Approche informée pour l'analyse du son et de la musique

Dominique Fourer

N° d'ordre: 4973

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Dominique FOURER**

pour obtenir le grade de

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Approche informée pour l'analyse du son et de la musique

Soutenance le: 11 décembre 2013

Après avis des rapporteurs:

Geoffroy Peeters	Maître de Conférences HDR	Rapporteur
Roland Badeau	Maître de Conférences HDR	Rapporteur

Devant la commission d'examen composée de:

Sylvain Marchand	Professeur des universités	Directeur de thèse
Geoffroy Peeters	Maître de conférences HDR	Rapporteur
Roland Badeau	Maître de conférences HDR	Rapporteur
Myriam Desainte-Catherine	Professeure des universités	Examinatrice
Frédéric Bimbot	Directeur de recherche	Examinateur
Charles Dossal	Maître de Conférences	Examinateur

- 2013 -

REMERCIEMENTS

« Ce n'est pas parce que les choses sont difficiles que nous n'osons pas, mais parce que nous n'osons pas qu'elles sont difficiles », Sénèque

Je tiens particulièrement à remercier mon directeur de thèse, Sylvain Marchand pour son expertise scientifique mais aussi pour ses qualités humaines et sa grande disponibilité. Je le remercie pour son accompagnement efficace et sa confiance qui m'aura permis de développer librement des approches originales dans le domaine émergeant de l'analyse informée des signaux audio.

Mes remerciements vont également à Isabelle Dutour qui m'a accueilli au sein du département informatique de l'IUT de Bordeaux 1, dans un cadre favorable qui m'a permis d'acquérir une expérience dans l'enseignement tout en menant à bien mes travaux.

Je remercie mes rapporteurs de thèse, Roland Badeau et Geoffroy Peeters ainsi que tous les autres membres du jury de m'avoir fait l'honneur de participer à ma soutenance : Frédéric Bimbot président du jury, Myriam Desainte-Catherine et Charles Dossal examinateurs.

Merci à tous les enseignants-chercheurs qui ont pris le temps d'échanger avec moi lors de mes premiers pas dans le monde de la recherche académique. En particulier Gilles Zemor professeur à l'Institut de Mathématiques de Bordeaux (IMB), Youssou Dieng et Romain Bourqui maîtres de conférences au LaBRI.

Enfin, ce travail aurait été bien plus difficile sans le soutien bienveillant de ma famille et de mes proches vers qui j'exprime ma plus grande gratitude.

TABLE DES MATIÈRES

Remerciements **iii**

Table des matières **iii**

Abréviations **viii**

Notations **x**

Introduction **1**

I Approche classique pour les problèmes d'analyse en traitement du son **5**

1 Représentation des signaux musicaux **7**
1.1 Son numérique 7
1.1.1 Échantillonnage d'un signal audio 8
1.1.2 Reconstruction des signaux échantillonnés 8
1.1.3 L'algorithme de Voronoï-Allebach 9
1.1.4 Le codage des signaux audio 11
1.2 Décomposition temps-fréquence 12
1.2.1 La transformée de Fourier fenêtrée 12
1.2.2 Modèles utilisés pour les signaux audio 14
1.3 Paramètres musicaux 16
1.3.1 Amplitude et volume 16
1.3.2 Enveloppe temporelle et transitoires 16
1.3.3 Fréquence fondamentale et hauteur 17
1.3.4 Enveloppe spectrale et timbre 20
1.3.5 Rythme et pulsation 21
1.4 Conclusions du chapitre 21

2 Problèmes d'analyse en traitement du signal audio **23**
2.1 L'estimation des paramètres sinusoïdaux des signaux musicaux 23
2.1.1 Le modèle sinusoïdal à court terme 24
2.1.2 Estimation des sons polyphoniques bruités 27

2.1.3 La non stationnarité . . . 28
2.1.4 Le modèle sinusoïdal à long terme : suivi de partiels . . . 29
2.1.5 Évaluation de la qualité des méthodes d'estimation . . . 30
2.2 La transcription automatique de la musique . . . 31
2.2.1 Détection des activations de note . . . 31
2.2.2 Modèle de signal d'un instrument tonal . . . 32
2.2.3 L'estimation de la fréquence fondamentale des sons monophoniques 33
2.2.4 Évaluation des systèmes de transcription . . . 35
2.3 La séparation de sources . . . 38
2.3.1 Description du problème . . . 38
2.3.2 Identification de la configuration du problème . . . 39
2.3.3 Formalisme et modèle du mélange . . . 40
2.3.4 Techniques usuelles pour la séparation de sources audio . . . 40
2.3.5 Évaluation des méthodes de séparation de sources . . . 43
2.4 Conclusions du chapitre . . . 46

II Approche informée pour l'analyse des signaux audio numériques 47

3 Estimation et codage de l'information 49
3.1 Théorie de l'estimation . . . 50
3.1.1 Définition d'un estimateur . . . 51
3.1.2 Approche fréquentiste (déterministe) de l'estimation . . . 52
3.1.3 Approche bayésienne de l'estimation . . . 53
3.1.4 Vers le codage de l'information manquante . . . 54
3.2 Théorie de l'information . . . 54
3.2.1 Mesure de l'information . . . 54
3.2.2 Codage de l'information . . . 55
3.2.3 Codage entropique sans perte . . . 58
3.2.4 Limitations du codage sans perte . . . 60
3.2.5 Quantification et codage de source . . . 61
3.3 Approche généralisée pour l'analyse informée . . . 65
3.3.1 État de l'art des techniques combinant estimation et codage . . . 65
3.3.2 Formulation du problème de l'analyse informée . . . 68
3.3.3 Analyse informée d'un seul paramètre . . . 69
3.3.4 Généralisation au cas vectoriel pour l'analyse informée . . . 71
3.4 Conclusions du chapitre . . . 72

4 L'analyse spectrale informée 75
4.1 La modélisation sinusoïdale des sons musicaux . . . 76
4.1.1 Extraction et estimation des composantes sinusoïdales . . . 76
4.1.2 La méthode de réallocation . . . 77
4.1.3 Bornes théoriques . . . 78
4.2 L'analyse spectrale informée dans le cas scalaire . . . 79
4.2.1 Principes . . . 79
4.2.2 Simulations . . . 79
4.3 Analyse spectrale informée dans le cas vectoriel . . . 84
4.3.1 Principes . . . 84
4.3.2 Quantification optimale . . . 84
4.3.3 Simulations . . . 87
4.4 Application à la séparation de sources informée . . . 91

4.4.1 Modèle de source et estimation des paramètres 91
4.4.2 Calcul et codage du masque pour chaque source 92
4.4.3 Tatouage audio numérique 93
4.4.4 Implémentation 94
4.4.5 Complexité calculatoire 94
4.4.6 Expérimentation et résultats 95
4.5 Conclusions du chapitre 97

5 Approche informée appliquée à l'extraction de données symboliques à partir d'un signal audio 101
5.1 Étude et proposition d'un système complet d'estimation F_0 multiple . . . 103
5.1.1 Extraction des composantes tonales 103
5.1.2 Construction des hypothèses F_0 107
5.1.3 Évaluation des candidats F_0 110
5.1.4 Estimation de la polyphonie 114
5.1.5 Évaluation comparative du système de transcription proposé . . . 115
5.2 La transcription polyphonique informée 120
5.2.1 Formulation du problème 120
5.2.2 Analyse des erreurs des systèmes de transcription polyphonique . 121
5.2.3 Systèmes de codage proposés 122
5.2.4 Évaluation comparative des systèmes de transcription proposés . . 132
5.3 Application à la séparation de sources informée 135
5.3.1 Aperçu de la méthode 135
5.3.2 Le tatouage audio numérique 136
5.3.3 La séparation de sources informée par la partition 136
5.3.4 Évaluation 138
5.4 Conclusion du chapitre 148

Conclusion 151

Annexe A Calculs détaillés 155
A.1 Preuve d'impossibilité de l'inversion du tatouage audio numérique basé sur la technique QIM 155
A.1.1 Principe de la quantification QIM 155
A.1.2 Démonstration 156
A.2 Calcul de la borne de Shannon pour les paramètres sinusoïdaux 157
A.2.1 Modèle sinusoïdal 157
A.2.2 Mesure de distorsion 157
A.2.3 Borne Shannon 158

Bibliographie de l'auteur 163

Bibliographie générale 165

Table des figures 179

Liste des tableaux 185

Abréviations

AM modulation d'amplitude, *Amplitude Modulation* 29

CASA *Computational Auditory Scene Analysis* 30

CQT transformée Q constant, *Constant Q Transform* 14

DPCM *Differential Pulse-Code Modulation* 66

EM Espérance-Maximisation 43

ECUSQ quantification sphérique avec dépendance des paramètres et contrainte d'entropie, *Entropy Constrained Unrestricted Spherical Quantization* 84

EQM erreur quadratique moyenne 51

ERB *Equivalent Rectangular Bandwidth* 17

FFT transformée de Fourier rapide, *Fast Fourier Transform* 79

FM modulation de fréquence, *Frequency Modulation* 29

HPS hypothèse de source harmonique, *Hypothetical Partial Sequence* 107

ICA analyse en composantes indépendantes, *Independant Component Analysis* 16

intMDCT transformée en cosinus discrète modifiée à valeurs entières, *integer Modified Discrete Cosinus Transform* 61

ISS séparation de sources informée, *Informed Sources Separation* 68

LSB bit de poids le plus faible, *Least Significant Bit* 136

MDCT transformée en cosinus discrète modifiée, *Modified Discrete Cosinus Transform* . . 93

MIDI *Musical Instrument Digital Interface* 2

MIR extraction d'informations musicales, *Music Information Retrieval* 101

MSB bit de poids le plus fort, *Most Significant Bit* 69

MSE erreur quadratique moyenne, *Mean Squared Error* 53

QIFFT interpolation quadratique de la transformée de Fourier rapide, *Quadratically Interpolated Fast Fourier Transform* 25

QIM quantification à modulation d'indice, *Quantization Index Modulation* 71

NMF factorisation en matrices non négatives, *Non-negative Matrix Factorization* 16

SNR rapport signal / bruit, *Signal-to-Noise Ratio* 30

STB sinusoïdes / transitoires / bruit 14

STFT transformée de Fourier à court terme, *Short-Time Fourier Transform* 12

TF temps-fréquence 12

WMSE erreur quadratique moyenne pondérée, *Weighted Mean Squared Error* 84

Notations

Ce document contient des expressions mathématiques pour lesquelles nous adopterons les conventions de notation suivantes.

$s(t)$	Signal continu
$s[n]$	Signal discret
$S_w(\omega)$	Transformée de Fourier (continue) à court terme du signal $s(t)$ utilisant la fenêtre d'analyse $w(t)$
$\hat{x}$	Estimation usuelle (non informée) de x
$\tilde{x}$	Estimation informée de x
x_i	Coefficient i du vecteur x
X	Matrice X formée des coefficients X_{ij}
$\mathbf{j}$	Unité imaginaire ($\mathbf{j}^2 = -1$)
$\mathbf{Im}\,(z)$	Partie imaginaire de $z \in \mathbb{C}$
$\mathbf{Re}\,(z)$	Partie réelle de $z \in \mathbb{C}$
z^*	Conjugué complexe de $z \in \mathbb{C} : z^* = \mathbf{Re}\,(z) - \mathbf{jIm}\,(z)$
$\angle z$	Argument du complexe z
$\lvert z \rvert$	Module du complexe $z : \lvert z \rvert = \sqrt{\mathbf{Re}\,(z)^2 + \mathbf{Im}\,(z)^2}$

Opérateurs

$\langle x, y \rangle$	Produit scalaire entre x et $y : \langle x, y \rangle = \sum_{i=1}^{I} x_i y_i^*$
$*$	Convolution $(f * g)(\tau) = \int f(t) g(\tau - t)\, dt$
$\lceil x \rceil$	Première valeur entière supérieure ou égale à $x : \min\{n \in \mathbb{Z} \lvert n \geq x\}$.
$\lfloor x \rfloor$	Première valeur entière inférieure ou égale à $x : \max\{n \in \mathbb{Z} \lvert n \leq x\}$.
$\lvert\lvert x \rvert\rvert_p$	Norme p du vecteur $x : \lvert\lvert x \rvert\rvert_p = \left(\sum_{i=1}^{I} \lvert x_i \rvert^p \right)^{\frac{1}{p}}$
$n!$	Factorielle de l'entier $n : n! = \prod_{i=1}^{n} i$

Fonctions

$\mathrm{E}[x]$	Espérance mathématique de la variable aléatoire $x \in \Omega$ calculée à partir de sa densité de probabilité $f(x)$ telle que $E[x] = \int_\Omega f(x) x\, dx$
$\mathrm{V}[x]$, $\mathrm{var}(x)$	Variance de la variable aléatoire $x : V[x] = E[x - E[x]]^2 = E[x]^2 - E[x^2]$
$\mathbf{e}^x$, $\exp(x)$	Fonction exponentielle où $\mathbf{e} \approx 2.71828$ est la constante de Neper
$\log(x)$	Fonction logarithme népérien
$\log_b(x)$	Fonction logarithme dans la base b tel que $\log_b(x) = \frac{\log(x)}{\log(b)}$
$\mathcal{F}[f](\omega)$	Transformée de Fourier de la fonction $f(t)$ (opérateur linéaire)
$\mathcal{F}^{-1}[\hat{f}](t)$	Transformée de Fourier inverse de $\hat{f}(\omega)$ (opérateur linéaire)
$\mathbf{1}_{[a,b]}$	Fonction indicatrice sur l'ensemble $[a, b] : \mathbf{1}_{[a,b]}(x) = \begin{cases} 1 & \text{si } x \in [a, b] \\ 0 & \text{sinon} \end{cases}$
$\mathrm{sinc}_T(t)$	Fonction sinus cardinal normalisé définie par $\mathrm{sinc}_T(t) = \frac{\sin(\pi t / T)}{\pi t / T}$

INTRODUCTION

L'évolution des techniques de traitement du son a suivi celle du traitement de l'information grâce à l'informatique [Roa96]. Jusqu'aux années 60, les studios d'enregistrement de musique effectuaient des prises de son directes où les musiciens devaient jouer simultanément. De tels systèmes étaient conçus pour obtenir la meilleure prise d'ensemble possible afin d'être enregistrée sur un support analogique comme un disque phonographique ou une bande magnétique. Dans une telle configuration, le mélange sonore final était fidèle à l'acquisition. Cependant, il était très difficile, voire impossible d'appliquer certains effets post-enregistrement, ou d'isoler un son particulier du reste du mélange.

De nos jours, les techniques modernes permettent d'effectuer des enregistrements multipistes qui sont stockés sur support numérique et traités par un ordinateur. Le mélange final est obtenu après une étape de mastérisation ou *mastering* durant laquelle des effets peuvent être appliqués sur certains instruments ou sur l'ensemble du mélange. La création d'un mélange sonore final à partir de pistes séparées est donc dépendante des choix subjectifs de l'ingénieur du son. Ainsi, les studios d'enregistrement actuels possèdent presque toujours les signaux correspondant à chaque partie d'un mélange. On parle alors plus généralement d'images ou de *stems*. Leur mise à la disposition de l'auditeur rend ainsi possible la création de nouveaux mélanges sur lesquels il est possible d'appliquer des effets, de spatialiser les instruments ou de supprimer certaines parties (effet karaoké). On parle désormais d'écoute active [Lep98] de la musique, notamment lorsque ces possibilités sont applicables en temps réel par l'auditeur. Malgré les questions de droits soulevées par cette nouvelle approche de la musique, nous étudions dans cette thèse une configuration où des informations connues avant la création d'un mélange musical peuvent être exploitées pour rendre possibles certaines transformations sur celui-ci.

L'analyse des sons musicaux

L'analyse dans le domaine du traitement du son regroupe l'ensemble des techniques ayant pour but de comprendre et de décrire le contenu des signaux audio. Ainsi, le son devient une dimension à part entière capable de véhiculer de l'information pouvant être traitée et interprétée par un être humain ou une machine. Parmi les nombreuses applications de l'analyse audio, nous pouvons citer :

- la manipulation et la création de contenu musical (*e.g.* écoute active [Lep98]) qui permet à un auditeur de manipuler en temps réel les entités sonores qui composent un mélange musical,
- le codage et la compression de signaux audio (*e.g.* MPEG [DM01]),
- la restauration de contenu musical [God93],
- la recherche dans une base de données et la classification,
- la musicologie et la pédagogie,
- etc.

La pertinence des informations extraites du son dépend fortement des connaissances *a priori* sur celui-ci. Par exemple, un être humain bien entraîné est en général capable d'identifier facilement une pièce musicale, les différents instruments présents et parfois même définir la hauteur des notes jouées ou le tempo. Au contraire, sans entraînement, ou dans le cas d'un traitement automatisé, ces tâches d'analyse s'avèrent beaucoup plus difficiles. Ainsi de nombreuses études proposent des modèles permettant de décrire et d'extraire des informations à partir d'un signal audio.

Le lecteur pourra par exemple consulter [Hai06] pour un état de l'art sur l'estimation du tempo. L'identification du timbre d'un instrument est traitée dans [HBKD06]. La transcription automatique permettant d'estimer la partition jouée à partir d'un signal audio est traitée par Klapuri dans [Kla03]. La séparation des signaux correspondants aux différents instruments composant une pièce musicale est présentée par exemple dans [DZZS08].
Ainsi, l'ensemble de ces travaux portant le nom d'extraction d'information musicales est traitée par la communauté pluridisciplinaire MIR (*Musical Information Retrieval*). Parmi cette communauté cohabitent des mathématiciens, physiciens et informaticiens qui s'intéressent respectivement à la musique sous sa représentation physique (signal correspondant aux variations de la pression de l'air ou de la membrane d'un capteur) et sous sa représentation symbolique (partition de musique, fichier MIDI[1] [Moo86], notation *piano Roll*, etc.). Dans ce document, nous revisitons certains des problèmes d'analyse grâce à l'approche informée qui sera développée dans la suite du document.

L'importance de l'efficacité pour l'analyse

L'analyse est la première étape de la chaîne de traitement classique utilisée en traitement du signal audio (*cf.* figure 1) dont elle est le plus souvent indissociable. Il est donc aisé de déduire que l'efficacité d'un tel système dépend en premier lieu de la précision de l'analyse.

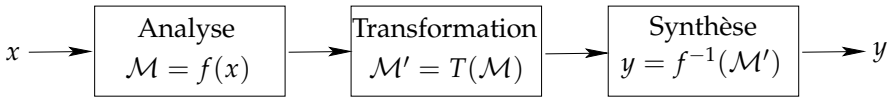


FIGURE 1 – Chaîne de traitement du signal audio permettant d'obtenir un signal de sortie y à partir d'une entrée x en passant par un modèle $\mathcal{M}$.

Cependant, malgré des efforts importants pour améliorer les méthodes d'analyse existantes, certaines techniques comme celles permettant la séparation de sources audio

[1] *Musical Instrument Digital Interface* (MIDI)

aveugle, ne possèdent que peu d'applications pratiques en raison de la qualité insuffisante pour les signaux résultants. C'est aussi la raison pour laquelle les approches plus récentes deviennent plus efficaces en exploitant de l'information supplémentaire *a priori* ou en exploitant des bases de données permettant un apprentissage supervisé [SRMS06]. Malgré les améliorations offertes par de tels systèmes, il existe toujours des limitations théoriques qui ne peuvent être franchies qu'en transmettant, lorsque c'est possible, l'information nécessaire et suffisante manquante permettant d'atteindre la qualité souhaitée.

Vers une approche informée pour les problèmes d'analyse

Nous avons choisi d'orienter nos travaux sur le traitement des signaux de musique (souvent les plus complexes à traiter que les signaux de parole). Notre étude se base sur deux domaines de représentation des signaux audio :

- La décomposition et la représentation physique des signaux (*e.g.* décomposition temps-fréquence, paramètres d'un modèle de forme d'onde du signal, etc.).
- L'extraction d'informations symboliques abstraites utilisées pour la représentation de la musique (*e.g.* hauteur musicale d'une note jouée, débuts et fins de note...).

Le premier domaine de représentation choisi a pour but d'analyser objectivement les signaux traités en utilisant les modèles existants permettant de décrire la forme d'onde des sons manipulés. Le second domaine fait intervenir certains descripteurs subjectifs qui dépendent des conventions utilisées par l'homme pour représenter et organiser la perception de la musique.

Structure du document

Nous avons choisi de séparer cette thèse en deux parties correspondant respectivement à l'approche classique et à l'approche informée appliquées aux problèmes d'analyse en traitement du signal audio.

La première partie est composée de deux chapitres introductifs :

- Nous présentons dans le chapitre 1 des éléments de base en traitement du signal audio ainsi que quelques notions de musique. Ce premier chapitre a pour objectif de présenter les domaines de recherche auxquels se rapportent les travaux de cette thèse. Nous y décrivons les outils ainsi que les modèles usuels développés par la suite.

- Le chapitre 2 a pour objectif d'introduire et de détailler les différents problèmes d'analyse du signal audio que nous traitons dans cette thèse : la modélisation spectrale des signaux audio, la transcription automatique et la séparation de sources. Le lecteur trouvera dans ce chapitre une formulation précise de ces problèmes, un état de l'art et une méthodologie permettant l'évaluation des solutions proposées.

Dans la seconde partie nous proposons des contributions originales portant sur le thème de l'analyse informée réparties en trois chapitres :

- Dans le chapitre 3 principalement bibliographique, nous nous efforçons d'établir un lien entre la théorie de l'estimation et la théorie de l'information. En effet, ces

deux domaines souvent traités séparément dans la littérature possèdent de nombreux liens permettant de les combiner dans le cadre d'une approche hybride comportant simultanément de l'estimation et du codage. Ainsi, ce chapitre apporte des éléments de solutions théoriques et pratiques au problème de l'analyse informée. Nous y décrivons une première contribution proposant un cadre d'application général permettant d'utiliser cette nouvelle approche.

- Le chapitre 4 propose un contribution utilisant l'approche informée appliquée à l'analyse spectrale des signaux audio. Les travaux présentés dans ce chapitre s'appuient principalement sur le modèle sinusoïdal aujourd'hui bien connu dans la littérature. Nous proposons ainsi une extension de l'étape d'analyse permettant de dépasser les limitations théoriques établies et offrant de nouvelles perspectives d'applications telles que l'écoute active de la musique ou la séparation de sources.

- Dans le chapitre 5, nous traitons le problème d'extraction d'informations symboliques musicales avec l'approche informée. Nous avons choisi de traiter dans ce chapitre le problème de la transcription automatique polyphonique de la musique à partir d'un signal audio. Il s'agit d'un problème difficile qui est aussi à la base de nombreuses applications musicales telles que l'estimation du rythme, la détection des accords, l'estimation de la structure d'une pièce musicale et même la séparation de sources.

Première partie

Approche classique pour les problèmes d'analyse en traitement du son

CHAPITRE 1

REPRÉSENTATION DES SIGNAUX MUSICAUX

Un son est une onde qui se caractérise par l'oscillation des particules du milieu à l'intérieur duquel ce son se propage (*e.g.* air, eau, matériau). Un son ne peut donc pas exister dans le vide. Lorsque l'on cherche à analyser un son, on mesure les variations de pression exercées dans son milieu environnant (le plus souvent, l'air) en fonction du temps. Dans ce cas, on parle de signal temporel. Ainsi, on définit un signal temporel par une fonction $s(t)$ définie sur $\mathbb{R}$, qui retourne une mesure de pression acoustique en fonction du temps t. Un signal est centré en 0 ou de moyenne nulle quand $E[s(t)] = 0$. Lorsque $s(t) = 0$ pour $t < 0$, on dit que ce signal est causal. Il s'agit d'une hypothèse qui s'applique aux signaux à durée limitée et qui se vérifie quand on tronque un signal (pour procéder à son analyse ou pour l'échantillonner).

1.1 Son numérique

Le son numérique est obtenu à partir d'un signal analogique naturel produit par une source sonore. Ainsi, le signal analogique continu (pression acoustique) capturé est transformé en un signal numérique discret (nombre de mesures fini) à l'aide d'un capteur qui mesure les variations de pression et d'un système d'acquisition qui procède à son échantillonnage. Un système d'acquisition classique est souvent composé d'un ou de plusieurs microphones reliés à une carte son. Celle-ci envoie à un ordinateur la valeur des échantillons en vue de leur enregistrement sur un système de stockage de données (*e.g.* disque dur, carte mémoire, etc.). Les échantillons sont ensuite enregistrés dans un fichier dont le format varie en fonction de l'utilisation. Les formats de données brutes non compressées (*e.g.* WAVE, AIFF, etc.) sont en général utilisés pour la manipulation des sons car ils permettent d'accéder directement aux valeurs des échantillons ce qui facilite l'analyse et les transformations appliquées sur le signal. Pour l'archivage, on privilégie les formats de données compressées sans perte (*e.g.* FLAC, ALAC, etc.) qui n'altèrent pas la qualité originale et qui permettent de retrouver la valeur exacte des échantillons au prix d'une étape de compression / décompression coûteuse en temps de calcul. Les formats de données compressées avec perte (*e.g.* MP3, OGG, AAC, etc.) sont obtenus par quantification

(*cf.* chapitre 3) des signaux échantillonnés et offrent un compromis entre la qualité perceptive et la taille des données. Ces formats utilisent un modèle psychoacoustique afin de réduire au maximum la taille des données permettant de représenter un signal tout en minimisant la dégradation engendrée sur la qualité sonore perçue. Ces formats offrent des taux de compression très supérieurs aux formats de compression sans perte et sont couramment utilisés pour l'archivage et les échanges sur des supports de stockage limités en espace et/ou en bande passante.

1.1.1 Échantillonnage d'un signal audio

On considère un signal continu noté $s(t)$. Le signal discret correspondant noté $s[t]$ pouvant être traité par un système numérique est obtenu en échantillonnant le signal $s(t)$. Le processus d'échantillonnage consiste à enregistrer des valeurs de la fonction $s(t)$ à des instants donnés. Dans le cas d'un échantillonnage régulier (le plus usuel), on fixe une période de temps notée T_s entre chaque mesure. L'échantillonnage consiste à associer à chaque valeur $s(nT_s)$ un Dirac localisé à l'instant $t = nT_s$. La relation entre le signal discret et le signal continu peut être exprimée par :

$$s[t] = \sum_{n=-\infty}^{\infty} s(nT_s)\delta(t - nT_s), \tag{1.1}$$

où $\delta(n) = \begin{cases} 1 & \text{si } n = 0 \\ 0 & \text{sinon.} \end{cases}$ est une fonction de Dirac discrète. Ainsi, la fréquence d'échantillonnage $F_s = \frac{1}{T_s}$ correspond au nombre d'échantillons mesurés en une seconde. Le théorème d'échantillonnage de Shannon-Nyquist initialement démontré par Whittaker [Whi35] définit la fréquence maximale F_{Nyquist} pouvant être représentée à partir d'un signal échantillonné comme la moitié de la fréquence d'échantillonnage :

$$F_{\text{Nyquist}} = F_s/2. \tag{1.2}$$

L'échantillonnage d'un signal périodise son spectre. En cas de sous-échantillonnage, il apparaît un phénomène de repliement du spectre ou *aliasing*. Le repliement du spectre engendre une distorsion du signal original qui ne peut alors plus être reconstruit sans information complémentaire.

Les limitations théoriques définies par le théorème de l'échantillonnage peuvent être contournées en effectuant de l'échantillonnage compressé ou *Compressive Sensing* [Don06]. Cette technique consiste à choisir une base de représentation parcimonieuse adaptée aux signaux analysés. Ainsi, le signal décomposé peut être décrit avec un nombre d'échantillons non nuls inférieur au nombre d'échantillons classique utilisant une base canonique. De telles propriétés sont évidemment utilisées pour la compression, la modélisation et la reconstruction des signaux. Ainsi, Dossal *et al.* [DPF09] propose une exploration théorique des conditions permettant la reconstruction exacte ou partielle des signaux obtenus par échantillonnage compressé. Parmi les nombreuses applications basées sur cette approche, nous pouvons citer la *superresolution* [KTL11] ou la restauration de signaux partiellement dégradés (*inpainting*)[AEJ+12]. Nous comprenons ainsi l'intérêt des modèles de signaux parcimonieux pour la représentation des signaux audio de musique (*cf.* section 2.1).

1.1.2 Reconstruction des signaux échantillonnés

D'après le théorème de Shannon-Whittaker [Sha49, Whi28], il est possible de reconstruire exactement le signal continu à partir de ses valeurs échantillonnées $\{s(nT_s)\}_{n\in\mathbb{Z}}$ si

sa bande fréquence est incluse l'intervalle $[-\pi/T_s; \pi/T_s]$ en appliquant l'équation (1.3). Même lorsque cette condition n'est pas vérifiée, l'équation (1.3) permet d'obtenir une approximation du signal reconstruit. Cette reconstruction s'obtient en appliquant un filtre passe-bas sur le signal discret dont la réponse impulsionnelle s'exprime à partir d'une fonction sinus cardinal :

$$s(t) = \sum_{n=-\infty}^{\infty} s(nT_s) \frac{\sin(\frac{\pi(t-nT_s)}{T_s})}{\frac{\pi(t-nT_s)}{T_s}} = (s_{T_s} * \mathrm{sinc}_{T_s})(t), \tag{1.3}$$

où s_{T_s} correspond au signal s échantillonné en utilisant une période T_s. Intuitivement, ce théorème se comprend lorsque l'on observe le spectre d'amplitude d'un signal échantillonné. En effet, l'échantillonnage d'un signal périodise son spectre en ajoutant des informations dans les hautes fréquences. La reconstruction consiste donc à projeter le signal discret sur son support fréquentiel $\Omega = [-\frac{F_s}{2}; \frac{F_s}{2}]$ (Bande passante définie par la fréquence de Nyquist $\frac{F_s}{2}$). Cette projection s'effectue en base de Fourier (domaine fréquentiel) par un produit avec la fonction porte $\mathbf{1}_\Omega$ qui correspond à appliquer un filtrage par la fonction sinus cardinal dans la base canonique (domaine temporel). En effet :

$$\mathcal{F}\left[\frac{\sin(\frac{\pi t}{T_s})}{\frac{\pi t}{T_s}}\right] = \mathbf{1}_\Omega. \tag{1.4}$$

1.1.3 L'algorithme de Voronoï-Allebach

Dans le cas d'un signal échantillonné irrégulièrement, l'algorithme Voronoï-Allebach [SA87, Str95] offre une solution pratique au problème de reconstruction des signaux. Cet algorithme repose sur le théorème de Shannon-Whittaker qui suppose le signal d'origine est limité en bande de fréquences. La reconstruction de signaux irrégulièrement échantillonnés s'obtient en appliquant un traitement itératif comprenant les étapes suivantes :

- reconstruction du signal à partir de ses échantillons en utilisant un opérateur d'approximation (*e.g.* interpolation linéaire, interpolation par courbe spline, interpolation de Voronoï),
- projection du signal sur le support Ω en effectuant un filtrage par la fonction sinc_{T_s} telle que $\Omega = [-\frac{1}{2T_s}; +\frac{1}{2T_s}]$,
- projection sur Ω de l'erreur d'interpolation (calculée pour les échantillons connus) puis addition avec la reconstruction précédente.

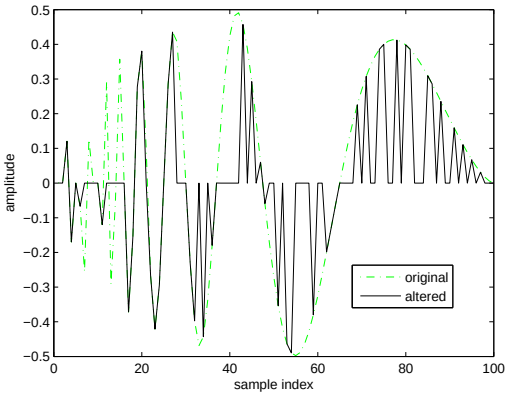
Ainsi, le signal reconstruit à l'itération i noté $s^{(i)}$ peut être formulé comme suit :

$$s^{(i)} = \mathrm{Vor}\left(s - s^{(i-1)}\right) * \mathrm{sinc}_{T_s} + s^{(i-1)} \tag{1.5}$$

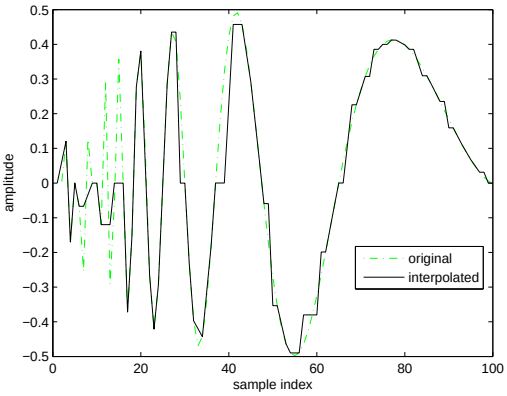
avec s le signal irrégulièrement échantillonné (échantillons manquants mis à 0) et $s^{(0)}$ la reconstruction initiale égale au vecteur nul. Ici $\mathrm{Vor}(.)$ est la fonction d'interpolation de Voronoï qui affecte aux échantillons interpolés la valeur de l'échantillon connu le plus proche (*cf.* figure 1.1) et pouvant être formulé comme suit :

$$\mathrm{Vor}(s[n]) = \begin{cases} s[a] & si \quad n < \frac{a+b}{2} \\ \frac{s[a]+s[b]}{2} & si \quad n = \frac{a+b}{2} \\ s[b] & si \quad n > \frac{a+b}{2} \end{cases}, \tag{1.6}$$

pour $\forall n \in [a;b]$ où a et b correspondent aux indices des échantillons connus les plus proches de l'indice n.

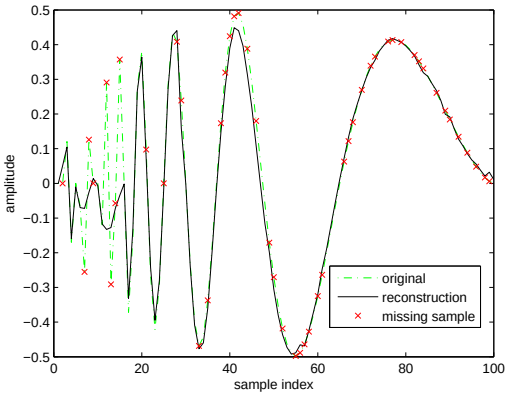


a) signal d'origine avec 50% des échantillons manquants (SNR=3dB)

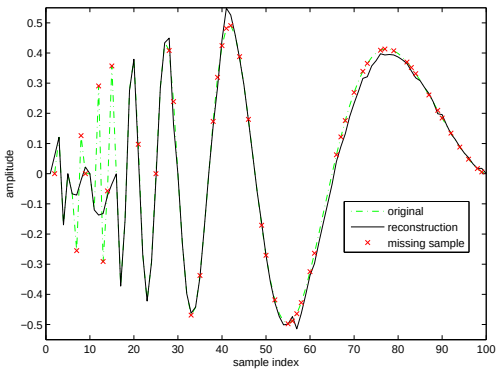


b) signal interpolé (SNR=11dB)

FIGURE 1.1 – Signal sinusoïdal modulé linéairement en fréquence et échantillonné aléatoirement 1.1a) et son interpolation de Voronoï 1.1b).



a) après 10 itérations (SNR=13dB)



b) après 50 itérations (SNR=13dB)

FIGURE 1.2 – Reconstruction d'un signal sinusoïdal modulé linéairement en fréquence échantillonné aléatoirement en utilisant l'algorithme de Voronoï-Allebach.

Comme le montre la figure 1.2, l'algorithme Voronoï-Allebach permet de reconstruire un signal échantillonné irrégulièrement. Ce type de reconstruction a pour effet de lisser la fonction reconstruite résultant du filtrage passe-bas tout en exploitant l'ensemble des échantillons connus de la fonction contrairement aux interpolations locales qui n'utilisent que le voisinage des segments à reconstruire. Cet algorithme présente quelques limitations : il converge lentement d'après [Str95] et le signal doit nécessairement être limité en bande. De plus, la solution obtenue dépend de la répartition des échantillons. Dans cet algorithme, la fonction $\mathrm{Vor}(.)$ peut être remplacée par une autre fonction interpolatrice ce qui a un effet sur la vitesse de convergence. Il existe dans la littérature d'autres méthodes équivalentes plus ou moins efficaces détaillées dans [FGS95, Str97, Mar01].

1.1.4 Le codage des signaux audio

Après avoir procédé à l'échantillonnage d'un signal audio, il est nécessaire d'enregistrer la valeur des échantillons sur un support de stockage numérique. Les seules données manipulées par un ordinateur étant des séquences de bits souvent regroupés par mots de 8 bits (les octets), 16 bits ou 32 bits, il convient de choisir une organisation adéquate. Ainsi, le codage d'un signal comprend en général plusieurs étapes :

- la **quantification** qui intervient au moment de l'échantillonnage a pour but de définir un ensemble fini de représentants (dictionnaire ou quantificateur) utilisé pour enregistrer la valeur des échantillons. La quantification s'accompagne toujours d'une perte d'information liée à la taille du dictionnaire choisi. Un dictionnaire de taille plus importante permet ainsi de représenter un plus grand nombre de valeurs distinctes et s'accompagne ainsi d'un gain en précision. Cependant, il faudra un nombre de bits plus important pour que chaque élément du dictionnaire possède un code distinct afin de rendre possible son décodage. La quantification peut aussi être appliquée sur des signaux déjà échantillonnés lorsque l'on souhaite compresser ces signaux en tolérant une perte d'information (et donc de qualité).

- **la compression** ou **codage de source** est l'étape permettant de définir un mot (séquence de bits) associé à chaque élément ou ensemble d'éléments du dictionnaire choisi lors de la quantification. Un codage simple mais peu efficace car sans compression consiste à utiliser la représentation binaire du numéro de chaque élément unique appartenant à un dictionnaire. Nous comprenons dans le chapitre 3 qu'un tel codage ne tient pas compte de la fréquence ou probabilité d'apparition de chaque élément du dictionnaire et suppose chaque élément comme équiprobable. Des systèmes plus efficaces sans perte (*cf.* section 3.2.3) permettent de prendre en compte cette probabilité d'apparition en définissant un code de taille variable tel que les éléments ayant une fréquence d'apparition plus importante correspondent à un mot de longueur plus faible. On parle alors de codage entropique décrit en détail dans la section 3.2.3. La compression peut également être obtenue par transformation en choisissant un dictionnaire (ou base de représentation) permettant de représenter un signal de manière plus parcimonieuse (*cf.* section 2.1).

- **la protection des données** peut être utilisée dans certains cas lorsque le support de stockage ou le canal de transmission est sujet à des perturbations aléatoires provoquant des erreurs de décodage. Les codes correcteurs d'erreurs permettent ainsi d'éviter les erreurs de substitution (confusion entre deux éléments d'un dictionnaire) en rendant les données plus robustes par l'ajout de redondance dans le codage. Ainsi, chaque élément d'un dictionnaire possède plusieurs codes ou représentants garantissant un décodage sans erreur lorsque les altérations ne sont pas trop importantes en fonction de la capacité du code. Les disques compacts audio utilisent par exemple le codage de Reed-Solomon [RS60] qui rend possible dans le meilleur des cas une reconstruction des données en cas d'effacement (lecture impossible de certaines zones en présence d'une rayure par exemple). L'utilisation des codes correcteurs a pour effet d'augmenter la taille des données et nécessite le remplacement du système de codage utilisé pour représenter certaines informations. N'entrant pas dans le cadre de notre problématique, nous avons choisi de ne pas approfondir la théorie des codes correcteurs d'erreur dans cette thèse.

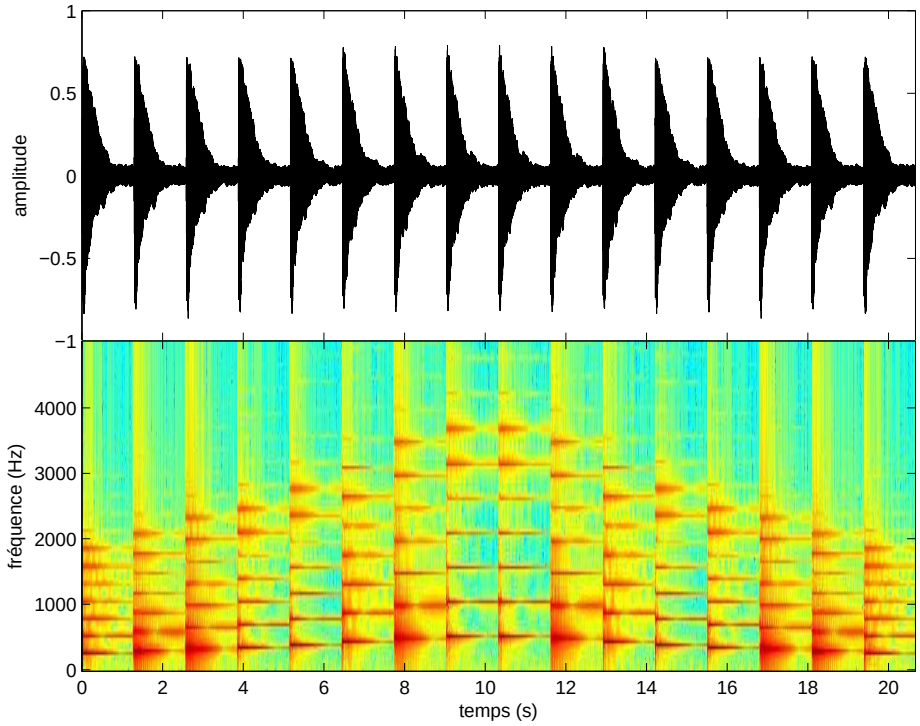


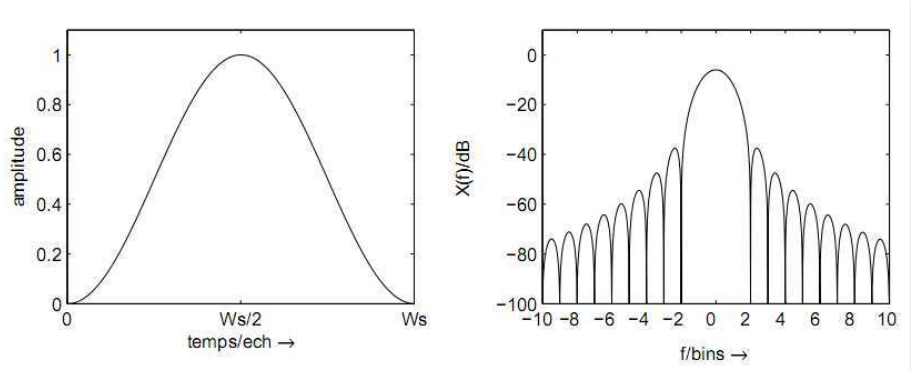
FIGURE 1.3 – Signal temporel et spectrogramme correspondant au son émis par un piano jouant la gamme de Do majeur montante puis descendante.

1.2 Décomposition temps-fréquence

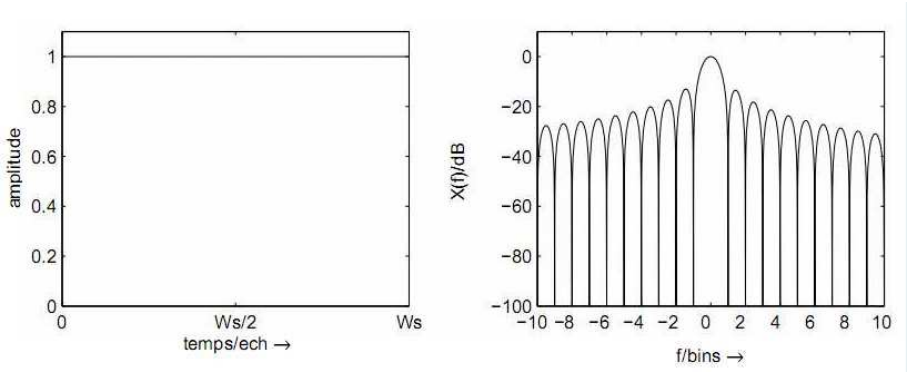
Jusqu'à présent, nous avons vu des signaux pouvant s'exprimer comme une fonction du temps. Pourtant cette représentation n'est pas très adaptée pour analyser les fréquences qui caractérisent un signal et auxquelles une oreille humaine est sensible. Ainsi, la structure particulière des signaux audio peut être étudiée en utilisant une décomposition en atomes temps-fréquence. Comme le présente la figure 1.3, une décomposition temps-fréquence (TF) comme le spectrogramme calculé grâce à la transformée de Fourier (*cf.* section 1.2.1) permet par exemple d'analyser le contenu fréquentiel d'un signal audio en fonction du temps.

1.2.1 La transformée de Fourier fenêtrée

Parmi les représentations temps-fréquence les plus utilisées en traitement du signal audio, nous pouvons citer le spectrogramme (*cf.* figure 1.3). Celui-ci se calcule en utilisant la transformée de Fourier à court terme, *Short-Time Fourier Transform* (STFT) combinée avec une fenêtre d'analyse. La fenêtre d'analyse notée $w(t)$ est une fonction de pondération qui est déplacée sur le signal analysé qui permet d'isoler un segment du signal sur une courte durée. Une fenêtre d'analyse permet ainsi de prendre en compte la proximité des échantillons au voisinage d'un instant t_0 situé au centre de cette fenêtre et pour laquelle la pondération est maximale. Une trame de signal est alors composée de tous les échantillons où la fonction $w(t)$ translatée à un instant t_0 est non nulle. Ainsi, à chaque coupe verticale du spectrogramme représenté sur la figure 1.3 correspond le spectre d'amplitude utilisant une échelle logarithmique et calculé pour une certaine du-



a) fenêtre de Hann



b) fenêtre rectangulaire

FIGURE 1.4 – Exemples de fenêtre d'analyse spectrale et spectre d'amplitude (en dB) correspondant [Mar08].

rée. Cette technique permet une analyse fréquentielle locale d'un signal sur une portion de temps limitée autour d'un instant donné.

Bien qu'il existe plusieurs types de fenêtre d'analyse, aucune d'entre elles n'est idéale car toutes obligent à faire un compromis entre la résolution temporelle et la résolution fréquentielle. Ce phénomène est mieux connu sous le nom du principe d'incertitude (*cf.* chapitre 4). Nous utiliserons principalement par la suite, sauf précision contraire, la fenêtre de Hann qui possède de bonnes propriétés pour le traitement de la musique et qui est définie par :

$$w_{\text{Hann},N}[n] = \begin{cases} \frac{1}{2}\left(1 - \cos(2\pi\frac{n}{N})\right) & \text{si } n \in [0;N] \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

Cette fenêtre en plus d'être bien adaptée à l'analyse de signaux musicaux, possède la particularité d'être plusieurs fois dérivable. Nous détaillons par la suite les raisons pour lesquelles cette propriété est utile dans le cadre de l'analyse spectrale (*cf.* section 2.1).

Le signal discret fenêtré de taille N à l'instant $t_0 = \frac{n_0}{F_s}$ où F_s est la fréquence d'échantillonnage, est obtenu par un produit terme à terme avec la fenêtre d'analyse translatée et centrée sur l'échantillon n_0 :

$$s_{t_0}^w[n] = s[n]w[n_0 - n + \frac{N}{2}], \tag{1.8}$$

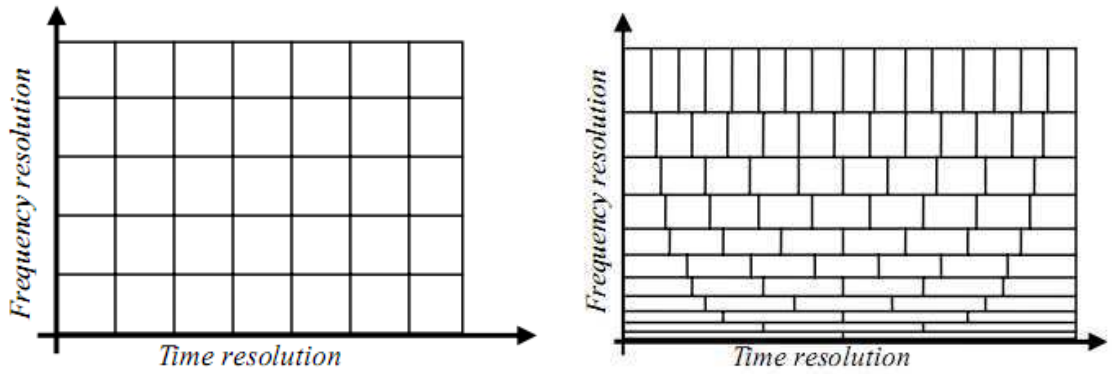
et sa STFT discrète est donnée par :

$$\text{STFT}^w[n_0, k] = \sum_{n=0}^{N-1} s_{t_0}^w[n]\, \mathbf{e}^{-\mathbf{j}2\pi kn/N}, \tag{1.9}$$

où k correspond à l'indice de fréquence discrète. La STFT peut s'écrire comme suit en continu :

$$\text{STFT}^w(t, \omega) = \int_{-\infty}^{\infty} s(\tau) w(t-\tau)\, \mathbf{e}^{-\mathbf{j}\omega\tau}\, d\tau, \tag{1.10}$$

pour $w(t)$ de durée T et définie sur $[-T/2; T/2]$. Il existe bien sûr d'autres outils permettant de faire de l'analyse temps-fréquence que nous ne détaillerons pas ici. Notamment la transformée en ondelettes [GM84] ou encore la transformée Q constant, *Constant Q Transform* (CQT). Le principal intérêt de ces techniques repose sur l'obtention d'une résolution non linéaire (*cf.* figure 1.5) parfois utile en fonction pour certaines applications.



a) résolution de la transformée de Fourier

b) résolution de la transformée en ondelettes

FIGURE 1.5 – Comparaison de la résolution temps-fréquence entre la transformée de Fourier 1.5a) et la transformée en ondelettes 1.5b) (source [PHC06]).

1.2.2 Modèles utilisés pour les signaux audio

La forme d'onde des signaux audio possède une structure particulière permettant de décrire leur signal avec un modèle adapté. De tels modèles sont utilisés couramment pour l'analyse mais aussi pour la synthèse des signaux audio.

a) Modèle sinusoïdes-transitoires-bruit

Le modèle sinusoïdes / transitoires / bruit (STB) [MQ86, SS87] est un modèle additif utilisé en synthèse et en analyse qui décompose un signal en 3 parties qui sont respectivement :

- **la partie déterministe** définie par une somme de fonctions périodiques, la plupart du temps de la forme :

$$s(t) = \sum_{i=1}^{I(t)} a_i(t) \cos\left(\Phi_i(t)\right) \text{ avec } \Phi_i(t) = \Phi_i(0) + \int_{t_0,i}^{t} \omega_i(\tau)\, d\tau. \tag{1.11}$$

  où $a_i(t)$, $\Phi_i(t)$ et $\omega_i(t)$ correspondent à l'amplitude, la phase et la fréquence de la composante d'indice i. Dans ce cas $I(t)$ correspond au nombre de composantes

en fonction du temps. Ce modèle est bien adapté pour représenter la partie tonale des sons et constitue la partie essentielle des instruments de musique jouant des notes (*e.g.* guitare, piano, voix, etc.). C'est un modèle suffisamment général pour représenter (moins efficacement) le reste du signal. C'est pour cela que nous avons choisi de concentrer nos efforts sur ce modèle dans la suite du document.

- **Les transitoires** correspondent à des changements brefs apparaissant sur un signal et souvent associés à des événements ponctuels tels que les débuts de note ou les sons produits par un instrument percussif. Les transitoires ne sont pas présents dans tous les modèles de signaux mais ils se caractérisent par des augmentations soudaines de l'énergie. Leur détection peut être utilisée comme pré-traitement en analyse par exemple pour ignorer les périodes de silence d'un signal. Ils peuvent cependant être utilisés dans des applications musicales pour faire de la détection de rythme ou pour détecter les débuts et fins de note dans le cadre d'une transcription automatique de la musique (*cf.* section 2.2). Certains codeurs [SOdBB03] distinguent deux types de transitoires : les transitoires de Meixner [Bri95] qui se traduisent par une enveloppe parabolique avec une hausse puis une baisse soudaine du niveau d'énergie. Les transitoires de niveaux se traduisent par une hausse sur une longue durée du niveau d'énergie. Les transitoires de Meixner, en raison de leur caractère bref et ponctuel compliquent en général l'estimation de la partie tonale des sons. Une réduction de la taille de la fenêtre d'analyse permet la plupart du temps d'améliorer la qualité des estimations.

- **Le bruit** est souvent modélisé par des processus stochastiques (ou aléatoires) [Han03]. Ne possédant pas de structure déterministe, le signal correspondant au bruit s'obtient souvent par soustraction de la partie déterministe depuis un signal de mélange. Le bruit peut également se modéliser en filtrant un bruit blanc correspondant à la réalisation d'un processus gaussien. On parle alors d'un modèle source-filtre (voir paragraphe suivant). Le filtre utilisé permet de définir l'enveloppe spectrale du bruit que l'on cherche à modéliser. L'estimation de l'enveloppe d'un bruit existant peut se faire en utilisant des techniques telles que [YR06, MHM06]. Sauf pour des sons spécifiques, le niveau de bruit est en général plus faible que celui des composantes tonales pour la musique non percussive. Cependant, le bruit est toujours omniprésent quelque soit la nature du signal analysé et il se combine en temps et en fréquence avec les composantes tonales déterministes. Cela a pour effet de dégrader la précision des techniques d'analyse utilisées pour décrire la partie tonale d'un signal (*cf.* section 2.1.2).

b) Modèle source-filtre

Le modèle source-filtre est un modèle utilisé principalement en synthèse sonore et qui considère tout signal comme étant l'interaction entre une source qui émet le son et un filtre jouant le rôle de résonateur. Bien que ne correspondant pas toujours à la réalité physique du son considéré, ce modèle est simple à mettre en oeuvre. Ainsi, en utilisant ce modèle, la partie tonale d'un son peut par exemple être décrite en utilisant une source harmonique filtrée (produit terme à terme dans le domaine spectral) avec une fonction décrivant son enveloppe spectrale (*e.g.* figure 1.7). Dans le cas d'un son bruité, on pourra par exemple utiliser un bruit blanc, un son composé de plusieurs sources ou tout signal existant comme source émettrice. Ce modèle est utilisé par exemple dans le chapitre 5 pour traiter le problème de séparation de sources par filtrage.

c) Modèle de factorisation en matrices non négatives

La factorisation en matrices non négatives, *Non-negative Matrix Factorization* (NMF) [LS99, LS00] est une méthode de réduction de rang applicable sur les données non négatives (matrices à valeurs positives ou nulles). Cette technique très répandue en traitement du signal audio permet de décomposer un spectrogramme en un produit de deux matrices non négatives qui correspondent à un dictionnaire d'atomes et à des activations temporelles.

La NMF permet notamment d'apprendre directement à partir des données ce dictionnaire d'atomes (*e.g.* dictionnaire de sources quasi-harmoniques) associés aux motifs répétitifs. Cette technique est donc comparable à d'autres méthodes de factorisation/apprentissage comme l'analyse en composantes principales [Jol02] ou l'analyse en composantes indépendantes, *Independant Component Analysis* (ICA) [Com94] dont elle se distingue de part la contrainte de non négativité.

Les auteurs de cette méthode [LS99] montrent que la non négativité permet d'isoler des éléments pertinents des signaux en relation avec la perception. Cependant, cette technique permet difficilement de modéliser les éléments qui varient au cours du temps pourtant omniprésents dans les signaux de musique (*e.g.* variations de hauteur ou de timbre des instruments).

1.3 Paramètres musicaux

1.3.1 Amplitude et volume

L'amplitude correspond à la pression acoustique mesurée par un récepteur sonore et est évaluée en pascals (Pa). Pour faire le lien avec la perception humaine, on calcule le volume en décibels (dB) sur une échelle de 0dB à 120dB (seuil de douleur) qui correspondent respectivement à une amplitude de 0 et de 1 sur une échelle linéaire normalisée. Le volume est donné par l'expression suivante :

$$\mathrm{Vol}(a) = 10\log_{10}\left(\frac{a}{A_{0dB}}\right)^2 = 20\log_{10}\left(\frac{a}{A_{0dB}}\right). \tag{1.12}$$

Dans la convention standard du dB SPL[1], $A_{0dB} = 10^{-6}$ correspond à une pression acoustique de $2 \cdot 10^{-5}$ Pa. Pour évaluer le volume sur une trame de taille N d'un signal $s[n]$, on peut utiliser l'amplitude efficace (RMS[2]) donnée par :

$$\mathrm{Amp} \approx \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}|s[n]|^2}. \tag{1.13}$$

Dans [Mar00], Marchand montre que cette formule s'applique aussi à la partie tonale des sons en ne considérant comme signal que l'amplitude de chaque partiel (*cf.* section 1.3.4).

1.3.2 Enveloppe temporelle et transitoires

En général, les transitoires [BDA+05] sont associés à des événements brefs se traduisant par des changements significatifs sur la nature du signal étudié. En musique, ils surviennent essentiellement lors des débuts et fins de note (*onsets* et *offsets*) et possèdent une enveloppe particulière [FR98] pouvant être schématisée par la figure 1.6.

[1] *Sound Pressure Level*
[2] *Root Mean Square*

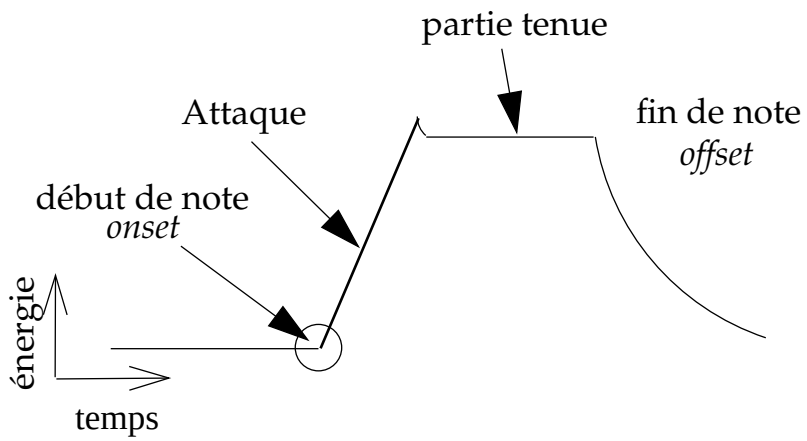


FIGURE 1.6 – Schéma décrivant l'évolution temporelle de l'enveloppe d'un signal musical.

Le début de note se situe au commencement de l'attaque et se traduit par une augmentation du niveau d'énergie du signal. La note peut ensuite être maintenue soit par résonance, soit volontairement par l'action du musicien. Il s'en suit une période d'affaiblissement caractéristique de la fin de note.

Cette enveloppe caractéristique peut être décrite par le modèle *Attack Decay Sustain Release* (ADSR) [TC83] dont les implémentations sont encore utilisées à ce jour dans le domaine de la synthèse musicale.

1.3.3 Fréquence fondamentale et hauteur

La perception du son émis par une source harmonique ou quasi-harmonique (*e.g.* instrument tonal) correspond généralement à la fréquence fondamentale appelée aussi fréquence F_0 de cette source. La fréquence F_0 est associée au premier partiel du modèle théorique d'une source harmonique (que nous détaillerons plus loin) identifiable par un pic dans le spectre d'amplitude du signal considéré (*cf.* figure 1.7). Les autres partiels sont des multiples entiers ou presque (source quasi-harmonique) de cette fréquence F_0.

En général, une fréquence est mesurée en hertz (Hz) ou en radians par seconde (vitesse angulaire) bien qu'il existe d'autres échelles de mesure dites perceptives. Notamment les échelles *Equivalent Rectangular Bandwidth* (ERB) [MG83], Mel [SVN37] et Bark[3] [Zwi61] ont été définies par rapport à un modèle psychoacoustique relatif à la perception humaine.

La hauteur musicale ou *pitch* est une grandeur qui permet d'associer une note de musique à une fréquence. La hauteur est calculée pour une fréquence de référence par la formule suivante :

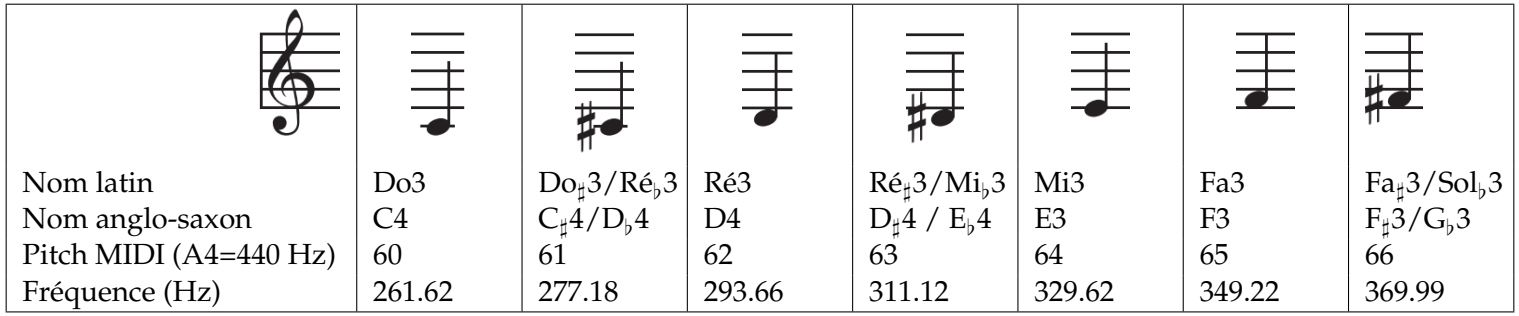
$$\text{Pitch}(f) = P_{\text{ref}} + O \log_2 \left(\frac{f}{F_{\text{ref}}} \right), \tag{1.14}$$

avec P_{ref} et F_{ref} respectivement le *pitch* de référence et sa fréquence correspondante. La constante O permet de définir le nombre de subdivisions d'une octave[4]. Dans la musique occidentale basée sur le tempérament égal, on définit 12 notes de musique distinctes par octave (*cf.* table 1.1), soit $O = 12$. Dans la norme MIDI utilisée dans le monde entier par tous les musiciens, on fixe $P_{\text{ref}} = 69$, $F_{\text{ref}} = 440$Hz.

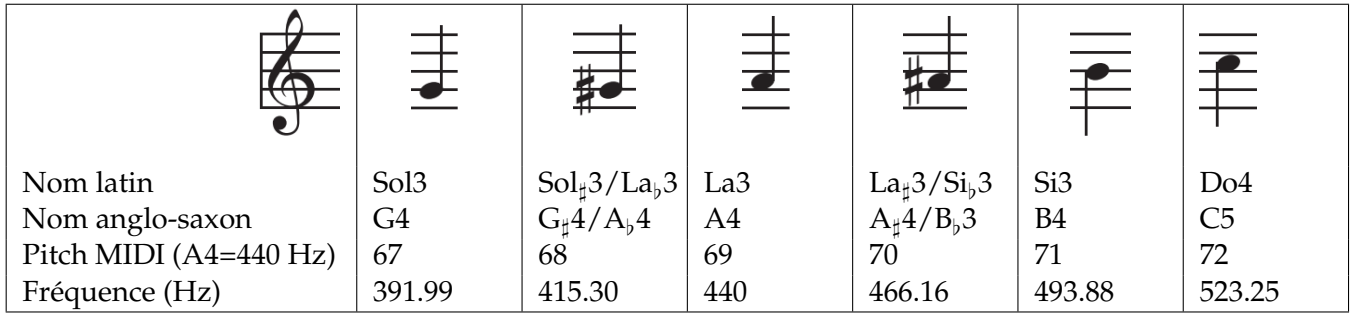
[3] L'échelle Bark subdivise en 24 bandes de taille minimale $\Delta f = 100$Hz les fréquences comprise entre 0 Hz et 15500 Hz.

[4] L'octave est l'intervalle séparant 2 sons de fréquence F_0 et $2F_0$ (*e.g.* La3 : 440Hz et La4 : 880Hz).

Le problème d'estimation de la hauteur pour la transcription automatique de la musique (*cf.* section 2.2) est associé à l'estimation de la fréquence fondamentale correspondante. En fonction des applications, on peut chercher à estimer uniquement leu nom de la note jouée en se ramenant sur une échelle de 12 possibilités. Cela revient à identifier la fréquence F_0 ou un de ses multiples de la forme $2^n F_0$ pour $n \in \mathbb{Z}$. Si on souhaite trouver en même temps le numéro de l'octave correspondante, c'est que l'on souhaite identifier presque exactement la fréquence fondamentale. Dans les applications musicales, la connaissance *a priori* du diapason utilisé (*e.g.* A4=440 Hz) peut permettre d'isoler les régions du spectre où l'on recherche une fréquence fondamentale. Ainsi, les fréquences détectées voisines sont ramenées sur l'échelle musicale correspondante. Un exemple de calcul de cette échelle est proposé dans la table 1.1.



Nom latin	Do3	Do♯3/Ré♭3	Ré3	Ré♯3/Mi♭3	Mi3	Fa3	Fa♯3/Sol♭3
Nom anglo-saxon	C4	C♯4/D♭4	D4	D♯4 / E♭4	E3	F3	F♯3/G♭3
Pitch MIDI (A4=440 Hz)	60	61	62	63	64	65	66
Fréquence (Hz)	261.62	277.18	293.66	311.12	329.62	349.22	369.99



Nom latin	Sol3	Sol♯3/La♭3	La3	La♯3/Si♭3	Si3	Do4
Nom anglo-saxon	G4	G♯4/A♭4	A4	A♯4/B♭3	B4	C5
Pitch MIDI (A4=440 Hz)	67	68	69	70	71	72
Fréquence (Hz)	391.99	415.30	440	466.16	493.88	523.25

TABLE 1.1 – Différentes représentations de la gamme chromatique du Do3 au Do4 sur une échelle basée sur le tempérament égal. Le système du tempérament égal subdivise une octave en 12 demi-tons égaux. Ce système est celui utilisé dans la musique occidentale par quasiment tous les instruments chromatiques actuels (guitare, piano, etc.). Le calcul des fréquences s'effectue par rapport au La3 (440 Hz), ainsi la fréquence du Do3 située 9 demi-tons en dessous est obtenue par le calcul $440 \times 2^{-9/12} \approx 261.62$Hz et celle du Do4 situé 3 demi-tons au-dessus par $440 \times 2^{+3/12} \approx 523.25$Hz.

1.3.4 Enveloppe spectrale et timbre

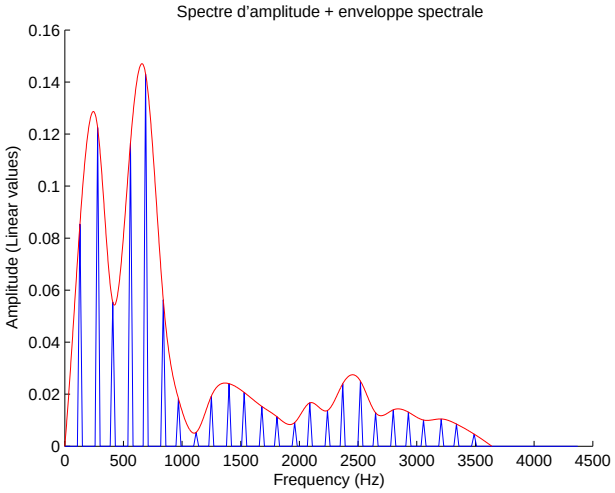


FIGURE 1.7 – Spectre d'amplitude d'un son de saxophone jouant la note Do$\sharp$2 ($F_0 \approx$ 138.6Hz). Nous voyons clairement les différents partiels (pics) ainsi que l'enveloppe spectrale correspondante (en rouge) construite par interpolation spline.

L'enveloppe spectrale d'une source sonore (*cf.* figure 1.7) est une fonction qui interpole la valeur de l'amplitude de chaque partiel en fonction de la fréquence. Ainsi, les partiels sont les composantes spectrales qui composent un son harmonique. Chaque partiel peut être décrit par un triplet de trois paramètres (*cf.* partie déterministe du modèle STB décrit dans la section 1.2.2). Ces paramètres peuvent être utilisés en analyse et en synthèse du signal sonore. L'ensemble des partiels permet de déduire l'enveloppe spectrale qui est une information très importante. En effet, elle est utile pour définir les caractéristiques suivantes :

- **le timbre** (ou couleur pour les musiciens) est une caractéristique du son directement lié à la perception et à l'identification d'une source sonore [HBKD06, SWT04]. Cette information peut être utile pour des taches telles que la séparation de sources ou la transcription automatique polyphonique. En effet, certaines propriétés de l'enveloppe son lissage ou sa douceur (*spectral smoothness*), sa centroïde ou la décroissance en amplitude des partiels ainsi que leur nombre sont des informations importantes pour détecter ou isoler des sources sonores présentes dans un mélange.
- **Les formants** sont des caractéristiques couramment utilisés dans les problématiques de reconnaissance vocale [Wol09]. Ils correspondent aux maxima locaux de l'enveloppe spectrale et permettent notamment l'identification de la prononciation des voyelles. Les consonnes étant plus fréquemment associées à un modèle de bruit ou de transitoire.

L'estimation de l'enveloppe spectrale d'une source en milieu bruité est un problème difficile. En effet, cette tâche est souvent associée à d'autres applications plus complexes telles que le rehaussement de signal, la transcription musicale automatique ou la séparation de sources. La connaissance de l'enveloppe d'une source permet également lorsque c'est nécessaire d'effectuer une reconstruction de l'amplitude des partiels manquants, mal estimés ou interférant avec d'autres signaux. La plupart du temps, cette enveloppe

n'est jamais exactement connue et ne peut être que approximée par interpolation en utilisant les partiels connus. Cette reconstruction pourra alors s'effectuer en appliquant une technique similaire à celle proposée pour la reconstruction des signaux échantillonnés (*cf.* section 1.1.2).

1.3.5 Rythme et pulsation

Le rythme et la pulsation sont des informations musicales qui permettent de structurer dans le temps les événements musicaux. Ainsi, pour la transcription automatique des signaux de musique, ces informations peuvent constituer un *a priori* utile permettant par exemple d'augmenter la précision de détection des débuts et des fins de note. On définit la pulsation comme une accentuation intervenant à chaque temps. Ainsi, la régularité de la pulsation permet de définir le tempo calculé en pulsations par secondes. Le rythme et la métrique sont des notions musicales plus abstraites qui consistent à regrouper les temps pour constituer des mesures. Le rythme est alors donné en nombre de temps par mesure. Sa valeur dépend du contenu musical (organisation des phrases), de la volonté du compositeur ainsi que des conventions d'écriture musicale. Il est donc tout à fait possible d'écrire des thèmes musicaux identiques avec des métriques différentes mais à pulsation équivalente. Ces informations peuvent se déduire à partir d'une représentation symbolique de la musique (*e.g.* partition MIDI) en utilisant par exemple la technique proposée par Takeda *et al.*[TNS04].

1.4 Conclusions du chapitre

Ce premier chapitre introductif nous a permis de présenter des éléments de traitement du signal audio ainsi que des notions de musique auxquels nous faisons régulièrement référence dans les chapitres qui suivent. Nous avons également jugé utile de présenter de manière brève certains outils et modèles qui ne feront pas l'objet d'approfondissements par la suite (*e.g.* ondelettes, codes correcteurs d'erreurs, etc.). Ces éléments sont présents, soit parce qu'ils sont implicitement liés à notre domaine d'étude et donc nécessaires à la compréhension de la thèse. Soit pour évoquer des pistes de travail explorées durant nos travaux de recherche que nous souhaitions rapidement écarter du contexte de cette étude. Les chapitres suivant décrivent de manière plus précise les problèmes que nous traitons dans ce document.

CHAPITRE 2

PROBLÈMES D'ANALYSE EN TRAITEMENT DU SIGNAL AUDIO

L'analyse du son est un domaine très vaste qui regroupe de nombreuses techniques dont la méthodologie dépend des applications et du champ disciplinaire. Dans ce travail de thèse, nous avons choisi de concentrer nos efforts sur plusieurs problèmes importants en traitement du signal audio reposant à la fois sur une représentation physique mais aussi symbolique des sons manipulés. Parmi les problèmes traités, nous avons choisi l'analyse spectrale des signaux audio basée sur l'estimation des paramètres sinusoïdaux, la transcription automatique de la musique polyphonique et la séparations de sources audio.

