous-bandes des sources, par exemple avec un modèle sinusoïdal harmonique.

Nous remarquons aussi que la performance d'extraction est largement liée à la performance de transcription. Le RSD moyen avec la transcription par oracle (20 dB) est meilleur que celui avec la transcription par modèle segmental (16 dB), lui-même meilleur que celui avec la transcription par modèle factoriel (9 dB). En particulier les notes de très courte durée estimées à la transcription par le modèle segmental engendrent une quantité importante d'artefacts à la séparation. Cette remarque semble contredire la similarité des performances d'extraction obtenues avec des modèles de sources par MG et par MMC notée par Benaroya [Ben03]. La différence pourrait être due à l'usage d'un modèle segmental au lieu d'un simple MMC, à la pondération de la loi de Bayes, et à l'augmentation du nombre d'atomes.

Le tableau 6.3 montre les résultats de modification de scène sonore. La modification choisie est la multiplication de l'amplitude de la clarinette par 2 (soit un ajout de 6 dB).

Nous notons que la performance est notablement supérieure à celle d'extraction de sources. Les zones temps-fréquence mal extraites pour cause de masquage se retrouvent pour la plupart à nouveau masquées après remixage. À l'écoute, la qualité des remixes estimés semble encore meilleure que ne le laissent entendre les critères numériques de RRD, en particulier avec la transcription par modèle factoriel. En effet, les erreurs réalisées sur les remixes estimés sont masquées auditivement, et donc moins dérangeantes que celles sur les sources estimées.

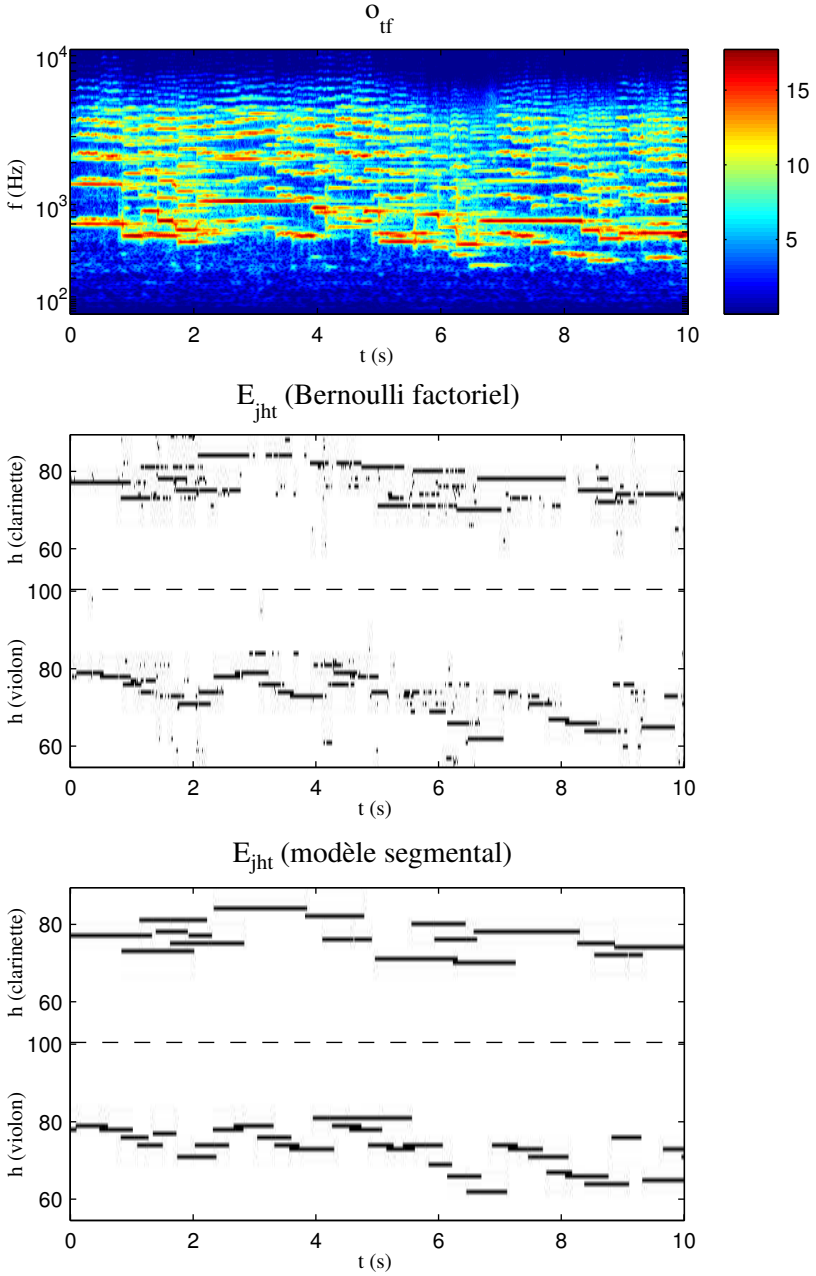

FIG. 6.2 – Notes identifiées sur le duo monocanal synthétique.

Méthode de transcription	Méthode de filtrage	Performance $\widehat{s_1}$ (dB) RSD	RSI	RSA	Performance $\widehat{s_2}$ (dB) RSD	RSI	RSA
Oracle	Wiener	22	39	22	19	37	19
	SP	21	34	22	17	30	19
Factoriel	Wiener	11	25	13	6,2	28	8,4
	SP	11	21	14	6,5	24	8,7
Segmental	Wiener	19	39	19	14	29	16
	SP	18	33	20	13	24	15

TAB. 6.2 – Performance d'extraction de sources sur le duo monocanal synthétique.

Méthode de transcription	Performance $\widehat{x_{\mathrm{rmx}}}$ (dB) RRD	RRES	RRA
Oracle	28	44	28
Factoriel	18	26	20
Segmental	25	40	27

TAB. 6.3 – Performance de modification de scène sonore sur le duo monocanal synthétique.

6.3.3 Duo monocanal réel de violoncelle et flûte

Dans notre troisième expérience, nous nous intéressons à un extrait d'un duo réel de violoncelle et flûte. Cette fois, les mélodies sont dans des zones de hauteur différentes, et les notes de flûte appartiennent à la tessiture des deux instruments. Ce mélange est néanmoins difficile pour les méthodes usuelles de transcription car dix notes de flûte (sur douze) forment un intervalle harmonique avec les notes de violoncelle présentes au même instant. Nous utilisons les mêmes hyper-paramètres que dans l'expérience précédente.

Les résultats d'identification d'instruments sont présentés dans la figure 6.3 et le tableau 6.4.

Lorsque l'effectif instrumental est connu, le couple violoncelle-flûte est celui qui engendre la meilleure probabilité de transcription. Conformément à nos remarques du paragraphe 6.3.1, nous voyons que la probabilité d'un couple d'instruments différents (hors diagonale) est toujours supérieure à celle des couples d'instruments identiques correspondants (sur la diagonale, même ligne ou même colonne). Nous remarquons aussi un grand écart de probabilité entre les couples contenant le violoncelle et les autres. En effet, les couples ne contenant que des instruments de tessiture aiguë ne peuvent pas modéliser la présence de notes graves, ce qui forme une erreur grossière. Les couples contenant le violoncelle mais pas la flûte peuvent modéliser la présence de toutes les notes, mais pas avec la bonne enveloppe spectrale, ce qui est une erreur plus subtile.

Lorsque l'effectif instrumental est inconnu, le violoncelle et la flûte sont les deux instruments qui obtiennent le plus d'états présents. Le nombre d'états de flûte présents reste beaucoup plus faible que celui du violoncelle, car la flûte joue *staccato* et sa partition comprend des silences.

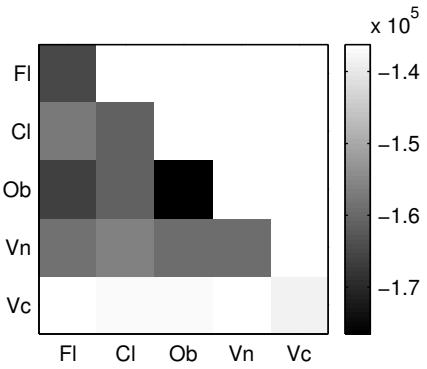

FIG. 6.3 – Probabilité des transcriptions obtenues avec divers couples d'instruments sur le duo monocanal réel.

La figure 6.4 montre les notes identifiées par les deux modèles de couche d'état. La transcription par modèle segmental correspond à la partition exacte de l'extrait, plus une note de violoncelle aiguë insérée au temps $t \simeq 7,3$ s. La transcription par modèle factoriel est similaire, avec plusieurs notes de très courte durée supplémentaires. Les algorithmes peuvent donc distinguer plusieurs notes au sein d'un

Instrument	Pourcentage d'états présents
Flûte	41%
Clarinette	8%
Hautbois	7%
Violon	17%
Violoncelle	124%

TAB. 6.4 – Pourcentage de notes présentes pour chaque instrument sur le duo monocanal réel par rapport au nombre total de trames.

spectre harmonique en fonction de son enveloppe spectrale.

Nous utilisons les résultats de transcription pour l'extraction de sources et la modification de scènes avec les deux méthodes de filtrage. L'écoute des résultats appelle différentes remarques.

Le violoncelle est bien séparé quelle que soit la méthode de transcription, et le filtrage pseudo-Wiener diminue un peu les interférences par rapport à la SP. La composante bruitée de souffle sur les attaques de flûte est attribuée au violoncelle, car elle n'est pas prise en compte dans les modèles d'instruments. Il est intéressant de noter que la fausse note insérée dans la partition estimée du violoncelle n'affecte pas la qualité de séparation.

La flûte est moins bien séparée et sonne de façon plus artificielle. La transcription par modèle segmental et la séparation par SP améliorent un peu les résultats en évitant des "trous" de courte durée au sein des notes ou une détérioration excessive du timbre de l'instrument.

Enfin, le remix estimé est toujours de bonne qualité quelle que soit la méthode utilisée.

6.3.4 Résumé des résultats

Nous résumons les résultats de ce chapitre en distinguant deux types de tâches.

Les résultats sur les tâches d'identification d'instruments et de modification de scène sont très satisfaisants. Pour ces tâches, l'utilisation de modèles d'instruments permet bien de reconnaître et de remixer des mélanges d'instruments de tessitures proches. De plus l'utilisation de modèles d'instruments complexes n'est pas nécessaire : les modèles factoriels suffisent.

Les résultats sur les tâches d'identification de notes et d'extraction de sources sont aussi satisfaisants, mais un peu moins bons dans l'absolu car ces tâches sont plus difficiles. Dans ce cas, les modèles d'instruments sont utiles pour identifier les hauteurs et les instruments associés aux notes et pour estimer les coefficients de filtrage sur des mélanges d'instruments de tessitures proches ou de fort recouvrement temps-fréquence. Une amélioration supplémentaire de la performance semble cependant nécessiter des modèles d'instruments plus complexes que les modèles segmentaux et des techniques de filtrage prenant en compte les relations de phase entre sous-bandes.

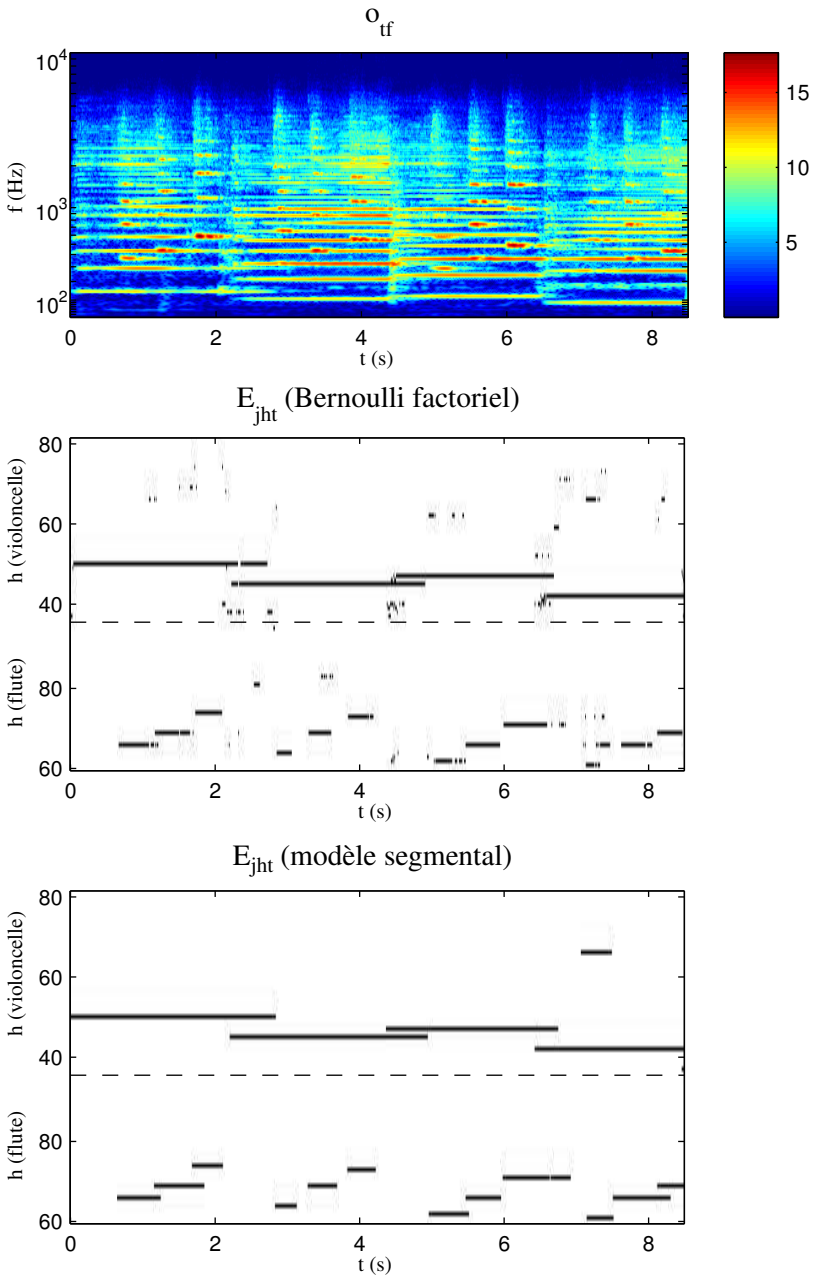

FIG. 6.4 – Notes identifiées sur le duo monocanal réel.

Chapitre 7

Transcription et séparation d'enregistrements multicanal

Ce septième et dernier chapitre décrit l'application des modèles d'instrument à la transcription et la séparation d'enregistrements de musique de chambre multicanal. Nous considérons trois types de mélanges : stéréo panoramiques, stéréo convolutifs de type AB étroit et multipistes. Nous proposons dans le paragraphe 7.1 divers moyens de combiner les modèles d'instruments en modèle de mélange selon le type de mélange considéré. Nous décrivons les méthodes de filtrage adaptées dans le paragraphe 7.2. Enfin, nous étudions des mélanges synthétiques dans le paragraphe 7.3. Nous confrontons les résultats de séparation à ceux des algorithmes usuels de séparation aveugle, et nous comparons la performance de transcription avec et sans informations spatiales en utilisant les algorithmes de transcription monocanal du chapitre précédent.

Les modèles proposés dans ce chapitre résultent d'une recherche originale. En particulier, l'utilisation conjointe d'informations spatiales et de modèles de timbre sur des enregistrements convolutifs longs n'a pas d'équivalent dans la littérature à notre connaissance. À nouveau, nos algorithmes ne sont pas comparables au seul algorithme existant d'ASA computationnelle de mélanges musicaux multicanal [Sak03], testé sur un seul mélange convolutif court d'instruments MIDI. Nous avons publié une méthode semblable de séparation de mélanges stéréo instantanés dans l'article [Vin04d]. L'utilisation d'un autre mélange test et d'autres critères de performance rend les résultats de l'article légèrement différents de ceux présentés ici.

