Transcription automatique de la musique de piano

Valentin Emiya

École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

Thèse

présentée pour obtenir le grade de docteur

de l'École Nationale Supérieure des Télécommunications

Spécialité : Signal et Images

Valentin EMIYA

Transcription automatique de la musique de piano

Soutenue le 8 octobre 2008 devant le jury composé de

Christophe d'Alessandro	Président
Alain de Cheveigné	Rapporteurs
Laurent Daudet	
Anssi Klapuri	Examinateurs
Gaël Richard	
Emmanuel Vincent	Invité
Bertrand David	Directeurs de thèse
Roland Badeau	

Remerciements

Ce doctorat a été l'occasion de nombreuses collaborations et rencontres. Je souhaite saluer et remercier l'ensemble des personnes qui m'ont aidé de près ou de loin à mener ce projet de recherche à terme.

Mes premiers remerciements vont bien évidemment à Bertrand David et à Roland Badeau qui ont supervisé cette thèse et avec qui j'ai pu étroitement travailler. Merci à eux deux pour tout ce j'ai pu découvrir et apprendre, pour les conseils qu'ils m'ont donnés, pour leur attention et leur soutien aux moments cruciaux, pour la complémentarité de leur supervision et pour la qualité de la relation qui s'est instaurée entre nous.

Merci à l'ensemble des membres du jury pour l'attention et l'intérêt portés à ces travaux, aussi bien dans la phase de lecture du mémoire que lors de la soutenance. En particulier, merci à Laurent Daudet et à Alain de Cheveigné pour la qualité de leurs rapports. Merci également à Anssi Klapuri pour avoir examiné le mémoire écrit en français et à Emmanuel Vincent pour ses commentaires et suggestions détaillés.

Merci à Nancy Bertin pour les travaux que nous avons menés ensemble et pour la relecture de ce mémoire, à Cédric Févotte pour les directions de recherche déterminantes qu'il m'a suggérées, à Adrien Daniel pour notre étude sur l'évaluation des transcriptions. Merci également à Jean-Louis Durrieu, Pierre Leveau, Laurent Oudre, Slim Essid, Sylvain Streiff, ainsi qu'à Laurent Daudet de l'équipe LAM de l'Institut Jean Le Rond D'Alembert, à Sylvain Marchand du LaBRI, Olivier Derrien du laboratoire STICS et à Antony Schutz d'Eurécom pour les collaborations que nous avons pu avoir et dont je garde un excellent souvenir.

Merci à Gaël Richard, à Yves Grenier et à toute l'équipe AudioSig du département de Traitement du Signal et des Images de TELECOM ParisTech pour m'avoir accueilli et pour faire de la recherche un travail d'équipe si captivant. Merci à ceux qui ont contribué à ce que le cadre de travail à TELECOM ParisTech soit le meilleur possible. Merci notamment à Bernard Robinet, à Marc Peyrade et à Henri Maître pour les échanges que nous avons eus. Merci également à Fabrice Planche, à Laurence Zelmar, à Patricia Friedrich, à Rahma Lamech et à Sophie-Charlotte Barrière pour leur assistance administrative et technique, à Peter Weyer-Brown et à James Benenson pour leurs conseils et leur aide en anglais, ainsi qu'à Clara, à Maryse, à François Auguste et à toute son équipe.

Merci aux chercheurs du LIMSI, et particulièrement à Christophe d'Alessandro, à François Signol, à Nicolas Sturmel, à François Rigaud et à Albert Rilliard de m'avoir accueilli et permis de profiter au maximum du peu de temps que j'ai pu consacrer au traitement de la parole et à mon séjour dans ce laboratoire.

Merci à tous les enseignants-chercheurs avec qui j'ai travaillé dans le cadre de mon monitorat à l'Université Denis Diderot - Paris 7 et pendant mon année d'ATER à l'Université Paris Sud-11, en particulier à Dominique Poulalhon et Anne Micheli (LIAFA), à Véronique Vèque, à Lionel Lacassagne, à Christian Clapier, à Hervé Mathias, à Hugues Mounier, à

Pavel Kalouguine, à Michèle Gouiffès (IEF) et à Abdelaziz Kallel.

Merci à Daniel Pressnitzer et Alain de Cheveigné de l'Équipe Audition de l'ENS, à Michèle Castellengo de l'équipe LAM de l'Institut Jean Le Rond D'Alembert, à Daniel Verba de Calypsociation, à Philippe Gal de Tropique du Cancer pour leur aide ponctuelle et précieuse.

Merci à Emmanuel Vincent de l'IRISA, à Nancy Bertin et à Miguel Alonso, à Matija Marolt de l'Université de Ljubljana, à Anssi Klapuri de l'Université de Tampere, à Alain de Cheveigné pour les codes informatiques, issus de leurs travaux, qu'ils ont mis à disposition.

Merci aux actifs du Bureau des Doctorants de TELECOM ParisTech, de la Confédération des Jeunes Chercheurs et du Pôle Île-de-France que j'ai pu côtoyer, merci en particulier à Loïs Rigouste, à Mathieu Guillaume, à Nancy Bertin, à Reda Kaced, à Zahir Larabi, à Charlotte Hucher, à Florent Olivier, à Jasmin Buco, à Maïwenn Corrignan, à Sylvain Collonge, à Olivier Béaslas, à Philippe Gauron, à Alexandra Naba et à Gwenaël Edeline.

Merci à tous les « copains dans la même galère », ceux que j'ai déjà cités, et également Antoine, Aurélia, Brahim, Caroline, Chloé, Cléo, Cyril C., Cyril J., Dora, Eduardo, Fabrice, Félicien, Grégoire, Ismaël, Jean, Julien, Lionel, Mathieu R., Maria, Nataliya, Nicolas, Rémi, Sarah, Simoné, Slim, Tabea, Thomas F., Thomas M., Viet-Son, Zaid, et ceux que j'oublie.

Merci enfin à toutes les personnes de mon entourage qui m'ont soutenu quotidiennement, remonté le moral et supporté dans mes nombreux moments de surmenage, sans qui je n'aurais pu achever cette thèse. Un immense merci à Nathalie et à ma famille qui ont été les plus exposés...

Table des matières

Liste des figures

NB : Dans la version électronique de ce document, il est possible de cliquer sur certaines figures (2.4(a), 5.7 et 5.8) pour voir une vidéo et éventuellement écouter l'extrait sonore associé aux morceaux originaux et transcrits. L'encodage a été réalisé par le codec xvid 1.1.3 (`http://www.xvid.org/`), qui doit être installé pour voir ces vidéos.

Liste des tableaux

Liste des algorithmes

Notations

Symboles, fonctions et opérateurs mathématiques

$\triangleq$	Définition
$[.]$	Arrondi à l'entier le plus proche
$\lfloor\ \rfloor$	Partie entière
$\lceil\ \rceil$	Partie supérieure
$[\![a;b]\!]$	Intervalle entier $\{a; a+1; \ldots; b-1; b\}$
x^*	Conjugué du nombre complexe x
M^{t}	Transposée de la matrice M
$M^\dagger$	Conjuguée hermitienne de la matrice M
M^+	Pseudo-inverse de la matrice M
$\#E$	Cardinal de l'ensemble E
$\otimes$	Convolution circulaire
$\mathcal{F}[x]$	Transformée de Fourier (éventuellement discrète) de x
$\mathcal{F}^{-1}[X]$	Transformée de Fourier (éventuellement discrète) inverse de X
$p_X(X=x)$, $p_X(x)$, $p_x(x)$ ou $p(x)$	Densité de probabilité de la variable aléatoire X pour une réalisation x. Dans un souci de clarté, la notation varie selon le contexte et la même variable désigne parfois la variable aléatoire et sa réalisation.
$\boldsymbol{\nabla} f$	Gradient d'une fonction f multivariée
$\boldsymbol{\nabla^2} f$	Hessien de la fonction f : $\boldsymbol{\nabla^2} f \triangleq \boldsymbol{\nabla}\left((\boldsymbol{\nabla} f)^{\mathrm{t}}\right)$

Acronymes

F_0	fréquence fondamentale
AR	*Auto-regressive*, [filtre/processus] autorégressif
HMM	*Hidden Markov model*, modèle de Markov caché
i.i.d.	Indépendant identiquement distribué
MA	*Moving average*, [filtre/processus à] moyenne ajustée

MAP	Maximum *a posteriori*
ML	*Maximum likelihood*, maximum de vraisemblance
SSL	Stationnaire au sens large
TP	*True positive*, détection correcte
FP	*False positive*, fausse alarme
FN	*False negative*, omission
EI	*Error Intensity*, intensité de l'erreur
MOR	*Mean Overlap Ratio*, taux de recouvrement moyen
MNR	*Modified Note Rate*, taux de notes modifiées
PCM	*Pulse Code Modulation*
MIDI	*Musical Instrument Digital Interface*
SMF	*Standard MIDI File*

Introduction

Pour l'être humain, le son n'a essentiellement d'intérêt qu'en tant que porteur de sens, et non comme vibration physique. La voix parlée porte le langage, la musique une intention artistique et les sons ambiants une image du milieu environnant. Le fonctionnement physiologique sous-jacent est hautement élaboré et notre compréhension du processus partielle. Il en est de même des capacités actuelles à reproduire cette analyse via l'outil informatique : sur bien des points, ses capacités n'égalent pas celles de l'être humain lorsqu'il s'agit de reconnaissance de la parole ou des instruments de musique à partir du son, pour ne citer que ces exemples. Dans cette thèse, nous nous intéressons au cas de la musique, et plus particulièrement de la musique de piano, pour lequel nous chercherons à extraire les notes jouées présentes dans un son en utilisant les outils informatiques et de traitement du signal.

La transcription de la musique

Nous appellerons *transcription musicale* une description symbolique de l'exécution d'un morceau de musique. Dans cette acception, transcrire consiste à analyser le son enregistré ou entendu pour en extraire des informations, c'est-à-dire le contenu faisant sens. Alors que dans le domaine voisin du traitement de la parole, la transcription d'une conversation ou d'un discours a pour but d'extraire les mots et phrases énoncés, la transcription musicale aura avant tout pour objectif d'estimer les notes jouées et leur paramètres : leurs hauteurs, instants d'attaque, durées, et éventuellement des informations de plus haut niveau telles que les figures rythmiques, la mesure ou l'armure.

Avant de rentrer dans le détail des travaux de cette thèse, il faut par ailleurs noter que dans un contexte musical, le terme de transcription peut également désigner une forme d'arrangement consistant en la ré-écriture d'une partition pour une instrumentation autre que l'originale : la réduction pour piano d'une pièce symphonique (comme celles de Liszt des symphonies de Beethoven) ou une orchestration (telle que celle réalisée par Ravel des *Tableaux d'une exposition* pour piano de Moussorgski) sont des exemples de transcriptions au sens d'arrangement. Dans cette thèse, nous écarterons ce sens du mot transcription pour ne garder que celui faisant référence au passage d'un enregistrement à une description symbolique.

La transcription humaine et la transcription automatique

La transcription d'un enregistrement constitue pour l'être humain une tâche difficile nécessitant généralement un apprentissage. La formation de musicien classique consacre un exercice spécifique au développement des facultés de transcription à travers les dictées musicales, première forme de transcription rencontrée par le musicien. L'exercice apparaît dès le début de la formation sous la forme simple de trancriptions de séquences tonales de notes, conjointes dans un premier temps, sans polyphonie, et dont le rythme est simple voire

inexistant. La difficulté croît ensuite sur plusieurs plans : augmentation de la complexité des mélodies, des rythmes et du contenu tonal/atonal, dictées à plusieurs voix, dictées d'accord, etc. Si cet apprentissage constitue une forme d'éducation de l'oreille du musicien qui lui est ensuite utile en situation de jeu, celui-ci rencontre d'autres occasions spécifiques de transcrire la musique qu'il entend. Le musicien de jazz est ainsi souvent amené à transcrire des solos, séquences improvisées pour lesquelles il n'existe pas de partition préalable, et en situation de jeu, doit avoir une oreille suffisamment entraînée pour reconnaître et suivre sur le vif les tonalités jouées par les musiciens qui l'entourent.

La transcription devient automatique lorsqu'elle est réalisée non plus par un être humain mais par un programme informatique. Dans ce cadre, la pièce à transcrire se présente comme un fichier son – de type .wav ou .mp3 par exemple – et la transcription générée prend la forme d'un fichier MIDI ou équivalent, approprié pour la représentation et le stockage de l'information extraite.

Quelles informations rechercher ?

La transcription automatique s'inscrit dans le cadre plus vaste de la recherche d'informations musicales, domaine de recherche à part entière, consacré en anglais sous l'appellation *Music Information Retrieval* (MIR).

Le support de la recherche d'informations musicales est le son produit par l'exécution d'un morceau. Ce son, émis *à un instant donné*, est enregistré ou directement traité. Il provient d'une ou plusieurs *sources sonores* et s'est propagé dans un *milieu physique*, le plus souvent l'air, et dans un *environnement*, salle de concert ou autre. La source sonore, un instrument de musique par exemple, est contrôlée – jouée – par un être humain, lui-même *interprétant* une *œuvre*, composée au préalable ou improvisée. Cette description, probablement restrictive, propose une approche sous la forme d'une chaîne de production dont chaque élément apporte sa propre contribution au résultat sonore et recèle des informations spécifiques pouvant faire l'objet de recherches en MIR. De la reconnaissance des instruments à la localisation des sources, en passant par la classification par genre musical pour la génération de listes de progammation (les fameuses *playlists*), chaque application s'attache à isoler les informations qui lui sont spécifiques, enfouies dans le support commun qu'est le son enregistré.

Les informations recherchées dans le cadre de la transcription automatique forment un champ qui se situe au niveau des sources sonores et qui comprend les notes jouées, avec leurs rythmes, leurs hauteurs, leurs nuances, leur articulation, l'instrument dont elles proviennent, les accords ou encore le suivi du tempo.

Les enjeux de la transcription automatique

Quelles applications pour la transcription ?

Si la transcription automatique présente un certain intérêt lié à la transcription humaine (pédagogie musicale, relevé automatique de solos de jazz, interaction musicien-ordinateur) et permet de réaliser des programmes de conversion audio/partition ou audio/MIDI, son utilité dépasse largement le cercle des musiciens. Son développement est en effet au cœur de celui du traitement du signal audio et des travaux en MIR. Le volume toujours croissant des fichiers audio a depuis quelques années établi le besoin de recourir à une indexation automatique plutôt que manuelle, la transcription automatique constituant alors un élément de base parmi les tâches développées. Elle intervient notamment dans des applications

comme l'extraction de mélodie ou du contenu tonal, l'identification des instruments, de la pièce, du style ou du compositeur, la détection des structures des morceaux (introduction, refrain, couplets, grilles, etc.) et la génération de résumés sonores, le suivi d'une partition en temps réel et son alignement sur un flux audio.

Alors que certaines d'entre elles font déjà partie de notre quotidien, la plupart de ces applications constituent des domaines de recherche très actifs, du fait de performances actuelles largement perfectibles et surtout de vastes perspectives d'innovations encore non exploitées, la plus convoitée (et générale) étant probablement la recherche et la navigation par le contenu dans de grandes bases de sons.

Quelle(s) représentation(s) pour la transcription ?

En fonction de l'objectif de la transcription, la nature des informations recherchées change ainsi que leur représentation. Les systèmes de notation musicale permettent de dresser une partition. Ils ont évolué au cours de l'histoire, depuis les premières traces de notations mélodiques et rythmiques antiques et les neumes du Moyen-Âge jusqu'à la notation classique, relativement récente et qui fait toujours l'objet d'innovations. L'avènement de l'informatique musicale a fait naître à la fois des formats spécifiques à l'édition de partition et des formats comme MIDI et SMF (Standard MIDI File) qui constituent également une représentation possible, avec comme avantage la communication avec d'autres instruments électroniques, à l'origine même de la norme. De son côté, l'indexation des signaux représente l'information sous forme codée, bien souvent à l'aide de *tags* ou d'annotations établis selon un language ou une norme donnés. On voit ainsi se développer des systèmes dédiés à la représentation du contenu sonore s'appuyant sur MPEG7 ou sur XML.

Problématique et structure du document

Nos travaux de thèse ont pour thème la transcription automatique de la musique pour piano solo. Le piano a la double particularité d'occuper une place majeure d'instrument solo dans le répertoire de musique occidentale, qu'elle soit classique ou de jazz, et d'être un instrument particulièrement difficile à transcrire par les systèmes actuels. La question de la transcription automatique en général étant elle-même une problématique vaste et difficile à traiter, nous essaierons de voir quels aspects spécifiques de l'instrument peuvent être modélisés et dans quelle mesure il peut être avantageux de spécialiser la tâche de transcription à un instrument comme le piano. Par ailleurs, nous aurons également l'occasion d'aborder des thématiques qui dépassent le cas du piano et sont liées à la transcription automatique en général, comme l'estimation de fréquences fondamentales multiples ou l'évaluation des résultats.

Ce mémoire de thèse est divisé en six chapitres inter-dépendants, accompagnés de quelques annexes. Nous y développerons ces motivations et aborderons plusieurs aspects liés à la transcription : nous proposerons en particulier une paramétrisation des diverses caractéristiques spectrales des sons de piano ; une méthode d'estimation de fréquences fondamentales simples performante lorsque les conditions d'analyse sont difficiles ; un modèle et une méthode d'estimation de fréquences fondamentales multiples ; un système de transcription effectif pour des enregistrements réels ; et enfin une réflexion et des travaux sur l'évaluation des transcriptions.

Le **chapitre 1** dressera un état de l'art des travaux liés à la transcription automatique. Nous y proposerons en prélude une discussion et une présentation des motivations de nos

travaux, établissant ainsi les spécificités de notre approche. L'état de l'art à proprement parler comportera ensuite quatre parties principales : la présentation de la problématique de l'estimation de hauteur, avec son ancrage dans la perception et les principes d'estimation les plus répandus ; la description des approches d'estimation de fréquences fondamentales multiples, module souvent central dans la transcription automatique ; un panorama des nombreux systèmes de transcription automatique proposés depuis une trentaine d'années ; enfin, dans le cadre de la transcription automatique de piano, un aperçu de la physique de l'instrument et des systèmes de transcriptions déjà proposés. Le chapitre se terminera par une série de questions qui posent notre problématique et auxquelles nous souhaiterions répondre.

Le **chapitre 2** abordera la question de la **caractérisation spectrale de sons de piano** pour la transcription. Après avoir présenté quelques modèles sinusoïdaux utiles par la suite pour décrire le contenu des sons, nous nous intéresserons à deux aspects spécifiques du piano – la distribution inharmonique des fréquences de ses partiels et la modélisation de l'enveloppe spectrale des notes – et à la modélisation du bruit. Nous proposerons alors des modèles et des algorithmes adaptés à l'instrument et à la tâche de transcription.

Dans le **chapitre 3**, nous nous intéresserons à **l'estimation de fréquences fondamentales** dans un contexte particulier rencontré avec le piano : la double contrainte d'une fenêtre d'analyse courte et d'un registre étendu. De manière générale, l'utilisation d'une petite fenêtre d'analyse est un défi lorsque l'on analyse des signaux audio, qu'ils soient de musique ou de parole, en raison de leur pseudo-stationnarité et du compromis temps-fréquence auquel on est rapidement confronté. Ces **conditions d'analyse difficiles** sont largement réunies dans le cas du piano où la musique peut être véloce, et les notes alors très courtes, alors que les fréquences fondamentales s'étendent, du grave à l'aigu, sur une des plus grandes tessitures que l'on puisse rencontrer. Dans ce cadre, nous proposerons une méthode d'estimation de notes isolées qui offre des performances satisfaisantes sur l'ensemble de la tessiture (fréquences fondamentales comprises entre $27,5$ et 4200 Hz, soit $7^1/_4$ octaves) en utilisant une fenêtre d'analyse de 60 ms (contre 93 ms en général).

Dans le **chapitre 4**, nous développerons une méthode d'**estimation de fréquences fondamentales multiples** reposant sur des modèles paramétriques d'enveloppe spectrale des notes (modèle autorégressif) et du bruit (modèle à moyenne ajustée). Ces modèles s'intègrent dans un cadre statistique à partir duquel nous proposons une résolution à base de maximum de vraisemblance. La technique proposée repose sur une estimation jointe des fréquences fondamentales multiples dans le domaine spectral et comporte l'estimation du degré de polyphonie (nombre de notes).

Dans le **chapitre 5**, nous proposerons un système complet de transcription automatique de la musique de piano. Nous y adopterons une stratégie de transcription liée aux spécificités du piano et de son répertoire. Il en résultera un cadre dans lequel le signal est segmenté en fonction des attaques détectées. Chaque segment sera ensuite analysé grâce à un modèle de Markov caché dans lequel la méthode d'estimation de fréquences fondamentales multiples du chapitre 4 est utilisée. Le système de transcription permettra d'analyser tout morceau de piano enregistré dans des conditions usuelles, tous styles confondus, avec des limites raisonnables, que nous détaillerons, en termes de polyphonie, de vélocité, et de tessiture.

Le **chapitre 6** sera consacré à l'évaluation des transcriptions automatiques. La question a été approfondie selon deux axes. Nous nous intéresserons tout d'abord aux méthodes d'évaluation, et proposons en particulier un raffinement des métriques usuelles. L'étude partira du constat des limites des systèmes d'évaluation courants et s'appuiera sur les résultats

d'un test perceptif pour proposer une pondération des erreurs typiques. Nous verrons également comment déterminer dans quelle mesure une métrique d'évaluation donnée remplit effectivement sa fonction. Dans la deuxième partie du chapitre, nous détaillerons le contenu d'une base de données que nous avons constituée spécifiquement pour l'estimation de fréquences fondamentales multiples et la transcription automatique de la musique de piano. Elle regroupera des séries de notes isolées, d'accords aléatoires et d'accords typiques, et de morceaux de musique, en faisant varier plusieurs paramètres tels que les nuances, les durées, ou l'utilisation de la pédale *forte*. Les enregistrements proviennent d'un piano ayant un dispositif d'entrée et sortie MIDI et de logiciels de synthèse de qualité, permettant dans les deux cas de disposer de références très précises sur le contenu des fichiers. Dans la dernière partie de ce chapitre, nous évaluerons notre algorithme d'estimation de fréquences fondamentales multiples et notre système de transcription et proposerons une analyse détaillée et comparative des résultats.

Pour terminer ce mémoire de thèse, notre conclusion établira un bilan de nos contributions avant de proposer quelques perspectives.

Chapitre 1

État de l'art

1.1 Motivations et questions

La transcription automatique constitue un domaine de recherche actif et compétitif, comme le montre le nombre de thèses [Moorer, 1975; Maher, 1989; Mellinger, 1991; Rossi, 1998; Godsmark, 1998; Sterian, 1999; Marolt, 2002; Ortiz-Berenguer, 2002; Bello, 2003; Klapuri, 2004; Vincent, 2004; Cemgil, 2004; Virtanen, 2006; Camacho, 2007; Kameoka, 2007] et d'ouvrages [Klapuri et Davy, 2006; Wang et Brown, 2006] qui s'y consacrent sous une forme ou une autre, et dont les points de vue pourront compléter celui exposé ici. Cet état de l'art a la double vocation de décrire les travaux liés à la transcription automatique de la musique et d'introduire les problématiques que nous développerons dans ce mémoire de thèse. Nous proposons en prélude à l'état de l'art quelques questions et discussions ayant pour objectif l'établissement d'une grille de lecture d'une bibliographie prolifique.

1.1.1 La transcription : information, variables et organisation

Considérons la transcription comme une tâche de recherche d'information, passage des données brutes, la suite d'échantillons d'un enregistrement audio numérique, à leur description symbolique par un fichier MIDI ou une partition. Les entités porteuses de sens musical – notes, instruments, tempo, tonalité, métrique – se caractérisent soit par une fonction du temps (pour le tempo par exemple), soit par l'ensemble des paramètres d'une note considérée comme un objet (fréquence fondamentale, instant d'attaque et d'extinction, nuance). Cette couche d'information peut être qualifiée de *principale* dans la mesure où son extraction constitue la finalité de la transcription. S'ajoute un ensemble de notions permettant de la caractériser : les fréquences des partiels, l'enveloppe spectrale d'une note, la réponse de la salle, le bruit de fond. Ces notions constituent une couche d'information *auxiliaire*, données *latentes* utiles à la transcription, voire nécessaires, mais n'apparaissant pas dans le résultat.

La question de la hiérarchie des informations pour l'analyse de sons est également abordée à travers la notion voisine des représentations mi-niveau (ou intermédiaires). La problématique consiste alors à introduire un niveau intermédiaire de représentation entre une représentation bas-niveau (forme d'onde, représentations temps-fréquence, temps-échelle, cepstre, etc.) et l'information recherchée afin de diviser le problème en tâches plus simples. Le choix de cette représentation est alors au centre de la problématique : certaines transformées élaborées [Ellis et Rosenthal, 1995], les chromagrammes [Shepard, 1964; Lee, 2008] ou encore les atomes et molécules associés à des instruments [Leveau *et al.*, 2008] sont

autant de représentations possibles pour sonder cette hiérarchie d'informations.

1.1.2 Quelles méthodes pour la transcription ?

Le choix d'une stratégie d'analyse des signaux dépend à la fois de l'application visée et de la facilité et de la robustesse de l'estimation de chaque type d'information. Comme nous le verrons dans ce chapitre, les techniques de transcription sont influencées par deux pôles :

- l'estimation de hauteur : les hauteurs peuvent être estimées à chaque instant ou trame de signal, et les notes sont ensuite extraites en fonction des résultats obtenus dans les trames successives. L'idée sous-jacente à cette approche est que l'on peut se fier essentiellement à l'information instantanée (de l'ordre d'une trame) pour estimer la hauteur des notes présentes, en prenant ensuite une décision sur l'ensemble de ces estimations. Comme argument en faveur de ce postulat, on peut évoquer les processus de perception qui permettent souvent de reconnaître des notes lorsqu'elles sont tronquées ou, de manière causale, avant qu'elles soient achevées.
- les modèles de notes : on ne se contente plus d'un simple modèle de hauteur mais d'un modèle de note, souvent associé à une technique de décomposition (approches bayésiennes, dictionnaires de formes d'ondes). Les dimensions temporelle et fréquentielle sont alors conjointement considérées afin de rendre l'approche plus robuste. Celle-ci est par ailleurs non causale, puisque l'on a besoin de considérer tout le signal pour identifier une note.

Nous verrons donc quelles stratégies peuvent être adoptées en nous intéressant en particulier au rôle de l'estimation de hauteur, aux modèles de notes ainsi qu'aux connaissances *a priori* et à leur apprentissage éventuel.

1.1.3 Pourquoi se restreindre au piano ?

Pourquoi élaborer un système de transcription spécifique au piano plutôt qu'un système de transcription générique ?

Plusieurs raisons justifient cette démarche :

1. en rendant plus spécifique la tâche à accomplir, on peut espérer obtenir des solutions plus spécialisées et donc potentiellement plus efficaces ;
2. la tâche n'est cependant pas plus aisée : le piano est un des instruments les plus difficiles à transcrire, notamment en raison de la forte polyphonie, de la complexité des sons de piano, de la taille du registre ($7^1/_4$ octaves), de la technicité des pièces et de la virtuosité des interprètes. Nous verrons comment les multiples difficultés liées au piano constituent un enjeu scientifique de taille ;
3. l'enjeu est également musical, étant données la place de cet instrument dans la musique occidentale et la taille du répertoire pour piano solo (enrichi des réductions de pièces orchestrales et autres adaptations).

Par ailleurs, on peut considérer la transcription automatique comme un regroupement de problématiques d'extraction d'information musicale. La transcrition du piano trouve alors sa place parmi d'autres tâches de transcription partielle :

- transcription de batterie ou des percussions [Goto et Muraoka, 1994b; Gillet, 2007] ;
- transcription de la mélodie [Goto, 2004; Ellis et Poliner, 2006; Ryynänen et Klapuri, 2008] ;
- transcription de la ligne de basse [Goto, 2004; Ryynänen et Klapuri, 2008] ;

– transcription d'accords [Lee, 2008; Ryynänen et Klapuri, 2008] ;
– transcription d'ornements [Gainza et Coyle, 2007] ;
– estimation de la tonalité [Lee, 2008] ;
– estimation de la pulsation, du tempo et de la mesure [Goto et Muraoka, 1994a; Klapuri *et al.*, 2006; Alonso *et al.*, 2007] ;
– reconnaissance des instruments [Eronen et Klapuri, 2000; Essid *et al.*, 2006].

1.2 Estimation de hauteur

Afin d'aborder la question de l'estimation de hauteurs multiples – problématique centrale pour la transcription –, nous nous intéressons ici à la perception et à l'estimation de hauteur dans le cas d'une hauteur unique. Deux motivations expliquent cette démarche. Premièrement, la **perception de la hauteur** des sons fait intervenir suffisamment de phénomènes pour être bien souvent étudiée dans le cas d'une seule hauteur. Le processus de discrimination de hauteurs simultanées s'appuie ensuite sur ces résultats. Deuxièmement, **l'estimation de hauteurs simples** a fait l'objet de nombreux travaux, dont nous verrons quelques aspects.

1.2.1 Perception de la hauteur

La hauteur est généralement définie comme l'une des quatre caractéristiques d'un son avec son intensité, sa durée et son timbre, la hauteur étant la grandeur relative à l'échelle grave-aigu. On distingue deux types de sensations de hauteur : la hauteur tonale et la hauteur spectrale. La **hauteur tonale** est associée aux sons périodiques ou quasi-périodiques, et est caractérisée par sa *période* ou sa **fréquence fondamentale** (F_0), inverse de la période. L'oreille a la capacité de quantifier très précisément cette grandeur, dont la valeur numérique est exprimée sur une échelle physique, en hertz (c'est précisément la fréquence fondamentale), ou sur l'échelle subjective de tonie, c'est-à-dire de sensation de hauteur, dont l'unité est le mel. La **hauteur spectrale** désigne une sensation de hauteur correspondant au centre de gravité spectral, et ne se restreignant pas à la classe des sons quasi-périodiques. La sensation grave/aigu peut par exemple être attribuée à des bruits tels qu'un souffle, un choc sur divers matériaux, un frottement. Dans certaines conditions, il peut arriver que les deux sensations de hauteur entrent en concurrence au point de faire apparaître des ambiguïtés ou des illusions auditives [Shepard, 1964; Houtsma *et al.*, 1988]. Dans le cadre de la transcription automatique, nous ne nous intéresserons qu'à la hauteur tonale, que nous appellerons simplement *hauteur* dans la suite du document, et qui est le phénomène de hauteur le plus répandu dans la musique, en particulier pour le piano.

La perception de la hauteur tonale s'explique en s'appuyant sur la physiologie de l'oreille. L'onde sonore se propage dans l'air au niveau de l'oreille externe et fait vibrer le tympan. Celui-ci agit comme un transducteur en transformant l'onde de pression en vibration mécanique. Elle est ensuite transmise dans l'oreille moyenne par la chaîne ossiculaire, composée des trois osselets – marteau, enclume, étrier –, à l'extrémité de laquelle se trouve la fenêtre ovale. Un « pattern d'excitation » dépendant du contenu spectral du signal se forme alors le long de la membrane basilaire à l'intérieur de la cochlée, organe de forme hélicoïdale rempli d'un liquide appelé endolymphe. Les cellules ciliées, portées par la membrane basilaire, transforment l'excitation mécanique en impulsions électriques transmises aux fibres nerveuses. On considère généralement que la perception de la hauteur tonale résulte de la contribution de deux phénomènes : le codage tonotopique et le codage

temporel. Le codage tonotopique consiste en la génération du pattern d'excitation dans la cochlée, véritable analyse spectrale sous forme spatiale, qui permet de sélectionner les fibres nerveuses à solliciter en fonction du contenu fréquentiel. Le codage temporel correspond au signal nerveux transmis par ces fibres, dont la cadence et la forme varient en fonction du contenu fréquentiel. Le cerveau exploite alors l'information sur les fibres excitées et sur la nature de l'excitation transmise pour générer la sensation de hauteur tonale.

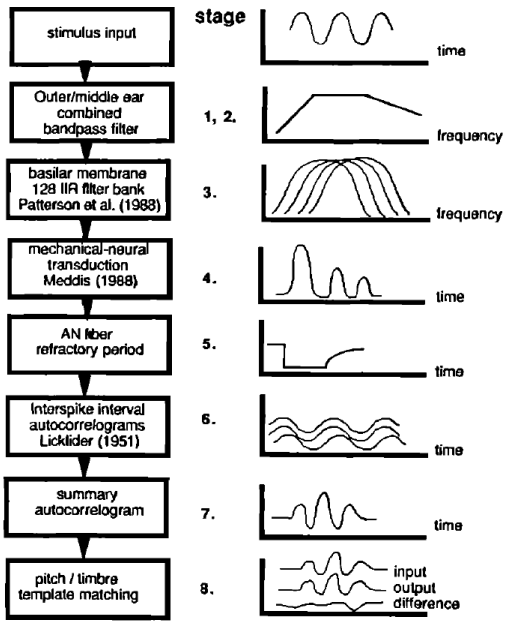


FIGURE 1.1 – Modèle de Meddis et Hewitt (Source : Meddis et Hewitt [1991a])

Le modèle de Meddis et Hewitt [1991a,b] établit un lien entre les connaissances sur la perception de la hauteur et les méthodes de traitement du signal sur l'estimation de hauteur. Pour concilier les deux approches, les auteurs proposent une modélisation de la chaîne perceptive par une suite de blocs fonctionnels qui possèdent à la fois un sens physiologique et une réalisation possible par des techniques de traitement de signal (cf. figure 1.1). Les deux premiers étages du système modélisent l'oreille externe et l'oreille moyenne par plusieurs filtrages élémentaires correspondant à la réponse au niveau du tympan (filtrage passe-bande entre 2 et 5 kHz) et à celle de l'oreille moyenne (filtrage passe-bande asymétrique autour de 1 kHz). L'étage suivant est un banc de filtres auditifs modélisant le phénomène de décomposition tonotopique au niveau de la membrane basilaire. Un modèle de cellule ciliée simule ensuite la transduction neuronale du signal mécanique en décharge électrique ainsi que l'effet d'inhibition des neurones. Les écarts entre décharges sont alors quantifiés par une fonction d'autocorrélation (ACF) du signal, qui est finalement sommée sur tous les canaux pour obtenir une unique fonction de détection, dont le maximum correspond à la hauteur perçue. Le modèle est finalement appliqué à plusieurs situations particulières dans lesquels il s'avère robuste : phénomène de fondamentale absente, ambiguïtés de hauteur, modulations, hauteur créée à partir de bruit, etc.

Ces résultats expliquent la place importante que peuvent prendre un pré-filtrage, un banc de filtres et une autocorrélation dans les sous-bandes, que nous retrouvons dans la littérature sur l'estimation de hauteur. Le modèle de Meddis et Hewitt a par la suite été souvent repris dans les travaux sur l'estimation de hauteur, soit en s'en inspirant [Tolonen et Karjalainen, 2000; de Cheveigné et Kawahara, 2002], soit en le reprenant dans sa quasi-totalité [Klapuri, 2008]. On remarque cependant que la théorie expliquant la perception de la hauteur par l'autocorrélation ne fait pas forcément consensus, des travaux ayant été menés dans d'autres directions [de Cheveigné, 1998].

1.2.2 Méthodes d'estimation

L'estimation de hauteur simple (monopitch) consiste à estimer la fréquence fondamentale d'un son périodique ou quasi-périodique. Le problème a été maintes fois étudié pour les signaux de musique, mais surtout pour ceux de parole [Rabiner *et al.*, 1976; Hess, 1983;

de Cheveigné et Kawahara, 2002]. Dans ce cas, l'estimation de la fréquence fondamentale est étroitement liée à la détection de voisement, c'est-à-dire la présence de fréquence fondamentale, conséquence de la vibration de cordes vocales, dans les divers segments de paroles [Atal et Rabiner, 1976; McAulay et Quatieri, 1990; Rouat *et al.*, 1997; Thomson et Chengalvarayan, 2002], voire à la localisation des instants de fermeture glottique [Strube, 1974; Cheng et O'Shaughnessy, 1989; Tuan et d'Alessandro, 1999; Kawahara *et al.*, 2000].

La grande majorité des techniques d'estimation de hauteur se répartissent en deux classes : les approches dans le domaine spectral et les approches dans le domaine temporel. Dans le domaine spectral, le son présente des pics répartis suivant une distribution quasi-harmonique : il s'agit alors d'estimer ce peigne pour obtenir la fréquence fondamentale. Dans le domaine temporel, c'est plutôt la période fondamentale, grandeur équivalente, qui est estimée, en cherchant le plus petit décalage temporel non nul pour lequel la forme d'onde et sa version décalée coïncident.

Les travaux de Schroeder [1968] font historiquement référence. L'auteur propose deux approches, temporelle et spectrale, chacune dans une version simple, l'histogramme, et dans une version plus élaborée, l'histogramme pondéré. L'approche spectrale commence par une étape de détection des pics spectraux. La fréquence fondamentale étant un sous-multiple de la fréquence de chaque pic, on l'estime comme étant la fréquence rassemblant le plus grand nombre de sous-multiples de fréquences détectées. Pour ce faire, on construit un histogramme en fonction de la fréquence, en dénombrant le nombre de sous-multiples des fréquences de chaque pic. Le maximum de l'histogramme correspond alors à la fréquence fondamentale (cf. figure 1.2). Dans le domaine temporel, l'histogramme des périodes est construit de manière analogue, en estimant le plus petit multiple commun des périodes des pics détectés. Pour aller plus loin, l'auteur propose de remplacer le simple dénombrement par une pondération des contributions des partiels en fonction de leurs amplitudes, et de généraliser la notion d'histogramme en supprimant la détection de pics pour obtenir une fonction « continue ». Il en résulte en particulier deux fonctions de détection de hauteur dans le domaine spectral :

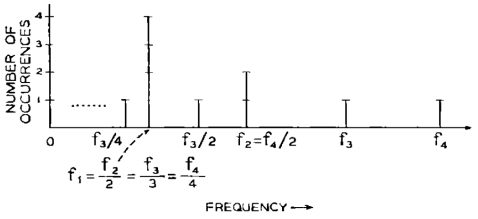


FIGURE 1.2 – Exemple d'histogramme des fréquences (Source : Schroeder [1968]).

– la **somme spectrale** $S(f)$, dans le cas d'une pondération par les amplitudes des H premiers partiels à la fréquence fondamentale f :

$$S(f) \triangleq 20 \log \sum_{h=1}^{H} |X(hf)| \tag{1.1}$$

– le **produit spectral** $P(f)$, dans le cas d'une pondération par le logarithme des amplitudes :

$$P(f) \triangleq \sum_{h=1}^{H} 20 \log |X(hf)| \tag{1.2}$$

$$= 20 \log \prod_{h=1}^{H} |X(hf)| \tag{1.3}$$

où X est la transformée de Fourier discrète du signal. Ces méthodes sont particulièrement efficaces lorsque l'on peut fixer le nombre H de partiels, c'est-à-dire lorsque la plage de variation de f est relativement restreinte.

Il est intéressant de remarquer que bon nombre de méthodes développées depuis aboutissent à des variantes de la somme spectrale ou du produit spectral et que ces méthodes s'appuient parfois sur des concepts très différents de celui d'histogramme pour aboutir à un résultat similaire. On peut ainsi considérer les estimateurs de fréquence fondamentale à base de reconnaissance de motifs spectraux (« pattern matching ») comme dans les travaux de Brown [1992] ou de Doval et Rodet [1991] : la somme pondérée prend la forme d'un produit scalaire, les poids correspondant à un spectre de référence. De même, l'estimateur de Klapuri [2005] conçu à partir du modèle perceptif de Meddis et Hewitt [1991a] est une fonction dite de saillance qui s'écrit

$$\lambda(\tau) = \frac{f_s}{\tau} \sum_{j=1} \max_{k \in \kappa_{j,\tau}} \left(H_{\mathrm{LP}}(k)\, U(k) \right) \tag{1.4}$$

où τ est la période fondamentale, f_s la fréquence d'échantillonnage et où les différents coefficients d'indice k du spectre prétraité $U(k)$ sont sélectionnés sur un ensemble $\kappa_{j,\tau}$, voisinage du partiel d'ordre j, et ajoutés avec une pondération $H_{\mathrm{LP}}(k)$. Cet exemple montre le cas d'une fonction temporelle de type autocorrélation qui peut-être facilement interprétée comme une variante de la somme spectrale.

La somme spectrale, calculée cette fois sur le module au carré du spectre, est également obtenue comme l'estimateur du maximum de vraisemblance en considérant un modèle de signal composé de la somme d'un signal périodique $s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t)$ et d'un bruit blanc gaussien centré $w(t)$:

$$x(t) = s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t) + w(t) \tag{1.5}$$

avec

$$s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t) = \sum_{h=1}^{H} 2a_h \cos\left(2\pi h f_0 t + \varphi_h\right) \tag{1.6}$$

Les paramètres du modèle sont les amplitudes réelles positives $\mathbf{a} \triangleq (a_1, \ldots, a_H)$ et les phases initiales $\boldsymbol{\varphi} \triangleq (\varphi_1, \ldots, \varphi_H)$ sur $[0; 2\pi[$ des composantes, la fréquence fondamentale f_0, et la variance σ_w^2 du bruit. Si l'on observe le signal sur un nombre entier de périodes $T = \frac{m}{f_0}$, la log-vraisemblance du vecteur $\mathbf{x} \triangleq (x(0), \ldots, x(T-1))$ d'observations est alors

$$L\left(\mathbf{x}|\mathbf{a}, \boldsymbol{\varphi}, f_0, \sigma_w^2\right) = -\frac{T}{2} \log\left(2\pi\sigma_w^2\right) - \frac{1}{2\sigma_w^2} \sum_{t=0}^{T-1} \left(x(t) - s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t)\right)^2 \tag{1.7}$$

On montre alors (cf. annexe B.1 (p. 173)) que la solution au sens du maximum de vraisemblance revient à maximiser la fonction

$$f_0 \mapsto \sum_{h=1}^{H} \left|X(h f_0)\right|^2 \tag{1.8}$$

Dans le domaine temporel, l'autocorrélation du signal, dont une étude de référence a été établie par Rabiner [1977], est la méthode temporelle élémentaire pour l'estimation de la

hauteur, étant donnée sa place dans l'explication de la perception de la hauteur. Wise *et al.* [1976] justifient par ailleurs son utilisation via le principe de maximum de vraisemblance d'une façon très similaire à celle présentée précédemment dans le cas spectral.

Comme dans le cas précédent avec la somme spectrale et le produit spectral, de nombreuses méthodes temporelles peuvent être interprétées comme des variantes de l'autocorrélation. C'est en particulier le cas de plusieurs approches qui partent de l'autocovariance comme transformée de Fourier inverse du périodogramme, et qui proposent de substituer d'autres fonctions au périodogramme. En prenant le logarithme du module de la transformée de Fourier, on obtient ainsi le cepstre [Noll, 1967], alors qu'en modifiant son exposant, quadratique à l'origine, Indefrey *et al.* [1985] et Tolonen et Karjalainen [2000] étudient les conséquences d'une telle compression. Ross *et al.* [1974] comparent l'autocorrelation et l'AMDF (Average Magnitude Difference Function), de Cheveigné et Kawahara [2002] montrent ensuite la relation qui les unit et développent l'estimateur de hauteur YIN, et Klapuri [2005, 2008] s'appuie sur le principe d'autocorrélation en modifiant complètement l'étape d'inversion de la transformée de Fourier.

Pour conclure cette partie, nous pouvons nous pencher sur les limites d'une classification des méthodes d'estimation de hauteur en deux classes, les méthodes temporelles et spectrales. Considérons une fonction temporelle $g(t)$ comme l'autocorrélation ou le cepstre, s'exprimant comme la tranformée de Fourier inverse d'une quantité spectrale $S(f)$, qui est l'estimée du périodogramme pour l'autocorrélation ou le logarithme de l'amplitude du spectre pour le cepstre : on a $g(t) = \mathcal{F}^{-1}[S(f)]$. On peut alors dire que $g(t)$ mesure donc la présence de pics régulièrement espacés dans $S(f)$, ou encore que, selon une vision « pattern matching », $g(t)$ mesure la similarité de $S(f)$ avec le motif $e^{2i\pi ft}$ périodique de période $\frac{1}{t}$. On peut ainsi avoir des interprétations fréquentielles de méthodes temporelles. Cette vision unificatrice de l'approche temporelle et fréquentielle reste discutable, car s'il y a des cas où elle s'applique très bien, comme celui de Klapuri [2005] dont nous avons que la méthode temporelle s'interprète aussi sous forme spectrale, il existe aussi des moyens de tirer parti d'une méthode temporelle et d'une méthode fréquentielle de façon complémentaire [Peeters, 2006].

1.3 Estimation de fréquences fondamentales multiples

Lorsque plusieurs notes sont jouées simultanément, l'oreille est capable d'appréhender le son dans sa globalité ou d'isoler les notes qui le composent. L'écoute synthétique dans le premier cas et l'écoute analytique dans le second sont utilisées dans divers contextes de la vie courante : lors d'un concert (focalisation sur un instrument ou appréciation du tout), lors de l'effet « cocktail party » (séparation de la voix d'un interlocuteur parmi plusieurs), etc. L'objet de l'estimation de hauteurs (ou fréquences fondamentales) multiples est de réaliser cette séparation/identification des fréquences fondamentales simultanées. Si l'estimation de hauteurs simples a plutôt été étudiée pour la parole, le problème de l'estimation de hauteurs multiples a fait l'objet de moins de travaux dans ce domaine [Wu *et al.*, 2003; Le Roux *et al.*, 2007]. La problématique concerne davantage les signaux musicaux étant donné le caractère polyphonique de la musique.

Un son comportant plusieurs hauteurs est un mélange additif de sons à hauteur simple. Étant donnée la simplicité de cette relation entre sons à hauteur unique et sons à hauteurs multiples, la problématique de l'estimation de fréquences fondamentales multiples peut au premier abord sembler très proche de celle de l'estimation de hauteurs uniques. Cette simplicité apparente est démentie dès que l'on essaie de poser le problème. Ainsi, si détecter

une hauteur unique revient à détecter la plus petite périodicité d'un signal, on ne peut transposer la méthode dans le cas de hauteurs multiples, détecter « plusieurs périodicités » n'ayant pas de sens. De même, d'un point de vue spectral, estimer la fréquence fondamentale d'un peigne harmonique ne présente pas d'ambiguïté, alors qu'une somme de peignes harmoniques peut être associée à plusieurs ensembles de fréquences fondamentales (on a par exemple la liberté d'ajouter les octaves) et donne lieu au mieux à un problème arithmétique non trivial [Klapuri, 1998], et au pire, dans le cas de l'octave par exemple, à un problème mal posé. Ainsi, deux difficultés liées aux rapports harmoniques entre fréquences fondamentales se conjuguent dans le cas de l'estimation de fréquences fondamentales multiples. La première, héritée du cas monophonique, est la tendance à confondre la véritable fréquence fondamentale avec les fréquences fondamentales en rapport harmonique. La seconde s'ajoute à la première dans le cas polyphonique : il s'agit non seulement de pouvoir choisir entre deux notes (ou davantage) en rapport harmonique, mais également d'être capable de déterminer si les deux notes sont présentes simultanément.

Il apparaît donc que poser le problème de l'estimation de fréquences fondamentales multiples est une tâche délicate qui consiste à caractériser le mélange obtenu à partir de plusieurs notes avec toutes les ambiguïtés que l'opération de mélange peut introduire, et avec une nouvelle dimension à traiter, le degré de polyphonie. Nous allons maintenant voir comment les nombreux travaux qui s'y sont consacrés abordent la question, en continuant de s'appuyer sur le caractère harmonique des notes, et en exploitant plus en profondeur d'autres informations, en particulier la notion d'enveloppe spectrale.

1.3.1 Estimation itérative des fréquences fondamentales

Dans le cas d'un mélange polyphonique, il est parfois aisé de trouver une note prédominante parmi toutes celles présentes, en utilisant par exemple un critère énergétique. Si l'on parvient à soustraire cette note du mélange, le résiduel contient théoriquement une note de moins que l'original, et l'on peut à nouveau chercher la nouvelle note prédominante. D'où le principe des méthodes itératives [Klapuri, 2003; Karjalainen et Tolonen, 1999; Ortiz-Berenguer et Casajús-Quirós, 2002; Yeh *et al.*, 2005; Klapuri, 2008] décrit par l'algorithme 1.1.

Entrées: signal audio polyphonique x
  $r \leftarrow x$ {Initialisation}
  Tant que r contient une note
    Sélectionner la fréquence fondamentale prédominante f_0 dans r
    Estimer le signal s correspondant à f_0
    $r \leftarrow r - s$
  Fin Tant que
Sorties: liste des f_0 successivement estimées

Algorithme 1.1: Méthodes itératives d'estimation de fréquences fondamentales multiples-principe général.

Une telle approche nécessite la réalisation de quatre tâches : la sélection d'une note prédominante dans le signal, l'estimation de sa contribution, sa soustraction au signal et l'évaluation de la condition d'arrêt. La sélection d'une note prédominante repose souvent sur un critère énergétique, qui correspond parfois à une méthode d'estimation conçue pour le cas monophonique. Par exemple, le maximum du produit spectral (cf. équation (1.3))

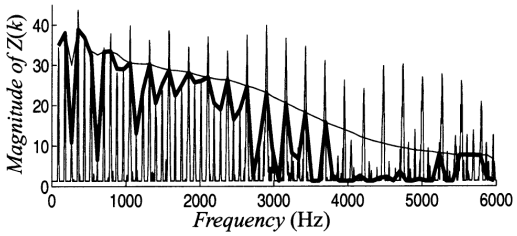


FIGURE 1.3 – *Spectral smoothness* : le lissage de l'enveloppe spectrale (trait épais) permet d'estimer la contribution d'une note (trait fin) (Source : Klapuri [2003]).

est un bon candidat. Une fois la fréquence fondamentale prédominante estimée, les autres tâches constituent le véritable défi introduit lors du passage de la monophonie à la polyphonie. L'estimation du signal correspondant à la fréquence fondamentale sélectionnée et sa soustraction se font en général approximativement, car le recouvrement spectral entre les notes ne permet pas de séparer complètement la contribution de la note à extraire du reste du mélange. Une solution consiste à exploiter l'information portée par **l'enveloppe spectrale**, courbe qui relie, dans le domaine fréquentiel, les amplitudes ou les énergies des partiels d'une note. Le principe de *spectral smoothness* (régularité de l'enveloppe spectrale) introduit par Klapuri [2003] fait référence. Il permet d'estimer cette contribution en s'appuyant sur le caractère régulier de l'enveloppe spectrale des notes de la majorité des instruments. Une hauteur n'est plus caractérisée uniquement en fonction de l'énergie des partiels pris individuellement, mais également selon une contrainte d'énergie relative entre partiels. De cette façon, en cas de recouvrement spectral, l'amplitude ou l'énergie d'un partiel est déterminée comme une valeur moyenne ou médiane des amplitudes des partiels d'ordres voisins. Les amplitudes ainsi estimées (cf. figure 1.3) sont alors soustraites. Yeh *et al.* [2005] reprennent ce principe et y ajoute une contrainte d'enveloppe temporelle régulière. Par ailleurs, la qualité du résiduel est améliorée par Klapuri [2008] qui le calcule à chaque itération en utilisant le signal original, plutôt que le résiduel obtenu à travers les itérations précédentes, et en lui soustrayant les contributions réestimées de l'ensemble des notes extraites successivement. Une soustraction partielle permet également de ne pas soustraire totalement la contribution estimée, afin d'éviter que le résiduel soit trop « creusé ».

