Multipitch estimation methods for the separation of speech and musical signals

Julie Rosier

▶ **To cite this version:**

Julie Rosier. Multipitch estimation methods for the separation of speech and musical signals. domain_other. Télécom ParisTech, 2003. English. NNT : . pastel-00000723

HAL Id: pastel-00000723
https://pastel.hal.science/pastel-00000723

Submitted on 24 Sep 2004

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

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

Thèse

présentée pour obtenir le grade de docteur

de l'Ecole nationale supérieure
des télécommunications

Spécialité : Traitement du signal

Julie Rosier

Méthodes d'estimation de fréquences fondamentales multiples pour la séparation de signaux de parole et de musique

soutenue le 17 décembre 2003 devant le jury composé de

Xavier Rodet	Président
Frédéric Bimbot	Rapporteurs
Dirk Slock	
Gaël Richard	Examinateur
Yves Grenier	Directeur de thèse

Ecole nationale supérieure des télécommunications

Remerciements

Je tiens à remercier en premier lieu Yves Grenier qui m'a permis de réaliser cette thèse. Je le remercie pour son encadrement tout au long de ces années et pour la confiance qu'il m'a accordé pour mener à termes ce projet.
Je remercie également les membres du jury, pour l'intérêt qu'ils ont porté à mon travail et tout particulièrement Gaël Richard, pour son aide précieuse et ses remarques constructives.
Enfin, je remercie tous ceux qui ont participé de près et de loin : Slim et Eduardo sans qui, plusieurs chapitres de ce manuscrit n'auraient pas vu le jour, Thomas sans qui, bien des figures auraient disparues à l'impression, mais également le coeur de TSI " Raphaël, Cléo, Chloé et les Cloportes, Emmanuèle ..." et ses poumons " Catherine Vazza, Patricia Friedrich..." pour leur amitié et leur soutient.
Le reste de mes remerciements vont à ma mère, pour son amour et sa patience, à ma soeur pour son soutient et son écoute, et bien évidemment à Assane pour ce qu'il est simplement.

Resume

L'objet de cette thèse est l'étude du problème d'estimation de fréquences fondamentales multiples. Nous proposons de nouvelles méthodes d'estimation de fréquences fondamentales applicables à des mélanges de parole et des mélanges musicaux dont le nombre de sources est inconnu. Dans le cadre de la parole, nous développons tout d'abord une méthode robuste qui permet d'estimer la fréquence fondamentale d'une source pour des mélanges complexes contenant jusqu'à cinq sources. Afin de caractériser la nature voisée/non-voisée des mélanges de signaux de parole, ces derniers sont paramétrés à l'aide d'un modèle mixte du type "somme de sinusoïdes harmoniques plus bruit autorégressif". L'estimation de la fréquence fondamentale est basée sur le principe du critère d'ordre qui consiste à maximiser un terme de Vraisemblance, pénalisé par une quantité qui dépend du nombre de paramètres du modèle. Cet estimateur est ensuite appliqué itérativement pour déterminer les différentes fréquences fondamentales du mélange. Cette nouvelle méthode d'estimation à l'avantage, contrairement aux méthodes itératives, de n'utiliser aucune étape de soustraction, généralement le point faible de ce type de méthode. Elle permet également de définir un détecteur de voisement permettant de stopper les itérations et d'estimer ainsi le nombre de sources. Dans le cadre de mélanges musicaux, nous proposons trois nouvelles méthodes, qui contrairement à la méthode utilisée pour la parole, déterminent conjointement les fréquences fondamentales du mélange. Basées sur une classification des pics spectraux du mélange, elles consistent à déterminer les classes représentatives des spectres de chaque note. Les trois méthodes diffèrent par la classification utilisée pour obtenir ces spectres séparés. Toutes permettent d'estimer le nombre de sources du mélange. La prise en compte des recouvrements spectraux entre notes, caractéristiques des mélanges musicaux, les rende applicable au traitement d'accords pour des instruments polyphoniques, tels que les accords de piano. Dans ce travail, les performances de ces nouvelles méthodes d'estimation sont évaluées sur de nombreux mélanges synthétiques et réels.

Abstract

This thesis deals with the multipitch estimation problem. We propose new estimation methods for speech and musical mixtures where the number of sources is unknown. In the speech context, we first propose a robust method which allows to estimate the fundamental frequency of one source from mixtures holding five harmonic sources. The complex "unvoiced/voiced" nature of speech mixtures is characterized using a two-part model where the first part consists in sums of harmonic sinusoids and the second part is an autoregressive process. The pitch estimation is based on the maximization of a penalized likelihood term. The penalization is a quantity which depends on the number of parameters used for modelling. This estimator is iteratively applied in order to determine all the pitches. This new estimation method does not use any subtraction step which was the weak point of iterative methods. It also allows to define a voicing detector used to stop the iterations and to estimate the number of sources. In the musical context, we propose three new methods, which simultaneously estimate the pitches. Based on a spectral peak classification of the mixture, they consist in determining the classes corresponding to the spectrum of each note. The three methods differ from the classification used to construct these separate spectra. All the methods allow to estimate the number of sources of the mixture. Moreover they are able to take into account the spectral overlaps between note spectra which characterize musical mixtures and can be used to process on musical chords of the polyphonic instruments like piano chords. The perfomance of these new multipitch estimation methods is evaluated using various synthetic and real mixtures.

Table des matières

Remerciements **i**

Resume **i**

Abstract **iii**

Table des matières **v**

Introduction **1**

I Etat de l'art 3

1 Méthodes d'estimation et techniques pour la séparation 5
1.1 Description générale du problème de séparation 6
1.2 Cas monophonique : Séparation de sources harmoniques 7
1.2.1 Le caractère harmonique des signaux de parole et de musique 7
1.2.2 Les étapes de la séparation monophonique 11
1.3 L'estimation de fréquences fondamentales multiples 14
1.3.1 Les méthodes d'estimation spectrale 15
1.3.2 Les méthodes d'estimation temporelles 20
1.3.3 La modélisation sinusoïdale 24
1.3.4 Introduction de modèles auditifs 28
1.4 Des techniques pour améliorer les estimations 32
1.4.1 Les techniques utilisées pour pallier les problèmes de résolution 33
1.4.2 Poursuites des paramètres harmoniques le long des trames 37
1.5 Des techniques pour la synthèse des sources 40
1.5.1 Méthodes de synthèse 41
1.5.2 Techniques de séparation 43
1.6 Objectifs et description des méthodes d'estimations proposées . . 46
1.6.1 Le cadre d'étude et les objectifs fixés 46
1.6.2 Description des méthodes d'estimation proposées 46

II Estimation de fréquences fondamentales pour la parole 47

2 Estimation par le principe du critère d'ordre 49
2.1 Modélisation du signal de parole 50
2.1.1 Le modèle CARD 50
2.1.2 Estimation des paramètres du modèle CARD par la méthode du Maximum de Vraisemblance 50
2.1.3 Le modèle CARD harmonique - Ambiguïté d'octave 53
2.2 Estimation de la fréquence fondamentale 55
2.2.1 Principe du MDL 55
2.2.2 Hypothèse asymptotique - réduction de la complexité . . . 59
2.3 Etude de la méthode d'estimation 62
2.3.1 Résumé de la méthode d'estimation 62
2.3.2 Simulations - mise en oeuvre 62
2.3.3 Le principe MDL - résolution des ambiguïtés d'octave . . . 65
2.3.4 Performance de l'estimateur LS 67
2.3.5 Réduction de la complexité 72
2.3.6 Choix d'un critère de sélection 72
2.4 Conclusion 77

3 Estimation pour la séparation de signaux de parole 79
3.1 Schéma général de la méthode de séparation 80
3.2 82
3.3 Ecart de vraisemblance Δ - Mesure de voisement 84
3.3.1 Définition de l'écart de vraisemblance 85
3.3.2 Mesure de voisement et poursuite des fréquences 85
3.3.3 Ecart de vraisemblance pour les signaux synthétiques . . . 87
3.4 Application à la séparation de signaux réels 91
3.4.1 Mesure de voisement 92
3.4.2 Etude comparative 94
3.4.3 Application à des mélanges synthétiques voisés 96
3.4.4 Application à des mélanges "voisé/non-voisé" 105
3.4.5 Application à la séparation de deux locuteurs 111
3.5 Conclusion 115

III Estimation de fréquences fondamentales pour la musique 117

4 Estimation par classification des pics spectraux 119
4.1 Présentation générale 121
4.2 Analyse spectrale pour la séparation 123

4.2.1 Représentation spectrale 123
4.2.2 Détection des pics spectraux 124
4.2.3 Estimation des fréquences fondamentales 124
4.2.4 Résolution spectrale - conséquences 125
4.3 Classification des pics spectraux 127
4.3.1 Le principe des formes fortes 128
4.3.2 Algorithme des Nuées Dynamiques 129
4.3.3 Distances pour la classification de pics spectraux 132
4.3.4 Evaluation des performances des Nuées Dynamiques 138
4.4 Estimation par pénalisation sur la classe 145
4.4.1 Construction des formes fortes 145
4.4.2 Etude de la pénalisation sur la classe 147
4.4.3 Résumé des performances 159
4.5 Estimation par pénalisation sur l'harmonicité 161
4.5.1 Définition de l'harmonicité de deux fréquences 162
4.5.2 Construction des formes fortes 163
4.5.3 Etude de la pénalisation sur l'harmonicité 164
4.5.4 Résumé des performances 168
4.6 Estimation par pénalisation sur la classe et l'harmonicité 169
4.6.1 Construction des formes fortes 169
4.6.2 Etude de la pénalisation sur la classe et l'harmonicité . . . 170
4.6.3 Résumé des performances 177
4.7 Applications aux signaux musicaux 177
4.7.1 Etude de quelques mélanges particuliers 178
4.7.2 Recouvrement d'harmoniques : cas des accords 183
4.7.3 Estimation sur une base de données de signaux musicaux . 186
4.8 Conclusions 191

Conclusions et perspectives **193**

AnnexeA **197**

AnnexeB **204**

Bibliographie **209**

Introduction

Nous nous intéressons dans ce travail de thèse, au problème de l'estimation de fréquences fondamentales multiples pour des mélanges de signaux de parole ou de signaux musicaux. Les mélanges étudiés proviennent exclusivement d'enregistrements monophoniques, qui ne donnent aucune information spatiale sur les sources et pour lesquels le contenu fréquentiel reste l'unique information exploitable.

A l'instar de l'oreille humaine qui est capable d'associer une hauteur tonale à des sons émis simultanément, l'estimation vise à caractériser les signaux d'un mélange à l'aide de leur fréquence fondamentale. Ce paramètre, défini comme l'équivalent mathématique de la hauteur tonale, caractérise la nature harmonique de la parole et de nombreux signaux musicaux.

L'estimation de fréquences fondamentales multiples est au coeur de nombreuses applications. Etroitement liées à des problèmes de séparation, on la retrouve dans l'identification de notes de musique pour la transcription de partitions ou encore la séparation de locuteurs. Toutefois, elle suscite de plus en plus d'intérêt dans des domaines connexes, qui se limitent encore à ce jour au traitement de sources isolées, comme celui de la reconnaissance vocale ou en encore de l'indexation musicale.

Une classification simple des méthodes d'estimation de fréquences fondamentales multiples, consiste à distinguer les méthodes d'estimation dites *itératives* des méthodes d'estimation dites *conjointes*. Les premières ont l'avantage d'estimer le nombre de sources, mais leurs performances sont limitées par l'extraction successive de ces sources, qui dégrade au fur et à mesure le mélange. Les secondes méthodes, quant à elles, permettent l'estimation conjointe des fréquences fondamentales, mais nécessitent la connaissance du nombre de sources.

Nous proposons dans ce travail de nouvelles méthodes d'estimation de fréquences fondamentales multiples, qui s'inscrivent dans cette classification selon la nature du mélange analysé. Ainsi dans le cadre de la parole, nous développons une méthode d'estimation itérative, tandis que nous ferons appel à des méthodes d'estimation conjointes pour traiter des mélanges musicaux. Les principales contributions techniques de cette thèse sont alors les suivantes :
Dans le cadre de la parole, la méthode d'estimation itérative proposée est construite sur une modélisation temporelle du mélange qui tient compte de la présence simultanée de sources voisées et non-voisées. Cette modélisation permet de pallier le problème d'extraction successive des sources. Elle est également à la base d'un détecteur de voisement, utilisé pour estimer le nombre de sources du mélange.
Dans le cadre de la musique, les trois méthodes proposées estiment simultanément les fréquences fondamentales du mélange, évitant ainsi d'introduire toute extraction de sources. Basées sur une classification des pics spectraux du mélange, elles tiennent compte des recouvrements éventuels entre les spectres de chaque note et permettent ainsi de traiter des mélanges complexes contenant des notes joués à la tierce, à la quinte tels que des accords de piano.

Le travail qui suit se scinde en trois parties indépendantes :

la première partie (chapitre 1) est un état de l'art présentant un ensemble non exhaustif des méthodes d'estimation de fréquences fondamentales multiples existantes à ce jour, ainsi que des techniques développées pour améliorer ces dernières.

La seconde partie est consacrée à l'estimation fréquences fondamentales multiples pour la parole : le chapitre 2 présente un estimateur de fréquences fondamentales basé sur l'utilisation d'un critère de sélection d'ordre, capable d'estimer la fréquence fondamentale de la source prédominante pour des mélanges contenant plusieurs sources. Une méthode d'estimation fréquences fondamentales multiples basée sur cet estimateur est proposée et étudiée dans le chapitre 3.

La troisième partie (chapitre 4) traite le cas des signaux musicaux. Elle présente trois méthodes d'estimation de fréquences fondamentales multiples toutes basées sur une classification des pics spectraux du mélange.

Les performances de l'ensemble des méthodes d'estimations, pour la parole et la musique, sont mesurées en terme de pourcentage de fréquences fondamentales correctement, ainsi qu'en nombre estimé moyen de sources du mélange. Le cas de mélanges synthétiques et de mélanges de signaux réels sont traités.

Première partie

Etat de l'art

Chapitre 1

Méthodes d'estimation de fréquences fondamentales et techniques pour la séparation de signaux harmoniques

La séparation de signaux audionumériques est un vaste domaine de recherche qui recoupe entre autres, celui de l'Analyse/Synthèse, de la Reconnaissance de locuteurs, de l'Analyse de scènes auditives ou encore de l'Indexation musicale.
Ce travail de thèse s'intéresse plus particulièrement à l'aspect "analyse" de la séparation de sources, et traite le problème de l'estimation de fréquences fondamentales multiples.
Ce premier chapitre propose un état de l'art des méthodes d'estimation de fréquences fondamentales multiples, mais également un ensemble de techniques développées pour la séparation de signaux de parole et de signaux musicaux. Une description générale de la problématique sur la séparation de sources est tout d'abord présentée. Par la suite, nous nous intéresserons plus particulièrement au cas de mélanges de sources harmoniques provenant d'enregistrements monophoniques.
Une partie importante de ce chapitre sera consacrée à la présentation de méthodes d'estimation de fréquences fondamentales multiples, développées dans le cadre de la séparation de signaux de parole et de signaux musicaux.
Face aux divers problèmes rencontrés dans la mise en oeuvre de ces méthodes, en partie dus à la pluralité des fréquences fondamentales à estimer, plusieurs techniques ont été proposées pour améliorer les performances de ces estimateurs, dont les plus pertinantes seront également présentées.
Pour compléter cet état de l'art, une dernière partie sera consacrée à la présen-

tation de quelques techniques de séparation basées sur la synthèse des sources. Ces dernières ont un intérêt particulier, car elles sont à l'origine des algorithmes d'estimation de fréquences fondamentales multiples de type itératif.
Enfin, nous conclurons ce chapitre en définissant le cadre d'étude dans lequel s'inscrit ce travail de thèse, ainsi que ses objectifs.

1.1 Description générale du problème de séparation

Etant donné que la résolution du problème de séparation peut prendre des formes bien différentes, "la séparation de sources" apparaît comme une problématique qu'il est nécessaire de définir précisément, notamment concernant la nature des mélanges à traiter et les objectifs recherchés. Ainsi par exemple, le problème de séparation qui consiste à extraire d'un mélange une source sonore à remixer nécessite généralement une étape de synthèse sonore. Dans le cas de la transcription automatique de partitions, cette étape n'est pas utile et le problème de séparation se résume à l'extraction de paramètres pertinents relatifs aux différentes sources du mélange.
Les approches diffèrent également suivant le type d'enregistrements analysés : les enregistrements stéréophoniques ou multicapteurs mettent à disposition une information spatiale supplémentaire sur les sources, totalement absente dans une prise de son monophonique.
Pour la parole, ainsi que pour de nombreux signaux musicaux, la caractéristique la plus exploitable est leur nature quasi-harmonique. Cette dernière est d'autant plus importante, voire essentielle dans le cas d'enregistrements monophoniques, pour lesquels la principale information disponible reste la périodicité des signaux. C'est la raison pour laquelle la plupart des approches font appel aux méthodes d'estimation de fréquences fondamentales.
Etant donné que tout type d'enregistrements se compose d'un ou des plusieurs enregistrements monophoniques, on comprend facilement que toutes les techniques valables pour ce cas, le seront à plus grande échelle pour le cas multicapteurs. La conséquence directe est que l'estimation de fréquences fondamentales multiples fait partie intégrante du problème de séparation. Son importance est telle que les deux termes ont même parfois tendance à se substituer l'un à l'autre, comme par exemple dans le cas de la musique où la séparation de notes n'est autre que l'estimation des multiples fréquences fondamentales d'un morceau.

Bien que le domaine de l'Analyse/Synthèse a mis à disposition de nombreuses méthodes d'estimation pour une unique fréquence fondamentale, la généralisation au cas de plusieurs fréquences fondamentales s'avère bien plus complexe qu'une simple extension du formalisme "monopitch". Le problème de séparation est à l'origine de nombreuses méthodes d'estimation de fréquences fondamentales multiples, plus ou moins efficaces et dont les plus usitées sont présentées dans la suite.

L'objectif principal de cette thèse est de développer de nouvelles méthodes d'estimation de fréquences fondamentales multiples assez robustes pour être utilisées dans les problèmes de séparation de mélanges de parole et de musique.

1.2 Cas monophonique : Séparation de sources harmoniques

Le cadre d'étude dans lequel nous nous plaçons est celui de l'analyse de mélanges de sources harmoniques, pour des enregistrements de type monophonique. La première partie de ce paragraphe est consacrée aux signaux de parole et de musique, et plus particulièrement à leur caractère quasi-harmonique. Nous détaillerons par la suite, les différentes étapes rencontrées dans un problème de séparation pour des enregistrements monophoniques.

1.2.1 Le caractère harmonique des signaux de parole et de musique

Il existe une classification assez simple des sons audibles qui consiste à distinguer les sons dits “musicaux” et les bruits. Les premiers résultent de vibrations périodiques et, d'un point de vue perceptif, peuvent être caractérisés par trois qualités physiologiques : l'intensité, la hauteur et le timbre. Les bruits, quant à eux sont dûs à des vibrations non-périodiques. Ils englobent les bruits impulsionnels ou percussifs, les sons transitoires dont les paramètres caractéristiques varient rapidement, et le bruit de fond formé par la superposition d'un très grand nombre de sons.
Nous nous intéresserons ici plus particulièrement aux sons de parole et de certains instruments musicaux. Dans le domaine de la parole, la classification évoquée auparavant est plus généralement présentée en termes de “voisement” : les sons

périodiques sont dits voisés et vont résulter de la mise en vibration des cordes vocales sous l'effet de l'air expulsé par les poumons. Les bruits sont des sons non voisés qui ne font pas intervenir les cordes vocales et qui sont le fruit de turbulences dans la cavité orale ou de constrictions au niveau des dents et des lèvres. Cette classification "périodique/apériodique" est également transposable aux sons musicaux. Les instruments à cordes (piano, guitare...) et à vent (flûte, saxophone...) offrent ainsi un grand choix de sons périodiques, tandis que les percussions sont l'exemple même de sources non périodiques.

Afin d'identifier la nature d'une source sonore quelconque, notre système auditif utilise l'ensemble de ces caractéristiques sonores. La sensation de hauteur permettra, par exemple, de distinguer la voix d'un homme de celle d'une femme. Couplées à des informations spatiales "récoltées" par nos deux oreilles, ces caractéristiques vont également permettre au système auditif de séparer différents sons émis simultanément ou de se focaliser sur un son particulier dans une ambiance bruitée (effet "cocktail party").
A l'instar de l'oreille humaine, les méthodes de séparation présentées ici cherchent à utiliser des informations caractéristiques des signaux audionumériques. Les périodicités ne sont alors plus représentées en terme de hauteur tonale, mais sont mesurées sous forme de fréquences fondamentales du signal périodique, également appelée par abus de langage "pitch". Les outils de la séparation sont ceux du Traitement du Signal.
Le problème de la séparation se faisant dans un cadre monophonique, aucune information spatiale n'est disponible, contrairement au cas du système auditif qui écoute de ses deux oreilles, et les informations sont simplement de nature temporelle ou spectrale.
On rappelle ci-dessous quelques propriétés des signaux périodiques et leurs conséquences dans le cas plus particulier de la parole et de la musique.

- **La modélisation source-fitre des signaux périodiques**

Pour un signal périodique $s(t)$ de période T_0, observé sur un grand nombre de périodes, la représentation spectrale $S(f)$ est un ensemble de raies régulièrement espacées d'une valeur $f_0 = \frac{1}{T_0}$ et dont les amplitudes différent les unes des autres. Il est possible de décomposer cette représentation en deux éléments multiplicatifs telle que :

$$S(f) = E(f).H(f) \tag{1.1}$$

$E(f)$ est un spectre de raies plat où l'écart entre deux raies est égal à f_0. $H(f)$ est une enveloppe spectrale. Cette dernière est définie à chaque fréquence multiple de f_0 par l'amplitude de cette même fréquence dans le signal original. La fréquence f_0 est appelée fréquence fondamentale et ses multiples sont les harmoniques du signal périodique.
Dans le domaine temporel, on peut considérer le signal périodique comme la convolution d'un train d'impulsions et d'un filtre de fonction de transfert $H(f)$:

$$s(t) = e(t) * h(t) \tag{1.2}$$

Cette dernière expression correspond à une modélisation dite "source-filtre" du signal $s(t)$. L'excitation est alors représentée par $e(t)$ et le filtre par $h(t)$.

- **L'analyse des signaux de parole**

Le caractère périodique des signaux voisés a permis d'introduire cette modélisation source-filtre pour l'analyse et le traitement de la parole. On parle alors de modèle simplifié de la production de la parole : la source vocale $e(t)$ est une série d'impulsions espacées de T_0, représentant les impulsions périodiques produites par les cordes vocales. Elle contient l'information fréquentielle du signal. La réponse impulsionnelle $h(t)$ du filtre modélise le conduit vocal. Ce modèle permet de représenter les sons voisés comme la sortie d'un filtre excité par une source périodique dont la période est appelée la "pitch-période". La représentation des sons non voisés est également possible. Ils sont équivalents à la sortie d'un filtre excité par un bruit blanc.
Dans le domaine spectral, l'enveloppe du signal possède des maxima d'amplitude appelés formants. Ces derniers correspondent aux fréquences de résonance des cavités couplées du conduit, qui sont excitées lors de la production du son. Les formants sont caractéristiques des voyelles et sont étroitement lié à la notion du timbre de voix. Ils permettent la distinction de deux voyelles émises à la même hauteur tonale.
Outre le caractère harmonique, l'observation de la forme temporelle a permis de mettre en évidence le caractère "quasi-stationnaire" de ce type de la parole : $h(t)$ peut être considéré comme un filtre linéaire invariant sur des intervalles de temps relativement courts, de l'ordre de 10 à 50 ms. Au delà de cette durée, cette hypothèse n'est plus valable et les impulsions glottiques peuvent, de plus devenir très irrégulières, entraînant des variations significatives de la "pitch-période".
Ainsi la parole est le plus souvent analysée par trames de signal pour lesquelles la période fondamentale et l'enveloppe spectrale varient très peu. Sous l'hypothèse de stationnarité, la période fondamentale et les amplitudes harmoniques

sont supposées constantes et le caractère harmonique de la parole se substitue au caractère quasi-harmonique précédemment évoqué.
Les fréquences fondamentales rencontrées pour la parole varient environ entre 50 Hz et 400 Hz. La transition entre deux trames voisées est douce, contrairement au cas des signaux musicaux.

• L'analyse des signaux de musique

L'analyse des signaux musicaux a mis en évidence de nombreuses similarités avec la parole, mais également quelques différences. Au regard de la diversité des instruments, il est difficile de généraliser le modèle source-filtre comme modèle de production du signal musical. Malgré tout, on notera que l'on retrouve la notion de formants pour la plupart des instruments à vent. Ainsi ces derniers, à l'instar des voyelles, vont pouvoir se distinguer dans un même registre musical, grâce à la forme de leur spectre.
Plus généralement, la notion de timbre pour les sons musicaux est couramment utilisée. Le timbre est en effet très souvent défini comme la signature acoustique d'un instrument. Comme pour la parole, ce sont les caractéristiques de l'enveloppe spectrale qui reflètent le timbre.

Les signaux musicaux ont également un caractère quasi-stationnaire, mais ils ont en général un comportement local plus stable : les zones de stationnarité sont de l'ordre de 100 ms. Comme pour la parole, on peut alors parler de sources harmoniques, car sur une telle durée d'analyse, la fréquence fondamentale et les amplitudes des harmoniques sont supposées constantes.
Une autre différence est celle relative à l'évolution temporelle de la fréquence fondamentale. Cette dernière peut varier très brutalement entre deux zones de stationnarité, les valeurs des fréquences fondamentales pouvant couvrir plusieurs octaves ou au contraire varier très peu comme dans le cas du vibrato.
Enfin, il est à noter que des instruments tels que la guitare ou le piano peuvent produire des sons polyphoniques sous forme d'accords : le mélange contient plusieurs notes émises par une source unique.
Les études sur la séparation de sons musicaux sont beaucoup plus récentes que pour la parole. Ce sont les similarités avec la parole et notamment leur structure quasi-harmonique, qui ont permis d'étendre le problème à la musique.

1.2.2 Les étapes de la séparation monophonique

Cette partie est consacrée à la description très générale des étapes rencontrées dans les problèmes de séparation. Bien qu'il soit difficile de présenter de manière exhaustive l'ensemble des méthodes et techniques développées pour la séparation, il est toutefois possible d'intégrer les plus importantes au sein d'une structure très générale. Cette dernière a ainsi l'avantage de mettre à disposition un schéma regroupant les diverses approches qui ont été proposées.
Les approches rencontrées peuvent se décomposer en trois étapes différentes : (1) l'**analyse** du signal composite qui cherche à extraire le maximum d'information utilisable pour la séparation.(2) L'**identification** qui permet d'assigner à chaque signal source un ensemble bien défini de paramètres (fréquences, amplitudes, phases). (3) La **synthèse** de chacune des sources à partir de leur ensemble de paramètres.
Comme on le verra par la suite, pour chacune des trois étapes différentes techniques ou méthodes entrent en jeu, la diversité des approches résidant alors dans les multiples combinaisons possibles sur l'ensemble des étapes (1), (2) et (3). On notera également que l'étape (3) reste facultative, puisque nous avons vu que pour certains problèmes de séparation, les résultantes ne sont pas forcément destinées à l'écoute.
La figure suivante (Fig. 1.1), présente un schéma général du problème de séparation. Chaque étape est détaillée par des techniques qui sont présentées par la suite. Le mélange est constitué de plusieurs signaux numériques résultant de sources de type parole ou musique.

- **Analyse du signal**

Cette première étape est primordiale, car elle vise à faire ressortir un maximum d'information sur la composition du signal. Cette information est exprimée le plus souvent sous forme de fréquences fondamentales. L'analyse pour la séparation utilise alors des méthodes d'estimation, qui en sortie mettent à disposition suivant la méthode, un ensemble plus ou moins complet de paramètres harmoniques.
Le caractère quasi-stationnaire des signaux composites évoqué précédemment implique une analyse par trames successives : le signal est découpé en segments sur lesquels les fréquences fondamentales des signaux sources, ainsi que leurs amplitudes harmoniques correspondantes sont supposées constantes.
Il existe un grand nombre de méthodes d'estimation pour la détermination des paramètres harmoniques. Les méthodes utilisées pour le problème de la séparation

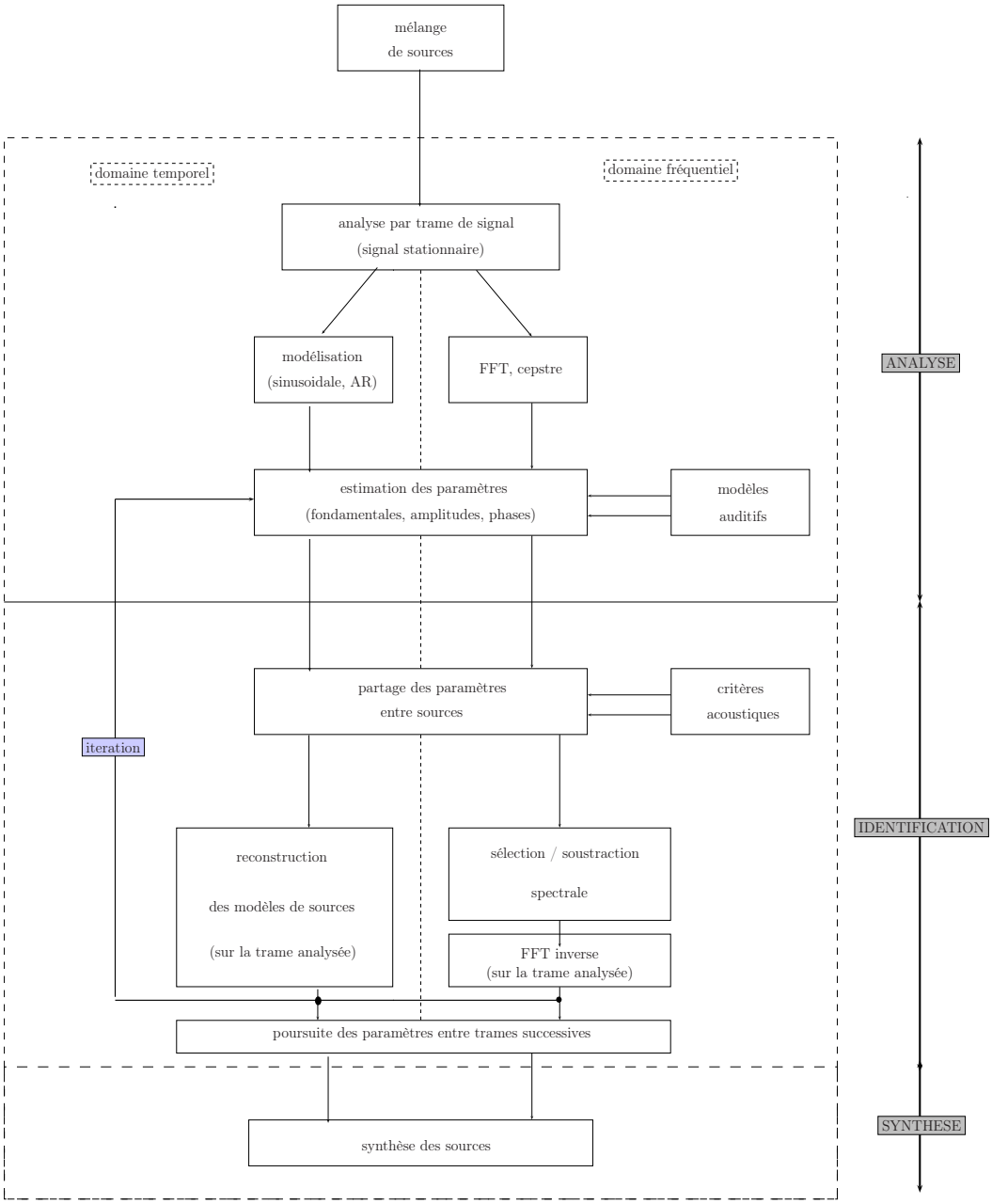


Fig. 1.1 – *Schéma général : les étapes de la séparation.*

sont détaillées par la suite et sont des méthodes dites “à court terme”. On verra alors que certaines estiment les paramètres dans le domaine spectral, à l'aide la transformation de Fourier, d'autres introduisent un modèle de signal temporel et maximisent par exemple, la vraisemblance de ce modèle. Des méthodes hybrides sont également utilisées. On verra également que certaines méthodes utilisent pour améliorer l'estimation, des modèles auditifs ou des modèles d'enveloppes spectrales de signaux connus.

• Identification des paramètres du signal

Cette étape consiste à identifier ou indexer les paramètres estimés lors de l'analyse. Dans certains cas, ce jeu de paramètres est complété à l'aide de critères ou d'hypothèses psychoacoustiques qui vont pallier les déficiences de l'estimateur, comme par exemple lorsque la résolution spectrale de la FFT n'est pas suffisante pour prendre en compte tous les partiels du mélange.
Par la suite, on cherche à attribuer à chaque source les paramètres qui sont susceptibles de lui correspondre. L'identification se fait à deux niveaux :

– Local : au sein d'une trame les paramètres sont classés en sous-ensembles. Le partage s'effectue généralement en déterminant les fréquences fondamentales du mélange et en leur associant les partiels ou les harmoniques relatives. Suivant l'approche, il peut s'appuyer sur la modélisation temporelle, ou encore sur des méthodes du type “soustraction spectrale”, “échantillonnage fréquentiel” ou “Peak Picking”.

– Global : sur la durée d'analyse du mélange, les sous-ensembles de trame ainsi constitués sont assignés aux sources à l'aide de méthodes de poursuite, certaines faisant appel à des critères psychoacoustiques.

On notera qu'il existe deux types d'identification. Ces derniers sont directement liés à la méthode d'estimation utilisée lors de l'analyse. Une première identification correspond alors à une estimation conjointe des paramètres harmoniques, qui sont par la suite séparés entre chaque source. La seconde identification est basée sur une méthode de type itératif, qui cherche à estimer les paramètres d'un premier signal source (généralement le plus fort) et réitère le processus d'estimation après élimination de ce jeu dans les données d'observation.

• **Synthèse des signaux sources**

Lorsque le problème de séparation nécessite l'écoute d'une ou plusieurs sources, la plupart des approches font appel aux méthodes couramment utilisées dans la synthèse sonore, telles que la synthèse harmonique ou la synthèse FFT/OLA. Toutes deux utilisent les paramètres estimés lors de l'analyse et assignés à la source à synthétiser.
La qualité sonore des sources reconstruites est donc directement liée aux performances des estimateurs, mais elle est également liée à celles des techniques employées par la suite, comme la soustraction spectrale ou la poursuite des paramètres le long des trames analysées.

1.3 L'estimation de fréquences fondamentales multiples

L'Analyse/Synthèse de la parole est à l'origine d'un grand nombre de méthodes d'estimation plus performantes les unes que les autres pour une fréquence fondamentale unique [1]. Ainsi, les premières approches pour traiter le problème de séparation ont cherché à exploiter ces méthodes pour le cas de deux ou plusieurs fréquences fondamentales.
Deux types d'approches ont été envisagées : certaines méthodes sont développées sur un schéma itératif dans lequel les fréquences fondamentales sont successivement estimées. Ce schéma fait alors appel à une méthode d'estimation "mono-pitch", et intègre une étape de soustraction des estimations courantes. D'autres effectuent l'estimation conjointe des fréquences fondamentales de l'ensemble des signaux sources du mélange.

Face au nombre conséquent de méthodes d'estimation de fréquences fondamentales existantes, il apparaît très difficile d'en faire une présentation exhaustive. Par conséquent, seules les méthodes utilisées ou développées dans le cadre de la séparation sont présentées dans la suite : les méthodes d'estimation spectrales, puis temporelles sont tout d'abord abordées, puis nous discutons le cas des méthodes basées sur des modèles auditifs.

1.3.1 Les méthodes d'estimation spectrale

La simplicité d'utilisation de la FFT a permis de développer un certain nombre de méthodes d'estimation spectrale. Ces dernières consistent à extraire les paramètres harmoniques du mélange, par estimation de sa densité spectrale de puissance (DSP). L'histogramme harmonique ou encore l'analyse cepstrale de la parole, initialement développés pour l'estimation d'une fréquence fondamentale unique, en sont des exemples type. Ces derniers, bien adaptés pour des signaux simples et peu bruités, sont limités par leur résolution spectrale. Ainsi par la suite, d'autres méthodes spectrales ont été proposées telles que les méthodes dites "AR".
L'ensemble de ces méthodes sont détaillées ci-dessous. Dans cette partie, nous présenterons également la méthode "Two-Way-Mismatch", un peu en marge des précédentes, mais qui comme elles, a été développée pour le problème de séparation.

• Histogramme harmonique, somme et produit spectral

"L'histogramme harmonique" est une méthode spectrale initialement proposée par Schroeder([2]) qui permet de déterminer la fréquence fondamentale d'un signal par compression de son spectre sur l'axe des fréquences.
Dans la pratique, les harmoniques sont détectées en localisant les maxima du spectre. Les fréquences correspondantes sont alors utilisées pour former un histogramme. Ce dernier présente la somme des contributions spectrales de chaque fréquence et de leurs sous-harmoniques (chaque fréquence est divisée par 2,3,4...). La fréquence fondamentale est alors choisie comme la fréquence qui recueille le maximum de contributions.
L'utilisation de cet estimateur sur des signaux composites est à l'origine d'une des premières méthodes de séparation de parole : " la sélection harmonique" proposée par Parson [3]. Appliquée au cas de deux signaux voisés, cette méthode utilise une version affinée de l'histogramme harmonique qui va prendre en compte d'éventuels recouvrements entre harmoniques, par l'intermédiaire de critères basés sur la forme du pic considéré (symmétrie, distance entre pics) et le comportement de la phase. L'estimation d'une première fréquence fondamentale est obtenue, puis les contributions harmoniques correspondantes sont annulées, permettant l'estimation de la seconde fréquence fondamentale.

Ce principe de compression du spectre est plus généralement recontré sous

le nom de "somme spectrale" et "produit spectral" [4]. La première méthode consiste à sommer les H premières harmoniques du périodogramme, tandis que la seconde effectue leur produit. Leur définition respective est donnée par :

$$S(f) = \sum_{h=1}^{H} |X(hf)| \qquad et \qquad P(f) = \prod_{h=1}^{H} |X(hf)| \tag{1.3}$$

Déterminer la fréquence fondamentale revient alors à maximiser par rapport à f, la somme S(f) dans le premier cas ou le produit P(f) dans le second cas.

• **Cepstre**

L'analyse cepstrale est une méthode bien connue du traitement de la parole ou encore des problèmes de dé-réverbération. Le cepstre est une transformation qui, lorsqu'elle est appliquée à un signal périodique, permet de déconvoluer l'enveloppe spectrale et l'excitation périodique de ce dernier, pour les traiter indépendamment [5]. Pour un signal voisé de période fondamentale T_0 et caractérisé par l'équation (eq. 1.1), l'expression de son cepstre est donnée par :

$$\begin{aligned} C(n) &= FFT^{-1}(log(|S(f)|)) \\ C(n) &= FFT^{-1}(log(|E(f)|)) + FFT^{-1}(log|(H(f)|)) \end{aligned} \tag{1.4}$$

L'enveloppe spectrale est alors représentée par des pics proches de l'origine, tandis que l'excitation harmonique $E(f)$ correspond à un pic à la période T_0.

Tout comme la méthode de l'histogramme harmonique, l'estimation de la fréquence fondamentale d'un signal périodique fût généralisée au cas de deux sources voisées [6]. Ainsi, pour un mélange contenant les sources voisées $s_1(n)$ et $s_2(n)$:

$$\begin{aligned} |S(f)|^2 = [|E_1(f)|\ |H_1(f)|]^2 + [|E_2(f)|\ |H_2(f)|]^2 \\ +2|E_1(f)|\ |H_1(f)|\ |E_2(f)|\ |H_2(f)| \times cos[\phi_1(f) - \phi_2(f)] \end{aligned} \tag{1.5}$$

où $|S(f)|$, $|E_i(f)|$, $|H_i(f)|$ et $\phi_i(f)$ sont respectivement les spectres d'amplitude du mélange, de l'excitation,de l'enveloppe et la phase pour la source $i = 1\, ou\, 2$. D'après l'équation (eq. 1.5), l'application de la transformation cepstrale ne permet plus d'obtenir une superposition linéaire des excitations et des enveloppes. Pour le cas particulier où il n'existe aucun recouvrement entre les harmoniques

des deux sources, il est tout de même possible d'utiliser l'analyse cepstrale pour déterminer leur fréquence fondamentale. En effet, si f_1 et f_2 sont respectivement les fréquences fondamentales de $s_1(n)$ et $s_2(n)$ on aura alors :

$$\begin{cases} S(f) = E_1(f)\, H_1(f) & \text{si} \quad f = m \times f_1,\ m = 1, 2, 3, .. \\ S(f) = E_2(f)\, H_2(f) & \text{si} \quad f = m \times f_2,\ m = 1, 2, 3, .. \end{cases} \tag{1.6}$$

Dans le cas le plus général, le cepstre résultant est alors la somme des contributions correspondant à des régions de recouvrement pour lesquelles le principe de superposition est invalide et des régions où les voix sont isolées.
Cet estimateur de fréquences fondamentales a été proposé par Stubb et Summerfield pour séparer des paires de voyelles. La séparation consiste à extraire le signal désiré en atténuant l'excitation harmonique du signal interférant.

• **Méthodes AR d'estimation spectrale**

Les méthodes AR sont des méthodes paramétriques qui, contrairement aux méthodes précédentes, utilisent une modélisation autorégressive des données, pour estimer la densité spectrale de puissance du signal. Développées pour le traitement de données courtes, elles permettent de pallier le problème de résolution spectrale, liée à la troncature en temps des méthodes qui estiment directement la DSP à partir du signal.
Ainsi les estimateurs basés sur le périodogramme, comme ceux présentés auparavant, ont des limitations dues à la résolution fréquentielle. Cette résolution qui est inversement proportionnelle à la durée du signal analysé va permettre, dans le cas de la parole de résoudre des fréquences distantes d'environ 20 Hz. Cette résolution largement suffisante dans le cas d'un locuteur unique, devient beaucoup trop faible dans le cas de deux locuteurs et plus.

Pour N observations $\{x(0), x(1), \cdots, x(N-1)\}$, dont la modélisation autorégressive est donnée par :

$$x(n) = -\sum_{k=1}^{p} a_k\, x(n-k) + e(n) \tag{1.7}$$

l'expression de sa densité spectrale de puissance se présente sous la forme :

$$DSP_{AR}(f) = \frac{\sigma^2}{|1 + \sum_{k=1}^{p} a_k e^{-j2\pi kf}|^2} \tag{1.8}$$

où σ^2 et les coefficients $\{a_k\}$ sont respectivement la variance d'un bruit blanc gaussien e(n) et les paramètres autorégressifs du modèle.

Il existe alors un grand nombre de méthodes d'estimation des coefficients autorégressifs [7], mais nous n'évoquerons ici que trois méthodes différentes, toutes développées dans le cadre de la séparation.
La première méthode est celle de Naylor et Porter [8] dont l'idée directrice est d'obtenir une meilleure analyse spectrale et plus particulièrement d'améliorer la finesse des pics. Ils utilisent le calcul de la covariance modifiée pour estimer les fréquences fondamentales de deux locuteurs à partir du spectre de mélange. Cette dernière permet d'ajuster un modèle au signal $x(n)$ en minimisant au sens des moindres carrés, les erreurs de prédiction directe et rétrograde ("forward and backward prediction").
L'estimation est effectuée à l'aide de l'algorithme de covariance modifiée développé par Marple (MC Algorithm) [9]. Par la suite, les pics spectraux sont regroupés à l'aide d'un algorithme de classification développé par les auteurs, mais non explicité.
La seconde méthode AR utilisée pour la séparation est celle de Childers et Lee [10], basée sur l'analyse spectrale par minimisation de l'entropie croisée multisignal ("Multisignal MCESA"). Ce type d'analyse a pour but d'estimer le spectre d'un signal à partir d'un nombre donné de valeurs de la fonction d'autocorrélation. Childers et Lee proposent d'utiliser la modélisation AR spectrale, non pas pour l'estimation directe des fréquences fondamentales des différentes sources, mais pour améliorer les estimations spectrales de chacune, séparées auparavant dans le domaine spectral par la technique HMS ("Harmonic Magnitude Suppression"). Cette technique fait l'objet d'un paragraphe complet dans la suite de cet état de l'art (cf. paragraphe 1.5.2).
Dans un même registre, une approche proposée par Mumolo, permet d'estimer les fréquences fondamentales à l'aide de la corrélation spectrale [11]. Cherchant à exploiter les périodicités du spectre de Fourier du signal composite, il calcule la fonction d'autocorrélation, puis détermine les fréquences fondamentales de chaque locuteur, en effectuant à nouveau une transformée de Fourier à court terme (TFCT). La corrélation spectrale permet ainsi d'obtenir des pics spectraux beaucoup plus fins et d'améliorer la résolution fréquentielle.

• La méthode du Two-Way-Mismatch

En 1989, Maher s'intéresse à des "duo musicaux" [12]. Il propose tout d'abord un nouvel algorithme pour l'estimation de la fréquence fondamentale d'un signal audio. Pour cela il associe les concepts de l'analyse spectrale à court-terme à l'approche de Mc Aulay et Quatiery [13] en parole : l'algorithme développé est basé sur la comparaison entre le spectre mesuré, pour lequel les paramètres du signal sont obtenus par une analyse TFCT, et un modèle spectral harmonique. Pour ce modèle, les partiels sont prédits à partir de fréquences fondamentales choisies.
La procédure, appelée "Two-Way-Mismatch", consiste à une double minimisation d'erreur : l'erreur à minimiser est ainsi calculée comme une somme de différences normalisées au carré, entre chaque partiel mesuré et son partiel prédit le plus proche et entre chaque partiel prédit et son partiel mesuré le plus proche.

Par la suite, il généralise cet algorithme au cas de deux fréquences fondamentales et s'intéresse alors au problème de séparation. Les duos considérés sont tels que les fréquences fondamentales appartiennent à deux intervalles fréquentiels $I_l = [fmin_l, fmax_l]$ et $I_h = [fmin_h, fmax_h]$ qui ne se recouvrent pas.
La procédure de recherche des fréquences fondamentales s'effectue comme suit :

(1) La paire de fréquences fondamentales est initialisée à :

$$\{f_l, f_h\} = \{fmin_l, \frac{1}{2}(fmin_h + fmax_h)\} \tag{1.9}$$

(2) l'erreur TWM est calculée pour chaque paire de fréquences $\{f_l, f_h\}$ où f_l est incrémentée par pas d'un demi ton jusqu'à la valeur $fmax_l$. Si un minimum local est détecté pour une paire de fréquence, le processus de recherche est répété dans le voisinage de cette paire, avec une valeur d'incrémentation réduite, pour affiner l'estimation. La fréquence $f_l \in I_l$ de la paire correspondant à l'erreur minimale est identifiée comme la première estimation $\widehat{f_l}$.

(3) l'erreur TWM est calculée pour chaque paire de fréquences $\{\widehat{f_l}, f_h\}$ où f_h est incrémentée par pas d'un demi ton jusqu'à la valeur $fmax_h$. De même que pour l'étape précédente, l'estimé $\widehat{f_h}$ est obtenu en recherchant l'erreur minimale.

(4) Le processus est finalement réitéré à partir de l'étape 2, jusqu'a l'obtention d'un minimum global pour l'erreur TWM.

1.3.2 Les méthodes d'estimation temporelles

Parallèlement aux méthodes spectrales, plusieurs approches temporelles ont été proposées. Pour la plupart, elles sont la généralisation mathématique de méthodes d'estimation monopitch et se distinguent par la prise en compte ou non d'un modèle paramétrique du mélange. Permettant de pallier le problème de résolution spectrale, ces dernières sont en général très coûteuses en calcul et sont très sensibles aux ambiguïtés d'octave, c'est-à-dire à l'estimation de multiples des périodes fondamentales.

• Les méthodes de corrélation

Il existe à ce jour un certain nombre de méthodes basées sur le calcul de la corrélation, développées pour l'estimation d'une unique fréquence fondamentale. Nous n'en définirons dans la suite que trois premières qui sont à l'origine de méthodes d'estimation de fréquences fondamentales multiples.
Les deux premières méthodes sont basées sur une recherche de similarités entre deux versions décalées du signal observé, tandis que la troisième utilise la périodicité de la fonction d'autocorrélation du signal.

La méthode AMDF : Certainement l'une des plus anciennes, elle correspond à la minimisation de la fonction dite AMDF ("Average Magnitude Difference Function"), définie par :

$$AMDF(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} |x(n) - x(n+k)| \tag{1.10}$$

La méthode ASDF : Cette méthode préfère la minimisation du critère énergétique qu'est la fonction ASDF ("Average Square Difference Function") :

$$ASDF(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} [x(n) - x(n+k)]^2 \tag{1.11}$$

L'autocorrélation : C'est sans aucun doute, la méthode la plus usitée. La périodicité du signal est obtenue en maximisant la fonction d'autocorrélation $r(k)$

du signal $x(n)$ définie par :

$$r(k) = \sum_{n=0}^{N-k-1} x(n)x(n+k) \tag{1.12}$$

On notera qu'il existe une variante à cette méthode, connue sous le nom de "méthode d'autocorrélation croisée". Pour cette dernière, la sommation (eq. 1.12) est calculée sur N termes. Elle correspond ainsi à la mesure du degré de similarité de versions décalées du signal pour deux fenêtres successives. L'intérêt de cette méthode est qu'elle permet de pallier les problèmes d'atténuation des pics aux grandes valeurs k, que l'on rencontre avec la méthode d'autocorrélation basique.

La généralisation de ces méthodes pour l'estimation de fréquences fondamentales multiples a été proposée par plusieurs auteurs. Cette dernière s'effectue facilement, en remarquant qu'une condition nécessaire et suffisante pour qu'un signal $x_s(n)$ soit une somme de K signaux périodiques, dont les périodes respectives sont T_1, T_2, $\cdots$,T_K, est qu'il s'annule par le filtre de réponse impulsionnelle $h(n)$:

$$\begin{cases} h(n) = h_1(n) * h_2(n) * \cdots * h_K(n) \\ \\ h_i(n) = [\delta(n) - \delta(n+T_i)] \end{cases} \tag{1.13}$$

h_i est la réponse impulsionnelle d'un filtre en peigne ou "filtre comb".

On remarque alors que :

- pour $K = 1$, le signal $x_s(n) = x_1(n)$ est périodique pour T_1, si $x_s(n) = x_s(n+T_1)$
- pour $K = 2$, le signal polyphonique $x_s(n) = x_1(n) + x_2(n)$ possède les périodes T_1 et T_2, si $x_s(n) = x_s(n+T_1) + x_s(n+T_2) - x_s(n+T_1+T_2)$. Il en résulte que x_s est lui-même périodique de période T_{12}, où T_{12} est le plus grand diviseur commun de T_1 et T_2.

Une généralisation de l'AMDF a été proposée par De Cheveigné [14] [15]. Pour une polyphonie à deux signaux caractérisés par T_1 et T_2, la fonction à minimiser est la DDF ("Double Difference Function") définie par :

$$DDF(k_1,k_2) = \frac{1}{(N-k_1-k_2)} \sum_{n=0}^{N-k_1-k_2-1} [x_s(n) - x_s(n+k_1) - x_s(n+k_2) + x_s(n+k_1+k_2] \tag{1.14}$$

L'estimation des deux périodes fondamentales est donc obtenue par minimisation de l'amplitude de sortie après application au signal de deux filtres comb en cascade.

Avec la DDF, De Cheveigné a pour principal but de développer une analyse temporelle permettant de modéliser le traitement auditif de voyelles émises simultanément. Quant à la généralisation de l'ASDF, on retrouve son utilisation dans le travail de Mumolo [16] qui permet d'estimer les périodes cette fois-ci par minimisation de l'énergie.

La DDF est une fonction symétrique, qui possède des zéros pour les couples de valeur (n_1T_1, n_2T_2) où n_1, n_2 sont des entiers. Mais elle possède également, d'après la remarque précédente, des zéros pour les couples du type (n_1T_{12}, n_2) et (n_1, n_2T_{12}) où l'on rappelle que T_{12} est le plus grand diviseur commun de T_1 et T_2. Il existe donc une ambiguïté d'octave lors de l'estimation, car il est impossible de déterminer si le minimum de la DDF est obtenu pour l'une des périodes fondamentales ou un multiple de l'une des deux. Cette indétermination est déjà présente dans les fonctions de distances AMDF et ASDF, mais il est à noter que l'ambiguïté entre les pics T, $2 \times T$, $3 \times T$, $\cdots$ est souvent atténuée par la non-stationnarité du signal : plus le décalage est grand, plus le signal de par sa non-stationnarité, intègre des différences par rapport à la trame de départ. Le signal sera donc "un peu plus différent" pour un décalage de $2 \times T$ que de T. Malgré tout, cette remarque ne résout pas le problème de l'estimation d'une valeur multiple de l'une des deux périodes pour le cas des signaux composites.
Dans ces approches, les estimations des périodicités se font conjointement et sont déterminées par une recherche exhaustive dans l'espace (T_1, T_2) qui nécessite en pratique un coût de calcul important. On notera tout de même que la symétrie des fonctions à minimiser permet de diminuer cet espace de recherche de moitié.

• L'estimateur du Maximum de Vraisemblance

Initialement proposée J.D Wise [4] [17], cette méthode d'estimation repose sur un modèle probabiliste du signal : le signal observé $x(n)$, sur N échantillons, est supposé être la somme d'un signal purement périodique $s(n)$ de période T_0 et d'un bruit blanc gaussien $e(n)$ de variance σ^2. Les paramètres du modèle sont alors estimés en maximisant la vraisemblance ou log-vraisemblance des observations.

La densité de probabilité du signal observé étant donnée par :

$$p(x|s,T_0,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\, exp[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-s(n))^2\,] \tag{1.15}$$

les estimations de $s(n)$ et σ^2 sont alors obtenues en maximisant l'expression de la log-vraisemblance correspondante :

$$L(s,T_0,\sigma^2) = -\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-s(n))^2 \tag{1.16}$$

Les maximisations successives par rapport à $s(n)$ et σ^2 déterminent les estimations correspondants :

$$\begin{cases} \widehat{s}(n) = \frac{1}{L}\sum_{l=0}^{L-1} x(n+lT_0) \\ \\ \widehat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x(n)-\widehat{s}(n))^2 \end{cases} \tag{1.17}$$

où L est le nombre de périodes observées tel que $N = LT_0$.

La réintroduction de (eq.1.17) dans l'expression de la densité de probabilité permet de montrer que la maximisation est équivalente à la minimisation de l'écart quadratique $\sum_{n=0}^{N-1}(x(n)-\widehat{s}(n))^2$.
L'estimation de la période fondamentale T_0 du signal est finalement obtenu par maximisation de l'expression :

$$E(T) = \frac{1}{K}[r_x(0) + 2\sum_{l=1}^{L-1} r_x(l\times T_0)] \tag{1.18}$$

où r_x est la fonction d'autocorrélation de $x(n)$:

$$r_x(l) = \frac{1}{N-l}\sum_{n=0}^{N-l-1} x(n)x(n+l) \tag{1.19}$$

La quantité E correspond alors à une moyenne des valeurs d'autocorrélation aux multiples de T_0. C'est une fonction qui possède des pics régulièrement espacés et d'amplitude voisine, lorsque le signal est purement périodique. En revanche, lorsque le signal n'est pas strictement périodique, les coefficients de covariance ne sont plus égaux à $r_x(0)$ pour les multiples de T_0 et la pondération implicite dans (eq.1.18) favorise les périodes élevées : les pics de E sont donc plus prononcés

pour les multiples de la période.

Tout comme pour les méthodes de corrélation, il est possible d'effectuer une généralisation de l'estimateur du maximum de vraisemblance. L'expression de la log-vraisemblance fait alors intervenir toutes les périodes $T_1, \cdots, T_K$ du signal composite :

$$L(s_1, \cdots, s_K, T_1, \cdots, T_K, \sigma^2) = -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x(n) - \sum_{i=1}^{K} s_i(n))^2 \tag{1.20}$$

Sa maximisation aboutit à celle d'une expression complexe de $E(T_1, \cdots, T_K)$, dont le coût de calcul augmente très fortement avec le nombre de sources.

Toutefois, des études passées ont montré que le Maximum de Vraisemblance est certainement l'un des estimateurs les plus robustes, permettant de traiter le cas d'observations bruitées [4] [17]. C'est principalement cette dernière raison qui est à l'origine de son utilisation dans les problèmes de séparation : chaque source d'un mélange peut être vue, en effet, comme noyée dans un environnement bruité. On retrouve ainsi dans [18], qui traite le problème de séparation de deux locuteurs, une estimation basée sur le principe du maximum de vraisemblance. Ce dernier est appliqué à l'aide d'une formulation utilisant le produit d'un filtre en peigne $C_{T_0}(k)$ et de l'autocovariance du signal d'entrée :

$$E'(T_0) = \frac{2T_0}{N} \sum_{l=1}^{L} (lT_0)^{-8} C_l(l \times T_0) r_x(l \times T_0) \tag{1.21}$$

où $L = (N - T_0)/T_0$ et le terme $(l \times T_0)^{-8}$ est introduit pour pallier le problème d'ambiguïté d'octave de l'estimateur. Les auteurs définissent alors (eq. 1.21) comme un détecteur de fréquence fondamentale qui possède un "effet de capture", permettant de détecter la fréquence fondamentale du locuteur le plus fort. L'idée proposée est alors de développer une méthode de séparation de type itératif, dans laquelle (eq. 1.21) est appliquée pour l'obtention des deux fréquences fondamentales.

1.3.3 La modélisation sinusoïdale

Les méthodes d'estimation basées sur la modélisation sinusoïdale sont aujourd'hui devenues des plus courantes et sont par exemple très utilisées dans le

Codage ou l'Analyse/Synthèse. L'idée est d'introduire un "à priori" sur l'harmonicité du signal, en le modélisant par une somme de sinusoïdes.
Les premiers gros travaux datent des années 80 et sont dûs à Quatieri et Mac Aulay : Ils s'intéressent alors à l'analyse et la synthèse d'un unique signal de parole voisée [19]. Ce modèle trés performant, est étendu peu après au cas de mélanges voisés par Quatieri et Danisewicz, chaque source étant modélisée par une somme de sinusoïdes caractérisées par leurs fréquences, amplitudes et phases[20].

• L'estimateur des moindres carrés

Soit un signal analysé sur une fenêtre de longueur N, pour laquelle l'hypothèse de stationnarité est vérifiée, la modélisation sinusoïdale est donnée par :

$$s(n) = \sum_{k=1}^{M} (A_k cos(2\pi f_k n) + B_k sin(2\pi f_k n)) \tag{1.22}$$

Les paramètres du modèle, supposés constants, sont alors les M couples d'amplitudes $\{A_k, B_k\}$ et les fréquences correspondantes f_k.

Hypothèse d'harmonicité : Il est possible d'introduire une hypothèse supplémentaire sur l'harmonicité du modèle, en supposant que les fréquences sont en relation harmoniques telles que $f_k = k \times f_0$. f_0 est alors la fréquence fondamentale du modèle.

Cette hypothèse à l'avantage de diminuer le nombre de paramètres à estimer, mais rend également le modèle moins flexible et dans certains cas plus irréaliste, comme par exemple pour des signaux musicaux qui sont clairement caractérisés par leur inharmonicité, tel que le piano.