7.1 Transcription

La loi de Bayes pondérée correspondant au critère MAP de transcription est inchangée par rapport aux mélanges monocanal, et exprimée dans l'équation 6.3. Cependant, le choix des observations et les hypothèses de combinaison des modèles d'instruments en modèle de mélange dépendent du type de mélange considéré.

7.1.1 Choix des observations

Mélange stéréo panoramique

Les mélanges panoramiques (instantanés) forment un cas idéal où les informations spatiales sont très précises et facilement utilisables.

Un mélange stéréo instantané est décrit par ses gains de mélange $(a_{ij})_{1\leq i\leq 2,\,1\leq j\leq n}$ et par les spectres de puissance modèles du bruit de fond sur les deux canaux $(\mathbf{n}'_i)_{1\leq i\leq 2}$. Nous posons comme contraintes

que les gains de mélange sont positifs et que le gain total $(a_{1,j}^2 + a_{2,j}^2)^{1/2}$ est unitaire pour chaque source j. Les spectres de log-puissance des deux canaux $(\mathbf{o}_{1,t})$ et $(\mathbf{o}_{2,t})$ ne sont pas indépendants sachant les descripteurs, ce qui complique l'expression de leur probabilité conjointe. Nous formons donc le vecteur des observations $\mathbf{o}_t$ à l'instant t à l'aide de deux quantités supposées conditionnellement indépendantes : le spectre de log-puissance total $\mathbf{o}_t^{\mathrm{tot}}$ défini par $o_{tf}^{\mathrm{tot}} = \log[(\|x_1^{tf}\|^2 + \|x_2^{tf}\|^2)/g_f^2 + 1]$ et le vecteur de différence de volume inter-canal $\mathbf{d}_{21,t}^{\mathrm{vol}}$ défini par l'équation 2.10.

Nous approchons les observations par

$$\mathbf{o}_t^{\mathrm{tot}} = \log\Big(\sum_{j=1}^{n} \mathbf{m}'_{jt} + \mathbf{n}'_1 + \mathbf{n}'_2\Big) + \boldsymbol{\epsilon}_t^{\mathrm{tot}}, \tag{7.1}$$

$$\mathbf{d}_{21,t}^{\mathrm{vol}} = \log\Big(\sum_{j=1}^{n} a_{2,j}^2 \mathbf{m}'_{jt} + \mathbf{n}'_2\Big) - \log\Big(\sum_{j=1}^{n} a_{1,j}^2 \mathbf{m}'_{jt} + \mathbf{n}'_1\Big) + \boldsymbol{\epsilon}_t^{\mathrm{vol}} \tag{7.2}$$

où $\boldsymbol{\epsilon}_t^{\mathrm{tot}}$ est un bruit gaussien à covariance diagonale isotrope d'écart-type $\sigma^{\epsilon\,\mathrm{tot}}$ et $\boldsymbol{\epsilon}_t^{\mathrm{vol}}$ un bruit gaussien à covariance diagonale dépendant de t. Nous notons $\sigma_{tf}^{\epsilon\,\mathrm{vol}}$ l'écart-type de $\boldsymbol{\epsilon}_t^{\mathrm{vol}}$ au point (t, f). La probabilité des observations en fonction des descripteurs vaut

$$P(\mathbf{o}_t|\boldsymbol{\Theta}, (\mathbf{p}_{jt}), (\mathcal{M}_j)) = \prod_{f=0}^{F-1} \mathcal{N}(\epsilon_{tf}^{\mathrm{tot}}; 0, \sigma^{\epsilon\,\mathrm{tot}})\, \mathcal{N}(\epsilon_{tf}^{\mathrm{vol}}; 0, \sigma_{tf}^{\epsilon\,\mathrm{vol}}). \tag{7.3}$$

Expérimentalement, l'écart-type de l'erreur résiduelle sur la différence de volume inter-canal $\sigma_{tf}^{\epsilon\,\mathrm{vol}}$ peut être relié à la cohérence inter-canal $d_{21,tf}^{\mathrm{coh}}$ au même point (t, f) par

$$\sigma_{tf}^{\epsilon\,\mathrm{vol}} = \sigma^{\epsilon\,\mathrm{vol}}\,(1 - d_{21,tf}^{\mathrm{coh}})^{\lambda^{\mathrm{vol}}}, \tag{7.4}$$

où λ^{vol} est un réel positif. Cette équation assure que l'écart-type est faible lorsque la cohérence est élevée et *vice-versa*. Cette propriété remarquable permet d'accorder plus d'importance aux informations spatiales fiables situées dans les zones temps-fréquence où une source masque les autres sources ainsi que sa propre réverbération. De plus nous remplaçons la définition de la cohérence inter-canal de l'équation 2.13 par $d_{21,tf}^{\mathrm{coh}} = (1 + \Re\langle x_1^{tf}, x_2^{tf}\rangle/\|x_1^{tf}\|\|x_2^{tf}\|)/2$. Cette définition est plus adaptée aux mélanges instantanés car elle tient compte de la différence de phase inter-canal, supposée nulle aux points (t, f) où le son est constitué d'une seule source.

Une autre définition de l'écart-type $\sigma_{tf}^{\epsilon\,\mathrm{vol}}$ consiste à le relier aux descripteurs par une approximation de type combinaison de modèles [Gal95]. L'écart-type dépend alors de la cohérence prédite par le modèle au lieu de la cohérence observée. Cette définition a donc des propriétés semblables, mais augmente le temps de calcul, raison pour laquelle nous ne l'utilisons pas par la suite.

Mélange stéréo "AB étroit"

À l'opposé des mélanges panoramiques, les mélanges stéréo enregistrés par une paire AB étroite sont assez difficiles à transcrire et à séparer, les informations spatiales étant moins facilement interprétables du fait de la réverbération. Lorsque les instruments sont suffisamment distants de la paire, les spectres de puissance des deux canaux peuvent être supposés égaux. L'information spatiale pertinente est alors fournie par la différence de phase inter-canal au lieu de la différence de volume.

Nous paramétrons les filtres de mélange à l'aide de leurs réponses fréquentielles en amplitude et en phase relative définies dans l'annexe B.1.4. Les réponses fréquentielles en amplitude sont considérées

égales à une réponse moyenne $\mathbf{a}^{\mathrm{amp}}$ pour toutes les sources. Les réponses fréquentielles en phase relative $(\mathbf{a}^{\mathrm{pha}}_{21,j})_{1\leq j\leq n}$ permettent d'estimer les délais entre les deux capsules à chaque fréquence pour chaque source liés aux azimuts.

Nous choisissons comme paramètres de mélange les réponses fréquentielles $\mathbf{a}^{\mathrm{amp}}$ et $(\mathbf{a}^{\mathrm{pha}}_{21,j})_{1\leq j\leq n}$, ainsi que le spectre de puissance $\mathbf{n}'$ et la phase relative $\mathbf{a}'^{\mathrm{pha}}_{21}$ du bruit de fond total. À première vue, l'attribution de propriétés spatiales au bruit de fond est valable surtout dans les mélanges synthétiques, et moins dans les mélanges réels contenant du bruit de fond diffus. Mais le modèle choisi par la suite utilise les informations spatiales principalement dans les zones temps-fréquence de cohérence inter-canal élevée. Le modèle de phase relative pour un bruit de fond diffus n'est donc presque pas pris en compte. Nous regroupons au sein des observations au temps t le spectre de log-puissance total $\mathbf{o}^{\mathrm{tot}}_t$ et le vecteur de différence de phase inter-canal $\mathbf{d}^{\mathrm{pha}}_{21,t}$ défini par l'équation 2.9.

Notre hypothèse habituelle d'additivité des spectres des sources en puissance se manifeste en terme de phase par une décorrélation de leurs formes d'onde dans chaque sous-bande. Nous utilisons la modélisation

$$\mathbf{o}^{\mathrm{tot}}_t = \log\Big(\sum_{j=1}^{n} \mathbf{a}^{\mathrm{amp}\,2} \bullet \mathbf{m}'_{jt} + \mathbf{n}'\Big) + \boldsymbol{\epsilon}^{\mathrm{tot}}_t, \tag{7.5}$$

$$\mathbf{d}^{\mathrm{pha}}_{21,t} = \angle\Big(\sum_{j=1}^{n} \mathbf{a}^{\mathrm{amp}\,2} \bullet \mathbf{m}'_{jt} \bullet \exp(i\,\mathbf{a}^{\mathrm{pha}}_{21,j}) + \mathbf{n}' \bullet \exp(i\,\mathbf{a}'^{\mathrm{pha}}_{21})\Big) + \boldsymbol{\epsilon}^{\mathrm{pha}}_t \mod 2\pi. \tag{7.6}$$

Dans ces équations, $\boldsymbol{\epsilon}^{\mathrm{tot}}_t$ est un bruit gaussien à covariance diagonale isotrope d'écart-type $\sigma^{\epsilon\,\mathrm{tot}}$ et $\boldsymbol{\epsilon}^{\mathrm{pha}}_t$ un bruit à valeurs dans $]-\pi,\,\pi]$ de densité proportionnelle à une gaussienne à covariance diagonale dépendant de t. En notant $\sigma^{\epsilon\,\mathrm{pha}}_{tf}$ l'écart-type de cette gaussienne au point (t, f), la probabilité des observations sachant les descripteurs prend la forme

$$P(\mathbf{o}_t|\boldsymbol{\Theta}, (\mathbf{p}_{jt}), (\mathcal{M}_j)) = \prod_{f=0}^{F-1} \mathcal{N}(\epsilon^{\mathrm{tot}}_{tf}; 0, \sigma^{\epsilon\,\mathrm{tot}}) \, \frac{\mathcal{N}(\epsilon^{\mathrm{pha}}_{tf}; 0, \sigma^{\epsilon\,\mathrm{pha}}_{tf})}{\displaystyle\int_{-\pi}^{\pi} \mathcal{N}(\epsilon; 0, \sigma^{\epsilon\,\mathrm{pha}}_{tf})\, d\epsilon}. \tag{7.7}$$

Comme précédemment nous relions l'écart-type $\sigma^{\epsilon\,\mathrm{vol}}_{tf}$ à la cohérence inter-canal $d^{\mathrm{coh}}_{21,tf}$ définie par l'équation 2.13 au même point (t, f) par la loi exponentielle

$$\sigma^{\epsilon\,\mathrm{pha}}_{tf} = \sigma^{\epsilon\,\mathrm{pha}}\,(1 - d^{\mathrm{coh}}_{21,tf})^{\lambda^{\mathrm{pha}}}. \tag{7.8}$$

Mélange multipistes

Un mélange multipistes contient à la fois des signaux enregistrés par un micro stéréo et par des micros d'appoint. L'utilisation des signaux des micros d'appoint suffit cependant pour la transcription, car le contraste spatial entre les sources est plus important et le niveau de réverbération plus faible que sur le micro stéréo (quel que soit le type de paire utilisé).

Nous numérotons les micros d'appoint de 3 à $n + 2$, la source j correspondant au micro $j + 2$. La directivité des micros d'appoint se manifeste principalement par le fait que la réponse fréquentielle en amplitude $\mathbf{a}^{\mathrm{amp}}_{ij}$ d'un filtre a_{ij} est plus élevée pour un filtre "direct" ($i = j + 2$) que pour un filtre "croisé" ($i \neq j + 2$). La réponse fréquentielle en phase n'apporte pas d'information pertinente. Un mélange multipistes peut donc être considéré en quelque sorte comme un mélange panoramique où les gains de mélange varient selon la fréquence.

Nous définissons le système de mélange par les réponses fréquentielles en amplitude des filtres de mélange $(\mathbf{a}_{ij}^{\mathrm{amp}})_{3\leq i\leq n+2,\,1\leq j\leq n}$ et par les spectres de puissance modèles du bruit de fond sur chaque canal $(\mathbf{n}'_i)_{3\leq i\leq n+2}$. Nous construisons les observations $\mathbf{o}_t$ à l'instant t par concaténation des spectres de log-puissance $(\mathbf{o}_{it})_{3\leq i\leq n+2}$ et des différences de volume inter-canal $(\mathbf{d}_{i'it}^{\mathrm{vol}})_{3\leq i\leq n+2,\,i+1\leq i'\leq n+2}$.

Nous modélisons ces observations par

$$\mathbf{o}_{it} = \log\Big(\sum_{j=1}^{n} \mathbf{a}_{ij}^{\mathrm{amp}\,2} \bullet \mathbf{m}'_{jt} + \mathbf{n}'_i\Big) + \boldsymbol{\epsilon}_{it}, \tag{7.9}$$

$$\mathbf{d}_{i'it}^{\mathrm{vol}} = \log\Big(\sum_{j=1}^{n} \mathbf{a}_{i'j}^{\mathrm{amp}\,2} \bullet \mathbf{m}'_{jt} + \mathbf{n}'_{i'}\Big) - \log\Big(\sum_{j=1}^{n} \mathbf{a}_{ij}^{\mathrm{amp}\,2} \bullet \mathbf{m}'_{jt} + \mathbf{n}'_i\Big) + \boldsymbol{\epsilon}_{i'it}^{\mathrm{vol}} \tag{7.10}$$

où les $(\boldsymbol{\epsilon}_{it})_{3\leq i\leq n+2}$ sont des bruits gaussiens à covariance diagonale isotrope d'écart-type σ^{ϵ} et les $(\boldsymbol{\epsilon}_{i'it}^{\mathrm{vol}})_{3\leq i\leq n+2,\,i+1\leq i'\leq n+2}$ des bruits gaussiens à covariance diagonale dépendant de t. L'équation 7.4 lie chaque covariance à la cohérence inter-canal correspondante $\mathbf{d}_{i'it}^{\mathrm{coh}}$, définie par l'équation 2.13. La probabilité des observations en fonction des descripteurs s'exprime par

$$P(\mathbf{o}_t|\boldsymbol{\Theta}, (\mathbf{p}_{jt}), (\mathcal{M}_j)) = \prod_{i=3}^{n+2}\prod_{f=0}^{F-1} \mathcal{N}(\epsilon_{itf}; 0, \sigma^{\epsilon}) \prod_{i'=i+1}^{n+2} \mathcal{N}(\epsilon_{i'itf}^{\mathrm{vol}}; 0, \sigma_{i'itf}^{\epsilon\,\mathrm{vol}}). \tag{7.11}$$

Autres types de mélange

La majorité des autres mélanges multicanal peuvent se modéliser en combinant les trois modèles de base que nous venons de décrire. Par exemple un mélange enregistré par un couple stéréo ORTF est à mi-chemin entre un mélange stéréo panoramique (avec gains de mélange variant légèrement en fréquence) et un mélange stéréo "AB étroit". Les observations pertinentes peuvent alors regrouper le spectre de log-puissance totale et les différences de volume et de phase inter-canal et être approchées par les équations 7.9, 7.10 et 7.6.

7.1.2 Estimation des gains ou des filtres de mélange

La transcription d'un mélange multicanal consiste à estimer conjointement les gains ou les filtres de mélange, les bruits de fond, les états et les descripteurs des sources.

Dans le cas d'un mélange panoramique, cette estimation peut être effectuée en deux étapes.