Enfin, la condition d'arrêt est en général difficile à élaborer et fait intervenir un seuil sur l'énergie ou le rapport signal à bruit dans le résiduel (cf. partie 1.3.3).

1.3.2 Estimation jointe des fréquences fondamentales

L'approche jointe est abordée par de Cheveigné [1993], par de Cheveigné et Kawahara [1999] et par Tolonen et Karjalainen [2000]. La méthode consiste à utiliser des filtres en peigne qui suppriment les partiels distribués suivant la distribution harmonique correspondant aux fréquences fondamentales supposées, puis à calculer l'énergie du résiduel. Elle est minimale lorsque les fréquences fondamentales ont été correctement choisies. C'est une façon de généraliser une méthode comme l'AMDF [Ross *et al.*, 1974] qui a été conçue pour le cas monophonique et ne peut pas être utilisée pour sélectionner un candidat prédominant dans le cas polyphonique.

Lors d'une estimation de type itératif, l'extraction du candidat prédominant est ap-

proximative, comme le montrent les taux de soustraction introduits [Klapuri, 2003, 2008]. Le résiduel obtenu à chaque itération est donc imparfait : il peut comporter des composantes parasites provenant de candidats mal soustraits ou au contraire avoir été privé de composantes lors d'une soustraction trop importante. Cela est sans doute inévitable lorsque l'on veut estimer un candidat prédominant en l'absence d'informations sur le résiduel. Estimer les hauteurs de façon conjointe peut alors sembler plus efficace.

Cependant, au premier abord, l'approche jointe ne peut être appliquée avec une polyphonie élevée : sa complexité est importante en raison du grand nombre de combinaisons de fréquences fondamentales à examiner. En effet, s'il y a N notes potentielles et une polyphonie maximale $P_{\max}$, l'estimation itérative consistera à chercher une note parmi N à chaque itération, impliquant au maximum $NP_{\max}$ évaluations, alors que l'estimation jointe devra tester toutes les combinaisons de $1, 2, \ldots, P_{\max}$ notes parmi N, soit $\sum_{p=0}^{P_{\max}} \binom{N}{p}$. Une telle complexité ($2^N$ dans le cas $P_{\max} = N$) reste abordable pour des polyphonies faibles [de Cheveigné et Kawahara, 1999] mais n'est pas envisageable dans le cas de la musique en général. Le cadre des approches bayésiennes (cf. partie 1.4.4 (p. 36)) offre des stratégies plus efficaces pour converger vers la solution. Une autre possibilité est d'estimer conjointement la contribution des N notes par moindres carrés [Bello *et al.*, 2006].

1.3.3 Estimation de la polyphonie

Seuls quelques travaux s'intéressent à l'estimation de la polyphonie, c'est-à-dire du nombre de notes présentes, comme tâche faisant partie intégrante de l'estimation de fréquences fondamentales multiples. Les auteurs préfèrent souvent s'intéresser aux performances en termes de reconnaissance de hauteur, en supposant le nombre de notes ou de sources connues. Il s'avère que l'estimation de la polyphonie est une tâche difficile, y compris pour un humain entraîné qui a tendance à sous-estimer le nombre de notes présentes [Klapuri, 2003].

Parmi les travaux qui présentent explicitement des techniques, citons ceux de Klapuri [2003, 2008] et de Yeh [2008] qui, dans le cadre d'une estimation itérative des fréquences fondamentales multiples, proposent un critère d'estimation ou une condition d'arrêt de nature énergétique sur le résiduel. D'autres méthodes présentent l'avantage de calculer implicitement la polyphonie comme une conséquence de la reconnaissance des notes, avec un seuil ou un nombre total de notes à fixer éventuellement [Smaragdis et Brown, 2003; Kameoka *et al.*, 2005].

1.4 Systèmes de transcription automatique

La transcription automatique de la musique a été introduite par les travaux de Moorer [1975], dans lesquels l'auteur définit en détail cette nouvelle problématique et propose le premier système de transcription. L'objectif théorique est de générer une partition, par un ordinateur, à partir d'un enregistrement musical. Un certain nombre d'hypothèses simplificatrices permettent alors de restreindre ce vaste cadre et de concevoir un système de transcription, les deux principales étant la limitation de la polyphonie à deux voix et l'absence de notes simultanées en rapport harmonique.

Ces travaux posent explicitement la problématique de la transcription et proposent différentes questions essentielles pour traiter la transcription automatique :

– la séparation du problème en deux étapes : dans un premier temps, le choix des techniques d'analyse dites de *bas-niveau*, menant à ce que l'on appelle aujourd'hui

les *représentations de mi-niveau* (cf. partie 1.1.1 (p. 23)). À travers cette première étape, la forme d'onde du signal est analysée pour obtenir une représentation mettant en valeur des informations caractéristiques du signal telles que son contenu fréquentiel. Dans un second temps, les techniques de *niveau intermédiaire* qui, à partir de l'analyse précédente, traitent le problème, c'est-à-dire l'estimation des notes dans le cas présent ;
- le problème des fréquences fondamentales en rapport harmonique et du recouvrement spectral entre notes ;
- la question de l'estimation des notes à partir de l'intégration temporelle du contenu fréquentiel du son ;
- la définition de la transcription automatique et des éléments à transcrire : de la note à l'instrument, en passant par la tonalité, la mélodie, le tempérament, les instruments percussifs, etc.

Les nombreux systèmes proposés après celui de Moorer ont traité le problème de la transcription automatique en essayant de supprimer progressivement les hypothèses restrictives de Moorer. Nous décrivons ici les idées principales de ces systèmes, dont la diversité des approches et des techniques nous a conduit à un classement en quatre catégories : les approches reposant sur une paramétrisation et des heuristiques pour la détection de fréquences fondamentales, les approches avec un apprentissage hors-ligne de modèles, les approches avec un apprentissage en ligne de modèles, et les approches bayésiennes.

1.4.1 Approches à base de paramétrisation et d'heuristiques

Ce type d'approche repose sur l'utilisation d'une technique d'estimation de fréquences fondamentales à partir d'une trame de signal. Cette première étape d'analyse s'apparente à une paramétrisation du signal, dans la mesure où l'on extrait une information partielle – par exemple une mesure de la saillance des notes dans la trame – sur laquelle les étapes suivantes s'appuieront, laissant de côté le signal proprement dit. La transcription est ensuite obtenue par une étape d'intégration temporelle de cette paramétrisation, qui est une analyse verticale, c'est-à-dire fréquentielle, du signal. Il s'agit de faire correspondre des notes à une suite de fréquences fondamentales à l'aide de critères tels que la variation de ces fréquences fondamentales dans les trames successives.

Ryynänen et Klapuri [2005] proposent un système qui illustre très bien ce type d'approche. Il utilise l'estimateur de fréquences fondamentales multiples de Klapuri [2005], qui fournit une *saillance* des fréquences fondamentales multiples estimées dans une trame. La paramétrisation d'une trame consiste alors à prendre les cinq fréquences fondamentales les plus saillantes, avec les valeurs de saillance associées, ainsi que les cinq fréquences fondamentales prédominantes pour les attaques (au sens de la dérivée temporelle de la saillance), avec les intensités d'attaques associées. Pour chaque note n et chaque trame t, un vecteur d'observations $\mathbf{o}_{n,t}$ est construit à partir des paramètres de la trame pour être utilisé dans un HMM (Hidden Markov Model, modèle de Markov caché). Le vecteur $\mathbf{o}_{n,t} \triangleq (\Delta x_{n,t}, s_{jt}, d_{n,t})$ est composé de trois paramètres : l'écart $\Delta x_{n,t}$ entre la fréquence fondamentale trouvée et celle théorique, la saillance s_{jt} associée et l'intensité éventuelle $d_{n,t}$ d'une attaque dont la fréquence fondamentale est proche de n.

Le HMM d'une note est représenté sur la figure 1.4. Les trois états utilisés sont associés à l'attaque, au *sustain* et au bruit, avec des contraintes sur les transitions possibles entre états. La vraisemblance des observations est modélisée par un GMM (Gaussian Mixture Model, modèle de mélange de gaussiennes). L'avantage de cette approche réside dans la

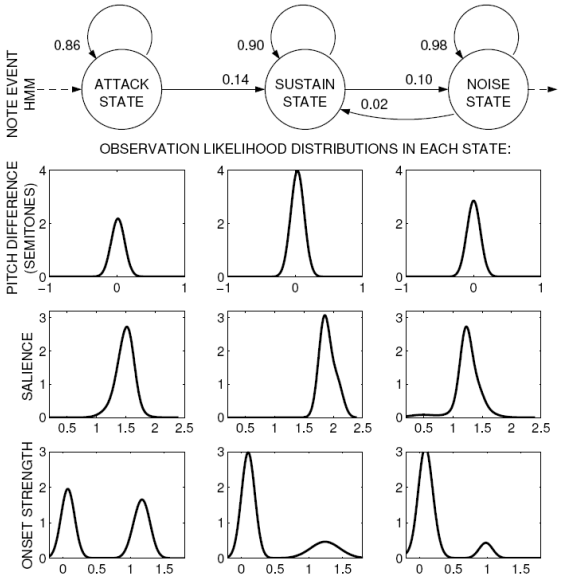


FIGURE 1.4 – HMM pour la reconnaissance de notes : chaîne de Markov (en haut) et densité de probabilité des observations par états (en bas). (Source : Ryynänen et Klapuri [2005])

simplification qu'elle propose. Grâce à cette division en deux tâches distinctes, la décision s'appuie sur un nombre restreint de paramètres, et sur un modèle de note à trois états : l'attaque caractérisée par une forte variation de saillance, le sustain pour lequel la saillance est élevée, stable et l'écart de fréquence fondamentale faible, et le silence dans les autres cas (cf. figure 1.4).

D'autres approches à base de paramétrisation et d'estimation de hauteur ont été élaborées. L'idée est déjà présente dans les travaux de Moorer [1975]. Martin [1996] propose un système de décision à plusieurs couches hiérarchisant l'information sur une échelle bas-niveau/haut-niveau, avec une influence mutuelle des diverses couches. Raphael [2002] propose une structure de HMM que l'on retrouve chez Ryynänen et Klapuri [2005]. La décomposition en deux couches – analyse fréquentielle et suivi temporel – est également adoptée dans le système de Poliner et Ellis [2007], dont la paramétrisation se fait cette fois par des SVM (Support Vector Machine, machines à vecteurs supports), et qui repose donc sur un apprentissage préalable.

1.4.2 Approches avec apprentissage préalable

Nous abordons maintenant les méthodes dites avec apprentissage, pour lesquelles la transcription est obtenue à l'aide d'informations issues d'une base de référence pour reconnaître les notes, et éventuellement les instruments. Les questions liées à cette problématique sont alors :

- quelle information apprendre : notes, timbres, dans le domaine temporel ou spectral ?
- quelle technique de reconnaissance utiliser ?
- comment concevoir la base d'apprentissage pour obtenir une reconnaissance robuste lors du passage à des données inconnues ?

Un premier principe d'apprentissage consiste à construire un **dictionnaire** dont les éléments correspondent en général à des notes, et à chercher à minimiser une distance donnée entre une combinaison d'éléments du dictionnaire et le morceau à transcrire. C'est la démarche adoptée par exemple par Rossi [1998], Ortiz-Berenguer et Casajús-Quirós [2002], Plumbley *et al.* [2006] et Cont *et al.* [2007] qui s'appuient sur un dictionnaire de spectres de notes. L'emploi d'un dictionnaire présente l'avantage, par rapport à l'utilisation de modèles, d'avoir pour références des données réelles, dont le niveau de détail et de complexité n'est pas toujours atteignable par un modèle. En revanche, il faut pouvoir tolérer une certaine variation autour des éléments du dictionnaire. Ainsi, l'utilisation d'un dictionnaire de spectres de notes atteint vite sa limite lorsque l'on considère que l'apparition d'une note correspond à une simple multiplication de son spectre de référence par un coefficient d'amplitude : en pratique, les variations d'amplitude d'un spectre en fonction de la nuance sont non-linéaires, tout comme l'évolution temporelle des partiels.

L'utilisation de techniques de **classification automatique** ajoute à l'apprentissage des données un apprentissage de leurs variations et une capacité de généralisation, c'est-à-dire la construction d'une classe à partir de plusieurs échantillons. Les réseaux de neurones ont ainsi été utilisés par Marolt [2004] à plusieurs niveaux : pour la détection d'attaques (« Integrate-and-Fire Neural Network »), la reconnaissance de notes à partir de partiels (« Time Delay Neural Network ») et la détection de notes répétées (« Multilayer Perceptron Neural Network »). En utilisant les machines à vecteurs supports, Poliner et Ellis [2007] effectuent un apprentissage non plus sur des notes isolées mais sur des mélanges de notes, avec pour effet d'apprendre à la fois les caractéristiques des notes et les conséquences d'un mélange telles que le recouvrement spectral. Le système aura par exemple appris à reconnaître une note lorsqu'elle est isolée et lorsqu'elle se trouve en présence de son octave. La capacité de généralisation du système dépend ensuite du choix de la base d'apprentissage et de la paramétrisation du système, étapes délicates qui peuvent mener à une dégradation des performances dans le cas d'un sous-apprentissage (base trop petite, manquant de diversité) ou, à l'opposé, dans le cas d'un sur-apprentissage.

1.4.3 Approches avec apprentissage en ligne

En général, les méthodes avec apprentissage préalable ont une capacité de généralisation limitée : les résultats se dégradent lorsque le signal à traiter présente des différences importantes avec les données apprises. La variabilité des timbres, la réverbération ou les effets de mixage peuvent par exemple être à l'origine de cette dégradation puisque la base d'apprentissage ne peut refléter tous les cas de signaux rencontrés ensuite. Plusieurs approches récentes proposent d'effectuer un apprentissage directement sur le signal à analyser. La NMF (Non-negative Matrix Factorization, factorisation en matrices non-négatives, c'est-à-dire à coefficients positifs, introduite par Lee et Seung [1999]) est une technique très appropriée de ce point de vue. Smaragdis et Brown [2003] ont été les premiers à l'appliquer à la transcription : une représentation temps-fréquence positive telle que le spectrogramme est modélisée et décomposée comme un produit de deux matrices non-négatives, l'une contenant un dictionnaire de spectres de notes, et l'autre des instants d'activation de ces notes. Les deux matrices sont estimées à partir du signal, le dictionnaire de notes constituant une sorte de base pour décrire l'observation. Ces approches furent ensuite perfectionnées en ajoutant des contraintes [Virtanen, 2007; Bertin *et al.*, 2007] ou en assouplissant le modèle de spectre qui était jusque-là figé pour chaque note [Vincent *et al.*, 2008]. Ces derniers travaux montrent la proximité que l'on peut trouver entre la

séparation de sources et la transcription automatique. Dans le même esprit, Duan *et al.* [2008] proposent une autre méthode dans laquelle les enveloppes spectrales des sources sont apprises de manière adaptative sur le signal à analyser, dans une approche s'appuyant sur une distance entre peignes harmoniques et des modèles gaussiens associés.

Dans un tout autre registre, l'approche de Bello *et al.* [2006] propose également un apprentissage en ligne. Il s'agit cette fois de constituer un dictionnaire de formes d'onde de notes de piano. Dans une première analyse du signal, les notes isolées sont repérées à l'aide d'une méthode spectrale d'estimation. Les formes d'onde des notes sélectionnées servent alors à construire un dictionnaire de formes d'onde, dont les signaux des notes manquantes sont obtenues par interpolation de ceux des notes existantes. La transcription est finalement obtenue en appliquant la méthode des moindres carrés afin d'identifier l'apparition des formes d'onde du dictionnaire dans le signal.

1.4.4 Approches bayésiennes

Récemment, l'introduction des approches bayésiennes pour la transcription automatique a permis de donner au problème un cadre statistique théorique et d'élargir les capacités de modélisation. Le signal est décrit de manière unifiée et systématique par des variables aléatoires représentant chaque élément à modéliser : distribution des fréquences, amplitudes, bruit, mais aussi polyphonie, évolution des partiels, notes, instruments, tonalité, etc. Ces variables aléatoires interdépendantes forment ainsi un réseau caractérisé par le choix de leurs distributions probabilistes. Le problème est alors résolu via la théorie de l'inférence bayésienne, en estimant les modèles et paramètres optimaux à partir du signal observé. Schématiquement, lorsque l'on modélise le signal observé x par un modèle statistique θ, résoudre le problème consiste à trouver les paramètres les plus probables, c'est-à-dire à maximiser la loi *a posteriori* $p(\theta|x)$ par rapport à θ. En appliquant la règle de Bayes, cette loi est proportionnelle à $p(x|\theta)\,p(\theta)$, produit de la *vraisemblance* $p(x|\theta)$ des données et de la loi *a priori* $p(\theta)$. Ces deux fonctions sont précisément celles qui ont été choisies pour modéliser le signal : la loi *a priori* traduit la connaissance que l'on a sur les paramètres – comme la distribution des amplitudes, des phases ou du nombre de composantes – tandis que la vraisemblance établit le rapport entre paramètres et signal – le signal est par exemple une somme de sinusoïdes et de bruit.

Comme nous l'avons vu jusque-là, la transcription automatique fait intervenir un grand nombre de paramètres et de variables interdépendantes. L'approche bayésienne offre l'avantage de proposer un cadre statistique rigoureux pour modéliser l'ensemble de ce système. Les techniques de résolution telles que les méthodes de Monte-Carlo sont alors particulièrement efficaces pour estimer les paramètres recherchés en prenant en compte conjointement toutes les interdépendances. Nous analysons ici les modèles utilisés dans quelques approches bayésiennes pour la transcription afin de les confronter au reste de l'état de l'art. Le lecteur intéressé par les techniques de modélisation bayésienne et de résolution pourra par exemple se reporter aux travaux de Cemgil [2004] et de Doucet et Wang [2005].

À notre connaissance, le premier modèle bayésien de signal pour la transcription a été proposé par Walmsley *et al.* [1999]. Le signal est considéré par trames comme la somme de signaux de notes et d'un bruit blanc. Chaque signal de note est lui-même une somme de sinusoïdes dont les fréquences suivent une distribution harmonique. La fréquence fondamentale, les amplitudes, le nombre de composantes d'une note sont des variables aléatoires. Il est intéressant de noter que les amplitudes sont *i.i.d.* (indépendantes identiquement distribuées) suivant une loi uniforme. Il n'y a donc pas de contrainte sur l'enveloppe spectrale, à

l'exception de la coupure passe-bas imposée par le nombre de composantes. Une contrainte de continuité des fréquences d'une trame à l'autre est en revanche introduite dans la loi suivie par la fréquence fondamentale.

La modélisation de Davy et Godsill [2002] enrichit la précédente sur plusieurs points. Le bruit est un modèle autorégressif (AR) d'ordre p fixe. Pour une note donnée, le vecteur des amplitudes des partiels est gaussien centré, de covariance diagonale. Une contrainte d'enveloppe spectrale est ici imposée en rendant le terme m de cette diagonale proportionnel à $\frac{1}{m^2}$. Autrement dit, le partiel d'ordre m est contraint d'avoir une amplitude de l'ordre de $\frac{1}{m}$, le coefficient de proportionnalité étant égal pour toutes les notes (et, accessoirement, proportionnel à la puissance du bruit). Une autre amélioration du modèle précédent consiste à autoriser une variation de fréquence (gaussienne centrée) autour de la fréquence théorique multiple de la fondamentale. Enfin, l'utilisation d'une fenêtre temporelle de pondération des trames permet d'obtenir de faibles variations des amplitudes en fonction du temps.

La modélisation de Davy et Godsill [2002] est reprise par Davy *et al.* [2006] avec quelques changements sur les points détaillés précédemment. D'une part, le modèle AR de bruit est abandonné au profit d'un bruit blanc gaussien car les pôles du processus AR peuvent modéliser certains partiels. D'autre part, la covariance diagonale des amplitudes des partiels ne décroît plus en fonction de l'ordre du partiel, elle ne porte que l'information d'un « rapport signal à bruit » entre les notes et le bruit blanc. Ainsi, sur ces deux points, le modèle utilisé tend à se rapprocher de celui de Walmsley *et al.* [1999].

La modélisation bayésienne introduite par Cemgil *et al.* [2006] pour la transcription automatique de la musique présente davantage la tâche comme un suivi d'oscillateurs. La note est modélisée de façon parfaitement harmonique, avec un facteur d'amortissement multiplicatif $\rho \in [0; 1]$. Cette décroissance est une contrainte plus forte que la fenêtre de pondération utilisée dans le modèle précédent et s'explique par l'allure générale d'enveloppes spectrales d'instruments à cordes vibrant librement. Le facteur d'amortissement (temporel) est également utilisé pour modéliser une décroissance exponentielle de l'enveloppe spectrale, le terme d'ordre h étant proportionnel à ρ^h. Ce modèle impose donc une contrainte très forte d'enveloppe spectrale et temporelle. Le bruit est quant à lui gaussien centré, et sa covariance est apprise sur les données.

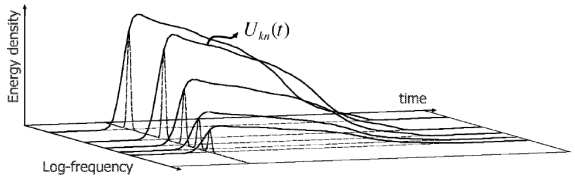


FIGURE 1.5 – Profil temps-fréquence du modèle de source HTC (Source : Kameoka *et al.* [2007])

Alors que les modélisations bayésiennes présentées jusqu'ici analysent des trames successives de signal sans transformation temps-fréquence, et peuvent donc être qualifiées à ce titre de méthodes temporelles, la modélisation de Kameoka *et al.* [2007] utilise comme observation une représentation temps-fréquence quadratique, obtenue à partir d'une transformée en ondelettes de Gabor. Il ne s'agit pas à proprement parler d'une modélisation bayésienne, mais elle s'en rapproche par son formalisme statistique et l'introduction de lois *a priori* sur certains paramètres. Dans cette modélisation, on considère que la transfor-

mée temps-fréquence observée est la somme de contributions de sources correspondant à des notes. La décomposition, baptisée HTC (Harmonic Temporal Structured Clustering), consiste à minimiser une distance (celle de Kullback-Leibler, mesurant l'information mutuelle, dans un cadre probabiliste) entre le modèle de source et sa contribution estimée. Le modèle de source, représenté sur la figure 1.5, consiste en le produit de l'énergie de la note et d'un profil temps-fréquence, quasi-harmonique, normalisé. La définition de ce profil à l'aide de contraintes temps-fréquence permet une grande souplesse dans la modélisation des partiels. La loi de la fréquence des partiels est une gaussienne centrée autour d'une distribution harmonique, dont la fréquence fondamentale peut varier de manière polynomiale au cours du temps. L'énergie des partiels est le produit d'une énergie relative, par rapport au premier partiel, et d'un profil normalisé d'évolution temporelle. L'énergie relative est centrée autour d'une valeur proportionelle à $\frac{1}{n^2}$, où n est l'ordre du partiel. Le profil d'évolution temporelle du partiel se compose quant à lui d'une somme pondérée de gaussiennes centrées aux instants d'analyse. Leurs poids sont des variables aléatoires de moyenne décroissant exponentiellement selon $e^{\alpha y}$, où y est l'indice de la trame et α un coefficient arbitrairement fixé.

Nous voyons qu'un autre intérêt des approches bayésiennes décrites ci-dessus réside dans le **modèle additif de notes** qu'elles proposent pour les mélanges polyphoniques. Nous avons constaté dans la partie 1.3 qu'il n'est pas évident de séparer des sources qui se recouvrent à la fois temporellement et fréquentiellement, et que cela donne lieu à des approximations, en particulier dans les méthodes d'estimation itératives ou lorsque l'on travaille sur des représentations temps-fréquence quadratiques (NMF, etc.) ne prenant donc pas en compte l'information de phase. De ce point de vue, l'approche bayésienne temporelle permet une décomposition et une estimation des amplitudes et des phases des composantes sans approximation. Quant à la méthode de Kameoka *et al.* [2007], elle repose sur une représentation temps-fréquence énergétique, et ne peut donc être qu'approximative, mais elle explicite cette approximation. Les contributions de chaque source k sont calculées comme une proportion $m_k(t,x)$ de l'énergie $W(t,x)$ du signal à l'instant t et à la fréquence x, avec la contrainte de conservation de l'énergie $\sum_k m_k(t,x) = 1$. $m_k(t,x)$ est alors estimé comme le rapport $\frac{q_k(t,x)}{\sum_{k'} q_{k'}(t,x)}$ entre l'énergie du modèle de la source k en ce point temps-fréquence et la somme des énergies de tous les modèles de sources. Il est intéressant de noter que ce rapport est analogue à la réponse du filtre de Wiener construit en considérant $q_k(t,x)$ comme la densité spectrale de puissance de la source k, à une différence près : la réponse fréquentielle du filtre de Wiener serait $\frac{q_k(t,x)}{\sum_{k'} q_{k'}(t,x)}$, entraînant une proportion d'énergie $m_k(t,x)$ égale à $\left(\frac{q_k(t,x)}{\sum_{k'} q_{k'}(t,x)}\right)^2$, soit le carré de la proportion utilisée dans la modélisation HTC. Cette différence provient de la contrainte de conservation de l'énergie dans le cas HTC alors que le filtrage de Wiener assure l'égalité entre le signal original et la somme des signaux des sources estimées.

Cette description des méthodes bayésiennes nous amène à revenir sur le problème de la modélisation de l'enveloppe spectrale. Les contraintes utilisées, telles que la décroissance exponentielle des amplitudes des partiels, fixent un modèle très approximatif, loin de la réalité des sons d'instruments de musique en général et de piano en particulier (cf. la description de ces sons dans la partie 1.5.1). On peut ainsi opposer ces modèles paramétriques d'enveloppe spectrale aux méthodes adaptatives comme le principe de *spectral smoothness* de Klapuri [2003] et aux modèles figés contenus dans les dictionnaires, et souligner la difficulté de la modélisation de l'enveloppe spectrale.

1.4.5 Traitement de l'information de haut-niveau

Nous n'avons pour l'instant considéré que l'information de bas-niveau que sont les hauteurs, les attaques et les extinctions de notes. L'information de haut-niveau est parfois prise en compte, soit pour être transcrite en tant que telle (extraction du tempo, reconnaissance des instruments, etc.), soit pour améliorer les performances globales du système (en s'aidant par exemple du contexte tonal pour estimer les hauteurs présentes). La transcription de l'information de haut-niveau est une tâche qui se trouve à la limite des travaux présentés ici. Citons tout de même Gómez [2006] et Lee [2008] qui s'intéressent à la transcription spécifique de la tonalité et des accords. Pour ce qui est de l'utilisation de l'information de haut-niveau, Kashino *et al.* [1998] introduisent ainsi de la connaissance *a priori* sur les accords et les transitions entre accords. Pour améliorer la transcription, Ryynänen et Klapuri [2005] reprennent le modèle musicologique proposé par Viitaniemi *et al.* [2003] et Ryynänen et Klapuri [2004]. Il comprend une estimation de la tonalité, puis l'utilisation des probabilités de transition entre notes sachant cette tonalité.

1.5 Transcription automatique de piano

Nous abordons maintenant les spécificités de la transcription automatique de la musique de piano. Dans un premier temps, nous dégagerons les caractéristiques des sons de piano à travers un tour d'horizon des études acoustiques du piano qui permettent d'expliquer les signaux à transcrire et de modéliser leurs spectres. Nous verrons ensuite dans quelle mesure les systèmes de transcription automatique de piano proposent des solutions adaptées à cet instrument.

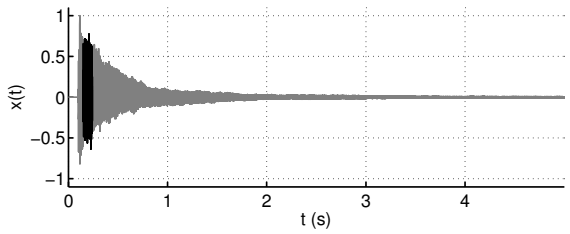
1.5.1 Éléments de physique du piano, caractérisation des sons

Le piano est un instrument à cordes libres. À ce titre, la production d'un son de piano repose sur les mécanismes suivants :

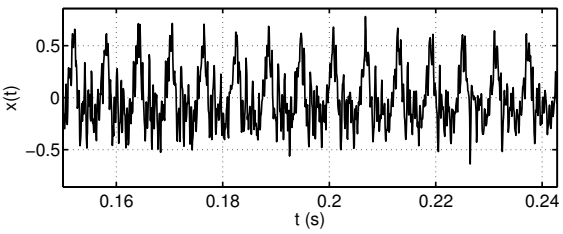
- les cordes subissent une excitation initiale sous l'effet de la frappe du marteau ; pour chaque note de piano, entre une et trois cordes, légèrement désaccordées, sont excitées ;
- elles entrent en vibration libre ;
- la vibration est transmise, via le chevalet, à la table d'harmonie qui rayonne, c'est-à-dire transforme sa vibration mécanique en onde acoustique ;
- l'onde rayonnée se propage dans l'air jusqu'aux auditeurs ou microphones éventuels.

La figure 1.6 représente un exemple de forme d'onde et de spectrogramme de son de piano. La forme d'onde montre une décroissance caractéristique des sons produits par les instruments à vibrations libres, pour lesquels l'énergie apportée au moment de l'excitation se dissipe progressivement, contrairement aux sons des instruments à son entretenu tels que les vents ou les cordes frottées. Sur le détail de la forme d'onde (figure 1.6(b)) apparaît la pseudo-périodicité du son, de fréquence fondamentale voisine de 165 Hz. La représentation temps-fréquence apporte davantage de détails sur le contenu du son. Un bruit large bande est visible au niveau de l'attaque et se dissipe rapidement. La pseudo-périodicité du son se traduit par la prédominance de partiels dont les fréquences suivent une distribution pseudo-harmonique, sans modulation fréquentielle apparente. Entre 2300 et 4000 Hz, quelques partiels ont des fréquences qui ne suivent pas cette distribution. Le spectrogramme, et plus particulièrement sa vue en trois dimensions, montre que l'évolution de l'amplitude

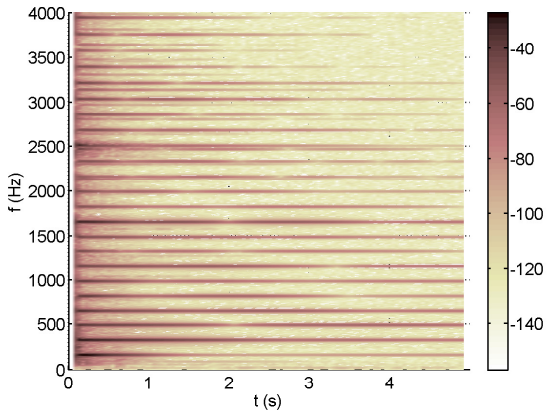
des partiels est relativement complexe. Elle présente des battements qui ne semblent pas corrélés entre partiels.



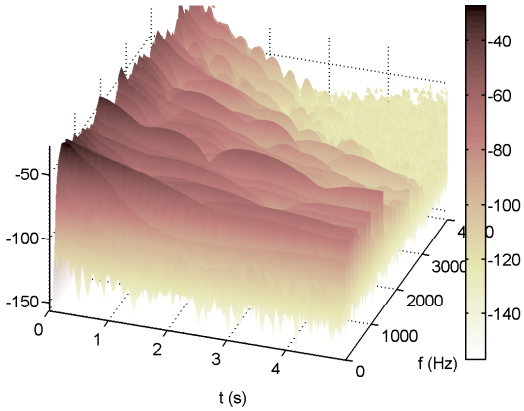
(a) Forme d'onde entière



(b) Zoom sur une partie de la forme d'onde



(c) Spectrogramme



(d) Vue 3D du spectrogramme

FIGURE 1.6 – Exemple de son de piano (Mi 2, 165 Hz)

Fréquence des partiels

La fréquence des partiels n'est pas influencée, en première approximation, par toute la chaîne de production (transmission à la table d'harmonie et rayonnement) et ne dépend que de la vibration des cordes [Fletcher et Rossing, 1998]. L'équation d'une corde simple (sans raideur) est

$$\rho \frac{\partial^2 y}{\partial t^2} = T \frac{\partial^2 y}{\partial x^2} \tag{1.9}$$

où y est le déplacement selon la direction transverse, ρ la densité linéique, T la tension, x la position le long de la corde et t le temps, pour laquelle les solutions ont pour fréquences les multiples de la fréquence fondamentale (en supposant que les extrémités de la corde sont en appui). La **fréquence fondamentale** f_0 s'exprime en fonction de la longueur L de la corde, de ρ et de T :

$$f_0 \triangleq \frac{1}{2L}\sqrt{\frac{T}{\rho}} \tag{1.10}$$

Les cordes de piano étant caractérisées par une tension et une raideur importantes, on ne peut leur appliquer les résultats précédents. Pour une corde avec raideur, l'équation de

la corde devient

$$\rho\frac{\partial^2 y}{\partial t^2} = T\frac{\partial^2 y}{\partial x^2} - EI\frac{\partial^4 y}{\partial x^4} \tag{1.11}$$

où E est le module de Young, $I \triangleq \frac{\pi d^4}{64}$ et d est le diamètre de la corde. Les solutions ont alors pour fréquences

$$f_h = hf_0\sqrt{1+\beta h^2} \tag{1.12}$$

où $h \in \mathbb{N}^*$, $\beta \triangleq \frac{\pi^3 E d^4}{64TL}$ est le **coefficient d'inharmonicité** et f_0 est la fréquence fondamentale définie par l'équation (1.10) pour la corde sans raideur.

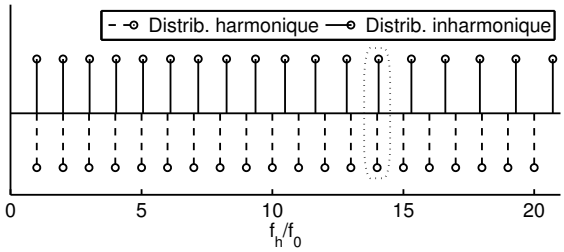


FIGURE 1.7 – Exemple de distribution inharmonique avec $\beta = 10^{-3}$: le 13e partiel se retrouve à la fréquence du 14e partiel d'une distribution harmonique de même fréquence fondamentale.

Le coefficient d'inharmonicité est propre à chaque piano [Young, 1952] et à chaque note, avec des coefficients de l'ordre de 10^{-4} dans le grave à 10^{-2} dans l'aigu. L'inharmonicité est suffisamment faible pour ne pas perturber la perception d'une hauteur à l'écoute d'un son de piano. Elle a en revanche un certain nombre de conséquences non négligeables. Ainsi, accorder un piano repose notamment sur les battements créés entre une note et son octave. Le tempérament du piano [Schuck et Young, 1943; Martin et Ward, 1954; Lattard, 1993; Conklin Jr., 1996b] intègre donc les écarts entre partiels dus à l'inharmonicité. Il en résulte un étirement de la répartition des fréquences fondamentales : celles des notes graves sont en-deçà du tempérament égal, tandis que celles de notes aiguës sont au-delà. L'écart des fréquences de partiels par rapport à une distribution harmonique est également loin d'être négligeable si l'on veut localiser ces partiels. Par exemple, avec un coefficient d'inharmonicité égal à 10^{-3}, le 13e partiel se retrouve à la fréquence du 14e partiel d'une répartition harmonique de même fréquence fondamentale (cf. figure 1.7). Estimer le coefficient d'inharmonicité [Rauhala *et al.*, 2007] se révèle donc utile voire nécessaire pour caractériser la hauteur d'une note de piano. Aussi, comme nous le verrons en détail dans la partie 1.5.2 (p. 45), la plupart des systèmes de transcription automatique de piano prennent en compte l'inharmonicité.

Amplitude des partiels

Lors de la frappe des cordes par le marteau, les modes propres ou partiels sont excités différemment les uns des autres, la localisation du point de frappe sur la corde étant un paramètre déterminant. Pour comprendre le phénomène, un modèle simple [Fletcher et Rossing, 1998] consiste à considérer une corde de longueur L vibrant librement sans

amortissement et sans raideur. Le mouvement de la corde peut se décomposer suivant ses modes propres Y_n :

$$y(x,t) = \sum_{n=1}^{+\infty} Y_n(x) \left(A_n \cos \omega_n t + B_n \sin \omega_n t\right) \tag{1.13}$$

avec $x \in [0, L]$ et $\omega_n = 2\pi n f_0$.

L'expression des modes propres est

$$Y_n(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} \tag{1.14}$$

Dans le cas d'une corde frappée ponctuellement au point $x = \alpha L$ ($\alpha \in [0; 1]$) à l'instant $t = 0$, on peut choisir comme conditions initiales

$$y(x, t = 0) = 0 \tag{1.15}$$

$$\frac{\partial y}{\partial t}(x, t = 0) = V_0 \delta(x - \alpha L) \tag{1.16}$$

Par projection de $y(x, t = 0)$ sur la base des modes propres, la condition (1.15) donne

$$\forall n, A_n = 0 \tag{1.17}$$

et

$$\forall n, B_n = \frac{V_0}{\sqrt{2L} f_0} \frac{\sin n\pi\alpha}{n\pi} \tag{1.18}$$

Ces amplitudes de partiels sont représentées sur la figure 1.8, après normalisation par $\frac{V_0}{\sqrt{2L} f_0}$.

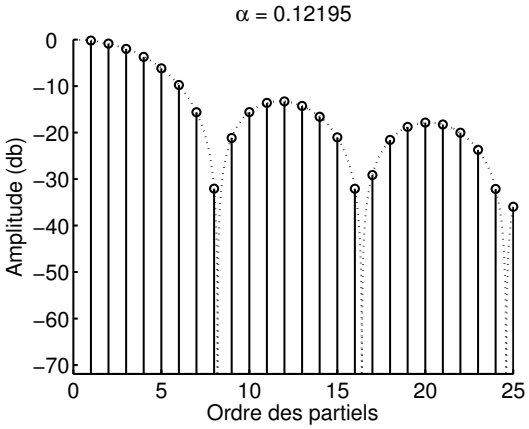


FIGURE 1.8 – Amplitudes normalisées, des partiels, en db, au niveau du chevalet et à l'instant initial, pour $\alpha = 0,12195$.

La position du marteau est donnée habituellement par le coefficient α après normalisation par la longueur de la corde. Sa principale conséquence est que les modes qui ont un nœud de vibration au voisinage du point de frappe sont peu excités. Cette position est particulièrement étudiée et discutée par Hall et Clark [1987] et Conklin Jr. [1996a]. Il en

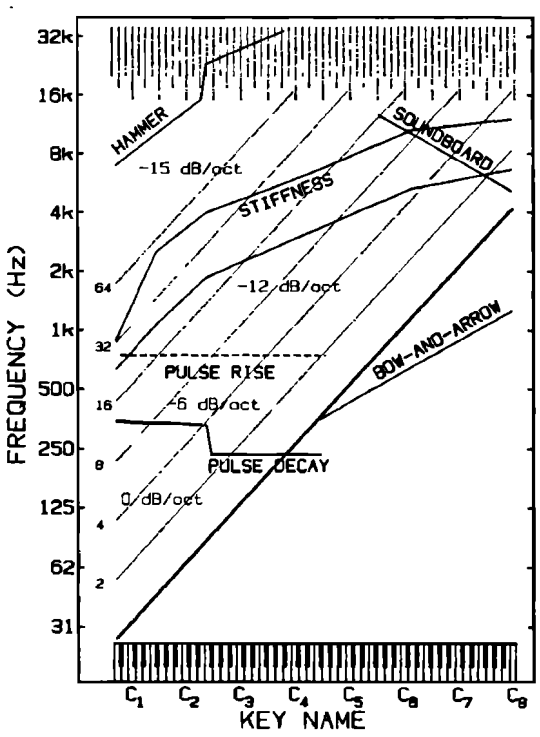


FIGURE 1.9 – Domaine de validité des modèles d'excitation des cordes, avec les notes en abscisse et les fréquences des partiels en ordonnée (Source : Hall et Askenfelt [1988]).

ressort qu'elle varie selon l'instrument et l'époque. En pratique, on considère couramment que la valeur de α est comprise entre $1/9 \approx 0,11$ et $1/7 \approx 0,14$. Cette règle assez répandue est très approximative, les valeurs pouvant par exemple descendre à $0,08$ dans l'aigu sur un piano moderne.

Le modèle présenté précédemment se révèle assez simpliste en pratique. Cela est regrettable car une modélisation de type MA (filtre à moyenne ajustée, cf. annexe A.2.2 (p. 170)) conviendrait naturellement si l'enveloppe spectrale des sons suivait ce modèle. L'absence d'excitation au niveau d'un nœud n'est en général pas vérifiée pour deux raisons principales. D'une part, la corde n'est pas excitée sur une zone ponctuelle compte tenu de la largeur du marteau. D'autre part, étant données les pertes au niveau du chevalet dues à l'admittance finie de la table d'harmonie, les ondes réfléchies sont atténuées par rapport aux ondes incidentes, mettant en défaut l'hypothèse d'ondes stationnaires et donc l'existence des nœuds. Aussi, le modèle précédent a été nettement enrichi par l'étude physique de l'excitation des cordes par un marteau dans le cas du piano, réalisée dans une série d'articles [Hall, 1986, 1987a,b; Hall et Clark, 1987; Hall et Askenfelt, 1988; Hall, 1992]. Une synthèse proposée dans Hall et Askenfelt [1988] dégage des zones de validité de divers régimes, représentées sur la figure 1.9, en fonction de la note et des fréquences des partiels :

- le modèle d'excitation ponctuelle par une impulsion de dirac est valable pour les premiers modes propres des notes les plus graves ;
- pour ces mêmes notes, au-delà de la ligne « Pulse Decay », une pente de -6dB/oct est à introduire pour prendre en compte la durée du contact marteau-corde, qui n'est plus négligeable devant la période du partiel ;

– au-delà de la ligne « Pulse Rise », une autre pente de −6dB/oct s'ajoute (soit −12dB/oct) car l'élasticité du marteau devient importante devant une élasticité critique ;
– pour les notes aiguës, un modèle où la masse du marteau est supérieure à celle de la corde donne une pente de −12dB/oct dès les premiers modes.

Il convient de préciser que les amplitudes des partiels ainsi modélisées ne sont valables que pour la vibration du chevalet et qu'à l'instant de frappe. En effet, avant d'être perçu, le son subit des transformations successives, modélisables sous forme de filtrages et produites lors de la transmission à la table d'harmonie, du rayonnement et de la propagation dans le milieu ambiant. Quant aux phénomènes qui suivent l'instant de frappe, ils sont maintenant examinés à travers l'étude de l'évolution des partiels.

Évolution des partiels

L'allure des partiels, observable sur la figure 1.6(d), fait intervenir deux phénomènes principaux : d'une part, une décroissance globale qualifiée de double décroissance en raison d'une pente forte au début suivie d'une pente plus faible, et d'autre part, des battements. Ces phénomènes sont des conséquences des couplages entre les modes des cordes, qui ont notamment été étudiés par Weinreich [1977]. Dans la configuration d'une note avec deux ou trois cordes, celles-ci ne vibrent pas indépendamment mais sont couplées au niveau du chevalet. De plus, chaque corde a deux polarisations, horizontale et verticale. On obtient ainsi six modes propres couplés. En étudiant le cas plus simple de deux cordes dont la polarisation horizontale est négligée, et dont les pulsations propres sont voisines en raison du désaccord entre les cordes qu'introduit l'accordeur de piano, les phénomènes de battement sont mis en évidence. Plusieurs cas sont étudiés, suivant la nature du couplage – purement résistif ou pas – et le désaccord entre les cordes. Le phénomène de double décroissance est ensuite expliqué en considérant cette fois une seule corde et ses deux polarisations, puis deux cordes. L'excitation lors de la frappe du marteau impose une vibration initiale en phase des cordes. La force appliquée au chevalet est donc importante, donnant lieu à une forte dissipation. Les oscillateurs se mettent ensuite rapidement en opposition de phase pour atteindre un régime moins dissipatif : les déplacements se compensent et la force appliquée au chevalet est alors minimisée, d'où un changement de pente.

Dans le cas général, l'évolution des partiels est donc relativement complexe, dépendant notamment de paramètres tels que le désaccord entre cordes, l'impédance caractéristique des cordes et l'admittance du chevalet. Ces valeurs varient en fonction de la note et de l'instrument et sont donc difficilement accessibles dans les conditions de la transcription automatique, c'est-à-dire en l'absence d'accès direct au piano. Ceci explique sûrement la rareté des méthodes d'analyse de l'évolution des partiels de piano [Rauhala, 2007], alors que ces phénomènes ont été largement utilisés pour la synthèse [Smith et Van Duyne, 1995; Bank, 2000; Bank et Sujbert, 2002; Bensa, 2003; Rauhala et Valimaki, 2006].

Autres partiels

Seuls les modes transverses, c'est-à-dire dont le déplacement est dans un plan orthogonal à l'axe de la corde, ont été considérés jusqu'à présent. Ils sont responsables du caractère pseudo-harmonique du son. Les modes longitudinaux (ou de compression) des cordes sont également excités et audibles [Conklin Jr., 1996a,b; Fletcher et Rossing, 1998; Galembo et Askenfelt, 1999; Bank et Sujbert, 2005], ainsi que des modes dits « fantômes » [Conklin Jr., 1997]. Ces partiels ont des fréquences qui ne font *a priori* pas partie de la distribution pseudo-harmonique relative à la note. Ils agissent donc plutôt comme des fréquences

parasites dans un contexte d'estimation de fréquences fondamentales.

Pédales

Le pédalier du piano à queue comporte traditionnellement trois pédales. Celle de droite, la pédale *forte*, relève les étouffoirs, laissant l'ensemble des cordes libres. Il en résulte une légère modification du son et la possibilité pour l'instrumentiste de laisser sonner une note après avoir relâché la touche. Le son produit lorsque cette pédale est enfoncée a été étudié, sur un plan perceptif [Martin et Ward, 1954] et physique [Fletcher *et al.*, 1962; Lehtonen *et al.*, 2007], plusieurs phénomènes ayant été mis en avant. Le principal est un rehaussement du bruit de fond dû à la vibration de l'ensemble des cordes. Celles-ci ne sont pas excitées directement par les marteaux mais elles vibrent néanmoins, soit par sympathie avec les notes jouées, soit en fonction des bruits impulsionnels transmis par le piano (relèvement des étouffoirs, chocs des marteaux, etc.). Les autres effets observés par Lehtonen *et al.* [2007] concernent une modification de l'évolution des partiels. Leur amplitude a tendance à décroître moins rapidement lorsque la pédale est enfoncée. L'explication proviendrait du couplage entre une corde jouée et l'ensemble des cordes, ces dernières dissipant davantage l'énergie transmise lorsqu'elles sont étouffées. La part plus importante de couplages non dissipatifs expliquerait également l'augmentation des battements observés avec la pédale enfoncée, tout comme un affaiblissement du phénomène de double décroissance.

La pédale du milieu, dite tonale ou de soutien, permet de tenir les notes jouées au moment où elle est enfoncée, en gardant les autres étouffées. Il n'y a à notre connaissance aucune étude à son sujet susceptible de nous intéresser, probablement du fait de sa similitude avec la pédale *forte* et d'une utilisation plus réservée à un contexte d'étude qu'à une interprétation musicale. La pédale de gauche, dite *una corda*, déplace latéralement le bloc constitué du clavier et des marteaux afin que toutes les cordes ne soient pas frappées. Le son obtenu en est ainsi modifié et sa production est évoquée dans plusieurs travaux [Weinreich, 1977; Fletcher et Rossing, 1998; Bank, 2000].

1.5.2 Systèmes de transcription de piano

Nous allons maintenant voir dans quelle mesure les systèmes de transcription automatique de piano allient les spécificités du piano aux techniques de transcription évoquées dans les parties 1.3 et 1.4.

La variété des techniques utilisées pour transcrire la musique de piano est aussi grande que pour la transcription automatique générique. On retrouve ainsi des méthodes reposant sur une paramétrisation [Dixon, 2000; Monti et Sandler, 2002; Raphael, 2002; Kobzantsev *et al.*, 2005; Wen et Sandler, 2005b], des méthodes avec apprentissage préalable [Rossi, 1998; Ortiz-Berenguer et Casajús-Quirós, 2002; Marolt, 2004; Poliner et Ellis, 2007] ou en ligne [Bello *et al.*, 2006; Vincent *et al.*, 2008] et des méthodes statistiques [Godsill et Davy, 2005].

L'inharmonicité des sons de piano est en général prise en compte dans les systèmes de transcription. L'utilisation explicite de la loi d'inharmonicité donnée par l'équation (1.12) (p. 41) permet de saisir au mieux les partiels dans une représentation fréquentielle [Ortiz-Berenguer et Casajús-Quirós, 2002; Vincent *et al.*, 2008]. L'inharmonicité peut également être implicite lorsque l'apprentissage l'englobe [Rossi, 1998; Poliner et Ellis, 2007; Bello *et al.*, 2006]. Enfin certains systèmes proposent une approximation ou une généralisation de l'inharmonicité en permettant que l'écart entre les partiels varie, sans s'appuyer sur la loi d'inharmonicité du piano [Klapuri, 2003; Godsill et Davy, 2005; Wen

et Sandler, 2005a].

La détection d'attaques semble être également un critère largement répandu. Il peut faire l'objet d'un module dédié [Barbancho *et al.*, 2004; Marolt, 2004; Monti et Sandler, 2002] ou d'un état de HMM [Raphael, 2002].

Comme dans la majorité des systèmes de transcription polyphonique, l'enveloppe spectrale constitue une source d'information importante. Il est assez rare de trouver des systèmes qui n'utilisent pas cette information [Raphael, 2002; Monti et Sandler, 2002]. L'apprentissage préalable de l'enveloppe des notes se retrouve dans les systèmes performants de Marolt [2004] et Poliner et Ellis [2007]. Les systèmes d'apprentissage en ligne [Bello *et al.*, 2006; Vincent *et al.*, 2008] ont en plus l'avantage de disposer de dictionnaires adaptés au morceau à transcrire. Ces systèmes avec apprentissage souffrent néanmoins de deux défauts : les variations de l'enveloppe spectrale au cours du temps, notamment en raison des battements, et, pour les systèmes avec apprentissage préalable, les différences entre les enveloppes spectrales apprises et celles rencontrées par la suite.

Enfin, signalons qu'à notre connaissance, seul le système de Barbancho *et al.* [2004] prend en compte la pédale *forte* pour réaliser une transcription.

1.6 Problématiques

À la lumière de cet état de l'art, voici maintenant les différentes questions que nous traiterons dans cette thèse.

Quel modèle de note utiliser pour l'estimation de fréquences fondamentales multiples ?

La question de l'enveloppe spectrale nous apparaît comme centrale pour aborder l'analyse de sons polyphoniques. Alors qu'une grande variété de solutions originales ont déjà été proposées, cet élément-clé reste souvent un point faible des systèmes de transcription.

Comment modéliser la superposition des notes ?

En particulier, comment traiter le problème du recouvrement spectral entre notes en rapport harmonique ? Il nous semble important d'utiliser l'information sur l'enveloppe spectrale pour lever l'ambiguïté de ces cas fréquents.

Quelle stratégie adopter pour transcrire la musique de piano ?

Parmi toutes les approches possibles, certaines sont-elles plus adaptées aux spécifités du piano ?

Quelles sont les conséquences de l'inharmonicité des sons de piano sur la tâche de transcription ?

Ce paramètre entraîne une augmentation de la complexité des modèles et des systèmes puisque les fréquences des partiels sont déterminées par deux paramètres, le coefficient d'inharmonicité et la fréquence fondamentale. Dans quelle mesure ce paramètre supplémentaire est-il une contrainte et introduit-il une incertitude sur la fréquence des partiels ? L'inharmonicité peut-elle être au contraire utilisée à profit pour identifier des notes jouées simultanément ?