Application aux mélanges : Une manière d'estimer les paramètres harmoniques des modèles de sources est d'utiliser la méthode d'estimation par les moindres carrés. Cette dernière minimise l'écart quadratique entre valeurs observées et valeurs attendues. Pour un mélange $x(n)$, constitué de N_s sons harmo-

niques, on a alors ;

$$J(A_k^i, B_k^i, f_k^i) = \sum_{n=0}^{N-1} |x(n) - \sum_{i=1}^{N_s} s_i(n)|^2 \tag{1.23}$$

Sous forme matricielle cette expression devient :

$$J(A_k^i, B_k^i, f_k^i) = |\underline{x} - \Omega\underline{\mu}|^2 \tag{1.24}$$

$\underline{x}$ est le vecteur des N données d'observation. $\underline{\mu} = [\underline{\mu_1}, \underline{\mu_2}, \cdots, \underline{\mu_{N_s}}]^T$ est telle que le vecteur des amplitudes $\underline{\mu_i}$ s'écrit :

$$\underline{\mu_i} = [A_1^i, .., A_{M_i}^i, B_1^i, .., B_{M_i}^i] \tag{1.25}$$

Ω est la matrice des cosinus-sinus dépendant des fréquences $\{f_k^i\}$. son expression est mise sous la forme $\Omega = [\,\Omega_1\,\Omega_2\,\cdots\,\Omega_{Ns}]$ avec :

$$\mathbf{\Omega_i(f)} = \begin{bmatrix} cos(2\pi f_1^i(p)) & \cdots & cos(2\pi f_M^i(p)) & sin(2\pi f_1^i(p)) & \cdots & sin(2\pi f_M^i(p)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1^i(p+1)) & \cdots & cos(2\pi f_M^i(p+1)) & sin(2\pi f_1^i(p+1)) & \cdots & sin(2\pi f_M^i(p+1)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1^i(N-1)) & \cdots & cos(2\pi f_M^i(N-1)) & sin(2\pi f_1^i(N-1)) & \cdots & sin(2\pi f_M^i(N-1)) \end{bmatrix} \tag{1.26}$$

L'estimation par la méthode des moindres carrés des amplitudes est obtenue à l'aide :

$$\underline{\widehat{\mu}} = (\Omega^H\Omega)^{-1}\Omega^H\underline{x} \tag{1.27}$$

Les fréquences f_k^i sont estimées en réintroduisant $\underline{\widehat{\mu}}$ dans la fonction coût (eq. 1.24). On montre alors que la minimisation de l'écart quadratique par rapport aux fréquences revient à maximiser la quantité suivante :

$$K(f_k^i) = \underline{x}\,[\Omega(\Omega^H\Omega)^{-1}\Omega^H]\,\underline{x} \tag{1.28}$$

et on a :

$$\underline{\widehat{f}_k^i} = argmax[K(f_k^i)] \tag{1.29}$$

La matrice $(\Omega^H\Omega)$ est inversible lorsque les fréquences $\{f_k^i\}$ sont toutes différentes, car les vecteurs colonnes de Ω sont ainsi indépendants. L'estimation par la méthode des moindres carrés ne peut donc s'appliquer telle quelle pour des

sources possédant des harmoniques communes.
La fonction (eq. 1.28) est une fonction multivariables qui possède en général plusieurs maxima locaux et qui rend le problème d'optimisation difficile à résoudre. Toutefois, on peut montrer que pour des écarts fréquentiels supérieurs à la limite de Fourier, la méthode est équivalente à l'utilisation dans le domaine fréquentiel du périodogramme.
En pratique, les fréquences fondamentales sont estimées pour chaque segment en localisant les maxima de leur FFT, et les amplitudes correspondantes sont déterminées par (eq. 1.27). On notera d'ailleurs que la simplicité d'utilisation de la FFT a contribué à la popularité de cette méthode d'estimation.

Une des premières utilisations de cet estimateur pour la séparation de deux locuteurs, a été proposée par Quatieri et Danisewicz [20]. Ils s'intéressent plus particulièrement aux performances de l'estimation des moindres carrés en la comparant à deux autres méthodes : la première estime tous les paramètres (amplitudes et fréquences) du modèle à l'aide du périodogramme, tandis que la seconde méthode estime les amplitudes par échantillonnage fréquentiel du spectre du signal à partir de fréquences supposées connues.
Les conclusions de cette étude comparative sont obtenues par le biais de tests d'écoute. Ces derniers montrent que l'approche moindres carrés avec ou sans information a priori sur les fréquences, apparaît dans les deux cas beaucoup plus performante que la méthode d'échantillonnage fréquentielle. Lorsqu'aucune information a priori n'est disponible, c'est-à-dire lorsque l'estimation de la fréquence fondamentale et de l'amplitude sont toutes deux requises, la suppression de l'interférence apparaît réalisable pour des sources de même intensité.
Pour déterminer les pertes de performance dues à l'estimation de la fréquence fondamentale elle-même, une autre série de tests a été effectuée, comparant ce cas à celui intégrant la connaissance a priori de la fréquence fondamentale. Aucune différence essentielle dans la suppression n'a été observée, mais d'un point de vue perceptif l'estimation semble rendre le signal désiré légèrement moins naturel.

A la suite du travail de Quatieri et Danisewicz , d'autres méthodes utilisant l'analyse-synthèse sinusoïdale ont été proposées. Etant donné que la qualité sonore des signaux séparés va dépendre très fortement de l'estimation des paramètres de modèle, des estimateurs de fréquence fondamentale se voulant plus performants que celui de la minimisation par les moindres carrés ont été présentés :
En 1997, Benincasa propose d'améliorer la méthode de séparation de Quatieri en déterminant les paramètres moindres carrés du modèle à l'aide d'une méthode d'optimisation non-linéaire. Il effectue alors cette dernière en développant une programmation quadratique séquentielle [21].

• **La modélisation quasi-sinusoïdale**

En 1993, Chazan, Stettiner et Malah proposèrent une méthode de séparation qui utilise non plus un modèle exactement périodique, mais un modèle où la fréquence fondamentale peut varier linéairement dans l'intervalle d'analyse [22]. Cette hypothèse de quasi-périodicité permet d'utiliser des trames de signal plus longues tout en palliant les problèmes d'inharmonicité des signaux de paroles. L'expression du modèle quasi-périodique est alors donnée par :

$$h_k(t) = \sum_{k=1}^{L(\phi(t))} C_k \exp\left(jk\phi(t)\right) \tag{1.30}$$

$\phi(t)$ est une fonction de distorsion. Sa dérivée représente la fréquence fondamentale instantanée $\varphi(t)$ du signal voisé :

$$\begin{aligned} \phi(t) &= \tfrac{1}{2}\alpha t^2 + 2\pi\beta t \\ \varphi(t) &= \alpha t + \beta \end{aligned} \tag{1.31}$$

Le problème d'estimation de pitch correspond finalement à rechercher simultanément les solutions moindres carrés de α et β pour chaque locuteurs. Les auteurs montrent tout d'abord que cette optimisation non-linéaire équivaut à déterminer la fonction de distorsion $\phi(t)$ qui maximise l'énergie de la sortie d'un filtre en peigne appliqué au signal composite. Ils utilisent par la suite l'algorithme EM (Expectation-Maximization algorithm) pour effectuer l'estimation de fréquences fondamentales multiples, ce dernier permettant de décomposer le problème d'estimation simultanée de (2×2) paramètres en deux problèmes séparés de dimension 2, résolus itérativement.

1.3.4 Introduction de modèles auditifs

Comme le concept de pitch se réfère à la perception auditive et est également en relation complexe avec les propriétés du signal, il apparaît naturel d'appliquer des méthodes de séparation qui simulent ou s'inspirent de la perception humaine. Ainsi un ensemble de méthodes de séparation s'appuyant sur une modélisation du traitement auditif ont été proposées. L'intérêt porté à la compréhension du système d'audition a permis d'introduire de nouvelles méthodes d'estimation pour

les fréquences fondamentales multiples. Une liste non-exhaustive des études psychophysiques du système auditif est proposée par De Cheveigné dans [23].

• **Modèle de perception de pitch**

Dans [24], Tolonen propose une méthode de séparation dans laquelle l'estimation des pitch s'effectue à partir d'un modèle simplifié de perception de pitch.
Il s'inspire pour cela du modèle de perception développé par Meddis et O'mard [25]. Ce dernier utilise un banc de filtres qui simule le comportement de la cochlée : le son à traiter est séparé à l'aide de canaux fréquentiels dont la largeur de bande correspond à la résolution fréquentielle de la cochlée.
Le signal d'entrée est divisé en signaux multicanaux (40 à 128 canaux suivant l'implémentation), puis filtré afin de détecter l'enveloppe du signal pour chaque canal. A partir de ces enveloppes, une mesure de périodicité utilisant la fonction d'autocorrélation (ACF) est effectuée. Les ACF obtenues sont alors sommées et fournissent ainsi une fonction d'autocorrélation globale à partir de laquelle l'estimation de fréquences fondamentales multiples sera faite.
Les diverses expérimentations de ce modèle ont montré que ce dernier possède une certaine habilité à simuler de nombreuses caractéristiques connues de la perception, mais possède également un inconvénient pratique majeur lié au coût de calcul.
Le coût de calcul étant directement lié aux nombre de canaux du banc de filtres utilisé lors de l'analyse, Tolonen propose de développer un modèle simplifié de la perception auditive. Ce modèle ne cherche pas, contrairement à celui de Meddis et O'Mard, à simuler le système auditif en détails. Il utilise alors pour l'estimation mutipitch un banc de filtres à 2 canaux.

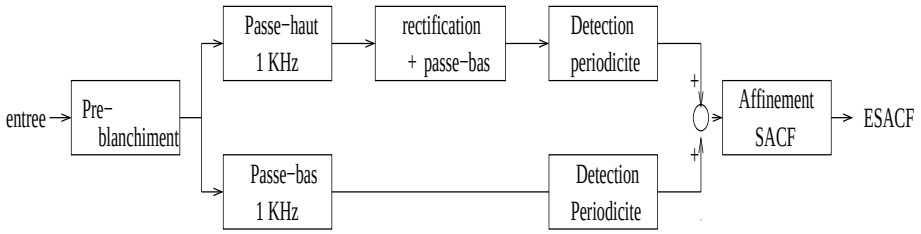


Fig. 1.2 – Séparation harmonique par banc de filtre à deux canaux.

Dans une première étape, le signal d'entrée est pré-traité à l'aide d'un filtre de blanchiment, afin de supprimer la corrélation court-terme du signal. Par la suite, le signal résultant est séparé en deux canaux par filtrage en dessous et au dessus

de 1 Khz. La détection périodique est basée sur "l'autocorrélation généralisée", qui consiste à calculer la transformée de Fourier Discrète du signal, suivie d'une compression d'amplitude de la représentation spectrale, puis d'un transformée inverse.
La méthode "d'autocorrélation généralisée" a l'avantage de pouvoir être calculée par FFT et est donc, d'un point de vue de temps de calcul, beaucoup plus rapide que l'autocorrélation classique.
Les fonctions d'autocorrélation de chaque canal sont finalement sommées pour former une fonction appellée SACF (Summed Autocorrelation Function). Parce que la SACF obtenue génère des pics à toutes les fréquences multiples de la fréquence fondamentale qui ne sont pas toujours bien résolus, Tolonen cherche à améliorer cette dernière en utilisant une méthode développée par Ellis [26]. L'estimation de fréquence fondamentale est alors effectuée à partir de cette SACF affinée (ESACF), dans laquelle les pics correspondant aux harmoniques sont supprimés.
D'un point de vue expérimental, Tolonen montre que la fonction d'autocorrélation de son modèle est équivalente à celle découlant du modèle de Meddis et O'Mard : il obtient les mêmes informations de périodicité avec une modélisation considérablement simplifiée de la perception de pitch.
Appliqué à des combinaisons sonores de deux ou trois signaux, sa méthode de séparation permet de résoudre assez facilement les sources, si les niveaux d'amplitude ne sont pas trop différents. Pour les accords avec 4 sources ou plus, des signaux vont être masqués par d'autre et les sources correspondantes ne seront pas résolues. Pour ce dernier cas, Tolonen évoque l'idée d'utiliser un algorithme de soustraction itératif afin de supprimer au fur et à mesure les sources déjà séparées.

- **Modèle de perception sonore**

Une autre approche, basée sur le principe du traitement par bandes de fréquences a été proposée par Klapuri et Virtanen [27] [28]. Cette méthode de séparation utilise le travail initial de klapuri, qui s'est longuement intéressé au problème d'estimation de fréquence fondamentale pour les sons naturels [29] [30] [31] : il a ainsi développé une méthode d'estimation capable de déterminer la fréquence fondamentale prédominant dans une polyphonie contenant jusqu'à six voix. Cette méthode s'appuie sur une version très simplifiée du modèle de perception sonore proposé par Moore et qui consiste à modéliser l'intensité sonore perçue par notre système auditif [32]. Il implante alors un algorithme d'estimation de fréquences fondamentales multiples qui associe à chaque fréquence candidate une vraisemblance. L'estimation est alors effectuée par maximisation de cette vraisemblance. Outre ses performances, cette méthode d'estimation permet également de mettre

en évidence le phénomème d'inharmonicité des sons naturels et offre ainsi une localisation plus réaliste des partiels d'un son. L'inconvénient majeur de cet estimateur réside dans le fait qu'il utilise des fenêtres d'analyse relativement longues et donc peu réalistes pour l'étude des signaux musicaux. Ces dernières sont de l'ordre de 200 ms.

Klapuri et Virtanen étendent cette méthode au problème de l'estimation de fréquences fondamentales multiples [28] et développent une méthode de séparation de signaux musicaux polyphoniques (Fig. 1.3). Le système utilise deux propriétés différentes des sons : leur structure harmonique pour l'estimation fréquentielle et la continuité de l'enveloppe spectrale pour l'estimation des amplitudes.
Le signal d'entrée est passé à travers un détecteur multipitch, basé sur la méthode d'estimation évoquée précédemment, qui permet d'estimer le nombre de sources et les fréquences de leurs composantes. Ces estimations supposées correctes sont utilisées comme le point de départ d'un système d'analyse sinusoïdale qui estime les fréquences et amplitudes du modèle de manière itérative en utilisant des modèles d'enveloppes spectrales. La raison de cette approche est liée d'une part, au problème de recouvrement des harmoniques de plusieurs sources, qui se traduit dans le spectre par la présence de pics de forte amplitude. D'autre part, elle vise à corriger les estimations imprécises obtenues par l'estimateur de fréquences fondamentales multiples appliqué à des trames d'analyse très longues.
Pour une itération donnée, le spectre de chaque signal source est lissé tout en respectant la continuité de l'enveloppe spectrale d'un son naturel. La précision des estimations pour les fréquences et amplitudes est alors obtenue grâce à une analyse itérative effectuée sur des trames courtes. Ces dernières étant déterminées par la méthode des moindres carrées. Lorsqu'il y a recouvrement d'harmoniques, la valeur pour les amplitudes est écartée car fausse.
Dans une première étude [27], Klapuri et Virtanen proposent de lisser le spectre d'amplitude sur une échelle fréquentielle de type bande critique détaillée dans [33], puis proposent par la suite d'utiliser différents types de modélisation linéaire du vecteur d'amplitudes. Les modèles sont alors construits en introduisant des transformations matricielles. Celles-ci permettent de passer d'un espace sous-dimensionné (car les amplitudes estimées pour des harmoniques se recouvrant ont été précédemment écartées) à un espace de dimension K_n (égal au nombre d'harmoniques du n^{ieme} signal source).
Ainsi dans [28], plusieurs matrices sont comparées : la matrice identité, une matrice banc de filtres, une matrice polynomiale et une matrice de type transformée de Fourier.
Appliqué à des polyphonies contenant deux à six voix, dont les pitch varient entre 65 et 2100 Hz, ils obtiennent par la première méthode de lissage proposée, une bonne qualité sonore pour les signaux synthétisés après séparation. La comparaison des modèles d'enveloppe spectrale améliorent ces résultats pour les polyphonies les plus complexes, sauf dans le cas de la transformation polynomiale.

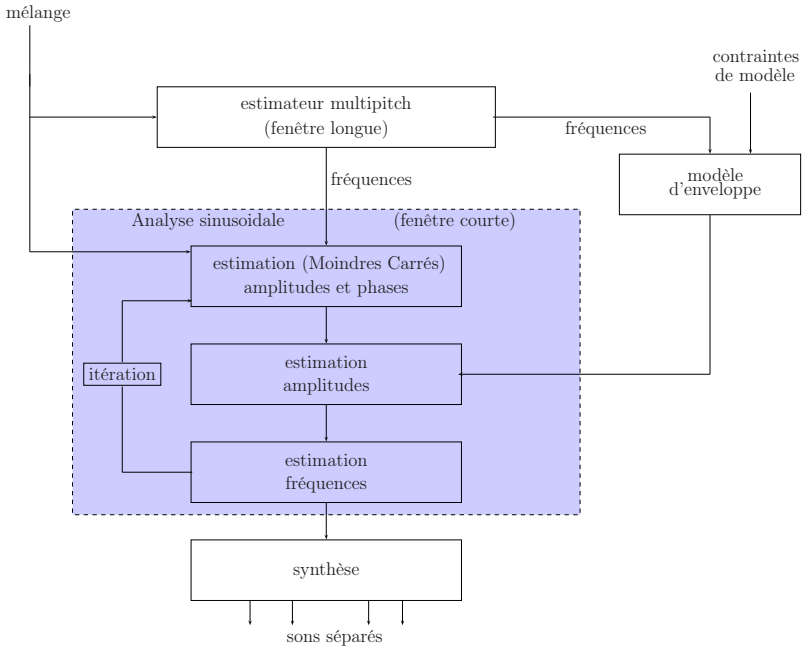


Fig. 1.3 – *Méthode de séparation développée par Klapuri et Virtanen.*

1.4 Des techniques pour améliorer les estimations

Les difficultés rencontrées pour l'estimation de plusieurs fréquences fondamentales sont à l'origine de diverses techniques dont le but est d'améliorer ces estimations. Ces difficultés sont en général caractéristiques des méthodes d'estimation de type itératif dont les performances se dégradent avec les estimations successives. Les améliorations sont faites à deux niveaux : les premières techniques présentées ci-dessous sont appliquées au sein d'une trame d'analyse et consistent en général à ajuster des modèles de spectres. Les secondes, quant à elles, utilisent l'information passée et future pour améliorer les estimations obtenues à un temps donné. On parle alors de techniques de poursuite de paramètres.

1.4.1 Les techniques utilisées pour pallier les problèmes de résolution

Lorsque les paramètres sont estimés à l'aide de la transformée de Fourier à court terme [20] [27], les limites dues à la résolution fréquentielle ne permettent pas toujours de séparer les harmoniques des sources d'un mélange : des partiels appartenant à des sources différentes sont susceptibles de se recouvrir, produisant ainsi des distorsions non négligeables au niveau des pics spectraux. Ainsi lorsque deux composantes spectrales sont proches, la plus importante peut par exemple cacher la seconde, impliquant des amplitudes et des phases estimées erronées, qui sont par la suite utilisées dans la modélisation sinusoïdale des sources.

Plusieurs techniques ont été proposées pour estimer les amplitudes et les phases du modèle sinusoïdal. Toutes basées sur l'ajustement spectral entre le modèle et les observations, elles diffèrent par la longueur de l'intervalle fréquentiel sur lequel les recouvrements éventuels sont pris en compte.
Ainsi, pour le i^{ieme} modèle d'un mélange, exprimé à l'aide de ses phases $\{\phi_k^i\}_{k=1,\cdots,M_i}$ tel que :

$$s_i(n) = \sum_{k=1}^{M_i} \alpha_k^i cos(2\pi f_k^i n + \phi_k^i) \tag{1.32}$$

l'analyse à court terme du mélange de Ns sources, à l'aide d'une fenêtre $w(n)$ est telle que :

$$S(f) = \sum_{i=1}^{Ns} \sum_{k=1}^{M_i} \frac{1}{2} \alpha_k^i (e^{j\,\phi_k^i} W(f - f_k^i) + e^{-j\,\phi_k^i} W(f + f_k^i)) \tag{1.33}$$

La transformée de Fourier S(f) du mélange est une somme de versions décalées de la transformée de Fourier $W(f)$ de la fenêtre d'analyse aux fréquences f_k^i.

Ajustement spectral : L'ajustement spectral consiste à minimiser l'erreur quadratique moyenne entre le spectre du modèle S(f) donné par (eq. 1.33) et celui des observations $X(f)$.
En posant $p_k^i = \frac{1}{2}\alpha_k^i cos(\phi_k^i)$ et $p_{(k+M_i)}^i = \frac{1}{2}\alpha_k^i sin(\phi_k^i)$ le modèle de spectre s'écrit :

$$\widehat{S}(f) = \sum_{i=1}^{Ns} \sum_{k=1}^{M_i} [p_k^i R_k^i(f) + p_{(k+M_i)}^i R_{k+M_i}^i(f)] \tag{1.34}$$

les expressions de $R_k^i(f)$ et $R_{(k+M_i)}^i(f)$ étant donnée par :

$$\begin{cases} R_k^i(f) = W(f - f_k^i) + W(f + f_k^i) \\ \\ R_{(k+M_i)}^i(f) = j\,[W(f - f_k^i) - W(f + f_k^i)] \end{cases} \tag{1.35}$$

La solution moindres carrés pour les phases et les amplitudes est alors donnée sous forme vectorielle par :

$$\underline{p} = (H^T H)^{-1} H^T S \tag{1.36}$$

$\underline{p}$ est le vecteur contenant les amplitudes et phases du modèle, H est une matrice dont les vecteurs colonnes sont les R_k^i aux fréquences f_k^i.

• Interpolation multitrame

Dans leur approche du problème de séparation de deux sources de parole voisée, Quatieri et Danisewicz montrent que ce problème de minimisation est équivalent à la résolution du système d'équations :

$$\begin{cases} \widetilde{H}\underline{\alpha} = 2\,\Re[\underline{X}(\underline{f})] \\ \\ \widetilde{H}\underline{\beta} = -2\,\Im[\underline{X}(\underline{f})] \end{cases} \tag{1.37}$$

$\underline{f} = [f_1, f_2, \cdots, f_M]$ est le vecteur contenant l'ensemble des $M = M_1 + M_2$ fréquences du mélange, rangées par ordre croissant.

L'expression de la matrice, lorsque les contributions des fréquences négatives sont omises est donnée par :

$$\widetilde{H} = \begin{bmatrix} W(0) & W(f_1 - f_2) & \cdots & W(f_1 - f_M) \\ W(f_2 - f_1) & W(0) & \cdots & W(f_2 - f_M) \\ \vdots & \vdots & \vdots & \vdots \\ W(f_M - f_1) & W(f_M - f_2) & \cdots & W(0) \end{bmatrix} \tag{1.38}$$

Le problème de la proximité des fréquences se traduit par un mauvais conditionnement de la matrice $\widetilde{H}$. Il en résulte que l'inversion de cette matrice, nécessaire à la résolution de (*eq.*1.37) devient instable.
Mac Aulay et Danisewicz proposent alors de détecter les recouvrements éventuels en imposant une distance fréquentielle minimale entre deux partiels, en dessous

de laquelle une unique sinusoïde sera utilisée pour représenter les deux voix.
Cette limite étant fixée à 25 Hz, l'estimation des amplitudes et des phases correspondantes pour chaque signal source est par la suite obtenue à l'aide d'une "interpolation multi-trame". Cette dernière utilise les fonctions d'interpolation proposées par Mac Aulay et Quatieri dans la synthèse sinusoïdale d'un signal de parole [13] : Si la k^{ime} trame contient des fréquences difficiles à résoudre, les amplitudes et les phases correspondantes sont interpolées entre la trame $(k-q)$ et la trame $(k+p)$ où p et q sont choisis de telle sorte que leurs paramètres correspondants sont résolus. Les fonctions d'interpolation utilisées pour les amplitudes et les phases sont respectivement de type linéaire et cubique. Ces dernières sont présentées par la suite, dans le paragraphe (eq. 1.5.1) consacré à la synthèse harmonique.

• **Analyse "Stationary Least squares"**

Initialement développée dans un contexte de codage de parole, une autre méthode qui cherche à prendre en compte le phénomène d'interférence harmonique pour l'estimation des amplitudes et phases du modèle sinusoïdal fût proposée par Silva et Almeida [34] : Pour des fréquences harmoniques connues, ils développent une procédure d'analyse appelée "Stationary Least squares" (SLS). Cette dernière résout un système d'équations linéaires faisant intervenir des termes de corrélation entre spectre du signal composite et transformée de la fenêtre d'analyse.
Afin de réduire fortement le coût de calcul, Silva et Almeida diagonalisent la matrice du système en négligeant les interférences entre harmoniques séparées d'une valeur supérieure à un certain seuil fixé.
Ils appliquent finalement cette méthode à la séparation de deux locuteurs :
les interférences harmoniques pour une différence fréquentielle supérieures à 300Hz sont négligées. Pour les harmoniques très proches ($\leq 10Hz$), les paramètres sont estimés comme si une unique harmonique était présente, puis l'amplitude de cette harmonique pour chacune des sources, est déterminée à partir des estimations faites sur les trames précédentes.
Des tests d'écoute informels montrent alors une amélioration de l'intelligibilité pour les signaux ainsi synthétisés. Elle se traduit notamment par le fait que le signal interférant est complètement inaudible dans l'écoute du signal cible et seules des distorsions de caractère tonal sont finalement audibles .

• **Couplage de trois stratégies de séparation**

Toujours dans le cas de deux signaux voisés, Maher a développé dans [12] trois stratégies de séparation des paramètres, qu'il utilise de façon complémentaire . Toutes visent à déterminer la contribution la plus probable de chaque source à partir du spectre court terme du signal composite.

Ainsi après avoir estimé les fréquences fondamentales de chaque voix par la méthode du Two-Way-Mismatch (cf. paragraphe (eq. 1.3.1)), les partiels susceptibles d'être "en collision" avec des partiels du second signal sont repérés. L'espace minimum entre partiels adjacents est déterminé, puis les différentes stratégies sont utilisées comme suit :

– si un partiel est à au moins 50 hz de tout autre partiel, la composante correspondante est supposée "non-corrompue".

– si deux partiels sont séparés de moins de 50 Hz, mais l'écart entre les deux reste supérieur à 25 Hz, une première stratégie de séparation est appliquée. Elle s'appuie sur la résolution d'un jeu d'équations linéaires similaire à celui de Quatieri et Danisewicz. La différence est que la matrice $\widetilde{H}$ utilisée par Maher dans le système (eq. 1.37), inclue seulement l'effet des transformées W(f) des fréquences les plus proches, plutôt que de tenir compte de l'ensemble des W(f).

– si les deux partiels sont séparés de moins de 25 Hz, une analyse modélisant le phénomène de battement est utilisée.

– cette seconde stratégie est limitée et n'est plus valable lorsque la collision traitée est inférieure à deux périodes de battement. Dans ce cas une troisième stratégie de séparation qui utilise des formes spectrales prédéfinies est appliquée. Elles correspondent à des modèles pré-calculés à partir de l'enveloppe spectrale d'instruments ou de voix.

• **Lissage spectral**

Une troisième méthode d'estimation des amplitudes et phases, au sens des moindres carrés a dernièrement été proposée par Virtuanen et Klapuri dans le cadre de la séparation de signaux musicaux [27]. Supposant que l'enveloppe spec-

trale des sources étudiées est continue, ces derniers proposent d'utiliser des modèles d'enveloppe pour obtenir les amplitudes des partiels corrompus :
Les composantes trop proches sont détectées en fixant une distance fréquentielle limite, puis les amplitudes des partiels se recouvrant sont lissées à l'aide de modèles proches du spectre idéal.

Dans une première étude [27], Klapuri et Virtanen proposent de lisser le spectre d'amplitude sur une échelle fréquentielle en bandes critiques détaillée dans [33]. Par la suite, ils proposent d'utiliser différents types de modélisation linéaire du vecteur d'amplitudes. Les modèles sont alors construits en introduisant des transformations matricielles. elles permettent de passer d'un espace sous-dimensionné (car les amplitudes estimées pour des harmoniques se recouvrant ont été précédemment écartées) à un espace de dimension K_n (égal au nombre d'harmoniques du n^{ime} signal source). Ainsi dans [28], différentes matrices ont été comparées : la matrice identité, une matrice banc de filtres, une matrice polynomiale et une matrice de type transformée de Fourier.
Appliqué à des polyphonies contenant de 2 à 6 voix, dont les fréquences fondamentales varient entre 65 et 2100 Hz, ils obtiennent par la première méthode de lissage proposée, une bonne qualité sonore des signaux synthétisés après séparation. La comparaison des modèles d'enveloppe spectrale améliorent ces résultats pour les polyphonies les plus complexes, sauf dans le cas de la transformation polynomiale.

1.4.2 Poursuites des paramètres harmoniques le long des trames

Les méthodes de poursuite des paramètres harmoniques sont utilisées pour relier les estimations des trames successives de signal. Elles permettent alors d'assigner à chaque source une trajectoire temporelle des paramètres d'un modèle.
La poursuite a pour but d'éliminer les changements brutaux au bord des fenêtres d'analyse, mais également d'interpoler les paramètres non résolus lors de l'analyse du mélange.

• Appariement des pics harmoniques

Une méthode de poursuite bien connue et utilisée pour l'Analyse/synthèse sinusoïdale a été développée par Mac Aulay et Quatieri [13]. Cette dernière cherche

à apparier les pics spectraux de trames adjacentes : pour les M^k fréquences fondamentales estimées de la trame k, la procédure de poursuite, consiste à assigner à chaque fréquence de la trame k, l'une des fréquences de la trame $k+1$.

• prédiction linéaire de la fréquence fondamentale

Cette méthode a été proposé par Parsons en 1976 pour la séparation de deux locuteurs [3]. Elle s'appuie sur l'hypothèse que pour des signaux de parole, la fréquence fondamentale varie peu entre deux trames consécutives. Ainsi pour la k^{ieme} trame courante, une procédure en deux étapes est utilisée pour l'appariement :

1. les deux fréquences fondamentales $\widehat{f}_1^k$ et $\widehat{f}_2^k$ estimées sont assignées à l'une des deux voix, à l'aide de fréquences prédites sur la $(k-1)^{ime}$ trame, notées $\widetilde{f}_1^{k-1}$ et $\widetilde{f}_2^{k-1}$.

2. à partir des fréquences fondamentales assignées, deux nouvelles prédictions $\widetilde{f}_1^k$ et $\widetilde{f}_2^k$ sont calculées.

L' étape 1 qui fait alors intervenir 2 prédictions et 2 estimations, correspond au calcul de quatre distances entre fréquences estimées et fréquences prédites sur la trame précédente. L'affectation des fréquences estimées à une voix se fait en considérant la distance minimale. Le prédicteur utilisé est un filtre à mémoire fixe qui ajuste une fonction affine aux trois plus récentes valeurs de fondamentales et donne une prédiction de la fréquence par extrapolation.

• Poursuite par la methode cepstrale ou la corrélation spectrale

Cette autre méthode de poursuite développée dans le cadre de la séparation de deux signaux de parole voisée, a été proposée par Mumolo [11]. Pour chaque segment de signal, un nombre de fréquences fondamentales candidates est estimé. Le but de la poursuite fréquentielle est alors de sélectionner parmi ces valeurs celles qui représentent les fréquences fondamentales des deux locuteurs et de déterminer les contours ou trajectoires temporelles correspondantes. L'algorithme de poursuite proposé est composé de 5 étapes :

1. Détermination de points fixes : à partir des fréquences fondamentales estimées sur toute la durée d'analyse, les plus grandes valeurs des pics correspondant sont sélectionnées comme points d'ancrage.

2. Normalisation : parce que les amplitudes des pics peuvent varier fortement pour plusieurs raisons, l'algorithme normalise les amplitudes et on a pour le pic n d'amplitude $a(n)$:

$$a(n) = \frac{a(n)}{0.2 \sum_{i=n+1}^{n+5} a(i)} \tag{1.39}$$

3. Elimination des harmoniques : Pour chaque trame, le pic de plus grande amplitude normalisée est trouvé et ses multiples sont marqués.

4. Poursuite : en commençant par les points fixes, une poursuite de fréquence de fondamentale est appliquée à droite et à gauche de la trame courante, en choisissant la valeur de la fréquence fondamentale dont l'amplitude normalisée est la plus proche de celle de la trame précédente.

5. Répétition : On recommence à partir de l'étape 4) jusqu'à ce qu'un critère choisi soit satisfait.

Le critère correspond à une valeur de seuil qui stoppe l'algorithme lorsque l'amplitude du pic se trouve en dessous. Il se rapproche d' une mesure de voisement.

• Utilisation de critères psychoacoustiques

Virtanen et Klapuri ont développé une approche quelque peu différente, qui cherche à décomposer les composantes spectrales du mélange pour les assigner aux sources à l'aide de critères psychoacoustiques [35]. Ils utilisent la modélisation sinusoïdale pour représenter les sources par des trajectoires sinusoïdales. Les fréquences et amplitudes des sinusoïdes sont obtenues avec une analyse de Fourier à court terme et une méthode statistique initialement développée par Thomsom appelée "F-test" [36]. Les interruptions courtes dans les trajectoires, causées par des modulations d'amplitude et de fréquence, des transitoires ou du bruit sont supprimées par interpolation linéaire.
Le système effectue une classification des trajectoires en calculant la distance "perceptive" entre paires à l'aide de deux critères psychoacoustiques différents :

les changements synchrones des composantes et la concordance harmonique.
Ces critères, introduit par Bregman, sont liés aux principes d'association dans l'organisation de l'audition humaine [37].
La distance "perceptive" utilisée pour la classification est alors constituée de trois distances pondérées. Les deux premières permettent de mesurer les changements de fréquence et d'amplitude, tandis que la troisième effectue une mesure de la relation harmonique entre trajectoires. Les pondérations sont différentes et sont choisies en tenant compte du fait que les fréquences ne vont pas varier autant que les amplitudes et que l'organisation perceptive est largement basée sur la distance harmonique.
Après estimation de la proximité entre paires de trajectoires, les sources sont séparées en utilisant une méthode de classification qui, à partir de deux jeux initiaux représentant les sources, va ajouter les trajectoires une à une en choisissant la trajectoire minimisant la distance.
Cette méthode de classification reste très sommaire, car elle nécessite des signaux sources ayant des temps d'attaque différents pour permettre l'initialisation des jeux de départ.
Il est à noter que cette méthode prend en compte les recouvrements d'harmoniques, qui se traduisent par des trajectoires commune aux deux sources. Le phénomène est détecté en utilisant la distance harmonique, qui est dans ce cas très faible pour les deux son. Le système utilise alors une valeur seuil et interpole les amplitudes des trajectoires communes à l'aide des autres trajectoires.

1.5 Des techniques pour la synthèse des sources

Bien que la synthèse n'est pas l'unique finalité du problème de séparation, comme nous l'avons évoqué au début de ce chapitre, nous lui consacrons une partie de notre état de l'art, d'une part parce qu'elle est intervenue dès les premières méthodes proposées en parole et d'autre part parce qu'elle est étroitement liée aux méthodes d'estimation itératives de fréquences fondamentales, qui font appel à une étape de soustraction des paramètres ou des modèles de sources successivement estimés.
Nous présentons, tout d'abord deux méthodes courantes pour la synthèse de signaux harmoniques, puis nous décrirons des techniques qui tout comme la synthèse, permettent d'extraire du mélange une source dans le but bien précis d'une écoute. L'extraction se fait alors par atténuation des signaux indésirables et par rehaussement des signaux cibles.

1.5.1 Méthodes de synthèse

Il existe plusieurs méthodes permettant la synthèse de ces signaux. Quelque soit le type de synthèse utilisé, il est nécessaire de pallier les discontinuités aux frontières des fenêtres analysées qui ont tendance à dégrader fortement la qualité du signal synthétisé.
Pour cela, Mac Aulay et Quatieri proposent d'effectuer une interpolation des paramètres. Une seconde solution tout aussi connue, est la synthèse de type Overlap-Add qui effectue un "fondu" entre les modèles de fenêtres successives pour assurer la continuité des estimations. Ces deux méthodes sont présentées ci-dessous.

• Synthèse "Overlap Add" (OLA)

De la manière la plus générale, la parole peut être synthétisée par addition/recouvrement des différents modèles $\widetilde{s_k}$ obtenus sur chaque trame analysée à l'aide de l'expression :

$$\widetilde{s}(n) = \sum_{-\infty}^{+\infty} w(n - pR)\widetilde{s_k}(n - pR) \tag{1.40}$$

$\widetilde{s}$ est le signal synthétisé, p l'index de la trame analysée et R le pas de l'analyse. La fonction $w(n)$ est une fenêtre de synthèse qui obéit à la contrainte $\sum_{-\infty}^{+\infty} w(n - pR) = 1$.

Lorsque le signal est analysé à l'aide de la transformée de Fourier à court terme, les modèles s_k sont obtenus par transformation inverse et on a :

$$s(n) = \sum_{-\infty}^{+\infty} w(n - pR)\frac{1}{N}\sum_{k=0}^{N-1} \widetilde{S}(p, \omega_k)e^{j\frac{2\pi}{N}kn} \tag{1.41}$$

où $\widetilde{S}(p, \omega_k)$ est la transformée de Fourier de $\widetilde{s_k}$ à la fréquence ω_k.

• Synthèse harmonique

Ce type de synthèse est basé sur la représentation sinusoïdale et a été initialement développé par Mac Aulay et Quatieri pour la parole voisée.
Le modèle de la k^{ieme} trame du signal à synthétiser est donné par :

$$\widetilde{s}_k(n) = \sum_{l=1}^{L(k)} \widehat{A}_l^k cos(2\pi n \widehat{f}_l^k + \widehat{\theta}_l^k) \tag{1.42}$$

Les amplitudes $\widehat{A}_l^k$ sont alors interpolées linéairement telles que :

$$\widetilde{A}_l = A_l^k + \frac{A_l^{k+1} - A_l^k}{T} n \tag{1.43}$$

où T est le nombre d'échantillons séparant les instants t_k et t_{k+1}.

La continuité des phases se fait par une interpolation cubique, qui permet de prendre également en compte la continuité des fréquences :

$$\widetilde{\Psi}(t) = a + bt + ct^2 + dt^3 \tag{1.44}$$

les contraintes de continuité étant données par :

$$\begin{cases} \widetilde{\Psi}(0) = \theta^k \text{ et} & \frac{d\widetilde{\Psi}}{dt}(0) = \omega^k \\ \widetilde{\Psi}(T) = \theta^{k+1} + 2D\pi \text{ et} & \frac{d\widetilde{\Psi}}{dt}(T) = \omega^{k+1} \end{cases} \tag{1.45}$$

Le paramètre D est le déroulement de phase que l'on détermine en minimisant la fonction :

$$f(D) = \int_0^T (\frac{d^2\widetilde{\Psi}(t)}{dt^2})^2 dt \tag{1.46}$$

Synthèse harmonique + bruit : Parce que les méthodes de synthèse sont étroitement liées à l'analyse du signal, l'analyse de Quatieri et Mac Aulay développée pour la modélisation seule des parties voisées de la parole, ne permet pas de reconstruire les parties non voisées qui sont indispensables à une synthèse de qualité.
Pour cette raison et plus particulièrement à cause de la généralisation de l'Analyse/synthèse harmonique au cas des signaux musicaux, une représentation plus réaliste a été développée par Serra [38]. Le signal audio est alors représenté par un modèle dit "sinusoïdes plus bruit". Le bruit permet alors d'inclure les parties non-voisées du signal de parole, l'attaque de certains sons musicaux ou encore le souffle d'une flûte ou le frottement de l'archet d'un violon.

La partie sinusoïdale du ce modèle, correspond également à une interpolation des paramètres de trame, mais dans laquelle l'information de phase n'est plus

prise en compte, des tests d'écoute ayant montré le faible d'impact de cette dernière sur la qualité de la synthèse. En revanche, si l'on souhaite obtenir le résiduel d'un signal après synthèse, il est nécessaire de prendre en compte la phase lorsque les sinusoïdes sont retirées. Cette dernière est obtenue par simple intégration :

$$\widetilde{\Psi}_l(t) = \widetilde{\Psi}_l(t_k) + \int_{t_k}^{t_{k+1}} \widetilde{\Psi}_l(u - t_k)du \tag{1.47}$$

Les fréquences sont interpolées telles que :

$$\widetilde{\omega_l^k} + \frac{t}{T}(\omega_l^{k+1} - \omega_l^k) \tag{1.48}$$

et le signal synthétisé est finalement donné par $\widetilde{s}(t) = cos(\widetilde{\Psi}_l(t))$.

La synthèse de la partie bruitée est quant à elle, obtenue en générant un bruit de densité spectrale donnée dans chaque fenêtre de synthèse, puis par Overlap Add on obtient le signal stochastique final qui doit être ajouté à la partie déterministe sinusoïdale.

1.5.2 Techniques de séparation

Nous présentons ci-dessous deux techniques de séparation qui permettent la "synthèse" de sources d'un mélange par atténuation des sources indésirables ou amélioration des sources cibles. Toutes deux ont initialement été développées pour la parole. La première technique s'applique dans le domaine spectral et est à l'origine d'un nombre important de travaux de recherche, la seconde est une technique temporelle.

• La soustraction spectrale :

Le principe de la soustraction spectrale pour le problème de séparation fût initialement proposé par Hanson et Wong [39]. S'intéressant plus particulièrement à l'amélioration de l'intelligibilité d'un signal de parole noyé dans un autre beaucoup plus fort, ils développèrent un système nommé HMS (Harmonic Magnitude Suppression) qui permet de supprimer dans le domaine spectral le signal

interférant. Le problème de la séparation des deux sources est ainsi abordé en termes de signal désiré et signal interférant et pour lequel le bruit à supprimer est un signal de parole.

Appliqué à des rapports de puissance entre signal désiré et signal interférant né-

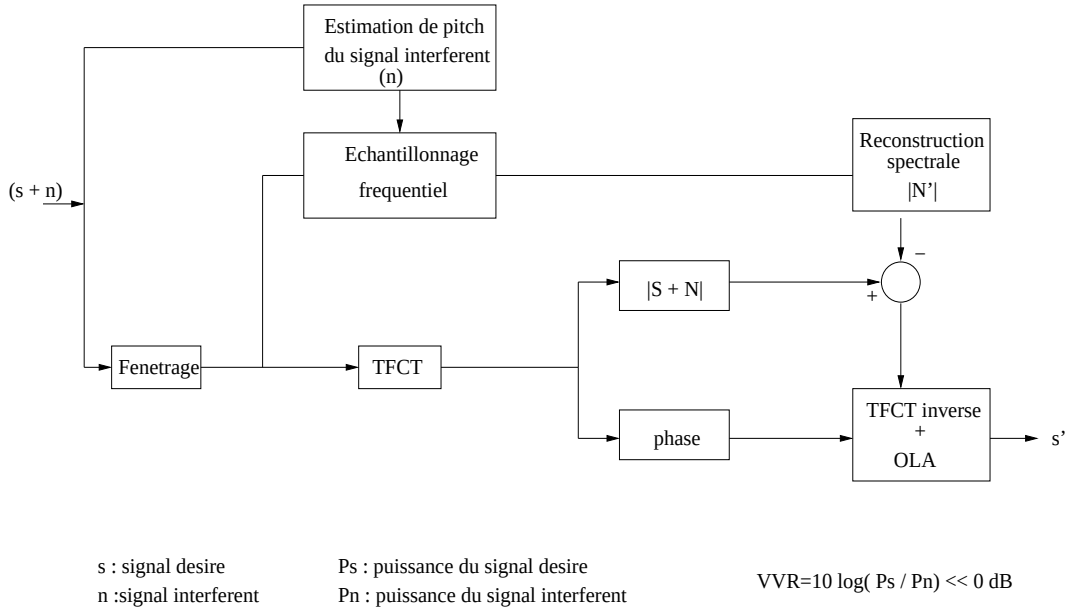


Fig. 1.4 – Principe de la séparation par soustraction spectrale

gatifs, cette technique cherche à soustraire le spectre de bruit a celui du signal composite. L'estimation du spectre de bruit est obtenue par échantillonnage fréquentiel, qui dans sa version initiale utilise une fréquence fondamentale connue "à priori".

Amélioration du système HMS : Il a fallu attendre le travail de Naylor et Boll [40], pour obtenir une méthode HMS plus réaliste dans laquelle tous les paramètres du modèle sont estimés à partir du signal du parole composite : un algorithme d'estimation de l'état de source (voisé, non-voisé, silence) est intégré, puis la fréquence fondamentale des parties voisées est extraite à l'aide d'un estimateur de type Maximum de Vraisemblance.
Cherchant toujours à améliorer le système HMS, Childers et Lee couplent la soustraction spectrale avec une analyse MCESA multisignal (Minimum Cross Entropy Analysis) [10]. Ce type d'analyse a alors pour but d'estimer le spectre d'un signal

à partir d'un nombre donné de valeurs de la fonction d'autocorrélation.
Une dernière version améliorée du système HMS fut proposée dans les années 90 par Naylor et Porter [8]. Pour effectuer la séparation, ils introduisent un modèle de spectre complexe de meilleure qualité. Un algorithme de Covariance Modifiée (MC Algorithm [9]) est également développé pour estimer la fréquence fondamentale des deux locuteurs.

Le système HES (Harmonic Enhancement Suppression) : Le système HES, développé dans [18] est peu différent du système HMS. La différence réside dans le fait que l'on vise à retrouver les deux locuteurs. Le signal de parole du locuteur le plus fort est alors obtenu en améliorant ses fréquences harmoniques et ses formants. L'autre signal de parole, correspondant au locuteur le plus faible, est obtenu par suppression de ses harmoniques et formants.

• **Séparation temporelle par filtrage en peigne**

Ce type de séparation a principalement été étudié par De cheveigné [23] [15]. Tout comme précédemment, les deux stratégies qui consistent à rehausser le signal cible et à annuler le signal interférant, qui peuvent de plus être utilisées de façon complémentaire, sont développées dans le domaine. Toutes deux utilisent la connaissance de la période fondamentale du signal cible ou interférant.

Le rehaussement est effectué à l'aide d'un filtre de réponse impulsionnelle dépendant de la période fondamentale T_0 du signal désiré :

$$h_r(t) = \frac{1}{N} \sum_{n=0}^{N-1} \delta(t - nT_0) \tag{1.49}$$

Ce qui, dans le domaine fréquentiel va correspondre à la sélection des harmoniques du signal cible par rapport aux autres.

Lorsqu'on choisit d'annuler un signal interférant, le filtre utilisé est le filtre Comb donné par (eq. 1.13).
C'est à partir de cette stratégie d'annulation que De Cheveigné a pu développer la méthode d'estimation de fréquences fondamentales utilisant la DDF (eq. 1.14) et qui peut être vue comme à l'origine d'une séparation par estimation-annulation conjointe.

1.6 Objectifs et description des méthodes d'estimations proposées

Nous consacrons pour finir un court paragraphe dans lequel nous définissons le cadre d'étude dans lequel ce travail de thèse s'inscrit, nous précisons les objectifs fixés et nous présentons les méthodes proposées.

1.6.1 Le cadre d'étude et les objectifs fixés

Dans le travail qui suit, nous nous intéressons au problème de séparation des signaux de parole et de musique. Les mélanges de parole et les mélanges musicaux sont étudiés séparément, mais proviennent dans les deux cas d'enregistrements de type monophonique. On considère également qu'aucune information sur le nombre de sources contenues dans le mélange n'est disponible.
Pour chaque type de mélanges considéré, nous proposons une méthode d'estimation de fréquences fondamentales multiples. Ces dernières sont développées dans le cadre de la séparation et ont pour but d'exploiter au mieux le caractère harmonique des sources des mélanges.

1.6.2 Description des méthodes d'estimation proposées

Une première partie sera donc consacrée aux mélanges de type parole. L'approche proposée est basée sur un modèle "somme de sinusoïdes + bruit autorégressif", qui cherche à prendre en compte la nature mixte des mélanges contenant des sources voisées et non-voisées. Les fréquences fondamentales sont déterminées itérativement par une méthode d'estimation utilisant un critère de sélection d'ordre. La seconde partie sera consacrée aux mélanges musicaux. Les méthodes d'estimation de fréquences fondamentales multiples proposées n'ont aucun rapport direct avec celle développée dans le cadre de la parole : ce sont des méthodes fréquentielles non paramètriques qui permettent d'estimer conjointement les fréquences fondamentales du mélange. Ces estimations sont alors obtenus par classification des pics spectraux du mélange étudié.
Les performances d'estimation de l'ensemble des méthodes proposées, sont évaluées par le calcul de pourcentages de fréquences fondamentales correctement estimées et par le nombre estimé de sources d'un mélange.

Deuxième partie

Estimation de fréquences fondamentales pour la parole

Chapitre 2

Estimation par le principe du critère d'ordre

Les nombreuses études sur l'analyse/synthèse de la parole ont permis de mettre en évidence l'efficacité de la modélisation sinusoïdale pour les parties voisées. Toutefois, la nécessité de modéliser également les parties non-voisées, pour une bonne intelligibilité de la parole, nous incite à penser qu'un modèle mixte du type "somme de sinusoïdes harmoniques + processus autorégressif" est un outil très approprié au traitement des signaux de parole. Dans le cadre de la séparation, cette modélisation est d'autant plus justifiée, en remarquant que chaque signal source pris indépendamment, peut être vue comme un signal bruité. Le bruit étant alors constitué par la superposition des autres sources présentes.
Cette idée d'appliquer un modèle mixte aux signaux de parole a initialement été évoquée par Kay et Nagesha. Ils se sont plus particulièrement intéressés au problème d'extraction d'un signal harmonique noyé dans du bruit coloré [41]. Le modèle mixte qu'ils définissent est plus connu sous le nom de modèle CARD.
Les performances de la méthode qu'ils proposent pour estimer la fréquence fondamentale, nous incitent à penser que cette méthode est exploitable pour l'estimation de fréquences fondamentales multiples.
Nous développons dans ce chapitre une nouvelle méthode d'estimation de la fréquence fondamentale, dérivant du travail de Kay et Nagesha. Cette dernière, intitulée *Méthode d'estimation du critère d'ordre*, est à la base d'une méthode d'estimation de fréquences fondamentales multiples itérative présentée dans le chapitre suivant.
La première partie de ce chapitre est consacrée à la modélisation CARD, puis la méthode d'estimation de la fréquence fondamentale basée sur le principe du critère d'ordre est présentée.

2.1 Modélisation du signal de parole

2.1.1 Le modèle CARD

Le modèle CARD ("Canonical AutoRegessive Decomposition"), proposé par Kay et Nagesha est défini pour N observations $\{y(0), \cdots, y(N-1)\}$ tel que [42] [43] :

$$y(n) = s(n) + v(n) \tag{2.1}$$

y(n) est la somme d'un signal sinusoïdal s(n) composé de M harmoniques et d'un processus autorégressif v(n) d'ordre p :

$$\begin{cases} s(n) = \sum_{k=1}^{M} (A_k cos(2\pi f_k n) + B_k sin(2\pi f_k n)) \\ \\ v(n) = -\sum_{m=1}^{p} a_m v(n-m) + u(n) \end{cases} \tag{2.2}$$

u(n) est un bruit blanc gaussien, de variance σ^2 et non corrélé avec les amplitudes $\{A_k, B_k\}_{k=1}^{M}$ des fréquences $\{f_k\}_{k=1}^{M}$. Les $\{a_m\}_{m=1}^{p}$ sont les p coefficients autorégressifs de *v(n)*.

2.1.2 Estimation des paramètres du modèle CARD par la méthode du Maximum de Vraisemblance

Le Maximum de Vraisemblance est une méthode permettant d'estimer les paramètres de la modélisation CARD et qui consiste à maximiser la vraisemblance du modèle. L'expression de cette vraisemblance est déterminée à partir de la densité de probabilité du modèle.
On trouvera dans l'annexe A, le détail des calculs permettant d'obtenir l'expression de la vraisemblance pour la modélisation CARD, et complétant ainsi la brève présentation qui suit :
Pour un nombre fini d'observations $\underline{\mathbf{y}} = [y(0)y(1)\cdots y(N-1)]^T$, la vraisemblance dans le cas d'une modélisation CARD s'écrit sous la forme
$p(\ \underline{\mathbf{y}}|\{a_m\}, \{A_k, B_k, f_k\}\ )$. Les estimés des différents paramètres s'obtiennent en maximisant la log-vraisemblance. On peut alors montrer que la maximisation est équivalente à la minimisation d'une fonction de coût $C(\{f_k\})$, qui ne dépend

explicitement que des fréquences $\{f_k\}_{k=1}^{M}$ et dont l'expression est donnée ci-après.

• Expression de la vraisemblance du modèle CARD

Pour un modèle CARD de paramètres $\{\{a_m\}, \{A_k, B_k, f_k\}\}$, l'expression de la vraisemblance est donnée par (cf. Annexe A) :

$$p(\ \underline{\mathbf{y}}|\{a_m\}, \{A_k, B_k, f_k\}\ ) \simeq \frac{1}{(\pi\sigma^2)^{(N-p)}}\ exp[-\frac{1}{\sigma^2}\sum_{n=p}^{N-1}|u(n)|^2] \tag{2.3}$$

$$u(n) = \sum_{m=0}^{p}\{a_m(y(n-m) - s(n-m)\} \tag{2.4}$$

• Estimation des fréquences par ML

L'estimation des fréquences $\{f_k\}_{k=1}^{M}$ du modèle est obtenue en minimisant la fonction de coût C d'expression :

$$C(\{f_k\}_{k=1}^{M}) = \underline{\mathbf{y}}_p[P_Y^{\dagger} - P_Y^{\dagger}\,\Omega(\Omega^H\,P_Y^{\dagger}\,\Omega)^{-1}\Omega^H P_Y^{\dagger}]\underline{\mathbf{y}}_p \tag{2.5}$$

Le termes matriciels définissant C ont pour expression :

$$\underline{\mathbf{y}}_p = [y(p)\,y(p+1)\cdots y(N-1)]^T \tag{2.6}$$

$$\mathbf{\Omega(f)} = \begin{bmatrix} cos(2\pi f_1(p)) & \cdots & cos(2\pi f_M(p)) & sin(2\pi f_1(p)) & \cdots & sin(2\pi f_M(p)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1(p+1)) & \cdots & cos(2\pi f_M(p+1)) & sin(2\pi f_1(p+1)) & \cdots & sin(2\pi f_M(p+1)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1(N-1)) & \cdots & cos(2\pi f_M(N-1)) & sin(2\pi f_1(N-1)) & \cdots & sin(2\pi f_M(N-1)) \end{bmatrix} \tag{2.7}$$

$P_Y^{\dagger} = I - Y(Y^HY)^{-1}Y^H$ est un opérateur de projection sur le sous-espace engendré par les vecteurs colonnes de la matrice $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} y(p-1) & y(p-2) & \cdots & y(0) \\ y(p) & y(p-1) & \cdots & y(1) \\ \vdots & \vdots & \ddots & \cdots \\ y(N-2) & y(N-3) & \cdots & y(N-1-p) \end{bmatrix} \tag{2.8}$$

Les fréquences estimées sont finalement données par :

$$\{\widehat{f}_1,\, \widehat{f}_2,\, \cdots,\, \widehat{f}_M\} = argmin_{\{f_1,\, f_2,\, \cdots,\, f_M\}}[C(\{f_k\}_{k=1}^{M})] \tag{2.9}$$

• Estimation des amplitudes et coefficients AR

Les expressions des amplitudes $\{(A_k, B_k)\}_{k=1}^{M}$ et des coefficients autorégressifs $\{a_m\}_{m=1}^{p}$ sont déduites de l'annulation des dérivées partielles de la vraisemblance (cf. Annexe A). Les amplitudes $\{A_k, B_k\}$ sont obtenues indirectement par l'introduction de paramètres c_k et s_k définis comme :

$$\begin{cases} c_k = -A_k \ \sum_{m=0}^{p} a_m cos(2\pi f_k m) + B_k \ \sum_{m=0}^{p} a_m sin(2\pi f_k m) \\ \\ s_k = -A_k \ \sum_{m=0}^{p} a_m sin(2\pi f_k m) - B_k \ \sum_{m=0}^{p} a_m cos(2\pi f_k m) \end{cases} \tag{2.10}$$

Les estimés $\{\widehat{A}_k, \widehat{B}_k\}$ sont alors déduits du vecteur $\underline{\widehat{\mu}} = [\underline{\widehat{c}}, \underline{\widehat{s}}]$, dont les M premiers paramètres correspondent au vecteur contenant les estimés des $\{c_k\}_{k=1,\cdots,M}$ et les M suivants à ceux des $\{s_k\}_{k=1,\cdots,M}$. L'expression de $\underline{\widehat{\mu}}$ est telle que :

$$\underline{\widehat{\mu}}(\mathbf{f}) = -(\,\Omega^H P_Y^{\dagger} \Omega\,)^{-1} \Omega\ P_Y^{\dagger}\ \underline{\mathbf{y}}_p \tag{2.11}$$

De même les estimés des coefficients AR sont obtenus à partir de $\underline{\widehat{\mu}}$:

$$\underline{\widehat{\mathbf{a}}}(\underline{\widehat{\mu}}, \mathbf{f}) = -(Y^H Y)^{-1} Y^H (\underline{\mathbf{y}}_p + \Omega\ \underline{\widehat{\mu}}) \tag{2.12}$$

où $\underline{\widehat{\mathbf{a}}}$ est le vecteur de paramètres contenant les p coefficients estimés $\{\widehat{a}_m\}$.

2.1.3 Le modèle CARD harmonique - Ambiguïté d'octave

• Hypothèse d'harmonicité

Parce que le signal de parole voisée est de nature dite "quasi-périodique", la modélisation sinusoïdale de la parole est très souvent supposée harmonique : les fréquences du modèle sont multiples entiers de la fréquence fondamentale et la fréquence du k^{ieme} partiel pour la fréquence fondamentale f_0 est donnée par :

$$f_k = k \times f_0 \tag{2.13}$$

Cette hypothèse, bien que restrictive, influe peu sur l'intelligibilité du signal lorsque ce dernier est synthétisé à partir de ce modèle harmonique [13]. Elle a également l'avantage de modéliser la parole à l'aide d'un nombre de paramètres plus restreint. Ainsi, les paramètres fréquentiels à estimer se réduisent à la fréquence fondamentale du modèle CARD harmonique (ou CARDH) :

$$\widehat{f_0} = argmin_{f_0}[C'(f_0)] \tag{2.14}$$

$C'(f_0)$ est la fonction de coût à minimiser dépendant de la matrice $\Omega(f_0)$:

$$C'(f_0) = \underline{\mathbf{y}}_p \, [P_Y^\dagger - P_Y^\dagger \, \Omega(f_0)(\Omega(f_0)^H \, P_Y^\dagger \, \Omega(f_0))^{-1} \Omega(f_0)^H P_Y^\dagger] \, \underline{\mathbf{y}}_p \tag{2.15}$$

$$\mathbf{\Omega(f)} = \begin{bmatrix} cos(2\pi f_0(p)) & \cdots & cos(2\pi M f_0(p)) & sin(2\pi f_0(p)) & \cdots & sin(2\pi M f_0(p)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_0(p+1)) & \cdots & cos(2\pi M f_0(p+1)) & sin(2\pi f_0(p+1)) & \cdots & sin(2\pi M f_0(p+1)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_0(N-1)) & \cdots & cos(2\pi M f_0(N-1)) & sin(2\pi f_0(N-1)) & \cdots & sin(2\pi M f_0(N-1)) \end{bmatrix} \tag{2.16}$$

De la même manière, la détermination des amplitudes et des coefficients de la partie autorégressive se fait en utilisant les formules développées pour le modèle CARD (eq. 2.10), (eq. 2.11) et (eq. 2.12), en posant $\Omega = \Omega(f_0)$.

• **Ambiguïté d'octave**

Le bon ajustement d'un modèle dépend de son ordre. Dans le cas d'une modélisation CARDH de p paramètres autorégressifs et M harmoniques, cet ordre est donné par le couple (M, p). On suppose dans la suite que M est tel que :

$$M \leq \frac{1}{2\, f_0} \quad (f_0 \quad \textit{fréquence réduite}) \tag{2.17}$$

On notera $(\, M(f_0),\, p\,)$, l'ordre d'une modélisation CARDH ayant pour fréquence fondamentale f_0.

Cette condition joue un rôle important dans le problème d'estimation de fréquence fondamentale. Elle est à l'origine du phénomène d'ambiguïté d'octave qui se définit comme la détection de valeurs sous-multiples de la véritable fréquence fondamentale. Ce phénomène, pour le cas de la modélisation CARDH, est illustré ci-dessous.
Considérons un signal synthétique harmonique de fréquence fondamentale réduite $f_0 = 0.0265$ mélangé à un bruit autorégressif dont l'ordre $p = 10$ est supposé connu. Le rapport "signal à bruit" (RSB) entre les 2 signaux est de 15 dB.
Les estimés des paramètres du modèle CARDH correspondant sont obtenus en minimisant la fonction de coût représentée sur la figure (Fig. 2.1).
La fonction $C'(f_0)$ possède un minimum local à la valeur 0.0265, mais le minimum global pour l'espace fréquentiel de recherche se situe à la valeur sous-multiple 0.0132. Par cet exemple, on voit clairement que l'application de l'estimateur du Maximum de Vraisemblance sur la modélisation CARDH est sensible au phénomène d'ambiguïté d'octave.

En général, la qualité d'une modélisation donnée augmente avec le nombre de paramètres utilisés. Dans le cas du modèle CARDH, de par l'inégalité (eq. 2.17), l'ordre $(\, M(f_0), p\,)$ du modèle est par construction dépendant de la fréquence fondamentale estimée. Ainsi, parce que l'ensemble des modèles d'ordre $(\, M(f_0), p\,)$ est inclus dans l'ensemble des modèles d'ordre $(\, M(f_0/2), p\,)$, on a $\forall f_0$:

$$C'(f_0/2) \leq C'(f_0)$$

En d'autres termes, la minimisation de $C^{'}(f_0)$ peut être vue comme une minimisation de $C^{'}(f_0/2)$, sous la contrainte que les amplitudes impaires localisées en $\{q \times (f_0/2)\}$ pour $q = 1, 3, 5, \cdots$ soient mises à zéro.