Dans un premier temps, nous appliquons l'algorithme de Abrard, Deville et White [Abr01] pour retrouver les gains relatifs de mélange dans un ordre non fixé et sans *a priori* sur les sources. Le principe de cet algorithme est que si la source j masque les autres au point (t, f), alors la différence de volume inter-canal $d_{21,tf}^{\mathrm{vol}}$ est égale au gain relatif $\theta_j = \log(a_{2,j}^2/a_{1,j}^2)$ et la cohérence inter-canal $d_{21,tf}^{\mathrm{coh}}$ est proche de 1. Ces points sont majoritaires lorsque le recouvrement des sources en temps-fréquence est limité. Nous commençons par estimer grossièrement les gains relatifs en localisant les pics de l'histogramme des différences de volume inter-canal. Puis nous sélectionnons autour de chaque pic 1% des points pour lesquels la cohérence inter-canal est la plus élevée et nous estimons le gain relatif plus précisément par la médiane des différences de volume inter-canal sur ces points.

Dans un deuxième temps, puisque les paramètres de mélange sont estimés dans un ordre quelconque, il reste à associer à chaque source les paramètres correspondants. Pour cela, nous proposons de tester toutes les configurations spatiales possibles par permutation des sources et de sélectionner celle qui maximise la probabilité de transcription P^{trans}.

Dans le cas d'un mélange stéréo "AB étroit", nous avons constaté que la méthode d'estimation aveugle des azimuts des sources proposée par Roman, Wang et Brown [Rom03] donnait de mauvais résultats en présence de réverbération, même après sélection des points temps-fréquence de cohérence inter-canal élevée. Dans le cadre de cette étude, nous nous limitons à approcher les réponses en phase relative des filtres de mélange par les réponses de délais purs fixés *a priori* et non réestimés. La réponse d'amplitude moyenne est initialisée par $a_f^{\text{amp}} = 1$ à toutes les fréquences f et réestimée lors de la transcription par un algorithme de Newton avec une contrainte de continuité identique à celle sur le spectre modèle du bruit de fond.

De même, dans le cas d'un mélange multipistes, l'estimation des réponses fréquentielles en amplitude des filtres de mélange pour les micros d'appoint est difficile sans *a priori* sur les sources, car ces réponses varient de façon importante selon la fréquence. Pour l'expérience effectuée dans ce chapitre, nous choisissons de les initialiser à chaque fréquence f par $a_{ijf}^{\text{amp}} = 1$ pour les filtres "directs" et $a_{ijf}^{\text{amp}} = 0,02$ pour les filtres "croisés", puis de les réestimer lors de la transcription par un algorithme de Newton avec une contrainte de continuité. Les réponses correspondant à des sources différentes ont des valeurs suffisamment différentes, de sorte qu'une légère erreur à l'initialisation ne compromet pas la qualité de la transcription estimée.

Dans cette expérience le micro stéréo utilisé en sus des micros d'appoint est une paire AB étroite. Afin d'estimer les images spatiales des sources sur le micro stéréo, nous estimons les paramètres de mélange pour le micro stéréo ($\mathbf{a}^{\text{amp}}$ et $\mathbf{n}'$, plus un gain de mélange pour chaque source) avec les algorithmes utilisés pour les mélanges "AB étroits", mais en utilisant les descripteurs fixés après la transcription basée sur les observations aux micros d'appoint uniquement.

7.1.3 Algorithmes de transcription

Les algorithmes de transcription sont similaires pour tous les types de mélange. Les états sont estimés à l'aide des algorithmes de saut et de recherche en faisceaux déjà décrits dans les chapitres précédents. Les descripteurs, les bruits de fond et éventuellement les réponses en amplitude des filtres de mélange sont estimés à l'aide d'un algorithme de Newton au second ordre approché.

L'algorithme de Newton pour les descripteurs est précédé d'une réestimation aléatoire, car leur probabilité *a posteriori* contient plus de maxima locaux que dans le cas monocanal. Dans un souci de limitation du temps de calcul, nous utilisons uniquement des modèles d'instruments sans composantes de variation ($K_j = 0$), cette réestimation aléatoire devenant très coûteuse lorsque le nombre de descripteurs augmente.

L'algorithme de réestimation aléatoire fonctionne comme suit. Dans l'algorithme de saut à l'étape 3d, la log-énergie e_{jht} d'une nouvelle note est initialisée par l'équation 4.24 en remplaçant $\mathbf{o}_t$ par $\mathbf{o}_t^{\text{tot}}$ ou par $\mathbf{o}_{j+2,t}$, et la probabilité P_t^{trans} est calculée. Ensuite une nouvelle valeur $e_{jht} + \Delta e_{jht}$ est tirée aléatoirement avec $P(\Delta e_{jht}) = \mathcal{N}(\Delta e_{jht}; 0, \sigma_{jh}^e)$. Si cette valeur conduit à une augmentation de P_t^{trans}, elle est conservée en effectuant $e_{jht} \leftarrow e_{jht} + \Delta e_{jht}$, sinon rien ne change. Ce tirage aléatoire est itéré $N_{\text{aléa}}$ fois. Dans la suite, une valeur de $N_{\text{aléa}} = 10$ est utilisée. Dans la recherche par faisceaux, les étapes 4b et 5b sont modifiées de façon similaire. Chaque itération est divisée en autant d'itérations qu'il y a de notes présentes dans l'état testé, et les log-puissances des notes sont réestimées successivement par tirage aléatoire.

Notons que cet algorithme est relativement simpliste et n'entend pas être optimal. Des méthodes plus complexes de type Monte Carlo par chaînes de Markov [Fév04b] pourraient se révéler plus efficaces.

7.2 Séparation

7.2.1 Filtrage pseudo-Wiener sur chaque canal

L'extension la plus directe des méthodes de filtrage monocanal consiste à les utiliser sur chaque canal pour extraire les images spatiales des sources ou effectuer une modification de scène sonore. Par exemple, dans le cas d'un mélange "AB étroit" l'image de la source j sur le canal i peut être estimée par filtrage pseudo-Wiener en chaque point temps-fréquence par $\widehat{s^{tf}_{\mathrm{img}\,ij}} = \Pi_{ijtf}\, x_i^{tf}$ où

$$\Pi_{ijtf} = \frac{a_f^{\mathrm{amp}\,2}\, m'_{jtf}}{\sum_{j=1}^{n} a_f^{\mathrm{amp}\,2}\, m'_{jtf} + n'_f}, \tag{7.12}$$

puis l'image globale $\widehat{s_{\mathrm{img}\,ij}}$ est reconstruite à partir des $(\widehat{s^{tf}_{\mathrm{img}\,ij}})_{0\leq f\leq F-1,\, 0\leq t\leq T-1}$ par mise côte-à-côte des trames et inversion du banc de filtres. Des équations de filtrage similaires existent pour les autres types de mélanges. Dans la situation optimale où les sources sont disjointes en temps-fréquence, ces équations permettent d'estimer les images spatiales des sources sans indétermination de filtrage car les gains de filtrage (Π_{ijtf}) sont des réels positifs.

7.2.2 Pseudo-inversion locale

Comme dans le cas monocanal, la performance de cette méthode est limitée lorsque les sources possèdent un certain recouvrement en temps-fréquence. Dans ce cas, il est possible de déterminer la phase locale des sources masquées en combinant les phases observées sur tous les capteurs à la fois au lieu d'un seul. Le filtrage variant dans le temps consiste alors à estimer en chaque point (t, f) le vecteur des sources $\mathbf{s}^{tf}$ en fonction du vecteur des canaux du mélange $\mathbf{x}^{tf}$ par $\widehat{\mathbf{s}^{tf}} = \mathbf{\Pi}_{tf}\, \mathbf{x}^{tf}$, où la matrice de démixage $\mathbf{\Pi}_{tf}$ remplace un simple gain [Gri03a]. Les sources estimées sont ensuite reconstruites comme précédemment à partir de leurs estimations dans chaque sous bande et dans chaque trame.

Pour les mélanges panoramiques, le pendant du filtrage de Wiener est la pseudo-inversion locale (en chaque point temps-fréquence) de la matrice formée des amplitudes estimées des sources sur chaque capteur. Il est possible de montrer que les deux méthodes reposent sur les mêmes hypothèses probabilistes, à savoir que les signaux à estimer sont de distribution gaussienne et décorrélés. Le vecteur des sources "étendu" $\mathbf{s}^{+tf} = [s_1^{tf}, \ldots, s_n^{tf}, n_1^{tf}, n_2^{tf}]^T$, comprenant également les formes d'onde des bruits de fond, est calculé par $\widehat{\mathbf{s}^{+tf}} = \mathbf{\Pi}_{tf}\, \mathbf{x}^{tf}$ avec

$$\mathbf{\Pi}_{tf} = \mathbf{A}_{tf}^T (\mathbf{A}_{tf} \mathbf{A}_{tf}^T)^{-1} \tag{7.13}$$

et

$$\mathbf{A}_{tf} = \begin{pmatrix} |a_{11}| {m'}_{1,tf}^{1/2} & \ldots & |a_{1,n}| {m'}_{n,tf}^{1/2} & {n'}_1^{1/2} & 0 \\ |a_{21}| {m'}_{1,tf}^{1/2} & \ldots & |a_{2,n}| {m'}_{n,tf}^{1/2} & 0 & {n'}_2^{1/2} \end{pmatrix}. \tag{7.14}$$

Lorsque la reconstruction du banc de filtres est exacte, la reconstruction parfaite du mélange est garantie : $x_1 = \sum_{j=1}^{n} a_{1,j}\, \widehat{s_j} + \widehat{n_1}$ et $x_2 = \sum_{j=1}^{n} a_{2,j}\, \widehat{s_j} + \widehat{n_2}$. Cette méthode s'étend aux autres types de mélange en remplaçant les coefficients de la matrice $\mathbf{A}_{tf}$ par des valeurs complexes prenant en compte la phase des filtres de mélange.

La qualité des sources extraites par cette méthode dépend fortement du type de mélange et de la précision d'estimation des gains ou des filtres de mélange. En effet les sources estimées peuvent prendre des valeurs très différentes de celles attendues en fonction de leurs descripteurs pour satisfaire la contrainte de reconstruction parfaite du mélange. En pratique, nous avons constaté que cette méthode apportait une

amélioration de la performance d'extraction uniquement pour les mélanges panoramiques. Pour les autres types de mélange, elle aboutissait plutôt à une dégradation, même en utilisant les réponses fréquentielles des filtres de mélange connues *a priori*. Par exemple, dans un mélange "AB étroit" avec deux sources, la matrice de démixage est identique pour toutes les trames d'une sous-bande donnée. Les résultats des algorithmes d'ACI basés sur ce principe de démixage invariant dans le temps montrent que cela n'est pas applicable aux mélanges réverbérants (voir paragraphe 2.4.1).

7.3 Exemples

7.3.1 Trio stéréo panoramique synthétique de violoncelle, clarinette et violon

Dans notre première expérience, nous traitons un trio synthétique formé des extraits de violoncelle (s_1), clarinette (s_2) et violon (s_3) utilisés dans le paragraphe 5.3.2. Le recouvrement des sources en temps-fréquence est limité. Mais la mélodie jouée par le violoncelle fait partie de la tessiture du violon et a un timbre proche du violon, et *vice-versa*. La distinction des deux instruments à cordes est donc difficile uniquement à partir des informations de timbre. Les formes d'onde sont mélangées respectivement à gauche, à droite et au milieu par la matrice

$$\mathbf{A} \approx \begin{pmatrix} 0,866 & 0,500 & 0,707 \\ 0,500 & 0,866 & 0,707 \end{pmatrix} \tag{7.15}$$

La figure 7.1 montre les trois quantités observées.

Nous fixons les hyper-paramètres à $Z = 0,96$, $\sigma^{\epsilon\,\mathrm{tot}} = 1,4$, $\sigma^{\epsilon\,\mathrm{vol}} = 1,4$, $\lambda^{\mathrm{vol}} = 0,2$, $w_{\mathrm{comb}} = 0,5$, $w_{\mathrm{desc}} = 0,5$ et $w_{\mathrm{état}} = 1$. Nous fixons également une valeur minimale de $\sigma_{tf}^{\epsilon\,\mathrm{vol}} = 0,1$. Les trois gains relatifs de mélange (θ_j), égaux respectivement à -1,099, 1,099 et 0, sont estimés à partir des différences de volume inter-canal avec une précision égale à 10^{-4} environ.

Nous commençons par valider les hypothèses de modélisation en calculant les erreur résiduelles à partir des descripteurs et des bruits de fond oracle estimés sur les sources dans le paragraphe 5.3.2. La figure 7.2 montre que l'équation 7.4 liant l'écart-type de $\boldsymbol{\epsilon}_t^{\mathrm{vol}}$ à la cohérence inter-canal correspond bien aux données expérimentalement.

Nous transcrivons alors le mélange à partir du modèle factoriel de couche d'état en testant diverses associations des gains relatifs aux instruments. Le tableau 7.1 montre que la probabilité de transcription ne permet pas de distinguer la configuration spatiale réelle (violoncelle à gauche, clarinette à droite et violon au milieu) de la configuration obtenue en échangeant les positions du violon et du violoncelle. Cependant les autres configurations apparaissent moins probables.

Directions spatiales			Probabilité de transcription
Vc	Cl	Vn	P^{trans}
G	D	M	$-1,273.10^5$
G	M	D	$-1,405.10^5$
D	G	M	$-1,397.10^5$
D	M	G	$-1,422.10^5$
M	G	D	$-1,375.10^5$
M	D	G	$-1,273.10^5$

TAB. 7.1 – Probabilités de transcription obtenues pour diverses directions spatiales estimées des sources sur le trio stéréo panoramique synthétique (G : gauche, D : droite, M : milieu).

La figure 7.3 décrit les notes identifiées après sélection manuelle de la configuration spatiale réelle. En sus de la transcription par modèles d'instruments factoriels et segmentaux, nous effectuons la trans-

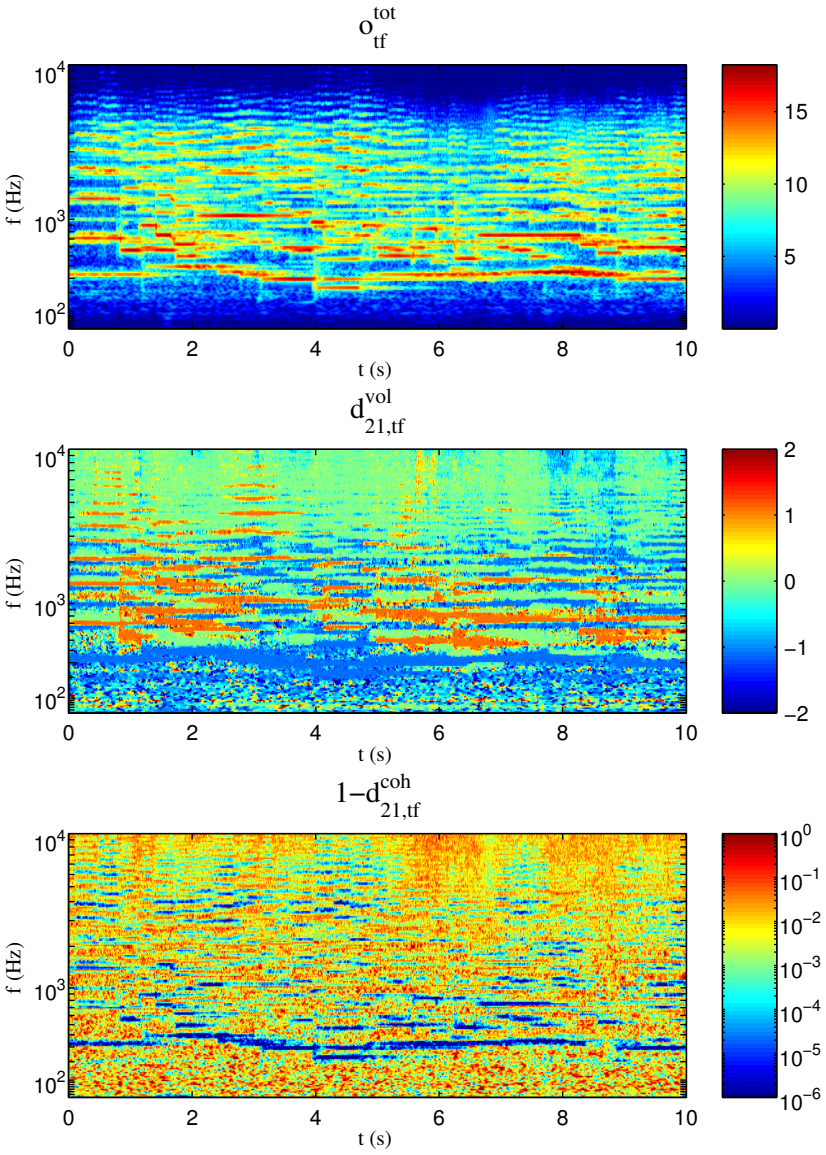

FIG. 7.1 – Quantités observées sur le trio stéréo panoramique synthétique.