Comment estimer le degré de polyphonie ?

Nous avons vu que bien peu de systèmes estiment le nombre de notes simultanées à un instant donné. Sur quels critères peut-on s'appuyer pour le déterminer ?

Qu'est-ce qu'une bonne transcription ?

Comment quantifier la qualité d'une transcription ? Peut-on comparer et ordonner des transcriptions suivant leur qualité ? Quels sont les critères à évaluer ? Quelles sont les erreurs de transcription typiques ?

Chapitre 2

Paramétrisation spectrale des sons de piano

Ce chapitre est consacré à la caractérisation des spectres de sons de piano, dans la perspective de l'estimation de la hauteur des notes. À ce titre, quatre aspects seront étudiés distinctement. Nous aurons tout d'abord besoin d'identifier les composantes sinusoïdales des sons. La modélisation du contenu tonal fera donc l'objet de la première partie. Pour faire le lien entre les composantes estimées et les notes présentes, nous devrons étudier la distribution des fréquences des partiels d'une note. Cette question est d'un intérêt particulier dans le cas du piano du fait de l'inharmonicité caractéristique des sons. Nous y consacrerons la deuxième partie, où nous présenterons nos travaux sur l'estimation de l'inharmonicité et sur l'impact d'une telle caractérisation. La troisième partie aura pour thème la modélisation de l'enveloppe spectrale des sons de piano pour la transcription. Nous avons vu l'intérêt, voire la nécessité de prendre en compte l'enveloppe spectrale pour l'estimation de fréquences fondamentales multiples. Dans cette optique, nous proposerons une modélisation de type autorégressif de cette enveloppe spectrale et un cadre statistique approprié. Enfin, nous nous pencherons sur la question de la modélisation du bruit dans la dernière partie.

2.1 Modélisation du contenu tonal des sons de piano

Si l'on entend souvent dire que l'on peut modéliser à l'aide de sinusoïdes les sons présentant une hauteur – les sons de piano par exemple –, on se rend rapidement compte qu'il existe un nombre important de façons de procéder. L'origine de cette multiplicité est double – dépendant à la fois de la variété des modèles et, pour chaque modèle, des méthodes d'estimation – et fait écho aux questions sous-jacentes relatives à la caractérisation des sons : doit-on modéliser des sinusoïdes modulées en fréquence et/ou en amplitude ? Si c'est le cas, comment modéliser les modulations, ou les fameux « chirps » : une sinusoïde modulée peut-elle être modélisée par plusieurs sinusoïdes non modulées (sur de courtes trames successives ou, de façon simultanée, par projection sur une base) ? Un modèle donné garantit-il l'unicité de la représentation ? Que peut-on considérer comme aléatoire : les phases initiales, le bruit, les amplitudes ? Quels modèles permettent d'optimiser un rapport qualité/nombre de paramètres ?

À défaut d'essayer de répondre à ces questions et d'approfondir la problématique de la modélisation sinusoïdale, nous présentons ici quelques approches que nous utiliserons par

la suite.

Le **modèle de McAulay et Quatieri [1986]** est fondateur dans le champ de la modélisation sinusoïdale à court terme des signaux audio. Proposé à l'origine pour des signaux de parole, il est également très utilisé pour les signaux de musique. Il consiste à considérer qu'une trame de longueur N d'un signal $x(t)$, avec $t \in [\![0; N-1]\!]$, est modélisable par une somme de K sinusoïdes, les trois paramètres de la k^{e} sinusoïde étant son amplitude $a_k > 0$, sa fréquence $f_k > 0$ et sa phase initiale $\varphi_k \in [0; 2\pi]$:

$$x(t) \triangleq \sum_{k=1}^{K} 2a_k \cos(2\pi f_k t + \varphi_k) \tag{2.1}$$

La variante complexe s'écrit

$$x_c(t) \triangleq \sum_{k=1}^{K} \alpha_k e^{i2\pi f_k t} \tag{2.2}$$

où les $\alpha_k \in \mathbb{C}^*$ sont les amplitudes complexes (de module l'amplitude réelle et d'argument la phase initiale).

Dans ce modèle, les amplitudes et les fréquences sont constantes sur la durée de la trame. Le modèle permet néanmoins de modéliser des signaux présentant des modulations en amplitude et en fréquence, à condition que ces variations soient suffisamment lentes pour pouvoir être négligées sur la durée d'une trame. Dans ce cas, les auteurs proposent un algorithme pour relier les composantes détectées d'une trame à la suivante, sur la base d'une distance entre leurs fréquences et de la possibilité qu'une sinusoïde puisse apparaître ou disparaître.

L'apport du **modèle de Serra et Smith [1990]** sur le modèle précédent réside dans l'introduction explicite d'un bruit additif. La partie bruit ne peut être ignorée qu'en première approximation et l'estimation de la partie non sinusoïdale a par la suite donné lieu à des études approfondies [d'Alessandro *et al.*, 1998a; David *et al.*, 2006]. En reprenant les notations précédentes, le signal est cette fois défini par

$$x(t) \triangleq \sum_{k=1}^{K} 2a_k \cos(2\pi f_k t + \varphi_k) + b(t) \tag{2.3}$$

ou, sous forme complexe,

$$x_c(t) \triangleq \sum_{k=1}^{K} \alpha_k e^{i2\pi f_k t} + b(t) \tag{2.4}$$

où $b(t)$ désigne la partie bruit. Le bruit est alors défini comme un processus stochastique, en l'occurrence le résultat du filtrage d'un bruit blanc par un filtre, variant temporellement, permettant de contrôler la forme de la densité spectrale de puissance du bruit. Nous aurons l'occasion de reprendre ce modèle de bruit pour l'estimation de fréquences fondamentales.

Pour ces modèles, l'estimation des paramètres se fait en général dans le domaine spectral. Considérons par exemple une trame de signal $[x(0), \ldots, x(N-1)]$ de longueur N et sa transformée de Fourier discrète $X(\nu_k)$ définie par

$$X(\nu_k) \triangleq \sum_{n=0}^{N-1} x(n)\, w(n)\, e^{-i2\pi\nu_k n} \tag{2.5}$$

où $\nu_k = \frac{k}{K}$ est la $k^{\text{ème}}$ des K fréquences considérées (avec $K \geq N$) et w est une fenêtre de pondération au choix de l'utilisateur. Le choix de w permet d'ajuster l'étalement spectral, et en particulier le compromis entre la largeur du lobe principal des composantes sinusoïdales et le niveau des lobes secondaires. Le nombre K de fréquences correspond à l'échantillonnage désiré de la transformée de Fourier continue à temps discret. On parle de *zero-padding* lorsque $K > N$, c'est-à-dire lorsque l'échantillonnage fréquentiel est ainsi augmenté. Pour aller plus loin en se rapprochant du cas continu et trouver la valeur du spectre à une fréquence f, Serra et Smith [1990] et Abe et Smith [2005] tirent parti d'une interpolation quadratique du spectre en considérant les trois fréquences discrètes les plus proches de f et les valeurs du spectre associées. Plus précisément, l'interpolation donne des résultats optimaux si elle est effectuée sur le logarithme de l'amplitude du spectre, et en utilisant une fenêtre de pondération gaussienne. Cette technique est particulièrement utile lorsqu'il s'agit de localiser les sinusoïdes comme des maxima du spectre (cf. figure 2.1) et nous l'utiliserons par la suite, en particulier pour estimer finement le coefficient d'inharmonicité d'une note.

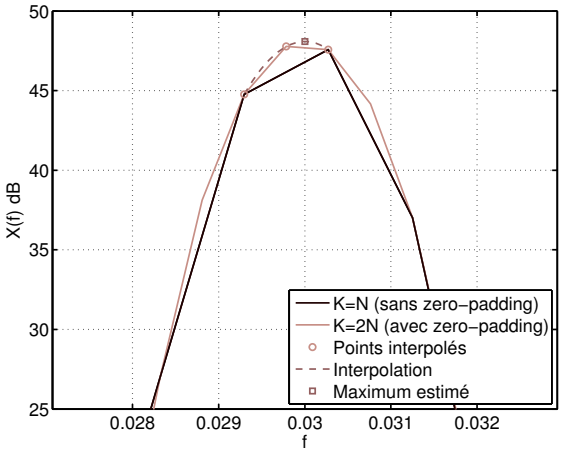


FIGURE 2.1 – Détail d'un spectre autour d'une sinusoïde de fréquence $0,03$: le maximum est estimé par interpolation des trois points les plus proches.

Suivant les modèles considérés, d'autres techniques d'estimation de paramètres existent, telle que la réallocation de la transformée de Fourier à court terme [Flandrin, 1993], les représentations parcimonieuses [Daudet et Torrésani, 2006], ou les techniques dites à Haute-Résolution [Badeau, 2005]. Parmi ces dernières, **l'algorithme ESPRIT** (*Estimation of Signal Parameters via Rotational Invariance Techniques* [Roy *et al.*, 1986]) est particulièrement performant. Nous le décrivons maintenant et aurons l'occasion de l'utiliser dans le chapitre 3 sur l'estimation de hauteurs simples. Le modèle considéré est composé d'une somme K de sinusoïdes exponentiellement amorties et de bruit. La sinusoïde k est notée $\alpha_k z_k$, où $\alpha_k \in \mathbb{C}^*$ est l'amplitude complexe et $z_k = e^{d_k + 2i\pi f_k} \in \mathbb{C}^*$ est le pôle, de fréquence $f_k \in \mathbb{R}$ et de facteur d'amortissement $d_k \in \mathbb{R}$. Ce dernier constitue un paramètre supplémentaire par rapport aux modèles présentés précédemment. Le bruit $b(t)$ est quant à lui supposé blanc additif gaussien, de variance σ_b^2. Le signal observé s'écrit donc comme la somme d'une partie sinusoïdale $s(t)$ et de la partie bruit sous la forme

$$x(t) = s(t) + b(t) \tag{2.6}$$

avec

$$s\left(t\right)=\sum_{k=1}^{K}\alpha_k z_k^t \tag{2.7}$$

L'algorithme ESPRIT permet d'estimer les pôles z_k. On suppose pour cela que le signal est défini pour $t \in [\![0; N-1]\!]$, en choisissant une longueur de trame N impaire, avec $n = \frac{N+1}{2}$. On définit alors la matrice de données

$$X \triangleq \begin{pmatrix} x(0) & x(1) & \dots & x(n-1) \\ x(1) & x(2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ x(n-1) & \dots & \dots & x(N-1) \end{pmatrix} \tag{2.8}$$

La matrice de covariance $C \triangleq \frac{1}{n}\mathbb{E}\left[XX^\dagger\right]$, où $X^\dagger$ désigne le conjugué hermitien de X, a K valeurs propres supérieures à σ_b^2 et $N-K$ valeurs propres égales σ_b^2. On estime C par la matrice d'autocorrélation empirique $\hat{C} = \frac{1}{n}\hat{X}\hat{X}^H$, où $\hat{X}$ est la réalisation de X que l'on observe. Les K vecteurs propres associés aux K plus grandes valeurs propres sont calculés puis regroupés dans une matrice W, de dimensions $n \times K$. On en extrait les matrices $W_\uparrow$ et $W_\downarrow$ en ne gardant respectivement que les $(n-1)$ dernières et les $(n-1)$ premières lignes. On prouve que $W_\uparrow$ et $W_\downarrow$ satisfont la propriété dite d'invariance rotationnelle $W_\uparrow = W_\downarrow \Phi$, où Φ est une matrice $K \times K$ dont les valeurs propres sont égales aux pôles $\{z_1, \dots, z_K\}$. La matrice Φ est estimée par moindres carrés : $\Phi = W_\downarrow^+ W_\uparrow$, où $W_\downarrow^+$ est la pseudo-inverse de $W_\downarrow$. Les pôles sont ensuite estimés en diagonalisant Φ.

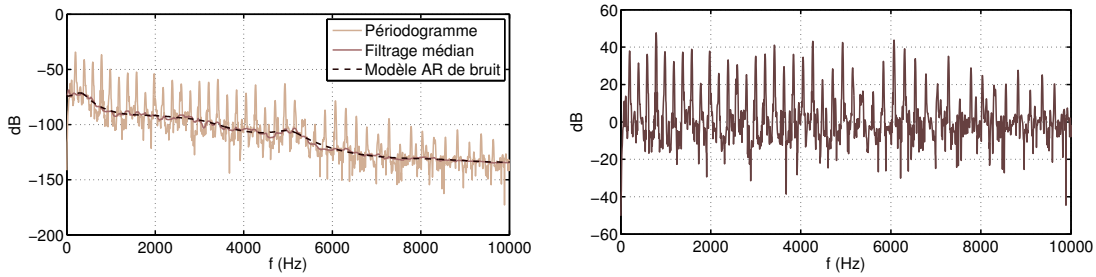
L'intérêt de cet algorithme est qu'il n'est pas limité par le compromis temps-fréquence que l'on rencontre avec la transformée de Fourier discrète. On peut donc en particulier estimer des fréquences très proches l'une de l'autre. Une fois les pôles obtenus, l'estimation des amplitudes complexes s'effectue par moindres carrés :

$$\begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_K \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ \hat{z}_1^1 & \dots & \hat{z}_K^1 \\ \vdots & \vdots & \vdots \\ \hat{z}_1^{N-1} & \dots & \hat{z}_K^{N-1} \end{pmatrix}^+ \begin{pmatrix} x(0) \\ \vdots \\ x(N-1) \end{pmatrix} \tag{2.9}$$

où $\hat{\alpha}_k$ et $\hat{z}_k$ sont les estimées de α_k et z_k pour $k \in [\![1; K]\!]$.

Cette méthode d'estimation des paramètres est performante à condition de prendre quelques précautions. Il faut tout d'abord s'assurer que l'hypothèse de bruit blanc est vérifiée avant d'appliquer ESPRIT. Le bruit large bande des signaux audio étant en pratique souvent coloré, une étape de blanchiment du bruit est souvent nécessaire comme prétraitement. Pour ce faire, nous estimerons approximativement le spectre du bruit en appliquant un filtre médian sur le périodogramme du signal afin d'éliminer les pics des composantes sinusoïdales, puis en estimant les paramètres d'un filtre autorégressif (AR) qui modélise ce niveau de bruit. L'inverse de ce filtre est ensuite appliqué sur le signal, comme l'illustre la figure 2.2, pour obtenir un niveau de bruit plat, c'est-à-dire blanc [Badeau, 2005].

L'algorithme ESPRIT a également besoin de connaître le nombre de pôles à estimer. La méthode ESTER [Badeau *et al.*, 2006] fournit un critère $J(p)$ permettant d'estimer ce nombre par



(a) Spectre du signal avant blanchiment et estimation du niveau de bruit.

(b) Spectre du signal après blanchiment.

FIGURE 2.2 – Blanchiment du bruit (Sol 2 (196 Hz) de piano analysé sur 93 ms, niveau de bruit estimé avec un filtrage médian de longueur 500 Hz environ et un filtre AR d'ordre 20).

$$\arg\max_{p \in P} \left(J(p) > \delta_J \right) \tag{2.10}$$

P étant l'ensemble des nombres de pôles possibles et δ_J un seuil arbitrairement ajusté à $\delta_J = 10$ dans notre cas. Cette méthode a tendance à fournir un résultat juste ou légèrement sur-estimé. Dans ce dernier cas, Badeau *et al.* [2006] ont montré que cela ne perturbait pas l'analyse par ESPRIT et que les pôles parasites se voyaient attribuer de faibles amplitudes.

L'utilisation d'un banc de filtres en prétraitement fournit par ailleurs un mécanisme pour diminuer le coût global de l'estimation. Pour cela, le signal est réparti dans $D = 32$ sous-bandes de largeur 500 Hz à l'aide de filtres en cosinus modulés [Vaidyanathan, 1993]. L'ordre de grandeur du coût cubique lié à ESPRIT passe alors de $\mathcal{O}\left(N^3\right)$ à $\mathcal{O}\left(N^3/D^2\right)$ ($\mathcal{O}\left(N^3/D^3\right)$ pour chaque bande).

En résumé, l'estimation des paramètres suit les étapes suivantes :

- filtrage en sous-bandes ;
- blanchiment du signal dans chaque sous-bande ;
- estimation du nombre de pôles dans chaque sous-bande ;
- estimation des pôles par ESPRIT dans chaque sous-bande ;
- estimation des amplitudes dans chaque sous-bande ;
- correction des effets des filtres de prétraitement sur les amplitudes (les pôles ne sont pas modifiés par les prétraitements).

Nous aurons l'occasion d'utiliser l'algorithme ESPRIT et les traitements qui l'accompagnent dans la méthode d'estimation de fréquences fondamentales du chapitre 3, où nous verrons comment tirer parti de cette estimation paramétrique.

2.2 Inharmonicité des sons de piano

Après avoir décrit quelques modèles sinusoïdaux, intéressons-nous maintenant à la localisation des composantes sinusoïdales, c'est-à-dire à la distribution des fréquences des partiels des sons. Nous allons étudier dans cette partie la distribution particulière qui caractérise les sons de piano (et également d'autres instruments à cordes oscillant librement tels que la guitare ou le clavecin).

2.2.1 Fréquences des partiels d'une note

Comme nous l'avons vu dans le chapitre 1 (équation (1.12) (p. 41)), la résolution de l'équation du mouvement pour une corde avec raideur donne l'expression de la fréquence f_h du partiel d'ordre h d'une note de piano :

$$f_h = h f_0 \sqrt{1 + h^2 \beta} \tag{2.11}$$

où β est le coefficient d'inharmonicité et f_0 la fréquence fondamentale d'une corde sans raideur (*i.e.* $\beta = 0$). Le coefficient β dépend en particulier de la note considérée et la fréquence f_h s'éloigne d'autant plus de l'harmonique exact hf_0 que h est élevé.

Dans la littérature, deux expressions légèrement différentes sont utilisées pour exprimer la fréquence f_h du partiel d'ordre h d'une note de piano : certains (Galembo et Askenfelt [1999]; Ortiz-Berenguer et Casajús-Quirós [2002]) utilisent l'expression (2.11) alors que d'autres (Klapuri [1999b,a]; Bello *et al.* [2006]) optent pour

$$f_h = h f_1 \sqrt{1 + (h^2 - 1)\beta} \tag{2.12}$$

où f_1 est la fréquence du premier partiel de la corde avec raideur. Il convient d'éclaircir ce point, les deux expressions n'étant pas strictement équivalentes et les raisons du choix de l'une ou l'autre n'étant en général pas explicitées. Il est vraisemblable que cette ambiguïté provienne de la référence souvent citée [Fletcher et Rossing, 1998], dans laquelle l'expression (2.11) est établie p. 62 alors que l'expression (2.12) est mentionnée p. 363. En réalité, cette dernière est obtenue à partir de la première en considérant une approximation sur $\beta \ll 1$:

$$\begin{aligned} f_h &= h f_0 \sqrt{1 + h^2 \beta} && (2.13)\\ &= h f_1 \sqrt{\frac{1 + h^2 \beta}{1 + \beta}} && (2.14)\\ &\approx h f_1 \sqrt{(1 + h^2 \beta)\,(1 - \beta)} && (2.15)\\ &\approx h f_1 \sqrt{(1 + (h^2 - 1)\,\beta)} && (2.16) \end{aligned}$$

L'expression (2.12) est donc bien une approximation de (2.11), présentant l'avantage de s'exprimer en fonction de la fréquence du premier partiel, observée sur le spectre, plutôt que de la fréquence fondamentale, légèrement différente et non observée. Pour plus d'exactitude, cet avantage étant en outre minime, nous ne considérerons que l'expression (2.11) dans la suite de ce document. Ce choix n'est pas déterminant : par exemple, pour $f_0 = 196$ Hz (sol 2) et $\beta = 2.10^{-4}$, la différence de fréquence est de l'ordre de $0,6$ Hz pour les partiels autour de 7000 Hz.

Par ailleurs, en inversant l'équation (2.11), nous établissons l'expression du nombre maximal H de partiels dont les fréquences sont comprises entre 0 et la fréquence de Nyquist $\frac{f_s}{2}$:

$$H = \left\lfloor \frac{f_s}{2 f_0} \sqrt{\frac{2}{\sqrt{1 + \beta \frac{f_s^2}{f_0^2}} + 1}} \right\rfloor \tag{2.17}$$

où $\lfloor . \rfloor$ désigne la fonction partie entière.

2.2.2 Estimation du coefficient d'inharmonicité

Nous nous intéressons maintenant à l'estimation du coefficient d'inharmonicité d'une note présente dans une trame de signal x. Nous supposons que la note est donnée, c'est-à-dire que sa fréquence fondamentale est approximativement connue, avec une précision d'un quart de ton, et nous cherchons à estimer précisément la valeur de la fréquence fondamentale et du coefficient d'inharmonicité, en utilisant la loi d'inharmonicité (2.11). Nous verrons que cette estimation n'est pas triviale [Galembo et Askenfelt, 1994, 1999; Rauhala *et al.*, 2007] et proposerons deux méthodes d'estimation : la première consiste à extraire les fréquences présentes puis à réaliser l'estimation en utilisant directement la loi d'inharmonicité alors que la seconde est une optimisation, par rapport aux deux paramètres à estimer, d'une fonction de type produit spectral inharmonique.

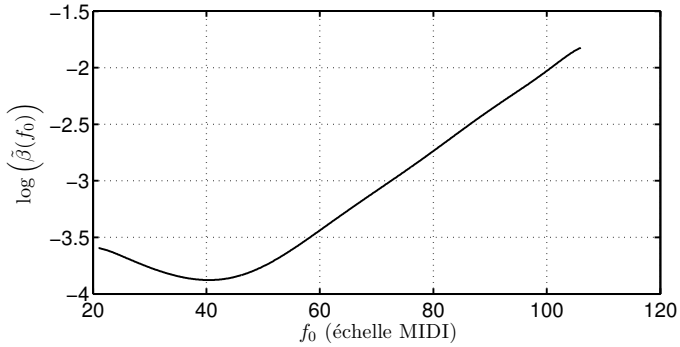
Auparavant, nous introduisons une courbe d'inharmonicité moyenne $f_0 \mapsto \tilde{\beta}(f_0)$, représentée sur la figure 2.3(a). Elle provient de l'interpolation d'une courbe extraite de [Fletcher et Rossing, 1998, p. 365], représentant un ordre de grandeur de l'inharmonicité en fonction des notes du piano. À part dans l'extrême grave, l'inharmonicité augmente avec la fréquence fondamentale. L'augmentation a une allure linéaire si l'on considère le logarithme des deux quantités, comme le montre la figure 2.3(a), en raison de la relation entre la longueur et le diamètre des cordes et la fréquence fondamentale [Young, 1952]. L'inharmonicité des notes les plus graves ne suit pas cette évolution en raison de différences dans la facture de leurs cordes (longueur, filage, tension).

La figure 2.3(b) montre un exemple de spectre de note de piano pour lequel nous avons fait coïncider un peigne inharmonique en utilisant les valeurs de la courbe moyenne $\tilde{\beta}(f_0)$. L'optimisation a été faite en prenant le maximum du produit spectral inharmonique $f_0 \mapsto \prod_{h=1}^{H} \left| X\left(hf_0\sqrt{1+\tilde{\beta}(f_0)\,h^2}\right)\right|^2$, où $X(f)$ est le spectre à la fréquence f et H est le nombre de partiels. Un peigne harmonique optimal, obtenu en maximisant la fonction $f_0 \mapsto \prod_{h=1}^{H} |X(hf_0)|^2$ a également été représenté. Le peigne inharmonique optimal parvient à sélectionner correctement les 17 premiers partiels sur un total de 37 partiels. À partir du 18e ou 19e (env. 3600 Hz), ses branches ne s'apparient plus aux lobes principaux. En comparaison, et alors que l'inharmonicité est relativement faible (environ 2.10^{-4}), le peigne *harmonique* optimal ne parvient pas à se superposer aux partiels au-delà du 12e (2500 Hz), malgré une compensation sur la fréquence fondamentale (197,9 Hz, soit 3 Hz de plus qu'avec le peigne inharmonique).

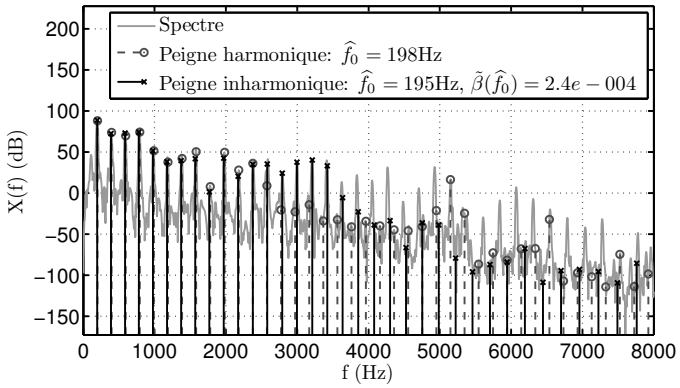
Cet exemple montre que même si l'introduction d'une inharmonicité moyenne améliore l'adéquation entre modèle de spectre et données, le gain est limité en raison des écarts, dus à des différences de facture, entre les valeurs moyennes $\tilde{\beta}(f_0)$ et l'inharmonicité réelle. Il convient alors de mesurer les coefficients d'inharmonicité au cas par cas et de quantifier ces erreurs.

Méthode 1 : régression sur la loi d'inharmonicité

La solution la plus directe consiste à évaluer conjointement f_0 et β dans la relation (2.11) à partir de la donnée des fréquences présentes et de l'ordre associé des partiels. Plus précisément, en supposant que l'on a extrait les fréquences $\{f_j\}_{j\in[\![1;j_{\max}]\!]}$ des sinusoïdes présentes, considérées comme des partiels, et qu'on leur a associé un ordre de partiel h_j, la relation $f_j = h_j f_0\sqrt{1+\beta h_j^2}$ (équation (2.11)) se réécrit, après élévation au carré,



(a) Courbe d'inharmonicité moyenne de référence $\tilde{\beta}(f_0)$ en fonction de la fréquence fondamentale.



(b) Estimation d'un peigne harmonique et d'un peigne inharmonique d'inharmonicité $\tilde{\beta}(f_0)$ sur une note de piano (sol 2) analysée sur 93 ms.

Figure 2.3 – Inharmonicité moyenne.

$$\forall j \in [\![1; j_{\max}]\!], \quad y_j = a x_j + b \tag{2.18}$$

avec

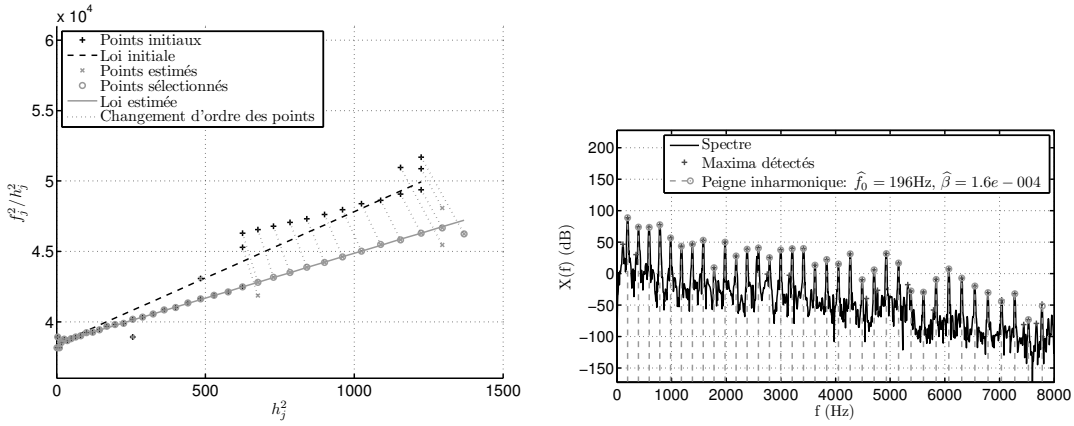
$$x_j \triangleq h_j^2, \quad y_j \triangleq \frac{f_j^2}{h_j^2}, \qquad a \triangleq f_0^2 \beta, \qquad b \triangleq f_0^2 \tag{2.19}$$

En d'autres termes, il existe une relation linéaire entre les données y_j et x_j. La pente $\widehat{a}$ et l'ordonnée à l'origine $\widehat{b}$ obtenues par régression linéaire conduisent alors à une estimation $\widehat{f_0} = \sqrt{\widehat{b}}$ et $\widehat{\beta} = \frac{\widehat{a}}{\widehat{b}}$.

Cependant, nous avons supposé qu'à chaque fréquence f_j était associé un ordre de partiel h_j connu. Le calcul de h_j pose en pratique une difficulté : d'après (2.11), son expression étant

$$h_j \triangleq \left[\frac{f_j}{f_0} \sqrt{\frac{2}{\sqrt{1 + 4\beta \frac{f_j^2}{f_0^2}} + 1}} \right] \tag{2.20}$$

où [.] désigne l'arrondi à l'entier le plus proche, il présuppose la connaissance de (f_0, β). Des valeurs approximatives de (f_0, β) induisent des erreurs pour les ordres h_j élevés, comme illustré sur la figure 2.4(a), où l'on voit le placement initial des points (x_j, y_j) en choisissant le couple $\left(\widehat{f_0}, \tilde{\beta}\left(\widehat{f_0}\right)\right) = \left(195, 2, 4.10^{-4}\right)$ obtenu précédemment avec une courbe d'inharmonicité moyenne, et la droite associée. Les points sont alignés jusqu'au 25e partiel ($h_j^2 = 625$), avant que l'estimation de l'ordre des partiels soit erronée : globalement, la régression linéaire mène alors à un résultat faussé. Afin d'éviter ce phénomène, une solution consisterait à choisir plusieurs valeurs *a priori* de (f_0, β) et à effectuer plusieurs régressions à partir de cette grille pour ne garder que l'optimale (au sens de l'erreur quadratique moyenne par exemple). Nous proposons une autre solution moins coûteuse qui s'appuie sur le fait que plus la connaissance des paramètres f_0 et β est précise, plus le premier ordre erroné intervient pour un ordre élevé. Il suffit de ne considérer d'abord que les premières fréquences, dont le calcul des ordres est fiable, pour estimer des valeurs plausibles de (f_0, β), puis d'itérer le processus en incluant davantage de fréquences et en recalculant les ordres. C'est ce qui est réalisé par l'algorithme 2.1, illustré sur la figure 2.4(a) où l'on constate un alignement des points et une régression corrects.



(a) Régression linéaire dans le plan $\left(h_j^2, \frac{f_j^2}{h_j^2}\right)$.

(b) Représentation du peigne inharmonique.

FIGURE 2.4 – Optimisation de la loi d'inharmonicité par rapport à f_0 et β (même signal que sur la figure 2.3(b)). Dans la version électronique de ce document, il est possible de cliquer sur la figure 2.4(a) pour visualiser le déroulement de l'algorithme.

Le peigne résultant (figure 2.4(b)) s'ajuste parfaitement sur le spectre et la totalité des 37 partiels a été correctement identifiée. L'optimisation a pu se faire malgré la présence de quelques pics spectraux parasites qui apparaissent sur les figures 2.4(a) et 2.4(b).

Méthode 2 : optimisation d'une fonction de détection

La première méthode d'estimation de l'inharmonicité se révèle efficace mais nécessite une extraction fiable des fréquences et souffre d'une complexité un peu élevée en raison du nombre de régressions à effectuer (environ H régressions avec un coût cubique, soit une complexité totale en $\mathcal{O}\left(H^4\right)$). Nous proposons donc une seconde approche pour l'estimation conjointe de β et de f_0, en laissant de côté l'aspect paramétrique lié à l'estimation des fréquences présentes dans le signal. Elle consiste à maximiser le produit spectral inharmo-

Entrées: Fréquences des partiels potentiels $\{f_j\}_{j=1,\dots,j_{\max}}$, valeurs initiales de fréquence fondamentale f_0^i et d'inharmonicité β^i, nombre minimal de partiels $H_{\min}$ pour réaliser une régression

{Initialisation}

$f_0 \leftarrow f_0^i$

$\beta \leftarrow \beta^i$

$H \leftarrow H_{\min}$

Pour $j \in [\![1; j_{\max}]\!]$

$h_j \leftarrow h(f_j, f_0, \beta)$ {via équation (2.20)}

Fin Pour

{Itérations}

Tant que $H \leq \#\{h_j\}_{j=1,\dots,j_{\max}}$

{Sélection des fréquences}

$J \leftarrow \left\{ j_0 \in [\![1; j_{\max}]\!] \ / \ \#\{h_j \ / \ h_j \leq h_{j_0}\}_{j=1,\dots,j_{\max}} \leq H \right\}$ {Sélection des H premiers ordres distincts présents}

$J \leftarrow \left\{ j_0 \in J \ / \ j_0 = \underset{\substack{j \in J \\ h_j = h_{j_0}}}{\arg\min} \left| f_j - h_j f_0 \sqrt{1 + \beta h_j^2} \right| \right\}$ {En cas d'occurrences multiples d'un ordre, sélection de la fréquence la plus proche de la fréquence théorique}

{Estimation des paramètres}

Pour $j \in J$

$x_j \leftarrow n_j^2, y_j \leftarrow \frac{f_j^2}{n_j^2}$

Fin Pour

$\begin{pmatrix} a \\ b \end{pmatrix} \leftarrow \left([x_j, 1]_{j \in J} \right)^+ [y_j]_{j \in J}$

$f_0 \leftarrow \sqrt{b}, \beta \leftarrow \frac{a}{b}$

Pour $j \in [\![1; j_{\max}]\!]$ {Mise à jour des ordres}

$h_j \leftarrow h(f_j, f_0, \beta)$ {via équation (2.20)}

Fin Pour

$H \leftarrow H + 1$

Fin Tant que

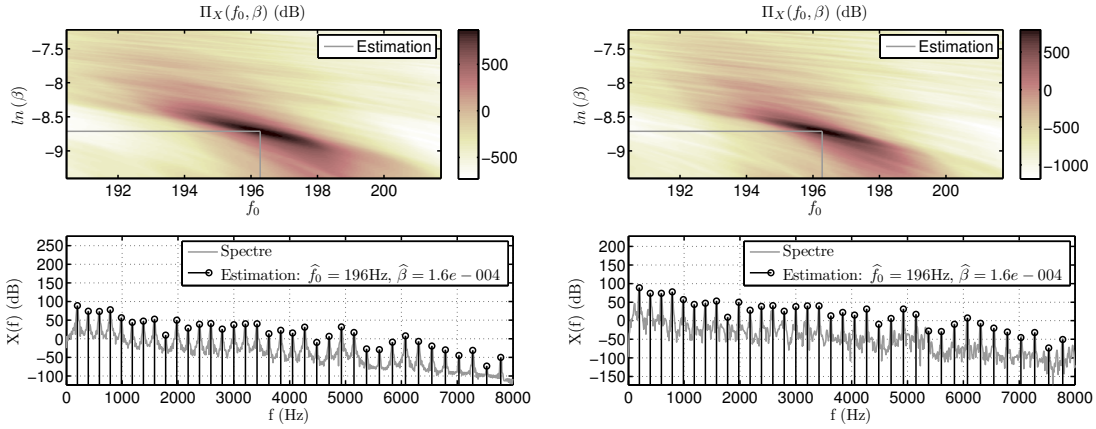
Sorties: f_0, β

Algorithme 2.1: Régression sur la loi d'inharmonicité

nique (ou toute autre fonction équivalente) défini par

$$\Pi_X : (f_0, \beta) \mapsto \prod_{h=1}^{H} \left| X\left(h f_0 \sqrt{1 + \beta\,(f_0)\, h^2}\right)\right|^2 \tag{2.21}$$

Cette fonction est particulièrement bien appropriée ici car elle présente des maxima très marqués. Les techniques d'optimisation numérique classiques [1] peuvent être appliquées en raison de la régularité de cette fonction, localement, sur les domaines de variations de f_0 et β. Cette régularité est due aux lobes principaux de X autour de chaque partiel, et dont la multiplication dans (2.21) crée un maximum local au niveau de la fréquence fondamentale et de l'inharmonicité optimales. La figure 2.5 montre l'allure de ce produit spectral en deux dimensions autour de son maximum. Alors que la fonction est un peu plus régulière dans le cas d'un signal synthétique composé uniquement de sinusoïdes aux fréquences théoriques des partiels (figure 2.5(a)) que dans le cas d'un son réel contenant des sinusoïdes parasites et du bruit (figure 2.5(b)), l'estimation donne des résultats satisfaisants dans les deux cas.



(a) Signal synthétique

(b) Son réel

FIGURE 2.5 – Optimisation du produit spectral par rapport à f_0 et β sur un signal synthétique (à gauche) et sur un son réel (à droite, même signal que sur la figure 2.3(b)). Dans chaque cas, le produit spectral est calculé et maximisé localement (en haut) pour aligner le peigne avec le spectre (en bas).

À la différence de la première méthode, cette approche n'utilise pas la loi d'inharmonicité comme un critère à optimiser explicitement mais l'intègre dans la fonction Π_X. Le calcul de Π_X étant peu coûteux, l'estimation se fait ici à moindre coût. Par ailleurs, notons que l'emploi du produit spectral semble être particulièrement approprié pour l'estimation de β : il a déjà été utilisé à cet effet par Galembo et Askenfelt [1994], qui calculent un produit spectral harmonique et étudient l'élargissement de ses pics en fonction de l'inharmonicité.

2.2.3 Impact de la prise en compte de l'inharmonicité

Nous souhaitons mesurer l'apport de la prise en compte de l'inharmonicité lors de la modélisation du spectre d'une note de piano, dans la perspective de toute application impliquant une indexation de sons pseudo-harmoniques (la transcription mais également le

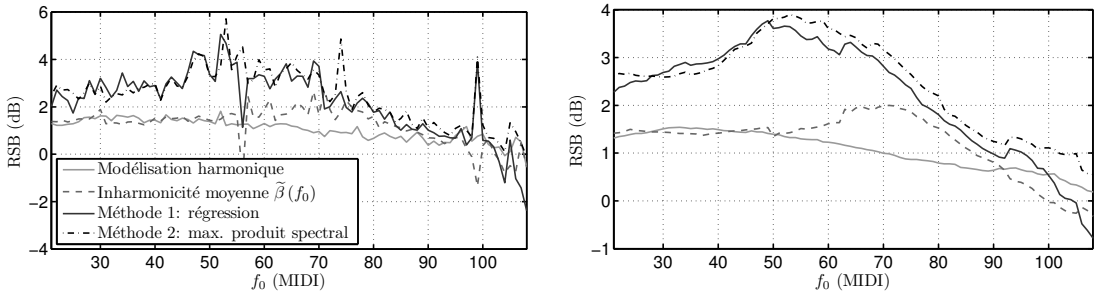
1. Nous utilisons ici la fonction `fminsearch` sous Matlab.

codage par exemple). Pour ce faire, nous appliquons les méthodes d'estimation précédemment étudiées sur des enregistrements de notes isolées de piano, sur toute la tessiture. Chacune permet d'estimer un peigne correspondant aux partiels présents, et nous cherchons à quantifier l'adéquation du peigne avec le signal, c'est-à-dire la validité du modèle et la performance de la méthode d'estimation utilisée. Les résultats sont établis pour les modèles et méthodes suivants :

- estimation d'un peigne harmonique (inharmonicité nulle) ;
- estimation d'un peigne d'inharmonicité moyenne $\tilde{\beta}(f_0)$;
- estimation d'un peigne inharmonique par régression sur la loi d'inharmonicité (méthode 1) ;
- estimation d'un peigne inharmonique par optimisation du produit spectral en deux dimensions (méthode 2).

Une fois les paramètres de fréquence fondamentale et d'inharmonicité estimés par une méthode donnée, les fréquences des partiels sont obtenues par l'équation (2.11) (p. 54) puis utilisées pour calculer les amplitudes associées par moindres carrés sur le signal original. Le résiduel est alors déduit en soustrayant le signal estimé au signal original. Le signal estimé contient donc les partiels qui ont été correctement identifiés, alors que le résiduel contient à la fois les partiels mal estimés et le reste du signal (bruit ambiant, bruits impulsionnels, modes longitudinaux). Un rapport signal à bruit (RSB), ratio entre l'énergie du signal estimé et du bruit résiduel, est ensuite calculé pour déterminer l'efficacité de la méthode.

La figure 2.6 représente les résultats obtenus. Ils proviennent de l'analyse de sons de 7 pianos (extraits de la base présentée dans la partie 6.2 (p. 138)), avec trois nuances différentes, soit 21 sons par note, et 1848 sons au total. Chaque analyse est réalisée sur une trame de 93 ms, échantillonnée à 16 kHz et prise 50 ms après l'attaque. L'optimisation du produit spectral a été effectuée sur une grille 25×25 pour des valeurs logarithmiquement réparties de la fréquence fondamentale sur un demi-ton et du coefficient d'inharmonicité sur un intervalle allant d'un tiers à trois fois l'inharmonicité moyenne $\tilde{\beta}(f_0)$ de la note. Un algorithme d'optimisation numérique (fonction `fminsearch` sous Matlab) est ensuite appliqué pour affiner la valeur maximale obtenue sur la grille.



(a) Résultats par note.

(b) Résultats moyennés sur une octave glissante.

Figure 2.6 – Rapport signal à bruit obtenu pour la séparation du contenu pseudo-harmonique et du bruit résiduel de notes de piano avec plusieurs méthodes d'estimation de la fréquence fondamentale et de l'inharmonicité. Pour plus de clarté, les résultats lissés sur une octave sont également représentés.

Alors qu'il n'était pas évident d'évaluer *a priori* le gain d'une prise en compte de l'inharmonicité dans nos modèles, nous voyons maintenant que l'amélioration du RSB en utilisant les méthodes 1 et 2 est significative. Dans le milieu du registre, ces méthodes

permettent de bien identifier l'ensemble des partiels, contrairement à celles utilisant un spectre purement harmonique ou une inharmonicité moyenne qui, comme dans l'exemple vu plus haut, laissent échapper les partiels au-delà d'un certain ordre. Il est intéressant de noter que ces méthodes sont efficaces dans le registre grave du piano, alors que la taille de la trame utilisée pourrait provoquer des problèmes de résolution temps-fréquence. Dans l'aigu du registre, les méthodes restent efficaces jusqu'à la dernière octave où le faible nombre de partiels met toutes les méthodes à égalité. L'emploi d'une inharmonicité moyenne n'a qu'une efficacité limitée à une partie du registre, entre les notes 60 et 90, où l'on voit une amélioration par rapport au cas harmonique, amélioration qui reste cependant en dessous des performances des méthodes 1 et 2.

2.3 Modélisation des enveloppes spectrales de sons de piano

La modélisation sinusoïdale que nous avons abordée dans la première partie de ce chapitre considère les composantes comme des entités individuelles sans rapport entre elles. Dans le cas d'une note, nous avons vu dans la deuxième partie qu'elles pouvaient être liées par la distribution de leurs fréquences. La question de l'enveloppe spectrale introduit un autre rapport entre les composantes, celui qu'entretiennent leurs amplitudes.

Nous abordons cette question dans la perspective de l'estimation de fréquences fondamentales multiples que nous verrons au chapitre 4. En effet, comme nous l'avons vu dans le chapitre 1, si la seule information sur la (quasi-)harmonicité des composantes suffit à poser le problème de l'estimation de fréquences fondamentales simples, le problème est mal posé dans le cas de fréquences fondamentales multiples en raison de l'indétermination d'octave. L'information d'enveloppe spectrale permet alors de mieux l'aborder.

Nous présenterons dans un premier temps (partie 2.3.1) le modèle du processus harmonique, ou comment considérer les amplitudes comme des variables aléatoires dont les paramètres statistiques peuvent porter l'information sur les enveloppes spectrales. Dans un second temps (partie 2.3.2), nous proposerons un modèle autorégressif d'enveloppe spectrale, expliciterons la méthode d'estimation associée et l'appliquerons aux sons de piano.

2.3.1 Le processus harmonique

Contrairement à ce que son nom peut faire croire, le *processus harmonique* ne désigne pas forcément une entité périodique, présentant une distribution *harmonique* de fréquences. Pour éviter toute confusion, il est utile de préciser que dans cette appellation consacrée, le terme *harmonique* fait référence aux sinusoïdes qui composent le signal, sans hypothèse sur leurs fréquences.

Définition et propriétés

Le **processus harmonique** désigne un modèle dans lequel le processus observé $s(n)$ est une somme de H sinusoïdes, dont les amplitudes complexes sont des variables aléatoires :

$$s(n) = \sum_{h=1}^{H} \alpha_h e^{2i\pi f_h n} \tag{2.22}$$

où les amplitudes complexes [2] α_h sont des variables aléatoires décorrélées, centrées et de variance σ_h^2. La version réelle du processus consiste à considérer le signal $2\mathrm{Re}\,(s(n))$.

$s(n)$ est centré, stationnaire au sens large de covariance

$$\gamma_s(k) \triangleq \mathbb{E}\left[s\left(n\right)s^*\left(n+k\right)\right] \tag{2.23}$$

$$= \sum_{h=1}^{H} \sigma_h^2 e^{-2i\pi f_h k} \tag{2.24}$$

En particulier, dans le cas où les fréquences des composantes sont les multiples d'une même fréquence fondamentale f_0 (par exemple si $f_h = hf_0$), les maxima de la partie réelle de la covariance sont situés aux multiples de la période fondamentale $\frac{1}{f_0}$. On peut ainsi déduire les estimateurs de fréquence fondamentale à base d'autocorrélation tels que YIN [de Cheveigné et Kawahara, 2002].

Par ailleurs, soulignons que $\gamma_s(k)$ n'étant pas sommable, le processus n'a pas de densité spectrale de puissance.

Périodogramme, estimation des variances des amplitudes

Soit $S(f)$ la transformée de Fourier discrète de N échantillons successifs $s(0), \ldots, s(N-1)$ pondérés par une fenêtre $w(n)$:

$$S(f) = \sum_{n=0}^{N-1} s(n)w(n)e^{-2i\pi fn} \tag{2.25}$$

$$= \sum_{h=1}^{H} \alpha_h W(f - f_h) \tag{2.26}$$

Le périodogramme $|S(f)|^2$ a alors comme propriété

$$\mathbb{E}\left[|S(f)|^2\right] = \sum_{h=1}^{H} \sigma_h^2 \left|W(f - f_h)\right|^2 \tag{2.27}$$

Ainsi, pour $h \in [\![1; H]\!]$, si pour tout $h' \in [\![1; H]\!]$ les fréquences f_h est $f_{h'}$ sont suffisamment espacées (ou de manière équivalente si N est assez élevé) pour que l'on puisse négliger $W\left(f_h - f_{h'}\right)$ devant $W(0)$, alors $\left|\frac{S(f_h)}{W(0)}\right|^2$ est un estimateur sans biais de σ_h^2.

Prédictibilité, loi du processus harmonique et de sa transformée de Fourier discrète

Plaçons-nous dans le cas, généralement utilisé, où le nombre N d'échantillons observés est strictement supérieur au nombre $2H$ de variables aléatoires réelles définissant le processus (deux variables aléatoires réelles par amplitude complexe). Nous avons, sous forme matricielle,

$$s = E\alpha \text{ avec } \begin{cases} s & \triangleq (s(0), \ldots, s(N-1))^{\mathrm{t}} \\ [E]_{n,h} & \triangleq e^{2i\pi f_h n} \\ \alpha & \triangleq (\alpha_1, \ldots, \alpha_H)^{\mathrm{t}} \end{cases} \tag{2.28}$$

2. Une autre définition équivalente consiste à prendre des amplitudes réelles, étant toujours décorrélées, centrées et de variance σ_h^2, et des phases initiales φ_h *i.i.d.* selon une loi uniforme sur $[0; 2\pi[$, indépendantes des amplitudes.

et

$$S = W\alpha \text{ avec } \begin{cases} S & \triangleq \left(S(0), \ldots, S\left(\frac{N-1}{N}\right)\right)^t \\ [W]_{f,h} & \triangleq W\left(f - f_h\right) \end{cases} \tag{2.29}$$

α étant un vecteur gaussien de loi $\mathcal{N}(0, \Sigma)$ avec $\Sigma \triangleq \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_H^2 \end{pmatrix}$, s et S sont des vecteurs gaussiens centrés de covariances respectives $E\Sigma E^\dagger$ et $W\Sigma W^\dagger$. Leur rang étant borné par $\mathrm{rg}(\Sigma) = H < N$, leur déterminant est nul. Ces deux vecteurs n'ont donc pas de densité de probabilité.

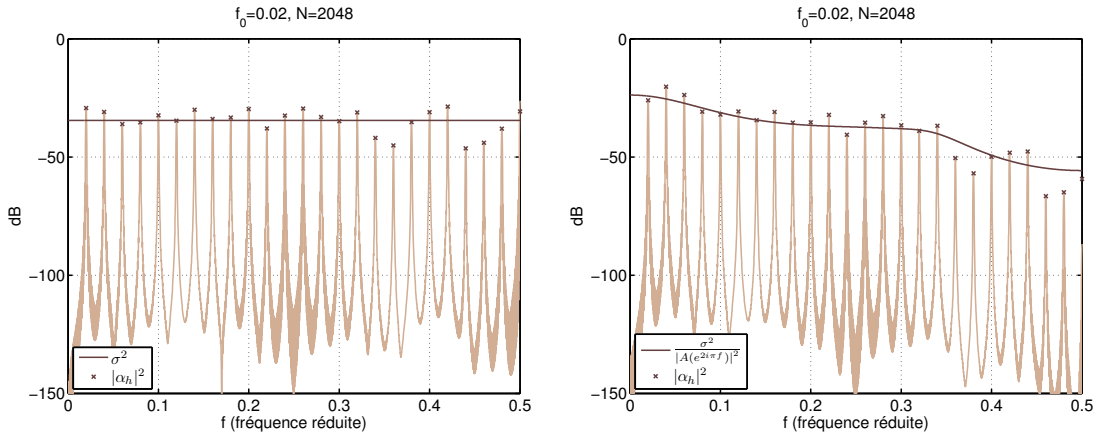
Exemple de réalisation d'un processus harmonique

Un exemple de réalisation est représenté sur la figure 2.7. Nous avons généré une réalisation pour deux processus harmoniques en distribuant les fréquences des composantes selon les multiples d'une fréquence fondamentale f_0. La modélisation d'une enveloppe spectrale se fait par le choix sur les variances des amplitudes : la variance peut par exemple être constante, ou suivre une enveloppe autorégressive choisie au préalable, de paramètres $\left(\sigma^2, A(z)\right)$ (tel que définis dans l'annexe A.2.1 (p. 165)), afin d'obtenir une enveloppe spectrale variable et régulière. Dans ce cas, pour chaque composante h, l'amplitude est alors tirée selon une loi normale centrée, de variance fixée à $\sigma_h^2 \triangleq \frac{\sigma^2}{\left|A\left(e^{2i\pi f_h}\right)\right|^2}$. Cet exemple illustre la raison pour laquelle nous utiliserons le processus harmonique : il permet d'introduire une connaissance sur l'enveloppe spectrale tout en laissant un degré de liberté quant aux réalisations de ce modèle. Les amplitudes obtenues forment ainsi une enveloppe spectrale dont l'allure suit le modèle d'enveloppe avec quelques écarts. La conjonction de cette contrainte et de ce degré de liberté nous sera utile dans le cadre de l'estimation de fréquences fondamentales multiples (chapitre 4), pour les cas de recouvrement entre les spectres de note.