L'analyse spectrale est un problème largement étudié et dont les limitations théoriques des techniques existantes sont aujourd'hui bien établies. La transcription automatique de la musique et la séparation de sources sont deux problèmes difficiles pour lesquels les approches classiques obtiennent des résultats souvent insuffisants. Dans ce chapitre, nous introduisons chaque problème en décrivant ses enjeux. Par la suite nous présentons des approches existantes permettant de traiter ces problèmes afin de constituer un état de l'art. Enfin, nous proposons une méthodologie permettant d'évaluer les performances des méthodes proposées dans la seconde partie de ce document.

2.1 L'estimation des paramètres sinusoïdaux des signaux musicaux

Le modèle sinusoïdal introduit par McAulay et Quatieri [MQ86] pour les signaux de parole et par Smith et Serra [SS87] pour la musique possède de bonne propriétés de parcimonie pour représenter la partie déterministe (tonale) des signaux audio. C'est pour cela que ce modèle paramétrique est utilisé dans certaines techniques de codage audio avancées [SOdBB03, PM00], souvent dans le cadre du modèle STB (*cf.* section 1.2.2). De plus, le modèle sinusoïdal repose sur l'analyse de Fourier qui propose un cadre théorique solide permettant à la fois d'analyser, de transformer et de synthétiser de nouveaux signaux audio. D'autre part, ce modèle peut servir à améliorer la qualité de certaines transformations telles que la transposition [Fed98] (*pitch shifting*) ou l'étirement temporel [RM07]

(*time stretching*). Bien que ce modèle ne permette pas une représentation parcimonieuse des signaux bruités ou percussifs, nous avons tout de même choisi de concentrer nos efforts sur son étude. Les transitoires pourront être analysés par ce modèle en utilisant des fenêtres d'analyse adaptées de plus courte durée. Un signal bruité pourra être approché avec un nombre plus important de composantes.

Ainsi, la qualité des signaux synthétisés et la pertinence des informations extraites utilisant ce modèle dépend donc fortement de la précision des paramètres dont on dispose. Par exemple, l'oreille humaine est capable d'entendre des incohérences de phase (discontinuité du signal) engendrées par une erreur très faible sur un des paramètres du modèle utilisé pour synthétiser un son. Le modèle sinusoïdal peut être décrit selon différentes hypothèses qui dépendent de la nature des signaux considérés.

2.1.1 Le modèle sinusoïdal à court terme

Ce modèle permet de décrire un signal périodique déterministe et convient pour représenter les partiels d'un signal tonal quelconque (*cf.* section 1.3.4). La version stationnaire de ce modèle s'inspire directement de l'analyse de Fourier et ne peut être utilisé que pour décrire des signaux audio de courte durée pour lesquels les paramètres ne varient pas dans le temps. Un signal à valeur dans $\mathbb{R}$ composé d'une seule sinusoïde s'écrit :

$$s(t) = a\cos(\omega t + \phi) = a\sin(\omega t + \phi + \pi/2), \tag{2.1}$$

et possède trois paramètres : l'amplitude a, la fréquence angulaire ω mesurée en rad.s^{-1} reliée à la fréquence f mesurée en Hz par la relation : $\omega = 2\pi f$.
En général les sons acoustiques que l'on trouve dans la nature sont composés de plusieurs sinusoïdes ayant des paramètres différents qu'il suffit d'additionner pour obtenir le signal de mélange correspondant.

Ainsi, ce modèle permet une manipulation paramétrique des signaux audio où chaque signal est représenté par un ensemble de triplets dont chacun est associé à une composante sinusoïdale. Lorsque l'on ne dispose que d'un signal audio seul, ces paramètres doivent alors être estimés à partir d'une ou de plusieurs observations de ce signal (signaux de mélange bruités). Bien que la littérature propose de nombreuses approches, nous avons choisi de n'en présenter ici que quelques une basées principalement sur l'utilisation de la transformée de Fourier. Nous ne développons donc pas dans ce chapitre les approches haute résolution [Bad06, BDR06] souvent plus complexes à mettre en oeuvre.

a) Utilisation de la transformée de Fourier discrète seule

Le signal $s[n]$ analysé est fenêtré comme décrit dans la section 1.2.1 en appliquant un produit terme à terme avec la fenêtre d'analyse pour obtenir $s^w_{t_0}[n]$. Le spectre discret (à l'instant t_0) noté $S_w[k]$ est obtenu après l'application de la transformée de Fourier discrète. La fréquence estimée est donnée par :

$$\hat{\omega} = 2\pi\hat{f} = 2\pi k_m \frac{F_s}{N}, \tag{2.2}$$

pour $k_m \in [0, \frac{N}{2}]$ l'indice du spectre discret correspondant au maximum du spectre d'amplitude $|S_w[k]|$. Ici F_s et N correspondent respectivement à la fréquence d'échantillonnage et à la taille de la transformée de Fourier discrète. L'amplitude et la phase sont données

en fonction de l'indice k_m par [KM02] :

$$\hat{a} = 2\frac{|S_w[k_m]|}{\sum_{n=0}^{N-1} w[n]}\,, \tag{2.3}$$

$$\hat{\phi} = \angle\left(\frac{S_w[k_m]}{\sum_{n=0}^{N-1} w[n]}\right). \tag{2.4}$$

On remarque que dans le cas d'un signal réel il faut multiplier par 2 pour obtenir l'amplitude, en effet la fonction cosinus est composée de 2 composantes sinusoïdales complexes associées à 2 pics détectés dans le spectre d'amplitude. Le spectre doit ensuite être normalisé par la transformée de Fourier de la fenêtre d'analyse pour la fréquence nulle $W[0] = \sum_{n=0}^{N-1} w[n]$ pour retrouver l'amplitude initiale.

Cet estimateur peut être amélioré en utilisant par exemple l'interpolation de Fourier par la technique de *zero padding* ou "bourrage de zéros" qui consiste à augmenter la précision de l'estimation en ajoutant des échantillons de valeur nulle à la suite de la trame d'analyse. Cela a pour effet d'augmenter par interpolation la taille du spectre discret obtenu.

b) Interpolation du spectre d'amplitude

Comme la précision de l'analyse utilisant transformée de Fourier du signal est le plus souvent insuffisante pour une estimation correcte des paramètres sinusoïdaux, le spectre d'amplitude peut être interpolé au voisinage d'un maximum local dans le but d'estimer plus précisément la fréquence pour laquelle ce maximum est atteint.

Par exemple la méthode d'interpolation parabolique décrite dans [SS87] ou la technique d'interpolation quadratique de la transformée de Fourier rapide, *Quadratically Interpolated Fast Fourier Transform* (QIFFT) [AS05] consistent à calculer l'équation d'une parabole passant par les points du spectre d'amplitude au voisinage de l'indice k_m correspondant au maximum local détecté. La fréquence est alors estimée plus précisément à partir du maximum de la fonction interpolant le spectre. Une autre approche désignée par l'algorithme du triangle décrit dans [KZ01] qui consiste à faire correspondre dans le spectre d'amplitude au voisinage de k_m, deux segments de droite associés à deux cotés d'un triangle (*cf.* figure 2.1). Le signal analysé aura été préalablement fenêtré par une fenêtre triangulaire adéquate. La fréquence estimée est alors déduite à partir du point d'intersection entre les deux segments.

A partir de la fréquence estimée par une technique d'interpolation du spectre $\hat{\omega}$, l'amplitude et la phase peuvent alors être déduites par :

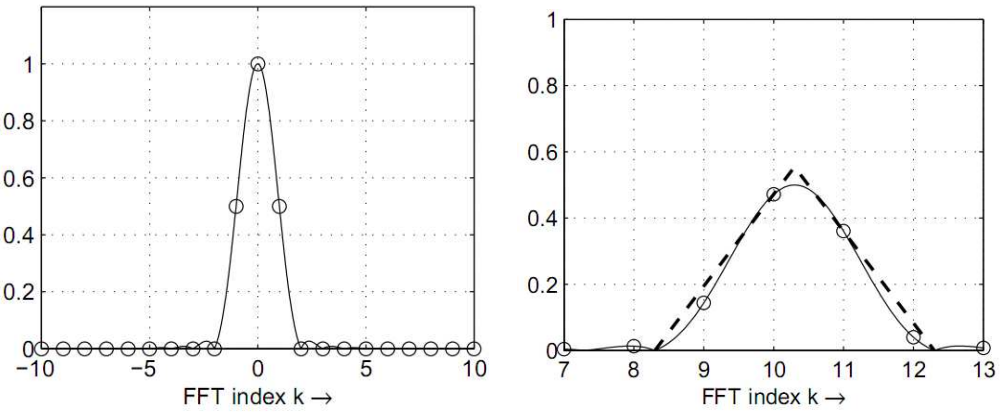
$$\hat{a} = 2\left|\frac{S_w[k_m]|}{W(\Delta_\omega)}\right| \tag{2.5}$$

$$\hat{\phi} = \angle\left(\frac{S_w[k_m]}{W(\Delta_\omega)}\right), \tag{2.6}$$

où $\Delta_\omega = \hat{\omega} - 2\pi k_m \frac{F_s}{N}$ correspond à la différence entre la fréquence estimée $\hat{\omega}$ et la fréquence discrète associée à l'indice k_m.

c) Méthodes utilisant le calcul de la dérivée

La méthode dite "des dérivées" et la méthode de réallocation sont des techniques d'analyse qui font intervenir respectivement l'expression analytique de la dérivée du signal observé ou de celle de la fenêtre d'analyse utilisée dans la formulation de l'estimateur.



a) spectre d'une fenêtre triangle

b) spectre d'un signal fenêtré par une fonction triangle

FIGURE 2.1 – Application de l'algorithme du triangle [KZ01] pour l'estimation des paramètres sinusoïdaux.

c.1) La réallocation spectrale

Cette méthode introduite par Kodera, Gendrin et de Villedary [KdVG76, KGdV78] utilise l'expression analytique du signal temporel reconstruit à partir de son spectre afin de calculer un spectre relocalisé (ou réalloué) sur le maximum local que l'on cherche estimer.

Ainsi, si on considère la transformée à de Fourier à court terme d'un signal $s(t)$ fenêtré par une fonction $w(t)$, nous avons :

$$\begin{aligned} S(t,\omega) &= \int s(\tau) w(\tau - t)\, \mathbf{e}^{-\mathbf{j}\omega\tau}\, d\tau \\ &= \mathbf{e}^{-\mathbf{j}\omega t} \int s(\tau) w(\tau - t)\, \mathbf{e}^{-\mathbf{j}\omega(\tau - t)}\, d\tau. \end{aligned} \tag{2.7}$$

Le spectre donné par (2.7) étant complexe, nous pouvons le reformuler par :

$$S(t,\omega) = A(t,\omega)\, \mathbf{e}^{\mathbf{j}(\phi(t,\omega) - \omega t)}, \tag{2.8}$$

où $A(t,\omega)$ et $\phi(t,\omega)$ correspondent respectivement au spectre d'amplitude et au spectre de phase. Le signal temporel peut être reconstruit en appliquant une transformée de Fourier inverse fenêtrée sur $S(t,\omega)$:

$$\begin{aligned} s(t) &= \int \left(\frac{1}{2\pi} \int S(\tau,\omega)\, \mathbf{e}^{\mathbf{j}\omega t} d\omega \right) w(t - \tau)\, d\tau \\ &= \frac{1}{2\pi} \iint A(\tau,\omega) w(t - \tau)\, \mathbf{e}^{\mathbf{j}(\phi(\tau,\omega) - \omega\tau + \omega t)}\, d\omega\, d\tau. \end{aligned}$$

Les conditions de stationnarité du terme $[\phi(\tau,\omega) - \omega\tau + \omega t]$ sont suffisantes pour déduire une expression permettant d'estimer ω et t :

$$\frac{\partial}{\partial t}(\phi(\tau,\omega) - \omega\tau + \omega t) = 0 \Rightarrow \tilde{\omega}(\tau,\omega) = \omega + \frac{\partial \phi(\tau,\omega)}{\partial t}. \tag{2.9}$$

La réallocation a été généralisée aux spectrogrammes par Auger et Flandrin dans [AF95] et permet de trouver une expression plus efficace :

$$\hat{\omega}(\tau,\omega) = \omega + \frac{\partial\phi(\tau,\omega)}{\partial t} = \omega - \mathbf{Im}\left(\frac{S_{w'}(\tau,\omega)}{S_w(\tau,\omega)}\right), \tag{2.10}$$

$$\hat{t}(\tau,\omega) = \tau - \frac{\partial\phi(\tau,\omega)}{\partial\omega} = t + \underbrace{\mathbf{Re}\left(\frac{S_{tw}(\tau,\omega)}{S_w(\tau,\omega)}\right)}_{G_d(\omega)}, \tag{2.11}$$

avec $S_{w'}$ la transformée de Fourier à court terme utilisant la dérivée de la fenêtre w et S_{tw} utilisant la fenêtre w pondérée par les instants t des échantillons.

c.2) La méthode des dérivées

La méthode de dérivée a été proposée initialement par Marchand [Mar98]. Cette approche considère la dérivée du signal temporel donnée par l'équation (2.1) et qui peut s'écrire :

$$s'(t) = -a\omega\sin(\omega t + \phi) = a\omega\cos\left(\omega t + \phi - \frac{\pi}{2}\right). \tag{2.12}$$

et par induction, on déduit facilement l'expression de la dérivée d'ordre m du signal $s(t)$ donnée par :

$$s^{(m)}(t) = a\omega^m\cos\left(\omega t + \phi - m\frac{\pi}{2}\right). \tag{2.13}$$

Dans le cas d'un signal composé de plusieurs sinusoïdes, l'utilisation du spectre permet d'isoler le pic pour lequel on cherche à estimer la fréquence. Ainsi, une estimation plus précise de la fréquence correspondante au maximum local d'indice k_m est donnée par :

$$\hat{\omega} = \frac{S^{(m+1)}[k_m]}{S^{(m)}[k_m]}, \tag{2.14}$$

où $S^{(m)}$ correspond à la transformée de Fourier discrète de la dérivée d'ordre m du signal s. En pratique, cette dérivée du signal peut être approximée par une simple différence :

$$s'[n] \approx F_s\left(s[n] - s[n-1]\right), \tag{2.15}$$

ou de manière plus précise par convolution avec un filtre différentiateur :

$$s'[n] \approx F_s\left(s[n] * h[n]\right) \text{ avec } h[n] = \begin{cases} \frac{(-1)^n}{n} & \text{pour } n \neq 0 \\ 0 & \text{sinon.} \end{cases} \tag{2.16}$$

2.1.2 Estimation des sons polyphoniques bruités

La plupart des sons naturels sont composés de plusieurs sources sonores qui peuvent se combiner en temps et en fréquence. De plus, certaines de ces sources peuvent être des processus stochastiques perçus comme des sons bruités tels que décrits dans [Han03].

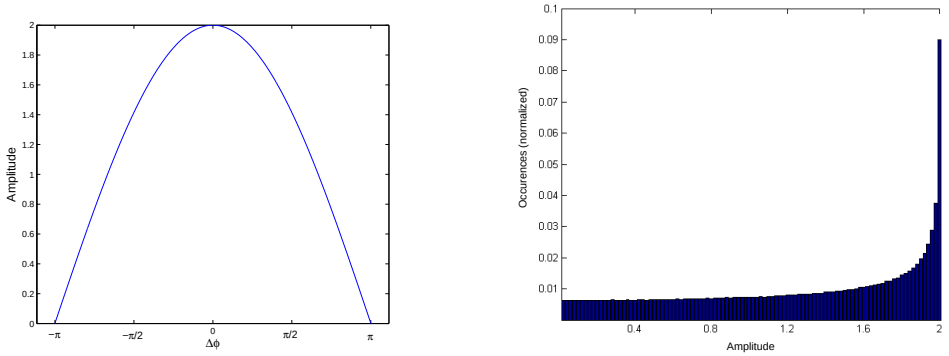
La combinaisons de plusieurs sons ajoute une difficulté au problème de l'analyse des signaux audio. En effet, les signaux qui composent un mélange ne sont pas toujours disjoints et peuvent entrer en collision dans le plan TF.

Par exemple, si on considère 2 signaux stationnaires périodiques de même fréquence ω_0, d'amplitude respective a_1 et a_2 et de phase ϕ_1 et $\phi_2 = \phi_1 + \Delta\phi$ modélisés par 2 composantes sinusoïdales complexes :

$$s_1(t) = a_1 e^{\mathbf{j}(\omega_0 t + \phi_1)} \text{ et } s_2(t) = a_2 e^{\mathbf{j}(\omega_0 t + \phi_2)}, \tag{2.17}$$

l'amplitude mesurée dans le spectre d'amplitude est une combinaison de ces 2 composantes sinusoïdales :

$$|S(\omega_0)| = \left|a_1 + a_2\, \mathbf{e}^{\mathbf{j}\Delta_\phi}\right| = \sqrt{(a_1 + a_2\cos(\Delta_\phi))^2 + (a_2\sin(\Delta_\phi))^2} \tag{2.18}$$



a) valeur des amplitudes

b) histogramme de l'amplitude

FIGURE 2.2 – Combinaison de 2 sinusoïdes d'amplitude $a_1 = a_2 = 1$ en fonction de leur différence de phase $\Delta_\phi \in [-\pi, \pi]$.

Ce résultat montre que la combinaison des amplitudes de 2 sinusoïdes de même fréquence dépend de leur différence de phase. L'amplitude résultante est maximale lorsque $\phi_1 = \phi_2 \Leftrightarrow \Delta\phi = 0$. En pratique, la différence de phase est suffisante pour modéliser les interactions lorsque les deux sinusoïdes ont des fréquences très proche.

Comme nous pouvons le voir sur la figure 2.2, cette valeur maximale correspond à la plus forte distribution si on suppose que $\Delta\phi$ suit une loi uniforme. Cette information peut donc être utilisée pour formuler une incertitude ou faire des approximations sur la valeur de l'amplitude d'un pic sinusoïdal. Cependant, la répartition des autres valeurs est non nulle et vient compliquer le problème de l'estimation des paramètres sinusoïdaux.

En général ce problème est traité en utilisant une approximation basée sur une hypothèse d'additivité sans tenir compte de la distribution de la phase. Dans [YR09], les auteurs proposent une étude comparative des approches existantes ainsi qu'une technique de résolution statistique sous la forme d'un algorithme itératif généralisé pour un nombre quelconque de composantes.

2.1.3 La non stationnarité

Lorsque les paramètres des composantes sinusoïdales évoluent rapidement, le modèle stationnaire même pour des trames d'analyse de petite taille peut s'avérer insuffisant. Le modèle non stationnaire suppose que l'amplitude et la fréquence varient en fonction du temps et se propose de décrire les fonction $a(t)$ et $\omega(t)$ à partir de fonctions polynomiales. Les coefficients des polynômes correspondants sont donnés par l'évolution temporelle de ces fonctions décrite par leur fonctions dérivées. Ainsi, en considérant les dérivées première de $a(t)$ et de $\omega(t)$, une composante sinusoïdale peut se formuler comme suit :

$$s(t) = \underbrace{\exp\left(\lambda_0 + \mu t\right)}_{a(t)} \exp\left(\mathbf{j}\,\underbrace{\left(\phi_0 + \omega t + \frac{\psi}{2}t^2\right)}_{\phi(t)}\right), \tag{2.19}$$

avec $\mu = \frac{\partial \log(a(t))}{\partial t}$ et $\psi = \frac{\partial \omega(t)}{\partial t}$ qui correspondent aux paramètres de modulation d'amplitude, *Amplitude Modulation* (AM) et de modulation de fréquence, *Frequency Modulation* (FM) qui doivent désormais être estimés lors de l'analyse. Pour cela il existe des estimateurs efficaces tels que la méthode de réallocation généralisée [MD08] ou la méthode du vocodeur de phase généralisée proposée plus récemment par Marchand [Mar12a]. Le modèle non stationnaire permet d'améliorer la précision et la robustesse pour l'analyse spectrale des signaux audio réalistes dont les paramètres peuvent varier rapidement en fonction du temps.

2.1.4 Le modèle sinusoïdal à long terme : suivi de partiels

Le modèle sinusoïdal permet d'estimer les paramètres de signaux de durée limitée dans le temps. Lorsque l'on souhaite observer l'évolution d'une entité sonore sur une durée plus longue qu'une trame de signal, il est nécessaire de construire la trajectoire des partiels en combinant les estimations effectuées sur des trames de signal successives. La modélisation sinusoïdale à long terme permet ainsi d'estimer la trajectoire fréquentielle des entités sonores qui composent un mélange comme le montre la figure 2.3.

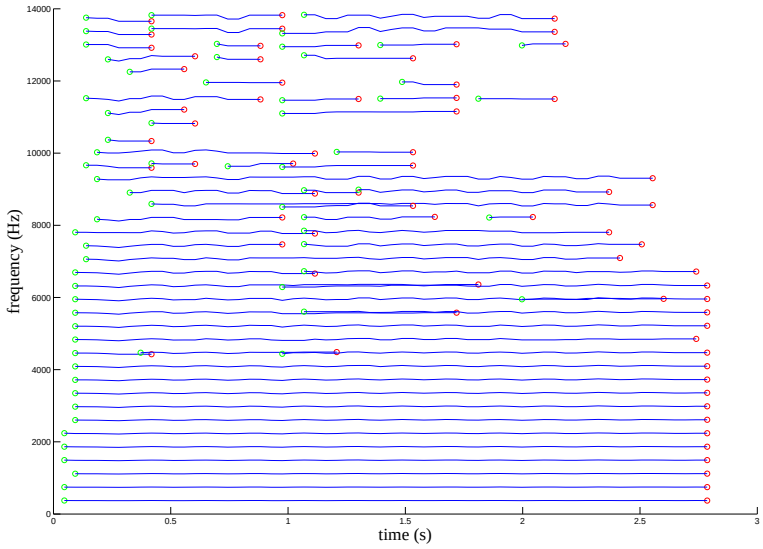


FIGURE 2.3 – Estimation des partiels d'un son de saxophone jouant avec un vibrato. La trajectoire est estimée par prédiction linéaire utilisant un modèle auto-régressif (AR) pour la fréquence et l'amplitude [LMR04b].

La construction des partiels à long terme s'effectue par identification en associant à chaque partiel instantané une trajectoire en combinant plusieurs trames de signal successives. Les principales difficultés rencontrées par les méthodes de suivi sont engendrées par la non-stationnarité (variations de l'amplitude et de la fréquence) et la polyphonie (présence de bruit, de transitoires ou de plusieurs sources sonores). Parmi les méthodes existantes utilisées pour le suivi le suivi de partiels, nous pouvons citer [LMR04a], où Lagrange *et al.* proposent d'explorer l'ensemble des trajectoires possibles pour chaque partiel. La trajectoire retenue est celle qui maximise la vraisemblance. Dans [LMR04b], les mêmes auteurs utilisent la prédiction linéaire à l'ordre n appliquée sur la trajectoire estimée à partir des trames précédentes. Les paramètres du modèle AR sont estimés par la méthode dite de Burg qui repose sur une hypothèse de stationnarité pour chaque tra-

jectoire estimée. Ainsi, chaque partiel est associé avec celui de la trame suivante dont les paramètres sont les plus proches des paramètres prédits. Les applications des méthodes de suivi de partiel sont par exemple la détection des débuts et fins de note, l'application de certains effets ainsi que l'analyse de la scène auditive désignés sous le terme *Computational Auditory Scene Analysis* (CASA) [Bre90]. Le suivi de partiel peut permettre également une représentation compacte d'un signal musical. Cette propriété est utilisée dans la section 4.4.2.

2.1.5 Évaluation de la qualité des méthodes d'estimation

L'évaluation des techniques d'estimation des paramètres sinusoïdaux consiste à mesurer la précision obtenue par les méthodes en utilisant la connaissance dont on dispose sur les paramètres de références. Bien qu'une évaluation subjective soit souvent possible en effectuant des tests d'écoute (*cf.* section 2.3.5) depuis les signaux synthétisés, ce type d'évaluation ne donne en réalité que peu d'information sur la précision réelle d'une méthode (*e.g.* l'oreille humaine est insensible aux translation de la phase). Ainsi, dans les travaux proposés, nous avons choisi d'utiliser les mesures suivantes comme indicateur de performance :

- une mesure d'erreur (ou fonction de distorsion) objective, reposant sur un modèle de signal donné, calculée à partir des signaux synthétisés et des signaux de référence,
- une mesure statistique reposant en pratique sur un nombre important d'expérimentations (présentant un facteur aléatoire comme l'ajout de bruit) réalisées pour un modèle d'observation donné. Ce type de mesure utilise la variance et l'espérance de l'estimateur. Cela donne une indication asymptotique objective sur l'efficacité de cet estimateur, notamment sur la dispersion des valeurs estimées et sur la présence d'un biais. Ces points sont détaillés dans le chapitre 3.

a) L'évaluation en pratique d'un estimateur

Afin de comparer plusieurs estimateurs, il est nécessaire de disposer de données de références permettant de mesurer une erreur. La fonction d'erreur la plus couramment utilisée est l'erreur quadratique calculée entre les signaux synthétisés d'après le modèle de signal en utilisant les paramètres estimés et de références. Pour une fenêtre d'analyse w en utilisant le modèle sinusoïdal stationnaire, cette erreur s'écrit :

$$\begin{aligned} d(a,\omega,\phi,\hat{a},\hat{\omega},\hat{\phi}) &= ||w\,(s-\hat{s})\,||_2^2 \\ &= \sum_{n=n_0}^{n_0+N-1} \left| w[n]\left(a\exp\left(\mathbf{j}(\omega n+\phi)\right) - \hat{a}\exp\left(\mathbf{j}(\hat{\omega}n+\hat{\phi})\right)\right)\right|^2 \\ &= \sum_{n=n_0}^{n_0+N-1} \left| w[n]a\exp\left(\mathbf{j}(\omega n+\phi)\right) - w[n]\hat{a}\exp\left(\mathbf{j}(\hat{\omega}n+\hat{\phi})\right)\right|^2 \\ &= ||w||^2(a^2+\hat{a}^2) - 2a\hat{a}\sum_{n=n_0}^{n_0+N-1} w[n]^2\cos\left((\omega-\hat{\omega})n+(\phi-\hat{\phi})\right). \end{aligned}$$

Sans connaissance des paramètres, il est parfois plus simple d'utiliser le rapport signal / bruit, *Signal-to-Noise Ratio* (SNR) qui nous donne une mesure qualitative en dB. Dans ce cas, on définit la différence entre le signal de référence s et son estimation $\hat{s}$ comme étant du bruit. Cette mesure s'exprime alors par :

$$\text{SNR} = 20\log_{10}\left(\frac{||s||_2}{||s-\hat{s}||_2}\right). \tag{2.20}$$

Dans tous les cas, ces mesures objectives sont couramment utilisées dans la littérature et nous donnent une indication sur la précision d'un estimateur. Ces mesures peuvent être généralisées à tout type de modèle même pour des signaux complexes, cependant ces mesures ne prennent pas en compte la perception humaine. Dans ce cas, il existe d'autres solutions qui sont proposées pour certaines applications spécifiques telles que la compression ou la séparation de sources audio (*cf.* section 2.3). Nous montrons dans la section 2.3.5 qu'il existe des fonctions d'erreur objectives qui sont corrélées avec la perception de certaines altérations présentes sur les signaux audio.

b) Performances statistiques d'un estimateur

Pour évaluer l'efficacité asymptotique d'un estimateur, il est parfois possible de recréer par simulation les conditions dans lesquelles cet estimateur sera utilisé (*e.g.* base de données de test, sons synthétiques, niveau de bruit donné, etc.). Dans ce cas, il est possible d'utiliser certaines mesures statistiques telles que la variance, l'espérance ou l'erreur quadratique moyenne qui sont en général de bon indicateurs de performance. En effet, la variance donne une indication sur la dispersion des estimations et donc sur la précision d'un estimateur. L'espérance permet de déterminer si un estimateur converge vers le paramètre que l'on cherche à estimer. Cela permet en d'autres termes de mesurer l'efficacité de l'estimateur et de mettre en évidence son biais. En effet, nous voyons dans le chapitre 3 que cette étude statistique peut être approfondie à partir d'un modèle d'observation pour définir les limitations théoriques du meilleur estimateur, même si on ne dispose pas de celui-ci. Ainsi, la borne de Cramér-Rao [Rao45] détaillée dans la section 3.1.2 définit la meilleure précision pouvant être atteinte en théorie par le meilleur estimateur sans biais. Ainsi, lorsque qu'un estimateur atteint cette borne, on dit que cet estimateur est efficace.

2.2 La transcription automatique de la musique

La transcription automatique de la musique a pour objectif d'obtenir une représentation symbolique du contenu musical à partir de l'observation d'un ou de plusieurs signaux audio. Dans le cas de la musique, cette transcription devra contenir des informations spécifiques essentielles telles que la hauteur des notes jouées, le rythme, le tempo et une métrique. Ces informations essentielles peuvent être utilisées pour déduire d'autres informations musicales plus abstraites comme la tonalité, les accords et la structure d'une pièce musicale.

Lorsque l'on s'intéresse uniquement à la partie tonale des signaux de musique, c'est-à-dire la partie du signal correspondante aux instruments jouant des notes (*e.g.* voix chantée, piano, etc.), la transcription consiste à détecter les débuts et fins de note appelées aussi activations (*onset*/*offset*) en plus de la hauteur de chacune des notes qui composent le signal analysé. De telles informations sont suffisantes pour obtenir une représentation symbolique précise des événements musicaux en utilisant par exemple le protocole MIDI ou le solfège.

2.2.1 Détection des activations de note

Les débuts de notes correspondent à des changements apparaissant sur le signal observé se traduisant par une hausse de l'énergie parfois brusque et caractérisée par un transitoire (*cf.* section 1.3.2). Ces événements se traduisent par une modification de la nature

du signal caractérisée par l'apparition de partiels associés à une nouvelle source sonore. Parmi les approches existantes, les plus usuelles [Dix01, Lar01] cherchent à détecter les variations soudaines de l'énergie en calculant une fonction d'enveloppe à partir du signal observé :

$$P_b[n] = \sum_{k \in K_b} |S[n,k]|^2, \tag{2.21}$$

où $S[n,k]$ est la transformée de Fourier discrète du signal s utilisant une fenêtre rectangulaire centrée à l'instant n pour l'indice fréquentiel k. Ici, K_b désigne l'ensemble des indices fréquentiels associés à la bande de fréquence b. En pratique on peut découper le spectre en 5 à 10 bandes couvrant l'intervalle 20 Hz - 15 kHz calculé pour des trames de 20 ms utilisant un recouvrement de 50% ou de 75%.

Ainsi, la détection des débuts et fins de note s'effectue en calculant le gradient de $P_b[n]$ approximé par :

$$D_b[n] = P_b[n] - P_b[n-1], \tag{2.22}$$

qui est ensuite comparé à un seuil fixé. Cette approche permet de détecter les variations d'énergie mais ne permet pas de déterminer si elles sont associées à l'apparition d'un instrument tonal, d'un instrument percussif ou d'un son bruité. L'utilisation d'une fonction de détection de ce type peut être suffisante pour faire de la détection du rythme ou du tempo. Elle peut donc aussi être utilisée pour structurer dans le temps une transcription existante, cependant celle-ci n'est pas discriminante pour détecter la présence d'un instrument tonal associé à un modèle de source quasi-harmonique.

2.2.2 Modèle de signal d'un instrument tonal

L'estimation F_0 multiple a pour objectif d'estimer la hauteur des notes de musique associées aux sources quasi-harmonique présentes dans un signal de mélange. Il s'agit donc de la tâche la plus importante d'un système de transcription cherchant à obtenir une transcription MIDI ou solfège à partir d'un signal audio.

Dans le cadre d'une approche non supervisée, on se sert uniquement d'un modèle d'observation permettant de décrire la forme d'onde du signal de tout instrument jouant des notes perçues comme ayant une hauteur (*pitched instrument*).

Ce modèle, décrivant une source sonore harmonique ou quasi-harmonique est caractérisé par une ou plusieurs (dans le cas d'un instrument polyphonique) fréquences fondamentales (F_0) et par ses partiels (multiples entiers ou presque de chaque fréquence F_0).

Comme expliqué dans le chapitre 1, la fréquence fondamentale est liée à la hauteur musicale perçue (ou *pitch*) pouvant être calculée à partir d'une fréquence en hertz en utilisant l'équation (1.14).

Il s'agit d'un modèle [MSW83, FR98] qui permet de décrire l'essentiel de la forme d'onde correspondante au signal d'un instrument de musique tonal (*e.g.* piano, trompette, violon, etc.). En fonction du timbre de l'instrument, le nombre de partiels et la forme de l'enveloppe spectrale peuvent varier. De plus, certains instruments comme le piano présentent un facteur d'inharmonicité qui provoque des écarts de fréquence pour chaque partiel.

Le modèle de signal proposé permettant de décrire dans le domaine temporel le signal d'un instrument jouant une ou plusieurs notes (monophonique ou polyphonique) peut s'exprimer par :

$$s(t) = \sum_{l=1}^{L} \sum_{h=1}^{H_l} a_{h,l}(t) \exp\left(\mathbf{j}\left(2\pi d_\beta(h) h F_l \cdot t + \phi_{h,l}\right)\right), \tag{2.23}$$

où F_l est la fréquence fondamentale de la source l et $d_\beta(h) = \sqrt{1 + h^2\beta}$ permet de modéliser la différence de fréquence de chaque partiel par rapport au modèle harmonique (multiple entier de la fréquence fondamentale). Cette différence fréquentielle est liée au facteur d'inharmonicité défini par le paramètre β et spécifique à chaque instrument. La source est dite harmonique quand $\beta = 0$ ou quasi-harmonique pour $\beta \neq 0$.

Le paramètre L correspond à la polyphonie (nombre de notes simultanées) et H_l est le nombre d'harmoniques de la fréquence fondamentale F_l. Ce modèle réutilise le modèle sinusoïdal introduit dans la section 2.1.1 qui convient pour décrire les signaux périodiques déterministes. Ainsi les paramètres $\phi_{h,l}$ et $a_{h,l}(t)$ sont respectivement la phase initiale et l'amplitude.

Un exemple de spectre d'amplitude correspondant à une source quasi-harmonique naturelle est présenté dans la figure 2.4. Dans cette représentation, le signal peut être modélisé par un peigne de Dirac dont les pics sont également espacés approximativement de la fréquence fondamentale. Dans la suite, nous voyons comment cette structure particulière peut être exploitée par les méthodes d'analyse. La figure 2.4 présente également un exemple de son inharmonique provenant d'un piano. Le facteur d'inharmonicité lié à la physique de l'instrument provoque un écart entre chaque partiel et le modèle parfaitement harmonique.

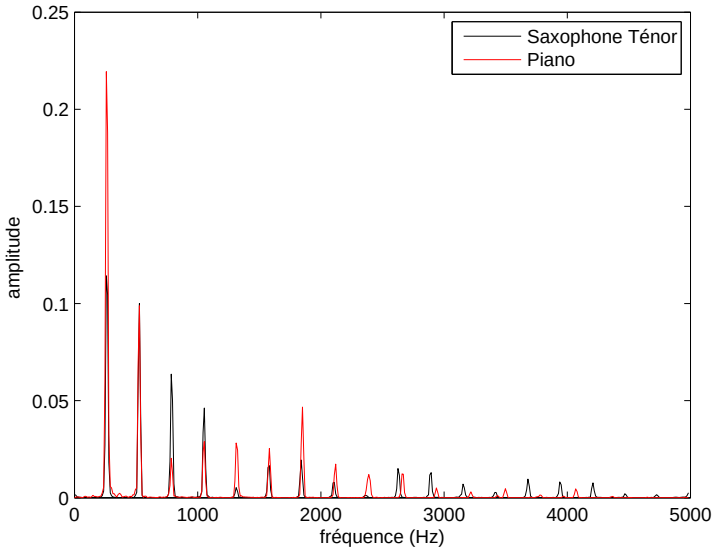


FIGURE 2.4 – Spectre d'amplitude d'un son naturel de saxophone ténor et d'un son naturel de piano superposés et jouant la même note Do3. Le premier partiel correspond à la fréquence $F_0 \approx 261,62$ Hz. Les autres partiels sont des multiples entiers (ou presque) de F_0. On remarque que les partiels du pianos ne sont plus confondus avec ceux du saxophone dans les hautes fréquences en raison de l'inharmonicité.

2.2.3 L'estimation de la fréquence fondamentale des sons monophoniques

Le cas des sons monophoniques ($L = 1$) est un problème qui ne suscite que très peu de travaux actuellement. En effet, des méthodes comme RAPT (*Robust Algorithm for Pitch*

Tracking) [Tal95], YIN (d'après la philosophie du Yin Yang) proposée par Chevigné et Kawahara [dCK02] et plus récemment la méthode SWIPE (*Sawtooth Waveform Inspired Pitch Estimator*) [Cam07] comptent aujourd'hui parmi les méthodes les plus efficaces. En effet, ces méthodes en plus d'être précises (généralement plus de 80% pour des signaux monophoniques), demeurent en général simple à mettre en oeuvre. Par exemple, la méthode YIN repose sur l'utilisation de la fonction d'autocorrélation du signal qui ne nécessite pas de décomposition TF du signal analysé.

Nous décrivons dans cette section quelques outils et approches couramment utilisés pour l'estimation de la fréquence F_0 des sons monophoniques.

a) L'autocorrélation

La fonction d'autocorrélation effectue la convolution d'un signal avec lui même et revient à calculer la projection de ce signal avec chacun de ses translatés. L'autocorrélation d'un signal s est donnée par l'équation suivante :

$$\mathrm{ACF}_s(\tau) = \frac{1}{N} \sum_{t=0}^{N-\tau-1} s(t)s(t+\tau). \tag{2.24}$$

Cette fonction est maximale pour $\tau = 0$ quand le signal est parfaitement en phase avec lui-même. On observe des pics quand τ correspond à une période (ou à des multiples de cette période) présente dans l'échantillon. La plupart des méthodes sélectionnent la période $\tau_0 \neq 0$ telle que :

$$F_0 = \frac{1}{\tau_0}\,,\ \tau_0 = \arg\max_{\tau>0} \left(\mathrm{ACF}_s(\tau)\right), \tag{2.25}$$

L'algorithme YIN [dCK02] d'estimation monophonique utilise une fonction de différence inspirée de la fonction d'autocorrélation appelée fonction de différence de magnitude moyenne[1] définie dans sa version quadratique[2] par :

$$\mathrm{SDF}_s(\tau) = \frac{1}{N} \sum_{t=0}^{N-\tau-1} \left|s(t) - s(t+\tau)\right|^2. \tag{2.26}$$

Les méthodes basées sur l'autocorrélation n'utilisent pas explicitement de modèle de source quasi-harmonique. C'est pour cela qu'elles restent très sensibles aux variations d'amplitude du signal et du bruit et commettent souvent des erreurs de hauteur (choix d'un multiple de F_0). Ces méthodes sont uniquement utilisées dans un cadre de transcription monophonique et sont inefficaces pour la polyphonie.

b) Cepstre

Étant donnée la structure de la forme d'onde d'un son harmonique (peigne de Dirac dans le spectre d'amplitude), il est raisonnable d'appliquer une seconde transformée de Fourier au spectre d'amplitude du signal analysé afin d'y rechercher une périodicité (nous avons vu qu'une source harmonique est composée de partiels également espacés). Une méthode complète a été proposée reposant sur ce principe [Nol67]. La fréquence F_0 est déterminée pour le pic maximal de plus faible quéfrence[3].
Le cepstre réel d'un signal peut être défini à partir de son spectre par :

[1] *Average Magnitude Difference Function*
[2] *Squared Difference Function*
[3] Composante du cepstre réel. Les transformations dans le domaine cepstral portent le nom de "liftrage"

$$C_r(\tau) = \mathcal{F}^{-1}[\log(|S(\omega)|)], \tag{2.27}$$

où τ correspond à la quéfrence mesurée en secondes. Le cepstre permet également d'obtenir des informations sur l'enveloppe spectrale et peut être calculé de différente manières en utilisant respectivement la transformée cosinus discrète pour le calcul des coefficients MFCC[4] ou la prédiction linéaire dans le cas des coefficients LPCC[5] [CSK77].

c) Autocorrélation spectrale

Comme pour le cepstre, il est aussi possible d'appliquer une fonction d'autocorrélation au spectre d'amplitude [LNK87]. Cette technique est sensible à l'alignement des partiels, la fréquence est alors directement donnée pour le décalage non nul maximisant la fonction d'autocorrélation du spectre.

d) La corrélation avec un modèle harmonique : *Harmonic Matching*

L'*Harmonic Matching* trouve les candidats F_0 en calculant la corrélation entre le spectre observé et un modèle de source harmonique. Gribonval [GB03] propose une approche basée sur l'algorithme *Matching Pursuit* [MZ93] et l'utilisation d'un dictionnaire d'atomes harmoniques composé de tous les candidats F_0 possibles appartenant à un intervalle $[F_{min}, F_{max}]$.
D'autres approches moins coûteuses en terme de complexité calculatoire reconstruisent les candidats de source quasi-harmoniques par affectation des partiels détectés à partir du signal observé. Cet affectation des partiels repose souvent sur des critères d'harmonicité et de timbre [Kla06] [Yeh08].

La robustesse de ces techniques dépend évidemment d'autres critères permettant de trier les candidats F_0 en tenant compte par exemple du nombre de partiels, de la qualité des pics, de la distance par rapport au modèle harmonique ou de la régularité des écarts entre chaque pic (*spectral peak inter-spacing*).

A ces critères peuvent s'en ajouter d'autres plus spécifiques à certains instruments qui utilisent des informations *a priori* sur les modèles physiques correspondants [FR98]. L'utilisation de ces informations supplémentaires est souvent essentielle pour améliorer la qualité des estimations. C'est aussi pour cela que nous avons choisi de développer l'approche "informée" dans le cadre de cette thèse.

2.2.4 Évaluation des systèmes de transcription

Pour mesurer l'efficacité d'une technique de transcription, il est nécessaire de comparer la transcription estimée avec une transcription exacte dite de référence.

Ce type de protocole expérimental est souvent très difficile à mettre en oeuvre, notamment pour obtenir une transcription de référence précise lorsque l'on travaille sur des sons réalistes ou naturels pour plusieurs raisons :

- le timbre de certains instruments peut compliquer la détection des débuts et fins de notes (partiels manquants ou mélangés au bruit),

[4] Mel-Frequency Cepstral Coefficients
[5] Linear Prediction Cepstral Coefficients

- le choix de l'interprétation des musiciens cause des différences parfois très importantes entre la partition écrite et la partition jouée, les techniques d'alignement automatique de partition avec l'audio [HDT03] étant souvent insuffisantes.

Ces problèmes peuvent être contournés de plusieurs manières :

- en générant les sons à partir de la transcription en utilisant par exemple un expandeur MIDI pour la base d'évaluation,
- en effectuant un alignement manuel ou semi-automatique (coûteux en temps et pouvant comporter des erreurs) entre le son et sa transcription.

Une fois que l'on dispose d'une base d'évaluation dont on possède la transcription de référence, il est alors nécessaire de définir une métrique capable de caractériser les différentes erreurs de transcription possibles.

a) Fonctions d'évaluation utilisées

Dans le travail décrit dans cette thèse, nous avons choisi d'utiliser la métrique [BED09] utilisée pour l'évaluation MIREX[6] pour la tâche d'estimation F_0 multiple. Le calcul de chaque score s'effectue à partir des mesures définies dans le tableau 2.1 obtenues par comparaison entre la transcription de référence et la transcription estimée.

N_{sys}	Nombre de candidats F_0 estimés par le système évalué.
N_{ref}	Nombre de candidats F_0 réels (référence).
N_{corr}	Nombre de candidats F_0 estimés correctement $N_{corr} \leq N_{ref}$.
N_{miss}	Nombre de candidats F_0 manquants en cas de sous-estimation de la polyphonie.
N_{subs}	Nombre de candidats F_0 substitués.
N_{ins}	Nombre de candidats F_0 insérés en cas de sur-estimation de la polyphonie.

TABLE 2.1 – Définition des mesures utilisées pour l'évaluation d'un système de transcription automatique.

a.1) La précision

Elle donne une indication sur le nombre d'estimations correctes sans tenir compte du type d'erreur.

$$\text{Precision} = \begin{cases} \frac{N_{\text{corr}}}{N_{\text{sys}}} & \text{si } N_{\text{sys}} > 0 \\ 0 & \text{sinon.} \end{cases}, \tag{2.28}$$

La valeur la plus proche de 1 est la meilleure.

a.2) Le rappel (*Recall*)

Il correspond au rapport entre le nombre d'estimations correctes et le nombre exact de F_0.

$$\text{Recall} = \frac{N_{\text{corr}}}{N_{\text{ref}}}. \tag{2.29}$$

La valeur la plus proche de 1 est la meilleure.

[6] Music Information Retrieval EXchange : http://www.music-ir.org/mirex/wiki/MIREX_HOME

a.3) L'indicateur *F-Measure*

Il s'agit d'une fonction combinant simultanément le rappel et la précision définie par :

$$\text{F-measure} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.30}$$

La valeur la plus grande est la meilleure.

a.4) La précision globale (*Accuracy*)

Il s'agit de la précision absolue tenant compte de toutes les erreurs commises. C'est en général l'indication la plus significative prise en compte pour évaluer l'efficacité d'un système de transcription.

$$\text{Accuracy} = \frac{N_{\text{corr}}}{N_{\text{corr}} + N_{\text{miss}} + N_{\text{subs}} + N_{\text{inst}}}. \tag{2.31}$$

La valeur la plus proche de 1 est la meilleure.

a.5) L'erreur totale

C'est l'erreur commise par la méthode (Nombre de F_0 présents - Nombre de F_0 estimés correctement) exprimée par :

$$E_{\text{tot}} = \frac{\sum_{t=1}^{T} \max\left(N_{\text{ref}}(t), N_{\text{sys}}(t)\right) - N_{\text{corr}}(t)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}, \tag{2.32}$$

avec T le nombre de trames considérées.
La valeur la plus proche de 0 est la meilleure.

a.6) L'erreur d'oubli

C'est la différence entre le nombre de F_0 exacts et le nombre de F_0 trouvés sans tenir compte de la justesse, donnée par :

$$E_{\text{miss}} = \frac{\sum_{t=1}^{T} \max\left(0, N_{\text{ref}}(t) - N_{\text{sys}}(t)\right)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \tag{2.33}$$

La valeur la plus proche de 0 est la meilleure.

a.7) L'erreur de substitution

C'est le nombre de notes détectées fausses. En fonction du type d'évaluation, les erreurs d'octave peuvent parfois être tolérées. Son expression est :

$$E_{\text{subs}} = \frac{\sum_{t=1}^{T} \min\left(N_{\text{ref}}(t), N_{\text{sys}}(t) - N_{\text{corr}}(t)\right)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \tag{2.34}$$

La valeur la plus proche de 0 est la meilleure.

a.8) L'erreur d'insertion

C'est l'erreur correspondant aux notes supplémentaires détectées lorsque la polyphonie est sur-estimée. Son expression est :

$$E_{\text{ins}} = \frac{\sum_{t=1}^{T} \max\left(0, N_{\text{sys}}(t) - N_{\text{ref}}(t)\right)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \tag{2.35}$$

La valeur la plus proche de 0 est la meilleure.

2.3 La séparation de sources

Il s'agit d'un des problèmes inverses [Idi01] les plus difficiles qui suscite toujours actuellement un nombre considérable de travaux depuis les années 90 dans différentes disciplines [CJ10, Jut07] (*e.g.* biologie, statistique, traitement du signal et des images, physique, etc.). Comme il serait très difficile de décrire exhaustivement l'ensemble de ces travaux dans ce seul chapitre, nous nous contentons de formuler le problème de séparation de sources appliqué au traitement des signaux audio de parole et/ou de musique). Enfin, nous présentons succinctement les principales approches de l'état de l'art.

2.3.1 Description du problème

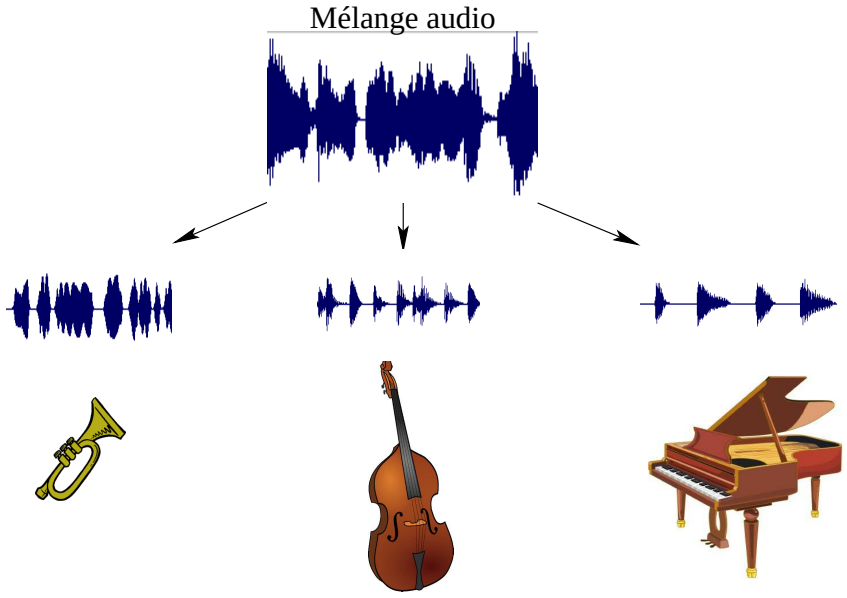


FIGURE 2.5 – Illustration de la séparation de sources en traitement du signal audio à partir d'un signal de mélange. Sa principale application en musique est le démixage qui consiste à retrouver le signal isolé de chacun des instruments qui composent un mélange.

La plupart des signaux audio sont des mélanges composés de plusieurs sources sonores. La séparation de sources consiste à estimer à partir d'un signal (mono ou multicanal) de mélange, les différents signaux qui le composent. La séparation de sources est un problème qui présente un grand intérêt en traitement du son. En effet, disposer des sources isolées qui composent un mélange rend possible de nombreuses applications telles que :

- l'écoute active et la manipulation des entités sonores qui composent un mélange,
- l'application d'effets de haute qualité localisés sur une seule entité sonore, ou sa suppression (effet karaoké),
- la suppression du bruit et le rehaussement d'un signal d'intérêt,
- la création de nouveaux mélanges à partir des sources séparées.

La principale difficulté de ce problème est engendrée par le fait que la création d'un mélange s'accompagne d'une perte d'information causée par la combinaison en temps et en fréquences des sources. Ce point est expliqué en détail dans la section 2.1.2. Cela

s'explique aussi intuitivement dans le cas de la musique parce que les différentes parties instrumentales jouent ensembles. Ainsi, les activations de note sont souvent synchronisées et les fréquences fondamentales jouées par les différents instruments possèdent des relations harmoniques se traduisant par un nombre important de collisions dans le plan TF.

Un signal de mélange brut ne donne *a priori* aucune information structurelle ni sur les signaux que l'on cherche à séparer (modèle de source), ni sur la nature du mélange (modèle d'observation). Ainsi, pour être traité efficacement, ce problème nécessite une identification de la configuration du mélange analysé ainsi que le choix de certaines hypothèses *a priori* pour isoler les signaux des sources qui composent ce mélange. Les techniques existantes comprennent en général les étapes de traitement suivantes :

- l'identification du mélange (nombre de sources, modèle d'observation, matrice de mélange),
- l'apprentissage à partir d'un modèle (estimation de la structure des sources à partir de l'observation),
- la séparation (la synthèse des signaux source à partir du modèle et/ou par transformation du mélange).

2.3.2 Identification de la configuration du problème

Concernant le signal de mélange observé, le problème de séparation de sources considère les différentes configurations et hypothèse suivantes.

- Le problèmes est sous-déterminé lorsque le nombre de sources K est supérieur aux nombre de signaux de mélanges J distincts dont on dispose. Dans le cas où $J \geq K$, le problème est alors déterminé ($J = K$) ou sur-déterminé ($J > K$) et consiste simplement à inverser l'opérateur qui est à l'origine de la création du mélange (*e.g.* la matrice de mélange).
- La nature du mélange considéré peut être linéaire instantané ou convolutif lorsqu'on applique des effets par filtrage sur chaque source. Dans certains cas, on considère le cas non-linéaire en présence d'effets tels que la compression dynamique plus difficile à traiter.
- Les paramètres du mélange peuvent aussi varier ou non au cours du temps (*e.g.* sources sonores en mouvement).

Une configuration très couramment traitée dans la littérature et intéressante d'un point de vue applicatif est le cas sous-déterminé (convolutif ou non) qui s'intéresse à inverser le processus de mélange de signaux musicaux perceptivement crédibles pour un auditeur.

Les méthodes existantes autorisent parfois de l'information *a priori* sous forme de modèle statistique ou physique, de paramètre d'un modèle supposé connus ou comme nous verrons par la suite sous forme de flux d'information supplémentaire disponible pour le processus de séparation. On utilise plus souvent la terminologie suivante pour décrire les différentes classes de problèmes de séparation de sources :

- la séparation de source aveugle (*Blind Audio Source Separation*) regroupe l'ensemble des méthodes pour lesquelles on ne dispose que du mélange sans aucune donnée sur les sources. La méthode de séparation repose donc uniquement sur des

hypothèses définies à l'avance (*e.g.* parcimonie, orthogonalité, structure de la forme d'onde des sources, etc.) et dont la principale est l'indépendance entre les sources[CJ10],

- la séparation de source semi-aveugle, semi-informée qui introduit de l'information partielle sur les sources afin de guider un estimateur (*e.g.* [ES06, SRMS06]),
- la séparation de source informée récemment introduite dans [Knu05, PGB09] qui consiste à utiliser de l'information codée disponible lors du processus de séparation.

2.3.3 Formalisme et modèle du mélange

En traitement audio, une source sonore désigne toujours une entité qui est à l'origine d'un son. Il n'y a cependant pas d'unicité concernant la définition formelle d'une source qui peut désigner aussi bien un signal vocal, un signal de bruit, une source spatiale (signal quelconque provenant d'une position donnée), une source quasi-harmonique (telle que définie dans la section 2.2.2) ou une combinaison de plusieurs modèles élémentaires. Les méthodes de séparation de sources existantes utilisent soit un modèle physique pour représenter la forme d'onde de chaque entité sonore [Vir06], soit un modèle perceptif pour définir ce qu'un être humain perçoit comme étant une seule source [Bre90]. Dans tous les cas, l'observation dont on dispose est un signal de mélange obtenu en combinant les signaux mono-canaux provenant des différentes sources. Lorsque plusieurs sources sont présentes simultanément dans un mélange, leurs signaux respectifs s'ajoutent linéairement. Il s'agit d'une l'hypothèse que l'on retrouve couramment dans la littérature permettant d'exprimer simplement le modèle de mélange comme suit :

$$x_j(t) = \sum_{k=1}^{K} a_{jk} s_k(t), \tag{2.36}$$

où chaque x_j correspond au signal de mélange du canal $j \in [1; J]$. Les $s_k(t)$ sont les signaux correspondant aux sources et a_{jk} sont les coefficients de la matrice de mélange de dimension $J \times K$. Ce modèle devient plus compliqué lorsque l'on cherche à estimer les sources après avoir appliqué des transformations spécifiques sur chacune d'elles ou sur le mélange et en présence d'un bruit d'observation. Dans le cas d'un filtrage linéaire appliqué sur chaque source et en présence de perturbations, on parle alors d'un mélange convolutif bruité pouvant s'exprimer comme suit :

$$x_j(t) = \sum_{k=1}^{K} \sum_{\tau=-\infty}^{+\infty} a_{jk}(\tau) s_k(t-\tau) + b(t), \tag{2.37}$$

où les coefficients de mélange a_{jk} correspondent à un filtre de réponse impulsionnelle infinie. Il existe bien sûr des mélanges plus complexes dont les mélanges non-linéaires qui sont obtenus après l'application de certains effets tels que la compression dynamique [McN84].

Malgré l'existence de ces modèles de mélanges souvent plus compliqués à traiter, Sturmel *et al.* dans [SLP+12] montrent les liens entre les modèles de mélange linéaires souvent plus simples (utilisés pour la séparation de sources) et les pratiques des studios d'enregistrement.

2.3.4 Techniques usuelles pour la séparation de sources audio

Nous présentons ici quelques approches existantes dans la littérature qui permettent lorsque certaines hypothèses sont vérifiées, d'isoler les signaux qui composent un ou

plusieurs signaux de mélange observés.

a) L'approche par filtrage

Cette approche inclue la technique de filtrage spatial (*Beamforming*) et la technique de masquage temps-fréquence. Ces méthodes isolent le signal de chaque source en appliquant sur le mélange un filtre distinct correspondant à la source et calculé sous certaines hypothèses.

a.1) Le filtrage spatial

Le filtrage spatial ou *beamforming* [VB88] exploite la répartition des sources dans l'espace se traduisant par des distinctions entre les différents canaux du signal observé. Cette approche n'est donc pas applicable dans le cas d'un mélange mono-canal. La diversité spatiale permet le calcul d'un filtre permettant de rehausser le signal correspondant à une source provenant d'une direction donnée. Cette technique suppose donc que chaque source possède une distribution spatiale spécifique distincte des autres des autres sources présentes dans l'espace. Connaissant la matrice de mélange qui définit la relation entre les différents canaux du signal observé ainsi que les paramètres spatiaux de chaque source, le meilleur filtre permettant d'isoler une source dans un signal de mélange est la solution du problème d'optimisation qui consiste à minimiser l'énergie directionnelle des autres sources. Cette approche est utilisée pour la séparation de sources sonore dans [PA02] mais peut aisément être appliquée dans d'autres disciplines (*e.g.* traitement du signal radar, réseaux sans-fil, etc.).

a.2) Le masquage temps-fréquence

Cette technique consiste à estimer chaque point TF du spectrogramme où une source donnée est active. Une première estimation de la source contenant des interférences avec les autres sources est obtenue en multipliant le spectrogramme par un masque binaire (0 lorsque la source est inactive ou 1 le cas contraire). La contribution des autres sources est réduite en appliquant un filtrage. Le filtrage de Wiener [Wie49] est par exemple la solution optimale au problème qui consiste à minimiser l'erreur quadratique moyenne du signal estimé. Dans le cas d'un mélange mono-canal le filtre de Wiener peut s'exprimer simplement comme le rapport entre l'énergie de la source considérée et l'ensemble des contributions du signal composé des autres sources et du bruit. Le spectre d'une source estimée peut s'exprimer simplement comme suit (dans le cas d'un mélange linéaire instantané mono-canal) :

$$\hat{S}_k(t,\omega) = \frac{|S_k(t,\omega)|^2}{\sum_{l=1}^{K} |S_l(t,\omega)|^2 + |B(t,\omega)|^2} X(t,\omega), \tag{2.38}$$

où $|S_k(t,\omega)|^2$ et $|B(t,\omega)|^2$ correspondent respectivement au spectre de puissance de chaque source k et du bruit. Comme les spectres des sources et du bruit sont généralement inconnus (cas aveugle), on utilise fréquemment un modèle stochastique (*e.g.* processus gaussien) pour les décrire. Il est alors aisé d'établir une relation de linéarité entre le spectre de puissance et le paramètre de variance de chacune des sources. Le problème de séparation de sources peut alors se résumer à un problème d'estimation des densités spectrales de puissance de chaque source [PC08, LBR11].

b) Approche statistique et par décomposition parcimonieuse

Cette approche regroupe les techniques qui se basent sur une réorganisation des données observées en optimisant un critère utilisant l'hypothèse initiale sur la nature des sources recherchées. Cette nouvelle représentation permet généralement d'isoler les sources en partitionnant les données dans cette nouvelle représentation. Les signaux sources sont retrouvés en appliquant la transformation inverse à partir des éléments isolés.

b.1) L'analyse en composantes indépendantes

L'ICA a été introduite par Comon dans [Com94] pour une problématique de décomposition de signaux, initialement non spécifique au problème de séparation de sources. Ainsi, l'ICA peut aussi être vue comme une extension de la décomposition en composantes principales [Jol02] couramment utilisée pour l'analyse de données statistiques.

L'ICA appliquée à la séparation de sources ne peut être utilisée que dans le cas des mélanges déterminés ou sur-déterminés. Cette technique consiste à décomposer un mélange comme une somme de signaux statistiquement indépendants. Les sources estimées sont les solutions d'un problème d'optimisation dont le but est de minimiser leur dépendance statistique. La dépendance entre les sources X et Y se mesure en général par l'information mutuelle donnée par :

$$I(X,Y) = \sum_{x,y} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right), \tag{2.39}$$

où $P_X(x)$ et $P_Y(y)$ sont les densités de probabilités associées aux sources X et Y. Les signaux des sources x et y sont considérées ici comme des réalisations des variables aléatoires X et Y. Cette information mutuelle donnée par (2.39) correspond à la divergence de Kullback-Leibler $D_{\mathrm{KL}}(P_{X,Y}, P_X P_Y)$.

Bien que l'hypothèse d'indépendance statistiques des sources soit discutable pour l'analyse de mélanges musicaux (car les instruments jouent ensemble), l'ICA a été appliquée avec succès pour la séparation de sources audio par exemple dans [SSWY01].

b.2) Les techniques DUET

Les techniques DUET (*Degenerate Unmixing Estimation Technique*) sont bien adaptées au traitement des mélanges sous-déterminés dont on possède au moins deux canaux. Cette approche [JRY00] suppose que les sources sont distinctes dans l'espace et qu'elles sont orthogonales dans le plan TF. Ainsi, on suppose que seule la source prédominante est active pour chaque point TF. Il est alors possible d'estimer la matrice du mélange en utilisant l'approximation suivante pour chaque source k dans le cas d'un mélange linéaire instantané stéréophonique (2 canaux) :

$$20\log_{10}\left(a_{2k}/a_{1k}\right) \approx 20\log_{10}\left(|X_2(t,\omega)|/|X_1(t,\omega)|\right), \tag{2.40}$$

où a_{1k} et a_{2k} sont les coefficients de la matrice de mélange pour les canaux 1 et 2 affectés à la source k. $X_1(t,\omega)$ et $X_2(t,\omega)$ correspondent aux spectrogrammes du mélange pour les canaux 1 et 2. Les approches DUET utilisent parfois d'autres indices binauraux comme la différence de phase entre canaux [VE02] et peuvent aussi se baser sur un modèle binaural pour estimer en même temps la localisation spatiale de chaque source active [MV09].

c) Approches perceptives par analyse de la scène musicale

L'analyse de la scène musicale ou les méthodes CASA [Bre90, Kas06] cherchent à décrire le contenu d'un signal musical en organisant d'après un modèle perceptif un ensemble

d'informations estimées depuis ce mélange.

Parmi ces informations comptent les fréquences fondamentales, les débuts et fins de note, le rythme, la localisation spatiale de chaque point TF et la synchronicité qui sont extraites par des techniques distinctes spécifiques. Chaque système CASA combine ces informations distinctes pour le calcul de la vraisemblance entre le mélange et chaque source estimée. Chaque source est alors estimée en maximisant cette vraisemblance dépendante du modèle choisi et de la précision des informations estimées. Les techniques CASA possèdent souvent une coût calculatoire important en raison de la complexités des systèmes mis en oeuvre (*e.g.* apprentissage, algorithme Espérance-Maximisation (EM), etc.).

Une technique de séparation de sources audio aveugle efficace pour les instruments quasi-harmoniques a été introduite par Duan *et al.*dans [DZZS08]. Les auteurs proposent une technique estimant simultanément les fréquences fondamentales présentes dans un mélange et les associe à une source en utilisant un modèle statistique sur le timbre de chaque instrument. Les signaux correspondant à chaque source sont finalement reconstruits par synthèse à partir d'un modèle de source quasi-harmonique similaire à celui décrit par l'équation (2.23).

2.3.5 Évaluation des méthodes de séparation de sources

Mesurer les performances d'un système de séparation de sources est une opération difficile qui dépend à la fois des exigences applicatives et des modèles de sources considérés. Ainsi les métriques qualitatives couramment utilisées peuvent être regroupées en deux catégories : l'évaluation objective du signal qui prend uniquement en compte un modèle formel de signal d'erreur (projection ou différence) et l'évaluation perceptive qui tend à quantifier les problèmes audibles depuis les signaux estimés.

a) L'évaluation objective

En raison de la subjectivité des tests d'écoute, il est nécessaire d'utiliser une métrique qualitative objective formelle qui ne dépend que des signaux de référence et de leur estimation. Parmi les mesures les plus usuelles on peut citer l'erreur l_2 entre le signal de la source de références s_k et son estimation $\hat{s}_k$ ou le rapport d'énergie entre le signal et son erreur d'estimation (*Signal-to-Error Ratio*) donnée en dB par :

$$\mathrm{SER}_k = 10 \log_{10} \left(\frac{||s_k||^2}{||s_k - \hat{s}_k||^2} \right). \tag{2.41}$$

Bien que précise, cette mesure d'erreur ne permet pas de comprendre la cause ou la nature de cette erreur. C'est la raison pour laquelle Gribonval *et al.*proposent dans [VGF06] de décomposer le signal d'erreur en plusieurs composantes additives :

$$s_k - \hat{s}_k = e_k^{(\mathrm{art})} + e_k^{(\mathrm{inter})} + e_k^{(\mathrm{dist})}, \tag{2.42}$$

permettant de dissocier plusieurs types d'erreurs spécifiques au problèmes de séparation de sources. Les termes $e_k^{(\mathrm{art})}$, $e_k^{(\mathrm{inter})}$ et $e_k^{(\mathrm{dist})}$ sont associés aux erreurs décrites ci-après.

- Les interférences sont dues à la présence de plusieurs sources dans le signal correspondant à une seule source estimée. Ce type d'erreur met en évidence une mauvaise séparation lorsque les sources sont mal isolées. Elle est définie comme l'ensemble des contributions des autres sources projetées sur la source estimée :

$$e_k^{(\mathrm{inter})} = \sum_{k' \neq k} \langle \hat{s}_k, s_{k'} \rangle \frac{s_{k'}}{||s_{k'}||^2}. \tag{2.43}$$

- La distorsion du timbre définie comme le résidu entre le signal de référence et la projection de la source estimée sur celui-ci. Cette erreur s'exprime comme suit :

$$e_k^{(\text{dist})} = s_k - \langle \hat{s}_k, s_k \rangle \frac{s_k}{||s_k||^2}. \tag{2.44}$$

- Les erreurs d'artefact correspondent à des sons additionnels résultant des transformations appliquées sur le mélange lors de la séparation de sources. Elle peut être déduite en calculant le résidu après l'évaluation des termes d'erreur précédents.

A partir de ces signaux d'erreur, on définit les fonction d'évaluation suivantes données en dB. Le SDR (*Sources-to-Distortion Ratio*) est donné par :

$$\text{SDR}_k = 10 \log_{10} \left(\frac{||s_k + e_k^{(\text{dist})}||^2}{||e_k^{(\text{art})} + e_k^{(\text{inter})}||^2} \right), \tag{2.45}$$

le SIR (*Sources-to-Interferences Ratio*) est donné par :

$$\text{SIR}_k = 10 \log_{10} \left(\frac{||s_k + e_k^{(\text{dist})}||^2}{||e_k^{(\text{inter})}||^2} \right), \tag{2.46}$$

et le le SAR (*Sources-to-Artifacts Ratio*) est donné par :

$$\text{SAR}_k = 10 \log_{10} \left(\frac{||s_k + e_k^{(\text{dist})}||^2}{||e_k^{(\text{art})}||^2} \right). \tag{2.47}$$

Ces mesures sont couramment utilisées pour l'évaluation des méthodes de séparation de sources. Une implémentation MATLAB de ces fonctions d'évaluation est proposée sous forme d'une application libre BSS Eval[7]. Les mesures d'erreurs précédentes ont été intégrées dans l'application PEASS [EVHH11] (décrite plus loin) en combinaison avec un modèle psychoacoustique afin de décomposer l'évaluation perceptive des signaux estimés dans le cadre de la séparation de sources.

b) Les mesures perceptives

Ce type de mesure a pour objectif d'évaluer les dégradations audibles des signaux estimés par un système en donnant une mesure significative sur la qualité d'un son perçu. L'évaluation perceptive est essentielle car elle n'est pas toujours corrélée avec des mesures objectives telle que l'erreur l_2 ou le SNR. En effet, deux signaux identiques mais en opposition de phase seront perçus comme identiques alors que l'erreur l_2 résultante sera très importante. Bien qu'une évaluation perceptive puisse être réalisée de manière informelle par un simple test, leur prise en compte dans le cadre d'expérimentations nécessite une méthodologie rigoureuse faisant intervenir un nombre significatif de sujets humains. Dans d'autre cas, une évaluation automatique est possible sans sujets humains grâce à un système utilisant un modèle psychoacoustique valide.

b.1) Les tests d'écoute

Les tests d'écoute font dans le meilleur des cas intervenir plusieurs sujets chargés d'évaluer la qualité des sons perçus, souvent en octroyant une note ou en répondant à

[7] http://bass-db.gforge.inria.fr/bss_eval/

certaines questions. Ils peuvent être coûteux, long à mettre en place et peuvent manquer de précision en fonction du nombre et de l'expérience des sujets. Ils donnent cependant une indication statistique de la qualité sonore perçue dont la précision dépend d'une méthodologie spécifique. Une formulation rigoureuse de la méthodologie des tests d'écoute est proposée sous forme de recommandations par l'Union Internationale des Radio-Télécommunications (UIT-R).

- La note d'opinion moyenne[8] donnée sur une échelle de 1 (qualité médiocre) à 5 (erreurs imperceptibles) est fréquemment utilisée comme mesure subjective pour évaluer la qualité des signaux audio. Cette mesure nécessite un nombre important de participants pour être significative. Cette technique de sondage convient pour la majorité des tests d'écoute mais nécessite une description détaillée du protocole expérimental (nombre et nature des sujets, interface de test, etc.) pour que l'on puisse "relativiser" sur la précision des résultats.

- La méthodologie MUSHRA[9] (*MUltiple Stimuli with Hidden Reference and Anchor*) utilisée dans l'industrie pour évaluer la qualité des sons compressés peut aussi être utilisée pour évaluer la qualité des résultats d'un système de séparation de sources. En effet, cette méthodologie convient pour des erreurs perceptibles moyennes quelle que soit l'expérience des sujets. En effet, les résultats sont pondérés en fonction de la capacité de chaque sujet à identifier le son de référence et l'ancre (son de dégradation maximale) qui sont cachés lors du test.

  Une évaluation MUSHRA possède les caractéristiques suivantes :

  - chaque participant donne une note sur l'intervalle $[0; 100]$,
  - le son de référence est explicitement accessible pour le test,
  - les tests d'écoute contiennent un nombre aléatoire de sons de référence et de sons "ancres" (de qualité médiocre) cachés. La pertinence des réponses données par chaque participant permet de pondérer chaque note donnée pour le calcul de l'évaluation globale.

  Par défaut, on définit comme ancre des version filtrées passe-bas à 3500Hz du son de référence. Les ancres permettent aussi de s'assurer que les sons évalués qui ne présentent que de faibles défauts ne se voient pas octroyés une note trop faible par le sujet. La méthodologie MUSHRA est souvent privilégiée à la note d'opinion moyenne car permet d'obtenir des résultats pertinents avec un nombre faible de participants.

b.2) L'évaluation perceptive automatique

En raison des difficultés liées à la mise en place des tests d'écoute précis, des systèmes d'évaluation automatiques donnant des mesures de qualité corrélées avec la perception humaine ont été proposés dans la littérature.