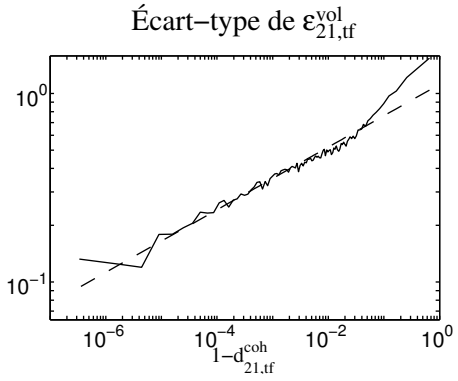

FIG. 7.2 – Écart-type empirique de l'erreur résiduelle sur la différence de volume inter-canal du trio stéréo panoramique synthétique (trait plein) comparé à sa définition en fonction de la cohérence inter-canal (tirets).

cription du spectre de log-puissance total ($\mathbf{o}_t^{\mathrm{tot}}$) par l'algorithme de saut monocanal (avec les mêmes hyper-paramètres sauf $w_{\mathrm{comb}} = 0,25$). Ces résultats peuvent être comparés aux transcriptions exactes des sources dans les figures 5.10, 5.11 et 5.12.

Nous voyons que la qualité de transcription sans exploitation des informations spatiales est assez faible. La plupart des notes de clarinette sont correctement identifiées, mais de nombreuses notes de violoncelle sont attribuées au violon. La prise en compte des différences de volume inter-canal observées améliore les résultats en permettant de distinguer les notes de violoncelle et de violon. Comme c'est le cas généralement, les quelques notes de très courte durée estimées avec les modèles factoriels sont supprimées dans la transcription par modèles segmentaux. Il subsiste une erreur d'insertion pour le violoncelle, deux pour la clarinette, ainsi qu'une erreur de suppression pour la clarinette et le violon (sur un total de soixante notes). Les erreurs d'insertion pour le violoncelle et la clarinette, aux temps $t \simeq 1,1\,\mathrm{s}$ et $t \simeq 9,7\,\mathrm{s}$, sont dues à des "trous" de courte durée au sein des notes. Ces erreurs pourraient probablement être évitées avec des modèles d'instruments plus complexes modélisant la durée des silences.

Le tableau 7.2 présente les résultats d'extraction de sources par pseudo-inversion locale. Les algorithmes d'extraction utilisant les modèles d'instruments sont comparés à deux algorithmes aveugles (ne nécessitant pas d'information *a priori* détaillée sur les sources) reposant sur une hypothèse de parcimonie des sources. L'algorithme de la "source la plus proche" attribue le point temps-fréquence (t, f) avec un gain unitaire à la source j pour laquelle le gain relatif de mélange θ_j est le plus proche de la différence de volume inter-canal observée $d^{\mathrm{vol}}_{21,tf}$, et avec un gain nul aux autres sources. L'algorithme des "sources les plus proches" attribue le point (t, f) aux deux sources j et j' lorsque $d^{\mathrm{vol}}_{21,tf}$ est comprise entre θ_j et $\theta_{j'}$ [Vin04d].

Nous constatons que la performance des nos algorithmes est comparable ou légèrement inférieure à celle des algorithmes aveugles. Ce résultat s'explique de plusieurs façons.

Premièrement, les sources choisies pour le mélange ont un recouvrement en temps-fréquence limité, ce qui rend l'hypothèse de parcimonie largement valable. La performance obtenue avec une transcription par oracle est ainsi à peine supérieure à celle des algorithmes aveugles. Cet écart de performance se creuserait probablement sur des mélanges plus réalistes, avec des sources jouant en harmonie.

Deuxièmement, les critères de performance ne rendent pas entièrement compte de la qualité des sources estimées. À l'écoute, les artefacts d'extraction apparaissent de deux natures différentes. Les artefacts produits par les algorithmes aveugles contiennent surtout du bruit musical léger et continu, alors que ceux produits par nos algorithmes sont faibles la plupart du temps et correspondent à des erreurs plus grossières localisées en temps. Cela s'entend particulièrement sur la source estimée de violon. Les erreurs grossières d'extraction sont dues à des erreurs de transcription, par exemple des notes supprimées, insérées ou de durées estimées plus courtes qu'en réalité. La comparaison de la figure 7.3 aux figures 5.10, 5.11 et 5.12 montre que la durée de la réverbération des notes estimées sur le mélange est réduite par rapport à celle estimée sur les sources. Ce problème pourrait être partiellement résolu en utilisant un modèle de couche d'état plus complexe qui impose une durée minimale de réverbération.

Méthode de transcription	Performance $\widehat{s_1}$ (dB)			Performance $\widehat{s_2}$ (dB)			Performance $\widehat{s_3}$ (dB)		
	RSD	RSI	RSA	RSD	RSI	RSA	RSD	RSI	RSA
Source la plus proche	15	41	15	21	42	22	15	32	15
Sources les plus proches	17	34	17	23	37	23	12	26	12
Oracle	18	43	18	24	48	24	18	35	18
Factoriel	16	40	16	20	44	20	15	32	15
Segmental	16	40	16	19	43	19	14	32	14

TAB. 7.2 – Performance d'extraction de sources sur le trio stéréo panoramique synthétique.

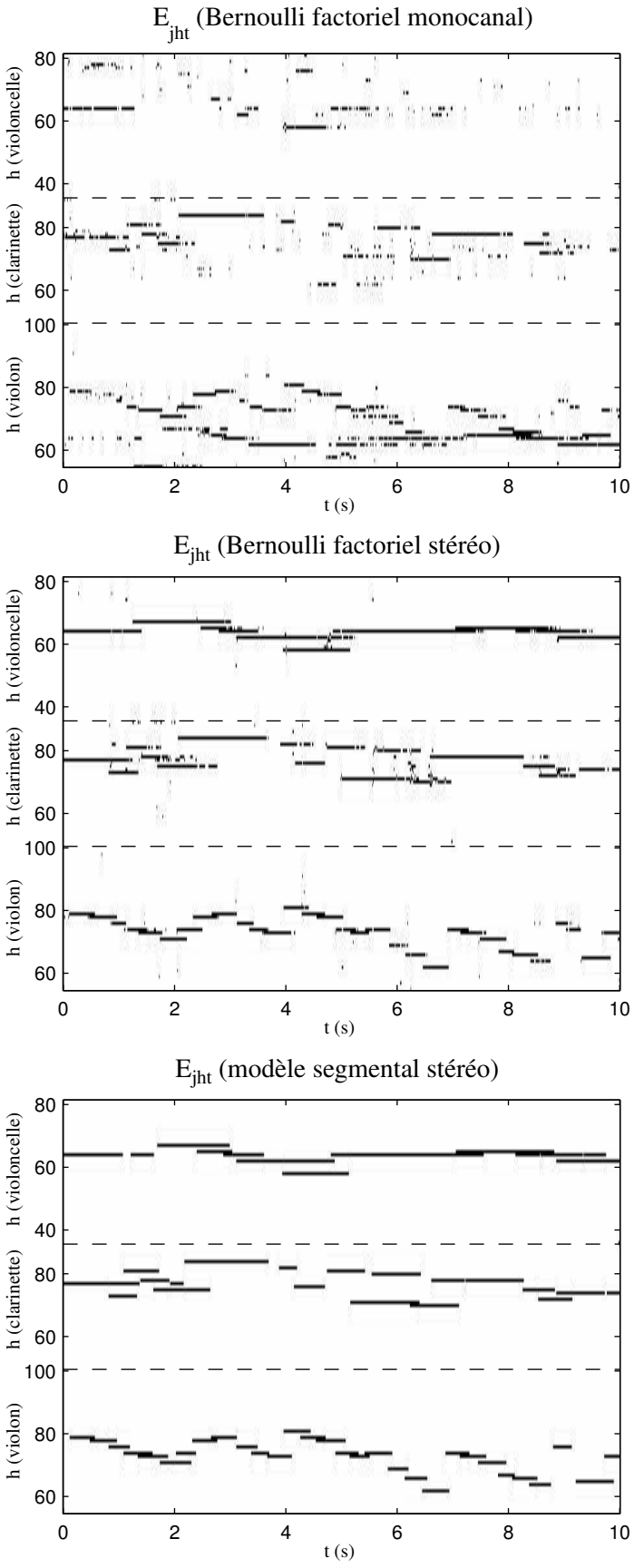

FIG. 7.3 – Notes identifiées sur le trio stéréo panoramique synthétique.

7.3.2 Duo stéréo “AB étroit” synthétique de violoncelle et violon

Notre deuxième expérience étudie un duo synthétique réalisé à partir des mêmes extraits de violoncelle (s_1) et de violon (s_2), convolués par des réponses impulsionnelles fixées correspondant à un couple stéréo lointain “AB étroit”, dont l'obtention est décrite dans l'annexe A.3. Les observations sont montrées dans la figure 7.4.

Nous modélisons les réponses fréquentielles de phase relative des filtres de mélange par des réponses de deux délais purs égaux à 4.83 et -1.14 échantillons respectivement. La figure 7.5 montre que cette approximation est globalement valable. Néanmoins des écarts importants existent dans certaines sous-bandes, en particulier dans la zone de fréquence autour de $f \approx 350$ Hz correspondant aux fréquences fondamentales des notes de violoncelle .

Nous fixons les nouveaux hyper-paramètres à $\sigma^{\epsilon\,\mathrm{tot}} = 1,4$, $\sigma^{\epsilon\,\mathrm{pha}} = 2,4$ et $\lambda^{\mathrm{pha}} = 0,2$, les autres restant inchangés, et nous imposons une valeur minimale de $\sigma_{tf}^{\epsilon\,\mathrm{pha}} = 0,1$.

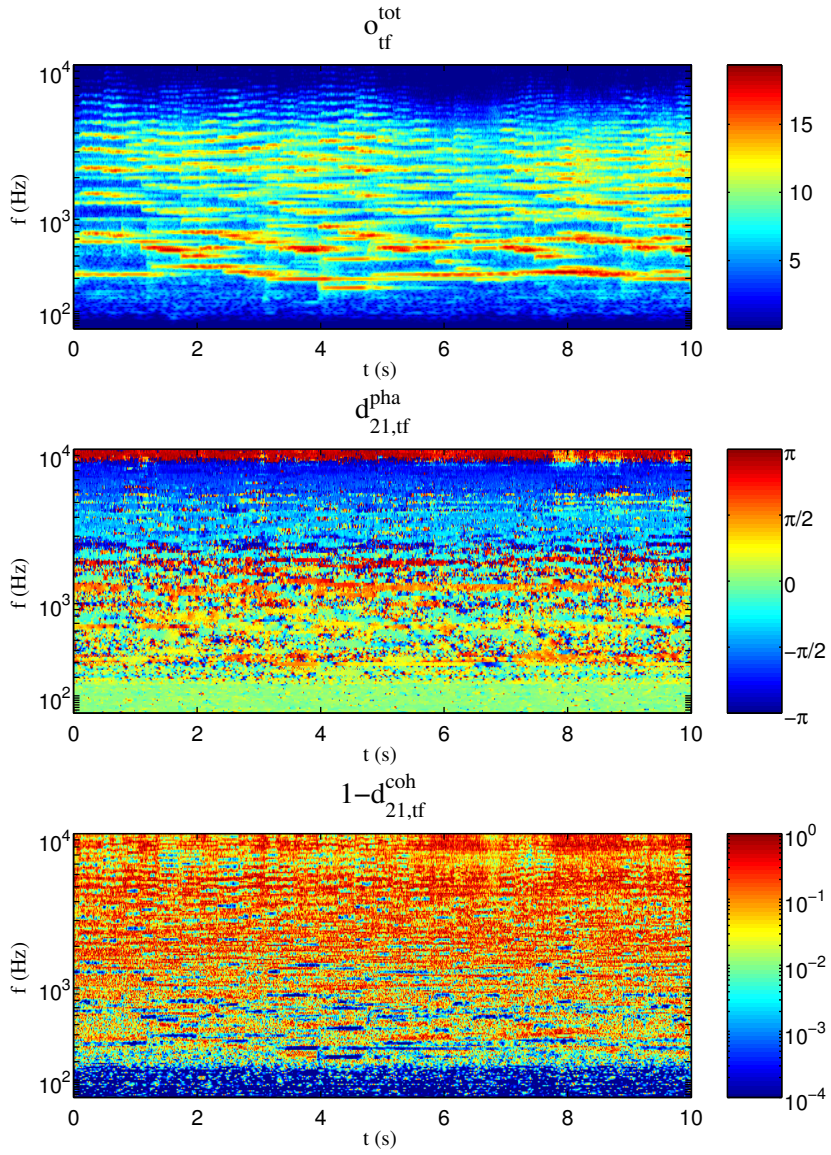

FIG. 7.4 – Quantités observées sur le duo stéréo “AB étroit” synthétique.

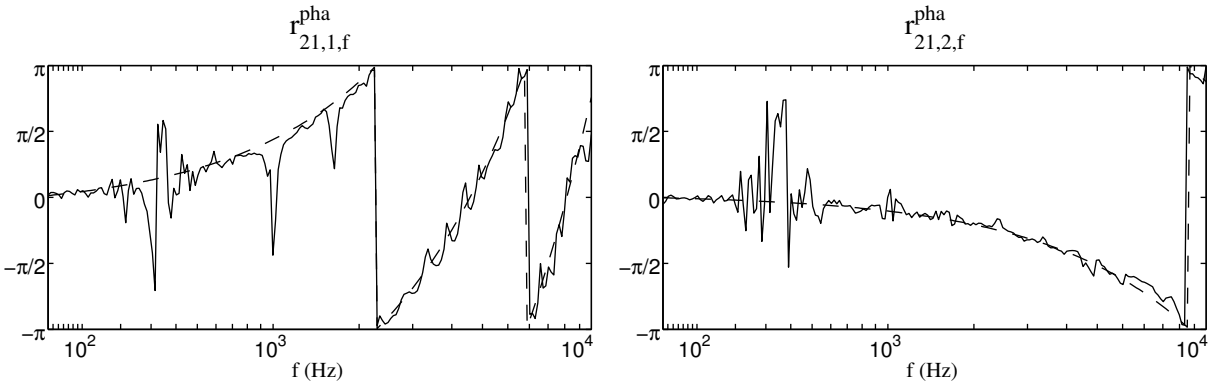

FIG. 7.5 – Réponses fréquentielles de phase relative des filtres de mélange "AB étroits" (trait plein) comparées aux réponses de délais purs (tirets).