2.3.2 Modèle autorégressif d'enveloppe spectrale

Dans la partie 1.5.1 (p. 39), nous avons vu l'état de l'art sur la modélisation physique de l'enveloppe spectrale pour le piano : d'une part, l'excitation d'un partiel dépend de plusieurs régimes de fonctionnement en fonction de son ordre et de la fréquence fondamentale ; d'autre part, l'évolution temporelle de son amplitude fait intervenir simultanément deux phénomènes principaux, conséquences du couplage entre les modes : les battements et la double décroissance. Pour ces deux raisons, il n'est pas envisageable de modéliser l'enveloppe spectrale de façon simple, par exemple par un filtre à moyenne ajustée (MA) comme le suggère, en première approximation, l'allure de l'amplitude des partiels au niveau du chevalet à l'instant de la frappe (équation (1.18) (p. 42)). L'utilisation d'un modèle physique général (et donc plus complexe) de l'amplitude d'un partiel en fonction de son ordre, de sa fréquence fondamentale et du temps, pose alors plusieurs difficultés dans le cadre de l'analyse et de l'indexation de morceaux de piano :

- les modèles physiques impliqués sont régis par un nombre important de paramètres physiques variables, en fonction de l'instrument ou des conditions d'enregistrement par exemple, et dont l'estimation n'est pas triviale ;
- la polyphonie introduit une dimension supplémentaire qui rend le problème hautement plus complexe ;



(a) (b)

Figure 2.7 – Périodogrammes de réalisations de deux processus harmoniques observés sur N échantillons. L'information sur l'enveloppe spectrale est introduite via les variances des amplitudes. À gauche, les amplitudes du processus sont *i.i.d.* normales centrées de variance σ^2, le modèle d'enveloppe spectrale est une constante. À droite, pour obtenir une enveloppe spectrale variable avec un aspect régulier, des pôles d'un filtre autorégressif $A(z)$ ont été choisis aléatoirement dans le disque de rayon $1/2$, et l'enveloppe autorégressive paramétrée par σ^2 et $A(z)$ a été utilisée pour définir les variances des composantes : pour $h \in [\![1; H]\!]$, on a $\sigma_h^2 = \frac{\sigma^2}{\left|A\left(e^{2i\pi f_h}\right)\right|^2}$. Les amplitudes α_h sont des réalisations de lois gaussiennes $\mathcal{N}\left(0, \sigma_h^2\right)$.

– des données peuvent être amenées à manquer. Par exemple, un morceau ne présente *a priori* pas l'ensemble des notes du piano de manière exhaustive.

Une telle modélisation ne garantit donc pas que l'on puisse déterminer la solution du problème inverse ainsi posé. Aussi, il semble plus réaliste d'utiliser un modèle plus simple. Plusieurs solutions sont envisageables. Apprendre un dictionnaire d'enveloppes spectrales ou temporelles de partiels est une méthode qui présente l'inconvénient d'être insensible, et donc peu robuste, à la variabilité que nous venons de décrire. Pour la même raison, nous évitons l'utilisation d'une loi *a priori* sur les amplitudes comme ceux décrits dans la partie 1.4.4.

La modélisation d'une enveloppe spectrale par un filtre autorégressif (AR) est largement répandue dans le cas de la voix [Atal et Hanauer, 1971]. On peut alors interpréter ce filtre et ses pôles comme un modèle physique du conduit vocal et de ses résonances. Dans le cas du piano, une modélisation AR ne traduit pas un tel caractère physique. Elle présente cependant l'intérêt de bien modéliser une enveloppe lisse avec un ordre faible, et d'être assez générique. Elle définit ainsi un cadre pour modéliser une notion équivalente à la *spectral smoothness* [Klapuri, 2003].

La méthode par prédiction linéaire [Makhoul, 1975] permet d'estimer les paramètres $\left(\sigma^2, a_1, \ldots, a_P\right)$ d'un processus AR d'ordre P tel que défini dans l'annexe A.2.1 (p. 165) en résolvant les équations de Yule-Walker pour minimiser l'erreur de prédiction. En notant $\mathbf{R}$ la matrice de Toeplitz dont l'élément (m, n) est la valeur $R\left(m-n\right)$ d'une autocorrélation empirique R et $\mathbf{a} \triangleq \left(\frac{1}{\sigma^2}, -\frac{a_1}{\sigma^2}, \ldots, -\frac{a_P}{\sigma^2}\right)^{\mathrm{t}}$, les équations de Yule-Walker s'écrivent

$$\mathbf{R}\mathbf{a} = (1, 0, \ldots, 0)^{\mathrm{t}} \tag{2.30}$$

Dans cette expression, l'autocorrélation empirique a été substituée à l'autocorrélation et la résolution de l'équation obtenue donne une bonne estimation des paramètres AR lorsque l'on observe effectivement un processus AR.

Considérons maintenant un signal x, et son spectre de puissance $|X(f)|^2$, composé de H partiels aux fréquences $\{f_1, \ldots, f_H\}$. Leur enveloppe spectrale est modélisée par un modèle AR d'ordre P. Comme précédemment, celui-ci est paramétré par les coefficients $\left(\sigma^2, a_1, \ldots, a_P\right)$, la réponse fréquentielle du filtre AR étant $\frac{\sigma}{1-\sum_{k=1}^{P} a_k e^{-i2\pi fk}}$. Dans la mesure où l'on n'observe de la réalisation de ce processus AR qu'un échantillonnage de son périodogramme aux fréquences des partiels, on peut exprimer une autocorrélation empirique du signal par

$$R(k) \triangleq \frac{1}{H} \sum_{h=1}^{H} 2 \left|X\left(f_h\right)\right|^2 \cos\left(2\pi f_h k\right) \tag{2.31}$$

L'estimation AR par la méthode précédente souffre alors d'un défaut : dans le domaine spectral, le spectre observé est un échantillonnage de la réponse fréquentielle du modèle AR au niveau des fréquences des partiels ; il en résulte un repliement dans le domaine temporel, affectant en particulier la fonction d'autocorrélation utilisée lors de la résolution. Ce phénomène a été étudié par El-Jaroudi et Makhoul [1991] qui ont montré comment intégrer ce repliement pour modifier les équations de Yule-Walker. La relation (2.30) devient alors

$$\mathbf{R}\mathbf{a} = \widehat{\mathbf{R}}_{\mathbf{a}}\mathbf{a} \tag{2.32}$$

$\widehat{\mathbf{R}}_{\mathbf{a}}$ étant la matrice de Toeplitz dont l'élément (m, n) est l'autocorrélation $\widehat{R}_{\mathbf{a}}(m-n)$ estimée à partir de la version échantillonnée de la réponse en fréquence du modèle AR et définie par

$$\widehat{R}_{\mathbf{a}}(k) \triangleq \frac{1}{H} \sum_{h=1}^{H} 2 \frac{\sigma^2}{\left|1 - \sum_{k=1}^{P} a_k e^{-i2\pi f_h k}\right|^2} \cos\left(2\pi f_h k\right) \tag{2.33}$$

La résolution de l'équation (2.32) s'effectue en général via un algorithme itératif reposant sur le fait que la solution $\mathbf{a}^*$ est un point fixe de la fonction $\varphi : \mathbf{a} \mapsto \mathbf{R}^{-1}\widehat{\mathbf{R}}_{\mathbf{a}}\mathbf{a}$. Il consiste alors à construire et à estimer la limite de la suite $\left(\mathbf{a}^{(n)}\right)$ définie par

$$\mathbf{a}^{(0)} \triangleq (1, 0, \ldots, 0) \tag{2.34}$$

$$\forall n > 0, \mathbf{a}^{(n)} \triangleq \mathbf{R}^{-1}\widehat{\mathbf{R}}_{\mathbf{a}^{(n-1)}}\mathbf{a}^{(n-1)} \tag{2.35}$$

$\mathbf{R}$ étant une matrice de Toeplitz, la résolution du système par l'algorithme de Levinson est rapide. La convergence de la suite $\left(\mathbf{a}^{(n)}\right)$ peut être accélérée de plusieurs façons. Dans une approche à base de gradient, El-Jaroudi et Makhoul [1991] introduisent un coefficient $\alpha \in [0; 1]$ de mise à jour (typiquement, $\alpha = 0,5$) et redéfinissent la suite $\left(\mathbf{a}^{(n)}\right)$ par

$$\mathbf{a}^{(0)} \triangleq (1, 0, \ldots, 0) \tag{2.36}$$

$$\forall n > 0, \mathbf{a}^{(n)} \triangleq (1-\alpha)\,\mathbf{a}^{(n-1)} + \alpha\mathbf{R}^{-1}\widehat{\mathbf{R}}_{\mathbf{a}^{(n-1)}}\mathbf{a}^{(n-1)} \tag{2.37}$$

Badeau et David [2008] proposent une convergence encore plus rapide en redéfinissant la suite $\left(\mathbf{a}^{(n)}\right)$ par

$$\mathbf{a}^{(0)} \triangleq (1, 0, \ldots, 0) \tag{2.38}$$

$$\forall n > 0, \mathbf{a}^{(n)} \triangleq \mathbb{P}\left(\mathbf{R}^{-1}\widehat{\mathbf{R}}_{\mathbf{a}^{(n-1)}}\mathbf{a}^{(n-1)}\right) \tag{2.39}$$

l'opérateur $\mathbb{P}(\mathbf{a})$, facultatif si l'on ne cherche pas une solution causale stable, faisant correspondre à tout modèle de paramètre $\mathbf{a}$ un modèle causal stable également solution (en inversant les modules des pôles situés à l'extérieur du cercle unité). Dans l'absolu, cette technique est plus rapide à condition que le coût lié à l'application de $\mathbb{P}$ soit faible, c'est-à-dire lorsque l'ordre du modèle AR est faible (le coût est *a priori* cubique en l'ordre du modèle AR).

La figure 2.8 illustre cette méthode sur deux exemples, pour lesquels le ratio entre le nombre de composantes sinusoïdales et le nombre de pôles est égal à 8 et à 2 respectivement. Dans les deux cas, l'estimation obtenue en prenant en compte le repliement temporel est meilleure que l'estimation traditionnelle. Lorsque le nombre de partiels diminue, l'estimation se dégrade relativement peu.

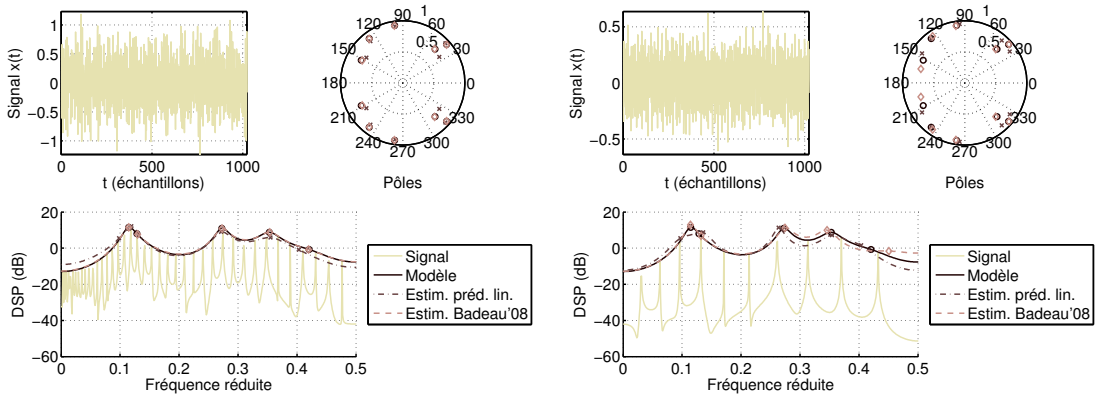
La méthode s'applique naturellement à un son de piano en considérant que les amplitudes des partiels coïncident avec un échantillonnage très clairsemé de la densité spectrale de puissance empirique du processus AR. Le caractère régulier de l'enveloppe spectrale est alors relativement bien modélisé par un processus AR, comme illustré sur la figure 2.9. Nous disposons ainsi d'un modèle d'enveloppe spectrale, voisin de celui couramment utilisé pour la parole, qui présente l'avantage de bien modéliser la variabilité des spectres rencontrés tout en introduisant une contrainte de régularité.

2.4 Modélisation de l'enveloppe du bruit

Nous avons vu au début de ce chapitre (partie 2.1 p. 49) que certains modèles sinusoïdaux se composent non seulement de sinusoïdes mais également de bruit additif. Il est en effet utile d'introduire un tel modèle afin de caractériser la partie non sinusoïdale, que l'on appelle selon le contexte bruit, résiduel, partie stochastique, etc. Elle contient en général le bruit de fond, le bruit de mesure – enregistrement, échantillonnage et traitements –, mais également les composantes large bande produites par les sources – un cas emblématique étant celui de la parole [Richard et d'Alessandro, 1996], dont la composante apériodique porte une grande quantité d'information. On modélise couramment le tout par un bruit coloré, c'est-à-dire résultant du filtrage d'un bruit blanc par un filtre donné.

Dans le cadre de l'estimation de fréquences fondamentales et plus généralement de la modélisation *sinus plus bruit*, le modèle de bruit doit présenter une qualité de discrimination vis-à-vis des sinusoïdes. Par exemple, un modèle AR de bruit n'est pas un choix judicieux car il est capable de bien modéliser une sinusoïde avec un pôle de module proche de 1. Dans ces conditions, il paraît difficile de garantir que les sinusoïdes seront identifiées en tant que telles et non comme du bruit. Ainsi, parmi les trois grandes familles de filtres – AR, MA et ARMA –, seuls les filtres MA sont discriminants à l'égard des sinusoïdes. C'est la raison pour laquelle nous proposons de modéliser la partie bruit par un processus MA, dont l'ordre sera suffisamment faible pour ne pas lui permettre de modéliser les sinusoïdes.

Nous définissons donc le bruit $x_b(t)$ comme résultant du filtrage d'un bruit blanc gaussien centré de puissance σ_b^2 par un filtre MA unitaire d'ordre Q_b dont la transformée en z



(a) Estimation avec 80 composantes complexes (b) Estimation avec 20 composantes complexes

FIGURE 2.8 – Exemple d'estimation d'enveloppe spectrale AR sur un signal synthétique : signal observé sur 1024 points, enveloppe spectrale AR d'ordre 10. Les fréquences des composantes suivent la loi $f_h = hf_0\sqrt{1+0,01h^2}$, $f_0 = 0,003$ (à gauche) et $f_0 = 0,03$ (à droite). L'estimation par la méthode de Badeau et David [2008] se confond avec le modèle original alors que l'estimation par prédiction linéaire s'en écarte.

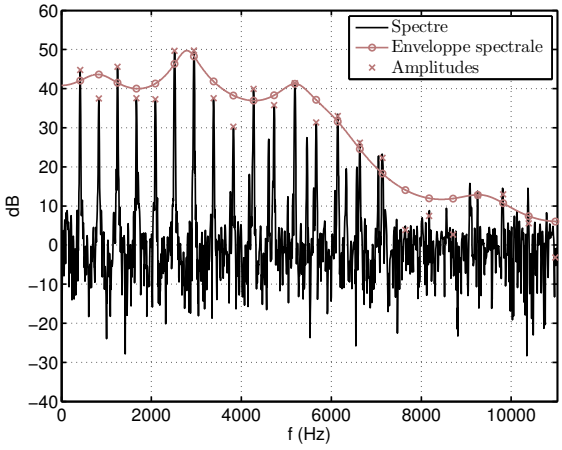


FIGURE 2.9 – Exemple d'estimation d'enveloppe spectrale AR sur un son de piano : les fréquences des partiels ont été présélectionnées pour extraire leurs amplitudes (croix) et estimer le modèle AR.

est

$$B(z) \triangleq \sum_{k=0}^{Q_b} b_k z^{-k} \tag{2.40}$$

avec $b_0 = 1$.

Nous réalisons l'estimation des paramètres du modèle par la méthode suivante, qui a le mérite d'être particulièrement rapide. Écrivons l'autocorrélation du processus MA (cf. équation (A.42) p. 170) sous la forme

$$\underline{r_b} = \mathbf{B}\,\underline{b} \tag{2.41}$$

avec

$$\underline{r_b} \triangleq \left(\frac{\mathbb{E}\left[x_b(t)\,x_b(t)\right]}{\mathbb{E}\left[x_b(t)^2\right]}, \ldots, \frac{\mathbb{E}\left[x_b(t)\,x_b(t+Q_b)\right]}{\mathbb{E}\left[x_b(t)^2\right]} \right)^{\mathrm{t}} \tag{2.42}$$

$$\mathbf{B} \triangleq \begin{pmatrix} b_0 & b_1 & \ldots & b_{Q_b} \\ 0 & b_0 & \ldots & b_{Q_b-1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & b_0 \end{pmatrix} \tag{2.43}$$

$$\underline{b} \triangleq (b_0, \ldots, b_{Q_b})^{\mathrm{t}} \tag{2.44}$$

En remplaçant $\underline{r_b}$ par l'autocorrélation empirique $\widehat{\underline{r_b}}$ calculée à partir des observations, l'algorithme 2.2 permet d'estimer $\underline{b}$ de manière itérative : il consiste à inverser dans la relation (2.41) l'estimation de la matrice $\mathbf{B}$ obtenue à l'itération précédente pour trouver une nouvelle estimation de $\underline{b}$.

Entrées: autocorrélation empirique $\widehat{r_b}$.
  $\widehat{\underline{b}} \leftarrow (1, 0, \ldots, 0)^{\mathrm{t}}$ {initialisation}
  Pour chaque itération n
    Pour $1 \leq i < j \leq Q_b + 1$,
      $\left[\widehat{\mathbf{B}}\right]_{i,j} \leftarrow \left[\widehat{\underline{b}}\right]_{j-i}$, $\left[\widehat{\mathbf{B}}\right]_{j,i} \leftarrow 0$.
    Fin Pour
    Résoudre $\widehat{\mathbf{B}}\,\widehat{\underline{b}} = \widehat{r_b}$ en $\widehat{\underline{b}}$ {résolution rapide du système triangulaire}
    $\widehat{\underline{b}} \leftarrow \widehat{\underline{b}} / \left[\widehat{\underline{b}}\right]_0$
  Fin Pour
Sorties: estimation $\widehat{\underline{b}}$ de $\underline{b}$.

Algorithme 2.2: Estimation itérative des paramètres du bruit

La résolution rapide de $\widehat{\mathbf{B}}\,\widehat{\underline{b}} = \widehat{r_b}$ en $\widehat{\underline{b}}$ par élimination récursive dans le système triangulaire (coût en $\mathcal{O}\left(Q_b^2\right)$) permet d'éviter l'inversion classique (en $\mathcal{O}\left(Q_b^3\right)$) de la matrice $\mathbf{B}$ et rend chaque itération peu coûteuse. La convergence a été observée pour une vingtaine d'itérations. L'estimation $\widehat{\sigma_b^2}$ de σ_b^2 est ensuite obtenue en fonction de l'autocovariance empirique $\hat{\gamma}(m)$ prise en 0 :

$$\widehat{\sigma_b^2} \triangleq \frac{\hat{\gamma}\left(0\right)}{\sum_{k=0}^{Q_b} b_k^2} \tag{2.45}$$

Un exemple d'estimation est représenté sur la figure 2.10. Un modèle MA réel d'ordre 20 a été généré de façon aléatoire puis estimé à partir d'une réalisation du processus sur $N = 1024$ points.

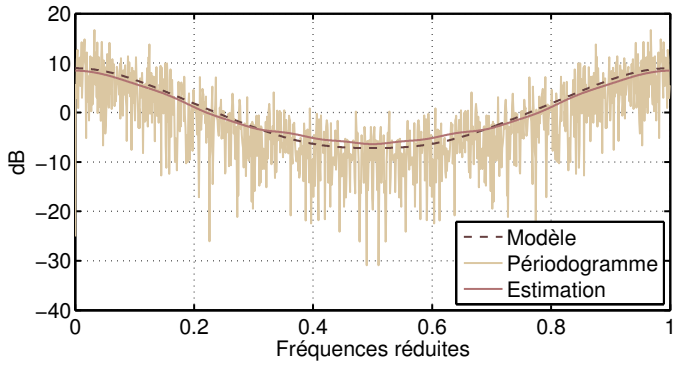


FIGURE 2.10 – Exemple d'estimation MA.

2.5 Conclusion

Au terme de ce chapitre, nous sommes en mesure de caractériser les sons de piano pour l'estimation de hauteur. Cette caractérisation intervient sur quatre axes : la modélisation du signal comme une somme de sinusoïdes et de bruit ; l'identification et la localisation fréquentielle précise des partiels de sons de piano ; la modélisation de l'enveloppe spectrale ; et enfin la modélisation du bruit. Nous disposons ainsi d'une palette d'outils que nous allons utiliser dans les deux prochains chapitres pour l'estimation de fréquences fondamentales.

Chapitre 3

Estimation à court terme de hauteur simple sur un registre étendu

L'estimation de hauteur constitue un sujet de recherche auquel de nombreux travaux ont été consacrés, mais qui continue à susciter des efforts et à donner lieu à de nouvelles méthodes. Pour aborder ce vaste problème, nous choisirons ici d'adopter un angle d'attaque particulier reposant sur le constat que l'efficacité de la plupart des méthodes d'estimation de hauteur chute lorsque l'on diminue la taille de la fenêtre d'analyse ou que l'on élargit l'intervalle des fréquences fondamentales possibles. Nous nous intéresserons donc à la robustesse des méthodes d'estimation de hauteur dans le contexte difficile que constituent ces deux conditions prises conjointement. Dans ce cadre, nous proposerons une méthode dont l'efficacité reste satisfaisante pour une fenêtre d'analyse plus courte qu'habituellement (60 ms contre 93 ms en général pour des signaux de musique) et sur les $7^1/_4$ octaves que constitue la tessiture du piano. Pour ce faire, nous utiliserons une approche paramétrique qui tire parti de considérations temporelles et spectrales, ainsi que de la nature inharmonique des sons de piano.

Ces travaux ont fait l'objet d'une publication [Emiya *et al.*, 2007b].

3.1 Introduction

Nous avons vu dans le chapitre 1 (partie 1.2.2 (p. 26)) que les méthodes élémentaires pour l'estimation de hauteur s'appuient sur des considérations temporelles ou spectrales. Dans le premier cas, il s'agit d'analyser les périodicités de la forme d'onde – par exemple via l'ACF [Rabiner, 1977], l'AMDF [Ross *et al.*, 1974] ou le cepstre [Noll, 1967]) – alors que dans le second, le principe sous-jacent est la détection d'un peigne harmonique [Schroeder, 1968]. Appliquées à des sons réels, ces méthodes s'avèrent limitées par divers facteurs : présence de bruit, stationnaire ou non, écart par rapport à l'harmonicité supposée, non-stationnarité des composantes, large tessiture, variabilité des timbres et des enveloppes spectrales, et bien sûr, présence de plusieurs hauteurs simultanées dans les mélanges polyphoniques. Les erreurs typiques offrent un bon aperçu des difficultés rencontrées et des défauts de chaque approche. Les méthodes temporelles ont tendance à commettre des erreurs de sous-octave – un signal T-périodique étant également $2T$-périodique – alors que les approches spectrales sont sujettes à des erreurs d'octave – l'énergie d'un peigne de fréquence fondamentale f_0 donnant lieu à la détection d'un peigne de fréquence fondamentale $2f_0$. Par ailleurs, les deux types d'approches sont sensibles à la taille de la tessiture, aux variations des timbres

et à l'inharmonicité éventuelle. Enfin, une fenêtre d'analyse courte constitue un facteur limitant dans le cas spectral pour la discrimination des partiels des notes graves tandis que l'échantillonnage uniforme des méthodes temporelles les rend inefficaces dans l'aigu au-delà d'une certaine fréquence fondamentale.

Pour faire face à ces difficultés et proposer des solutions robustes, les travaux sur l'estimation de hauteur reposent souvent sur la réutilisation des principes de base cités précédemment couplée à l'introduction de mécanismes spécifiques : un traitement en sous-bandes, en particulier pour étendre les méthodes temporelles au cas polyphonique ou pour se rapprocher du fonctionnement de l'oreille [Meddis et Hewitt, 1991a; Klapuri, 2005], l'élaboration de fonctions de détection plus robustes [de Cheveigné et Kawahara, 2002; Klapuri, 2003, 2005] ou encore l'utilisation conjointe de méthodes temporelles et spectrales [Peeters, 2006].

L'algorithme que nous introduisons vise à améliorer les résultats d'estimation de fréquences fondamentales dans le cas d'une fenêtre d'analyse courte et d'une grande tessiture. Ces difficultés se rencontrent dans le cas des sons du piano étant données sa tessiture – 88 notes, soit $7^1/_4$ octaves, avec des fréquences fondamentales comprises entre 27 et 4200 Hz – et la vélocité avec laquelle ces notes peuvent être jouées, contraignant la pseudo-stationnarité des signaux, et donc l'analyse, à des durées restreintes. De plus, nous pouvons souligner que le piano est de manière générale l'un des instruments donnant lieu au taux d'erreur les plus élevés pour l'estimation de hauteur (voir par exemple Peeters [2006]).

La hauteur d'un son périodique ou quasi-périodique ne dépend que des composantes sinusoïdales de ce son. L'estimation de fréquences fondamentales ne requiert donc que les paramètres de ces composantes : fréquences, amplitudes, et éventuellement facteurs d'amortissement et phases initiales. Par conséquent, pour autant que nous le sachions, les méthodes d'estimation n'utilisent pas l'autre partie du son constituée du bruit de fond, des transitoires et autres composantes non sinusoïdales. C'est pourquoi la première étape de notre méthode d'estimation consiste à extraire les paramètres des composantes sinusoïdales. Nous élaborerons ensuite des versions paramétriques de méthodes temporelles et spectrales d'estimation de fréquences fondamentales, que nous combinons à la manière des travaux de Peeters [2006]. L'aspect paramétrique permettra non seulement de s'affranchir du bruit pour ne garder que l'information relative aux composantes sinusoïdales, mais également de corriger l'effet de l'inharmonicité des sons.

3.2 Estimation de hauteur

Notre approche s'appuie sur l'estimation des paramètres des composantes sinusoïdales du son, réalisée à partir de l'algorithme ESPRIT [Roy *et al.*, 1986] présenté dans la partie 2.1 (p. 51). Pour $t \in [\![0; N_a - 1]\!]$, la forme d'onde $s(t)$, analysée sur une trame de longueur N_a, est modélisée par

$$s(t) = \sum_{k=1}^{K} \alpha_k z_k^t + b(t) \tag{3.1}$$

comme une somme de K sinusoïdes complexes exponentiellement modulées $\alpha_k z_k^t, k \in [\![1; K]\!]$ et de bruit additif coloré $b(t)$. Les $\alpha_k = A_k e^{i\Phi_k} \in \mathbb{C}^*$ sont les amplitudes complexes, composées d'une amplitude réelle positive A_k et d'une phase initiale Φ_k. Les $z_k = e^{d_k + i2\pi f_k}$ sont les pôles, deux-à-deux distincts, les f_k étant les fréquences et les d_k les facteurs d'amortissement. Ce modèle de signal est adapté aux sons de piano, qui ne présentent pas de mo-

dulations de fréquence remarquables et peuvent en revanche être modulés en amplitude. L'estimation des paramètres $\{\alpha_k\}$ et $\{z_k\}$ se fait par l'algorithme ESPRIT tel que nous l'avons décrit dans la partie 2.1 (p. 51).

L'estimation de fréquences fondamentales s'appuie sur une méthode temporelle et une méthode spectrale. Elles sont successivement présentées dans les parties 3.2.1 et 3.2.2. Si chacune d'elles peut être considérée comme un estimateur de fréquences fondamentales, leur combinaison permet de tirer parti de leurs avantages respectifs pour obtenir un estimateur plus efficace, que nous décrirons ensuite (partie 3.2.3).

3.2.1 Méthode temporelle

Comme nous l'avons vu dans la partie 1.2.2 (p. 26) de l'état de l'art, une façon très répandue d'analyser la périodicité consiste à considérer un signal comme l'observation d'un processus y réel stationnaire au sens large (SSL) et à estimer sa fonction d'autocorrélation $R_y(\tau) = \mathbb{E}\left[y(t)y(t+\tau)\right]$. Lorsque le signal est périodique, les maxima de $R_y(\tau)$ se situent à $\tau = 0$ et à tous les multiples de la période fondamentale. Considérons donc un processus SSL y composé de K sinusoïdes non amorties de fréquences ν_k, d'amplitudes réelles $2a_k$, de phases initiales φ_k supposées indépendantes et uniformément distribuées sur $[0; 2\pi[$ et d'un bruit blanc de variance $\sigma^2_{b_y}$. L'expression de la fonction d'autocorrélation de y est alors

$$R_y(\tau) = \sum_{k=1}^{K} 2a_k^2 \cos\left(2\pi\nu_k\tau\right) + \delta(\tau)\sigma^2_{b_y} \tag{3.2}$$

Le modèle de son que l'on considère n'est pas stationnaire en raison notamment de la présence de facteurs d'amortissement (équation (3.1)), qui font varier l'amplitude des sinusoïdes entre le début et la fin de la trame (dans l'équation (3.1), les amplitudes en début et en fin de trame étant respectivement $|\alpha_k|$ et $|\alpha_k|\, e^{d_k(N_a-1)}$). Pour construire une fonction temporelle d'estimation de hauteur en nous inspirant du cas SSL ci-dessus et en intégrant les modulations d'amplitude, nous proposons de calculer la puissance de chaque composante sur une trame (cf. figure 3.1) et de considérer, à puissances égales, l'autocorrélation d'un processus SSL équivalent. Nous définissons ainsi une fonction temporelle $R(\tau)$ pour l'estimation de fréquence fondamentale à partir des paramètres estimés par l'analyse HR :

$$R(\tau) \triangleq \sum_{k=1}^{K} p_k \cos\left(2\pi f_k \tau\right) \tag{3.3}$$

$$p_k \triangleq \begin{cases} |\alpha_k|^2 & \text{si } |z_k| = 1 \\ \frac{|\alpha_k|^2}{N_a} \frac{1-|z_k|^{2N_a}}{1-|z_k|^2} & \text{sinon} \end{cases} \tag{3.4}$$

avec $\tau > 0$, $f_k = \frac{\arg(z_k)}{2\pi}$ étant la fréquence normalisée de la composante k, et p_k sa puissance. Pour simplifier cette expression, le dirac en 0 correspondant au bruit a été supprimé car nous ne considérons que la partie signal.

Pour un son légèrement inharmonique, l'écart fréquentiel par rapport au cas parfaitement harmonique a pour conséquence d'atténuer les pics de $R(\tau)$ aux multiples de la

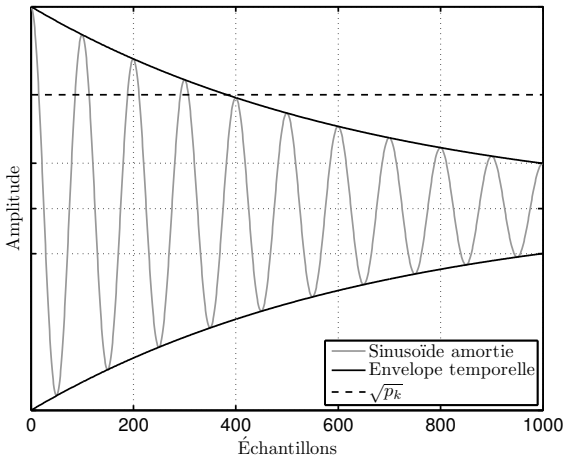


Figure 3.1 – Sinusoïde exponentiellement amortie et puissance p_k associée.

période fondamentale. Pour prendre en compte l'inharmonicité des sons de piano (cf. partie 2.2 (p. 53)), nous considérons le phénomène comme le résultat de l'étirement d'un spectre harmonique, comme illustré sur la figure 3.2, et appliquons l'opération inverse. Pour ce faire, l'ensemble des fréquences estimées $\{f_k, k \in [\![1; K]\!]\}$ est transformé en un ensemble de fréquences $\{g_{f_0,k}, k \in [\![1; K]\!]\}$, avec

$$g_{f_0,k} \triangleq \frac{f_k}{\sqrt{1 + \beta\left(f_0\right) h^2\left(f_0, f_k\right)}} \tag{3.5}$$

où $\beta\left(f_0\right)$ est le coefficient d'inharmonicité moyenne représenté en fonction de la fréquence fondamentale sur la figure 2.3(a) (p. 56). L'utilisation d'une valeur moyenne d'inharmonicité est suffisante dans notre cas pour obtenir des résultats satisfaisants, comme nous le verrons lors de l'évaluation de notre algorithme.

L'estimation de l'ordre du partiel $h\left(f_0, f_k\right)$ relatif à la fréquence f_k a pour expression (cf. équation (2.20) p. 56)

$$h\left(f_0, f_k\right) = \frac{f_k}{f_0} \sqrt{\frac{2}{\sqrt{1 + 4\beta\left(f_0\right) \frac{f_k^2}{f_0^2}} + 1}} \tag{3.6}$$

Par cette opération, les fréquences $g_{f_0,k}$ des partiels sont des multiples de la fréquence fondamentale f_0. En remplaçant les fréquences f_k par leurs corrections $g_{\frac{1}{\tau},k}$ dans l'équation (3.3), nous en déduisons une fonction temporelle $R_{\text{inh}}(\tau)$ qui est maximale pour $\tau = \frac{1}{f_0}$ dans le cas de sons de piano :

$$R_{\text{inh}}(\tau) \triangleq \sum_{k=1}^{K} p_k \cos\left(2\pi g_{\frac{1}{\tau},k} \tau\right) \tag{3.7}$$

3.2.2 Méthode spectrale

Nous définissons maintenant de façon paramétrique un spectre d'amplitude à partir de l'estimation des fréquences f_k et des énergies $E_k = N_a p_k$ des composantes, pour $k \in [\![1; K]\!]$.

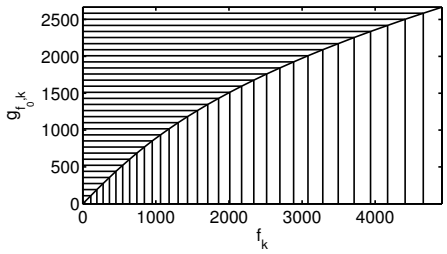


FIGURE 3.2 – Correction de l'inharmonicité : pour chaque fréquence fondamentale, les fréquences f_k sont transformées en $g_{f_0,k}$, afin de corriger les effets de l'inharmonicité. La figure représente un partiel sur cinq, avec $f_0 = 27,5$ Hz et $\beta = 2,54.10^{-4}$.

Il est constitué d'une somme de K gaussiennes centrées en f_k, d'écart-type constant σ, pondérées par une amplitude moyenne, racine carrée de l'énergie des composantes

$$S(f) \triangleq \sum_{k=1}^{K} \frac{\sqrt{E_k}}{\sqrt{2\pi}\sigma} e^{-\frac{(f-f_k)^2}{2\sigma^2}} \tag{3.8}$$

σ étant en pratique arbitrairement fixé à $f_{0\min}/4$ où $f_{0\min}$ est la fréquence fondamentale la plus basse, afin d'éviter tout recouvrement entre les partiels.

Nous définissons ensuite une méthode spectrale d'estimation de hauteur qui repose sur la maximisation d'un produit scalaire $U(f)$ entre le spectre paramétrique $S(f)$ et des motifs spectraux harmoniques normalisés pour chaque note candidate :

$$U(f) \triangleq \sum_{h=1}^{H_f} w_{f,h} S\left(hf\right) \tag{3.9}$$

où H_f est le nombre maximal de partiels à la fréquence fondamentale f et $\{w_{f,h}, h \in [\![1, H_f]\!]\}$ est le motif harmonique relatif à f. Le choix de ce motif s'appuie sur une approximation exponentielle de la décroissance de l'enveloppe spectrale des composantes. Pour ce faire, nous estimons la pente p d'une régression linéaire entre $\log(\sqrt{E_k})$ et f_k et définissons les poids $w_{f,h}$ par

$$w_{f,h} \triangleq w_0 e^{phf} \tag{3.10}$$

où $w_0 = \left(\sum_{h=1}^{H_f} e^{2phf}\right)^{-\frac{1}{2}}$ est un terme de normalisation tel que $\sum_{h=1}^{H_f} w_{f,h}^2 = 1$.

La fonction $U(f)$ est ensuite adaptée au cas inharmonique des sons de piano en redéfinissant le produit scalaire sur des valeurs du spectre prises selon une échelle inharmonique au lieu de l'échelle harmonique précédente. La nouvelle fonction est alors définie par

$$U_{\text{inh}}(f) \triangleq \sum_{h=1}^{H_f} w_{f,h} S\left(hf\sqrt{1+\beta(f)h^2}\right) \tag{3.11}$$

Par ailleurs, nous avons remarqué que les résultats étaient améliorés en supprimant toutes les fréquences en-deçà d'une fréquence de coupure passe-haut fixée à 100 Hz en raison de l'impédance au niveau du chevalet [Fletcher et Rossing, 1998] qui crée des écarts significatifs de fréquence avec la loi d'inharmonicité dans le grave, là où les poids $w_{f,h}$ des motifs spectraux sont précisément les plus importants.

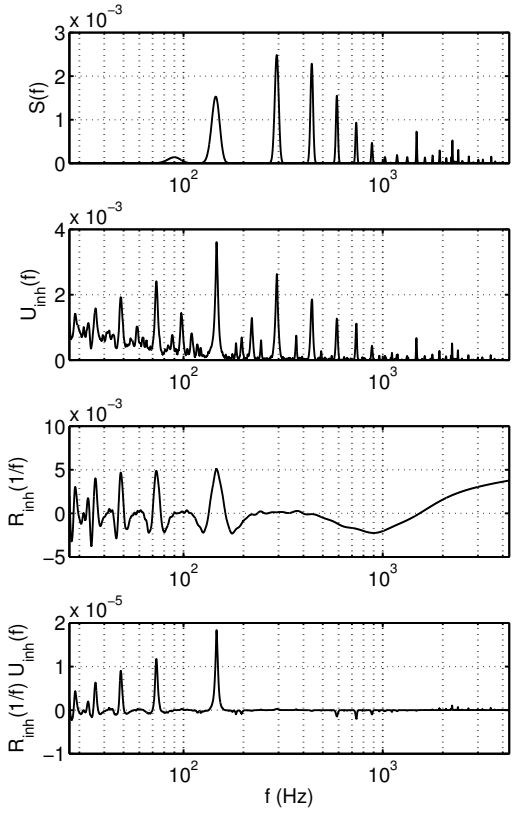


Figure 3.3 – Exemple d'analyse d'un Ré 2 (147 Hz) de piano sur 60 ms : de haut en bas, sur une échelle fréquentielle logarithmique, spectre paramétrique, fonction d'estimation spectrale $U_{\text{inh}}(f)$, fonction d'estimation temporelle $R_{\text{inh}}\left(\frac{1}{f}\right)$, fonction combinée pour l'estimation de la hauteur.

3.2.3 Estimation de la hauteur

Comme nous l'avons souligné dans la partie 3.1, les méthodes temporelles et spectrales s'opposent dans leurs tendances à commettre des erreurs harmoniques et sous-harmoniques. Ce phénomène est relevé par Peeters [2006] qui propose un moyen simple d'en tirer parti. Il consiste à multiplier une fonction temporelle avec une fonction spectrale sur une échelle commune de fréquences fondamentales afin de préserver les pics communs aux deux fonctions (en particulier celui correspondant à la hauteur à estimer) et d'atténuer, voire de supprimer les autres (les pics pouvant induire en erreur) comme illustré sur la figure 3.3. En suivant ce principe, nous estimons la hauteur présente en maximisant le produit des fonctions $R_{\text{inh}}\left(\frac{1}{f}\right)$ et $U_{\text{inh}}(f)$:

$$\hat{f}_0 \triangleq \text{argmax}_f \left(R_{\text{inh}}\left(\frac{1}{f}\right) U_{\text{inh}}\left(f\right) \right) \tag{3.12}$$

Grâce au caractère analytique des expressions (3.7) et (3.11), les valeurs $R_{\text{inh}}\left(\frac{1}{f}\right)$ et $U_{\text{inh}}(f)$ peuvent être directement et précisément évaluées pour f quelconque. La distribu-

tion des fréquences fondamentales étant logarithmique dans le cas du tempérament égal, nous choisissons d'échantillonner le support fréquentiel de recherche des fréquences fondamentales suivant N_f points logarithmiquement espacés sur l'intervalle de recherche. La liberté de choisir cet échantillonnage constitue un avantage de taille car nombre de méthodes ne permettent pas naturellement ce découpage logarithmique (cf. de Cheveigné et Kawahara [2002]; Peeters [2006]). En effet, les méthodes temporelles sont contraintes par un échantillonnage linéaire de l'axe des temps, qui a pour effet un manque de précision dans les hautes fréquences fondamentales et une résolution inutilement fine dans les basses fréquences fondamentales. À l'inverse, les méthodes reposant sur une analyse de Fourier présentent un découpage linéaire de l'axe des fréquences, avec les inconvénients opposés. Dans les deux cas, l'approche doit souvent faire intervenir une interpolation de la fonction d'estimation pour atteindre, de façon limitée, la précision nécessaire.

Avec une mise en œuvre sous Matlab et un processeur cadencé à $2,4$GHz, le traitement d'une trame de 60 ms nécessite environ $6,5$ s. La phase d'estimation des paramètres dure environ 1 s. Environ 95% du temps restant sert au calcul de l'estimateur spectral et pourrait être optimisé en C pour obtenir une mise en œuvre efficace.

3.3 Évaluation des résultats

L'algorithme a été évalué sur des sons isolés de piano provenant de plusieurs sources : 3168 notes (trois pianos) de la base RWC [Goto *et al.*, 2003], 270 notes (cinq pianos) d'une base PROSONUS et 264 notes d'une base de sons privées (piano droit Yamaha). Chaque base contient plusieurs versions de chacune des 88 notes de la tessiture du piano (à l'exception de la base PROSONUS pour laquelle les notes sont échelonnées par quartes) avec des nuances variables. La base RWC offre par ailleurs des enregistrements avec plusieurs modes de jeu (normal, staccato, avec la pédale *forte*). La recherche des fréquences fondamentales s'étale sur $N_f = 8192$ valeurs logarithmiquement distribuées entre $f_{0\min} = 26,73$ Hz et $f_{0\max} = 4310$ Hz. L'estimation est réalisée à partir de l'analyse d'une unique trame de 60 ms ou 93 ms : 60 ms correspond à une durée très courte, inférieure à deux périodes pour les notes les plus graves, alors que 93 ms est une durée d'analyse standard pour l'estimation de hauteur de signaux musicaux. Chaque fréquence fondamentale est arrondie au demi-ton le plus proche sur une échelle bien tempérée, le La 3 accordé à 440 Hz (cf. annexe C (p. 177)). Une erreur est alors définie comme l'estimation d'une mauvaise note.

Les résultats sont comparés avec ceux de deux autres estimateurs. Le premier a été construit de la manière la plus similaire possible au nôtre, mais en remplaçant l'analyse par ESPRIT par des procédés plus communs : l'autocorrélation est estimée à partir du signal par l'expression

$$r(\tau) = \frac{N_a}{N_a - \tau} \mathrm{DFT}^{-1}\left[\left|\mathrm{DFT}\left[s\right]\right|^2\right] \tag{3.13}$$

le facteur $\frac{N_a}{N_a-\tau}$ étant la correction du biais ; l'estimateur spectral $U_{\mathrm{inh}}(f_0)$ est obtenu en remplaçant le spectre paramétrique par le module de la transformée de Fourier discrète du signal, avec un *zero-padding* de $8N_f$ points ; le support temporel de $r(\tau)$ est transformé en support fréquentiel par interpolation tel que le décrit Peeters [2006] ; la hauteur est finalement estimée en maximisant le produit des deux fonctions sur le support fréquentiel commun. La seconde méthode utilisée est l'algorithme YIN [de Cheveigné et Kawahara,

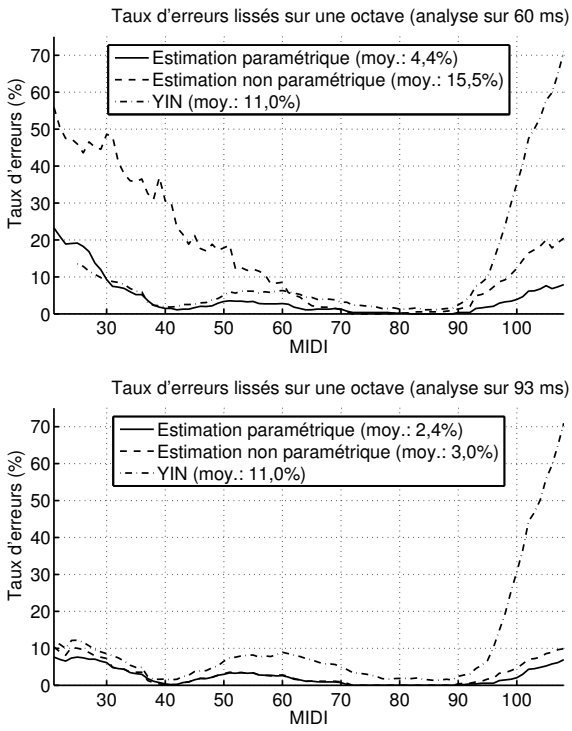


Figure 3.4 – Taux d'erreurs par note, moyennés sur une octave, pour deux durées d'analyse différentes. Résultats pour la méthode présentée et, à titre comparatif, pour deux autres méthodes : un algorithme similaire mais non paramétrique et non HR, et l'algorithme YIN. Le taux d'erreur moyen sur tout le registre figure entre parenthèses dans la légende.

2002], généralement considéré comme un estimateur de hauteur très performant. Nous avons utilisé le code que les auteurs fournissent sur leur site Internet.

Les résultats de l'évaluation comparative sont représentés sur la figure 3.4. Pour une fenêtre d'analyse de 60 ms, le taux d'erreur global de notre estimateur est d'environ $4,4\%$, soit au moins deux fois moindre que ceux obtenus avec les autres estimateurs. Cela s'explique par un taux d'erreur faible sur un grand intervalle de fréquences fondamentales ($1,1\%$ sur l'intervalle $65 - 2000$ Hz) et une faible augmentation aux extrêmes grave et aigu du registre. En comparaison, l'estimateur non paramétrique n'atteint un taux d'erreur de $1,1\%$ que sur l'intervalle $240 - 2000$ Hz. Sa baisse d'efficacité en dehors de cet intervalle montre que les bons résultats obtenus par notre approche sont à la fois dus à l'analyse haute-résolution et à l'avantage d'avoir des méthodes paramétriques et des formules analytiques. Quant à l'algorithme YIN, il est un peu moins efficace dans le registre médium et offre des résultats comparables aux nôtres dans les basses (au niveau de la première octave, les deux courbes devraient être au même niveau, mais celle relative à notre estimateur se situe plus haut car elle intègre les résultats des quatres notes les plus graves, que l'algorithme YIN ne peut pas estimer sur 60 ms). Dans le registre aigu, YIN s'avère moins efficace et présente à ce titre le comportement typique des méthodes temporelles d'estimation de hauteur. De façon générale, les résultats s'améliorent lorsque l'on passe à une analyse sur 93 ms. Néanmoins, l'analyse HR n'améliore pas significativement l'estimation de fréquences fondamentales, même si l'algorithme reste celui qui commet le moins d'erreurs.

Il est intéressant de se pencher sur les erreurs typiques commises, que nous traitons dans le cas de l'analyse sur 60 ms. Lorsqu'ils se trompent, les algorithmes ont logiquement tendance à surestimer les fréquences fondamentales basses et à sous-estimer les aiguës. Environ 18% des erreurs commises par chaque méthode sont des erreurs d'octave ou de sous-octave. Dans le cas de la nôtre, les autres erreurs correspondent à des intervalles de tous types, avec seulement 5% d'erreurs de demi-ton, alors que ce taux atteint 10% avec les deux autres méthodes. Les erreurs de YIN sont plutôt sous-harmoniques (13% à l'octave inférieure, 8% à la 19^{e} inférieure). Ainsi, bien que le nombre d'erreurs harmoniques et sous-harmoniques de notre algorithme soit réduit, il reste visiblement difficile d'éviter ce genre d'erreurs. En revanche, le faible taux d'erreurs de demi-ton montre l'efficacité de la méthode, alors que les autres algorithmes souffrent d'un manque de précision, dans l'aigu, dû à leur approche temporelle. Enfin, nous avons constaté que la prise en compte de l'inharmonicité contribuait à faire baisser le taux d'erreurs global de $4,9$ à $4,4$% (soit 10% d'erreurs en moins). L'amélioration se situe plutôt dans le registre grave : le taux d'erreur sur l'intervalle MIDI $[\![21, 37]\!]$ passe ainsi de $16,6$ à $14,1$%.

3.4 Conclusion

La méthode présentée pour l'estimation de fréquences fondamentales parvient à des taux d'erreurs significativement meilleurs que l'état de l'art dans le contexte d'une fenêtre d'analyse courte et d'une tessiture étendue. L'analyse à Haute-Résolution, l'utilisation conjointe d'une méthode temporelle et d'une méthode spectrale, ainsi que l'approche paramétrique contribuent à réduire le nombre d'erreurs, en particulier les erreurs typiques d'octave, de sous-octave et de demi-ton, et à rendre la méthode robuste à ces conditions d'analyse peu favorables.

Chapitre 4

Estimation de fréquences fondamentales multiples

Nous abordons à présent la question de l'estimation de fréquences fondamentales multiples. Nous avons vu dans l'état de l'art (partie 1.3 (p. 29)) que cette problématique posait des difficultés supplémentaires importantes par rapport à celle de l'estimation de hauteurs simples, en particulier en raison du recouvrement des spectres et du nombre inconnu de notes présentes. Nous avons également vu que la notion d'enveloppe spectrale pouvait être la source d'informations utiles pour faire face à ces difficultés. Nous proposons ici une approche qui tire parti de ce constat et explorons le cadre théorique statistique pour utiliser le modèle d'enveloppe présenté dans la partie 2.3 (p. 61).

Le chapitre développera en premier lieu la problématique de l'estimation de fréquences fondamentales multiples (partie 4.1). Le modèle utilisé et son cadre statistique seront ensuite introduits (partie 4.2). Puis l'estimation de fréquences fondamentales multiples sera décrite dans les parties 4.3 à 4.8.

Une version antérieure de ces travaux a fait l'objet d'une publication [Emiya *et al.*, 2007a].

4.1 Problématique

Observons la transformée de Fourier discrète d'un mélange de deux notes dont les fréquences fondamentales sont en rapport harmonique. À titre d'illustration, un exemple de deux notes à la quinte est présenté sur la figure 4.1. Plusieurs types de coefficients spectraux sont visibles :

1. les coefficients correspondant à un pic isolé, comme ceux relatifs aux trois premiers partiels sur la figure 4.1 ; l'amplitude du partiel associé est alors mesurable directement et de façon fiable ;
2. les coefficients issus d'un recouvrement de plusieurs partiels, dont les fréquences coïncident exactement (pic autour de 800 Hz sur la figure 4.1) ou approximativement (pic autour de 2400 Hz) ; il est alors plus difficile d'estimer les amplitudes de chaque composante compte tenu des interférences dues à la phase de leurs spectres ;
3. les coefficients situés dans le lobe principal autour d'un pic ;
4. les coefficients à des fréquences éloignées de tout pic, résultant du bruit de fond et des lobes secondaires.

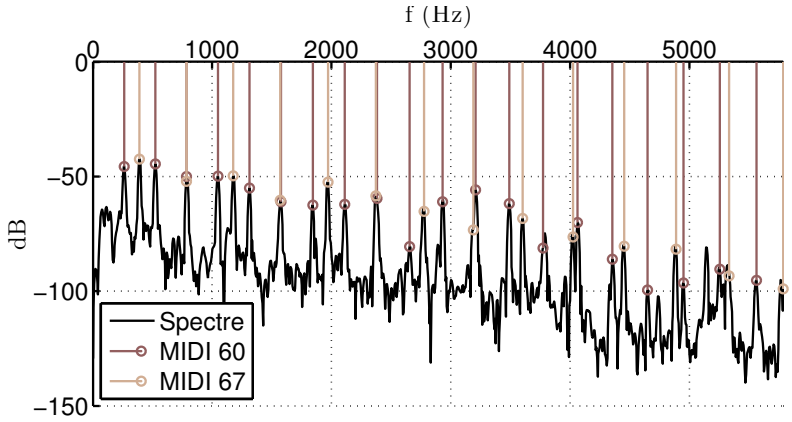


Figure 4.1 – Spectre de deux notes à la quinte. Recouvrement des partiels du Do 3 (note MIDI 60) dont l'ordre est un multiple de 3 avec les partiels du Sol 3 (note MIDI 67) dont l'ordre est un multiple de 2 (son enregistré sur un piano Bechstein D 280).