Ainsi Huber et Kollmeier décrivent dans [HK06] un système payant baptisé PEMO-Q[10] aujourd'hui accepté et utilisé par la communauté du traitement du signal audio. Ce système intègre un modèle psychoacoustique qui retourne une mesure de similarité appelée PSM (*Perceptual Similarity Measure*) sur $[0, 1]$. Dans [EVHH11], Emiya *et al.* proposent une amélioration de l'évaluation PEMO-Q en établissant un lien entre des tests d'écoute et

[8] http://www.itu.int/rec/T-REC-P.800-199608-I/en
[9] http://www.itu.int/rec/R-REC-BS.1534/en
[10] http://www.hoertech.de/web_en/produkte/pemo-q.shtml

l'utilisation de mesures objectives décrites ci-après. Ainsi, une alternative libre permettant d'effectuer une évaluation perceptive automatique est proposée sous la forme d'une application libre distribuée sous l'appellation PEASS[11].

Les mesures perceptives donnent une indication de qualité audible par un humain mais sont par définition subjectives et non formelles. De plus, elles ne donnent pas (ou très peu) d'information sur la nature des erreurs d'estimation commises. Ainsi, les mesures perceptives peuvent difficilement être exploitées pour améliorer une méthode existante ou pour expliquer les erreurs commises par une technique de séparation de sources. C'est la raison pour laquelle les mesures qualitatives objectives décrites dans la section suivante ont été privilégiées dans les travaux décrits dans cette thèse.

2.4 Conclusions du chapitre

Nous venons de présenter trois problèmes importants d'analyse bien définis dans littérature du traitement audio. Nous avons également décrit quelques techniques de l'état de l'art permettant de les traiter ainsi qu'une méthodologie permettant d'évaluer leur performances. Malgré les avancées récentes effectuées pour traiter ces problèmes, la précision des meilleures techniques utilisant une approche "classique" (non informée) est encore insuffisante pour les applications les plus exigeantes. En effet, la précision du meilleur estimateur permettant de calculer les paramètres sinusoïdaux est limitée par sa borne de Cramér-Rao (*cf.* section 2.1.5) qui dépend du niveau de perturbation d'un signal. Pour la séparation de sources aveugle, les sons sont en général mal isolés et présentent de nombreux artefacts audibles dans le cas de la musique en raison du nombre important de collisions dans le plan TF (*cf.* section 2.1.2) lié à la nature de ces signaux. Enfin, les meilleures techniques de transcription polyphoniques automatiques nécessitent toujours une correction manuelle en post-traitement pour obtenir une transcription fiable. C'est en raison de ces limitations que l'on propose d'introduire dans les chapitres suivants l'approche informée ayant pour but d'améliorer les performances des techniques existantes en les combinant avec de l'information complémentaire.

[11] disponible en ligne http://pageperso.lif.univ-mrs.fr/~valentin.emiya/?page=soft_data

Deuxième partie

Approche informée pour l'analyse des signaux audio numériques

CHAPITRE 3

ESTIMATION ET CODAGE DE L'INFORMATION

L'analyse des signaux de musique s'effectue grâce à des estimateurs dont le but est d'exploiter au mieux l'information présente dans un signal. La théorie de l'estimation [Fri04, KJ92] propose des outils permettant de comparer les performances des estimateurs et d'estimer la quantité d'information disponible dans un échantillon accessible pour le meilleur estimateur théorique.

Dans le cas où l'information présente dans le signal est insuffisante, ou lorsqu'aucune technique existante ne peut atteindre la précision requise par certaines applications, la théorie de l'information propose des outils permettant une transmission efficace de l'information manquante sous forme de données codées.

Pour pouvoir combiner estimation et codage, il est nécessaire de se placer dans une configuration où les deux approches sont applicables simultanément et non triviales (puisque le codage pur permet de coder toute l'information sans utiliser d'estimateur). Une configuration non triviale suppose donc d'avoir accès à la fois au signal de référence x sur lequel on applique un estimateur et aux paramètres d'intérêt θ_i (regroupés dans un vecteur θ) se rapportant au signal x et pouvant être partiellement ou entièrement codés.

Le point de départ de l'approche développée dans cette thèse repose sur le système de communication décrit par la figure 3.1. Dans ce système on souhaite transmettre à un récepteur un signal de mélange audio x et un vecteur de paramètres θ associé avec la meilleure précision possible tout en minimisant le débit nécessaire, c'est-à-dire la taille des informations codées. On suppose connu θ se rapportant au signal x avant l'étape d'encodage. L'encodage consiste à calculer une information $\mathcal{I}$ de taille minimale et à choisir une représentation de x qui seront transmises au décodeur en utilisant deux canaux de communication distincts. Nous verrons plus loin qu'il est possible de combiner les canaux 1 et 2 en utilisant par exemple une technique de tatouage audio numérique qui consiste à cacher de façon inaudible $\mathcal{I}$ dans le signal $\hat{x}$. A la sortie de chaque canal, $\hat{x}' = \hat{x} + \epsilon_x$ et $\mathcal{I}' = \mathcal{I} + \epsilon_{\mathcal{I}}$ sont récupérés et comportent des erreurs aléatoires liées aux perturbations éventuelles sur les canaux 1 et 2.

Le décodeur a donc pour objectif de corriger les erreurs sur $\hat{x}'$ et $\mathcal{I}'$, puis d'appliquer une estimation sur $\hat{x}$ en utilisant $\mathcal{I}$ pour obtenir $\tilde{\theta}$. On souhaite garantir que l'estimation

informée $\tilde{\theta}$ soit toujours plus précise que son estimation classique $\hat{\theta}$ obtenue sans utilisation de $\mathcal{I}$. On souhaite également minimiser le débit nécessaire permettant de transmettre $\mathcal{I}$ pour éviter le cas trivial (codage pur sans utilisation de l'estimateur). On veut ainsi définir au décodage un système permettant d'utiliser efficacement $\mathcal{I}$ dans le processus d'estimation.

En résumé, on souhaite utiliser un tel système pour transmettre (x, θ) en utilisant la connaissance de $\hat{x}$ identique ou presque au codeur et au décodeur et en minimisant la taille des données supplémentaires permettant de retrouver θ à partir de $\hat{x}$ avec la précision désirée.

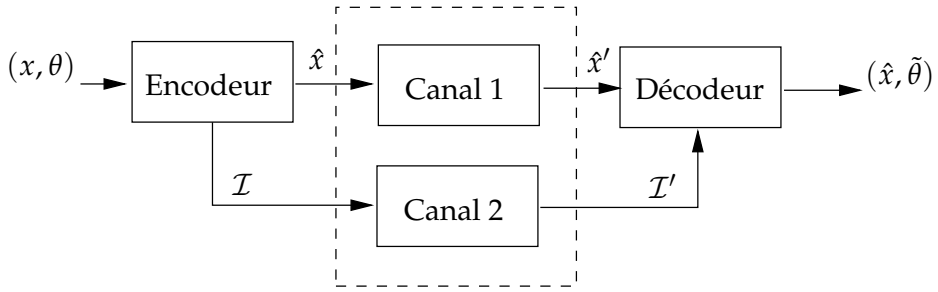


FIGURE 3.1 – Système de communication considéré dans le cadre de notre étude dans lequel on souhaite transmettre un signal de mélange x et un vecteur de paramètres θ.

Compte tenu des contraintes fixées par la formulation de ce problème, nous souhaitons apporter des éléments de réponse aux questions formulées ci-après.

- Comment peut on mesurer et comparer la précision entre plusieurs estimateurs ?
- Quelles sont les limitations théoriques d'un estimateur ?
- Comment mesurer et coder efficacement de l'information pour minimiser sa taille ?
- Comment combiner efficacement simultanément estimation et codage ?

Pour cela, nous avons choisi de développer dans ce chapitre des notions de la théorie de l'estimation et de la théorie de l'information qui sont des domaines d'étude très vastes avec un champ d'application pluridisciplinaire. L'utilisation conjointe de ces deux théories est à ce jour très peu courante dans le domaine du traitement du signal audio où elles sont en général traitées de manière isolée. Pourtant, nous montrons dans ce chapitre qu'elles soient parfaitement applicables au problème que nous venons de formuler. Quelques travaux récents montrent pourtant un intérêt croissant pour la combinaison de ces deux domaines [GSV05, Ver10] dans les problématiques liées au traitement du signal en général.

3.1 Théorie de l'estimation

La théorie de l'estimation s'intéresse aux problèmes de décision qui se posent lorsqu'il s'agit de caractériser un phénomène étudié et donc d'extraire de l'information à partir d'observations. Dans le cadre de notre étude, on considère le cas où l'on cherche à obtenir un paramètre θ qui se rapportent à une observation x. On considère que l'observation x est une réalisation d'une variable aléatoire X de densité de probabilité $p(x|\theta)$.

D'après la théorie de l'estimation, la construction d'un estimateur permettant de retrouver un paramètre θ à partir de x peut être formulée selon les deux hypothèses distinctes décrite ci-après.

- Le cas déterministe basé sur une approche fréquentiste qui suppose que θ est un paramètre fixe.
- Le cas stochastique utilisant l'approche bayésienne lorsque θ est une variable aléatoire décrit par sa densité de probabilité $p(\theta)$.

Dans les deux cas, la théorie de l'estimation permet de construire et d'évaluer des estimateurs permettant d'obtenir la valeur la plus précise possible de θ à partir de x.

3.1.1 Définition d'un estimateur

En traitement du son, on est souvent amené à chercher un ou plusieurs paramètres permettant de caractériser un signal d'observation $x(t)$ généralement bruité. On considère l'observation x qui correspond à un nombre fini de N échantillons regroupés dans un vecteur $x = [x_0, x_1, x_2, \ldots, x_{N-1}]^T$.

Si on considère le cas d'un paramètre scalaire θ et un bruit aléatoire représenté par un vecteur b. Alors on note le modèle d'observation (ou de mesure) comme une fonction h de ces deux paramètres :

$$x = h(\theta, b). \tag{3.1}$$

Comme b est une variable aléatoire, il est impossible de définir exactement θ à partir de x même si h est connue. On peut cependant calculer une estimation $\hat{\theta}$ en définissant une fonction $\hat{\theta}(x)$ aussi proche que possible de θ.

Le paramètre estimé $\hat{\theta}$ est donc une variable aléatoire qui possède une espérance et une variance. La qualité d'un estimateur se mesure par son biais, sa variance et son erreur quadratique moyenne (EQM) qui sont définis par :

$$\text{Biais}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta, \tag{3.2}$$

qui correspond à la différence entre l'espérance du paramètre estimé et sa valeur exacte. Lorsque cette différence est nulle, l'estimateur est dit non biaisé.

La variance d'un estimateur donne une indication sur la dispersion des valeurs estimées et donc sur la précision de cet estimateur :

$$\text{V}[\hat{\theta}] = \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^2\right]. \tag{3.3}$$

Lorsque θ est un vecteur, $\text{V}[\hat{\theta}]$ correspond à la matrice de variance-covariance où $\text{E}[\hat{\theta}]$ est le vecteur des espérances de chaque composante θ_i :

$$\text{V}[\hat{\theta}] = \text{E}\left[\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)\left(\hat{\theta} - \text{E}[\hat{\theta}]\right)^T\right]. \tag{3.4}$$

L'EQM est liée aux mesures précédentes et est donnée par :

$$\begin{aligned} \text{EQM}(\hat{\theta}) &= E\left[|\hat{\theta} - \theta|^2\right] && (3.5)\\ &= ||\text{Biais}(\hat{\theta})||_2^2 + \text{V}[\hat{\theta}]. && (3.6) \end{aligned}$$

Évidemment, plus $\text{EQM}(\hat{\theta})$ et $\text{V}[\hat{\theta}]$ sont faibles, et plus l'estimateur est précis. L'EQM est une information importante car il peut exister des estimateurs biaisés ayant une variance

plus faible que certains estimateurs non biaisés.

L'estimateur $\hat{\theta}(x)$ est une statistique de x. En effet, on définit une statistique $S(x)$ se rapportant à l'échantillon x comme un ensemble d'opérations appliquées sur cet échantillon. Une statistique est exhaustive pour un paramètre θ lorsque $p(x|S(x),\theta) = p(x|S(x))$ c'est-à-dire lorsque la probabilité d'observer x sachant $S(x)$ ne dépend pas de θ. D'après [Fis25], cette statistique contient toute l'information se rapportant au paramètre θ contenue dans l'échantillon x. Les statistiques exhaustives peuvent servir à construire de meilleurs estimateurs. En effet, le théorème de Rao-Blackwell [Bla47] énonce qu'il est possible de construire un estimateur plus précis à partir d'une statistique exhaustive $\mathcal{S}$ et d'un estimateur initial sans biais. Le nouvel estimateur obtenu noté $\hat{\theta}(x|\mathcal{S})$ possède alors une variance plus faible. On peut ainsi déduire intuitivement que cet estimateur augmenté exploite l'information complémentaire fournie par la statistique $\mathcal{S}$. On verra plus loin qu'il est possible définir si un estimateur est optimal en mesurant la quantité d'information contenue dans x.

3.1.2 Approche fréquentiste (déterministe) de l'estimation

Cette approche souvent plus intuitive que l'approche bayésienne suppose que le paramètre θ inconnu est déterministe. D'après le modèle de mesure, l'observation x suit une densité de probabilité paramétrée par θ notée $p(x|\theta)$. Ainsi, cette approche permet de définir des critères qu'il convient d'optimiser lors de la construction de la fonction $\hat{\theta}(x)$.

a) Méthode des moindre carrés

Lorsque l'on possède un modèle mathématique théorique noté $f(y;\theta)$ de variable muette y permettant de décrire les données observées en fonction du paramètre recherché θ, la méthode des moindres carrés consiste à calculer le paramètre $\hat{\theta}$ qui minimise l'EQM. Cet estimateur est formulé comme suit :

$$\hat{\theta} = \arg\min_{\theta} ||x - f(y;\theta)||_2^2 = \arg\min_{\theta} \sum_{i=0}^{N-1} |x_i - f(y_i|\theta)|^2 . \tag{3.7}$$

Cette minimisation peut se faire de manière empirique par ajustement de $\hat{\theta}$ ou de manière analytique en résolvant l'équation :

$$\frac{\partial ||x - f(y;\theta)||_2^2}{\partial \theta} = 0. \tag{3.8}$$

Lorsque le modèle d'observation est linéaire et que l'espérance de la matrice de variance-covariance de l'erreur est nulle, alors le théorème de Gauss-Markov [Pla50] énonce que le meilleur estimateur linéaire sans biais est l'estimateur des moindres carrés.

b) Estimateur du maximum de vraisemblance

Un estimateur du maximum de vraisemblance est un estimateur qui maximise la probabilité $p(x|\theta)$. Cet estimateur peut s'exprimer comme suit :

$$\hat{\theta} = \arg\max_{\theta} \left(p(x|\theta) \right). \tag{3.9}$$

Cet estimateur cherche à obtenir le paramètre $\hat{\theta}$ qui explique au mieux l'observation. Tout comme l'estimateur des moindres carrés, cette maximisation peut s'effectuer empiriquement ou en minimisant le gradient de l'opposé de la fonction $\log\left(p(x|\theta)\right)$ si celle-ci

est différentiable. L'utilisation de la fonction logarithme permettant de simplifier le problème. Parmi les techniques calculatoire d'ajustement les plus connues, nous pouvons citer l'algorithme itératif EM proposé initialement par Dempster *et al.* [DLR77] pour les modèles incluant des variables latentes.

c) Efficacité et bornes théoriques d'un estimateur

Tout estimateur non biaisé d'un paramètre θ de densité de probabilité $p(x|\theta)$ et vérifiant :

$$\forall\theta, \mathrm{E}\left[\frac{\partial \log(p(x|\theta))}{\partial\theta}\right] = 0, \tag{3.10}$$

admet une borne minimale sur sa variance définie comme suit :

$$\mathrm{V}[\hat{\theta}] \geq I(\theta)^{-1} = \mathrm{CRB} \text{ avec } I(\theta) = -\mathrm{E}[\frac{\partial^2 \log(p(x|\theta))}{\partial\theta^2}], \tag{3.11}$$

où $I(\theta)$ est l'information de Fisher [Fis25] relative à θ et où $I(\theta)^{-1}$ correspond à la borne de Cramér-Rao (CRB) [Rao45]. Cette borne définit ainsi la variance minimale atteinte par le meilleur estimateur théorique. Un estimateur sans biais est dit efficace si il atteint la borne de Cramér-Rao. Un estimateur est dit asymptotiquement efficace lorsqu'il atteint cette borne pour un nombre de mesures qui tend vers l'infini.

3.1.3 Approche bayésienne de l'estimation

L'approche bayésienne suppose que le paramètre à estimer θ est une variable aléatoire admettant une densité de probabilité *a priori* notée $p(\theta)$. La loi de Bayes permet d'exprimer la densité jointe entre l'observation x et le paramètre inconnu grâce à l'expression suivante :

$$p(x, \theta) = p(x|\theta)p(\theta). \tag{3.12}$$

Connaissant la densité de probabilité jointe, il est alors possible de calculer :

$$p(\theta|x) = p(x, \theta)/p(x). \tag{3.13}$$

Ainsi, l'estimateur non biaisé optimal minimisant le critère erreur quadratique moyenne, *Mean Squared Error* (MSE) est donné par l'expression suivante :

$$\hat{\theta} = E_{\theta|x}[\theta] = \int \theta p(\theta|x)\, d\theta, \tag{3.14}$$

ce qui correspond à la moyenne du paramètre θ d'après sa loi *a posteriori*.

a) Estimateur du maximum *a posteriori*

L'estimateur du maximum *a posteriori* ou MAP est un estimateur qui maximise la probabilité *a posteriori* du mélange en connaissant la loi $p(\theta)$. Cet estimateur est défini par :

$$\hat{\theta} = \arg\max_{\theta} p(x|\theta)p(\theta) = \arg\max_{\theta} p(x, \theta). \tag{3.15}$$

Ainsi, lorsque $p(\theta)$ suit une loi uniforme, cet estimateur correspond exactement à l'estimateur du maximum de vraisemblance.

b) Bornes de Cramér-Rao *a posteriori*

Dans un contexte bayésien, il est aussi possible de définir une borne théorique appelée aussi borne de Cramér-Rao de Van Trees ou borne de Cramér-Rao stochastique. Pour un estimateur $\hat{\theta}$ sans biais, son EQM vérifie l'inégalité :

$$\mathrm{EQM}(\hat{\theta}) \geq I(\theta)^{-1}, \tag{3.16}$$

où l'information de Fisher correspondante est donnée par :

$$I(\theta) = -\mathrm{E}_{x,\theta}\left[\frac{\partial^2 \log\left(p(x,\theta)\right)}{\partial\theta^2}\right]. \tag{3.17}$$

3.1.4 Vers le codage de l'information manquante

Que l'on choisisse une approche déterministe ou bayésienne, la théorie de l'estimation nous permet de calculer la précision maximale pouvant être atteinte par le meilleur estimateur pour un modèle d'observation donné. Cette borne calculée à partir de l'information de Fisher tend à mesurer la quantité d'information relative à un paramètre disponible dans un échantillon. Il est donc théoriquement impossible d'obtenir une meilleure précision quel que soit l'estimateur choisi sans apporter une information complémentaire sous une autre forme. La théorie de l'information que nous introduisons par la suite fournit des outils théoriques et pratiques pour évaluer et coder cette information complémentaire en vue de son utilisation dans le cadre du problème de l'analyse informée.

3.2 Théorie de l'information

La théorie de l'information définit des fondements permettant de résoudre certains problèmes de représentation des données manipulées pouvant être rencontrés en traitement du son ou plus généralement en traitement du signal. D'après le problème initial décrit par la figure 3.1, on souhaite transmettre en plus du signal x une information complémentaire $\mathcal{I}$ en minimisant la taille des données nécessaires en s'assurant de pouvoir retrouver $\mathcal{I}$ au décodeur.

Les mises en oeuvre théoriques et pratiques de la théorie de l'information que nous décrivons dans ce chapitre concernent le codage appliqué à la compression de données avec et sans perte d'information. Nous ne nous intéresserons donc pas au problème de la cryptographie [Zem00] ni aux codes correcteurs d'erreurs [LC83].

3.2.1 Mesure de l'information

Shannon a défini l'entropie comme mesure de l'information d'un point de vue statistique et physique. Si on considère un message m quelconque, l'entropie propre ou l'incertitude $h(m)$ est définie en fonction de la probabilité d'apparition de ce message. Cette probabilité dépend bien sûr d'un modèle sur la source M qui est à l'origine de ce message et qui suppose certaines connaissances *a priori*. Ainsi, l'entropie de ce message ou sa quantité d'information est donnée par :

$$h(m) = -\log_b\left(p(m)\right). \tag{3.18}$$

D'après cette mesure (*cf.* figure 3.2), plus un message m est improbable et plus il contient d'information, ce qui a pour effet d'augmenter la taille des données nécessaires permettant de le représenter.

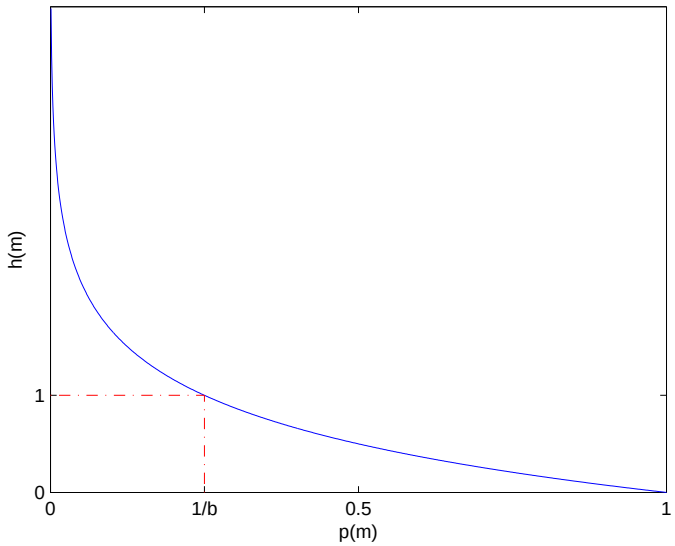


FIGURE 3.2 – Valeur de l'entropie en fonction de la probabilité d'un message m. Plus la probabilité de recevoir un message m est importante, plus la quantité d'information qu'il transporte est faible. Cette figure est obtenue en utilisant l'équation (3.18) avec $b = 4$.

Lorsque l'on souhaite mesurer la quantité d'information portée par une source M, c'est-à-dire une variable aléatoire dont m est une réalisation sur un ensemble Ω de messages possibles, l'entropie correspond à l'information moyenne donnée par :

$$H(M) = -E\left[\log_b(p(m))\right] = -\int_{\Omega} p(m)\log_b\left(p(m)\right)\, dm, \tag{3.19}$$

qui dans le cas d'une source M discrète se calcule par :

$$H(M) = -\sum_{m=1}^{\text{card}(\Omega)} p(m)\log_b\left(p(m)\right). \tag{3.20}$$

Comme il a été proposé initialement par Shannon, cette information est mesurée en bits Shannon (*binary unit*) lorsque $b = 2$ (on rappelle que $\log_2(x) = \frac{\log(x)}{\log(2)}$). Cette quantité correspond théoriquement au nombre moyen de questions ayant 2 réponses possibles (vrai ou faux) qu'il est nécessaire de poser pour deviner un mot m. Il n'existe pas de lien trivial entre les bits Shannon et les bits utilisés en informatique pour coder l'information, en effet l'entropie est une mesure physique qui peut être négative ou avoir des valeurs non entières alors qu'un ordinateur manipule toujours un nombre entier positif de bits. Nous montrons plus loin qu'il est impossible de trouver un code permettant de représenter M dont sa longueur moyenne est inférieure à $H(M)$.

3.2.2 Codage de l'information

Le codage est l'opération qui consiste à associer à chaque élément (ou symbole) m d'une source une séquence d'éléments appartenant à un alphabet $\mathcal{A}$ q-aire. La séquence c_m associée à un élément m constitue un mot du code. L'ensemble des séquences constitue un code C. La longueur moyenne d'un code pour coder une source M dépend de sa densité de probabilité et s'exprime comme suit :

$$L(C) = \sum_{m \in \Omega} p(m) l(c_m) \tag{3.21}$$

où $l(c)$ est la fonction qui retourne la longueur (nombre entier de bits) du mot c donné en paramètre.

a) Déchiffrabilité

On définit $\mathcal{C} : \Omega \to \mathcal{A}^n$ et $\mathcal{D} : \mathcal{A}^n \to \Omega$ les applications permettant de coder un symbole m ou de décoder un mot du code C. On a ainsi $\mathcal{C}(m) = c_m$ et $\mathcal{D}(c_m) = m$.

Un code C est non-singulier ou régulier si tous les mots du code sont distincts. Tous les mots d'un code peuvent être décodés de manière unique si celui-ci est régulier et si il vérifie au moins une des propriétés suivantes :

- tous les mots sont de longueur fixe,
- si un mot du code est utilisé comme séparateur,
- le code est préfixe (ou instantané), c'est-à-dire qu'aucun mot du code n'est préfixe d'un autre.

Dans ce cas, ce code régulier est dit déchiffrable ou uniquement décodable.

b) Inégalité de Kraft-McMillan

Les codes préfixes permettent de définir des codes déchiffrables et de taille variable, ce qui présente un grand intérêt lorsque l'on cherche à réduire la longueur moyenne d'un code pour représenter une source M. Ainsi l'inégalité de Kraft-McMillan [McM56] fournit une condition nécessaire et suffisante pour l'existence d'un code préfixe en fonction des mots qui composent un code.

Théorème 1. *Soit $n_i = l(c_{m_i})$ pour $i \in [1; I]$ la longueur de chaque mot d'un code C permettant de coder les $I = \text{card}(C)$ états d'une source en utilisant un alphabet de destination q-aire. Une condition nécessaire et suffisante d'existence d'un code préfixe ayant ces longueurs de mots est donnée par :*

$$\sum_{i=1}^{I} q^{-n_i} \leq 1. \tag{3.22}$$

Lorsque cette inégalité est une égalité, alors le code est dit complet. Lorsque cette inégalité n'est pas respectée, le code n'est pas déchiffrable (et n'est donc pas préfixe).

Démonstration. On considère un arbre q-aire de profondeur maximale n_I en fixant $n_i < n_j$ si $i < j$ et possédant q^{n_I} sommets terminaux (*cf.* figure 3.3). La condition préfixe impose qu'un mot de longueur n_i exclut $q^{n_I - n_i}$ sommets terminaux. En effet, ce mot ne peut pas être le préfixe d'un autre mot de longueur plus importante. Le nombre total de sommets exclus est borné par le nombre total de sommets q^{n_I} et définit la **condition nécessaire** du théorème :

$$\sum_{i=1}^{I} q^{n_I - n_i} \leq q^{n_I}. \tag{3.23}$$

En divisant chaque membre par q^{n_I} on obtient (3.22). □

La conséquence de l'inégalité de Kraft-McMillan est le théorème suivant :

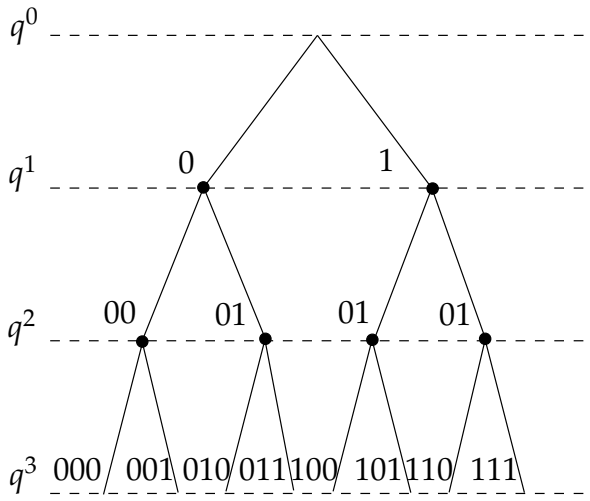


FIGURE 3.3 – Représentation d'un arbre binaire complet correspondant à un code complet utilisant un alphabet q-aire pour $q = 2$. Dans cet exemple la profondeur de l'arbre vaut la taille maximale d'un mot $n_I = 3$. Il existe q^{n_I} mots différents pouvant être codés avec n_I symboles et correspondant chacun à un sommet terminal.

Théorème 2. *La longueur moyenne $\bar{L}$ des mots de tout code préfixe déchiffrable permettant de coder une source M est bornée par :*

$$\frac{H(M)}{\log(q)} \leq \bar{L}, \tag{3.24}$$

en fixant $b = e$ dans l'équation (3.19).

Démonstration. Soit une source M sans mémoire possédant I états et soit $p_i = p(m_i)$ la probabilité d'apparition de m_i associé à un mot de code déchiffrable q-aire et de longueur n_i. En fixant :

$$r_i = \frac{q^{-n_i}}{\sum_{j=1}^{I} q^{-n_j}} \tag{3.25}$$

puis en appliquant l'inégalité de Gibbs[1] à p_i et à r_i on obtient :

$$
\begin{aligned}
-\sum_{i=1}^{I} p_i \log\left(p_i\right) &\leq -\sum_{i=1}^{I} p_i \log\left(r_i\right) \\
\sum_{i=1}^{I} p_i \log\left(p_i^{-1}\right) &\leq -\sum_{i=1}^{I} p_i \left(\log\left(q^{-n_i}\right) - \log\left(\sum_{j=1}^{I} q^{-n_j}\right)\right) \\
\sum_{i=1}^{I} p_i \log\left(p_i^{-1}\right) &\leq -\sum_{i=1}^{I} p_i \log\left(q^{-n_i}\right) + \underbrace{\sum_{i=1}^{I} p_i}_{1} \log\left(\sum_{j=1}^{I} q^{-n_j}\right) \\
\sum_{i=1}^{I} p_i \log\left(p_i^{-1}\right) + \sum_{i=1}^{I} p_i \log\left(q^{-n_i}\right) &\leq \log\left(\sum_{j=1}^{I} q^{-n_j}\right) \\
H(M) - \bar{L}\log(q) &\leq \log\left(\sum_{j=1}^{I} q^{-n_j}\right)
\end{aligned}
$$

D'après l'inégalité de Kraft-McMillan on a $\log\left(\sum_{j=1}^{I} q^{-n_j}\right) \leq 0$ ce qui permet d'écrire :

$$H(M) - \bar{L}\log(q) \leq 0. \tag{3.26}$$

En développant cette expression on obtient bien (3.24). □

En cas d'égalité avec la borne, le code est dit absolument optimum. C'est le cas lorsque $p_i = q^{-n_i}$ et lorsque $n_i = \frac{\log(p_i^{-1})}{\log(q)}$ ce qui est rarement vérifié. Cependant il est possible de construire des codes vérifiant :

$$n_i = \left\lceil \frac{\log(p_i^{-1})}{\log(q)} \right\rceil. \tag{3.27}$$

On parle alors de code compact de Shannon. Le théorème de Shannon que nous énonçons ici sans preuve affirme que :

Théorème 3. *Pour toute source M stationnaire, il existe un code uniquement déchiffrable dont la longueur moyenne $\bar{L}$ est aussi proche que l'on souhaite de sa borne inférieure donnée par l'équation (3.24). De plus ce code vérifie :*

$$\frac{H(M)}{\log(q)} \leq \bar{L} < \frac{H(M)}{\log(q)} + 1. \tag{3.28}$$

Ce thèorème qui énonce une propriété fondamentale du codage, n'est pourtant pas constructif dans la mesure où il n'indique pas comment définir un code permettant d'atteindre cette borne inférieure.

3.2.3 Codage entropique sans perte

Pour tenter d'atteindre la borne inférieure de Shannon, plusieurs techniques de codage ont été proposées dans la littérature. En pratique ces techniques permettent de se rapprocher de la borne théorique cependant aucune n'est idéale car aucune ne parvient à

[1] Inégalité de Gibbs énonce que pour $P = \{p_1, p_2, ..., p_n\}$ et $Q = \{q_1, q_2, ..., q_n\}$ des densités de probabilité, alors on a $-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i)$

l'atteindre. Ces techniques utilisées en général pour effectuer de la compression de données sans perte d'information portent le nom de codage entropique. Parmi ces méthodes compte l'algorithme de Shannon-Fano [Sha48] qui fut l'une des premières techniques exploitant la redondance d'une source. Nous présentons ici des méthodes plus efficaces encore utilisées aujourd'hui [Say06].

a) Le codage de Huffman

Le codage de Huffman [Huf52] est une amélioration du codage de Shannon-Fano qui permet de construire un code préfixe dont la longueur de chaque mot associé à chaque événement m_i d'une source M dépend de sa probabilité $p(m_i)$ ou de son poids p_i (nombre d'occurences ou poids associé à une probabilité d'apparition $p_i = p(m_i)$). Cet algorithme se base sur la structure d'un arbre dont la construction s'effectue en plusieurs étapes :

- On trie les poids $p_i = p(m_i)$ par ordre croissant.
- On construit un arbre q-aire en partant des feuilles où chaque couple (m_i, p_i) correspond à une feuille et où chaque noeud est associé à un poids p_i. Chaque noeud de l'arbre est construit itérativement en sélectionnant systématiquement les q noeuds précédents ayant le poids le plus faible. Le poids associé au nouveau noeud correspond alors à la somme des poids de ses fils. Les fils du nouveau noeud pouvant être une feuille ou un noeud calculé précédemment.
- On réitère la construction des noeuds jusqu'à atteindre la racine (de poids $\sum_{i \in I} p_i$).

L'encodage et le décodage s'effectuent en parcourant l'arbre des feuilles à la racine (pour l'encodage) ou des racines jusqu'à atteindre une feuille pour le décodage. On aura préalablement associé chaque arête à un symbole q-aire de l'alphabet de destination (*e.g.* arête gauche codée par 0, arête droite codée par 1 dans le cas $q = 2$). Le code obtenu affecte des mots de code de taille plus faible aux m_i ayant une probabilité d'apparition plus élevée.

Par exemple, le message **commencement** est associé au tableau de poids suivant :

m_i :	c	o	m	e	n	t
p_i :	2	1	3	3	2	1

qui permet de construire l'abre binaire ($q = 2$) décrit par la figure 3.4.

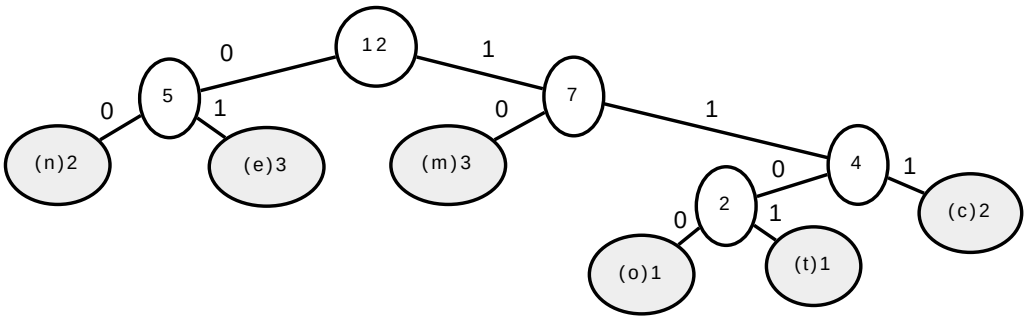


FIGURE 3.4 – Arbre binaire ($q = 2$) du message **commencement** utilisant la convention gauche :0, droite :1 pour le codage des arêtes.

Le code binaire de chaque symbole pour le message **commencement** est donné par :

c	o	m	e	n	t
111	1100	10	01	00	1101

.

Ainsi, le code correspondant au message **commencement** obtenu par concaténation est `111110010100100111011010010011101`.

En pratique, les p_i sont souvent approximés par le nombre d'occurrences normalisé (ou fréquence d'apparition) observé pour chaque mot m_i (*e.g.* le nombre de fois qu'on observe une valeur d'amplitude pour le codage d'un signal audio qui aura été préalablement échantillonné et quantifié sur un nombre fini de valeurs). Ainsi, le codage de Huffman nécessite une connaissance de l'arbre à la fois au codeur et au décodeur qui dans certains cas peut être de taille très importante.

Ces limitations peuvent être contournées en utilisant des variantes, telles que l'algorithme de Huffman adaptatif qui construit l'arbre au fur et à mesure que les m_i sont lus. Le codage de Huffman tronqué qui permet de réduire la complexité liée au calcul de l'arbre lorsque le nombre d'états possibles de la source est important. Dans ce cas, une autre solution peut consister à utiliser de l'indexation récursive ou *Recursive Indexing* [SN92] pour contraindre en pratique la taille du code permettant la représentation des états possibles d'une source.

La longueur moyenne $\bar{L}$ du code obtenu vérifie l'inéquation (3.28). Cependant ce code est limité car l'information est codée sur un nombre entier de bits ce qui l'empêche d'atteindre la borne inférieure.

b) Le codage arithmétique

Le codage arithmétique [Say06] est une alternative intéressante au codage de Huffman souvent plus efficace que ce dernier. Il s'agit d'une technique statistique qui code chaque mot m_i en l'associant à un segment distinct de l'intervalle de $[0;1[$ qui correspond à sa probabilité d'occurrence. Cette approche nécessite pour le codeur et le décodeur la connaissance d'une table d'associations entre chaque m_i et son intervalle $[b_i^{\text{inf}}, b_i^{\text{sup}}[$ correspondant. Chaque symbole m_i est codé par tout nombre réel appartenant à son intervalle associé. Plus un intervalle est petit, et plus le choix d'un représentant nécessitera de la précision (et donc un code de taille plus importante) pour pouvoir être codé.

c) Le codage par dictionnaire : algorithme Lempel-Ziv

L'algorithme Lempel-Ziv [ZL77] est le premier algorithme de type dictionnaire permettant de compresser sans perte et sans connaissance *a priori* de la distribution des symboles à coder. Contrairement aux méthodes précédentes qui utilisent l'hypothèse que tous les éléments qui composent une séquence à coder sont indépendants, cette approche tient compte de la structure de la séquence à coder en tirant bénéfice de la redondance de ces blocs. En pratique, cela se traduit par des meilleures performances que le codage de Huffman lorsque l'on souhaite coder un nombre conséquent de données structurées. Ce codage est plus efficace lorsque ces données présentent une redondance par bloc et celui-ci fonctionne sans utiliser d'*a priori* sur la structure ni sur la distribution de probabilité de ces données. Compte tenu de ses propriétés intéressantes, cet algorithme muni de quelques adaptations peut être utilisé pour réduire la taille de n'importe quel flux de données sans transmission préalable d'un dictionnaire (celui-ci étant reconstruit au décodeur).

3.2.4 Limitations du codage sans perte

Jusqu'à présent nous avons vu des techniques permettant de coder des données sans perte d'information tout en minimisant la taille des codes utilisés. Comme nous avons

vu pour l'algorithme ZL77, la prise en compte de la structure des données permet d'améliorer les performances du codage. Ainsi, dans un domaine spécifique comme le traitement audio, il est également possible de combiner les techniques de codage précédentes avec des représentations parcimonieuses comme l'utilisation des paramètres d'un modèle adéquat ou l'utilisation des échantillons du signal dans un domaine transformé (*e.g.* transformée en cosinus discrète modifiée à valeurs entières, *integer Modified Discrete Cosinus Transform* (intMDCT) ou la transformée de Fourier à court terme, *Short-Time Fourier Transform* STFT) pour améliorer les performances du codage [SOdBB03, BB97].

Malgré l'utilisation de codes optimaux, le codage sans perte n'est pas toujours applicable et possède des limitations, notamment pour le codage d'un signal avec contrainte de taille (support de stockage limité) ou pour certains systèmes de communication (capacité d'un canal limité en bande). Ainsi, dans ses travaux Shannon propose le codage de source ou *Source Coding* [Sha59] et le codage de canal *Channel Coding* [Sha48] qui se distinguent l'un de l'autre par leur domaine d'application. Le codage de canal utilisé en communication et permet le contrôle des erreurs pour la transmission d'informations fiables sur un canal bruité tandis que le codage de source permet de représenter un signal de manière efficace en contrôlant sa taille et son erreur. Le codage de canal et la théorie des codes correcteurs associée n'a pas été développée dans cette thèse. En effet, les codes correcteur ont pour objectif de rendre les données robustes aux altérations en y ajoutant de la redondance. Cela a pour effet d'augmenter la taille des données et de modifier leur format initial. Le codage de source que nous avons choisi de développer ci-après apporte des éléments de réponse pour la minimisation du débit nécessaire permettant de représenter les signaux audio.

3.2.5 Quantification et codage de source

Lorsqu'il existe des contraintes (souvent matérielles) limitant la taille des données il est parfois nécessaire de limiter le nombre d'états pouvant être représentés pour une source. Dans ce cas, il faut trouver un compromis entre la taille des données et l'erreur engendrée par cette perte d'information. Par exemple un signal audio dont les échantillons sont quantifiés en utilisant 8 bits peut représenter chaque valeur d'amplitude en utilisant seulement 256 valeurs distinctes possibles. La quantification consiste donc à définir une fonction Q qui effectue la projection d'un élément x (scalaire ou vecteur) vers son représentant $\hat{x}_q$ qui est l'élément le plus proche de x appartenant à l'ensemble des représentants de Q. Cette opération revient donc à définir les cellules ou intervalles de quantification permettant à la fonction Q d'appliquer la décision concernant le choix du représentant $\hat{x}_q$. Dans le cas scalaire (*cf.* figure 3.5), chaque cellule est définie par un intervalle dont les bornes t_q et t_{q+1} définissent les frontières de la cellule de quantification q. Dans l'exemple de la figure 3.5, la convention suivante est appliquée pour $Q(x)$: si $x \in [t_q; t_{q+1}[$, alors $Q(x) = \hat{x}_q$.

Un quantificateur est caractérisé par un ensemble de cellules de quantification munies de leur représentant respectif. Leur construction est effectuée en utilisant une contrainte sur le nombre total de cellules tout en minimisant l'erreur résultante lors de la quantification d'une source M. Cette optimisation s'effectue en connaissant la densité de probabilité de la source $p(m_i)$ pour une fonction de distorsion (fonction mesurant l'erreur de quantification) donnée. Cette discipline est parfois mieux connue sous le terme de "théorie débit-distorsion" [Gra89] utilisée pour le codage de source.

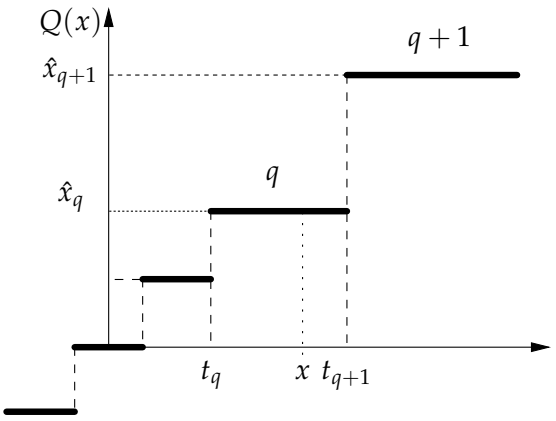


FIGURE 3.5 – Exemple de représentation graphique d'une fonction de quantification d'un scalaire x.

a) La quantification classique

Nous traitons ici dans un premier temps la quantification scalaire pour des raisons de simplification. Le cas vectoriel étant une généralisation multidimensionnelle du problème de quantification.

a.1) La quantification scalaire uniforme

La quantification uniforme utilise une taille de cellule constante (on parle aussi de pas de quantification noté Δ constant). Ainsi, les cellules de quantification sont uniformément répartie sur l'espace des états possibles de la source quantifiée. Ce quantificateur est optimal lorsque la source suit une distribution uniforme ou lorsque l'hypothèse haute résolution est vérifiée (suppose que l'erreur de quantification est répartie uniformément sur chaque cellule). En pratique ce pas de quantification correspond à la différence entre deux représentants ou à la résolution. Pour un pas de quantification Δ, ce quantificateur peur s'écrire :

$$Q(x) = \text{sign}(x) \cdot \Delta \cdot \left(\left\lfloor \frac{|x|}{\Delta} \right\rfloor + \frac{1}{2} \right). \tag{3.29}$$

L'EQM pour une source distribuée selon une loi $p(x)$ est donnée par :

$$\mathrm{E}[(x - \hat{x})^2] = \int_{\mathcal{R}} p(x)(x - \hat{x})^2 \, dx. \tag{3.30}$$

Dans le cas d'une source distribuée uniformément, en appliquant l'équation (3.30) pour une cellule de quantification donnée ayant son représentant situé au centre. L'erreur mesurée est symétrique et prend une valeur sur $[0; \Delta/2]$:

$$\begin{aligned} \mathrm{E}[(x - \hat{x})^2] &= 2 \cdot \int_0^{\Delta/2} x^2 \frac{1}{\Delta} \, dx \\ &= \frac{2}{\Delta} \cdot \frac{\Delta^3}{24} \\ &= \frac{\Delta^2}{12}. \end{aligned} \tag{3.31}$$

a.2) La quantification non uniforme

Pour une source distribuée selon une loi quelconque (surtout non uniforme), le quantificateur optimal permettant de minimiser l'EQM peut être calculé en utilisant l'algorithme de Lloyd [Llo82]. Il s'agit d'un algorithme itératif ayant pour objectif de construire un quantificateur qui minimise l'EQM pour un nombre fixe de cellules de quantification. L'algorithme comprend les étapes suivantes :

(1) On initialise l'algorithme en définissant un quantificateur uniforme avec un nombre T de cellules (et donc $T-1$ seuils de décision).

(2) On calcule les seuils de décision placés entre chaque représentant $t_q = \frac{1}{2}(\hat{x}_{q-1} + \hat{x}_q)$.

(3) On calcule pour chaque cellule son représentant optimal donné en fonction de sa distribution de probabilité : $\hat{x}_q = \dfrac{\int_{t_q}^{t_{q+1}} x p(x)\, dx}{\int_{t_q}^{t_{q+1}} p(x)\, dx}$.

(4) On réitère (2) et (3) jusqu'à ce que les représentants se stabilisent.

Le quantificateur obtenu est alors optimal pour un nombre fixe de cellules en fonction de la distribution $p(x)$ de la source quantifiée.

b) La quantification vectorielle

La quantification vectorielle est une généralisation du problème de quantification aux espaces multidimensionnels (images, volumes, etc.). Pour la quantification d'un élément de dimension k, on considère des éléments $x \in \mathbb{R}^k$. Chaque élément quantifié est représenté par son numéro de cellule $q \in \mathbb{N}$ tel que $Q(x) = q$. Pour la déquantification, on fait correspondre à un numéro de cellule un représentant choisi dans un dictionnaire de $\mathbb{R}^k$: $Q^{-1}(q) = \hat{x}_q$.

La quantification vectorielle a pour objectif de minimiser une fonction de distorsion globale dépendante de chaque dimension. Ce problème peut être traité de plusieurs manières.

- Pour la première approche, chaque dimension est traitée de manière indépendante. Ainsi pour chaque composante du vecteur à quantifier, on applique un quantificateur scalaire optimal, dans ce cas il est nécessaire de répartir l'information disponible sur chaque dimension séparée pour un budget global (en bits) donné. Le problème se ramène à effectuer une allocation en bits optimale permettant de minimiser une fonction d'erreur globale dépendante des k dimensions. Cette allocation peut se faire par l'utilisation d'un algorithme glouton itératif qui détermine bit par bit la quantité d'information affectée à chaque dimension en recalculant pour chaque combinaison la fonction de distorsion globale [GG91].

- Pour la seconde approche, on partitionne l'espace k-dimensionnel afin de construire directement les cellules permettant de minimiser la distorsion globale. Le budget en bits est utilisé pour coder l'indice de chaque cellule. Cette approche nécessite la construction et la connaissance d'un dictionnaire (ou *codebook*) simultanément

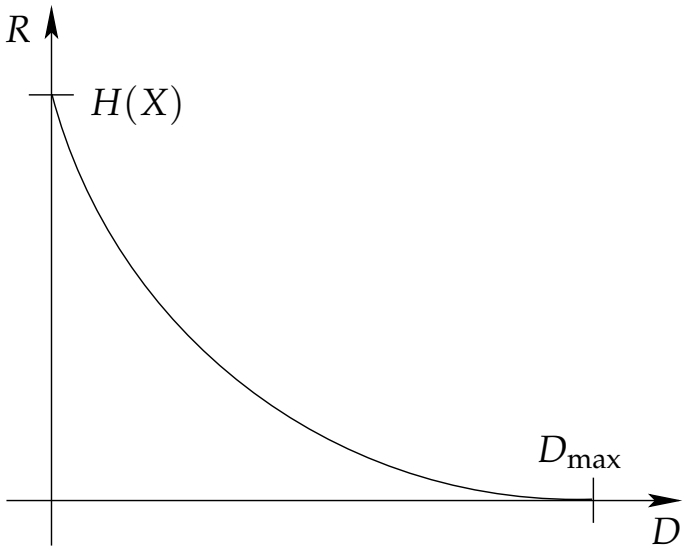


FIGURE 3.6 – Fonction débit-distorsion typique utilisée pour le codage d'une source X (quantification avec contrainte d'entropie).

au codeur et au décodeur afin d'associer chaque indice au représentant de la cellule correspondante. Ce problème peut être résolu en utilisant l'algorithme Linde-Buzo-Gray (LBG) [LBG80] qui est une généralisation de l'algorithme de Lloyd à la quantification vectorielle.

c) La quantification avec contrainte d'entropie

L'algorithme de Lloyd et l'algorithme LBG décrits précédemment proposent des solutions optimales permettant de construire un quantificateur pour un nombre de cellules de quantification fixe (donc à débit constant). Pourtant, nous avons vu dans la section 3.2.3 que l'utilisation d'un code de taille variable peut permettre de réduire le débit global utilisé tout en préservant l'entropie d'une source.

Ce nouveau problème a donc pour objectif de minimiser simultanément la distorsion moyenne $D = \mathrm{E}[d(X, \hat{X})]$ et le débit R. D'après le théorème de Shannon, le débit minimal correspond à l'entropie de la source $H(X)$. La méthode des multiplicateurs de Lagrange nous permet de formuler ce problème par la fonction de coût de Lagrange suivante :

$$J = D + \lambda R = \mathrm{E}[d(X, \hat{X})] + \lambda \int_{\mathbb{R}} p_X(x) \log \left(\frac{1}{p_X(x)} \right) dx, \tag{3.32}$$

où λ est le multiplicateur de Lagrange, $\hat{X} = Q(X)$ et $p_X(x)$ est la probabilité d'apparition du symbole x connaissant la loi de la source X.

Ainsi, si on veut atteindre le débit minimal permettant de représenter une source pour une distorsion moyenne donnée ou si on souhaite atteindre la meilleure qualité possible (distorsion minimale) pour un débit donné il est nécessaire d'établir la relation débit-distorsion de la source X que l'on souhaite quantifier. La fonction débit-distorsion (*cf.* figure 3.6) donnant le débit minimal correspondant à une distorsion moyenne D donnée correspond à la borne inférieure théorique de la meilleure performance de codage avec perte possible. Par définition, la fonction débit-distorsion correspond à l'information mutuelle minimale entre X et $\hat{X}$ telle que la distorsion moyenne est inférieure à une

distorsion moyenne maximale D donnée. Cela peut être formulé comme suit :

$$R(D) = \min_{D^* \leq D} I(X; \hat{X}) \tag{3.33}$$

où l'équation suivate :

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) = \iint_{\mathbb{R}^2} p(x, \hat{x}) \log \left(\frac{p(x, \hat{x})}{p(x)p(\hat{x})} \right) dx\, d\hat{x}, \tag{3.34}$$

correspond à l'information mutuelle entre X et $\hat{X}$.

Pour le problème de quantification initial, il suffit de fixer une entropie cible H_t et de la substituer par le débit dans l'équation (3.32). Le calcul de D se fait plus simplement en utilisant l'hypothèse de haute-résolution qui suppose que l'erreur est distribuée uniformément sur chaque cellule de quantification. Le quantificateur obtenu est un quantificateur uniforme optimal lorsqu'il est combiné à un codage entropique. Une généralisation de l'algorithme de Lloyd appliqué à la construction d'un quantificateur optimal avec contrainte d'entropie a été proposé par Chou, Lookabaugh et Gray dans [CLG89].

3.3 Approche généralisée pour l'analyse informée

Estimation et codage sont souvent traités comme deux domaines distincts, pourtant il existe dans la littérature des applications qui exploitent, souvent de manière empirique, simultanément l'information de Fisher et l'information de Shannon. Des travaux théoriques récents [GSV05, Ver10] montrent pourtant l'intérêt suscité d'établir formellement un lien entre estimation et codage.

3.3.1 État de l'art des techniques combinant estimation et codage

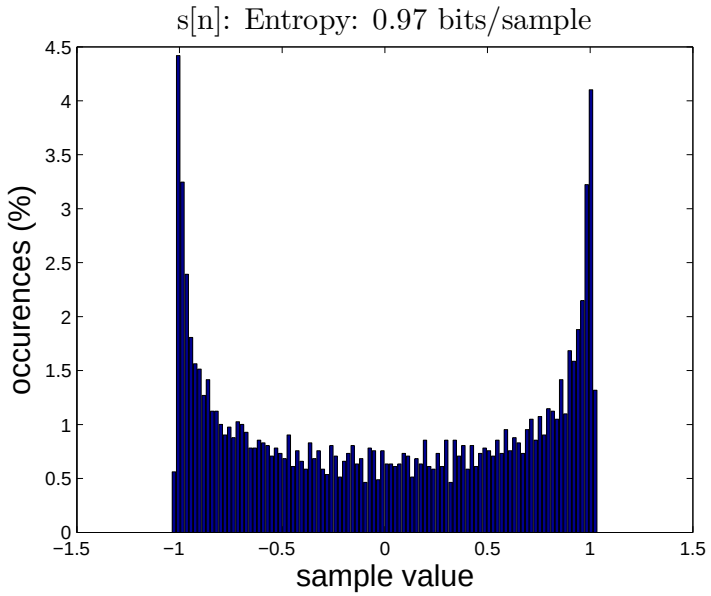
a) Le codage différentiel

Le codage différentiel est une technique qui consiste à coder la différence entre chaque échantillon et une valeur de référence qui peut être par exemple la valeur d'un échantillon voisin, une moyenne locale, une estimation ou une prédiction. Comme le montre la figure 3.7, le codage différentiel permet de réduire la variance de l'ensemble des échantillons que l'on souhaite représenter. En effet, lorsque la valeur choisie comme référence est suffisamment proche de l'état que l'on souhaite représenter, cela a pour effet de concentrer les valeurs autour de 0 (ou du biais de l'estimateur choisi). Cela se traduit notamment par une réduction significative du nombre d'états pouvant être représentés et donc du débit nécessaire permettant de coder une source sans perte d'information.

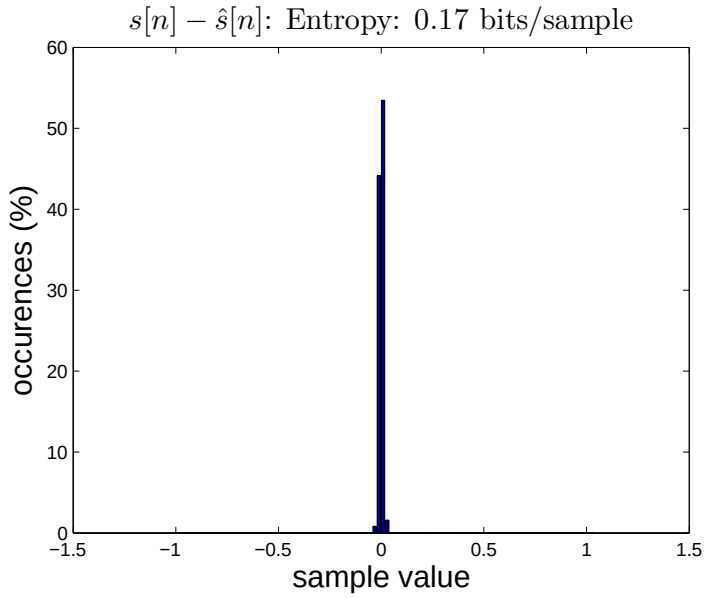
b) Boucle fermée

La boucle fermée est un système de codage qui exploite les propriétés du codage différentiel en utilisant la prédiction. Il s'agit donc d'un système combinant simultanément estimation et codage. L'objectif de la boucle fermé décrit par la figure 3.8 a pour objectif de réduire le débit nécessaire permettant de coder une suite d'échantillons quantifiés x_n. Le principe de la boucle fermée est le suivant.

Au codeur, on calcule l'erreur de prédiction $e_n = x_n - \tilde{x}_n$ (on pourra fixer une valeur de prédiction initiale nulle ou égale à l'espérance de la source si on possède une loi *a priori* sur x_n) que l'on quantifie et que l'on code pour obtenir c_n. La prédiction suivante $\tilde{x}_{n+1}$ est calculée à partir de $\hat{x}_n = \hat{e}_n + \tilde{x}_n$ où $\hat{x}_n$ correspond à la valeur de x_n quantifiée qui sera retrouvée au décodeur.



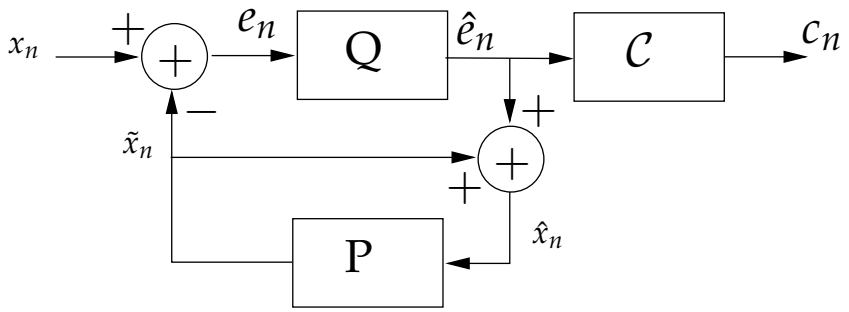
a) signal original



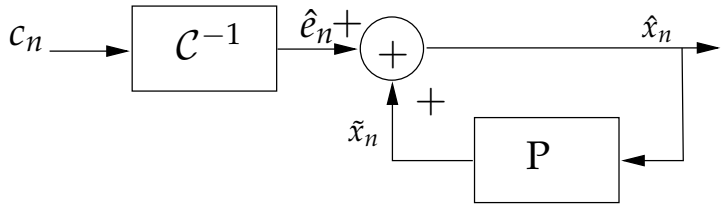
b) différence

FIGURE 3.7 – Histogramme d'un signal temporel discret $s[n]$ normalisé composé d'une sinusoïde et d'un bruit gaussien, quantifié uniformément sur 100 valeurs 3.7a) et histogramme du signal de différence calculé à partir de son estimation en utilisant la méthode de réallocation (*cf.* section 2.1.1) 3.7b). On remarque que l'écart type de la distribution du signal $s[n] - \hat{s}[n]$ est plus faible et possède une entropie plus faible. Le signal $s[n] - \hat{s}[n]$ nécessitera un code de taille plus faible pour être représenté.

Au décodeur, on utilise le même prédicteur pour reconstruire $\hat{x}_n$ à partir de l'erreur $\hat{e}_n$ codée. En fonction du prédicteur et du modèle choisi, un tel système permet d'obtenir un gain significatif par rapport au codage classique. Des variantes du système de boucle fermé mieux connus sous l'appellation *Differential Pulse-Code Modulation* (DPCM) ou *Adaptative Differential Pulse-Code Modulation* [CJF73] sont couramment utilisés dans l'industrie pour le codage compressé du son, de l'image et de la vidéo.



a) codeur



b) décodeur

FIGURE 3.8 – Système de boucle fermée utilisant un quantificateur Q, un prédicteur P et une fonction de codage $\mathcal{C}$ d'après [GG91].

c) Codage et estimation avec information complémentaire

Le problème initial traité dans ce chapitre et décrit dans la figure 3.1 a déjà été considéré dans la communauté du traitement de l'information. Notamment dans des applications de codage avec information complémentaire au décodeur [Wyn75, SW73] qui sont principalement utilisées pour les problèmes de communication où le codeur et le décodeur partagent une information mutuelle sur les données transmises. Les systèmes proposés étudient comment il est possible de réduire le débit nécessaire pour coder une source en tirant profit de cette information *a priori* que possède un décodeur (récepteur) sur les données transmises. Dans ce type de configuration (proche du problème initial décrit par la figure 3.1), la borne débit-distorsion calculée met en évidence un gain théorique significatif [WZ76] lorsque l'information complémentaire peut être exploitée pour le codage d'une source. Toujours appliqué aux communications et à la stéganographie (*e.g.* tatouage de signaux), le *Dirty Paper Coding* [Cos83] [FGLS05] propose de résoudre le problème du codage de source sur un canal bruité en utilisant la connaissance *a priori* sur les dégradations subies par le signal transitant sur ce canal. Cette approche étant plus flexible que la simple utilisation d'un code correcteur dont la capacité de correction et de détection des erreurs est constante.

Dans le cadre des problèmes d'estimation à partir de signaux audio, Knuth [Knu05] propose pour la première fois un cadre théorique pour l'approche informée appliquée au problème de séparation de sources (*cf.* section 2.3). Cette idée est combinée au tatouage audio numérique puis utilisée en pratique par Liu [Liu07] puis par Parvaix [PG11] dans le cadre du projet ANR DReaM[2]. D'autres méthodes plus efficaces ont alors été proposées depuis pour le problème de la séparation de sources informée par Gorlow [GM13], Liutkus [LPB+11] et Sturmel [SD13]. L'approche informée a également été proposée pour la première fois pour d'autres problèmes difficiles comme l'inversion de la compression dynamique [LD08, GR13].

[2]Projet ANR DReaM (ANR-09-CORD-006), le Disque Repensé pour l'écoute active de la Musique.http://dream.labri.fr

3.3.2 Formulation du problème de l'analyse informée

En raison des limitations des approches classiques (non informées), des méthodes plus récentes considèrent l'utilisation d'une information complémentaire pour améliorer les résultats obtenus pour les problèmes d'estimation existants (*cf.* section 3.3.1). Dans cette section, nous proposons de généraliser cette idée à tout problème d'analyse où il est nécessaire d'estimer des paramètres d'un modèle de signal à partir d'une observation détériorée du signal d'origine. Ainsi, le problème de l'estimation de paramètres avec une information complémentaire est formulé et résolu avec la méthode proposée.

a) Approche classique pour l'estimation de paramètres

Soit s un signal réel pouvant s'exprimer comme une fonction d'un paramètre déterministe $p \in \mathbb{R}^{\nu}$ ($\forall \nu \geq 1$) combinée avec un bruit b résultant d'un processus stochastique. Ainsi, le signal observé peut s'exprimer comme suit :

$$s = \mu(p, b), \tag{3.35}$$

où μ est la fonction du modèle du signal observé. Le problème d'estimation classique consiste à retrouver p à partir de s avec une erreur minimale. La valeur estimée $\hat{p}$, résultante de l'utilisation d'un estimateur appliqué sur s noté $\hat{p}(s)$, est un processus stochastique en raison de la présence de b. Ainsi, nous avons :

$$\hat{p}(s) = \hat{p} = p + \epsilon, \tag{3.36}$$

où ϵ correspond à l'erreur d'estimation. D'après la théorie de l'estimation, la borne de Cramér-Rao introduite précédemment définit la variance minimale pour le meilleur estimateur non biaisé (vérifiant $\mathrm{E}[\hat{p} - p] = 0$). Ainsi nous avons :

$$\mathrm{V}[\hat{p} - p] = \mathrm{V}[\epsilon] \geq \mathrm{CRB} \text{ où } \mathrm{CRB} = F^{-1}. \tag{3.37}$$

La matrice de Fisher peut s'exprimer comme la dérivée seconde de la fonction de log-vraisemblance exprimée comme suit :

$$F = -\mathrm{E}\left[\frac{\partial^2}{\partial p^2} \log\left(f(s|p)\right)\right], \tag{3.38}$$

avec $f(s|p)$ la fonction de densité de probabilité de s sachant p. L'inégalité (3.37) signifie que la variance minimale du meilleur estimateur $\hat{p}(s)$ est bornée pour le meilleur estimateur. Le seul moyen d'atteindre une variance cible notée $V_{\text{cible}} \leq \mathrm{CRB}$ pour un modèle donné est d'utiliser de l'information complémentaire.

b) Approche informée pour l'estimation de paramètres

On suppose la configuration suivante décrite dans par la figure 3.9 similaire aux techniques séparation de sources informée, *Informed Sources Separation* (ISS) existantes où p est supposé exactement connu avant la synthèse de s obtenue en appliquant l'équation (3.35).

Nous souhaitons a présent minimiser simultanément l'erreur d'estimation ϵ ainsi que le débit utilisé permettant de coder l'information complémentaire notée $\mathcal{I}$.

Dans cette configuration (*cf.* figure 3.9), l'information minimale $\mathcal{I}$ est calculée au codeur à partir de p et de V_{cible} en utilisant I_σ qui dépend de la précision de l'estimateur. au décodeur, $\mathcal{I}$ est combiné avec $\hat{p}$ pour obtenir l'estimation informée vérifiant

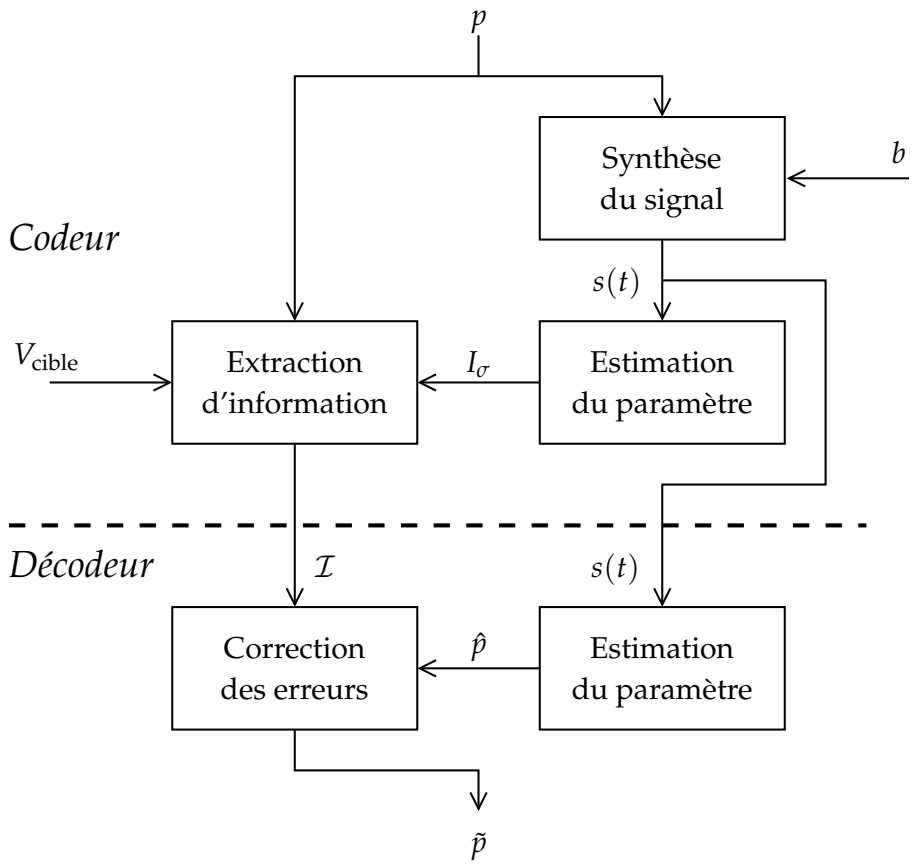


FIGURE 3.9 – Schéma de l'approche informée pour l'estimation d'un paramètre p.

$\mathrm{V}[\tilde{p} - p] = V_{\text{cible}} \leq \mathrm{V}[\hat{p} - p]$. Pour un estimateur non biaisé, on remarque que la variance est égale à l'EQM. En effet, $\mathrm{V}[\hat{p} - p] = \mathrm{E}[(\hat{p} - p)^2] = \mathrm{E}[\epsilon^2]$.

La méthode proposée pour traiter ce problème considère deux configurations distinctes qui sont respectivement le cas scalaire $p \in \mathbb{R}$ et le cas vectoriel généralisé pour $p \in \mathbb{R}^\nu$ pour $\nu > 1$.

3.3.3 Analyse informée d'un seul paramètre

On suppose que l'on doit estimer un paramètre réel normalisé $p \in [0; 1[$. Le signal s est obtenu d'après (3.35) d'après p et b. L'information permettant de retrouver p en utilisant son estimation $\hat{p}$ est obtenue comme suit.

D'abord, nous définissons $\mathcal{C}_d : [0; 1[\rightarrow \{0, 1\}^d$ l'application de codage binaire à virgule fixe et $\mathcal{D}$ l'application de décodage correspondante telle que : $C = (C_1, C_2, ..., C_d)$ est la représentation de p et $\tilde{p} = \mathcal{D}(C) = \sum_{i=1}^{d} C_i 2^{-i}$ est la valeur quantifiée sur d bits de p. Les applications de codage et de décodage sont associées à un quantificateur scalaire uniforme (*cf.* section 3.2.5) utilisant un pas de quantification $\Delta = 2^{-d}$. La précision en bits d permettant de construire le quantificateur peut être déduite simplement à partir de la fonction de distorsion choisie et de l'erreur de quantification résultante. Par exemple, dans le cas de l'EQM, Δ peut être déduit en utilisant l'équation (3.31) et $d = \lceil \log_2(\Delta) \rceil$.

Dans un deuxième temps, on définit I_σ le bit de poids le plus fort, *Most Significant Bit* (MSB) de la représentation binaire de la borne supérieure de l'intervalle de confiance pour l'estimateur choisi. I_σ est une valeur asymptotique qui peut être déduite analytiquement

de la fonction de densité de probabilité du bruit et de la variance de l'estimateur. Comme nous verrons plus loin (*cf.* chapitre 4), I_σ peut être calculé en pratique pour un nombre significatif d'estimations $\hat{p}$ pour une fonction de densité de probabilité du bruit fixée dans le cas où le bruit peut être mesuré ou simulé.

Ainsi, on peut séparer la représentation binaire de l'estimation $\hat{p}$ (*cf.* expérimentation figure 4.1) telle que I_σ correspond à l'indice séparant la partie "fiable" de la partie "non fiable" de $\mathcal{C}(\hat{p})$ obtenue par l'estimateur. Cela peut être formulé comme suit :

$$\mathcal{C}_d(p) = \underbrace{C_1, C_2, \ldots, C_{I_\sigma-1}}_{\text{partie fiable}}, \underbrace{C_{I_\sigma}, \ldots, C_d}_{\text{partie non fiable}} . \tag{3.39}$$

Ainsi, $\tilde{p}$ peut être retrouvé exactement à partir de $\hat{p}$ en utilisant $\mathcal{I}$:

$$\mathcal{I} = C_{I_\sigma-1}, C_{I_\sigma}, \ldots, C_d \tag{3.40}$$

à condition que $I_\sigma \leq \text{msb}\left(\mathcal{C}(|p-\hat{p}|)\right)$. L'information supplémentaire notée $\mathcal{I}$ est définie comme la partie de $\mathcal{C}(p)$ comprise entre les indices $I_\sigma - 1$ et d (partie "non fiable" de $\mathcal{C}(\hat{p})$). La valeur de $C_{I_\sigma-1}$ est utilisée lors de la correction des erreurs basée sur un principe de substitution binaire appliqué par l'algorithme 1. L'estimation informée notée $\tilde{p}$ est alors finalement retrouvée pour $\forall \hat{p} \in [p - 2^{-I_\sigma}, p + 2^{-I_\sigma}]$ en utilisant $\mathcal{I}$ dans l'algorithme 1, où "inc" et "dec" sont les fonctions d'incrémentation et de décrémentation appliquées

Algorithme 1 Correction des erreurs de $\hat{p}$ en utilisant $\mathcal{I}$

$C \leftarrow \mathcal{C}(\hat{p})$
$l \leftarrow \text{taille}(\mathcal{I})$
si $l \geq 2$ **alors**
  $C(I_\sigma : I_\sigma + l - 2) \leftarrow \mathcal{I}(2 : l)$
fin si
$\tilde{p} \leftarrow \mathcal{D}(C)$
si $\mathcal{I}(1) \neq C(I_\sigma - 1)$ **alors**
  $C^{\text{ante}} \leftarrow C(1 : I_\sigma - 1)$
  $C^{\text{post}} \leftarrow C(I_\sigma : d)$
  $p^+ \leftarrow \mathcal{D}(\text{inc}(C^{\text{ante}}), C^{\text{post}})$
  $p^- \leftarrow \mathcal{D}(\text{dec}(C^{\text{ante}}), C^{\text{post}})$
  si $|\hat{p} - p^+| < |\hat{p} - p^-|$ **alors**
    $\tilde{p} \leftarrow p^+$
  sinon
    $\tilde{p} \leftarrow p^-$
  fin si
fin si
return $\tilde{p}$

sur la représentation binaire fournie en paramètre. Pour l'écriture de cet algorithme, nous avons choisi la notation MATLAB telle que $C(i)$ correspond à C_i et $C(i : j)$ représente le vecteur $[C_i, C_{i+1}, \ldots, C_j]^T$.