Les résultats de transcription sont décrits dans la figure 7.6 et à nouveau comparés à ceux d'une transcription du spectre de log-puissance total par l'algorithme de saut monocanal (avec les mêmes hyperparamètres sauf $w_{\mathrm{comb}} = 0,25$). Comme dans l'expérience précédente, nous constatons que l'utilisation des informations spatiales permet de distinguer les mélodies des deux instruments, qui sinon restent confondues à cause de leur timbre et de leur tessiture semblables.

La transcription par modèles d'instruments factoriels est cependant un peu moins bonne que dans l'expérience précédente. Par exemple, certaines notes de violoncelle sont transcrites par des états des deux instruments au temps $t \simeq 2,5$ s. Cette baisse de performance semble due au fait que peu de points temps-fréquence offrent une information spatiale pertinente, cette dernière étant perturbée par la réverbération. Nous avons obtenu des résultats bien meilleurs en remplaçant le mélange par un mélange synthétique à base de délais purs, mais aucune amélioration en gardant le même mélange et en utilisant les réponses réelles des filtres de mélange au lieu de leur approximation par des délais purs.

La transcription par modèles d'instruments segmentaux corrige partiellement ce manque d'information spatiale en exploitant l'information temporelle. Les suites de notes estimées contiennent deux insertions pour le violoncelle (dont un "trou" de courte durée au sein d'une note) et une substitution et une insertion pour le violon, sur un total de quarante-trois notes.

Nous extrayons les images spatiales des sources par filtrage pseudo-Wiener sur les deux capteurs et nous évaluons la performance d'extraction sur le canal gauche ($i = 1$) dans le tableau 7.3. Nous comparons nos algorithmes à l'algorithme d'ACI aveugle pour mélanges convolutifs déterminés de Murata, Ikeda et Ziehe [Mur01] (utilisé avec ses paramètres par défaut) et à un algorithme de "source la plus proche" attribuant le point temps-fréquence (t, f) à la source j pour laquelle la réponse en phase relative $a^{\mathrm{pha}}_{21,jf}$ est la plus proche (modulo 2π) de la différence de phase inter-canal observée $d^{\mathrm{pha}}_{21,tf}$ [Rom03]. Nous calculons également la performance maximale des algorithmes d'ACI par un algorithme d'ACI oracle connaissant les images spatiales à estimer. Ce dernier minimise l'erreur quadratique sur chaque image spatiale estimée (ce qui est légèrement différent de la minimisation du SDR) en effectuant une projection orthogonale de l'image spatiale oracle sur le sous-espace engendré par les sous-bandes des canaux du mélange, de façon similaire aux projections utilisées pour calculer les critères de performance dans le paragraphe 1.3.1. Les tailles des filtres de démixage estimés par les algorithmes d'ACI et d'ACI oracle sont de 512 et 2048 échantillons respectivement.

Les résultats prouvent que la performance des algorithmes d'ACI est limitée par la présence d'un temps de réverbération élevé. Même avec des filtres de démixage oracle de taille 2048, les images spatiales extraites contiennent encore une quantité importante d'interférences. Or cette taille correspond à la limite maximale pour les algorithmes d'ACI existants. De plus la performance effective de ces algo-

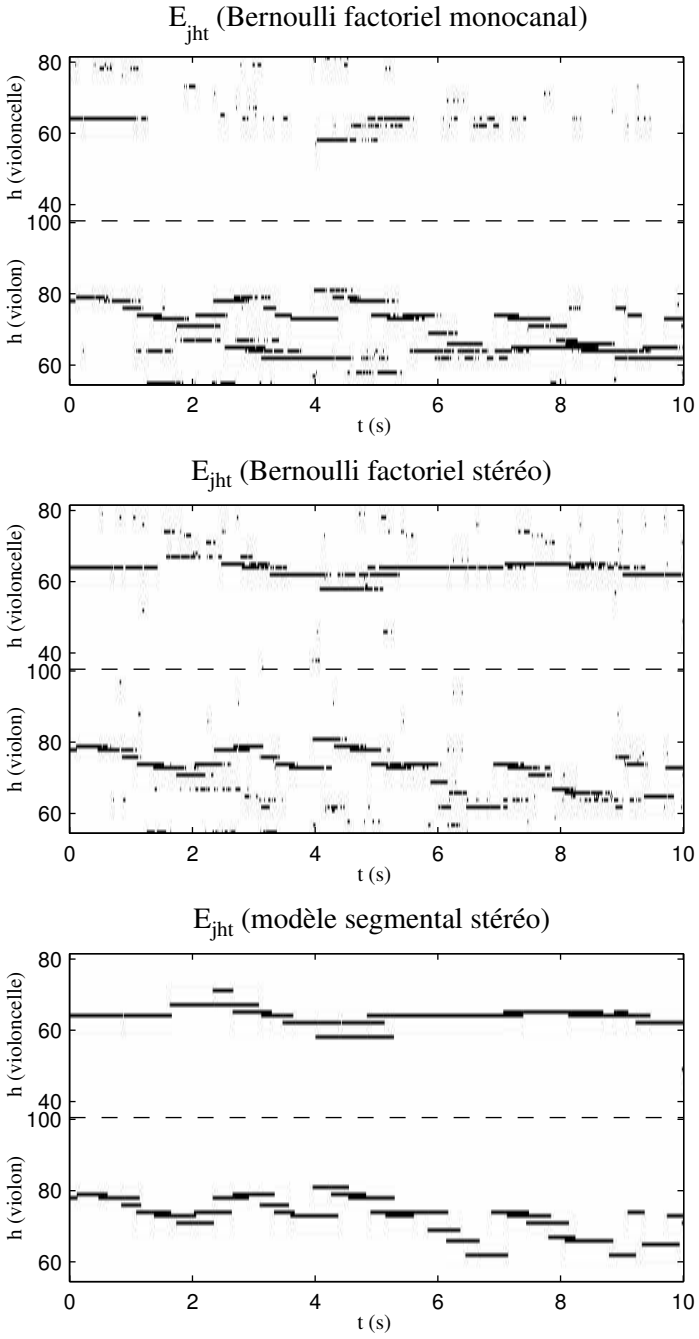

FIG. 7.6 – Notes identifiées sur le duo stéréo “AB étroit” synthétique.

rithmes est nettement inférieure à celle de l'oracle. Ainsi l'algorithme testé ne réduit quasiment pas les interférences dans les images spatiales estimées par rapport aux interférences préexistant dans le mélange.

L'algorithme de "source la plus proche" résout relativement bien le problème des interférences mais engendre en contrepartie beaucoup de bruit musical.

Nos algorithmes surpassent les algorithmes d'ACI et de "source la plus proche" à la fois en terme d'interférences et d'artefacts. Le filtrage après transcription par modèles d'instruments factoriels conserve une sonorité artificielle due à des interférences de très courte durée ou à des "trous" au sein de certaines notes. L'utilisation de modèles d'instruments segmentaux améliore globalement la performance en imposant une durée minimale aux notes séparées. Comme dans l'expérience précédente, la qualité des images spatiales extraites est bonne la plupart du temps et mauvaise sur les trames contenant des erreurs de transcription, par exemple autour de $t \simeq 9$ s pour le violon.

Nous avons également modifié la scène sonore en multipliant l'amplitude du violoncelle par 2 (soit un ajout de 6 dB). L'écoute des résultats montre que les remixes estimés par nos algorithmes et par l'algorithme de "source la plus proche" ne contiennent pas de bruit musical et sont tous proches du remix attendu. Par contre, le remix estimé par ACI n'est pas perçu comme différent de la scène initiale.

Méthode de transcription	Performance $\widehat{s_{\mathrm{img}\,11}}$ (dB)			Performance $\widehat{s_{\mathrm{img}\,12}}$ (dB)		
	RSD	RSI	RSA	RSD	RSI	RSA
Oracle ACI taille 2048	13	26	14	11	23	12
ACI taille 512	-4	3	1	-6	0	3
Source la plus proche	8	17	9	5	18	6
Oracle	17	38	17	16	33	16
Factoriel	9	27	10	10	20	10
Segmental	13	38	16	14	33	14

TAB. 7.3 – Performance d'extraction des images spatiales des sources sur le canal gauche pour le duo stéréo "AB étroit" synthétique.

7.3.3 Duo multipistes synthétique de violoncelle et violon

Notre troisième et dernière expérience reprend le mélange de la deuxième expérience en ajoutant deux canaux synthétiques supplémentaires correspondant à des micros d'appoint, dont les réponses impulsionnelles sont obtenues dans les mêmes circonstances que celles de la paire AB étroite (voir annexe A.3). Les figures 7.7 et 7.8 montrent les quatre nouvelles quantités observées et les réponses fréquentielles en amplitude des filtres de mélange correspondant aux micros d'appoint. Nous constatons que les filtres ont une réponse faible en basse fréquence, et que la directivité des micros d'appoint augmente avec la fréquence.

Nous fixons les hyper-paramètres à $\sigma^{\epsilon_1} = \sigma^{\epsilon_2} = 1,4$, $\sigma^{\epsilon\,\mathrm{vol}} = 2,4$ et $\lambda^{\mathrm{vol}} = 0,15$. Le poids w_{comb} est remplacé par deux poids distincts pour donner la même importance aux informations spatiales et spectro-temporelles : un poids $w_{\mathrm{comb}} = 0,25$ est attribué à $\mathcal{N}(\epsilon_{1,tf}; 0, \sigma^{\epsilon_1})\mathcal{N}(\epsilon_{2,tf}; 0, \sigma^{\epsilon_2})$ et un poids $w_{\mathrm{comb}} = 0,5$ à $\mathcal{N}(\epsilon^{\mathrm{vol}}_{tf}; 0, \sigma^{\epsilon\,\mathrm{vol}}_{tf})$. Nous imposons également une valeur minimale de $\sigma^{\epsilon\,\mathrm{vol}}_{tf} = 0,2$.

La figure 7.9 décrit les résultats d'identification de notes. Nous observons que la combinaison des modèles d'instruments et des informations spatiales fournies par les micros d'appoint permet à nouveau de distinguer les mélodies jouées par le violoncelle et le violon. Les résultats sont même meilleurs que dans l'expérience précédente, car la directivité des micros d'appoint introduit un contraste spatial plus visible entre les sources. La transcription estimée par modèles d'instruments segmentaux contient pour

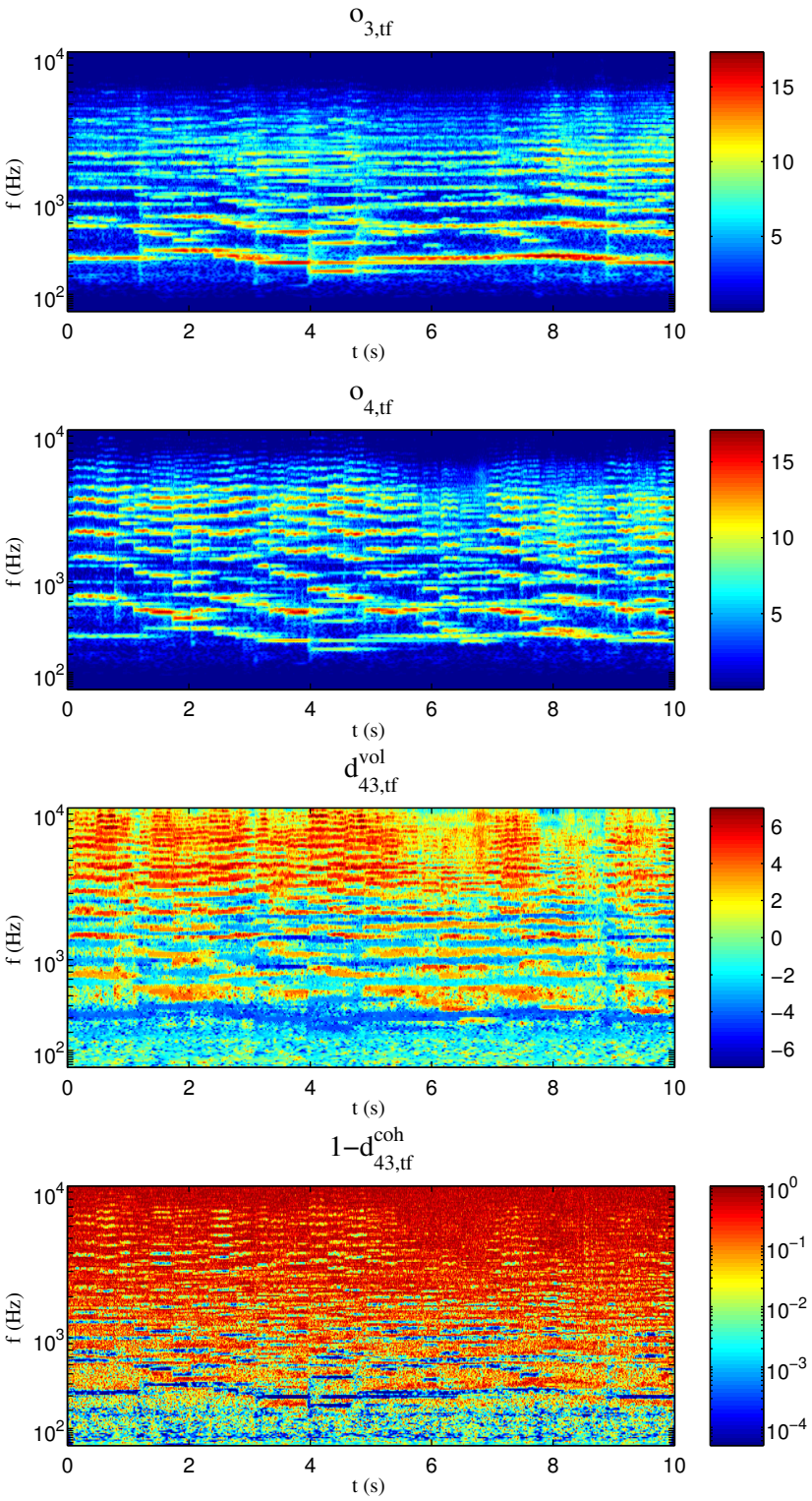

FIG. 7.7 – Quantités observées sur le duo multipistes synthétique.

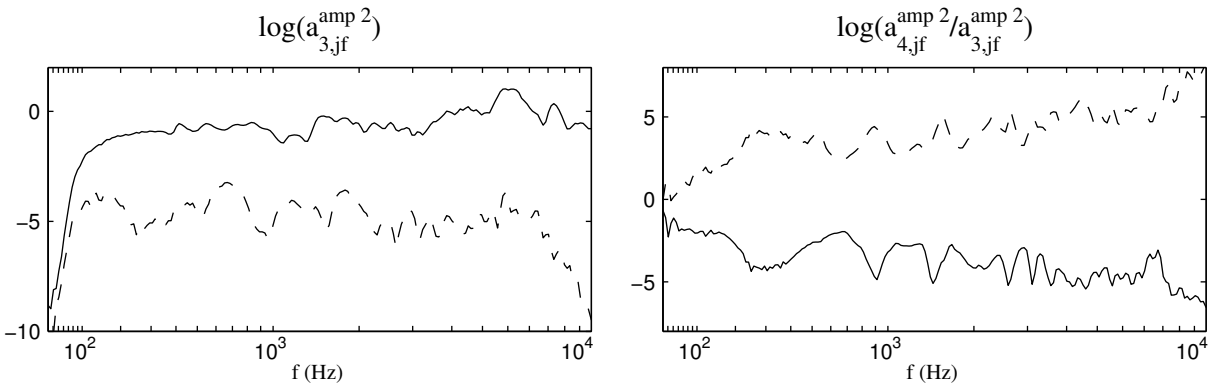

FIG. 7.8 – Réponses fréquentielles de log-puissance des filtres de mélange correspondant aux micros d'appoint (trait plein : violoncelle, tirets : violon).

erreur unique une note de violon supprimée sur un total de quarante-trois notes.