Nous proposons alors une approche statistique pour l'estimation de fréquences fondamentales multiples qui s'appuie sur les principes suivants :

- l'information relative aux notes et au bruit est concentrée dans des ensembles distincts de coefficients spectraux (les pics relatifs à une note, les coefficients résiduels pour le bruit) ;
- en cas de recouvrement spectral, certains coefficients portent l'information provenant de plusieurs composantes, dont il s'agit d'estimer les contributions ;
- les notes de piano ont une enveloppe spectrale relativement lisse, que nous modéliserons par un modèle autorégressif (AR), tel que nous l'avons décrit dans la partie 2.3.2 (p. 63) ; l'utilisation d'un modèle paramétrique permet ainsi un certain nombre de développements analytiques ;
- le bruit sera modélisé par un processus à moyenne ajustée (MA), tel que nous l'avons décrit dans la partie 2.4 (p. 66). N'ayant pas de pôles, ce modèle présente l'avantage de ne pas être adapté aux sinusoïdes contenues dans un bruit résiduel mal estimé et d'être à ce titre discriminant lorsqu'il s'agit, comme dans le cas présent, de distinguer les partiels du bruit.

La démarche, illustrée sur la figure 4.2, consiste alors à estimer les paramètres des différents modèles et à maximiser une fonction de détection par rapport aux mélanges de notes possibles.

4.2 Cadre statistique

Nous définissons dans un premier temps le modèle de son utilisé pour l'estimation de fréquences fondamentales multiples puis expliquons le principe de l'estimation. Dans un souci de lisibilité, les variables aléatoires introduites sont en général notées de la même manière que leurs réalisations. Les densités de probabilité sont également notées de façon simplifiée ($p_y(y)$ ou $p(y)$) lorsque le contexte ne laisse pas d'ambiguïté.

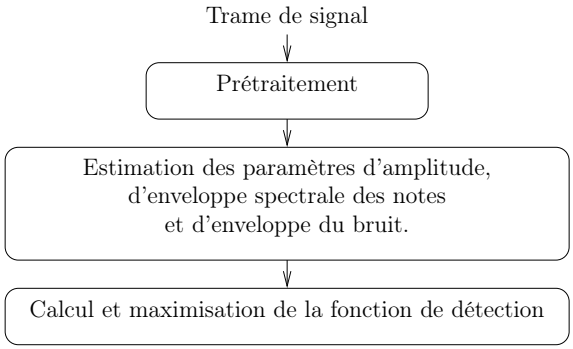


FIGURE 4.2 – Principe de l'estimation de fréquences fondamentales multiples.

4.2.1 Modèle génératif de son

Considérons un **mélange** additif de $P \in \mathbb{N}$ notes et de bruit, dont on observe une trame de N échantillons. Nous modélisons ce signal en utilisant un modèle statistique, représenté sur la figure 4.3 et décrit ci-dessous.

Pour $p \in [\![1; \dots, P]\!]$, la **note** p se caractérise avant tout par sa **fréquence fondamentale** $f_{0,p}$ et son **inharmonicité** β_p qui forment le couple

$$\mathcal{C}_p \triangleq (f_{0,p}, \beta_p) \tag{4.1}$$

On associe à ces paramètres le nombre H_p de partiels observables, dont l'expression est donnée par l'équation (2.17) (p. 54), et l'ensemble $\mathcal{F}_p$ des fréquences des partiels défini par

$$\mathcal{F}_p \triangleq \left\{ f_h^{(p)} / h \in [\![1; H_p]\!] \right\} \tag{4.2}$$

avec

$$f_h^{(p)} \triangleq h f_{0,p} \sqrt{1 + \beta_p h^2} \tag{4.3}$$

En utilisant le modèle de la partie 2.3 (p. 61), nous introduisons ensuite un **modèle autorégressif d'enveloppe spectrale** de paramètre

$$\theta_p \triangleq \left(\sigma_p^2, A_p(z)\right), \tag{4.4}$$

composé d'une puissance σ_p^2 et d'un filtre unitaire d'ordre Q_p et de réponse $A_p(z)$, tel que défini dans l'annexe A.2.1 (p. 165). La valeur de Q_p est choisie suffisamment faible par rapport au nombre de partiels pour que l'enveloppe spectrale soit lisse, et suffisamment élevée pour bien modéliser les enveloppes rencontrées en pratique : $Q_p = H_p/2$ constitue un bon compromis, que nous utiliserons (la moitié des Q_p pôles ont des fréquences positives, l'autre moitié étant leurs conjugués). Il est important de remarquer que nous ne définissons pas une enveloppe spectrale mais un *modèle* d'enveloppe spectrale. L'enveloppe spectrale, c'est-à-dire les amplitudes des partiels, est alors une réalisation de ce modèle. Nous définissons les **amplitudes complexes des partiels** en tant que variables aléatoires complexes gaussiennes indépendantes telles que pour $h \in [\![1; H_p]\!]$,

$$\alpha_h^{(p)} \sim \mathcal{N}\left(0, \frac{\sigma_p^2}{\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2}\right) \tag{4.5}$$

Signaux temporels

$x(t)$	mélange observé
$x_b(t)$	bruit
$x_p(t)$	note p
$e_{h,p}(t)$	partiel h (note p)

Variables de trame

$\mathcal{C}_p = (f_{0,p}, \beta_p)$	F_0 et inharmonicité (note p)
$\theta_p = \left(\sigma_p^2, A_p\right)$	paramètres AR d'enveloppe spectrale (note p)
$\alpha_{h,p}$	amplitude du partiel h (note p)
$\theta_b = \left(\sigma_b^2, B\right)$	paramètres MA de bruit

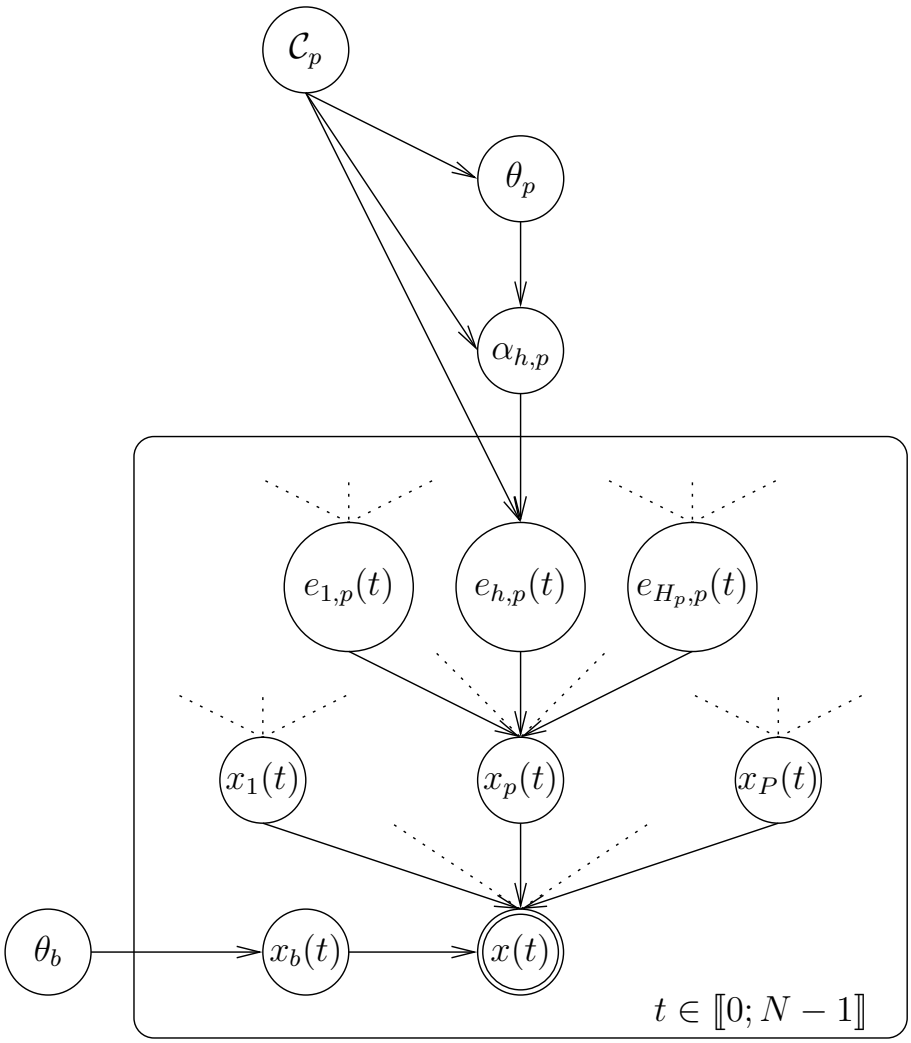


Figure 4.3 – Réseau bayésien représentant le modèle de son

Ainsi, le signal de la note p est le processus harmonique (cf. partie 2.3.1 (p. 61)) résultant de la somme de H_p partiels $e_{1,p}, \ldots, e_{H_p,p}$

$$x_p(t) \triangleq 2\mathrm{Re}\left(\sum_{h=1}^{H_p} e_{h,p}(t)\right) \tag{4.6}$$

$$= 2\mathrm{Re}\left(\sum_{h=1}^{H_p} \alpha_h^{(p)} e^{2i\pi f_h^{(p)} t}\right) \tag{4.7}$$

La transformée de Fourier discrète fenêtrée de x_p est

$$X_p(f) = \sum_{h=1}^{H_p} \left(\alpha_h^{(p)} W\left(f - f_h^{(p)}\right) + \alpha_h^{(p)*} W^*\left(f + f_h^{(p)}\right)\right) \tag{4.8}$$

où W est la transformée de Fourier discrète de la fenêtre de pondération w utilisée.

Pour plus de clarté dans la suite, nous introduisons les notations synthétiques suivantes :

$$\mathcal{C} \triangleq (\mathcal{C}_1, \ldots, \mathcal{C}_P) \tag{4.9}$$

$$\theta \triangleq (\theta_1, \ldots, \theta_P) \tag{4.10}$$

$$\text{Pour } p \in [\![1; P]\!],\ \alpha^{(p)} \triangleq \left(\alpha_1^{(p)}, \ldots, \alpha_{H_p}^{(p)}\right) \tag{4.11}$$

$$\alpha \triangleq \left(\alpha^{(1)}, \ldots, \alpha^{(P)}\right) \tag{4.12}$$

$\mathcal{C}$ est donc l'ensemble des fréquences fondamentales et inharmonicités des notes du mélange considéré, θ est l'ensemble des paramètres des modèles d'enveloppe spectrale des notes de ce mélange, α est l'ensemble des amplitudes des partiels des notes et $\alpha^{(p)}$ les amplitudes des partiels de la note p.

Enfin, le **bruit** est modélisé par un processus x_b à moyenne ajustée (MA) d'ordre Q_b décorrélé des amplitudes, de paramètre

$$\theta_b \triangleq \left(\sigma_b^2, B(z)\right) \tag{4.13}$$

composé de la puissance σ_b^2 et du filtre MA $B(z)$ avec $B(0) = 1$.

Le **signal observé** est donc une réalisation du processus

$$x(t) \triangleq \sum_{p=1}^{P} x_p(t) + x_b(t) \tag{4.14}$$

$$\tag{4.15}$$

De manière équivalente, on peut prendre comme observation sa transformée de Fourier discrète

$$X(f) \triangleq \sum_{p=1}^{P} X_p(f) + X_b(f) \tag{4.16}$$

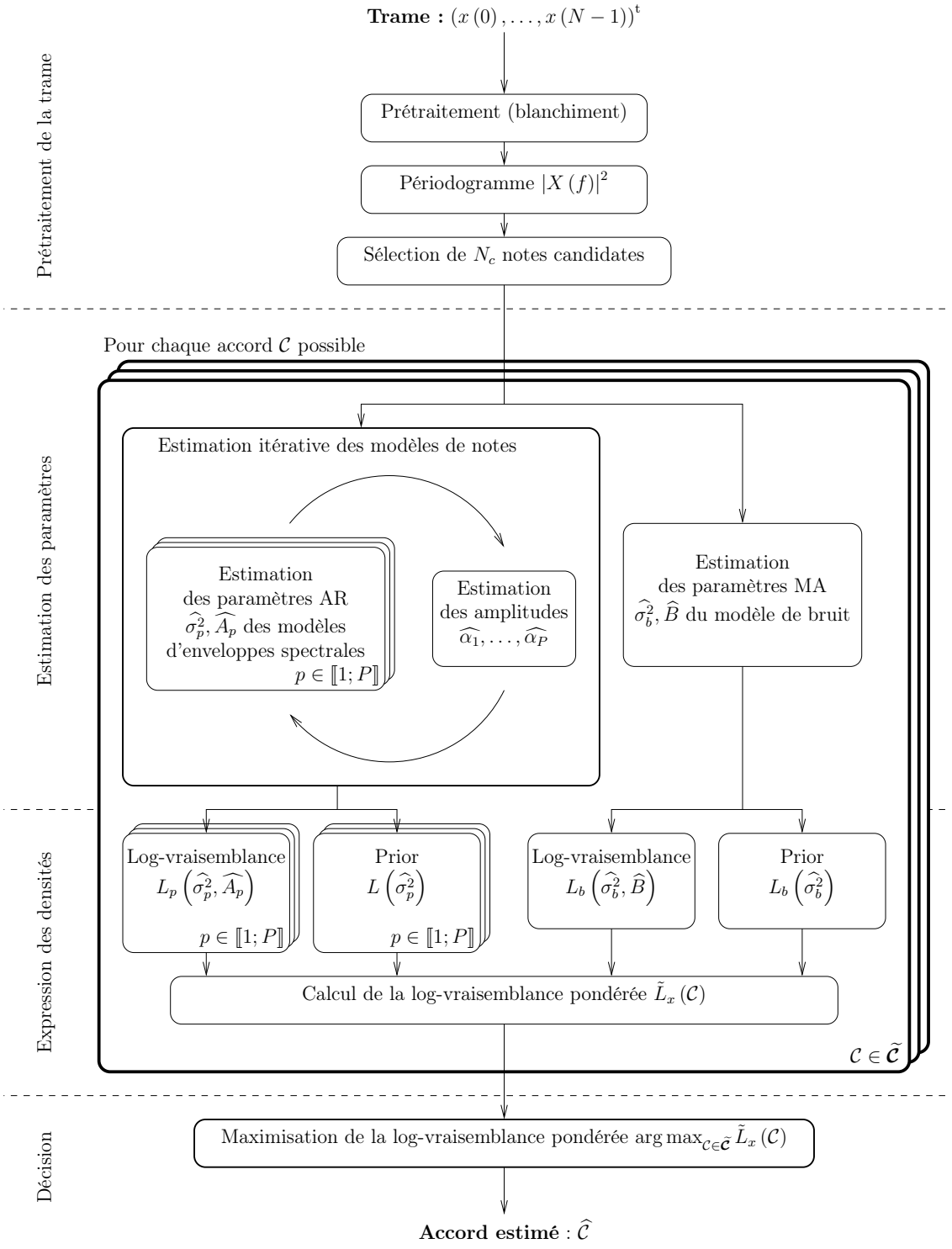


Figure 4.4 – Diagramme en blocs de l'algorithme d'estimation de fréquences fondamentales multiples.

4.2.2 Description inférentielle : méthode de résolution

Après une description de type génératif du modèle de son dans la partie précédente, nous abordons maintenant l'estimation de fréquences fondamentales multiples. La démarche est donc inférentielle : elle part des observations et de la problématique, et mène à l'estimation des paramètres du modèle. Nous décrivons ici ce cheminement qui conduit à isoler plusieurs blocs pour l'estimation des différentes quantités (amplitudes, enveloppes spectrales, bruit, etc.). Ces blocs s'agencent alors pour obtenir le diagramme représenté sur la figure 4.4.

Mélange le plus probable

L'observation est une trame x de longueur N du processus $x(t)$ prétraité. Le prétraitement consiste à estimer grossièrement – par filtrage médian sur le spectre – le niveau de bruit et à le blanchir afin de réduire la dynamique spectrale. x désigne ici le résultat de ce blanchiment, opération qui ne change pas les hypothèses sur le modèle introduit dans la partie précédente. En termes statistiques, x a été généré par le modèle $\mathcal{C}^\circ$ du mélange original. L'ensemble des modèles de mélanges possibles est noté $\boldsymbol{\mathcal{C}}$. Idéalement, l'estimation de fréquences fondamentales multiples consiste à déterminer le mélange le plus probable d'après x, c'est-à-dire le maximum *a posteriori* (MAP)

$$\hat{\mathcal{C}} = \underset{\mathcal{C}\in\boldsymbol{\mathcal{C}}}{\arg\max}\, p\left(\mathcal{C}|x\right) \tag{4.17}$$

En considérant tous les mélanges comme étant équiprobables, c'est-à-dire que la fonction $\mathcal{C} \mapsto p\left(\mathcal{C}\right)$ est constante [1], on a

$$\begin{aligned}\hat{\mathcal{C}} &= \underset{\mathcal{C}\in\boldsymbol{\mathcal{C}}}{\arg\max}\, \frac{p\left(x|\mathcal{C}\right)p\left(\mathcal{C}\right)}{p\left(x\right)} \\ &= \underset{\mathcal{C}\in\boldsymbol{\mathcal{C}}}{\arg\max}\, p\left(x|\mathcal{C}\right)\end{aligned} \tag{4.18}$$

Autrement dit, l'estimé du MAP $\hat{\mathcal{C}}$ correspond à l'estimé du maximum de vraisemblance (ML). L'estimation des fréquences fondamentales présentes consiste donc à calculer la valeur de la vraisemblance $p\left(x|\mathcal{C}\right)$ pour tout $\mathcal{C} \in \boldsymbol{\mathcal{C}}$ et à sélectionner celui pour lequel la vraisemblance est maximale. Étant donnée la taille de $\boldsymbol{\mathcal{C}}$, cette démarche est inconcevable en termes de complexité. La procédure que nous décrirons dans la partie 4.3 opère une réduction de l'espace de recherche $\boldsymbol{\mathcal{C}}$ par sélection d'un nombre restreint de notes candidates et estimation des paramètres de fréquence fondamentale et de coefficient d'inharmonicité. Le sous-ensemble obtenu est noté $\widetilde{\boldsymbol{\mathcal{C}}}$ et l'équation (4.18) devient donc

$$\hat{\mathcal{C}} = \underset{\mathcal{C}\in\widetilde{\boldsymbol{\mathcal{C}}}}{\arg\max}\, p\left(x|\mathcal{C}\right) \tag{4.19}$$

L'espace de recherche $\widetilde{\boldsymbol{\mathcal{C}}}$ contient alors un nombre suffisamment faible d'éléments pour pouvoir être entièrement exploré en un temps raisonnable.

1. Dans un contexte musical, il est possible de traiter le cas où $\mathcal{C}$ n'est pas distribué selon une loi uniforme et de généraliser l'approche présentée.

Expression intégrale de la vraisemblance des données

Nous nous intéressons donc, pour un mélange $\mathcal{C}$ quelconque, à l'expression et au calcul de $p\left(x|\mathcal{C}\right)$. La densité $p\left(x|\mathcal{C}\right)$ est une marginale que l'on peut écrire comme une intégrale en utilisant les variables latentes de notre modèle et leurs interdépendances (cf. figure 4.3) :

$$p\left(x|\mathcal{C}\right) = \iint p\left(x,\alpha,\theta_b|\mathcal{C}\right) d\alpha d\theta_b \tag{4.20}$$

$$= \iint p\left(x|\alpha,\theta_b,\mathcal{C}\right) p\left(\alpha|\mathcal{C}\right) p\left(\theta_b\right) d\alpha d\theta_b \tag{4.21}$$

$$= \iint p\left(x|\alpha,\theta_b,\mathcal{C}\right) \left(\int p\left(\alpha,\theta|\mathcal{C}\right) d\theta\right) p\left(\theta_b\right) d\alpha d\theta_b \tag{4.22}$$

$$= \iiint p\left(x|\alpha,\theta_b,\mathcal{C}\right) p\left(\alpha|\theta,\mathcal{C}\right) p\left(\theta\right) p\left(\theta_b\right) d\theta d\alpha d\theta_b \tag{4.23}$$

Dans cette expression, $p\left(x|\alpha,\theta_b,\mathcal{C}\right) = p_x\left(x|\alpha,\theta_b,\mathcal{C}\right)$ se réécrit à l'aide de la densité de probabilité du bruit p_{x_b} en soustrayant au signal observé x sa partie sinusoïdale, c'est-à-dire en utilisant le changement de variable $x \mapsto x - 2\mathrm{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_h^{(p)}e^{2i\pi f_h^{(p)}\mathbf{t}}\right)$, $\mathbf{t}$ étant le vecteur d'instants relatifs à la trame

$$p_x\left(x|\alpha,\theta_b,\mathcal{C}\right) = p_{x_b}\left(x - 2\mathrm{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_h^{(p)}e^{2i\pi f_h^{(p)}\mathbf{t}}|\theta_b\right)\right) \tag{4.24}$$

L'équation (4.23) devient donc

$$p\left(x|\mathcal{C}\right) = \iiint p_{x_b}\left(x - 2\mathrm{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_h^{(p)}e^{2i\pi f_h^{(p)}\mathbf{t}}\right)|\theta_b\right) p\left(\alpha|\theta,\mathcal{C}\right) p\left(\theta\right) p\left(\theta_b\right) d\theta d\alpha d\theta_b \tag{4.25}$$

L'intégrande est composée des termes suivants :

- $p\left(\alpha|\theta,\mathcal{C}\right)$: la vraisemblance des amplitudes α des partiels, associée aux modèles d'enveloppes spectrales θ ;
- $p_{x_b}\left(x - 2\mathrm{Re}\left(\sum_{p=1}^{P}\sum_{h=1}^{H_p}\alpha_h^{(p)}e^{2i\pi f_h^{(p)}\mathbf{t}}\right)|\theta_b\right)$: la vraisemblance associée au modèle de bruit θ_b ;
- $p\left(\theta\right), p\left(\theta_b\right)$: les densités *a priori* des paramètres des enveloppes spectrales et du bruit.

Estimation des paramètres optimaux et de la vraisemblance maximale

Le calcul de l'intégrale (4.25) est délicat et nécessite une approximation. Une solution efficace consisterait à faire appel aux méthodes de type Monte-Carlo [Doucet et Wang, 2005], pour lesquelles l'intégrande est calculée pour un certain nombre de valeurs. Une autre méthode couramment utilisée, en particulier en statistiques, est l'approximation de Laplace (cf. annexe A.3 (p. 172)), qui permet de ne prendre en compte que le maximum de l'intégrande. Nous adoptons cette approche dans la mesure où nous nous intéressons à l'estimation des paramètres des modèles de note et de bruit plutôt qu'au parcours de l'espace des paramètres. Dans le cas présent, elle consiste, sous certaines conditions, à considérer l'intégration sur les paramètres Θ du modèle et l'ordre $n\left(\mathcal{C}\right)$ associé (c'est-à-dire le nombre de paramètres contenus dans le vecteur Θ, dépendant du modèle $\mathcal{C}$ considéré)

pour exprimer $p\left(x|\mathcal{C}\right)$ sous la forme

$$p\left(x|\mathcal{C}\right) = \int p\left(x,\Theta|\mathcal{C}\right) d\Theta \tag{4.26}$$

$$\approx p\left(x,\Theta^*|\mathcal{C}\right) g\left(x,\Theta^*,n\left(\mathcal{C}\right)\right) \tag{4.27}$$

avec

$$\Theta^* \triangleq \arg\max_{\Theta} p\left(x,\Theta|\mathcal{C}\right) \tag{4.28}$$

Dans l'expression obtenue (équation (4.27)), la fonction g s'exprime notamment à partir du hessien de l'intégrande de la ligne précédente. Dans le cadre statistique des méthodes de sélection de modèles telles que BIC, AIC ou GIC [Stoica et Selen, 2004], nous verrons qu'elle peut se simplifier pour ne dépendre que de l'ordre $n\left(\mathcal{C}\right)$ du modèle et du nombre d'observations $\#x$.

La démarche nécessite donc deux étapes :

1. la maximisation de l'intégrande (résolution de l'équation (4.28)) ;
2. l'estimation de l'intégrale en fonction de l'optimum Θ^* et de l'équation (4.27).

Dans notre cas, nous appliquerons ce principe à l'intégrale (4.25). La première étape – l'estimation des paramètres du modèle – fera l'objet des parties 4.4 et 4.5 tandis que la seconde sera détaillée dans les parties 4.6 et 4.7. Auparavant, nous décrivons les grandes lignes de cette résolution.

L'étape de maximisation de l'intégrande se formule ainsi : étant donné un ensemble de notes $\mathcal{C}$ dont les fréquences fondamentales et les coefficients d'inharmonicité sont connus (et, par extension, les fréquences des partiels), quels sont les paramètres optimaux d'amplitudes des partiels, d'enveloppes spectrales de note, et de bruit ?

Nous nous intéressons tout d'abord à l'optimisation par rapport aux amplitudes α des partiels. Pour ce faire, nous reprenons les considérations évoquées dans la partie 4.1 (p. 81) sur le spectre observé. Nous supposons qu'au niveau de la fréquence d'un partiel, la valeur du spectre du bruit est négligeable devant celle du partiel, hypothèse sur le rapport signal à bruit au niveau des partiels qui est en pratique largement vérifiée. En l'absence de recouvrement spectral entre partiels (par exemple dans le cas monophonique), nous pouvons alors estimer l'amplitude $\alpha_h^{(p)}$ d'un partiel comme étant égale à la valeur du spectre observé $X\left(f_h^{(p)}\right)$.

Il est ensuite possible d'opérer la maximisation de l'intégrande par rapport aux paramètres d'enveloppe spectrale des notes, c'est-à-dire de maximiser $p\left(\alpha|\theta,\mathcal{C}\right)p\left(\theta\right)$ par rapport à θ. Nous introduisons à ce niveau une approximation consistant à prendre comme solution l'estimée de θ au sens du maximum de vraisemblance, c'est-à-dire à maximiser $p\left(\alpha|\theta,\mathcal{C}\right)$, sans prendre en compte l'influence de la loi *a priori* $p\left(\theta\right)$. Nous verrons dans la partie 4.4.2 que cette maximisation consiste en une estimation des paramètres des modèles AR d'enveloppes spectrales de notes à partir d'observations partielles. Nous justifions l'approximation utilisée [2] en considérant que l'information sur les puissances σ_p^2 ne sera utilisée que comme une pénalisation dans l'expression de $p\left(x|\mathcal{C}\right)$ (cf. sections 4.6 et 4.7).

2. Par ailleurs, dans la littérature, l'argument asymptotique est couramment utilisé pour ce genre d'approximation : lorsque la taille de l'observation tend vers l'infini, l'estimation du maximum de vraisemblance est « équivalente ».

Le raisonnement s'applique parfaitement en l'absence de recouvrement spectral. Dans le cas contraire, les coefficients spectraux observés résultent de la contribution de plusieurs partiels, la contribution du bruit étant toujours négligeable. Les amplitudes des partiels concernés ne sont alors pas directement identifiables sur le spectre observé. En utilisant ce dernier et à l'aide de l'information sur l'enveloppe spectrale, nous chercherons à estimer les amplitudes manquantes. Nous détaillerons dans la partie 4.4 une méthode pour estimer itérativement les amplitudes des partiels et les modèles d'enveloppes spectrales des notes afin de converger vers une solution.

L'hypothèse d'un rapport signal à bruit élevé au niveau des composantes sinusoïdales présente l'avantage de ne pas faire intervenir le modèle de bruit et la densité p_{x_b} associée dans l'estimation des modèles de notes (amplitudes et enveloppes spectrales). Une fois ceux-ci estimés, l'optimisation par rapport aux paramètres du modèle de bruit peut être réalisée, au sens du maximum de vraisemblance, comme nous l'expliquerons dans la partie 4.5.

Dans la partie 4.6, nous introduirons les lois *a priori* sur les paramètres d'enveloppes spectrales et de bruit. Ces lois n'entrent pas en compte dans l'estimation des paramètres (par maximum de vraisemblance) mais interviennent en revanche dans l'expression de la vraisemblance $p\left(x|\mathcal{C}\right)$.

La maximisation par rapport à tous les paramètres terminée, nous aborderons finalement dans la partie 4.7 l'estimation de l'intégrale (4.25), c'est-à-dire de la vraisemblance $p\left(x|\mathcal{C}\right)$, pour laquelle nous proposerons une version équivalente qui sera alors détaillée.

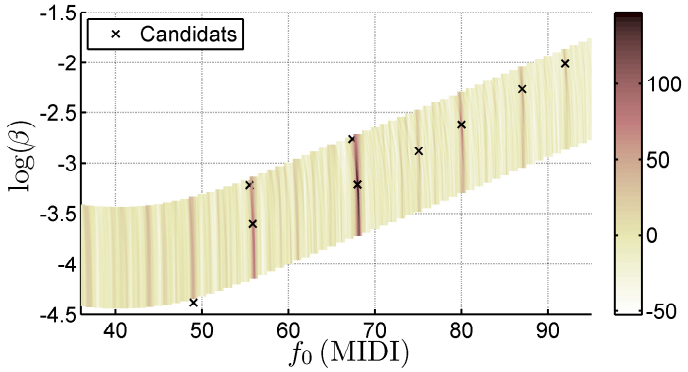
4.3 Sélection des notes candidates

Nous détaillons ici la réduction du nombre de mélanges candidats, c'est-à-dire le passage de l'équation (4.18) à l'équation (4.19). D'après l'équation (4.18), lors de l'analyse d'une trame x de signal, la vraisemblance $p\left(x|\mathcal{C}\right)$ doit normalement être évaluée sur l'ensemble $\boldsymbol{\mathcal{C}}$ des mélanges $\mathcal{C}$ possibles. Leur nombre étant $\binom{Q}{P}$ pour une polyphonie P donnée et une tessiture comprenant Q notes, il s'élève au total à $\sum_{P=0}^{Q}\binom{Q}{P}=2^Q$, avec $Q=88$ dans le cas du piano. Dans l'hypothèse où l'on limite par exemple la tessiture à $Q=60$ notes et la polyphonie à $P_{\max}=6$, le nombre de combinaisons atteint $\sum_{P=0}^{P_{\max}}\binom{Q}{P}\approx 56.10^6$ et reste une taille d'espace à explorer trop élevée. De plus, les fréquences fondamentales des notes composant un mélange peuvent également varier, en raison de l'incertitude sur l'accordage du piano, augmentant ainsi la taille de l'espace estimée précédemment. Cette complexité algorithmique inhérente aux méthodes d'estimation conjointe de fréquences fondamentales est bien connue et a déjà été évoquée dans la partie 1.3.2 (p. 31). Pour l'amener à un niveau convenable, nous proposons de réduire le nombre de notes possibles à travers une étape préliminaire de sélection de notes candidates. En outre, nous verrons que cette étape permet d'estimer précisément les valeurs de fréquence fondamentale et d'inharmonicité des notes sélectionnées. Il faudra ensuite s'assurer que le coût de chaque évaluation de la vraisemblance $p\left(x|\mathcal{C}\right)$ est suffisamment faible pour permettre ces évaluations sur l'ensemble restreint $\widetilde{\boldsymbol{\mathcal{C}}}$ des mélanges formés de notes candidates.

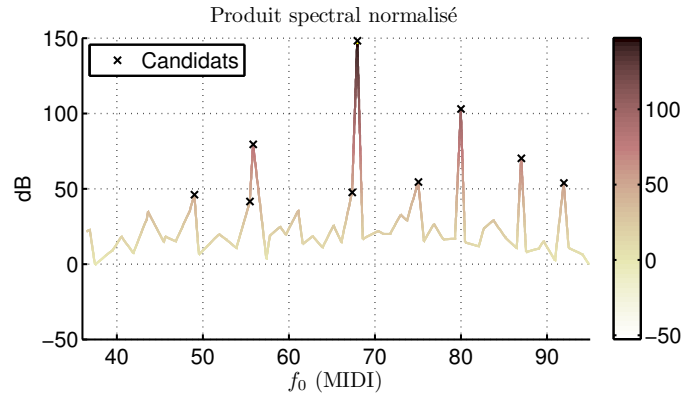
Pour sélectionner les candidats, nous utilisons le produit spectral normalisé (par le nombre de partiels) dont l'expression en décibels est

$$\Pi_X : (f_0,\beta) \mapsto \frac{1}{H\left(f_0,\beta\right)^{\nu}} 10\log \prod_{h=1}^{H(f_0,\beta)} \left| X\left(hf_0\sqrt{1+\beta\left(f_0\right)h^2}\right)\right|^2 \tag{4.29}$$

où $H(f_0, \beta)$ est le nombre de composantes de la note de fréquence fondamentale f_0 et d'inharmonicité β, ν est un paramètre fixé à $0,38$ et X est le spectre observé. Le facteur de normalisation permet de comparer le produit spectral des différentes notes, c'est-à-dire lorsque le nombre de partiels varie. Cette fonction présente des maxima locaux au niveau des notes présentes dans le signal, en raison du caractère énergétique sur lequel elle repose. Les autres maxima locaux correspondent à des fréquences fondamentales en rapport harmonique avec les notes présentes. Pour chaque note de la tessiture, nous appliquons la méthode d'optimisation utilisée dans la partie 2.2.2 (p. 55) pour maximiser le produit spectral à partir d'une grille dans la région du plan (f_0, β) concernée. La sélection des notes candidates consiste ensuite à retenir celles correspondant aux N_c plus grandes valeurs. Les deux étapes sont illustrées sur la figure 4.5.



(a) Optimisation locale du produit spectral normalisé pour chaque note de la tessiture.



(b) Sélection de candidats à partir des valeurs optimales obtenues pour chaque note.

FIGURE 4.5 – Sélection de notes candidates par le produit spectral normalisé $\Pi_X(f_0)$. Exemple sur un La♭ 3 (MIDI 68).

Comme nous l'avons vu dans la partie 2.2.2 (p. 55), la méthode est particulièrement efficace et rapide pour estimer précisément la fréquence fondamentale et l'inharmonicité. L'estimation reste bonne dans le cas d'un mélange de notes. Nous fixons en pratique le nombre de candidats à $N_c = 9$. Le nombre de mélanges passe alors de 56.10^6 à $\sum_{P=0}^{P_{\max}} \binom{N_c}{P} = 466$, pour $Q = 60$ et $P_{\max} = 6$ (nous verrons dans la partie 6.3.1 (p. 145) les performances de cette sélection de candidats).

4.4 Estimation itérative des modèles d'enveloppe spectrale et des amplitudes des partiels

4.4.1 Principe

Nous disposons à ce stade de notes candidates, dont les fréquences fondamentales et les coefficients d'inharmonicité ont été estimés. Nous sélectionnons alors un mélange $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_P)$ possible, qui est composé de P notes (cf. partie 4.2.1 (p. 83)), la note p ayant une fréquence fondamentale et une inharmonicité $\mathcal{C}_p = (f_{0,p}, \beta_p)$. Nous souhaitons alors trouver les paramètres optimaux du modèle associé à ce mélange, comme l'explique la partie 4.2.2. Nous décrivons ici l'estimation des paramètres des modèles d'enveloppe spectrale θ et des amplitudes des partiels α. Nous allons voir qu'il est possible d'estimer θ en connaissant α, et α en connaissant θ, puis proposerons un algorithme itératif qui alterne ces deux étapes et affine progressivement les estimations.

4.4.2 Estimation des modèles d'enveloppe spectrale

Supposons l'ensemble des amplitudes α connu et intéressons-nous à l'estimation des paramètres des modèles d'enveloppe spectrale θ.

Nous avons vu que l'optimisation par rapport au paramètre d'enveloppe spectrale θ consistait à maximiser la vraisemblance des amplitudes qui, compte tenu de l'indépendance entre les modèles, s'écrit

$$p\left(\alpha|\theta,\mathcal{C}\right) = \prod_{p=1}^{P} p\left(\alpha^{(p)}|\theta_p,\mathcal{C}_p\right) \tag{4.30}$$

L'optimisation par rapport à θ consiste donc à maximiser indépendamment la vraisemblance $p\left(\alpha^{(p)}|\theta_p,\mathcal{C}_p\right)$ des amplitudes de chaque note p par rapport à $\theta_p = \left(\sigma_p^2, A_p(z)\right)$. D'après le modèle donné par l'équation (4.5) (p.83), la log-vraisemblance à maximiser est

$$L_p\left(\sigma_p^2, A_p\right) \triangleq \ln p\left(\alpha_1^{(p)}, \ldots, \alpha_{H_p}^{(p)}|\sigma_p^2, A_p\right) \tag{4.31}$$

$$= -\frac{H_p}{2}\ln 2\pi\sigma_p^2 + \frac{1}{2}\sum_{h=1}^{H_p}\ln\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2 - \frac{1}{2\sigma_{(p)}^2}\sum_{h=1}^{H_p}\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2 \tag{4.32}$$

La maximisation par rapport à σ_p^2 s'obtient en annulant la dérivée de cette expression par rapport à σ_p^2, et donne

$$\widehat{\sigma_p^2} = \frac{1}{H_p}\sum_{h=1}^{H_p}\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2 \tag{4.33}$$

En injectant cette valeur dans l'expression (4.32), nous obtenons

$$L_p\left(\widehat{\sigma_p^2}, A_p\right) = -\frac{H_p}{2}\ln 2\pi e - \frac{H_p}{2}\ln\frac{1}{H_p}\sum_{h=1}^{H_p}\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2 + \frac{1}{2}\sum_{h=1}^{H_p}\ln\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2 \tag{4.34}$$

$$= c + \frac{H_p}{2}\ln\rho\left(A_p\right) \tag{4.35}$$

avec

$$c \triangleq -\frac{H_p}{2}\ln 2\pi e - \frac{1}{2}\sum_{h=1}^{H_p}\ln\left|\alpha_h^{(p)}\right|^2 \tag{4.36}$$

$$\rho\left(A_p\right) \triangleq \frac{\left(\prod_{h=1}^{H_p}\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2\right)^{\frac{1}{H_p}}}{\frac{1}{H_p}\sum_{h=1}^{H_p}\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2} \tag{4.37}$$

L'optimisation consiste donc à maximiser le terme $\rho\left(A_p\right)$, c étant constant par rapport à A_p. $\rho\left(A_p\right)$ est le rapport de la moyenne géométrique et de la moyenne arithmétique d'une quantité spectrale. Ce rapport est couramment appelé *platitude spectrale*, et mesure la *blancheur* de la quantité concernée. C'est un réel compris entre 0 et 1 qui atteint sa valeur maximale lorsque les données sont égales à une constante, et des valeurs plus faibles si elles ne sont pas « plates ». La platitude spectrale mesurée ici est celle de $\left|\alpha_h^{(p)}\right|^2\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2$, c'est-à-dire des amplitudes $\left|\alpha_h^{(p)}\right|^2$ après filtrage par $A_p(z)$, soit l'inverse du filtre du modèle d'enveloppe. $\rho\left(A_p\right)$ mesurant ainsi la capacité de $A_p(z)$ à blanchir les amplitudes, le filtre $\widehat{A_p}(z)$ optimal est celui qui modélise au mieux les amplitudes. Badeau et David [2008] ont montré que la solution optimale était justement fournie par l'algorithme présenté dans la partie 2.3 (p. 61). Il suffit donc d'appliquer cette méthode d'estimation afin d'obtenir une estimation optimale des paramètres du modèle autorégressif d'enveloppe spectrale de la note à partir de l'observation des carrés des amplitudes $\left|\alpha_1^{(p)}\right|^2, \ldots, \left|\alpha_{H_p}^{(p)}\right|^2$ de ses partiels.

4.4.3 Estimation des amplitudes des partiels

Nous supposons ici que les paramètres $(\theta_1, \ldots, \theta_P) = \left(\left(\sigma_1^2, A_1\right), \ldots, \left(\sigma_P^2, A_P\right)\right)$ des modèles d'enveloppe spectrale sont connus. Nous allons voir qu'en cas de recouvrement spectral, l'estimation des amplitudes des partiels est un problème d'estimation de variables aléatoires latentes à partir de l'observation du spectre et des caractéristiques statistiques de ces variables aléatoires, fournies par les paramètres d'enveloppe spectrale. Nous établissons alors l'expression d'un estimateur linéaire optimal des amplitudes.

En l'absence de recouvrement spectral et sous l'hypothèse que la puissance du bruit est négligeable devant celle des composantes sinusoïdales, l'estimation de l'amplitude du partiel h de la note p est simplement la valeur $X\left(f_h^{(p)}\right)$ du spectre observé à la fréquence du partiel. En présence de recouvrement spectral, identifier $\alpha_h^{(p)}$ à $X\left(f_h^{(p)}\right)$ conduit à une

mauvaise estimation, le spectre observé résultant de la contribution de plusieurs partiels aux fréquences proches. Ces contributions sont difficiles à déterminer en raison des phases inconnues des composantes et la tâche relève de la séparation de sources, dont nous nous inspirons ici en reprenant la technique du filtrage de Wiener et en l'adaptant. L'approche consiste à utiliser l'information statistique disponible sur les amplitudes pour concevoir des estimateurs.

D'après l'équation (4.5) (p. 83), l'amplitude $\alpha_h^{(p)}$ est une variable aléatoire telle que

$$\alpha_h^{(p)} \sim \mathcal{N}\left(0, v_{p,h}\right) \tag{4.38}$$

avec

$$v_{p,h} \triangleq \frac{\sigma_p^2}{\left|A_p\left(e^{2i\pi f_h^{(p)}}\right)\right|^2} \tag{4.39}$$

C'est cette information que nous allons utiliser pour estimer les amplitudes. Nous commençons par analyser le cas le plus général, pour lequel nous construisons l'estimateur d'une amplitude en prenant en compte l'influence de toutes les composantes présentes, puis expliquons comment simplifier les calculs en ne considérant que les composantes dont le recouvrement spectral est significatif.

Dans le cas général, nous pouvons réécrire le modèle de son x comme une somme de K sinusoïdes et de bruit :

$$x(t) = \sum_{k=1}^{K} \alpha_k e^{2i\pi f_k t} + b(t) \tag{4.40}$$

avec

$$\alpha_k \sim \mathcal{N}\left(0, v_k\right) \tag{4.41}$$

$$K = 2\sum_{p=1}^{P} H_p \tag{4.42}$$

Cette réécriture permet de passer des couples d'indices (p, h) relatifs aux notes et aux partiels à un seul indice k. L'hypothèse déjà évoquée sur le rapport signal à bruit permet par ailleurs de négliger les coefficients spectraux du bruit aux fréquences considérées. Cette hypothèse permet de simplifier les calculs et est en pratique vérifiée puisque nous ne considérerons que les fréquences f_k relatives à des pics spectraux.

Nous observons la transformée de Fourier discrète X de x sur une trame de longueur N, avec une fenêtre de pondération $w(n)$:

$$X\left(f\right) = \sum_{k=1}^{K} \alpha_k W\left(f - f_k\right) \tag{4.43}$$

où W est la transformée de Fourier discrète de w.

Estimation des α_k

Dans le cas de notre modèle, pour $1 \leq k_0 \leq K$, l'estimateur linéaire $\hat{\alpha}_{k_0}$ de α_{k_0} en fonction de $X(f_{k_0})$, optimal au sens de la minimisation de l'erreur quadratique moyenne est

$$\hat{\alpha}_{k_0} = \frac{W^*(0)\, v_{k_0}}{\sum_{k=1}^{K} |W(f_{k_0} - f_k)|^2 v_k} X(f_{k_0}) \tag{4.44}$$

et l'erreur d'estimation est

$$\begin{aligned} \epsilon_{k_0} &= \mathbb{E}\left[|\alpha_{k_0} - \hat{\alpha}_{k_0}|^2\right] \\ &= \left(1 - \frac{|W(0)|^2 v_{k_0}}{\sum_{k=1}^{K} |W(f_{k_0} - f_k)|^2 v_k}\right) v_{k_0} \end{aligned} \tag{4.45}$$

La démonstration des équations (4.44) et (4.45) figure en annexe B.2 (p. 174).

Analogie avec le filtrage de Wiener

La démarche pour déterminer l'estimateur des amplitudes (équation (4.44)) s'inspire du filtrage de Wiener [Wiener, 1964]. Rappelons son principe afin de comprendre l'analogie avec notre démarche.

Soit un signal $s(t)$ composé d'une somme de P sources $s_1(t), \ldots, s_P(t)$. Pour $1 \leq p \leq P$, la source p est une réalisation d'un processus réel stationnaire au sens large. On suppose que ce processus est paramétrique. On cherche à estimer $s_p(t)$ à partir de l'observation de $s(t)$ et de la connaissance des propriétés statistiques du second ordre des sources.

L'estimateur optimal, au sens de l'erreur quadratique moyenne $\mathbb{E}\left[(s(t) - s_p(t))^2\right]$, est l'espérance conditionnelle $s_{\text{opt}}(t) = \mathbb{E}[s_p(t)\,|s(\tau), \tau \in \mathbb{N}]$. Cet estimateur n'est pas toujours aisément calculable, on se contente alors d'un estimateur linéaire. L'estimateur linéaire optimal au sens de l'erreur quadratique moyenne résulte du filtrage du signal $s(t)$ par le filtre de Wiener dont la réponse en fréquence est

$$H_p(f) = \frac{S_{s,s_p}(f)}{S_s(f)} \tag{4.46}$$

où $S_s(f)$ est la densité spectrale de puissance de s et S_{s,s_p} l'interspectre de s et s_p. Dans le cas où les sources sont décorrélées, on a

$$H_p(f) = \frac{S_p(f)}{\sum_{q=1}^{P} S_q(f)} \tag{4.47}$$

$S_q(f)$ étant la densité spectrale de puissance de $s_q(t)$.

Le filtrage de Wiener permet donc d'estimer (de « séparer ») une source à partir de l'observation du mélange $s(t)$ et de la connaissance des propriétés statistiques des sources présentes, en particulier de leur densité spectrale de puissance. À une fréquence f donnée, le coefficient de Fourier de l'estimée correspond ainsi au coefficient de Fourier du mélange atténué par un gain correspondant au rapport des densités spectrales de puissance de la source et du mélange.

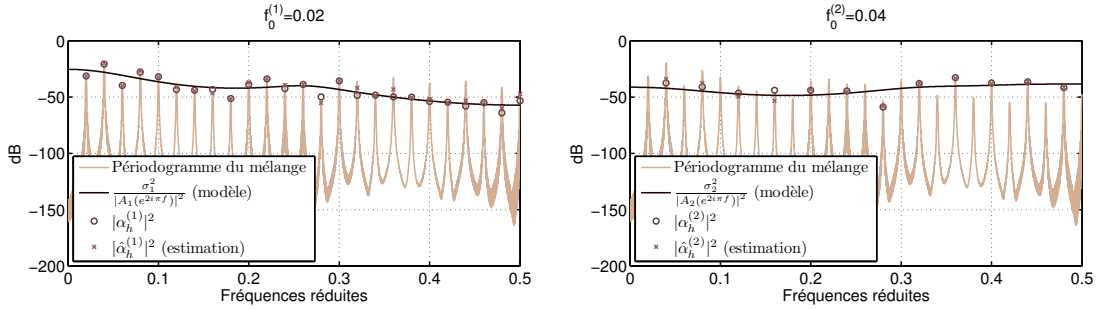
Notre approche est similaire à celle du filtrage de Wiener à plusieurs égards. Les deux cas présentent un problème d'estimation de variables aléatoires dont on observe une statistique – c'est-à-dire une combinaison – et dont on connaît les propriétés statistiques du

second ordre. Il en résulte une démarche similaire pour trouver l'estimateur linéaire optimal. Les résultats sont également comparables, dans la mesure où le « gain » trouvé dans les expressions (4.44) et (4.47) est un rapport de deux quantités quadratiques relatives à la source que l'on estime et au signal observé. Ces quantités correspondent aux propriétés statistiques du second ordre des modèles : densité spectrale de puissance dans le cas du filtrage de Wiener, et variances dans celui du processus harmonique.

La différence entre les deux approches réside dans le fait que le filtrage de Wiener s'applique dans le cas stationnaire et en présence de densités spectrales de puissance alors que dans notre cas, nous cherchons à modéliser le phénomène de recouvrement, dû à une analyse à court terme et à l'étalement spectral qui en résulte, absent du cas précédent. Ainsi, notre approche répond à un problème d'estimation des amplitudes, différent d'une problématique de filtrage (lors de l'utilisation d'un filtre de Wiener pour la séparation de sources par exemple).

Application à la séparation de deux notes avec recouvrement spectral

À titre d'illustration, l'estimateur est utilisé sur un mélange synthétique de deux processus harmoniques, chacun ayant des composantes à des fréquences multiples d'une fréquence fondamentale. Un recouvrement partiel entre les spectres des deux sources est obtenu en choisissant des fréquences fondamentales en rapport harmonique. La figure 4.6 donne un aperçu du résultat. L'estimation et l'enveloppe spectrale originale se superposent lorsqu'il n'y a pas de recouvrement spectral (pics d'ordres impairs de la source 1). En cas de recouvrement, l'estimation est bonne pour le pic prédominant éventuel (cf. pic de la source 1 à $f = 0,04$ ou de la source 2 à $f = 0,04$), ainsi que lorsque les deux amplitudes sont du même ordre (cf. pics de la source 1 à $f \approx 0,04$ et de la source 2 à $f > 0,3$). Lorsque l'amplitude d'une source est petite devant celle de l'autre, l'estimation peut se dégrader (cf. pic de la source 1 $f = 0,36$).



(a) Note 1. (b) Note 2.

Figure 4.6 – Estimation des amplitudes avec recouvrement spectral sur un signal synthétique représentant deux notes. Les fréquences fondamentales $f_0^{(1)}$ et $f_0^{(2)}$ des deux notes sont en rapport d'octave et l'estimation est faite sur une observation de $N = 2048$ échantillons. Pour chaque note, les amplitudes (cercles) sont générées à partir du modèle d'enveloppe spectrale (traits noirs) via les équations (4.38) et (4.39). Les croix représentent l'estimation de ces amplitudes (équation (4.44)).

Estimation des amplitudes des partiels

D'après l'équation (4.45), l'erreur d'estimation est grande lorsqu'il existe au moins une valeur $k \neq k_0$ telle que $|W(f_{k_0} - f_k)|^2 v_k$ n'est pas négligeable devant $|W(0)|^2 v_{k_0}$. C'est

le cas lorsque la distance $|f_{k_0} - f_k|$ est petite devant la largeur du lobe principal de w et que la variance v_k de l'amplitude de la composante parasite est de l'ordre de grandeur de v_{k_0} : il y a alors recouvrement spectral. À l'inverse, il est possible de négliger l'influence des composantes dont la fréquence est éloignée. En revenant au problème initial de l'estimation des amplitudes des partiels, nous définissons l'estimateur de $\alpha_h^{(p)}$ par

$$\widehat{\alpha_h^{(p)}} \triangleq \frac{W^*(0)\, v_{p,h}}{\sum_{\left|f_{h'}^{(p')} - f_h^{(p)}\right| < \Delta_w} \left|W\left(f_{h'}^{(p')} - f_h^{(p)}\right)\right|^2 v_{p',h'}} X\left(f_h^{(p)}\right) \tag{4.48}$$

où Δ_w est la largeur du lobe principal de w.