L'algorithme 1 a pour objectif de substituer la partie "non fiable" de $\mathcal{C}(\hat{p})$ avec $\mathcal{I}(2 : l)$ où $l = \min(d, d - I_\sigma + 2)$ correspond à la taille du vecteur $\mathcal{I}$. La valeur du bit situé à la position $I_\sigma - 1$ est utilisée pour effectuer une comparaison avec $\mathcal{I}(1)$ pour vérifier si la substitution a suffi pour la correction. En effet, lorsque $C(I_\sigma - 1) \neq \mathcal{I}(1)$, une opération arithmétique complémentaire est requise pour résoudre un problème d'incompatibilité de la représentation binaire lorsque $C(1 : I_\sigma - 1) \neq \hat{C}(1 : I_\sigma - 1)$ pour $\hat{C} = \mathcal{C}(\hat{p})$ et $C =$

$\mathcal{C}(p)$. Cette différence peut survenir en raison du mécanisme arithmétique de retenue. Dans ce cas, la représentation binaire de $\hat{p}$ est séparée en deux parties C^{ante} et C^{post} qui sont utilisées pour calculer deux candidats possibles p^+ et p^-. Le candidat le plus proche de $\hat{p}$ est utilisé pour le calcul de la valeur corrigée de $\tilde{p}$.

Dans les applications audio, I_σ peut être estimé directement à partir du mélange en utilisant une technique d'estimation du bruit (*e.g.* [MHM06]) ou peut être déduit en utilisant d et la taille de $\mathcal{I}$. Dans les autres cas, I_σ doit être transmis en utilisant un maximum de $\lceil \log_2(d) \rceil$ bits.

On remarque que dans la configuration considérée, la valeur exacte de $\hat{p}$ est supposée différente au codeur et au décodeur car résulte d'un processus stochastique. Cela est d'autant plus réaliste lors de l'utilisation d'une technique de tatouage où le mélange analysé dépend de l'information complémentaire qui y est cachée de manière inaudible. Cette configuration rend impossible la mise en place d'une boucle fermée telle que décrite dans la section 3.3.1. En effet, il est impossible de prédire la valeur de $\hat{p}$ au décodeur en raison du bruit aléatoire ajouté au mélange analysé.

De plus, dans le cas des techniques de tatouage basées sur la quantification à modulation d'indice, *Quantization Index Modulation* (QIM) [CW01], il a été démontré qu'il est impossible de retrouver les valeurs initiales du mélange avant le tatouage. La preuve est formulée dans l'annexe A.1.

Les travaux présentés dans cette section ont fait l'objet d'une publication [MF10] dans laquelle nous avons montré qu'il était possible de combiner un estimateur classique (non informé) estimant un seul paramètre scalaire avec de l'information complémentaire afin d'en améliorer la précision.

3.3.4 Généralisation au cas vectoriel pour l'analyse informée

A présent nous devons estimer un paramètre $P \in [0;1[^\nu$ un vecteur réel de dimension ν. Comme nous souhaitons minimiser la taille des données complémentaires ainsi que l'erreur résultante, P doit être quantifié. La théorie débit-distorsion [Gra89] permet de résoudre ce problème en utilisant la technique de quantification vectorielle avec contrainte d'entropie afin d'obtenir $\tilde{P}$ (*cf.* section 3.2.5).

Ainsi, pour une distorsion moyenne maximale cible $D = \mathrm{E}[\delta(P, \tilde{P})]$, le théorème de Shannon énonce qu'il existe un code de débit minimal $R = H(\tilde{P})$. Le problème de quantification peut alors être formulé comme un problème de minimisation de la fonction de coût de Lagrange suivante :

$$J = D + \lambda R, \tag{3.41}$$

où λ est le multiplicateur de Lagrange. La solution de ce problème de minimisation permet d'obtenir la fonction débit-distorsion $R(D)$ qui est définie comme la borne inférieure du débit nécessaire permettant de coder $\tilde{P}$ avec une distorsion moyenne D.

Une solution calculatoire consiste à utiliser l'algorithme de Lloyd généralisé appliqué à la quantification vectorielle sous contrainte d'entropie proposée initialement par Chou *et al.* in [CLG89]. Le quantificateur obtenu est presque uniforme d'après la théorie débit-distorsion [Gra89] et peu aisément être combiné avec n'importe quelle technique de codage entropique (*cf.* section 3.2.3).

Après l'étape de construction du quantificateur optimal, il est nécessaire de calculer l'information complémentaire permettant de retrouver $\tilde{P}$ à partir de n'importe quelle estimation $\hat{P}$. Comme chaque composante P_i du vecteur a une contribution qui lui est propre sur la distorsion globale D, il en résulte une précision variable relative à chaque composante du vecteur (et donc une répartition de l'entropie globale sur chaque composante du vecteur). Sous l'hypothèse de haute-résolution utilisée couramment pour la construction des quantificateurs vectoriels à contrainte d'entropie, il résulte (d'après [GG91]) un

quantificateur uniforme appliqué sur chaque dimension du vecteur. Il est alors simple de déduire à partir du pas de quantification de chaque composante un budget en bits d_i entiers permettant de coder chaque composante P_i. Il est alors possible d'appliquer l'analyse informée scalaire (*cf.* section 3.3.3) sur chaque composante du vecteur P. La méthodologie permettant d'appliquer l'analyse informée dans le cas vectoriel peut être résumée comme suit pour le codeur et pour le décodeur.

.1) Codeur

- Synthèse du signal s à partir de P d'après le modèle d'observation (3.35),
- quantification vectorielle optimale sous contrainte d'entropie de P en utilisant [CLG89] (ou n'importe quelle méthode équivalente) pour une distorsion cible D^{cible},
- calcul de $I_{\sigma,i}$ permettant de définir pour chaque composante P_i la partie "fiable" et la partie "non fiable" en utilisant un estimateur $\hat{P}(s)$ (*cf.* équation (3.39) dans la section 3.3.3),
- extraction de l'information complémentaire définie par le ν-uplet : $\mathcal{I} = (\mathcal{I}_1, ..., \mathcal{I}_i, ..., \mathcal{I}_\nu)$ (*cf.* équation (3.40) dans la section 3.3.3) qui peut être codé *a posteriori* en utilisant une technique de codage entropique (*cf.* section 3.2.3),
- Le signal s et l'information complémentaire $\mathcal{I}$ sont transmis au décodeur (*cf.* figure 3.1).

.2) Décodeur

- Estimation de $\hat{P}$ à partir de s,
- quantification vectorielle optimale sous contrainte d'entropie de $\hat{P}$ à partir de la même technique et des mêmes paramètres qu'au codeur (pour une distorsion cible D^{cible}),
- application de l'algorithme 1 en utilisant les $\mathcal{I}_i$ pour chaque composante pour retrouver les $\tilde{P}_i$,
- resynthèse du signal s à partir de $\tilde{P}$ d'après le modèle (3.35) sans le bruit.

Nous proposons dans le chapitre suivant d'exploiter la méthode proposée dans le cadre d'une application pratique de traitement du signal audio.

3.4 Conclusions du chapitre

Dans ce chapitre nous avons défini la problématique de cette thèse puis nous l'avons placée dans un contexte théorique général bien établi dépassant le simple cadre du traitement des signaux audio. Nous avons rappelé des éléments essentiels de la théorie de l'estimation et de la théorie de l'information sur lesquels s'appuient les travaux de cette thèse. Enfin, nous avons proposé une première contribution originale en formulant un cadre de travail ainsi que des éléments de solution pratiques permettant d'appliquer l'approche informée aux problèmes d'analyse audio (développés dans les chapitres suivants). La technique introduite utilise un mécanisme de substitution qui rend plus robuste le processus de correction des erreurs dans les configurations où il est nécessaire de corriger des erreurs aléatoires. Les expérimentations présentées dans les chapitres suivants montrent que cette approche permet d'améliorer le rapport débit-distorsion du

codage tout en améliorant la précision de l'estimation classique. Nous avons tenté de rendre ce chapitre suffisamment général afin d'ouvrir d'autres perspectives d'applications dépassant le cadre du traitement du signal audio (*e.g.* traitement d'images, traitement vidéo, télécommunications) où les méthodes décrites restent applicables.

CHAPITRE 4

L'ANALYSE SPECTRALE INFORMÉE

Les deux domaines principaux permettant de représenter et de comprendre les signaux musicaux sont le temps et les fréquences (*cf.* section 1.2). Le domaine temporel est bien adapté pour représenter directement les variations de pression du milieu à l'intérieur duquel un son se propage. Le domaine fréquentiel permet de représenter la hauteur des sons perçus ou de décrire le timbre des différentes entités sonores qui composent un mélange.

Cependant, les fréquences contenues dans un extrait musical peuvent varier dans le temps. Ces évolutions ne peuvent être observées qu'en choisissant une décomposition du signal en fonctions élémentaires bien concentrées à la fois en temps et en fréquence.

La STFT basée sur la transformée de Fourier à fenêtre compte parmi les outils les plus couramment utilisés en traitement audio. Cette approche bien qu'efficace en pratique pour observer des fréquences instantanées possède des limitations. En effet, quelle que soit la méthode choisie, la résolution TF est limitée car soumise au principe d'incertitude d'Heisenberg qui montre que l'énergie d'une fenêtre d'analyse $w_{t,f}$ est localisée dans un rectangle ayant pour côtés σ_t, σ_f et pour centre (t, f) tel que :

$$\sigma_t \sigma_f \geq \frac{1}{2}. \tag{4.1}$$

Malgré cette contrainte, la décomposition des signaux en composantes TF élémentaires reste largement utilisée pour les problèmes d'analyse et permet d'introduire des modèles paramétriques associés à des méthodes d'analyse qui peuvent être séparées en deux classes :

- les méthodes analyse/synthèse qui permettent grâce à un ensemble de paramètres ou de coefficients de reconstruire le signal temporel d'origine de façon exacte ou approchée,
- les méthodes d'analyse pures permettant d'estimer de l'information utile pour décrire un signal mais sont insuffisantes pour permettre une reconstruction du signal analysé (*e.g.* tempo, descripteurs du timbre, fréquence fondamentale F_0, etc.).

Ainsi, ce chapitre s'intéresse à la classe des modèles dits analyse/synthèse. En effet, un modèle de cette classe permet en théorie de décrire de façon exacte un signal dont on

possède les paramètres. Une telle approche combinée avec de l'information complémentaire pourrait ainsi permettre de compenser les limitations théoriques dues au principe d'incertitude.

Dans ce chapitre, nous avons choisi de nous baser sur les travaux de McAulay et Quatieri [MQ86] ainsi que de ceux de Smith et Serra [SS87] sur la modélisation sinusoïdale appliquée aux signaux de parole et de musique. Nous ne faisons donc aucune distinction pour les transitoires et le bruit même si il existe des techniques spécifiques plus efficaces pour les traiter [SOdBB03].

4.1 La modélisation sinusoïdale des sons musicaux

Le modèle sinusoïdal implique une chaîne complète de traitement comportant l'analyse, la transformation et la synthèse des signaux traités. Ainsi, cette chaîne de traitement commune à la plupart des applications audio consiste à apprendre les paramètres du modèle d'un signal audio, à les modifier puis à synthétiser un nouveau signal en utilisant les nouveaux paramètres grâce à l'expression analytique du modèle. Le modèle sinusoïdal est particulièrement bien adapté pour représenter la partie tonale des sons musicaux qui est aussi perceptuellement la plus importante [VM00]. Ainsi la qualité perçu du signal décomposé dépend fortement de la précision des paramètres estimés, du choix et du nombre de composantes sinusoïdales.

D'après le théorème de Fourier, tout signal périodique peut être décomposé en une somme de fonctions élémentaires périodiques bien concentrées en temps et en fréquence. Dans son expression la plus générale, un signal périodique s'exprime comme une somme de sinusoïdes appelées aussi les partiels. En pratique, on se satisfait d'une approximation qui considère un nombre fini L de composantes sinusoïdales avec des paramètres variant dans le temps permettant de décrire un signal de la manière suivante :

$$\begin{aligned} x(t) &= \sum_{l=1}^{L} s_l(t) + r(t) \\ &= \sum_{l=1}^{L} a_l(t) \exp\left(\mathbf{j}\left(\omega(t) + \phi_0\right)\right) + r(t) \end{aligned} \tag{4.2}$$

où $a(t)$, $\omega(t) = 2\pi f(t)$ et ϕ_0 correspondent respectivement à l'amplitude, la fréquence et la phase à la naissance du partiel. Cette équation considère aussi le signal résiduel $r(t)$ résultant de cette approximation.

4.1.1 Extraction et estimation des composantes sinusoïdales

Le modèle sinusoïdal a la particularité d'être un modèle parcimonieux pour un grand nombre de signaux musicaux. Cela signifie qu'un son peut être représenté par un nombre relativement faible de composantes sinusoïdales. Pour cela, on privilégie les composantes de plus forte énergie détectées par un pic (maximum local) dans le spectre d'amplitude. Keiler et Marchand proposent dans [KM02] un aperçu des méthodes classiques pouvant être utilisées pour estimer les paramètres sinusoïdaux (a_0, ω_0, ϕ_0) qui se basent sur la STFT du signal observé. Certaines de ces méthodes ont été détaillées dans la section 2.1. Dans tous les cas, ces techniques permettent d'affiner en priorité l'estimation de la fréquence associée à pic du spectre d'amplitude. Ainsi, la chaîne de traitement classique, considérant un signal observé bruité, comporte une étape de détection des pics dans le spectre d'amplitude (estimation du masque d'activation du spectre). L'étape de détection

des pics est parfois associée à des techniques d'estimation de l'enveloppe [YR06] permettant ainsi de dissocier les pics correspondant au bruit d'observation. L'estimateur choisi est appliqué localement sur chaque pic indépendant en limitant le modèle au cas $L = 1$, c'est-à-dire pour un signal ne contenant qu'une seule composante sinusoïdale. Dans le cas où on utilise un modèle stationnaire (paramètres supposés constants), considérant une fenêtre d'analyse centrée à l'instant $t = 0$, le signal considéré peut être exprimé comme suit :

$$s(t) = a_0 \exp\left(\mathbf{j}\left(\phi_0 + \omega_0 t\right)\right), \tag{4.3}$$

où a_0, ω_0 et ϕ_0 sont les paramètres sinusoïdaux instantanés. En général, le modèle stationnaire est suffisant pour les sons variant lentement en fonction du temps pour des trames de signal suffisamment courtes. Un ordre plus élevé du modèle utilisant les dérivées d'ordre p de la phase et de l'amplitude permettent de prendre en compte les variations temporelles des paramètres (cas non stationnaire) et permet d'améliorer légèrement la précision de l'estimation. Des études comparatives sur le gain observé par rapport à l'ordre du modèle sinusoïdal utilisé sont proposées dans [MD08] et [GMdM$^+$03] respectivement pour l'analyse et la synthèse.

4.1.2 La méthode de réallocation

La réallocation compte actuellement parmi les meilleurs techniques d'analyse utilisées pour l'estimation des paramètres sinusoïdaux [MB10]. Cette méthode fut d'abord proposée par Kodera, Gendrin et de Villedary [KdVG76, KGdV78] puis généralisée pour les analyses temps-fréquence par Auger et Flandrin [AF95]. Ainsi, cette technique améliore considérablement la précision pour l'estimation du temps et de la fréquence des méthodes classiques basées sur la transformée de Fourier discrète à court terme. Cependant cette technique ne permet pas d'améliorer la résolution temps-fréquence qui demeure limitée par le principe d'incertitude. Le principe de cette méthode consiste (*cf.* section 2.1.1) à relocaliser ou recentrer les données du spectre sur les coordonnées TF les plus proches du support du signal analysé (*e.g.* partiel). Ainsi, cette technique se base sur l'expression analytique de la transformée de Fourier à court terme d'un signal s que l'on rappelle ici :

$$S_w(t,\omega) = \int_{-\infty}^{+\infty} s(\tau) w(\tau - t) \exp\left(-\mathbf{j}\omega(\tau - t)\right)\, d\tau. \tag{4.4}$$

Dans notre utilisation de cette technique, nous supposons l'utilisation d'une fenêtre d'analyse w, limitée en bande telle que chaque fréquence correspond à un partiel spécifique correspondant à un pic dans le spectre d'amplitude. Ainsi, nous avons choisi d'utiliser la fenêtre de Hann symétrique centrée en 0 et de durée N et définie sur $[-N/2; +N/2]$:

$$w(t) = \frac{1}{2}\left(1 + \cos\left(2\pi\frac{t}{N}\right)\right). \tag{4.5}$$

En considérant (4.4), on peut déduire :

$$\frac{\partial}{\partial t}\log\left(S_w(t,\omega)\right) = \mathbf{j}\omega - \frac{S_{w'}(t,\omega)}{S_w(t,\omega)} \tag{4.6}$$

Ainsi la fréquence estimée peut s'exprimer par :

$$\hat{\omega}(t,\omega) = \frac{\partial}{\partial t}\mathbf{Im}\,\left(\log\left(S_w(t,\omega)\right)\right) = \omega - \underbrace{\mathbf{Im}\,\left(\frac{S_{w'}(t,\omega)}{S_w(t,\omega)}\right)}_{-\Delta\omega}. \tag{4.7}$$

En pratique, un partiel correspond à un maximum local m dans le spectre d'amplitude discret localisé à la fréquence $\omega_m = 2\pi\frac{m}{N}F_s$. Ainsi, la fréquence estimée est donnée par :

$$\hat{\omega}_0 = \hat{\omega}(t, \omega_m). \tag{4.8}$$

L'amplitude et la phase $\hat{\phi}_0$ peuvent alors être estimé par les expressions suivantes :

$$\hat{a}_0 = \left| \frac{S_w(\omega_m)}{W(\Delta_\omega)} \right|, \tag{4.9}$$

$$\hat{\phi}_0 = \angle \left(\frac{S_w(\omega_m)}{W(\Delta_\omega)} \right) \tag{4.10}$$

avec $W(\omega)$ le spectre de la fenêtre de Hann w donné par :

$$W(\omega) = \int_{-\infty}^{+\infty} w(t) \exp\left(-\mathbf{j}\omega t\right)\, dt \tag{4.11}$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2}\left(1 + \cos\left(2\pi\frac{t}{N}\right)\right) \exp\left(-\mathbf{j}\omega t\right)\, dt \tag{4.12}$$

$$= \frac{1}{4}\left(2W_{R,N+1}(\omega) + W_{R,N+1}(\omega - \Omega_{F_s,N}) + W_{R,N+1}(\omega + \Omega_{F_s,N})\right) \tag{4.13}$$

telle que $W_{R,N} = \frac{\sin(N\omega/2)}{\sin(\omega/2)}$ est le spectre d'une fenêtre rectangulaire de taille N et $\Omega_{F_s,N} = 2\pi F_s/N$ avec F_s la fréquence d'échantillonnage.

Actuellement la méthode de réallocation compte parmi les meilleures techniques utilisant la transformée de Fourier à court terme en terme d'efficacité et de précision [BCRD08]. Les méthodes à haute résolution [Bad06, BDR06] sont plus coûteuses en temps de calcul et optimisent uniquement la résolution (limitée par le principe d'incertitude) mais pas la précision (limitée par la borne de Cramér-Rao).

4.1.3 Bornes théoriques

Lorsque l'on veut évaluer les performances d'un estimateur en présence de bruit, la borne de Cramér-Rao indique la variance minimale de l'erreur commise par l'estimateur optimal théorique. Ainsi, la borne de Cramér-Rao donne la meilleure performance théorique pouvant être atteinte en fonction du modèle d'observation (qui dépend ici du niveau de bruit). Pour le modèle donné par l'équation (4.3), utilisant les 3 paramètres, cette borne a été calculée par Zhou *et al.*[ZGS96]. Dans notre cas, nous considérons la version asymptotique de la borne qui suppose un nombre infini d'observations.

Djurić et Kay [DK90] ont démontré que la borne de Cramér-Rao dépend de l'échantillon temporel n_0 qui correspond à l'échantillon pour lequel les paramètres sont estimés, c'est-à-dire $t = 0$ dans l'équation (4.3). En pratique, il apparaît que le meilleur choix pour n_0 est le centre de la trame d'analyse. Ainsi la borne de Cramér-Rao s'exprime grâce à la fonction suivante dans le cas stationnaire :

$$\epsilon_k(N) = \sum_{n=0}^{N-1} \left(\frac{n - n_0}{N}\right)^k. \tag{4.14}$$

Ainsi, les bornes inférieures pour l'amplitude a, la fréquence ω et la phase ϕ sont données par [ZGS96] :

$$\mathrm{CRB}_a(a, N, \sigma) \approx \frac{\sigma^2 \epsilon_2}{2(\epsilon_0\epsilon_2 - {\epsilon_1}^2)}, \tag{4.15}$$

$$\mathrm{CRB}_{\omega}(a, N, \sigma) \approx \frac{\sigma^2 \epsilon_0}{2a_0^2 N^2 (\epsilon_0 \epsilon_2 - {\epsilon_1}^2)}, \tag{4.16}$$

$$\mathrm{CRB}_{\phi}(a, N, \sigma) \approx \frac{\sigma^2 \epsilon_2}{2a^2 N^2 (\epsilon_0 \epsilon_2 - {\epsilon_1}^2)}. \tag{4.17}$$

La précision pour l'estimation des paramètres de chaque sinusoïde est limitée par cette borne tant qu'aucune information supplémentaire n'est ajoutée. Les figures 4.2a), 4.2c) et 4.2b) (le protocole expérimental est décrit dans la section 4.2.2) comparent la borne de Cramér-Rao avec la variance de l'erreur d'estimation pour chaque paramètre sinusoïdal en utilisant la méthode de réallocation décrite précédemment. On remarque que cette borne est presque atteinte pour la méthode de réallocation utilisant la fenêtre de Hann. Le seul moyen d'obtenir une variance de l'erreur plus faible que cette borne en conservant le même modèle de signal consiste à introduire de l'information complémentaire comme décrit dans la section suivante.

4.2 L'analyse spectrale informée dans le cas scalaire

La meilleure précision pouvant être obtenue par un estimateur peut s'avérer insuffisante pour les applications les plus exigeantes (*e.g.* écoute active, effets haute qualité, etc.). Ainsi, nous proposons dans cette section d'appliquer l'approche informée (introduite dans la section 3.3.1) à l'estimation des paramètres sinusoïdaux. On se propose dans un premier temps de traiter de manière isolée chaque paramètre du modèle sinusoïdal en utilisant la méthodologie décrite dans la section 3.3.3 dans le cas scalaire.

4.2.1 Principes

L'approche que nous proposons ici (*cf.* figure 3.9) consiste à effectuer une analyse en deux étapes dans une configuration codeur / décodeur qui se situent respectivement avant et après la création du mélange. Au codeur, où l'on dispose des signaux d'origine avant le mixage, on extrait de l'information en utilisant la connaissance *a priori* sur les erreurs d'estimation qui seront commises pour retrouver les signaux d'origine à partir du mélange. Au décodeur (après le processus de mixage), le même estimateur est appliqué sur le signal de mélange altéré où les erreurs d'estimation sont corrigées à partir des informations complémentaires codées.

4.2.2 Simulations

Dans cette expérimentation on souhaite comparer la précision de la méthode de réallocation dans sa version classique (non informée) et informée respectivement avec la borne de Cramér-Rao. Pour cela, on considère un signal discret s échantillonné à $F_s = 44.1$kHz et composé d'un seul partiel ($L = 1$) synthétisée d'après l'équation (4.2) en utilisant une amplitude fixée $a_0 = 1$. La fréquence ω et de phase ϕ sont sélectionnées uniformément en générant toutes les combinaisons possibles.

Le signal s est mélangé avec un bruit blanc gaussien de variance σ^2 connue. Ainsi, le SNR de référence est donné en décibels (dB) par :

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{a_0^2}{\sigma^2} \right). \tag{4.18}$$

Pour rendre les paramètres indépendants de la fréquence d'échantillonnage, ω est normalisée (par F_s). La longueur des trames d'analyse est de taille impaire et comporte

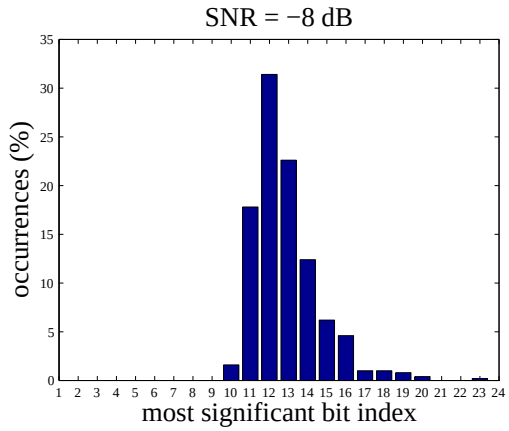
$N = 2H + 1 = 513$ échantillons (la durée en secondes est donnée par $T = N/F_s$). En pratique, l'estimation des paramètres est effectuée au centre de chaque trame de taille N à l'instant $t = 0$ (échantillon n_0). Les indices de chaque échantillons varient entre $-H$ et $+H$ et les calculs sont effectués à l'aide de la transformée de Fourier rapide, *Fast Fourier Transform* (FFT) basé d'après l'équation 4.4 où l'intégrale continue est transformée en somme discrète sur N valeurs.

Sur la figure 4.1, on observerve la valeur du MSB de la représentation binaire de la valeur absolue de l'erreur d'estimation de la fréquence ($\mathcal{C}_d(\frac{|\omega-\hat{\omega}|}{\pi F_s})$) où $\mathcal{C}_d$ est l'application de codage binaire sur d bits, telle que définie dans la section 4.2. Dans cet expérimentation, chacun des paramètres est quantifié uniformément sur $d = 16$bits. Ainsi, on constate que le MSB (associé à I_σ) varie en fonction du SNR. Un SNR élevé correspond aussi à une valeur de I_σ plus grande. Cela a pour effet de réduire la quantité d'information nécessaire permettant d'atteindre la précision maximale fixée par d.

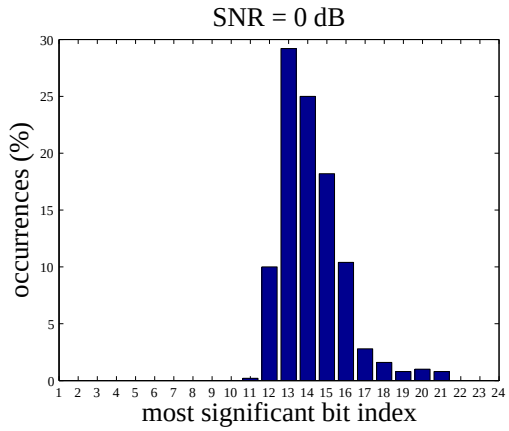
Sur les figures 4.2 et 4.3, les variances d'erreur obtenues sont comparées à la borne inférieure d'information ou ILB (*Informed Lower Bound*) qui suppose que chaque bit transmis permet de diviser l'erreur par 2 (et donc la variance de l'erreur par 4). Ainsi, cette borne peut s'exprimer comme une fonction de la borne de Cramér-Rao et du nombre i de bits informants :

$$\mathrm{ILB}(i) = \mathrm{CRB} \cdot 2^{-2i}. \tag{4.19}$$

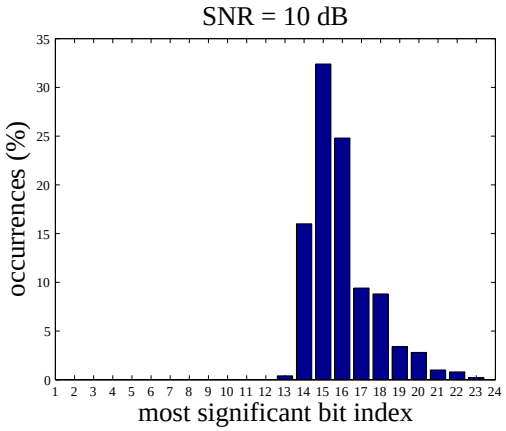
Dans le cas de la figure 4.2, cette borne a été modifiée pour prendre en compte la précision maximale donnée en fonction de d. Lorsque la borne est atteinte, ou lorsque la quantification obtient une erreur supérieure à l'estimation, il n'est plus nécessaire de transmettre de l'information supplémentaire. Dans ce cas, seule l'estimation classique (non informée) est prise en compte (cela explique pourquoi les droites ILB et CRB ne sont plus parallèles). On remarque qu'en pratique cette nouvelle borne ILB n'est pas atteinte, cela s'explique en raison de l'hypothèse "optimiste" qui suppose que chaque bit permet de réduire l'erreur d'estimation. Cependant en pratique, il arrive lors de l'application de l'algorithme 1 que certains bits substitués aient la même valeur que le bit d'origine dans la représentation binaire de l'estimation. Cela a pour conséquence d'annuler l'effet du bit pour la réduction de l'erreur résultante.



a)

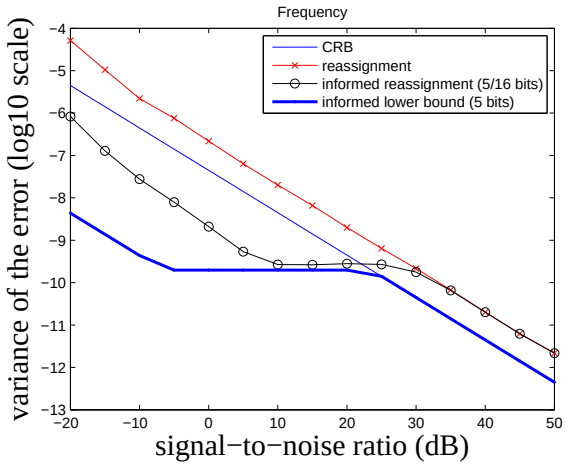


b)

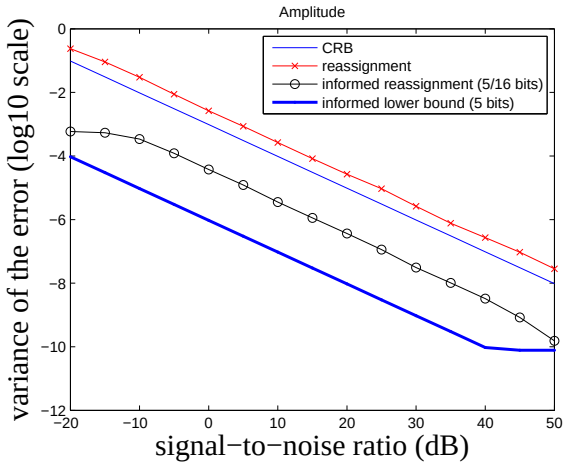


c)

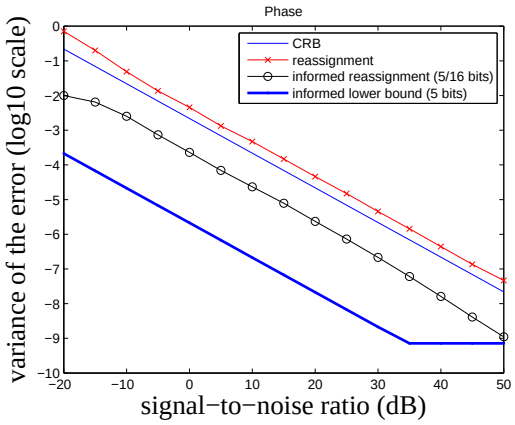
FIGURE 4.1 – Histogramme du MSB (bit de poids le plus fort) de la représentation binaire de la valeur absolue de l'erreur d'estimation utilisant la méthode de réallocation appliquée au signal mélangé avec un bruit blanc additif. Le premier indice non nul augmente en même temps que le SNR, le nombre de bits à corriger pour retrouver le paramètre de référence est donc plus faible lorsque le SNR augmente.



a) frequency

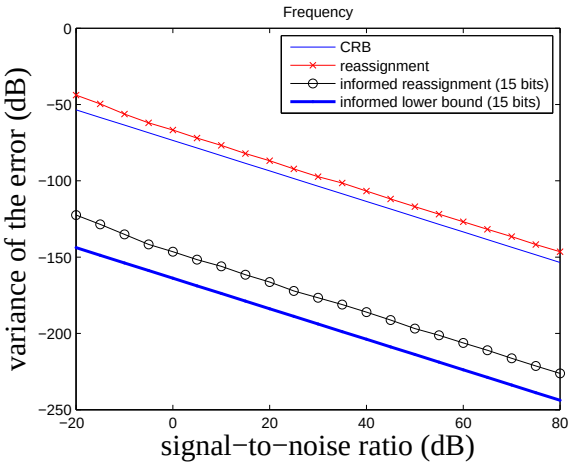


b) amplitude

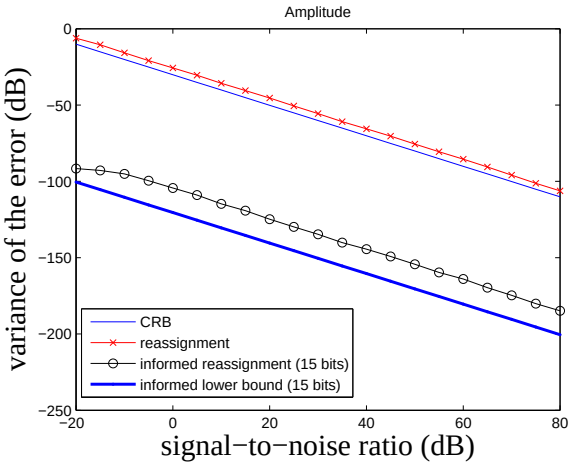


c) phase

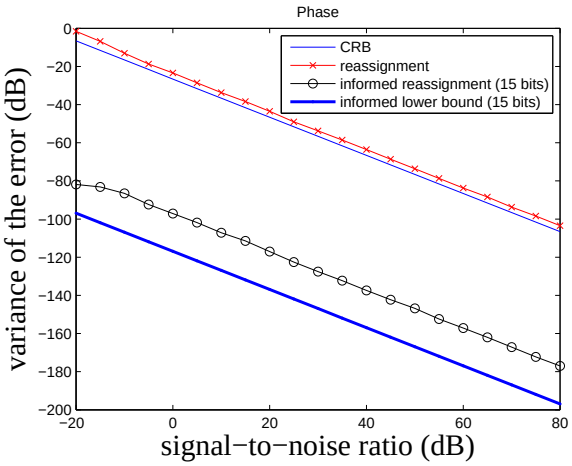
FIGURE 4.2 – Variance de l'erreur obtenue en fonction du SNR d'un signal composé d'une seule sinusoïde mélangée dans du bruit pour l'estimation de la fréquence 4.2a), de l'amplitude 4.2b) et de la phase 4.2c). La variance obtenue reste toujours supérieure à la borne de Cramér-Rao sauf dans le cas informé où on informe 5 bits pour chaque paramètre quantifié sur 16 bits.



a) frequency



b) amplitude



c) phase

FIGURE 4.3 – Variance de l'erreur obtenue en fonction du SNR d'un signal composé d'une seule sinusoïde mélangée dans du bruit pour l'estimation de la fréquence 4.3a), de l'amplitude 4.3b) et de la phase 4.3c). La variance obtenue reste toujours supérieure à la borne de Cramér-Rao sauf dans le cas informé où on informe 15 bits pour chaque paramètre quantifié.

4.3 Analyse spectrale informée dans le cas vectoriel

Dans le cas des sinusoïdes utilisant le modèle stationnaire, on doit estimer un triplet de paramètres $P = (a, \omega, \phi)$ qui est un vecteur de $\mathbb{R}^3$. Comme a, ω et ϕ correspondent à des grandeurs physiques différentes, la quantification optimale du vecteur P pour une fonction de distorsion et un budget global de d bits permet d'obtenir une précision relative différente sur chacune des composantes. Le rôle de l'analyse informée consiste répartir de manière optimale l'information complémentaire sur chacune des composantes du vecteur P.

4.3.1 Principes

La quantification optimale de P est obtenue grâce à la quantification sphérique avec dépendance des paramètres et contrainte d'entropie, *Entropy Constrained Unrestricted Spherical Quantization* (ECUSQ) [KJH07] décrite en détail dans la section 4.3.2. L'optimalité en terme débit-distorsion a été montrée pour cette approche en utilisant comme mesure de distorsion l'erreur quadratique moyenne pondérée, *Weighted Mean Squared Error* (WMSE) entre les signaux synthétisés d'après le modèle (4.3). Le budget global d de bits affectés au vecteur $P = (a, \omega, \phi)$ est désormais variable et dépend désormais d'une mesure d'entropie cible constante notée H_t (*Target Entropy*).

Il en résulte une valeur de d (variable) dépendant de la valeur des composantes de P et de H_t (fixe). Ainsi dans le cas où $a \approx 0$, il n'est pas nécessaire de coder la phase ni la fréquence, d'où une valeur de d plus faible. La fonction qui retourne le nombre de bits alloués à chaque composante de P est $\lceil \log_2 \gamma \rceil$ où γ est la fonction de densité calculée par la méthode ECUSQ.

La méthode d'analyse spectrale informée proposée dans la section précédent est appliquée séparément sur chaque composante du vecteur P de la même façon que dans le cas scalaire (*cf.* section 4.2). L'application de codage devient alors $\mathcal{C}_d : [0;1[^3 \rightarrow \{0,1\}^d$ en utilisant une concaténation simple telle que :

$$\mathcal{C}_d(P) = \big(\mathcal{C}_e(a), \mathcal{C}_f(\omega), \mathcal{C}_g(\phi)\big) \text{ avec } e + f + g = d. \tag{4.20}$$

Ainsi, l'information supplémentaire considérée est $\mathcal{I} = \big(\mathcal{I}_a, \mathcal{I}_\omega, \mathcal{I}_\phi\big)$. Pour le décodage, la précision relative de chaque paramètre est requise (e, f, g) afin de pouvoir appliquer la correction d'erreur. Cependant, comme le quantificateur optimal pour les paramètres ϕ et ω dépend de la valeur de l'amplitude. Ainsi, $\tilde{a}$ est calculé dans un premier temps grâce à $\mathcal{I}_a$. En suite, $\tilde{\omega}$ et $\tilde{\phi}$ sont calculés respectivement en utilisant $\mathcal{I}_\omega$ et $\mathcal{I}_\phi$ grâce à f et g donnés par ECUSQ en fonction de $\tilde{a}$. Nous détaillons les calculs de la technique ECUSQ dans la section suivante.

4.3.2 Quantification optimale

D'après Shannon [Gra89], la fonction débit-distorsion appelée aussi borne inférieure de Shannon (*Shannon Lower Bound*) définit la distorsion minimale pouvant être atteinte pour coder une source (on parle ici d'une variable aléatoire). Dans le cas d'une composante sinusoïdale à 3 paramètres, le calcul détaillé de la fonction débit-distorsion pourra être trouvé en annexe section A.2. En pratique, la borne de Shannon n'est jamais atteinte car il s'agit d'un résultat asymptotique donné pour des données supposées de taille infinie. De plus, cette borne ne donne pas d'indication sur la manière de construire un quantificateur optimal permettant d'atteindre cette borne (même problème que pour le codage). En pratique on préférera utiliser la distorsion théorique donnée par ECUSQ qui donne de bonnes performances de quantification pour une plus faible complexité calculatoire.

En effet le quantificateur est donné directement par une expression analytique contrairement à [CLG89] qui est itératif. Cet aspect est important pour permettre de traiter dans un temps raisonnable un nombre important de composantes sinusoïdales présentes dans un signal.

Pour la construction du quantificateur ECUSQ, nous définissons d'abord la distorsion moyenne D qui correspond à l'espérance d'une fonction de distorsion notée ∂ calculée à partir du signal analytique correspondant aux paramètres de référence et les paramètres quantifiés tels que :

$$D = E[\delta(s, \hat{s})]. \tag{4.21}$$

D peut se calculer d'après le modèle sinusoïdal défini par (4.3), en utilisant l'erreur quadratique moyenne pondérée (*Weighted Mean Square Error*). Ainsi, pour un signal discret de taille N, la fonction de distorsion peut s'exprimer par :

$$\begin{aligned}
\delta(s, \tilde{s}) =& \delta(a, \omega, \phi, \hat{a}, \hat{\omega}, \hat{\phi}) \\
=& \sum_{n=\nu}^{\nu+N-1} \left| w[n] \left(s[n] - \tilde{s}[n] \right) \right|^2 \\
=& \sum_{n=\nu}^{\nu+N-1} \left| w[n] \left(a e^{\mathbf{j}(\omega n + \phi)} - \tilde{a} e^{\mathbf{j}(\tilde{\omega} n + \tilde{\phi})} \right) \right|^2 \\
=& ||w||^2 \underbrace{(a^2 + \tilde{a}^2)}_{\Delta_a^2 + 2a\tilde{a}} - 2a\tilde{a} \sum_{n=\nu}^{\nu+N-1} w[n]^2 \cos\left(\underbrace{(\omega - \tilde{\omega})}_{\Delta_\omega} n + \underbrace{(\phi - \tilde{\phi})}_{\Delta_\phi} \right),
\end{aligned} \tag{4.22}$$

avec $||w||^2 = \sum_{n=\nu}^{\nu+N-1} w[n]^2$, $n = \nu, \ldots, \nu + N - 1$. Ici w correspond à la fenêtre d'analyse (*i.e.* fenêtre de Hann) permettant de définir une trame du signal. D'après [KJH07], la valeur optimale de ν permettant de minimiser la distorsion (4.21) est $\nu = -(N-1)/2$.

En utilisant $\sigma^2 = \frac{1}{||w||^2} \sum_{n=\nu}^{\nu+N-1} w[n]^2 n^2$, le développement limité de la fonction cos ainsi que l'approximation $a\tilde{a} \approx \tilde{a}^2$, l'équation (4.22) peut être approximée par (*cf.* détails dans l'annexe A.2.2) :

$$\delta(a, \omega, \phi, \tilde{a}, \tilde{\omega}, \tilde{\phi}) \approx ||w||^2 \left(\Delta_a^2 + \tilde{a}^2 (\Delta_\phi^2 + \sigma^2 \Delta_\omega^2) \right). \tag{4.23}$$

Ainsi, la distorsion moyenne $\bar{\delta}$ calculée pour chaque cellule de quantification de taille Δ_a, Δ_ω, Δ_ϕ peut être déduite de (4.23) en calculant l'espérance à partir de la fonction de la densité de probabilité d'un triplet (a, ω, ϕ). Chaque triplet est donc considéré comme une réalisation de 3 variables aléatoires notées respectivement A, Ω et Φ. Ainsi nous obtenons :

$$\bar{\delta}(\tilde{a}, \tilde{\omega}, \tilde{\phi}, \Delta_a, \Delta_\omega, \Delta_\phi) = \frac{\iiint f_{A,\Omega,\Phi}(a, \omega, \phi) \delta(a, \omega, \phi, \tilde{a}, \tilde{\omega}, \tilde{\phi}) \, da \, d\omega \, d\phi}{\iiint f_{A,\Omega,\Phi}(a, \omega, \phi) \, da \, d\omega \, d\phi}, \tag{4.24}$$

avec $f_{A,\Omega,\Phi}(a, \omega, \phi)$ la fonction de densité de probabilité jointe de chaque composante du vecteur P.

En utilisant l'hypothèse de haute résolution, $f_{A,\Omega,\Phi}(a, \omega, \phi)$ est supposé constant sur chaque cellule de quantification. Ainsi, le représentant de chaque cellule est situé au

centre de l'intervalle correspondant sur chaque dimension. Cela permet d'approximer $\bar{\delta}$ en utilisant le résultat précédent (4.23) par :

$$\begin{aligned}\bar{\delta}(\tilde{a},\Delta_a,\Delta_\omega,\Delta_\phi) \approx & ||w||^2 \left(2\int_0^{\Delta_a/2}\frac{1}{\Delta_a}x^2\,dx + 2\tilde{a}^2\sigma^2\int_0^{\Delta_\omega/2}\frac{1}{\Delta_\omega}y^2\,dy + 2\tilde{a}^2\int_0^{\Delta_\phi/2}\frac{1}{\Delta_\phi}z^2\,dz\right)\\ \approx & ||w||^2\left(\frac{2}{\Delta_a}\frac{\Delta_a^3}{24} + \tilde{a}^2\sigma^2\frac{2}{\Delta_\omega}\frac{\Delta_\omega^3}{24} + \tilde{a}^2\frac{2}{\Delta_\phi}\frac{\Delta_\phi^3}{24}\right)\\ \approx & \frac{||w||^2}{12}(\Delta_a^2 + \tilde{a}^2(\sigma^2\Delta_\omega^2 + \Delta_\phi^2)).\end{aligned} \tag{4.25}$$

La distorsion moyenne D globale est obtenue en calculant l'espérance sur l'ensemble des cellules de quantification d'indices $\iota_a, \iota_\omega, \iota_\phi$ associés aux alphabets I_a, I_ω, I_ϕ :

$$\begin{aligned}D &= \sum_{\iota_a\in I_a}\sum_{\iota_\omega\in I_\omega}\sum_{\iota_\phi\in I_\phi} p_{I_aI_\omega I_\phi}(\iota_a,\iota_\omega,\iota_\phi)\bar{\delta}(\tilde{a},\Delta_a,\Delta_\omega,\Delta_\phi)\\ &\approx \frac{||w^2||}{12}\iiint f_{A,\Omega,\Phi}(a,\omega,\phi)\left(\gamma_A^{-2}(a,\omega,\phi) + a^2(\gamma_\Phi^{-2}(a,\omega,\phi) + \sigma^2\gamma_\Omega^{-2}(a,\omega,\phi))\right)\,da\,d\omega\,d\phi,\end{aligned} \tag{4.26}$$

où $p_{I_aI_\omega I_\phi}(\iota_a,\iota_\omega,\iota_\phi)$ est la probabilité de la cellule d'indices $\iota_a,\iota_\omega,\iota_\phi$ et $\gamma = \Delta^{-1}$ correspond à la fonction de densité qui donne le nombre total de cellules de quantifications lorsqu'elle est intégrée sur un intervalle [GN98, Llo82]. Cette équation peut à nouveau être approximée en utilisant l'hypothèse haute résolution à partir de $f_{A,\Omega,\Phi}$ [KJH07] pour obtenir (4.26).

Maintenant, nous souhaitons définir la fonction de densité qui minimise D pour une entropie cible donnée H_t. Ici H_t désigne l'entropie de Shannon qui est l'information moyenne et permettant en théorie de coder de façon optimale une composante sinusoïdale. Ce débit pourrait éventuellement être approché en utilisant une technique de codage entropique [Say06] (*cf.* section 3.2.3).

En utilisant l'hypothèse de haute résolution pour la quantification, qui suppose que l'erreur de quantification est répartie uniformément sur chaque cellule, on peut approximer l'expression de l'entropie jointe :

$$\begin{aligned}H_t \approx & h(A,\Omega,\Phi)\\ &+ \iiint f_{A,\Omega,\Phi}(a,\omega,\phi)\log_2(\gamma_A(a,\omega,\phi))\,da\,d\omega\,d\phi\\ &+ \iiint f_{A,\Omega,\Phi}(a,\omega,\phi)\log_2(\gamma_\Omega(a,\omega,\phi))\,da\,d\omega\,d\phi\\ &+ \iiint f_{A,\Omega,\Phi}(a,\omega,\phi)\log_2(\gamma_\Phi(a,\omega,\phi))\,da\,d\omega\,d\phi.\end{aligned} \tag{4.27}$$

En fixant $\tilde{H}_t = H_t - h(A,\Omega,\Phi)$, ce problème d'optimisation peut être résolu en appliquant la méthode des multiplicateurs de Lagrange :

$$J = D + \lambda\tilde{H}_t.$$

D'après [KJH07] on obtient :

$$\gamma_A(a,\phi,\omega) = \left(\frac{||w||^2}{6\lambda \log_2(\mathbf{e})}\right)^{\frac{1}{2}}, \tag{4.28}$$

$$\gamma_\Phi(a,\phi,\omega) = a\gamma_A(a,\phi,\omega), \tag{4.29}$$

$$\gamma_\Omega(a,\phi,\omega) = a\sigma\gamma_A(a,\phi,\omega) \tag{4.30}$$

avec :

$$\lambda = \frac{||w||^2 2^{\frac{-2}{3}(\tilde{H}_t - 2b(A) - \log_2(\sigma))}}{6\log_2(\mathbf{e})}, \tag{4.31}$$

avec $b(A) = \int f_A(a)\log_2(a)\ da$, on déduit alors :

$$\gamma_A(a,\phi,\omega) = 2^{\frac{1}{3}(\tilde{H}_t - 2b(A) - \log_2(\sigma))}, \tag{4.32}$$

$$\gamma_\Phi(a,\phi,\omega) = a2^{\frac{1}{3}(\tilde{H}_t - 2b(A) - \log_2(\sigma))}, \tag{4.33}$$

$$\gamma_\Omega(a,\phi,\omega) = a\sigma 2^{\frac{1}{3}(\tilde{H}_t - 2b(A) - \log_2(\sigma))}. \tag{4.34}$$

En substituant (4.32), (4.33) et (4.34) dans (4.26), la distorsion théorique minimale donnée par le quantificateur est donnée par :

$$D_{\text{ECUSQ}} = \frac{||w||^2}{4} 2^{-\frac{2}{3}(\tilde{H}_t - 2b(A) - \log_2(\sigma))}. \tag{4.35}$$

Cette mesure est utilisée dans nos expérimentations décrites dans la section 4.4 et est comparée la borne de Shannon dont les calculs sont décrit en annexe A.2.

4.3.3 Simulations

Afin de valider l'efficacité de notre approche, 10000 signaux aléatoires discrets composés d'un seul partiel qui est synthétisé d'après (4.2), en combinaison avec un bruit blanc gaussien additif.

La variance du bruit est fixée de manière à obtenir un SNR en décibels (dB) connu compris dans l'intervalle $[-20;50]$. Les paramètres d'amplitude a et de fréquence ω des signaux simulés sont générés d'après une loi de Rayleigh de paramètre $\sigma_a = 0.2$ et $\sigma_\omega = \pi/11$. La phase ϕ est générée à partir d'une loi uniforme $U(0,2\pi)$. Ces paramètres et ces lois ont été choisis car ils sont relativement proches de ceux observés dans les sons naturels, de plus ce sont les mêmes que ceux proposés dans [KJH07].

Pour l'analyse spectrale, nous utilisons une fenêtre de Hann de taille impaire $N = 1023$ pour laquelle l'estimation s'effectue sur l'échantillon central. L'entropie cible H_t est calculée d'après [KJH07] pour un SNR cible donné (ici nous avons choisi les valeurs de 45dB et de 100dB). Pour cela nous utilisons l'expression suivante obtenue à partir de (4.35) :

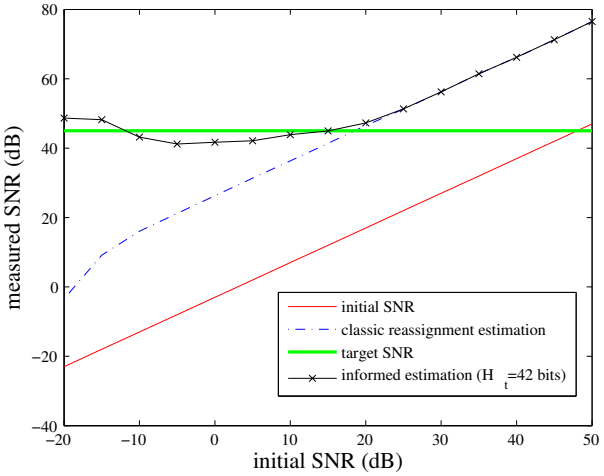
$$H_t = -\frac{3}{2}\log_2\left(\frac{4D^{\text{cible}}}{||w||^2}\right) + h(A,\Phi,\Omega) + 2b(A) + \log_2(\sigma) \tag{4.36}$$

avec D^{cible} la distorsion moyenne cible déduite du SNR moyen cible. D^{cible} est approximé à partir de l'énergie du signal théorique dépendant des paramètres simulés. I_σ est estimé en utilisant la connaissance du SNR initial quantifié uniformément avec 4 bits sur l'intervalle $[-20;\text{SNR}^{\text{cible}}]$. Ainsi, l'équation 4.36 permet de de retrouver l'entropie cible H_t permettant de calculer le quantificateur ECUSQ correspondant.

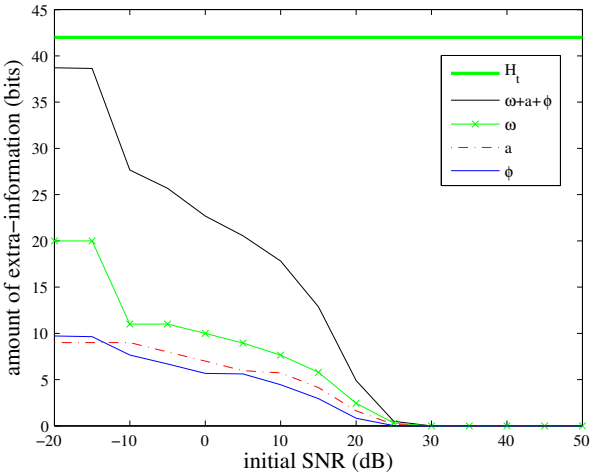
Les figures 4.4a) et 4.5a) présentent le SNR atteint en utilisant l'analyse spectrale informée et les figures 4.4b) et 4.5b) présentent la quantité d'information moyenne utilisée ainsi que sa répartition sur chacune des composantes du vecteur (a, ω, ϕ). Ainsi, la quantité d'information transmise est proportionnelle au taux d'erreur de l'estimateur utilisé (ici il s'agit de la méthode de réallocation).

Le SNR obtenu en pratique reste très proche du SNR cible tandis que le nombre de bits alloués par paramètre décroît lorsque le SNR initial croît. Vu que le SNR cible et l'entropie cible restent constants pour un débit décroissant, l'information manquante fournie par l'estimateur est donc correctement exploitée.

Quand le SNR cible est atteint, la quantité d'information (instantanée) transmise est nulle. On remarque sur la figure 4.4b) que le débit moyen nul ne correspond pas exactement à l'intersection entre le SNR cible et le SNR de la figure 4.4a). Cela s'explique parce que ces figures représentent des valeurs moyennes, en effet même pour un SNR initial suffisant, l'estimateur peut parfois obtenir des erreurs instantanées nécessitant une correction pour atteindre le SNR cible ce qui a pour effet d'augmenter le débit moyen correspondant. Cette expérimentation démontre que la technique proposée permet d'atteindre n'importe quelle précision souhaitée avec un estimateur existant en le combinant avec l'information complémentaire de taille minimale.

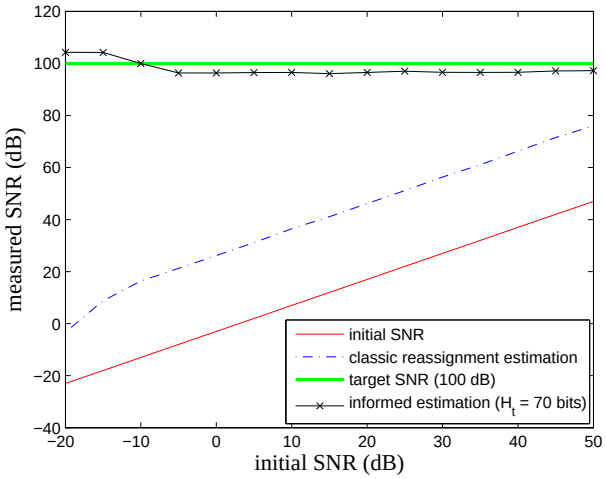


a) SNR moyen résultant

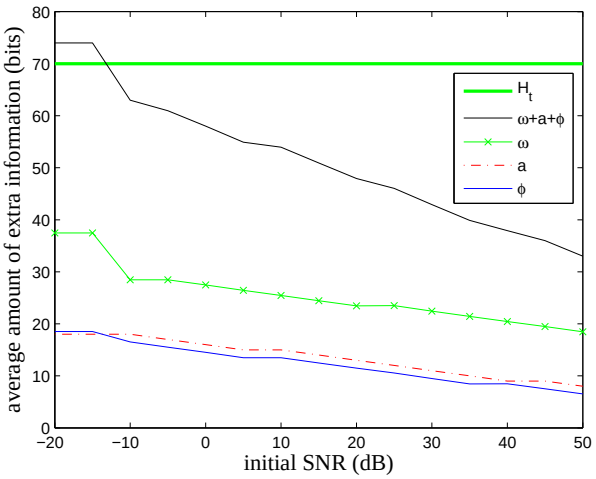


b) allocation des bits aux paramètres sinusoïdaux

FIGURE 4.4 – Comparaison du SNR moyen obtenu 4.4a) entre un estimateur classique et le même estimateur informé. 4.4b) montre la répartition du débit entre les paramètres sinusoïdaux. Ici le SNR cible est fixé à 45dB.



a) SNR moyen résultant

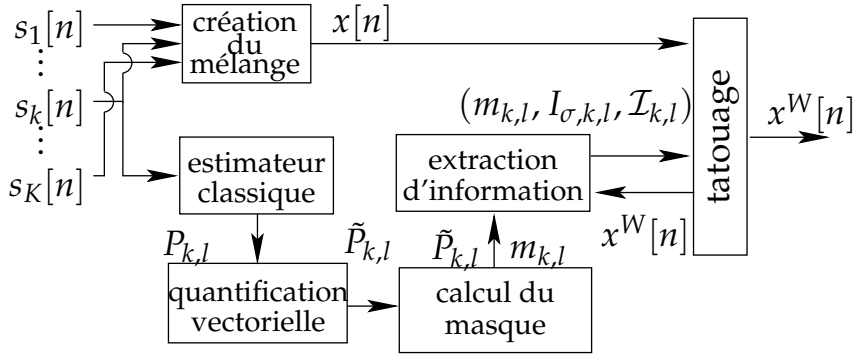


b) allocation des bits aux paramètres sinusoïdaux

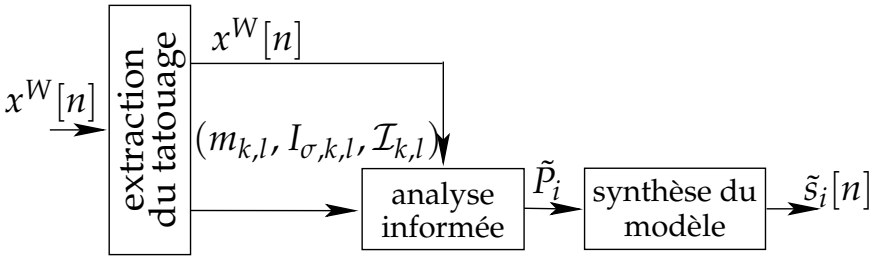
FIGURE 4.5 – Comparaison du SNR moyen obtenu 4.5a) entre un estimateur classique et le même estimateur informé. 4.5b) montre la répartition du débit entre les paramètres sinusoïdaux. Ici le SNR cible est fixé à 100dB.

4.4 Application à la séparation de sources informée

Dans cette section, nous proposons un système basé sur l'analyse informée permettant d'estimer les paramètres sinusoïdaux de chaque source isolée présente dans un mélange. A partir de leur modèle, il est alors aisé de resynthétiser les signaux correspondant à chaque source pour les retrouver. Dans ce système, nous nous plaçons dans une configuration de type codeur / décodeur combinée à une technique de tatouage audio numérique telle que décrite par la figure 4.6.



a) codeur



b) décodeur

FIGURE 4.6 – Structure du système proposé pour l'estimation des paramètres sinusoïdaux de chaque source séparée à partir d'un mélange mono-canal.

On suppose les signaux originaux correspondant à chaque source $s_k[n]$ connus au codeur. Les paramètres de référence de chaque source notés P_k sont estimés à partir des signaux $s_k[n]$ isolés en utilisant un estimateur classique (non informé) avant la création du mélange. L'information nécessaire permettant de retrouver P_k à partir de $x[n]$ en utilisant le même estimateur est calculée et cachée de façon inaudible dans le mélange en utilisant une technique de tatouage [PGBP10]. Au décodeur, l'information cachée est extraite et combinée avec l'estimateur classique comme proposé dans la section 4.3 pour l'analyse informée.

4.4.1 Modèle de source et estimation des paramètres

Nous considérons à présent un mélange mono-canal discret composé de K sources exprimé comme suit :

$$x[n] = \sum_{k=1}^{K} s_k[n] + r[n], \tag{4.37}$$

avec $r[n]$ un signal résiduel considéré comme un bruit additif. Les signaux source $s_k[n]$ sont décomposés comme une somme de L sinusoïdes réelles. D'après [MQ86], chaque

source décomposée en utilisant la modélisation sinusoïdale peut s'écrire comme :

$$s_k[n] = \sum_{l=1}^{L} a_l \cos\left(\omega_l n + \phi_l\right), \tag{4.38}$$

ce qui correspond à la partie réelle de (4.3) où a, ω, et ϕ sont respectivement l'amplitude, la fréquence et la phase supposées localement constantes.

Pour l'estimation de ces paramètres, on choisit d'utiliser l'estimateur de réallocation décrit dans la section 4.1.2. Comme discuté dans [MF10], tout estimateur pourrait convenir. Un estimateur (efficace) atteignant la borne de Cramér-Rao aura donc pour effet de réduire le débit utilisé pour l'analyse informée.

4.4.2 Calcul et codage du masque pour chaque source

Comme expliqué dans la section 4.1.1, les méthodes d'analyse basées sur l'analyse de Fourier à court terme sont précédées d'une étape de sélection des pics dans le spectre d'amplitude afin de trouver l'indice temps-fréquence correspondant à chaque composante sinusoïdale présente dans le mélange. De plus, chacune des composantes est affectée à une où plusieurs sources difficiles à identifier en utilisant la connaissance seule d'un mélange mono-canal.

Ainsi le masque temps-fréquence d'activation noté $m_{k,l}$ de chaque source doit être transmis avant l'estimation des paramètres à partir du spectre. Pour cela, nous proposons d'utiliser une technique de suivi de partiels basée sur la modélisation sinusoïdale à long terme (*cf.* section 2.1.4) reposant sur la prédiction [LMR04b]. Cette technique associe chaque partiel instantané à une trajectoire temporelle couvrant plusieurs trames du signal. Ainsi, les composantes sinusoïdales les plus proches situées sur des trames voisines successives sont liées pour former une trajectoire. Par exemple, pour une composante l situé à la trame $\mathring{t}$, on recherche à la trame $\mathring{t}+1$ la composante sinusoïdale ayant les paramètres les plus proches de ceux prédits depuis la trame $\mathring{t}$ pour la composante l. En pratique, on recherche la composante sinusoïdale de la trame suivante qui minimise une fonction de distance (*i.e.* fonction de distorsion donnée par (4.22)). Dans notre implémentation nous utilisons un prédicteur simplifié (pour des raisons d'efficacité) qui suppose l'amplitude et la fréquence de chaque composante sinusoïdale constantes entre deux trames de signal successives. La phase de la trame suivante est alors extrapolée à partir de la fréquence et du temps séparant deux trames successives.

En utilisant notre algorithme de suivi de partiel, une nouvelle trajectoire de partiel est créée (naissance du partiel) quand on détecte un partiel isolé (non associé à une trajectoire existante). On termine une trajectoire (mort du partiel) lorsqu'il n'est plus possible d'associer un nouveau partiel instantané à une trajectoire existante. Cependant, comme l'application de l'algorithme de suivi de partiels donne des résultats différents lorsqu'il est appliqué depuis les signaux des sources isolées s_k ou sur le signal de mélange x, l'information sur les trajectoires de référence doit être transmise. Or, comme l'analyse informée garantit que les paramètres estimés au décodeur sont les même que ceux utilisés au codeur, il n'est pas nécessaire de transmettre en détail toute la trajectoire de chaque partiel. En effet, il suffit de coder chaque trajectoire par un triplet $(\mathring{k}, \alpha, \beta)$ où $\mathring{k}$ correspond à l'indice STFT de la première trame de la trajectoire du partiel considéré (fréquence à la naissance du partiel). Les paramètres α et β permettent de coder les numéros des trames associées respectivement à la naissance et à la mort du partiel considéré. Ainsi la trajectoire de chaque partiel est retrouvée en combinant le prédicteur de l'algorithme de suivi avec les paramètres des composantes sinusoïdales.

Cette stratégie permet un codage efficace du masque qui dépend du nombre de trajectoires codées et non plus du nombre total de composantes sinusoïdales. En effet, il n'est pas nécessaire de transmettre l'indice STFT à chaque instant sur toute la durée de la trajectoire du partiel puisque celle-ci est retrouvée à partir des paramètres sinusoïdaux estimés. Ainsi, chaque indice de fréquence $\mathring{k}$ peut être codé avec au plus $\lceil \log_2(N/2) \rceil$ bits où N correspond à la taille de la STFT. Chaque numéro de trame peut être codé avec $\lceil \log_2(T) \rceil$ où T est le nombre total de trames du signal. Dans nos expérimentations, nous avons utilisé $N = 1024$ avec un chevauchement de trame de 50% et une fréquence d'échantillonnage $F_s = 44.1$kHz. Comme le montre la figure 4.9, le débit correspondant permettant de coder le masque de chaque source est négligeable comparé au débit total utilisé.

4.4.3 Tatouage audio numérique

Pour cacher les informations de façon inaudible dans le mélange, nous utilisons la technique de tatouage décrite dans [PGBP10]. Cette méthode est basée sur la quantification à modulation d'indice appelée aussi *Quantization Index Modulation* (QIM) [CW01] utilise dans ce cas la transformée en cosinus discrète modifiée, *Modified Discrete Cosinus Transform* (MDCT). Ainsi, les coefficients MDCT du signal à tatouer sont quantifiés et l'information est cachée en choisissant comme valeur le représentant le plus proche d'un quantificateur Q_i. Ainsi, le choix du quantificateur le plus proche de la valeur représentée permet ainsi le codage de l'information. Une description plus détaillée de la technique de tatouage QIM se trouve en annexe dans la section A.1.1.

Les performances de tatouage sont fixées par une contrainte d'inaudibilité qui définit un nombre maximal de quantificateurs utilisables pour chaque coefficient $X(\omega, t)$. Ainsi l'erreur maximale de quantification est contrainte à rester sous un seuil de masquage qui est calculé à partir d'un modèle psychoacoustique utilisé en compression audio avec pertes [PS00, BB97]. Cette technique de tatouage a été choisie en raison de ses performances en terme de débit souvent supérieur à 200kbits/s par canal pour la plupart des signaux musicaux de styles variés. De plus, cette méthode offre une certaine flexibilité permettant de paramétrer le débit en fonction de l'audibilité ou de la probabilité d'obtenir des erreurs d'extraction au décodeur. Dans nos expérimentations, nous fixons la contrainte de pouvoir extraire l'information cachée du mélange systématiquement sans aucune erreur. Cela a pour effet de réduire la capacité de tatouage des signaux de mélange traités. Compte tenu de sa capacité élevée, cette méthode n'est pas robuste aux attaques, notamment lors de l'application d'un codage compressé avec pertes. Ainsi, cette méthode ne peut être utilisé qu'avec des formats audio sans perte tels que FLAC, AIFF ou WAVE.

4.4.4 Implémentation

Le système complet décrit dans les figures 4.6a) et 4.6b) peut être implémenté à l'aide des algorithmes 2 et 3, qui correspondent respectivement au codeur et au décodeur. Les résultats obtenus avec notre implémentation sont présentés dans la section 4.4.6.

Algorithme 2 Codeur

données : $s_k[n]$: signaux source originaux
sorties : $x^W[n]$: mélange tatoué

- Estimer $P_{k,l}$ à partir de $s_k[n]$ avec la méthode de réallocation (*cf.* section 4.1.2).
- Calculer les paramètre quantifiés $\tilde{P}_{k,l}$ en utilisant la méthode ECUSQ (*cf.* section 4.3.2).
- Calculer le masque d'activation $m_{k,l}$ à partir de $\tilde{P}_{k,l}$ en utilisant la modélisation sinusoïdale à long terme (*cf.* section 4.4.2).
- Estimer $I_{\sigma,k,l}$ et $\mathcal{I}_{k,l}$ à partir $\hat{P}_{k,l}$ grâce à l'analyse informée dans le cas vectoriel (*cf.* section 4.3) en simulant le processus de mélange d'après (4.37) et en ajoutant le tatouage (*cf.* section 4.4.3).
- Calculer le mélange $x^W[n]$ final en utilisant la technique de tatouage [PGBP10] contenant $(m_{k,l}, I_{\sigma,k,l}, \mathcal{I}_{k,l})$.

Algorithme 3 Décodeur

données : $x^W[n]$: mélange tatoué
sorties : $\tilde{s}_k[n]$, $\tilde{P}_{k,l}$: signaux et paramètres sinusoïdaux des sources isolées estimés

- Retrouver $(m_{k,l}, I_{\sigma,k,l}, \mathcal{I}_{k,l})$ par le décodage du tatouage provenant de $x^W[n]$ en utilisant [PGBP10].
- Estimer $\hat{P}_{k,l}$ grâce à $m_{k,l}$ en utilisant la méthode de réallocation (*cf.* section 4.1.2).
- Calculer $\tilde{P}_{k,l}$ avec $I_{\sigma,k,l}$ et $\mathcal{I}_{k,l}$ en utilisant l'analyse spectrale informée (voir section 4.3).
- Synthétiser $\tilde{s}_k[n]$ à partir de $\tilde{P}_{k,l}$ d'après (4.2).

4.4.5 Complexité calculatoire

La technique proposée dépend du nombre de source K, de la taille N de la transformée STFT et du nombre $M < L$ de composante sinusoïdales non négligeables ($\tilde{a} > 0$) après l'application de la technique ECUSQ pour la quantification optimale des paramètres. Dans notre implémentation, le nombre maximum pour M a été fixé à 50 par fenêtre d'analyse. Nous considérons un nombre λ d'itérations utilisées au codeur afin de mettre à jours la valeur de I_σ qui nécessite la création du mélange après l'application du tatouage. Il s'agit de l'étape la plus complexe de l'algorithme proposé.

La table 4.1 détaille la complexité calculatoire exprimée par unité de temps respectivement au codeur et au décodeur. Ces complexités sont exprimées pour le pire cas en

utilisant la notation "grand O" dite de Landau telle que $\lambda < K < M < N$. Dans la notation proposée, nous supposons que chaque opération arithmétique nécessite exactement une unité de temps pour être exécutée. La complexité de la technique de tatouage n'est pas prise en compte dans les calculs de complexité présentés. Les résultats montrent cependant que le codeur est plus coûteux que le décodeur en terme de temps de calcul. Cela s'explique par le processus d'estimation des paramètres de chaque source qui nécessite l'application de K transformées STFT ainsi que le processus itératif utilisé pour le calcul de $I_{\sigma,k,l}$.

TABLE 4.1 – Complexité calculatoire en unité de temps pour la technique d'estimation des paramètre sinusoïdaux de sources isolées à partir d'un mélange mono-canal

Processus	Nombre d'opérations en unités de temps
STFT & estimation des paramètres des sources	$\mathrm{O}(KN\log(N))$
ECUSQ	$\mathrm{O}(KM)$
Calcul du masque	$\mathrm{O}(M^2)$
Extraction d'informations	$\mathrm{O}(\lambda N\log(N))$
Complexité totale du codeur	$T_{\mathrm{enc}}(\lambda,K,M,N)=\mathrm{O}((\lambda+K)N\log(N)+M^2)$
STFT & estimation des paramètres	$\mathrm{O}(N\log(N))$
Correction des erreurs	$\mathrm{O}(KM)$
Synthèse des signaux sources	$\mathrm{O}(KN\log(N)))$
Complexité totale du décodeur	$T_{\mathrm{dec}}(K,M,N)=\mathrm{O}(KN\log(N))$

Le temps d'exécution est exprimé en fonction du nombre de sources K, du nombre de composantes sinusoïdales M, le nombre d'itérations λ utilisées pour l'extraction d'information et la taille N de la STFT.