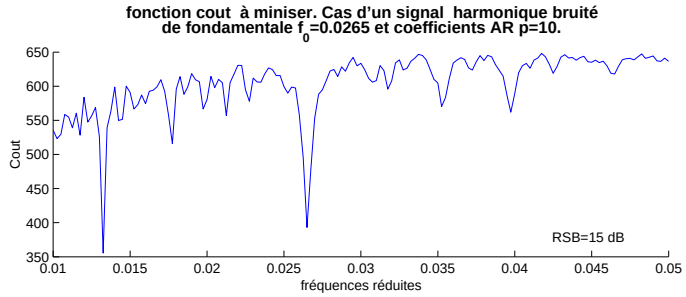


Fig. 2.1 – *Exemple de fonction de coût à minimiser pour l'estimation de la fréquence fondamentale par le Maximum de Vraisemblance. Cas d'un signal harmonique de fondamentale* $f_0 = 0.0265$ *mélangé à un bruit autorégressif d'ordre* $p = 10$ *avec un* $RSB = 15$ *dB.*

Ce phénomène d'ambiguïté d'octaves, inhérent à la modélisation CARDH, s'avère peu important dans le cas de l'analyse/synthèse simple d'une source unique, car hormis le fait de représenter le signal à l'aide d'un plus grand nombre de paramètres (dont certains sont quasi-nulles), il est tout de même possible de resynthétiser ce dernier. Une simple méthode de poursuite de la fréquence fondamentale le long des fenêtres d'analyse peut éventuellement être introduite pour obtenir un contour régulier. Dans le cas du problème de séparation, ce petit inconvénient prend de l'importance avec la pluralité des sources. Ainsi, plus il y a de sources, plus les phénomènes d'ambiguïté apparaissent, et plus il sera nécessaire d'effectuer des poursuites de fréquences fondamentales efficaces pour séparer les sources. Un moyen de pallier ce problème serait de résoudre les ambiguïtés d'octave directement lors de l'estimation des fréquences. La méthode d'estimation que nous proposons par la suite a été développée dans cette optique.

2.2 Estimation de la fréquence fondamentale à l'aide d'un critère d'ordre

2.2.1 Principe du MDL

Le MDL ou "Minimum Description Length" est à l'origine un critère développé par Rissanen, pour sélectionner l'ordre d'un processus autorégressif [44]. La raison d'être des critères de sélection, s'appuie sur le fait qu'une modélisation s'améliore au fur et à mesure que l'on augmente son nombre de paramètres. En

pratique, il est possible de déterminer un ordre optimal, au dessus duquel l'erreur de modélisation varie très peu. Ce dernier représente alors l'ordre suffisant pour rendre compte du comportement d'un signal.
Etant donné que pour le modèle CARDH d'ordre ($M(f_0), p$), le nombre d'harmoniques est directement lié à la fréquence fondamentale f_0 par le biais de (eq. 2.17), l'utilisation d'un critère de sélection du type MDL, apparaît comme un moyen de pallier le problème d'ambiguïté d'octave. Cette technique, plus connue sous le nom du "Principe du MDL", fut initialement proposée par Kay et Naghesa pour l'extraction d'un signal périodique dans du bruit coloré [41].

• Expression de critères d'ordre

Il existe un grand nombre de critères d'ordres différents. Parmi les plus connus, on trouve le critère MDL de Rissanen, le critère AIC (ou "Akaike Information criterion") proposé par Akaike [45] et le critère d'Hannan-Quinn [46].
Les approches utilisées par les auteurs pour définir ces trois critères sont quelque peu différentes et sont présentées ci-dessous.

Soit $p(\mathbf{y}, \underline{\mathbf{\Theta}})$, la vraisemblance associée à la modélisation de N observations $\underline{\mathbf{y}} = [y(0), \cdots, y(N-1)]^T$, où $\underline{\mathbf{\Theta}}$ est le vecteur des paramètres du modèle. Les différents critères sont définis comme suit :

Critère AIC : L'expression de ce critère développé par Akaike est donnée par :

$$AIC = -\, ln\, p(\mathbf{y}, \underline{\widehat{\mathbf{\Theta}}}) + Q \tag{2.18}$$

$\underline{\widehat{\Theta}}$ est l'estimé du vecteur de paramètres $\underline{\Theta}$ obtenu par la méthode du Maximum de Vraisemblance et Q est le nombre de paramètres libres du modèle.
L'AIC est un estimateur non biaisé de la distance Kulback-Liebler entre la densité du modèle $p(\mathbf{x}, \underline{\mathbf{\Theta}})$ et la densité estimée $p(\mathbf{x}, \underline{\widehat{\mathbf{\Theta}}})$. Il a été montré que le critère AIC est un estimateur inconsistant qui a tendance à surestimer l'ordre du modèle et cela même pour des rapports signaux sur bruit important.

Critère MDL : L'approche de Rissanen a été développée sur l'idée que chaque modèle peut être utilisé pour encoder les données. Le critère MDL sélectionne alors le modèle qui va fournir la longueur de code minimale. Son expression est

donnée par :

$$MDL = -ln\, p(\underline{\mathbf{y}}, \underline{\widehat{\mathbf{\Theta}}}) + \frac{1}{2}\, Q\; ln(N) \tag{2.19}$$

Le critère de Rissanen, contrairement à l'AIC, est un estimateur consistant, fournissant ainsi l'ordre véritable du modèle avec la probabilité 1 lorsque la taille des échantillons augmente à l'infini. Il a été montré que le facteur de pénalité utilisé dans le MDL permet de régler la probabilité de surestimation à zéro. En revanche, la probabilité de sous-estimation n'est pas nulle lorsque le nombre d'échantillons est faible et le MDL a finalement tendance à surpénaliser.

Critère de Hannan-Quinn : Pour pallier ces différents problèmes de sous-estimation et sur-estimation, un troisième critère développé par Hannan et Quinn, peut être utilisé. Son expression générale est donnée par :

$$HQ = -ln\, p(\mathbf{y}, \underline{\widehat{\mathbf{\Theta}}}) +\; Q\; \beta \;\; avec \quad \beta = 2c \times ln\,(ln\, N) \quad ,\; c \geq 1 \tag{2.20}$$

β est une variable arbitraire, nommé par la suite *facteur de pénalité*. La quantité $ln\,(ln\, N)$ permet d'obtenir d'un côté la propriété de consistance de l'estimateur et d'un autre côté de réduire le risque de sous-ajustement. On retrouve ce dernier type de critère dans les travaux de Stoica [47] , sous le nom de critère GAIC (ou "General Akaike Information Criterion").

• **Approche bayésienne**

Il existe un autre critère proposé par Schwarz [48]. Son expression est équivalente au critère MDL (eq. 2.19), mais l'approche permettant son élaboration est quelque peu différente. Elle consiste à discriminer les modèles possibles à l'aide d'une règle de décision bayésienne : Schwarz suppose qu'à chaque modèle compétitif peut être assigné une probabilité a priori. Le critère de sélection correspond alors à la sélection du modèle qui fournira la probabilité a posteriori maximale. Cette approche, possède un intérêt non négligeable pour les problèmes de séparation de signaux de parole, dans lesquels on souhaite discriminer les parties voisées des parties non-voisées. On verra par la suite qu'elle nous permettra de développer une mesure de voisement fort utile pour la séparation de ce type de signaux.

On retrouve également cette idée de discrimination dans le travail de Chatterjee [49] qui utilise l'approche bayésienne, d'une part pour estimer les fréquences de deux sinusoïdes proches, et d'autre part pour déterminer la nature réelle du signal observé.

• **Estimation par le principe d'un critère d'ordre**

L'estimation des paramètres du modèle CARD par le principe du MDL, consiste à minimiser la log-vraisemblance de ce modèle, pénalisée d'un terme dépendant du critère de Rissanen. Nous généralisons ici ce principe à l'ensemble des critères présentés ci-dessus.
Par la suite, nous parlerons plus généralement du "principe du critère d'ordre".

Pour un modèle CARDH d'ordre ($M(f_0)$, p), l'estimation de la fréquence fondamentale et du nombre de coefficients autorégressifs p par le principe du critère d'ordre est telle que :

$$[\widehat{f}_0, \widehat{p}] = argmin_{f,p}[L(f,p)] \tag{2.21}$$

$$L(f,p) = (N-p)\ \ ln(\frac{\pi e\, C'(f)}{N-p}) + \beta\, Q \tag{2.22}$$

Où C' est donné par l'expression (eq. 2.15).

Le nombre de paramètres Q est égal à $(2M+p+1)$ et les valeurs prises par β pour les trois critères sont présentées dans le tableau suivant (2.2.1) :

	AIC	MDL	HQ
β	1	$0.5 \times ln(N-p)$	$2c \times \ln(\ln(N-p))\ c \geq 1$

Tab. 2.1 – **Facteur de pénalité β pour les critères AIC, MDL et HQ**

2.2.2 Hypothèse asymptotique - réduction de la complexité

Dans le domaine de la parole, les fréquences fondamentales sont généralement comprises entre 80 Hz et 400 Hz. En terme de précision fréquentielle, le cadre de la séparation impose l'utilisation d'une précision élevée, car dans un mélange quelconque de signaux de parole, les fréquences des harmoniques peuvent être très proches. Un écart raisonnable entre deux fréquences peut être pris égal à 2 Hz. Ainsi, quelque soit le mélange à analyser, la recherche des fréquences fondamentales est limitée à un nombre fini de valeurs.

Soit E_f l'ensemble des valeurs possibles de la fréquence fondamentale f_0. On définit également E_p, l'ensemble contenant les valeurs possibles de paramètres p pour la modélisation CARDH. L'estimation par le principe du critère d'ordre nécessite alors le calcul de $(dim[E_f] \times dim[E_p])$ termes de vraisemblances pénalisées, $dim[E]$ étant la dimension de l'ensemble E. Le coût de calcul d'une vraisemblance seule est $dim[E_f]$. Il dépend principalement de l'intervalle fréquentiel de recherche de f_0 et de la précision fréquentielle voulue.
Il en résulte finalement que le principe du critère d'ordre pour l'estimation de fréquence fondamentale, ou plus précisément la minimisation (eq. 2.21) peut s'avérer très coûteuse en terme de calcul.
Dans cette partie nous proposons alors d'utiliser le comportement asymptotique de l'estimateur du Maximum de Vraisemblance pour réduire considérablement le coût de calcul de cette méthode d'estimation, la rendant ainsi opérationnelle pour le problème de séparation.

• Propriétés asymptotiques

Les propriétés asymptotiques de l'estimateur du Maximum de Vraisemblance sont obtenues en étudiant son comportement, lorsque le nombre d'observations N à modéliser devient "grand".
Un travail conséquent effectué par Stoica sur le problème d'estimation de sinusoïdes dans du bruit coloré [50] [51], a mis en évidence la possibilité d'estimer les paramètres harmoniques indépendamment des paramètres de bruit pour une modélisation CARD. Cette estimation, dite découplée, se justifie en démontrant que le comportement asymptotique de l'estimateur du Maximum de Vraisemblance (ML) est équivalent à celui de la méthode des Moindres Carrés (LS). L'estimateur LS ainsi obtenu, a l'avantage d'être beaucoup moins coûteux en termes de calcul,

car il suppose le nombre de paramètres autorégressifs nul ($p = 0$).
On trouvera dans l'annexe B des détails, ainsi que de nombreuses références sur l'équivalence asymptotique des deux estimateurs. Nous rappelons simplement ci-dessous l'expression de l'estimateur LS. Les conséquences de son utilisation dans le principe du critère d'ordre, en termes de coût de calcul, sont ensuite présentées.

L'expression de la fréquence fondamentale estimée par la méthode des Moindres Carrés pour un modèle CARDH est donnée par :

$$\widehat{f}_{LS} = argmin_f[\sum_{n=0}^{N-1}(y(n) - s(n))^2] \tag{2.23}$$

Cette expression correspond à la minimisation de l'erreur quadratique moyenne entre les observations $\{y(0), \cdots, y(N-1)\}$ et la partie harmonique $\{s(0), \cdots, s(N-1)\}$ du modèle CARDH. Elle est équivalente à la minimisation (eq. 2.14) pour $p = 0$. On notera que la méthode des Moindres Carrés lorsqu'elle est appliquée seule, est sensible aux ambiguïtés d'octave, puisqu'elle est un cas particulier de la méthode du Maximum de Vraisemblance.
La conséquence directe de l'estimation découplée des paramètres est que pour une fenêtre d'analyse suffisamment longue, il est possible d'appliquer le principe du critère d'ordre en remplaçant suivant le critère considéré, dans les équations (eq. 2.18), (eq. 2.19) ou (eq. 2.20), l'estimé ML des paramètres harmoniques par leur estimé LS.

Pour les signaux de parole, la longueur des fenêtres d'analyse est limitée par l'hypothèse de stationnarité, nécessaire à la modélisation CARDH. En tenant compte de cette hypothèse, la durée est alors choisie la plus grande possible, c'est-à-dire environ 50 ms. Nous verrons alors par la suite que cette durée fournit un nombre d'observations suffisant pour utiliser l'estimation découplée.

Finalement, pour N grand, l'estimation par le principe du critère d'ordre peut s'effectuer en déterminant au préalable l'estimé LS de f_0, puis en introduisant ce dernier dans l'expression (eq. 2.14). Ainsi l'hypothèse asymptotique permet de réduire considérablement le coût de calcul, en passant du calcul de $(dim[E_f] \times dim[E_p])$ termes de log-vraisemblance, suivi de $(dim[E_f] \times dim[E_p])$ termes de pénalité, au calcul de $(dim[E_f] \times 1)$ termes de log-vraisemblance, suivi de $(dim[E_f] \times dim[E_p])$ termes de pénalité.

• Prise en compte explicite des ambiguïtés d'octave

Nous avons vu auparavant que pour N grand, le coût de calcul nécessaire à l'estimation de la fréquence fondamentale par le principe du critère d'ordre, se réduit considérablement. En supposant que l'estimateur LS fournit un estimé correct de f_0 ou d'une sous-harmonique quelconque de f_0, l'introduction des termes de pénalités par le calcul de l'équation (eq. 2.21) détermine finalement l'ordre harmonique de cette fréquence estimée. Cette simple constatation nous permet de réduire à nouveau la complexité de la méthode d'estimation, par la prise en compte explicite des ambiguïtés d'octave :

soit $\widehat{f}_{LS}$, la fréquence estimée par la méthode des Moindres Carrés. L'estimé de f_0 est calculée à partir d'un jeu réduit de modèle CARDH où les parties harmoniques sont déterminées pour $f\epsilon\{\widehat{f}_{LS}, 2.\widehat{f}_{LS}, 3.\widehat{f}_{LS}, \cdots\}$. L'ensemble $E_k = \{0, 1, 2, 3, \cdots\}$ définit alors l'espace des ordres harmoniques de la fréquence $\widehat{f}_{LS}$.
La prise en compte explicite des ambiguïtés d'octave, entraîne le calcul de $(dim[E_f] \times 1)$ termes de log-vraisemblance, suivi cette fois-ci de seulement $(dim[E_k] \times dim[E_p])$ termes de pénalité.
On obtient finalement un estimateur de fréquence fondamentale dont la complexité est considérablement réduite par rapport à l'estimateur initialement proposé par Kay et Nagesha (voir tableau (2.2)).
Dans la suite de ce chapitre, la validité de l'hypothèse "N grand" sera illustrée par un grand nombre de simulations sur des mélanges synthétiques de signaux harmoniques, ainsi que sur des mélanges réels de signaux de parole.

méthode	coût
principe du MDL (Kay & Nagesha)	$(dim[E_f] \times dim[E_p])$ $+(dim[E_f] \times dim[E_p])$ E_f : ensemble des fréquences E_p : ensemble des paramètres AR
Principe du critère d'ordre	$(dim[E_f] \times 1) + (dim[E_k] \times dim[E_p])$ E_k : ensemble de l'ordre harmonique

Tab. 2.2 – **Complexité de la méthode** : comparaison avec la méthode initiale de Kay et Nagesha [41].

2.3 Etude de la méthode d'estimation à l'aide d'un critère d'ordre

Cette partie est consacrée à l'estimation de fréquence fondamentale à l'aide d'un critère d'ordre. Un résumé de l'estimation basée sur le principe du critère d'ordre est tout d'abord présenté, puis à partir d'outils de simulations, son étude détaillée est entreprise. Cette dernière consiste alors à mettre en évidence l'utilité du critère d'ordre pour pallier les ambiguïtés d'octave et la validité de l'hypothèse asymptotique. L'estimation de valeurs sous-multiples de la fréquence fondamentale par la méthode des Moindres Carrés est également étudiée.
L'évaluation des performances de la méthode d'estimation, lorsque le critère MDL est utilisé, montre que ce dernier ne permet pas de réduire à zéro le taux de sous-harmoniques estimées. Plusieurs critères d'ordre, de la famille des critères proposées par Hannan-Quinn (eq. 2.20), sont alors testés.

2.3.1 Résumé de la méthode d'estimation

La méthode d'estimation proposée ici est résumée ci-dessous :

- étape 1 : on accepte le phénomène d'ambiguïté en cherchant à estimer une valeur sous-harmonique de la fréquence fondamentale, plutôt que f_0. Cette sous-harmonique, notée $\widehat{f}_{0_{LS}}$ est obtenue sans prendre en compte le bruit autorégressif, par l'estimateur des Moindres Carrés (eq. 2.23).

- étape 2 : l'ordre harmonique $\widehat{k}$ de $\widehat{f}_{0_{LS}}$ est déterminé par application du principe du critère d'ordre à un jeu de modèles CARDH, dont les fréquences fondamentales sont $\{(k \times \widehat{f}_{0_{LS}})\}_{k=0,1,2,\cdots}$ et pour lesquels le nombre de paramètres AR varie.

2.3.2 Simulations - mise en oeuvre

Afin d'étudier en détails la méthode proposée ci-dessus, nous développons un ensemble de simulations basées sur des mélanges de signaux synthétiques. Les performances de la méthode sont alors illustrées par le calcul de pourcentages de fréquences correctement estimées et par l'intermédiaire du biais et de l'écart

quadratique moyen des estimateurs utilisés.

• **mélanges synthétiques**

Deux types différents de mélange vont être utilisés :

– Un signal harmonique dans un bruit autorégressif d'ordre $p = 10$

– Un mélange du plusieurs signaux harmoniques de fréquence fondamentale différente dans un bruit blanc de variance σ^2

Ces deux mélanges ont été choisis pour représenter des modèles de parole de type "voisée/non-voisée" et de parole "voisée/voisée". En conséquence, les fréquences fondamentales utilisées, appartiennent à l'intervalle fréquentiel [80,400] Hz relatif à la parole. Le nombre d'harmoniques de la partie sinusoïdale est alors déduit de l'équation (eq. 2.17) et les harmoniques sont supposées de même amplitude.
La densité spectrale du processus AR(10) est illustrée par la figure (Fig. 2.2).
Le caractère aléatoire de la partie bruit de chaque mélange est utilisé pour construire des jeux de simulation, constitué de N_b réalisations indépendantes, à partir desquelles nous pourrons entre autre déterminer les pourcentages de fréquences fondamentales correctement estimées.

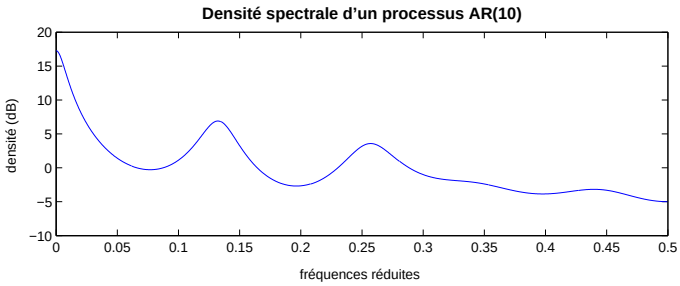


FIG. 2.2 – *Densité spectrale de puissance d'un processus AR d'ordre 10.*

Dans la suite, les mélanges sont échantillonnés à une fréquence $f_e = 4000Hz$ et l'ensemble des résultats de simulation utilisent la notation en fréquence réduite. Ainsi, l'intervalle de recherche des fréquences fondamentales est donnés par [0.02,0.1]. La longueur d'analyse des mélanges est choisie égale à 50ms, durée

pour laquelle l'hypothèse de quasi-stationnarité du signal de parole reste valable.

• Définition d'une fréquence correctement estimée

Pour l'ensemble des simulations, l'estimée $\widehat{f}$ d'une fréquence fondamentale réduite est supposée correcte lorsqu'on a :

$$|\widehat{f} - f_0| \leq 5.10^{-4} \tag{2.24}$$

La valeur de cet écart correspond à la précision fréquentielle utilisée dans l'estimation des fréquences fondamentales.

Dans le cas d'une sous-harmonique d'ordre k, l'estimé est supposé correcte lorsque :

$$|\widehat{f} - \frac{f_0}{k}| \leq \frac{5.10^{-4}}{k} \tag{2.25}$$

Les performances de la méthode d'estimation, pour les différents types de mélanges synthétiques considérés, sont évaluées à l'aide des pourcentages de fréquences correctement estimées. Ces pourcentages sont calculés en déterminant le nombre de fréquences correctement estimées sur les N_b réalisations caractéristques du mélange.

• Définition de la moyenne et de l'écart quadratique moyen

Afin de mettre en évidence les performances et de faciliter la comparaison des estimateurs utilisés par la suite, deux caractéristiques statistiques sont utilisées : la moyenne et l'écart quadratique moyen de l'estimé $\widehat{f_0}$.

$$moyenne = \frac{1}{N_b} \sum_{i=1}^{N_b} \widehat{f_0^i} \tag{2.26}$$

$$rmse = 20.\log_{10}(\sqrt{\frac{1}{N_b} \sum_{i=1}^{N_b} (\widehat{f_0^i} - f_0)^2}) \tag{2.27}$$

$\widehat{f_0^i}$ est la i^{ieme} estimation. N_b est le nombre de réalisations utilisé dans un jeu de simulations. On choisit par la suite $N_b = 200$.

2.3.3 Le principe MDL - résolution des ambiguïtés d'octave

A partir de mélanges du type "signal harmonique dans un bruit AR", on détermine successivement les fréquences fondamentales estimées par la méthode du Maximum de Vraisemblance et par le principe du MDL.
On cherche à mettre en évidence les phénomènes d'ambiguïté d'octave omniprésents dans les estimations ML et la réduction de ces derniers lorsqu'un critère de sélection du type Rissanen est appliqué. Trois fréquences fondamentales réduites des mélanges $f_{0_1} = 0.0945$, $f_{0_2} = 0.053$ et $f_{0_3} = 0.024$ sont prises en compte, permettant ainsi de traiter trois mélanges qui différent par le nombre d'harmoniques éventuellement détectables. Par exemple, dans l'intervalle d'étude [0.02,0.1], f_{0_1} possède les trois valeurs sous-harmoniques $\{0.0473, 0.0, 315, 0.0236\}$, tandis que f_{0_3} n'en possède aucune.
Les figures (Fig. 2.3), (Fig. 2.4) et (Fig. 2.5) illustrent les résultats obtenus. Elles présentent les pourcentages de fréquences estimées correspondant à la fréquence fondamentale, ainsi qu'à ses valeurs sous-harmoniques (ou sur-harmoniques dans le cas du signal de fréquence fondamentale f_{0_3}), en fonction du rapport "signal à bruit" noté RSB.

Lorsque l'estimateur du Maximum de Vraisemblance est appliqué, les fréquences correctement estimées correspondent à des valeurs sous-harmoniques d'ordre 3 pour f_{0_1} (Fig. 2.3 -a)) et d'ordre 2 pour f_{0_2} (Fig. 2.4 -a)). Aucune fréquence fondamentale n'est estimée par cette méthode, quel que soit le RSB.
Lorsque le principe du MDL est appliqué (Fig. 2.3 -b)) (Fig. 2.4 -b)), la plupart des fréquences estimées sont des fréquences fondamentales, bien qu'un pourcentage non nul de sous-harmoniques estimées reste présent. On note également que le critère de sélection augmente globalement le pourcentage de fréquences correctement estimées, qu'elles soient relatives à une fondamentale ou à une sous-harmonique. L'augmentation est particulièrement visible pour les rapports "signal à bruit" ou RSB inférieurs à 10 dB.
Dans le cas particulier de la fréquence fondamentale f_{0_3} (Fig. 2.5)), les pourcentages déterminés correspondent aux valeurs multiples $\{f_{0_3}, 2 \times f_{0_3}, \cdots\}$. Parce qu'aucune valeur sous-harmonique n'entre en jeu dans ce cas, le Maximum de Vraisemblance fournit directement l'estimé de f_{0_3}. On notera tout de même que l'application du critère de Rissanen fait apparaître des pourcentages très faibles correspondant à des surharmoniques d'ordre 2.

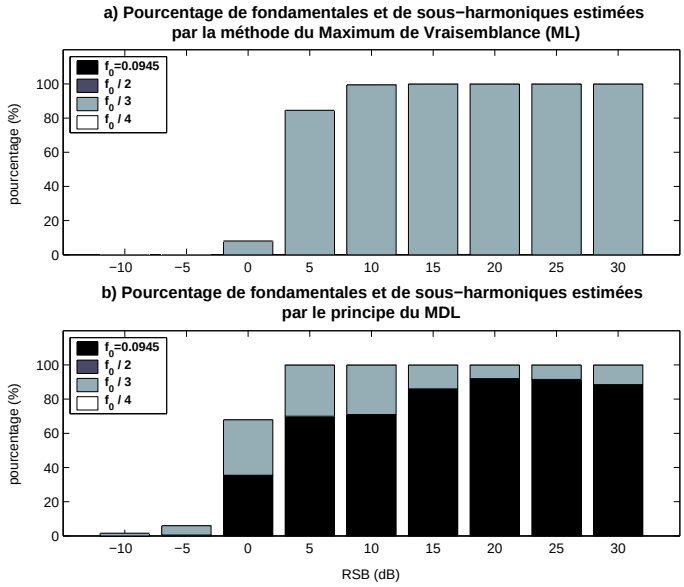


Fig. 2.3 – *Pourcentage de fréquences fondamentales et de sous-harmoniques estimées en fonction du RSB (dB) : a) par le Maximum de Vraisemblance (ML) - b) par le principe du MDL. Application à un signal harmonique caractérisée par* $f_{0_1} = 0.0945$ *dans un bruit AR d'ordre 10.*

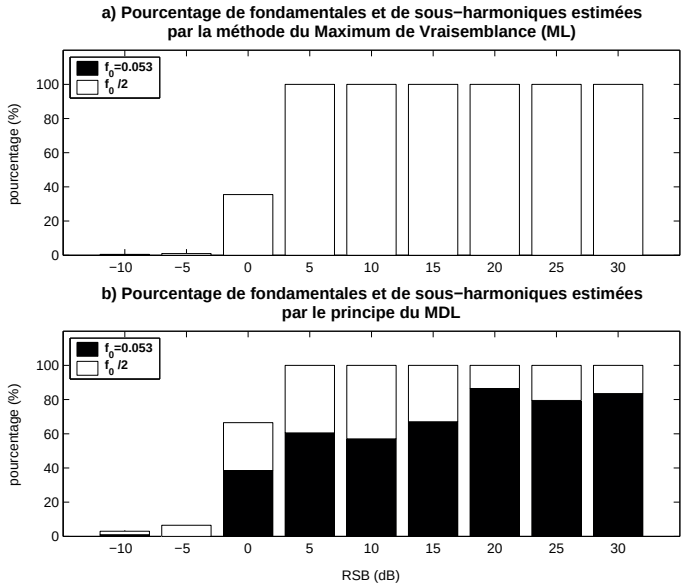


Fig. 2.4 – *Pourcentage de fréquences fondamentales et de sous-harmoniques estimées en fonction du RSB (dB) : a) par le Maximum de Vraisemblance (ML) - b) par le principe du MDL. Application à un signal harmonique caractérisée par* $f_{0_2} = 0.053$ *dans un bruit AR d'ordre 10.*

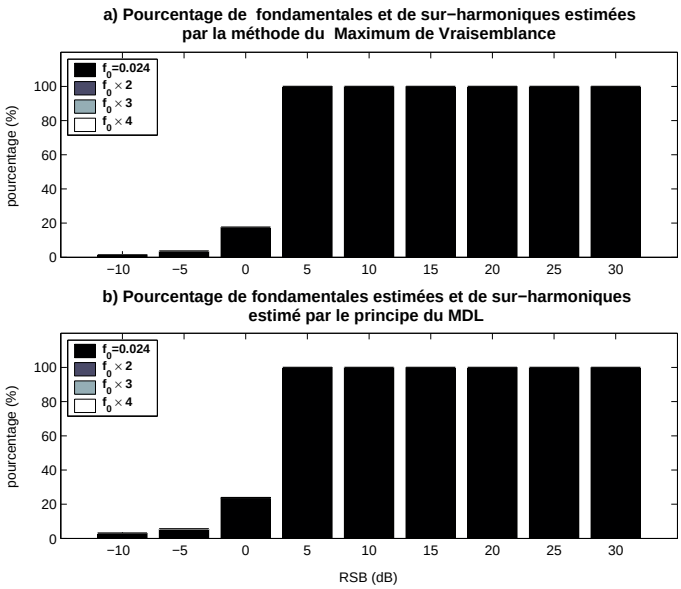


FIG. 2.5 – *Pourcentage de fréquences fondamentales et de sous-harmoniques estimées en fonction du RSB (dB) : a) par le Maximum de Vraisemblance (ML) - b) par le principe du MDL. Application à un signal harmonique caractérisée par* $f_{0_3} = 0.024$ *dans un bruit AR d'ordre 10.*

2.3.4 Performance de l'estimateur LS

Comme nous avons vu précédemment, l'hypothèse asymptotique suppose que pour un nombre d'observations N grand, l'estimation des paramètres harmoniques du modèle CARDH peut être obtenue sans prendre en compte les paramètres de bruit. Dans cette approximation, on peut alors substituer l'estimateur du Maximum de Vraisemblance de la fréquence fondamentale (ML) par celui de la méthode des Moindres Carrés (LS). Ce paragraphe illustre les performances de ces deux estimateurs lorsqu'ils sont déterminés sur des fenêtres de longueur égale $N = 200$.

• **Cas d'un mélange "signal harmonique dans un AR (10)"**

Les performances des estimateurs LS et ML sont illustrées par l'intermédiaire du biais (eq. 2.26) et de l'écart quadratique moyen (eq. 2.27) : la figure (Fig. 2.6) comparent les résultats obtenus pour les deux méthodes, dans le cas d'un signal harmonique de fréquence fondamentale $f_0 = 0.053$ dans un bruit AR.
L'estimateur ML suppose le nombre de paramètres autorégressifs $p = 10$ connu, tandis que pour la méthode des Moindres Carrés, qui ne prend pas en compte la partie bruitée, on a $p = 0$. Les RSB varient de $-10\ dB$ à $+30\ dB$.

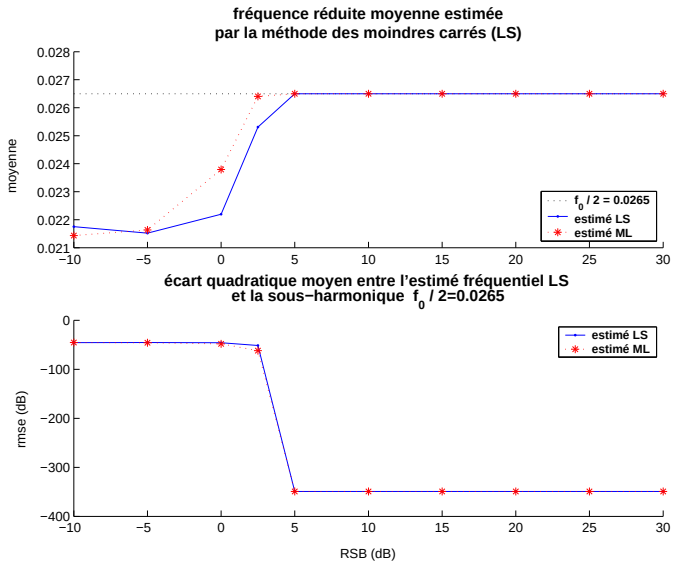


Fig. 2.6 – *Moyenne et écart quadratique moyen de la fréquence estimée par la méthode des Moindres Carrés (·) et le Maximum de Vraisemblance (⋆) en fonction du RSB(dB). Cas d'une fondamentale recherchée dans l'intervalle [0.02 0.1]. Application à un signal harmonique dans un AR caractérisé par* $f_0 = 0.0053$ *et* $p = 10$.

Tout comme l'estimateur ML, l'estimateur LS est capable de déterminer une valeur sous-harmonique de f_0. Il reste tout de même un peu moins performant que l'estimateur ML pour des RSB inférieurs à 3 dB. Ce résultat est quelque peu attendu, puisque l'estimateur LS ne prend pas en compte la partie bruitée du modèle, que l'on ne peut cependant plus négliger pour de tels niveaux de bruit.

Nous avons vu précédemment que l'intervalle de recherche de la fréquence fondamentale utilisé est [0.02,0.1]. Puisque la probabilité d'estimation d'une valeur sous-harmonique est supérieure à celle de la fréquence fondamentale, on peut réduire cet intervalle fréquentiel pour rechercher directement une sous-harmonique. Dans ce cas, il est alors également intéressant de regarder la moyenne et l'écart quadratique moyen obtenu lorsque l'estimation par les Moindres Carrés est appliqué à un intervalle fréquentiel réduit à une octave : les fréquences que l'on cherche à estimer sont comprises entre 0.02 et 0.04. Afin d'obtenir une précision équivalente à l'estimation LS précédente, le pas entre deux fréquences est également réduit.
La figure (Fig. 2.7) présente alors la moyenne et l'écart quadratique obtenu pour l'intervalle de recherche [0.02,0.04], en fonction du RSB.
Les résultats obtenus sont équivalents à ceux obtenus sur l'intervalle [0.02, 0.1] et montrent ainsi qu'il est possible d'estimer directement une sous-harmonique de la fréquence fondamentale par la méthode des Moindres Carrés, en réduisant l'intervalle de recherche à l'octave [0.02,0.04].

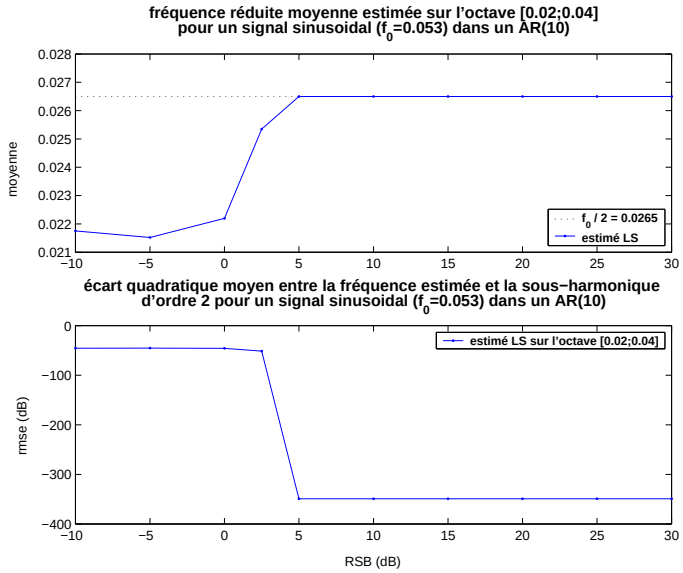


FIG. 2.7 – *Moyenne et écart quadratique moyen de la fréquence estimée par la méthode des Moindres Carrés (·) en fonction du RSB(dB). Cas d'une fondamentale recherchée dans l'intervalle* [0.020.04]. *Application à un signal harmonique dans un AR caractérisé par* $f_0 = 0.0053$ *et* $p = 10$.

• **Cas d'un mélange de deux sources harmoniques**

Dans cette application, on considère deux mélanges harmoniques différents : le premier est un mélange de deux signaux harmoniques de fréquences fondamentales assez éloignées $f_{01} = 0.074$ et $f_{02} = 0.0945$, tandis que le second correspond à deux fréquences fondamentales proches $f_{01} = 0.092$ et $f_{02} = 0.0945$.
Leur proximité est à prendre au sens de la résolution de Fourier : pour $N = 200$ observations, l'écart entre les deux dernières fréquences reste inférieur à la résolution $(1/N)$ et ces dernières ne peuvent alors être résolues par des méthodes basées sur la transformée de Fourier.
La figure (Fig. 2.8) présente le pourcentage des sous-harmoniques estimées par la méthode des Moindres Carrés en fonction du rapport de puissance des deux signaux harmoniques, noté RSS (dB). Ce dernier est défini à partir des puissances P_1 et P_2 respectives aux fréquences fondamentales f_{0_1} et f_{0_2} comme :

$$RSS\,(dB) = 10\log_{10}(\frac{P_1}{P_2}) \tag{2.28}$$

Le pourcentage est calculé pour chaque ordre sous-harmonique $\{1, 2, 3, 4\}$ et une fréquence est supposée correctement estimée lorsqu'elle est égale à l'une des deux sous-harmoniques (eq. 2.25).

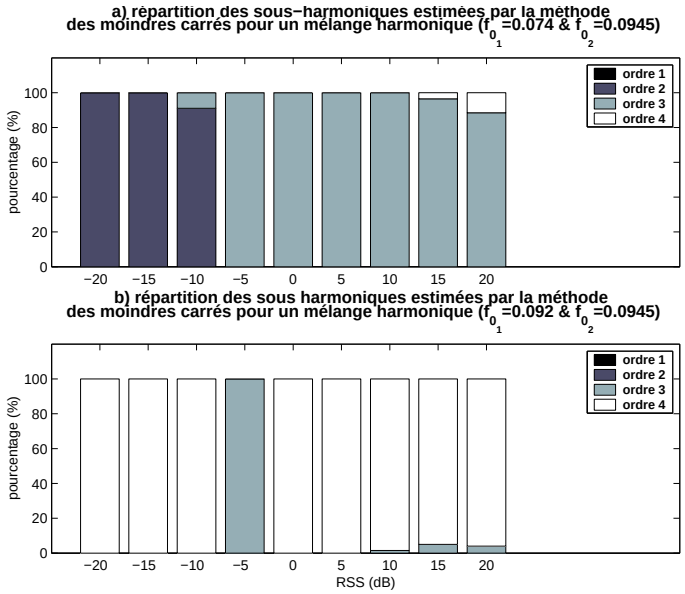


Fig. 2.8 – *Pourcentage de fondamentales et de sous-harmoniques estimées par la méthode des Moindres Carrés. Cas de mélanges de 2 sources harmoniques où : a)* $f_{01} = 0.074$ *et* $f_{02} = 0.0945$. *b)* $f_{01} = 0.092$ *et* $f_{02} = 0.0945$.

L'estimateur des Moindres Carrés permet l'estimation correcte de l'une des fréquences fondamentales dans un mélange harmonique de deux sources et ce même pour des fréquences fondamentales proches.
En règle général, c'est-à-dire pour des $RSS \neq 0dB$, l'estimé LS correspond au signal harmonique prédominant dans le mélange.

• Cas d'un mélange de 5 sources harmoniques

Le cas d'un mélange de 5 sources, dans lequel le rapport de puissance entre la source prédominante et l'ensemble des quatre autres sources varient de 0 dB à 30 dB est ici traité. Cette simulation a pour but de montrer que l'estimation d'une fréquence fondamentale ou de l'un de ses sous-multiples par la méthode des Moindres Carrés reste performante en présence d'autres sources harmoniques.
La figure (Fig. 2.9) présente alors le pourcentage de fréquences estimées correspondant à la source de plus grande énergie dans le mélange, en fonction du rapport de puissance RSS (dB). Les fréquences du mélange sont $\{f_{0_1} = 0.0230, f_{0_2} = 0.0255, f_{0_3} = 0.0305, f_{0_4} = 0.0355, f_{0_5} = 0.0405\}$. f_{0_1} est choisie comme la fréquence fondamentale du signal prépondérant et le rapport de puissance RSS est donné par :

$$RSS\,(dB) = 10.\log_{10}(\frac{P_1}{P_t}) \tag{2.29}$$

où P_t est la puissance de l'ensemble des 4 autres signaux.

La figure (Fig. 2.9) met en évidence la possibilité d'estimer la fréquence fondamentale d'une source, ou bien l'une de ses sous-harmoniques, même lorsque cette source est contenue dans un mélange allant jusqu'à 5 sources harmoniques et dont le RSS reste faible.

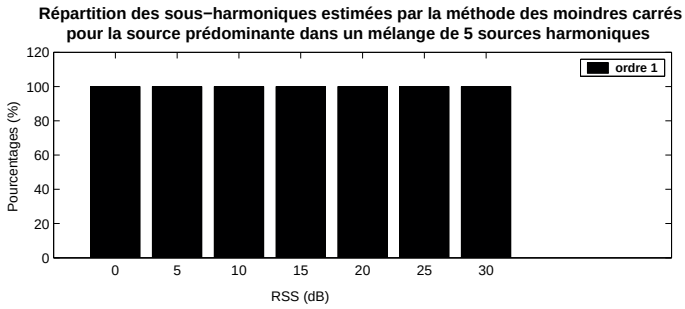


Fig. 2.9 – *Pourcentage de fondamentales et de sous-harmoniques estimées pour la source prédominante (f_{01}) d'un mélange de 5 sources harmoniques par la méthode des Moindres Carrés (LS) :* $f_{0_1} = 0.0230$, $f_{0_2} = 0.0255$, $f_{0_3} = 0.0305$, $f_{0_4} = 0.0355$ *et* $f_{0_5} = 0.0405$.

2.3.5 Réduction de la complexité

Les espaces des fréquences réduites E_f et des paramètres E_p utilisés dans l'ensemble des simulations de ce chapitre sont tels que :

$$E_f = \{\ (0.02 :\ 1.25\, e^{-4} :\ 0.04)\ \} \quad et \quad E_p = \{0, 1, \cdots, 20\} \tag{2.30}$$

Parce que la dimension de l'espace des ordres harmoniques E_k dépend de la sous-harmonique estimée par l'estimateur des Moindres Carrés, la valeur maximale que peut prendre $dim[E_k]$ est telle que :

$$dim[E_k] \leq 5$$

Le tableau qui suit présente alors la complexité de notre méthode comparée à celle de Kay et Nagesha pour ces espaces. Il se réfère au tableau plus général présenté dans le paragraphe (2.2.2) (voir tableau (2.2)).

2.3.6 Choix d'un critère de sélection

On s'intéresse ici plus particulièrement à l'étude du critère de sélection. Ce dernier est utilisé pour déterminer l'ordre harmonique d'une fréquence estimée

méthode	coût
principe du MDL (Kay & Nagesha)	(161×21) +(161×21)= 6762
principe du critère d'ordre	(161×1) + (5×21)= 3486

TAB. 2.3 – *Complexité de la méthode* : comparaison avec la méthode initiale de Kay et Nagesha [41].

par la méthode des Moindres Carrés.

Les simulations présentées dans le paragraphe (2.3.3), utilisent le critère de Rissanen. Elles montrent une diminution notable des sous-harmoniques détectées, mais le critère ne permet pas de réduire à zéro ce taux de détection.
Cette constatation reste en accord avec les résultats théoriques du critère MDL, selon lesquels il aurait tendance à surestimer l'ordre d'un modèle. Pour le modèle CARDH, une surestimation du nombre d'harmonique se traduit alors par une fréquence estimée égale à une sous-harmonique. On peut alors supposer que le choix du critère de Rissanen pour l'estimation de la fréquence fondamentale n'est pas optimal.
On se propose donc d'étudier un second type de critère, afin d'améliorer la détection de l'ordre harmonique. Les expressions des deux critères, en fonction du nombre d'observations N sont rappelées dans le tableau suivant :

critère	MDL	Hannan-Quin
β	$0.5 \ln N$	$2c \times \ln(\ln N),\ c > 1$

TAB. 2.4 – *Facteur de pénalité β pour les critères de Rissanen et d'Hannan-Quinn.*

La figure (Fig. 2.10) présente la valeur de la pénalisation en fonction du nombre de paramètres d'un modèle pour le critère de Hannan-Quinn où le coefficient β est tel que $2c = \{1.5, 2, 2.5, 3, 3.5, 4\}$ et pour le critère de Rissanen.
Les deux critères sont équivalent pour $2c \approx 1.68$.
Dans la suite, le principe du critère d'ordre est appliqué selon la méthode présentée au paragraphe (2.3.1).

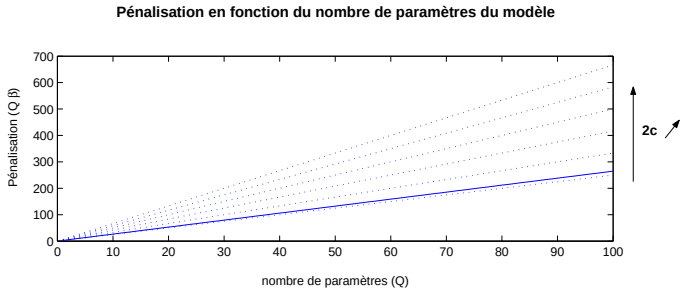


Fig. 2.10 – *Pénalisation en fonction du nombre de paramètres du modèle : (-) critère de Rissanen - (⋯) critère de sélection du type "Hannan-Quinn" pour* $2c = \{1.5, 2, 2.5, 3, 3.5, 4\}$

• **Cas du signal harmonique dans un AR(10)**

Afin de rendre compte des différents comportements des critères, on détermine pour $2c = \{1.5, 2, 2.5, 3, 3.5, 4\}$, les pourcentages de fréquences fondamentales, de sous-harmoniques et de sur-harmoniques estimées par la méthode basée sur le critère d'ordre. Les figures (Fig. 2.11) et (Fig. 2.12) présentent les résultats obtenues en fonction du RSB, pour deux mélanges différents. Les fréquences fondamentales réduites de la partie harmonique sont respectivement $f_0 = 0.0945$ et $f_0 = 0.024$. La valeur $2c$ et plus précisément le critère, est alors choisie de telle sorte que le pourcentage de fréquences fondamentales estimées soit maximum, c'est-à-dire que l'on cherche à minimiser la détection de sous-harmoniques due à une pénalisation trop faible de la log-vraisemblance (eq. 2.22), ainsi que la détection de sur-harmoniques due à une pénalisation trop forte. Les résultats obtenus pour les différentes pénalisations, montrent que le cas $2c = 3$ permet d'obtenir le meilleur compromis entre les détections de sous-harmoniques pour $f_0 = 0.0945$ et de sur-harmoniques pour $f_0 = 0.024$.

Il également possible de comparer la répartition des fréquences estimées présentée sur la figure (Fig. 2.11), a celle obtenue par l'estimateur de Kay et Nagesha et illustrée par la figure (Fig. 2.3). Ainsi, si les ambiguïtés d'octave ne sont pas prises en compte, les deux estimations sont équivalentes pour des RSB supérieurs à 0 dB. En dessous de cette valeur, les performances de notre méthode apparaissent un peu moins bonnes. Ce résultat s'explique par l'introduction de la méthode des Moindres Carrés, dont les performances se dégradent pour ces faibles valeurs de RSB (cf. paragraphe (2.3.4)).

Tout comme pour les fréquences fondamentales estimées, on peut également

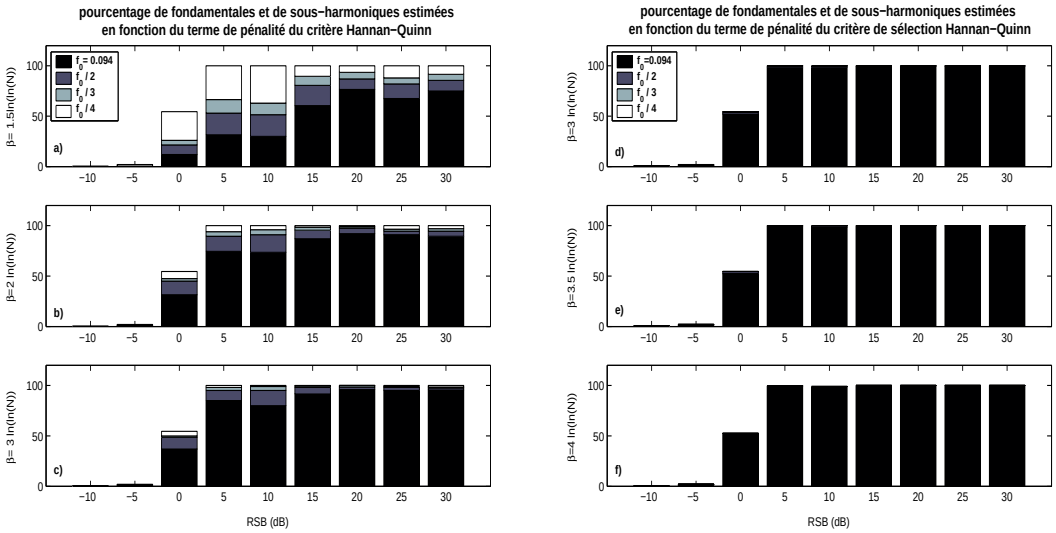


FIG. 2.11 – *Pourcentage de fréquences fondamentales et de sous-harmoniques estimées en fonction du RSB. Cas d'un signal harmonique de fréquence fondamentale* $f_0 = 0.0945$ *dans un bruit AR et pour différents critères de sélection : a)* $2c = 1.5$ *- b)* $2c = 2$ *- c)* $2c = 2.5$ *- d)* $2c = 3$ *- e)* $2c = 3.5$ *- f)* $2c = 4$.

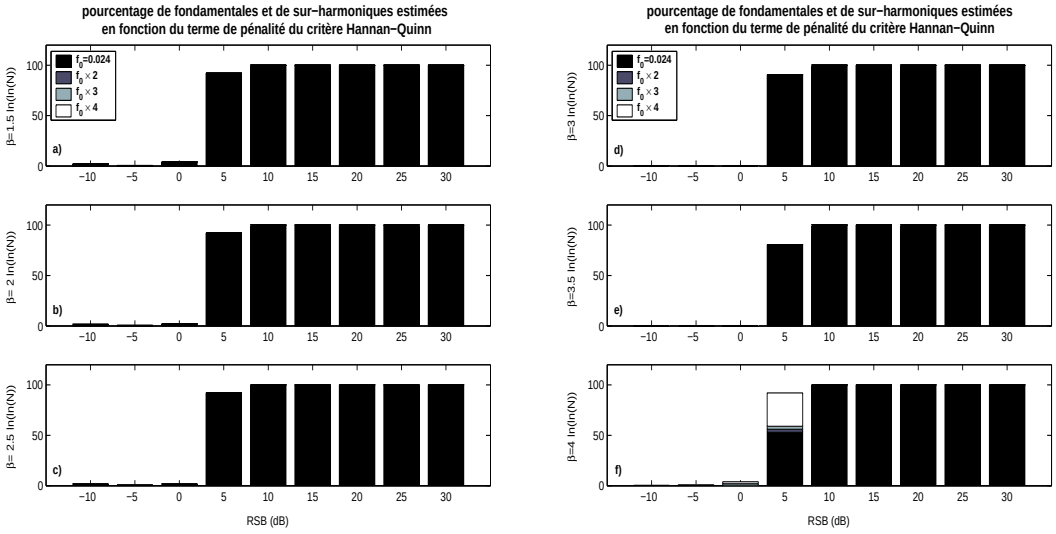


FIG. 2.12 – *Pourcentage de fréquences fondamentales et de sous-harmoniques estimées en fonction du RSB. Cas d'un signal harmonique de fréquence fondamentale* $f_0 = 0.024$ *dans un bruit AR et pour différents critères de sélection : a)* $2c = 1.5$ *- b)* $2c = 2$ *- c)* $2c = 2.5$ *- d)* $2c = 3$ *- e)* $2c = 3.5$ *- f)* $2c = 4$.

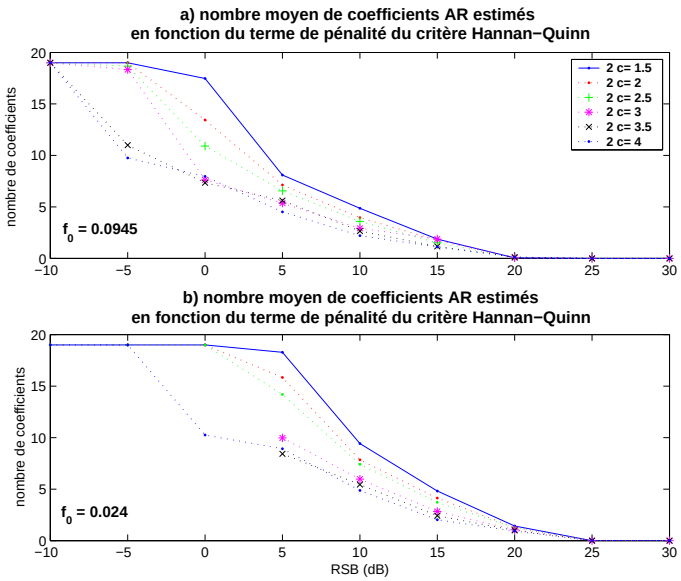


Fig. 2.13 – *Nombre moyen de coefficients AR estimés par la méthode du critère d'ordre, lorsque la fréquence fondamentale du signal source est correctement estimée. Cas d'un signal harmonique dans un bruit AR ($P = 10$) : a) $f_0 = 0.0945$ - b) $f_0 = 0.024$.*

regarder l'influence du critère de sélection sur le nombre de coefficients autorégressifs estimé. Les résultats sont illustrés sur la figure (Fig. 2.13).

• Cas d'un mélange de deux signaux harmoniques

Pour un mélange harmonique dont les fréquences fondamentales sont $f_{01} = 0.0945$ et $f_{02} = 0.074$, on détermine à nouveau le pourcentage de fréquences fondamentales et de sous-harmoniques estimées, lorsqu'un critère de sélection du type Hannan-Quinn est utilisé. Le terme de pénalité appliqué est tel que $2c = 3$. Les pourcentages obtenus, ainsi que le nombre moyen de coefficients AR estimé, en fonction du RSS des deux signaux, sont présentés sur la figure (Fig. 2.14).
La fréquence estimée par la méthode basée sur le critère d'ordre, permet d'obtenir l'une des deux fréquences fondamentales, quelque soit la puissance entre les deux sources. En général, la fréquence fondamentale obtenue correspond au signal prédominant du mélange.

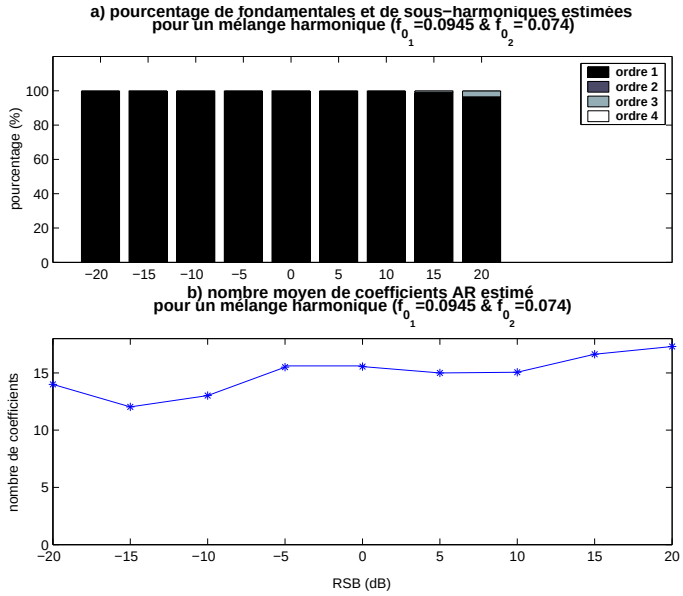


Fig. 2.14 – *a) Fréquences fondamentales estimées et sous-harmoniques estimées par la méthode du critère d'ordre en fonction du RSS des deux signaux. b) Nombre moyen de coefficients AR estimé pour les fréquences égales à l'une des fréquences fondamentales ou à une de leur sous-harmonique.*

2.4 Conclusion

Ce chapitre présente une méthode d'estimation de la fréquence fondamentale applicable aux signaux de parole. Basée sur une modélisation "harmonique + autorégressive" du signal, ou modélisation CARDH, elle estime la fréquence fondamentale par une méthode basée sur le principe du critère d'ordre.
Cette méthode dérive du "principe du MDL" initialement proposé par Kay et Nagesha, pour l'extraction d'un signal périodique dans du bruit coloré.
Dans ce chapitre, nous avons tout d'abord développé le principe du MDL pour l'estimation de la fréquence fondamentale, dans le cas de mélanges de signaux de parole. L'évaluation de ses performances sur des mélanges synthétiques, ont montré qu'il permet de diminuer les problèmes de détection de sous-harmoniques. Toutefois, le principe du MDL s'avère relativement coûteux, en termes de complexité de calcul.
Nous proposons alors d'utiliser les propriétés asymptotiques de l'estimateur du Maximum de Vraisemblance, afin de diminuer cette complexité. Par ailleurs, l'étude de plusieurs critères de sélection d'ordre, autres que le MDL, montre qu'un critère du type Hannan-Quinn donne de meilleurs résultats.
La méthode basée sur le critère d'ordre, estime la fréquence fondamentale f_0 du

modèle en deux étapes. D'abord une valeur sous-harmonique de f_0 est estimée par la méthode des Moindres Carrés, puis le principe du critère d'ordre, est utilisé pour déterminer son ordre harmonique.
L'évaluation de cette méthode montre des performances supérieures à une estimation utilisant le principe du MDL. Elle permet d'estimer la fréquence fondamentale du signal prédominant dans un mélange contenant jusqu'à cinq sources harmoniques. De plus, son faible coût de calcul permet d'envisager son exploitation dans une méthode d'estimation itérative de fréquences fondamentales multiples.

Chapitre 3

Estimation pour la séparation de signaux de parole

Après avoir mis en oeuvre une méthode d'estimation de fréquence fondamentale basée sur le principe du critère d'ordre, nous nous intéressons à présent à la séparation de signaux de parole. La séparation proposée dans la suite, analyse un mélange de signaux de parole à l'aide de la modélisation CARDH.
La méthode du critère d'ordre est utilisée sous la forme d'un estimateur itératif qui extrait successivement les différentes fréquences fondamentales d'un mélange de parole.
Les méthodes itératives, lorsqu'elles sont couplées à un détecteur de voisement, ont l'avantage d'estimer le nombre de sources d'un mélange. En revanche, les performances d'estimation, sont directement liées aux techniques permettant d'extraire les sources du mélange et qui sont le point faible de ce type de méthode.
Ce chapitre propose une nouvelle méthode d'estimation itérative, qui résout ce problème d'extraction. De plus, elle permet d'estimer directement le nombre de sources du mélange.
Le schéma général de la méthode de séparation est tout d'abord présenté. Par la suite, nous traiterons le problème de soustraction. Enfin, une dernière partie sera consacrée aux vraisemblances pénalisées, calculées par le critère de sélection. Nous présenterons l'intérêt de ces dernières, d'un part pour déduire la mesure de voisement, utilisable notamment pour estimer le nombre de sources voisées du mélange, et d'autre part pour la poursuite des fréquences fondamentales le long des fenêtres d'analyse.