4.4.4 Algorithme itératif d'estimation des paramètres des notes

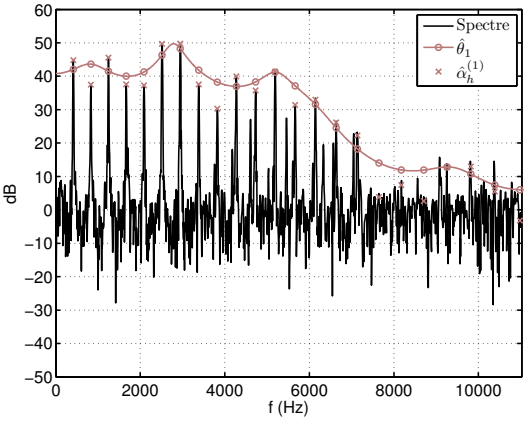
Nous proposons alors l'algorithme itératif suivant, qui alterne les phases d'estimation des amplitudes et des modèles d'enveloppe spectrale, en ajoutant un seuillage des amplitudes trop faibles :

Entrées: spectre X, notes $\mathcal{C}_1, \ldots, \mathcal{C}_P$.
Initialiser les amplitudes des partiels $\alpha_h^{(p)(0)}$ aux valeurs du spectre $X\left(f_h^{(p)}\right)$ pour $p \in [\![1; P]\!]$ et $h \in [\![1; H_p]\!]$.
Pour chaque itération i
Pour chaque note p
Estimer $\theta_p^{(i)}$ à partir de $\alpha^{(p)(i-1)}$. {estimation AR}
Fin Pour
Estimer conjointement tous les $\alpha_h^{(p)(i)}$ à partir des $X\left(f_h^{(p)}\right)$ et des $\theta_p^{(i)}$. {éq. (4.48)}
$\alpha_h^{(p)(i)} \leftarrow \max\left(\alpha_h^{(p)(i)}, \min\left(|X|\right)\right)$. {seuillage des amplitudes trop faibles}
Fin Pour
Sorties: estimation des $\alpha_h^{(p)}$ et θ_p pour $p \in [\![1; P]\!]$ et $h \in [\![1; H_p]\!]$.

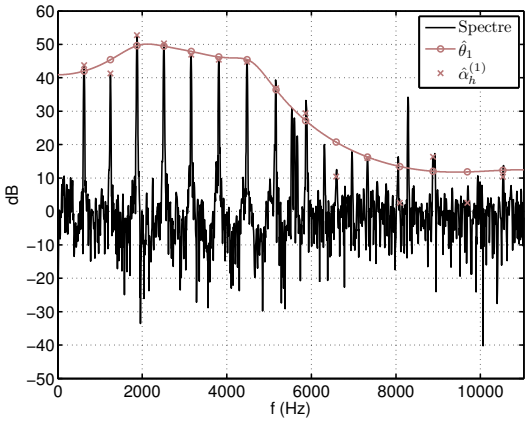
Algorithme 4.1: Estimation itérative des paramètres et des sources

Pour $p \in [\![1; P]\!]$ et $h \in [\![1; H_p]\!]$, les estimations $\widehat{\alpha_h^{(p)}}$ de $\alpha_h^{(p)}$ et $\widehat{\theta_p}$ de θ_p sont alors définies comme les valeurs respectives des suites $\left(\alpha_h^{(p)(i)}\right)$ et $\left(\theta_p^{(i)}\right)$ après un certain nombre d'itérations. La convergence de l'algorithme n'est pas prouvée mais a été constatée et l'on peut voir que la technique s'apparente à l'algorithme EM [Dempster *et al.*, 1977] dans sa manière de considérer des variables latentes – les amplitudes et les enveloppes – et des observations – les coefficients spectraux.

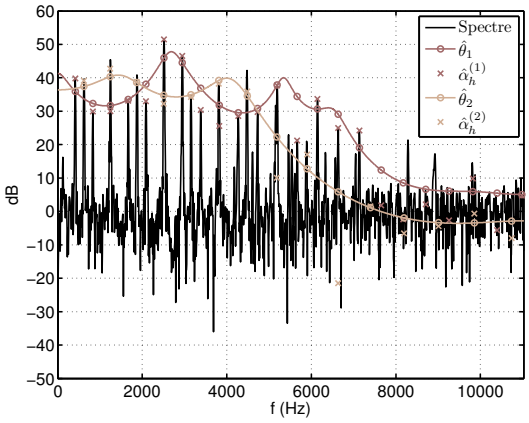
Le fonctionnement de l'algorithme est illustré sur les figures 4.7 et 4.8. Lorsque l'on analyse un signal à partir d'un modèle de note seule, les amplitudes des partiels sont directement accessibles à partir du spectre. Il n'y pas lieu de réaliser plusieurs itérations : une seule estimation AR suffit. C'est le cas sur les figures 4.7(a), 4.7(b), 4.7(d), 4.8(a) et 4.8(b). Dans les trois premiers cas, le modèle considéré correspond à la note contenue dans le signal ; l'enveloppe spectrale et les amplitudes des partiels sont alors correctement estimés. Lorsque l'on cherche les paramètres d'un modèle de mélange de plusieurs notes (figures 4.7(c) et 4.7(e)), l'algorithme itératif permet de prendre en compte le recouvrement spectral. Dans le cas d'une quinte (figure 4.7(c)), les partiels isolés de chaque note assurent une bonne estimation des deux enveloppes spectrales, permettant à leur tour d'estimer les



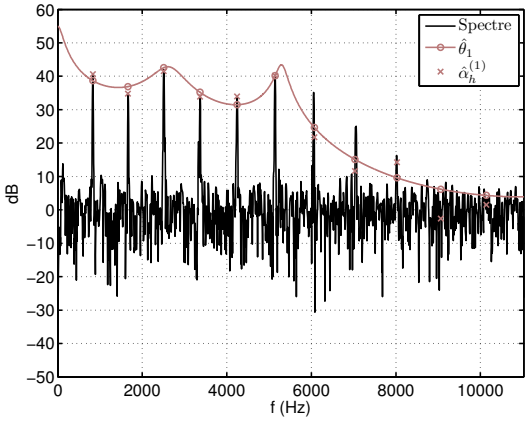
(a) Lab 3 (415 Hz) analysé par un modèle de Lab 3.



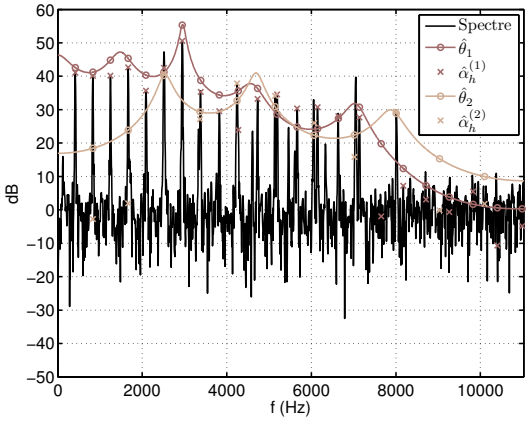
(b) Mib 4 (622 Hz) analysé par un modèle de Mib 4.



(c) Mélange Lab 3 (415 Hz) + Mib 4 (622 Hz) analysé par un modèle de Lab 3 + Mib 4.

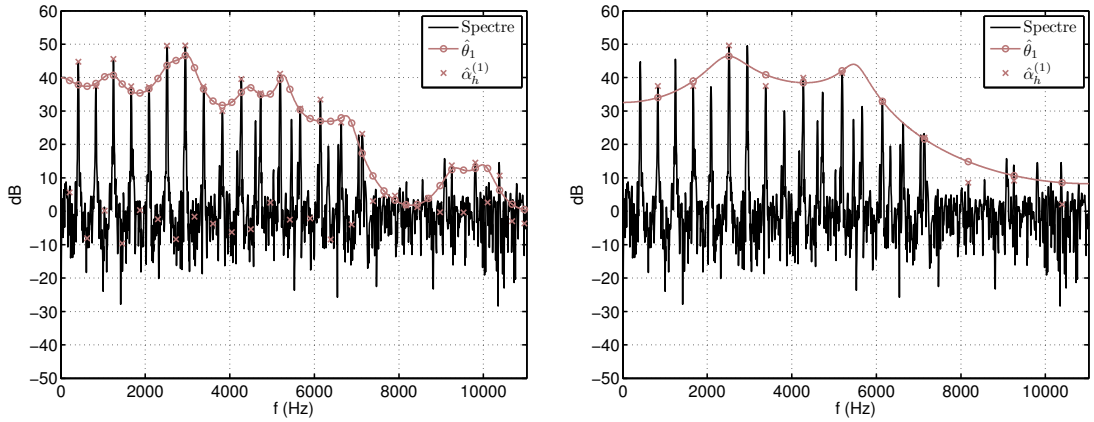


(d) Lab 4 (831 Hz) analysé par un modèle de Lab 3.



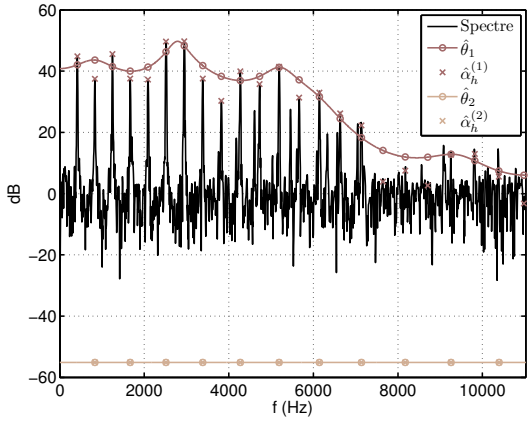
(e) Mélange Lab 3 (415 Hz) + Lab 4 (831 Hz) analysé par un modèle de Lab 3 + Lab 4.

Figure 4.7 – Estimation des amplitudes $\widehat{\alpha_p}$ et des enveloppes spectrales $\widehat{\theta}_p$ lorsque le modèle de mélange correspond au contenu du son. Illustration sur des notes isolées, un intervalle de quinte et une octave.



(a) Lab 3 (415 Hz) analysé par un modèle de Lab 2 (208 Hz).

(b) Lab 3 (415 Hz) analysé par un modèle de Lab 4 (831 Hz).



(c) Lab 3 (415 Hz) analysé par un modèle de Lab 3 (415 Hz) + Lab 4 (831 Hz).

FIGURE 4.8 – Estimation des amplitudes $\widehat{\alpha_p}$ et des enveloppes spectrales $\widehat{\theta_p}$ lorsque le modèle de mélange ne correspond pas au contenu du son.

amplitudes des partiels qui se recouvrent. L'estimation est plus difficile mais donne des résultats intéressants lorsqu'il s'agit d'une octave, cas pour lequel seule la note la plus grave possède des partiels isolés.

Avec les exemples de la figure 4.8, nous nous penchons sur le comportement de l'algorithme lorsque le modèle considéré ne correspond pas aux notes contenues dans le signal. La figure 4.8(a) représente le cas d'un modèle à l'octave inférieure de la véritable note : les amplitudes estimées sont alternativement dans le bruit et sur une composante sinusoïdale. Le modèle d'enveloppe estimé est alors sous-optimal, une enveloppe AR d'ordre faible ne permettant pas de modéliser une alternance de coefficients spectraux faibles et forts. En particulier, la platitude spectrale $\rho(A_p)$ obtenue (équation 4.37) est faible : le modèle a alors peu de chance d'être sélectionné, au profit du modèle associé à la véritable note. Dans le cas d'une erreur d'octave (figure 4.8(b)), l'enveloppe spectrale est certes bien estimée, mais un partiel sur deux reste dans le résiduel. Dans le dernier exemple (figure 4.8(c)), le signal contient une note unique alors que l'on essaie de le modéliser par cette note et son octave. L'algorithme itératif parvient à détecter des amplitudes négligeables pour la note absente, les amplitudes des partiels d'ordre pair de la note grave ne laissant pas présager l'octave.

4.5 Estimation des paramètres du modèle de bruit

En établissant notre modèle dans la partie 4.2.1, nous avons introduit le bruit $x_b(n)$ comme un processus MA paramétré par une puissance σ_b^2 et un filtre $B(z)$. Plus précisément, on suppose que la trame $x_b \triangleq (x_b(0), \ldots, x_b(N-1))$ résulte du filtrage circulaire d'un bruit blanc gaussien x_{bb} de puissance σ_b^2 par $B(z)$, dont l'ordre Q_b est supposé inférieur à N et que l'on écrit sous la forme

$$B(z) = \sum_{k=0}^{Q_b} b_k z^{-k} \tag{4.49}$$

avec $b_0 = 1$. Cette hypothèse sur le processus MA est valable asymptotiquement. Elle permet de simplifier certains calculs comme expliqué en annexe (cf. partie A.2.2 (p. 170)).

D'après la proposition A.2.14 (p. 171), l'expression de la log-vraisemblance est alors

$$L_b\left(\sigma_b^2, B\right) = -\frac{N}{2}\ln 2\pi\sigma_b^2 - \frac{1}{2}\sum_{k=0}^{N-1}\ln\left|B\left(e^{2i\pi\frac{k}{N}}\right)\right|^2 - \frac{1}{2\sigma_b^2 N}\sum_{k=0}^{N-1}\left|\frac{X_b\left(\frac{k}{N}\right)}{B\left(e^{2i\pi\frac{k}{N}}\right)}\right|^2 \tag{4.50}$$

La log-vraisemblance peut donc se calculer à partir de l'observation du spectre du bruit. Dans notre cas, nous avons supposé que les coefficients spectraux du bruit étaient négligeables devant ceux des notes au niveau des fréquences des partiels. Ils ne sont donc pas facilement observables et nous nous contentons de l'information portée par les coefficients dont les fréquences se situent en dehors des lobes principaux des partiels, c'est-à-dire sur le support fréquentiel défini par

$$\mathcal{F}_b \triangleq \left\{ \frac{k}{N} \,\middle/\, \forall f' \in \bigcup_{p=1}^{P} \mathcal{F}_p, \left|\frac{k}{N} - f'\right| > \frac{\Delta_w}{2} \right\} \tag{4.51}$$

où Δ_w désigne la largeur d'un lobe principal de la fenêtre w ($\Delta_w = \frac{4}{N}$ pour une fenêtre de Hann). Cette approximation est asymptotiquement valable, lorsque $N \to +\infty$, c'est-à-dire

lorsque le nombre d'éléments supprimés du vecteur d'observation est petit devant la taille de ce vecteur. L'expression de la log-vraisemblance (4.50) devient [3]

$$L_b\left(\sigma_b^2, B\right) \approx -\frac{\#\mathcal{F}_b}{2}\ln 2\pi\sigma_b^2 - \frac{1}{2}\sum_{f\in\mathcal{F}_b}\ln\left|B\left(e^{2i\pi f}\right)\right|^2 - \frac{1}{2\sigma_b^2}\sum_{f\in\mathcal{F}_b}\frac{1}{N}\left|\frac{X_b(f)}{B\left(e^{2i\pi f}\right)}\right|^2 \quad (4.52)$$

En raison du caractère gaussien des coefficients spectraux du bruit, l'expression (4.52) est relativement similaire à l'expression (4.32) (p. 92) de la log-vraisemblance des amplitudes. Aussi, la maximisation de (4.52) par rapport à σ_b^2 puis à B s'obtient de manière analogue. L'estimée de la puissance du bruit est

$$\widehat{\sigma_b^2} = \frac{1}{\#\mathcal{F}_b}\sum_{f\in\mathcal{F}_b}\frac{1}{N}\left|\frac{X_b(f)}{B\left(e^{2i\pi f}\right)}\right|^2 \quad (4.53)$$

En injectant cette valeur dans l'expression (4.52), nous obtenons

$$L_b\left(\widehat{\sigma_b^2}, B\right) = c_b + \frac{\#\mathcal{F}_b}{2}\ln\rho_b(B) \quad (4.54)$$

avec

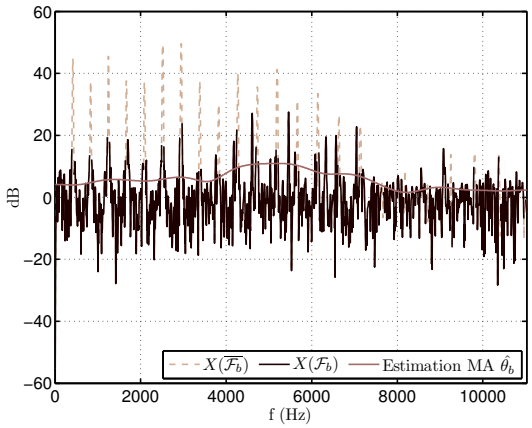
$$c_b \triangleq -\frac{\#\mathcal{F}_b}{2}\ln 2\pi e - \frac{1}{2}\sum_{f\in\mathcal{F}_b}\ln\frac{|X_b(f)|^2}{N} \quad (4.55)$$

$$\rho_b(B) \triangleq \frac{\left(\prod_{f\in\mathcal{F}_b}\left|\frac{X_b(f)}{B\left(e^{2i\pi f}\right)}\right|^2\right)^{\frac{1}{\#\mathcal{F}_b}}}{\frac{1}{\#\mathcal{F}_b}\sum_{f\in\mathcal{F}_b}\left|\frac{X_b(f)}{B\left(e^{2i\pi f}\right)}\right|^2} \quad (4.56)$$

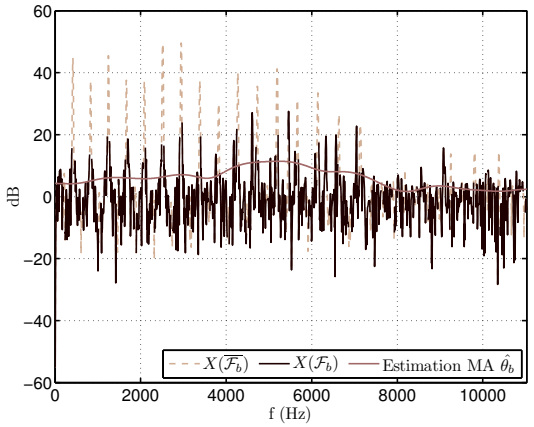
Nous retrouvons la maximisation d'une platitude spectrale, $\rho_b(B)$. La solution est cette fois-ci obtenue en effectuant une estimation des paramètres du filtre MA à partir des observations partielles du spectre X aux fréquences $\mathcal{F}_b$. De même que dans le cas des AR, l'estimation par l'algorithme de Badeau et David [2008] est optimale au sens de la maximisation de (4.56). Cependant, les coefficients spectraux omis étant peu nombreux par rapport aux coefficients observés, nous optons pour la technique d'estimation à partir du spectre entier, un peu moins efficace mais plus rapide, présentée dans la partie 2.4 (p. 66). Nous appliquons donc l'algorithme 2.2 (p. 68) en considérant l'autocorrélation empirique obtenue à partir des observations du spectre sur le support $\mathcal{F}_b$. Une bonne modélisation du bruit est obtenue en fixant l'ordre du modèle MA à $Q_b = 20$.

La figure 4.9 illustre cette étape d'estimation des paramètres du modèle de bruit. Le même spectre est successivement analysé avec le modèle de la véritable note, puis ceux de la sous-octave et de l'octave. Lorsque l'on a sélectionné le modèle de la note originale (figure 4.9(a)) ou celui de la sous-octave (figure 4.9(b)), les fréquences des partiels ont été enlevées du support fréquentiel $\mathcal{F}_b$ du modèle de bruit : les coefficients observés sont alors bien modélisés par un modèle MA. Lorsqu'il reste des sinusoïdes dans le support fréquentiel $\mathcal{F}_b$ du modèle de bruit, dans le cas de l'octave (figure 4.9(c)) mais également de

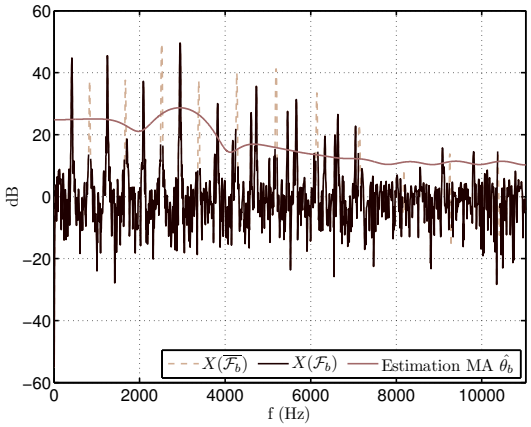
3. Dans cette expression, la taille de l'observation est $\#\mathcal{F}_b$, contre N dans l'expression (4.50). Nous notons donc que ces deux log-densités ne sont pas homogènes entre elles (ce qui n'est pas gênant pour la suite).



(a) Estimation avec le modèle Mi♭ 3 (311 Hz, note originale).



(b) Estimation avec le modèle Mi♭ 2 (156 Hz, sous-octave).



(c) Estimation avec le modèle Mi♭ 4 (622 Hz, octave).

Figure 4.9 – Exemple d'estimation des paramètres du modèle de bruit sur un Mi♭ 3 (311 Hz).

toute note qui n'est pas un sous-harmonique, la modélisation MA n'est pas adaptée : les paramètres estimés modélisent mal les données et la vraisemblance associée, faible, permet de rejeter ce modèle qui ne correpond pas à la véritable note. On remarque par ailleurs que les coefficients spectraux attribués à du bruit sont parfois des lobes secondaires de partiels. Le phénomène n'est cependant pas gênant en pratique car l'enveloppe spectrale du « bruit » résultant, qui inclut les lobes secondaires, conserve les qualités de régularité déterminantes pour notre modèlisation.

Nous voyons ainsi à travers la modélisation des spectres de notes et du bruit résiduel que les données sont bien modélisées lorsque l'on sélectionne le modèle de mélange original (figures 4.7(a) (p. 98) et 4.9(a) par exemple), tandis qu'un des modèles au moins – celui de spectre de note ou celui de bruit – permet de rejeter le mélange testé (figures 4.8(a) (p. 99) et 4.9(b) pour la sous-octave, 4.8(b) (p. 99) et 4.9(c) pour l'octave).

4.6 Lois *a priori* sur les modèles AR et MA

Nous n'avons pour l'instant pas spécifié les lois *a priori* $p(\theta)$ et $p(\theta_b)$, l'estimation des paramètres θ et θ_b ayant été réalisées par maximum de vraisemblance. L'introduction de ces lois permet néanmoins de fournir une information supplémentaire qui, même si elle est négligée dans l'étape d'estimation des paramètres, est utile dans l'expression de la vraisemblance $p(x|\mathcal{C})$. En l'absence de telles informations, le modèle proposé est très général puisque la seule contrainte sur les enveloppes des notes et du bruit est leur caractère lisse lié aux modélisations AR et MA d'ordres faibles. En particulier, les coefficients spectraux du bruit peuvent tout à fait être bien modélisés par un modèle de note, donnant lieu à la détection de fausses alarmes. Nous avons ici la possibilité de pénaliser les modèles pour lesquels la puissance du bruit est élevée ou la puissance des notes faible. Les lois *a priori* s'écrivent

$$p(\theta) = \prod_{p=1}^{P} p\left(\sigma_p^2\right) p(A_p) \tag{4.57}$$

$$p(\theta_b) = p\left(\sigma_b^2\right) p(B) \tag{4.58}$$

Le choix des lois *a priori* sur les puissances (variances) ne dépend que de la contrainte d'une puissance faible pour le bruit et forte pour les enveloppes spectrales. Dans ces conditions, on choisit habituellement des lois *a priori* telles que les lois Gamma ou Gamma Inverse, dont les définitions et propriétés sont données en annexe A.1.3 (p. 164). La famille de lois Gamma favorise les valeurs faibles de variable aléatoire, comme l'illutrent leurs densités représentées sur la figure A.1 (p. 165). Nous choisirons donc une loi Gamma pour la distribution de la puissance σ_b^2 du bruit. Si une variable aléatoire suit une loi Gamma, son inverse suit une loi Gamma Inverse. Pour favoriser des valeurs élevées, nous choisirons donc une loi Gamma Inverse pour la distribution des puissances des modèles d'enveloppes spectrales.

Les puissances suivent donc les lois suivantes :

$$\forall p \in [\![1;P]\!], \sigma_p^2 \sim \mathcal{IG}\left(k_{\sigma_p^2}, E_{\sigma_p^2}\right) \tag{4.59}$$

$$\sigma_b^2 \sim \Gamma\left(k_{\sigma_b^2}, E_{\sigma_b^2}\right) \tag{4.60}$$

où les paramètres de forme $k_{\sigma_p^2}$ et $k_{\sigma_b^2}$ et d'échelle $E_{\sigma_p^2}$ et $\theta_{E_b^2}$ ont été fixés empiriquement à $\left(k_{\sigma_p^2}, E_{\sigma_p^2}\right) \triangleq \left(1, 10^{-4}\right)$ et $\left(k_{\sigma_b^2}, E_{\sigma_b^2}\right) \triangleq \left(2, 10^{-3}\right)$.

Nous ne souhaitons pas introduire d'information supplémentaire via une loi *a priori* sur les filtres normalisés A_p et B. Une solution consisterait à considérer une loi non informative [Gelman *et al.*, 2004], par exemple de densité constante, associée à une distribution uniforme des pôles du filtre AR et zéros du filtre MA dans le disque ouvert de rayon unité. Nous allons voir dans la section 4.7 que nous pouvons tout simplement nous passer de ces lois *a priori* non informatives dans l'estimation de la vraisemblance $p\left(x|\mathcal{C}\right)$.

4.7 Fonction de détection de fréquences fondamentales

Nous avons vu dans la partie 4.2.2 (p. 87) que si l'on note Θ l'ensemble des paramètres d'un modèle $\mathcal{C}$ de mélange, l'estimation optimale des fréquences fondamentales consistait à maximiser la log-vraisemblance

$$L_x\left(\mathcal{C}\right) \triangleq \ln p\left(x|\mathcal{C}\right) \tag{4.61}$$

$$= \ln \int p\left(x, \Theta|\mathcal{C}\right) d\Theta \tag{4.62}$$

par rapport à tous les mélanges possibles $\mathcal{C} \in \widetilde{\boldsymbol{\mathcal{C}}}$, mais que le calcul de cette intégrale n'était pas réalisable en pratique. Par ailleurs, si l'on note $\widehat{\Theta} \triangleq \left(\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta_P}, \widehat{\theta}_b\right)$ l'estimation des paramètres pour un mélange $\mathcal{C}$, on ne peut se contenter de remplacer la maximisation de (4.62) par celle – toujours par rapport à $\mathcal{C}$ – de la fonction

$$\ln p\left(x, \widehat{\Theta}|\mathcal{C}\right) = \sum_{p=1}^{P} L_p\left(\widehat{\theta}_p\right) + L_b\left(\widehat{\theta}_b\right) + \sum_{p=1}^{P} \ln p\left(\widehat{\sigma_p^2}\right) + \ln p\left(\widehat{\sigma_b^2}\right) + \sum_{p=1}^{P} \ln p\left(\widehat{A_p}\right) + \ln p\left(\widehat{B}\right) \tag{4.63}$$

En effet, lorsque $\mathcal{C}$ varie, le nombre de notes P, de partiels observés H_p pour chaque note p, l'ordre du modèle d'enveloppe spectrale et le nombre d'observations $\#\mathcal{F}_b$ varient également ; les termes de la somme (4.63) ainsi que $\ln p\left(x|\widehat{\Theta}, \mathcal{C}\right)$ ne sont donc pas comparables d'un mélange $\mathcal{C} \in \widetilde{\boldsymbol{\mathcal{C}}}$ à l'autre. Nous allons voir qu'il est possible de corriger ces termes en s'appuyant sur la problématique de l'estimation d'ordre de modèles. Nous proposerons alors une vraisemblance pondérée dont l'expression est

$$\tilde{L}_x\left(\mathcal{C}\right) \triangleq w_1 \sum_{p=1}^{P} \tilde{L}_p\left(\widehat{\theta}_p\right)/P + w_2 \tilde{L}_b\left(\widehat{\theta}_b\right) + w_3 \sum_{p=1}^{P} \ln p\left(\widehat{\sigma_p^2}\right)/P + w_4 \ln p\left(\widehat{\sigma_b^2}\right) - \mu_{\text{pol}}\, P \tag{4.64}$$

où $\tilde{L}_p$ et $\tilde{L}_b\left(\widehat{\theta}_b\right)$ sont des versions « corrigées » de L_p et L_b, w_1, w_2, w_3, w_4 sont les coefficients de pondération et μ_{pol} est une pénalité sur la polyphonie P. Les corrections et la pondération introduites visent à rendre les valeurs de $\tilde{L}_x\left(\mathcal{C}\right)$ comparables lorsque $\mathcal{C}$ varie : l'estimation de fréquences fondamentales est alors obtenue en maximisant la fonction de détection $\mathcal{C} \mapsto \tilde{L}_x\left(\mathcal{C}\right)$.

Problématique de la sélection de modèle

Comme nous l'avons évoqué dans la partie 4.2.2 (p. 87), il est possible d'approcher $p\left(x|\mathcal{C}\right)$ par l'approximation de Laplace (cf. annexe A.3 (p. 172)), qui se réécrit sous forme logarithmique

$$\ln p\left(x|\mathcal{C}\right) = \ln \int p\left(x,\Theta|\mathcal{C}\right) d\Theta \tag{4.65}$$

$$\approx \ln p\left(x,\Theta^*|\mathcal{C}\right) + \ln g\left(x,\Theta^*,n\left(\mathcal{C}\right)\right) \tag{4.66}$$

avec

$$\Theta^* \triangleq \arg\max_{\Theta} p\left(x,\Theta|\mathcal{C}\right) \tag{4.67}$$

$$g\left(x,\Theta^*,n\left(\mathcal{C}\right)\right) = \left(2\pi\right)^{\frac{n(\mathcal{C})}{2}} \det\left(-\boldsymbol{\nabla^2}\left\{\log p\left(x,\Theta|\mathcal{C}\right)\right\}_{\Theta=\Theta^*}\right)^{-\frac{1}{2}} \tag{4.68}$$

$n\left(\mathcal{C}\right)$ étant l'ordre du modèle, c'est-à-dire le nombre de paramètres d'enveloppes spectrales, d'amplitudes et de bruit dans le cas présent.

Cette approximation est à la base des méthodes dites d'estimation (d'ordre) de modèles [Stoica et Selen, 2004] telles que BIC, AIC ou GIC. Plus précisément, ces critères consistent à développer l'approximation précédente – sur la base de principes propres à chaque critère – résultant en la simplification de la fonction g, de sorte que l'on obtient une expression de la forme

$$\ln p\left(x|\mathcal{C}\right) \approx \ln p\left(x|\Theta^*,\mathcal{C}\right) + \ln g\left(n\left(\mathcal{C}\right),\#x\right) \tag{4.69}$$

La fonction g ne dépend plus que de deux nombres, l'ordre du modèle $n\left(\mathcal{C}\right)$ et la taille de l'observation $\#x$. De plus, le logarithme de g est linéaire en $n\left(\mathcal{C}\right)$, avec une pente différente d'un critère à l'autre. C'est alors cette expression (équation (4.69)) qui est maximisée par rapport à $\mathcal{C}$ pour estimer le modèle le plus vraisemblable.

En pratique, les critères sont en général exprimés, de manière équivalente, comme la minimisation d'une expression de la forme

$$-2\ln p\left(x|\Theta^*,\mathcal{C}\right) + G\left(n\left(\mathcal{C}\right),\#x\right) \tag{4.70}$$

avec

$$G \triangleq -2\ln g \tag{4.71}$$

Par exemple, les critères définis précédemment donnent

$$G\left(n\left(\mathcal{C}\right),\#x\right) = \begin{cases} 2n\left(\mathcal{C}\right) & \text{pour AIC ;} \\ \ln\left(\#x\right) n\left(\mathcal{C}\right) & \text{pour BIC ;} \\ \eta n\left(\mathcal{C}\right) & \text{pour GIC, où } \eta \text{ est un paramètre.} \end{cases} \tag{4.72}$$

Ainsi, on peut interpréter l'approximation de l'intégrale (4.65) comme l'ajout d'une pénalité G à $-2\ln p\left(x|\Theta^*,\mathcal{C}\right)$ pour compenser les effets de la variation de l'ordre du modèle. Le choix du critère utilisé – et donc de la pénalité – dépend en général du contexte, et permet de choisir la pente de la fonction G. Nous pourrons ainsi fixer cette pente de manière empirique.

Approximation de la log-vraisemblance $L_x(\mathcal{C}) = \ln p(x|\mathcal{C})$

Dans notre cas, le type de méthode décrit précédemment ne peut être appliqué directement pour plusieurs raisons :

- en ne considérant que la vraisemblance $p(x|\Theta^*, \mathcal{C})$, elles ignorent toute information portée par une densité *a priori* $p(\Theta)$. En particulier, elles reviendraient à ne considérer ici que $p(x|\alpha, \theta_b, \mathcal{C})$ en négligeant $p(\alpha|\theta, \mathcal{C})$, $p(\theta)$ et $p(\theta_b)$ (cf. équation (4.23) (p. 88)) ;
- la structure des dépendances de notre modèle est composée de deux couches, la première étant les paramètres du modèle d'enveloppe spectrale et la seconde les amplitudes des partiels ;
- dans les vraisemblances p_{x_b} et p_α que nous considérons, le nombre effectif d'observations varie d'un mélange $\mathcal{C}$ à l'autre, alors que les méthodes de sélection de modèle reposent sur l'hypothèse que le nombre d'observations ne dépend pas du modèle ;

Nous proposons une approche pour intégrer des pénalisations sur l'ordre des modèles à la manière des critères de sélection de modèles afin d'estimer approximativement $p(x|\mathcal{C})$ à partir du modèle optimal $\widehat{\Theta} \triangleq \left(\widehat{\alpha}, \widehat{\theta}_1, \ldots, \widehat{\theta_P}, \widehat{\theta}_b\right)$ estimé précédemment, tout en prenant en compte les contraintes ci-dessus.

L'expression à pénaliser est celle de la log-densité des données et des paramètres estimés, donnée par l'équation (4.63), où L_p s'obtient par l'équation (4.35) (p. 93) et L_b par l'équation (4.54) (p. 101). Cette expression contient plusieurs vraisemblances dont le nombre d'observations est variable en fonction du modèle $\mathcal{C}$: la vraisemblance des amplitudes L_p de chaque note p, le nombre d'observations étant le nombre H_p de partiels, et la vraisemblance L_b du bruit résiduel, la taille de l'observation étant $\#\mathcal{F}_b$. Par conséquent, la log-densité (4.63) ne peut être comparée d'un modèle $\mathcal{C}$ à l'autre. Pour nous affranchir de cette variabilité, nous proposons le principe de normalisation suivant. Si deux variables aléatoires multivariées X_m et $X_{m'}$ sont de tailles respectives m et m', leurs densités $p(X_m)$ et $p(X_{m'})$ (si elles existent) ne sont pas de même dimension. En revanche, les versions normalisées $p(X_m)^{\frac{1}{m}}$ et $p(X_{m'})^{\frac{1}{m'}}$ le sont : elles sont homogènes à la densité d'une variable aléatoire univariée. Sous forme logarithmique, ceci revient à diviser chaque log-densité par la taille de la variable aléatoire. L'application de ce principe aux vraisemblances de l'équation (4.63) consiste donc à diviser la log-vraisemblance des amplitudes L_p de la note p par le nombre de partiels H_p et la log-vraisemblance du bruit L_b par le nombre d'observations $\#\mathcal{F}_b$.

Les log-vraisemblances normalisées $\frac{L_p}{H_p}$ et $\frac{L_b}{\#\mathcal{F}_b}$ étant homogènes à des lois univariées, nous appliquons alors un critère de sélection d'ordre. Comme nous l'avons vu précédemment, cela consiste à leur ajouter un terme correctif linéaire. Les corrections respectives apportées à $\frac{L_p}{H_p}$ et $\frac{L_b}{\#\mathcal{F}_b}$ sont linéaires de la forme $\mu_{\text{env}} H_p$ et $\mu_b \#\mathcal{F}_b$. Les quantités résultantes de ces opérations sont

$$\tilde{L}_p \triangleq \frac{L_p}{H_p} - \mu_{\text{env}} H_p \tag{4.73}$$

$$\tilde{L}_b \triangleq \frac{L_b}{\#\mathcal{F}_b} - \mu_b \#\mathcal{F}_b \tag{4.74}$$

La fonction $H_p \mapsto \mu_{\text{env}} H_p$ (resp. $\#\mathcal{F}_b \mapsto \mu_b \#\mathcal{F}_b$) peut être vue comme une correction de la pente de $\frac{L_p}{H_p}$ (resp. $\frac{L_b}{\#\mathcal{F}_b}$) en fonction de H_p (resp. $\#\mathcal{F}_b$). Les paramètres μ_{env} et μ_b

sont alors fixés empiriquement à

$$\mu_{\text{env}} \triangleq -8,9.10^{-3} \tag{4.75}$$

$$\mu_b \triangleq -2,2.10^{-4} \tag{4.76}$$

Les normalisations des vraisemblances et la correction d'ordre sont illustrées sur la figure 4.10.

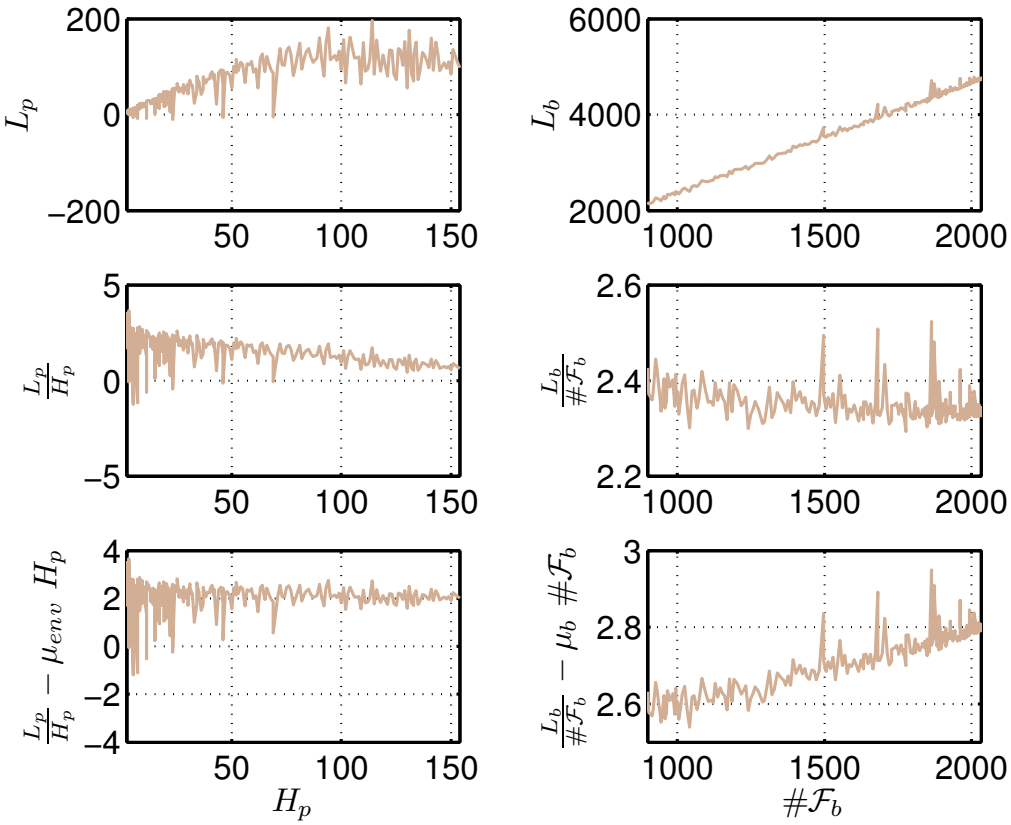


FIGURE 4.10 – Normalisation et corrections des vraisemblances des amplitudes (à gauche) et du bruit (à droite) : fonctions originales en haut ; versions normalisées au milieu ; versions normalisées et corrigées (pénalisées) $\tilde{L}_p$ et $\tilde{L}_b$ en bas. La note analysée est un Lab 4, les ordres correspondant à cette note étant $H_p = 22$ et $\#\mathcal{F}_b = 1873$. Dans un proche voisinage de $H_p = 22$, les courbes de gauche prennent successivement des valeurs élevées (lorsque les partiels ont été sélectionnés) et faibles (lorsqu'une partie des partiels n'a pas été sélectionnée). De même, on distingue un pic à $\#\mathcal{F}_b = 1873$ sur les courbes de droite, correspondant à l'élimination des sinusoïdes du résiduel.

Du fait de la normalisation, les quantités obtenues $\tilde{L}_p$, $\tilde{L}_b$, $\ln p\left(\widehat{\sigma_p^2}\right)$, $p\left(\widehat{\sigma_b^2}\right)$, $\ln p\left(\widehat{A_p}\right)$ et $\ln p\left(\widehat{B}\right)$ sont chacune individuellement homogènes lorsque le modèle $\mathcal{C}$ varie, c'est-à-dire lorsque les H_p et $\#\mathcal{F}_b$ changent. Cependant, elles ne sont plus homogènes entre elles et l'on ne peut les ajouter dans une expression analogue à (4.63). Nous corrigeons l'effet en pondérant ces quantités pour les rendre homogènes. Nous définissons donc la vraisemblance pondérée suivante :

$$\tilde{L}_x(\mathcal{C}) \triangleq w_1 \sum_{p=1}^{P} \tilde{L}_p\left(\widehat{\theta_p}\right)/P + w_2 \tilde{L}_b\left(\widehat{\theta_b}\right) + w_3 \sum_{p=1}^{P} \ln p\left(\widehat{\sigma_p^2}\right)/P + w_4 \ln p\left(\widehat{\sigma_b^2}\right) - \mu_{\text{pol}} P \tag{4.77}$$

où w_1, w_2, w_3, w_4 sont les poids des vraisemblances normalisées et des densités *a priori* sur σ_p^2 et σ_b^2, et μ_{pol} est une correction d'ordre relative à la polyphonie P. Les sommes des log-densités $\tilde{L}_p$ et $\ln p\left(\widehat{\sigma_p^2}\right)$ sont normalisées par P selon le même principe que pour la normalisation de L_p et L_b par H_p et $\#\mathcal{F}_b$. Par ailleurs, les log-densités *a priori* sur les filtres A_p et B ont été supprimés, en leur imposant un poids nul, car elles ne portent pas d'information utile (ou, de manière équivalente, en considérant $p(A_p)$ et $p(B)$ constantes). Dans l'expression (4.77), les coeffcients sont fixés de manière empirique à

$$w_1 \triangleq 8,1.10^{-1} \tag{4.78}$$

$$w_2 \triangleq 1,4.10^{4} \tag{4.79}$$

$$w_3 \triangleq 6,2.10^{2} \tag{4.80}$$

$$w_4 \triangleq 5,8 \tag{4.81}$$

$$\mu_{\text{pol}} \triangleq 25 \tag{4.82}$$

La fonction $\tilde{L}_x(\mathcal{C})$ obtenue peut alors être utilisée comme fonction de détection pour l'estimation des fréquences fondamentales multiples. Le principal mérite de la normalisation est de proposer une solution homogène, en terme de dimensions. Ce problème se pose dès lors que le nombre d'observations varie, comme c'est le cas ici. Quant à l'efficacité – au sens commun du terme – de ces approximations, nous verrons qu'elle est en pratique bonne lors de l'évaluation de l'algorithme (partie 6.3).

Estimation des fréquences fondamentales multiples

L'estimation des fréquences fondamentales multiples consiste finalement à déterminer

$$\widehat{\mathcal{C}} \triangleq \underset{\mathcal{C}\in\tilde{\boldsymbol{\mathcal{C}}}}{\arg\max}\, \tilde{L}_x(\mathcal{C}) \tag{4.83}$$

La figure 4.11 illustre la maximisation (4.83) de la vraisemblance pondérée $\tilde{L}_x(\mathcal{C})$ en montrant les modèles de mélanges conduisant aux vraisemblances les plus élevées dans le cas de l'analyse d'un La♭ 3. Le modèle correspondant à cette note (MIDI 68) apparaît en première position : il est bien estimé comme celui qui maximise la vraisemblance pondérée $\tilde{L}_x(\mathcal{C})$. Suivent ensuite des mélanges d'une ou plusieurs notes, souvent composés de la note originale, de l'octave ou de la sous-octave. Dans la partie supérieure de la figure, nous pouvons remarquer que les diverses contributions composant la vraisemblance pondérée ne varient pas de la même façon d'un modèle de mélange à l'autre, certains étant davantage pénalisés par la vraisemblance du bruit et d'autres par le paramètre de puissance des enveloppes spectrales par exemple.

4.8 Résumé de l'algorithme d'estimation de fréquences fondamentales multiples

L'algorithme 4.2 détaille le déroulement de l'estimation de fréquences fondamentales multiples, dont nous avons déjà vu un diagramme sur la figure 4.4 (p. 86). L'estimation est

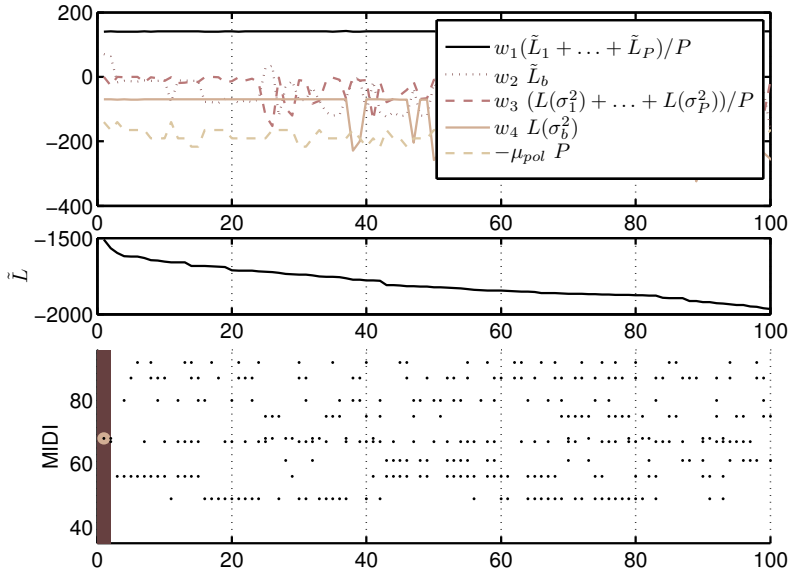


FIGURE 4.11 – Log-vraisemblance pondérée pour un La♭ 3 (note MIDI 68) : en haut, les cinq composantes pondérées (un décalage vertical a été introduit dans un souci de confort visuel) ; au milieu, la log-vraisemblance pondérée, somme des composantes pondérées ; en bas, les notes correspondantes. Sur les trois graphiques, les abscisses représentent l'indice des mélanges testés, triés par ordre décroissant selon leur log-vraisemblance pondérée. Seuls les 100 premiers mélanges sur les 466 testés sont représentés. Le mélange estimé est surligné sur la figure du bas (premier mélange, à gauche, composé de la note 68).

réalisée sur une trame de 93 ms (soit 2048 points échantillonnés à 22050 Hz). Le temps de calcul sur un ordinateur du commerce est de l'ordre de 150 fois le temps réel. L'algorithme est donc plus coûteux que certains tels que celui de Klapuri [2006] mais ce coût reste raisonnable (en particulier si l'on considère le coût de l'approche naïve de l'estimation jointe, sans sélection de candidats, discutée plus haut). La partie la plus coûteuse est l'estimation des paramètres du modèle qu'il faut réaliser pour les 466 mélanges à tester. Elle a été mise en oeuvre en C alors que le reste du programme est mis en œuvre en Matlab.

Entrées: trame x de signal audio
  Estimation du niveau de bruit et blanchiment {pré-traitement}
  Calcul du périodogramme $|X|^2$
  Sélection de N_c candidats
  Estimation de μ et μ_b
  Pour chaque mélange candidat $\mathcal{C}$
    Estimation itérative des amplitudes de partiels et des enveloppes spectrales de notes {algorithme 4.1 (p. 97)}
    Pour chaque note p, calcul de L_p {équation (4.35) (p. 93)}
    Pour chaque note p, calcul de $\ln p\left(\widehat{\sigma_p^2}\right)$ {équation (4.59) (p. 103)}
    Estimation des paramètres du bruit {algorithme 2.2 (p. 68)}
    Calcul de L_b {équation (4.54) (p. 101)}
    Calcul de $\ln p\left(\widehat{\sigma_b^2}\right)$ {équation (4.59) (p. 103)}
    Calcul de la vraisemblance pondérée associée au mélange {équation (4.77) (p. 108)}
  Fin Pour
  Extraction du mélange le plus vraisemblable {équation (4.83) (p. 108)}
Sorties: notes présentes dans la trame

Algorithme 4.2: Estimation de fréquences fondamentales multiples.

4.9 Conclusion

Nous avons décrit un algorithme d'estimation de fréquences fondamentales multiples adapté aux sons de piano en construisant un modèle de son, en détaillant des méthodes pour estimer les paramètres du modèle et en proposant une fonction d'estimation du mélange de notes le plus vraisemblable. L'approche, de nature spectrale, prend en compte l'inharmonicité des sons, considère conjointement les notes en supposant que leurs enveloppes spectrales sont régulières et propose une modélisation du recouvrement entre les spectres. L'ensemble s'intègre dans un cadre statistique, la décision sur l'estimation des fréquences fondamentales multiples présentes étant prise à partir d'une approximation de la vraisemblance des observations étant donné un mélange de notes. Les performances de cet algorithme sont mesurées dans la partie 6.3.1 (p. 145) et des perspectives sur ces travaux sont dressées dans la partie 6.4 (p. 159).

Chapitre 5

Système de transcription

Nous avons vu dans le chapitre 4 comment estimer les hauteurs des notes présentes dans une trame de signal donnée. Un mode de transcription élémentaire consiste à appliquer l'algorithme sur des trames successives pour obtenir les hauteurs présentes dans un morceau entier en fonction du temps. Il s'agit alors d'une transcription de « bas-niveau » dans la mesure où elle ne considère pas les notes comme des entités : le résultat obtenu n'est qu'une fragmentation des notes selon le tramage d'analyse. Dans ce chapitre, nous utiliserons la méthode d'estimation de fréquences fondamentales multiples présentée précédemment et introduirons les mécanismes nécessaires pour élaborer des notes à partir d'une analyse par trames. Après avoir détaillé notre problématique dans la partie 5.1, nous montrerons comment les modèles de Markov cachés offrent un cadre approprié pour suivre les mélanges de hauteurs et en déduire des notes. Le système de transcription complet sera finalement capable d'analyser l'enregistrement d'une pièce pour piano solo et d'en estimer les notes, c'est-à-dire leur hauteur, leur instant d'attaque et leur durée.

Une version antérieure des travaux présentés ici a fait l'objet d'une publication [Emiya *et al.*, 2008].

5.1 Principe et stratégies de transcription

Nos travaux sur la transcription automatique de la musique de piano reposent sur l'idée que la prise en compte des spécificités du piano doivent pouvoir simplifier la tâche de transcription et/ou mener à de meilleures performances par rapport à un système générique. La démarche consiste d'une part à déterminer comment simplifier le problème du fait de la restriction au piano que l'on s'impose, et d'autre part à enrichir le modèle en introduisant les éléments propres à l'instrument.

Dans cet esprit, la stratégie adoptée s'appuie sur le recensement de quelques principes relatifs au piano :

- le caractère fortement polyphonique de la musique de piano ;
- la grande virtuosité des pièces du répertoire classique de piano solo ;
- l'évolution complexe des amplitudes des partiels ;
- la grande variabilité inter- et intra-instrument des enveloppes spectrales ;
- les modulations de fréquence négligeables ;
- la nature percussive de l'attaque des notes ;
- la distribution inharmonique des fréquences des partiels.