4.4.6 Expérimentation et résultats

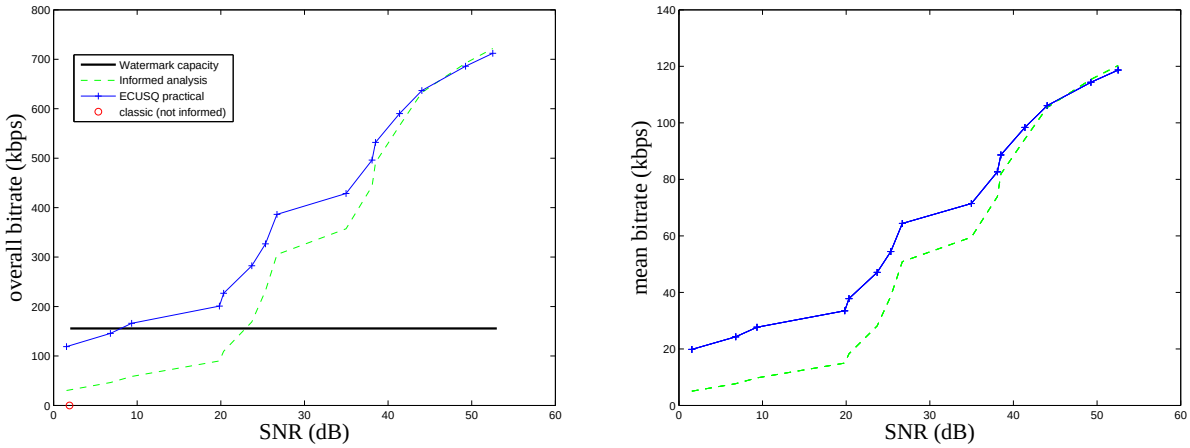
Dans cette section, nous présentons les résultats obtenus en appliquant le système implémenté décrit dans la section 4.4.4 sur un signal de mélange musical mono-canal réaliste dont on dispose les signaux des sources isolées avant la création du mélange.

Dans cette expérimentation, nous utilisons une pièce musicale composée de $K = 6$ sources musicales qui sont respectivement, une voix chantée de femme (Norah Jones), deux guitares, une contrebasse, un clavier et une batterie. Les paramètres de référence $P_{k,l}$ sont d'abord estimés au codeur à partir des signaux isolés de référence s_k dont on dispose. En fonction de la qualité cible fixée, les paramètres $P_{k,l}$ sont quantifiés pour obtenir $\tilde{P}_{k,l}$ et les trajectoires des partiels sont construites afin de déduire le masque d'activation $m_{k,l}$ de chaque source. Finalement, l'information constituée du masque et de l'information complémentaire extraite codée, est tatouée de façon inaudible dans le mélange. Après le codage, le mélange tatoué est vérifié en s'assurant que la valeur de I_σ de chaque composante est correcte pour permettre le décodage et la correction des paramètres estimés au décodeur. Si ce n'est pas le cas, un nouveau mélange est recréé à partir de la mise à jour de I_σ. Dans nos expérimentations, moins de trois itérations étaient suffisantes la plupart du temps. Cette étape est importante, en effet nous expliquons dans la section 4.2 que pour l'analyse informée, une petite valeur pour I_σ augmente la quantité d'information transmise mais cela a aussi pour effet d'améliorer la robustesse aux erreurs d'estimations. Cependant, une valeur surestimée de I_σ rend toute correction d'erreur impossible. Or l'application du tatouage sur le signal (dépendant lui-même de l'analyse du mélange) génère des erreurs aléatoires pouvant nécessiter un réajustement de la variable I_σ utilisée pour l'extraction d'information.

Les figures 4.7 et 4.9 comparent les débits utilisés en pratique par le codage pur basé sur ECUSQ et par l'estimation classique et informée permettant d'atteindre le SNR calculé à partir des signaux de référence (synthétisés en utilisant $P_{k,l}$). La courbe ECUSQ théorique suppose que chaque composante sinusoïdale est codée avec un budget moyen correspondant à une entropie cible H_t déduite du SNR en utilisant (4.36). Pour l'analyse informée, dans la figure 4.9, nous présentons deux courbes d'estimation informée permettant de visualiser le débit supplémentaire utilisé pour coder le masque $m_{k,l}$. Les figures présentent les SNR réels mesurés pour un signal de mélange tatoué, cependant lorsque le débit nécessaire dépasse la capacité du tatouage, l'information est transmise séparément et l'estimation est calculée sur un mélange simulé utilisant la capacité maximale du tatouage. Sur ces figures apparaissent les résultats obtenus utilisant le codage pur en théorie et en pratique ainsi que les résultats de l'estimation classique non informée (cercle rouge). Les résultats obtenus peuvent être expliqués comme suit.

- L'approche classique non informée représentée par un cercle rouge correspond à l'application de la méthode de réallocation et utilise un débit de 0 kbps.
- Le codage pur basé sur ECUSQ présente les résultats théoriques calculés à partir de l'équation (4.35) en supposant que chaque sinusoïde est codée en utilisant la même entropie cible H_t. Le nombre de composantes sinusoïdales non négligeables ($\tilde{a} > 0$) en fonction de l'entropie est décrite dans la figure 4.8. Ce nombre de composante devient plus important en fonction de l'entropie lorsque le modèle sinusoïdal est mal adapté (*e.g.* signal de batterie ou sources avec certains effets). La différence entre la courbe théorique et pratique utilisant la technique ECUSQ s'explique par l'approximation basée sur l'hypothèse de haute résolution utilisée pour la construction du quantificateur.
- L'approche informée présente le débit permettant de transmettre l'information complémentaire avec et sans le masque. Le débit résultant reste plus faible que le codage pur pour des SNR moyens. Pour les SNR élevés, le gain de l'approche informée par rapport au codage pur est réduit. En effet, la part du débit utilisée pour coder le masque devient plus importante et le gain apporté par l'estimateur devient négligeable.

Pour une application pratique utilisant la capacité totale du tatouage, d'après la figure 4.7, l'analyse informée permet de réduire significativement le débit nécessaire pour atteindre la même qualité que le codage pur. Le gain observé sur la figure 4.7 pour un mélange réaliste avec une capacité de tatouage limitée, est de l'ordre de 15 dB sur le SNR du mélange resynthétisé. Les signaux d'exemples relatifs à cette expérimentation sont accessibles en démonstration sur ma page personnelle http://www.fourer.fr.



a) débit global utilisé pour le mélange

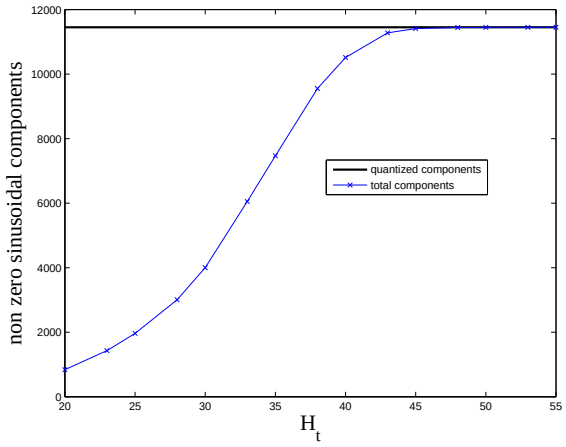
b) débit moyen par source

FIGURE 4.7 – Comparaison entre le débit global (toutes sources confondues) utilisé par la quantification ECUSQ et l'analyse informée. Il s'agit du débit total nécessaire pour estimer toutes les sources à partir de l'observation du mélange.

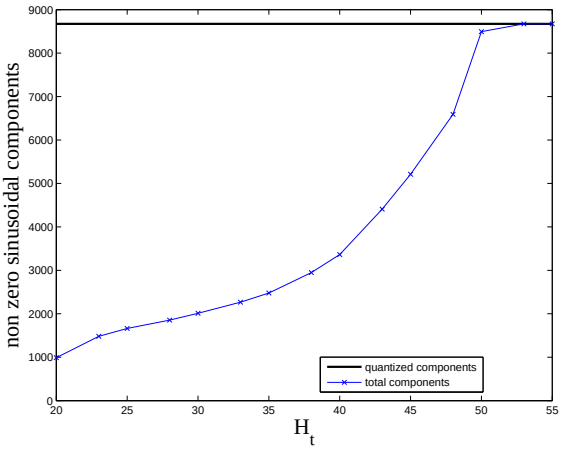
4.5 Conclusions du chapitre

Dans ce chapitre nous avons appliqué l'approche informée à l'analyse spectrale. Dans un premier temps, nous avons présenté les outils qui s'appliquent à la problématique de l'estimation des paramètres sinusoïdaux. Enfin nous avons décrit les spécificité permettant de mettre en oeuvre l'approche informée dans le cadre d'une application pratique. Les résultats montrent que l'approche informée permet désormais de mettre en place des applications où il est possible de configurer une qualité d'estimation cible différente du codage pur. Les résultats montrent que cette approche permet d'améliorer simultanément les performances obtenues en codage pur (débit) mais aussi d'analyse (qualité).

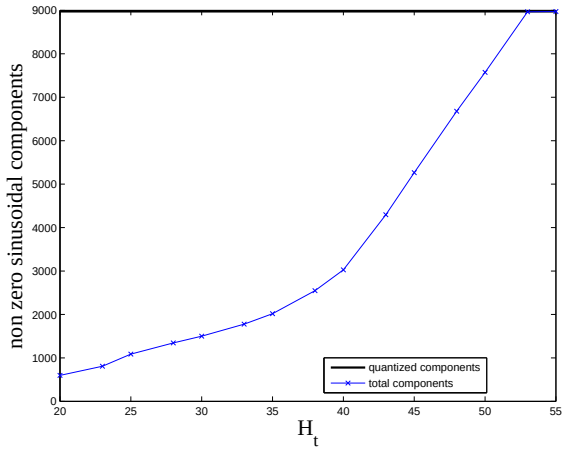
De plus, les résultats que nous obtenons dans les simulations décrites dans ce chapitre sont très encourageants. En effet, ils montrent que l'approche informée permet de repousser des limitations théoriques précédemment établies telle que la borne de Cramér-Rao ainsi que la borne de Shannon lorsque nous combinons à la fois estimation et codage. Nous espérons ainsi voir émerger prochainement des liens plus étroits entre estimation et codage d'un point de vue théorique et pratique.



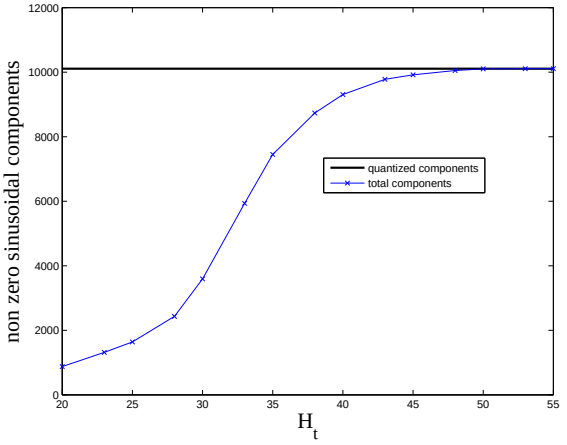
a) guitare 1



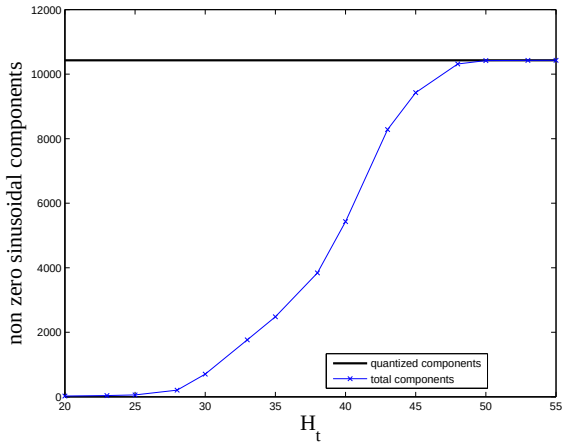
d) synthétiseur



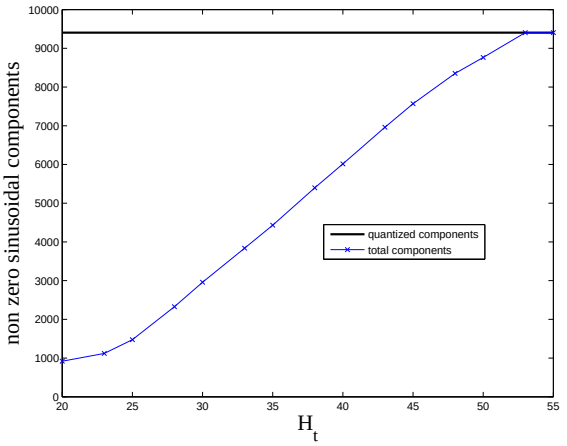
b) basse



e) voix

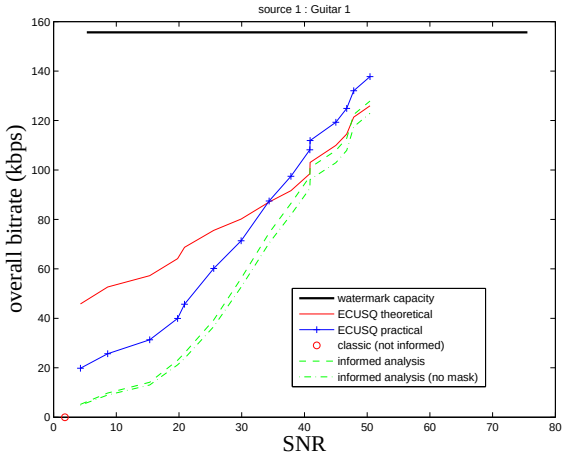


c) batterie

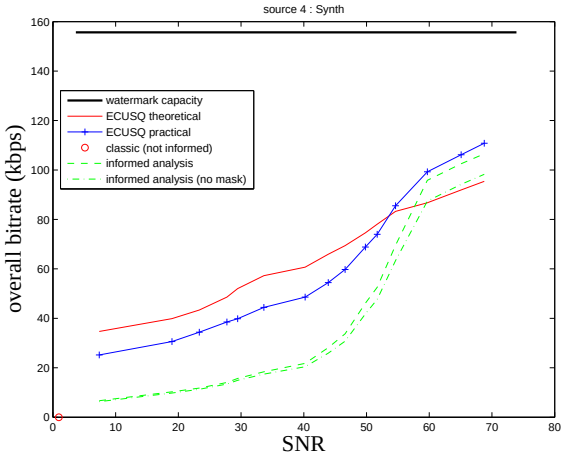


f) guitare 2

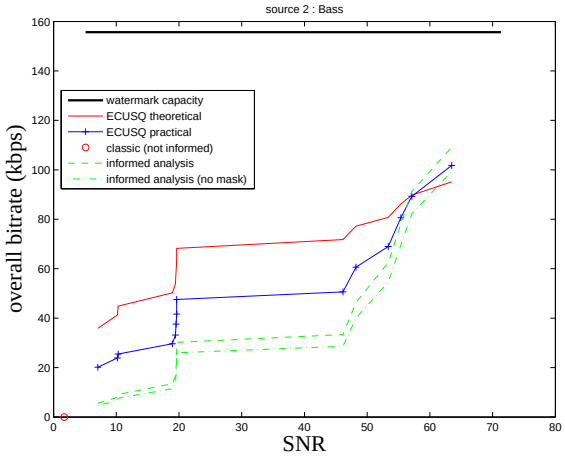
FIGURE 4.8 – Nombre de composantes d'amplitude non nulle pour chaque source en fonction de l'entropie cible H_t. Ce nombre augmente en fonction du SNR cible et dépend de la quantification calculée par la méthode ECUSQ.



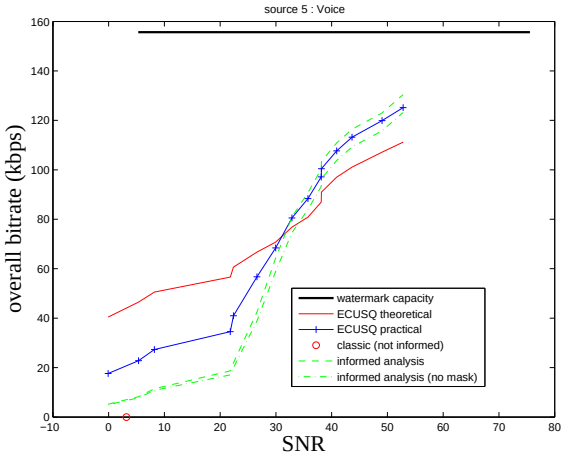
a) guitare 1



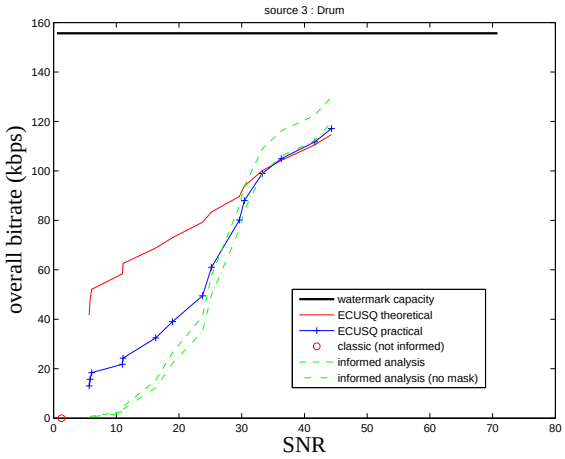
d) synthétiseur



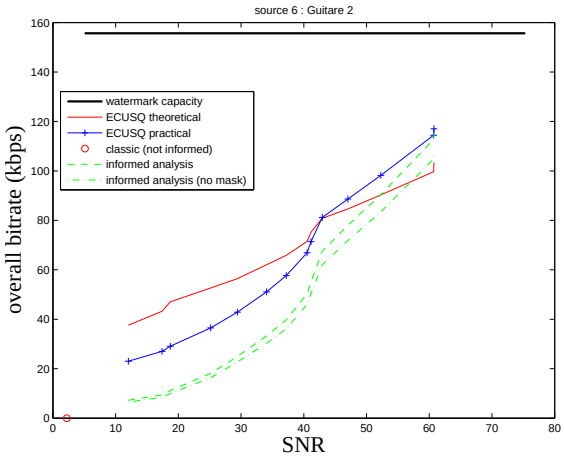
b) basse



e) voix



c) batterie



f) guitare 2

FIGURE 4.9 – Comparaison du débit obtenu par l'estimation classique (cercle rouge), par le codage pur (utilisant la quantification ECUSQ) et par l'analyse informée en fonction du SNR. Le débit augmente avec la qualité obtenue mais reste plus faible pour l'analyse informée pour des valeurs de SNR raisonnables. Pour les SNR très élevés, le codage seul semble plus efficace.

CHAPITRE 5

APPROCHE INFORMÉE APPLIQUÉE À L'EXTRACTION DE DONNÉES SYMBOLIQUES À PARTIR D'UN SIGNAL AUDIO

L'extraction d'informations musicales, *Music Information Retrieval* (MIR) est un domaine de recherche pluridisciplinaire qui a pour objectif d'estimer les paramètres musicaux permettant de décrire le contenu d'une oeuvre musicale. Cette estimation se fait en général à partir d'un signal audio ou d'une représentation symbolique (*e.g.* partition) de la pièce musicale analysée.

Ce domaine de recherche propose des solutions, la plupart du temps sous la forme d'une application logicielle ou d'un système permettant de traiter des problèmes distincts portant sur l'analyse des signaux de musique. Parmi ces applications, nous pouvons citer : la détection du tempo [Hai06], l'estimation de la hauteur des notes jouées par un instrument monophonique [dCK02] ou polyphonique [YRR10, Kla03], la détection de la tonalité, du nom des accords et de la structure d'une pièce musicale (couplet, refrain, etc.) [RRHO10].

Contrairement aux problèmes d'estimation de paramètres physiques utilisant un modèle mathématique (*e.g.* chapitre 4), les systèmes MIR s'intéressent à estimer des informations symboliques abstraites.

Cela rend d'autant plus difficile toute évaluation objective, d'abord en absence de données exactes et précises de référence d'une part. D'autre part, parce que les données recherchées sont la plupart du temps subjectives (*e.g.* tonalité, débuts et fins de note) ce qui rend la théorie de l'estimation (*cf.* chapitre 3) difficilement applicable. C'est aussi la raison pourquoi les techniques proposées dans le domaine MIR sont souvent statistiques voire empiriques et introduisent de nombreuses astuces difficilement démontrables mais pourtant issues des observations expérimentales ou tout simplement du "bon sens".

Les travaux décrits dans cette thèse portent sur la transcription automatique des instruments jouant des hauteurs (*e.g.* voix, piano, guitare, etc.) à partir d'un signal audio de mélange. Ces instruments dits quasi-harmoniques constituent la partie tonale des si-

gnaux de musique qui correspond presque toujours à la partie essentielle d'une musique (mélodie, accords, etc.). Ainsi, la transcription des instruments percussifs sans hauteur tonale (*e.g.* batterie) n'est pas développée ici. Le lecteur pourra cependant si il le souhaite consulter des travaux spécifiques sur ce sujet [FP06, APF09, GR08].

La transcription d'un signal musical est une tâche essentielle permettant d'obtenir la partition musicale sur laquelle s'appuient de nombreuses autres applications MIR. En effet, une partition musicale précise obtenue à partir d'un signal audio permet de déduire d'autres informations sur celui-ci telles que la tonalité, le nom des accords et la structure musicale de l'oeuvre. D'autres tâches plus difficiles s'en retrouvent simplifiées comme le calcul de similarité [Mar12b] entre deux pièces musicales qui intervient par exemple dans des applications telles que la détection de *cover* (reprise) [RE10] ou dans le cadre de la pédagogie musicale.

L'estimation F_0 multiple est une tâche essentielle à l'origine de tout système de transcription musical à partir d'un signal audio qui consiste à estimer les hauteurs jouées par les différents instruments quasi-harmoniques qui composent un signal. Le cas polyphonique est le cas le plus général qui permet de transcrire un orchestre ainsi que les instruments jouant plusieurs notes à la fois comme la guitare ou le piano. Ce cas est donc aussi le plus difficile à traiter en raison des problèmes évoqués dans le chapitre 2.

Ainsi, en plus de l'estimation des fréquences fondamentales de chaque note présente dans un mélange, l'estimation F_0 multiple doit aussi estimer la polyphonie, c'est-à-dire le nombre de fréquences à estimer simultanément à chaque instant qui peut varier en fonction du temps. Cette tache est rendue plus compliquée par lorsque des timbres différents composent le mélange en présence de bruit et de transitoires créés par certains instruments comme la batterie. Il est intéressant de noter que le problème de transcription polyphonique est très proche du problème de séparation de sources [YR04, BM01] (*cf.* section 2.3). Plusieurs combinaisons de ces deux problèmes ont ainsi déjà été proposés dans la littérature, comme la séparation de sources informée par la partition [HDB11, GSMA10].

En raison de ses nombreuses applications, l'estimation F_0 multiple suscite toujours aujourd'hui de nombreux travaux [Kla03, Mar04, BBV10, YRR10]. A ce jour, la technique la plus précise évaluée a obtenu lors de la compétition MIREX (*Music Information Retrieval eXchange*) 2011 une précision de 68% [YR11] sur la base de données d'évaluation. Il est donc évident qu'une telle précision reste très insuffisante pour les applications plus exigeantes nécessitant nécessitant une transcription exacte ou presque.

De plus, en raison des limitations des estimateurs sur lesquels s'appuient les systèmes de transcription polyphonique, il est très probable que les meilleures techniques aveugles qui seront proposées dans l'avenir n'arrivent jamais à atteindre une précision de 100% quelle que soit la pièce musicale analysée. En effet, nous avons montré dans le chapitre 4 que le processus de mélange engendre une perte d'information sur les signaux des sources. De plus, la théorie de l'estimation nous permet de mesurer cette perte d'information par le calcul des bornes de Cramér-Rao.

C'est la raison pour laquelle nous proposons dans ce chapitre l'approche "informée" appliquée au problème d'estimation F_0 multiple à partir d'un signal audio. Comme pour l'estimation paramétrique informée décrite dans le chapitre 4, cette approche consiste à combiner les techniques d'analyse classiques existantes avec de l'information complémentaire codée de taille minimale. Cette approche a donc pour objectif d'atteindre dans tous les cas la précision souhaitée quelle que soit la pièce musicale analysée lorsque l'approche informée est possible.

5.1 Étude et proposition d'un système complet d'estimation F_0 multiple

Nous avons choisi de commencer notre étude en implémentant un système d'estimation F_0 multiple "efficace". Pour cela, nous proposons une description détaillée d'un système de transcription polyphonique basé sur la méthode de Yeh [Yeh08] évaluée comme étant la plus précise au MIREX 2011. Notre intuition, comme dans le cas de l'analyse spectrale informée, est que plus une estimation est précise, plus la quantité d'information complémentaire codée utilisée dans le cadre d'une approche informée sera faible.

La méthode que nous proposons ici n'utilise aucune connaissance musicale *a priori* et permet d'obtenir une représentation symbolique (*e.g.* partition solfège ou MIDI) à partir d'un signal de musique ou de parole. Dans ce cas, on s'intéresse uniquement à la partie tonale générée par les instruments jouant des notes caractérisées par une hauteur et pouvant être décrits par le modèle de source quasi-harmonique formulé dans la section 2.2.2 par l'équation (2.23).

L'information que nous cherchons à extraire à partir du signal observé se base sur les descripteurs suivants :

- les activations temporelles liées aux débuts et aux fins de note (*onset / offset*),
- la hauteur des notes associées à chaque activation et caractérisée par une fréquence fondamentale F_0.

Contrairement à l'estimation F_0 simple qui suppose que le nombre de notes jouées simultanément vaut au maximum $L = 1$ à chaque instant (d'après l'équation 2.23. L'estimation F_0 multiple estime la polyphonie $L \geq 1$ qui varie au cours du temps et dépend de la nature des signaux traités.

L'estimation F_0 multiple est traitée par des méthodes se présentant sous forme de système combinant plusieurs étapes de traitement appliquée chacune à un sous-problème tel que l'extraction des composantes tonales, l'estimation de la polyphonie et l'estimation des candidats F_0 présents dans le mélange. Pour traiter le problème de la polyphonie, Klapuri propose dans [Kla03] de soustraire chaque source harmonique détectée itérativement en appliquant un lissage spectral. Par ce mécanisme, il effectue une reconstruction partielle de l'enveloppe des sources harmoniques résiduelles permettant de ne pas altérer leur détection lors des itérations suivantes. Bertin dans [BBV10] propose une approche basée sur la factorisation du spectrogramme utilisant la technique de NMF [LS99], reposant sur un dictionnaire d'atomes de sources harmoniques. Enfin, le système proposée ici est basé sur la méthode de Yeh [YRR10] utilisant la modélisation sinusoïdale et un modèle de source quasi-harmonique générique permettant de décomposer le signal émis par un instrument tonal. Nous décrivons dans cette section un système complet (*cf.* figure 5.1) reprenant la chaîne de traitement décrite dans la figure 5.1 et intégrant les techniques les plus avancées à notre connaissance de l'état de l'art. Ainsi, cette implémentation nous permettra d'introduire par la suite l'approche "informée" développée dans la section 5.2.

5.1.1 Extraction des composantes tonales

L'objectif de cette première étape consiste à séparer du signal de mélange $x[n]$ les composantes sinusoïdales qui constituent la partie tonale du mélange. Pour cela on propose d'appliquer le traitement suivant sur le spectre d'amplitude du signal discret noté $|X[k]|$

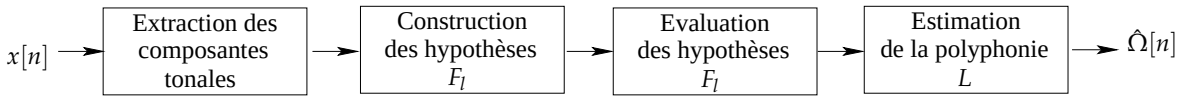


FIGURE 5.1 – Schéma descriptif de la méthode d'estimation F_0 multiple proposée permettant d'estimer un ensemble $\hat{\Omega}[n]$ de candidats F_0 actifs à un instant n à partir d'un signal de mélange discret $x[n]$.

obtenu en appliquant la STFT sur une trame du signal $x[n]$. Ce traitement comprend les étapes suivantes :

- Estimation d'un seuil pour l'amplitude minimale noté $\hat{T}[k]$ dépendant de la fréquence discrète k pour la trame de signal considérée.

- Estimation itérative des paramètres sinusoïdaux pour les pics spectraux sélectionnés par ordre décroissant d'amplitude.

- Soustraction de la sinusoïde estimée du mélange : le pic est ignoré si celui-ci dégrade la qualité (c'est-à-dire fait augmenter l'erreur quadratique) du signal resynthétisé à partir de la sommes des composantes sinusoïdales estimées.

- Les pics suivants sont estimés et soustraits à leur tour jusqu'à ce que le seuil du niveau de bruit soit atteint ou que l'énergie résiduelle soit inférieure à une constante (*e.g.* -60dB).

Contrairement à l'algorithme de *Matching Pursuit* [MZ93], le traitement proposé n'utilise pas de dictionnaire et permet une estimation automatique du critère d'arrêt. Chacune des étapes est détaillée ci-après.

a) Seuillage

Ce traitement a pour objectif de réduire le nombre de composantes sinusoïdales détectées comme un pic (maximum local) puis extraites à partir du spectre d'amplitude. En effet, comme notre objectif est d'estimer uniquement la partie tonale du signal (contrairement au chapitre 4 qui ne faisait pas de distinction), il est alors nécessaire de fixer un seuil définissant l'amplitude minimale de chaque composante sinusoïdale recherchée dans le spectre d'amplitude. Cette approche suppose que les composantes tonales possèdent une amplitude plus importante que les pics de bruits, plus nombreux et d'amplitude plus faible. Pour cela nous avons choisi d'utiliser une fonction de seuillage automatique basée sur la méthode utilisée dans [ES06].

La fonction de seuillage utilisée dépendant de la fréquence discrète notée $T[k]$ est obtenu en filtrant le spectre d'amplitude $|X[k]|$ par une fenêtre de Hamming définie par :

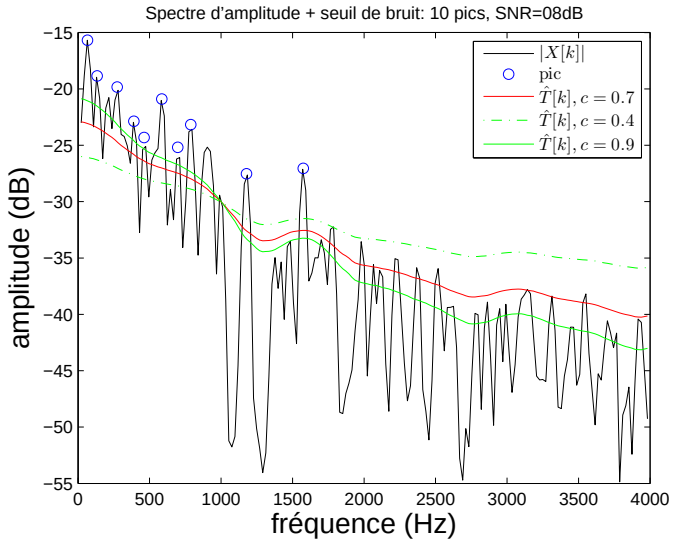
$$w_{\text{Hamming}}[n] = \begin{cases} 0.54 - 0.46\cos 2\pi \frac{n}{\dot{N}} & \text{si } n \in [0;\dot{N}] \\ 0 & \text{sinon,} \end{cases} \tag{5.1}$$

de taille $\dot{N} = 1 + \frac{N}{64}$ où N correspond à la taille initiale de la STFT. La fenêtre de Hamming utilisée est donc de taille impaire pour des raisons de symétrie (composée d'une valeur centrale et de deux segments de même taille). Un facteur de compression $c \in [0.5;1[$

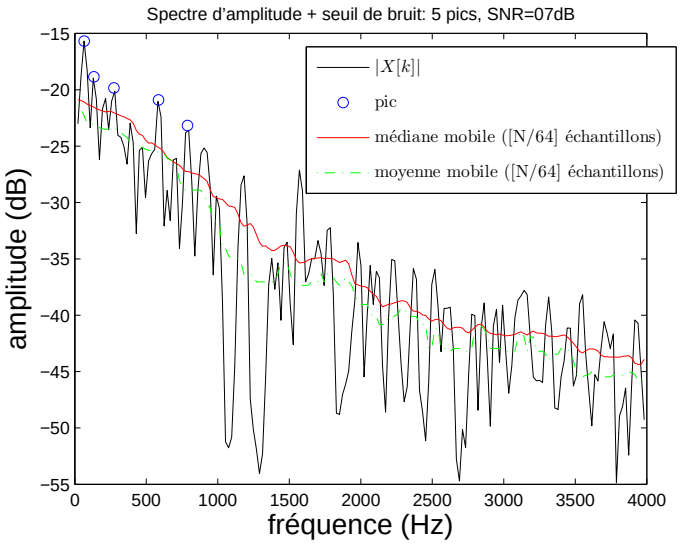
est ensuite appliqué afin d'obtenir $\hat{T}[k] = T[k]^c$. Dans notre implémentation nous avons choisi $c = 0.7$. Le seuil d'amplitude noté $\hat{T}[k]$ est donc défini par :

$$\hat{T}[k] = [w * |X[k]|]^c. \tag{5.2}$$

Dans les figures 5.2a) et 5.2b), nous présentons un exemple d'application de la fonction de seuillage proposée que nous comparons avec la médiane mobile. Sur cette figure nous indiquons le nombre de pics extraits ainsi que le SNR du signal reconstruit à partir des paramètres sinusoïdaux estimés (détails ci-après). Sur la figure 5.2a), nous pouvons visualiser la valeur de $\hat{T}[k]$ avec différentes valeur de c. Nous présentons également la valeur de la moyenne mobile sur la figure 5.2b).



a) seuillage automatique utilisant l'équation (5.2)



b) seuillage utilisant la médiane mobile

FIGURE 5.2 – Extraction des composantes tonales (*peak*) à partir d'un son polyphonique en utilisant respectivement le seuillage automatique proposé 5.2a) et un seuillage utilisant la médiane mobile 5.2b).

b) Estimation des paramètres sinusoïdaux

L'estimation des paramètres pour chaque composante sinusoïdale suppose ici un modèle non stationnaire (déjà introduit dans la section 2.1.3) que nous rappelons ici par l'équation (5.3) :

$$s(t) = \underbrace{\exp\left(\lambda_0 + \mu t\right)}_{a(t)} \underbrace{\exp\left(\mathbf{j}(\phi_0 + \omega t + \frac{\psi}{2}t^2)\right)}_{\phi(t)}. \tag{5.3}$$

Ce modèle, plus général que celui utilisé dans le chapitre 4 permet de prendre en compte les variations temporelles d'amplitude et de fréquence sur une courte durée de signal (sur une trame du signal). Dans le cas d'une analyse spectrale non informée, ce modèle combiné avec la méthode de réallocation généralisée permet d'améliorer la robustesse et la précision d'estimation. Cette meilleure précision est nécessaire pour la précision de l'évaluation des candidats F_0. De plus, ce modèle permet de compenser l'utilisation d'une fenêtre d'analyse de grande taille.

Ainsi, les pics de plus forte amplitude sont sélectionnés après le seuillage décrit précédemment par ordre décroissant d'amplitude dans la bande des fréquences considérées. Nous choisissons des fréquences comprises entre 27.5Hz et 12543.85Hz qui correspondent respectivement aux fréquences des notes *La -1* et *Sol 8*.

Les paramètres sont estimés par la méthode de réallocation généralisée au modèle non stationnaire [MD08]. Comme montré dans le chapitre 4, l'estimateur de réallocation utilisant une analyse de Fourier et la fenêtre de Hann atteint presque la borne de Cramér-Rao et compte actuellement parmi les meilleures techniques d'analyse spectrale sans information complémentaire. Ainsi, la fréquence $\hat{\omega}$ et le temps réalloué $\hat{t}$ sont estimés pour chaque point (t, ω) à partir du spectre (*cf.* équation (4.4)) :

$$\hat{\omega} = \omega - \underbrace{\mathbf{Im}\,\frac{S_{w'}(t,\omega)}{S_w(t,\omega)}}_{-\Delta_\omega}, \tag{5.4}$$

$$\hat{t} = t + \underbrace{\mathbf{Re}\,\frac{S_{tw}(t,\omega)}{S_w(t,\omega)}}_{\Delta_t}. \tag{5.5}$$

Dans le cas d'un signal discret de fréquence d'échantillonnage F_s, t et ω peuvent être calculés à partir des indices discrets du spectrogramme du signal noté $S[n,k]$ tel que $t = n/F_s$ avec n l'indice temporel situé au centre de la trame de taille N correspondante. $\omega = 2\pi F_s \frac{k}{N}$ est la fréquence discrète correspondant à l'indice k.

Les paramètres de modulation de l'amplitude $\hat{\mu}$ et de modulation de la fréquence $\hat{\psi}$ sont obtenus par la généralisation de la méthode proposée dans [MD08, Hai04, Roe02] :

$$\begin{aligned}\hat{\mu} =& \frac{\partial}{\partial t}\mathbf{Re}\,\left(\log\left(S_w(t,\omega)\right)\right)\\ =& -\mathbf{Re}\,\left(\frac{S_{w'}(t,\omega)}{S_w(t,\omega)}\right),\end{aligned} \tag{5.6}$$

$$\begin{aligned}\hat{\psi}(t,\omega) =& \frac{\partial\hat{\omega}}{\partial\hat{t}} = \frac{\frac{\partial\hat{\omega}}{\partial t}}{\frac{\partial\hat{t}}{\partial t}} \\ =& \frac{\mathbf{Im}\left(\frac{S_{w''}(t,\omega)}{S_w(t,\omega)}\right) - \mathbf{Im}\left(\left(\frac{S_{w''}(t,\omega)}{S_w(t,\omega)}\right)^2\right)}{\mathbf{Re}\left(\frac{S_{tw}(t,\omega)S_{w'}(t,\omega)}{S_w(t,\omega)^2}\right) - \mathbf{Re}\left(\frac{S_{tw'}(t,\omega)}{S_w(t,\omega)}\right)}.\end{aligned} \tag{5.7}$$

La fonction $G_w(\Delta_\omega, \hat{\mu}, \hat{\psi})$ utilisée pour le calcul de $\hat{a}$ et de $\hat{\phi}_0$ est définie par [MD08] :

$$G_w(\Delta_\omega, \hat{\mu}, \hat{\psi}) = \int_{-\infty}^{+\infty} w(t) \exp\left(\hat{\mu}t + \mathbf{j}\left(\Delta_\omega t + \frac{\hat{\psi}}{2}t^2\right)\right) dt, \tag{5.8}$$

et correspond au facteur de correction appliqué à la valeur de la sinusoïde à l'instant $t = 0$ notée $s_0 = \mathbf{e}^{\lambda_0 + \mathbf{j}\phi_0}$.
Ainsi, l'amplitude $\hat{a}$ et la phase $\hat{\phi}_0$ sont estimés à partir de la fonction $G_w(\Delta\omega, \hat{\mu}, \hat{\psi})$ comme suit :

$$\hat{a} = \left|\frac{S_w(t,\omega)}{G_w(\Delta_\omega, \hat{\mu}, \hat{\psi})}\right|, \tag{5.9}$$

$$\hat{\phi}_0 = \angle\left(\frac{S_w(t,\omega)}{G_w(\Delta_\omega, \hat{\mu}, \hat{\psi})}\right). \tag{5.10}$$

Ces résultats sont donnés avec S_w, $S_{w'}$ et $S_{w''}$ la transformée de Fourier à court terme du signal $s(t)$ calculée en utilisant respectivement la fenêtre d'analyse $w(t)$, sa dérivée première $w'(t) = \frac{dw}{dt}(t)$ et sa dérivée seconde $w''(t) = \frac{d^2w}{dt^2}(t)$.
S_{tw} est la transformée de Fourier à court terme utilisant la fenêtre $w(t)$ pondérée terme à terme par l'axe de temps : $t \cdot w(t)$.
Ainsi, chaque composante sinusoïdale est désormais décrite par un quintuplet $(a, \mu, \phi_0, \omega, \psi)$, dont le signal correspondant peut être synthétisé grâce à l'expression donnée par (5.3).

5.1.2 Construction des hypothèses F_0

Afin de définir les candidats F_0, on fait correspondre à chaque composante sinusoïdale extraite précédemment une hypothèse de source harmonique, *Hypothetical Partial Sequence* (HPS) construite à partir du modèle de source harmonique donné par (2.23).

Ainsi, chaque sinusoïde est traitée comme un candidat F_0 pour lequel on recherche les sinusoïdes voisines dont la fréquence est proche d'un multiple hF_0 ($h \in \mathbb{N}^*$) suggéré par le modèle harmonique décrit par l'équation (2.23) (*cf.* chapitre 2).

Afin de rendre la méthode plus robuste en cas d'inharmonicité dans le cas polyphonique bruité, un intervalle de fréquence est toléré pour l'affectation des sinusoïdes. On définit un paramètre α tel que $\alpha = 2^{\frac{1}{12}} - 1 \approx 0.06$. Ainsi, nous recherchons les sinusoïdes dans l'intervalle $[(1-\alpha)hF_0; (1+\alpha)hF_0]$ en vue de les affecter au partiel h du HPS considéré. La valeur du paramètre α choisie correspond approximativement au facteur séparant 2 notes de musiques successives en utilisant l'échelle de tempérament égal (gamme chromatique dans la musique occidentale). Sur cette échelle, chaque octave est divisée en 12 demi-tons comme expliqué dans la section 1.3.3. Ainsi, la différence entre deux fréquences espacées d'un demi-ton vaut $f \cdot 2^{1/12} - f = f \cdot \underbrace{(2^{1/12} - 1)}_{\alpha}$.

La construction des HPS peut être schématisée par la figure 5.3 pour laquelle la recherche des partiels s'effectue itérativement par ordre croissant de fréquence. On choisit d'utiliser un modèle qui s'adapte aux observations afin de déterminer l'intervalle de tolérance des pics suivants. Ainsi, dans le cas où un pic observé correspondant à la fréquence $f_{obs}^{(h)}$ a été affecté de manière fiable au HPS dans l'intervalle considéré, la fréquence de référence utilisé pour la recherché à l'itération suivante devient $f_{model}^{(h+1)} = f_{obs}^{(h)} + F_0$.

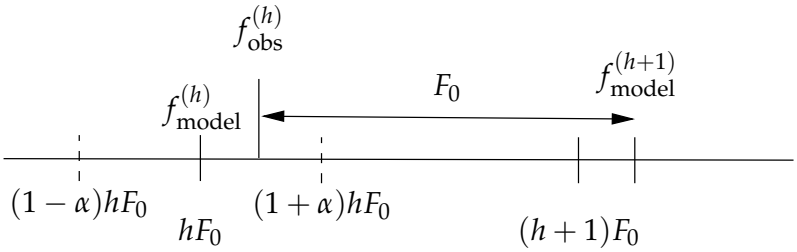


FIGURE 5.3 – Construction des hypothèses de source harmonique par sélection des pics observés depuis le spectre d'amplitude.

Lors de l'affectation des sinusoïdes à chaque HPS, les cas suivants sont considérés :

- Un seul pic est présent dans l'intervalle : il est alors affecté au HPS.

- Aucun pic n'a été détecté : dans ce cas le pic est ignoré (amplitude nulle). Le nombre de sinusoïdes affectées à l'HPS ainsi la forme de son enveloppe spectrale sont de toute manière des facteurs discriminant permettant de valider ou non une hypothèse sur un candidat F_0.

- Plusieurs pics sont présents sur l'intervalle : on affecte alors le pic minimisant le critère MBW[1] proposé dans la partie b.1) qui correspond au lissage de l'enveloppe spectrale. Pour des raisons de simplification en temps de calcul, nous avons choisi de considérer le partiel le plus proche du modèle harmonique. Dans ce cas, on favorise uniquement le critère d'harmonicité décrit dans la section 5.1.3.

L'ordre d'extraction des candidats F_0 joue un rôle important sur l'efficacité globale de la méthode. En effet, l'ordre d'affectation des sinusoïdes a un effet sur les erreurs d'estimations obtenues pour les itérations suivantes. Il est donc important d'estimer les sources quasi-harmoniques les plus importantes en priorité.

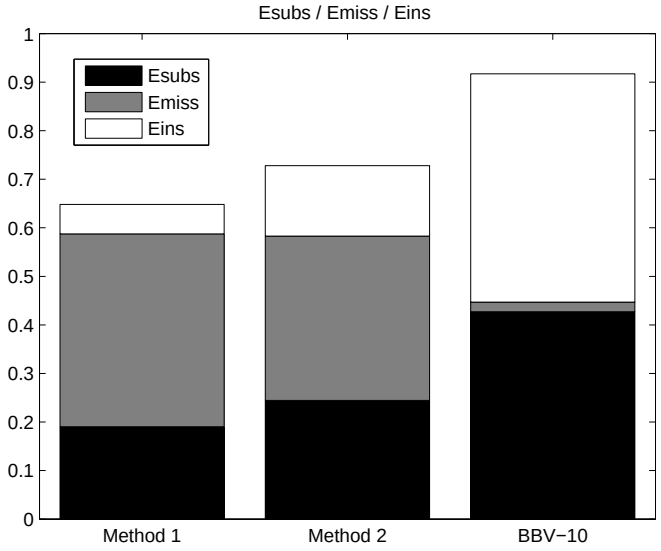
Deux méthodes ont été comparées :

- Méthode 1 : choix prioritaire du candidat F_0 de plus faible fréquence et comparaison itérative avec le candidat $2F_0$ pour maximiser la vraisemblance. Le calcul de la vraisemblance pour un HPS s'effectue avec une fonction de score donnée par l'équation (5.19) dont le calcul est détaillé ci-après.

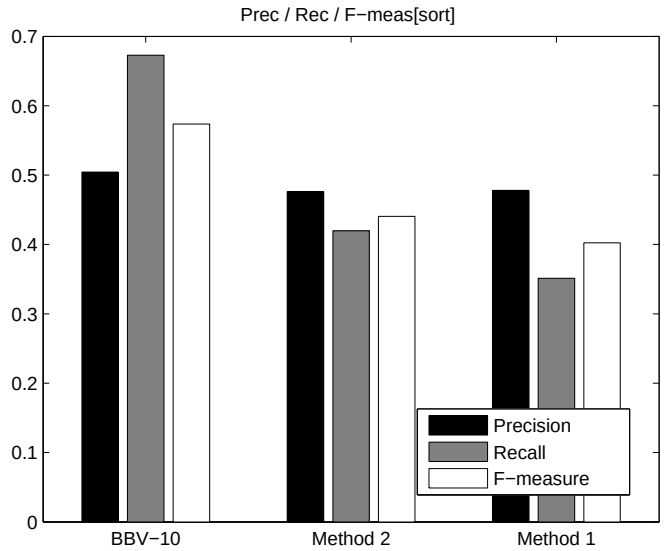
- Méthode 2 : choix prioritaire du candidat F_0 correspondant à la sinusoïde de plus forte amplitude et comparaison itérative avec les hypothèses $F_0/2$ puis $2F_0$ afin de maximiser la vraisemblance.

[1] *Mean BandWidth* (MBW) correspond à la bande passante moyenne de l'enveloppe spectrale du signal considéré

Pour cette comparaison, nous avons utilisé une base d'évaluation composée de 10 pièces musicales polyphoniques synthétisées à partir du MIDI en utilisant MIDITOOLBOX [ET04]. Nous avons ensuite comparé les performances obtenues avec chacune des deux approches en utilisant la métrique décrite dans la section 2.2.4 pour obtenir la figure 5.4.



a) comparatif des erreurs d'estimation (candidats substitués, oubliés ou insérés) triées par ordre croissant de l'erreur totale



b) comparatif de la précision triée par ordre décroissant de *F-measure*

FIGURE 5.4 – Comparaison des performances moyennes d'estimation F_0 multiple entre les deux méthodes proposées pour la construction des HPS et la technique de transcription polyphonique **BBV-10** [BBV10] sur la base de données de test.

D'après les résultats de cette étude comparative présentés dans la figure 5.4, la méthode 1 est celle qui obtient l'erreur totale la plus faible. On remarque aussi que les deux

méthodes ont tendance à sous-évaluer la polyphonie (E_{miss} plus faible) alors que la méthode **BBV-10** [BBV10] a tendance à la sur-évaluer (erreurs de substitution et d'insertion plus importantes). Les résultats montrent que sur cette base d'évaluation, la méthode **BBV-10** est celle qui obtient le meilleur score de précision suivie ensuite par la méthode 2 qui choisit en priorité le candidat F_0 correspondant à la sinusoïde de plus forte amplitude. Ces résultats montrent une différence en précision assez faible entre les deux méthodes choisies pour la construction des HPS. Nous avons cependant choisit d'utiliser dans notre implémentation la méthode 1 qui obtient au total un nombre d'erreurs plus faible.

5.1.3 Évaluation des candidats F_0

Afin de comparer les hypothèses de candidats F_0 entre elles ainsi que pour leur sélection, il est nécessaire d'estimer leur vraisemblance qui a pour objectif de mesurer la pertinence entre l'observation et le modèle d'une source quasi-harmonique. Pour son calcul, nous avons choisi de combiner simultanément plusieurs critères que nous détaillons ci-après.

a) L'harmonicité

Dans le cas d'un instrument parfaitement identifié, il est possible grâce à un modèle physique spécifique [FR98] de déterminer avec une grande précision la fréquence de chacun des partiels qui composent la forme d'onde de son signal. L'utilisation de l'information *a priori* sur la forme d'onde des sources quasi-harmoniques rentre dans le cadre des systèmes de transcription polyphonique spécifiques (*e.g.* transcription du piano [Emi08], dictionnaire spécifique pour le *Matching Pursuit* [GB03] ou la NMF [SB03]).

Dans le cas le plus général que nous considérons ici, nous ne possédons aucune information *a priori* sur la forme d'onde de chaque source quasi-harmonique qui compose un signal de mélange polyphonique. Ainsi, nous avons choisi d'utiliser la mesure d'inharmonicité proposée par Yeh [YRR10] qui consiste à mesurer l'écart entre le modèle quasi-harmonique proposé dans la section 2.2.2 et les partiels affectés au HPS. Ce critère d'harmonicité s'exprime comme suit :

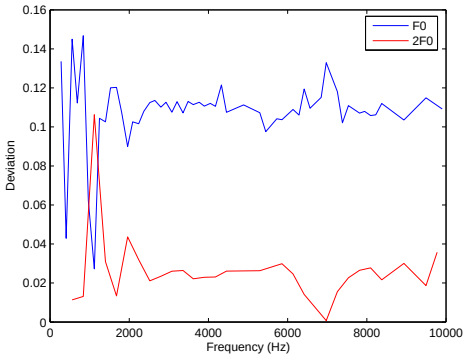
$$\text{HAR} = \frac{\sum_{h \in \text{HPS}} d_h A_h^\beta}{\sum_{h \in \text{HPS}} A_h^\beta} \tag{5.11}$$

avec A_h l'amplitude de la sinusoïde correspondant à l'harmonique h associé au HPS. Ici β est un facteur de compression permettant d'équilibrer l'influence des pics de plus forte amplitude en général présents dans les basses fréquences. Ainsi, dans notre implémentation nous avons fixé $\beta = 0.5$ comme proposé initialement dans [YRR10]. Le facteur d_h correspond à l'écart calculé pour chaque partiel h en utilisant :

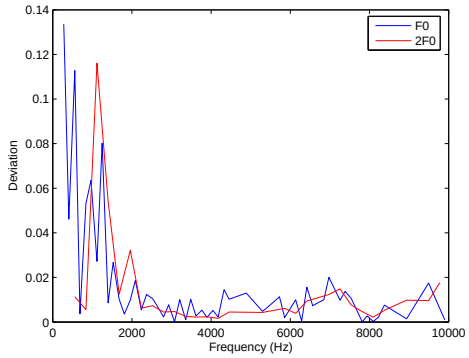
$$d_h = \begin{cases} \frac{|f_{obs}^{(h)} - f_{model}^{(h)}|}{\alpha f_{model}^{(h)}} & \text{si } |f_{obs}^{(h)} - f_{model}^{(h)}| < \alpha f_{model}^{(h)} \\ 1 & \text{sinon} \end{cases} \tag{5.12}$$

où $f_{obs}^{(h)}$ et $f_{model}^{(h)}$ sont respectivement la fréquence observée et la fréquence théorique du modèle quasi-harmonique. Nous réutilisons ici la variable α définie précédemment dans la section 5.1.2 pour la construction des HPS.

Ainsi, ce descripteur donne une indication sur la qualité de construction des HPS cependant comme le montrent les figures 5.5 et 5.6, ce critère n'est pas discriminant lorsque l'on compare F_0 avec $2F_0$. Cela s'explique car les partiels sont supposés être également

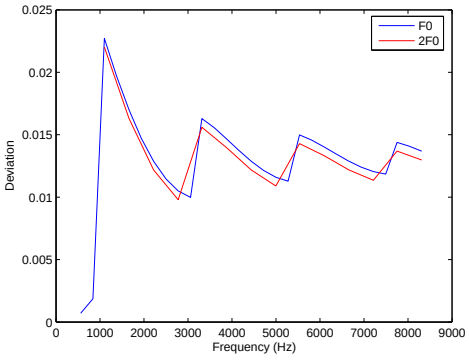


a) par rapport au modèle harmonique sur la source (HAR(F_0) = 0.103 et HAR($2F_0$) = 0.021)

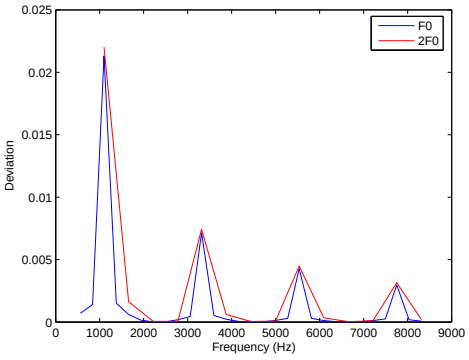


b) modèle évolutif sur la source (HAR(F_0) = 0.030 et HAR($2F_0$) = 0.013) (Le facteur d'inharmonicité est mis à jour en fonction des observations)

FIGURE 5.5 – Valeur de l'écart d_h pour les partiels d'un son de saxophone.



a) par rapport au modèle harmonique sur la source (HAR(F_0) = 0.0126 et HAR($2F_0$) = 0.0127)



b) modèle évolutif sur la source (HAR(F_0) = 0.0014 et HAR($2F_0$) = 0.0026) (Le facteur d'inharmonicité est mis à jour en fonction des observations)

FIGURE 5.6 – Valeur de la l'écart d_h pour un son sinusoïdal harmonique composé de 30 partiels d'amplitude égale.

espacés. L'utilisation seule du descripteur HAR peut engendrer musicalement des problèmes d'estimation de hauteur : note exacte mais octave fausse. Il doit donc être utilisé conjointement avec d'autre critères pour permettre une détection correcte des sources harmoniques à partir d'un signal de mélange polyphonique.

b) L'enveloppe spectrale

L'enveloppe spectrale nous fournit des informations essentielles sur le timbre de la source harmonique considérée (*cf.* section 1.3.4). Les caractéristiques qui nous intéressent dans le cadre de l'estimation F_0 multiple pour caractériser une source quasi-harmonique sont :

- Le lissage de l'enveloppe : caractérisé par une bande passante fréquentielle étroite de l'enveloppe spectrale.
- La décroissance en amplitude des partiels : malgré la présence de formants (maxima

locaux), l'enveloppe spectrale d'une source harmonique a tendance à décroître lorsque que l'on se rapproche des hautes fréquences.

Ces deux caractéristiques peuvent être formalisées par deux descripteurs utilisés par [YRR10] qui sont respectivement la bande passante moyenne et la centroïde spectrale.

b.1) La bande passante spectrale moyenne

Ce critère a pour objectif de donner une indication sur le lissage de l'enveloppe spectrale du HPS considéré. Ainsi, son calcul fait intervenir l'utilisation d'une seconde transformée de Fourier appliquée sur le spectre d'amplitude du signal considéré (*cf.* section 1.3.4). Ainsi, une enveloppe spectrale lisse sera caractérisée par une concentration de son énergie dans les basses fréquences et donc une bande passante plus faible.

Le calcul de ce critère s'effectue comme suit :

- Construction d'un signal g_h symétrique définit pour $h \in [-H; H]$ et correspondant à l'enveloppe dans le spectre d'amplitude défini en fonction du modèle de source quasi-harmonique. Ainsi, pour une hypothèse F_0 donnée, cette fonction est définie comme suit :
$$g_h = \begin{cases} A_h & \text{si } h \in \text{HPS} \\ 0 & \text{sinon} \end{cases} \tag{5.13}$$
où A_h est l'amplitude de la sinusoïde détectée correspondant au partiel h.
- On calcule $G[k] = |\mathcal{F}[g_h]|$ le spectre d'amplitude du signal g_h.
- le critère est donné par :
$$\text{MBW} = \frac{1}{K/2}\sqrt{2\frac{\sum_{k=1}^{K/2} k|G[k]|^2}{\sum_{k=1}^{K/2}|G[k]|^2}} \tag{5.14}$$
où K est la taille de la transformée de Fourier discrète de g_h.

Le critère MBW étant une indication sur la bande passante moyenne, une valeur faible correspond à une enveloppe spectrale lisse et donc une concentration de l'énergie dans les basses fréquences. Ce critère est discriminant lorsque l'hypothèse sur le candidat F_0 est fausse et correspond à un sous-harmonique (*e.g.* $F_0/2$). En effet, cela aura pour effet d'ajouter des 0 dans la fonction g_h ce qui aura pour effet d'augmenter la valeur de MBW.

Ce critère peut aussi être utilisé comme critère d'optimisation pour la reconstruction de l'amplitude des partiels manquants d'une source harmonique.

b.2) Centroïde spectrale

La centroïde spectrale[2] est associée à la brillance d'un son d'après [SWT04]. La brillance est un descripteur du timbre d'un son qui correspond à la *clarté* ou la précision de la perception de la fréquence fondamentale. Le calcul de ce descripteur est donnée par :

$$\text{SPC} = (2/B)\sqrt{2\frac{\sum_{h=1}^{H} hA_h^2}{\sum_{h=1}^{H} A_h^2}} \tag{5.15}$$

avec $B = \frac{F_{90}}{2F_{\min}}$ ou F_{90} correspond à la fréquence où 90% de l'énergie du spectre est concentrée (*Spectral Rolloff*) et $F_{\min}$ est la plus petite fréquence observée dans le spectre.

[2] *SPectral Centroid* (SPC)

Ce critère est un descripteur du timbre qui a donc pour effet de pénaliser les enveloppes spectrales anormales (*e.g.* enveloppe croissante ou formants dans les hautes fréquences).

c) Synchronicité

La synchronicité mesure la différence entre le temps de chaque partiel et le centre de gravité temporel du HPS. Ainsi, si toutes les sinusoïdes appartiennent bien à la même source, on s'attend que tous les partiels soient synchronisés et que la valeur de ce descripteur soit proche de 0. Il s'agit donc d'un indicateur sur la qualité des pics affectés au HPS considéré. Ce descripteur se calcule par l'expression suivante :

$$\text{SYNC} = \sqrt{\sum_{h \in HPS} \left\{ (\bar{t}_h - T_{HPS})^2 w[h] \right\}} \tag{5.16}$$

avec $\bar{t}_h$ le temps moyen du partiel h donné par :

$$\bar{t}_h = \frac{\int_{-\infty}^{+\infty} t|s(t)|^2 dt}{\int_{-\infty}^{+\infty} |s(t)|^2 dt} \tag{5.17}$$

$$= \frac{\int_{-\infty}^{+\infty} -\phi'(\omega)|S(\omega)|^2 \, d\omega}{\int_{-\infty}^{+\infty} |S(\omega)|^2 d\omega}, \tag{5.18}$$

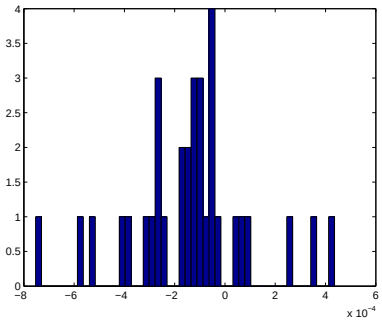
où $\phi = \angle S(\omega)$ et $-\phi'(\omega) = -\frac{d\phi(\omega)}{d\omega}$ correspond au délai de groupe donné en secondes. T_{HPS} est la moyenne des $\bar{t}_h$ appartenant au HPS considéré pondérée par l'énergie correspondant à chaque partiel.
w est un vecteur normalisé de pondération permettant d'accorder plus ou moins d'importance à chaque partiel en fonction de son amplitude et de sa fiabilité. Pour les partiels incertains (affectés à plusieurs sources harmoniques) on fixe $w[h] = 0$.

Comme on peut voir sur la figure 5.7, les instants t_h de chaque partiel sont centrés en zéro et s'éloignent du centre de gravité lorsque plusieurs sources sont combinées (*cf.* figure. 5.7c)). Malheureusement les valeurs des t_h sont très faibles et ce critère est donc très sensible. Pour être discriminant, il dépend de l'importance accordée à chaque partiel (d'où la fenêtre de pondération w) utilisée pour calculer ce critère.

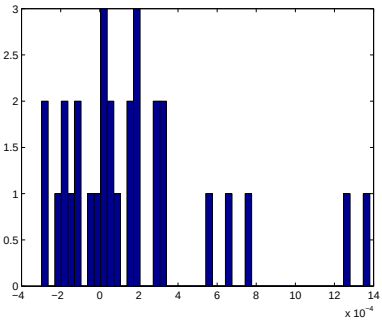
Pour être calculé, nous avons comparé les approches suivantes :

- Le centre de gravité dans le domaine temporel, basé sur l'équation (5.17) obtenu en synthétisant le partiel à partir de ses paramètres sinusoïdaux,
- la réallocation spectrale, le temps t_h est directement donné par Δ_t à partir de l'équation (5.5),
- l'expression dans le domaine spectral donnée par (5.18) calculée pour tous les indices appartenant au partiel comme proposé dans la méthode de Yeh[YRR10].

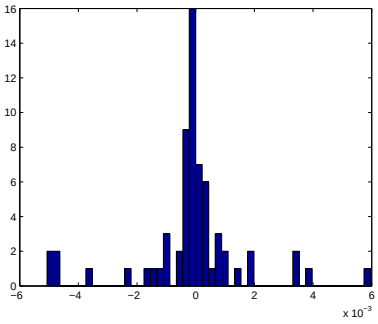
Comme attendu, la méthode de réallocation donne les valeurs les plus précises centrées en 0 (*cf.* figure 5.7), la méthode temporelle est très proche car basée sur l'estimation



a) pour la première source harmonique ($\Delta t \approx 1.1 \cdot 10^{-3}$)



b) pour la seconde source harmonique ($\Delta t \approx 1.6 \cdot 10^{-3}$)



c) combinaison des deux sources harmoniques ($\Delta t \approx 1.1 \cdot 10^{-2}$)

FIGURE 5.7 – Histogramme des t_h en fonction des combinaisons entre deux sources harmoniques. Les figures présentent le nombre d'occurrences pour chaque valeur du temps t_h observé mesuré en secondes et (Sons de saxophone jouant respectivement les notes Ré$_\flat$ et Mi).

des paramètre sinusoïdaux dont le délai de groupe. La méthode spectrale proposée initialement par Yeh dépend de la résolution du spectre discret et est basée sur l'approximation de la dérivée du spectre de phase $\hat{\phi}'[k] = \frac{\phi[k+1]-\phi[k]}{F_s/N}$ obtient des valeurs moins précises. Nous avons donc choisi d'utiliser la réallocation spectrale dans notre implémentation.

5.1.4 Estimation de la polyphonie

Le nombre de sources harmoniques actives simultanément est estimé itérativement en optimisant la fonction de score qui est une combinaison linéaire des critères proposés précédemment.

$$S(\mathrm{HPS}_m) = p_1\mathrm{HAR} + p_2\mathrm{MBW} + p_3\mathrm{SPC} + p_4\mathrm{SYNC}. \tag{5.19}$$

Le choix optimal pour les facteurs appliqués à chaque descripteur est donné par $p = (0.3774, 0.2075, 0.2075, 0.2075)$. Ce résultat a été obtenu en effectuant des tests sur une base de données d'évaluation [YRR10]. Dans notre implémentation, nous n'avons pourtant pas observé de différence significative sur les résultats après l'application de légère modification sur ces paramètres (*e.g.* $p_i = 1/4 \forall i \in [1;4]$). Ne disposant pas la base

de test pour reproduire les résultats décrits dans [YRR10], nous avons conservé la valeur proposé pour ce paramètre.

Ainsi, la fonction (5.19) doit être minimisée pour obtenir le meilleur candidat F_0 donné par $\arg\min_{\text{HPS}}(S(HPS_m))$.

L'estimation de la polyphonie s'effectue de manière itérative en sélectionnant les candidats F_0 détectés par ordre croissant du score obtenu en appliquant la fonction (5.19). Pour chaque candidat évalué, la différence de score avec celui du candidat précédent $\Delta S = S_M - S_{M-1}$ est calculée. Si ΔS est trop important et dépasse un seuil fixé, alors le candidat F_0 correspondant est ignoré et l'algorithme s'arrête. Il en es de même si l'énergie portée par le dernier HPS est trop faible (inférieur à un seuil de bruit fixé).

Dans notre implémentation, nous avons choisi d'ignorer l'estimation des multiples d'un candidat F_0 (présence simultanée de F_0 et de hF_0 pour $h \geq 2$). Ainsi, même si, une fréquence F_0 et un ou plusieurs de ses multiples entier son présents en même temps dans le signal, seul F_0 est considéré. Une solution pour l'estimation de la présence des F_0 confondus consisterait à optimiser le descripteur MBW (en lissant l'enveloppe) du HPS correspondant. Cette optimisation pourrait se faire par soustraction de l'énergie portée par les candidats hF_0 pour $h \geq 2$ correspondants aux partiels du HPS de F_0.

5.1.5 Évaluation comparative du système de transcription proposé

a) Protocole expérimental

On dispose d'une base de données[3] utilisée dans [HDB11] composée de 10 pièces musicales contenant chacune 4 instruments de musique distincts. Le signal correspondant à chaque instrument a été obtenu par synthèse à partir de la transcription de référence enregistrée au format MIDI et est donc parfaitement alignée. Les signaux de musique utilisés sont monophoniques codés sur 16bits et ont été échantillonnés à une fréquence $F_s = 11025$Hz. Les mélanges sonores composés par tous les instruments ont été calculés par addition des signaux.

Pour toutes nos évaluations, nous utilisons la méthodologie d'évaluation décrite dans la section 2.2.4. Les fonctions d'erreurs qui y sont proposées sont couramment utilisées par la communauté MIR [BED09].

Nous comparons ainsi les scores obtenus avec les méthodes suivantes :

- **FM** (Fourer-Marchand) est une implémentation inspirée des travaux de Yeh [Yeh08] détaillée précédemment,

- **YIN** est une méthode F_0 simple (monophonique) proposée par de Cheveigné et Kawahara [dCK02]. Nous avons utilisé le code de référence accessible sur la page de l'auteur[4],

- **KL-03** est une implémentation personnelle de la méthode d'estimation F_0 multiple proposée par Klapuri en 2003 [Kla03]. Bien que non officielle, cette implémentation se base sur l'article de l'auteur qui est suffisamment détaillé pour permettre sa mise en oeuvre. Nous n'avons donc apporté aucune modification ni optimisation par rapport à la description faite dans [Kla03],

[3] http://www.romain-hennequin.fr/database.zip

[4] http://audition.ens.fr/adc/sw/yin.zip

- **BBV-10** est la méthode d'estimation F_0 multiple proposée par Bertin *et al.*dans [BBV10] et repose sur l'utilisation de la NMF. Une implémentation officielle est disponible librement avec la boîte à outils DESAM [LBD+10].

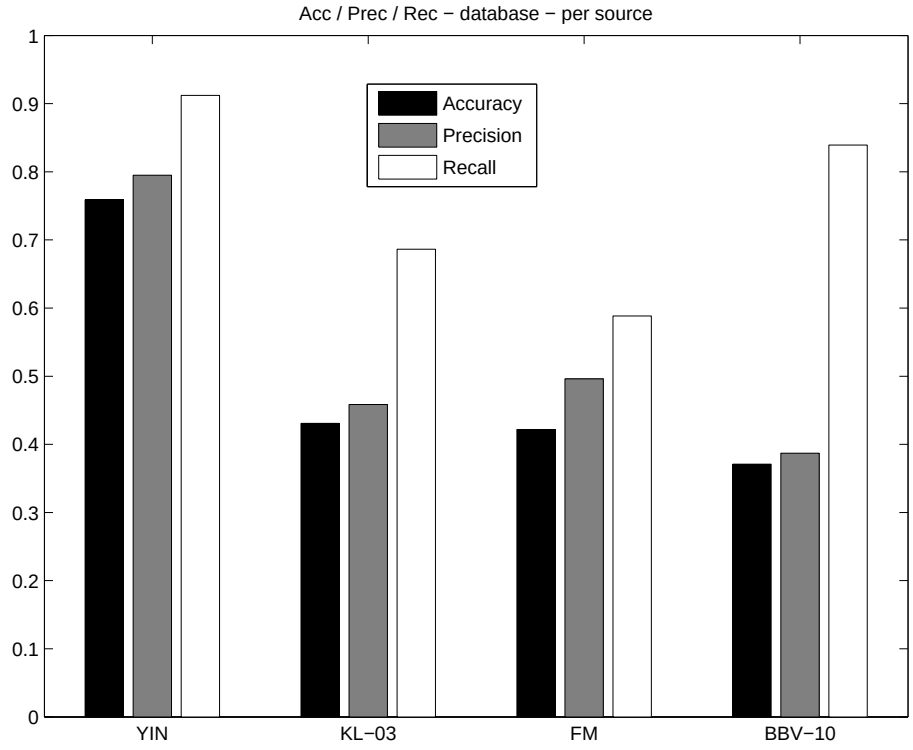
Pour l'utilisation de chaque méthode, nous avons tenté lorsque c'était possible, de choisir les réglages permettant d'obtenir globalement la meilleure précision. Dans le cas de la méthode **FM**, nous utilisons par exemple des tailles de fenêtre d'analyse de 92ms combinée avec un recouvrement de 50%. Pour la méthode **YIN**, celle-ci ne retourne qu'un seul candidat F_0 même si le signal analysé est polyphonique.

b) Présentation des résultats

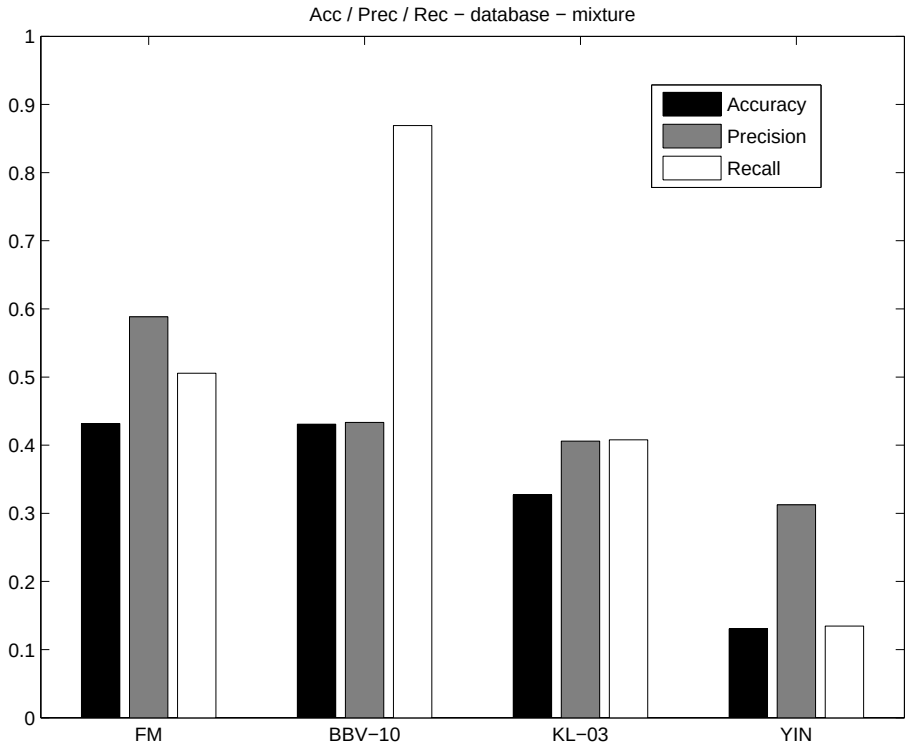
Sur les figures 5.8, 5.9 et 5.10, nous présentons les scores de performance (*cf.* section 2.2.4) obtenus par les méthodes choisies.