Comme dans l'expérience précédente, nous extrayons les images spatiales des sources sur les deux canaux du couple stéréo lointain par filtrage pseudo-Wiener sur chaque canal et nous comparons la performance de nos algorithmes aux algorithmes d'ACI et de "source la plus proche". L'algorithme d'ACI oracle utilise des filtres de démixage de 1024 échantillons. L'algorithme de "source la plus proche" attribue le point temps-fréquence (t, f) au violoncelle si $d^{\text{vol}}_{43,tf} < 0$ et au violon sinon. Ce seuil *ad hoc* trace une limite valable entre les réponses en amplitude des filtres de mélange des deux sources, d'après la figure 7.8.

Les mesures de performance listées dans le tableau 7.4 montrent que nos algorithmes sont légèrement moins performants que l'algorithme de "source la plus proche". Comme dans la première expérience, le recouvrement spatial limité des sources mène à une bonne performance des méthodes basées sur la parcimonie. Cependant la source de violoncelle extraite par nos algorithmes est un peu meilleure perceptivement car elle contient moins de bruit musical.

L'algorithme d'ACI oracle donne de meilleurs résultats que dans l'expérience précédente avec des filtres de démixage de taille inférieure. Pour évaluer la performance des algorithmes d'ACI en pratique, nous avons appliqué l'algorithme d'ACI utilisé précédemment aux deux canaux correspondant aux micros d'appoint pour estimer les images spatiales des sources sur ces canaux. Une fois de plus nous n'avons constaté aucune diminution des interférences par rapport à celles préexistant dans le mélange (RSI de l'ordre de 25 dB). De plus, les images spatiales des sources sur les autres canaux ne peuvent être estimées directement par cette méthode.

Enfin, comme dans l'expérience précédente, la modification de la scène sonore des canaux de stéréo donne des résultats de très bonne qualité à l'écoute, à la fois pour nos algorithmes et pour l'algorithme de "source la plus proche".

7.3.4 Résumé des résultats

Les résultats de ce chapitre peuvent se résumer par deux affirmations.

Premièrement, la prise en compte des informations spatiales en sus des informations spectro-temporelles dans les mélanges multicanal semble faciliter l'identification des notes jouées par des instruments de tessiture et de timbre semblables. L'amélioration par rapport au cas monocanal est particulièrement sensible pour les mélanges fournissant des informations spatiales faciles à interpréter, comme les mélanges panoramiques ou multipistes.

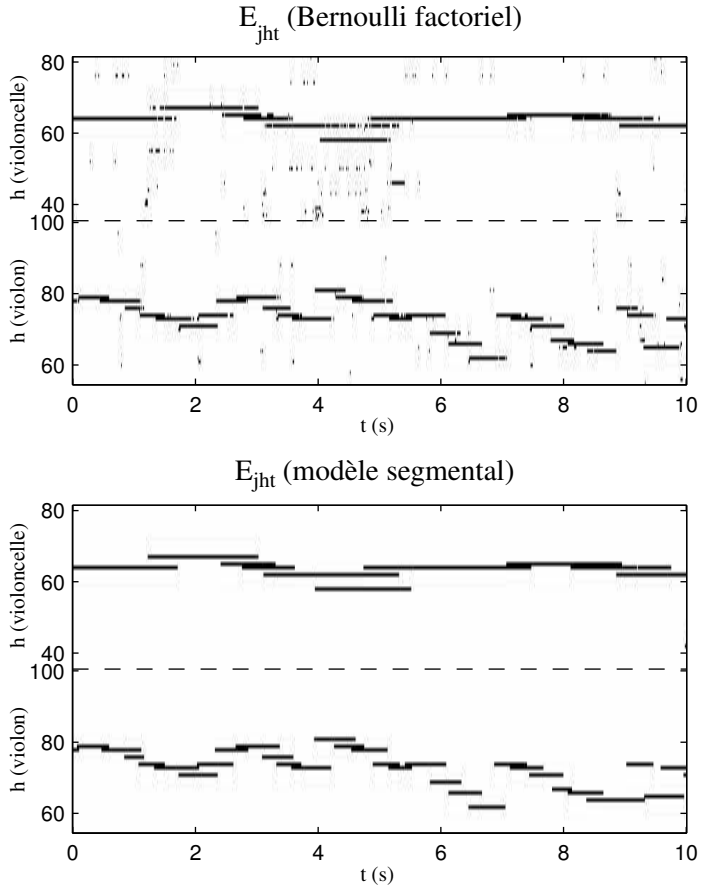

FIG. 7.9 – Notes identifiées sur le duo multipistes synthétique.

Méthode de transcription	Performance $\widehat{s_{\text{img}\,11}}$ (dB)			Performance $\widehat{s_{\text{img}\,12}}$ (dB)		
	RSD	RSI	RSA	RSD	RSI	RSA
Oracle ACI taille 1024	21	39	21	18	38	18
Source la plus proche	19	39	19	19	41	19
Factoriel	15	40	15	14	31	14
Segmental	16	38	16	15	33	15

TAB. 7.4 – Performance d'extraction des images spatiales des sources sur le canal gauche pour le duo multipistes synthétique.

Deuxièmement, l'utilisation de modèles d'instruments peut aider à augmenter la performance d'extraction de sources par rapport aux méthodes aveugles exploitant uniquement les informations spatiales. Dans ce cas, l'augmentation est d'autant plus significative que les informations spatiales sont difficiles à interpréter, comme c'est le cas dans les mélanges stéréo "AB étroits". Par contre, la performance de nos algorithmes est semblable à celle des algorithmes aveugles basés sur une hypothèse de parcimonie en ce qui concerne la modification de scène sonore.

Conclusion

Nous avons développé au cours de ce travail un cadre probabiliste de construction et d'utilisation de modèles de sources instrumentales, ensuite appliqué à la transcription et la séparation d'enregistrements de musique de chambre. Nous avons d'abord montré que ces modèles permettent de reconnaître avec une performance satisfaisante les divers instruments présents dans un mélange monocanal et les notes jouées par chacun, dans la mesure où leurs timbres sont assez différents. Ensuite, nous avons constaté une augmentation de la performance d'identification de notes en combinant ces modèles aux informations spatiales dans les mélanges multicanal. Enfin, nous avons utilisé les résultats de transcription pour l'extraction de sources et la modification de scènes sonores, avec de meilleurs résultats que les méthodes aveugles existantes sur certains mélanges difficiles.

D'un point de vue théorique, nous avons combiné dans les modèles d'instruments proposés deux types de modèles de sources habituellement considérés comme différents : les décompositions parcimonieuses spectrales et les modèles de séries temporelles de type MMC. Nous avons décrit aussi une méthode originale pour utiliser ces modèles efficacement dans le cadre d'observations multicanal. Nous proposons trois extensions théoriques différentes à ce travail.

Premièrement, dans le cas de mélanges monocanal, il serait sans doute profitable de rendre les modèles d'instruments plus complexes pour augmenter la quantité d'information *a priori* disponible. Par exemple, il serait possible de prendre en compte dans les modèles d'instruments harmoniques les relations de phase entre sous-bandes fréquentielles, dès la transcription ou uniquement durant le filtrage. Les modèles de sources gouverneraient l'amplitude des partiels et du bruit large bande entre les partiels, mais la phase des partiels et du bruit serait contrôlée par des modèles sinusoïdaux harmoniques classiques. Une question difficile se pose alors : comment mesurer la probabilité conditionnelle des observations sachant les descripteurs ? L'erreur quadratique sur les formes d'onde utilisée par les modèles sinusoïdaux classiques ne correspond pas en effet à l'erreur quadratique sur les spectres de log-puissance que nous avons proposée (et pas à l'"erreur perceptive" non plus). Une autre possibilité consisterait à modéliser les informations de type musical (rythme, intervalles, *etc*). Dans cas la complexité supplémentaire pourrait nécessiter de revoir les algorithmes de transcription proposés.

Deuxièmement, il serait intéressant de proposer des solutions dans les cas où l'information sur les sources est au contraire plus limitée. Par exemple, les enregistrements de musiques électroniques contiennent souvent des sons originaux sans équivalent ailleurs. Le caractère souvent répétitif de ces enregistrements pourrait être mis à profit pour un apprentissage non supervisé de modèles d'instruments, en fixant un seuil permettant de distinguer entre deux notes d'instruments différents et une note d'un même instrument répétée de façon légèrement différente. Il serait aussi possible de modéliser les instruments inconnus par un "modèle de réjection", comme cela se fait en traitement de la parole.

Troisièmement, il serait utile pour des applications industrielles de combiner les modèles d'instruments à des modèles de voix chantée ou parlée, les musiques actuelles contenant en majorité de la voix. Les modèles usuels de parole ne sont pas applicables à des mélanges de plusieurs sources car ils modélisent le cepstre des observations. Les modèles d'instruments pourraient inspirer des modèles de parole plus complets. En effet ils peuvent actuellement modéliser les structures principales d'un signal de voix, comme le caractère sinusoïdal harmonique des voyelles et bruité des consonnes et les variations de l'en-

veloppe spectrale en fonction du phonème. Reste à modifier les dépendances entre variables cachées et les procédures de partage de paramètres pour rendre cette modélisation réellement efficace. Des modèles de parole similaires à l'ASI ont déjà été étudiés sous le nom de Décomposition Temporelle Généralisée.

D'un point de vue appliqué, nos résultats les plus originaux ont concerné l'identification d'instruments multiples avec effectif instrumental inconnu et la séparation de mélanges convolutifs longs. Cependant, nous avons testé la performance de nos algorithmes dans des conditions assez limitées.

La première limitation vient du fait que nous avons utilisé en majorité des mélanges synthétiques. Dans le cas monocanal, nous avons montré sur un mélange réel que nos algorithmes pouvaient fonctionner en présence de notes à intervalle harmonique. Dans le cas multicanal, nous faisons aussi l'hypothèse que la performance de nos modèles ne se dégraderait pas de façon brutale en présence de mouvements des sources, car le modèle choisi pour prendre en compte les informations spatiales se focalise uniquement sur les informations fiables déterminées à l'aide de la cohérence inter-canal. Les résultats sont donc prometteurs, mais cela ne remplace en rien des tests approfondis sur des mélanges réels. De même, hormis pour l'identification d'instruments, nous avons testé toutes les tâches sur un nombre de mélanges trop limité pour être généralisable. La collecte d'un nombre suffisant de mélanges étiquetés pose un véritable problème. Pour l'identification de notes, il est possible de trouver des fichiers MIDI ou des enregistrements libres de droits de musique classique sur internet ou dans les bases de données existantes, mais jamais les deux simultanément. Sur les deux ou trois enregistrements que nous avons trouvé de cette façon, nous avons constaté que les musiciens se permettaient des libertés par rapport à la partition, ce qui fait perdre toute utilité à la partition dans un but d'évaluation ! Le problème est encore plus manifeste en ce qui concerne l'extraction de sources et la modification de scène sonore. Dans le cadre de la collaboration au sein de l'Action Jeunes Chercheurs mentionnée dans le chapitre 1, nous avons essayé de créer une base de données adaptées contenant les sources nécessaires à l'évaluation des résultats. Mais nous avons été confrontés à des problèmes juridiques assez inextricables qui nous ont conduit à repousser le projet d'une base libre de droits.

La deuxième limitation de nos conditions de test est due au choix manuel des hyper-paramètres des algorithmes. Souvent nous avons pris les mêmes valeurs de paramètres pour des mélanges semblables ou de même type, mais un réglage automatique reste souhaitable. Notons que cela ne constitue pas forcément un problème important pour les tâches de séparation. En effet, les résultats de modification de scène sont généralement voués à être écoutés par un opérateur humain en temps différé. Il est donc pensable que cet opérateur effectue plusieurs réglages manuels de paramètres avant de sélectionner celui qui lui paraît le meilleur perceptivement et musicalement. Le fonctionnement des logiciels actuels repose sur ce principe. Nos algorithmes pourraient donc simplement contribuer à diminuer le temps de travail de l'opérateur en proposant un simple réglage de paramètres au lieu d'une sélection manuelle des zones temps-fréquence à modifier.

Annexe A

Données d'apprentissage et de test

Cette première annexe décrit une partie des données intervenant au long de ce travail. Nous présentons dans le paragraphe A.1 la base de données de notes isolées utilisée dans la phase d'apprentissage des modèles d'instruments, et dans le paragraphe A.2 les extraits solo utilisés dans la phase de test. Nous décrivons ensuite dans le paragraphe A.3 les conditions d'enregistrement des réponses impulsionnelles de salle utilisés pour créer des mélanges convolutifs synthétiques.

A.1 Base de données de notes isolées

Les modèles d'instruments sont appris à partir des enregistrements de notes isolées de la base de données musicales *RWC* [RWC-MDB] pour cinq instruments : flûte, clarinette, hautbois, violon et violoncelle.

Chaque instrument est enregistré en trois sessions, par des musiciens différents sur des instruments de facture différente, mais dans la même salle d'enregistrement. À chaque session, le musicien joue toutes les notes de l'échelle des demi-tons avec des styles de jeu et des nuances différents. Pour le violon et le violoncelle, certaines notes sont jouées sur plusieurs cordes. Nous sélectionnons une seule session par instrument (correspondant aux numéros 331, 311, 291, 151 et 171) et les styles de jeu les plus courants. Le tableau A.1 présente les styles de jeu et les nuances sélectionnées. Sauf mention contraire, trois enregistrements correspondant aux nuances *piano*, *mezzo* et *forte* sont associés à chaque style de jeu, soit un total de neuf enregistrements pour les instruments à vent et de huit pour les cordes frottées.

Tous les enregistrements sont échantillonnés à 22050 Hz. La puissance locale de chaque enregistrement est calculée sur des trames rectangulaires de taille 256, puis tous les enregistrements d'un instrument sont normalisés par le même gain, de sorte que la trame la plus puissante ait une puissance locale unitaire.

	Flûte	Clarinette	Hautbois	Violon	Violoncelle
Styles de jeu	normal *staccato* *vibrato*	normal *staccato* *vibrato*	normal *staccato* *vibrato*	normal *spiccato* *non-vibrato mezzo sordino mezzo*	normal, *spiccato* *non-vibrato mezzo sordino mezzo*
Tessiture	MIDI 60-96 do4–do7	MIDI 50-89 ré3–fa6	MIDI 58-91 la#3–sol6	MIDI 55-100 sol3–mi7	MIDI 36-81 do2–la5

TAB. A.1 – Liste des enregistrements de notes isolées.

A.2 Extraits solo réels

Les résultats d'identification d'instruments et de séparation de mélanges synthétiques sont obtenus à partir d'enregistrements solo réels. Pour chacun des cinq instruments considérés, nous collectons dix extraits d'une minute dans dix CD différents de styles musicaux allant du classique au contemporain.

Pour l'identification d'instruments, nous sélectionnons deux extraits de cinq secondes dans chaque enregistrement, en évitant les zones de silence ou les extraits identiques et en cherchant à couvrir toute la tessiture de l'instrument et les différents *tempi*.

Pour la création de mélanges synthétiques, nous sélectionnons trois extraits de dix secondes de violoncelle, clarinette et violon.

Chaque extrait est échantillonné à 22050 Hz et normalisé indépendamment des autres selon la méthode décrite ci-dessus.