Pour donner une idée du niveau relativement élevé de **polyphonie** que l'on rencontre

dans le répertoire pour piano solo, un ensemble de morceaux [Krueger, 2008] a été analysé en extrayant le nombre de notes présentes simultanément. Chaque niveau de polyphonie trouvé est représenté sur la figure 5.1, en proportion de la durée totale des 232 pièces de la base utilisée. La polyphonie varie généralement entre 0 (silence) et 10, avec une moyenne à $4,5$. Grâce à la pédale *forte*, un pianiste peut produire plus de dix notes simultanément, le maximum relevé dans la base de morceaux, lors d'une montée chromatique rapide, étant de 60 notes.

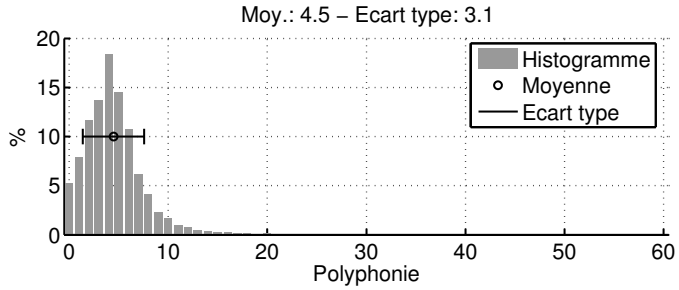


Figure 5.1 – Niveau de polyphonie de pièces du répertoire classique, en proportion de la durée totale.

La figure 5.2 permet de donner un ordre de grandeur de la complexité engendrée par la **virtuosité** dans le cas du répertoire de piano solo, par rapport aux répertoires d'autres formations classiques. En analysant des fichiers MIDI disponibles sur Internet, nous avons calculé le débit moyen de notes de chaque fichier (rapport du nombre de notes par la durée totale). Les pièces analysées appartiennent aux répertoires classique et romantique et ont été choisies au hasard. Chaque fichier correspond à un mouvement d'une œuvre, tous les mouvements étant analysés. La base de fichiers comporte ainsi une alternance de mouvements lents et rapides, et d'œuvres de virtuosités variables. Nous distinguons trois sortes de formations : musique de chambre sans piano – du trio au quatuor, à cordes en général –, piano solo, et musique symphonique. Pour chaque formation, les statistiques sont établies tous compositeurs confondus ainsi que dans quelques cas particuliers – celui de la musique de Beethoven, choisie car elle présente dans un même style toutes les formations comparées, et celui de la musique de Chopin, musique pianistique virtuose de référence. De manière générale, le débit de notes de la musique pour piano solo est plus élevé que celui de la musique de chambre sans piano, et moins élevé que celui de la musique symphonique. Plus précisément, le débit médian de la musique pour piano solo coïncide avec le troisième quartile du débit de musique de chambre, alors que la première moitié des morceaux de musique symphonique a un débit correspondant à l'ensemble de la musique pour piano solo. Ces tendances fortes nous permettent d'affirmer qu'en raison de la vituosité des œuvres, la transcription de la musique de piano est d'une difficulté comparable à la transcription de musique d'ensembles conséquents pouvant aller jusqu'à l'orchestre. Nos affirmations s'en tiendront à ce constat, la comparaison effectuée étant très limitée : elle fait abstraction de phénomènes comme l'unisson de plusieurs instruments ou la diversité des timbres dans la musique d'ensemble (qui rendent la musique de piano plus facile à transcrire), la tessiture ou la concentration spatiale du piano (qui la rend plus difficile).

Si polyphonie élevée et virtuosité constituent deux défis pour la transcription automatique de la musique de piano, nous allons maintenant voir comment d'autres aspects spécifiques évoqués plus haut se rangent plutôt dans la catégorie des éléments qui peuvent

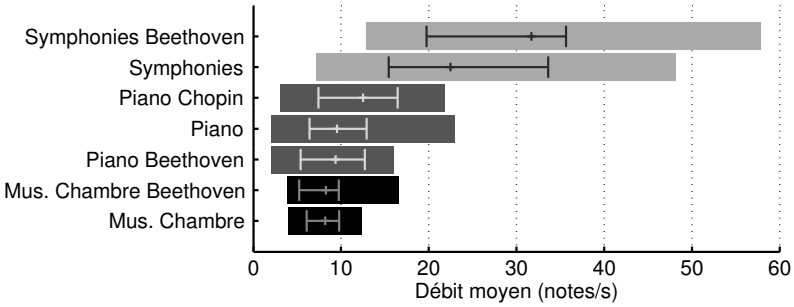


FIGURE 5.2 – Débit moyen de notes en fonction de plusieurs formations classiques (musique de chambre sans piano, piano solo, musique symphonique), tous compositeurs confondus et pour quelques compositeurs emblématiques. Les limites trouvées sont représentées par les zones pleines, les quartiles par les traits. Les statistiques ont été calculées par rapport au débit moyen des mouvements de chaque morceau.

faciliter la transcription. Dans un contexte musical le plus large possible, le début d'une note (voire d'un événement musical si l'on considère les musiques qui n'utilisent pas seulement des notes comme matériau, mais par exemple des bruits) survient soit de manière discontinue dans le cas d'une note détachée, soit de manière davantage continue, par exemple lors de passages liés ou de glissandi joués par des vents, des cordes frottées ou chantés [d'Alessandro *et al.*, 1998b]. Les notes de piano contrastent avec cette diversité puisqu'elles ne peuvent commencer que par une attaque, plus ou moins franche. Nous pouvons ainsi simplifier la tâche de caractérisation du début des notes en nous concentrant sur ce genre d'attaques et en excluant toute forme de continuité entre deux notes successives. La nature percussive des sons et l'absence de modulation fréquentielle significative nous permettent même d'aller plus loin : le contenu sonore dans un segment compris entre deux attaques consécutives se compose d'une ou plusieurs notes, toutes présentes au début du segment, de fréquences fondamentales constantes le long du segment et pouvant éventuellement se terminer, de manière indépendante les unes des autres, au cours de ce laps de temps. Cette évolution relativement simple exclut en particulier l'apparition d'une note à l'intérieur du segment, ainsi que tout phénomène de type vibrato ou glissando.

5.2 Description du système

5.2.1 Segmentation de l'extrait analysé

Dans notre système de transcription, l'estimation des notes repose sur une analyse synchrone avec les attaques. La première étape de la transcription consiste donc à estimer ces attaques. Pour ce faire, nous utilisons la méthode proposée par Alonso *et al.* [2005] s'appuyant sur le principe du flux d'énergie spectrale (SEF, *Spectral Energy Flux*). Le SEF consiste à calculer les variations temporelles d'une représentation énergétique du signal, telle que le spectrogramme représenté sur la figure 5.3. Dans ce genre de représentations, le bruit s'étale sur une partie du spectre alors que les composantes sinusoïdales sont très localisées : ce pouvoir de séparation rend l'apparition d'une note plus facile à détecter que dans le domaine temporel où l'on ne peut détecter qu'une variation globale d'énergie. La méthode de Alonso *et al.* [2005] propose en outre d'appliquer des traitements spécifiques à la transformée de Fourier à court terme (compression, pondération psycho-acoustique) avant

de calculer le SEF. Les attaques sont ensuite détectées comme les maxima de la fonction de détection, extraits en utilisant un seuil adaptatif, comme illustré sur la figure 5.3.

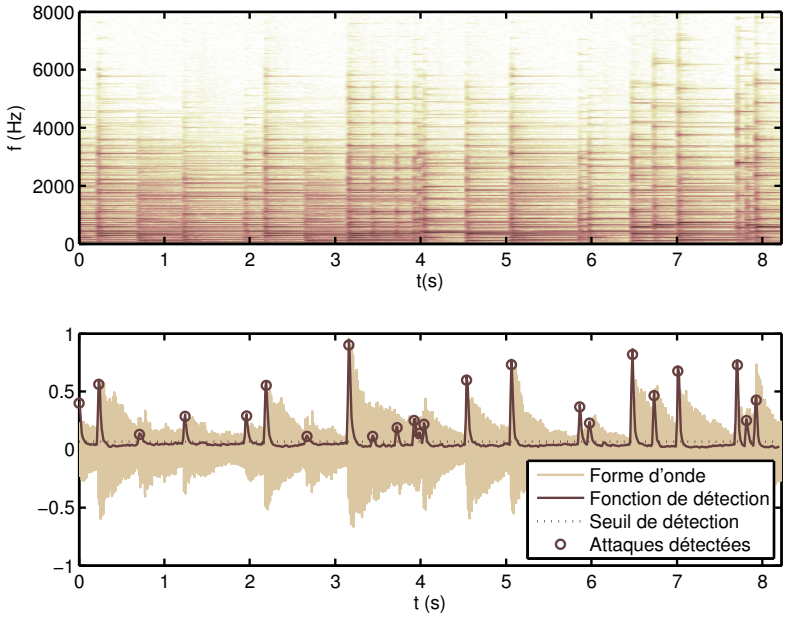


Figure 5.3 – Exemple de détection des attaques sur un extrait d'*España-Tango*, Op. 165 d'I. Albéniz : en haut, spectrogramme de l'extrait, et, en bas, forme d'onde avec fonction de détection.

Dans toute la suite, nous appellerons *segment* l'intervalle de taille variable compris entre deux attaques consécutives, et qui sera décomposé en trames de durée fixe.

5.2.2 Modélisation du suivi des mélanges de notes par HMM

Dans la partie 4.2.2 (p. 87), nous avons vu que le principe de l'estimation de fréquences fondamentales multiples à partir d'une trame de signal x donnée s'exprime sous forme statistique comme la maximisation de la probabilité *a posteriori* $p\left(\mathcal{C}|x\right)$ (équation (4.17) p. 87) par rapport à tous les mélanges de notes $\mathcal{C}$ possibles ou, de manière équivalente et sous certaines conditions, comme la maximisation de la vraisemblance $p\left(x|\mathcal{C}\right)$ (équation (4.18) p. 87). Nous avons également vu comment réaliser la maximisation sur un ensemble restreint $\widetilde{\boldsymbol{\mathcal{C}}}$ de mélanges (équation (4.19) p. 87) plutôt que sur l'ensemble de tous les mélanges possibles $\boldsymbol{\mathcal{C}}$, dont la taille est trop grande.

Nous étendons maintenant ce principe à plusieurs trames consécutives afin d'estimer et de suivre un mélange de notes, au fil des trames, sur un segment arbitrairement long entre deux attaques. Nous décrivons ici comment ce suivi s'intègre naturellement dans le cadre des modèles de Markov cachés (HMM) [Rabiner, 1989].

Lorsqu'une ou plusieurs notes sont jouées, elles s'étalent sur un certain nombre de trames consécutives au cours desquelles chaque note peut à tout moment se terminer. Ainsi, sur un horizon relativement court comme celui constitué par la durée d'une note, le contenu polyphonique d'une trame donnée dépend fortement du passé à court terme. Nous allons voir comment ce processus peut être décrit par un HMM, comme illustré sur la figure 5.4. Nous considérons pour cela que l'étape de détection d'attaques permet de diviser le signal en ce que nous appellerons **segments**, et qui sont les portions, de longueur variable,

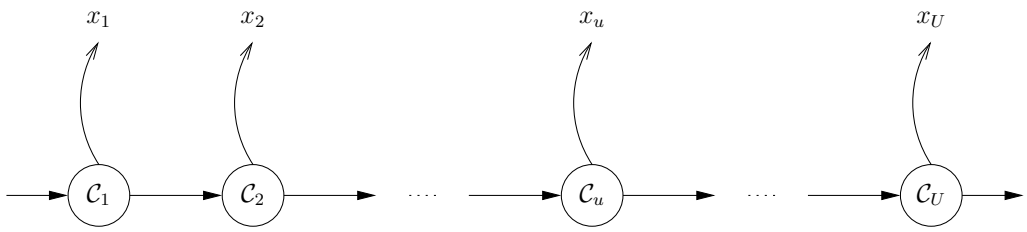


FIGURE 5.4 – Processus de génération des observations par HMM : entre deux attaques, la suite des mélanges de notes forme une chaîne de Markov, dont les états (cachés) permettent de générer les trames de signal observées.

délimitées par deux attaques consécutives. Un segment donné se compose de U **trames** de longueur constante numérotées de 1 à U, U dépendant de la longueur du segment. Le signal x_u d'une trame $u \in [\![1; U]\!]$ est un vecteur aléatoire qui dépend du mélange de notes sous-jacent, noté $\mathcal{C}_u$. $\mathcal{C}_u$ est lui-même considéré comme une variable aléatoire : étant données nos considérations sur la dépendance entre les contenus polyphoniques des trames successives, nous supposerons qu'un processus de Markov du premier ordre sous-tend la génération de la suite $\mathcal{C}_1 \dots \mathcal{C}_U$, mélanges de notes successifs dans le segment. Formellement, pour $u \geq 2$, le mélange $\mathcal{C}_u$ ne dépend que de $\mathcal{C}_{u-1}$ et ne dépend pas de u, de sorte que

$$\begin{aligned} p\left(\mathcal{C}_u | u, \mathcal{C}_1 \dots \mathcal{C}_{u-1}\right) &= p\left(\mathcal{C}_u | u, \mathcal{C}_{u-1}\right) \\ &= p\left(\mathcal{C}_u | \mathcal{C}_{u-1}\right) \end{aligned} \tag{5.1}$$

La transcription du segment consiste donc à trouver la séquence optimale d'états cachés $\hat{\mathcal{C}}_1 \hat{\mathcal{C}}_2 \dots \hat{\mathcal{C}}_U$ étant données les observations successives $x_1 x_2 \dots x_U$. Le critère d'optimalité n'est plus le maximum de vraisemblance dans chaque trame comme dans le cas de l'estimation de fréquences fondamentales multiples (équation (4.83) p. 108), mais prend en compte l'enchaînement des trames grâce au processus markovien énoncé ci-dessus. En termes statistiques, $\hat{\mathcal{C}}_1 \hat{\mathcal{C}}_2 \dots \hat{\mathcal{C}}_U$ est défini par

$$\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U = \underset{\mathcal{C}_1 \dots \mathcal{C}_U}{\arg\max}\, p\left(\mathcal{C}_1 \dots \mathcal{C}_U | x_1 \dots x_U\right) \tag{5.2}$$

qui devient, en appliquant la règle de Bayes :

$$\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U = \underset{\mathcal{C}_1 \dots \mathcal{C}_U}{\arg\max}\, \frac{p\left(\mathcal{C}_1 \dots \mathcal{C}_U\right) p\left(x_1 \dots x_U | \mathcal{C}_1 \dots \mathcal{C}_U\right)}{p\left(x_1 \dots x_U\right)} \tag{5.3}$$

Le dénominateur pouvant être supprimé sans changer le résultat de la fonction arg max, il vient :

$$\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U = \underset{\mathcal{C}_1 \dots \mathcal{C}_U}{\arg\max}\, p\left(\mathcal{C}_1 \dots \mathcal{C}_U\right) p\left(x_1 \dots x_U | \mathcal{C}_1 \dots \mathcal{C}_U\right) \tag{5.4}$$

La trame observée x_u ne dépendant que du mélange sous-jacent $\mathcal{C}_u$, on a :

$$\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U = \underset{\mathcal{C}_1 \dots \mathcal{C}_U}{\arg\max}\, p\left(\mathcal{C}_1 \dots \mathcal{C}_U\right) \prod_{u=1}^{U} p\left(x_u | \mathcal{C}_u\right) \tag{5.5}$$

En utilisant l'hypothèse markovienne (équation (5.1)), l'expression se simplifie alors en :

$$\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U = \underset{\mathcal{C}_1 \dots \mathcal{C}_U}{\arg\max}\, p\left(\mathcal{C}_1\right) \prod_{u=2}^{U} p\left(\mathcal{C}_u | \mathcal{C}_{u-1}\right) \prod_{u=1}^{U} p\left(x_u | \mathcal{C}_u\right) \tag{5.6}$$

Cette équation représente précisément l'optimisation qui est réalisée dans le cadre de l'utilisation de HMM. À condition de définir et de pouvoir calculer chacun des termes de l'expression à maximiser, nous pouvons alors trouver la suite $\hat{\mathcal{C}}_1 \dots \hat{\mathcal{C}}_U$ grâce à l'algorithme de Viterbi [1967]. La vraisemblance d'une observation étant donné un état, $p\left(x_u | \mathcal{C}_u\right)$, correspond à la vraisemblance pondérée définie, sous forme logarithmique, pour l'estimation de fréquences fondamentales multiples par l'équation (4.77) (p. 108). Les deux autres termes restent maintenant à définir : il s'agit des probabilités initiales notée $p\left(\mathcal{C}_1\right)$ (probabilité de se trouver dans l'état $\mathcal{C}_1$ dans la première trame) et des probabilités de transition notées $p\left(\mathcal{C}_u | \mathcal{C}_{u-1}\right)$ (probabilité de passer de l'état $\mathcal{C}_{u-1}$ à l'état $\mathcal{C}_u$).

5.2.3 Initialisation et transitions de la chaîne de Markov

Nous avons vu dans la partie 4.3 (p. 90) le moyen de réduire l'espace de recherche des mélanges de notes en sélectionnant des notes candidates dans chaque trame analysée. Le principe est ici mis en œuvre en n'extrayant les notes candidates qu'au début du segment et non dans chaque trame, puisque toutes les notes présentes dans le segment le sont forcément au début de celui-ci. Nous procédons donc comme auparavant, en extrayant les N_c premiers pics du produit spectral inharmonique normalisé défini par l'équation (4.29) (p. 90). Pour plus de robustesse vis-à-vis de l'étalement temporel de l'amorce des partiels et du bruit percussif des attaques, nous moyennons ce produit spectral sur les trois premières trames du segment pour extraire ses maxima.

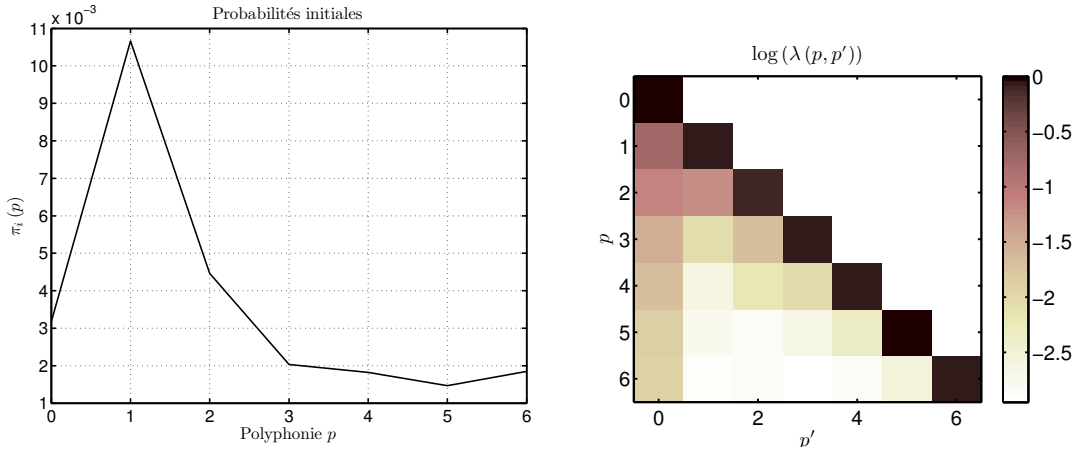
En composant des mélanges à partir des N_c notes candidates et en limitant la polyphonie à un nombre maximal $P_{\max}$, nous constituons l'ensemble $\widetilde{\boldsymbol{\mathcal{C}}}$ des mélanges possibles dans chaque trame du segment considéré. Ces mélanges, au nombre de $\#\widetilde{\boldsymbol{\mathcal{C}}} = \sum_{p=0}^{P_{\max}} \binom{N_c}{p} = 466$ dans notre cas, constituent les états cachés possibles. Nous définissons la probabilité initiale d'un mélange comme ne dépendant que du nombre de notes du mélange :

$$p\left(\mathcal{C}\right) \triangleq \begin{cases} \pi_i\left(\#\mathcal{C}\right) & \text{si } \mathcal{C} \in \widetilde{\boldsymbol{\mathcal{C}}} \\ 0 & \text{sinon} \end{cases} \tag{5.7}$$

où les valeurs de la fonction $p \mapsto \pi_i\left(p\right)$ seront fixées par une phase d'apprentissage, comme l'explique la prochaine partie.

La probabilité de transition d'un mélange $\mathcal{C}$ à un mélange $\mathcal{C}'$ est ensuite définie comme suit :

- pour $2 \leq u \leq U$, l'apparition d'une note dans la trame u est interdite du fait de la définition même d'un segment comme étant délimité par deux attaques consécutives ; nous en déduisons que $p\left(\mathcal{C}' | \mathcal{C}\right) \triangleq 0$ si $\mathcal{C}'$ n'est pas un sous-ensemble de $\mathcal{C}$;
- en conséquence, la seule transition possible depuis l'état « silence » ($\mathcal{C} = \emptyset$) est vers lui-même : $p\left(\emptyset | \emptyset\right) \triangleq 1$;
- dans les autres cas ($\mathcal{C}' \subset \mathcal{C}$), les transitions depuis $\mathcal{C}$ sont possibles vers $\mathcal{C}$ lui-même ($\mathcal{C}' = \mathcal{C}$) ou vers un sous-ensemble strict $\mathcal{C}'$ (il y a alors extinction des notes qui ont disparu de $\mathcal{C}$). Nous choisissons de ne faire dépendre la probabilité de transition que du nombre de notes dans les accords $\mathcal{C}$ et $\mathcal{C}'$. Cette probabilité, notée $\lambda\left(\#\mathcal{C}, \#\mathcal{C}'\right)$, est fixée par la phase d'apprentissage détaillée dans la prochaine partie.



(a) Apprentissage des probabilités initales (b) Apprentissage des probabilités de transition

FIGURE 5.5 – Résultat de l'apprentissage des paramètres des HMM pour $N_c = 9$ notes candidates et un polyphonie maximale $P_{max} = 6$.

5.2.4 Apprentissage des paramètres du HMM

L'apprentissage des paramètres du système se réduit à celui des probabilités initiales et des probabilités de transition. Le processus d'apprentissage est réalisé sur des fichiers MIDI issus de la base de Krueger [2008]. En utilisant une telle base qui contient directement l'information que l'on souhaite apprendre, l'apprentissage est relativement simple et ne nécessite pas de recourir à l'algorithme EM [Dempster *et al.*, 1977].

Pour apprendre la probabilité initiale $\pi_i(p)$ associée à une polyphonie $p \in [\![0; P_{max}]\!]$, il suffit de dénombrer le nombre de cas d'attaques où la polyphonie est égale à p et de rapporter le résultat au nombre total d'attaques où la polyphonie est comprise entre 0 et P_{max}, multiplié par le nombre total d'états de polyphonie p. En toute logique, il n'existe pas de cas d'attaque pour lequel la polyphonie est nulle, cet apprentissage mène donc à $\pi_i(0) = 0$, mais en pratique, la phase de détection d'attaques peut commettre quelques fausses alarmes. Pour essayer de corriger ces erreurs, nous permettons à notre algorithme de détecter un silence dès la première trame d'un segment en attribuant une valeur non nulle à $\pi_i(0)$. Cette valeur est arbitrairement fixée à la moyenne des probabilités initiales apprises pour les autres polyphonies. Le résultat est représenté sur la figure 5.5(a). Il faut noter que $\sum_{p=0}^{P\text{max}} \pi_i(p) \neq 1$ car c'est la probabilité de l'ensemble des états initiaux, et non de l'ensemble des polyphonies possibles, qui doit valoir 1. On a alors

$$\sum_{p=0}^{P\text{max}} \binom{N_c}{p} \pi_i(p) = 1 \tag{5.8}$$

En conséquence, contrairement à ce que la figure 5.5(a) peut laisser croire à première vue, la polyphonie 1 n'est pas la plus probable car il n'existe que N_c mélanges d'une seule note. L'allure de la distribution statistique initiale des polyphonies, non représentée ici, se rapproche donc de celle de la figure 5.1 (p. 114).

En suivant le même principe de dénombrement, les probabilités de transition d'un mélange de polyphonie $p \in [\![0; P_{max}]\!]$ à un mélange de polyphonie $p' \in [\![0; p]\!]$ sont calculées en découpant en trames l'axe temporel des fichiers MIDI et en calculant la proportion

de passage d'une polyphonie p à une polyphonie p' à partir de toutes les trames dont la polyphonie est p. Les résultats de cet apprentissage sont représentés sur la figure 5.5(b). Pour chaque mélange de polyphonie p, la probabilité de l'ensemble des transitions possibles doit valoir 1. Les transitions étant possibles vers tout sous-ensemble de cet accord, nous obtenons

$$\forall p \in [\![0; P_{\max}]\!], \sum_{p'=0}^{p} \binom{p}{p'} \lambda(p, p') = 1 \tag{5.9}$$

Contrairement aux probabilités initiales, les probabilités de transitions dépendent non seulement du nombre de notes candidates sélectionnées et de la polyphonie maximale, mais également de la durée des trames d'analyse. Ce paramètre, fixé à 93 ms dans notre cas, a une influence particulière sur la probabilité de rester dans le même état.

5.2.5 Estimation du mélange de notes le plus vraisemblable

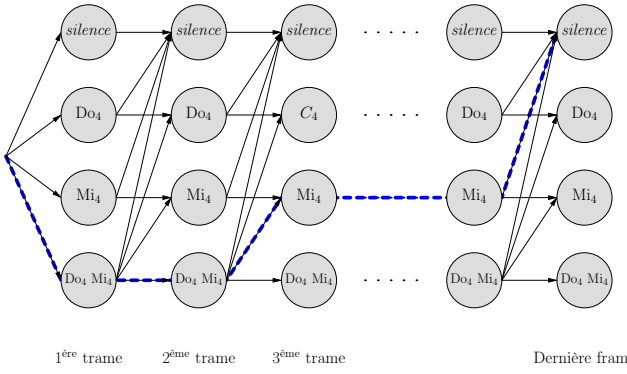
En pratique, une fois la vraisemblance des observations calculées pour chaque état possible et pour chaque trame, la mise en œuvre de l'algorithme de Viterbi [1967] donne lieu à une estimation très rapide. En particulier, elle s'affranchit de tout calcul relatif aux états ainsi qu'aux transitions dont la probabilité est nulle du fait de la sélection de candidats et des règles de transition. L'ordre de grandeur de la complexité de l'estimation est alors en $\mathcal{O}\left(\left(\#\widetilde{\boldsymbol{\mathcal{C}}}\right)^2 \#U\right)$. Le temps d'exécution propre de l'algorithme de Viterbi est en pratique petit devant le temps de calcul de la vraisemblance du signal dans chaque état.

La figure 5.6 représente un exemple d'estimation du mélange de notes et de son suivi sur un segment, dans un cas simplifié. Cette estimation est réalisée en pratique pour chaque segment, en utilisant un HMM avec $\#\widetilde{\boldsymbol{\mathcal{C}}}$ états, et donne lieu à la transcription du segment : une même note estimée dans des trames successives sera alors transcrite comme une unique note présente dont la durée correspond à celle des trames.

5.2.6 Détection des notes répétées et génération de la transcription

Les transcriptions de chaque segment une fois réalisées, il reste une étape avant de générer la transcription du morceau complet. Lorsqu'une note a été détectée dans la totalité d'un segment et au début du suivant, il convient de déterminer si la même note s'étale sur les deux segments (auquel cas l'attaque au début du second segment correspond au début d'une autre note), ou bien si la note dans le second segment est une répétition de celle du premier. Pour distinguer les deux cas, nous estimons une variation d'intensité en utilisant la fonction définie par l'équation (4.29) (p.90) pour la sélection de candidats, et calculée ici sur une version non blanchie du signal. Nous considérons alors qu'une note est répétée si cette variation dépasse un seuil arbitrairement fixé à 3 dB.

La transcription du morceau complet peut être alors générée. Elle se présente sous la forme d'un fichier MIDI contenant les notes estimées avec leur hauteur et leurs instants d'attaque et de fin. L'information de nuance (*velocity* en MIDI) n'ayant pas été estimée, elle est fixée à une valeur moyenne (64 sur une échelle variant de 1 à 127). L'estimation de cette intensité n'a pas fait l'objet d'une étude car il nous a semblé que le caractère non-linéaire de la perception de l'intensité et l'absence d'une normalisation précise en MIDI



(a) Estimation de la séquence d'états : la matrice de transition étant creuse, les transitions depuis un état ne sont pas possibles vers tous les autres états mais uniquement vers les états relatifs à des sous-ensembles du mélange de départ.

$\mathcal{C}$ \ $\mathcal{C}'$	$\emptyset$	Do_4	Mi_4	Do_4 Mi_4
$\emptyset$	1	0	0	0
Do_4	.17	.83	0	0
Mi_4	.17	0	.83	0
Do_4 Mi_4	.07	.06	.06	.80

(b) Matrice de transition : probabilité de transition d'un accord $\mathcal{C}$ à un accord $\mathcal{C}'$.

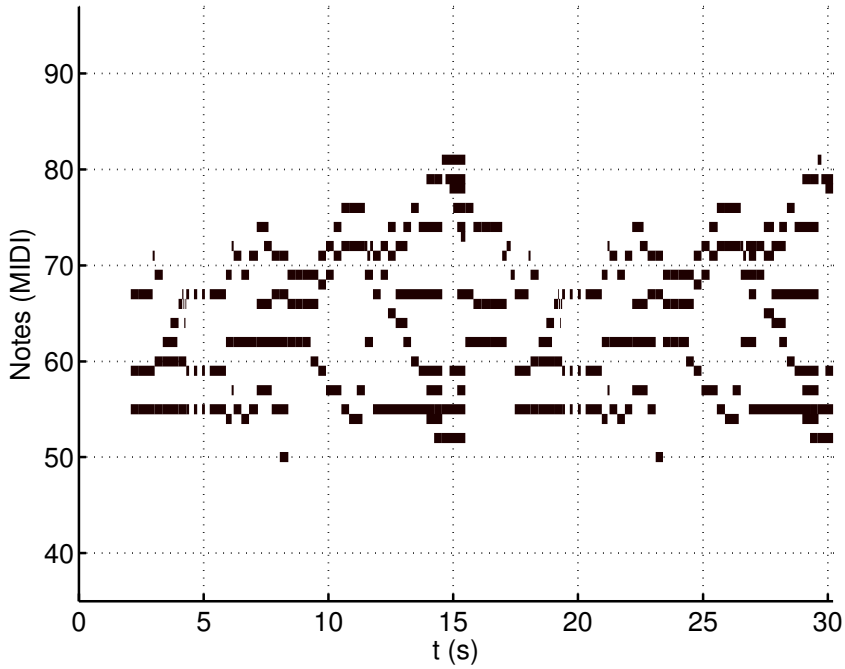
FIGURE 5.6 – Exemple de transcription d'un segment par HMM : dans un souci de lisibilité, seules deux notes (Do 4 et Mi 4) ont été sélectionnées. La ligne en gras et en pointillés représente le chemin estimé : le mélange $\{\text{Do } 4, \text{Mi } 4\}$ est détecté, le Do 4 se termine à la trame 3 alors que le Mi 4 dure jusqu'à l'avant-dernière trame du segment.

rendait cette étude trop complexe dans le cadre de cette thèse [1].

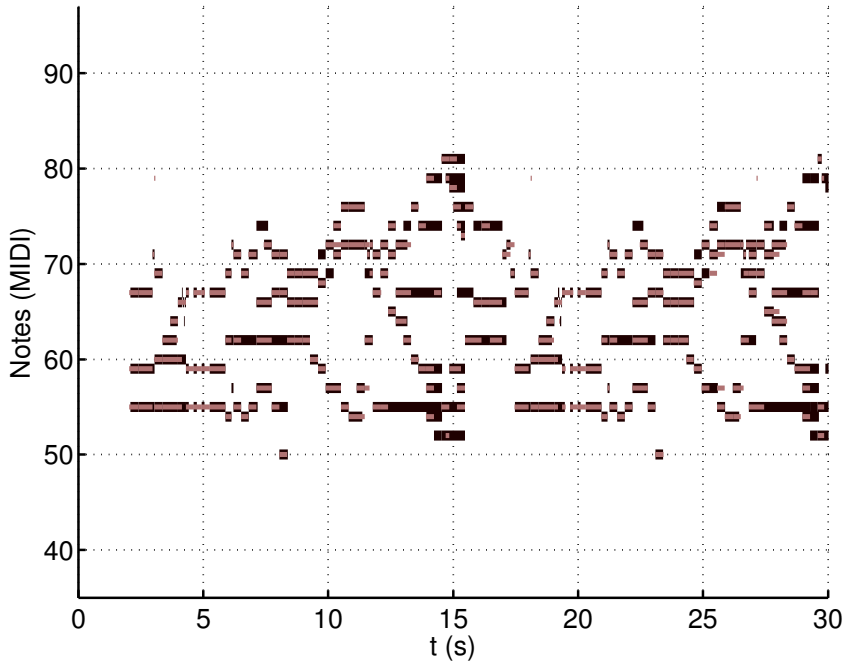
Des exemples de transcriptions sont représentés sur les figures 5.7 et 5.8. Ils ont été obtenus dans les conditions suivantes. Le morceau à analyser se présente sous la forme d'un enregistrement monaural échantillonné à 22 kHz. Une fois le signal segmenté en fonction des attaques détectées, l'analyse se déroule sur des trames de 93 ms se recouvrant de moitié. Le système détecte les notes comprises entre le Do 1 (note MIDI 36) et le Si 5 (note MIDI 95). $N_c = 9$ notes sont sélectionnées comme candidates pour chaque segment et la polyphonie maximale est fixée à $P_{\max} = 6$. La précision temporelle pour des notes est de $11,6$ ms (pas utilisé pour la détection des attaques). La détection de la fin des notes, bien moins déterminante pour la qualité de la transcription, dépend de l'estimation par HMM réalisée par pas de 46 ms. Le programme a été mis en œuvre en Matlab et en C, et son temps d'exécution est inférieur à 200 fois le temps réel sur un PC récent.

La transcription représentée sur la figure 5.7(b) est issue d'un morceau assez lent, avec des notes tenues. Le morceau transcrit sur la figure 5.8(b) est plus technique dans la mesure où il est plus rapide et qu'il présente des accords plus fournis en notes. Sur l'ensemble de ces exemples, qui sont de « bonnes » transcriptions, les hauteurs et durées des notes sont bien estimées, avec quelques erreurs. Nous verrons dans le chapitre 6 une évaluation plus détaillée du comportement de notre algorithme.

1. Si la perception de l'intensité a été largement étudiée dans le champs de la perception des sons en général, peu de travaux s'y sont intéressés dans le cas de la transcription automatique. Citons tout de même Marolt [2004] qui évalue l'intensité d'une note en fonction de l'énergie de son premier partiel et [Klapuri et Davy, 2006, p. 8 et 172] qui mentionnent l'utilisation d'une échelle logarithmique appliquée au niveau RMS (Root Mean Square) estimé.

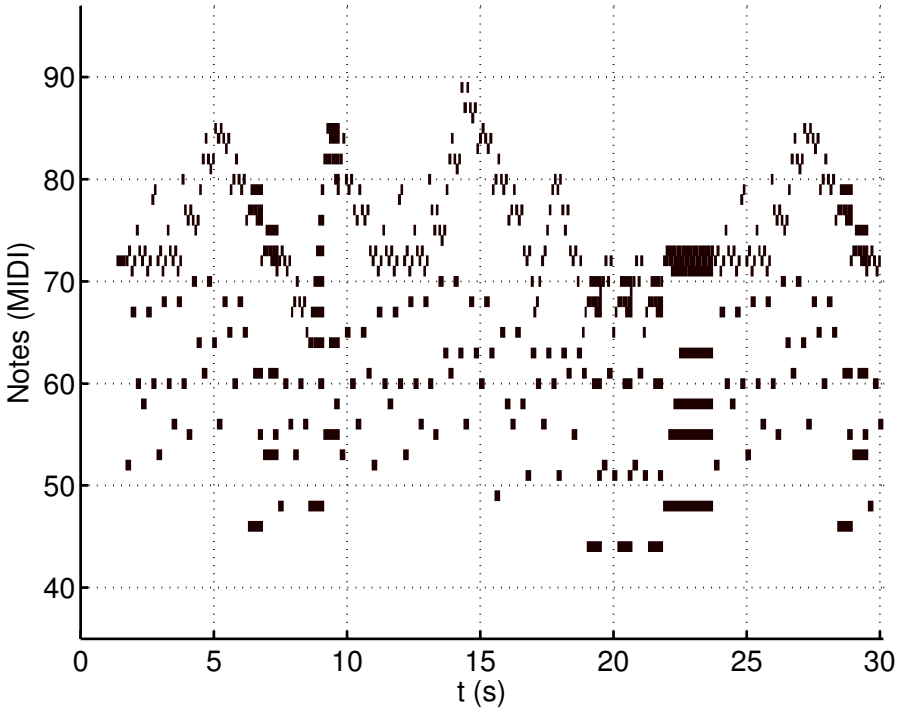


(a) Original.

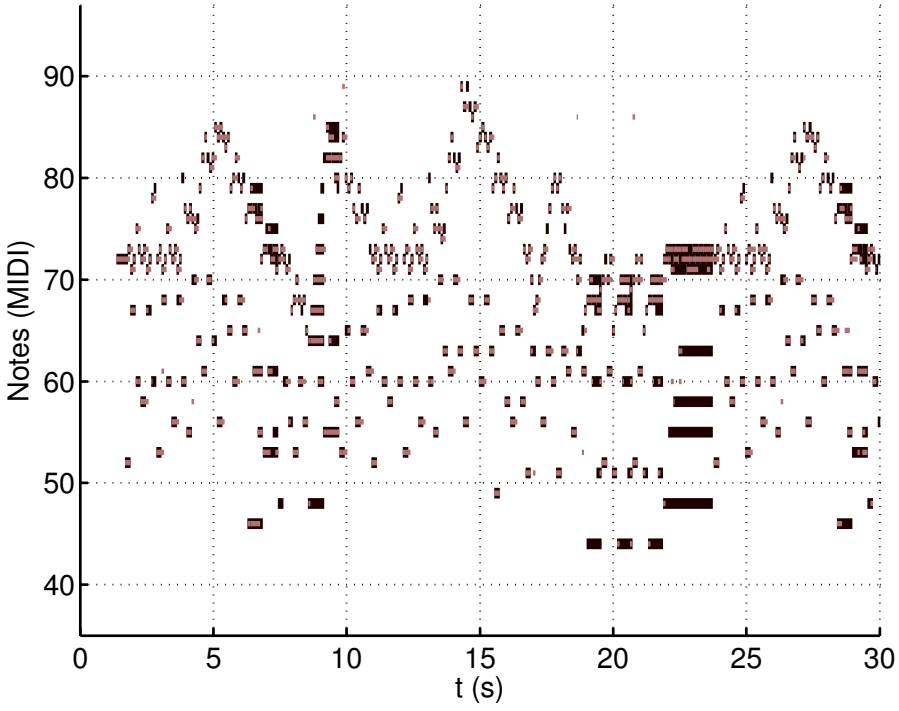


(b) Transcription (et original, en noir).

Figure 5.7 – Exemple de transcriptions : début de la *Sonate n° 54* pour piano en Sol Majeur Hob. XVI :40 de J. Haydn. Sous la forme d'un *piano roll* (hauteurs des notes en fonction du temps), les figures représentent la référence (traits noirs) et la transcription (traits clairs et plus fins). Dans la version électronique de ce document, il est possible de cliquer sur ces figures pour voir une vidéo et écouter l'extrait sonore associé aux morceaux originaux et transcrits.



(a) Original.



(b) Transcription (et original, en noir).

FIGURE 5.8 – Exemple de transcriptions : début de l'*Étude n° 2 (Presto)* Op. 25 de F. Chopin. Dans la version électronique de ce document, il est possible de cliquer sur ces figures pour voir une vidéo et écouter l'extrait sonore associé aux morceaux originaux et transcrits.

5.3 Conclusion

Dans ce chapitre, nous avons décrit la conception d'un système de transcription opérationnel pour l'extraction des notes contenues dans un enregistrement de piano. Le système est capable d'estimer conjointement le niveau de polyphonie, dans la limite de 6 notes simultanées, et d'identifier les notes présentes. Il utilise l'algorithme d'estimation de fréquences fondamentales multiples présenté précédemment et assure le suivi des mélanges de notes par modèles de Markov cachés. Les performances et résultats du système seront analysés au chapitre 6.

Chapitre 6

Évaluation

Nous nous intéressons maintenant aux résultats pratiques obtenus avec notre système. Si l'évaluation d'un système de transcription constitue une démarche nécessaire, la tâche n'en est pas moins complexe, comme nous le montrerons ici. L'évaluation met en jeu deux composantes : les **critères d'évaluation** et la **base d'évaluation**. Toutes deux doivent être fixées pour pouvoir établir et comparer les performances de plusieurs systèmes.

En pratique, par manque de consensus sur les méthodes d'évaluation et parce que les bases de données appropriées sont rares, la plupart des auteurs préfèrent utiliser une base et des critères qui leur sont propres, avec lesquels ils réalisent une évaluation comparative de plusieurs systèmes. En espérant contribuer à faciliter et enrichir cette étape, nous aborderons ici l'évaluation en développant la question des critères d'évaluation, puis proposerons une base de données adaptée à l'évaluation des tâches de transcription automatique et d'estimation de fréquences fondamentales multiples, dans le cas du piano.

Nous terminerons ce chapitre par une évaluation détaillée et comparative de nos algorithmes d'estimation de fréquences fondamentales multiples et de transcription, en utilisant les outils et sons élaborés.

Ces travaux ont partiellement fait l'objet d'une publication [Daniel *et al.*, 2008]. Par ailleurs, une première comparaison détaillée de résultats sur l'estimation de fréquences fondamentales multiples de sons de piano a été présentée par David *et al.* [2007].

6.1 Méthodes d'évaluation

6.1.1 Introduction

La problématique de l'évaluation des systèmes de transcription soulève deux questions :
- quelles sont les méthodes utilisées habituellement pour évaluer les systèmes de transcription ?
- dans quelle mesure l'évaluation par ces méthodes est-elle valide ?

Nous allons donc dans un premier temps nous intéresser aux méthodes d'évaluation habituellement utilisées. Nous verrons qu'elles se concentrent essentiellement sur un dénombrement des notes détectées ou manquantes et discuterons ensuite de la validité de ces méthodes et de leurs limites.

Évaluation qualitative

Certains auteurs se contentent d'illustrer la présentation de leur système à l'aide de quelques exemples de transcription [Moorer, 1975; Martin, 1996; Rossi, 1998; Walmsley

et al., 1999; Smaragdis et Brown, 2003; Cemgil *et al.*, 2006; Davy *et al.*, 2006]. Ils montrent ainsi le type de résultats auquel on peut s'attendre. L'avantage de ce type d'évaluations réside dans la possibilité de mettre en avant quelques erreurs typiques, comme les erreurs d'octave ou les notes répétées, et de les relier à la technique utilisée. En revanche, elles ne donnent pas de résultats sur un nombre significatif de transcriptions, et ne proposent pas d'évaluation quantitative ou comparative.

Critères quantitatifs

Les critères quantitatifs constituent le type d'évaluation le plus largement répandu. Ils consistent à compter le nombre de détections correctes et d'erreurs, en fonction desquels plusieurs taux sont calculés. Ce système d'évaluation n'est toutefois pas unifié. Certains effectuent le décompte dans chaque trame analysée [Plumbley *et al.*, 2006; Poliner et Ellis, 2007] alors que d'autres s'appuient sur les notes [Dixon, 2000; Marolt, 2004; Bello *et al.*, 2006; Ryynänen et Klapuri, 2005; Bertin *et al.*, 2007; Vincent *et al.*, 2008]. Dans ce dernier cas, la définition d'une note correctement estimée dépend d'un seuil de tolérance sur la fréquence fondamentale (le demi-ton en général, cf. correspondance entre fréquence fondamentale et notes dans l'annexe C (p. 177)), sur l'instant d'attaque (50 ms pour Bello *et al.* [2006]; Vincent *et al.* [2008], 70 ms pour Dixon [2000], 128 ms pour Bertin *et al.* [2007], 150 ms pour Ryynänen et Klapuri [2005]) et éventuellement sur l'instant d'extinction de la note. Enfin, plusieurs systèmes concurrents sont proposés, avec des critères légèrement différents.

Un premier système d'évaluation quantitative est utilisé sur la base d'un décompte par note [Bello *et al.*, 2006; Ryynänen et Klapuri, 2005; Bertin *et al.*, 2007; Vincent *et al.*, 2008], par trame Plumbley *et al.* [2006]; Poliner et Ellis [2007], ou des deux [International Music Information Retrieval Systems Evaluation Laboratory, 2007]. On détermine d'abord l'ensemble TP des notes correctement estimées (*true positive*), l'ensemble FP des notes ajoutées (*false positive*, ou fausses alarmes), et l'ensemble FN des notes oubliées (*false negative*). En fonction des cardinaux de ces ensembles, on définit alors deux critères complémentaires, le rappel (*recall*) r et la précision (*precision*) p [Van Rijsbergen, 1979] :

$$r \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \tag{6.1}$$

$$p \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}} \tag{6.2}$$

Le rappel donne la proportion de notes correctes parmi les notes originales alors que la précision donne la proportion de notes correctes parmi les notes transcrites. Les deux critères peuvent être synthétisés en un seul pour obtenir une note globale, par exemple via la F-mesure f [Van Rijsbergen, 1979] définie par

$$f \triangleq 2\frac{rp}{r+p} \tag{6.3}$$

De manière relativement équivalente, on peut également définir une note globale a, appelée *score* [Dixon, 2000] ou *accuracy* [Poliner et Ellis, 2007; Bertin *et al.*, 2007], par

$$a \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN} + \#\mathrm{FP}} \tag{6.4}$$

$$= \frac{1}{\frac{2}{f} - 1} \tag{6.5}$$

Un autre système d'évaluation quantitative, utilisé par Raphael [2002]; Poliner et Ellis [2007] et de façon plus simplifiée par Kameoka *et al.* [2007], repose non plus sur deux mais sur trois critères complémentaires : les taux de notes manquantes, de notes substituées, et de fausses alarmes (dont la définition diffère du cas précédent). Le décompte s'effectue par trame et les trois critères sont définis par

$$E_{\text{miss}} \triangleq \frac{\sum_{t=1}^{T} \max\left(0, \#\text{FN}_t - \#\text{FP}_t\right)}{\sum_{t=1}^{T} \left(\#\text{TP}_t + \#\text{FN}_t\right)} \tag{6.6}$$

$$E_{\text{subs}} \triangleq \frac{\sum_{t=1}^{T} \min\left(\#\text{FP}_t, \#\text{FN}_t\right)}{\sum_{t=1}^{T} \left(\#\text{TP}_t + \#\text{FN}_t\right)} \tag{6.7}$$

$$E_{\text{fa}} \triangleq \frac{\sum_{t=1}^{T} \max\left(0, \#\text{FP}_t - \#\text{FN}_t\right)}{\sum_{t=1}^{T} \left(\#\text{TP}_t + \#\text{FN}_t\right)} \tag{6.8}$$

$$\tag{6.9}$$

où T est le nombre de trames et TP_t, FN_t et FP_t désignent respectivement l'ensemble des notes correctes, des notes oubliées et des notes ajoutées dans la trame t. Un taux d'erreur global est alors

$$E_{\text{tot}} \triangleq \frac{\sum_{t=1}^{T} \max\left(\#\text{FP}_t, \#\text{FN}_t\right)}{\sum_{t=1}^{T} \left(\#\text{TP}_t + \#\text{FN}_t\right)} \tag{6.10}$$

$$\tag{6.11}$$

Évaluation de la transcription de la durée

Ryynänen et Klapuri [2005] proposent un critère d'évaluation de la durée transcrite. Pour chaque note n correctement transcrite, on définit le taux de recouvrement (*overlap ratio*) o_n entre la note originale et la note transcrite comme étant le rapport entre la longueur de l'intersection des supports temporels des deux notes et celle de leur union :

$$o_n \triangleq \frac{\min t_n^{\text{off}} - \max t_n^{\text{on}}}{\max t_n^{\text{off}} - \min t_n^{\text{on}}} \tag{6.12}$$

où t_n^{on} et t_n^{off} sont les couples d'instants, respectivement d'attaque et d'extinction, pour les notes originale et transcrite d'indice n. Le taux de recouvrement moyen (MOR, *mean overlap ratio*) o est ensuite obtenu en prenant la moyenne des taux de recouvrement de toutes les notes transcrites :

$$o \triangleq \frac{1}{N} \sum_{n=1}^{N} o_n \tag{6.13}$$

où N est le nombre total de notes transcrites.

Vers des critères subjectifs

Les méthodes d'évaluation habituellement répandues offrent un réel aperçu des performances d'un système, sous un angle permettant un jugement objectif. On retrouve ce type d'évaluations pour les algorithmes d'estimation de fréquences fondamentales multiples pour lesquels les erreurs sont dénombrées, soit en fonction de la polyphonie, soit de manière globale. Une évaluation objective est alors adaptée dans la mesure où la tâche n'est

pas associée à une application musicale particulière, et où elle permet d'examiner plusieurs taux d'erreurs de base, tels que la précision et le rappel. En revanche, lorsqu'il s'agit de la transcription de morceaux de musique, l'évaluation objective semble insuffisante, puisque la qualité musicale d'une transcription n'est pas en rapport direct avec ce type de décompte. Par exemple, la qualité d'une transcription est en général jugée meilleure lorsque les erreurs sont des oublis plutôt que des ajouts. Par ailleurs, l'aspect rythmique et le contexte harmonique sont déterminants pour évaluer la gravité d'une erreur, comme le montrent les études sur l'influence de la tonalité [Bigand *et al.*, 1999].

En considérant la problématique de l'évaluation des résultats en général, un parallèle peut être dressé entre le problème posé pour la transcription et les pratiques répandues dans les domaines du codage audio ou vidéo et de la séparation de sources. Comme illustré dans le tableau 6.1, on peut considérer que les métriques utilisés actuellement pour l'évaluation des transcriptions correspondent à un niveau 0 d'évaluation, au même titre que des critères de rapport signal à bruit (SNR) en codage, des métriques légèrement plus détaillées comme le rapport signal à interférences (SIR), le rapport signal à artéfacts (SAR) et le rapport signal à distortion (SDR) utilisés en séparation de sources [Vincent *et al.*, 2006]. De ce point de vue, les travaux dans le domaine du codage ont menés à des systèmes d'évaluation plus évolués. Le test d'écoute est ainsi considéré comme la méthode d'évaluation optimale et est utilisée comme ultime critère. Le protocole étant difficile à mettre en oeuvre matériellement et financièrement, des travaux sont menés pour proposer des métriques dites « objectives » qui permettent de remplacer les jugements subjectifs [Winkler, 2005; Huber et Kollmeier, 2006; Creusere *et al.*, 2008]. Elles permettent ainsi une évaluation de qualité intermédiaire, entre l'évaluation que nous qualifions de niveau 0 et l'évaluation de qualité optimale par des sujets.

niveau	transcription	codage audio/video
0	F-mesure $f \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP}+\frac{1}{2}\#\mathrm{FN}+\frac{1}{2}\#\mathrm{FP}}$	SNR/PSNR
0^+	rappel r, precision p, $f = 2\frac{rp}{r+p}$	SNR, SIR, SAR, SDR (séparation de sources)
intermédaire	?	métriques plus élaborées
optimal	?	tests d'écoute

Table 6.1 – classement qualitatif des méthodes d'évaluation

S'il peut sembler inconcevable de créer un système d'évaluation pouvant se substituer au jugement et à la sensibilité humains, on peut néanmoins espérer pouvoir en intégrer quelques aspects élémentaires dans des critères d'évaluation. Les travaux sur la similarité entre deux extraits musicaux ont abordé cette problématique, plutôt dans le cadre de la comparaison de mélodies que dans celui de la transcription. La distance d'édition, ou distance de Levenshtein [Levenshtein, 1965], qui minimise la distance entre deux séquences quelconques de symboles en utilisant trois opérations élémentaires (insertion, suppression et remplacement d'un symbole), a ainsi été adaptée par Mongeau et Sankoff [1990] à la comparaison de deux séquences monodiques, c'est-à-dire sans polyphonie. Les symboles considérés sont alors les notes, prises dans leur ordre chronologique d'apparition. L'adaptation au contexte musical consiste en l'ajout de deux opérations élémentaires – la fusion de plusieurs notes consécutives de même hauteur et son opération inverse, la fragmenta-

tion d'une note – et en l'assignation de poids à chaque opération en fonction du contexte tonal des notes modifiées. La distance d'édition a ensuite été étendue au cas polyphonique par Hanna et Ferraro [2007]. Elle s'applique à la classe restreinte des pièces musicales dans lesquelles les notes simultanées commencent au même instant, à la manière d'accords. Un morceau adopte alors une structure de séquence quotientée à laquelle on peut appliquer la distance d'édition.