3.1 Schéma général de la méthode de séparation

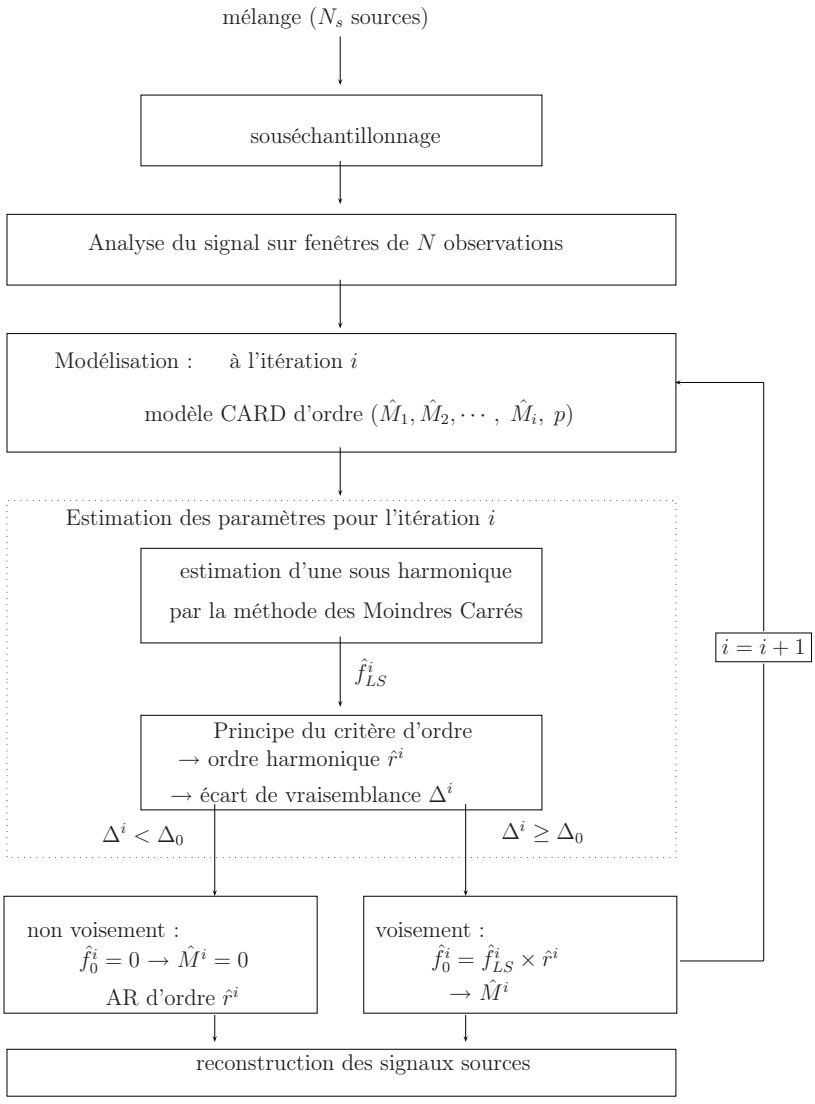


Fig. 3.1 – *Méthode de séparation basée sur le principe du critère d'ordre.*

• **sous-échantillonnage :**

Parce que l'intervalle fréquentiel dans lequel évolue la fréquence fondamentale f_0 des signaux étudiés est [80,400] Hz, le signal composite est préalablement sous-échantillonné. Ceci nous permet de limiter le nombre d'harmoniques de la

modélisation et de réduire la complexité : un signal de parole échantillonné par exemple à 8000 Hz, sera décimé d'un facteur 2 et le nombre maximum M d'harmoniques utilisables dans une modélisation sera tel que :

$$M \times 80 \leq 2000$$

• **Analyse :**

Le signal composite à traiter estanalysé à l'aide de fenêtres de Hamming se recouvrant avec un taux de 75%. La longueur de chaque fenêtre est de 50ms. Cette durée choisie deux fois plus longue que la durée généralement utilisée dans l'analyse d'un signal de parole simple (cas de la source unique), permet d'augmenter le nombre N d'observations, tout en tenant compte de la validité de l'hypothèse de stationnarité de la parole.

• **Modélisation du signal :**

Chaque fenêtre du signal d'entrée est analysée à l'aide de la modélisation CARDH définie par les expressions (eq. 2.1) et (eq. 2.2).
A chaque itération i, on estime les paramètres M^i et p du modèle CARDH dont l'ordre est donné par $(\widehat{M}^1, \widehat{M}^2, \cdots, M^{i-1}, \hat{M}^i, p)$. Les paramètres $(\widehat{M}^1, \widehat{M}^2, \cdots, \widehat{M}^{i-1})$ sont les estimations obtenues aux itérations $1, \cdots, (i-1)$. Ainsi, à chaque nouvelle itération, les fréquences fondamentales estimées aux itérations précédentes sont conservées.

• **Estimation des paramètres pour l'itération i :**

A l'itération i, la fréquence fondamentale $\hat{f}_0^i$ est obtenue par la méthode d'estimation basée sur le principe du critère d'ordre. Elle est telle que :

$$\widehat{f}_0^i = \widehat{f}_{LS}^i \times \widehat{r}^i$$

$\hat{f}_{LS}^i$ est la fréquence estimée par la méthode des moindres carrés (LS), recherchée en tant que sous-harmonique, sur l'intervalle réduit [80,160] Hz. $\hat{r}^i$ est l'ordre harmonique déterminé par le principe du critère d'ordre.

A partir des vraisemblances pénalisées calculées Δ^i, on déduit l'écart de vraisemblance entre le modèle CARDH estimé d'ordre $(\widehat{M}^1, \widehat{M}^2, \cdots, \widehat{M}^i, \widehat{p})$ et le modèle d'ordre $(\widehat{M}^1, \widehat{M}^2, \cdots, 0, \widehat{p})$.

• **Voisement/non-voisement :**

L'écart de vraisemblance est équivalent, comme on le verra par la suite, à une mesure de voisement. Ainsi, en introduisant une valeur seuil Δ_0, il est possible de détecter les parties purement voisées et purement non-voisées du signal :

– Voisement : pour $\Delta \geq \Delta_0$, les paramètres $\{\hat{f}_0^i, M^i\}$ sont validés comme étant les paramètres du modèle i et sont réutilisés à l'itération $(i+1)$. La fenêtre sur laquelle se fait l'analyse est supposée de nature voisée.

– Non-voisement : pour $\Delta < \Delta_0$, les paramètres $\{\hat{f}_0^i, M^i\}$ sont mis à zéro et la fenêtre analysée est supposée ne plus contenir de sources harmoniques à extraire. Le nombre total de sources harmoniques estimées est égal à $\hat{N}_s = (i-1)$.

• **Reconstruction :**

Par la suite, les $\hat{N}_s$ modèles sont synthétisés séparément à partir de leurs paramètres estimés : les paramètres harmoniques sont utilisés pour synthétiser la partie déterministe, tandis que l'ordre $\widehat{p}$ est utilisé pour la reconstruction des parties non-voisées.

3.2 Estimation itérative et sans soustraction

La méthode d'estimation de fréquence fondamentale utilisée dans la séparation est la méthode du critère d'ordre proposée dans le chapitre précédent. On rappelle alors ci-dessous, la manière dont cette dernière est utilisée sur une fenêtre d'analyse du mélange pour l'estimation de la fréquence fondamentale à l'itération i. Nous discuterons par la suite des avantages d'une telle méthode dans le cadre de la séparation de signaux de parole.

A l'itération i, la fréquence fondamentale f_0^i est estimée en deux étapes successives :

- **Estimation d'une valeur sous-harmonique de f_0^i dans l'intervalle [80,160]Hz** par la méthode des moindres carrés et qui consiste à minimiser l'expression (eq. 2.23).

- **Estimation de l'ordre harmonique** r^i. Le critère de sélection est utilisé pour déterminer l'ordre de $\widehat{f}^i_{0_{LS}}$:

  • on construit un jeu de R modèles CARDH où l'ordre du r^{ieme} modèle est donné par $(\{\widehat{M}^i\}_r\ ,\ p_r\ )$. L'ensemble $\{\widehat{M}^i_r\} = \{\widehat{M}^1_r, \widehat{M}^2_r, \cdots, \widehat{M}^i_r\}$ dépend, par l'intermédiaires de $\widehat{M}^1_r$, $\widehat{M}^2_r$, $\cdots$, $\widehat{M}^{(i-1)}_r$, des fréquences fondamentales estimées aux itérations précédentes.

  • le nombre d'harmoniques pour la i^{ieme} fréquence fondamentale est donné par :

$$\widehat{M}^i_r = \frac{1}{2\widehat{f}^i_{0_{LS}} \times r^i} \tag{3.1}$$

  • l'entier r^i et le nombre de paramètres AR p^r_i sont tels que :

$$\begin{cases} r_i\ \epsilon[0,\ r_{max}] & et\ r_{max} \leq (400/\widehat{f}^i_{0_{LS}}) \\ \\ p^i_r\ \epsilon[0,\ p_{max}] & et\ p_{max} = 20 \end{cases} \tag{3.2}$$

  • l'ordre harmonique est déterminé par sélection du modèle dont le terme de log-vraisemblance pénalisée sera minimale, dans le jeu des R modèles :

$$(\widehat{r}^i, \widehat{p}^i_r) = arg \min_{r,p_r}[L(r, p_r)] \tag{3.3}$$

$$L(r, p_r) = (N - p_r)\, ln(\pi e \frac{C'(\widehat{f}_{0_{LS}} \times r)}{N - p_r}) + Q_r\ \beta \tag{3.4}$$

  • le paramètre β lié au critère d'ordre est donné par :

$$\beta = 3ln(ln(N - p_r)) \tag{3.5}$$

  • Q_r est le nombre de paramètres du modèle, tel que :

$$Q_r = (2M_r + 1 + p_r) \tag{3.6}$$

– **La fréquence fondamentale estimée à l'itération i** est finalement donnée par :

$$\widehat{f}_0^i = \widehat{r}^i \times \widehat{f}_{0_{LS}}^i \tag{3.7}$$

Nous avons vu dans le premier chapitre, présentant l'état de l'art des méthodes d'estimation de fréquences fondamentales multiples, que ces dernières se rangent en deux classes : les méthodes itératives et les méthodes d'estimation conjointe. Les premières sont généralement moins lourdes en calcul et ont l'avantage d'être utilisées sans connaissance à priori du nombre N_s de sources à extraire. N_s peut être estimé simplement en utilisant par exemple un détecteur de voisement, qui stoppe l'estimation à l'itération pour laquelle le mélange est supposé ne plus contenir de sources voisées.
L'inconvénient des méthodes itératives est qu'elles font appel à une étape de soustraction qui consiste à extraire du mélange, les paramètres du modèle avant chaque nouvelle itération. Bien qu'il existe plusieurs types de soustraction, aussi bien dans le domaine temporel que fréquentiel, cette étape est dépendante des estimations et dégrade à chaque nouvelle itération le mélange. Les méthodes d'estimations conjointe, quant à elles, évitent le problème de soustraction en estimant conjointement les paramètres de toutes les sources du mélange. L'inconvénient majeur de ce type de méthodes est qu'elles nécessitent la connaissance du nombre de sources N_s.
L'estimation de fréquences fondamentales proposée ici, est un compromis de ces deux types d'estimation : sa structure itérative permet de combler le manque d'information sur le nombre de sources harmoniques mélangées, tandis que l'introduction des fréquences estimées $\{\hat{f}_0^1, \cdots, \hat{f}_0^{(i-1)}\}$ à l'itération i, permet d'éviter toute étape de soustraction.

3.3 Ecart de vraisemblance Δ - Mesure de voisement

Lorsque le nombre N_s de sources voisées du mélange est inconnu, il est possible d'utiliser conjointement à une méthode d'estimation itérative de fréquences fondamentales multiples, un détecteur de voisement qui analyse le signal étudié et stoppe les itérations lorsque ce dernier ne possède plus de sources voisées.
Dans cette section, nous proposons d'utiliser un détecteur de voisement qui uti-

lisent les vraisemblances pénalisées calculées lors de l'estimation des fréquences fondamentales. Dans cette section, ce détecteur, appelé "écart de vraisemblance" est tout d'abord défini, puis son utilisation pour les mesures de voisement, ainsi que pour la poursuite des fréquences fondamentales le long des fenêtres d'analyse est détaillée. Pour finir, quelques illustrations de cette mesure pour des mélanges de signaux synthétiques sont présentées.

3.3.1 Définition de l'écart de vraisemblance

L'écart de vraisemblance Δ lié à l'estimation de la fréquence fondamentale d'un modèle CARDH d'ordre harmonique $\hat{r}$ et dont la partie autorégressive possède $\hat{p}_r$ coefficients, est défini par :

$$\Delta = |L(\widehat{r}, \widehat{p}_r) - L(0, \widehat{p}_r)| \tag{3.8}$$

où $L(r, p_r)$ est la log-vraisemblance d'expression (eq. 3.4).

Δ calcule l'écart de vraisemblance entre le modèle CARDH estimé possédant $\widehat{p}_r$, et un simple modèle autorégressif de même ordre.

3.3.2 Mesure de voisement et poursuite des fréquences

Il est possible d'utiliser un seuil de discrimination Δ_0, en dessous duquel le modèle estimé possède une vraisemblance pénalisée trop proche de celle d'un bruit autorégressif de même ordre, pour considérer ses paramètres harmoniques comme des paramètres valides. Cette non-validité se justifie de deux manières différentes :

- si le signal analysé est de nature non-voisée, la fréquence estimée ne représente rien et peut être mise à zéro.

- si le signal analysé est de nature voisée, la fréquence estimée est très certainement incorrecte. Dans ce cas, nous ne pouvons pas nous fier à cette valeur.

De ces deux suppositions découle alors une mesure de voisement qui sera utilisée pour estimer le nombre de sources harmoniques du mélange et un critère de

fiabilité de l'estimation, utilisé pour la poursuite des fréquences fondamentales le long des trames d'analyse du signal.

• **Mesure de voisement**

A l'itération i et pour un seuil Δ_0 fixé :
– si $(\Delta^i < \Delta_0)$, alors le mélange est non-voisé et $\hat{f}_0^i = 0$.
On stoppe l'extraction des sources voisés.

– si $(\Delta^i \geq \Delta_0)$, alors le mélange est de nature voisé et le modèle CARDH estimé est validé. On effectue une nouvelle itération qui prend en compte les fréquences fondamentales $\{\hat{f}_0^1, \cdots, \hat{f}_0^i\}$.

• **Poursuite des fréquences**

Sur la fenêtre d'analyse j, à l'itération i et pour un seuil Δ_0' fixé :
– si $((\Delta^i)_j < \Delta_0')$, alors la fréquence $(\hat{f}_0^i)_j$ est supposée peu fiable et ne peut être utilisée pour déterminer la fréquence fondamentale $(f^i_{0_{(j+1)}})$de la fenêtre suivante.

– si $((\Delta^i)_j \geq \Delta_0')$, alors la fréquence $(\hat{f}_0^i)_j$ est supposée fiable et est utilisée pour déterminer la fréquence fondamentale de la fenêtre suivante :
$(f_0^i)_{j+1}$ est obtenue à partir d'une valeur sous-harmonique $(f_{LS}^i)_{j+1}$ recherchée au voisinage de la fréquence $(f_{LS}^i)_j$, telle que :

$$(f_{LS}^i)_{j+1} \in [(f_{LS}^i)_j - 10Hz\ ;\ (f_{LS}^i)_j + 10Hz]$$

La fiabilité d'une estimation étant supposée proportionnelle à la valeur de l'écart de vraisemblance correspondant, Δ_0' doit être choisi le plus grand possible, pour ne garder que les fréquences les plus fiables.

Comme de nombreuses méthodes d'analyse de signaux de parole, la méthode de séparation par le critère d'ordre, utilise l'hypothèse de variation lente des fréquences fondamentales pour introduire une étape de poursuite le long du signal analysé. En revanche, la méthode du critère d'ordre effectue cette poursuite en

utilisant non pas les paramètres estimés, mais la fiabilité accordée aux estimations données par l'écart de vraisemblance.

3.3.3 Ecart de vraisemblance pour les signaux synthétiques

Nous nous intéressons ici plus particulièrement au comportement de l'écart de vraisemblance pour des mélanges de signaux synthétiques. Une présentation détaillée de ces mélanges a été proposée dans le chapitre sur l'estimation de la fréquence fondamentale (cf. paragraphe 2.3.2).
Les simulations qui suivent, mettent en évidence l'information donnée par l'écart de vraisemblance sur la fiabilité d'une estimation fréquentielle, ainsi que sur la nature de la source correspondante. L'écart de vraisemblance est tout d'abord déterminé pour le signal prédominant d'un mélange donné. On s'intéressera ensuite, à l'écart de vraisemblance obtenu pour la seconde fréquence estimée par la méthode du critère d'ordre, dans un mélange de deux sources.

• Ecart de vraisemblance d'un signal harmonique bruité

On considère un signal harmonique de fréquence fondamentale réduite $f_0 = 0.074$ dans un bruit autorégressif d'ordre 10 et dont la densité spectrale est illustrée par la figure (Fig. 2.2). Pour un rapport signal à bruit fixé (RSB en dB), on détermine à partir d'un jeu de $N_b = 200$ simulations, la valeur moyenne de l'écart de vraisemblance Δ obtenue pour les fréquences fondamentales estimées par la méthode du critère d'ordre.
La figure (Fig. 3.2) présente les valeurs moyennes de Δ pour les fréquences correctement estimées et les fréquences incorrectes, en fonction du rapport signal à bruit (RSB). On rappelle qu'une fréquence est supposée correctement estimée lorsque l'écart entre sa valeur et la véritable fréquence fondamentale reste inférieure à la quantité 5.10^{-4} (cf. paragraphe 2.3.2).

Pour une analyse du signal utilisant des fenêtres de $N = 200$ observations, les fréquences sont toutes correctement estimées lorsque le RSB reste supérieur ou égal à 5 dB (les pourcentages de fréquences correctes sont égaux à 100% pour ces RSB et il en résulte qu'aucun écart pour les fréquences incorrectes n'est défini lorsque le $RSB \geq 5$ dB). La valeur moyenne de Δ pour ces fréquences

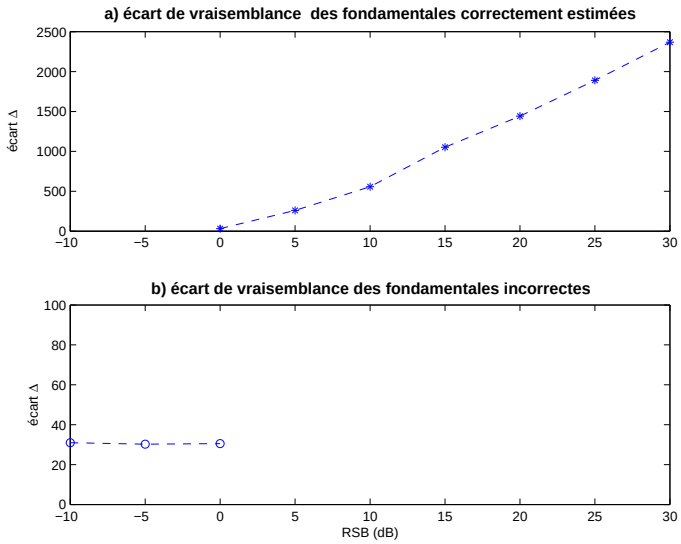


Fig. 3.2 – *écart de vraisemblance moyen : a) pour les fréquences fondamentales correctement estimées (⋆)- b) pour les fréquences fondamentales incorrectes (o). Cas d'un signal harmonique de fondamentale* $f_0 = 0.074$ *dans un bruit AR d'ordre 10.*

reste supérieure à 200. Pour les fréquences dont le RSB est nul, ainsi que pour les fréquences incorrectes, le Δ moyen est aux alentours de 30. Cette dernière remarque implique que pour des niveaux de bruit importants, il est très difficile de discriminer une fréquence incorrecte d'une fréquence correcte par le biais de Δ.

• Ecart de vraisemblance d'un mélange de signaux harmoniques

On s'intéresse ici à l'écart de vraisemblance de la fréquence fondamentale estimée lors de la première itération pour un mélange de plusieurs sources harmoniques. On présente le cas d'un mélange de 2, puis de 5 sources.
Les mélanges sont caractérisés par leur RSS défini comme le rapport de puissance entre un premier signal, dit prédominant et l'ensemble des autres sources du mélange (cf. (eq. 2.29)).
Les fréquences fondamentales sont respectivement $f_{0_1} = 0.074$ et $f_{0_2} = 0.0205$ pour le mélange contenant 2 sources. Dans le cas du mélange contenant 5 sources, les fréquences sont $f_{01} = 0.074$, $f_{02} = 0.0715$, $f_{03} = 0.0665$, $f_{04} = 0.059$, $f_{05} = 0.0515$ et $f_{06} = 0.0465$.
Pour les deux mélanges considérés, on détermine la valeur de l'écart de vraisemblance Δ correspondant aux fréquences fondamentales du signal prédominant

correctement estimées par la méthode du critère d'ordre, pour des RSS variant de 0 à 30dB. La figure (Fig. 3.3) présente les résultats obtenus.

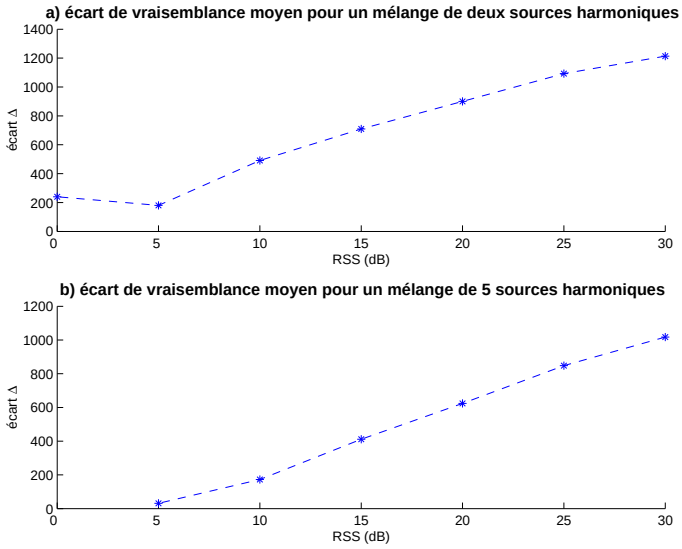


Fig. 3.3 – *écart de vraisemblance moyen en fonction du rapport de puissance RSS : a) cas d'un mélange de 2 sources harmoniques - b) cas d'un mélange de 5 sources harmoniques.*

Au fur et à mesure que le niveau du signal prédominant augmente, son écart de vraisemblance augmente, et ce aussi bien dans le cas d'un mélange de 2 signaux que de 5. Le minimum est alors obtenu, dans les deux cas pour un $RSS = 0$ dB. En revanche sa valeur diffère suivant le type de mélange : plus il y a de sources, plus l'écart de vraisemblance de la source prédominante diminue. Ainsi, pour un mélange de 2 sources cette valeur reste très proche de 200 lorsque les sources sont au même niveau. En comparaison, il faut un RSS d'au moins 10 dB dans le mélange de 5 sources pour obtenir un Δ proche de 200.

L'ensemble de ces simulations, montrent finalement que quelle que soit la nature des signaux sources perturbant un signal, c'est-à-dire que ce soit du bruit ou un mélange de plusieurs sources harmoniques, l'écart de vraisemblance est porteur d'information. Les fréquences fondamentales correctement estimées du signal prédominant, se distinguent par de fortes valeurs de Δ, pour des RSS supérieurs à 5 dB : pour une analyse sur $N = 200$ observations d'un mélange de 2 sources, on obtient des $\Delta \geq 200$ pour les fréquences correctement estimées, alors que l'écart associé aux fréquences incorrectes reste en dessous de 50.
En revanche, pour des rapports signaux à bruit faibles, l'écart de vraisemblance ne

permet plus de discriminer une fréquence fondamentale correcte, d'une fréquence fondamentale incorrecte. De plus, le fait que ces valeurs d'écart interviennent aussi bien dans le cas où le signal perturbateur est de nature harmonique ou non, implique qu'on ne peut supposer la nature réelle du signal détecté. Pour des $\Delta \geq 200$, on peut finalement supposer que la source détectée est harmonique et que sa fréquence fondamentale a été estimée correctement.

• **Ecart de vraisemblance du second signal extrait**

On considère le cas du mélange de 2 signaux harmoniques de fréquences $f_{01} = 0.074$ et $f_{02} = 0.0205$. La figure (Fig. 3.4) présente tout d'abord le pourcentage de fréquences estimées par la méthode du critère d'ordre pour deux itérations successives, en fonction du rapport de puissance entre les sources.

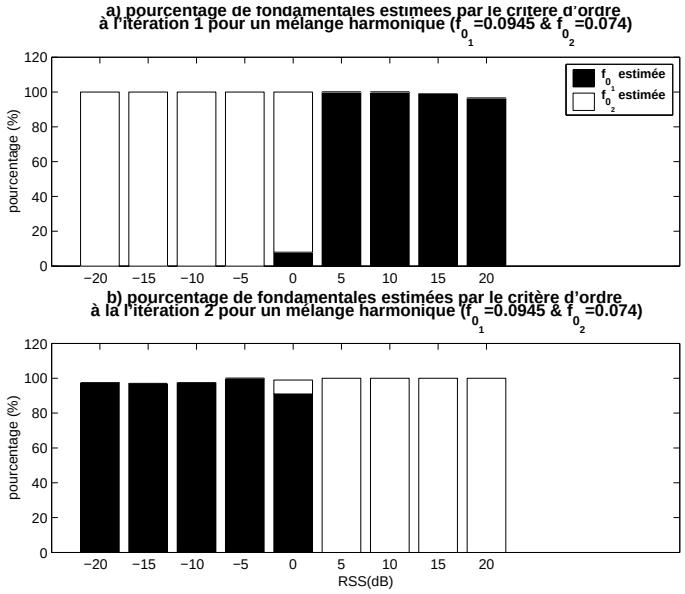


Fig. 3.4 – fréquences estimées par la méthode du critère d'ordre pour un mélange de deux signaux de fréquences $f_{01} = 0.074$ et $f_{02} = 0.0205$: a) fréquences estimées à l'itération 1 - b) fréquences estimées à l'itération 2.

Les deux fréquences fondamentales du mélange sont estimées sans difficulté par la méthode du critère d'ordre et ce quelque soit la valeur du RSS. La figure (Fig. 3.4) illustre bien le fait que la première estimation va généralement correspondre au signal prédominant dans le mélange. Les écarts de vraisemblance correspondant aux fréquences fondamentales estimées sont présentés sur la figure (Fig. 3.5)

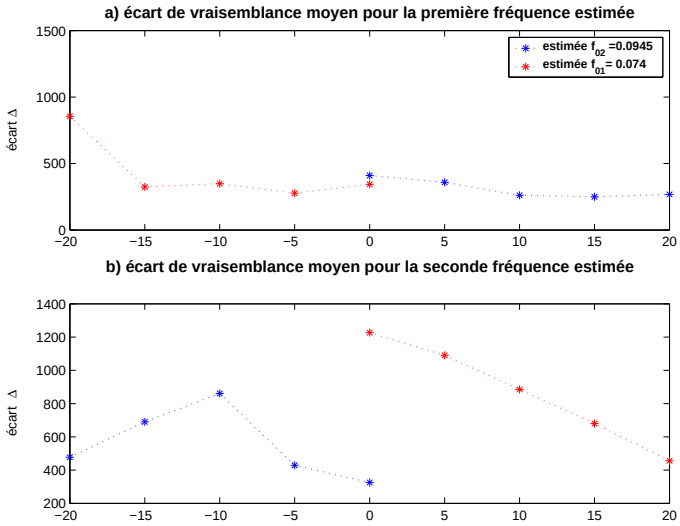


Fig. 3.5 – *écart de vraisemblance moyen en fonction du rapport de puissance RSS : a) cas du mélange de deux sources harmoniques , b) cas du mélange de cinq sources harmoniques.*

Tout comme précédemment, on remarque que l'écart de vraisemblance obtenu aux deux itérations ne descend jamais en dessous de la valeur 200. Ainsi, l'écart de vraisemblance permet de mettre en évidence la présence d'au moins deux sources voisées dans le mélange étudié.

3.4 Application à la séparation de signaux réels

Après avoir traité le cas de mélanges de signaux synthétiques, nous consacrons la fin de ce chapitre à la séparation de signaux de parole proprement dits. Nous nous intéresserons dans un premier temps à la détection des parties voisées et non-voisées d'un signal de parole, à l'aide de l'écart de vraisemblance. Par la suite, nous étudierons la méthode de séparation sur des mélanges de deux ou trois signaux de parole, en tenant compte des rapports de puissance entre les sources. Pour finir, nous présenterons le cas d'un mélange réel de parole, pour lequel le rapport de puissance entre les deux reste inconnu.

3.4.1 Mesure de voisement

Afin de mettre en évidence les capacités du détecteur de voisement qu'est l'écart de vraisemblance Δ, la méthode du critère d'ordre est appliquée à un signal de parole simple, possédant de parties voisées et de parties non-voisées.
Le signal présenté sur la figure (Fig. 3.6) est analysé sur des fenêtres d'une durée de 50 ms, se recouvrant à 75 %. L'écart de vraisemblance est déterminé pour chacune des estimations de fréquence fondamentale à l'aide de l'expression (eq. 3.8).

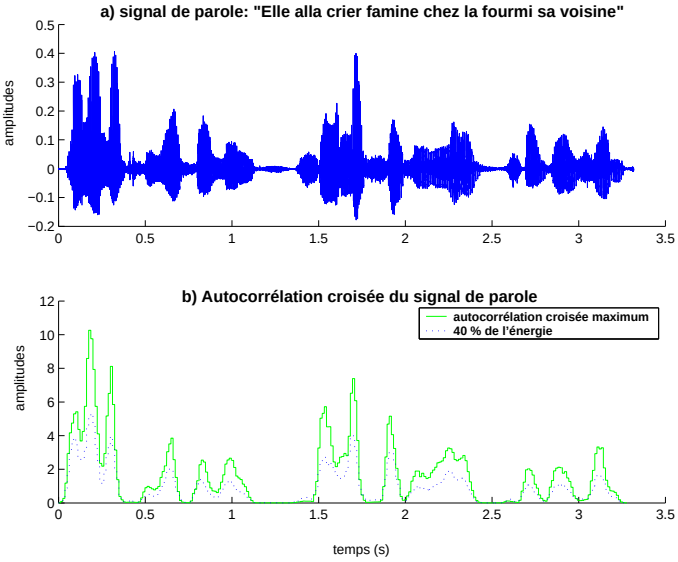


Fig. 3.6 – *a) Signal de parole comportant des parties voisées et non-voisées - b) Autocorrélation croisée du signal et seuil énergétique* à 40%.

Une méthode basique de détection de voisement est utilisée pour calculer la valeur moyenne des écarts finalement obtenue pour chaque type de voisement. Cette dernière utilise un seuil énergétique, au dessus duquel le signal est considéré comme voisé. Plus précisément, sur chaque fenêtre d'analyse le maximum de l'autocorrélation croisée du signal de parole est déterminée. On rappelle que l'autocorrélation croisée, permet de mesurer le degré de similarité entre deux versions décalées du signal appartenant à des fenêtres d'analyse différentes (cf. Etat de l'art). La détection est telle que si l'autocorrélation croisée est supérieure à 40% de l'énergie du signal, alors le signal est supposé de nature voisée. La figure (Fig. 3.6) présente également l'autocorrélation croisée du signal, ainsi que le seuil énergétique de 40%.

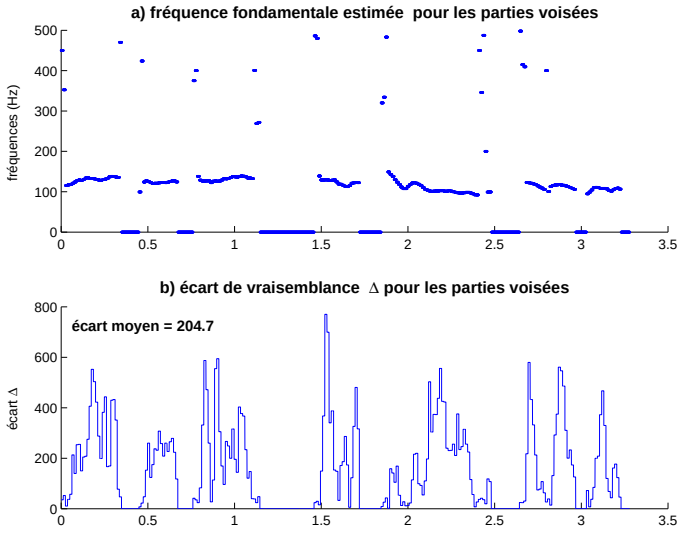


Fig. 3.7 – *fréquences fondamentales estimées et écarts de vraisemblance calculés par la méthode du critère d'ordre pour les parties voisées d'un signal de parole.*

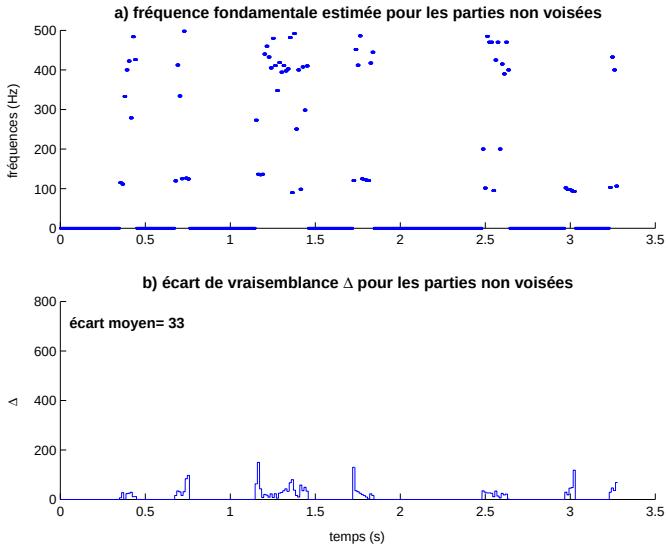


Fig. 3.8 – *fréquences fondamentales estimées et écarts de vraisemblance calculés par la méthode du critère d'ordre pour les parties non-voisées d'un signal de parole.*

Les figures (Fig. 3.7) et (Fig. 3.8) présentent les fréquences fondamentales et écarts de vraisemblance, respectivement pour les parties voisées et les parties non-voisées du signal. Les valeurs moyennes des écart de vraisemblance pour les deux cas sont respectivement $\Delta_v = 204.7$ et $\Delta_{nv} = 33$.

La comparaison de Δ_v et Δ_{nv} aux écarts de vraisemblance des signaux synthétiques, montrent que l'on obtient pour les parties voisées, des écarts plus faibles dans le cas des signaux réels. Ceci peut s'expliquer par le fait que l'on traite à présent des signaux quasi-harmoniques. On remarquera, malgré tout que le Δ des parties voisées atteint très souvent des valeurs supérieures à 200.

3.4.2 Etude comparative

Dans la suite de ce chapitre, la méthode de séparation proposée est appliquée à des mélanges synthétiques et réels de signaux de parole. Les mélanges synthétiques correspondent à des mélanges de signaux de parole, enregistrés séparément et puis mélangés.
Afin de mettre en évidence les performances et les limitations de la méthode, nous développons également dans cette partie, une méthode de séparation témoin basée sur le calcul de l'autocorrélation. Cette dernière sera également appliquée tout au long de l'étude aux différents mélanges traités.

• Méthode de séparation témoin basée sur l'autocorrélation

Cette méthode témoin est basée sur l'estimation itérative des fréquences fondamentales du mélange par la méthode de l'autocorrélation (cf. Etat de l'art).
A l'itération 1, une première fréquence fondamentale est estimée. Un modèle harmonique est alors synthétisé. Les amplitudes du modèle sont estimées par la méthode des moindres carrés. Le modèle est par la suite soustrait au mélange, donnant un signal résiduel à partir duquel une nouvelle estimation est faite.

• **"Pitch" de référence**

Dans le cas de mélanges synthétiques, les signaux sources sont à notre disposition. Une première étape consiste à déterminer les fréquences fondamentales de chaque source. Ces dernières sont choisies comme référence et vont permettre de quantifier la qualité des estimations calculées par la méthode du critère d'ordre et par conséquent celle de la séparation.

Les pitch de référence sont estimés par la méthode de l'Autocorrélation, méthode "monopitch" très utilisée et simple à mettre en oeuvre. Parce que cette dernière est également sensible aux erreurs d'octaves (cf. Etat de l'art), un post-traitement a été effectué. Ce dernier permet de lisser les contours de pitch afin d'obtenir une référence correcte [52].

On notera pour finir que dans le cas des mélanges instantanés, nous ne disposons pas des signaux sources. Il nous est impossible d'utiliser cette notion de pitch de référence.

• **Etude quantitative - mise en oeuvre**

L'utilisation d'une méthode de séparation témoin a pour but de quantifier les performances de la séparation basée sur le principe du critère d'ordre.
L'étude qui suit fait intervenir deux types de mélanges différents. On étudiera, tout d'abord des mélanges de signaux entièrement voisés, puis un mélange de deux sources qui possèdent toutes deux des parties non-voisées.
Quel que soit le mélange, l'aspect quantitatif de cette étude est présenté au moyen de différents pourcentages d'estimations correctes, calculés à partir des pitch de référence. On déterminera ainsi le pourcentage de fréquences fondamentales correctement estimées aux diverses itérations de la méthode de séparation, ainsi que dans le cas de mélanges mixtes "voisé/non voisé", le pourcentage de fausses alarmes et de non-détections.
Une estimation par la méthode du critère d'ordre sera supposée correcte lorsque l'écart entre l'estimation et le pitch de référence est inférieure à 3% de cette référence. Ce pourcentage est celui généralement choisi pour l'analyse des signaux musicaux. Il correspond en effet à un intervalle égal au quart de ton dans l'échelle musicale et permet de discriminer deux notes séparées d'un demi-ton. Ce pourcentage reste dans le cas de la parole, une bonne approximation pour l'estimation correcte de fréquences fondamentales.
Afin de mettre en évidence l'influence de l'écart de vraisemblance Δ, directement

lié au taux de détection des fréquences fondamentales, différentes valeurs seuil Δ_0 variant entre 0 et 140 seront prises en compte. On rappelle qu'une fréquence possédant un écart de vraisemblance inférieur Δ_0 est alors décidée non-voisée et sa valeur est mise à zéro.

3.4.3 Application à des mélanges synthétiques voisés

On s'intéresse ici à un ensemble de voix chantées : une voix de soprano, un alto et un ténor. Les formes d'onde de chaque voix sont présentées sur la figure (Fig. 3.9). Ce type de signal, met à notre disposition des sons de nature presque entièrement voisée sur une durée assez longue, facilitant ainsi l'étude détaillée du problème de séparation.
Bien que ces trois signaux de parole sont, contrairement aux études précédentes des signaux réels, les mélanges sont dits synthétiques, car ils sont obtenus par sommation des sources et ajustement de leur niveau.

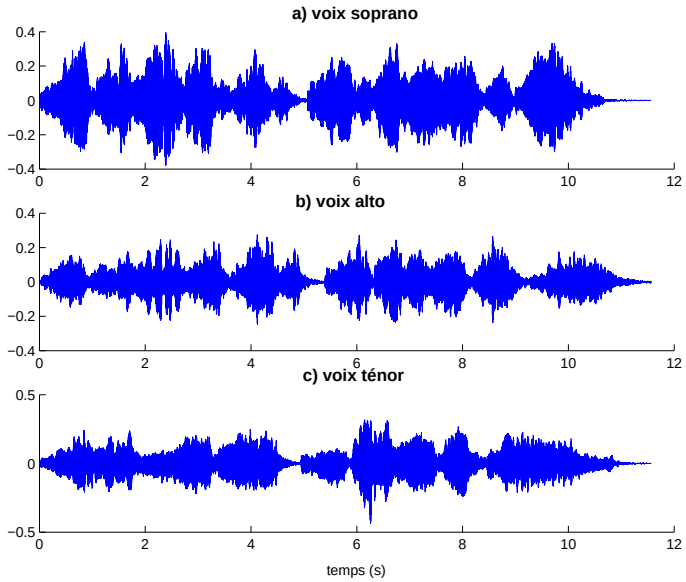


Fig. 3.9 – **Forme d'onde :** a) voix soprano b) voix alto c) voix ténor.

• **Estimation de la fréquence fondamentale des sources isolées**

Dans un premier temps, les différents signaux source sont étudiés séparément. La fréquence fondamentale de chaque signal est estimée par la méthode du critère d'ordre. Les résultats de cette méthode d'estimation sont comparés à ceux obtenus par la méthode de l'autocorrélation.
La figure (Fig. 3.10) présente les fréquences fondamentales estimées par la méthode de l'autocorrélation pour les trois voix. Ces estimations sont prises comme pitch de référence.
Les fréquences fondamentales estimées par la méthode du critère d'ordre sont présentées sur les figures (Fig. 3.11) et (Fig. 3.12), qui correspondent respectivement à une estimation "sans suivi" et "avec suivi" des fréquences fondamentales le long des fenêtres d'analyse. Ce suivi fréquentiel ou poursuite, consiste à prendre en compte la fiabilité d'une fréquence estimée sur une fenêtre donnée, pour estimer celle de la fenêtre suivante. Elle est effectuée à partir des écarts de vraisemblance associés aux estimations. La mise en oeuvre de cette poursuite est détaillée dans le paragraphe (3.3.2).

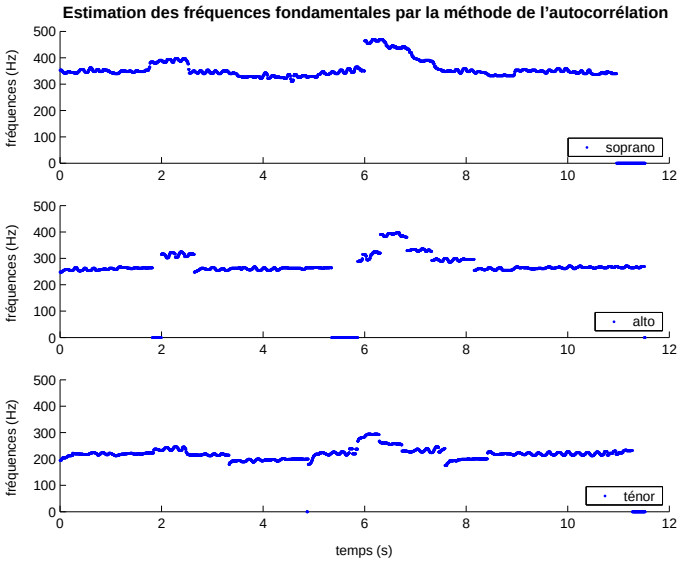


Fig. 3.10 – *Pitch de référence fréquences fondamentales estimées par la méthode de l'autocorrélation : a) voix soprano b) voix alto c) voix ténor.*

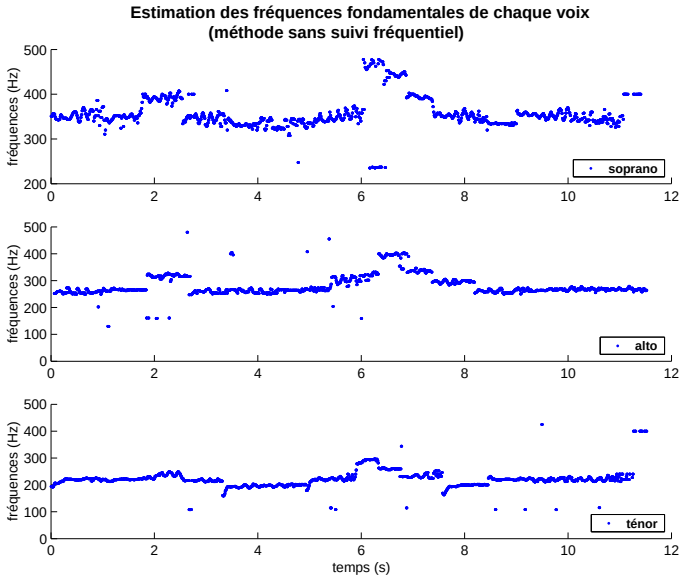


Fig. 3.11 – *fréquences fondamentales estimées par la méthode du critère d'ordre sans poursuite des fréquences fondamentales : a) voix soprano b) voix alto c) voix ténor.*

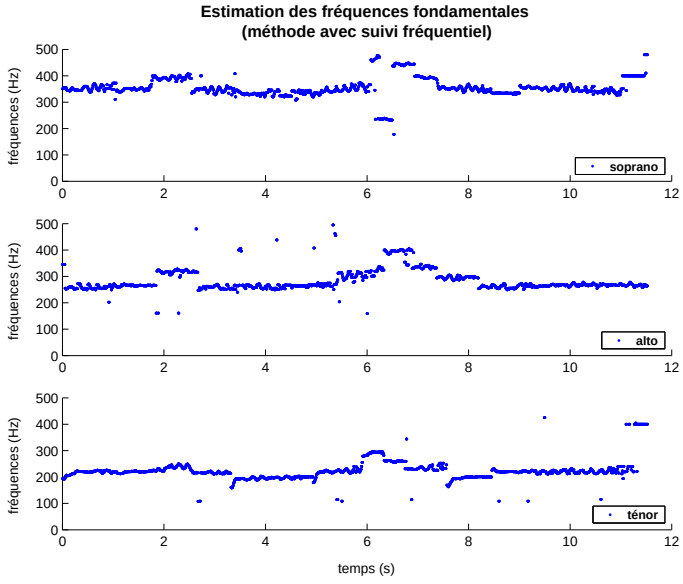


Fig. 3.12 – *fréquences fondamentales estimées par la méthode du critère d'ordre avec poursuite des fréquences fondamentales : a) voix soprano b) voix alto c) voix ténor.*

On remarque alors qu'avec ou sans suivi, la méthode du critère d'ordre permet d'estimer correctement les pitch des trois voix, bien que l'on voit apparaître quelques artefacts dûs en général à une mauvaise estimation de l'ordre harmonique.

Afin de quantifier les performances de la méthode de séparation par le critère d'ordre, on détermine le pourcentage $P(\%)$ de fréquences fondamentales correctes. Ce pourcentage est défini comme suit :

$$P(\%) = \frac{(\textit{nombre de fondamentales} = \textit{pitch référent})}{\textit{nombre total de fondamentales}} \times 100\% \qquad (3.9)$$

L'égalité entre fréquence fondamentale et pitch référent est valable pour une précision de 3% par rapport à la fréquence de référence.

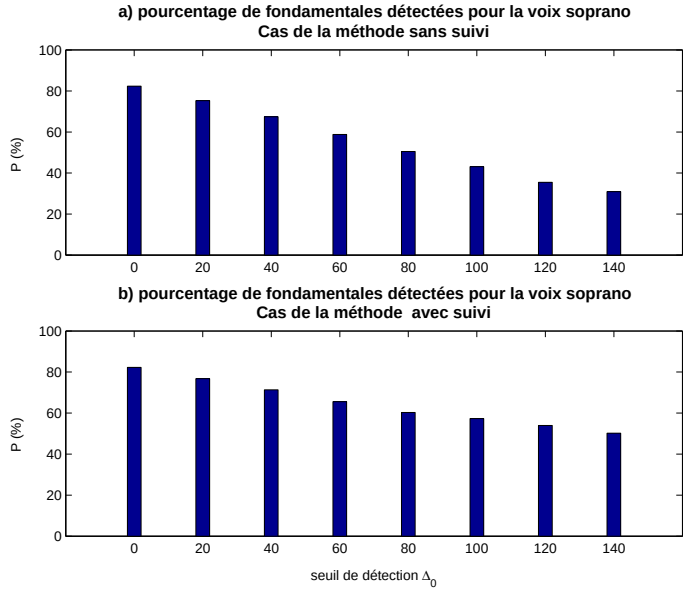


Fig. 3.13 – *Comparaison de la méthode du critère d'ordre : a) sans suivi - b) avec suivi. Pourcentage de fréquences correctes $P(\%)$ pour la voix soprano.*

L'ensemble des pourcentages $P(\%)$ obtenus pour les différentes voix montre que si Δ_0 est choisi nul, toutes les fréquences sont détectées et le pourcentage de

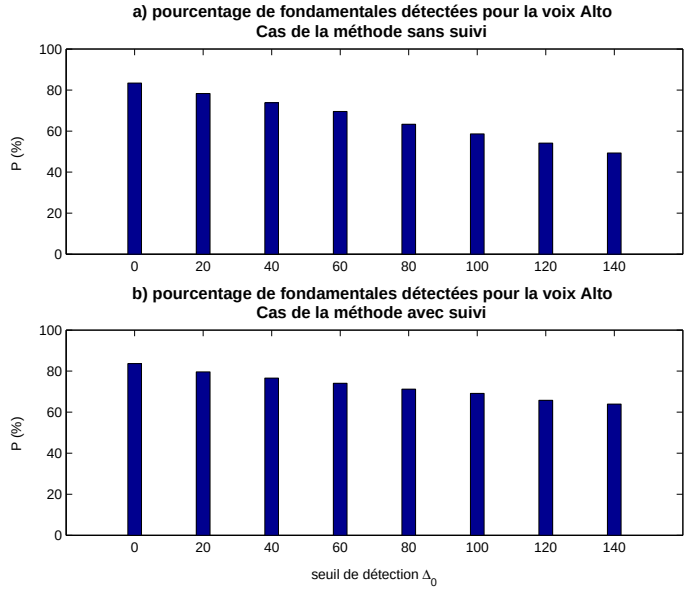


Fig. 3.14 – *Comparaison de la méthode critère d'ordre : a) sans suivi - b) avec suivi. Pourcentage de fréquences correctes $P(\%)$ pour la voix alto.*

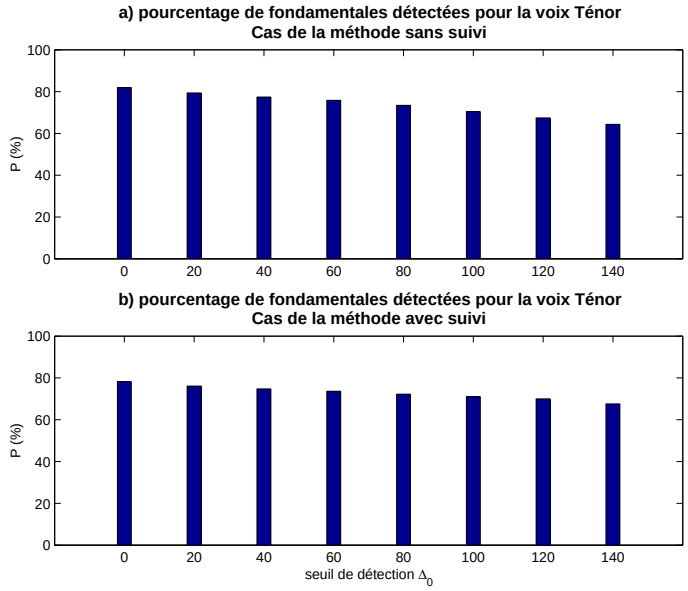


Fig. 3.15 – *Comparaison de la méthode critère d'ordre : a) sans suivi - b) avec suivi. Pourcentage de fréquences correctes $P(\%)$ pour la voix ténor.*

fréquences correctes est aux alentours de 80% par rapport au pitch de référence. Le choix d'un Δ_0 trop grand, limite la détection de fréquences, car le nombre de fréquences qui ont un écart $\Delta \geq 140$ est beaucoup plus faible que celles ayant un écart $\Delta \geq 40$. En revanche, la comparaison des cas avec et sans suivi, montre que lorsqu'on utilise la poursuite des fréquences, les pourcentages P(%) sont plus élevés lorsque le seuil Δ_0 augmente.
La méthode avec "suivi fréquentiel" permet donc d'utiliser des seuils plus fort qui vont faciliter la discrimination des parties voisées et non-voisées. L'utilisation de la poursuite des fréquences, permet également d'éliminer des erreurs d'octave : si deux fréquences estimées sur des fenêtres successives ont la même estimation de sous-harmonique (f_{LS}), mais des ordres harmoniques différents, alors l'ordre choisi pour les deux fréquences est celui qui possède l'écart de vraisemblance le plus fort.

• **Mélange de 2 sources**

Dans le cas d'un mélange de 2 des voix chantées, on peut comme précédemment comparer les estimations obtenues par la méthode du critère d'ordre avec les pitch de référence. Le cas du mélange des voix soprano et ténor est présenté ci-dessous. La poursuite des fréquences à l'aide de l'écart de vraisemblance est utilisée.

Comparaison avec une version utilisant la soustraction : On s'intéresse plus particulièrement au caractère de non-soustraction de la méthode du critère d'ordre : la version "sans soustraction", proposée dans le paragraphe (3.2) est comparée à sa version "avec soustraction". Cette dernière consiste à synthétiser le premier modèle harmonique estimé, puis à le soustraire au mélange. Le signal résiduel ainsi obtenu est à nouveau analysé pour estimer la seconde source harmonique.
La figure (Fig. 3.16) présente alors les pourcentages $P_2(\%)$ de fréquences fondamentales estimées pour la seconde source, dans les cas "sans soustraction" et "avec soustraction". Le pourcentage $P_1(\%)$ de fréquences fondamentales de la première source harmonique, obtenu lors de la première itération de la méthode du critère d'ordre est celui obtenu pour $\Delta_{0_1} = 0$ et est égal à $P_1(\%) = 66.3$.
La comparaison des deux cas, montre que l'étape de soustraction entraîne une diminution du pourcentage de fréquences correctes $P_2(\%)$, lorsque l'écart augmente. Ainsi pour un même seuil de détection, la méthode sans soustraction va permettre l'estimation d'un plus grand nombre de fréquences fondamentales.

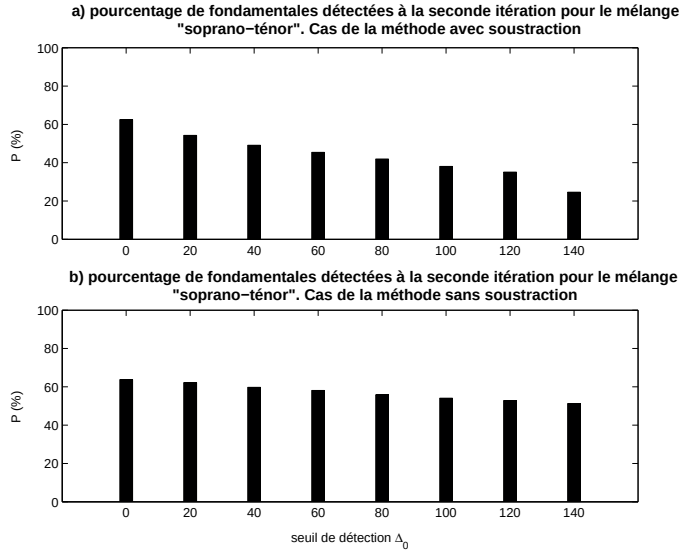


Fig. 3.16 – *Comparaison de la méthode du critère d'ordre : a) sans soustraction - b) avec soustraction. Pourcentage de fréquences estimées pour le mélange vocal "soprano/ténor". Le pourcentage obtenu à la première itération pour* $\Delta_{01} = 0$ *est* $P_1(\%) = 66.3$.

Comparaison avec la méthode de séparation témoin : la méthode basée sur le critère d'ordre est ici comparée à la méthode témoin. La figure (Fig. 3.17) présente les résultats obtenus pour des seuils de détection lors de chaque estimation itérative $\Delta_{0_1} = 0$ et $\Delta_{0_2} = 0$: les performances de cette méthode pour la seconde itération, apparaissent alors très supérieures à celles de la méthode témoin.
Bien que la méthode témoin utilise une étape de soustraction qui est susceptible de dégrader les performances de cet estimateur à chaque nouvelle itération, la comparaison avec la méthode du critère d'ordre utilisant cette même étape de soustraction montre qu'elle n'est pas l'origine principale du faible pourcentage de fréquences estimées. On peut en effet voir sur la figure (Fig. 3.16), que l'on obtient toujours un meilleur pourcentage à la seconde estimation pour un seuil $\Delta_{0_2} = 0$. On peut finalement supposer que l'estimateur basée sur la méthode de l'autocorrélation et utilisé dans la méthode témoin, est peu performant dans le cas présent.

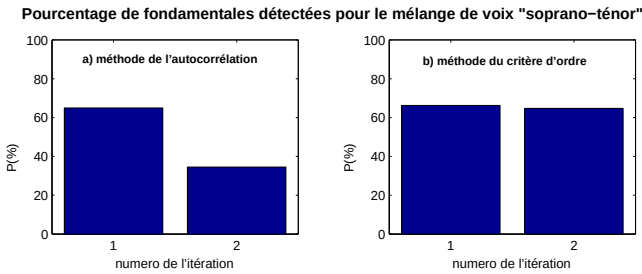


Fig. 3.17 – *Comparaison avec la méthode témoin. Cas du mélange des voix "soprano/ténor" Pourcentage de fréquences fondamentales estimées aux itérations successives : a) par la méthode de l'autocorrélation - b) par la méthode du critère d'ordre.*

Etude en fonction du Rapport de puissance des deux sources : Etant donné que les signaux de voix chantées sont à l'origine enregistrés séparément, il est possible de traiter des mélanges pour lesquels les rapports de puissance ont été préalablement choisis. On s'intéresse alors au comportement de la méthode du critère d'ordre, lorsque ce rapport varie dans des ordres de grandeur allant de -20 dB à +20 dB. Pour des seuils de détection nuls à chaque itération, on présente les pourcentages de fréquences correctement estimées. Ces derniers sont comparés aux pourcentages obtenus par la méthode de séparation témoin.
La comparaison des deux méthodes montre que quelque soit le niveau de puissance entre les deux sources, la méthode de séparation par le critère d'ordre, est plus performante et permet l'estimation d'un plus grand nombre de fréquences fondamentales à la seconde itération. On notera que les résultats pour un $RSS = 0$ dB, sont équivalent à ceux présentés sur la figure (Fig. 3.17), les deux sources ayant à l'origine des niveaux quasi égaux.

• **Mélange des trois sources**

L'étude du mélange des trois sources se fait à l'instar du mélange de 2 sources "soprano/ténor", à partir des pitch de référence. On détermine les pourcentages de fréquences correctement estimées par la méthode du critère d'ordre et on les compare à la séparation témoin. Les seuils de détection Δ_0 sont choisis nuls pour chaque itération. Les résultats sont présentés sur la figure (Fig. 3.18).

Comme précédemment la méthode basée sur le critère d'ordre s'avère la plus

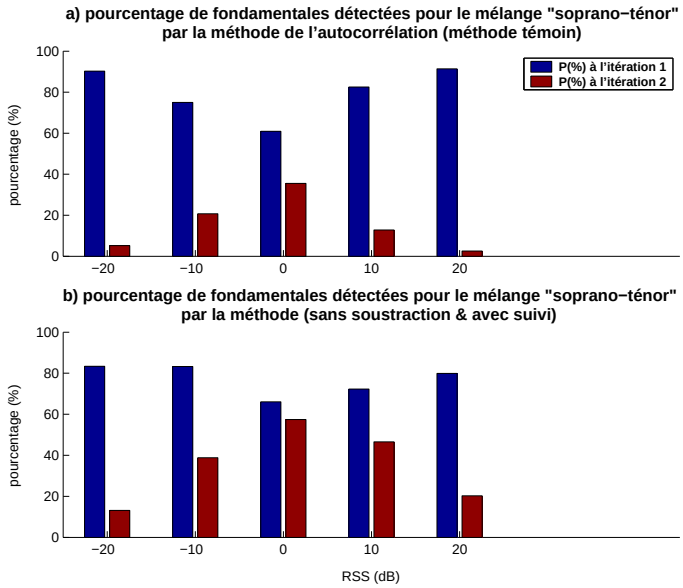


Fig. 3.18 – *Comparaison avec la méthode témoin en fonction du RSS entre les 2 sources du mélange. Pourcentage de fréquences estimées aux itérations successives : a) par la méthode de l'autocorrélation - b) par la méthode du critère d'ordre.*

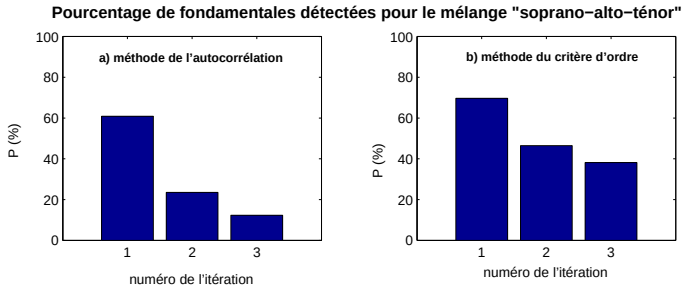


Fig. 3.19 – *Comparaison avec la méthode témoin. Cas du mélange des voix "soprano-alto-ténor" Pourcentage de fréquences fondamentales estimées aux itérations successives : a) par la méthode de l'autocorrélation - b) par la méthode du critère d'ordre.*

performante. Elle permet d'augmenter le pourcentage de fréquences correctes pour la troisième itération d'environ 25%. De plus, l'estimateur de cette méthode est plus robuste au "bruit" que l'estimateur basé sur l'autocorrélation, puisque pour un mélange de trois sources, son pourcentage de fréquences correctes à la première itération est supérieur de presque 10%.

3.4.4 Application à des mélanges "voisé/non-voisé"

On s'intéresse maintenant au mélange de 2 signaux contenant des parties voisées et des parties non-voisées. La forme d'onde de chaque signal source est présentée sur la figure (Fig. 3.20). Tout comme pour le cas des mélanges entièrement voisés, la méthode de séparation par le principe du critère d'ordre est comparée à la méthode témoin présentée au paragraphe (3.4.2). La comparaison des deux méthodes est tout d'abord effectuée pour le cas des sources isolées, puis le mélange des 2 signaux est traité. On cherchera alors à quantifier les performances de la méthode du critère d'ordre, en déterminant le pourcentage $P(\%)$ de fréquences fondamentales correctement estimées. La nature mixte "voisée/non-voisée" des signaux, nous permettra également de déterminer les pourcentages de fausses alarmes et de non-détection dont les définitions sont données ci-après.