A.3 Réponses impulsionnelles de salle

Les réponses impulsionnelles utilisées pour la création de mélange convolutifs synthétiques font partie de la base de données créée en collaboration avec l'IRISA et l'IRCCyN [BASS-dB]. Elles ont été enregistrées dans l'Espace de Projection de l'IRCAM avec le dispositif décrit dans la figure A.1 puis sous-échantillonnées à 22050 Hz.

Dans un mélange de deux sources, la source de gauche est attribuée au haut-parleur 2 et celle du milieu au haut-parleur 3. Ensuite l'image des sources est simulée sur les micros d'appoint 8 et 9 et sur les capsules 1 et 2 formant une paire AB étroite. Les azimuts des sources par rapport à la paire stéréo sont de -20° et 0° environ. La distance entre les hauts-parleurs et les micros est de 50 cm environ pour les micros d'appoint et de 4 m environ pour la paire stéréo. La distance entre les deux capsules omnidirectionnelles formant la paire stéréo est de 30 cm environ.

Le temps de réverbération perçu sur chaque réponse impulsionnelle peut être approché en calculant la puissance de la réponse sur des trames à court terme, puis en trouvant le temps que la puissance met à décroître entre son maximum et une valeur 60 dB plus faible (le temps de réverbération exact dépend de la fréquence). En utilisant des trames rectangulaires disjointes de 256 échantillons, nous obtenons un temps de réverbération de l'ordre de 0,8 s entre le haut-parleur 3 et la capsule 1 et de l'ordre de 0,3 s entre le même haut-parleur et le micro d'appoint 9 correspondant. Dans le premier cas les premières trames de réverbération sont environ 20 dB plus faibles que la trame correspondant au chemin direct, alors qu'elles sont environ 40 dB plus faibles dans le deuxième cas.

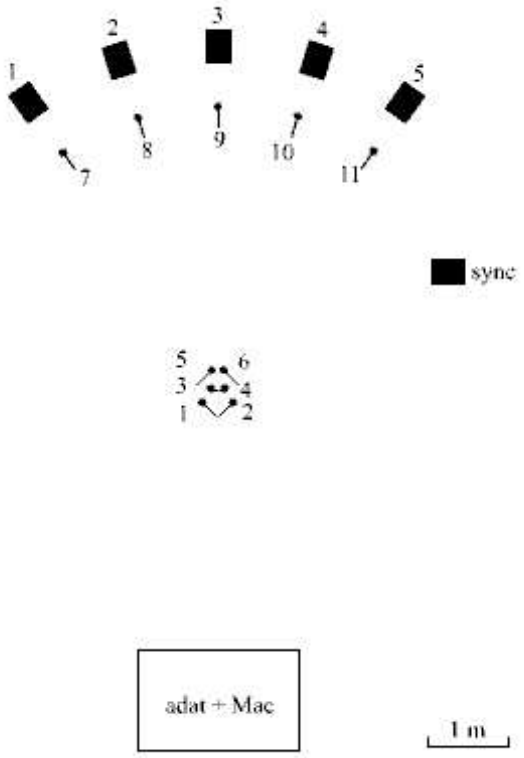

FIG. A.1 – Dispositif expérimental d'enregistrement de réponses impulsionnelles de salle.

Annexe B

Calcul du spectre et inversion du banc de filtres

Cette deuxième annexe traite le calcul des différents types de spectres utilisés dans ce travail : spectre à court terme, spectre d'une sinusoïde et réponse fréquentielle des filtres de mélange. Nous définissons le banc de filtres utilisé et les autres paramètres de calcul dans le paragraphe B.1, puis nous rappelons dans le paragraphe B.2 les procédures d'inversion approchée du banc de filtres utilisées pour la séparation par filtrage.

B.1 Calcul du spectre

B.1.1 Définition des filtres

Suite à la discussion du paragraphe 4.7, nous choisissons de calculer le spectre à court terme par un banc de filtres passe-bande $(H_f)_{0\leq f\leq F-1}$ espacés linéairement sur l'échelle ERB [Rom03], définie par

$$f_{ERB} = 9.26 \log(0.00437 f_{Hz} + 1). \tag{B.1}$$

Cette échelle modélise l'échelle fréquentielle caractéristique de l'appareil auditif : asymptotiquement elle est linéaire en-deçà de 200 Hz environ et logarithmique au-delà.

Nous définissons $F = 204$ filtres répartis entre 31 Hz et 10900 Hz par

$$H_f(u) = \frac{2}{\sqrt{3}} (L_f + 1)^{-1/2} w_F^f(u) \exp(2i\pi f u), \tag{B.2}$$

où w_F^f est une fenêtre de Hanning centrée en $u = 0$ de taille impaire $2L_f + 1$. Nous utilisons la même notation f pour l'indice de sous-bande et la fréquence centrale correspondante. Le gain attribué aux filtres implique $\|H_f\| = 1$, ainsi le spectre d'amplitude d'une impulsion de Dirac est bien un spectre plat normalisé.

Nous fixons la largeur du lobe principal des filtres à quatre fois l'écart de fréquence entre deux filtres voisins. Cette largeur relative équivaut à celle d'une Transformée de Fourier à Court Terme (TFCT) avec une fenêtre de Hanning. Pour la plupart des instruments, elle est suffisamment faible pour discriminer des notes de hauteurs semblables d'après la localisation de leurs premiers partiels et pour séparer ces partiels dans un accord. Pour les instruments de tessiture très grave (comme la contrebasse), ce n'est plus le cas, mais la discrimination des notes reste possible sur les partiels plus aigus.

La figure B.1 montre les valeurs de f et L_f selon le canal fréquentiel.

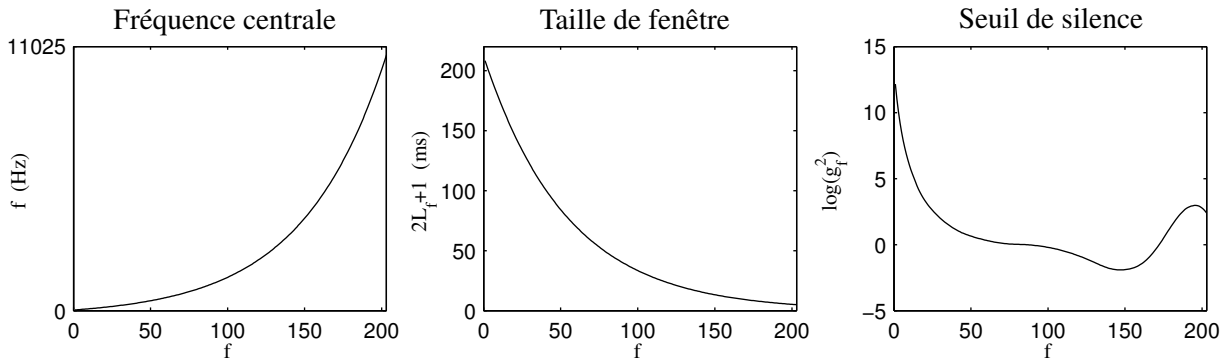

FIG. B.1 – Paramètres du banc de filtres.

B.1.2 Spectre à court terme

Le banc de filtres est utilisé pour l'analyse spectro-temporelle et spatiale des mélanges comme expliqué dans les paragraphes 2.2.2 et 2.2.3. Un signal y est découpé en sous-bandes $(y^f)_{0\leq f\leq F-1}$ définies par

$$y^f = H_f \star y, \tag{B.3}$$

puis en trames temporelles disjointes $(y^{tf})_{0\leq f\leq F-1,\, 0\leq t\leq T-1}$ de 256 échantillons (environ 12 ms), ce qui correspond à une fenêtre w_T rectangulaire et à un pas L de cette taille.

Le filtrage est implémenté de façon rapide en filtrant d'abord le signal par une cascade de décimateurs de facteur 2, puis en appliquant chaque filtre du banc au signal décimé adéquat.

Le seuil de silence est fixé égal au seuil absolu d'audition. Ce seuil dépend du volume auquel le signal est diffusé et de la réponse en amplitude de l'oreille externe et médiane $(g_f)_{0\leq f\leq F-1}$ [Rom03, Vir03b], tracée dans la figure B.1. La log-puissance est alors liée grossièrement au volume sonore perçu, et une valeur nulle est attribuée à tous les points temps-fréquence perçus comme du silence (en réalité les courbes d'isosonie ne sont pas parallèles).

B.1.3 Spectre d'une sinusoïde

Pour définir des spectres harmoniques dans les paragraphes 4.5.3 et 4.6.1, il est utile de pouvoir calculer rapidement le spectre d'une sinusoïde. Le spectre d'amplitude d'une sinusoïde x^{f_0} de fréquence f_0 est défini à une constante multiplicative près par

$$\|x_{tf}^{f_0}\| = \left|\operatorname{sinc}\big((f-f_0)(2L_f+1)\big) + \frac{1}{2}\operatorname{sinc}\big((f-f_0)(2L_f+1)+1\big) + \frac{1}{2}\operatorname{sinc}\big((f-f_0)(2L_f+1)-1\big)\right|. \tag{B.4}$$

En pratique, cette équation doit être complétée par un lissage des spectres de fréquences voisines, car la valeur du spectre de log-puissance correspondant varie beaucoup dans les sous-bandes de faible puissance pour des petites variations de f_0. Nous choisissons de calculer les spectres de log-puissance $\log\|x_{tf}^{f_0}\|^2$ par cette équation pour des fréquences f_0 espacées d'un seizième de ton entre 31 et 10900 Hz. Puis nous redéfinissons chaque spectre $\log\|x_{tf}^{f_0}\|^2$ en transposant tous les spectres voisins $\log\|x_{tf}^{f_0'}\|^2$ à la fréquence f_0 par interpolation linéaire et en les pondérant par une fenêtre gaussienne (d'écart-type 3.6 seizièmes de ton environ).

Le spectre de puissance d'une note harmonique est maintenant défini comme une somme pondérée des spectres de puissance de ses partiels sinusoïdaux. La fréquence fondamentale f_h d'une note de

hauteur h sur l'échelle MIDI s'exprime en Hertz (Hz) par

$$f_h = 440 \times 2^{\frac{h-69}{12}}, \tag{B.5}$$

et les fréquences des partiels sont les multiples de cette fréquence fondamentale.

B.1.4 Réponse fréquentielle des filtres de mélange

La réponse fréquentielle des filtres de mélange se calcule aussi à l'aide du banc de filtres. Chaque filtre de mélange a_{ij} est découpé en sous-bandes $(a_{ij}^f)_{0 \leq f \leq F-1}$ définies par

$$a_{ij}^f = H_f \star a_{ij}. \tag{B.6}$$

La réponse fréquentielle en amplitude de a_{ij} à la fréquence f vaut

$$a_{ijf}^{\text{amp}} = \|a_{ij}^f\|. \tag{B.7}$$

La réponse fréquentielle en phase relative entre $a_{i'j}$ et a_{ij} à la fréquence f vaut

$$a_{i'ijf}^{\text{pha}} = \angle\langle a_{ij}^f, a_{i'j}^f \rangle. \tag{B.8}$$

B.2 Inversion approchée du banc de filtres

B.2.1 Définition des filtres de reconstruction

Dans le cas général, il n'existe pas de formule permettant de reconstruire exactement un signal y à partir de ses sous-bandes $(y^f)_{0 \leq f \leq F-1}$. De plus, les sous-bandes forment une représentation redondante : des sous-bandes quelconques ne correspondent pas forcément à un signal existant. L'inversion d'un banc de filtres consiste donc à estimer un signal $\widehat{y}$ tel que ses sous-bandes $(\widehat{y}^f)$ soient les plus proches possibles de sous-bandes fixées (y^f) pour une certaine distance.

Nous adoptons la formule d'inversion proposée dans [Sla94]

$$\widehat{y} = \frac{2}{3} \sum_{f=0}^{F-1} 2\Re\big(H'_f \star y^f\big), \tag{B.9}$$

où

$$H'_f(u) = \frac{\sqrt{3}}{2} \, (L_f + 1)^{-3/2} w_F^f(u) \exp(2i\pi f u). \tag{B.10}$$

Le gain attribué aux filtres de reconstruction $(H'_f)_{0 \leq f \leq F-1}$ garantit que pour chaque sous-bande le filtre global $H'_f \star H_f$ a une réponse en amplitude unitaire et une réponse en phase nulle à la fréquence centrale f. Le coefficient $2/3$ normalise le résultat pour tenir compte du recouvrement entre sous-bandes.

Le filtrage et la somme sont implémentés de façon rapide à partir des versions décimées des sous-bandes en utilisant une cascade d'interpolateurs de facteur 2.

Cette formule correspond à la formule d'inversion optimale au sens des moindres carrées pour une TFCT [Gri84]. Dans notre cas, elle n'offre qu'une inversion approchée. Généralement l'erreur d'inversion correspond à un RSD de l'ordre de 30 dB à 50 dB, et elle est parfaitement inaudible (car masquée auditivement par le signal estimé). Dans le cadre de l'extraction de sources, cette erreur est donc négligeable par rapport aux autres erreurs d'estimation des sources.

B.2.2 Réestimation de phase

Un autre critère d'inversion consiste à estimer le signal $\widehat{y}$ dont le spectre d'amplitude à court terme ($\|\widehat{y}^{tf}\|$) est le plus proche possible d'un spectre fixé ($\|y^{tf}\|$).

L'inversion est alors basée sur un algorithme itératif exploitant la redondance [Gri84, Cas00]. Les sous-bandes ($\widehat{y}^{f}$) sont initialisées de sorte que $\|\widehat{y}^{tf}\| = \|y^{tf}\|$ en chaque point (t, f). À chaque itération, ces sous-bandes sont normalisées sur chaque trame de sorte que $\|\widehat{y}^{tf}\| = \|y^{tf}\|$, puis elles sont transformées en un signal $\widehat{y}$ par l'équation B.9, puis ce signal est redécomposé en sous-bandes ($\widehat{y}^{f}$) par l'équation B.3. Dans le cas d'une TFCT il est prouvé que cet algorithme réduit à chaque itération l'erreur quadratique entre ($\|\widehat{y}^{tf}\|$) et ($\|y^{tf}\|$).

Cela peut-être utile par exemple pour l'extraction d'une source par soustraction d'énergie. Dans ce cas le spectre d'amplitude des sources est connu, mais leur phase est mal estimée dans les zones temps-fréquence masquées. En partant de l'estimation de phase fournie par le mélange, cet algorithme permet de réestimer la phase dans ces zones en fonction de celle des zones voisines [Cas00]. Cette réestimation n'améliore pas forcément la performance d'extraction lorsque le spectre d'amplitude est mal estimé ou lorsque l'estimation de phase initiale est mauvaise. Dans les expériences présentées dans ce travail, nous avons constaté que l'amélioration de performance éventuelle était trop faible pour justifier son utilisation.

Bibliographie

[Abd01] S.A. Abdallah and M.D. Plumbley. Sparse coding of music signals. 2001. Submitted.

[Abr01] F. Abrard, Y. Deville, and P. White. From blind source separation to blind source cancellation in the underdetermined case : a new approach based on time-frequency analysis. In *Proc. ICA*, pages 734–739, 2001.

[Alb03] B. Albouy and Y. Deville. Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. In *Proc. ICA*, pages 361–366, 2003.

[Ama97] S.I. Amari, S. Douglas, A. Cichocki, and H. Yang. Novel online algorithms for blind deconvolution using natural gradient approach. In *Proc. SYSID*, pages 1057–1062, 1997.

[Ane00] J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. ICA*, 2000.

[Att99] H. Attias. Independent factor analysis. *Neural Computation*, 11 :803–851, 1999.

[Ave02] C. Avendano and J.-M. Jot. Frequency domain techniques for stereo to multichannel upmix. In *Proc. AES 22nd Conference on Vitual, Synthetic and Entertainment Audio*, 2002.