Comme attendu, la méthode **YIN** est la plus efficace pour l'estimation depuis les signaux des sources isolées (sources monophoniques la plupart du temps). Pour l'estimation polyphonique à partir du mélange composé de toutes les sources, la technique que nous proposons obtient le meilleur score *Precision* ainsi qu'une erreur plus faible dans le cas polyphonique que l'ensemble des méthodes évaluée. La méthode **BBV-10** a tendance a sur-évaluer la polyphonique. Cela a pour effet de trouver plus de candidats F_0 que les autres méthodes avec un taux d'erreurs d'insertion plus élevé.

Ainsi, pour cette évaluation, **YIN** est la méthode la plus précise dans le cas monophonique. Dans le cas polyphonique **FM** semble être la méthode la plus précise mais obtient des résultats très proches de ceux obtenus par **BBV-10**.

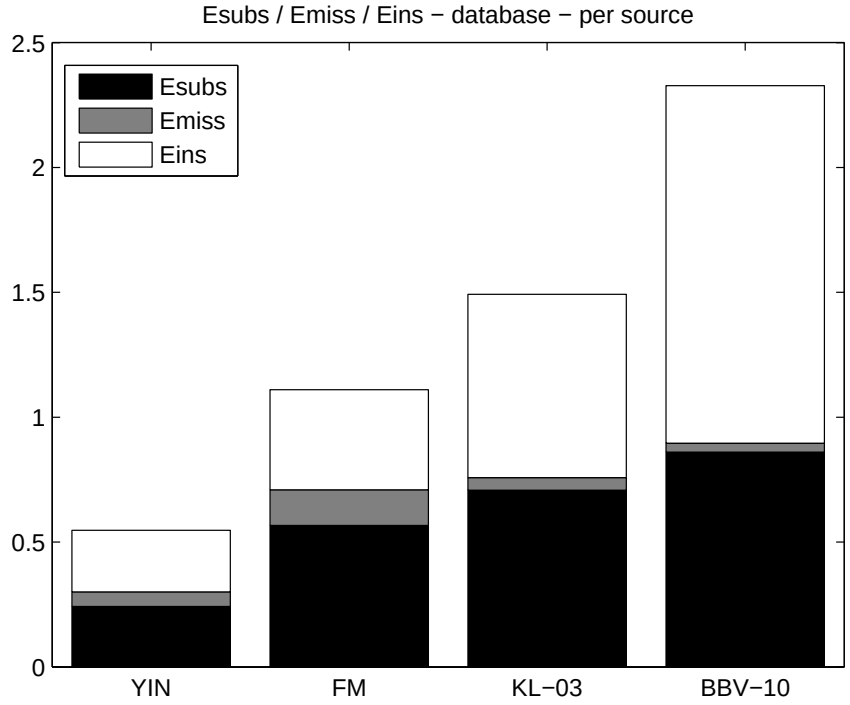


a) depuis les source isolées

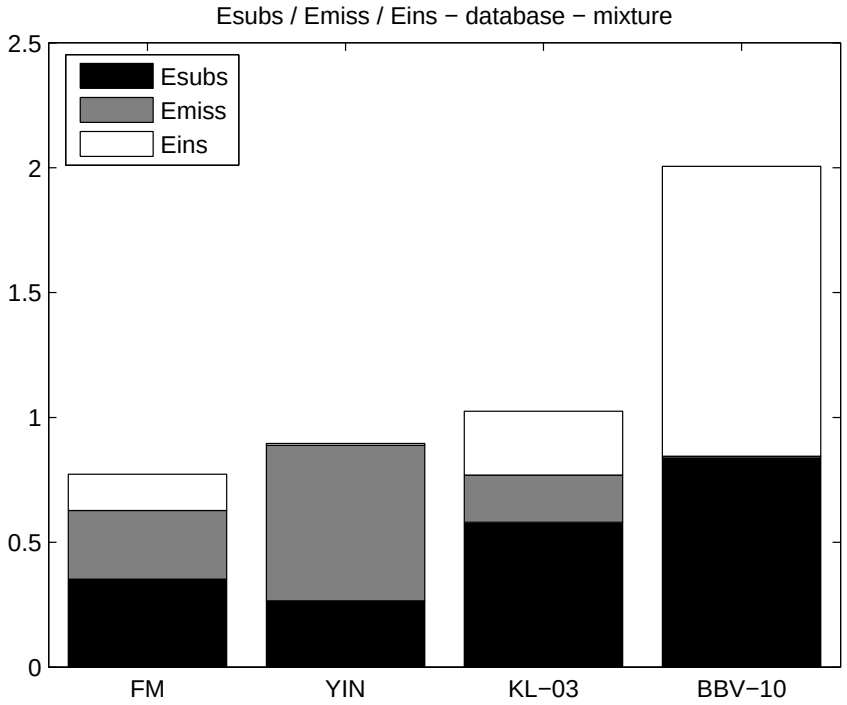


b) depuis le mélange

FIGURE 5.8 – Comparaison des scores *Precision, Accuracy* et *Recall* obtenus pour l'estimation des F_0 en utilisant les méthodes proposées. Les résultats sont triés par ordre décroissant du score *Accuracy*.

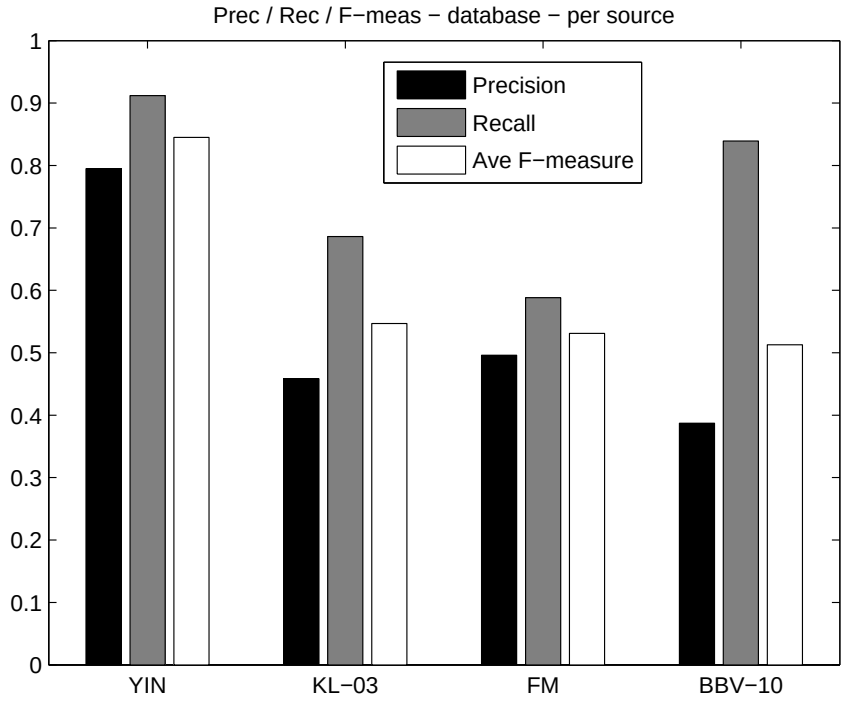


a) depuis les source isolées

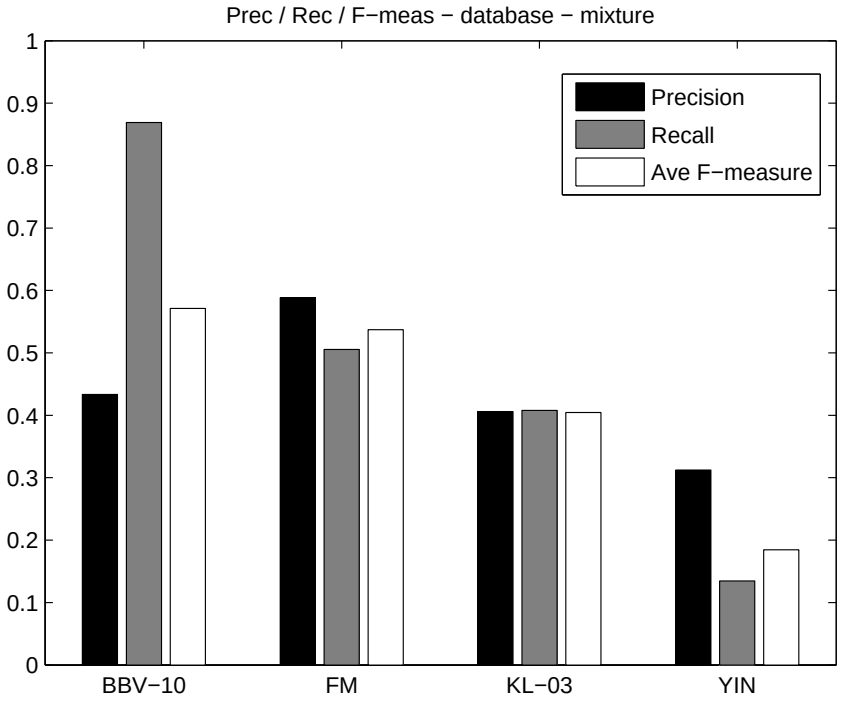


b) depuis le mélange

FIGURE 5.9 – Comparaison du taux d'erreur d'édition (substitution, oubli et insertion) mesuré par les méthodes d'estimation F_0. Les résultats sont triés par ordre croissant de l'erreur totale.



a) depuis les source isolées



b) depuis le mélange

FIGURE 5.10 – Comparaison des scores *Accuracy*, *Recall* et *F-Measure* obtenus pour l'estimation F_0 multiple en utilisant les méthodes proposées. Les résultats sont triés par ordre décroissant de *F-Measure*.

5.2 La transcription polyphonique informée

Comme vu précédemment, il est quasiment impossible avec les méthodes de l'état de l'art d'obtenir une transcription sans erreur à partir d'un signal audio. C'est pourquoi, on se propose désormais de définir une stratégie permettant de combiner une méthode de transcription avec de l'information supplémentaire, nécessaire et si possible de taille minimale, permettant de trouver la transcription exacte Nous nous plaçons ainsi dans la configuration où nous supposons connues la transcription exacte de chaque source harmonique avant le processus de mélange (on décrira par la suite comment une telle configuration peut être mise en oeuvre en pratique). Le problème de transcription est ainsi posé dans une configuration codeur / décodeur similaire à celle proposée dans la figure 3.9 exploitée dans le chapitre 4. Il s'agit désormais de minimiser l'information codée qui sera transmise avec le mélange analysé tout en garantissant une transcription exacte au décodeur. Dans cette optique, il convient de définir une stratégie de codage efficace exploitant au maximum l'information apportée par les méthodes d'analyse classique tout en corrigeant les erreurs les plus fréquentes avec le minimum d'information.

5.2.1 Formulation du problème

On considère un mélange linéaire instantané discret composé de plusieurs sources (ou instruments) harmoniques mélangées à du bruit pouvant être exprimé par :

$$x[n] = \sum_{k=1}^{K} s_k[n] + b[n]. \tag{5.20}$$

D'après (2.23), à chaque signal de source $s_k[n]$ correspond un ensemble de notes actives à chaque instant. A chaque note de musique perçue correspond une fréquence fondamentale F_0. Chaque note est codée par son *pitch* MIDI (*cf.* section 1.3.3). L'échelle MIDI permet de représenter 128 notes différentes codées chacune avec 7 bits. Cette résolution s'avère suffisante pour effectuer des opérations sur le signal comme il a été proposé dans [HDB11, ES06, GSMA10] pour la séparation de sources informée par la partition (*cf.* section 5.3).

Nous avons vu dans la section 5.1 qu'un estimateur F_0 multiple peut, malgré ses erreurs, donner une approximation à chaque instant de l'ensemble des notes actives dans un mélange musical sans aucune information *a priori* sur les règles harmoniques utilisées en musique. Cependant un tel estimateur n'est pas capable sans information *a priori* de séparer la partition correspondante à chaque instrument présent dans le mélange.

Comme un système de transcription obtient toujours de meilleurs résultats (*cf.* section 5.2.2) lorsque la polyphonie est plus faible (*cf.* section 2.1.2), il est possible lorsque l'on possède les signaux isolés qui composent un mélange, de les utiliser pour obtenir une estimation plus précise de la transcription du mélange. Ainsi, en appliquant sur chaque signal isolé une estimation de la transcription, on obtient une transcription de référence qui pourra être utilisée dans une configuration de type codeur / décodeur similaire aux approches proposées dans la section 3.3.1.

Notre objectif est de proposer un système combinant simultanément un estimateur F_0 multiple classique (non informé) avec du codage pour améliorer les performances globales. Ce système basé sur une configuration codeur / décodeur possède les objectifs suivants :

- garantir que l'estimation instantanée des notes actives dans le mélange noté $\hat{\Omega}[n]$ (sans séparer les instruments) est identique à l'ensemble de référence noté $\Omega[n]$ (calculé à partir des pistes séparées),

- affecter à chacun des instruments un ensemble de notes qui lui correspond noté $\Pi_k[n]$,
- utiliser pour l'information complémentaire une quantité d'information plus faible que celui nécessaire pour le codage de la transcription (car dans ce cas, l'estimateur F_0 multiple appliqué sur le mélange devient inutile).

L'estimateur F_0 multiple classique est utilisé à la fois au codeur et au décodeur et permet d'obtenir une approximation de la transcription attendue avec des erreurs. L'information transmise codée sert à corriger les erreurs et à affecter les candidats F_0 à chacune des sources correspondantes.

Pour la suite, nous utilisons les notations suivantes. $\Pi_k^{(t)}$ représente l'ensemble des notes (code du pitch MIDI) actives à l'instant t pour la source k et $|\Pi_k^{(t)}| \in [0, L_k]$ est le cardinal (ou le nombre d'éléments) de cet ensemble. $\Omega^{(t)} = \bigcup_{k=1}^{K} \Pi_k^{(t)}$ est l'ensemble de toutes les notes actives (chaque élément est unique) pour toutes les sources confondues et $\hat{\Omega}^{(t)}$ est son estimation calculée à partir d'un estimateur F_0 multiple appliqué sur le signal de mélange $x[n]$. D'après la configuration proposée pour l'analyse informée (*cf.* figure 3.9), les paramètres de référence $\Pi_k^{(t)}$ sont supposés connus au codeur. Ils peuvent être plus facilement estimés à partir d'un estimateur F_0 multiple appliqué sur chaque signal source $s_k[n]$ avant le processus de mélange.

L'information supplémentaire notée $\mathcal{I}^{(t)}$ calculée au codeur et utilisée au décodeur est calculée à l'aide de l'algorithme prop dans la section b.1). Cette information, est ensuite combinée avec l'estimation $\hat{\Omega}^{(t)}$ obtenue en analysant le mélange $x[n]$ avec l'estimateur F_0 multiple choisi.

5.2.2 Analyse des erreurs des systèmes de transcription polyphonique

Les codeurs efficaces utilisés en compression ou plus généralement pour le codage de source (*cf.* chapitre 3) tiennent compte des propriétés statistiques des données représentées. Ainsi, les données ayant une fréquence (ou une probabilité) d'apparition plus élevée ont une entropie plus faible et peuvent être codées avec moins de bits que les données moins fréquentes.

Comme dans le chapitre 4, on cherche ici à minimiser le nombre de bits utilisés permettant de corriger les erreurs tout en tenant compte de l'information apportée par un estimateur.

Pour cela, on se propose d'analyser en détails les erreurs obtenues avec les systèmes de transcription proposés précédemment. Pour cela nous choisissons le protocole expérimental décrit dans la section 5.1.5 utilisant la même base de données ainsi que les même paramètres pour les 4 méthodes d'estimation F_0. Ainsi les figures 5.11, 5.12, 5.13 et 5.14 présentent pour chacune des méthodes la répartition des candidats F_0 estimés en distinguant les erreurs d'octave (*Overtone*) pour lesquelles la note estimée porte le même nom mais se trouve à une octave différente et les erreurs de substitution mesurées en demi-tons qui ne sont pas des erreurs d'octave. En effet, certaines applications comme la détection d'accord peuvent tolérer les candidats F_0 présentant une erreur d'octave pour lesquels le nom de la note correspondante a bien été estimé. Les erreurs de candidat manquant (*Missing*) ou les fausses détections (*False Alarm*) sont spécifiques à l'estimation F_0 multiple et permettent de déterminer les erreurs d'estimation de la polyphonie.

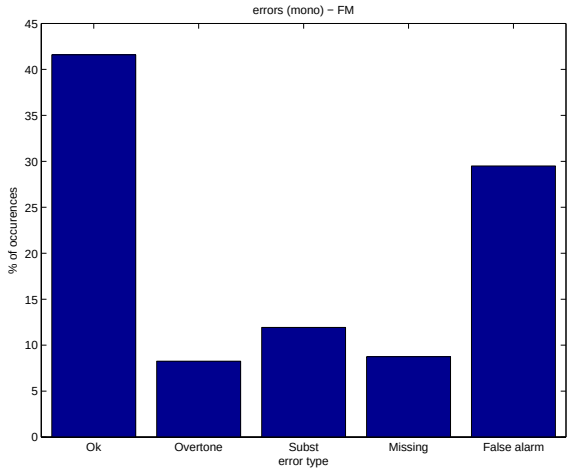
D'après les histogrammes obtenus (*cf.* figures 5.11, 5.12, 5.13 et 5.14), on remarque que la distribution des erreurs de substitution et d'octave varie en fonction de la méthode utilisée. On note aussi que certaines méthodes comme **BBV-10** et **KL-10** ont tendance à surestimer la polyphonie avec un taux de fausses détections (*False Alarm*) souvent supérieur au taux de candidats correctement estimés (*OK*). D'après ces observations, la mise en place d'une stratégie de codage générique pour la correction des erreurs d'estimation de chaque candidat F_0 basée sur l'application de transformations simples ne semble donc pas évidente. En effet, leur application nécessite d'identifier le type d'erreur, puis d'appliquer la transformation correspondante, ce qui risque d'augmenter la taille des données nécessaires. On se propose donc de résoudre ce problème par 2 systèmes de codage décrits ci-après permettant de corriger les erreurs les plus fréquentes (fausses détections) avec un codage minimal (1 seul bit).

5.2.3 Systèmes de codage proposés

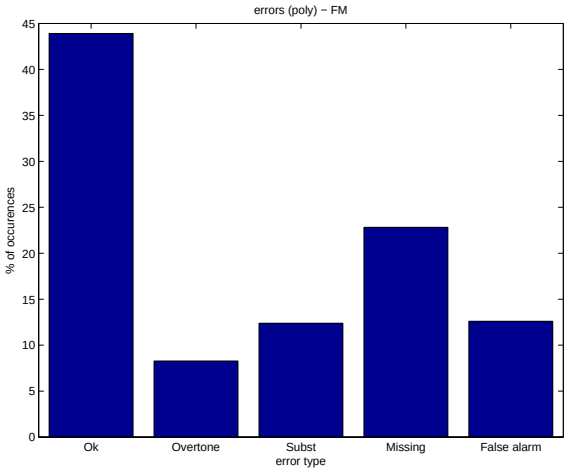
Les principales erreurs sont les insertions et les suppressions [PE06] qui correspondent respectivement à l'ajout de mauvais candidats F_0 ou le fait que certains candidats F_0 soient absents. En pratique ces type d'erreurs peuvent être corrigés de la manière suivante :

- Suppressions : tout candidat F_0 manquant doit être inséré en codant la valeur de la hauteur correspondante. Ce codage peut être effectué de manière relative en substituant la valeur de la hauteur d'une note présente non valide, soit de manière absolue en indiquant la hauteur de la note.
- Insertions : les mauvais candidats F_0 doivent être supprimés par invalidation, par exemple en utilisant un bit pour coder l'information (vrai/faux). La quantité totale d'information utilisée dépend ainsi du nombre de candidats F_0 détectés par le système de transcription. Sa quantité peut être limitée en fixant une borne sur le nombre maximum de candidats détectés avec le risque d'ignorer d'éventuels candidats valides qui devront être insérés par la suite.

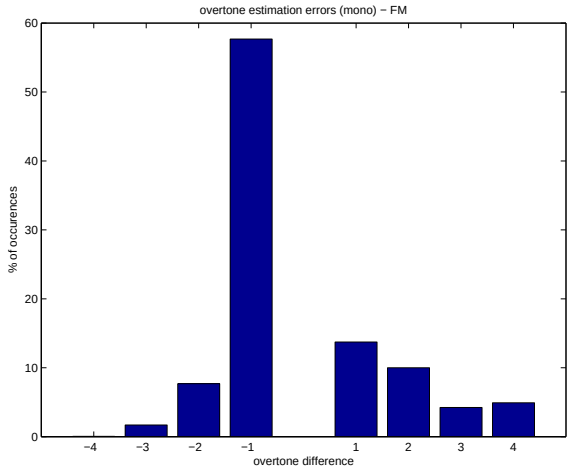
D'après ces observations, on propose deux systèmes de codage basés sur une analyse à court terme des signaux musicaux. On suppose que l'on se place dans la configuration de type codeur / décodeur décrite dans la figure 3.9 où l'on dispose de la transcription exacte uniquement au codeur et d'un système de transcription classique (non informé). Chacun des système de codage proposé génère au codeur un code binaire noté $\mathcal{I}^{(t)}$ qui est ensuite transmis au décodeur en même temps que le signal d'observation. Le signal analysé est composé de plusieurs instruments. On note $\Omega^{(t)}$ l'ensemble des F_0 de référence et $\hat{\Omega}^{(t)}$ l'ensemble des F_0 estimés actifs dans le mélange, toutes sources confondues à l'instant t. $\tilde{\Omega}^{(t)}$ est l'ensemble F_0 calculé au décodeur obtenu par transformation du couple $(\hat{\Omega}^{(t)}, \mathcal{I})$.



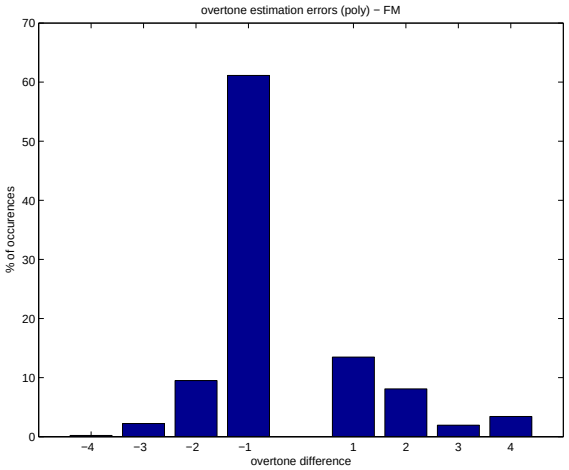
a) répartition des candidats F_0 estimés à partir des source séparées (monophonique)



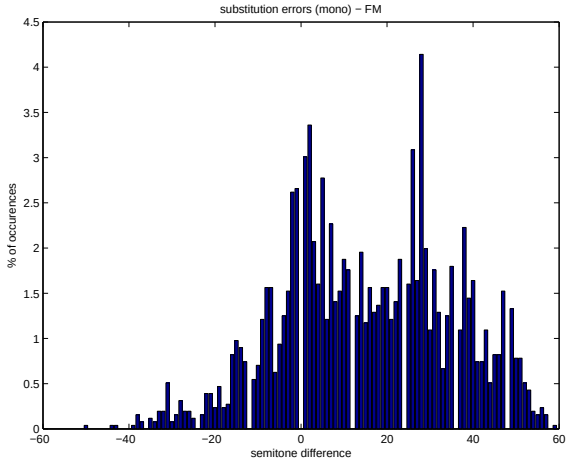
d) répartition des candidats F_0 estimés depuis le mélange(polyphonique)



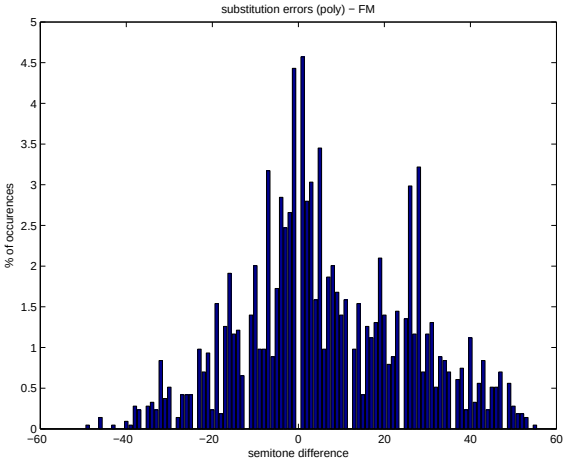
b) erreurs d'octave (monophonique)



e) erreurs d'octave (polyphonique)

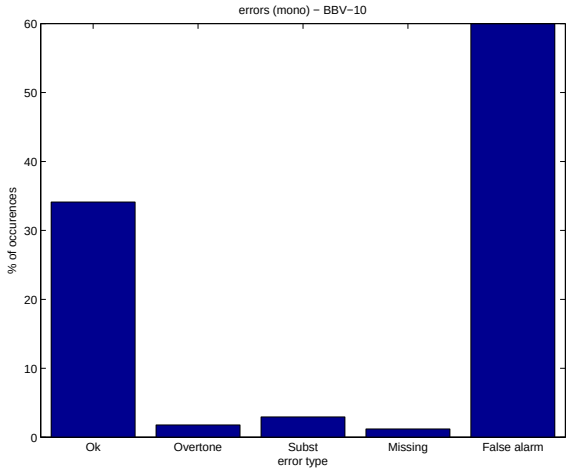


c) erreurs de substitution (monophonique)

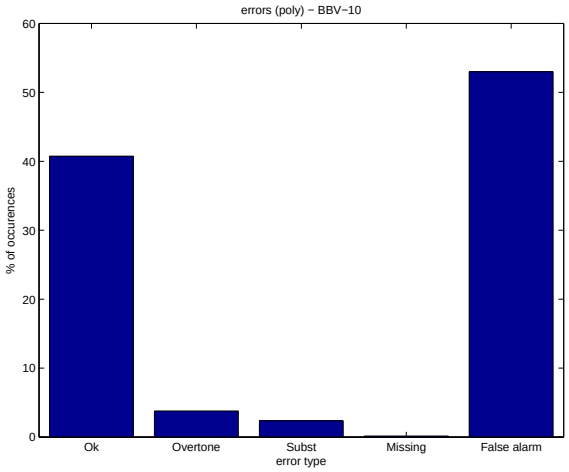


f) erreurs de substitution (polyphonique)

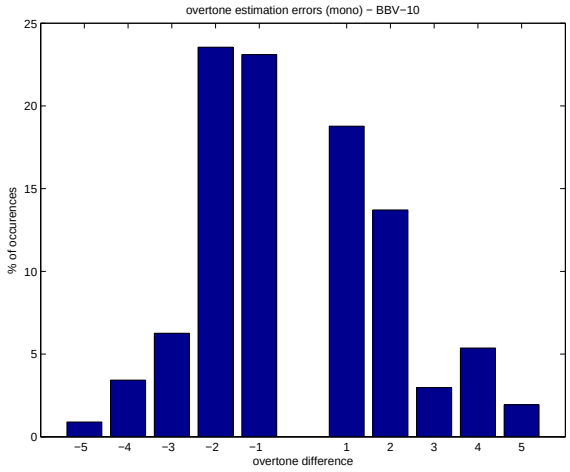
FIGURE 5.11 – Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **FM** proposée. Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite.



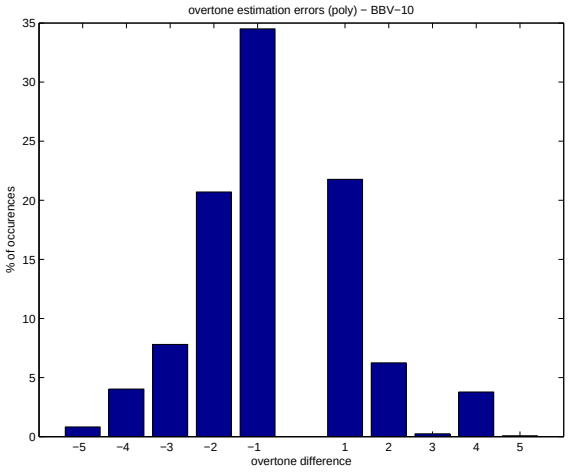
a) répartition des candidats F_0 estimés à partir des source séparées (monophonique)



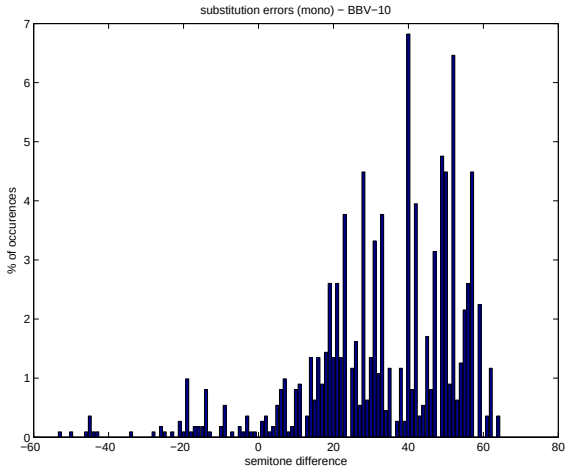
d) répartition des candidats F_0 estimés depuis le mélange (polyphonique)



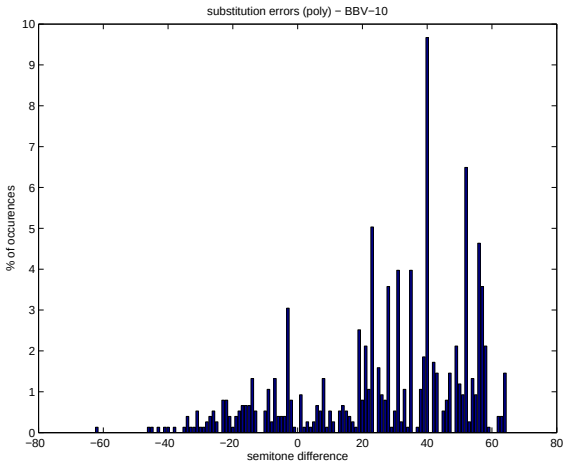
b) erreurs d'octave (monophonique)



e) erreurs d'octave (polyphonique)

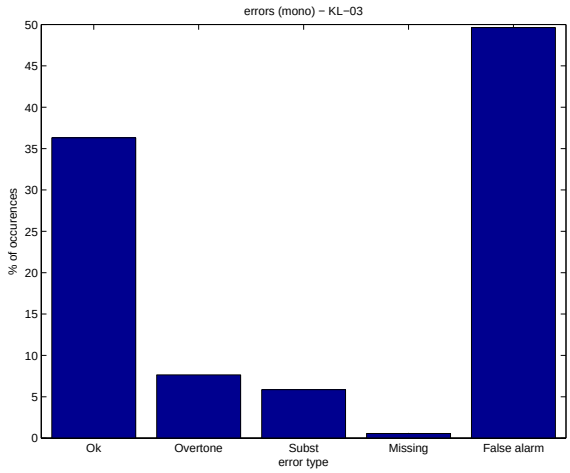


c) erreurs de substitution (monophonique)

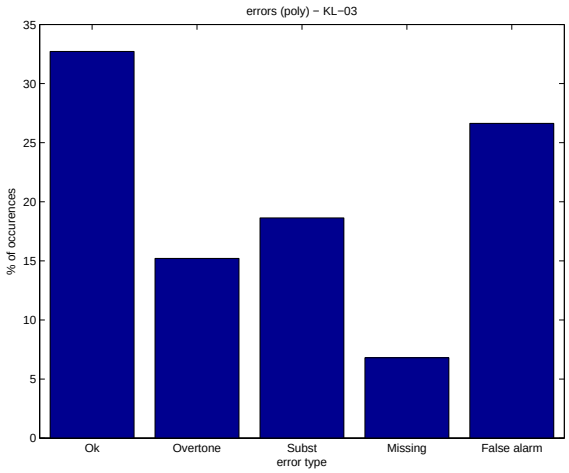


f) erreurs de substitution (polyphonique)

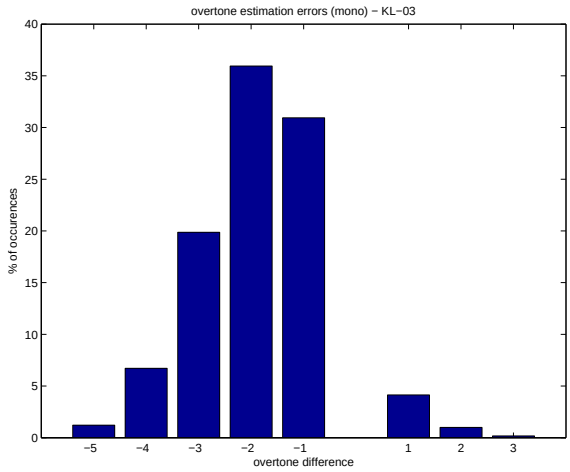
FIGURE 5.12 – Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **BBV-10** [BBV10].Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite.



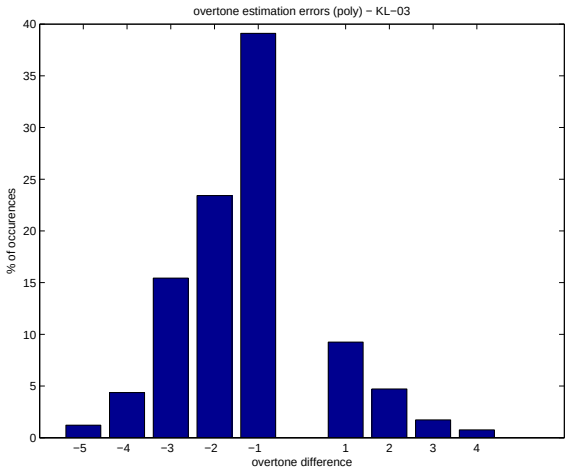
a) répartition des candidats F_0 estimés à partir des source séparées (monophonique)



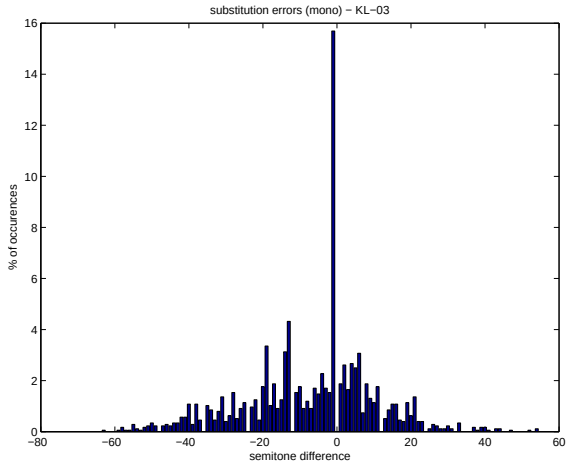
d) répartition des candidats F_0 estimés depuis le mélange (polyphonique)



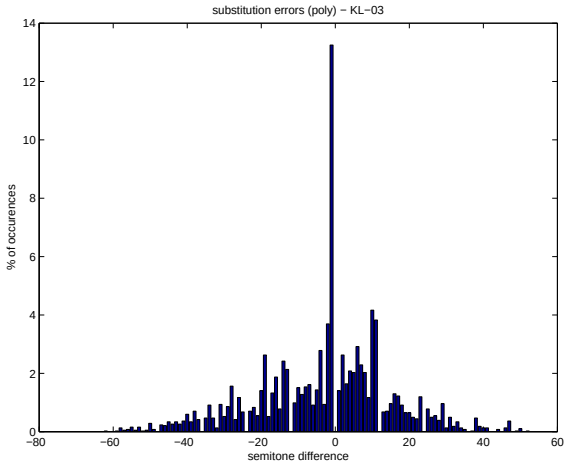
b) erreurs d'octave (monophonique)



e) erreurs d'octave (polyphonique)

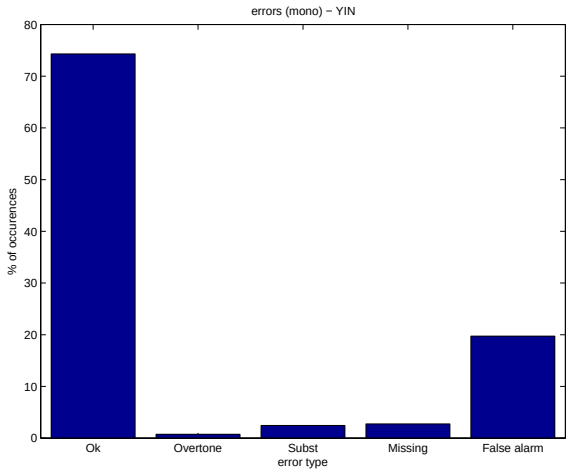


c) erreurs de substitution (monophonique)

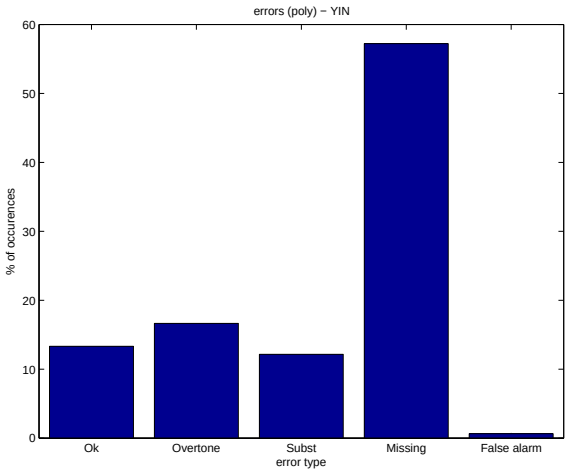


f) erreurs de substitution (polyphonique)

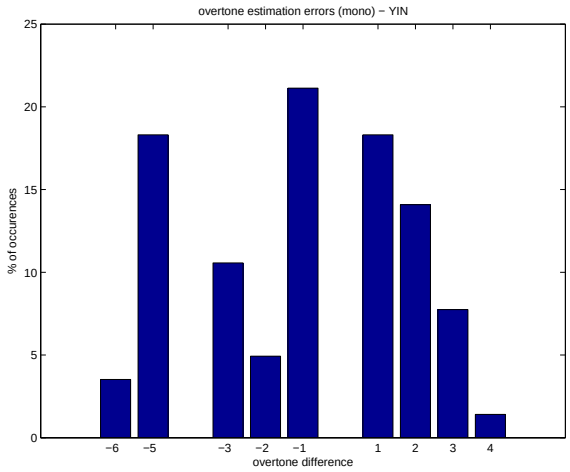
FIGURE 5.13 – Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **KL-03** [Kla03]. Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite.



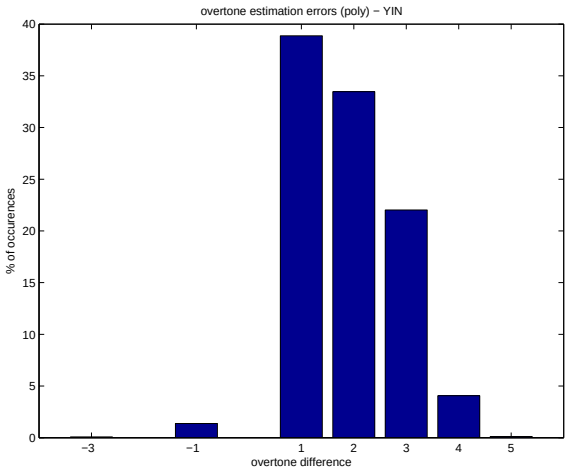
a) répartition des candidats F_0 estimés à partir des source séparées (monophonique)



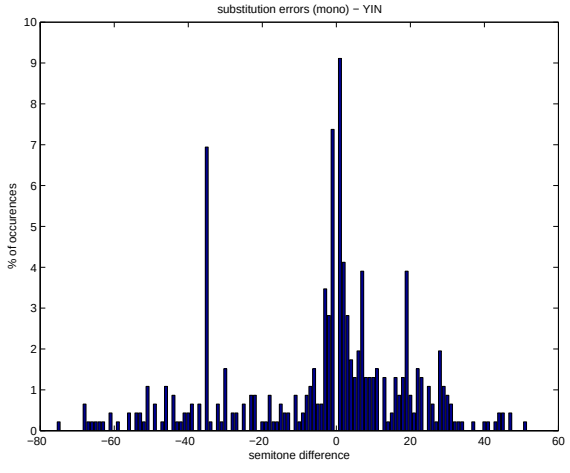
d) répartition des candidats F_0 estimés depuis le mélange (polyphonique)



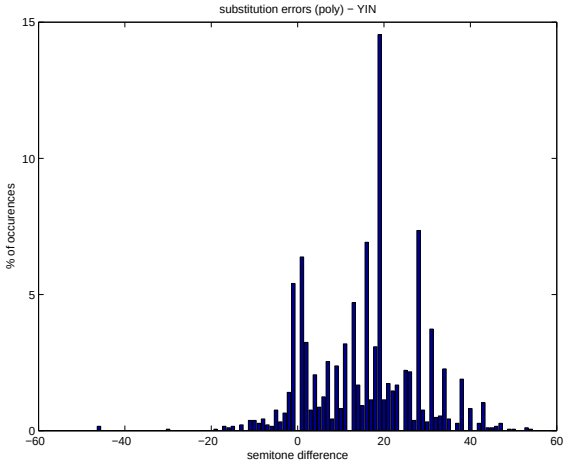
b) erreurs d'octave (monophonique)



e) erreurs d'octave (polyphonique)



c) erreurs de substitution (monophonique)



f) erreurs de substitution (polyphonique)

FIGURE 5.14 – Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **YIN** [dCK02].Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite.

a) Système M1 par validation, suppression et insertion des candidats F_0 à court terme

Ce codage proposé considère un ensemble des candidats F_0 actifs à chaque instant noté $\Omega^{(t)}$. En fonction des candidats proposés par l'estimateur, un seul bit est utilisé pour valider ou non chacune des transformations suggérées par l'estimateur. Les ordres d'insertion/suppression manquants sont ensuite ajoutés en dernier par codage. Le code ainsi généré dépend de l'estimation $\hat{\Omega}^{(t)}$ supposée identique au codeur et au décodeur.

Le codeur proposé est décrit par l'algorithme 4 qui génère un mot $\mathcal{I}$ qui est envoyé au décodeur en même temps que le signal de mélange $x(t)$. L'information $\mathcal{I}$ garantit de pouvoir retrouver $\Omega^{(t)}$ à partir du signal de mélange combiné à un estimateur.

a.1) Codeur

Cet algorithme calcule à chaque instant t les ensembles P^{insert} et P^{suppr} en utilisant l'estimation $\hat{\Omega}^{(t)}$ obtenue en appliquant le système de transcription sur le signal de mélange $x(t)$. Les ensembles P^{insert} et P^{suppr} permettent d'effectuer toutes les transformations (d'insertion et de suppression de candidats F_0) permettant d'obtenir $\hat{\Omega^{(t)}}$ à partir de $\tilde{\Omega}^{(t-1)}$. Pour chaque transformation, le code $\mathcal{I}$ est mis à jour en y ajoutant 1 ou 0 permettant d'indiquer si l'acceptation ou non de chaque transformation.

a.2) Décodeur

Le décodeur reçoit l'information $\mathcal{I}$ calculée par l'algorithme 4 ainsi que le signal de mélange $x(t)$. Le décodage consiste à estimer la transcription à partir du mélange puis de corriger les estimations grâce à $\mathcal{I}$. Ainsi, l'algorithme de décodage proposé reprend les même étapes que le codeur mais en lisant cette fois $\mathcal{I}$.

Algorithme 4 Codeur basé sur les événements insertion/suppression

données : $\Omega^{(t)}$:transcription de référence, $x[n]$ signal de mélange
sorties : $\mathcal{I}$: information transmise au décodeur
variables : $\tilde{\Omega}^{(t)}$: transcription corrigée

Calcul de $\hat{\Omega}^{(t)}$ à partir de $x[n]$ le signal d'observation
$\mathcal{I} \leftarrow ()$
$t \leftarrow 0$
$\tilde{\Omega}^{(t)} \leftarrow \varnothing$
tant que $t \leq$ durée **faire**
  {génère les ensembles de candidats P^{insert} et P^{suppr}}
  $P^{\text{insert}} \leftarrow \hat{\Omega}^{(t)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t)}$
  $P^{\text{suppr}} = \tilde{\Omega}^{(t)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t)}$
  {validation des candidats $p \in P^{\text{suppr}}$ et mise à jour de $\tilde{\Omega}^{(t)}$}
  pour $p \in P^{\text{suppr}}$ **faire**
    si $p \notin \Omega^{(t)}$ **alors**
      $\mathcal{I} \leftarrow (\mathcal{I}, 1)$
      $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t)} - p$
    sinon
      $\mathcal{I} \leftarrow (\mathcal{I}, 0)$
    fin si
  fin pour
  {validation des candidats $p \in P^{\text{insert}}$ et mise à jour de $\tilde{\Omega}^{(t)}$}
  pour $p \in P^{\text{insert}}$ **faire**
    si $p \in \Omega^{(t)}$ **alors**
      $\mathcal{I} \leftarrow (\mathcal{I}, 1)$
      $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t)} \bigcup p$
    sinon
      $\mathcal{I} \leftarrow (\mathcal{I}, 0)$
    fin si
  fin pour
  {Codage et mise à jour des candidats manquants P^{manqu}}
  si $\Omega^{(t)} \neq \tilde{\Omega}^{(t)}$ **alors**
    $P'^{\text{insert}} \leftarrow \{p | (p \in \Omega^{(t)}) \wedge (p \notin \tilde{\Omega}^{(t)})\}$
    $P'^{\text{suppr}} \leftarrow \{p | (p \notin \Omega^{(t)}) \wedge (p \in \tilde{\Omega}^{(t)})\}$
    $\mathcal{I} \leftarrow \left(\mathcal{I}, \mathcal{C}\left(P'^{\text{insert}}, P'^{\text{suppr}}\right)\right)$
    $\tilde{\Omega}^{(t)} \leftarrow \Omega^{(t)}$
  fin si
  $\tilde{\Omega}^{(t+1)} \leftarrow \tilde{\Omega}^{(t+1)}$
  $t \leftarrow t + 1$
fin tant que

Algorithme 5 Décodeur corrigeant les événements insertion/suppression générés à partir de l'estimation de la transcription $\hat{\Omega}^{(t)}$.

données : $x[n]$ signal de mélange, $\mathcal{I}$
sorties : $\tilde{\Omega}^{(t)}$: transcription corrigée ($\tilde{\Omega}^{(t)} = \Omega^{(t)}$)

Calcul de $\hat{\Omega}^{(t)}$ à partir de $x(t)$ le signal d'observation
$t \leftarrow 0$
$\tilde{\Omega}^{(t)} \leftarrow \varnothing$
tant que $t \leq$ durée **faire**
  {génère les ensembles de candidats P^{insert} et P^{suppr}}
  $P^{\text{insert}} \leftarrow \hat{\Omega}^{(t)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t)}$
  $P^{\text{suppr}} = \tilde{\Omega}^{(t)} - \hat{\Omega}^{(t)} \bigcap \tilde{\Omega}^{(t)}$
  {validation des candidats $p \in P^{\text{suppr}}$ et mise à jour de $\tilde{\Omega}^{(t)}$}
  pour $p \in P^{\text{suppr}}$ **faire**
    si lecture$(\mathcal{I}) = 1$ **alors**
      $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t)} - p$
    fin si
  fin pour
  {validation des candidats $p \in P^{\text{insert}}$ et mise à jour de $\tilde{\Omega}^{(t)}$}
  pour $p \in P^{\text{insert}}$ **faire**
    si lecture$(\mathcal{I}) = 1$ **alors**
      $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t)} \bigcup p$
    fin si
  fin pour
  {Mise à jour des candidats manquants P^{manqu}}
  $P'^{\text{insert}} \leftarrow \mathcal{D}(\text{lecture}(\mathcal{I}))$
  $P'^{\text{suppr}} \leftarrow \mathcal{D}(\text{lecture}(\mathcal{I}))$
  $\tilde{\Omega}^{(t)} \leftarrow (\tilde{\Omega}^{(t)} \bigcup P'^{\text{insert}}) - P'^{\text{suppr}}$
  $\tilde{\Omega}^{(t+1)} \leftarrow \tilde{\Omega}^{(t+1)}$
  $t \leftarrow t + 1$
fin tant que

b) Système M2 utilisant un suivi temporel

Dans un soucis de réduire la taille de $\mathcal{I}$ nous proposons une version modifiée du système de codage précédent intégrant les deux modifications suivantes :

- codage prédictif de la transcription : on suppose $\tilde{\Omega}^{(t+1)} = \tilde{\Omega}^{(t)}$,
- prise en compte de la polyphonie : chaque source k possède une transcription propre notée $\Pi_k^{(t)}$ pouvant être calculée à partir de $\Omega^{(t)}$.

Le nouvel algorithme proposé considère un mélange $x(t)$ composé de plusieurs sources $s_k(t)$. L'ensemble $\Omega^{(t)}$ correspond à l'ensemble des notes distinctes actives à l'instant t dans le mélange $x(t)$ et l'ensemble $\Pi_k^{(t)}$ est l'ensemble instantané des notes actives pour la source k.

b.1) Codeur

Comme pour l'algorithme 4, on calcule un code binaire $\mathcal{I}^{(t)}$ basé sur les transformations de type *insertion/suppression* ou *prediction* qui doivent être appliquées pour retrouver $\Omega^{(t)}$ à partir de l'estimation $\hat{\Omega}^{(t)}$ donnée par un système de transcription classique. Le codeur utilise comme variable $\tilde{\Omega}^{(t)}$ qui est la transcription corrigée pouvant être calculée à partir de $\mathcal{I}$. Il n'y a donc pas d'autres informations à coder lorsque l'on obtient l'identité $\tilde{\Omega}^{(t)} = \Omega^{(t)}$.

Dans un premier temps (*cf.* figure 5.15), on code sur 1 bit l'égalité entre $\Omega^{(t)}$ et $\Omega^{(t-1)}$. Cette information est utilisée par le décodeur pour déterminer si la prédiction $\tilde{\Omega}^{(t)} \leftarrow \tilde{\Omega}^{(t-1)}$ doit être utilisée ou si $\tilde{\Omega}^{(t)}$ doit être calculé en utilisant l'estimation. Ainsi, l'hypothèse de prédiction utilisée suppose que chaque note active est maintenue à l'instant suivant et qu'aucune nouvelle note n'est ajoutée. Dans ce cas, l'estimation $\hat{\Omega}^{(t)}$ est ignorée, ce qui réduit significativement le temps de calcul et la taille des données nécessaires pour la correction des erreurs. Quand cette égalité n'est pas vérifiée, le même mécanisme basé sur les insertions/suppressions proposé par l'algorithme 4 est utilisé. Cette première partie de l'algorithme est décrit par la figure 5.15. Tous les bits calculés (0 = faux, 1 = vrai) sont concaténés à $\mathcal{I}^{(t)}$. Lorsque toutes les transformations sont traitées, si $\tilde{\Omega}^{(t)} \neq \Omega^{(t)}$, les opérations restantes sont codées dans $\mathcal{I}^{(t)}$.

Dans un deuxième temps (*cf.* figure 5.16), l'ensemble des notes actives pour chaque source noté $\Pi_k^{(t)}$ est codé en utilisant une stratégie similaire. On utilise 1 bit pour déterminer si on utilise la prédiction $\tilde{\Pi}_k^{(t)} \leftarrow \tilde{\Pi}_k^{(t-1)}$. Sinon, les candidats pour *insertion* et *suppression* sont générés à partir de $\tilde{\Omega}^{(t)}$. Ainsi, les candidats qui n'appartiennent pas à l'intersection $\tilde{\Omega}^{(t)} \bigcap \Pi_k^{(t)}$ sont insérés ou supprimés en fonction de leur présence dans $\tilde{\Omega}^{(t)}$ (insertion) ou dans $\Pi_k^{(t)}$ (suppression). Chaque opération est alors validée ou annulée en utilisant un codage sur 1 bit. Il n'est pas nécessaire de coder les candidats F_0 pour chaque source car l'ensemble de tous les F_0 actifs $\tilde{\Omega}^{(t)}$ a été corrigé lors de la première étape.

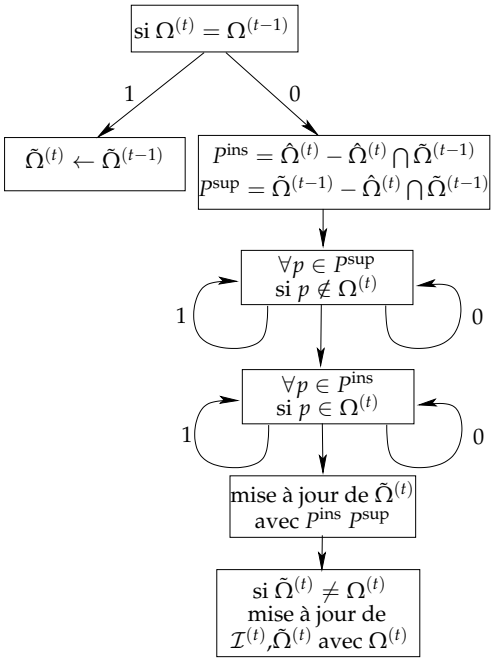


FIGURE 5.15 – Schéma du système de codage utilisé pour calculer $\mathcal{I}^{(t)}$ nécessaire pour obtenir $\tilde{\Omega}^{(t)} = \Omega^{(t)}$ à partir de l'estimation $\hat{\Omega}^{(t)}$. La réponse à chaque test "*si ...*" est codée par un seul bit (1 ou 0) excepté pour l'étape finale de mise à jour de $\mathcal{I}^{(t)}$ plus coûteuse.

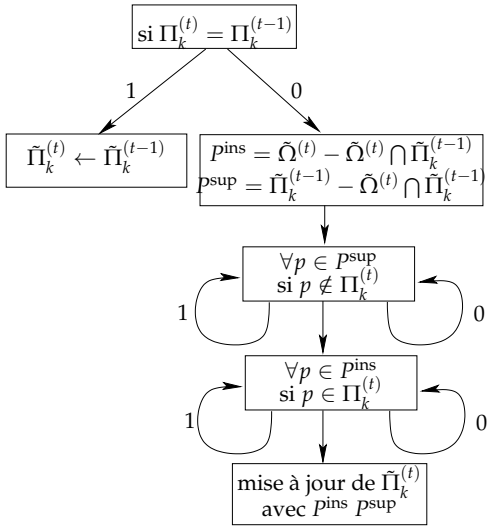


FIGURE 5.16 – Schéma de l'algorithme utilisé pour calculer le code qui permet au décodeur de retrouver $\tilde{\Pi}_k^{(t)}$ à partir de $\tilde{\Omega}^{(t)}$. La réponse à chaque test "*si ...*" est codée par un seul bit (1 ou 0) qui s'ajoute au débit initial permettant de calculer $\tilde{\Omega}^{(t)}$.

b.2) Décodeur

Au décodeur, on dispose uniquement du mélange $x(t)$ et de l'information codée $\mathcal{I}^{(t)}$ (les signaux $s_k(t)$ et la transcription $\Omega^{(t)}$ sont inconnus). On initialise $\tilde{\Omega}^{(0)} = \tilde{\Pi}_k^{(0)} = \varnothing$ puis on utilise le système de transcription sur $x(t)$ pour obtenir $\hat{\Omega}^{(t)}$. Les opérations de prédiction / insertion / suppression générées pour mettre à jour $\tilde{\Omega}^{(t)}$ sont corrigées en utilisant $\mathcal{I}^{(t)}$.

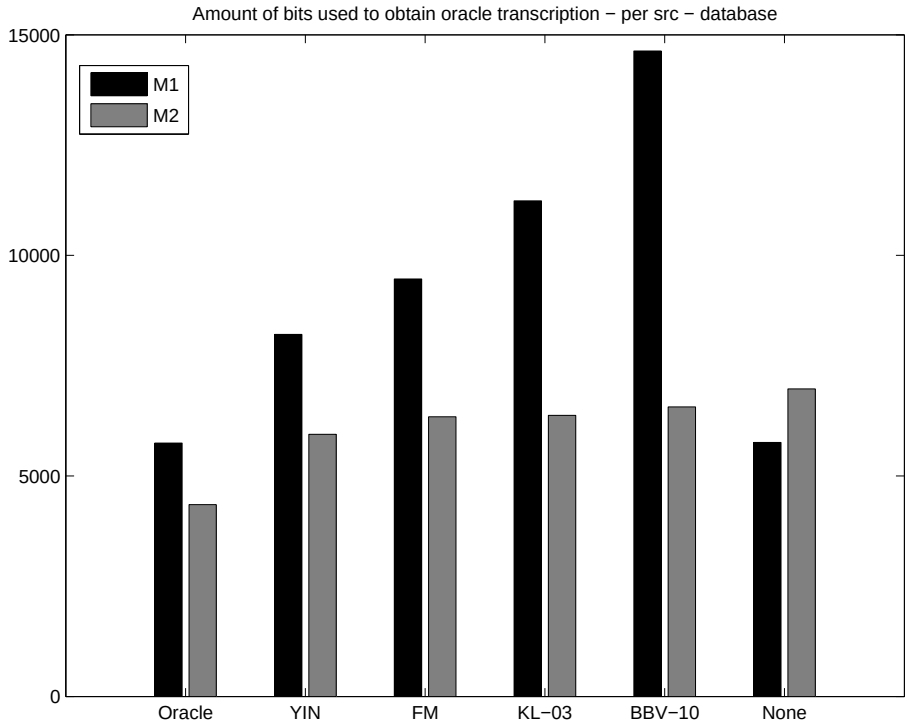
Ce traitement est effectué par une lecture séquentielle de $\mathcal{I}^{(t)}$ où chaque bit permet de définir si l'opération de mise à jour doit être effectuée ou ignorée. Une fois $\tilde{\Omega}^{(t)}$ obtenu, la même opération est appliquée pour chaque source $s_k(t)$.

Ainsi, le processus de décodage utilise le même schéma décrit pour le codeur présentés dans les figures 5.15 et 5.16. Le principe de prédiction utilisant un seul bit lu permet d'augmenter la robustesse de correction de la transcription qui suppose que $\hat{\Omega}^{(t)}$ est identique au codeur et au décodeur. La prédiction permet de tolérer certaines différences qui peuvent se produire en cas d'ajout de bruit ou lors de l'utilisation d'un système de tatouage audio numérique comme proposé dans la méthode décrite dans la section 5.3. Dans ce cas, l'estimateur est utilisé pour introduire les nouveaux événements (fréquence des candidats F_0 pour les débuts et fins de note), ainsi les erreurs se produisant lorsque les notes maintenues n'ont plus à être corrigées car un bit suffit pour indiquer au décodeur d'utiliser la prédiction.

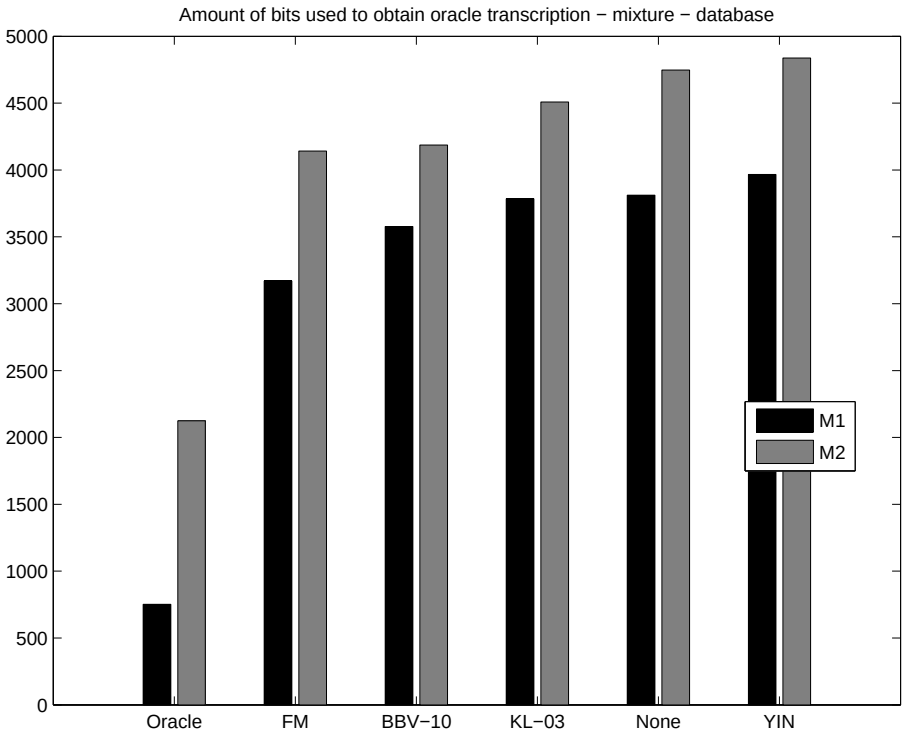
5.2.4 Évaluation comparative des systèmes de transcription proposés

On se propose désormais de comparer sur nos exemples sonores la taille des données codées permettant de retrouver la transcription exacte. Pour cela on réutilise le même protocole expérimental que celui décrit précédemment. La figure 5.17 compare le nombre de bits utilisés pour informer les méthodes de transcription évaluées sur la base de test en utilisant chacun des système de codage (M1 et M2 détaillés dans la section 5.2.3). Pour la figure 5.17, on compare ce nombre de bits avec celui nécessaire dans le cas sans estimateur (**None**) et celui nécessaire si on disposait d'un estimateur oracle parfait. Dans le cas sans estimateur, on fixe $\hat{\Omega}^{(t)} = \varnothing$. Nous avons bien sûr vérifié que dans tous les cas, la taille des données utilisée par les systèmes M1 et M2 reste bien inférieure à celle du fichier MIDI correspondant. L'ordre de grandeur de cette différence est illustré dans les résultats présentés dans la table 5.2.

Pour chaque méthode, la figure 5.18 présente le gain en nombre de bits observé pour chaque système M1 et M2 en utilisant un estimateur. Dans le cas où le gain est négatif, cela signifie que la correction des erreurs est plus coûteuse que le codage de la transcription seule sans estimateur. On observe donc que la méthode M2 permet d'obtenir la plupart du temps un gain supérieur à la méthode M1.

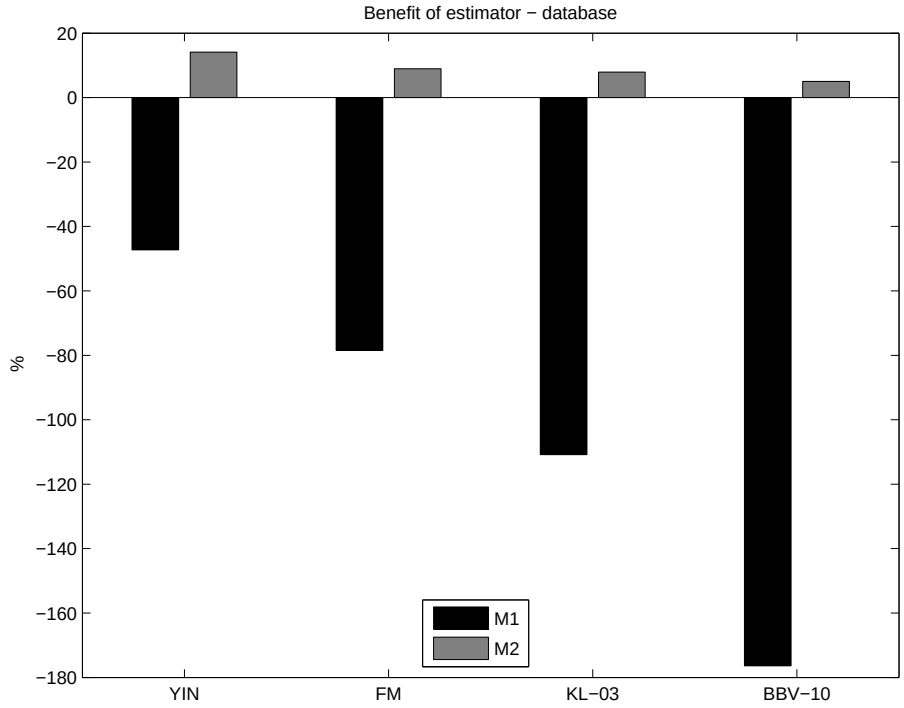


a) depuis les source isolées

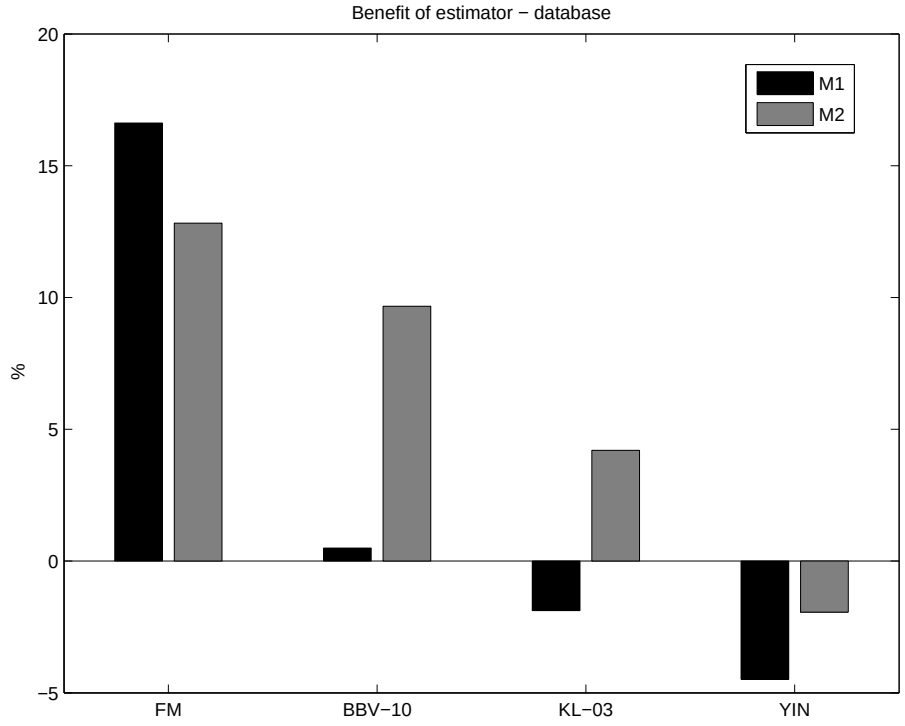


b) depuis le mélange

FIGURE 5.17 – Comparaison du nombre de bits utilisés pour chaque système de codage pour informer les méthodes d'estimation F_0 appliquées sur la base de données d'évaluation. Les résultats sont triés par ordre croissant de bits utilisés pour la méthode M2.



a) depuis les source isolées



b) depuis le mélange

FIGURE 5.18 – Comparaison du gain en pourcentage de nombre de bits apporté par l'utilisation d'un estimateur. Les résultats sont triés par ordre décroissant du pourcentage de gain sur M2.

5.3 Application à la séparation de sources informée

Le problème de la séparation de sources à partir d'un mélange musical décrit en détail dans le chapitre 2 consiste à retrouver les signaux s_k correspondant à chacune des entités qui composent un mélange en ne disposant que d'une ou plusieurs observations de celui-ci. Dans notre cas, on cherche à estimer les signaux des instruments de musique pouvant jouer des notes (*e.g.* saxophone, guitare, piano,...) qui composent un signal de mélange mono-canal. Il s'agit d'un problème difficile à résoudre sans information complémentaire, car il s'agit d'une configuration sous-déterminée (nombre de sources supérieur au nombre d'observations). Dans notre cas, nous souhaitons exploiter la connaissance de la transcription et de la structure particulière du signal de chacune des sources pour effectuer la séparation. Comme la transcription d'un mélange polyphonique est difficile à obtenir, on choisi ici de l'estimer grâce au système de codage proposé en se plaçant dans une configuration particulière où la séparation de sources audio informée [PG11, Knu05] est applicable.

5.3.1 Aperçu de la méthode

On se place dans une configuration codeur / décodeur comme décrite par la figure 5.19 qui caractérise l'approche informée.

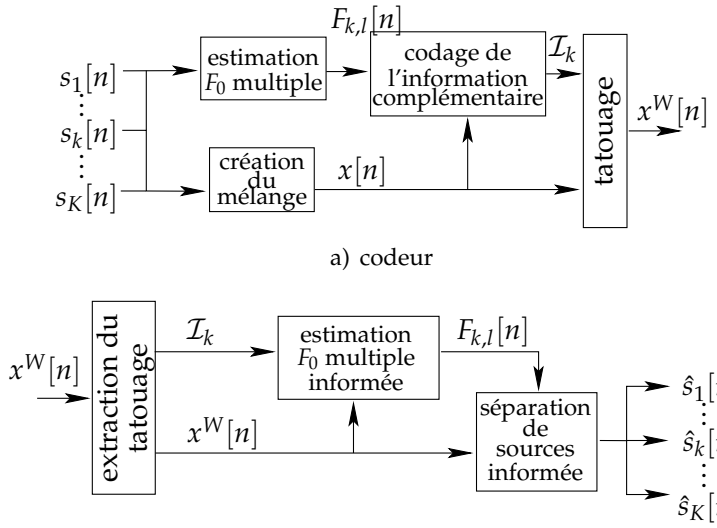


FIGURE 5.19 – Schéma d'un système de séparation de sources informée utilisant une configuration codeur / décodeur.

Au codeur, on suppose que le signal de chaque source $s_k[n]$ est exactement connu avant le processus de mélange. La transcription de référence $\Pi_k^{[n]}$ de chaque source isolée est estimée en utilisant le système de transcription proposé (*cf.* section 5.1.5). L'information nécessaire permettant de retrouver chaque source isolée à partir du mélange est codée en utilisant une méthode de codage (M1 ou M2) décrite précédemment dans la section 5.2.3. Pour le système proposé, l'information supplémentaire $\mathcal{I}$ résultant du codage est tatouée de façon inaudible dans le signal du mélange.

Au décodeur où les sources $s_k[n]$ sont inconnues, on extrait l'information du mélange en inversant le processus de tatouage. L'information extraite est utilisée pour la correction des erreurs de transcription obtenue à partir du mélange tatoué $x^W[n]$ tel que décrit dans la section 5.2.3 pour le décodage. A l'aide de la transcription corrigée, on applique une

méthode de séparation de sources informée par le MIDI classique telle que [HDB11] pour estimer les différents signaux $\hat{s}_k[n]$ from $x^W[n]$.