• Estimation de la fréquence fondamentale des sources isolées

La première étape de l'étude pour l'application à la séparation de mélanges "voisé/non-voisé" consiste à déterminer les contours de pitch de chacune des sources du mélange étudié. Ces contours serviront par la suite de pitch de référence. En appliquant l'estimateur basé sur la méthode d'autocorrélation, on obtient les pitch de référence présentés sur la figure (Fig. 3.21).
On détermine le pourcentage de fréquences détectées correctes défini par l'équation (eq. 3.9). Ce pourcentage $P(\%)$ prend en compte aussi bien les fréquences estimées des parties voisées, que celles des parties non-voisées. Pour illustrer les performances obtenues sur chaque partie, on définit également les pourcentages $P_v(\%)$ et $P_{nv}(\%)$. Ces derniers sont représentatifs du nombre de fréquences estimées qui ont été décidées voisées ou décidées non-voisées par l'intermédiaire de leur écart de vraisemblance Δ :

$$P_v(\%) = \frac{\textit{(nombre de fondamentales décidées voisées)}}{\textit{nombre de fondamentales voisées}} \times 100\% \qquad (3.10)$$

$$P_{nv}(\%) = \frac{(\textit{nombre de fondamentales décidées non-voisées})}{\textit{nombre de fondamentales non-voisées}} \times 100\% \qquad (3.11)$$

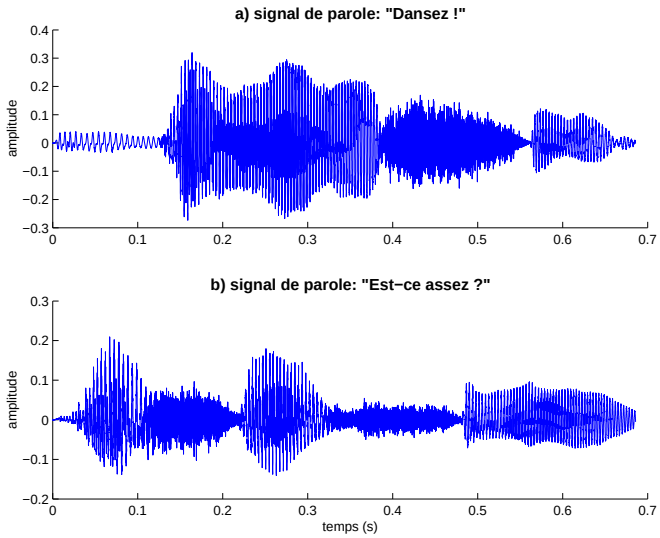


Fig. 3.20 – *Signaux de parole de nature mixte "voisé/non-voisé" : a) le signal de parole " Dansez ! " - b) le signal de parole " Est-ce assez ?".*

La figure (Fig. 3.22) présente l'ensemble des résultats obtenus pour les deux signaux pris séparément : quelque soit le signal, le pourcentage global $P(\%)$ de fréquences estimées correctement en comparaison aux pitch de référence, sont plus faibles que ceux obtenus pour les mélanges de sources voisées présentés dans le paragraphe précédent. Toutefois, les pourcentages des parties voisées et non-voisées $P_v(\%)$ et $P_{nv}(\%)$ obtenus, illustrent bien les capacités de l'écart de vraisemblance en tant que détecteur de voisement : lorsque la valeur seuil Δ_0 augmente la détection des fréquences nulles sur la parties non-voisées augmente également, sans pour autant diminuer celles des parties voisées.

Dans la méthode de séparation proposée, l'écart de vraisemblance Δ associé à une fréquence fondamentale estimée est utilisé sous forme de détecteur de voisement. Les fréquences pour lesquelles Δ reste inférieur à une certaine valeur seuil Δ_0 sont mises à zéro. Afin de rendre compte des performances de Δ en tant

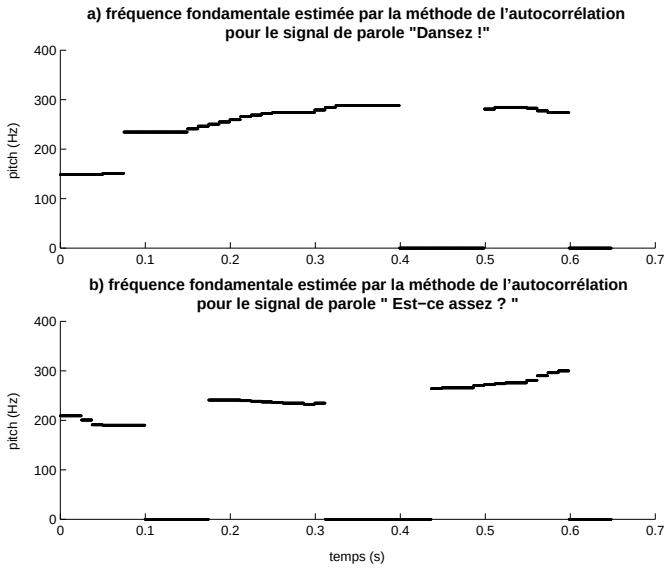


Fig. 3.21 – *Pitch de référence : fréquences fondamentales estimées par la méthode de l'autocorrélation : a) le signal de parole " Dansez ! " - b) le signal de parole " Est-ce assez ?".*

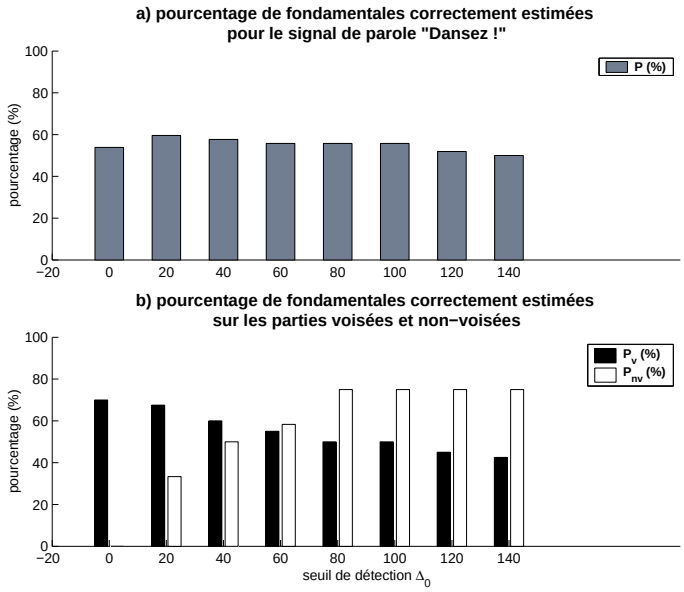


Fig. 3.22 – *a) pourcentage global $P(\%)$ de fondamentales estimées par la méthode du critère d'ordre. b) pourcentage $P_v(\%)$ des parties voisées et pourcentage $P_{nv}(\%)$ des parties non voisées : cas du signal de parole " Dansez ! ".*

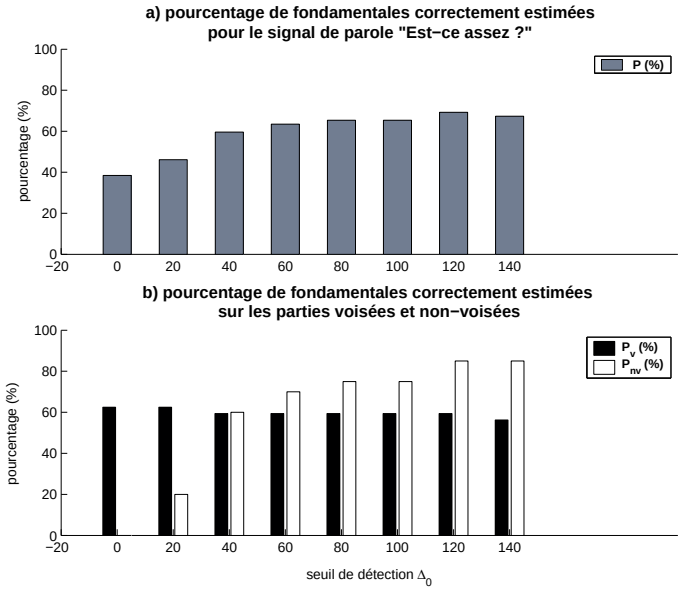


FIG. 3.23 – *a) pourcentage global $P(\%)$ de fondamentales estimées par la méthode du critère d'ordre. b) pourcentage $P_v(\%)$ des parties voisées et pourcentage $P_{nv}(\%)$ des parties non voisées : cas du signal de parole " Est-ce assez ?".*

que détecteur de voisement, nous déterminerons également les pourcentages de fausses alarmes et de non-détection, définis tels que :

$$FA(\%) = \frac{\textit{(nombre de fondamentales décidées voisées)}}{\textit{nombre de fondamentales non-voisées}} \times 100\% \qquad (3.12)$$

$$ND(\%) = \frac{\textit{(nombre de fondamentales décidées non-voisées)}}{\textit{nombre de fondamentales voisées}} \times 100\% \qquad (3.13)$$

Lorsque $\Delta_0 = 0$, toutes les fréquences estimées par la méthode du critère d'ordre sont non nulles. Il en résulte que le pourcentage de fausses alarmes est maximal et celui de non-détection est nul : $FA(\%) = 100$ et $ND(\%) = 0$. Quand Δ_0 augmente, $ND(\%)$ augmente également, tandis que $FA(\%)$ diminue. Le choix d'un taux de fausses alarmes équivalent au taux de non-détection nécessite alors de prendre un seuil de détection autour de la valeur 60.

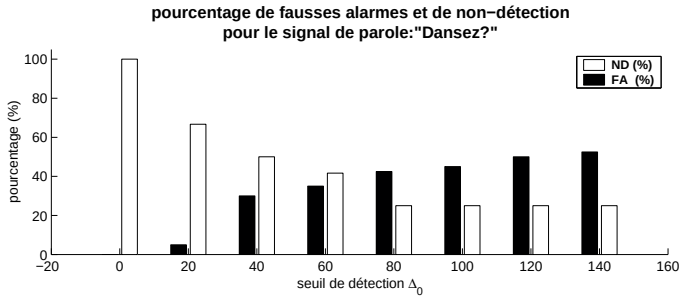


Fig. 3.24 – *Pourcentage de fausses alarmes $FA(\%)$ et de non-détection $ND(\%)$ en fonction du seuil de détection choisi Δ_0 : cas du signal de parole " Dansez ! ".*

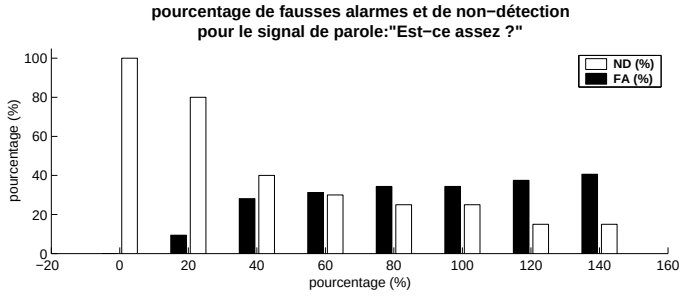


Fig. 3.25 – *Pourcentage de fausses alarmes $FA(\%)$ et de non-détection $ND(\%)$ en fonction du seuil de détection choisi Δ_0 : cas du signal de parole " Est-ce assez ? ".*

• **Mélange des deux sources**

Lorsque les deux signaux de parole "Dansez !" et "Est-ce-assez ?" sont mélangés avec un rapport de puissance égal à 0 dB, on applique la méthode du critère d'ordre pour déterminer itérativement les fréquences fondamentales des deux sources. La figure (Fig. 3.26) présente les pourcentages de fréquences fondamentales $P(\%)$ obtenus pour chaque itération. La figure (Fig. 3.26) illustre le cas où $\Delta_{01} = 0$ et $\Delta_{02} \in [0, 120]$, tandis que pour la figure (Fig. 3.27) $\Delta_{01} = 60$ et $\Delta_{02} \in [0, 120]$.

Le pourcentage P(%) obtenu à la première itération est beaucoup plus faible

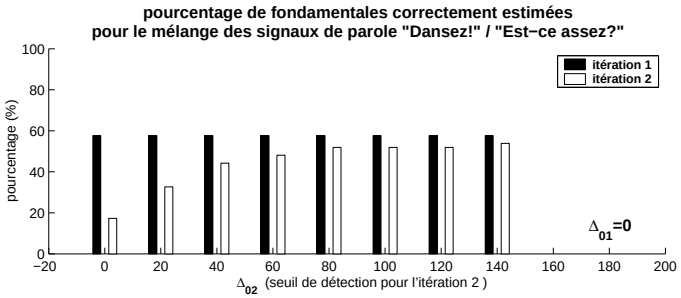


Fig. 3.26 – *Pourcentage de fondamentales correctement estimées à chaque itération par la méthode du critère d'ordre en fonction du seuil de détection choisi* Δ_{02}*, pour le mélange "voisé/non-voisé" : pour le choix d'un seuil à l'itération 1 tel que* $\Delta_{01} = 0$*.*

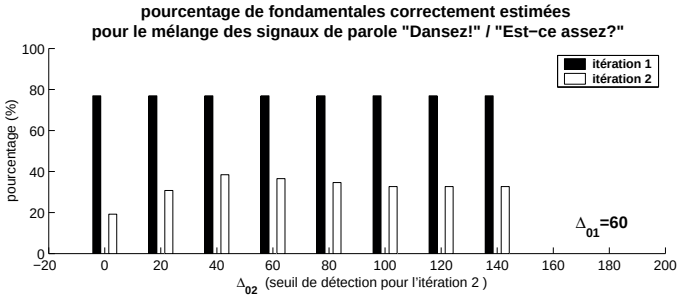


Fig. 3.27 – *Pourcentage de fondamentales correctement estimées à chaque itération par la méthode du critère d'ordre en fonction du seuil de détection choisi* Δ_{02}*, pour le mélange "voisé/non-voisé" : pour le choix d'un seuil à l'itération 1 tel que* $\Delta_{01} = 60$*.*

pour $\Delta_{01} = 0$ que pour $\Delta_{01} = 60$, car le premier cas ne permet pas la détection des parties non-voisées du signal. A la seconde itération, le pourcentage correspondant est quant à lui beaucoup plus fort que celui obtenu pour $\Delta_{01} = 60$,

lorsque la valeur seuil Δ_{02} devient supérieure à 20. On atteint alors globalement les mêmes performances dans les deux cas considérés. Ce qui diffère entre ces deux cas est finalement l'ordre des fondamentales estimées : dans le premier cas, comme $\Delta_{01} = 0$ les fréquences fondamentales des parties non-voisées sont détectées à la seconde itération, alors que pour le second cas une partie de ces fréquences est obtenue dès la première itération.

Comparaison avec la méthode de séparation témoin : la comparaison avec la méthode de l'autocorrélation montre à nouveau les performances de la méthode basée sur le critère d'ordre. Comme la méthode témoin ne permet pas de prendre en compte les parties non-voisées du mélange, nous comparons les deux méthodes pour des seuils de détection $[\Delta_{01}, \Delta_{02}] = [0, 0]$. Les résultats sont illustrés sur la figure (Fig. 3.28) : à l'itération 1, le meilleur pourcentage $P(\%)$ est obtenu par la méthode témoin, tandis qu'à l'itération 2 la méthode du critère d'ordre fournit de meilleurs résultats. Globalement, les performances des deux méthodes sont alors équivalentes.
On notera tout de même qu'une augmentation, même faible, des seuils Δ_{01} et Δ_{02} permet à la méthode du critère d'ordre de détecter les fréquences des parties non-voisées et dès lors d'obtenir de meilleures performances.

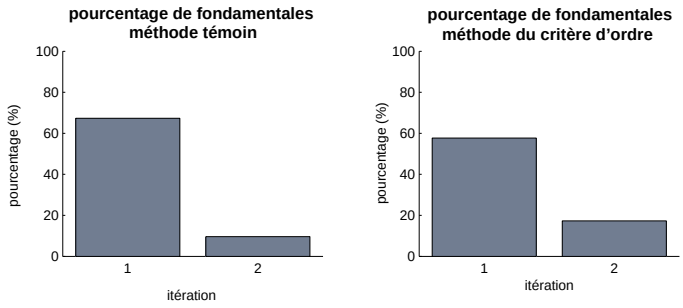


Fig. 3.28 – *Comparaison avec la méthode de séparation témoin. Pourcentage de fréquences fondamentales estimées : a) par la méthode de l'autocorrélation - b) par la méthode du critère d'ordre.*

3.4.5 Application à la séparation de deux locuteurs

Une dernière étude de la méthode de séparation par le critère d'ordre est présentée ci-dessous. Le mélange utilisé est cette fois-ci un mélange réel, qui consiste

en l'enregistrement monophonique de deux locuteurs comptant de un jusqu'à dix simultanément. La première voix est anglaise et la seconde espagnole. Ce signal est disponible à l'adresse "*www.cnl.salk.edu/ tewon/Blind/blind_audio.html*".

Comme les signaux source ne sont pas à notre disposition, aucun pitch de référence ne peut être utilisé dans cette partie. Deux types de comparaison sont alors présentés : une première comparaison est effectuée entre la méthode du critère d'ordre "avec soustraction" et la méthode témoin. Par la suite, la méthode du critère d'ordre "avec soustraction" est comparée à la méthode du critère d'ordre "sans soustraction". Les résultats sont illustrés par les figures (Fig. 3.29) et (Fig. 3.30) qui présentent les contours de pitch obtenus pour les deux locuteurs. Les seuils utilisés pour la méthode basée sur le critère d'ordre sont $\Delta_{0_1} = 90$ et $\Delta_{0_2} = 65$.

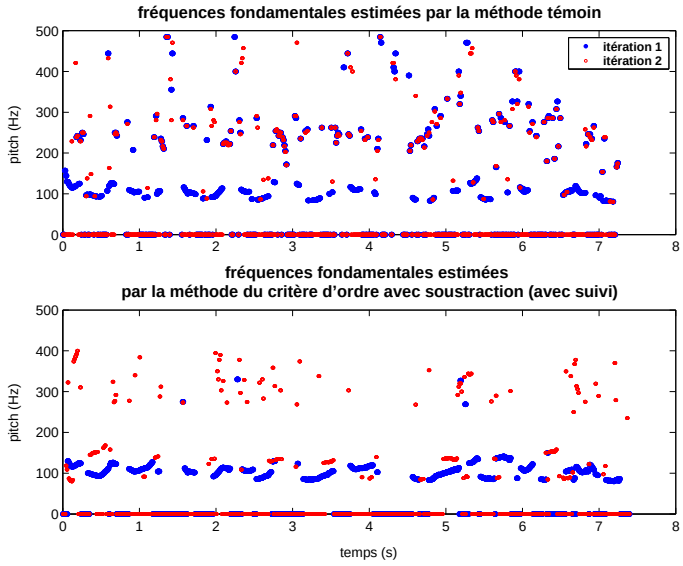


Fig. 3.29 – *Comparaison des méthodes de séparation appliquées à un mélange réel contenant deux locuteurs : a) méthode de l'autocorrélation - b) méthode du critère d'ordre avec soustraction.*

Les résultats obtenus par la méthode de séparation témoin ne pas satisfaisant : bien qu'à la première itération la fréquence fondamentale du locuteur le plus fort reste correctement estimé, la seconde obtenue après soustraction n'est

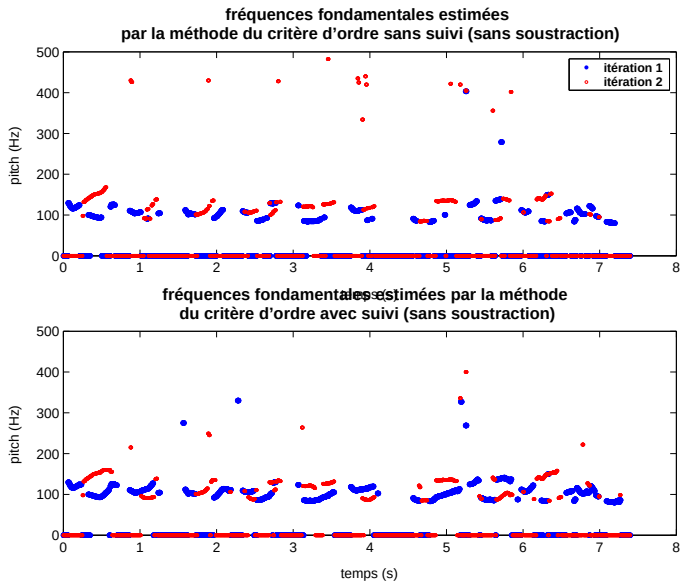


Fig. 3.30 – *étude de la poursuite des fréquences fondamentales pour un mélange réel contenant deux locuteurs. Comparaison de la méthode de séparation basée sur le critère d'ordre : a) sans suivi fréquentiel - b) avec suivi fréquentiel.*

quasiment jamais retrouvé.
La comparaison avec la méthode basée sur le critère d'ordre utilisant également une étape de soustraction, permet d'obtenir de meilleurs résultats, aussi bien lors de la première estimation que lors de la seconde. Ainsi bien que l'étape de soustraction implique très certainement une dégradation des estimés de la seconde itération, l'estimateur utilisant le principe du critère d'ordre apparaît beaucoup plus performant que l'estimateur basé sur l'autocorrélation.
Afin de pallier les problèmes liés à l'étape de soustraction, nous avons développée une méthode de séparation dite "sans soustraction". A cette dernière, nous avons montré qu'il est possible d'associer une poursuite des fréquences fondamentales tout au long des trames analysées. La figure (Fig. 3.30) présente alors les estimations obtenues par la méthode du critère d'ordre "sans suivi fréquentiel" et celles obtenues "avec suivi fréquentiel" le long des fenêtres d'analyse. La comparaison des deux cas montre des résultats bien meilleurs pour la méthode utilisant la poursuite : on voit apparaître pour la seconde itération, un grand nombre de fréquences correctement estimées.
On peut également comparer les résultats de la méthode du critère d'ordre "avec soustraction" présenté sur la figure (Fig. 3.29) et ceux de la méthode "sans soustraction" de la figure (Fig. 3.30). Ils confirment bien que l'étape de soustraction peut dégrader très fortement le mélange étudié, impliquant ainsi de faibles performances d'estimation pour chaque nouvelle itération.

• **Estimation du nombre de sources**

On s'intéresse ici pour finir, à l'estimation du nombre de sources contenues dans un mélange. La méthode de séparation du critère d'ordre estime ce dernier, au moyen de l'écart de vraisemblance, qui comme on l'a vu auparavant est équivalent à une mesure de voisement. Ainsi pour un écart de vraisemblance trop faible à l'itération (Ns+1) de la méthode, le mélange est supposé constitué de $\widehat{N}s$ sources. La figure (Fig. 3.31) présente les écarts de vraisemblances obtenus à chaque nouvelle itération de la méthode de séparation du critère d'ordre, lorsque cette dernière est appliquée au mélange de voix des deux locuteurs. Ainsi, au fur et à mesure des itérations la valeur des écarts de vraisemblance diminue. Cette diminution devient flagrante à la troisième itération.
Il est donc possible d'utiliser l'écart de vraisemblance pour déterminer le nombre de sources harmoniques extraites du mélange.
On notera que les fréquences relatives aux écarts de vraisemblance supérieurs à 200, pour les itérations 3 et 4, correspondent généralement à des estimés de valeurs fréquentielles qui restent très proches des fréquences estimées à l'itération précédente.

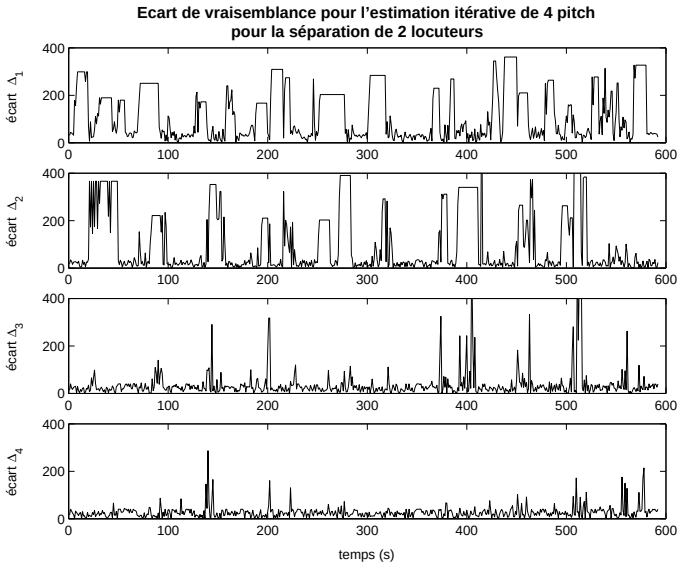


Fig. 3.31 – *écart de vraisemblance pour 4 itérations successives : application au mélange de 2 locuteurs.*

3.5 Conclusion

A partir de l'estimateur de fréquence fondamentale basé sur le principe du critère d'ordre, une nouvelle méthode pour la séparation de signaux de parole est proposée. Cette dernière est une méthode de séparation itérative, qui ne nécessite pas la connaissance du nombre de sources voisées du mélange et qui contrairement aux méthodes itératives existantes, n'utilise aucune étape de soustraction. La comparaison avec une méthode de séparation témoin, basée sur l'estimation itérative des fréquences fondamentales par la méthode de l'autocorrélation, nous permet de mettre en évidence les performances de notre méthode et cela pour différents types de mélanges. Ainsi, bien que son application dans le cas de sources isolées reste très proche de la méthode témoin, cette dernière à l'avantage dans le cas de mélange, d'être plus performante. Ces performances sont présentées en terme de pourcentage de fréquences fondamentales correctement estimées, et montrent une meilleure robustesse face au nombre de sources du mélange.
Un autre avantage de la méthode basée sur le critère d'ordre est qu'elle intègre directement un détecteur de voisement, ainsi qu'une méthode de poursuite des fréquences fondamentales le long des fenêtres d'analyse. Le premier est utilisé pour stopper les itérations et estimer ainsi le nombre de sources du mélange. La poursuite, quant à elle, permet de réduire les erreurs d'octaves éventuelles et d'augmenter le pouvoir de discrimination des parties voisées et non-voisées.

Troisième partie

Estimation de fréquences fondamentales pour la musique

Chapitre 4

Estimation par classification des pics spectraux

La méthode d'estimation de fréquences fondamentales multiples basée sur le principe du critère d'ordre (cf. Partie 1), possède de nombreux avantages pour le traitement de mélange de sources harmoniques. Cette méthode paramétrique utilise la modélisation CARD ("Canonical AutoRegressive Decomposition") qui permet de rendre compte de la nature "voisée/non-voisée" des mélanges et d'obtenir une résolution supérieure à la résolution de Fourier pour l'estimation des fréquences fondamentales des différentes sources. Développée précédemment dans le cadre de la séparation de signaux de parole, elle reste applicable à un grand nombre de signaux musicaux, ces derniers devant posséder le même caractère de "quasi-harmonicité" que l'on trouve dans la parole voisée (sons non percussifs, entretenus $\cdots$). Toutefois, comme la plupart des méthodes "haute résolution", la méthode d'estimation du critère d'ordre a l'inconvénient d'être relativement coûteuse en calcul. Bien que ce coût de calcul est considérablement réduit par la prise en compte des propriétés asymptotiques de l'estimateur du Maximum de Vraisemblance, ainsi que des ambiguïtés d'octave, il reste lié à l'intervalle dans lequel les fréquences fondamentales sont recherchées (cf. tableau(2.3)). Pour la parole, les fréquences fondamentales sont supposées appartenir à l'intervalle [80,400] Hz. Dans le cas des signaux musicaux, la diversité des instruments impose un intervalle de recherche beaucoup plus grand de l'ordre de [27,2500] Hz. Le traitement de mélanges musicaux par la méthode d'estimation du critère d'ordre, apparaît dès lors relativement plus coûteuse.
Outre les valeurs possibles de la fréquence fondamentale, les signaux musicaux et les signaux de parole diffèrent sur d'autres points : on peut généralement supposer que les signaux musicaux sont plus stationnaires et qu'ils peuvent être analysés sur des fenêtres d'une durée de 100 ms. De plus, bien que la plupart de ces si-

gnaux ont à l'instar des parties non-voisées, des parties non-harmoniques caractéristiques des bruits musicaux d'instruments (frottements de l'archet d'un violon, bruits de souffle pour les instruments à vent $\cdots$), nous pouvons supposer que ces parties sont moins importantes, en terme de puissance, que les parties non-voisées dans la parole. Cette remarque nous incite à penser que le modèle CARD n'est pas le plus approprié pour l'analyse de mélanges musicaux. Un dernier point important qui différencie ces deux types de mélange se rapporte à l'harmonie. Les mélanges musicaux sont généralement des suites d'accords ou d'intervalles de notes consonants et dissonants. Comme la consonance de deux sons entraîne un recouvrement de leur spectre, la séparation de ces derniers s'avère généralement plus complexe du fait de la présence d'harmoniques communes. Les méthodes d'estimation de fréquences fondamentales multiples pour les signaux musicaux se doivent d'être robustes face au problème de recouvrement spectral. Ce problème intervient peu dans le traitement des mélanges de parole et n'est pas explicitement pris en compte par la méthode d'estimation basée sur le critère d'ordre.

Pour toutes ces raisons, nous souhaitons développer de nouvelles méthodes d'estimation de fréquences fondamentales multiples pour la séparation des mélanges musicaux. L'approche que nous choisissons consiste à estimer les différentes fréquences fondamentales du mélange, à partir de "spectres séparés" obtenus par classification des pics spectraux du mélange. Nous proposons alors trois méthodes d'estimation de fréquences fondamentales multiples qui diffèrent par la méthode de classification utilisée pour obtenir les spectres séparés.
Comme la méthode basée sur le critère d'ordre, ces 3 méthodes permettent d'estimer le nombre de sources du mélange. Contrairement aux méthodes itératives qui sont actuellement les plus performantes, l'estimation des fréquences fondamentales se fait simultanément, évitant ainsi le traitement des nombreux problèmes propres à ces méthodes (reconstruction, soustraction des paramètres de source $\cdots$). De plus, les 3 méthodes proposées tiennent compte du problème de recouvrement spectral. Elles sont alors capables de séparer des notes jouées à la quinte, à la tierce et même d'identifier les notes d'un accord.

Ce dernier chapitre est consacré à ces 3 méthodes d'estimation de fréquences fondamentales multiples dites "par classification des pics spectraux du mélange". Nous donnerons tout d'abord une présentation de ce type de méthode, puis nous développerons trois méthodes de classification des pics spectraux : la première classification utilise l'algorithme des Nuées Dynamiques, la seconde est basée sur le calcul du rapport harmonique entre les fréquences des pics à classer, tandis que la troisième est une version hybride des deux autres, utilisant à la fois les Nuées Dynamiques et le calcul de l'harmonicité.
L'étude détaillée des performances de ces méthodes d'estimation sera par la suite

présentée, aussi bien pour des mélanges synthétiques que des mélanges réels.

4.1 Présentation générale

Les méthodes d'estimation de fréquences fondamentales multiples proposées dans ce chapitre, sont basées sur une classification des pics spectraux du mélange analysé. Les fréquences fondamentales de chaque source sont estimées à partir d'un ensemble restreint de fréquences, qui correspond aux pics d'une classe donnée. Le schéma général de ces méthodes est présenté ci-dessous :

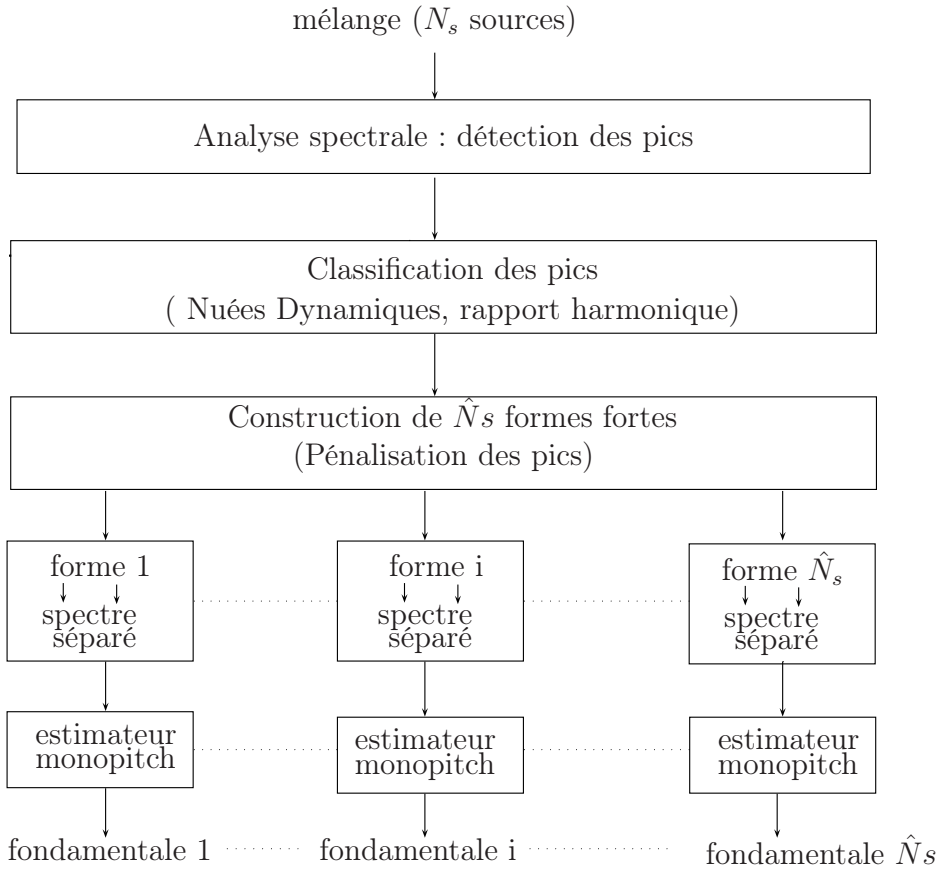


Fig. 4.1 – *Principe de l'estimation de fréquences fondamentales multiples basé sur la classification des pics spectraux du mélange*

• **Analyse spectrale**

Le mélange, dont le nombre de sources N_s est inconnu, est tout d'abord analysé à l'aide de fenêtres se recouvrant à 50% et dont la durée est choisie égale à 100 ms. Pour chaque fenêtre de signal, les pics spectraux du mélange sont détectés à partir des maxima du périodogramme. L'ensemble des fréquences des pics détectés pour la p^{ieme} fenêtre est noté E_p.

• **Classification des pics par le principe des Formes Fortes**

La classification des pics spectraux est obtenue par *le principe des formes fortes*. Ce principe est à l'origine une pratique développée pour limiter l'arbitraire des méthodes de classification non-hiérarchiques, telles que les Nuées Dynamiques. L'algorithme des Nuées Dynamiques nécessite en effet une initialisation aléatoire, dont va dépendre la partition de l'ensemble E_p. Le principe des formes fortes consiste à appliquer plusieurs fois l'algorithme des Nuées Dynamiques, puis à identifier les éléments les plus souvent regroupés ensembles dans les différentes partitions. La partition finale est constituée de $\hat{N}_s$ classes appelées *formes fortes*. En pratique, les formes fortes sont construites en associant à chaque couple d'éléments $(\nu_i, \nu_{j\neq i}) \in E_p^2$, un indice de pénalité $\delta(\nu_i, \nu_j)$. La valeur de $\delta(\nu_i, \nu_j)$ est égale au nombre de partitions obtenues par les Nuées Dynamiques, pour lesquelles ν_i et ν_j sont dans des classes différentes. Les fréquences telles que $\delta < \Delta$, où Δ est un seuil arbitraire à définir, sont ensuite regroupées dans une même forme forte.

Le principe des formes fortes reste applicable à des méthodes autres que les Nuées Dynamiques. Nous proposons alors d'utiliser ce dernier pour deux autres classifications : la première est une classification basée sur le calcul du rapport harmonique entre deux fréquences à classer. Le principe des formes fortes est alors utilisé pour pénaliser deux fréquences qui ne sont pas en relation harmonique, plutôt que deux fréquences qui n'appartiennent pas à une même classe. La seconde classification est une fusion des deux précédentes, pour laquelle le principe des formes fortes pénalise les fréquences en tenant compte à la fois des partitions obtenues par l'algorithme des Nuées Dynamiques et du rapport harmonique des fréquences.
Ce type de classification met finalement à disposition $\hat{N}_s$ formes fortes. Elles sont utilisées pour définir des "spectres séparés" sur desquels un estimateur "monopitch" est appliqué. Les $\hat{N}_s$ fréquences fondamentales estimées sont finalement associées aux $\hat{N}_s$ sources du mélange.

4.2 Analyse spectrale pour la séparation

4.2.1 Représentation spectrale

• Périodogramme

La représentation spectrale du mélange utilisée est celle du périodogramme. Elle est basée sur le calcul de la Transformée de Fourier à court terme, dont l'expression pour un signal x à analyser est donnée par :

$$X[k] = \sum_{n=0}^{N-1} W[n]\, x[n]\, exp\{\frac{-j.2\pi kn}{N}\} \tag{4.1}$$

$x[n]$ est l'échantillon à l'instant n, $2\pi k/N$ la fréquence et N/k la période en échantillons. W est la fenêtre d'analyse.
$W[n]$ choisie égale à 1 sur l'intervalle $[0, (N-1)]$, correspond à une fenêtre d'analyse rectangulaire.
L'analyse spectrale consiste à détecter les maxima du spectre en amplitude $|X[k]|$. Les fréquences associées à ces maxima vont constituer l'ensemble des pics spectraux à classifier.

• Longueur de la fenêtre d'analyse

Nous avons vu dans le cadre de la parole, que l'hypothèse de quasi-stationnarité des signaux implique l'utilisation de fenêtres d'analyse dont la longueur maximale est d'environ 50 ms. Dans le cas des signaux musicaux, on peut supposer que cette hypothèse reste valable pour des durées inférieures ou égales à 100 ms. Des durées deux fois plus longues sont fréquemment utilisées, comme par exemple dans le cas de l'estimateur de fréquences fondamentales multiples proposé par Klapuri [53] [28]. Toutefois, elles sont à l'origine d'estimations fréquentielles "grossières" ou exploitables dans le cas de sons très longuement entretenus.

L'analyse spectrale proposée dans ce travail, utilise des fenêtres d'une durée de 100 ms. Elle permet d'obtenir une résolution spectrale de 10 Hz, impliquant

notamment que deux fréquences espacées d'une quantité inférieure à cette résolution, ne pourront être séparées par des méthodes basées sur la FFT.

Il serait envisageable d'utiliser des représentations spectrales "haute résolution", dans le but de réduire la taille des fenêtres, jugées encore trop importantes pour de nombreuses applications musicales. Toutefois, nous souhaitons avant tout mettre en avant dans ce travail, le principe de classification des pics spectraux qui est indépendant de la représentation spectrale. Les conséquences liées au choix d'une représentation spectrale basée sur la FFT, pour des durées d'analyse de 100 ms, sont discutées par la suite.

4.2.2 Détection des pics spectraux

Avant d'effectuer la classification des pics spectraux, les différents paramètres fréquentiels sont estimés à partir de la représentation spectrale du mélange. Cette estimation est alors réalisée en détectant les fréquences correspondant aux maxima du périodogramme. La détection se fait à partir d'une valeur seuil σ_0 choisie de façon arbitraire telle que :

$$\sigma_0 = p \times |X_{max}[k]| \quad avec \quad 0 < p < 1 \tag{4.2}$$

Une fréquence f_k est détectée si son amplitude $|X[k]|$ est supérieure à $(p \times 100)\%$ de la valeur maximale $|X_{max}[k]|$.
Les fréquences $\{f_k\}_{k=1,2,\cdots}$ ainsi détectées vont être par la suite classifiées, puis séparées pour estimer les fréquences fondamentales du mélange.

4.2.3 Estimation des fréquences fondamentales

L'étape de classification des pics spectraux a pour but de mettre à disposition N_s classes (ou formes fortes) qui contiennent les fréquences correspondant aux pics d'une même note. On parle alors de "spectres séparés" à partir desquels une méthode d'estimation est appliquée pour déterminer la fréquence fondamentale de la note.
La méthode d'estimation utilisée par la suite est une méthode de compression

spectrale, similaire à celle de l'histogramme harmonique déjà présentée dans l'état de l'art. La différence avec l'histogramme harmonique est qu'elle ne prend pas en compte la contribution des sous-harmoniques (cf. paragraphe (1.3.1)). Le choix de cet estimateur fréquentiel est justifié par sa simplicité de mise en oeuvre dans cette application. Toutefois, il est à noter que l'efficacité de cette méthode d'estimation spectrale dépend de la représentation fréquentielle du signal analysé. Ainsi, si aucun pic n'est détecté à la fréquence fondamentale de la note analysée, la compression spectrale est incapable de déterminer cette dernière et estime finalement une fréquence fondamentale erronée correspondant généralement à des octaves supérieures.

4.2.4 Résolution spectrale - conséquences

Comme nous l'avons évoqué précédemment, l'utilisation de fenêtres d'analyse de 100 ms, implique que deux fréquences séparées de moins de 10 Hz, ne peuvent être théoriquement résolues. Dans le cas idéal d'un mélange de notes, où l'on considère qu'il n'existe aucun recouvrement entre les harmoniques du spectre, on peut alors remarquer qu'un tel écart intervient pour des notes situées en dessous de la note **G3** dans la gamme musicale occidentale, c'est-à-dire pour des fréquences inférieures à 191 Hz.
Cette limitation de la résolution se traduit dans le cadre musical par deux conséquences directes :

- parce que l'écart fréquentiel relatif au demi-ton pour des fréquences au dessous de 191 Hz reste inférieur à la résolution de Fourier, une note musicale dans ce domaine fréquentiel peut être confondue après estimation spectrale, avec sa note voisine. En prenant l'exemple de la note **G2** de fréquence fondamentale 97.99 Hz, on remarque que si l'estimation de cette dernière est égale à 102 Hz, valeur inclus dans l'intervalle de résolution, alors la note finalement identifiée sera **G#2**.

- dans le cas d'un mélange contenant deux notes dont les fréquences fondamentales sont écartées d'une valeur inférieure à 10 Hz, les pics correspondant ne peuvent être résolus et un unique pic de fréquence proche des deux fréquences fondamentales est généralement détecté. Deux possibilités sont à prendre en compte : Soit le pic détecté est associé à la note de fréquence la plus proche et dans ce cas la seconde note ne possède plus de fréquence fondamentale. Soit une note intermédiaire ou "note fantôme", qui à l'origine n'appartient pas au mélange, est détectée. Dans ce cas l'une des notes, voire les deux, ne possède plus de fréquence fondamentale. On parlera par

la suite du phénomène de fondamental absent pour ce type de notes.

La figure (Fig. 4.2) illustre le cas d'un mélange de deux notes dans le domaine des basses fréquences, séparées d'un demi-ton. La première note jouée par un tuba est un **F#3** (185 Hz), tandis que la seconde est un **F3** (174 Hz) d'un basson français. La représentation spectrale correspondante sur une fenêtre d'analyse de 100 ms, met en évidence deux pics aux fréquences 166 Hz et 181 Hz. Ces dernières correspondent respectivement aux notes **E3** et **F#3**. Ainsi, seule la première note du mélange est détectée. Le pic détecté à 166 Hz peut quant à lui s'apparenter à une note "fantôme".

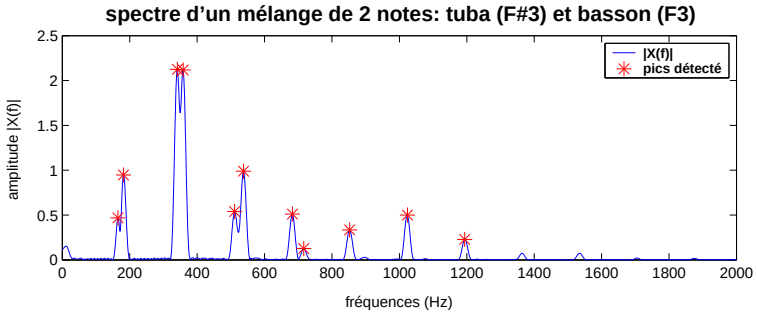


Fig. 4.2 – *Représentation spectrale d'un mélange contenant 2 notes proches jouées par un tuba (***F#3***) et un basson (***F3***) : La faible résolution entraîne la détection d'une note fantôme* **E3** *à la fréquence 166 Hz. (⋆) fréquences détectées pour un seuil de* 10%.

• Phénomène du fondamental absent

Si le phénomène du fondamental absent est lié à la faible résolution de la FFT dans le domaine des basses fréquences, il est également propre à certains instruments tels que le basson ou le hautbois. La figure (Fig. 4.3) qui présente le spectre d'une note de hautbois, met en évidence un pic de très faible amplitude à la fréquence fondamentale. Dans un mélange quelconque, même relativement simple, ce pic reste indétectable la plupart du temps.

Quand l'absence de la fréquence fondamentale est causée par le manque de résolution, on peut supposer qu'une méthode d'estimation de type haute résolution va permettre de pallier ce problème. Malheureusement, cette solution ne permettra pas de traiter le cas de signaux tels que celui de la figure (Fig. 4.3).

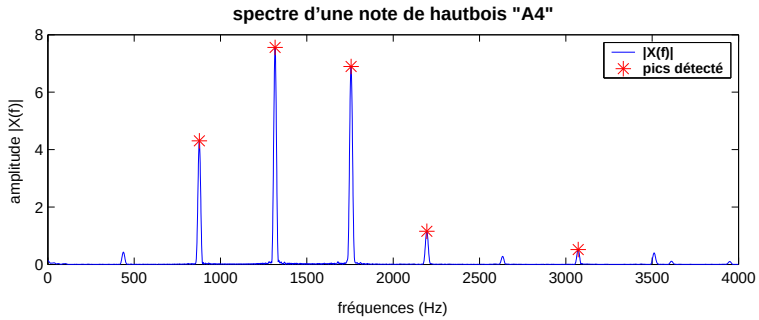


Fig. 4.3 – *Représentation spectrale d'une note de hautbois* **A4** *: le pic à la fréquence fondamentale 440 Hz n'est pas détecté. La note est caractéristique du phénomène de fondamental absent. (⋆) fréquences détectées pour un seuil de* 10%.

4.3 Classification des pics spectraux

Les méthodes d'estimation de fréquences fondamentales multiples proposées dans ce chapitre font appel à une classification des pics spectraux du mélange. Dans cette approche, la classification idéale correspond à une partition de l'ensemble des pics dans laquelle les fréquences de chaque note sont regroupées dans une même classe.
Il existe de nombreuses méthodes de classification. Elles se distinguent avant tout par leur procédure : les méthodes hiérarchiques regroupent les éléments les plus semblables en premier lieu, tandis que les méthodes non-hiérarchiques les groupent simultanément. Toutes ces méthodes dépendent également du choix des critères de "ressemblance" entre deux éléments, entre deux groupes ou encore entre un élément et un groupe.
Historiquement, les méthodes hiérarchiques furent les premières développées, principalement en raison de leur simplicité de calcul. De nos jours, l'avènement des puissants ordinateurs a permis l'utilisation courante des méthodes non-hiérarchiques. Ces dernières permettent généralement une meilleure description des zones denses du nuage des observations.

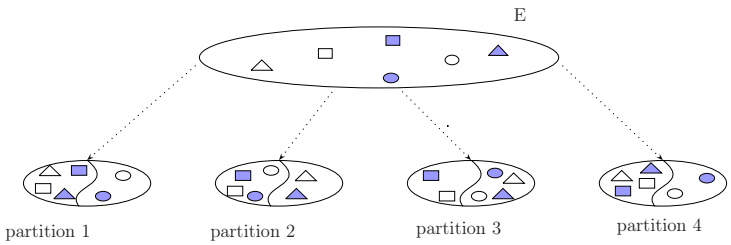
Pour la classification de pics spectraux, nous proposons d'utiliser une méthode basée sur le principe des formes fortes [54]. Elle consiste à calculer T partitions de l'ensemble des fréquences détectées à l'aide de la méthode de classification non-hiérarchique des Nuées Dynamiques, puis à déterminer les classes stables, dites *formes fortes*, en identifiant les éléments qui se regroupent toujours ensemble.
Dans cette partie, nous introduisons de manière très générale le principe des formes fortes. Les éléments constitutifs de l'algorithme des Nuées Dynamiques

seront ensuite présentés. Par la suite, nous proposerons deux distances définissant les critères de "ressemblance" de la méthode des Nuées Dynamiques. Les classifications ainsi obtenues seront évaluées sur des mélanges de notes synthétiques. L'évaluation des deux algorithmes de classification par les Nuées Dynamiques montre que les distances *moindres carrés* et *sinusoïdale* permettent de grouper les fréquences en tenant compte du caractère harmonique des signaux musicaux. Les performances sont présentées en termes de pourcentages de classifications correctes. Elles dépendent du nombre de sources dans le mélange, mais également du nombre de classes choisi pour la classification par les Nuées Dynamiques. Nous verrons finalement que la distance *sinusoïdale* donne généralement de meilleures performances.

4.3.1 Le principe des formes fortes

Le principe des formes fortes, appliqué à un ensemble E d'éléments à classer, consiste à déterminer les éléments qui se regroupent toujours ensemble, lorsqu'on a à disposition T partitions différentes de E. Le jeu de partitions est généralement obtenu en appliquant plusieurs fois une méthode de classification non-hiérarchique, telle que la méthode des Nuées Dynamiques, pour laquelle chaque partition dépend de son initialisation. Le principe donne une partition finale de E dont les classes sont appelées *formes fortes*. Il est illustré par un exemple simple sur la figure (Fig. 4.4) : on considère un ensemble E de 6 éléments que l'on souhaite regrouper en fonction de leur géométrie. Le nombre de géométries possibles est supposé inconnu. Une méthode de classification met à disposition 4 partitions de E en 2 classes. Le tableau de la figure (Fig. 4.4) résume l'ensemble de ces classifications en présentant le numéro de classe de chaque élément pour chaque partition. Il permet de distinguer 3 formes fortes qui contiennent les éléments toujours groupés ensemble : une première forme forte pour les triangles, une seconde pour les carrés et une troisième pour les cercles. La partition finale de l'ensemble E est ainsi constituée de trois classes (ou formes fortes) à deux éléments.

Le principe des formes fortes cherche à limiter l'arbitraire de la méthode de classification utilisée pour obtenir les T partitions de l'ensemble E. En pratique, les formes fortes sont construites en associant un indice de pénalité δ pour chaque paire d'éléments. δ est égal au nombre de partitions dans lesquelles ces éléments appartiennent à des classes différentes. Ainsi pour deux éléments $(i, j) \in E^2$, $\delta(i, j) = 0$ signifie que pour les T partitions, i et j sont toujours classés ensemble. La condition $\delta = 0$ pour la construction des formes fortes est très restrictive et



	partition 1	partition 2	partition 3	partition 4
△	1	2	2	1
▲	1	2	2	1
■	1	1	1	1
□	1	1	1	1
●	2	1	2	2
○	2	1	2	2

Fig. 4.4 – *Construction de formes fortes à partir de* $T = 4$ *partitions différentes d'un ensemble* E *d'éléments à classer.*

nécessite l'utilisation d'une méthode de classification très performante. Ces performances dépendent du nombre de classes choisi pour les T partitions de E et des critères de ressemblances utilisés. On peut toutefois supposer qu'une méthode de classification est performante lorsqu'elle fournit T partitions relativement similaires. Cette stabilité se traduit alors par des indices de pénalité faibles pour les éléments les plus souvent classés ensemble. Les formes fortes se construisent finalement en définissant un seuil de pénalité Δ, en dessous duquel les éléments sont supposés être dans la même classe.

4.3.2 Algorithme des Nuées Dynamiques

La méthode des Nuées Dynamiques est un procédé de classification basé sur le calcul d'une distance entre éléments à classer. Dans le cadre de la séparation de signaux ou notes de musique, nous proposons d'utiliser cette méthode de classification pour déterminer les T partitions du principe des formes fortes.
La méthode des Nuées Dynamiques possède de nombreuses variantes ("k-means" de Mac Queen [55], "isodata" de Ball et Hall [56] $\cdots$). Le principe de base de ces méthodes non-hiérarchiques est de déterminer une partition d'un ensemble E en k classes, ce nombre k étant fixé a priori. Nous développons ci-dessous la version

des Nuées Dynamiques proposée par Diday [57].

Soit E un ensemble fini muni d'une distance d. La distance d'un élément $x \in E$ à un sous-ensemble N de E est donnée par :

$$D(x, N) = \sum_{y \in N} d(x, y) \tag{4.3}$$

La distance entre deux sous-ensembles de E est donnée par :

$$D(N, M) = \sum_{x \in N} D(x, M) = \sum_{x \in N} \sum_{y \in M} d(x, y) \tag{4.4}$$

La méthode des Nuées Dynamiques, pour la classification en k classes des éléments de E, comporte les étapes suivantes :

1. la première étape consiste à prendre au hasard dans l'ensemble E, k sous-ensembles de q éléments, notés $\{N_1^0,\ N_2^0,\ \cdots,\ N_k^0\}$. Chacun des sous-ensembles est appelé "noyau". $N^0 = \{N_i^0\}_{i=1,\cdots,k}$ constitue l'ensemble des noyaux à l'étape d'initialisation.

2. la seconde étape associe à chaque noyau N_i^0 pour $i = 1 \cdots k$, un sous-ensemble $C_i^0 \in E$, contenant les éléments de E les plus proches de N_i^0 :

$$C_i^0 = \{x |\ D(x, N_i^0) < D(x, N_j^0),\ \forall j \neq i\} \tag{4.5}$$

   L'ensemble des classes $C^0 = \{C_i^0\}_{i=1,\cdots,k}$ constitue une partition de E.
   On définit alors une application f qui associe à l'ensemble des noyaux N^0, un ensemble de classes C^0 telle que :

$$f : N^0 \longrightarrow C^0 = f(N^0) \tag{4.6}$$

   .

3. Dans une troisième étape, on associe à la partition C^0 ainsi obtenue, un nouvel ensemble de noyaux N^1. Pour cela, on se donne un critère à valeurs positives qui permet de mesurer l'adéquation entre partition et noyau. L'ensemble $N^1 = \{N_i^1\}_{i=1,\cdots,k}$ est alors tel que pour tout $i = 1, \cdots, k$, le noyau

N_i^1 contient q éléments et minimise la distance $D(N_i^1, C_i^0)$.

On note alors g la fonction qui associe l'ensemble N^1 à la partition C^0 :

$$g : C^0 \longrightarrow N^1 = g(C^0) \tag{4.7}$$

.

4. Le procédé consiste ensuite à itérer les étapes 2 et 3, en appliquant successivement les fonctions f et g, à partir de l'ensemble d'initialisation N^0 choisi au hasard :

$$N^0 \xrightarrow{f} C^0 \xrightarrow{g} N^1 \xrightarrow{f} C^1 \xrightarrow{g} N^2 \xrightarrow{f} C^3 \xrightarrow{g} \cdots \tag{4.8}$$

• Convergence de l'algorithme

A la suite d'un nombre suffisant d'itérations, on obtient toujours les mêmes classes ou partitions C et les mêmes noyaux N. Une justification de la convergence de l'algorithme des Nuées Dynamiques est proposée par Jambu [54].
On remarquera tout de même, que les fonctions f et g, ainsi que le critère de minimisation de distance ne sont pas définis indépendamment : pour une partition C donnée, g associe une famille de noyaux N qui minimise le critère suivant défini par :

$$H(N, C) = \sum_{i=1}^{k} D(N_i, C_i) \tag{4.9}$$

De même pour N donné, f associe la partition qui minimise ce même critère H.
La méthode consiste alors :

- à se donner une famille de noyaux tirés aléatoirement N^0,
- à en déduire par f une famille de classes C^0. On a ainsi une première valeur de $H(N^0, C^0)$,
- à déterminer une nouvelle famille de noyaux N^1 donnée par l'application g, déduite de $H(N^0, C^0)$ telle que : $H(N^1, C^0) \leq H(N^0, C^0)$,
- à calculer de nouveau $C^1 = f(N^1)$ tel que : $H(N^1, C^1) \leq H(N^1, C^0)$,

– et ainsi de suite.

A chaque pas, le critère H mesurant l'adéquation entre C et N diminue. Comme ce critère est à valeurs positives, il atteint son minimum pour :

$$H(N^p, C^p) = H(N^p, C^{p+1}) = H(N^{p+1}, C^{p+1}) \tag{4.10}$$

et l'algorithme converge, soit encore : $C^p = f(N^p)$ et $N^p = g(C^p)$.

Pour un nombre fini d'individus de E, l'algorithme converge rapidement. La méthode des Nuées dynamiques a alors l'avantage de classifier de manière simple et rapide l'ensemble E en k classes, à partir de k noyaux pris au hasard. Son inconvénient est de comporter une part d'arbitraire : la partition obtenue après convergence, dépend du choix initial des noyaux.

4.3.3 Distances pour la classification de pics spectraux

Dans le cadre de la séparation de mélanges musicaux, les éléments que l'on souhaite classer sont les pics spectraux du mélange préalablement détectés.
Pour la p^{ieme} fenêtre de signal composite, l'étape de détection met à disposition un ensemble E_p de M_p pics spectraux :

$$E_p = \{\nu_1, \cdots, \nu_{M_p}\}, \quad \nu_1 < \cdots < \nu_{M_p} \tag{4.11}$$

Nous proposons ci-dessous, deux classifications par l'algorithme des Nuées Dynamiques. Elles diffèrent par la nature des noyaux représentant les classes, ainsi que par le critère de proximité utilisé, c'est-à-dire par les distances à minimiser lors de la classification. La première classification utilise une distance appelée *moindres carrés*. Les noyaux des classes sont des peignes harmoniques caractérisés par leur fréquence fondamentale. La seconde fait appel à une distance *sinusoïdale* caractérisée à la fois par une fréquence ω_0 et une période T_0. Le noyau associé à la i^{ieme} classe est donné par le couple (ω_0^i, T_0^i).
Pour les deux types de classification, le nombre de classes choisi est noté N_c.

• **"Classification** *moindres carrés*"

Cette méthode de classification a la particularité d'utiliser comme noyau de classe, des peignes harmoniques $\Pi(\omega)$, tels que pour la classe i :

$$\Pi(\omega_i) = \{\omega_i, 2\omega_i, 3\omega_i, \cdots\} \tag{4.12}$$

Cette classification correspond à une approche physique du problème, car elle consiste à regrouper les pics du spectre directement sous forme de peignes, traduisant ainsi le caractère harmonique des signaux musicaux.

Définition de la distance : Pour tout pic ν à classer, la distance utilisée est donnée par :

$$d(\nu, \Pi(\omega_i)) = \min_{k=1,2,\cdots} |\nu - k \times \omega_i|^2 \tag{4.13}$$

La distance d est l'écart minimal entre le pic fréquentiel ν et la k^{ime} harmonique du peigne $\Pi(\omega_i)$. Le pic ν est classé dans la classe C_i si :

$$\forall j \neq i, \qquad d(\nu, \Pi(\omega_i)) \leq d(\nu, \Pi(\omega_j)) \tag{4.14}$$

La distance d est appelée distance *moindres carrés*, car l'équation (eq. 4.13) peut être vue comme la minimisation d'une erreur quadratique moyenne entre le pic ν et la k^{ieme} harmonique de la i^{ieme} note.

Calcul du noyau : A chaque itération, le noyau de chaque classe doit être recalculé. Il doit alors "coller" au mieux avec les éléments de sa classe. Le calcul du nouveau peigne $\Pi(\omega_i)$ est obtenu par :

$$\omega_i = arg \min_{f \epsilon \Pi(\omega_i)} [\sum_{\nu \epsilon C_i} \min_k |\nu - k \times f|^2] \tag{4.15}$$

La double minimisation de l'expression (eq. 4.15) ne permet pas de calculer directement ω_i. On peut cependant supposer que la valeur k, notée k_{min}, qui minimise la quantité $|\nu - kf|^2$ est précisément l'ordre du pic ν dans la classe C_i.