[Bal01] R. Balan, J. Rosca, and S. Rickard. Robustness of parametric source demixing in echoic environments. In *Proc. ICA*, pages 144–148, 2001.

[Bar98] B. Bartlett and J. Bartlett. *Practical recording techniques : the step-by-step approach to professional recording*. Focal Press, 1998.

[BASS-dB] BASS-dB : Blind Audio Source Separation dataBase. Authors : R. Gribonval, E. Vincent and C. Févotte. URL : http ://www.irisa.fr/metiss/BASS-dB/.

[Bel02] J.P. Bello, L. Daudet, and M. Sandler. Time-domain polyphonic transcription using self-generating databases. In *Proc. AES 112th Convention*, 2002.

[Ben01] L. Benaroya. Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Proc. GRETSI*, 2001.

[Ben03] L. Benaroya. *Séparation de plusieurs sources sonores avec un seul microphone*. PhD thesis, Université Rennes I, 2003.

[Bon96] A. Bonafonte, J. Vidal, and A. Nogueiras. Duration modeling with expanded HMM applied to speech recognition. In *Proc. ICSLP*, 1996.

[Bre90] A.S. Bregman. *Auditory scene analysis*. MIT Press, 1990.

[Bro94] G.J. Brown and M.P. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8 :297–333, 1994.

[Bro01] J.C. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the ASA*, 109(3) :1064–1072, 2001.

[Cap93] O. Cappé. *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*. PhD thesis, ENST, 1993.

[Car92] M.-J. Caraty, C. Montacié, and C. Barras. Integration of frequential and temporal structurations in a symbolic learning system. In *Proc. ICSLP*, pages 475–478, 1992.

[Car98a] J.-F. Cardoso. Blind source separation : statistical principles. *Proceedings of the IEEE*, 9(10) :2009–2025, oct. 1998. Special issue on blind identification and estimation.

[Car98b] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. ICASSP*, 1998.

[Cas00] M.A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. ICMC*, 2000.

[Cas02] M.A. Casey. Generalized sound classification and similarity in MPEG-7. *Organized Sound*, 6(2), 2002.

[Cem03] A.T. Cemgil, B. Kappen, and D. Barber. Generative model based polyphonic music transcription. In *Proc. WASPAA*, pages 181–184, 2003.

[Che98] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 1998.

[Dav02a] M.E. Davies. Audio source separation. In *Mathematics in Signal Processing V*. Oxford University Press, 2002.

[Dav02b] M.E. Davies. A sparse mixture of Gaussians model for blind separation of more sources than sensors. In *Proc. N2SP Workshop*, 2002.

[Dav02c] M. Davy and S. Godsill. Bayesian harmonic models for musical pitch estimation and analysis. Technical Report 432, CUED/F-INFENG, 2002.

[Dow03] J.S. Downie. Music Information Retrieval. *Annual review of information science and technology*, 37 :295–340, 2003.

[Dre00] R. Dressler. Dolby Surround Pro Logic II decoder : principles of operation. Dolby Laboratories Information, 2000.

[Dub02] S. Dubnov. Extracting sound objects by independent subspace analysis. In *Proc. AES 22nd Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[Dud02] R. Dudas. Spectral envelope correction for real-time transposition : proposal of a "floating-formant" method. In *Proc. ICMC*, pages 126–129, 2002.

[Dup99] S. Dupuis. Le rôle des transitions legato dans la reconnaissance des instruments de musique. Master's thesis, DEA de Sciences Cognitives, june 1999.

[Egg03] J. Eggink and G.J. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. ISMIR*, 2003.

[Ell96] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, 1996.

[Ero03] A. Eronen. Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proc. ISSPA*, 2003.

[Fit03a] D. Fitzgerald, E. Coyle, and B. Lawlor. Independent subspace analysis using locally linear embedding. In *Proc. DAFx*, 2003.

[Fit03b] D. Fitzgerald, B. Lawlor, and E. Coyle. Prior subspace analysis for drum transcription. In *Proc. AES 114th Convention*, 2003.

[Fév04a] C. Févotte. Bayesian approach for blind separation of undetermined mixtures of sparse sources. In *Proc. ICA*, 2004.

[Fév04b] C. Févotte and C. Doncarli. Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Processing Letters*, 11(3), 2004.

[Gal95] M.J.F. Gales. *Model-based techniques for noise robust speech recognition*. PhD thesis, University of Cambridge, 1995.

[Gha97] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29 :245–273, 1997.

[Gha01] Z. Ghahramani. An introduction to hidden Markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1) :9–42, 2001.

[Gil04] O.K. Gillet and G. Richard. Automatic labelling of tabla signals. In *Proc. ICASSP*, 2004.

[God99] D. Godsmark and G.J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27 :351–366, 1999.

[Got99] M. Goto and S. Hayamizu. A real-time music description system : detecting melody and bass lines in audio signals. In *Proc. IJCAI Workshop on CASA*, pages 31–40, 1999.

[Gri84] D. Griffin and J. Lim. Signal estimation from modified short time Fourier transform. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32 :236–242, 1984.

[Gri99] R. Gribonval. *Approximations non linéaires pour l'analyse des signaux sonores*. PhD thesis, Université Paris IX Dauphine, 1999.

[Gri03a] R. Gribonval. Piecewise linear source separation. In *Proc. SPIE Conference 5207 "Wavelets : Applications in Signal and Image Processing"*, pages 297–310, 2003.

[Gri03b] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proc. ICA*, 2003.

[Hai03] S. Hainsworth and M. Macleod. Onset detection in musical audio signals. In *Proc. ICMC*, 2003.

[Han01] D.J. Hand and K. Yu. Idiot's bayes - not so stupid after all ? *International Statistical Review*, 69(3) :385–398, 2001.

[Hyv00] A. Hyvärinen. The independence assumption : analyzing the independence of the components by topography. In *Advances in Independent Component Analysis*, pages 45–62. Springer, 2000.

[Hyv01a] A. Hyvärinen. Fast ICA by a fixed-point algorithm that maximizes non-Gaussianity. In *Independent Component Analysis : Principles and Practice*, pages 71–94. Cambridge Press, 2001.

[Hyv01b] A. Hyvärinen, J. Karhunen, and E. Oja. Principal component analysis and whitening. In *Independent Component Analysis*. Wiley, 2001.

[Jan03] G.-J. Jang, T.-W. Lee, and Y.-H. Oh. Blind separation of single channel mixture using ICA basis functions. In *Proc. ICA*, pages 595–600, 2003.

[Jen99] K. Jensen. *Timbre models of musical sounds*. PhD thesis, Datalogisk Institut, Copenhague University, 1999.

[Jos99] L. Josifovsky, M. Cooke, P. Green, and A. Vizinho. State-based imputation of missing data for robust speech recognition and speech enhancement. In *Proc. Eurospeech*, 1999.

[Jut03] C. Jutten and J. Karhunen. Advances in nonlinear blind source separation. In *Proc. ICA*, pages 245–256, 2003.

[Kas95] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of bayesian probability network to music scene analysis. In *Working notes of IJCAI Workshop on CASA*, pages 52–59, 1995.

[Kas99] K. Kashino and H. Murase. A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication*, 27 :337–349, 1999.

[Kin99] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. IJCAI Worshop on CASA*, pages 18–24, 1999.

[Kit03] T. Kitahara, M. Goto, and H.G. Okuno. Musical instrument identification based on F0-dependent multivariate normal distribution. In *Proc. ICASSP*, pages 421–424, 2003.

[Kli00] J. Klingseisen and M.D. Plumbley. Towards musical instrument separation using multiple-cause neural networks. In *Proc. ICA*, pages 447–452, 2000.

[Lam99] R.H. Lambert. Difficulty measures and figures of merit for source separation. In *Proc. ICA*, pages 133–138, 1999.

[Lee99] T.-W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4), 1999.

[Mar99a] J. Marques and P. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. Technical report, Compaq Cambridge Research Lab, june 1999.

[Mar99b] K.D. Martin. *Sound-source recognition : a theory and computational model*. PhD thesis, MIT, 1999.

[Mar03] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg. The dependency of timbre on fundamental frequency. *Journal of the ASA*, 114(5) :2946–2957, 2003.

[McD03] L. McDonagh, F. Bimbot, and R. Gribonval. A granular approach for the analysis of monophonic audio signals. In *Proc. ICASSP*, 2003.

[Mer98] Y. Meron and K. Hirose. Separation of singing and piano sounds. In *Proc. ICSLP*, 1998.

[Mis01] J. Miskin and D. MacKay. Ensemble learning for blind source separation. In *Independent Component Analysis : Principles and Practice*, pages 209–233. Cambridge Press, 2001.

[Mit02] N. Mitianoudis and M. Davies. Intelligent audio source separation using independent component analysis. In *Proc. AES 112th Convention*, 2002.

[Mol03] S. Molla and B. Torrésani. An hybrid audio scheme using hidden Markov models of waveforms. 2003. Submitted.

[Mur01] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4) :1–24, 2001. URL : http ://www.ism.ac.jp/˜shiro/research/blindsep.html.

[Nak02] T. Nakatani. *Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition*. PhD thesis, Kyoto University, 2002.

[O'L03] S. O'Leary and N.J.J. Griffith. A hybrid approach to timbral consistency in a virtual instrument. In *Proc. DAFx*, 2003.

[Ort96] S. Ortmanns, H. Ney, and A. Eiden. Language-model look-ahead for large vocabulary speech recognition. In *Proc. ICSLP*, pages 2095–2098, 1996.

[Ost96] M. Ostendorf, V. Digalakis, and O.A. Kimball. From HMMs to segment models : a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5) :360–378, 1996.

[Par00] L.C. Parra and C. Spence. On-line convolutive blind source separation of non-stationary signals. *Journal of VLSI Signal Processing*, 26(1/2) :39–46, 2000.

[Par02] L.C. Parra and C.V. Alvino. Geometric source separation : merging convolutive source separation with geometric beamforming. *IEEE Trans. on Speech and Audio Processing*, 10(6), 2002.

[Pen00] W. Penny, R. Everson, and S. Roberts. Hidden Markov independent component analysis. In *Advances in Independent Component Analysis*. Springer, 2000.

[Pen01] W. Penny, S. Roberts, and R. Everson. ICA : model order selection and dynamic source models. In *Independent Component Analysis : Principles and Practice*. Cambridge Press, 2001.

[Pha01] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. on Signal Processing*, 49(9) :1837–1848, 2001.

[Pon03] N.H. Pontoppidan and M. Dyrholm. Fast monaural separation of speech. In *Proc. AES 23rd Conference on Signal Processing in Audio Recording and Reproduction*, 2003.

[Pul01] V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I : stereophonic panning. *Journal of the AES*, 2001.

[Rab89] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[Rad02] R. Radke and S. Rickard. Audio interpolation. In *Proc. AES 22nd Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[Rap02] C. Raphael. Automatic transcription of piano music. In *Proc. ISMIR*, 2002.

[RG03] M.J. Reyes-Gomez, B. Raj, and D.P.W. Ellis. Multi-channel source separation by factorial HMMs. In *Proc. ICASSP*, 2003.

[Rom03] N. Roman, D. Wang, and G.J. Brown. Speech segregation based on sound localization. *Journal of the ASA*, 114(4) :2236–2252, 2003.

[Ros02] J. Rosier and Y. Grenier. Pitch estimation for the separation of musical sounds. In *Proc. AES 112th Convention*, 2002.

[Row00] S. Roweis. One microphone source separation. In *Proc. NIPS*, pages 793–799, 2000.

[RWC-MDB] RWC Music Database : database of copyright-cleared musical pieces and instrument sounds for research purposes. Authors : M. Goto, H. Hashiguchi, T. Nishimura and R. Oka. URL : http ://staff.aist.go.jp/m.goto/RWC-MDB/.

[Sak03] Y. Sakuraba and H.G. Okuno. Note recognition of polyphonic music by using timbre similarity and direction proximity. In *Proc. ICMC*, 2003.

[Saw03] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In *Proc. ICA*, pages 505–510, 2003.

[Sch99] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proc. ICA*, pages 261–266, 1999.

[Sch00] E. Scheirer. *Music-listening systems*. PhD thesis, MIT, 2000.

[Ser90] X. Serra and J.O. Smith. Spectral modeling synthesis : a sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4) :12–24, 1990.

[Ser03] C. Servière. Separation of speech signals with segmentation of the impulse responses under reverberant conditions. In *Proc. ICA*, pages 511–516, 2003.

[Sla94] M. Slaney, D. Naar, and R.F. Lyon. Auditory model inversion for sound separation. In *Proc. ICASSP*, pages 77–80, 1994.

[SOL] Studio OnLine. URL : http ://forumnet.ircam.fr/.

[Str85] J.M. Strawn. *Modeling musical transitions*. PhD thesis, CCRMA, Stanford University, 1985.

[The03] F.J. Theis, C. Puntonet, and E.W. Lang. A histogram-based overcomplete ICA algorithm. In *Proc. ICA*, pages 1071–1076, 2003.

[Uhl03] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. ICA*, pages 843–848, 2003.

[Vie01] L. Vielva, D. Erdoğmuş, and J.C. Príncipe. Underdetermined blind source separation using a probabilistic source sparsity model. In *Proc. ICA*, pages 675–679, 2001.

[Vin01] E. Vincent. Séparation de signaux audio : principes statistiques de l'analyse en composantes indépendantes et applications au signal monophonique. Master's thesis, DEA ATIAM, 2001.

[Vin03a] E. Vincent, C. Févotte, and R. Gribonval. Comment évaluer les algorithmes de séparation de sources audio ? In *Proc. GRETSI*, 2003.

[Vin03b] E. Vincent, X. Rodet, A. Röbel, C. Févotte, R. Gribonval, L. Benaroya, and F. Bimbot. A tentative typology of audio source separation tasks. In *Proc. ICA*, 2003.

[Vin04a] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. 2004. Submitted to IEEE Trans. on Speech and Audio Processing.

[Vin04b] E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proc. ISMIR*, 2004.

[Vin04c] E. Vincent and X. Rodet. Music transcription with ISA and HMM. In *Proc. ICA*, 2004.

[Vin04d] E. Vincent and X. Rodet. Underdetermined source separation with structured source priors. In *Proc. ICA*, 2004.

[Vir03a] T. Virtanen. Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint. In *Proc. DAFx*, 2003.

[Vir03b] T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *Proc. ICMC*, 2003.

[Vis02] H. Viste and G. Evangelista. An extension for source separation techniques avoiding beats. In *Proc. DAFx*, 2002.

[Vis03] H. Viste and G. Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. DAFx*, 2003.

[Wes99] A. Westner and V.M. Bove. Blind separation of real world audio signals using overdetermined mixtures. In *Proc. ICA*, 1999.

[Wol00] P.J. Wolfe and S.J. Godsill. The application of psychoacoustic criteria to the restoration of musical recordings. In *Proc. AES 108th Convention*, 2000.

[Wol03] P.J. Wolfe and S.J. Godsill. A Gabor regression scheme for audio signal analysis. In *Proc. WASPAA*, pages 103–106, 2003.

[Wu03a] M. Wu and D. Wang. A one-microphone technique for reverberant speech enhancement. In *Proc. ICASSP*, 2003.

[Wu03b] M. Wu, D. Wang, and G.J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Trans. on Speech and Audio Processing*, 11(3) :229–241, 2003.

[Yil02] Ö. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 2002. Submitted.

[Zib01] M. Zibulevsky, B.A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition in a signal dictionary. In *Independent Component Analysis : Principles and Practice*, pages 181–208. Cambridge Press, 2001.