5.3.2 Le tatouage audio numérique

L'information supplémentaire $\mathcal{I}^{(t)}$ calculée par le codeur est cachée de manière inaudible grâce à une méthode de tatouage audio numérique. Compte tenu du faible débit nécessaire pour coder une partition musicale (en général moins de 2kbs pour un fichier MIDI) on choisit une technique de tatouage basé sur le bit de poids le plus faible, *Least Significant Bit* (LSB) afin de coder l'information dans la représentation binaire de chaque échantillon du signal. Malgré un débit plus faible que les techniques utilisant QIM, cette approche présente plusieurs avantages pour notre problème :

- la bande passante réservée au tatouage est suffisante, en effet un signal échantillonné à 44.1kHz permet de tatouer de l'information à un taux de 44.1kbps (1 bit par échantillon),
- le temps de calcul se retrouve amélioré, en effet il n'est pas utile d'estimer un masque perceptif pour connaître la capacité du signal. Le décodage s'en retrouve simplifié car on sait à l'avance où se trouve l'information tatouée. Des tests d'écoute informels que nous avons réalisé ont montré que l'utilisation d'un seul bit pour le tatouage était quasiment imperceptible,
- l'opération de codage/décodage est déterministe, bien qu'un système de transcription polyphonique soit plus robuste aux perturbations sur un signal et permette de retrouver dans certains cas exactement la même transcription à partir de deux versions bruitées provenant d'un même signal. Ce système de tatouage garantit l'existante d'un quantificateur Q permettant de vérifier l'égalité $Q(x) = Q(x^W)$. En effet, il suffit de récupérer la valeurs des échantillons en ignorant ou en fixant une valeur arbitraire pour les bits tatoués.

La capacité d'une telle approche peut évidemment être aisément augmentée en calculant un masque perceptif et en appliquant le tatouage sur une représentation mieux adaptée (plus parcimonieuse) du signal comme cela est couramment proposé dans [Bar06, GMai, Liu07].

5.3.3 La séparation de sources informée par la partition

Après l'étape de décodage permettant d'obtenir une transcription de chaque signal utilisant la quantification MIDI, les signaux correspondant à chaque source isolée peuvent être séparés du mélange en utilisant une méthode de séparation de sources adéquate permettant d'exploiter l'information *a priori* apportée par la transcription. Dans nos expérimentations, nous avons comparé une méthode de l'état de l'art [HDB11] utilisant la NMF avec d'autres approches que nous avons développé dans nos travaux utilisant respectivement la modélisation sinusoïdale ainsi que le filtrage du mélange.

a) Factorisation du spectrogramme

Cette méthode basée sur l'utilisation de la NMF a pour objectif de décomposer le spectrogramme d'amplitude qui correspond à une matrice notée X de dimension $F \times T$. Cette décomposition permet d'obtenir un produit de deux matrices dont les coefficients sont positifs ou nuls :

$$X \approx \hat{X} = WH \tag{5.21}$$

où $W \in \mathbb{R}^{F \times R}$ correspond à un dictionnaire d'atomes de sources quasi-harmoniques composé de l'ensemble des notes de musique possibles. La matrice $H \in \mathbb{R}^{R \times T}$ correspond à un dictionnaire d'activations temporelles. Cette décomposition s'accompagne aussi d'une réduction du rang de la matrice X telle que ($FR + RT \ll FT$). Les matrices W et H sont obtenues en initialisant ces matrices à partir de la transcription de référence (ce qui a pour effet de contraindre les activations et les atomes harmoniques correspondants aux notes actives).

Dans nos expérimentations, nous avons utilisée la méthode proposée par Hennequin [HDB11] qui a la particularité d'utiliser un dictionnaire W de sources harmoniques pouvant varier dans le temps. Cette méthode a montré des résultats comparables à l'approche PLCA (*Probabilistic Latent Component Analysis*) [GSMA10] qui compte à ce jour parmi les plus efficaces.

b) Séparation par modélisation sinusoïdale

D'après le modèle de source harmonique décrit par (2.23), il est possible d'affecter les composantes sinusoïdales qui correspodent à une source quasi-harmonique (*cf.* équation 2.23) pour une fréquence fondamentale donnée. Pour cela, on effectue un traitement en trois étapes qui consiste à :

- extraire et estimer les composantes sinusoïdales à partir du signal en utilisant la méthode proposée dans la section 5.1.1
- construire les HPS correspondant à chaque fréquence fondamentale en utilisant la méthode *Harmonic Matching* décrite dans la section 5.1.2,
- reconstruire l'amplitude et de la phase des composantes sinusoïdales en collision avec d'autres sources par interpolation.

Pour l'estimation de chaque signal source, nous avons utilisé un algorithme de suivi de partiels (*cf.* section 2.1.4) basée sur la prédiction linéaire comme proposé par Lagrange *et al.*[LMR04b]. Les détails d'implémentation sont identiques à ceux utilisés dans la section 4.4.2. Pour traiter les collisions (partiels affectés à plusieurs sources), nous effectuons une reconstruction des paramètres d'amplitude et de phase à partir des composantes "plus fiables" par interpolation (*cf.* section 1.1.2). Pour la resynthèse, nous utilisons l'interpolation polynomiale proposée dans [GMdM$^+$03].

c) Séparation par filtrage

Dans cette configuration, la transcription et le modèle d'une source harmonique (*cf.* section 2.2.2) sont utilisés pour déduire le masque TF de chaque source. Ainsi, les maxima locaux détectés dans le spectre sont affectés aux pics du peigne de Dirac associé à chaque note active de fréquence F_0. Pour chaque pic affecté à plus de deux signaux s_k différents, Every et Szymanski [ES06] proposent de calculer le spectre de chaque note p en filtrant localement le spectre du mélange au voisinage du pic détecté.

Nous avons choisi d'implémenter le filtre suivant qui isole le pic du reste du signal en appliquant une fonction de type gaussienne centrée sur la fréquence centrale interpolée $f_h^p = hF_0$ du pic. La fonction de transfert de ce filtre est définie par :

$$\hat{G}^p[k] = A_h^p \cdot \exp\left(-\frac{|f_k - f_h^p|}{\sigma}\right), \tag{5.22}$$

avec A_h^p l'amplitude interpolée (à partir des pics voisins) pour le partiel h de la source harmonique considérée. Ici $f_k = k\frac{F_s}{N}$ est la fréquence en Hertz de l'indice du pic dans le spectre discret $k \in [\lfloor \frac{f_h^p}{Fs} N \rfloor - 3; \lfloor \frac{f_h^p}{Fs} N \rfloor + 3]$. Le paramètre $\sigma = 0.25$ permet de paramétrer la largeur de la fonction gaussienne appliquée.

Le filtre final qui est obtenu après normalisation est défini pour l'ensemble Q des notes actives qui partagent le même pic est donné par l'expression suivante :

$$G^p[k] = \frac{\hat{G}^p[k]}{\sum_{q \in Q} \hat{G}^q[k]} \tag{5.23}$$

5.3.4 Évaluation

Pour évaluer le système proposé nous avons utilisé 3 extraits musicaux de 20 secondes, synthétisés à partir d'un fichier MIDI en utilisant un expandeur. Nous avons utilisé l'expandeur EXS24[5] qui permet de produire des sons réalistes. La composition de chaque extrait est décrite dans la table 5.1.

Extrait	nombre de sources	Instruments
1	3	flûte, piano et contrebasse
2	4	flûte, piano, contrebasse et batterie
3	4	B3, piano, contrebasse et batterie

TABLE 5.1 – Description des extraits sonores utilisés pour l'évaluation du système de séparation de sources informée par l'estimation F_0 multiple.

Nous comparons dans un premier temps (*cf.* figures 5.20, 5.21 et 5.22) les résultats de transcription obtenus en appliquant les techniques d'estimation des F_0 sur chaque source isolée puis sur le mélange. Pour cela, nous utilisons comme référence la transcription utilisée pour la synthèse des signaux.

Dans les figures 5.23 et 5.24 nous présentons les débits obtenus en combinant les méthodes de codage M1 et M2 avec chacun des systèmes de transcription. Nous comparons ces résultats avec le débit nécessaire pour informer la transcription exacte (l'oracle) et sans estimateur (*none*). En effet, les systèmes de codage M1 et M2 nécessitent un débit minimal permettant d'informer que la transcription utilisée est identique à celle de référence. Dans le cas du codage sans estimateur, chaque méthode doit insérer par codage la valeur de toutes les notes manquantes.

Les figures 5.25, 5.26 et 5.27 présentent les résultats de séparation de source en fonction de la transcription de référence puis détaillée pour chaque source en utilisant la meilleure transcription. Ces résultats sont présentés en comparant chacune des trois techniques de séparation de sources proposées.

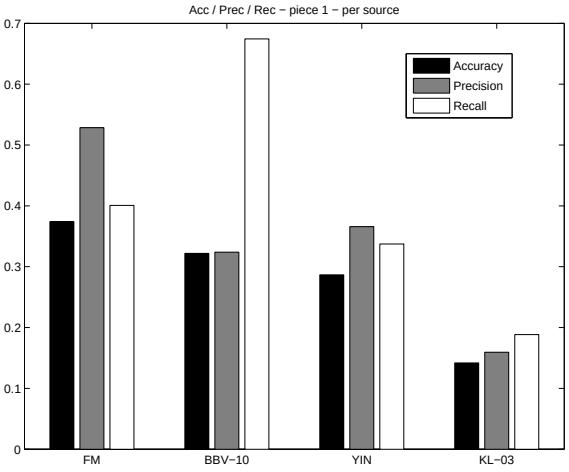
[5]EXS24 est un expandeur logiciel présent dans l'application professionnelle Logic Audio http://www.apple.com/fr/logic-pro/

a) Évaluation de l'estimation F_0 multiple informée

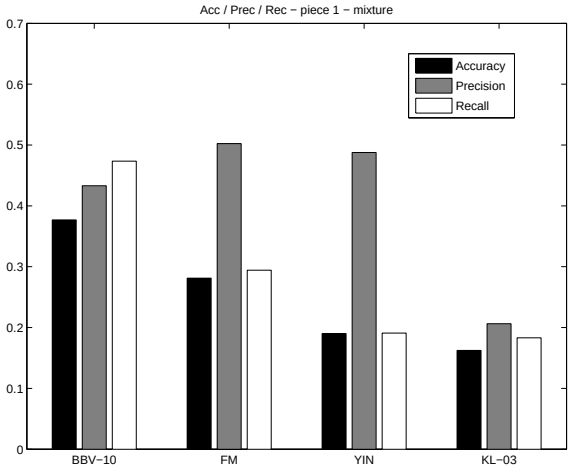
Nous présentons dans la table 5.2 les débits obtenus en combinant chaque système de codage avec le meilleur estimateur permettant de retrouver la transcription de référence. Comme le montre les expérimentations de la figure 5.23, la méthode **FM** est celle qui obtient toujours le débit le plus faible quelque soit le système choisi (M1 ou M2). Dans ces exemples, la système M1 obtient des débits plus faible que M2. On observe également un gain à chaque fois que l'estimateur **FM** est combiné avec un système de codage par rapport au codage pur. Dans tous les cas, le codage pur utilisant le format MIDI nécessite un nombre de bits plus important.

Codage	Extrait 1	Extrait 2	Extrait 3
M1 (avec le meilleur estimateur)	2877 bits (FM)	2123 bits (FM)	2959 bits (FM)
M1 (sans estimateur)	3373 bits	2531 bits	3553 bits
M2 (avec le meilleur estimateur)	3621 bits (FM)	2327 bits (FM)	3926 bits (FM)
M2 (sans estimateur)	4085 bits	2642 bits	4521 bits
fichier MIDI	16808 bits	14896 bits	20400 bits

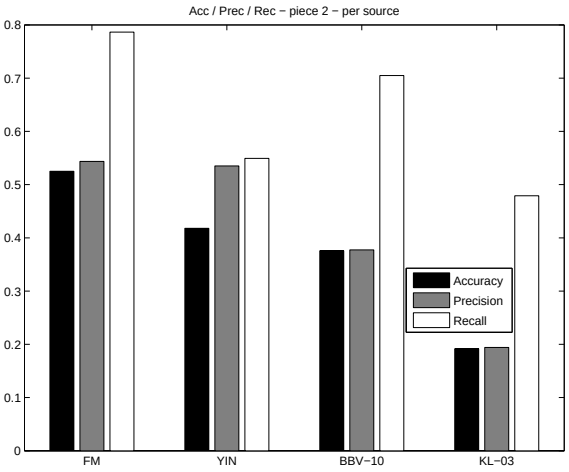
TABLE 5.2 – Nombre de bits utilisés en fonction du codage permettant d'obtenir la transcription exacte. Lorsque le codage est combiné avec un estimateur, celui-ci est appliqué sur le signal de mélange tatoué. Le meilleur estimateur considéré est celui qui permet de minimiser le nombre de bits utilisés.



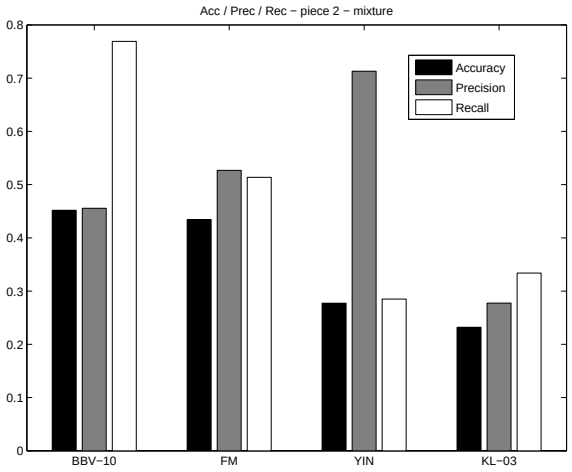
a) depuis les sources séparées de l'extrait 1



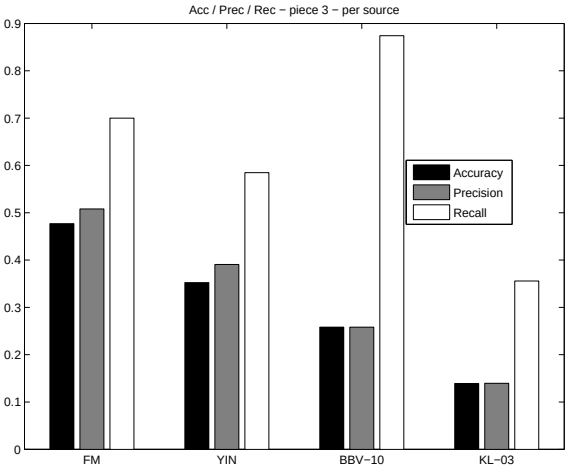
d) depuis le mélange de l'extrait 1



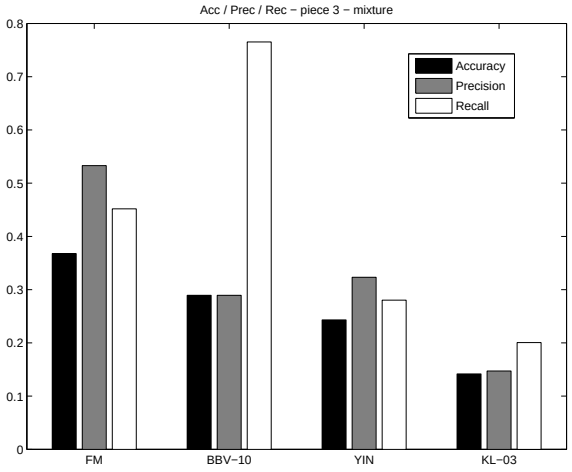
b) depuis les sources séparées de l'extrait 2



e) depuis le mélange de l'extrait 2

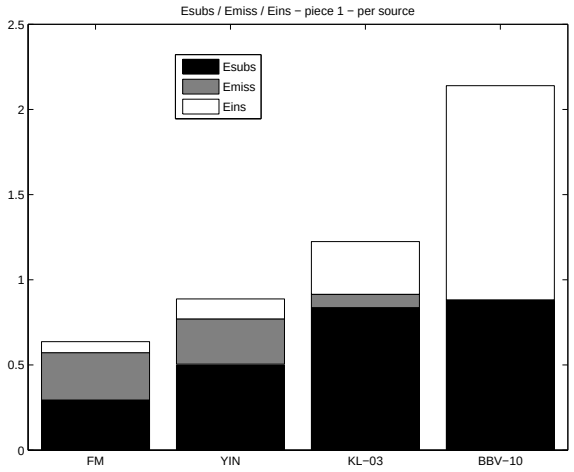


c) depuis les sources séparées de l'extrait 3

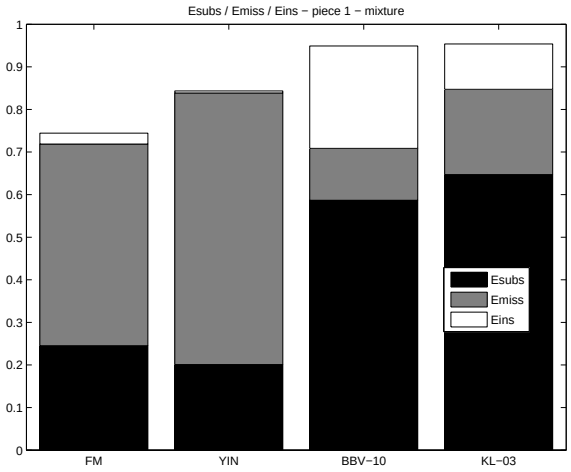


f) depuis le mélange de l'extrait 3

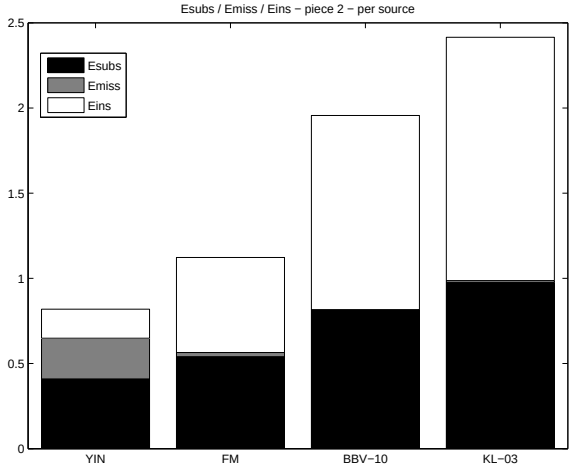
FIGURE 5.20 – Score *Accuracy*, *Precision* et *Recall* obtenu pour chaque technique d'estimation F_0 multiple.



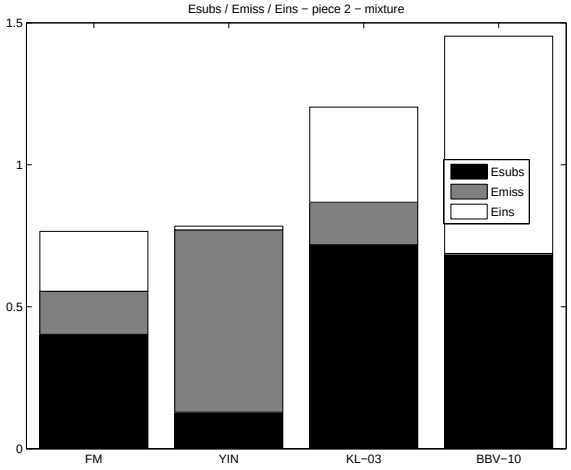
a) depuis les sources séparées de l'extrait 1



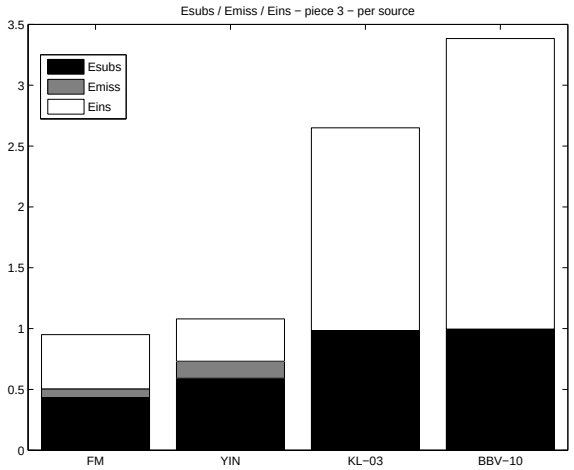
d) depuis le mélange de l'extrait 1



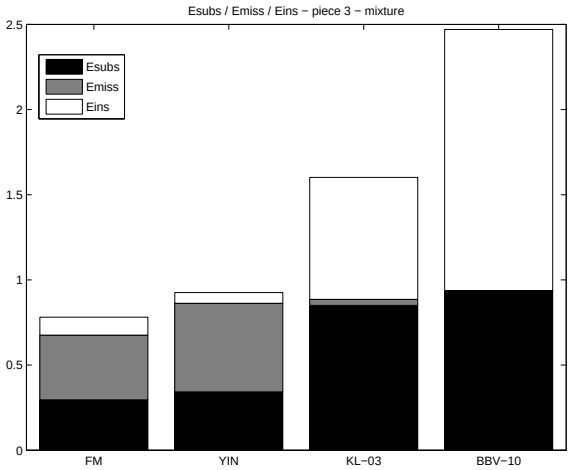
b) depuis les sources séparées de l'extrait 2



e) depuis le mélange de l'extrait 2

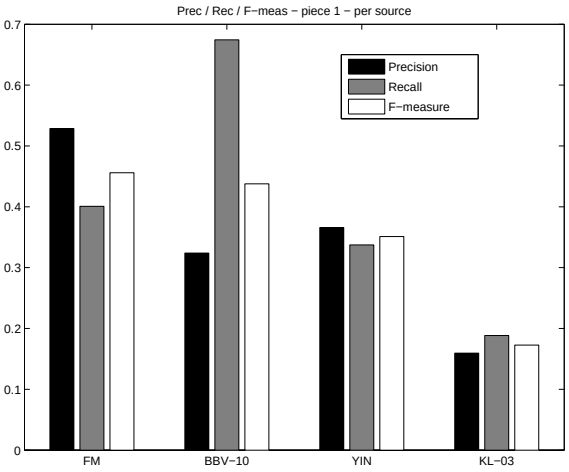


c) depuis les sources séparées de l'extrait 3

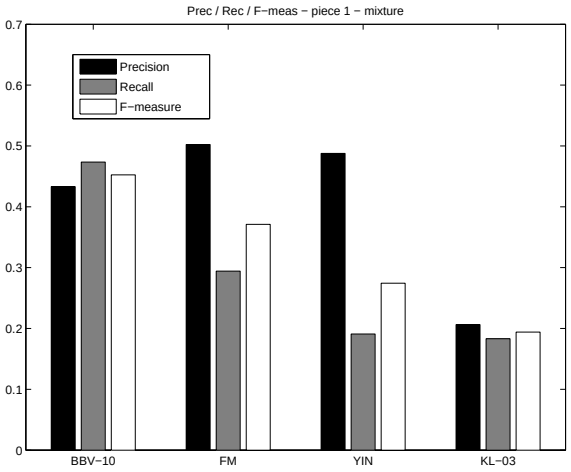


f) depuis le mélange de l'extrait 3

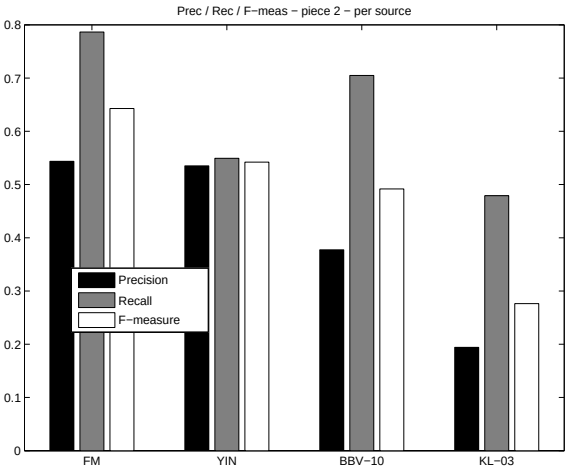
FIGURE 5.21 – Classification des erreurs pour chaque technique d'estimation F_0 multiple.



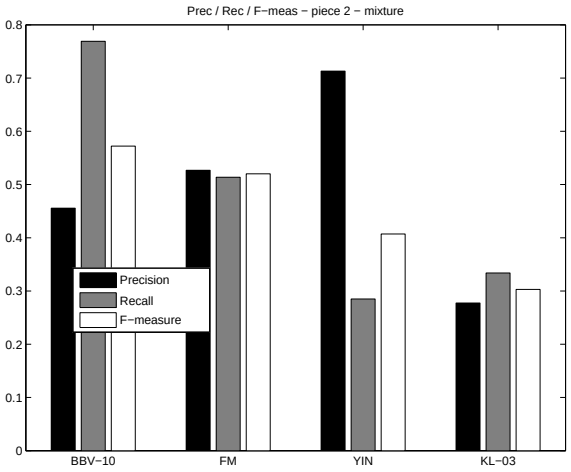
a) depuis les sources séparées de l'extrait 1



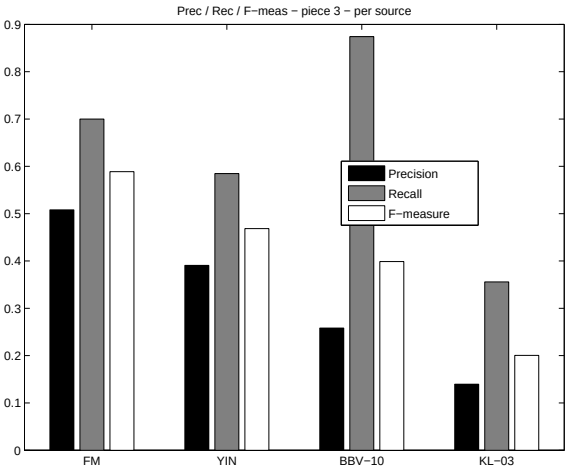
d) depuis le mélange de l'extrait 1



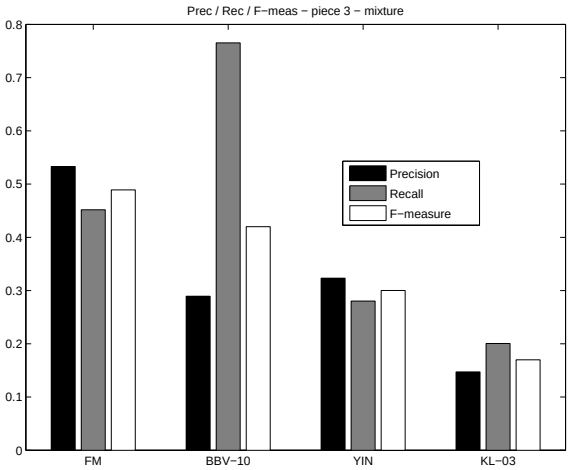
b) depuis les sources séparées de l'extrait 2



e) depuis le mélange de l'extrait 2

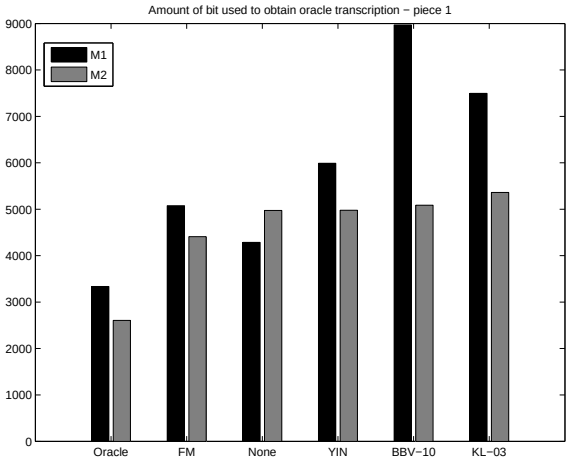


c) depuis les sources séparées de l'extrait 3

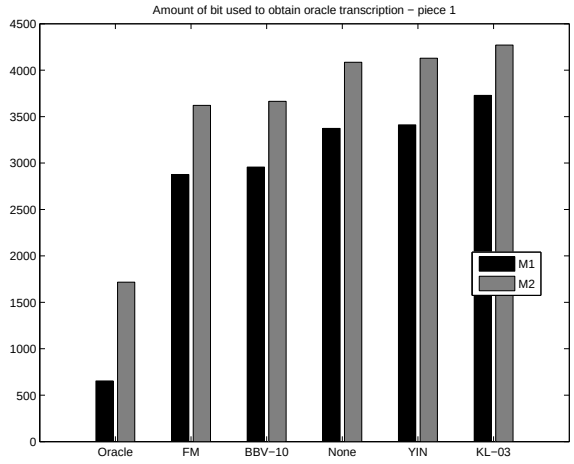


f) depuis le mélange de l'extrait 3

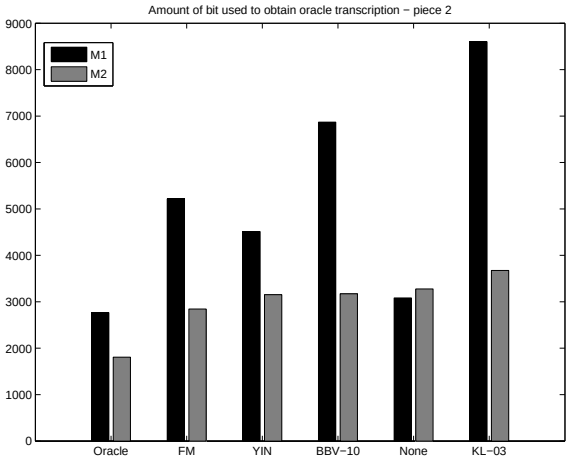
FIGURE 5.22 – Score *Precision*, *Recall* et *F-Measure* calculés pour chaque technique d'estimation F_0 multiple.



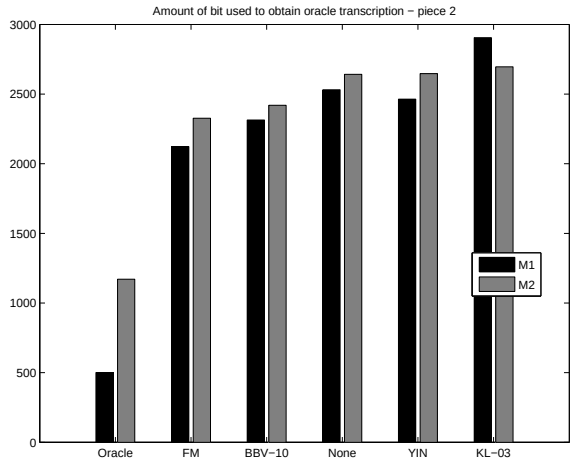
a) depuis les sources séparées de l'extrait 1



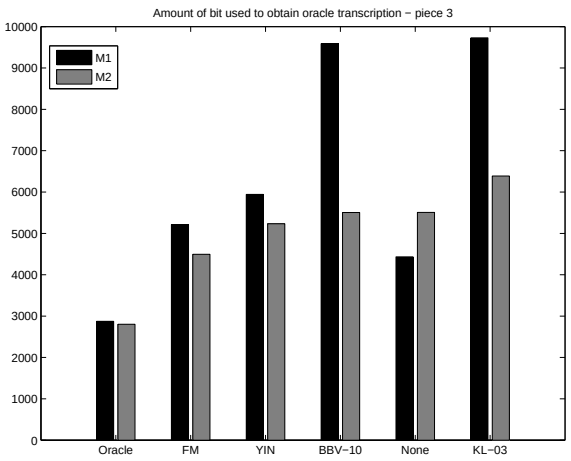
d) depuis le mélange de l'extrait 1



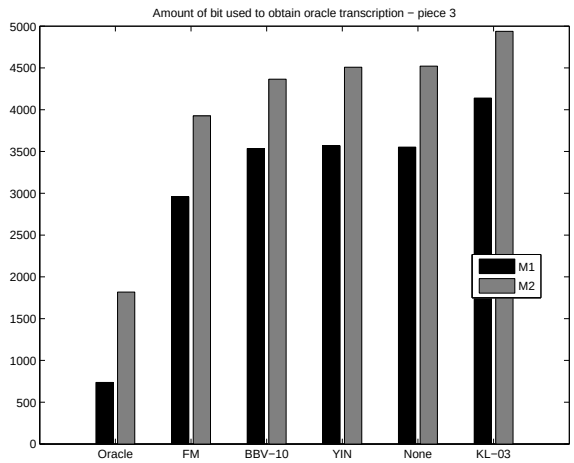
b) depuis les sources séparées de l'extrait 2



e) depuis le mélange de l'extrait 2

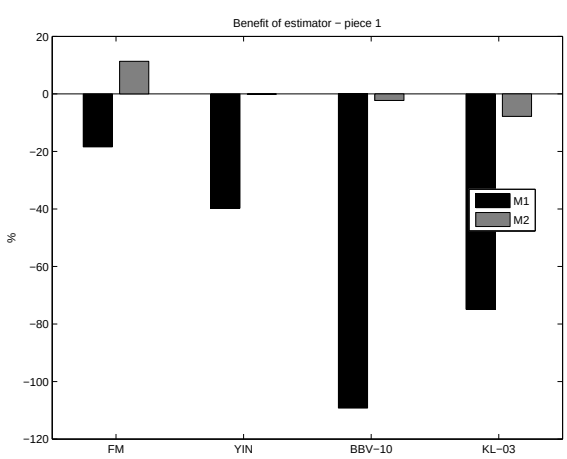


c) depuis les sources séparées de l'extrait 3

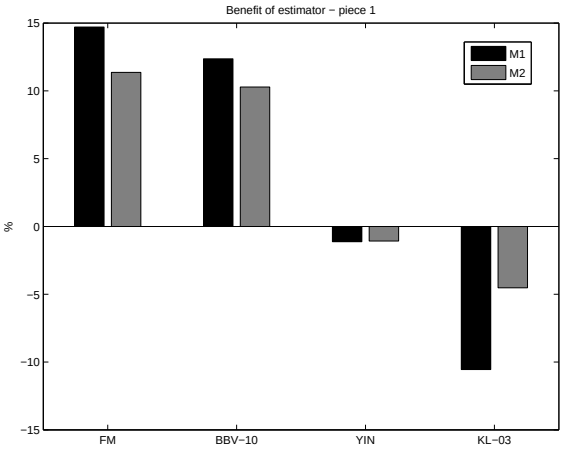


f) depuis le mélange de l'extrait 3

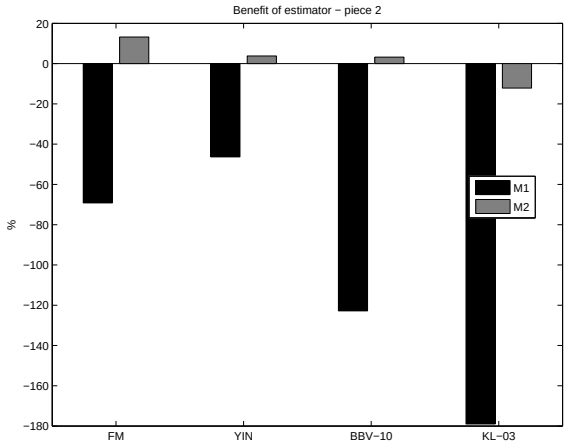
FIGURE 5.23 – Quantité d'information utilisée pour calculer la transcription de référence à partir de l'estimation calculé pour chacune des techniques d'estimation F_0 multiple.



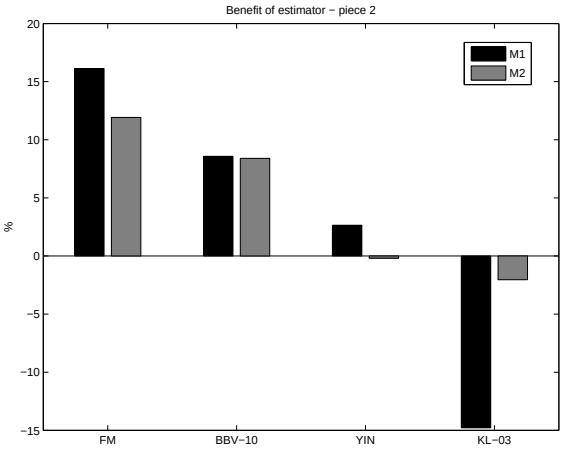
a) depuis les sources séparées de l'extrait 1



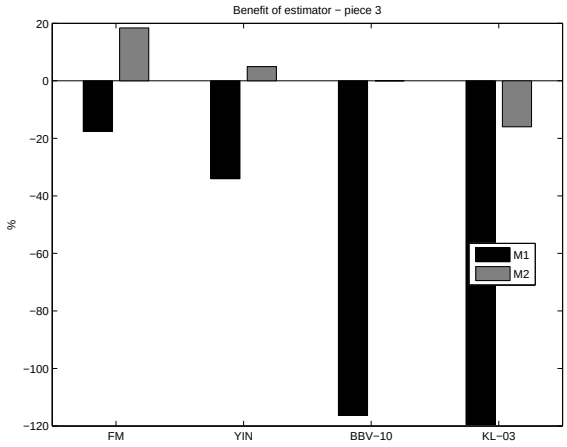
d) depuis le mélange de l'extrait 1



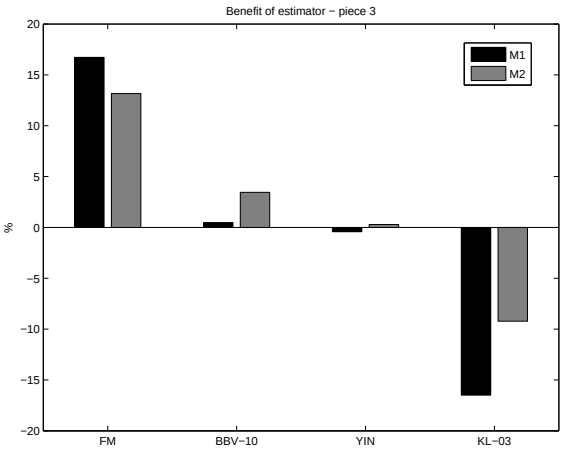
b) depuis les sources séparées de l'extrait 2



e) depuis le mélange de l'extrait 2



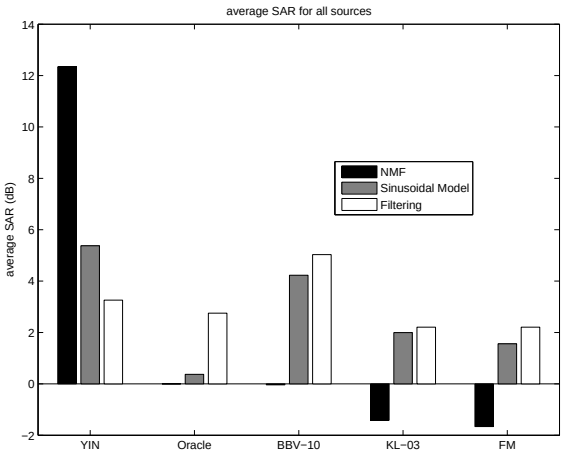
c) depuis les sources séparées de l'extrait 3



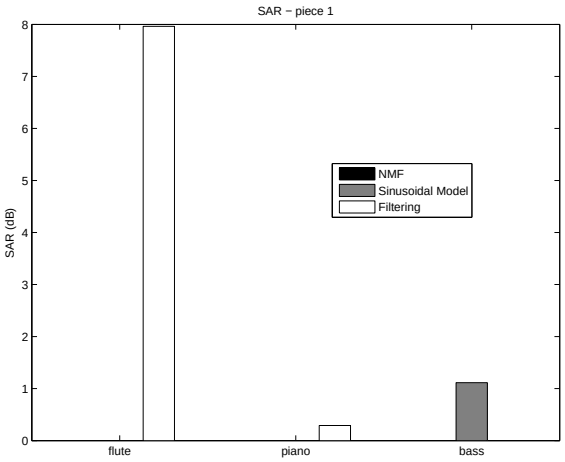
f) depuis le mélange de l'extrait 3

FIGURE 5.24 – Gain apporté par l'estimation F_0 multiple sur la quantité d'information codée permettant d'obtenir la transcription de référence.

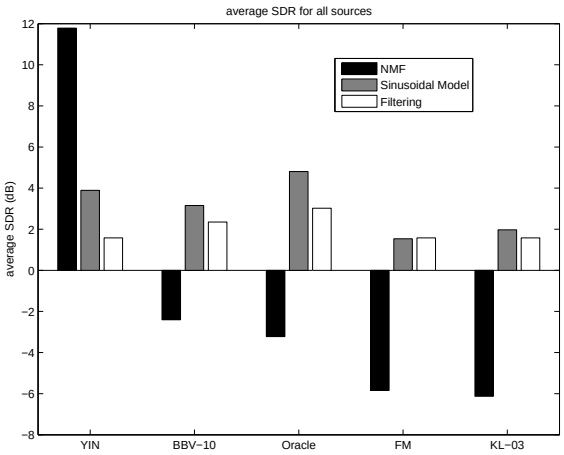
b) Évaluation de la qualité obtenue par la séparation de sources



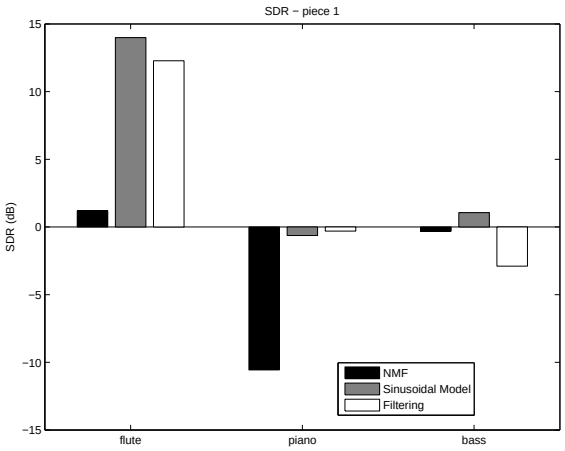
a) SAR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



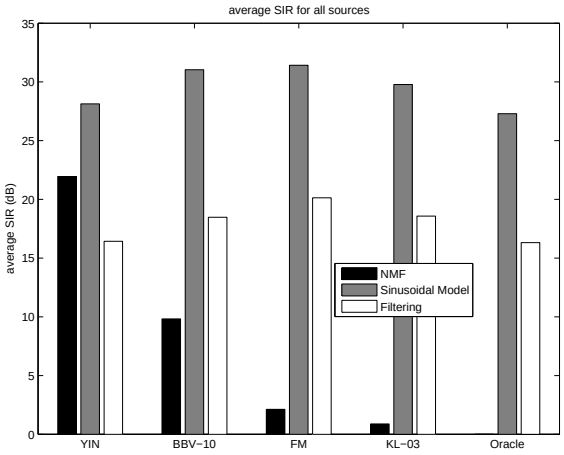
d) SAR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple



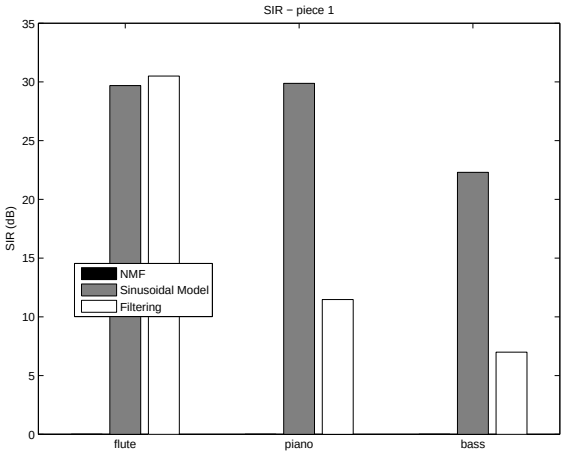
b) SDR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



e) SDR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple

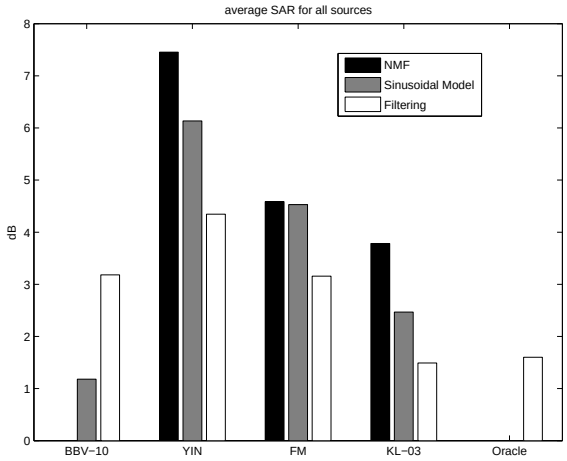


c) SIR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée

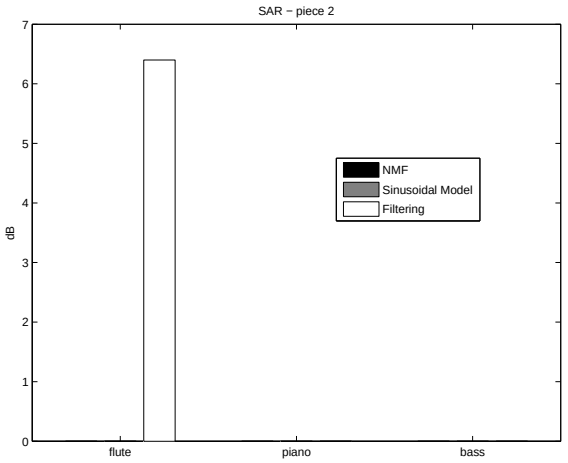


f) SIR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple

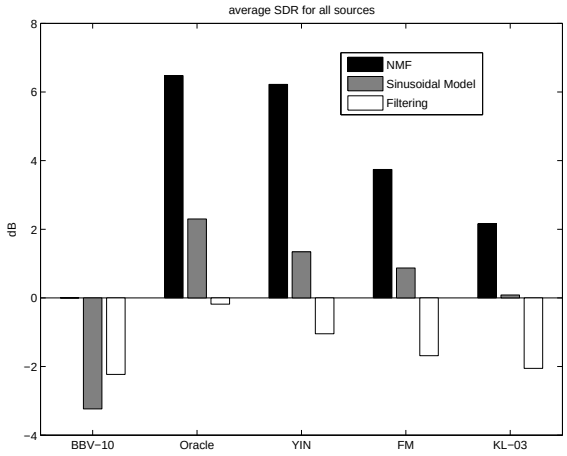
FIGURE 5.25 – Qualité de séparation mesurée pour l'extrait 1.



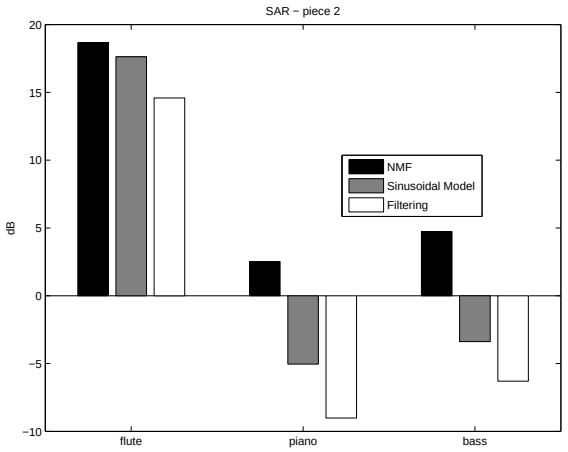
a) SAR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



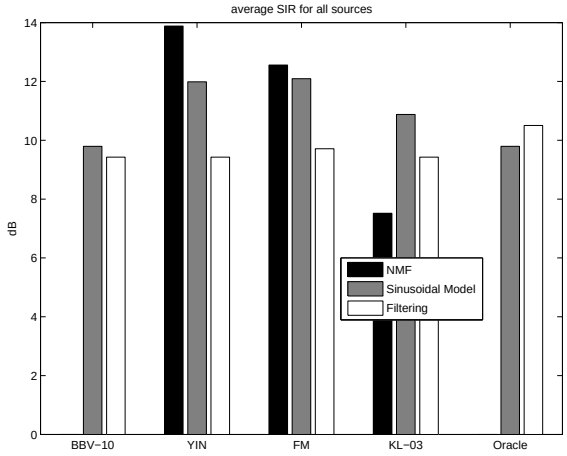
d) SAR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple



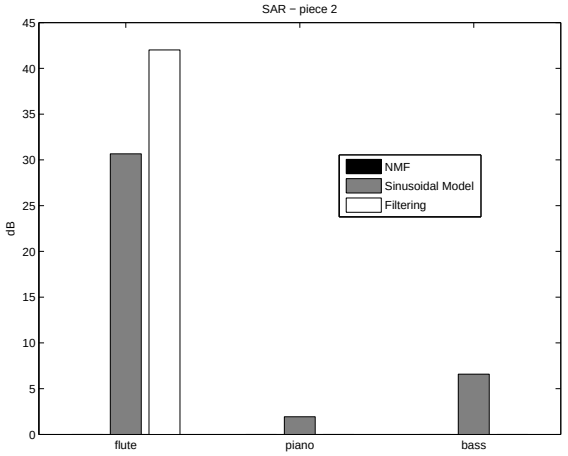
b) SDR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



e) SDR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple

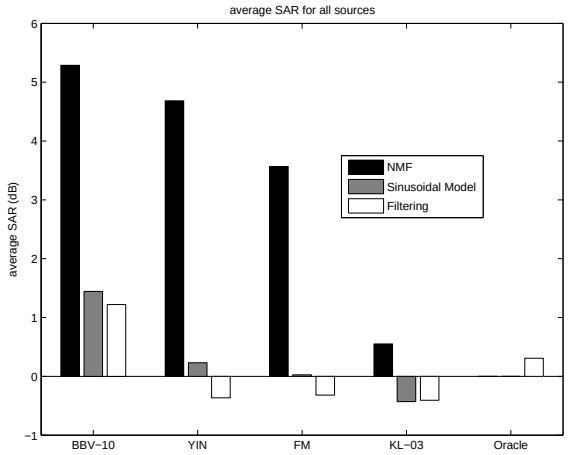


c) SIR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée

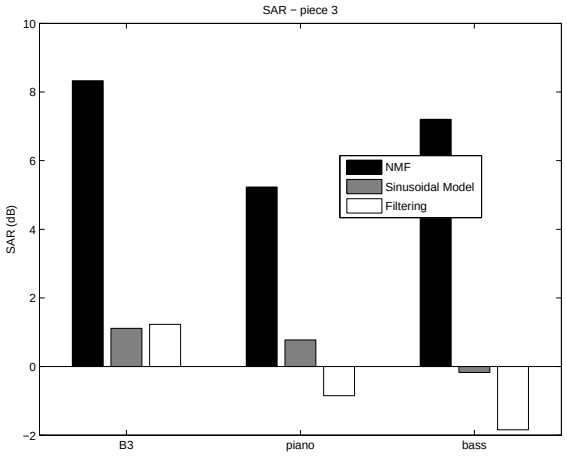


f) SIR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple

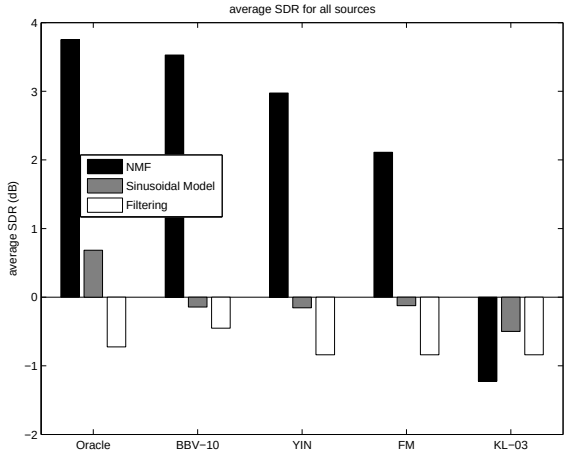
FIGURE 5.26 – Qualité de séparation mesurée pour l'extrait 2.



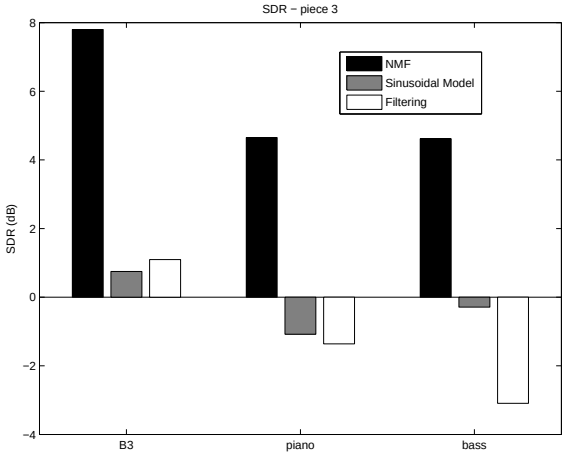
a) SAR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



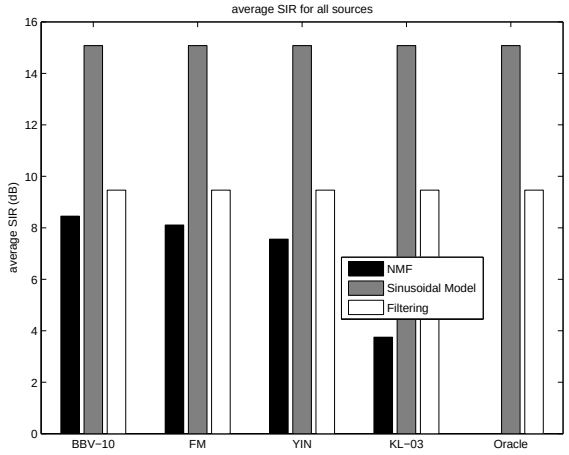
d) SAR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple



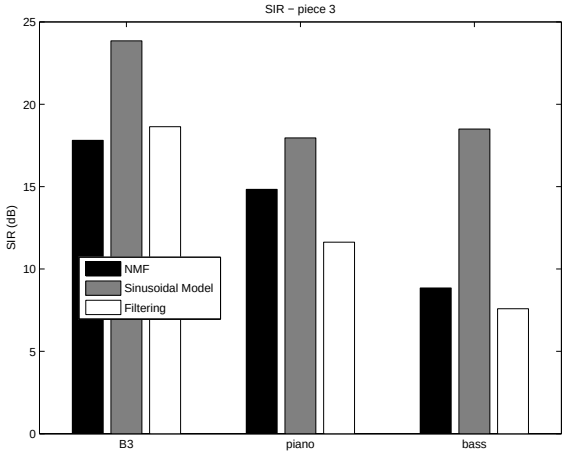
b) SDR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



e) SDR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple



c) SIR moyen pour toutes les sources en fonction de la méthode d'estimation F_0 multiple utilisée



f) SIR pour chaque source en utilisant la meilleure méthode d'estimation F_0 multiple

FIGURE 5.27 – Qualité de séparation mesurée pour l'extrait 3.

Les figures 5.25, 5.26 et 5.27 présentent les résultats qualitatifs obtenus sur chacun des 3 extraits musicaux pour la séparation de sources. Pour ces mesures nous utilisons la métrique proposées dans la section 2.3.5 pour estimer la qualité objective entre les signaux d'origine et les signaux estimés. La combinaison entre une méthode d'estimation F_0 et un méthode de séparation de source informée par la partition donne parfois des résultats surprenants, en effet la méthode NMF est plus sensible à la précision de la transcription. En effet, la factorisation du spectre échoue lorsque le dictionnaire d'atome est mal initialisé en cas d'erreurs de transcription. Les approches de séparation par modélisation sinusoïdale et par filtrages permettent de minimiser les interférences entre les sources au prix d'artefacts plus importants liés à l'approximation du signal de chaque source (peigne harmonique et filtrage passe-bas).

Sur ces figures, les données manquantes correspondent aux cas où les techniques de séparation ne fournissent pas une solution suffisante permettant une décomposition de l'erreur basée sur la métrique décrite dans la section 2.3.5. Ce problème se produit lorsque la séparation échoue (*e.g.* divergence de la NMF) ou lorsque les signaux estimés ne correspondent pas aux sources (*e.g.* signal résiduel). Nous avons choisi délibérément de présenter séparément les résultats obtenus pour chaque extrait afin de présenter en détail les résultats de séparation en fonction de la nature des sources.

En général les sources prédominantes dans les hautes fréquences (flûte et orgue B3) sont mieux estimées. La transcription du piano jouant des accords et présents dans les fréquence intermédiaires possède en général une transcription moins précise et de nombreuses collisions TF avec la contrebasse.

Une première interprétation de ces résultats semble indiquer que :

- La méthode **FM** combinée avec la méthode M2 permet de retrouver la transcription de référence (sources isolées) avec le débit le plus faible.
- La méthode M2 permet de compenser le débit supplémentaire requis pour corriger un "mauvais" estimateur contrairement à M1.
- La transcription oracle ayant servi à la synthèse des signaux dans cet expérimentation n'obtient pas toujours les meilleurs performances pour la séparation. Cela s'explique car les sons générés par l'expandeur ne sont pas précisément alignés sur la transcription, principalement lors des débuts et fins de notes. Ces différences posent des problèmes à la séparation par NMF qui est plus sensible à l'initialisation de l'algorithme.
- La plupart du temps, la meilleure combinaison permettant la séparation de sources semble atteinte en utilisant la transcription **FM** combinée avec la méthode M2 et la séparation par NMF.

5.4 Conclusion du chapitre

Dans ce chapitre, nous avons proposé l'implémentation détaillée d'un système de transcription polyphonique reposant sur la meilleure technique publiée au MIREX. Bien que le système proposé n'ait pas été a son tour évalué au MIREX, nous l'avons comparé avec plusieurs méthodes existantes publiées en utilisant la métrique d'évaluation décrite dans la section 2.2.4 couramment utilisée dans le domaine. Les principales contributions de ce chapitre reposant sur l'approche informée concernent la proposition de deux systèmes de codage combinant un estimateur F_0 multiple avec l'information complémentaire permettant de retrouver la transcription de référence. Nos expérimentations montrent qu'un

estimateur F_0 est capable d'apporter de l'information permettant de réduire le débit nécessaire pour le codage de la transcription de référence. Nous observons que lorsque les erreurs d'estimation sont trop nombreuses, leur correction par le système de codage nécessite comme attendu une quantité d'information codée plus importante.

La dernière contribution de ce chapitre porte sur la séparation de sources informée par la transcription elle-même informée à partir d'un signal de mélange mono-canal dans le cas sous-déterminé. Ainsi, nous décrivons un système de séparation de source complet basé sur une configuration codeur/décodeur et permettant de choisir la technique utilisée pour chacune des tâches qui composent ce système (estimation F_0, séparation, codage).

Ce chapitre a permis d'introduire l'approche informée pour les problèmes d'estimation d'information symbolique et doit être considéré comme une preuve de faisabilité (*proof of concept*). En effet, les résultats très récents que nous décrivons nécessitent d'être approfondis (base de données plus grande et déjà validée expérimentalement), notamment pour mieux comprendre les interactions entre les différentes approches utilisées respectivement pour la transcription et la séparation de sources. De plus, d'autres approches basées sur de l'information *a priori* sur les transcriptions ou même sur les signaux de référence doivent encore être explorées afin de guider un estimateur ou pour améliorer la correction des erreurs (*e.g.* distribution des F_0, polyphonie minimale/maximale, modèle d'enveloppe spectrale, etc.).

CONCLUSION

Bilan

Ce travail de thèse a permis une nouvelle exploration de certains problèmes d'analyse dans le domaine du traitement du signal audio grâce à l'approche informée. Ainsi, après avoir défini le cadre où cette nouvelle approche est applicable, nous avons proposé des solutions permettant de dépasser les performances des meilleures techniques classiques (non informées) existantes. Cette exploration établit des liens théoriques et pratiques entre la théorie de l'estimation et la théorie de l'information jusqu'alors traités séparément dans la littérature.

De plus, nous avons proposé des applications réalistes de séparation de sources informée permettant des manipulations de meilleure qualité des signaux audio. Dans le cadre des problèmes d'analyse utilisant une représentation symbolique du signal (*i.e.* transcription polyphonique d'un signal musical), la combinaison d'un estimateur et d'un algorithme de codage donne des résultats prometteurs caractérisé par une réduction du débit nécessaire pour coder la transcription de référence. Cependant des expérimentations complémentaires permettront certainement de mieux comprendre les interactions entre le codage et un estimateur MIR afin d'améliorer les performances des techniques proposées.

Dans ce travail, nous avons pu confronter des aspects à la fois théoriques en terme d'optimalité mais aussi pratiques sous forme d'applications concrètes. Enfin, nous avons généralisé l'approche informée lui permettant d'être appliquée dans d'autres domaines en exploitant des configurations où il est possible d'extraire de l'information et de la combiner avec des techniques d'analyse classiques.

Les principales contributions de ce travail de thèse présentées dans la seconde partie de ce document (chapitre 3, 4 et 5) peut être résumé comme suit.

Généralisation de l'estimation informée

L'analyse informée généralisée permettant de combiner un estimateur avec de l'information complémentaire a été introduire dans le chapitre 3. Nous avons placé cette nouvelle approche dans un cadre théorique confirmé préexistant (théorie de l'estimation et de l'information). La solution pratique que nous proposons dans nos travaux prend la forme d'un système de type codeur / décodeur applicable dans tout problème faisant interve-

nir de l'analyse et du codage (*e.g.* traitement du signal et des images, télécommunications, etc.). Il s'agit de la principale distinction de cette méthode par rapport à d'autres techniques plus spécifiques telles que la séparation de sources informée [Knu05, PGB09], le format de codage MPEG SAOC (*Spatial Audio Object Coding*) [HD07] ou l'inversion d'effets audio informée [LD08, GR13]. La nouvelle approche que nous proposons fait l'objet d'une publication dans le journal *EURASIP Journal on Advances in Signal Processing* [FM13].

Analyse Spectrale informée

Nous avons appliqué l'approche informé au problème d'estimation des paramètres sinusoïdaux et à la séparation de sources informée à partir d'un signal de mélange audio. Les résultats très encourageants ont permis de démontrer l'efficacité de l'approche d'un point de vue théorique et pratique. En effet, cette approches permet d'atteindre une précision d'estimation supérieure à la précision maximale indiquée par la borne de Cramér-Rao pour un estimateur classique. Nous avons également montré que cette approche permet d'utiliser un débit inférieur à celui indiqué par la borne de Shannon qui s'applique dans le cas du codage pur. Enfin, nous avons proposé une application originale et réaliste permettant de créer un mélange tatoué contenant de l'information complémentaire utilisable au décodeur pour guider un estimateur. Les résultats obtenus montrent un gain de qualité significatif par rapport à une approche classique non informée. Ce travail a fait l'objet de deux publications dans des conférences internationales [MF10, FM11].

Extraction d'informations musicales informée

Nous avons étendu l'utilisation de l'approche informée aux problématiques d'extraction d'informations symboliques à partir d'un signal audio. Ainsi, nous nous sommes intéressé plus précisément à l'estimation F_0 multiple qui est à l'origine de la transcription automatique ainsi que de nombreuses applications MIR reposant sur la représentation symbolique de la musique. Nous proposons d'abord une implémentation détaillée d'un système de transcription polyphonique que nous avons comparé à l'état de l'art. Nous proposons par la suite deux systèmes de type codeur / décodeur permettant d'appliquer l'approche informée combinée un système de transcription automatique. Enfin, nous proposons une application réaliste combinant estimation F_0 multiple informée et séparation de sources informée par la transcription MIDI à partir d'un signal de mélange tatoué. Cette contribution a fait l'objet d'une publication dans une conférence internationale [FM12].

Ces résultats montrent dans certains cas un gain significatif pour le débit utilisé en comparaison avec le codage simple. Cependant, le débit dépend évidemment du système de transcription utilisé et du type d'erreurs commises. Il peut donc arriver que la correction des erreurs soit plus coûteuse que le codage simple. Ainsi, ces derniers résultats sont donc préliminaires et ouvrent de nouvelles perspectives de travaux.

Perspectives

Nous avons vu dans ce document (principalement dans le chapitre 3) que l'estimation et le codage sont des domaines souvent traités indépendamment dans la littérature : pourtant cette thèse montre que ces deux approches peuvent être combinées pour traiter des problèmes variés, même dépassant le cadre du traitement du signal audio. Dans ce document, nous avons développé l'approche informée combinant estimation et codage selon

les trois axes décrits ci-après et qui nous l'espérons susciteront de nouveaux travaux de recherche.

Théorie

A l'heure actuelle, il n'existe que peu de travaux établissant les liens entre la théorie de l'estimation et la théorie de l'information [GSV05, Ver10]. Pourtant nous avons montré expérimentalement qu'il existait des configurations où ces deux approches peuvent être combinées pour offrir de nouvelles applications. Ainsi, la quantité d'information fournie par un estimateur peut donc être mesurée en bits Shannon et est donc compatible avec la théorie de l'information. Le fait d'établir formellement les liens qui existent entre la théorie de l'estimation et la théorie de l'information devrait permettre d'établir de nouvelles bornes théoriques telles que proposées dans [Wyn75]. Cela permettrait de généraliser le calcul du débit minimal nécessaire en fonction de la distorsion souhaitée et de la configuration traitée. On serait alors en mesure d'évaluer objectivement les performances de tous les systèmes basés sur l'approche informée et de démontrer leur optimalité.

Analyse paramétrique informée

Dans le chapitre 4 portant sur l'analyse spectrale informée, nous avons proposé une solution permettant d'estimer avec une grande précision les paramètres des composantes sinusoïdales appartenant à des sources sonores distinctes à partir d'un signal mono-canal de mélange. Ce travail ouvre donc nouvelles perspectives d'applications basées sur la manipulation des signaux audio (*cf.* écoute active de la musique [Lep98]).

La technique proposée pourrait être améliorée en combinant le modèle sinusoïdal avec d'autres modèles spécifiques de représentation des signaux [SOdBB03], permettant par exemple une représentation parcimonieuse de la partie bruitée ou des transitoires (*cf.* section 1.2.2). Cela permettrait une réduction du débit résultant, notamment lorsque le modèle sinusoïdal est mal adapté au signal traité (dans le cas d'une source sonore bruitée par exemple).

Ce changement de modèle devra idéalement s'accompagner de l'utilisation d'une autre mesure de distorsion prenant en compte la perception humaine [BFH+13]. A l'heure actuelle, les techniques décrites dans cette thèse restent spécifiques au modèle sinusoïdal utilisant une mesure de distorsion non perceptive.

Extraction d'informations musicales informée

Dans nos travaux, nous avons choisi de traiter la transcription automatique de la musique basée sur l'estimation F_0 multiple. La connaissance de la transcription exacte d'une pièce musicale permet d'obtenir d'autres informations souvent traitées par des applications MIR distinctes. La transmission d'un signal de mélange tatoué garantissant de retrouver la transcription de référence en utilisant une configuration codeur / décodeur peut permettre de nombreuses applications telles que la séparation des sources, l'indexation des signaux, la détection de *cover* [RE10], etc.

Nous avons montré que l'utilisation d'un estimateur F_0 multiple classique permet de réduire le débit nécessaire au codage de la transcription de référence. Cependant les gains observés sur les débits restent en pratique assez faible ($\leq 20\%$) lorsque l'on souhaite obtenir la transcription exacte. Ces premiers résultats doivent être considérés comme une preuve de faisabilité et nécessitent des améliorations.

Dans le cas où une transcription approximative est suffisante, il pourrait être intéressant de combiner un estimateur F_0 multiple avec d'autres types d'information *a priori*

(sous forme d'indices, de paramètres ou de descripteurs). Par exemple la connaissance de la distribution des F_0 (durée des notes, polyphonie, etc.), de la forme de l'enveloppe spectrale et de la tessiture de chaque source quasi-harmonique contenue dans un signal de mélange polyphonique devrait permettre une correction automatique des erreurs d'estimation sans l'utilisation d'un quelconque algorithme de codage.

ANNEXE A

CALCULS DÉTAILLÉS

A.1 Preuve d'impossibilité de l'inversion du tatouage audio numérique basé sur la technique QIM

Cette preuve a été réalisée grâce à l'aide de Jonathan Pinel, doctorant à l'université de Grenoble lorsque cette thèse a été rédigé. Cette preuve permet d'affirmer que le codage différentiel classique ne peut pas s'appliquer dans les problèmes d'analyse informée lorsque le signal analysé lui-même dépend de l'information cachée qu'il contient. Comme la plupart des techniques de tatouage audio numériques utilisent les bits de poids les plus faibles du code obtenu par quantification du signal analysé, on montre alors qu'à partir d'un signal x quantifié par QIM [CW01], il n'est pas possible d'appliquer une fonction de quantification permettant de se projeter sur un élément identique à celui qui serait obtenu en quantifiant le signal original. Cette preuve constitue donc un argument supplémentaire pour justifier le choix de l'utilisation de la technique d'analyse spectrale informée développée dans le chapitre 4 à la place d'un codage prédictif différentiel classique.

A.1.1 Principe de la quantification QIM

La quantification par QIM est utilisée pour transmettre de l'information cachée codée à l'intérieur des échantillons d'un signal. En général on choisit une base de représentation pour le signal à tatouer (par exemple la intMDCT) puis on calcule pour chaque échantillon la capacité de codage n correspondant au nombre de bits pouvant être utilisés. On définit alors 2^n quantificateurs (le plus souvent uniformes) entrelacés comme décrit dans la figure A.1.

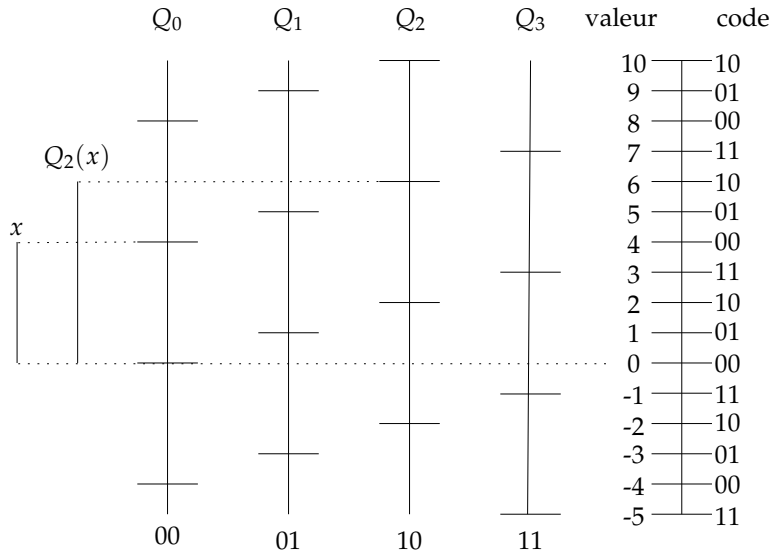


FIGURE A.1 – Exemple de quantification QIM codant un message c sur 2 bits avec 4 quantificateurs.

Pour coder un message c, il suffit alors de remplacer la valeur initiale d'un échantillon x par son représentant en utilisant le quantificateur correspondant (le message c correspond au choix du quantificateur).
Pour le décodage, on retrouve le quantificateur en identifiant l'origine du représentant utilisé pour coder l'échantillon. L'indice du quantificateur permet ainsi de retrouver le message codé. Une illustration du fonctionnement d'un quantificateur QIM utilisant 4 quantificateurs pour coder une message sur 2 bits est présenté dans la figure A.1. Un tel quantificateur combiné avec une base de représentation et un modèle perceptif adéquat permet d'atteindre de bonnes performances pour le tatouage des signaux audio [PGBP10].

A.1.2 Démonstration

Un quantificateur Q correspond à l'ensemble de ses représentants et $Q(x)$ est le représentant de x en utilisant le quantificateur Q. $Q(x)$ correspond donc à la valeur quantifiée de x en utilisant Q. On note $Q_{c,n}$ le quantificateur QIM codant c (tatouage) en utilisant n bits ($c \in [0, 2^n[$) et $Q_{n,c}(x)$. On note $\mathcal{S}_n = Q_{0,n}, Q_{1,n}, ..., Q_{2^n-1,n}$ l'ensemble des quantificateurs utilisant n bits (permettant de coder n'importe quel message $c \in [0, 2^n - 1]$.

.1) Proposition

On veut **nier** l'affirmation comme quoi il existe un quantificateur Q tel que $Q(Q_{c,n}(x)) = Q(x)$ pour $\forall c \forall n \in \mathbb{N}^*$. Cette affirmation peut se formuler comme suit :

$$\exists Q, \forall x \forall c \in [0, 2^n - 1], Q(Q_{c,n}(x)) = Q(x).$$

On cherche donc à démontrer sa négation pouvant s'exprimer comme :

$$\forall Q, \exists x \exists c \in [0, 2^n - 1], Q(Q_{c,n}(x)) \neq Q(x). \tag{A.1}$$

.2) Preuve

On suppose que tous les quantificateurs utilisés sont uniformes avec arrondi vers $-\infty$.

- Soit un quantificateur Q et soient r_1 et r_2 deux de ses représentants consécutifs ($r_1 < r_2$).
- Soit $f = \frac{r_1+r_2}{2}$ la "frontière" entre r_1 et r_2.
- Soit $a = \sup_{x \in \bigcup_{Q \in \mathcal{S}_n}} (Q)$ tel que $x \leq f$ le plus grande représentant d'un des $Q_{c,n}$ plus petit que f.
- Soit $b = \inf_{x \in \bigcup_{Q \in \mathcal{S}_n}} (Q)$ tel que $x > f$ le plus petit représentant d'un des $Q_{c,n}$ strictement supérieur à f.