Soit $0(\nu, \Pi(\omega_i))$ l'ordre harmonique d'un pic $\nu \in C_i$:

$$0(\nu, \Pi(\omega_i)) = arg \min_{k=1,2,\cdots} |\nu - k \times \omega_i|^2 \tag{4.16}$$

sous l'hypothèse $k_{min} = 0(\nu, \Pi(\omega_i))$, la minimisation devient :

$$\omega_i = arg \min_{f \,\epsilon\, \Pi(\omega_i)} [\sum_{\nu \,\epsilon\, C_i} |\nu - 0(\nu, \Pi(f))f|^2] \tag{4.17}$$

La fréquence ω_i du noyau $\Pi(\omega_i)$ de la classe C_i est finalement obtenue par annulation de la dérivé de l'expression (eq. 4.17) :

$$\omega_i = \frac{\sum_{\nu \epsilon C_i} 0(\nu, \Pi(\omega))\nu}{\sum_{\nu \epsilon C_i} 0(\nu, \Pi(\omega))^2} \tag{4.18}$$

Initialisation des noyaux : L'initialisation est effectuée par un tirage aléatoire de N_c pics dans l'ensemble E_p. En pratique, si un pic ν est classé dans C_i et est tel que $\nu < \omega_i$, alors l'ordre correspondant $0(\nu, \Pi(\omega_i))$ est mis à zéro. Dans ce cas, la classification n'est plus valable car l'équation (eq. 4.18) n'est plus définie. Ainsi, lorsque l'initialisation des noyaux fournit des fréquences élevées ou bien lorsque le bruit est important aux basses fréquences, l'algorithme de classification n'est plus applicable.
Un moyen de pallier ce problème consiste à définir lors de l'initialisation des noyaux, chaque fréquence $\{\omega_i\}$ pour $i = 1, \cdots, N_c$, comme étant la plus petite valeur sous-harmonique des N_c fréquences tirées aléatoirement.
Soient $\{\omega_i^o \,| i = 1, \cdots, N_c\}$, les N_c fréquences tirées aléatoirement dans l'ensemble des pics E_p. Les noyaux utilisés pour l'initialisation de l'algorithme de classification sont donnés par l'ensemble des sous-harmoniques $\{\omega_i \,| i = 1, \cdots, N_c\}$ telles que :

$$\omega_i = \frac{\omega_i^o}{k} \geq 27Hz \quad \text{avec} \quad k = 1, 2, \cdots \tag{4.19}$$

• **"Classification** *sinusoïdale*"

La seconde méthode de classification par les Nuées Dynamiques proposée utilise une distance de nature sinusoïdale caractérisée par une fréquence ω_0 et une période T_0. (ω_{0_i}, T_{0_i}) définissent le noyau de la classe i.

Définition de la distance : Pour un pic ν à classer, la distance sinusoïdale est définie par :

$$d(\nu, \omega_{0_i}) = sin^2[\pi T_{0_i}(\nu - \omega_{0_i})] \tag{4.20}$$

Le noyau (T_{0_i}, ω_{0_i}) est tel que :

$$T_{0_i} = \frac{1}{k_0.\omega_{0_i}} \qquad k_0 = 1, 2, \cdots \tag{4.21}$$

Comme le montre la figure (Fig. 4.5), on possède ainsi une sorte de peigne qui a l'avantage de ne pas prendre en compte tous les multiples de la fréquence ω_{0_i}. Cette distance vise ainsi à caractériser des spectres qui ne possèdent que des harmoniques paires ou impaires. Ces spectres très particuliers, interviennent notamment dans l'analyse de notes de clarinette ou encore de hautbois.

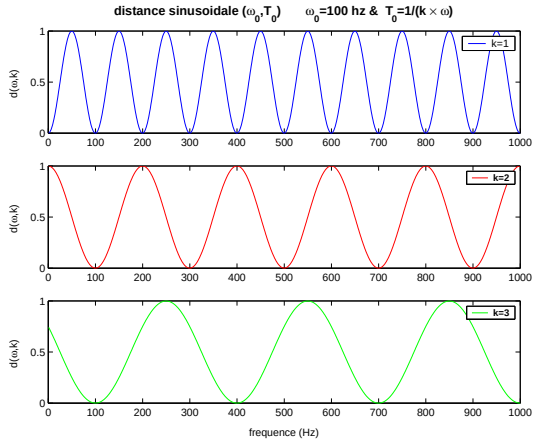


FIG. 4.5 – *Distance sinusoïdale pour* $\omega_0 = 100Hz$ *et* $k_0 = 1, 2\,et\,3$.

Calcul du noyau : Pour obtenir les paramètres ω_{0_i} et T_{0_i}, on minimise le critère suivant à chaque nouvelle itération :

$$J = \sum_{\nu \epsilon C_i} sin^2[\pi T_{0_i}(\nu - \omega_{0_i})] \tag{4.22}$$

En explicitant le paramètre k_{0_i}, on a alors :

$$(\omega_{0_i}, k_{0_i}) = arg \min_{f,k} \left[\sum_{\nu \epsilon C_i} sin^2[\frac{\pi}{k}(\frac{\nu}{f} - 1)] \right] \tag{4.23}$$

En pratique, le calcul du noyau de la classe C_i est effectué pour trois valeurs de k_{0_i}. On a alors $k \in \{1, 2, 3\}$.

Le calcul du noyau (ω_{0_i}, k_{0_i}) est fait sur une grille de valeurs déterminées (f, k). Cette méthode reste applicable à la classification de type "moindres carrés" présentée auparavant. Ainsi, plutôt que de supposer que l'ordre harmonique $0(\nu, \Pi(\omega_{0_i}))$ minimise l'expression (eq. 4.15), ce dernier peut être explicité et le noyau représentant la classe C_i est donné par le couple $(\omega_{0_i}, 0(\nu, \Pi(\omega_{0_i}))$.
L'inconvénient de cette méthode est que contrairement à la distance sinusoïdale, les valeurs que l'on doit considérer pour calculer l'ordre du noyau doivent être largement supérieures à 3, car les signaux musicaux possèdent généralement plus de 3 harmoniques. Le calcul du noyau sur une grille de valeurs déterminées, pour la distance *moindres carrés*, apparaît alors comme une méthode relativement lourde et nous préférerons utiliser l'hypothèse de l'ordre harmonique proposée auparavant.

Unicité de la solution : Il existe un problème majeur pour la minimisation définie par (eq. 4.23) : il n'y a pas unicité de la solution. En effet, la fonction sinusoïdale est $1/T_{0_i}$-périodique et on a :

$$sin^2[\pi T_{0_i}(\nu - \omega_{0_i})] = sin^2[\pi T_{0_i}(\nu - \omega_{0_i} - \frac{1}{T_{0_i}})]$$

Tous les couples de la forme $(\omega_{0_i} + \frac{k}{T_{0_i}}, T_{0_i})$ avec k entier relatif, minimisent le critère J de l'expression (eq. 4.22). Cette périodicité à l'inconvénient de favoriser les fréquences basses lors du calcul des noyaux $\{\omega_{0_i}\}_{i=1,\cdots,N_c}$. Ces dernières correspondent généralement à des pics de bruit détectés à des fréquences sous-multiples des fréquences fondamentales du mélange. Elles ont pour effet d'introduire aux itérations suivantes des fréquences qui ne devraient pas se trouver dans la classe qui leur est associée.
Ce phénomène est également présent dans le cas de la classification de type "moindres carrés", puisque des sous-harmoniques sont directement choisies à l'initialisation des noyaux. En revanche, dans le cas sinusoïdal, il existe un moyen simple de pallier le problème : la distance sinusoïdale eq(eq. 4.20) peut être modulée pour former une nouvelle distance utilisable pour la classification.

Distance sinusoïdale modulée : Pour obtenir un signal pseudo-périodique, la distance sinusoïdale est modulée par une fonction $x \longrightarrow \frac{1}{x^\alpha}$. La nouvelle distance est telle que :

$$d(\nu, \omega_{0_i}) = \begin{cases} \left| \frac{sin^2[\pi T_{0_i}(\nu - \omega_{0_i})]}{[\frac{\pi}{\omega_{0_i}}(\nu - \omega_{0_i})]^\alpha} \right| & si\ \nu \neq \omega_{0_i} \\ \\ 0 & si\ \nu = \omega_{0_i} \end{cases} \tag{4.24}$$

Plus le paramètre α est grand, plus les "lobes secondaires" de la fonction "sin^2" ont une amplitude faible. Plus α tend vers 0, plus on se rapproche de la distance sinusoïdale (eq. 4.20). Le phénomène est illustré sur la figure (Fig. 4.6)
Dans le cadre musical, cette pondération possède un second avantage : elle permet en effet, de tolérer plus d'erreurs sur les fréquences élevées, qui sont justement celles où l'inharmonicité pour les spectres musicaux est la plus importante.
Le choix de la valeur prise par α dans la classification, est un choix heuristique, déterminé à la suite de tests comparatifs et qui permet de pénaliser correctement les hautes fréquences.
Dans la suite, nous utiliserons $\alpha = 1/5$.

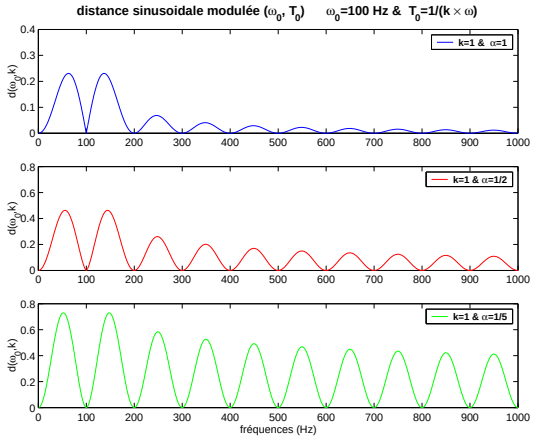


FIG. 4.6 – *Distance sinusoïdale modulée* pour $\omega_0 = 100Hz$, $k_0 = 1$ et $\alpha = 1, 1/2$ *et* $1/5$.

4.3.4 Evaluation des performances des Nuées Dynamiques

Cette section étudie et compare les performances de l'algorithme des Nuées Dynamiques, lorsque les distances *moindres carrés* et *sinusoïdale* sont utilisées. Les performances sont évaluées par le calcul de pourcentages de classifications correctes obtenus pour des jeux de mélanges synthétiques contenant jusqu'à six notes.

Le concept de "classification correcte" englobe deux types de classification : la première dite correcte au "Sens Strict" (CS), fait référence aux partitions de l'ensemble permettant d'estimer toutes les fréquences fondamentales du mélange. La seconde dite correcte au "Sens Large" (CL), correspond aux partitions permettant d'estimer au moins l'une des fréquences fondamentales du mélange. Le calcul des pourcentages de CS, permet de quantifier les problèmes de convergence de la méthode des Nuées Dynamiques pour les deux distances. Les pourcentages de CL, quant à eux, mettent en évidence la quantité d'information contenue dans les classes, qui sera ensuite exploitée par le principe des formes fortes.
Le protocole d'évaluation permettant le calcul des différents pourcentages est détaillé par la suite. Comme les performances des classifications vont dépendre du nombre de classes arbitraire N_c, trois cas seront considérés : le cas où le nombre de sources N_s est égal au nombre de classes ($N_c = N_s$), le cas où le nombre de classes est surestimé ($N_c > N_s$) et celui où le nombre de classes est sous-estimé ($N_c < N_s$).

• Critères pour une classification correcte

Afin de mettre en évidence les performances de la classification par les Nuées Dynamiques, deux critères différents pour le calcul des pourcentages de CS et CL sont proposés : le premier est basé sur l'estimation des fréquences fondamentales, tandis que le second s'intéresse directement au contenu des classes.

Critère basé sur l'estimation de fréquence fondamentale ($\forall N_c$) **:** Pour tous les types de classification considérés, les pourcentages de CS et CL sont calculés par l'intermédiaire du nombre de fréquences fondamentales correctement estimées après classification. Les définitions détaillées de ces classifications CS et CL sont présentées ci-dessous :

– une classification correcte au sens stricte ou $CS(f)$ se définit comme une classification pour laquelle les classes constituées permettent l'estimation de la fréquence fondamentale de toutes les notes du mélange.

– une classification correcte au sens large ou $CL(f)$ se définit comme une classification pour laquelle les classes constituées permettent l'estimation de la fréquence fondamentale d'au moins une note du mélange.

Ce critère basé sur l'estimation de fréquence fondamentale, permet de quantifier les performances de la classification, car plus les classes vont être erronées, plus l'estimation de fréquence fondamentale sera difficile : un faible pourcentage de fondamentales correctes va traduire une classification de mauvaise qualité.

Critère basé sur le contenu des classes ($N_c = 2$, $N_s = 2$) **:** Pour d'un mélange de 2 notes, on s'intéresse en détail au contenu des classes obtenues pour une classification par les Nuées Dynamiques utilisant 2 classes.

Ainsi, pour le cas ($N_c = 2$, $N_s = 2$), un second critère est proposé. Il consiste à déterminer les classifications pour lesquelles chaque classe contient tous les éléments fréquentiels d'une des 2 notes et éventuellement des pics de bruit. Ces classifications peuvent être vue comme des classifications correctes au sens strict et sont dites $CS(c)$.

A l'instar des $CL(f)$, nous calculerons des $CL(c)$. Elles sont définies comme les classifications pour lesquelles l'une des 2 classes contient tous les éléments d'une même note connue, auxquels viennent s'ajouter quelques éléments de la seconde. Afin de quantifier la présence de la seconde note, on associera à chaque $CL(c)$ un taux de recouvrement. Bien évidemment, plus le taux de recouvrement est faible, plus une $CL(c)$ se rapproche d'une $CS(c)$. Les différents taux de recouvrement pris en compte, appartiennent à l'intervalle $]0\%, 100\%[$.
On notera finalement qu'un taux de 0% correspond à une $CS(c)$, tandis que pour un taux de 100%, l'une des classes contient les pics des 2 notes connues et la classification ne peut être ni $CS(c)$, ni $CL(c)$.

Le tableau (Tab. 4.1), récapitule les définitions relatives au concept de "classification correcte" :

$CS(f)$:	Classification permettant l'estimation de la fréquence fondamentale de toutes les notes du mélange.
$CL(f)$:	Classification permettant l'estimation de la fréquence fondamentale d'au moins une note du mélange.
$CS(c)$:	Classification pour laquelle chaque classe contient tous les éléments fréquentiels d'une des 2 notes et éventuellement des pics de bruit. ($N_s = 2, N_c = 2$)
$CL(c)$:	Classification pour laquelle l'une des 2 classes contient tous les éléments d'une même note et des éléments de la seconde dont la note. ($N_s = 2, N_c = 2$)

Tab. 4.1 – *Définitions des classifications correctes*

• **Protocole d'évaluation**

On considère un jeu de N_b mélanges synthétiques, tous constitués par le même nombre de sources harmoniques N_s et d'un bruit blanc gaussien de variance connue. pour La i^{ieme} source harmonique est une somme de M_i sinusoïdes harmoniques dont la fréquence fondamentale est f_{0_i} ($i = 1, \cdots, N_s$). Les N_s fréquences fondamentales sont choisies telles qu'il n'existe aucun recouvrement entre leurs harmoniques. Le rapport de puissance entre les sources est supposé égal à 0 dB. En revanche, un jeu de N_b mélanges est caractérisé par son rapport signal à bruit en dB, noté RSB. La longueur d'analyse d'un mélange est de 100 ms.
Ainsi, la caractère aléatoire de la partie bruitée des mélanges, est utilisé pour A un RSB fixé correspond alors un jeu de N_b mélanges sur lesquels l'algorithme des Nuées Dynamiques est appliqué pour classifier les pics spectraux. Ces pics ont été préalablement détectés par la méthode basée sur le périodogramme. Les pourcentages de CS et de CL en fonction du RSB sont calculés, en déterminant pour chaque valeur de RSB le nombre de classifications correctes. L'étude qui suit utilise les valeurs de $RSB = \{0, 10, 20, 30\}$ dB. $N_b = 1000$ réalisations sont prises en compte dans le calcul des pourcentages. Les résultats obtenus pour la distance *moindres carrés* et la distance *sinusoïdale*, sont présentées en parallèle, tout au long de l'étude.

• **Cas d'une note synthétique unique** ($N_s = 1$)

On s'intéresse tout d'abord au cas d'une note synthétique à laquelle on applique l'algorithme des Nuées Dynamiques pour un nombre de classes N_c variant entre 2 et 6 classes. La figure (Fig. 4.7) présente alors le pourcentage de $CS(f)$ obtenu pour les trois cas $N_c = 2$, 3, 6.

Pour les deux distances, il est possible d'estimer la fréquence fondamentale de la note en appliquant sur la partition obtenue, la méthode de compression spectrale présentée dans le paragraphe (4.2.3). Cet estimateur reste efficace même lorsque le nombre de classes N_c surestime largement le nombre de source N_s.

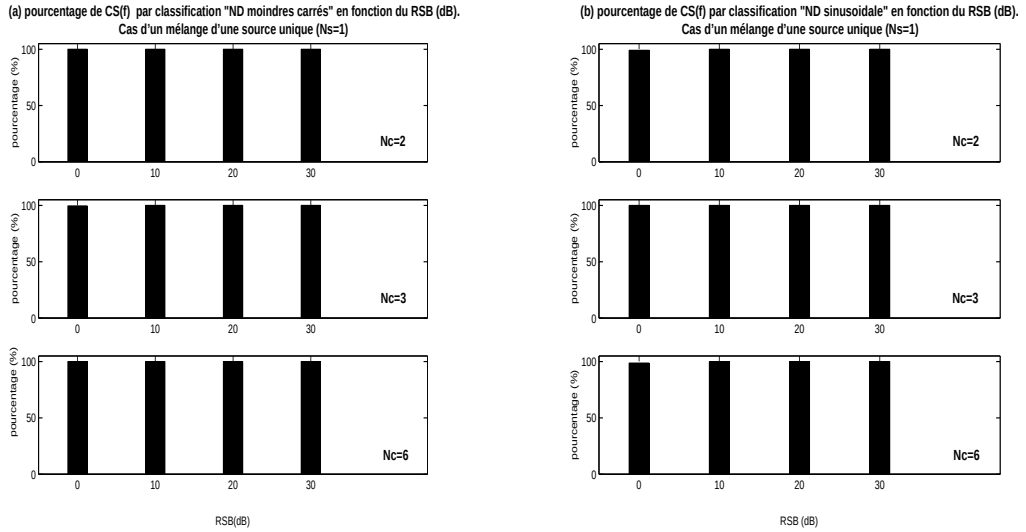


FIG. 4.7 – *Pourcentage de $CS(f)$ en fonction du RSB(dB) pour un mélange d'une note synthétique : a) Cas de la distance "moindres carrés" pour N_c= 2, 3, 6 - b) Cas de la distance "sinusoidale" pour N_c= 2, 3, 6.*

• **Cas d'un mélange de 2 notes synthétiques** ($N_s = 2$)

La figure (Fig. 4.8) présente les pourcentages de $CS(f)$ et de $CL(f)$ obtenus pour la distance "moindres carrés" et la distance "sinusoïdale" lorsque N_c varie entre 2 et 6 classes.
Quelque soit la distance et le nombre de classes choisi pour effectuer la classification, le pourcentage de $CL(f)$ est toujours égal à 100% et cela pour toutes les valeurs de RSB considérées. Ainsi, dans le cas d'un mélange de deux notes, on

peut supposer que la classification par les Nuées Dynamiques permet de regrouper un maximum d'information sur au moins l'une des deux notes.
A nombre de classes égal N_c, la comparaison des deux types de classification, montre que la distance "sinusoïdale" permet d'obtenir des pourcentages de $CS(f)$ plus importants pour les faibles valeurs de RSB, tandis que la distance "moindres carrées" aura tendance à donner plus de classifications $CS(f)$ pour des RSB forts.

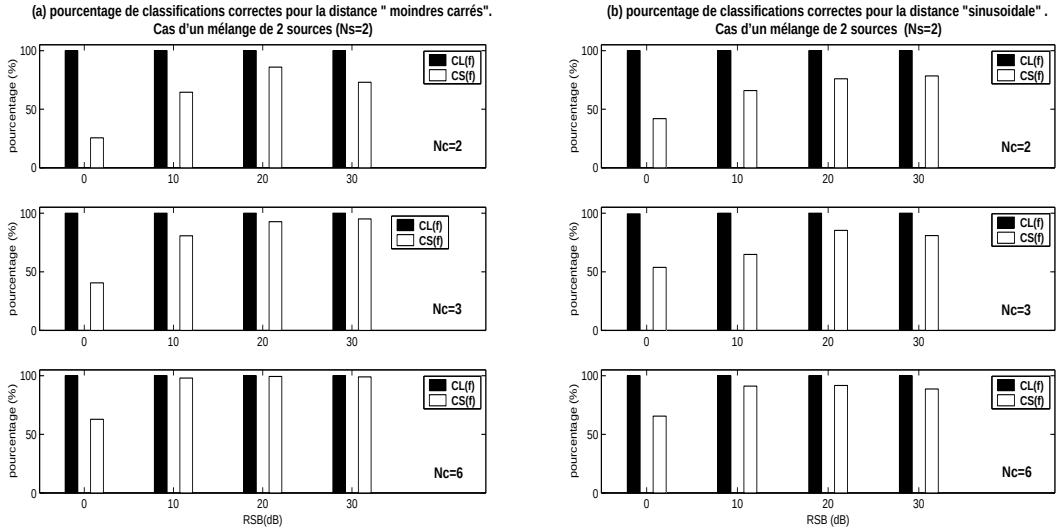


Fig. 4.8 – *Pourcentage de $CL(f)$ et de $CS(f)$ en fonction du RSB (dB) pour un mélange de 2 notes synthétiques : a) Cas de la distance "moindres carrés" pour N_c= 2, 3, 6 - b) Cas de la distance "sinusoidale" pour N_c= 2, 3, 6.*

Cas où $N_s = Nc = 2$ **:** Dans ce cas, on peut déterminer les pourcentages de classifications correctes en terme de contenu de classes. La figure (Fig. 4.9) présente alors les pourcentages de $CS(c)$ et de $CL(c)$ pour le mélange de 2 notes. Les différents taux de recouvrement correspondant aux $CL(c)$ sont, quant à eux illustrés par la figure (Fig. 4.10). Pour les deux distances, plus le RSB augmente plus le nombre de classes correctes est important, qu'elles soient correctes au sens strict ou au sens large. De même plus le RSB augmente, plus le nombre de classes faiblement recouvertes augmente. La distance "sinusoïdale" apparaît ici beaucoup plus performante que la distance "moindres carrés". Elle donne un plus grand nombre de classifications $CS(c)$ et permet ainsi d'obtenir un pourcentage de classifications correctes (pourcentage global) quasi-maximal pour un $RSB = 30dB$.

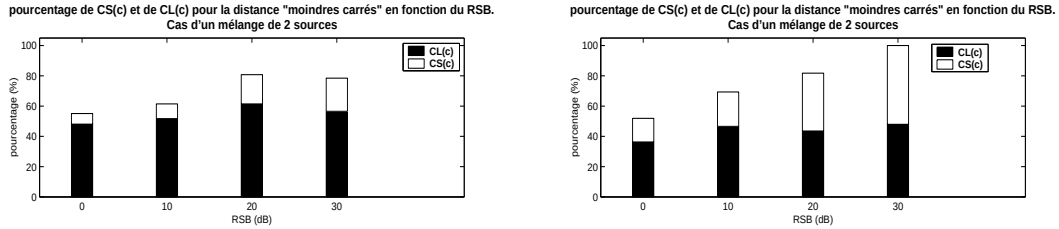


Fig. 4.9 – *Pourcentage de $CL(c)$ et de $CS(c)$ en fonction du RSB(dB) pour un mélange de 2 notes synthétiques : a) Cas de la distance "moindres carrés" - b) Cas de la distance "sinusoidale".*

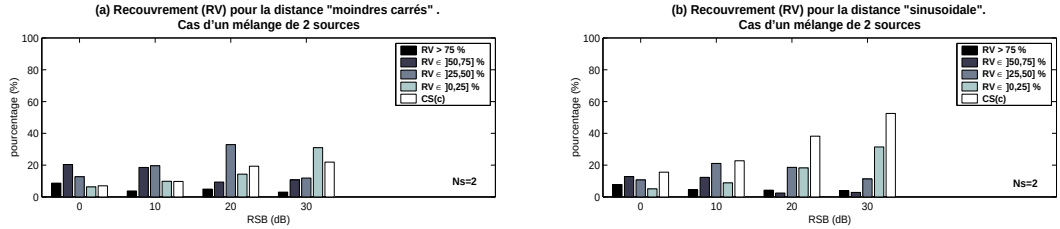


Fig. 4.10 – *Répartition des $CL(c)$ pour un mélange de 2 notes synthétiques en fonction de leur taux de recouvrement RV : a) Cas de la distance "moindres carrés" - b) Cas de la distance "sinusoidale".*

• **Cas d'un mélange $N_s > 2$ avec $N_c > N_s$**

Les pourcentages de classifications $CS(f)$ et $CL(f)$ sont calculés pour des mélanges synthétiques de 4 et 6 notes. On remarque alors que les deux distances donnent des pourcentages de $CS(f)$ nuls, quelque soient les valeurs de RSB considérées. Cela signifie qu'aucune des distances n'est capable de fournir des partitions à partir desquelles toutes les fréquences fondamentales du mélange sont estimables.

La figure (Fig. 4.11) présente finalement les pourcentages de $CL(f)$, ainsi que le pourcentage de classifications correctes permettant l'estimation d'au moins 2 des fréquences fondamentales. Ces classifications peuvent être vues comme des classifications correctes définies au sens large. Elles sont notées dans la suite $CL(2f)$.

L'augmentation du nombre de notes dans le mélange ne permet plus d'obtenir des pourcentages de $CL(f)$ maximum pour des RSB de 0 dB et 10 dB et cela quelque soit la distance considérée. On notera tout de même que la distance *sinusoïdale* donne des pourcentages plus importants que la distance *moindres carrés*. Cette tendance se retrouve également dans le cas des $CL(f)(2)$ ou la différence entre les deux distances considérées est cette fois-ci très nettement visible. La distance *sinusoïdale* apparaît de nouveau plus performante que la distance *moindres carrés*.

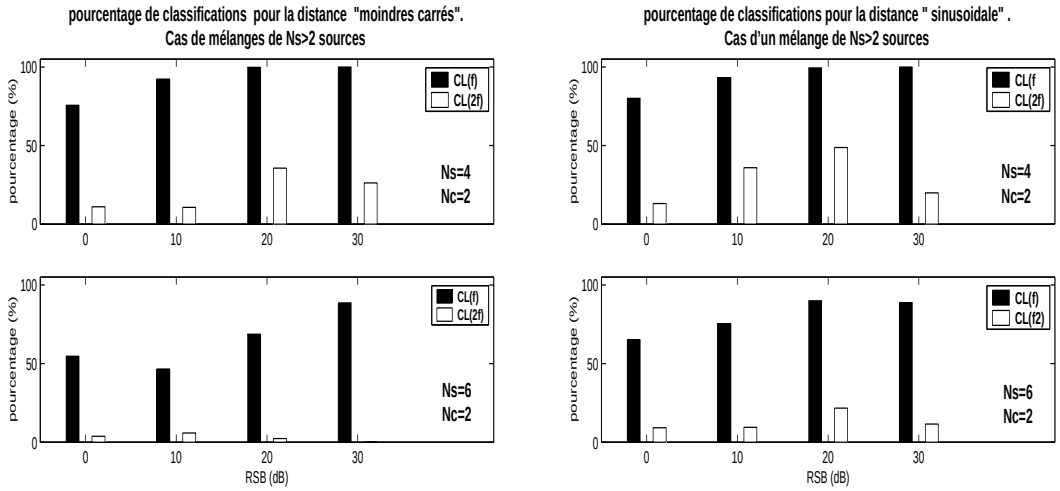


Fig. 4.11 – *Pourcentage de $CL(f)$ et de $CL(2f)$ en fonction du RSB(dB), pour des mélanges de 4 et 6 notes synthétiques : a) Cas de la distance "moindres carrés" pour $N_c = 2$ - b) Cas de la distance "sinusoidale" pour $N_c = 2$.*

• **Cas d'un mélange $N_s > 2$ avec $N_c > N_s$**

La figure (Fig. 4.12) présente les pourcentages de $CS(f)$ et de $CL(f)$ lorsque le nombre de classes $N_c = 10$ choisi pour les Nuées Dynamiques surestime le nombre de notes du mélange de 6 sources utilisé précédemment.

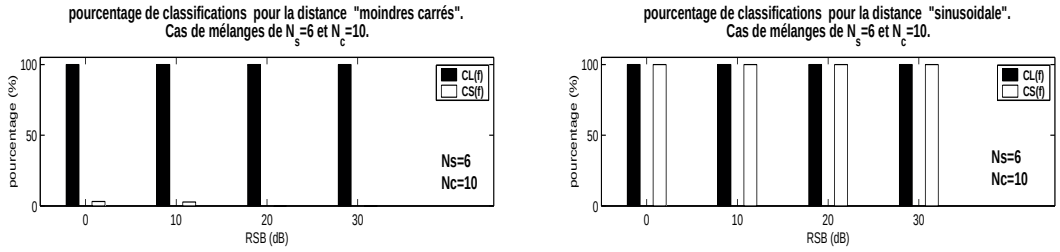


Fig. 4.12 – *Pourcentage de $CL(f)$ et de $CS(f)$ en fonction du RSB(dB), pour un mélange de 6 notes synthétiques : a) Cas de la distance "moindres carrés" pour $N_c = 10$ - b) Cas de la distance "sinusoidale" pour $N_c = 10$.*

La comparaison de ces résultats à ceux de la figure (Fig. 4.11) montrent que l'on a tout intérêt à surestimer le nombre de sources du mélange lorsque l'algorithme des Nuées Dynamiques est appliqué, puisque les pourcentages de $CS(f)$

et $CL(f)$ augmentent avec N_c. Pour la distance sinusoïdale, cette valeur de N_c permet même l'estimation de toutes les fréquences fondamentales du mélange et cela quelque soit le RSB.

• **Conclusions**

L'ensemble des simulations permet d'évaluer les performances des deux distances proposées précédemment. Les faibles pourcentages de CS montrent que la méthode des Nuées Dynamiques ne peut être utilisée directement pour la classification des pics spectraux.
En revanche, les classifications CL mettent en évidence toute l'information contenue dans les classes obtenues après application de la méthode des Nuées Dynamiques. La comparaison des classifications pour les deux distances proposées, montre que la distance *sinusoïdale* donne dans la plupart des cas de meilleures performances. C'est cette dernière qui sera par la suite utilisée pour le problème de classification des pics spectraux.

4.4 Estimation par pénalisation sur la classe

Nous développons dans ce paragraphe, une première méthode d'estimation de fréquences fondamentales multiples. Pour cette méthode, la classification des pics spectraux est effectuée par le principe des formes fortes, présenté dans le paragraphe (4.3.1). Il consiste à pénaliser les fréquences deux à deux, selon les classes qu'elles occupent, dans un jeu de partitions obtenue par l'Algorithme des Nuées Dynamiques. On parlera alors d'une estimation de fréquences fondamentales multiples *par pénalisation sur la classe*.
La construction des formes fortes est tout d'abord détaillée, puis les performances d'estimation de cette méthode d'estimation sont évaluées sur mélanges synthétiques de notes.

4.4.1 Construction des formes fortes

L'application du principe des formes fortes dans le cas de la classification des pics spectraux d'un mélange de notes est effectuée comme suit. Pour la p^{ieme}

fenêtre d'analyse d'un mélange, dont le nombre N_s de notes est inconnu, la construction des formes fortes, se fait à partir de T partitions de l'ensemble des pics détectés $E_p = \{\nu_1, \nu_2, \cdots, \nu_{M_p}\}$. Ces partitions sont déterminées à l'aide de l'algorithme des Nuées Dynamiques présenté dans le paragraphe (4.3.2). La distance utilisée est la distance *sinusoïdale* (eq. 4.24).
Pour chaque couple (ν_i, ν_j) de E_p, on associe l'indice de pénalité $\delta(\nu_i, \nu_j)$:

$$\begin{aligned} \delta(\nu_i, \nu_j) \quad = \quad & \textit{nombre de partitions pour lesquelles} \\ & \nu_i \textit{ et } \nu_j \textit{ appartiennent à des classes différentes} \end{aligned}$$

On détermine dans un premier temps les indices de pénalité pour tous les couples $(i, j) \in \{1, 2, \cdots, M_p\}$, puis à partir d'un seuil de pénalité Δ, on construit les formes fortes de la façon suivante :

- **Construction des formes :** pour chaque fréquence ν_i de l'ensemble E_p, on associe toutes les fréquences $\nu_{j \neq i}$ telles que $\delta(\nu_i, \nu_j) < \Delta$.
  A cette étape, les classes obtenues définissent M_p formes fortes.

- **Suppression des classes parasites :** en règle général, le principe des formes fortes fait apparaître un certain nombre de classes qui ne contiennent qu'un seul élément. Ces dernières peuvent alors être considérées comme parasites et sont éliminées. Dans le cas de la classification des pics spectraux, ces classes vont le plus souvent correspondre à des pics de bruit.

- **Suppression du recouvrement :** une dernière étape consiste à vérifier que deux formes fortes ne se recouvrent pas : si tous les éléments d'une forme forte sont déjà contenus dans une autre, alors la forme forte entièrement recouverte est éliminée.

Le nombre de formes fortes finalement obtenu correspond alors au nombre estimé de notes du mélange, noté $\hat{N}_s$.

Le principe des formes fortes, appliqué aux T partitions de E_p obtenues par les Nuées Dynamiques, permet d'exploiter toute l'information de ces classifications, qu'elles soient correctes aussi bien au sens strict (CS) qu'au sens large (CL). En effet, cette information n'est plus traitée globalement sur chaque classification, mais sur chaque couple de fréquences. Ainsi, même si une classification ne permet pas de regrouper tous les pics de chaque note dans des classes distinctes, on peut supposer qu'au fil des T partitions, les fréquences d'une même note seront le plus

souvent regroupées ensemble.

Outre sa raison initiale qui permet de réduire l'arbitraire des Nuées Dynamiques, la construction de formes fortes possède deux avantages supplémentaires, non négligeables pour le problème de séparation de notes de musique. Le premier avantage est qu'elle permet d'estimer le nombre de sources du mélange et ce, quelque soit le nombre de classes N_c choisi dans l'algorithme des Nuées Dynamiques. Le second avantage est qu'elle permet à un pic spectral d'appartenir à plusieurs formes fortes en même temps. Nous verrons ainsi par la suite, qu'un estimateur basé sur le principe des formes fortes, est capable de traiter des mélanges complexes, telles que des notes jouées à la tierce ou à la quinte. Ces mélanges sont caractérisés par un recouvrement spectral entre les deux notes, qui les rend généralement difficile à analyser.
En revanche, on constate d'ores et déjà que le cas de notes jouées à l'octave ne peut être traité par cette méthode d'estimation, puisque la dernière étape de la construction des formes fortes élimine les formes entièrement recouvertes par une autre.

4.4.2 Etude de la pénalisation sur la classe

La partition finale obtenue après pénalisation sur la classe, dépend du choix du seuil de pénalité Δ, du nombre T de partitions préalablement effectuées, ainsi que du nombre de classes N_c de ces partitions. L'étude qui suit, a pour objectif de déterminer Δ, T et N_c permettant d'estimer au mieux les fréquences fondamentales d'un mélange donné. Elle consiste à déterminer les pourcentages de fréquences fondamentales correctement estimées, ainsi que le nombre moyen de sources estimées, pour la méthode d'estimation de fréquences fondamentales multiples proposée ci-dessus. Le protocole d'évaluation est tout d'abord détaillé, puis les résultats obtenus pour des mélanges possédant jusqu'à 6 notes synthétiques sont présentés.

• Protocole d'évaluation

Afin de tester la méthode d'estimation proposée, nous proposons dans cette partie, un ensemble de simulations utilisant des mélanges synthétiques, analysés sur des durées de 100 ms. Un mélange synthétique est la somme de N_s sources harmoniques et d'un bruit blanc de variance connue. La i^{ieme} source harmonique

est une somme de M_i sinusoïdes harmoniques dont la fréquence fondamentale est f_{0_i} ($i = 1, \cdots, N_s$). Le rapport de puissance entre les sources est supposé égal à 0 dB. Le rapport signal à bruit, noté RSB, varie dans cette étude entre -15 dB et 30 dB.
Dans la suite, une fréquence fondamentale est supposée correctement estimée, lorsque l'écart entre l'estimé est égal à 3% à sa véritable valeur. Ce pourcentage, qui correspond à une précision au quart de ton dans l'échelle musiale, permet de discriminer deux notes séparées d'un demi-ton.

L'évaluation des performances d'estimation de la méthode est faite sur différents types de mélanges. Ils se distinguent par leur nombre de notes N_s, le choix des fréquences fondamentales qui permet ou non un recouvrement spectral de ces notes, ainsi que le niveau de bruit du mélange.
Pour un type de mélange fixé, les pourcentages de fréquences fondamentales estimées $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et le nombre moyen de sources estimées $\hat{N}_s$, sont calculés à partir $N_b = 100$ réalisations indépendantes.
Différentes valeurs de Δ, T et N_c sont prises en compte. Leur comparaison va permettre de justifier les choix fait sur ces paramètres, ces derniers devant optimiser les pourcentages $\{P_i(\%)\}_{i=1,\cdots,N_s}$, ainsi que le nombre de sources $\widehat{N}_s$.

Seuil de pénalité : Le seuil de pénalité Δ étant directement lié au nombre de partitions obtenues par les Nuées Dynamiques, on définit ce dernier en fonction de T :

$$\Delta(T) = \frac{T}{\beta} \quad \text{et} \quad \beta = \{2, 3, 5, 8, 10\} \tag{4.25}$$

Le paramètre β est appelé coefficient de pénalité.

Performances de l'estimation par compression spectrale : En plus du calcul des pourcentages $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et du nombre moyen de sources $\widehat{N}_s$, nous nous intéresserons aux performances mêmes de l'estimateur, utilisé sur les spectres séparés correspondant à chaque forme forte, pour déterminer les fréquences fondamentales du mélange. Cet estimateur, basée sur une méthode de compression spectrale est présenté dans le paragraphe (4.2.3). L'idée est alors de déterminer la quantité d'information contenue dans une forme forte, qui permet l'estimation de l'une des fréquences fondamentales du mélange.
Tout comme pour l'étude des classifications par les Nuées Dynamiques (cf. paragaphe (4.3.4)), nous définissons des taux de recouvrement pour les formes fortes.

Ces derniers vont permettre de faire correspondre aux pourcentages $\{P_i(\%)\}_{i=1,\cdots,N_s}$, la répartition des différents types de formes fortes à l'origine des ces estimations.

Les différents taux de recouvrement sont obtenus en associant à chaque forme forte, la note qui possède le plus de pics fréquentiels dans cette forme forte. Plus précisément, si pour une forme forte donnée, 75% des pics correspondent à une même note, tandis que 50% des pics correspondent à une seconde note, alors la forme forte est associée à la première note et le taux de recouvrement est supposé égal à 25%. Le tableau (Tab. 4.2) résume les différents recouvrements pris en compte dans l'étude.

$FF1$:	forme forte dont la note associée est recouverte à plus de 75%
$FF2$:	forme forte dont la note associée a un taux de recouvrement contenu dans l'intervalle [75,50[%
$FF3$:	forme forte dont la note associée a un taux de recouvrement contenu dans l'intervalle [50,25[%
$FF4$:	forme forte dont la note associée a un taux de recouvrement contenu dans l'intervalle [25,0[%
$FF5$:	forme forte dont la note associée a un taux de recouvrement égal à 0%

Tab. 4.2 – *formes fortes définies selon leur taux de recouvrement*

Ces formes fortes, pour lesquelles on définit un taux de recouvrement restent très proches du concept des classifications CS et CL, développées dans l'étude de la classification par la méthode des Nuées Dynamiques. Ainsi les formes fortes à recouvrement nul sont équivalentes à de formes fortes correctes au sens strict (ou FS), tandis que la somme des $FF1$, $FF2$, $FF3$ et $FF4$ définit les formes fortes correctes au sens large (ou FL).

• Influence du nombre de partitions T

On s'intéresse tout d'abord au nombre d'applications de l'algorithme des Nuées Dynamiques choisi pour construire les formes fortes. Pour cela, on considère un mélange de 2 notes analysé à l'aide de $T = 300$ classifications et $T = 30$ classifications. Les fréquences fondamentales sont choisies de telle sorte qu'il n'y ait pas de recouvrement spectral entre les deux notes. Les figures (Fig. 4.13) et (Fig. 4.14) présentent les pourcentages de fondamentales estimées $\{P_i(\%)\}_{i=1,2}$, ainsi que le nombre moyen de sources $\hat{N}_s$ en fonction du coefficient de pénalité β (eq. 4.26). Le nombre de classes N_c d'une partition est égal à 2 et le rapport signal à bruit est de 30 dB.

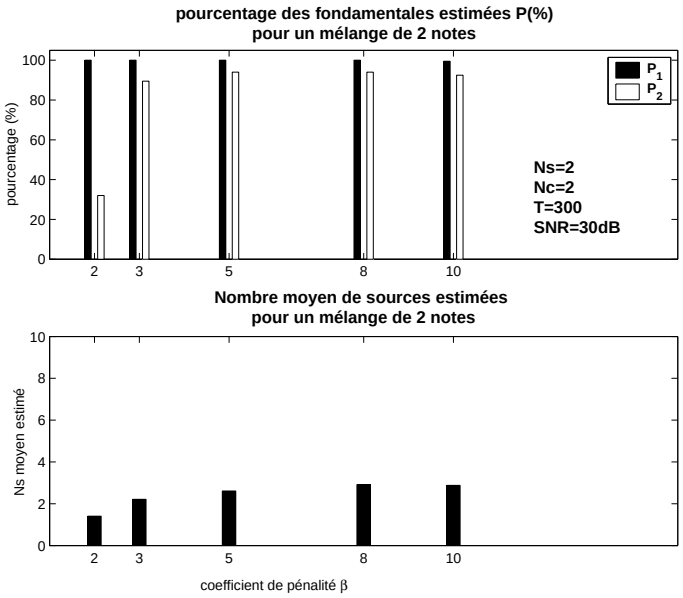


Fig. 4.13 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec $N_c = 2$ et $T = 300$ pour un mélange de 2 notes où le RSB=30dB.*

La comparaison des résultats montrent qu'il n'est pas nécessaire de prendre un nombre de classifications très grand pour obtenir de forts pourcentages d'estimation. Ceci est principalement dû au seuil Δ qui est défini en fonction de T et qui apparaît ainsi comme le seul paramètre dont les résultats vont dépendre. Ainsi, quelque soit T, les meilleurs pourcentages de fréquences fondamentales estimées et la meilleure estimation du nombre de source sont obtenus pour $\beta = 3, 5$.

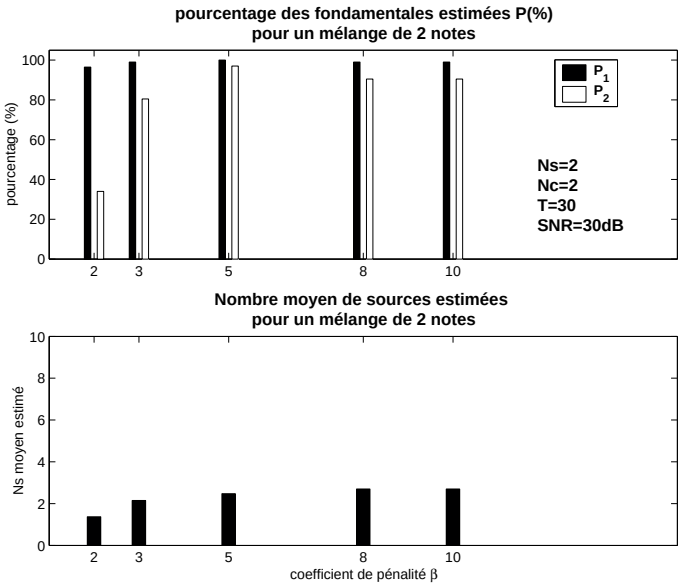


FIG. 4.14 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec* $N_c = 2$ *et* $T = 30$ *pour un mélange de 2 notes où le RSB=30dB.*

• Influence du RSB

Lorsque le rapport signal à bruit entre les notes synthétiques et le bruit blanc du mélange diminue, les performances de la méthode d'estimation de fréquences fondamentales diminue également. Les figures (Fig. 4.15) et (Fig. 4.16) illustrent les cas où le $RSB = -15$ dB et $RSB = 0$ dB. Dans cet exemple, la méthode d'estimation est à nouveau appliquée à un mélange de 2 notes synthétiques, qui ne se recouvrent pas spectralement. Le nombre de classes N_c pour une partition est choisi égal à 2.

Lorsque le RSB est très faible, les meilleurs pourcentages $P_1(\%)$ et $P_2(\%)$ de chaque note du mélange, restent inférieurs à 50% et le nombre estimé de sources correspondant est très largement supérieur à 2. Pour diminuer ce nombre $\hat{N}_s$, il est possible de prendre un coefficient β plus grand, mais cela implique une forte diminution des pourcentages d'estimation. Pour un $RSB = 0$ dB, on obtient des pourcentages d'estimation proches de 100% lorsque β est égal à 3 ou 5, mais là encore $\widehat{N}_s$ surestime très largement le nombre de sources.

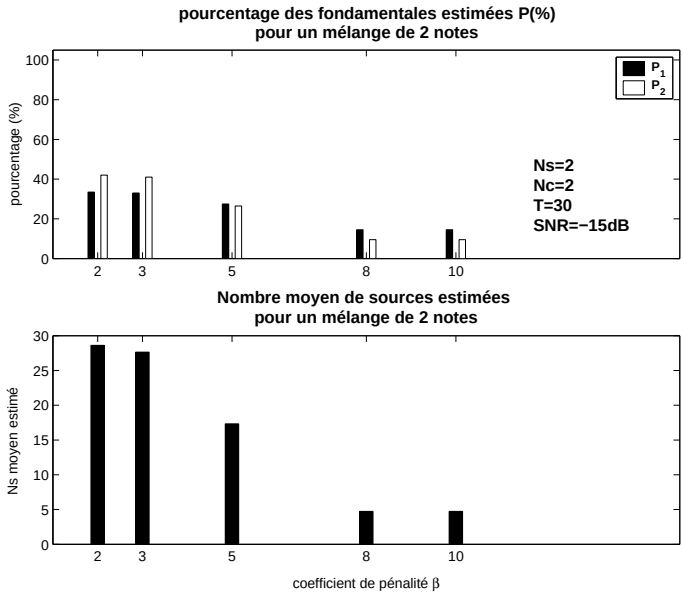


Fig. 4.15 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec $N_c = 2$ et $T = 30$ pour un mélange de 2 notes où le RSB= -15dB.*

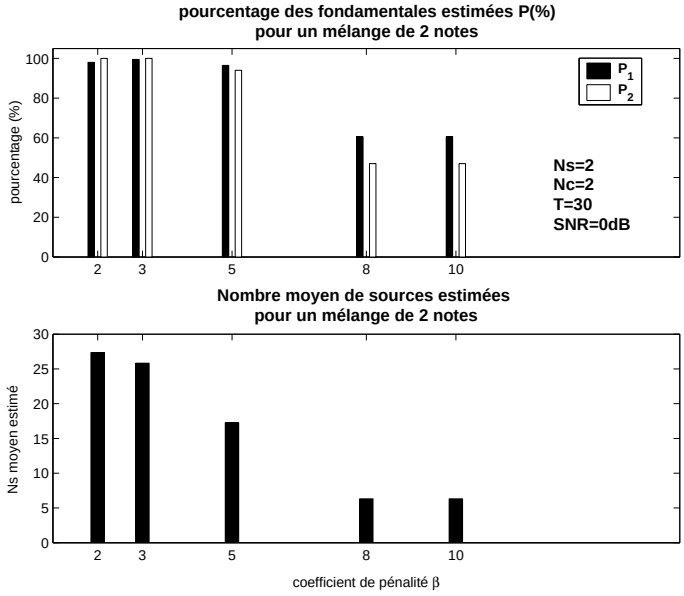


Fig. 4.16 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec $N_c = 2$ et $T = 30$ pour un mélange de 2 notes où le $RSB = 0$ dB.*

• **Influence du nombre de classes N_c**

On s'intéresse à l'influence du nombre de classes des partitions utilisées, dans le cas d'un mélange de 2 notes. Ces sources sont choisies, comme précédemment sans recouvrement spectrale et le rapport signal à bruit du mélange est égal à 30dB. On évalue ci-dessous les performances d'estimation de la méthode pour $N_c = 4$ et $N_c = 6$. Ces dernières sont à comparer à celles du cas $N_c = 2$ (Fig. 4.14). L'augmentation du nombre de classes N_c dans la méthode des Nuées Dynamiques permet d'améliorer les performances pour des valeurs de β inférieures à 5, mais cette augmentation a également tendance à faire croître $\hat{N}_s$. Pour $\beta = 8, 10$, le nombre moyen $\hat{N}_s$ varie peu lorsque N_c augmente, mais les pourcentages de fréquences fondamentales estimées diminuent fortement lorsque N_c passe de 2 à 6 classes. L'ensemble de ces résultats est illustré sur les trois figures (Fig. 4.14), (Fig. 4.17) et (Fig. 4.18).

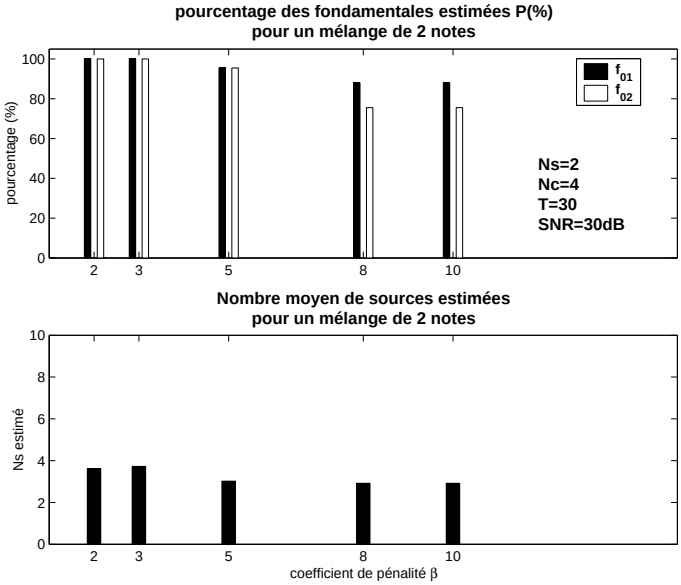


FIG. 4.17 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec $\mathbf{N_c = 4}$ et $T = 30$ pour un mélange de 2 notes où le $RSB = 30$ dB.*

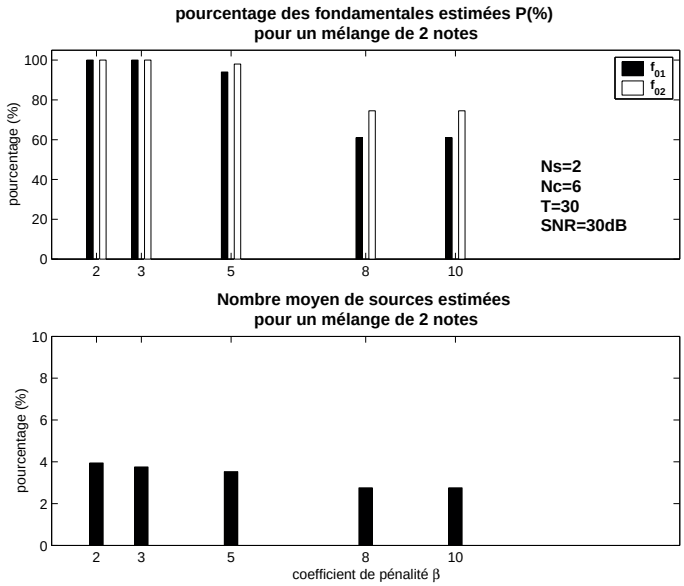


Fig. 4.18 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec* $\mathbf{N_c = 6}$ *et* $T = 30$ *pour un mélange de 2 notes où le* $RSB = 30$ *dB.*

• **Mélanges de plus de 2 sources**

On s'intéresse à présent à des mélanges pour lesquels $N_s > 2$. Deux types de mélange sont considérés : un mélange de 4 sources et un mélange de 6 sources. Les fréquences fondamentales $\{f_{0_i}\}_{i=1,\cdots,N_s}$ sont choisies, pour les deux types de mélanges, de telle sorte qu'il n'y ait aucun recouvrement spectral entre les notes. Le rapport signal à bruit de 30 dB et le nombre de partitions pour la construction des formes fortes est $T = 30$. Les performances de la méthode d'estimation de fréquences fondamentales multiples sont étudiées pour $N_c > N_s$ et pour $N_c < N_s$.

Cas où $N_c < N_s$ **:** Lorsque le nombre de classes des T partitions est inférieur au nombre de sources du mélange, la méthode d'estimation est capable d'estimer toutes les sources du mélange, comme le montrent les figures (Fig. 4.19) et (Fig. 4.20), pour lesquelles $N_c = 2$. En revanche le choix du coefficient de pénalité β est primordial. Le meilleur compromis entre pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et nombre moyen $\widehat{N}_s$ apparaît à nouveau pour les valeurs $\beta = 3, 5$.

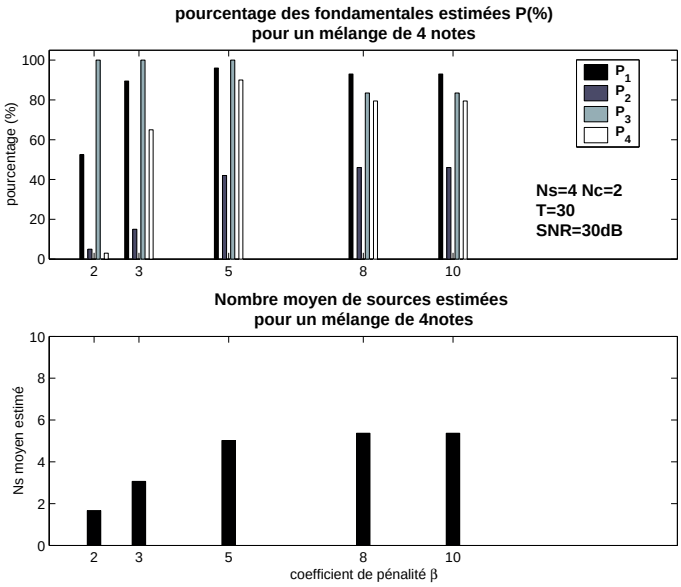


FIG. 4.19 – *Pourcentages de fondamentales estimées et nombre estimé moyende sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation de la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 4 notes où le* $RSB = 30$ *dB.*

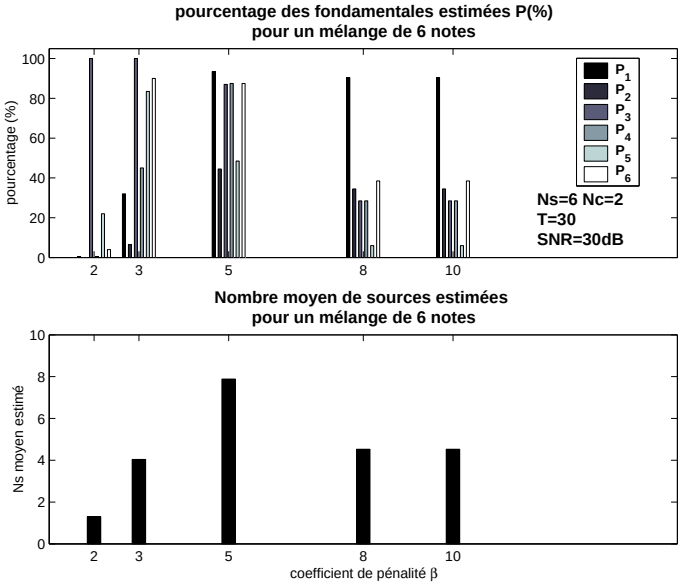


FIG. 4.20 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 6 notes où le* $RSB = 30$ *dB.*

Cas où $N_c > N_s$: Lorsque le nombre de classes N_c est cette fois-ci supérieur au nombre de sources N_s, les performances de la méthode d'estimation sont tout autre. La figure (Fig. 4.21) illustre le cas où $N_c = 10$. On remarque que pour obtenir des pourcentages d'estimation corrects, il est nécessaire de prendre un β relativement faible. En revanche, un tel choix implique un nombre de sources $\hat{N}s$ très largement surestimé.

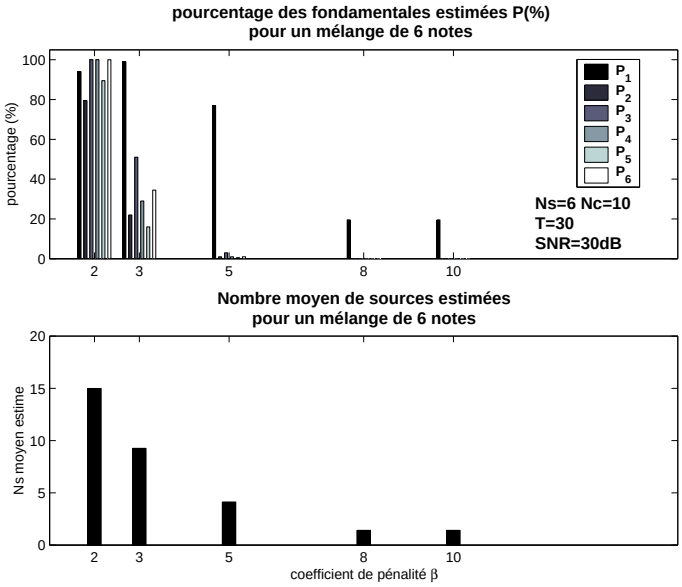


Fig. 4.21 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 10}$ *et* $T = 30$ *pour un mélange de 6 notes où* $RSB = 30$ *dB.*

• Répartition des FS et FL :

Afin d'évaluer correctement les performances de la méthode d'estimation par pénalisation sur les classifications ND, nous étudions également les différents types de formes fortes obtenues en fonction du seuil de pénalité choisi. On présente la répartition des formes fortes selon leur taux de recouvrement. On rappelle que le pourcentage de formes fortes à recouvrement nul (FF5) est équivalent au pourcentage de formes fortes correctes au sens strict ou FS, tandis que les pourcentages FF1,FF2, FF3 et FF4 correspondent au pourcentage de formes fortes correctes au sens large ou FL. Dans le cas d'un mélange de 2 et de 6 sources, l'ensemble de ces pourcentages a été calculé et représenté sur les figures (Fig. 4.22)

et (Fig. 4.23). Comme précédemment, les mélanges sont tels que le $RSB = 30$ dB et il n'y a pas de recouvrement spectral entre les notes. Le nombre de partitions utilisé pour la construction des formes fortes est $T = 30$.

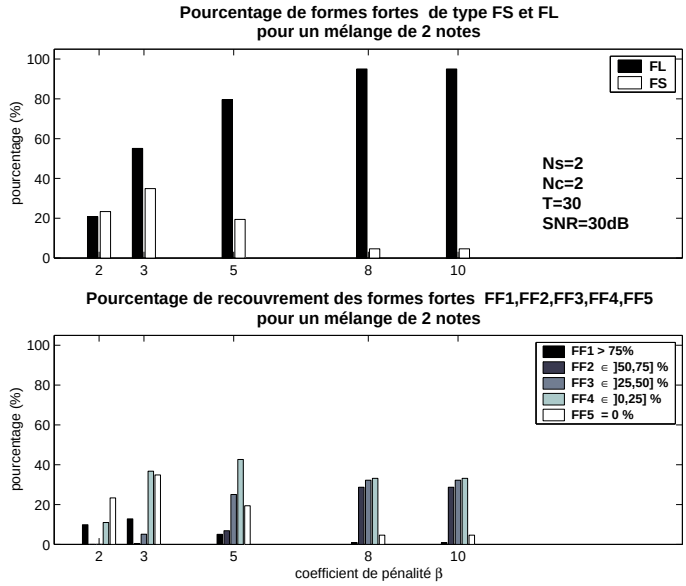


Fig. 4.22 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 2 notes où le* $RSB = 30$ *dB.*

Quel que soit le nombre de sources du mélange, plus le coefficient de pénalité β augmente, plus les formes fortes intervenant sont de type FL. Le plus grand nombre de formes fortes à recouvrement nul (FS) est obtenu pour $\beta = 3$, 5, valeurs pour lesquelles apparaissent les meilleurs compromis entre nombre moyen $\widehat{N}_s$ et pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$. Lorsque le mélange contient 2 sources, ces pourcentages restent relativement forts pour des $\beta \geq 8$ (cf. figure (Fig. 4.14)). Ils correspondent à une répartition dans laquelle les formes fortes à faible taux de recouvrement interviennent peu. Pour le mélange de 6 sources, les pourcentages d'estimation pour des valeurs $\beta \geq 8$ sont beaucoup moins forts que pour les autres seuil de pénalité (cf. figure (Fig. 4.20)). On remarque qu'ils correspondent à une répartition de formes fortes pour laquelle des taux de recouvrement supérieurs à 50% sont trés présents.

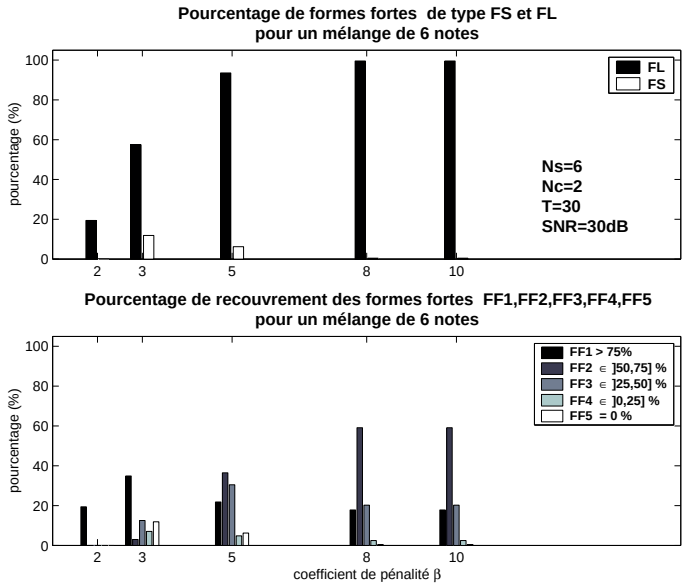


Fig. 4.23 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 6 notes où le* $RSB = 30$ *dB.*

• Cas de 2 notes à la quinte

Les simulations précédentes ont été réalisées à partir de notes synthétiques dont les harmoniques ne se recouvrent pas. Elles permettent d'illustrer les performances de la méthode d'estimation proposée pour de nombreux cas, mais elles ne prennent pas en compte l'une des caractéristiques majeures de la musique, celle de la consonance de notes jouées simultanément. Pour illustrer ce cas important, nous utilisons à présent un mélange de 2 notes synthétiques jouées à la quinte. Les recouvrements dans le spectre du mélange sont tels que les harmoniques d'ordre $\{2, 4, 6, \cdots\}$ de la première note, se confondent avec celles d'ordre $\{3, 6, 9, \cdots\}$ de la seconde note. La première note est alors recouverte à 50%, tandis que la second l'est à 33%. Un bruit blanc est ajouté au deux notes mélangées avec un rapport signal à bruit de 30 dB. L'estimation des fréquences fondamentales utilise 30 partitions à 2 classes. La figure (Fig. 4.24) présente les pourcentages d'estimation pour les deux fréquences fondamentales, ainsi que le nombre moyen de sources $\hat{N}_s$, en fonction du coefficient de pénalité β. Il apparaît dès lors que les performances d'estimation de la méthode sont très fortement dégradées par rapport au cas de 2 notes dissonantes. Ainsi, pour des valeurs de $\beta = 3, 5$, l'une des deux fréquences fondamentales est difficilement estimée et le nombre de sources est sous-estimé. On remarque qu'il faut alors augmenter β pour obtenir une estimation correcte de N_s. La figure (Fig. 4.25), qui présente les taux de recouvrement

pour les formes fortes obtenues, montre que pour $\beta \geq 8$, les taux compris dans l'intervalle $]25, 50]\%$ sont les plus importants.

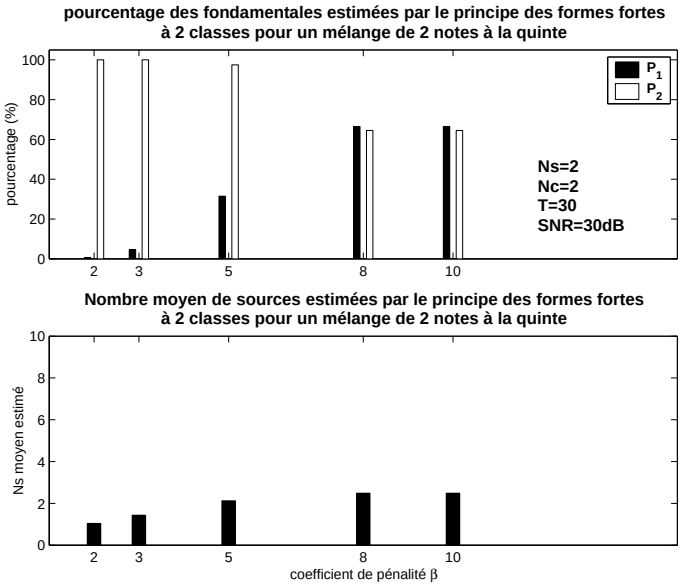


Fig. 4.24 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 2 notes jouées à la quinte où le* $RSB = 30$ *dB.*

4.4.3 Résumé des performances

L'étude de cette première méthode d'estimation de fréquences fondamentales multiples, basée sur la construction de formes fortes *par pénalisation sur la classe*, fournit un grand nombre de résultats que nous résumons ci-dessous :

– Les bonnes performances de la méthode dépendent principalement du seuil de pénalité Δ utilisé pour construire les formes fortes. Ce seuil prend implicitement en compte le nombre T de classifications utilisé dans la méthode des Nuées Dynamiques. Il n'est pas nécessaire d'utiliser un très grand nombre de classifications pour obtenir de bons pourcentages d'estimation des fréquences fondamentales et du nombre de sources. Par la suite, nous prendrons typiquement $T = 30$.

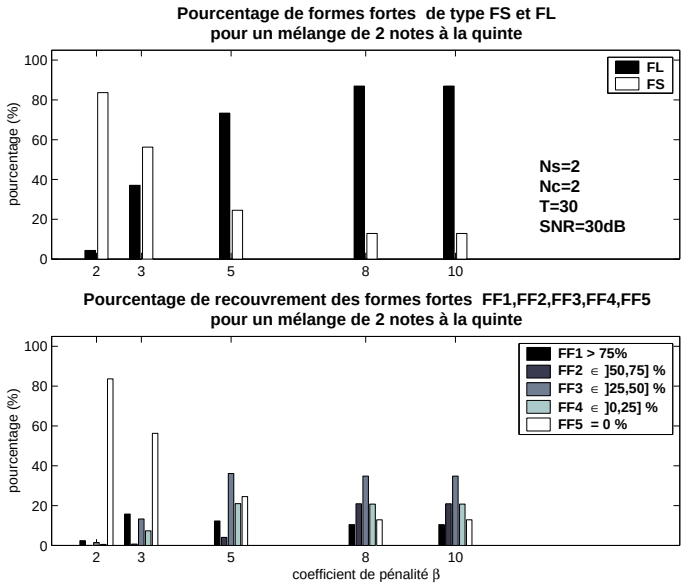


Fig. 4.25 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe avec* $\mathbf{N_c = 2}$ *et* $T = 30$ *pour un mélange de 2 notes jouées à la quinte où le* $RSB = 30$ *dB.*

- L'ensemble des simulations montre qu'il faut en règle générale faire un compromis entre forts pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et nombre moyen de sources $\hat{N}_s$ correct : plus β augmente, plus $\hat{N}_s$ se rapproche de sa valeur réelle, mais plus les pourcentages d'estimation diminuent.
  Lorsque $T = 30$, de bonnes performances sont finalement obtenues pour l'ensemble des simulations avec $\beta = 3,5$.

- On remarque que le nombre de classes N_c choisi dans la méthode des Nuées Dynamiques a peu d'influence sur les performances de la méthode. En d'autres termes, il n'est pas nécessaire de prendre un nombre de classes très supérieur à N_s pour obtenir de meilleurs résultats.

- La répartition des formes fortes selon leur taux de recouvrement montre que plus les formes sont de type FS, plus l'estimateur basé sur la compression spectrale est performant. Les FS sont généralement majoritaires pour $\beta = 3,5$.

- Les performances de la méthode diminuent avec le rapport signal à bruit du mélange mais également avec l'augmentation du nombre de sources. Toutefois, les pourcentages d'estimation et le nombre estimé moyen $\hat{N}_s$ restent très corrects, même dans le cas de 6 sources.

– La méthode proposée est capable de détecter des notes dont les harmoniques se recouvrent, mais les performances d'estimation sont nettement inférieures à celles obtenues pour le cas de notes dissonantes.

4.5 Estimation par pénalisation sur l'harmonicité

Les signaux musicaux possèdent un caractère quasi-harmonique que l'on souhaite exploiter au mieux pour résoudre le problème de l'estimation de fréquences fondamentales multiples. Cette caractéristique est prise en compte dans l'algorithme des Nuées Dynamiques, au moyen de la distance, qu'elle soit *moindres carrés* ou *sinusoïdale*. Toutefois, les faibles résultats de classification obtenus lorsque le principe des formes fortes n'est pas utilisé (cf. paragraphe (4.3.4)), laisse supposer que l'harmonicité n'est pas efficacement exploitée. Le principe des formes fortes cherche à pallier le problème en pénalisant les pics fréquentiels de classes différentes. Les résultats obtenus par application de ce principe nous incitent à penser que le traitement de l'information contenue dans les classes par pénalisation deux à deux des fréquences, est une méthode très prometteuse pour le problème d'estimation de fréquences fondamentales multiples. L'idée qui en découle, est qu'il est très certainement possible d'améliorer les performances d'estimation à partir des formes fortes, en introduisant une pénalisation supplémentaire, exploitant directement la relation harmonique entre les pics fréquentiels.

Dans cette partie, nous proposons une seconde méthode d'estimation de fréquences fondamentales multiples pour les mélanges musicaux. Cette méthode est également basée sur la classification des pics spectraux par le principe des formes fortes. Elle diffère de la méthode d'estimation *par pénalisation sur la classe*, par la nature de la pénalité appliquée aux différentes paires de fréquences. L'indice de pénalité utilisé dans le principe des formes fortes est calculé à partir du rapport harmonique de 2 fréquences.
Cette section est consacrée à cette nouvelle méthode d'estimation de fréquences fondamentales multiples dite *par pénalisation sur l'harmonicité*. L'harmonicité de deux fréquences est tout d'abord définie. Par la suite, nous détaillerons la construction des formes fortes pour une pénalisation sur l'harmonicité. Une étude similaire à celle proposée pour la méthode d'estimation *pénalisation sur la classe* est présentée. Le protocole d'évaluation et les types de mélanges utilisés sont ceux du paragraphe (4.4.2).

4.5.1 Définition de l'harmonicité de deux fréquences

Soit la fonction *modulo* définie pour tout couple $(f_i, f_j) \in E^2$ telle que :

$$modulo(f_i, f_j) = f_i - A[\frac{f_i}{f_j}] \times f_j \quad \forall \quad f_i > f_j \tag{4.26}$$

où $A[x]$ est l'arrondi à l'entier le plus proche de x.

Lorsque les deux fréquences f_i et f_j sont en relation harmonique telles que $f_i = k \times f_j$ avec k entier, on a :

$$A[\frac{f_i}{f_i}] = k \implies modulo(f_i, f_j) = 0 \tag{4.27}$$

Une première définition de l'harmonicité de deux fréquences est donnée par l'équation (eq. 4.27) : deux fréquences f_i et f_j sont supposées en relation harmonique si la valeur de leur *modulo* est égale à zéro.
Nous allons voir que cette définition est en pratique très restrictive. Elle sera remplacée par une seconde définition de l'harmonicité de deux fréquences, présentée ci-dessous.

En pratique, les fréquences que l'on désire classer sont des estimations dont la précision dépend de la méthode de détection utilisée pour obtenir ces dernières. Si les estimées des fréquences f_i et f_j, notées respectivement $\hat{f}_i$ et $\hat{f}_j$ sont précises à 3% près, elles appartiennent alors aux intervalles I_i et I_j :

$$\begin{cases} \hat{f}_i \in I_i = [f_i - 0.015.f_i, f_i + 0.015.f_i] \\ \\ \hat{f}_j \in I_j = [f_j - 0.015.f_j, f_j + 0.015.f_j] \end{cases}$$

Afin de mettre en évidence le comportement de la fonction *modulo*, en tenant compte de la précision, nous proposons un exemple numérique qui consiste à calculer les valeurs de la fonction *modulo* pour les fréquences $f_i = 3 \times 98$ et $f_j = 98$. Les fréquences estimées $\hat{f}_i$ et $\hat{f}_j$ sont supposées appartenir aux intervalles I_i et I_j :

$$I_i = [290, 291, 292, 293, 294, 295, 296, 297, 298] \quad et \quad I_j = [97, 98, 99]$$

Les valeurs de la fonction *modulo* pour chaque couple fréquentiel possible, sont présentées dans le tableau (Tab. 4.3) :

	290	291	292	293	294	295	296	297	298
97	1	0	1	2	3	4	5	6	7
98	4	3	2	1	0	1	2	3	4
99	7	6	5	4	3	2	1	0	1

Tab. 4.3 – *Valeurs de la fonction modulo pour les estimées des fréquences* $f_i = 3 \times 98$ *et* $f_j = 98$

Ainsi, si l'on réduit la définition d'harmonicité de 2 fréquences f_i et f_j à la relation (eq. 4.27), seuls trois couples de fréquences estimées sur les 27 couples possibles, vont correspondre à des fréquences supposées en relation harmonique. Les autres couples possèdent des rapports harmoniques non nuls, simplement parce que les fréquences ne sont pas parfaitement estimées.
En règle générale, on remarque que l'on a :

$$modulo(\hat{f}_i, \hat{f}_j) \leq \hat{f}_i.0.03 \quad \text{avec} \quad \hat{f}_i > \hat{f}_j \tag{4.28}$$

Cette dernière inéquation (eq. 4.28), beaucoup moins restrictive que la condition (eq. 4.27), servira par la suite de définition pour l'harmonicité : deux fréquences f_i et f_j seront supposées en relation harmonique si leur modulo est inférieur ou égal à la valeur $0.03.f_j$, pour $f_j > f_i$.

4.5.2 Construction des formes fortes

A l'instar de méthode d'estimation *par pénalisation sur la classe*, il est possible d'effectuer une classification très similaire, basée cette fois-ci sur l'harmonicité. Elle consiste à associer à chaque couple fréquentiel, un indice de pénalité qui sera fonction de l'harmonicité des 2 fréquences :

Pour la p^{ieme} fenêtre d'analyse d'un mélange, dont le nombre N_s de notes est inconnu, la classification se fait à partir de l'ensemble des pics détectés $E_p = \{\nu_1, \nu_2, \cdots, \nu_{Mp}\}$. Pour chaque couple (ν_i, ν_j) de E_p, on associe un indice de

pénalité $\delta(\nu_i, \nu_j)$ tel que :

$$\delta(\nu_i, \nu_j) = \begin{cases} 1 & \text{si} \quad \text{modulo}(\nu_i, \nu_j) > 0.03.\nu_i, \quad (\nu_i > \nu_j) \\ \\ 0 & \text{sinon} \end{cases}$$

On détermine dans un premier temps, les indices de pénalité $\delta(\nu_i, \nu_j)$ pour $i\,et\,j \in \{1, 2, \cdots, M_p\}$, puis à partir d'une valeur seuil de pénalité Δ choisie, on construit les formes fortes :

- Construction des formes fortes.
- Suppression des classes parasites.
- Suppression du recouvrement.

Pour le détail de ces trois étapes, on pourra se référer au paragraphe (4.4.1).
Afin de pénaliser les couples fréquentiels pour lesquels les deux fréquences ne sont pas en relation harmonique, on pose $\Delta = 1$. On ne garde ainsi que les couples dont l'indice $\delta(\nu_i, \nu_j)$ est nul, c'est-à-dire les fréquences en relation harmonique selon la définition (eq. 4.28).

Nous avons donc à disposition une seconde méthode de classification, développée à nouveau à partir du principe des formes fortes, c'est-à-dire par le biais d'un indice de pénalité δ. Toutefois cette seconde méthode est beaucoup plus simple en terme de complexité de calcul, puisqu'elle ne nécessite pas de T partitions de l'ensemble E_p obtenues par la méthode des Nuées Dynamiques, contrairement à la méthode d'estimation *par pénalisation sur la classe*.

4.5.3 Etude de la pénalisation sur l'harmonicité

Comme pour la classification obtenue *par pénalisation sur la classe*, nous étudions à présent les performances de la méthode d'estimation de fréquences fondamentales multiples, *par pénalisation sur l'harmonicité*. Le protocole d'évaluation est le même que celui du paragraphe (4.4.2). Ainsi pour des mélanges contenant de 1 à 6 notes, les performances de la classification sont présentées en évaluant les pourcentages de fréquences fondamentales correctement estimées $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et le nombre estimé moyen de sources $\hat{N}_s$.

• **Influence du RSB**

On considère un mélange de 2 notes, dont les harmoniques ne se recouvrent pas. La figure (Fig. 4.26) présente dans ce cas, les pourcentages d'estimation et le nombre moyen $\hat{N}_s$ associé, lorsque le rapport signal à bruit du mélange prend les valeurs -15, 0, 10, 30 dB :

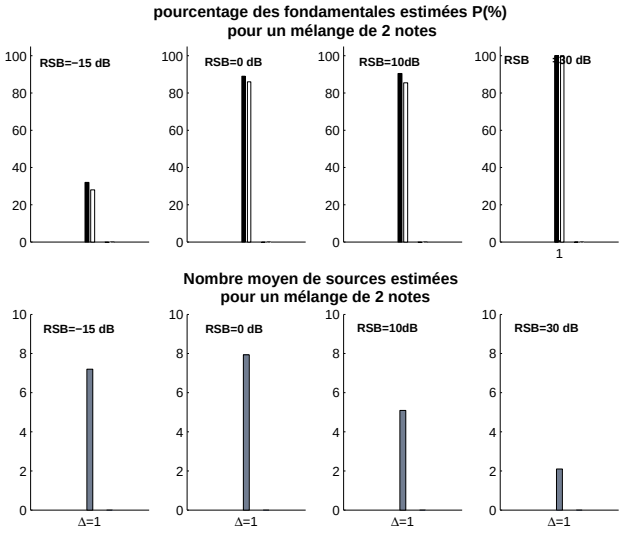


FIG. 4.26 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources* $\widehat{N}_s$ *en fonction du RSB. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur l'harmonicité pour un mélange de 2 notes. Le seuil de pénalité* $\Delta = 1$.

La comparaison des performances de cette méthode à celles obtenues par pénalisation sur la classe (Fig. 4.14) (4.15) (4.16), montre que les deux méthodes sont équivalentes pour un coefficient de pénalité $\beta = 5$, lorsque $RSB = 30$ dB. Pour des rapports signal à bruit plus faibles, il faut augmenter la valeur de β pour se rapprocher de l'équivalence. Toutefois, la méthode d'estimation *par pénalisation sur l'harmonicité* apparaît finalement plus performante, car elle permet de faire un bon compromis entre des pourcentages d'estimation corrects $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et une valeur de $\hat{N}_s$ qui surestime pas trop le nombre de sources.

On s'intéresse à présent à la répartition des formes fortes selon leur taux de recouvrement (figure (Fig. 4.27)). On remarque que le nombre de formes correctes au sens strict (FS), domine largement celui des formes correctes au sens large (FL). On voit à nouveau qu'une telle répartition permet d'obtenir de bons pourcentages pour l'estimation des fréquences fondamentales. Pour la méthode

d'estimation *par pénalisation sur la classe*, les pourcentages de FS les plus importants, sont obtenus lorsque $\beta = 3,5$ (Fig. 4.22). Ces coefficients de pénalité sont ceux pour lesquelles les performances atteintes sont généralement les meilleures.

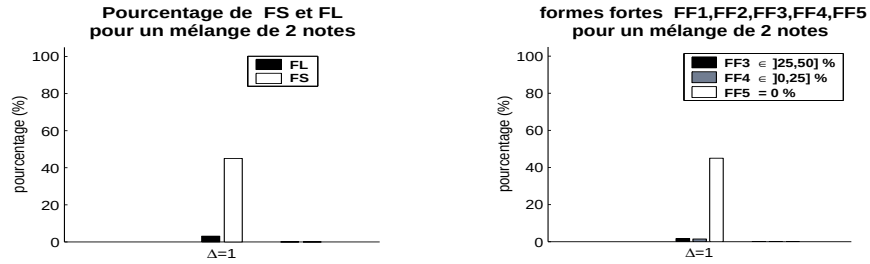


Fig. 4.27 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur l'harmonicité pour un mélange de 2 notes où le* $RSB = 30$ *dB.*

• Cas de plus de 2 sources

La figure (Fig. 4.28) présente les résultats obtenus pour des mélanges de 4 et 6 sources. Les fréquences fondamentales des notes sont choisies de telle sorte qu'aucun recouvrement spectral n'entre en jeu. Pour un $RSB = 30$ dB, les performances *par pénalisation sur l'harmonicité* sont très bonnes : la méthode d'estimation fournit des pourcentages d'estimation aux alentours de 100%, pour toutes les fréquences fondamentales du mélange. De plus, le nombre estimé moyen $\hat{N}_s$ reste très proche de sa véritable valeur.

• Cas de 2 notes à la quinte

Lorsqu'il existe un recouvrement entre les harmoniques de 2 notes mélangées, la méthode d'estimation *par pénalisation sur l'harmonicité* fournit également de très bons résultats. La figure (Fig. 4.29) présente les pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ des 2 fréquences fondamentales pour des notes jouées à la quinte. Ces derniers sont maximum pour un $RSB = 30$ dB et le nombre estimé moyen $\widehat{N}_s$ associé est proche de 2.
La répartition des FS et FL présentée sur la figure (Fig. 4.30), met en évidence la prédominance de classes avec un taux de recouvrement compris dans l'intervalle

$]25, 50]\%$, résultat déjà remarqué dans la première méhode de classification (Fig. 4.25).

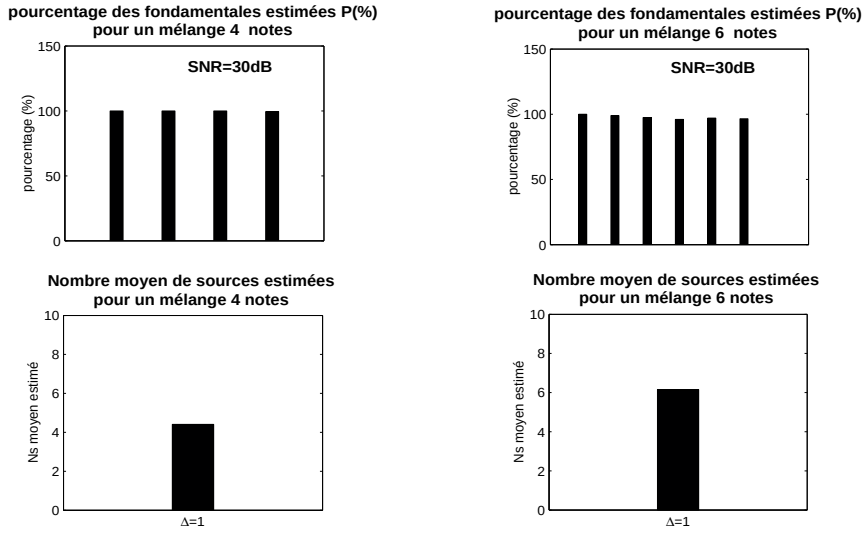


FIG. 4.28 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}s$. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur l'harmonicité pour un mélange de 4 notes et un mélange de 6 notes où le $RSB = 30$ dB. Le seuil de pénalisation $\Delta = 1$.*

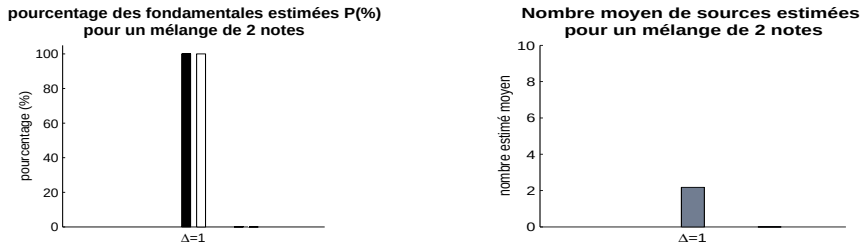


FIG. 4.29 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}s$. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur l'harmonicité pour un mélange de 2 notes jouées à la quinte où le $RSB = 30$ dB.*

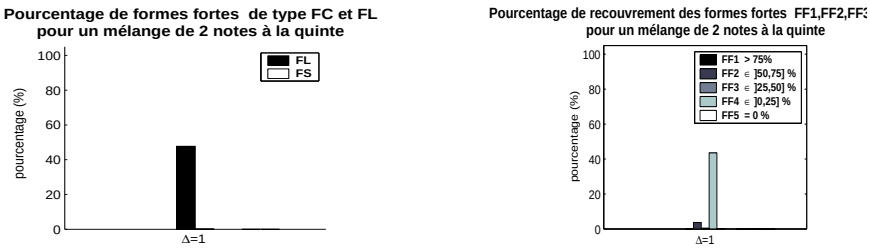


Fig. 4.30 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur l'harmonicité pour un mélange de 2 notes jouées à la quinte où le* $RSB = 30$ *dB.*

4.5.4 Résumé des performances

Comme pour la première méthode d'estimation de fréquences fondamentales multiples proposée, nous résumons ici les résultats obtenus pour une classification basée sur le calcul du rapport harmonique :

- Contrairement à la méthode précédente, aucune classification par la méthode des Nuées Dynamiques n'est effectuée au préalable. Ainsi, il n'existe aucun compromis à faire entre pourcentages d'estimation maximum $P(\%)$ et nombre estimé moyen de sources $\widehat{N}s$. La comparaison des deux méthodes montrent en général que les performances de l'estimation *par pénalisation sur l'harmonicité* sont équivalentes, voire supérieures aux performances optimales de de l'estimation *par pénalisation sur la classe*, c'est-à-dire celles obtenues pour $\beta = 3,5$.

- Les performances de la méthode augmentent avec le rapport signal à bruit du mélange. Elle apparaissent même supérieures à celles obtenues *par pénalisation sur la classe*, en terme de nombre estimé de sources.

- Cette nouvelle méthode d'estimation est capable de traiter le cas de notes dont les harmoniques se recouvrent et offrent même des très bons résultats, contrairement à la méthode précédente.

4.6 Estimation par pénalisation sur la classe et l'harmonicité

Les deux méthodes d'estimation précédentes ont la particularité commune d'estimer les fréquences fondamentales d'un mélange à partir de formes fortes. Leur différence réside dans la partition finale de l'ensemble des pics spectraux E : pour la première méthode les formes fortes sont construites par pénalisation sur T classifications obtenues par la méthode des Nuées Dynamiques, tandis que pour la seconde méthode les formes fortes sont construites par pénalisation sur l'harmonicité de 2 fréquences. Ces deux types de pénalisation n'étant pas incompatibles, nous proposons et développons une troisième méthode d'estimation de fréquences fondamentales multiples, dite *par pénalisation sur la classe et l'harmonicité*.
La construction des formes fortes pour cette troisième méthode est tout d'abord présentée, puis une évaluation de ses performances est proposée. Afin de comparer les performances des 3 méthodes d'estimation, le protocole d'évaluation de cette partie, est le même que celui utilisé pour les deux précédentes méthodes. Ce dernier est présenté dans le paragraphe (4.4.2).

4.6.1 Construction des formes fortes

Pour la p^{ieme} fenêtre d'analyse d'un mélange, dont le nombre N_s de notes est inconnu, la construction des formes fortes pour la séparation des notes se fait à partir de T partitions de l'ensemble des pics détectés $E_p = \{\nu_1, \nu_2, \cdots, \nu_{M_p}\}$.
Pour chaque couple de fréquences (ν_i, ν_j) de E, on associe un indice de pénalité $\delta(\nu_i, \nu_j)$ qui prend en compte à la fois les T partitions et l'harmonicité. Cet indice est calculé à partir du paramètre δ^t pour chaque partitions t obtenues par l'algorithme des Nuées Dynamiques et défini comme suit :

1. si ν_i et ν_j sont dans des classes différentes, $\delta^t(\nu_i, \nu_j) = 1$

2. si ν_i et ν_j sont dans la même classe, on vérifie leur rapport d'harmonicité. En d'autres termes :
   si pour $\nu_i > \nu_j$, $\quad \text{modulo}(\nu_i, \nu_j) > 0.03.\nu_i \quad$ alors $\quad \delta^t(\nu_i, \nu_j) = 1$

L'indice de pénalité finalement associé au couple (ν_i, ν_j) est obtenu en sommant les indices de chaque partition :

$$\delta(\nu_i, \nu_j) = \sum_{t=1}^{T} \delta^t(\nu_i, \nu_j)$$

On détermine dans un premier temps, les indices de pénalité $\delta(\nu_i, \nu_j)$ pour $i\,et\,j \in \{1, 2, \cdots, M_p\}$, puis à partir d'une valeur seuil de pénalité Δ choisie, on construit les formes fortes :

- Construction des formes fortes.
- Suppression des classes parasites.
- Suppression du recouvrement.

Pour le détail de ces trois étapes, on pourra se référer au paragraphe (4.4.1).

4.6.2 Etude de la pénalisation sur la classe et l'harmonicité

Pour cette dernière partie, on utilise à nouveau les mélanges synthétiques proposés lors de l'étude des 2 précédentes méthodes. Les performances sont illustrées à l'aide des pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ pour chaque fréquence fondamentale du mélange et du nombre estimé moyen de sources $\hat{N}_s$.

• Influence du RSB

Les figures (Fig. 4.31), (Fig. 4.32), (Fig. 4.33) et (Fig. 4.34) présentent, dans le cas d'un mélange de 2 notes, les pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et $\hat{N}_s$ pour des rapports signal à bruit de -15, 0, 10 et 30 dB. Aucun recouvrement spectral n'est supposé et pour ces quatre valeurs de RSB, le nombre de partitions est choisi égal à 30. Chacune de ces partitions possède $N_c = 2$ classes.

La prise en compte conjointe des partitions obtenues par les Nuées Dynamiques et de l'harmonicité pour la pénalisation, permet d'améliorer les performances d'estimation par rapport à une simple pénalisation sur la classe. En effet pour des $\beta = 3,5$, le nombre estimé moyen de sources est considérablement réduit pour les plus faibles RSB (Fig. 4.15). Ces valeurs de $\hat{N}_s$ restent tout de même deux fois plus grandes que celles de la méthode d'estimation *par pénalisation sur*

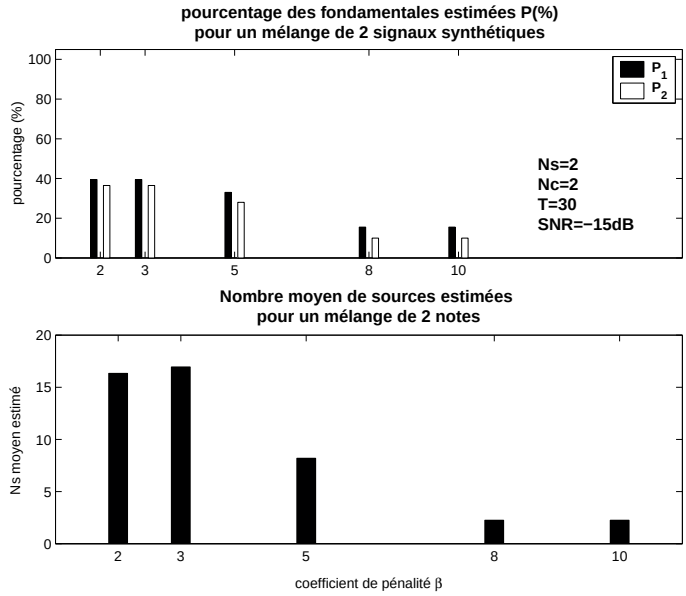


FIG. 4.31 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 2 notes où le $RSB = -15$ dB.*

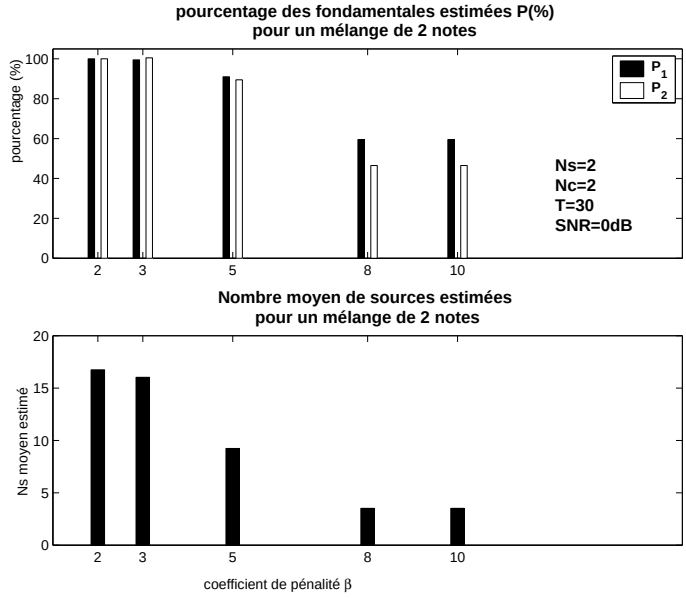


FIG. 4.32 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 2 notes où le $RSB = 0$ dB.*

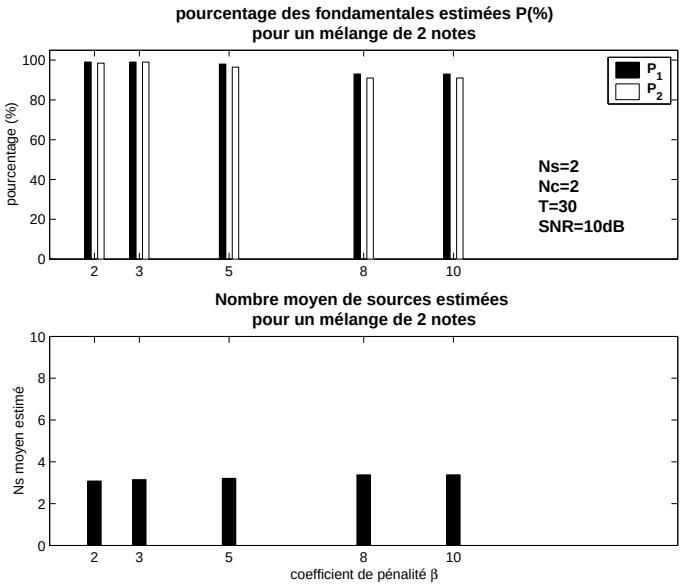


Fig. 4.33 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 2 notes où le $RSB = 10$ dB.*

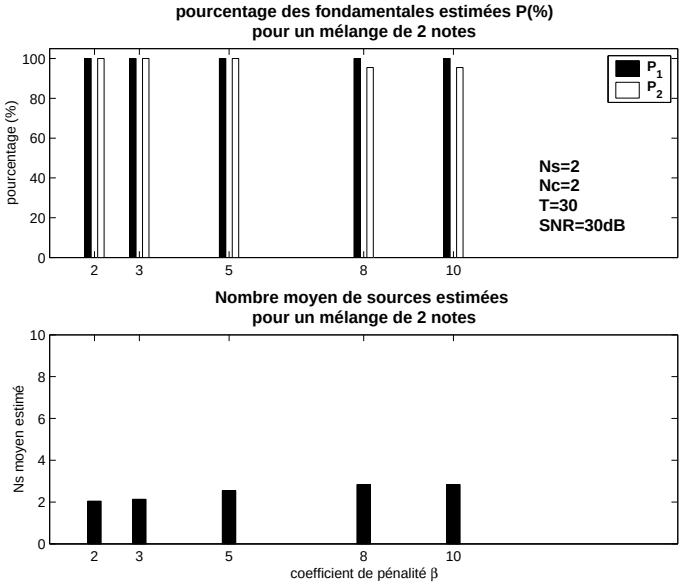


Fig. 4.34 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 2 notes où le $RSB = 30$ dB.*

l'harmonicité (Fig. 4.26). Pour les RSB plus forts, la double pénalisation améliore également les pourcentages d'estimation de la première méthode, tout en conservant une estimation du nombre de sources du mélange correcte (Fig. 4.16) (Fig. 4.14).

• Mélanges de plus de 2 sources

Pour des mélanges de 4 et 6 notes sans recouvrement spectral, nous évaluons les performances d'estimation de la méthode *par pénalisation sur la classe et l'harmonicité*, pour les cas où $N_c \leq N_s$ et $N_c \geq N_s$. Le rapport signal à bruit est de 30 dB et 30 partitions obtenues par les Nuées Dynamiques sont utilisées.

Cas où $N_c \leq N_s$: Lorsque $N_c = 2$, la pénalisation supplémentaire sur l'harmonicité de deux fréquence, permet d'améliorer de façon considérable les pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ obtenues pour la méthode *par pénalisation sur la classe* (Fig. 4.19) (Fig. 4.20). Les performances de cette troisième méthode d'estimation sont illustrées par les figures (Fig. 4.35) et (Fig. 4.36) pour des mélanges de 4 et 6 notes.

Cas où $N_c \geq N_s$: Lorsque $N_c = 10$, les performances d'estimation de la méthode sont présentées sur la figure (Fig. 4.37). Elles sont équivalentes à celles obtenues par une simple pénalisation sur la classe (Fig. 4.21). Dans ce cas, la prise en compte d'une pénalisation supplémentaire sur l'harmonicité de deux fréquences, n'améliore pas les estimations. Il est alors nécessaire de diminuer le coefficient de pénalité β pour obtenir de plus forts pourcentages $\{P_i(\%)\}_{i=1,\cdots,N_s}$. Cette amélioration se fait au dépend de $\widehat{N}_s$ va surestime alors très largement le nombre de sources.

• Cas de 2 notes à la quinte

On s'intéresse à présent au cas d'un mélange pour lequel 2 notes se recouvrent spectralement. Le rapport signal à bruit est de 30 dB et la méthode d'estimation utilise $T = 30$ partitions obtenues par les Nuées Dynamiques. Lorsque les deux

notes possèdent des harmoniques communes, les résultats obtenus par cette dernière méthode d'estimation sont très supérieurs à la méthode *par pénalisation sur la classe*. On obtient en effet des pourcentages $\{P_i(\%)\}_{i=1,\cdots,N_s}$ nettement plus élevés pour les valeurs de $\beta \leq 5$. Pour $\beta = 2$, les performances sont équivalentes à celles obtenues par la méthode d'estimation *par pénalisation sur l'harmonicité*. La comparaison des FS et FL montrent que la répartition de ces dernières pour une double pénalisation (Fig. 4.39), reste plus proche de celle obtenue *par pénalisation sur l'harmonicité* : le pourcentage total de formes fortes correctes est beaucoup plus faible que pour une simple pénalisation sur la classe et les FL sont beaucoup plus nombreuses que les FS (Fig. 4.25) (Fig. 4.30). On remarque de plus que les meilleures performances, qui sont obtenues dans le cas présent pour $\beta = 2$, correspondent à nouveau au cas où le taux de recouvrement $]25, 50]\%$ est le plus grand.

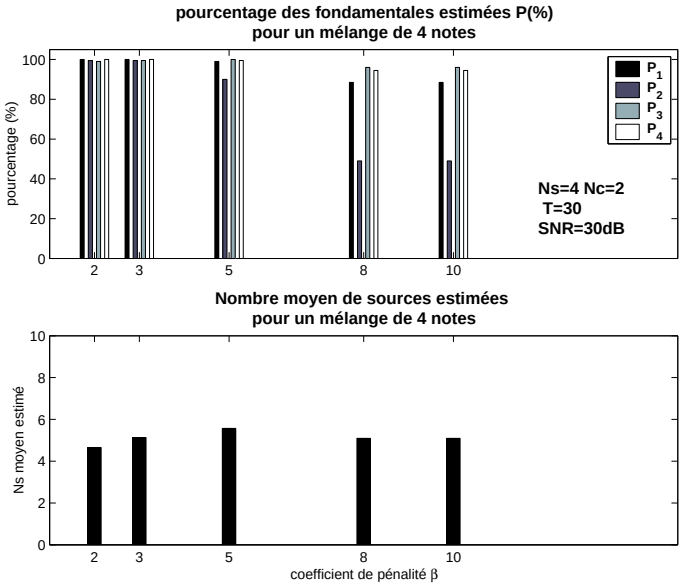


Fig. 4.35 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 4 notes où le $RSB = 30$ dB.*

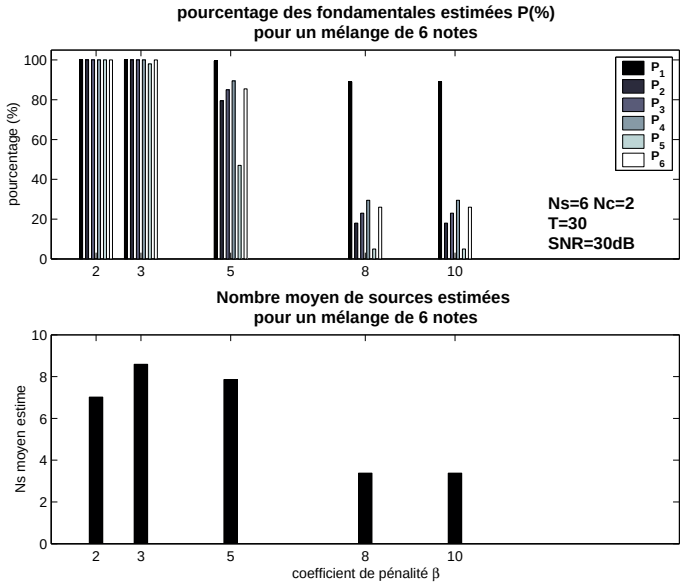


FIG. 4.36 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 2$ et $T = 30$, pour un mélange de 6 notes où le $RSB = 30$ dB.*

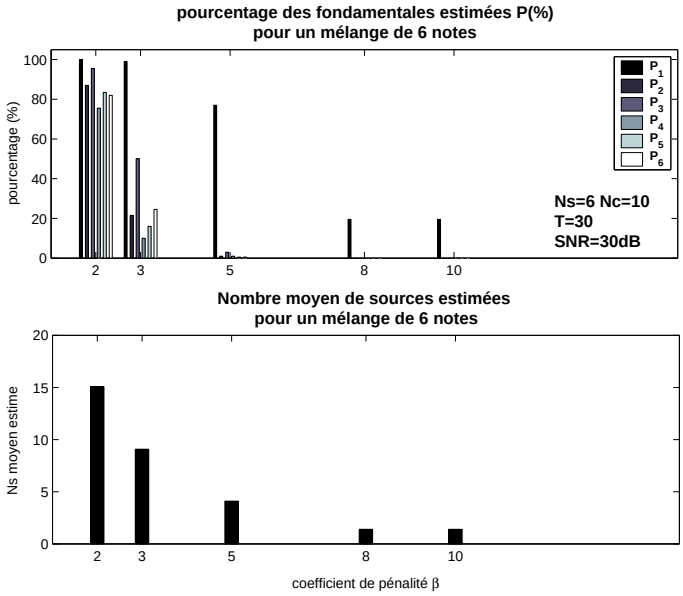


FIG. 4.37 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du coefficient de pénalité β. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec $N_c = 10$ et $T = 30$, pour un mélange de 6 notes où le $RSB = 30$ dB.*

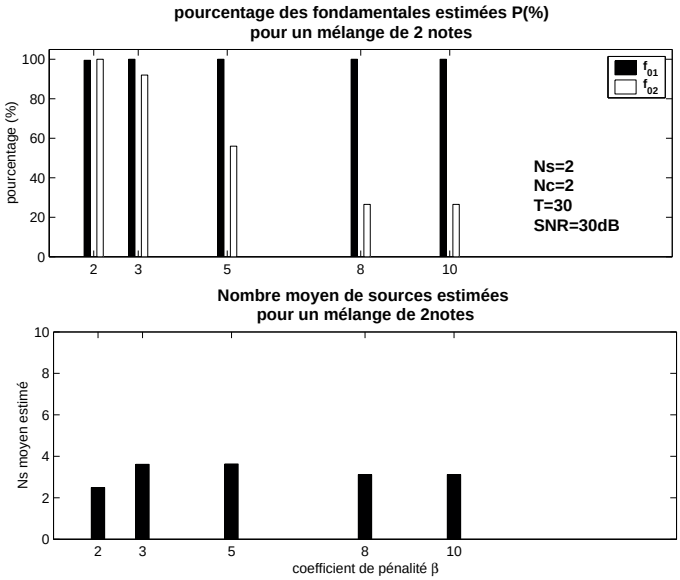


Fig. 4.38 – *Pourcentages de fondamentales estimées et Nombre estimé de sources* $\widehat{N}_s$ *en fonction du coefficient de pénalité* β*. Cas de l'estimation de fréquences fondamentales multiples par pénalisation sur la classe et l'harmonicité avec* $N_c = 2$ *et* $T = 30$*, pour un mélange de 2 notes jouées à la quinte où le* $RSB = 30dB$*.*

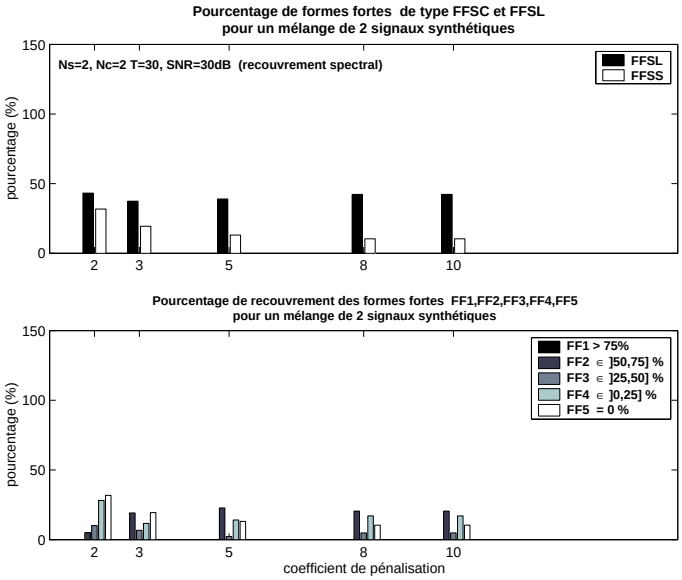


Fig. 4.39 – *Pourcentage de FS et FL et répartition des formes fortes selon leur taux de recouvrement en fonction du coefficient de pénalité* β*. Cas de l'estimation par pénalisation sur la classe et l'harmonicité, pour un mélange de 2 notes jouées à la quinte où le* $RSB = 30dB$*.*

4.6.3 Résumé des performances

Les différents résultats pour cette dernière méthode d'estimation sont résumés ci-dessous :

- La méthode d'estimation dépend du seuil de pénalité Δ, qui permet de construire les formes fortes. De bonnes performances sont synonymes de forts pourcentages d'estimation $\{P_i(\%)\}_{i=1,\cdots,N_s}$ et d'un $\hat{N}_s$ qui ne surestima pas trop fortement le nombre de sources N_s. L'ensemble des simulations montrent que le choix $\beta = 3$ satisfait ces exigences, pour la plupart des mélanges synthétiques étudiés.

- La prise en compte d'une pénalisation supplémentaire sur l'harmonicité de deux fréquence, permet d'améliorer les performances d'estimation obtenues *par pénalisation sur la classe*. Cela est particulièrement vrai lorsque le nombre de sources du mélange augmente ou lorsqu'il existe des recouvrements harmoniques entre les notes du mélange.

4.7 Applications aux signaux musicaux

Après avoir étudié les trois méthodes d'estimation de fréquences fondamentales multiples sur des mélanges synthétiques, nous nous intéressons à présent au cas de véritables mélanges de signaux musicaux.
Nous appliquons dans un premier temps, les trois méthodes d'estimation par classification des pics spectraux, à des mélanges particuliers : un mélange pour lequel la fréquence fondamentale de l'une des notes est absente, un mélange de 3 voix chantées déjà utilisé dans la partie parole, ainsi qu'un dernier mélange contenant une unique source polyphonique, correspondant à l'enregistrement d'une suite d'accords de piano. Les performances d'estimation des 3 méthodes sont comparées à celles obtenues par application de la méthode d'estimation basée sur le critère d'ordre et de la méthode de l'autocorrélation, toutes deux initialement développées dans le cadre de la séparation de mélanges de parole (cf. Partie 2). Enfin, afin de juger et de comparer les 3 méthodes d'estimation sur une grande diversité d'instruments, nous testerons ces dernières sur une base de données de sons musicaux.
Pour l'ensemble des méthodes d'estimation de fréquences fondamentales multiples, les mélanges sont analysés sur des fenêtres d'analyse d'une durée de 100 ms et le pas de recouvrement entre fenêtres successives est de 50 ms.

4.7.1 Etude de quelques mélanges particuliers

• Cas du fondamental absent

On s'intéresse au cas d'un mélange dans lequel la fréquence fondamentale d'une note est absente.
Nous avons vu auparavant que deux raisons peuvent être à l'origine de cette absence. La première est inhérente à la nature de l'instrument : pour des instruments tels que le basson ou encore le hautbois, la fréquence fondamentale peut avoir une amplitude tellement faible par rapport aux amplitudes harmoniques, qu'elle n'est dans ce cas que très rarement détectable. La seconde raison est directement liée au problème de résolution spectrale, dont nous avons discuté dans le paragraphe (4.2.4).
L'exemple qui suit illustre le comportement des méthodes d'estimation basée sur la classification des pics spectraux, lorsque la fréquence fondamentale de l'une des notes est absente : le mélange étudié est constitué de trois notes jouées simultanément par un hautbois, un basson français et une contrebasse, leur note respective étant **A4** ($f_0 = 440Hz$), **C#5** ($f_0 = 554Hz$) et **F4** ($f_0 = 349Hz$). Pour ce mélange présentée sur la figure (Fig. 4.40-a), la fondamentale absente est celle de la note **A4**. Les fréquences fondamentales sont estimées à l'aide de la méthode d'estimation *par pénalisation sur l'harmonicité*.

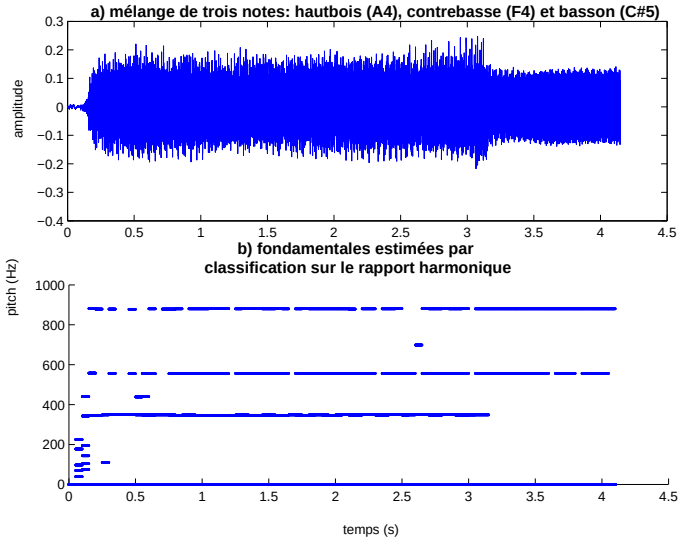


Fig. 4.40 – *a) Mélange où le fondamental de l'une des notes est absent - b) Pitch estimés par la méthode d'estimation basée sur une pénalisation sur l'harmonicité. Cas d'un mélange de 3 notes* **A4**, **C#5** *et* **F4**, *respectivement jouées par un hautbois, un basson français et une contrebasse.*

La méthode permet de détecter trois fréquences fondamentales différentes, illustrés sur la figure (Fig. 4.40-b) par la présence de trois trajectoires temporelles aux fréquences 349 Hz, 555 Hz et 881 Hz. Les deux premières fréquences correspondent aux notes **F4** et **C#5** jouées par la contrebasse et le basson. En revanche, la dernière fréquence correspond à l'octave supérieure de la note jouée par le hautbois.
Les figures (Fig. 4.41) et (Fig. 4.42) illustrent la classification des pics spectraux sur une fenêtre analysée du mélange pour ce type d'erreur d'estimation : la figure (Fig. 4.41) représente le spectre du mélange et ses pics détectés par la méthode de détection proposée dans le paragraphe (4.2.2). Elle met en évidence la non-détection de la fréquence fondamentale relative à la note **A4** : l'amplitude du pic à 440 Hz est en effet 2 fois plus faible que celle du plus petit pic détecté. La figure (Fig. 4.42) représente les pics spectraux classés, qui vont définir les "spectres séparés" à partir desquels les fréquences fondamentales du mélange sont estimées. Cette illustration est à l'origine de plusieurs remarques sur les méthodes d'estimation par classification des pics :

- Lorsque les fréquences fondamentales sont détectées, ces dernières appartiennent à des classes différentes, permettant une estimation de fréquence fondamentale correcte pour les notes correspondantes. C'est le cas dans cet exemple, des notes **F4** et **C#5** pour lesquelles les fréquences fondamentales sont détectées et séparées. La note **A4**, quant à elle, ne peut être estimée correctement puisqu'elle ne possède pas de fondamental.

- Un pic harmonique peut appartenir à deux classes différentes. C'est le cas du pic de fréquence 2784Hz qui est commun aux trois classes, comme le montre la figure (Fig. 4.42).

- La présence du pic de fréquence 2784 Hz se justifie pleinement pour les notes **F4** et **C#5**, ce dernier étant en relation harmonique avec les deux fréquences fondamentales. En revanche, ce n'est pas le cas pour la note jouée par le hautbois. Le spectre séparé de cette note contient en fait tout un ensemble de pics correspondant principalement à la note **F4**.
  Les méthodes de classification de pics peuvent donc entraîner simultanément la détermination de classes "très" correctes, contenant seulement les éléments d'une même note, mais également de classes "très" corrompues par la présence de pics d'une seconde note. Toutefois, on notera que dans l'exemple présenté ci-dessus, bien que l'absence d'une fréquence fondamentale rende la classification relativement difficile, il est possible de déterminer à une octave près la note jouée par le hautbois, à partir d'une classe de pics très corrompue.

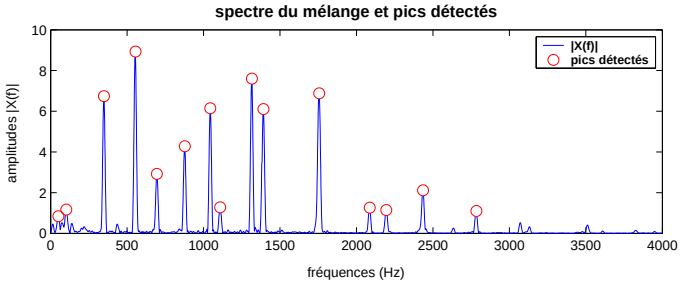


FIG. 4.41 – *Pics spectraux détectés pour une fenêtre de 100 ms de signal. Cas d'un mélange de 3 notes* **A4**, **C#5** *et* **F4**, *respectivement jouées par un hautbois, un basson français et une contrebasse.*

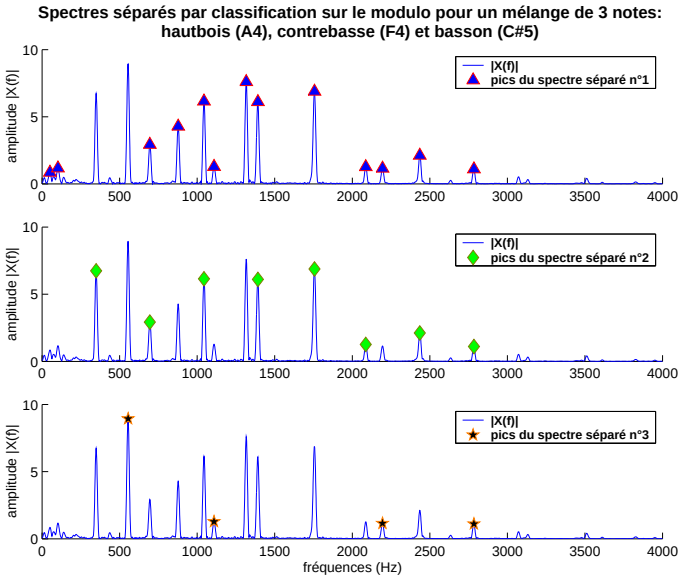


FIG. 4.42 – *"spectres séparés" pour une fenêtre de 100 ms de signal, obtenus par pénalisation sur l'harmonicité. Cas d'un mélange de 3 notes* **A4**, **C#5** *et* **F4**, *respectivement jouées par un hautbois, un basson français et une contrebasse.*

L'application de la méthode d'estimation basée sur le principe du critère d'ordre et proposée pour la séparation de mélange de parole (cf. Partie 1), permet d'obtenir les estimations fréquentielles présentées sur la figure (Fig. 4.43- a)). Pour le mélange considéré, la fréquence fondamentale absente est estimée la plupart du temps. En revanche, il apparaît de nombreuses trajectoires relatives à des sous-harmoniques des notes **C#5** et **F4**. Les trois fréquences fondamentales sont finalement correctement détectées lorsque la note **F4** s'éteint progressivement.
La figure (Fig. 4.43 -b)) présente les résultats d'estimation de la méthode l'autocorrélation. Cette méthode également développée dans le cadre de la parole (cf. paragraphe (3.4.2)), consiste à estimer itérativement les fréquences fondamentales du mélange, à partir du calcul de l'autocorrélation de ce signal. Elle utilise une étape de soustraction qui permet d'éliminer au fur et à mesure les sources estimées. Pour le mélange considéré dans cette partie, la méthode de l'autocorrélation n'est pas très performante : l'absence du fondamental à 440 Hz entraîne des valeurs de fréquences fondamentales estimées qui sont la plupart du temps erronées dès la première itération. L'étape de soustraction, qui est généralement le point faible des méthodes d'estimation itératives, ne permet pas d'obtenir des estimations correctes aux itérations suivantes.

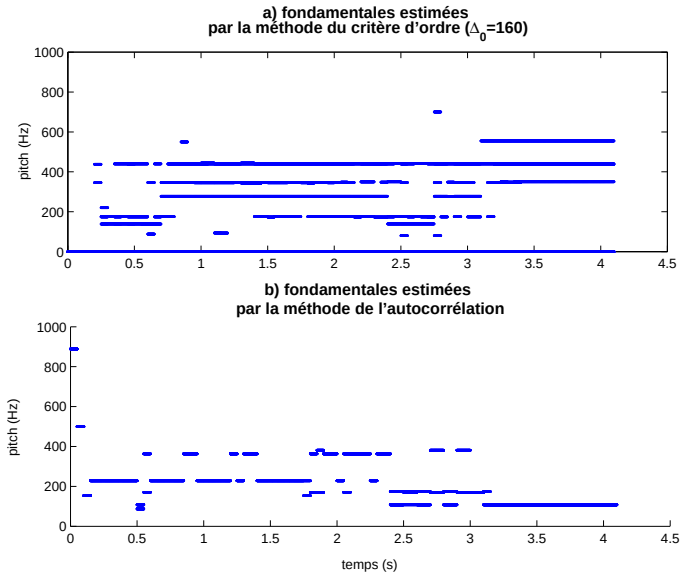


Fig. 4.43 – *fréquences fondamentales estimées par : a) la méthode d'estimation basée le critère d'ordre - b) la méthode de l'autocorrélation. Cas d'un mélange de 3 notes* **A4**,**C#5** *et* **F4**, *respectivement jouées par un hautbois, un basson français et une contrebasse.*

• **Cas des voix chantées**

Dans cette partie, nous comparons les méthodes d'estimation par classification des pics spectraux à la méthode d'estimation basée sur le critère d'ordre. Les méthodes sont appliquées au mélange des trois voix chantées déjà étudié dans le cadre de la séparation de la parole. Les trois sources du mélange sont présentées dans le paragraphe (3.4.3). Afin de comparer les performances des méthodes d'estimation par classification des pics spectraux, à celles obtenues précédemment par la méthode d'estimation du critère d'ordre, la durée d'analyse d'une fenêtre de mélange est réduite à 50 ms. Les pourcentages de fréquences fondamentales correctes $P(\%)$, ainsi que le nombre estimé moyen de sources $\widehat{N}_s$ sont présentés dans les tableaux (Tab. 4.4) et (Tab. 4.5). Le pourcentage de fréquences fondamentales correctement estimées $P(\%)$ est calculé en déterminant le nombre de fréquences estimées par les différentes méthodes, qui sont égales à l'une des fréquences fondamentales des sources. $\hat{N}_s$ est le nombre de sources moyen estimé sur chaque fenêtre de signal analysé. Pour la méthode basée sur le principe du critère d'ordre, les performances obtenues pour deux valeurs seuil de l'écart de vraisemblance $\Delta_0 = 60$ et $\Delta_0 = 20$ sont présentées. On rappelle que ce dernier permet d'estimer le nombre de sources du mélange : le nombre de sources sur une fenêtre analysée est obtenu en itérant les estimations jusqu'à ce que l'écart de vraisemblance associé passe en dessous du seuil Δ_0. Pour $\Delta_0 = 0$, toutes les fréquences estimées sont supposées correspondre à un son voisé.

	pénalisation sur la classe	pénalisation sur l'harmonicité	pénalisation sur la classe et l'harmonicité
$P(\%)$	71.0	75.6	73.3
$\widehat{N}_s$	2.38	3.04	2.08

Tab. 4.4 – *Performances des méthodes d'estimation basées sur la classification des pics spectraux*

	méthode du critère d'ordre (Δ_0=60)	méthode du critère d'ordre (Δ_0=20)
$P(\%)$	88.1	83.3
$\widehat{N}_s$	1.57	2.92

Tab. 4.5 – *Performances de la méthode d'estimation basée sur le principe du critère d'ordre pour 2 valeurs d'écart de vraisemblance* $\Delta_0 = 0$ *et* $\Delta_0 = 20$.