Par définition a et b sont distincts. De plus, si on fixe c et c' tels que :
$a \in Q_{c,n}$ et $b \in Q_{c',n}$
alors $c \neq c'$, car les quantificateurs $Q_{c,n}$ sont entrelacés par construction. De plus, il n'existe pas de représentant d'un des $Q_{c,n} \in \mathcal{S}_n$ qui soit entre a et b. On a donc

$$\begin{aligned} Q(b) &= r_2 \\ Q(Q_{c,n}(b)) &= Q(a) = r_1. \end{aligned}$$

On a donc $Q(b) \neq Q(Q_{c,n}(b))$ qui démontre l'affirmation décrite par l'équation (A.1).

A.2 Calcul de la borne de Shannon pour les paramètres sinusoïdaux

Ici, on s'intéresse à la borne Shannon débit-distorsion utile pour la quantification des paramètres sinusoïdaux.

A.2.1 Modèle sinusoïdal

Un signal composé d'une seule sinusoïde utilisant le modèle stationnaire peut s'exprimer de la façon suivante pour une trame de signal de taille N :

$$s[n] = a \exp\left(\mathbf{j}(\omega n + \phi)\right), \tag{A.2}$$

avec a, ω et ϕ respectivement l'amplitude, la fréquence et la phase (*cf.* section 2.1).

A.2.2 Mesure de distorsion

Pour quantifier optimalement une sinusoïde, on définit une mesure d'erreur, ici il s'agit de l'erreur quadratique pondérée par un vecteur de poids w définie comme suit :

$$\begin{aligned} d(a,\omega,\phi,\hat{a},\hat{\omega},\hat{\phi}) &= \sum_{n=n_0}^{n_0+N-1} \left| w[n] \left(a \exp\left(\mathbf{j}(\omega n + \phi)\right) - \hat{a} \exp\left(\mathbf{j}(\hat{\omega} n + \hat{\phi})\right)\right)\right|^2 \\ &= \sum_{n=n_0}^{n_0+N-1} \left| w[n] a \exp\left(\mathbf{j}(\omega n + \phi)\right) - w[n] \hat{a} \exp\left(\mathbf{j}(\hat{\omega} n + \hat{\phi})\right)\right|^2 \\ &= ||w||^2 (a^2 + \hat{a}^2) - 2a\hat{a} \sum_{n=n_0}^{n_0+N-1} w[n]^2 \cos\left((\omega - \hat{\omega})n + (\phi - \hat{\phi})\right), \end{aligned}$$

avec $n_0 = -N/2$.

En utilisant le développement limité d'ordre 2 de la fonction $\cos(x) \approx 1 - \frac{x^2}{2}$ et en introduisant la notation $\Delta_p = p - \hat{p}$ pour chaque paramètre p et en fixant :

$$\sigma^2 = \frac{1}{||w||^2} \sum_{n=n_0}^{n_0+N-1} w[n]^2 n^2, \tag{A.3}$$

on obtient :

$$\begin{aligned} d &= ||w||^2(\Delta_a^2 + 2a\hat{a}) - 2a\hat{a} \sum_{n=n_0}^{n_0+N-1} w[n]^2 \left(1 - \frac{1}{2}(\Delta_\omega^2 n^2 + 2n\Delta_\omega\Delta_\phi + \Delta_\phi^2)\right) \\ &= ||w||^2\Delta_a^2 + a\hat{a} \left(\underbrace{\sum_{n=n_0}^{n_0+N-1} w[n]^2 n^2}_{\sigma^2||w||^2} \Delta_\omega^2 + \sum_{n=n_0}^{n_0+N-1} w[n]^2 \Delta_\phi^2 + 2 \sum_{n=n_0}^{n_0+N-1} w[n]^2 n \Delta_\omega \Delta_\phi \right) \\ &= ||w||^2 \left(\Delta_a^2 + a\hat{a} \left(\sigma^2\Delta_\omega^2 + \Delta_\phi^2 + \underbrace{\frac{2}{||w||^2} \sum_{n=n_0}^{n_0+N-1} w[n]^2 n \Delta_\phi \Delta_\omega}_{\approx 0} \right) \right). \end{aligned}$$

En effet, pour une fenêtre symétrique et en utilisant $n_0 = -N/2$ le dernier terme s'annule. En utilisant l'approximation $a\hat{a} \approx a^2$ on obtient :

$$d_a(\Delta_a, \Delta_\omega, \Delta_\phi) \approx ||w||^2 \left(\Delta_a^2 + a^2 \left(\Delta_\phi^2 + \sigma^2\Delta_\omega^2\right)\right), \tag{A.4}$$

A.2.3 Borne Shannon

La borne Shannon[1] est définie comme le débit minimal théorique (en bits) permettant de coder une source (une variable aléatoire) avec une distorsion moyenne maximale D. Ainsi, si on considère une variable aléatoire p et sa version quantifiée $\hat{p}$, la borne de Shannon permet de minorer toutes les fonctions débit-distorsion minimisant l'information mutuelle entre p et $\hat{p}$ (*e.g.* ECUSQ). D'après [Gra89], cette borne de Shannon peut s'exprimer comme suit :

$$R_{\text{SLB}}(D) = h(p) + \log(\mathring{a}(D)) - D\mathring{b}(D), \tag{A.5}$$

avec $h(p)$ l'entropie de p et D la distorsion moyenne. Les 2 membres $\mathring{a}(D)$ et $\mathring{b}(D)$ sont solution des équations suivantes :

$$\mathring{a}(D) \int \mathbf{e}^{-\mathring{b}(D)L(x)}\, dx = 1, \tag{A.6}$$

et

$$\mathring{a}(D) \int L(x)\, \mathbf{e}^{-\mathring{b}(D)L(x)}\, dx = D, \tag{A.7}$$

où $L(x)$ est la mesure de la distorsion en fonction de l'erreur x entre deux réalisations provenant respectivement de p et de $\hat{p}$. Par la suite, on considère que la variable aléatoire p (resp. $\hat{p}$) est une combinaison de 3 variables aléatoires notées A, Ω et Φ et dont chaque réalisation est un triplet (a, ω, ϕ) (*i.e.* $x \in \mathbb{R}^3$).

[1] *Shannon Lower Bound (SLB)*

a) Entropie conjointe

Dans notre modèle, on suppose que a et ω suivent respectivement une loi de Rayleigh de paramètre σ_a et σ_ω et ϕ suit une loi uniforme $f_\Phi = \frac{1}{2\pi}$. En supposant que les 3 paramètres sont indépendants, l'entropie peut s'écrire :

$$\begin{aligned} h(p) &= h(A, \Omega, \Phi) \\ &= h(A) + h(\Omega) + h(\Phi) \\ &= 1 + \frac{\gamma}{2} + \log\left(\frac{\sigma_a}{\sqrt{2}}\right) + 1 + \frac{\gamma}{2} + \log\left(\frac{\sigma_\omega}{\sqrt{2}}\right) + \log\left(\frac{1}{2\pi}\right) \\ &= 2 + \gamma + \log\left(\frac{\sigma_a \sigma_\omega}{4\pi}\right). \end{aligned}$$

avec γ la constante d'Euler-Mascheroni.

b) Calcul de $\mathring{a}(D)$ et $\mathring{b}(D)$

En utilisant l'équation (A.4) comme mesure de distorsion, on obtient en utilisant (A.3) et en posant $\Delta_a = x_a$, $\Delta_\omega = x_\omega$ et $\Delta_\phi = x_\phi$:

$$L(x) = L_{a,\sigma}(x_a, x_\omega, x_\phi) = ||w||^2 \left(x_a^2 + a^2 x_\phi^2 + a^2 \sigma^2 x_\omega^2\right). \tag{A.8}$$

Ainsi l'équation (A.6) peut s'écrire :

$$\mathring{a}(D) \iiint \mathbf{e}^{-\mathring{b}(D)||w||^2(x_a^2 + a^2 x_\phi^2 + a^2\sigma^2 x_\omega^2)}\, dx_a\, dx_\phi\, dx_\omega = 1. \tag{A.9}$$

Par changement de variable ($X_\phi = a x_\phi$ et $X_\omega = a\sigma x_\omega$) on peut écrire :

$$\begin{aligned} 1 &= \mathring{a}(D) \iiint \mathbf{e}^{-\mathring{b}(D)||w||^2(x_a^2 + X_\phi^2 + X_\omega^2)} a^{-2}\sigma^{-1}\, dx_a\, dX_\phi\, dX_\omega \\ &= \mathring{a}(D) a^{-2}\sigma^{-1} \int_{\mathbb{R}^3} \mathbf{e}^{-\mathring{b}(D)||w||^2||x||^2}\, dx \\ &= \mathring{a}(D) a^{-2}\sigma^{-1} \left(\frac{\pi}{||w||^2 \mathring{b}(D)}\right)^{3/2}, \end{aligned}$$

donc $\mathring{a}(D) = a^2\sigma \left(\frac{||w||^2\mathring{b}(D)}{\pi}\right)^{3/2} = a^2\sigma||w||^3 \left(\frac{\mathring{b}(D)}{\pi}\right)^{3/2}$.

Pour le calcul de $\mathring{b}(D)$ on utilise l'équation (A.7) :

$$\begin{aligned} D &= \mathring{a}(D) \iiint ||w||^2(x_a^2 + X_\phi^2 + X_\omega^2)\, \mathbf{e}^{-\mathring{b}(D)||w||^2(x_a^2 + X_\phi^2 + X_\omega^2)} a^{-2}\sigma^{-1}\, dx_a\, dX_\phi\, dX_\omega \\ &= \underbrace{\mathring{a}(D)||w||^2 a^{-2}\sigma^{-1}}_{K} \iiint (x_a^2 + X_\phi^2 + X_\omega^2)\, \mathbf{e}^{\underbrace{-\mathring{b}(D)||w||^2}_{\alpha}(x_a^2 + X_\phi^2 + X_\omega^2)}\, dx_a\, dX_\phi\, dX_\omega \\ &= K \iiint (x_a^2 + X_\phi^2 + X_\omega^2)\, \mathbf{e}^{-\alpha(x_a^2 + X_\phi^2 + X_\omega^2)}\, dx_a\, dX_\phi\, dX_\omega. \end{aligned}$$

En séparant les termes x_a, X_Φ et X_ω on obtient :

$$
\begin{aligned}
D &= K\left(\iiint x_a^2\,\mathbf{e}^{-\alpha(x_a^2+X_\phi^2+X_\omega^2)}\,dx_a\,dX_\phi\,dX_\omega + \iiint X_\phi^2\,\mathbf{e}^{-\alpha(x_a^2+X_\phi^2+X_\omega^2)}\,dx_a\,dX_\phi\,dX_\omega\right.\\
&\left. + \iiint X_\omega^2\,\mathbf{e}^{-\alpha(x_a^2+X_\phi^2+X_\omega^2)}\,dx_a\,dX_\phi\,dX_\omega\right)\\
&= K\left(\int x_a^2\,\mathbf{e}^{-\alpha x_a^2}dx_a\cdot\int \mathbf{e}^{-\alpha X_\omega^2}\,dX_\omega\cdot\int \mathbf{e}^{-\alpha X_\phi^2}dX_\phi + \int X_\phi^2 e^{-\alpha X_\phi^2}\,dX_\phi\cdot\int \mathbf{e}^{-\alpha x_a^2}\,dx_a\cdot\int \mathbf{e}^{-\alpha X_\omega^2}\,dX_\omega\right.\\
&\left. + \int X_\omega^2\,\mathbf{e}^{-\alpha X_\omega^2}\,dX_\omega\cdot\int \mathbf{e}^{-\alpha x_a^2}\,dx_a\cdot\int \mathbf{e}^{-\alpha X_\phi^2}\,dX_\phi\right)\\
&= 3K\cdot\left(\frac{\pi}{4\alpha^3}\right)^{1/2}\cdot\left(\frac{\pi}{\alpha}\right)^{1/2}\cdot\left(\frac{\pi}{\alpha}\right)^{1/2} = \frac{3}{2}K\frac{\pi^{3/2}}{\alpha^{5/2}}.
\end{aligned}
$$

En remplaçant K et α on trouve :

$$
\begin{aligned}
D &= \frac{3\mathring{a}(D)||w||^2\pi^{3/2}}{2a^2\sigma\mathring{b}(D)^{5/2}||w||^5}\\
&= \mathring{a}(D)\frac{3\pi^{3/2}}{2a^2\sigma\mathring{b}(D)^{5/2}||w||^3}.
\end{aligned}
$$

En substituant $\mathring{a}(D)$ par le résultat obtenu précédemment on obtient :

$$
D = \frac{3}{2}\mathring{b}(D)^{-1}.
$$

On déduit par substitution que $\mathring{b}(D) = \frac{3}{2D}$ et que $\mathring{a}(D) = a^2\sigma||w||^3(\frac{2}{3}\pi D)^{-3/2}$.

c) Calcul de la borne de Shannon

D'après l'équation (A.5), la borne de Shannon vaut :

$$
\begin{aligned}
R_{\mathrm{SLB}}(D) &= h(p) + \log(\mathring{a}(D)) - D\mathring{b}(D)\\
&= 2 + \gamma + \log\left(\frac{\sigma_a\sigma_\omega}{4\pi}\right) + \log\left(a^2\sigma||w||^3\right) - \frac{3}{2}\log\left(\frac{2}{3}\pi D\right) - \frac{3}{2}\\
&= \frac{1}{2} + \gamma + \log\left(a^2\sigma||w||^3\frac{\sigma_a\sigma_\omega}{4\pi}\right) - \frac{3}{2}\log\left(\frac{2}{3}\pi D\right). \qquad \text{(A.10)}
\end{aligned}
$$

Par inversion, on en déduit la distorsion moyenne en fonction du débit :

$$
D(R) = \frac{3\,\mathbf{e}^{\frac{2\gamma+1-2R}{3}}}{2\pi}\cdot\left(\frac{a^2\sigma||w||^3\sigma_a\sigma_\omega}{4\pi}\right)^{2/3}. \qquad \text{(A.11)}
$$

d) Expérimentations

Dans cette section, on génère 1000 signaux synthétiques d'après le modèle sinusoïdal présenté à l'équation (A.2) dont chaque paramètre est tiré aléatoirement. L'amplitude et la fréquence suivent une loi Rayleigh de paramètres $\sigma_a = 0.25$ et $\sigma_\omega = 2\pi 1000$. La phase suit une loi uniforme sur $[0, 2\pi]$. La figure A.2 compare la distorsion en décibels (dB) obtenue par la quantification ECUSQ pratique (en rouge) et théorique (en bleu). En vert il s'agit de la distorsion obtenue par resynthèse des paramètres estimés par la méthode de réallocation non informée (*cf.* section 2.1.1) pour une sinusoïde mélangée avec un bruit additif tel que le SNR vaut 10dB. La courbe noire correspond à la distorsion obtenue par l'analyse informée utilisant une quantification maximale uniforme de 32 bits utilisant exactement le même débit que la méthode de quantification ECUSQ pratique. La quantification choisie pour l'estimation informée est différente car en utilisant les même cellules de quantification que ECUSQ il est impossible d'obtenir une distorsion plus faible.

Sur la figure (A.2) on observe que comme attendu, le quantificateur ECUSQ pratique atteint sa valeur théorique donnée par (4.35) en fonction du débit (on fixe ici $R = H_t$ où H_t correspond à l'entropie cible). Comme attendu, la borne de Shannon donnée par (A.10) qui n'est jamais atteinte en pratique est inférieure au débit théorique du quantificateur ECUSQ (*cf.* figure A.3). La combinaison de l'estimateur combiné avec le même débit que pour le quantificateur ECUSQ en pratique permet d'obtenir une distorsion plus faible que la distorsion ECUSQ théorique et plus faible que la distorsion théorique indiquée par la borne de Shannon.

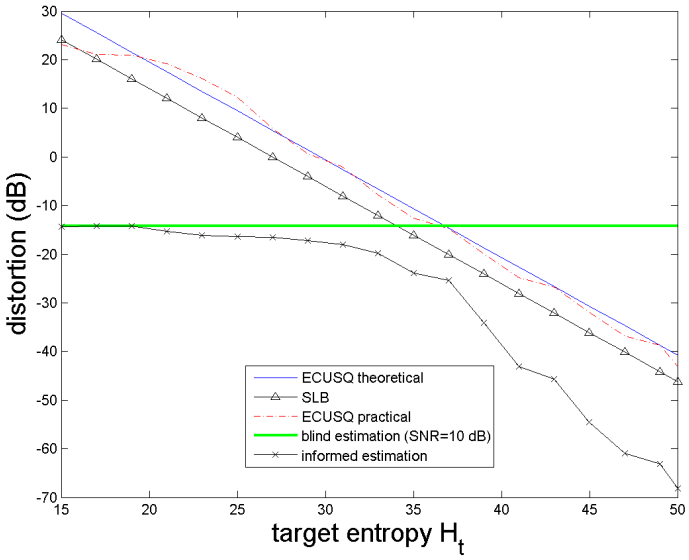


FIGURE A.2 – Comparatif des distorsions théoriques (SLB et ECUSQ) et pratiques (codage pur ECUSQ, estimation aveugle et estimation informée).L'estimation informée combine un estimateur avec le même débit que pour le codage pur. On remarque que combiner un estimateur avec de l'information permet d'obtenir une distorsion plus faible que la distorsion théorique de ECUSQ et de la borne de Shannon.

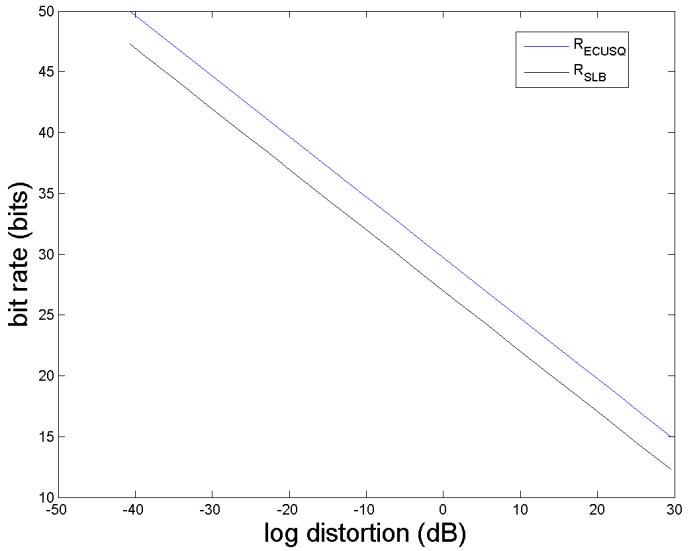


FIGURE A.3 – Comparaison entre la borne de Shannon et la borne ECUSQ théorique. Comme attendu, la distorsion moyenne théorique de la méthode ECUSQ est supérieure à la distorsion Shannon. La différence moyenne mesurée est d'environ 2.7bits.

BIBLIOGRAPHIE DE L'AUTEUR

Journaux internationaux avec comité de relecture

- **Dominique Fourer** and Sylvain Marchand.Informed Spectral Analysis : audio signal parameters estimation using side information. EURASIP, Journal on Advances in Signal Processing

Conférences internationales avec actes

- S. Marchand, R. Badeau, C. Baras, L. Daudet, **D. Fourer**, L. Girin, S. Gorlow, A. Liutkus, J. Pinel, G. Richard, N. Sturmel, S. Zhang. DReaM : a novel system for joint source separation and multi-track coding, in AES 133rd Convention, 2012.

- **Dominique Fourer** and Sylvain Marchand. Informed Multiple-F0 Estimation Applied to Monaural Audio Source Separation. In Proceedings of the European Signal Processing Conference (EUSIPCO'12), Bucharest, Romania, August 2012.

- **Dominique Fourer** and Sylvain Marchand. Informed Spectral Analysis for Isolated Audio Source Parameters Estimation. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11), New Paltz, New York, USA, Oct. 2011. Institute of Electrical and Electronics Engineers (IEEE).

- Sylvain Marchand and **Dominique Fourer**. Breaking the Bounds : Introducing Informed Spectral Analysis. In Proceedings of the Digital Audio Effects (DAFx'10) Conference, pages 359-366, Graz, Austria, Sept. 2010.http://dafx10.iem.at

Communications sans actes

- Informed Spectral Analysis for Under Determined Audio Source Separation, 6e école d'été de Peyresq en traitement du signal et des images.http://peyresq11.u-bourgogne.fr/. GRETSI, Peyresq, Juil. 2011.

- Breaking the bounds : introducing informed spectral analysis, JJCAAS 2010, 6e Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio. IRCAM, Paris, Nov. 2010.

Séminaires

- Dominique Fourer, Informed Spectral Analysis for Isolated Audio Source Parameters Estimation, Séminaire problèmes inverses, IMB, Bordeaux, Nov 2011.http://spi.labri.fr

BIBLIOGRAPHIE GÉNÉRALE

[AEJ+12] A. ADLER, V. EMIYA, M.G. JAFARI, M. ELAD, R. GRIBONVAL et M.D. PLUMBLEY : Audio inpainting. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 20(3):922–932, 2012. *Cité page 8*

[AF95] François AUGER et Patrick FLANDRIN : Improving the readibility of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, mai 1995. *2 citations pages 26 et 77*

[APF09] David S. ALVES, Jouni PAULUS et José FONSECA : Drum transcription from multichannel recordings with non-negative matrix factorization. *In Proc. IEEE European Signal Processing Conference (EUSIPCO'09)*, pages 894–898, Glasgow, Ecosse, août 2009. *Cité page 102*

[AS05] M. ABE et J.O. SMITH : Am/fm rate estimation for time-varying sinusoidal modeling. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'05)*, volume 3, pages 201–204, Philadelphia, PA, USA, 2005. *Cité page 25*

[Bad06] Roland BADEAU : *Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées. Application aux signaux de musique.* Thèse de doctorat, Telecom ParisTech, avril 2006. *2 citations pages 24 et 78*

[Bar06] Cléo BARAS : *Tatouage informé de signaux audio numériques*. Thèse de doctorat, Télécom Paris, 2006. *Cité page 136*

[BB97] Karlheinz BRANDENBURG et Marina BOSI : Overview of MPEG audio : Current and future standards for low bit-rate audio coding. *Journal of the Audio Engineering Society*, 45(1/2):4–21, 1997. *2 citations pages 61 et 93*

[BBV10] Nancy BERTIN, Roland BADEAU et Emmanuel VINCENT : Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, mars 2010. *8 citations pages 102, 103, 109, 110, 115, 124, 181 et 182*

[BCRD08] M. BETSER, P. COLLEN, G. RICHARD et B. DAVID : Estimation of frequency for am/fm models using the phase vocoder framework. *IEEE Transactions on Signal Processing*, 56(2):505–517, février 2008. *Cité page 78*

[BDA+05] J.P. BELLO, L. DAUDET, S. ABDALLAH, C. DUXBURY, M. DAVIES et Mark B. SANDLER : A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005. *Cité page 16*

[BDR06] R. BADEAU, B. DAVID et G. RICHARD : High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *IEEE Transactions on Signal Processing*, 54(4):1341–1350, avril 2006. *2 citations pages 24 et 78*

[BED09] Mert BAY, Andreas F. EHMANN et J. Stephen DOWNIE : Evaluation of multiple-f0 estimation and tracking systems. *In* Keiji HIRATA, George TZANETAKIS et Kazuyoshi YOSHII, éditeurs : *Proc. of the International Conference of Music Information Retrieval (ISMIR)*, pages 315–320, Kobe, Japon, octobre 2009. International Society for Music Information Retrieval. *2 citations pages 36 et 115*

[BFH+13] K. BRANDENBURG, C. FALLER, J. HERRE, J.D. JOHNSTON et W.B. KLEIJN : Perceptual coding of high-quality digital audio. *Proc. IEEE*, 101(9):1905–1919, 2013. *Cité page 153*

[Bla47] D. BLACKWELL : Conditional expectation and unbiased sequential estimation. *Ann.Math.Statist.*, 18:105–110, 1947. *Cité page 52*

[BM01] P. BOFILL et M.ZIBULEVSKI : Underdetermined blind source separation. *Signal Processing*, 81(11):2353–2362, 2001. *Cité page 102*

[Bre90] A. S. BREGMAN : *Auditory scene analysis*. MIT Press : Cambridge, MA, 1990. *3 citations pages 30, 40 et 42*

[Bri95] A. C. Den BRINKER : Meixner-like functions having a rational z-transform. *International Journal of Circuit Theory and Applications*, 23(3):237–246, 1995. *Cité page 15*

[Cam07] Arturo CAMACHO : *SWIPE : A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. Thèse de doctorat, University of Florida, USA, 2007. http://www.cise.ufl.edu/~acamacho/english/. *Cité page 33*

[CJ10] P. COMON et C. JUTTEN : *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. Academic Press, 1 édition, avril 2010. *2 citations pages 38 et 40*

[CJF73] P. CUMMISKEY, N. S. JAYANT et J. L. FLANAGAN : Adaptive quantization in differential pcm coding of speech. *Bell System Technical Journal*, 52:1105–1118, 1973. *Cité page 66*

[CLG89] P. A. CHOU, T. LOOKABAUGH et R. M. GRAY : Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(1):31–42, 1989. *4 citations pages 65, 71, 72 et 85*

[Com94] P. COMON : Independent component analysis, a new concept ? *Signal Processing*, 36:287–314, avril 1994. *2 citations pages 16 et 42*

[Cos83] M. H M COSTA : Writing on dirty paper (corresp.). *IEEE Transactions on Information Theory*, 29(3):439–441, 1983. *Cité page 67*

[CSK77] Donald G. CHILDERS, D.P. SKINNER et R.C. KEMERAIT : The cepstrum : A guide to processing. *Proc. IEEE*, 65(10):1428–1443, 1977. *Cité page 35*

[CW01] B. CHEN et G.W. WORNELL : Quantization index modulation : a class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4):1423–1443, mai 2001. *3 citations pages 71, 93 et 155*

[dCK02] A. de CHEVEIGNÉ et H. KAWAHARA : YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002. *6 citations pages 33, 34, 101, 115, 126 et 182*

[Dix01] S. DIXON : Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30:39–58, 2001. *Cité page 32*

[DK90] Petar M. DJURIĆ et Steven M. KAY : Parameter estimation of chirp signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(12):2118–2126, décembre 1990. *Cité page 78*

[DLR77] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. *Cité page 53*

[DM01] N. DAY et J.M. MARTINEZ : Introduction to MPEG-7. Iso/iec jtc1/sc29/wg11 n4325, International Organisation for Standartization (ISO), 2001. http ://ipsi.fraunhofer.de/delite/Projects/MPEG7/. *Cité page 2*

[Don06] D.L. DONOHO : Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, avril 2006. *Cité page 8*

[DPF09] C. DOSSAL, G. PEYRÉ et J. FADILI : A numerical exploration of compressed sampling recovery. *In Proc. SPARS'09*, Saint-Malo, France, avril 2009. *Cité page 8*

[DZZS08] Zhiyao DUAN, Yungang ZHANG, Changshui ZHANG et Zhenwei SHI : Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(4):766–778, mai 2008. *2 citations pages 2 et 43*

[Emi08] Valentin EMIYA : *Transcription automatique de la musique de piano*. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 37-39 rue Dareau, 75014 Paris, octobre 2008. *Cité page 110*

[ES06] M.R. EVERY et J.E. SZYMANSKI : Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(5):1845–1856, septembre 2006. *4 citations pages 40, 104, 120 et 137*

[ET04] Tuomas EEROLA et Petri TOIVIAINEN : *MIDI Toolbox : MATLAB Tools for Music Research*. University of Jyväskylä, Jyväskylä, Finland, 2004. *Cité page 109*

[EVHH11] V. EMIYA, E. VINCENT, N. HARLANDER et V. HOHMANN : Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(7):2046–2057, 2011. *2 citations pages 44 et 45*

[Fed98] Riccardo Di FEDERICO : Waveform preserving time stretching and pitch shifting for sinusoidal models of sound. *In Proc. Digital Audio Effects Conference (DAFx'98)*, novembre 1998. *Cité page 23*

[FGLS05] J. FRIDRICH, M. GOLJAN, P. LISONEK et D. SOUKAL : Writing on wet paper. *IEEE Transactions on Signal Processing*, 53(10):3923–3935, 2005. *Cité page 67*

[FGS95] Hans G. FEICHTINGER, Karlheinz GRÖCHENIG et Thomas STROHMER : Efficient numerical methods in non-uniform sampling theory. *Numer. Math*, 69:423–440, 1995. *Cité page 10*

[Fis25] R.A. FISHER : Theory of statistical estimation. *In Proc. Cambridge Philos. Soc. 22*, pages 700–725, 1925. *2 citations pages 52 et 53*

[FM11] D. FOURER et S. MARCHAND : Informed spectral analysis for isolated audio source parameters estimation. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'11)*, pages 57–60, New Platz, NY, USA, octobre 2011. *Cité page 152*

[FM12] D. FOURER et S. MARCHAND : Informed multiple-f0 estimation applied to monaural audio source separation. *In Proc. IEEE European Signal Processing Conference (EUSIPCO'12)*, pages 2158–2162, Bucharest, Roumanie, août 2012. *Cité page 152*

[FM13] D. FOURER et S. MARCHAND : Informed spectral analysis : audio signal parameters estimation using side information. *Journal on Advances in Signal Processing (EURASIP JASP)*, page accepté pour publication, novembre 2013. *Cité page 152*

[FP06] Derry FITZGERALD et Journi PAULUS : Unpitched percussion transcription. *In* KLAPURI et DAVY [KD06], pages 101–129. *Cité page 102*

[FR98] N. F. FLETCHER et T. D. ROSSING : *The Physics of Musical Instruments*. Springer-Verlag, 1998. *4 citations pages 16, 32, 35 et 110*

[Fri04] B. Roy FRIEDEN : *Science from Fisher Information : A Unification*. Cambridge University Press, juin 2004. *Cité page 49*

[GB03] R. GRIBONVAL et E. BACRY : Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51:101–111, janvier 2003. *2 citations pages 35 et 110*

[GG91] A. GERSHO et R. M. GRAY : *Vector quantization and signal compression*. Kluwer Academic Publishers, USA, 1991. *4 citations pages 63, 67, 72 et 180*

[GM84] A. GROSSSMANN et J. MORLET : Decomposition of hardy functions into square integrable wavelets of constant shape. *IAM Journal of Mathematical Analysis*, 15(4):723–736, 1984. *Cité page 14*

[GM13] S. GORLOW et S. MARCHAND : Informed audio source separation using linearly constrained spatial filters. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 21(1):3–13, 2013. *Cité page 67*

[GMdM+03] L. GIRIN, S. MARCHAND, J. di MARTINO, A. ROBEL et G. PEETERS : Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'03)*, pages 193–196, New Platz, NY, USA, octobre 2003. *2 citations pages 77 et 137*

[GMai] L. GIRIN et S. MARCHAND : Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'04)*, volume 1, pages I–633–6 vol.1, Montreal, Quebec, Canada, mai. *Cité page 136*

[GN98] R.M. GRAY et D.L. NEUHOFF : Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. *Cité page 86*

[God93] S.J. GODSILL : *The Restoration of Degraded Audio Signals*. Thèse de doctorat, Cambridge University, UK, 1993. *Cité page 2*

[GR08] O. GILLET et G. RICHARD : Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 16(3):529–540, mars 2008. *Cité page 102*

[GR13] S. GORLOW et J.D. REISS : Model-based inversion of dynamic range compression. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 21(7):1434–1444, 2013. *2 citations pages 67 et 152*

[Gra89] Robert M. GRAY : *Source coding theory*. Kluwer Academic Publishers, USA, octobre 1989. *4 citations pages 61, 71, 84 et 158*

[GSMA10] J. GANSEMAN, P. SCHEUNDERS, G. J. MYSORE et J. S ABEL : Evaluation of a score-informed source separation system. *In Proc. of the International Conference of Music Information Retrieval (ISMIR)*, pages 219–224, août 2010. *3 citations pages 102, 120 et 137*

[GSV05] Dongning GUO, S. SHAMAI et S. VERDU : Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. *3 citations pages 50, 65 et 153*

[Hai04] Stephen Webley HAINSWORTH : *Techniques for the Automated Analysis of Musical Audio*. Thèse de doctorat, University of Cambridge, UK, septembre 2004. *Cité page 106*

[Hai06] Stephen HAINSWORTH : Beat tracking and musical metre analysis. *In* KLAPURI et DAVY [KD06], pages 101–129. *2 citations pages 2 et 101*

[Han03] P. HANNA : *Statistical modeling of noisy sounds*. Thèse de doctorat, Université de Bordeaux I, Talence, France, décembre 2003. *2 citations pages 15 et 27*

[HBKD06] Perfecto HERRERA-BOYER, Anssi KLAPURI et Manuel DAVY : Automatic classification of pitched musical instruments sounds. *In* KLAPURI et DAVY [KD06], pages 101–129. *2 citations pages 2 et 20*

[HD07] J. HERRE et S. DISCH : New concepts in parametric coding of spatial audio : From SAC to SAOC. *In IEEE International Conference on Multimedia and Expo (ICME'07)*, pages 1894–1897, Beijing, Chine, juillet 2007. *Cité page 152*

[HDB11] R. HENNEQUIN, B. DAVID et R. BADEAU : Score informed audio source separation using a parametric model of non-negative spectrogram. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'11)*, pages 45–48, Prague, République tchèque, Mai 2011. *5 citations pages 102, 115, 120, 136 et 137*

[HDT03] R.B. Ning HU, DANNENBERG et G. TZANETAKIS : Polyphonic audio matching and alignment for music retrieval. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'03)*, pages 185–188, New Platz, NY, USA, octobre 2003. *Cité page 35*

[HK06] R. HUBER et B. KOLLMEIER : Pemo-q. a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(6):1902–1911, 2006. *Cité page 45*

[Huf52] D.A. HUFFMAN : A method for the construction of minimum-redundancy codes. *Proc. IRE*, 40(9):1098–1101, 1952. *Cité page 59*

[Idi01] J. IDIER : *Approche bayésienne pour les problèmes inverses*. Hermès - Lavoisier, 1re édition édition, janvier 2001. *Cité page 38*

[Jol02] L. T. JOLLIFFE : *Principal Component Analysis*. Springer-Verlag, New York, NY, USA, second edition édition, 2002. *2 citations pages 16 et 42*

[JRY00] A. JOURJINE, S. RICKARD et O. YILMAZ : Blind separation of disjoint orthogonal signals : demixing n sources from 2 mixtures. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'00)*, pages 2985–2988, Istanbul, Turquie, juin 2000. *Cité page 42*

[Jut07] C. JUTTEN : *Séparation de sources. : au-delà de l'aveugle et applications*, volume 2. Hermes Science Publications, 1 édition, février 2007. *Cité page 38*

[Kas06] Kunio KASHINO : Auditory scene analysis in music signals. *In* KLAPURI et DAVY [KD06], pages 299–325. *Cité page 42*

[KD06] Anssi KLAPURI et Manuel DAVY, éditeurs. *Signal Processing Methods for Music Transcription*. Springer US, 2006. *4 citations pages 168, 169, 170 et 177*

[KdVG76] Kunihiko KODERA, Claude de VILLEDARY et Roger GENDRIN : A new method for the numerical analysis of non-stationary signals. *Physics of the Earth and Planetary Interiors*, 12:142–150, 1976. *2 citations pages 26 et 77*

[KGdV78] Kunihiko KODERA, Roger GENDRIN et Claude de VILLEDARY : Analysis of time-varying signals with small bt values. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):64–76, février 1978. *2 citations pages 26 et 77*

[KJ92] Samuel KOTZ et NormanL. JOHNSON, éditeurs. *Breakthroughs in Statistics*. Springer Series in Statistics. Springer New York, 1992. *Cité page 49*

[KJH07] P. KORTEN, J. JENSEN et R. HEUSDENS : High-resolution spherical quantization of sinusoidal parameters. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 15(3):966–981, mars 2007. *4 citations pages 84, 85, 86 et 87*

[Kla03] A. KLAPURI : Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 11(6):804–816, novembre 2003. *7 citations pages 2, 101, 102, 103, 115, 125 et 182*

[Kla06] Anssi KLAPURI : Multiple fundamental frequency estimation by summing harmonic amplitudes. *In Proc. of the International Conference of Music Information Retrieval (ISMIR)*, pages 216–221, 2006. *Cité page 35*

[KM02] F. KEILER et S. MARCHAND : Survey on extraction of sinusoids in stationary sounds. *In Proc. Digital Audio Effects Conf. (DAFx'02)*, pages 51–58, septembre 2002. *2 citations pages 25 et 76*

[Knu05] Kevin H. KNUTH : Informed source separation : A bayesian tutorial. *In Proc. IEEE European Signal Processing Conference (EUSIPCO'05)*, 2005. *4 citations pages 40, 67, 135 et 152*

[KTL11] B.P. KEEGAN, S.K. TJOA et K.J.R. LIU : Super-resolution of musical signals using approximate matching pursuit. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'11)*, pages 81–84, New Platz, NY, USA, octobre 2011. *Cité page 8*

[KZ01] F. KEILER et U. ZÖLZER : Extracting sinusoids from harmonic signals. *Journal of New Music Research, Special Issue : "Musical Applications of Digital Signal Processing"*, 30(3):243–258, septembre 2001. *3 citations pages 25, 26 et 179*

[Lar01] J. LAROCHE : Estimating tempo, swing and beat locations in audio recordings. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'01)*, pages 135–138, New Platz, NY, USA, avril 2001. *Cité page 32*

[LBD+10] Mathieu LAGRANGE, Roland BADEAU, Bertrand DAVID, Nancy BERTIN, Jose ECHEVESTE, Olivier DERRIEN, Sylvain MARCHAND et Laurent DAUDET : The DESAM toolbox : spectral analysis of musical audio. *In Proc. Digital Audio Effects Conf. (DAFx'10)*, pages 254–261, Graz, Autriche, septembre 2010. *Cité page 115*

[LBG80] Y. LINDE, A. BUZO et R. M. GRAY : An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980. *Cité page 64*

[LBR11] A. LIUTKUS, R. BADEAU et G. RICHARD : Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155–3167, juillet 2011. *Cité page 41*

[LC83] Shu LIN et Daniel J. COSTELLO : *Error Control Coding : Fundamentals and Applications*. Prentice Hall, 1983. *Cité page 54*

[LD08] B. LACHAISE et L. DAUDET : Inverting dynamics compression with minimal side information. *In Proc. Digital Audio Effects Conf. (DAFx'08)*, pages 97–102, septembre 2008. *2 citations pages 67 et 152*

[Lep98] P. LEPAIN : *Recherche et applications en informatique musicale*, chapitre Ecoute interactive des documents musicaux numériques, pages 209–226. Hermes, Paris, France, 1998. *2 citations pages 1 et 153*

[Liu07] Yi-Wen LIU : Sound source segregation assisted by audio watermarking. *In Multimedia and Expo, 2007 IEEE International Conference on*, pages 200–203, juillet 2007. 10.1109/ICME.2007.4284621. *2 citations pages 67 et 136*

[Llo82] S. LLOYD : Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. *2 citations pages 63 et 86*

[LMR04a] M. LAGRANGE, S. MARCHAND et J.B. RAULT : Partial tracking based on future trajectories exploration. *In 116-th Convention of the Audio Engineering Society (AES)*, volume 4, mai 2004. *Cité page 29*

[LMR04b] M. LAGRANGE, S. MARCHAND et J.B. RAULT : Using linear prediction to enhance the tracking of partials. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'04)*, volume 4, pages 241–244, Montreal, Quebec, Canada, mai 2004. *4 citations pages 29, 92, 137 et 179*

[LNK87] M. LAHAT, R. NIEDERJOHN et D. KRUBSACK : A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 35(6):741–750, juin 1987. *Cité page 35*

[LPB+11] A. LIUKTUS, J. PINEL, R. BADEAU, L. GIRIN et G. RICHARD : Informed source separation through spectrogram coding and data embedding. *Signal Processing*, septembre 2011. *Cité page 67*

[LS99] Daniel D. LEE et H. Sebastian SEUNG : Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, octobre 1999. *2 citations pages 16 et 103*

[LS00] Daniel D. LEE et H. Sebastian SEUNG : Algorithms for non-negative matrix factorization. *In Proc. of Neural Information Processing Systems (NIPS)*, pages 556–562. MIT Press, novembre 2000. *Cité page 16*

[Mar98] Sylvain MARCHAND : Improving spectral analysis precision with an enhanced phase vocoder using signal derivatives. *In Proc. Digital Audio Effects Conf. (DAFx'98)*, pages 114–118, novembre 1998. *Cité page 27*

[Mar00] Sylvain MARCHAND : *Sound Models for Computer Music (analysis, transformation, synthesis)*. Thèse de doctorat, Université de Bordeaux, Talence, France, décembre 2000. *Cité page 16*

[Mar01] Farokh MARVASTI : *Nonuniform Sampling : Theory and Practice (Information Technology : Transmission, Processing and Storage)*. Springer, 1 édition, juin 2001. *Cité page 10*

[Mar04] M. MAROLT : A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, juin 2004. *Cité page 102*

[Mar08] Sylvain MARCHAND : *Habilitation à diriger des recherches : Avancées en modélisation spectrale du son musical*. Thèse de doctorat, Université de Bordeaux, décembre 2008. *2 citations pages 13 et 179*

[Mar12a] Sylvain MARCHAND : The simplest analysis method for non-stationary sinusoidal modeling. *In Proc. Digital Audio Effects Conf. (DAFx'12)*, pages 23–26, York, UK, septembre 2012. *Cité page 29*

[Mar12b] Benjamin MARTIN : *Analyse de structures répétitives dans les séquences musicales*. Thèse de doctorat, Université de Bordeaux, Talence, France, décembre 2012. *Cité page 102*

[MB10] S. MUSEVIC et J. BONADA : Comparison of non-stationary sinusoid estimation methods using reassignment and derivatives. *In Proc. Sound and Music Computing Conference (SMC'10)*, juillet 2010. *Cité page 77*

[McM56] B. MCMILLAN : Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116, 1956. *Cité page 56*

[McN84] Guy W. MCNALLY : Dynamic range control of digital audio signals. *Journal of the Audio Engineering Society*, 32(5):316–327, 1984. *Cité page 40*

[MD08] S. MARCHAND et P. DEPALLE : Generalization of the derivative analysis method to non-stationary sinusoidal modeling. *In Proc. Digital Audio Effects Conference (DAFx'08)*, pages 281–288, septembre 2008. *4 citations pages 29, 77, 106 et 107*

[MF10] S. MARCHAND et D. FOURER : Breaking the bounds : introducing informed spectral analysis. *In Proc. Digital Audio Effects Conf. (DAFx'10)*, pages 359–366, Graz, Autriche, septembre 2010. *3 citations pages 71, 92 et 152*

[MG83] B.C.J. MOORE et B.R. GLASBERG : Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753, 1983. *Cité page 17*

[MHM06] G. MEURISSE, P. HANNA et S. MARCHAND : A new analysis method for sinusoids+noise spectral models. *In Proc. Digital Audio Effects Conf. (DAFx'06)*, pages 139–144, septembre 2006. *2 citations pages 15 et 71*

[Moo86] Robert A. MOOG : MIDI : Musical instrument digital interface. *Journal of the Audio Engineering Society*, 34(5):394–404, 1986. *Cité page 2*

[MQ86] R. MCAULAY et T. QUATIERI : Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, août 1986. *4 citations pages 14, 23, 76 et 91*

[MSW83] M. E. MCINTYRE, R. T. SCHUMACHER et J. WOODHOUSE : On the oscillations of musical instruments. *Journal of the Acoustical Society of America*, 5(74):1325–1345, 1983. *Cité page 32*

[MV09] S. MARCHAND et A. VIALARD : The hough transform for binaural source localization. *In Proc. Digital Audio Effects Conf. (DAFx'09)*, pages 252–259, Como, Italie, septembre 2009. *Cité page 42*

[MZ93] S. G. MALLAT et Zhang ZHIFENG : Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(Issue 12):3397–3415, décembre 1993. *2 citations pages 35 et 104*

[Nol67] A. M. NOLL : Cepstrum pitch determination. *Journal of the Acoustical Society of America*, 41:293–309, 1967. *Cité page 34*

[PA02] L.C PARRA et C.V. ALVINO : Geometric source separation : merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, 2002. *Cité page 41*

[PC08] Sunho PARK et Seungjin CHOI : Gaussian processes for source separation. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'08)*, pages 1909–1912, Las Vegas, NV, USA, mars 2008. *Cité page 41*

[PE06] Graham E. POLINER et Daniel P. W. ELLIS : A discriminative model for polyphonic piano transcription. *Journal on Advances in Signal Processing (EURASIP JASP)*, octobre 2006. *Cité page 122*

[PG11] M. PARVAIX et L. GIRIN : Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 19(6):1721–1733, août 2011. *2 citations pages 67 et 135*

[PGB09] M. PARVAIX, L. GIRIN et J.-M. BROSSIER : A watermarking-based method for single-channel audio source separation. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'09)*, pages 101–104, Taipei, Taiwan, avril 2009. *2 citations pages 40 et 152*

[PGBP10] J. PINEL, L. GIRIN, C. BARAS et M. PARVAIX : A high-capacity watermarking technique for audio signals based on MDCT-domain quantization. *In Int. Congress on Acoustics*, octobre 2010. *4 citations pages 91, 93, 94 et 156*

[PHC06] Aliaksandr PARADZINETS, Hadi HARB et Liming CHEN : Use of continuous wavelet-like transform in automated music transcription. *In Proc. IEEE European Signal Processing Conference (EUSIPCO'06)*, septembre 2006. *2 citations pages 14 et 179*

[Pla50] R. L. PLACKETT : Some theorems in least squares. *Biometrika*, 37, 1950. *Cité page 52*

[PM00] H. PURNHAGEN et N. MEINE : HILN – the MPEG-4 parametric audio coding tools. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'00)*, pages 201–204, Juin 2000. *Cité page 23*

[PS00] T. PAINTER et A. SPANIAS : Perceptual coding of digital audio. *Proc. IEEE*, 88(4):451–515, avril 2000. *Cité page 93*

[Rao45] Radhakrishna C. RAO : Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945. *2 citations pages 31 et 53*

[RE10] S. RAVURI et D.P.W. ELLIS : Cover song detection : From high scores to general classification. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'10)*, pages 65–68, Dallas, TX, USA, 2010. *2 citations pages 102 et 153*

[RM07] M. RASPAUD et S. MARCHAND : Enhanced resampling for sinusoidal modeling parameters. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'07)*, pages 327–330, New Platz, NY, USA, octobre 2007. *Cité page 23*

[Roa96] Curtis ROADS : *The Computer Music Tutorial*. MIT Press, avril 1996. *Cité page 1*

[Roe02] Axel ROEBEL : Estimating partial frequency and frequency slope using reassignment operators. *In International Computer Music Conference*, pages 122–125, Göteborg, Suisse, septembre 2002. *Cité page 106*

[RRHO10] Thomas ROCHER, Matthias ROBINE, Pierre HANNA et Laurent OUDRE : Concurrent estimation of chords and keys from audio. *In Proc. of the International Society on Music Information Retrieval (ISMIR)*, pages 141–146, Utrecht, Pays-Bas, août 2010. *Cité page 101*

[RS60] I. S. REED et G. SOLOMON : Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 2(8):300–304, juin 1960. *Cité page 11*

[SA87] K.D. SAUER et J.P. ALLEBACH : Iterative reconstruction of bandlimited images from nonuniformly spaced samples. *IEEE Transactions on Circuits and Systems*, 34(12):1497–1506, 1987. *Cité page 9*

[Say06] K. SAYOOD : *Introduction to data compression*. Morgan Kaufmann, Elsevier, San Francisco, USA, third edition édition, 2006. *3 citations pages 59, 60 et 86*

[SB03] P. SMARAGDIS et J.C. BROWN : Non-negative matrix factorization for polyphonic music transcription. *In Proc. IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA'03)*, pages 177–180, New Platz, NY, USA, octobre 2003. *Cité page 110*

[SD13] N. STURMEL et L. DAUDET : Informed source separation using iterative reconstruction. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 21(1):178–185, 2013. *Cité page 67*

[Sha48] C. E. SHANNON : A mathematical theory of communication. *Bell System Technical Journal*, 7:379–423, juillet 1948. *2 citations pages 59 et 61*

[Sha49] C. E. SHANNON : Communications in the presence of noise. *In Proc. IRE*, volume 37, pages 10–21, janvier 1949. *Cité page 8*

[Sha59] C. E. SHANNON : Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 4:142–163, 1959. *Cité page 61*

[SLP+12] N. STURMEL, A. LIUTKUS, J. PINEL, L. GIRIN, S. MARCHAND, G. RICHARD, R. BADEAU et L. DAUDET : Linear mixing models for active listening of music productions in realistic studio conditions. *In Proc. 132nd AES Convention*, page Paper number : 8594, Budapest, Hongrie, avril 2012. *Cité page 40*

[SN92] K. SAYOOD et S. NA : Recursively indexed quantization of memoryless sources. *IEEE Transactions on Information Theory*, 38(5):1602–1609, 1992. *Cité page 60*

[SOdBB03] E. SCHUIJERS, W. OOMEN, B. den BRINKER et J. BREEBAART : Advances in parametric coding for high quality audio. *In Audio Engineering Society 114th Convention*, Amsterdam, Pays-Bas, 2003. *5 citations pages 15, 23, 61, 76 et 153*

[SRMS06] P. SMARAGDIS, B. RAJ, V. MADHUSUDANA et S. SHASHANKA : Supervised and semi-supervised separation of sounds from single-channel mixtures. *In 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 414–421, juillet 2006. *2 citations pages 3 et 40*

[SS87] Julius SMITH et Xavier SERRA : PARSHL an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. *In International Computer Music Conference*, pages 290–297, Urbana, Illinois, USA, août 1987. *4 citations pages 14, 23, 25 et 76*

[SSWY01] F. SATTAR, M. Y. SIYAL, L. C. WEE et L. C. YEN : Blind source separation of audio signals using improved ica method. *In Proc. IEEE International Workshop on Statistical Signal Processing (SSP'01)*, pages 452–455, août 2001. *Cité page 42*

[Str95] Thomas STROHMER : *Irregular Sampling, Frames and Pseudoinverse*. Thèse de doctorat, Department of Mathematics, University of Vienna, Strudlhofgasse 4, A-1090 Vienne, Autriche, 1995. e-mail : strohmer@tyche.mat.univie.ac.at. *2 citations pages 9 et 10*

[Str97] Thomas STROHMER : Computationally attractive reconstruction of band-limited images from irregular samples. *IEEE Transactions on Image Processing*, 6(4):540–548, avril 1997. Digital Object Identifier 10.1109/83.563319. *Cité page 10*

[SVN37] S. S. SMITH, J. VOLKMAN et B. E. NEWMAN : A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937. *Cité page 17*

[SW73] D. SLEPIAN et J.K. WOLF : Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973. *Cité page 67*

[SWT04] Emery SCHUBERT, Joe WOLFE et Alex TARNOPOLSKY : Spectral centroid and timbre in complex, multiple instrumental textures. *In Proc. of the 8th International Conference on Music Perception & Cognition (ICMPC)*, Evanston, août 2004. *2 citations pages 20 et 112*

[Tal95] D. TALKIN : *Speech coding and synthesis*, chapitre A Robust Algorithm for Pitch Tracking (RAPT). Elsevier, New York, USA, 1995. *Cité page 33*

[TC83] G. TORELLI et G. CAIRONI : New polyphonic sound generator chip with integrated microprocessor-programmable adsr envelope shaper. *IEEE Transactions on Consumer Electronics*, CE-29(3):203–212, 1983. *Cité page 17*

[TNS04] H. TAKEDA, T. NISHIMOTO et S. SAGAYAMA : Rhythm and tempo recognition of music performance from a probabilistic approach. *In Proc. of the International Conference of Music Information Retrieval (ISMIR)*, Barcelona, Espagne, octobre 2004. *Cité page 21*

[VB88] B.D. Van VEEN et K.M. BUCKLEY : Beamforming : a versatile approach to spatial filtering. *ASSP Magazine, IEEE*, 5(2):4–24, 1988. *Cité page 41*

[VE02] Harald VISTE et Gianpaolo EVANGELISTA : On the use of spatial cues to improve binaural source separation. *In Proc. Digital Audio Effects Conf. (DAFx'02)*, pages 209–213, 2002. *Cité page 42*

[Ver10] S. VERDU : Mismatched estimation and relative entropy. *IEEE Transactions on Information Theory*, 56(8):3712–3720, 2010. *3 citations pages 50, 65 et 153*

[VGF06] E. VINCENT, R. GRIBONVAL et C. FÉVOTTE : Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(4):1462–1469, juillet 2006. *Cité page 43*

[Vir06] Tuomas VIRTANEN : Unsupervised learning methods for source separation in monaural music signals. *In* KLAPURI et DAVY [KD06], pages 267–296. *Cité page 40*

[VM00] T.S. VERMA et T.H.Y. MENG : A 6kbps to 85kbps scalable audio coder. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'00)*, volume 2, pages 877–880, Istanbul, Turquie, juin 2000. *Cité page 76*

[Whi28] J. M. WHITTAKER : On the function which are. *In Proc. of the Edinburgh Mathematical Society (Series 2)*, volume 1, pages 169–176, juillet 1928. *Cité page 8*

[Whi35] J. M. WHITTAKER : *Interpolatory function theory*, volume 33. The University Press, 1935. *Cité page 8*

[Wie49] N. WIENER : *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, New York, USA, 1949. *Cité page 41*

[Wol09] Matthias WOLFEL : Signal adaptive spectral envelope estimation for robust speech recognition. *Speech Communication*, 51(6):551–561, 2009. *Cité page 20*

[Wyn75] A.D. WYNER : On source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 21(3):294–300, 1975. *2 citations pages 67 et 153*

[WZ76] A.D. WYNER et J. ZIV : The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976. *Cité page 67*

[Yeh08] Chunghsin YEH : *Multiple fundamental frequency estimation of polyphonic recordings*. Thèse de doctorat, Ph.D. in Computer Science, Université Pierre et Marie Curie - Paris 6, juin 2008. *3 citations pages 35, 103 et 115*

[YR04] O. YILMAZ et S. RICKARD : Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004. *Cité page 102*

[YR06] C. YEH et A. ROEBEL : Adaptive noise level estimation. *In Proc. Digital Audio Effects Conf. (DAFx'06)*, septembre 2006. *2 citations pages 15 et 76*

[YR09] C. YEH et A. ROEBEL : The expected amplitude of overlapping partials of harmonic sounds. *In Proc. IEEE International Conference on Acoust., Speech and Signal Process. (ICASSP'09)*, pages 3169–3172, 2009. *Cité page 28*

[YR11] Chunghsin YEH et Axel ROEBEL : Multiple-f0 estimation for MIREX 2011. *In Proc. of the International Conference of Music Information Retrieval (ISMIR)*, Miami, FL, USA, 2011. *Cité page 102*

[YRR10] C. YEH, A. ROEBEL et X. RODET : Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 18(6):1116–1126, 2010. *7 citations pages 101, 102, 103, 110, 112, 113 et 114*

[Zem00] Gilles ZEMOR : *Cours de cryptographie*. Cassini, 2000. *Cité page 54*

[ZGS96] G. ZHOU, G. GIANNAKIS et A. SWAMI : On polynomial phase signal with time-varying amplitudes. *IEEE Transactions on Signal Processing*, 44(4):848–860, avril 1996. *Cité page 78*

[ZL77] J. ZIV et A. LEMPEL : A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977. *Cité page 60*

[Zwi61] E. ZWICKER : Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, 33(issue 2):248, 1961. *Cité page 17*

TABLE DES FIGURES

1 Chaîne de traitement du signal audio permettant d'obtenir un signal de sortie y à partir d'une entrée x en passant par un modèle $\mathcal{M}$. 2

1.1 Signal sinusoïdal modulé linéairement en fréquence et échantillonné aléatoirement 1.1a) et son interpolation de Voronoï 1.1b). 10
1.2 Reconstruction d'un signal sinusoïdal modulé linéairement en fréquence échantillonné aléatoirement en utilisant l'algorithme de Voronoï-Allebach. 10
1.3 Signal temporel et spectrogramme correspondant au son émis par un piano jouant la gamme de Do majeur montante puis descendante. 12
1.4 Exemples de fenêtre d'analyse spectrale et spectre d'amplitude (en dB) correspondant [Mar08]. 13
1.5 Comparaison de la résolution temps-fréquence entre la transformée de Fourier 1.5a) et la transformée en ondelettes 1.5b) (source [PHC06]). 14
1.6 Schéma décrivant l'évolution temporelle de l'enveloppe d'un signal musical. 17
1.7 Spectre d'amplitude d'un son de saxophone jouant la note Do$_\sharp$2 ($F_0 \approx$ 138.6Hz). Nous voyons clairement les différents partiels (pics) ainsi que l'enveloppe spectrale correspondante (en rouge) construite par interpolation spline. 20

2.1 Application de l'algorithme du triangle [KZ01] pour l'estimation des paramètres sinusoïdaux. 26
2.2 Combinaison de 2 sinusoïdes d'amplitude $a_1 = a_2 = 1$ en fonction de leur différence de phase $\Delta_\phi \in [-\pi, \pi]$. 28
2.3 Estimation des partiels d'un son de saxophone jouant avec un vibrato. La trajectoire est estimée par prédiction linéaire utilisant un modèle autorégressif (AR) pour la fréquence et l'amplitude [LMR04b]. 29
2.4 Spectre d'amplitude d'un son naturel de saxophone ténor et d'un son naturel de piano superposés et jouant la même note Do3. Le premier partiel correspond à la fréquence $F_0 \approx 261,62$ Hz. Les autres partiels sont des multiples entiers (ou presque) de F_0. On remarque que les partiels du pianos ne sont plus confondus avec ceux du saxophone dans les hautes fréquences en raison de l'inharmonicité. 33

2.5 Illustration de la séparation de sources en traitement du signal audio à partir d'un signal de mélange. Sa principale application en musique est le démixage qui consiste à retrouver le signal isolé de chacun des instruments qui composent un mélange. 38

3.1 Système de communication considéré dans le cadre de notre étude dans lequel on souhaite transmettre un signal de mélange x et un vecteur de paramètres θ. 50
3.2 Valeur de l'entropie en fonction de la probabilité d'un message m. Plus la probabilité de recevoir un message m est importante, plus la quantité d'information qu'il transporte est faible. Cette figure est obtenue en utilisant l'équation (3.18) avec $b = 4$. 55
3.3 Représentation d'un arbre binaire complet correspondant à un code complet utilisant un alphabet q-aire pour $q = 2$. Dans cet exemple la profondeur de l'arbre vaut la taille maximale d'un mot $n_I = 3$. Il existe q^{n_I} mots différents pouvant être codés avec n_I symboles et correspondant chacun à un sommet terminal. 57
3.4 Arbre binaire ($q = 2$) du message **commencement** utilisant la convention gauche :0, droite :1 pour le codage des arêtes. 59
3.5 Exemple de représentation graphique d'une fonction de quantification d'un scalaire x. 62
3.6 Fonction débit-distorsion typique utilisée pour le codage d'une source X (quantification avec contrainte d'entropie). 64
3.7 Histogramme d'un signal temporel discret $s[n]$ normalisé composé d'une sinusoïde et d'un bruit gaussien, quantifié uniformément sur 100 valeurs 3.7a) et histogramme du signal de différence calculé à partir de son estimation en utilisant la méthode de réallocation (*cf.* section 2.1.1) 3.7b). On remarque que l'écart type de la distribution du signal $s[n] - \hat{s}[n]$ est plus faible et possède une entropie plus faible. Le signal $s[n] - \hat{s}[n]$ nécessitera un code de taille plus faible pour être représenté. 66
3.8 Système de boucle fermée utilisant un quantificateur Q, un prédicteur P et une fonction de codage $\mathcal{C}$ d'après [GG91]. 67
3.9 Schéma de l'approche informée pour l'estimation d'un paramètre p. . . . 69

4.1 Histogramme du MSB (bit de poids le plus fort) de la représentation binaire de la valeur absolue de l'erreur d'estimation utilisant la méthode de réallocation appliquée au signal mélangé avec un bruit blanc additif. Le premier indice non nul augmente en même temps que le SNR, le nombre de bits à corriger pour retrouver le paramètre de référence est donc plus faible lorsque le SNR augmente. 81
4.2 Variance de l'erreur obtenue en fonction du SNR d'un signal composé d'une seule sinusoïde mélangée dans du bruit pour l'estimation de la fréquence 4.2a), de l'amplitude 4.2b) et de la phase 4.2c). La variance obtenue reste toujours supérieure à la borne de Cramér-Rao sauf dans le cas informé où on informe 5 bits pour chaque paramètre quantifié sur 16 bits. 82
4.3 Variance de l'erreur obtenue en fonction du SNR d'un signal composé d'une seule sinusoïde mélangée dans du bruit pour l'estimation de la fréquence 4.3a), de l'amplitude 4.3b) et de la phase 4.3c). La variance obtenue reste toujours supérieure à la borne de Cramér-Rao sauf dans le cas informé où on informe 15 bits pour chaque paramètre quantifié. 83

4.4 Comparaison du SNR moyen obtenu 4.4a) entre un estimateur classique et le même estimateur informé. 4.4b) montre la répartition du débit entre les paramètres sinusoïdaux. Ici le SNR cible est fixé à 45dB. 89
4.5 Comparaison du SNR moyen obtenu 4.5a) entre un estimateur classique et le même estimateur informé. 4.5b) montre la répartition du débit entre les paramètres sinusoïdaux. Ici le SNR cible est fixé à 100dB. 90
4.6 Structure du système proposé pour l'estimation des paramètres sinusoïdaux de chaque source séparée à partir d'un mélange mono-canal. 91
4.7 Comparaison entre le débit global (toutes sources confondues) utilisé par la quantification ECUSQ et l'analyse informée. Il s'agit du débit total nécessaire pour estimer toutes les sources à partir de l'observation du mélange. 97
4.8 Nombre de composantes d'amplitude non nulle pour chaque source en fonction de l'entropie cible H_t. Ce nombre augmente en fonction du SNR cible et dépend de la quantification calculée par la méthode ECUSQ. 98
4.9 Comparaison du débit obtenu par l'estimation classique (cercle rouge), par le codage pur (utilisant la quantification ECUSQ) et par l'analyse informée en fonction du SNR. Le débit augmente avec la qualité obtenue mais reste plus faible pour l'analyse informée pour des valeurs de SNR raisonnables. Pour les SNR très élevés, le codage seul semble plus efficace. 99

5.1 Schéma descriptif de la méthode d'estimation F_0 multiple proposée permettant d'estimer un ensemble $\hat{\Omega}[n]$ de candidats F_0 actifs à un instant n à partir d'un signal de mélange discret $x[n]$. 104
5.2 Extraction des composantes tonales (*peak*) à partir d'un son polyphonique en utilisant respectivement le seuillage automatique proposé 5.2a) et un seuillage utilisant la médiane mobile 5.2b). 105
5.3 Construction des hypothèses de source harmonique par sélection des pics observés depuis le spectre d'amplitude. 108
5.4 Comparaison des performances moyennes d'estimation F_0 multiple entre les deux méthodes proposées pour la construction des HPS et la technique de transcription polyphonique **BBV-10** [BBV10] sur la base de données de test. 109
5.5 Valeur de l'écart d_h pour les partiels d'un son de saxophone. 111
5.6 Valeur de la l'écart d_h pour un son sinusoïdal harmonique composé de 30 partiels d'amplitude égale. 111
5.7 Histogramme des t_h en fonction des combinaisons entre deux sources harmoniques. Les figures présentent le nombre d'occurrences pour chaque valeur du temps t_h observé mesuré en secondes et (Sons de saxophone jouant respectivement les notes Ré$_\flat$ et Mi). 114
5.8 Comparaison des scores *Precision*, *Accuracy* et *Recall* obtenus pour l'estimation des F_0 en utilisant les méthodes proposées. Les résultats sont triés par ordre décroissant du score *Accuracy*. 117
5.9 Comparaison du taux d'erreur d'édition (substitution, oubli et insertion) mesuré par les méthodes d'estimation F_0. Les résultats sont triés par ordre croissant de l'erreur totale. 118
5.10 Comparaison des scores *Accuracy*, *Recall* et *F-Measure* obtenus pour l'estimation F_0 multiple en utilisant les méthodes proposées. Les résultats sont triés par ordre décroissant de *F-Measure*. 119

5.11 Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **FM** proposée. Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite. 123
5.12 Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **BBV-10** [BBV10].Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite. 124
5.13 Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **KL-03** [Kla03]. Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite. 125
5.14 Histogramme des erreurs d'estimation F_0 mesurées en utilisant la méthode **YIN** [dCK02].Les erreurs sont classées par type (bien estimé, erreur d'octave, erreur de substitution, candidats manquants ou insérés). Les erreurs d'octave et de substitutions sont détaillées par la suite. 126
5.15 Schéma du système de codage utilisé pour calculer $\mathcal{I}^{(t)}$ nécessaire pour obtenir $\tilde{\Omega}^{(t)} = \Omega^{(t)}$ à partir de l'estimation $\hat{\Omega}^{(t)}$. La réponse à chaque test "*si* ..." est codée par un seul bit (1 ou 0) excepté pour l'étape finale de mise à jour de $\mathcal{I}^{(t)}$ plus coûteuse. 131
5.16 Schéma de l'algorithme utilisé pour calculer le code qui permet au décodeur de retrouver $\tilde{\Pi}_k^{(t)}$ à partir de $\tilde{\Omega}^{(t)}$. La réponse à chaque test "*si* ..." est codée par un seul bit (1 ou 0) qui s'ajoute au débit initial permettant de calculer $\tilde{\Omega}^{(t)}$. 131
5.17 Comparaison du nombre de bits utilisés pour chaque système de codage pour informer les méthodes d'estimation F_0 appliquées sur la base de données d'évaluation. Les résultats sont triés par ordre croissant de bits utilisés pour la méthode M2. 133
5.18 Comparaison du gain en pourcentage de nombre de bits apporté par l'utilisation d'un estimateur. Les résultats sont triés par ordre décroissant du pourcentage de gain sur M2. 134
5.19 Schéma d'un système de séparation de sources informée utilisant une configuration codeur / décodeur. 135
5.20 Score *Accuracy*, *Precision* et *Recall* obtenu pour chaque technique d'estimation F_0 multiple. 140
5.21 Classification des erreurs pour chaque technique d'estimation F_0 multiple. 141
5.22 Score *Precision*, *Recall* et *F-Measure* calculés pour chaque technique d'estimation F_0 multiple. 142
5.23 Quantité d'information utilisée pour calculer la transcription de référence à partir de l'estimation calculé pour chacune des techniques d'estimation F_0 multiple. 143
5.24 Gain apporté par l'estimation F_0 multiple sur la quantité d'information codée permettant d'obtenir la transcription de référence. 144
5.25 Qualité de séparation mesurée pour l'extrait 1. 145
5.26 Qualité de séparation mesurée pour l'extrait 2. 146
5.27 Qualité de séparation mesurée pour l'extrait 3. 147

A.1 Exemple de quantification QIM codant un message c sur 2 bits avec 4 quantificateurs. 156

A.2 Comparatif des distorsions théoriques (SLB et ECUSQ) et pratiques (codage pur ECUSQ, estimation aveugle et estimation informée).L'estimation informée combine un estimateur avec le même débit que pour le codage pur. On remarque que combiner un estimateur avec de l'information permet d'obtenir une distorsion plus faible que la distorsion théorique de ECUSQ et de la borne de Shannon. 161
A.3 Comparaison entre la borne de Shannon et la borne ECUSQ théorique. Comme attendu, la distorsion moyenne théorique de la méthode ECUSQ est supérieure à la distorsion Shannon. La différence moyenne mesurée est d'environ 2.7bits. 162

LISTE DES TABLEAUX

1.1 Différentes représentations de la gamme chromatique du Do3 au Do4 sur une échelle basée sur le tempérament égal. Le système du tempérament égal subdivise une octave en 12 demi-tons égaux. Ce système est celui utilisé dans la musique occidentale par quasiment tous les instruments chromatiques actuels (guitare, piano, etc.). Le calcul des fréquences s'effectue par rapport au La3 (440 Hz), ainsi la fréquence du Do3 située 9 demi-tons en dessous est obtenue par le calcul $440 \times 2^{-9/12} \approx 261.62$Hz et celle du Do4 situé 3 demi-tons au-dessus par $440 \times 2^{+3/12} \approx 523.25$Hz. 19

2.1 Définition des mesures utilisées pour l'évaluation d'un système de transcription automatique. 36

4.1 Complexité calculatoire en unité de temps pour la technique d'estimation des paramètre sinusoïdaux de sources isolées à partir d'un mélange monocanal 95

5.1 Description des extraits sonores utilisés pour l'évaluation du système de séparation de sources informée par l'estimation F_0 multiple. 138

5.2 Nombre de bits utilisés en fonction du codage permettant d'obtenir la transcription exacte. Lorsque le codage est combiné avec un estimateur, celui-ci est appliqué sur le signal de mélange tatoué. Le meilleur estimateur considéré est celui qui permet de minimiser le nombre de bits utilisés. 139

Approche informée pour l'analyse du son et de la musique

Résumé :

En traitement du signal audio, l'analyse est une étape essentielle permettant de comprendre et d'interagir avec les signaux existants. En effet, la qualité des signaux obtenus par transformation ou par synthèse des paramètres estimés dépend de la précision des estimateurs utilisés. Cependant, des limitations théoriques existent et démontrent que la qualité maximale pouvant être atteinte avec une approche classique peut s'avérer insuffisante dans les applications les plus exigeantes (e.g. écoute active de la musique). Le travail présenté dans cette thèse revisite certains problèmes d'analyse usuels tels que l'analyse spectrale, la transcription automatique et la séparation de sources en utilisant une approche dite "informée". Cette nouvelle approche exploite la configuration des studios de musique actuels qui maîtrisent la chaîne de traitement avant l'étape de création du mélange. Dans les solutions proposées, de l'information complémentaire minimale calculée est transmise en même temps que le signal de mélange afin de permettre certaines transformations sur celui-ci tout en garantissant le niveau de qualité. Lorsqu'une compatibilité avec les formats audio existants est nécessaire, cette information est cachée à l'intérieur du mélange lui-même de manière inaudible grâce au tatouage audionumérique. Ce travail de thèse présente de nombreux aspects théoriques et pratiques dans lesquels nous montrons que la combinaison d'un estimateur avec de l'information complémentaire permet d'améliorer les performances des approches usuelles telles que l'estimation non informée ou le codage pur.

Mots-clés : analyse spectrale informée, modèle sinusoïdal, estimation, codage audio, séparation de sources, transcription automatique.

Informed approach for sound and music analysis

Abstract:

In the field of audio signal processing, analysis is an essential step which allows interactions with existing signals. In fact, the quality of transformed or synthesized audio signals depends on the accuracy over the estimated model parameters. However, theoretical limits exist and show that the best accuracy which can be reached by a classic estimator can be insufficient for the most demanding applications (e.g. active listening of music). The work which is developed in this thesis revisits well known audio analysis problems like spectral analysis, automatic transcription of music and audio source separation using the novel "informed" approach. This approach takes advantage of a specific configuration where the parameters of the elementary signals which compose a mixture are known before the mixing process. Using the tools which are proposed in this thesis, the minimal side information is computed and transmitted with the mixture signal. This allows any kind of transformation of the mixture signal with a constraint over the resulting quality. When the compatibility with existing audio formats is required, the side information is embedded directly into the analyzed audio signal using a watermarking technique. This work describes several theoretical and practical aspects of audio signal processing. We show that a classic estimator combined with the sufficient side information can obtain better performance than classic approaches (classic estimation or pure coding).

Keywords: informed spectral analysis, sinusoidal modeling, estimation, audio coding, source separation, automatic transcription.

Université de Bordeaux 1
LaBRI
UMR 5800
351, cours de la Libération
F-33400 TALENCE
FRANCE