La comparaison des performances des quatre méthodes d'estimation de fréquences fondamentales multiples, met en évidence de meilleurs pourcentages d'estimation pour la méthode basée sur le principe du critère d'ordre. C'est un résultat attendu, car cette dernière est une méthode temporelle paramétrique qui n'est pas, contrairement aux autres méthodes, limitée par la résolution de Fourier. Cette caractéristique lui permet d'être mieux appropriée à l'analyse de signaux de parole que les méthodes basées sur le périodogramme.
En revanche, parce que le nombre estimé de sources dans le cas de la méthode basée sur le principe du critère d'ordre dépend fortement du choix de Δ_0, les méthodes de classification peuvent être vues comme de meilleurs estimateurs pour N_s. On rappellera tout de même que le seuil de détection Δ_0 prend toute son importance pour des mélanges mixtes qui contiennent des signaux voisés et non-voisés.

4.7.2 Recouvrement d'harmoniques : cas des accords

Un autre type de mélange sur lequel nous proposons de tester les méthodes d'estimation par classification des pics est celui de l'accord. Le problème de l'estimation de fréquences fondamentales multiples pour ce dernier est relativement difficile, car les notes de la plupart des accords ont en général des harmoniques communes et sont joués simultanément.
L'évaluation des performances d'estimation sur des mélanges synthétiques a montré que les trois méthodes d'estimation sont capables de traiter le cas de 2 notes synthétiques à la quinte. A présent, nous nous intéressons au cas plus complexe d'une suite d'accords de piano. La figure (Fig. 4.44) présente le mélange traité. Il contient les 6 accords : $C - Cm - CM7 - C7 - Cm7 - Cm7$. Les figures (Fig. 4.45), (Fig. 4.46) et (Fig. 4.47) présentent les fréquences fondamentales estimées respectivement par les trois méthodes basées sur la classification des pics spectraux.

Quelle que soit la méthode de classification, les résultats obtenus pour les accords de piano sont de très bonne qualité :

- Les 3 méthodes d'estimation permettent d'analyser des accords et plus généralement les mélanges dans lesquelles les notes jouées ont des harmoniques communes (quinte, tierce, $\cdots$).

- Les 3 méthodes d'estimation permettent d'estimer correctement le nombre

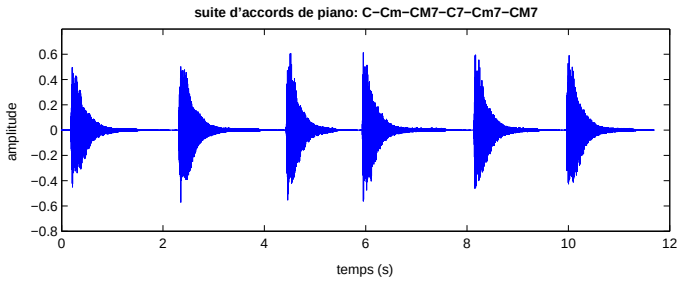


Fig. 4.44 – *mélange de notes jouées par un instrument polyphonique* Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.

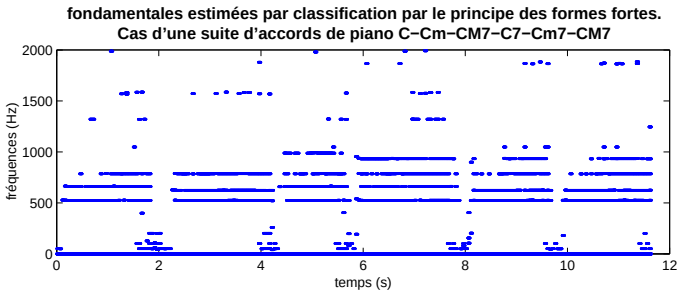


Fig. 4.45 – *fréquences fondamentales estimées pour la méthode d'estimation par pénalisation sur la classe. ($N_c = 2$, $T = 30$, $\beta = 3$). Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.*

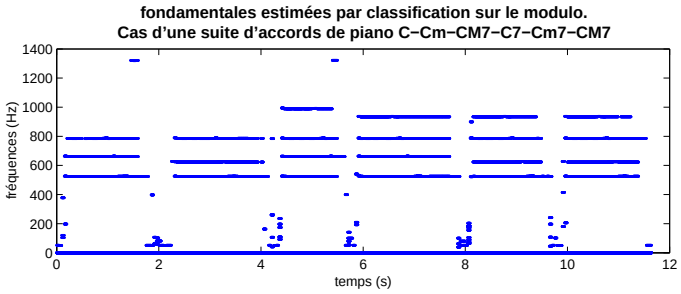


Fig. 4.46 – *fréquences fondamentales estimées pour la méthode d'estimation par pénalisation sur l'harmonicité. Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.*

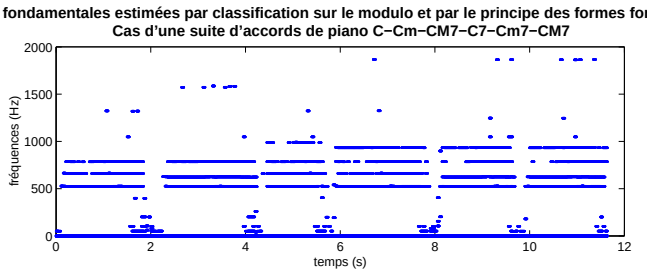


FIG. 4.47 – *fréquences fondamentales estimées pour la méthode d'estimation par pénalisation sur la classe et l'harmonicité. ($N_c = 2$, $T = 30$, $\beta = 3$). Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.*

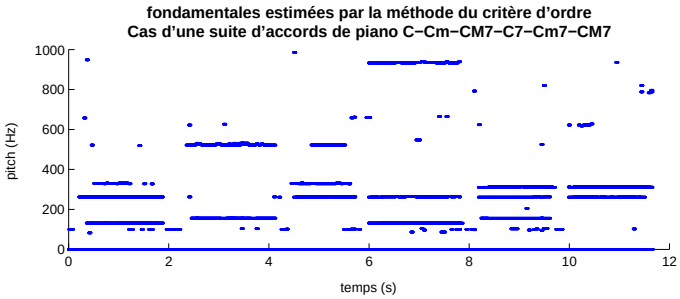


FIG. 4.48 – *fréquences fondamentales estimées pour la méthode d'estimation basée sur le critère d'ordre. Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.*

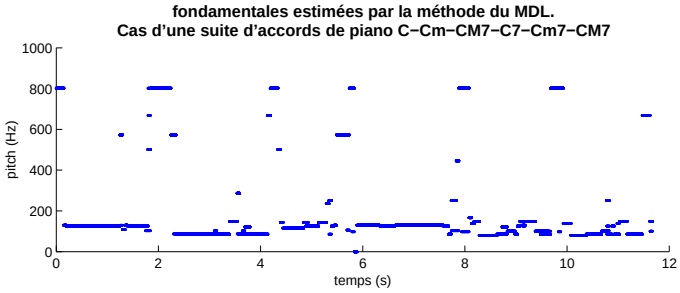


FIG. 4.49 – *fréquences fondamentales estimées pour la méthode d'estimation basée sur l'autocorrélation. Cas d'une suite d'accords de piano : $C - Cm - CM7 - C7 - Cm7 - Cm7$.*

de notes de chacun des accords. La seconde méthode apparaît à nouveau comme la plus performante.

– La stabilité des classifications permet dans les trois cas d'obtenir des trajectoires temporelles pour les fréquences fondamentales très régulières, ce qui est un avantage certain pour des applications de poursuite de fréquence fondamentale le long des trames d'analyse.

– La nature très inharmonique des sons de piano met en évidence la grande capacité d'adaptation de ces méthodes d'estimation de fréquences fondamentales multiples, qui ont été initialement développées dans le but d'exploiter l' harmonicité que l'on retrouve chez de nombreux signaux musicaux.

Les résultats d'estimation de la méthode basée sur le principe du critère d'ordre et la méthode de l'autocorrélation sur les accords de piano sont présentés sur les figures (Fig. 4.48) et (Fig. 4.49). Ils illustrent à nouveau les limites de ces deux méthodes dans le cadre de la séparation de mélanges musicaux. Ces dernières, contrairement aux méthodes basées sur la classification des pics spectraux, sont développées sans prendre en compte d'éventuels recouvrement spectral. Ainsi, la méthode d'estimation du critère d'ordre, pour ce type de mélange, permet la détection de plusieurs notes d'un même accord, mais les erreurs d'octave sont très fréquentes. La méthode de l'autocorrélation, quant à elle, est très peu performante pour ce type de mélange : à la première itération, la fréquence fondamentale estimée est une valeur erronée qui se rapproche généralement d'une sous-harmonique commune à deux notes de l'accord. Après soustraction de sa contribution dans le mélange, le résiduel obtenu est trop fortement dégradé pour obtenir des estimations correctes aux itérations suivantes.

4.7.3 Estimation sur une base de données de signaux musicaux

Les bonnes performances obtenues sur des mélanges de voix chantées et d'accords de piano nous incitent à penser que les méthodes d'estimation basées sur la classification des pics spectraux ont de grandes capacités d'adaptation face à la nature des instruments traités. Afin de s'en convaincre, une toute dernière application est proposée dans ce paragraphe. Elle consiste à tester les trois méthodes d'estimation de fréquences fondamentales multiples, sur des mélanges variés, contenant plusieurs sortes d'instruments. Les notes de chaque instrument sont tirées aléatoirement dans une base de données. La base de données utilisée

est celle proposée par l'IRCAM [58]. Elle contient entre autres, des notes de piano, d'instruments à vent tels que le saxophone, le basson, le hautbois ou encore la flûte, ainsi que d'instruments à cordes comme la guitare, le violoncelle, le violon ou la contrebasse.
Les performances d'estimation sont évaluées en déterminant les pourcentages de fréquences fondamentales $P(\%)$ ainsi que le nombre de sources estimés $\hat{N}_s$ sur des mélanges réalistes, qui peuvent contenir des notes jouées à l'octave. Sachant que les méthodes d'estimation par classification des pics spectraux, ne peuvent séparer deux notes à l'octave, nous évaluerons également les conséquences de ce type de mélanges sur le nombre moyen $\hat{N}_s$.

• Principe de l'étude

Des mélanges contenant de 1 à 6 notes sont engendrés aléatoirement à partir de la base de données, les niveaux sonores des notes étant ajustés pour être équivalents. Ils sont analysés sur des fenêtres de longueur égale à 100 ms. Afin de juger de la robustesse des méthodes face au bruit, le cas de mélanges "noyés" dans un bruit blanc avec un $RSB = -15$ dB sera également étudié. 6 sortes de mélange, caractérisés par leur nombre de source $N_s \in \{1, \cdots, 6\}$, sont étudiés. Pour chaque sorte de mélange, les pourcentages d'estimation et le nombre moyen $\hat{N}_s$ sont calculés à partir de 500 réalisations, c'est-à-dire 500 mélanges obtenus par tirage aléatoire des notes dans la base de données. Une fréquence fondamentale est supposée correcte lorsqu'elle est égale, à 3% près , à l'une des fréquences fondamentales du mélange. $P(\%)$ est alors le pourcentage global des fréquences fondamentales estimées. Le nombre estimé de sources $\hat{N}_s$ est défini comme le nombre moyen de fréquences fondamentales estimées sur un jeu de 500 réalisations.

• Evaluation des performances

Les deux méthodes d'estimation *par pénalisation sur la classe* et *par pénalisation sur la classe et l'harmonicité* qui utilisent T partitions de N_c classes, obtenues par l'algorithme des Nuées Dynamiques, sont appliquées pour $T = 30$ et $N_c = 2$. Le coefficient de pénalité β est choisi égal à 3.
La figure (Fig. 4.50) illustre les résultats obtenus pour l'ensemble des méthodes. $P(\%)$ et $\widehat{N}_s$ mettent en évidence de bonnes performances d'estimation pour les 3 méthodes. Lorsque $N_s = 6$, ces pourcentages sont supérieurs à 50% et la méthode

d'estimation *par pénalisation de la classe* apparaît comme la plus performante. Elle reste également la plus performante pour l'estimation du nombre de sources, les deux autres méthodes ayant tendance à sous-estimer N_s.
Lorsqu'un bruit blanc est ajouté aux mélanges musicaux, les performances des trois méthodes d'estimation se dégradent fortement. Ces dernières sont illustrées sur la figure (Fig. 4.51) : les pourcentages d'estimation $P(\%)$ pour $N_s = 6$, sont de l'ordre de 40% pour la première méthode qui reste la plus performante et de 20% pour la troisième méthode. Le nombre moyen de sources, dans les trois cas, est très largement surestimé, quel que soit le type de mélange considéré et ce plus particulièrement pour la méthode *par pénalisation de la classe*.

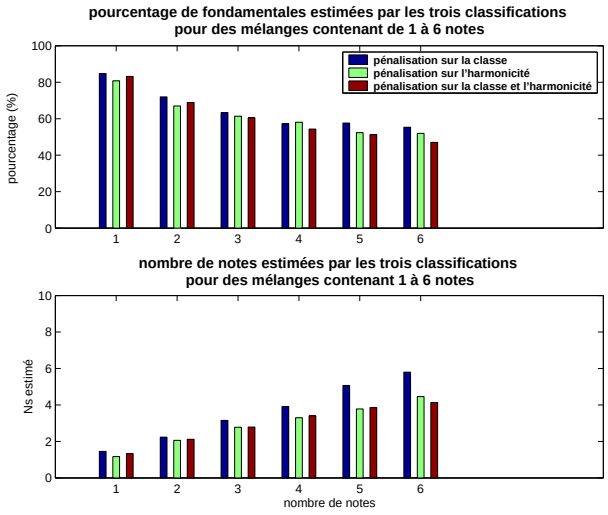


Fig. 4.50 – *Pourcentages de fondamentales estimées et nombre estimé moyen de sources $\widehat{N}_s$ en fonction du nombre de notes du mélange, pour les 3 méthodes d'estimation basées sur la classifications des pics spectraux.*

• Cas de notes à l'octave

Comme la plupart des méthodes d'estimation de fréquences fondamentales multiples, les méthodes basées sur la classification des pics spectraux, ne permettent pas de détecter deux notes jouées simultanément à l'octave. Cette caractéristique qui reste un inconvénient majeur dans le cadre de l'analyse de mélanges musicaux est prise en compte dans l'évaluation des performances d'estimation présentée ci-dessus. On s'intéresse à présent, à la conséquence directe de ce type de mélange, sur l'estimation du nombre de sources par les trois méthodes proposées.

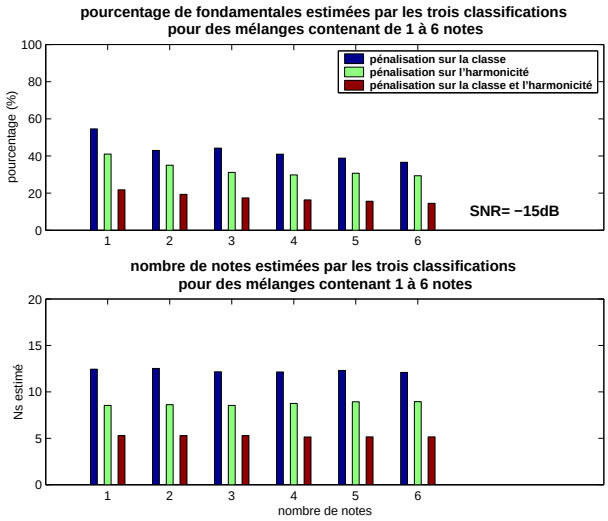


Fig. 4.51 – *Pourcentages de fondamentales estimées et nombre moyen de sources estimé $\widehat{N}_s$ en fonction du nombre de notes du mélange, pour les 3 méthodes d'estimation basées sur la classification des pics spectraux. Cas de mélanges de notes dans un bruit blanc avec un $RSB = -15$ dB.*

La présence de notes à l'octave dans un mélange donné de N_s sources, entraîne une sous-évaluation des performances d'estimation du nombre de sources pour les 3 méthodes : par exemple, pour un mélange de 6 notes contenant 2 fois 2 notes à l'octave, les notes détectables par les estimateurs sont au nombre de 4. Ce nombre est par la suite présenté en tant que " nombre de sources détectables". Afin de rendre compte le plus justement des performances des méthodes d'estimation pour le nombre de sources d'un mélange, nous déterminons le nombre moyen de sources réellement détectables pour chaque simulation faite sur la base de donnée. Ce dernier est alors comparé au nombre moyen de sources estimées $\hat{N}_s$ par les méthodes. Les figures (Fig. 4.52) et (Fig. 4.53) illustrent cette comparaison pour les différents types de mélange. Lorsque les mélanges ne contiennent pas de bruit blanc, la méthode d'estimation *par pénalisation de la classe*, qui donne le plus grand nombre de fréquences fondamentales correctement estimées, a tendance à surestimer le nombre de sources réellement détectables. Cela signifie que cette méthode a tendance à estimer des fréquences fondamentales erronées. Les deux autres méthodes, qui utilisent le calcul de l'harmonicité de deux fréquences, apparaissent dès lors beaucoup plus performantes pour le problème d'estimation du nombre de sources. Ce phénomène est particulièrement bien illustré dans le cas des accords de piano présenté dans le paragraphe (4.7.2). Lorsqu'un bruit blanc est ajouté au mélange, avec un $RSB = -15$ dB, la comparaison de $\hat{N}_s$ avec le nombre de sources détectables, montre que la méthode *par pénalisation sur la classe et sur l'harmonicité* reste dans ce cas la plus performante.

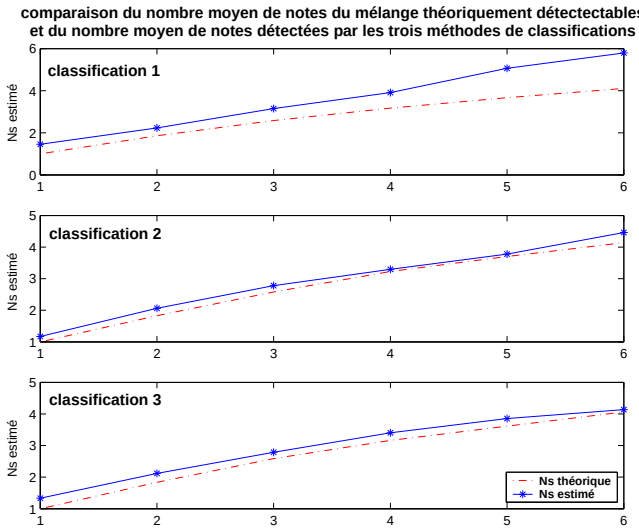


Fig. 4.52 – *Comparaison entre le nombre de notes détectables du mélange et le nombre estimé de sources par les 3 méthodes d'estimation basées sur la classification des pics spectraux.*

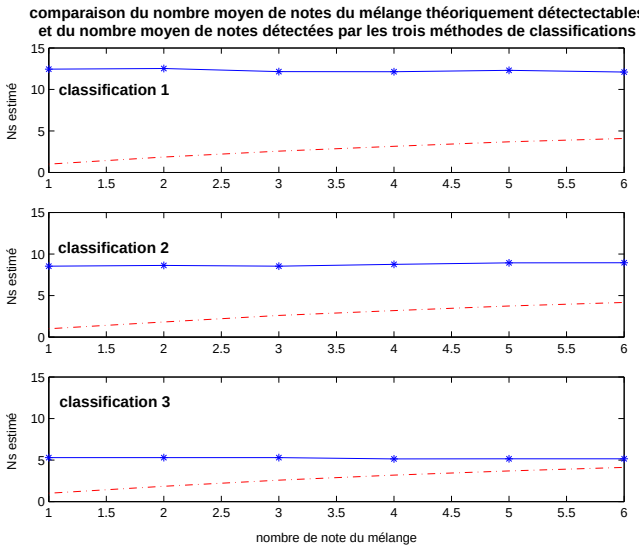


Fig. 4.53 – *Comparaison entre le nombre de notes détectables du mélange et le nombre estimé de sources par les 3 méthodes d'estimation basées sur la classification des pics spectraux. Cas de mélanges de notes dans un bruit blanc pour un* $RSB = -15$ *dB.*

4.8 Conclusions

Dans le cadre de la séparation des signaux musicaux, trois nouvelles méthodes d'estimation de fréquences fondamentales multiples sont proposées. Ces dernières utilisent toutes la même approche qui consiste à classifier les pics spectraux du mélange étudié, puis à estimer les fréquences fondamentales de chaque source à partir des classes déterminées.
Dans les trois cas, la classification fait appel au principe des formes fortes, qui pénalise les fréquences à classer en leur associant un indice de pénalité, puis regroupe ces dernières dans des classes appelées "formes fortes" à l'aide d'un seuil de pénalité. La partition du spectre ainsi obtenue permet de définir des "spectres séparés" sur lesquels est appliquée une méthode de compression spectrale. Elle permet de déterminer pour chaque spectre séparé, une fréquence fondamentale, finalement choisie comme étant la fréquence fondamentale de l'une des sources du mélange.

Les trois méthodes d'estimation diffèrent par la pénalisation utilisée pour la construction des formes fortes :

- La méthode d'estimation *par pénalisation sur la classe*, fait appel à l'algorithme des Nuées Dynamiques. Ce dernier est appliqué T fois à l'ensemble des pics à classer, donnant ainsi T partitions à partir desquelles les fréquences des pics vont être pénalisées.

- La méthode d'estimation *par pénalisation sur l'harmonicité*, fait appel à une classification plus simple, basée sur le calcul de l'harmonicité de deux fréquences, qui permet de définir si deux fréquences sont en relation harmonique.

- La méthode *par pénalisation sur la classe et sur l'harmonicité*, exploite un indice de pénalité tenant compte des deux types de pénalisation précédentes.

"Les Nuées Dynamiques" est une méthode de classification basée sur la minimisation d'une distance entre les éléments à classer. Dans le cas de pics spectraux, deux distances sont proposées. Bien que différentes, elles sont toutes deux basées sur le caractère harmonique des signaux musicaux. L'inconvénient majeur de ce type de classification est celui de la convergence : la partition finale obtenue dépend de l'étape d'initialisation dans laquelle N_c fréquences sont tirées au hasard. Le nombre N_c est arbitraire et correspond au nombre de classes de cette partition. Ainsi, après deux applications successives de l'algorithme, on obtiendra très

certainement deux partitions différentes. Il en résulte qu'une simple classification des pics spectraux par cette méthode ne permet pas d'obtenir de bonnes performances d'estimation. En revanche, une étude détaillée, basée sur des calculs de pourcentages de classifications correctes, montre que l'ensemble de ces classifications contient une quantité d'information, qu'il est possible d'exploiter à l'aide du principe des formes fortes.
Le principe des formes fortes, et plus généralement le principe de pénalisation des fréquences à classer, permet d'estimer le nombre de sources du mélange. Il permet également de prendre en compte les recouvrements d'harmoniques possibles entre les notes du mélange.

Les 3 méthodes d'estimation sont développées, étudiées et comparées, par le biais de nombreuses simulations sur des mélanges synthétiques de notes. Ces dernières permettent de justifier les différentes valeurs choisies pour le seuil de pénalité ou encore le nombre de classes utilisé dans l'algorithme des Nuées Dynamiques, pour construire les formes fortes.
Ces méthodes sont ensuite appliquées à de véritables mélanges musicaux. Il ressort de cette étude plusieurs résultats, résumés ci-dessous :

- Les méthodes d'estimation basées sur la classification des pics ne permettent pas d'estimer les notes dont le fondamental est absent. Ces dernières sont en général détectées, mais les fréquences fondamentales qui leurs sont associées correspondent à des octaves supérieures.

- L'application à un mélange de 3 voix chantées met en évidence les bonnes performances de ces méthodes pour ces signaux particuliers. La comparaison avec celles de la méthode d'estimation du critère d'ordre, développée pour la séparation de signaux de parole, montre que les pourcentages d'estimation pour ces trois méthodes sont plus faibles, mais qu'elles permettent d'estimer correctement le nombre de sources. L'estimation *par pénalisation sur l'harmonicité* s'avère être la meilleure.

- Les trois méthodes permettent d'estimer les fréquences fondamentales de mélange ne contenant qu'un unique instrument polyphonique. L'application à une suite d'accords de piano, permet d'obtenir les trajectoires temporelles de toutes les notes jouées. Elle illustre également les capacités de ces méthodes, d'une part pour le traitement de notes consonantes (quinte, tierce, $\cdots$) et d'autre part pour le traitement d'instruments très inharmoniques comme le piano.

Pour finir, une dernière application de ces trois estimateurs de fréquences fondamentales multiples est proposée. Elle vise à étudier le comportement de ces

derniers, face à la diversité des instruments musicaux que l'on peut rencontrer dans un mélange. Ainsi, à partir d'une base de données contenant des notes jouées aussi bien par du piano, de la flûte que du violoncelle, des jeux de mélanges contenant jusqu'à 6 notes sont engendrés aléatoirement. Ces derniers sont analysés à travers deux critères illustrant les performances de ces méthodes : le pourcentage de fréquences fondamentales correctement estimées $P(\%)$ et le nombre estimé moyen de sources $\hat{N}_s$. Les résultats montrent finalement que les méthodes basées sur une classification des pics ont de bonnes capacités d'adaptation à la nature des instruments du mélange. Les pourcentages d'estimation de fréquences fondamentales diminuent avec le nombre de sources et avec le bruit. La méthode *par pénalisation sur la classe*, donne de meilleurs pourcentages d'estimation, mais a tendance à surestimer le nombre de sources. La prise en compte supplémentaire d'une pénalisation sur l'harmonicité, permet de diminuer $\hat{N}_s$ et plus particulièrement dans le cas de mélange très bruités, mais il a également tendance à diminuer les pourcentages d'estimation $P(\%)$. Les performances de la méthode d'estimation *par pénalisation sur l'harmonicité* apparaît comme un bon compromis des deux autres méthodes : elle permet une estimation correcte du nombre de sources, sans pour autant avoir la faible qualité d'estimation de la dernière méthode.

Conclusion générale

Dans ce travail, nous proposons de nouvelles méthodes d'estimation de fréquences fondamentales multiples pour des mélanges provenant d'enregistrements monophoniques.
Une première partie est consacrée aux mélanges de signaux de parole. La méthode d'estimation de fréquences fondamentales multiples proposée pour ce type de mélange, est basée sur une approche itérative dans laquelle les fréquences fondamentales sont extraites du mélange les unes après les autres. Elle fait appel à un estimateur développé à partir d'un critère de sélection d'ordre, pour un modèle de mélange du type "sinusoïdes + bruit autoregressif". Cette modélisation, connue sous le nom de modélisation CARD ("Canonical AutoRegressive Decomposition") à l'avantage de prendre en compte la nature mixte "voisé/non voisé" des mélanges de parole.
La méthode d'estimation permet d'éviter l'utilisation d'une étape de soustraction après chaque estimation et qui limite généralement les performances des méthodes itératives. Elle permet également de définir un détecteur de voisement utilisé pour stopper les itérations successives et ainsi estimer le nombre de sources harmoniques du mélange.

La seconde partie est consacrée aux mélanges de signaux musicaux. L'approche développée pour ces derniers, est basée sur la classification des pics spectraux du mélange. Elle est à l'origine de trois nouvelles méthodes d'estimation de fréquences fondamentales multiples, qui diffèrent par leur classification : la première utilise l'algorithme des Nuées dynamiques, la seconde est une classification basée sur les relations harmoniques entre fréquences et la troisième est une fusion des deux précédentes. Toutes permettent l'estimation conjointe des fréquences fondamentales pour des mélanges dont le nombre de notes est supposé inconnu.
Elles permettent également de traiter des mélanges complexes, spécifiques à la musique, comme les accords pour lesquels il existe des recouvrements spectraux entre les sources. Elles s'adapte à la diversité des mélanges, qui peuvent contenir aussi bien des voix chantées, que du piano.

Les méthodes d'estimation de fréquences fondamentales multiples proposées dans ce travail offrent de nombreuses perspectives. Quelques unes sont présentées ci-dessous :

Le cadre très restrictif de l'analyse monophonique, dans lequel ces méthodes d'estimation ont été développées, font d'elles un atout pour les enregistrements multicapteurs : la méthode d'estimation est applicable à chaque signal enregistré par un microphone. Ainsi, plus le nombre de microphones mis à disposition est grand, plus la quantité d'information fréquentielle à extraire est importante. Dans le cas d'une prise de son stéréophonique, un second enregistrement dans lequel l'intensité relative entre les sources est différente de celle du premier enregistrement, permettra d'améliorer la détection et l'estimation des fréquences fondamentales du mélange.

L'ensemble des nouvelles méthodes proposées ont en commun la possibilité d'estimer le nombre de sources d'un mélange. Ce dernier est un paramètre descriptif qui prend toute son importance dans des applications telles que l'indexation sonore, la reconnaissance multi-locuteurs ou multi-instruments. L'information fournit par application de ces méthodes d'estimation apparaît dès lors, comme un moyen de généraliser au cas de plusieurs sources, des nombreuses méthodes de reconnaissance ou d'indexation initialement développées pour des signaux contenant une source isolée.

Bien que des méthodes d'estimation différentes ont été proposées pour traiter les mélanges de parole et de musique, elles restent applicables à tout signal modélisable par une somme de sinusoïdes harmoniques. Les deux approches bien distinctes que sont l'estimation temporelle itérative pour la parole et l'estimation spectrale conjointe pour la musique, peuvent alors s'appliquer de façon complémentaire : le détecteur de voisement développé dans le cadre de la parole est par exemple directement utilisable pour détecter les zones de silence d'un mélange musical. Ce dernier est alors calculé à partir de fréquences fondamentales estimées à l'aide de l'une des trois méthodes d'estimation conjointe. De même, la rapidité d'exécution de la méthode d'estimation par *pénalisation sur l'harmonicité*, permet d'utiliser les résultats de cette dernière comme des estimations "grossières" des différentes fréquences fondamentales contenues dans un mélange de parole, qui pourront être par la suite affinées par la méthode d'estimation basée sur le critère d'ordre.
Plus généralement, les bonnes performances obtenues sur l'ensemble des mélanges traités, nous incitent à penser que l'association de ces méthodes, permettrait de traiter des mélanges très complexes contenant à la fois de la parole et de la musique.

Bien que les méthodes d'estimation conjointes développées dans un cadre musical, ont l'avantage de prendre en compte les recouvrements spectraux, elles sont par construction, limitées au traitement d'intervalles de tierce ou encore de quinte.

Une prochaine étape dans le problème d'estimation de fréquences fondamentales multiples est alors celui du traitement de mélanges contenant des notes jouées à l'octave. Dans cet objectif, l'idée première consisterait à introduire des contraintes sur l'enveloppe spectrale qui permettrait de décider si un spectre de source donné correspond à une ou plusieurs notes.

Outre ces perspectives, les méthodes d'estimation de fréquences fondamentales soulèvent également d'autres questions qui pourront être à l'origine de nouveaux thèmes de recherche, comme par exemple celui de la reconnaissance ou de la poursuite de sources à partir d'un ensemble de fréquence fondamentales préalablement estimées, en vue d'une synthèse sonore. Plus précisément, on pourrait se demander à quelle source doit-on attribuer une fréquence fondamentale estimée.

Annexe A

Estimation du modèle CARD

Cette annexe est consacrée à l'estimation des paramètres du modèle CARD par le principe du maximum de vraisemblance. On rappelle tout d'abord la définition du modèle CARD. Par la suite l'expression de sa vraisemblance est déterminée. Une dernière partie détaille l'ensemble des calculs permettant l'estimation des paramètres.

On considère une modélisation CARD d'un nombre fini d'observations $\underline{\mathbf{y}} = [y(0)y(1)\cdots y(N-1)]^T$:

$$y(n) = s(n) + v(n) \tag{29}$$

où s(n) est une somme de M sinusoïdes et v(n) un processus autorégressif d'ordre p, généré par un bruit blanc gaussien u(n) de moyenne nulle et de variance σ^2 tels que :

$$\begin{cases} s(n) = \sum_{k=1}^{M} [A_k cos(2\pi f_k n) + B_k sin(2\pi f_k n)] \\ v(n) = -\sum_{m=1}^{p} a_m v(n-m) + u(n) \end{cases} \tag{30}$$

Les paramètres du modèle CARD sont les p coefficients AR $\{a_m\}_{m=1}^{p}$, les 2M amplitudes $\{A_k, B_k\}_{k=1}^{M}$ associées aux fréquences des partiels $\{f_k\}_{k=1}^{M}$.

• Expression de la vraisemblance du modèle CARD

On établie l'expression de la vraisemblance pour le modèle CARD défini ci-dessus. Cette dernière est obtenue à partir de la densité de probabilité du modèle CARD notée $p(\underline{\mathbf{y}}|\{a_m\},\{A_k,B_k,f_k\})$

Il est possible de factoriser cette densité de probabilité du modèle CARD en un produit d'une vraisemblance conditionnelle et d'une vraisemblance marginale :

$$p(\underline{\mathbf{y}}|\{a_m\},\{A_k,B_k,f_k\}) = p(\underline{\mathbf{y}}_{\mathbf{p}}|\mathbf{y(0)},\cdots,\mathbf{y(p-1)}) \times \mathbf{p(y(0)},\cdots,\mathbf{y(p-1)}) \tag{31}$$

où :
$\underline{\mathbf{y}}_p = [y(p), y(p+1),\cdots,y(N-1)]$,
$k \epsilon \{1,\cdots,M\}$,
$m \epsilon \{1,\cdots,p\}$.

En posant la condition $p(y(0),\cdots,y(p-1)) = 1$, on obtient une expression approchée de la vraisemblance :

$$p(\underline{\mathbf{y}}|\{a_m\},\{A_k,B_k,f_k\}) \simeq p(\underline{y}_p|y(0),\cdots,y(p-1)) \tag{32}$$

Le vecteur des observations $\underline{y}_p$ suit la distribution du bruit gaussien u(n) qui, pour $a_0 = 1$, s'écrit sous la forme :

$$u(n) = \sum_{m=0}^{p} \{a_m[y(n-m) - s(n-m)]\} \tag{33}$$

L'expression de la vraisemblance du modèle CARD est finalement donnée par :

$$p(\underline{\mathbf{y}}|\{a_m\},\{A_k,B_k,f_k\}) \simeq \frac{1}{(\pi\sigma^2)^{(N-p)}}\ exp[-\frac{1}{\sigma^2}\sum_{n=p}^{N-1}|u(n)|^2] \tag{34}$$

• Estimation des paramètres par la méthode du Maximum de Vraisemblance

Le problème d'optimisation

Les estimations de la variance σ^2, des amplitudes $\{A_k, B_k\}_{k=1}^{M}$ et des paramètres AR $\{a_m\}_{m=1}^{p}$ sont successivement obtenus à partir de l'expression de la vraisemblance (34). Ils sont ensuite réintroduits dans le problème d'optimisation, afin de déterminer l'expression d'une fonction coût ne dépendant que des fréquences du modèle :

La variance $\widehat{\sigma}^2$ estimée est obtenue en annulant la dérivée de $p(\underline{\mathbf{y}}|\{a_m\}, \{A_k, B_k, f_k\})$ par rapport à σ^2. Son expression est alors donnée par :

$$\widehat{\sigma}^2 = \frac{1}{(N-p)} \sum_{n=p}^{N-1} \Big| \sum_{m=0}^{p} a_m[y(n-m) - s(n-m)]\Big|^2 \tag{35}$$

En remplaçant σ^2 par son estimé dans l'expression (34), la maximisation de la vraisemblance est équivalente à la minimisation de la fonction coût suivante :

$$C_0(\{a_m\}, \{A_k, B_k, f_k\}) = \sum_{n=p}^{N-1} \Big| \sum_{m=0}^{p} a_m[y(n-m) - s(n-m)]\Big|^2 \tag{36}$$

Cette expression est équivalente à un problème des moindres carrés non-linéaire, causé par l'interaction entre les éléments $\{a_m\}$ et $\{A_k, B_k\}$. Un moyen de linéariser le problème est d'effectuer une transformation des paramètres d'amplitude. En effet, en explicitant le terme sinusoïdal $s(n-m)$ dans l'expression (36), on a :

$$C_0(\{a_m\}, \{A_k, B_k, f_k\}) = \sum_{n=p}^{N-1} | \sum_{m=0}^{p} a_m y(n-m) + \sum_{k=1}^{M} c_k \; cos(2\pi f_k n) + \sum_{k=1}^{M} s_k \; sin(2\pi f_k n)|^2 \tag{37}$$

c_k et s_k sont les paramètres transformés définis par :

$$\begin{cases} c_k = [-A_k \; \sum_{m=0}^{p} a_m cos(2\pi f_k m) + B_k \; \sum_{m=0}^{p} a_m sin(2\pi f_k m)] \\ s_k = [-A_k \; \sum_{m=0}^{p} a_m sin(2\pi f_k m) - B_k \; \sum_{m=0}^{p} a_m cos(2\pi f_k m)] \end{cases} \tag{38}$$

Ecriture matricielle

Soit le vecteur $\underline{\theta}$ contenant les paramètres du modèle CARD :

$$\underline{\theta} = [\sigma^2\, c_1 \cdots c_M\, s_1 \cdots s_M\, a_1 \cdots a_m\, f_1 \cdots f_k]^T \tag{39}$$

Les différents termes intervenant dans le problème de minimisation (37) mis sous forme vectorielle et matricielle sont définis par :

$$\begin{cases} \underline{a} = [a_1\, a_2 \cdots a_p]^T \\ \underline{f} = [f_1\, f_2 \cdots f_M]^T \\ \underline{\mu} = [c_1\, c_2 \cdots c_M\, s_1\, s_2 \cdots s_M]^T \end{cases} \tag{40}$$

Le vecteur $\underline{\mu}$ contient l'ensemble des paramètres d'amplitudes transformées. Le vecteur paramètre $\underline{\theta}$ défini par (39) s'écrit alors :

$$\underline{\theta} = [\sigma^2\ \underline{\mu}^T\, \underline{a}^T\, \underline{f}^T] \tag{41}$$

Les p dernières observations sont introduites vectoriellement par :

$$\underline{y}_p = [y(p)\, y(p+1) \cdots y(N-1)]^T \tag{42}$$

On définit également les matrices suivantes :

$$\mathbf{Y} = \begin{bmatrix} y(p-1) & y(p-2) & \cdots & y(0) \\ y(p) & y(p-1) & \cdots & y(1) \\ \vdots & \vdots & \ddots & \cdots \\ y(N-2) & y(N-3) & \cdots & y(N-1-p) \end{bmatrix} \tag{43}$$

$$\mathbf{\Omega(f)} = \begin{bmatrix} cos(2\pi f_1(p)) & \cdots & cos(2\pi f_M(p)) & sin(2\pi f_1(p)) & \cdots & sin(2\pi f_M(p)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1(p+1)) & \cdots & cos(2\pi f_M(p+1)) & sin(2\pi f_1(p+1)) & \cdots & sin(2\pi f_M(p+1)) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \\ cos(2\pi f_1(N-1)) & \cdots & cos(2\pi f_M(N-1)) & sin(2\pi f_1(N-1)) & \cdots & sin(2\pi f_M(N-1)) \end{bmatrix} \quad (44)$$

L'expression matricielle de la fonction C_0 à minimiser (36), est finalement donnée par :

$$C_0(\underline{a}, \underline{\mu}, \underline{f}) = \|\underline{y}_p + Y\underline{a} + \Omega(\underline{f})\underline{\mu}\|^2 = (\underline{y}_p + Y\underline{a} + \Omega\underline{\mu})^H(\underline{y}_p + Y\underline{a} + \Omega\underline{\mu}) \quad (45)$$

Estimation des paramètres

Pour obtenir les vecteurs paramètres estimés $\underline{\hat{a}}$ et $\underline{\hat{\mu}}$, on effectue la minimisation de C_0 successivement par rapport à $\underline{a}$ et $\underline{\mu}$.

La minimisation de C_0 par rapport $\underline{\mathbf{a}}$ donne (en omettant la dépendance aux fréquences $\underline{\mathbf{f}}$ de la matrice Ω pour plus de clarté) :

$$\underline{\widehat{a}}(\underline{\mu}, \underline{f}) = -(Y^HY)^{-1}Y^H(\underline{y}_p + \Omega\underline{\mu}) \quad (46)$$

et l'expression de C_1 devient :

$$C_0(\underline{\widehat{a}}(\underline{\mu}, \underline{f}), \underline{\mu}, \underline{f}) = (P_Y^\dagger \underline{y}_p + P_Y^\dagger \Omega\underline{\mu})^H(P_Y^\dagger \underline{y}_p + P_Y^\dagger \Omega\underline{\mu}) \quad (47)$$

où I est la matrice identité et $P_Y^\dagger = I - Y(Y^HY)^{-1}Y^H$ est un opérateur qui permet de projeter le vecteur des observations $\underline{y}_p$ sur le sous-espace engendré par les vecteurs colonnes de la matrice Y. Cette matrice a la propriété d'être hermitienne et telle que :

$$(P_Y^\dagger)^H = P_Y^\dagger = P_Y^\dagger\, P_Y^\dagger \quad (48)$$

Par la suite, la minimisation par rapport à $\underline{\mu}$ fournit l'expression de $\underline{\widehat{\mu}}$:

$$\underline{\widehat{\mu}}(\underline{f}) = -(\Omega^H P_Y^\dagger \Omega)^{-1}\Omega P_Y^\dagger \underline{y}_p \quad (49)$$

Remarque : Il est à noter que le choix de minimiser tout d'abord par rapport à $\underline{a}$,

puis par rapport à $\underline{\mu}$ est arbitraire et les expressions des estimés lorsque ce choix est inversé sont données par :

$$\begin{cases} \widehat{\underline{\mu}}(\underline{a},\ \underline{f}) = -(\Omega^H\Omega)^{-1}\Omega^H(\underline{y}_p + Y\underline{a}) \\ \widehat{\underline{a}}(\underline{f}) = -(Y^H P_\Omega^\dagger Y)^{-1} Y P_\Omega^\dagger \underline{y}_p \end{cases} \tag{50}$$

avec $P_\Omega^\dagger = I - \Omega(\Omega^H\Omega)^{-1}\Omega^H$.

En remplaçant $\underline{a}$ et $\underline{\mu}$, respectivement par (46) et (49), on obtient une expression de la fonction coût C_0 qui ne dépend plus que des fréquences $\{f_k\}$ du modèle, par le biais de la matrice Ω. On a :

$$C(\underline{f}) = \underline{y}_p[P_Y^\dagger - P_Y^\dagger\,\Omega(\Omega^H\,P_Y^\dagger\,\Omega)^{-1}\Omega^H P_Y^\dagger]\underline{y}_p \tag{51}$$

Finalement, la minimisation de cette dernière expression de C_0 est utilisée pour obtenir les estimés ML du vecteur $\underline{f}$.

Annexe B

Comportement asymptotique

On s'intéresse ici aux propriétés asymptotiques de l'estimateur de vraisemblance pour la modélisation CARD définie par (2.1) et (2.2).
Ces propriétés sont utilisées pour diminuer le coût de calcul du principe MDL sur lequel se base la méthode d'estimation multipitch proposée pour la séparation de signaux de parole.
L'utilisation courante de modèles "sinusoïdal plus bruit", dont fait partie le modèle CARD, dans des domaines d'application très diversifiés (Analyse sonore, Biomédical, $\cdots$) est à l'origine de nombreuses études sur la distribution asymptotique des paramètres estimés de ce type de modèle. Nous proposons ci-dessous, un résumé de ces dernières, ainsi qu'un ensemble de références renvoyant aux démonstrations longues ou aux détails de calcul.

• Modèle "sinusoïdes plus bruit"

Nous rappelons dans ce paragraphe la définition la plus générale du modèle "sinusoïdes plus bruit", ainsi que les diverses hypothèses faites sur ses composantes.

On considère le modèle suivant :

$$y(n) = s(n, \theta) + v(n) \tag{52}$$

$s(n, \theta)$ est constitué d'une somme de M sinusoïdes et d'une constante μ tel que :

$$s(n,\theta) = \mu + \sum_{m=1}^{M} A_m cos(2\pi n f_m) + B_m sin(2\pi n f_m) \tag{53}$$

L'ensemble des paramètres θ est donné par les amplitudes et fréquences :

$$\theta = \{A_1, \cdots, A_M, B_1, \cdots, B_M, f_1, \cdots, f_M\}$$

.

$v(n)$ est un processus aléatoire stationnaire à valeurs réelles, de moyenne nulle tel que :

$$v(n) = \sum_{-\infty}^{\infty} \alpha(l)\ \epsilon(n-l), \qquad \sum_{-\infty}^{\infty} |\alpha(l)| < \infty \tag{54}$$

où ϵ est un processus i.i.d, de moyenne nulle et de variance σ^2. Son spectre de puissance est donné par :

$$f_v(\lambda) = \frac{\sigma^2}{2\pi} \sum_{l} |\alpha(l) e^{-jl\lambda}|^2, \tag{55}$$

$$\tag{56}$$

• Analyse de séries temporelles

Dans les années 70, Hannan s'intéresse au comportement asymptotique des paramètres du modèle "sinusoïdes plus bruit", et cherche à étendre les résultats asymptotiques développés dans l'analyse des séries temporelles. Il obtient l'estimation des paramètres sinusoïdaux par les moindres carrés pondérés dans le domaine spectral en définissant le périodogramme résiduel :

$$I(\lambda;\theta) = \frac{1}{2\pi N} \Big| \sum_{n=1}^{N} \{y(n) - s(n,\theta)\} e^{(-j\lambda n)} \Big| \tag{57}$$

Les estimés $\hat{\theta}_N$ correspondent à la valeur θ qui minimise la quantité suivante :

$$R_N(\theta) = \frac{1}{T} \sum_{n=0}^{N-1} \phi(\lambda_n) I(\lambda_n;\theta) \tag{58}$$

où $\phi(\lambda_n)$ est une fonction de pondération strictement positive et $\lambda_n = 2\pi nf$.

Dans un premier temps, Hannan prouve pour le cas d'une sinusoïde unique, la forte consistance et la normalité asymptotique de $\hat{\theta}_N$ [59]. Il montre que la fonction de pondération optimale, pour laquelle la variance asymptotique des estimés est minimisée, est donnée par $\phi(\lambda) = f_v(\lambda)$. Il montre de plus que la distribution asymptotique de $\hat{\theta}_N$ reste inchangée si $f_v(\lambda)$ est remplacé par un estimé consistant. Par la suite, Hannan traite le cas où $\mu = 0$ et $d = 1$ pour l'équation (53). Il suppose que la distribution asymptotique est indépendante de $\phi(\lambda)$ et pose $\phi(\lambda) = 1$. Il étend alors ses résultats et établit la convergence et la consistance lorsque $\lambda = 0$ et $\lambda = 1/2$ [60].
Il s'intéresse également au cas de sommes de sinusoïdes harmoniques [61]. La partie sinusoïdale du modèle est alors telle que :

$$s(n,\theta) = \sum_{k=1}^{K} s_k(n,\theta_k) = \sum_{k=1}^{K} \Big[\sum_{m=1}^{M_k} A_{m,k} cos(2\pi n f_{m,k}) + B_{m,k} sin(2\pi n f_{m,k}) \Big] \quad (59)$$

où les fréquences de chaque signal $s_k(n,\theta_k)$ sont telles que$f_{m,k} = m.f_{1,k}$.

Hannan décrit la façon d'estimer les paramètres du modèle à partir de l'équation (58). Il donne également sans justification, l'expression de la variance asymptotique des estimés.
Ces travaux seront entre autres repris par Brown, qui constate une erreur dans l'expression variance asymptotique [62]. Ce dernier propose alors une nouvelle expression qui dépend cette fois-ci de $\phi(\lambda)$.

• Estimation découplée des paramètres

Quelques années plus tard, un travail conséquent sur le comportement asymptotique des modèles "sinusoïdes plus bruit" a été proposé par Stoica. Il fait appel à la notion de spectres mixtes, définis comme des spectres qui possèdent à la fois des composantes discrètes et continues.

Son travail consiste à l'évaluation des performances asymptotiques des estimations sinusoïdales obtenues par deux estimateurs non-linéaires de type moindres carrés. Le premier est l'estimateur des moindres carrés basique (BNLSE), qui ne prend pas en compte la coloration du bruit. Le second estimateur, encore appelé

estimateur des moindres carrés optimal (ONLSE), tient compte de cette coloration. Les paramètres sinusoïdaux estimés par ce dernier, sont équivalent à ceux de l'estimateur des moindres carrés pondérés évoqué précédemment.

Le bruit $\{v(n)\}$ est défini comme un processus linéaire stationnaire de moyenne nulle tel que :

$$v(n) = H(q^{-1})\,\epsilon(n) \tag{60}$$

où $H(q^{-1})$ est un filtre linéaire inversible exponentiellement stable et $\{\epsilon(n)\}$ est un bruit blanc de moyenne nulle et de variance unitaire.

Les paramètres sinusoïdaux estimés dans les deux cas sont donnés par :

$$\hat{\theta}_{BNLSE} = argmin_{\theta}\,[\sum_{n=0}^{N-1}\epsilon^2(n)] \tag{61}$$

$$\left\{\hat{\theta}_{ONLSE}; \hat{\beta}_{ONLSE}\right\} = argmin_{\theta}\,[\sum_{n=0}^{N-1}[H^{-1}(q^{-1})\,\epsilon(n)]^2] \tag{62}$$

où β_{ONLSE} est l'ensemble des paramètres de bruit associés à $H(q^{-1})$.

Lorsque le bruit $\{\epsilon(n)\}$ est supposé gaussien, l'expression (62) est alors équivalente à une constante près à la log-vraisemblance négative conditionnelle des échantillons $\{y(1), \cdots, y(N-1)\}$, par rapport aux valeurs initiales $\{y(0),\ y(-1),\ \cdots\}$. Lorsque pour ces dernières, on pose $y(n) = s(n)$ pour $n < 0$, on introduit une erreur dans la fonction de vraisemblance. Malgré tout, l'hypothèse d'un filtre stable $H(q^{-1})$, permet d'annuler cette erreur lorsque N augmente.

Stoica prouve l'équivalence asymptotique des deux estimateurs $\hat{\theta}_{ONLSE}$ et $\hat{\theta}_{BNLSE}$, en montrant qu'ils atteignent tous les deux la borne de Cramer-Rao pour N grand. Il traite tout d'abord le cas de sinusoïdes réelles dans du bruit coloré [63], puis celui des "cisoïdes" ou sinusoïdes complexes [50] [51].

L'équivalence asymptotique des estimateurs implique que pour un nombre d'observations suffisamment grand, il est préférable de déterminer les paramètres sinusoïdaux $\hat{\theta}$ du modèle à l'aide de l'estimateur des moindres carrés basique, beaucoup moins coûteux car il ne prend pas en compte les paramètres de bruit.

Parce que dans de nombreuses applications, il est nécessaire de modéliser la partie bruit du modèle, et plus particulièrement dans le cas de parole ou l'on cherche à modéliser des mélanges "voisés/non-voisé", la conséquence directe de cette équivalence asymptotique est qu'elle permet d'appliquer le principe de "l'estimation découplée" des paramètres sinusoïdaux et des paramètres de bruit. Ainsi dans

une première étape, les paramètres sinusoïdaux sont estimés sans tenir compte de la coloration du bruit. Ces derniers sont ensuite soustraient du signal et le résiduel ainsi obtenu est utilisé pour estimer les paramètres de bruit. C'est à l'aide de ce principe que Stoica développe l'algorithme RELAX pour l'estimation des spectres mixtes [47].
C'est également ce même principe qui est à l'origine de la diminution du coût de calcul dans l'application du principe MDL, pour le problème d'estimation multipitch proposé pour la parole. Il permet en effet, de développer une estimation de la fréquence fondamentale d'un modèle CARD harmonique en deux étapes distinctes où la première consiste à estimer une valeur sous-harmonique de la fréquence fondamentale par la méthode des moindres carrés. Les paramètres de bruit sont alors seulement pris en compte dans la seconde étape lors du calcul de la log-vraisemblance pénalisée (2.22). Ils permettent également de définir l'écart de vraisemblance (3.8), utilisé par la suite comme détecteur de voisement.

Bibliographie

[1] W. Hess, *Pitch Determination of Speech Signals*, Springer-verlag, 1983.

[2] M. R. Schroeder, "Period histogram and product spectrum : New methods for fundamental frequency measurements," *J. Acoust. Am. Soc.*, vol. 43, 1968.

[3] T.W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Am.*, vol. 60, pp. 911–918, 1976.

[4] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum , the harmonic sum spectrum and a maximum likelihood system," in *Symposium on Computer Processing in Communications*, Polytechnic Institute of Brooklyn, Ed., 1969.

[5] A. M. Noll, "Cepstrum Pitch Determination," *J. Acous. Soc. Am.*, vol. 41, 1967.

[6] R.J. Stubbs and Q. Summerfield, "Evaluation of Two Voice-Separation Algorithms Using Normal-Hearing and Hearing-Impaired Listeners," *J. Acoust. Soc. Am.*, vol. 84, no. 4, october 1988.

[7] S. M. Kay and S. L. Marple, "Spectrum analysis- a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, November 1981.

[8] J. Naylor and J. Porter, "An Effective Speech Separation System Which Requires No A Priori Information," *IEEE Trans. on ASSP*, 1991.

[9] S. L. Marple, *Digital Spectral Analysis with applications*, Prentice-Hall, Englewood Cliffs, 1984.

[10] C.K. Lee and D.G. Childers, "Cochannel Speech Separation," *J. Acoust. Soc. Am.*, vol. 83, pp. 274–280, 1988.

[11] E. Mumolo and M. Romanin, "New Speech Separation Algorithms," *Signal Processing*, 1994.

[12] R.C. Maher, *An Approach for the Separation of Voices in Composite Musical Signal,*, Ph.D. thesis, University of Illinois at Urbana-Champaign, June 1989.

[13] R. Mc Aulay and T. Quatieri, "Speech Analysis/Synthesis based on a Sinusoidal Representation," *IEEE Trans. on ASSP*, August 1986.

[14] A. De Cheveigne, "Separation of Concurrent Harmonics Sounds : Fundamental Frequency Estimation and Time-Domain Cancellation Model of Auditory Processing," *J. Acoust. Soc. Am.*, vol. 6, 1993.

[15] A ; De Cheveigné, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, no. 175 -185, 1999.

[16] E. Mumolo and L. Salvi, "New Speech Processing Algorithms Based on Sinusoidal Modeling," *Signal Processing*, 1994.

[17] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum Likelihood Pitch Estimation," *IEEE Trans. on ASSP*, vol. 24, no. 5, 1976.

[18] D.P. Morgan, B. George, and L.T. Lee, "Cochannel Speaker Separation by Harmonic Enhancement and Suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, september 1997.

[19] R. J. MacAulay and T.F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Trans. on ASSP*, vol. 34, no. 4, 1986.

[20] T.F. Quatieri and R.G. Danisewicz, "An Approach to Cochannel Talker Interference Suppression Using a Sinusoidal Model for Speech," *IEEE transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-38, no. 1, january 1990.

[21] D.S. Benincasa and M.I. Savic, "Co-Channel Speaker Separation Using Constrined Nonlinear Optimization," *Proc. of ICASSP*, 1997.

[22] D. Chazan, Y. Stettiner, and D. Malah, "Optimal Multi-Pitch Estimation Using The EM Algorithm for Co-Channel Speech Separation," *IEEE Trans. on ASSP*, 1993.

[23] De cheveigne, "Separation of Concurrent Harmonic Sounds : Fundamental Frequency Estimation and Time-Domain Cancellation Model of Auditory Processing," *J. Acoust. Soc. Am*, 1993.

[24] T. Tolonen, "Methods for Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in *AES 106th Convention,*, Munich, Germany, 1999.

[25] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.*, vol. 3, pp. pp 1811 and 1820, 1997.

[26] D. W. P. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. thesis, Mass. Inst. Technol., Cambridge, 1996.

[27] T. Virtanen and A. Klapuri, "Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation," *IEEE Proc. of WASPAA*, 2001.

[28] T. Virtanen and A. Klapuri, "Separation of harmonics sounds using linear models for the overtones series," in *Proceedings of ICASSP*, 2002.

[29] A. Klapuri, "Pitch estimation using multiple independent time-frequency windows," in *Proceedings of WASPAA*, 1999.

[30] A. Klapuri, "Number theorical means of resolving a mixture of several harmonics sounds," in *Proceedings of EUSIPCO*, 1998.

[31] A. Klapuri, "Wide-band pitch estimation for natural sound sources with inharmonicities," in *Proceedings of AES*, 1999.

[32] Moore, Glasberg, and Baer, "A model for the prediction of thresholds, loudness and partial loudness," *J.Acoust.Soc.Am.*, vol. 4, April 1997.

[33] T. Virtanen and A. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proceedings of ICASSP*, 2001.

[34] F. M. Silva and L. B. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," *Proceeding of ICASSP*, pp. 809, 812, 1990.

[35] T. Virtanen and A. Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," *IEEE Trans. on ASSP*, 2000.

[36] T. Virtuanen, "Audio signal modeling with sinusoid plus noise," M.S. thesis, Tampere University of Technology, 2000.

[37] A. S. Bregman, *Auditory Scene Analysis*, MIT press, 1990.

[38] X. Serra and J. Smith, "Spectral Modeling Synthesis : A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, 1990.

[39] B.A. Hanson and D.Y. Wong, "The Harmonic Magnetude Suppression (HMS) Technique for Intelligibility Enhancement in the Presence of Interfering Speech," *Proc. of ICASSP*, vol. 2, 1984.

[40] J.A. Naylor and S.F. Boll, "Techniques for Suppression of an Interfering Talker in Cochannel Speech," *Proc. of ICASSP*, 1987.

[41] S. Kay and V. Nagesha, "Extraction of Periodic Signals in Colored Noise," *Proc. of ICASSP*, vol. V, 1992.

[42] V. Nagesha and S. Kay, "Spectral Analysis based on the Canonical Autoregressive Decomposition," *IEEE Trans. on Signal Processing*, July 1996.

[43] S. M. Kay and V. Nagesha, "Maximum Likelihood Estimation of Signals in Autoregressive Noise," *IEEE Trans. on Signal Processing*, vol. 42, no. 1, January 1994.

[44] J. Rissanen, "Modeling by Shortest Data Description," *Automatica*, 1978.

[45] H. Akaike, "A new look at The Statistical Model Identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, December 1974.

[46] E.J Hannan and B.G. Quinn, "The determination of the order of an autoregression," *J. R. Statist. Soc. B*, , no. 41, N°2, 1979.

[47] J. Li and P. Stoica, "Efficient Mixed-Spectrum Estimation with Applications to Target Feature Extraction," *IEEE Trans. on Signal Processing*, vol. 44, no. 2, february 1996.

[48] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics,*, , no. 2, 1978.

[49] C. Chatterjee, R. L. Kashyap, and G.K. Boray, "Estimation of close Sinusoids in Colored Noise and Model Discrimination," *IEEE Trans. on ASSP*, vol. 35, 1987.

[50] P. Stoica, A. Jakobsson, and J. Li, "Cisoid Parameter estimation in the Colored Noise Case : Asymptotic Cramer-Rao Bound, Maximum Likelihood and Nonlinear Least-squares," *IEEE trans. on Signal Processing*, August 1997.

[51] A. Jakobsson P. Stoica and J. Li, "Asymptotic Results for Cisoid Parameter Estimation in The Colored Noise Case," *11th IFAC Symposium on System Identification*, July 1997.

[52] H. Ney, "A Dynamic Programming Technique for Nonlinear Smoothing," *IEEE Trans. on ASSP*, vol. 5, 1981.

[53] T. Virtanen and A. Klapuri, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proceedings of DAFx*, 2000.

[54] M. Jambu, *Exploration informatique et statistique des données*, DUNOD, coll. ENST, 1989.

[55] Mac Queen, "Some Methods for Classification and Analysis of Multivariate Observation," in *Univ. of California Press,*, Berkeley, 1967.

[56] G. H. Ball and D. J. Hall, *Isodata : a novel method of data analysis and pattern classification*, Technical report, Standford Research Institute, 1965.

[57] E. Diday, *Optimisation en classification automatique*, INRIA, 1980.

[58] IRCAM, "Studio on line," .

[59] E.J. Hannan, "Non-linear Time Series Regression," *Journal of Applied Probability*, vol. 8, 1971.

[60] E.J. Hannan, "The Estimation of Frequency," *Journal of Applied Probability*, vol. 10, 1973.

[61] E.J. Hannan, "Time Series Analysis," *IEEE Transactions on Auto Control*, vol. 19, 1974.

[62] E.N. Brown, "A note on the Asymptotic Distribution of the Parameter Estimates for the harmonic regression model," *Biometrika*, vol. 77, 1990.

[63] P. Stoica and A. Nehorai, "Statistical Analysis of Two Nonlinear Least-Squares Estimators of Sine-wave Parameters in the Colored-Noise Case," *Circuits Systems Signal Processing*, 1989.