L'aspect séquentiel imposé par la distance d'édition constitue un obstacle à son utilisation dans le cas général de morceaux polyphoniques quelconques du fait des chevauchements possibles entre les supports temporels des notes. Typke *et al.* [2003, 2007] proposent une autre approche en considérant les notes comme des points dans un espace métrique multidimensionnel, comme le plan temps-hauteur muni d'une distance euclidienne. Un poids est également attribué à chaque note, en l'occurrence sa durée. La similarité entre deux morceaux est alors définie comme l'effort minimum à fournir pour transformer, via la distance choisie, l'ensemble de points pondérés représentant l'un des morceaux en l'ensemble de points représentant l'autre. Pour ce faire, deux algorithmes sont proposés, l'EMD (*Earth Movers Distance*), et sa variante la PTD (*Proportional Transportation Distance*) pour laquelle il y a conservation de la masse totale des deux morceaux. Dans ces algorithmes, les déplacements de poids peuvent s'opérer partiellement, d'une note source vers plusieurs notes cibles et *vice-versa*, et aucune contrainte sur les morceaux n'est requise, en particulier sur l'aspect polyphonique.

Ces méthodes ont en commun d'intégrer, sous forme d'opérations spécifiques, des différences typiques que l'on peut rencontrer entre deux morceaux. Le choix des coûts relatifs de ces opérations est une étape incontournable dans l'élaboration de telles méthodes et repose sur la perception de la musique. Nous nous y intéressons maintenant à travers un test perceptif.

6.1.2 Évaluation subjective des erreurs typiques de transcription

6.1.2.1 Principe du test et protocole

L'objectif du test est d'obtenir un classement des erreurs typiques de transcription sur une échelle numérique de gêne. Nous l'avons réalisé au printemps 2007, via Internet, dans le cadre du stage de Master 2 ATIAM d'Adrien Daniel. Il repose sur l'écoute d'extraits musicaux et de leurs transcriptions et sur leur comparaison par les sujets. Ces morceaux sont générés à partir de fichiers MIDI que l'on peut modifier précisément. En particulier, les transcriptions sont artificiellement obtenues en insérant un et un seul type d'erreur typique parmi l'ensemble suivant :

- suppression d'une note ;
- insertion d'une note à l'octave d'une note originale ;
- insertion d'une note à la quinte d'une note originale ;
- insertion d'une note de hauteur quelconque (distante d'entre 1 et 11 demi-tons d'une note originale) ;
- remplacement d'une note à l'octave d'une note originale ;
- remplacement d'une note à la quinte d'une note originale ;
- remplacement d'une note de hauteur quelconque (distante d'entre 1 et 11 demi-tons d'une note originale) ;
- fragmentation d'une note en deux notes de même hauteur ;
- déplacement de l'instant d'attaque d'une note ;
- changement de la durée d'une note (déplacement de l'instant de fin) ;

Figure 6.1 – Test 1 : pour chaque paire de sons, le sujet désigne celui provoquant la plus grande gêne.

– modification de la nuance d'une note (paramètre MIDI *velocity*).

L'épreuve est réalisée sur trois séries correspondant à des extraits de l'*Étude, op. 10, No. 1 en Do Majeur* de F. Chopin (extrait de 8 s), du *Clair de Lune* de la *Suite bergamasque* de C. Debussy (extrait de 20 s), et de l'Allegro con spirito de la *Sonate en Ré Majeur KV. 311* de W.A. Mozart (extrait de 13 s). Pour chaque type d'erreur, plusieurs transcriptions sont générées en faisant varier le taux d'erreur. Le nombre de notes modifiées en pratique est déterminé par un taux de notes modifiées fixé à l'avance (MNR, Modified Note Rate), qui peut prendre les valeurs 10% et 33%. Les notes modifiées sont tirées au hasard. L'intensité de l'erreur (EI, Error Intensity) de durée, d'attaque et de nuance est contrôlée par un autre paramètre pouvant prendre les valeurs 25% et 75%. L'intensité des erreurs est aléatoire, uniformément répartie sur l'intervalle borné par le paramètre EI. Les durées et instants d'attaques sont modifiés en termes de proportion de la durée originale, et la nuance en proportion de la nuance originale.

Pour obtenir une quantification de la gêne occasionnée par chaque morceau, c'est-à-dire par chaque erreur typique à un taux donné, il n'est pas envisageable de demander aux sujets d'établir un tel classement en raison du nombre trop élevé de fichiers. La solution adoptée est le protocole BTL (Bradley-Terry-Luce) proposé par Bradley [1953], méthode statistique qui n'exige du sujet que des jugements binaires consistant à choisir la transcription occasionnant le plus de gêne dans chaque paire présentée (cf. figure 6.1). Les résultats sont ensuite traités pour obtenir l'échelle globale souhaitée.

Cette méthode repose sur l'hypothèse qu'il existe de véritables valeurs de gêne associées aux transcriptions et qu'elles constituent des variables latentes du problème. La réponse du sujet pour une paire donnée est le résultat d'une comparaison d'une version bruitée des valeurs vraies, le bruit modélisant l'incertitude introduite par le sujet. Ce modèle statistique permet ensuite de reconstruire l'échelle des valeurs vraies à partir des réponses binaires. Pour chacun des trois extraits musicaux, 20 paires de transcriptions ont ici été présentées à chaque sujet. On peut estimer l'incertitude sur les résultats en utilisant une méthode dite de bootstrap [Efron et Tibshirani, 1993], qui construit une distribution empirique à partir des données de l'expérience, en réalisant plusieurs rééchantillonnages (100 dans notre cas) de la distribution des observations.

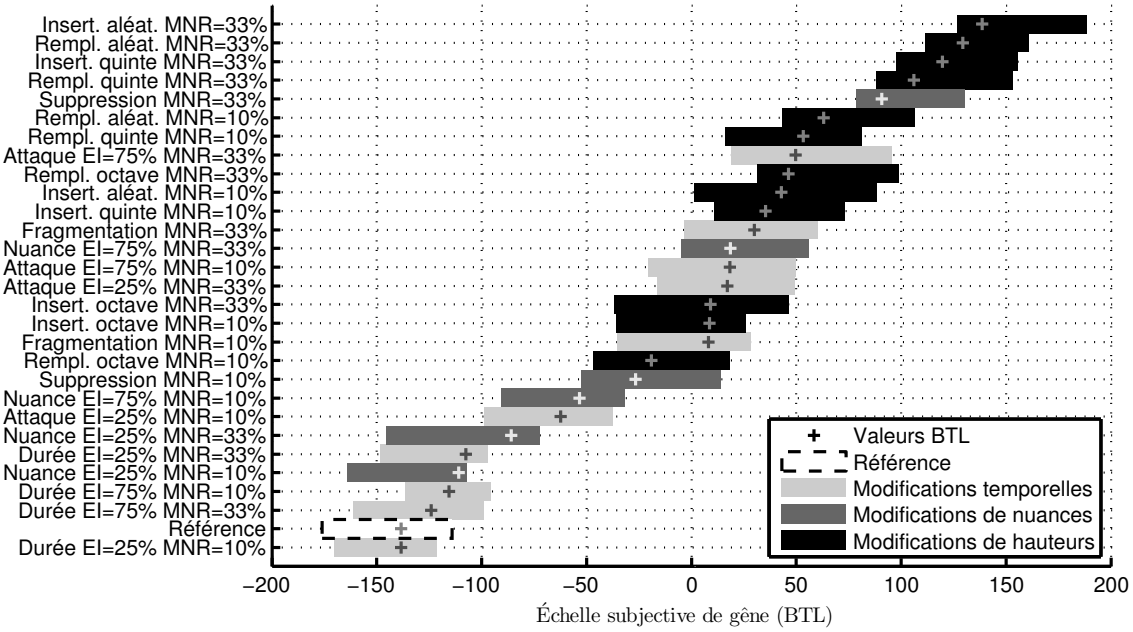


FIGURE 6.2 – Échelle subjective de gêne en fonction des erreurs typiques : les croix représentent les valeurs trouvées, les barres l'intervalle de confiance à 90% obtenu par une méthode de bootstrap [Efron et Tibshirani, 1993] (les intervalles ne sont pas centrés sur les valeurs BTL car la distribution des données n'est pas forcément gaussienne).

6.1.2.2 Résultats

Trente-sept sujets, dont 24 musiciens et 13 non-musiciens ont participé à ce test. Les commentaires des sujets à la fin du test montrent que les instructions ont bien été comprises. Lorsqu'on leur demande de verbaliser les types d'erreurs entendues, les sujets évoquent majoritairement les erreurs de hauteur. Rares sont ceux qui parviennent à nommer des types d'erreur comme les erreurs d'intensité ou de durée.

Les résultats du test sont représentés sur la figure 6.2, sous forme d'une échelle numérique de gêne que la méthode BTL permet de construire en fonction des jugements binaires des sujets.

Plusieurs éléments permettent de vérifier la consistance des résultats. Premièrement, pour un type d'erreur donné, la gêne croît avec les taux d'erreur – MNR ou EI – et décroît avec la consonance (intervalle d'octave, de quinte, quelconque). Deuxièmement, les intervalles de confiance sont suffisamment étroits pour permettre de distinguer les types d'erreur les uns des autres. Troisièmement, du fait de leur épaisseur tout de même non négligeable, la plupart des types d'erreur doivent être considérés comme subjectivement équivalent à leurs plus proches voisins. Quatrièmement, une gêne minimale a été attribuée à la référence (à l'incertitude des résultats près).

Comme on peut s'y attendre d'après les commentaires des sujets, les erreurs de hauteurs sont les plus gênantes. Les modifications temporelles et d'intensité causent quant à elles une gêne faible à moyenne. Parmi les erreurs de hauteur, celles d'octave sont les moins gênantes, suivies de celles de quinte puis de celles de hauteur quelconque. À taux égal, remplacements et insertions obtiennent le même score, ce qui indiquerait que la gêne d'un

remplacement est plutôt causée par la note ajoutée que par la note omise. Les faibles valeurs obtenues pour les suppressions confirment cette hypothèse, qui est couramment remarquée lors des travaux sur la transcription automatique : il vaut mieux oublier une note qu'en ajouter une, le résultat étant en général subjectivement meilleur.

Les résultats relatifs aux erreurs temporelles montrent que celles sur l'attaque sont plus gênantes que celles sur la durée. Les sujets semblent même insensibles à la majorité de ces dernières. La nature de l'instrument utilisé est une explication probable : les sons de piano, caractérisés par leurs oscillations libres ont une fin moins perceptible que des sons d'instruments à oscillations entretenues. De ce point de vue, les résultats de ce test ne sont pas généralisables à tous les instruments.

Enfin, une analyse complémentaire des résultats montre deux tendances qui ne sont pas visibles sur la figure 6.2. La première est que les résultats obtenus avec les musiciens et les non-musiciens sont similaires. La seconde est que les échelles séparées pour chaque morceau donnent également des résultats comparables, à l'exception de l'extrait de Debussy, pour lequel les erreurs de suppression sont plus faibles et celles de durée plus élevées, probablement en raison du tempo relativement faible.

6.1.3 Critères perceptifs d'évaluation

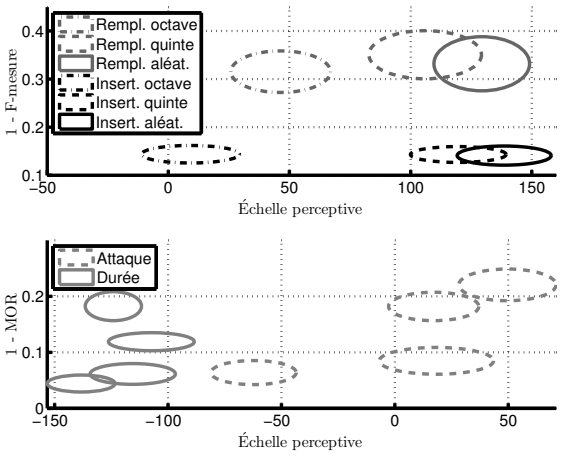


FIGURE 6.3 – Exemples de différences entre évaluations objective et subjective : les résultats perceptifs sont confrontés à la la quantité 1−F-mesure (en haut, pour des insertions et des remplacements à MNR= 33%) et au taux de recouvrement moyen (MOR, en bas, pour les modifications de durée et d'instants d'attaque). Chaque type d'erreur est représenté avec un tracé et une couleur propres, quel que soit le taux d'erreur (la tendance de celui-ci est d'augmenter de gauche à droite et de bas en haut), d'où la présence de plusieurs ellipses de même tracé. L'incertitude selon chaque dimension est représentée par les dimensions des ellipses. Les ellipses n'étant pas disposées selon une courbe croissante, la F-mesure et le MOR ne sont pas représentatifs de l'échelle perceptive.

Lorsque l'on compare les résultats perceptifs du test et ceux obtenus avec des méthodes d'évaluation objective introduites dans la partie 6.1.1, les deux types d'évaluation diffèrent sur certains points. La figure 6.3 en explicite quelques-uns. Ainsi, les modifications d'octave, de quinte et de hauteur aléatoire ont la même F-mesure alors que la gêne générée est

croissante. Au contraire, remplacements et insertions sont perceptivement équivalents alors que la F-mesure pénalise plus les remplacements (comptés comme une omission et une insertion). Quant à la différence perceptive déjà évoquée entre modifications d'instants d'attaque et de durée, elle n'est pas prise en compte par le taux de recouvrement moyen. On peut donc espérer prendre en compte ces résultats perceptifs pour pondérer les types d'erreur afin d'obtenir des mesures d'évaluation subjective. Les résultats du test présentent en outre l'avantage de donner des poids relatifs entre tous les types d'erreurs, y compris quand elles sont de dimensions différentes comme les erreurs temporelles et de hauteur, alors que la pondération entre temps et hauteur fait souvent l'objet d'un réglage de paramètre laissé à la discrétion de l'utilisateur dans les méthodes de type distance d'édition ou PTD.

Dans cette partie, nous estimons les coefficients de pondération des erreurs typiques à partir des résultats du test, puis nous adaptons deux méthodes d'évaluations – la F-mesure et la PTD – pour en donner des versions destinées à refléter une évaluation subjective. L'application de ces méthodes à des transcriptions réelles d'extraits musicaux permet de valider les résultats.

6.1.3.1 Extraction des coefficients de pondération

L'extraction des coefficients commence par une étape de normalisation entre 0 et 1 des résultats du test. Nous sélectionnons ensuite ceux dont le MNR vaut 33%, et les moyennons dans le cas des insertions et remplacements. Nous obtenons ainsi une réduction des erreurs typiques à 6 critères représentatifs [1] à intégrer dans les métriques, et dont les coefficients de pondérations associés figurent dans le tableau 6.2. Ces coefficients ont été normalisés de telle sorte que

$$\frac{1}{3}\sum_{i=1}^{3}\alpha_i + \sum_{i=4}^{6}\alpha_i = 1 \tag{6.14}$$

car les erreurs d'octave, de quinte et d'autres hauteurs sont des fausses alarmes complémentaires.

Critères	**Poids**
Octave	$\alpha_1 = 0,1794$
Quinte	$\alpha_2 = 0,2712$
Autres intervalles	$\alpha_3 = 0,2941$
Suppression	$\alpha_4 = 0,2475$
Durée	$\alpha_5 = 0,0355$
Instants d'attaque	$\alpha_6 = 0,4687$

TABLE 6.2 – Coefficients perceptifs associés à des erreurs typiques

1. Le critère sur la nuance a été éliminé car les résultats obtenus ne sont pas satisfaisants, probablement en raison de la difficulté de modéliser une échelle de perception des nuances. La fragmentation n'est pas utilisée non plus car elle était difficilement intégrable dans les métriques.

6.1.3.2 F-mesure perceptive

La F-mesure définie par l'équation (6.3) s'exprime en fonction du nombre de TP, de FN et de FP :

$$f = \left(\frac{1}{2} \times \frac{1}{p} + \frac{1}{2} \times \frac{1}{r}\right)^{-1} \tag{6.15}$$

$$= \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \frac{1}{2}\#\mathrm{FP} + \frac{1}{2}\#\mathrm{FN}} \tag{6.16}$$

Les erreurs, de deux types – FP et FN –, y sont comptabilisées avec des poids identiques égaux à $\frac{1}{2}$. En introduisant cette mesure, Van Rijsbergen [1979] a étudié l'hypothèse de pondérer différemment ces deux types d'erreurs grâce à un coefficient $\alpha \in [0; 1]$ tel que

$$f = \left(\frac{\alpha}{p} + \frac{1-\alpha}{r}\right)^{-1} \tag{6.17}$$

$$= \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \alpha\#\mathrm{FP} + (1-\alpha)\#\mathrm{FN}} \tag{6.18}$$

En nous inspirant de cette expression, nous définissons la F-mesure perceptive par

$$f_{\mathrm{percept}} \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \sum_{i=1}^{6} \alpha_i w_i \#E_i} \tag{6.19}$$

où $\#E_i$ est le nombre d'erreurs de type i (cf. algorithme 6.1), $w_1 = w_2 = w_3 = w_4 = 1$, w_5 est l'erreur moyenne de durée dans la transcription à évaluer, et w_6 est l'erreur moyenne d'instants d'attaque (ces erreurs moyennes sont en pratique calculées comme la racine de l'erreur quadratique moyenne). On peut remarquer que l'expression (6.19) est égale à la F-mesure lorsque l'on prend $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2}$ et $\alpha_5 = \alpha_6 = 0$. La classe d'erreurs FP contenant les sous-catégories octave/quinte/autres (erreurs de type $i \in \{1; 2; 3\}$, nous voyons que la normalisation introduite par l'équation (6.14) permet de retrouver la F-mesure originale en affectant des poids identiques à chaque type d'erreur, en considérant le poids global correspondant à l'ensemble de ces trois sous-catégories pour effectuer la sommation à 1.

Par ailleurs, on peut définir de manière équivalente une *accuracy* perceptive, qui est égale à l'*accuracy* (équation (6.5)) dans le même cas particulier :

$$a_{\mathrm{percept}} \triangleq \frac{\#\mathrm{TP}}{\#\mathrm{TP} + 2\sum_{i=1}^{6} \alpha_i w_i \#E_i} \tag{6.20}$$

$$= \frac{1}{\frac{2}{f_{\mathrm{percept}}} - 1} \tag{6.21}$$

L'extraction des erreurs à partir des fichiers MIDI du morceau original et de la transcription s'effectue selon l'algorithme 6.1.

Entrées: Fichiers MIDI du morceau original et de la transcription.
Les TP sont estimés comme des notes ayant une hauteur correcte (au demi-ton près, suivant la correspondance donnée en annexe C (p. 177)) et un instant d'attaque juste à 150 ms près. Pour chaque TP, on estime en outre l'erreur sur la durée et sur l'attaque (en proportion de la durée de la note originale). On fixe alors $\#E_5 = \#E_6 = \#\text{TP}$.
Les FP sont les notes transcrites qui ne sont pas des TP.
Pour chaque FP,
  Si il existe une note originale à l'octave supérieure ou inférieure au même moment (c'est-à-dire avec un recouvrement partiel des supports temporels), **alors**
    le FP est ajouté à l'ensemble E_1 des FP d'octave.
  Sinon si il existe une note originale à la quinte supérieure ou inférieure au même moment, **alors**
    le FP est ajouté à l'ensemble E_2 des FP de quinte.
  Sinon
    le FP est ajouté à l'ensemble E_3 des autres FP.
  Fin Si
Fin Pour
Les FN sont l'ensemble E_4 des notes originales qui ne sont pas des TP.
Sorties: Ensembles $E_1, E_2, E_3, E_4, E_5, E_6$ d'erreurs typiques.

Algorithme 6.1: Extraction des erreurs typiques.

6.1.3.3 PTD perceptive

La PTD introduite dans la partie 6.1.1 est à l'origine appliquée à l'évaluation de la similarité de mélodies. Dans ce contexte, prendre les durées des notes comme poids à transférer et utiliser la distance euclidienne dans le plan temps/hauteur semblent être des choix appropriés. Cependant, dans le cas générique de la comparaison de morceaux quelconques, ces deux choix doivent être revus. Idéalement, il conviendrait d'attribuer des poids de PTD en fonction de l'importance musicale des notes, comme le suggèrent Typke *et al.* [2007]. Cela reste en dehors du cadre que nous nous sommes fixé, et nous nous contentons ici d'un poids unitaire pour chaque note. En utilisant les coefficients du tableau 6.2, nous définissons une distance perceptive entre deux notes dans l'espace multidimensionnel composé de la hauteur (octave, quinte ou autre), la durée et l'instant d'attaque. L'algorithme est ensuite appliqué en utilisant les poids unitaires et la distance perceptive.

6.1.3.4 Application à l'évaluation subjective de transcriptions musicales

Nous souhaitons maintenant déterminer dans quelle mesure les deux métriques perceptives que nous avons définies permettent de s'approcher de l'évaluation subjective d'une transcription. Nous élaborons un test perceptif afin de disposer de notes fournies par des sujets, pour des transcriptions réelles de plusieurs pièces pour piano. Nous utilisons pour cela trois extraits provenant du *Prélude en do mineur BWV 847* de J.S. Bach (13 secondes), du *Clair de Lune* de la *Suite bergamasque* de C. Debussy (20 secondes), et de *Allegro con spirito* de la *Sonate en Ré Majeur KV 311* de W.A. Mozart (13 secondes). Cinq transcriptions de chaque extrait sont présentées au sujet. Il peut les écouter autant de fois qu'il le souhaite et doit donner à chacune un score représentant la gêne ressentie, en comparaison avec le morceau original (cf. figure 6.4).

Les mêmes morceaux sont présentés à tous les sujets, dans un ordre aléatoire pour

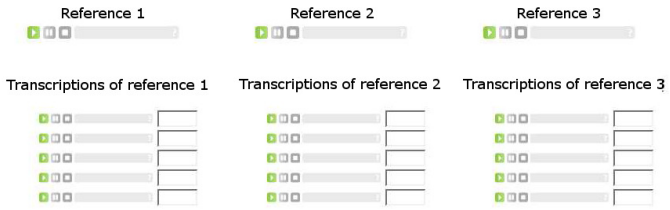


FIGURE 6.4 – Test 2 : le sujet attribue aux transcriptions un score (nombre positif).

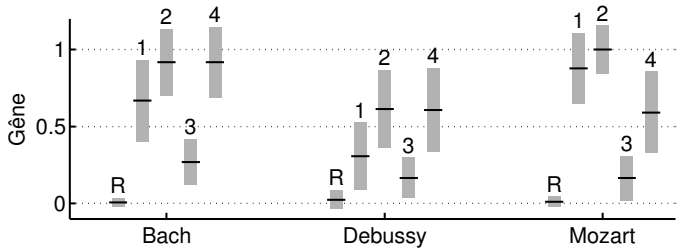


FIGURE 6.5 – Résultat de l'évaluation de transcriptions. Les traits noirs indiquent les valeurs moyennes, les barres grises l'étalement des réponses selon les sujets. Les chiffres font référence aux différents systèmes de transcription (rendus anonymes pour éviter de présenter les résultats comme ceux d'une comparaison de systèmes), et 'R' désigne la référence.

chacun. Pour chaque extrait, l'une des cinq transcriptions est en réalité l'original, afin de contrôler la cohérence des résultats. Les quatre autres ont été obtenues par des systèmes de transcription automatique : SONIC [Marolt, 2004], disponible sur le site Internet de l'auteur, le système de Bertin *et al.* [2007], un système de P. Leveau selon [Leveau *et al.*, 2008] et une version préliminaire de [Emiya *et al.*, 2008]. Les erreurs commises dépendent donc du comportement spécifique de chaque système.

Les résultats sont représentés sur la figure 6.5. Ils ont été normalisés par la note maximale donnée par chaque sujet, et ceux qui avaient donné une gêne supérieure à 20% à la référence ont été éliminés (6 sujets sur 37). La moyenne et l'écart-type par morceau, par rapport à tous les sujets restants, sont alors calculés. Ces résultats ont été validés par un test ANOVA factoriel 3×5 (nombre de compositeurs $\times$ nombre de systèmes de transcription). Le test est passé avec succès, avec un niveau $p = 0,01$ (c'est-à-dire un risque de 5%), le long de chaque dimension et suivant les interactions entre les dimensions. Les notes obtenues varient significativement en fonction de l'extrait, ce qui confirme que les performances dépendent du contenu musical des morceaux et de la base de données d'évaluation choisie. La largeur des écarts-type montre l'importance de critères subjectifs personnels dans l'évaluation d'une transcription, et le recouvrement qu'il en résulte entre les scores reflète la difficulté d'une entreprise d'évaluation de systèmes de transcription, même si l'on peut attribuer la première et la dernière places respectivement aux systèmes 3 et 2.

Nous pouvons à présent appliquer les métriques perceptives définies précédemment et comparer les notes de l'évaluation subjective ainsi qu'avec les résultats obtenus avec leurs versions originales. La figure 6.6 représente ces résultats. La F-mesure et la F-mesure perceptive ont subi l'opération $x \mapsto 1 - x$ afin de représenter des taux d'erreur, et non une

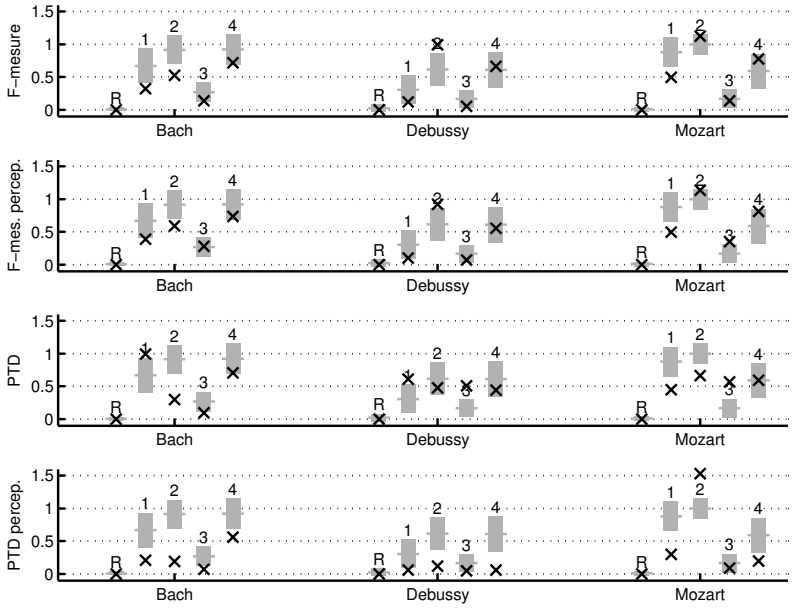


FIGURE 6.6 – Résultats d'évaluation de transcriptions (croix) avec plusieurs métriques objectives et subjectives : dans chaque cas, la gêne ressentie est représentée par les barres grises à titre de comparaison.

similarité, et d'être comparées à la gêne. La correspondance entre les échelles de chaque métrique et la gêne est par ailleurs obtenue par multiplication par un coefficient d'échelle obtenu en minimisant l'erreur quadratique moyenne.

Pour quantifier la qualité des métriques proposées par rapport aux métriques usuelles, nous proposons d'utiliser les critères établis pour un contexte similaire dans le domaine du codage vidéo [Winkler, 2005]. Pour comparer les résultats $y = (y_1, \ldots, y_N)$ d'une métrique avec les jugements subjectifs $x = (x_1, \ldots, x_N)$, trois critères sont ainsi introduits :

– la **précision de la prédiction**, donnée par le coefficient de corrélation linéaire de Pearson

$$r_P \triangleq \frac{\sum_n (x_n - \overline{x})(y_n - \overline{y})}{\sqrt{\sum_n (x_n - \overline{x})^2}\sqrt{\sum_n (y_n - \overline{y})^2}} \tag{6.22}$$

où $\overline{\cdot}$ désigne la moyenne empirique des valeurs.

– la **monotonie de la prédiction**, donnée par le coefficient de corrélation de Spearman à partir du rang $\mathrm{rg}\,()$ des scores

$$r_S \triangleq \frac{\sum_n \left(\mathrm{rg}\,(x_n) - \overline{\mathrm{rg}\,(x)}\right)\left(\mathrm{rg}\,(y_n) - \overline{\mathrm{rg}\,(y)}\right)}{\sqrt{\sum_n \left(\mathrm{rg}\,(x_n) - \overline{\mathrm{rg}\,(x)}\right)^2}\sqrt{\sum_n \left(\mathrm{rg}\,(y_n) - \overline{\mathrm{rg}\,(y)}\right)^2}} \tag{6.23}$$

– la **consistence de la prédiction**

$$R_O \triangleq \frac{N_O}{N} \tag{6.24}$$

où $N_O \triangleq \#\{n/\,|x_n - y_n| > 2\sigma_x\}$ est le nombre d'*outliers* (données aberrantes) calculé en fonction d'un seuil σ_x.

Les résultats, donnés dans le tableau 6.3, montrent une amélioration globale de la qualité lorsque l'on passe d'une métrique usuelle à sa version perceptive. On constate cette amélioration pour les critères de précision et de monotonie de la prédiction, la consistence étant ici non-significative (aucun outlier à l'exception d'un seul dans le cas de la PTD perceptive). L'amélioration est légère dans le cas de la F-mesure perceptive. Elle est plus significative pour la monotonie de la prédiction de la PTD perceptive, passant de $61,6\%$ à $89,6\%$. On remarque par ailleurs une valeur isolée très élevée obtenue pour la transcription dont la gêne est maximale (Mozart, Système 2), qui perturbe la mise à l'échelle des résultats, expliquant la présence d'un *outliers*, et surtout la faible précision de la PTD perceptive par rapport à la F-mesure perceptive. Ainsi, la relation entre valeurs subjectives et PTD perceptive semble non-linéaire, tout en conservant de très bonnes propriétés de monotonie.

	Précision	Monotonie	Consistence
F-mesure	$83,4\%$	$83,5\%$	0%
F-mesure perceptive	$\mathbf{84,1}\%$	$\mathbf{84,9}\%$	0%
PTD	$60,3\%$	$61,6\%$	0%
PTD perceptive	$\mathbf{64,8}\%$	$\mathbf{89,6}\%$	$6,7\%$

Table 6.3 – Qualité des métriques d'évaluation.

6.2 Base d'évaluation MAPS

Après avoir abordé la question des méthodes d'évaluation, nous nous intéressons maintenant à la deuxième composante nécessaire pour évaluer un système qu'est la base de données. Pour ce genre d'évaluation, une base de sons doit remplir deux conditions : fournir des sons de qualité en assurant leur diversité et les accompagner d'une référence précise, résultat optimal à atteindre par les systèmes testés. C'est dans cette optique que nous avons élaboré la base MAPS (*MIDI Aligned Piano Sounds*) présentée ci-après.

6.2.1 Vue d'ensemble de la base

La base MAPS a été spécifiquement conçue pour l'évaluation des algorithmes d'estimation de fréquences fondamentales multiples et des systèmes de transcription. Elle a été imaginée suite à une première tentative d'enregistrement avec un piano de type Yamaha Disklavier jouée par la pianiste Mélanie Desportes. À la suite de cet enregistrement, nous avons réalisé que dans l'optique de la transcription automatique, il était nécessaire d'automatiser la tâche de génération pour pouvoir disposer de grands volumes de données. La base que nous proposons se compose donc de plusieurs types de sons, des notes isolées les plus simples aux morceaux de musique du répertoire et de plusieurs conditions d'enregistrement. L'automatisation de la tâche est possible de deux manières. La première consiste à utiliser un piano Yamaha Disklavier, vérible instrument auquel a été ajouté un dispositif de moteurs permettant d'actionner les touches et les pédales. De cette façon, le piano est joué automatiquement, sans pianiste, le pilotage se faisant via une liaison MIDI. L'autre manière employée pour automatiser la production de sons de piano repose sur l'utilisation de logiciels de synthèse de qualité à base de banques de sons pré-enregistrés.

L'utilisation d'une dizaine de pianos et de conditions d'enregistrement différents confère à la base une diversité qualitative relativement importante. Le deux procédés de production

décrits précédemment permettent d'offrir cette diversité. En rendant possible la comparaison des transcriptions en fonction du mode de génération des sons, nous souhaitons aborder la question récurrente de la validité des bases synthétiques, souvent utilisées car plus faciles à créer. L'utilisation de plusieurs pianos droits et pianos à queue contribue également à la variété qualitative des sons. Enfin, la diversité des conditions d'enregistrement se traduit d'une part par l'utilisation de réverbérations différentes (dans le cas des synthétiseurs logiciels) et d'autre part, par une double prise de son, la première à proximité de l'instrument, la deuxième à plus grande distance, pour certains pianos. L'ensemble de ces combinaisons d'instruments et de conditions d'enregistrement sont représentées dans le tableau 6.4.

Abréviation	Piano réel / Logiciel	Instrument	Conditions d'enregistrement
StbgTGd2	The Grand 2 (Steinberg)	Hybride	Par défaut
AkPnBsdf	Akoustik Piano (Native Instruments)	Boesendorfer 290 Imperial	*Preset* « Gregorian »
AkPnBcht	Akoustik Piano (Native Instruments)	Bechstein D 280	*Preset* « Bechstein Bach »
AkPnCGdD	Akoustik Piano (Native Instruments)	Concert Grand D	*Preset* « Production »
AkPnStgb	Akoustik Piano (Native Instruments)	Steingraeber 130 (piano droit)	*Preset* « Modern Play Time »
SptkBGAm	The Black Grand (Sampletekk)	Steinway D,	"Ambient" (prise distante)
SptkBGCl	The Black Grand (Sampletekk)	Steinway D,	« Close » (à proximité)
ENSTDkAm	Piano réel (Disklavier)	Yamaha Mark III (piano droit)	« Ambient » (prise distante)
ENSTDkCl	Piano réel (Disklavier)	Yamaha Mark III (piano droit)	« Close » (à proximité)

Table 6.4 – MAPS : instruments et conditions d'enregistrement.

La diversité du contenu de la base a été la deuxième préoccupation lors de sa composition. Pour chaque instrument, quatre classes de sons ont été distinguées, le contenu de chacune étant détaillé dans la partie 6.2.2 :

- classe ISOL : notes isolées et autres extraits monophoniques ;
- classe RAND : accords tirés aléatoirement ;
- classe UCHO : accords usuels ;
- classe MUS : morceaux de musique.

Les références associées à chaque son ou morceau sont créées au format SMF (*Standard MIDI File*), le format de fichier associé au protocole MIDI. Elles sont d'une grande précision grâce au dispositif d'enregistrement employé, dans lequel l'intervention humaine est minimisée au profit de processus automatisés (cf. partie 6.2.3).

La nomenclature de chaque fichier comprend sous forme abrégée le nom de la base, de la classe de sons, de l'instrument et des conditions d'enregistrement, ainsi qu'un descriptif du contenu du fichier. Chaque fichier audio, au format PCM (.wav) stéréo échantillonné à $44,1$ kHz, porte le même nom que sa référence, extension exceptée (.mid pour la référence).

Le dimensionnement de la base a été ajusté pour faciliter sa diffusion : les sons et références de chaque couple instrument/conditions d'enregistrement tiennent sur un DVD. Le nombre de paramètres que nous souhaitons faire varier conduisant à un volume inutilement grand de sons, nous avons réduit la taille à un DVD par instrument en utilisant des tirages aléatoires, comme cela sera expliqué dans la partie 6.2.2.

6.2.2 Contenu détaillé

Pour chacune des quatre classes contenues dans la base, plusieurs types de sons ont été produits. Nous détaillons ici le contenu de ces classes. Chaque son est enregistré en laissant un laps de temps de silence avant et après afin d'inclure une portion de silence et de ne pas couper l'amorce d'un son ou la résonnance après la fin d'une note.

6.2.2.1 ISOL : base de notes isolées et autres extraits monophoniques

Cette classe contient exclusivement des sons monophoniques. Elle est donc particulièrement appropriée pour tester les algorithmes d'estimation de fréquences fondamentales simples. Chaque son se caractérise par un mode de jeu *mo*, une nuance *i0*, l'utilisation ou non de la pédale *forte*, et une hauteur de note. Le fichier correspondant à un tel son porte un nom du type

MAPS_ISOL_mo_i0_Ss_Mm_nomInstrument.wav

Les différents modes de jeux *mo* sont les suivants :
- NO : jeu normal, notes durant 2 secondes ;
- LG : notes longues, entre 3 secondes pour les notes aiguës et 20 secondes pour les notes graves ;
- ST : staccato ;
- RE : une même note répétée, en accélérant ;
- CHd : montée et descente chromatiques, avec différentes durées d de notes ;
- TRi : trilles, en accélérant, au demi-ton ($i = 1$) ou au ton ($i = 2$) supérieur.

La nuance *i0* peut prendre trois valeurs possibles : P (piano), M (mezzo-forte), F (forte). L'utilisation (s= 1) ou non (s= 0) de la pédale *forte* est décidée par tirage aléatoire. Dans le cas où elle est utilisée (50% des cas), la pédale est enfoncée 300 ms avant le début de la séquence et relâchée 300 ms après la fin [2]. Le champ *nomInstrument* est une abréviation définie dans le tableau 6.4 (p. 139).

À l'exception des montées et descentes chromatiques, la hauteur du son est codée par son code MIDI $m \in [\![21; 108]\!]$ (registre du piano), et toutes les notes sont enregistrées.

6.2.2.2 RAND : base d'accords tirés aléatoirement

Cette classe contient des accords dont les notes sont tirées aléatoirement. Elle a été conçue dans la perspective d'évaluer les algorithmes de manière objective, sans *a priori* musical sur le contenu de la base, procédé répandu dans la littérature. Chaque accord est stocké dans un fichier dont le nom est du type

MAPS_RAND_Px_Mm1-m2_Ii1-i2_Ss_nn_nomInstrument.wav,

2. Enfoncer la pédale avant de jouer la note n'est pas une pratique musicale courante. Nous procédons ainsi pour que l'effet d'enfoncement de la pédale n'interfère pas avec l'attaque des notes dans le son.

où x désigne la polyphonie, *m1-m2* la tessiture, *i1-i2* l'intervalle de nuances, s l'utilisation de la pédale et n le numéro de l'accord.

La polyphonie x est comprise entre 2 et 7. La tessiture *m1-m2* est une fourchette de hauteurs MIDI prenant la valeur $21 - 108$ ou $36 - 95$, dans laquelle les notes sont tirées de façon aléatoire, uniformément sur l'intervalle. Le premier intervalle correspond à la tessiture complète du piano alors que le second, plus restreint, est souvent utilisé dans l'évaluation des algorithmes d'estimation de fréquences fondamentales multiples. Les nuances sont tirées au hasard, indépendamment pour chaque note, dans l'intervalle *i1-i2* codé en MIDI (paramètre *velocity*), pouvant prendre les valeurs $60 - 68$ (mezzo-forte) ou $32 - 96$ (de piano à forte). La durée de l'accord est également tirée aléatoirement, de façon uniforme entre 300 ms et 700 ms. L'utilisation de la pédale *forte* se fait de manière identique à la classe ISOL.

Tous paramètres fixés par ailleurs, 50 réalisations sont tirées aléatoirement et générées, le paramètre $n \in [\![1; 50]\!]$ permettant de les distinguer. La base contient par exemple 50 accords de 3 notes de hauteurs comprises entre 36 (Do 2) et 95 (Si 6), de nuance mezzo-forte, la moitié environ jouée avec la pédale *forte*.

6.2.2.3 UCHO : base d'accords usuels

Cette classe contient des accords typiques rencontrés dans la musique occidentale, la musique classique et le jazz notamment. Son intérêt est double : mesurer les performances avec un *a priori* musical et proposer des sons dont les notes sont beaucoup plus souvent en rapport harmonique. Chaque accord est stocké dans un fichier dont le nom est du type

$$MAPS\_UCHO\_Cc_1\text{-}\ldots\text{-}c_p\_Ii1\text{-}i2\_Ss\_nn\_nomInstrument.\text{wav},$$

où c_1-…-c_p désigne la composition de l'accord considéré, *i1-i2* l'intervalle de nuances, s l'utilisation de la pédale et n le numéro de l'accord. La composition c_1-…-c_p d'un accord à p notes est codée par des entiers c_k, $k \in [\![1; n]\!]$ correspondant à la distance de la note k à la fondamentale, en demi-tons. De cette façon, un accord parfait majeur sera codé 0-4-7. La durée de l'accord est par ailleurs fixée à une seconde. Les paramètres de nuance *i1-i2* et de pédale *forte* s sont régis à l'identique à la classe RAND.

En polyphonie 2, la base contient tous les intervalles de 1 à 12 demi-tons, ainsi que la 19$^{\text{e}}$ (quinte à l'octave) et la 24$^{\text{e}}$ (double-octave). En polyphonie 3, la base contient les accords parfaits majeur et mineur, les accords de quinte augmentée et de quinte diminuée. En polyphonie 4, elle contient les sept espèces d'accords usuels en harmonie, et en polyphonie 5, les dix espèces d'accords usuels. Pour les polyphonies 3 à 5, la forme fondamentale des accords et tous les renversements sont générés. L'ensemble des accords enregistrés figure dans les tableaux 6.5 et 6.6 avec le codage utilisé.

Tous paramètres fixés par ailleurs, 10 réalisations sont tirées aléatoirement et générées, le paramètre $n \in [\![1; 10]\!]$ permettant de les distinguer. Pour les accords de polyphonie ≥ 4, le nombre de tirages est réduit de 10 à 5. Lors de chaque tirage, la note fondamentale de l'accord est choisie aléatoirement sur toute la tessiture du piano permettant la génération de l'accord (entre 21 et 101 (*i.e.* $108 - 7$) pour l'exemple de l'accord parfait majeur).

6.2.2.4 MUS : base de morceaux de musique

Pour générer la base de morceaux musicaux, nous avons utilisé les fichiers SMF que Krueger [2008] a mis à disposition, sous license Creative Commons, sur son site Internet. Ces fichiers de grande qualité ont été produits, ou plutôt écrits à la main pour réaliser

Nom	Fondamental	Renvers. 1	Renvers. 2
Seconde mineure	0-1		
Seconde majeure	0-2		
Tierce mineure	0-3		
Tierce majeure	0-4		
Quarte juste	0-5		
Quinte mineure	0-6		
Quinte juste	0-7		
Sixte mineure	0-8		
Sixte majeure	0-9		
Septième mineure	0-10		
Septième majeure	0-11		
Octave juste	0-12		
Dix-neuvième	0-19		
Double octave	0-24		
Parfait majeur	0-4-7	0-3-8	0-5-9
Parfait mineur	0-3-7	0-4-9	0-5-8
Quinte diminuée	0-3-6	0-3-9	0-6-9
Quinte augmentée	0-4-8	0-4-8	0-4-8

Table 6.5 – Accords usuels de 2 et 3 sons et nomenclature (écarts à la note fondamentale en demi-tons).

en quelque sorte une interprétation sous forme MIDI. La place, la durée et l'intensité de chaque note ont ainsi fait l'objet d'un ajustement par l'auteur (interprète). Lors de la finalisation de la base, 238 morceaux du répertoire classique et traditionnel de piano étaient proposés.

Pour chaque instrument et condition d'enregistrement (entrées du tableau 6.4 (p. 139)), 30 morceaux sont choisis au hasard et enregistrés. Nous disposons ainsi d'un choix varié de morceaux, dont certains sont enregistrés plusieurs fois dans des conditions différentes. La nomenclature des fichiers suit ici le modèle

MAPS_MUS_nomMorceau_nomInstrument.wav

6.2.3 Dispositif

La génération de la base de données a fait l'objet de deux dispositifs différents, l'un pour l'utilisation du piano Disklavier, l'autre pour la synthèse logicielle. Dans les deux cas, nous avons dû procéder de manière non triviale et prendre des précautions qu'il nous a semblé utile de rapporter ici. Auparavant, tous les fichiers MIDI ont été créés, en prenant soin de garantir qu'ils pouvaient réellement être générés (les morceaux de musique contenaient par exemple quelques notes à supprimer car injouables, et le Disklavier était limité dans la rapidité avec laquelle il peut jouer automatiquement).

La génération à partir de logiciels a été automatisée en concaténant les nombreux fichiers MIDI à générer en un petit nombre de longs fichiers, et en lançant l'enregistrement à partir d'un séquenceur (Cubase SX 3 de Steinberg). Les fichiers son ainsi générés sont ensuite segmentés. Ce procédé a été utilisé faute de pouvoir contrôler le séquenceur par un script et enregistrer ainsi les fichiers un par un.

Nom	Fondamental	Renvers. 1	Renvers. 2	Renvers. 3	Renvers. 4
Septième majeure	0-4-7-11	0-3-7-8	0-4-5-9	0-1-5-8	
Septième mineure	0-3-7-10	0-4-7-9	0-3-5-8	0-2-5-9	
Septième de dominante	0-4-7-10	0-3-6-8	0-3-5-9	0-2-6-9	
Septième mineure et quinte diminuée	0-3-6-10	0-3-7-9	0-4-6-9	0-2-5-8	
Septième diminuée	0-3-6-9	0-3-6-9	0-3-6-9	0-3-6-9	
Septième majeure et parfait mineur	0-3-7-11	0-4-8-9	0-4-5-8	0-1-4-8	
Septième majeure et quinte augmentée	0-4-8-11	0-4-7-8	0-3-4-8	0-1-5-9	
Neuvième majeure de dominante	0-4-7-10-14	0-3-6-8-10	0-3-5-7-9	0-2-4-6-9	0-2-5-8-10
Neuvième mineure de dominante	0-4-7-10-13	0-3-6-8-9	0-3-5-6-9	0-2-3-6-9	0-3-6-9-11
Neuvième majeure et septième mineure	0-3-7-10-14	0-4-7-9-11	0-3-5-7-8	0-2-4-5-9	0-1-5-8-10
Neuvième mineure et septième mineure	0-3-7-10-13	0-4-7-9-10	0-3-5-6-8	0-2-3-5-9	0-2-6-9-11
Neuvième mineure et quinte diminuée	0-3-6-10-13	0-3-7-9-10	0-4-6-7-9	0-2-3-5-8	0-2-5-9-11
Neuvième majeure et septième majeure	0-4-7-11-14	0-3-7-8-10	0-4-5-7-9	0-1-3-5-8	0-2-5-9-10
Neuvième augmentée	0-4-7-11-15	0-3-7-8-11	0-4-5-8-9	0-1-4-5-8	0-1-4-8-9
Neuvième mineure et septième diminuée	0-3-6-9-13	0-3-6-9-10	0-3-6-7-9	0-3-4-6-9	0-2-5-8-11
Neuvième majeure, septième majeure et parfait mineur	0-3-7-11-14	0-4-8-9-11	0-4-5-7-8	0-1-3-4-8	0-1-5-9-10
Neuvième majeure et quinte augmentée	0-4-8-11-14	0-4-7-8-10	0-3-4-6-8	0-1-3-5-9	0-2-6-9-10

TABLE 6.6 – Accords usuels de 4 et 5 sons et nomenclature (écarts à la note fondamentale en demi-tons).

Figure 6.7 – Dispositif d'enregistrement : la carte son (en bas à droite) envoie les fichiers MIDI vers le boîtier du piano (en haut à droite), reçoit les notes jouées via une liaison MIDI inverse et enregistre le son produit (les micros sont ici placés près du piano).

Les précautions à prendre avec le Disklavier concernent essentiellement la synchronisation entre les signaux MIDI et les sons enregistrés. D'une part, il faut être capable d'assurer la synchronisation du début de l'enregistrement en contrôlant précisément l'intervalle entre l'envoi d'une instruction MIDI et la production du son. D'autre part, il faut également prendre en compte les éventuels décalages qui peuvent ensuite s'insérer du fait de la différence des horloges de référence pour le MIDI et le son. Par ailleurs, pour des raisons mécaniques, le fonctionnement optimal du Disklavier est obtenu en réglant le dispositif de sorte qu'un délai d'environ 500 ms soit inséré entre l'envoi d'une instruction MIDI et la réponse du piano. Cette contrainte vient s'ajouter à celle de la synchronisation du début de l'enregistrement déjà évoquée. La solution finalement adoptée est illustrée sur la figure 6.7 et consiste à :

1. utiliser trois types de liaisons : une liaison MIDI de commande du Disklavier à travers laquelle chaque fichier MIDI est envoyé ; une liaison MIDI en sens inverse permettant d'enregistrer ce que le Disklavier joue avec le retard évoqué précédemment ; enfin, la liaison audio d'enregistrement, censée être synchronisée avec la liaison MIDI d'enregistrement ; ce sont les enregistrements synchronisés provenant de ces deux dernières liaisons (MIDI et audio) qui constituent la base de données ;
2. centraliser l'ensemble des trois canaux sur un même ordinateur, pour bénéficier de la même horloge ; il convient même de n'utiliser qu'une seule carte son avec une entrée MIDI, une sortie MIDI et des entrées audio ;
3. vérifier les décalages possibles, malgré les deux premières précautions, soit sur les morceaux générés eux-mêmes, soit en utilisant les possibilités de suivi des différents événements, offerts par les langages de programmation, lors de l'enregistrement.

6.3 Évaluation des algorithmes

Les algorithmes développés pendant la thèse sont ici évalués de manière comparative. La base d'évaluation MAPS est utilisée pour l'estimation de fréquences fondamentales multiples et pour le système de transcription. Dans le premier cas, nous dresserons une évaluation objective détaillée. Dans le second, nous utiliserons à la fois une évaluation objective et subjective, en utilisant la F-mesure perceptive que nous avons décrite précédemment.

6.3.1 Estimation de fréquences fondamentales multiples

L'algorithme d'estimation de fréquences fondamentales multiples a été évalué sur les notes isolées, les accords aléatoires et les accords usuels de la base MAPS, c'est-à-dire sur les classes de son ISOL, RAND et UCHO. Chaque son est analysé sur une trame unique, de longueur 93 ms (4096 échantillons à 44,1 kHz ou 2048 à 22,1 kHz) extraite environ 10 ms après l'attaque. Deux autres algorithmes d'estimation de fréquences fondamentales multiples ont été testés à titre comparatif : celui de Tolonen et Karjalainen [2000], dans la version de la MIR Toolbox [Lartillot et Toiviainen, 2007], et celui de Klapuri [2006], dont le code a été fourni par l'auteur. Ce dernier système nous a été présenté par son auteur comme le plus performant parmi ceux qu'il a développés. La base de sons contient des notes comprises entre le Do 0 (33 Hz) et le Si 5 (1865 Hz), réparties de façon uniforme. Elle se compose de 9344 sons, de polyphonie comprise entre 1 et 6. Les performances de l'algorithme de Tolonen et Karjalainen [2000] se dégradant lorsque les fréquences fondamentales sont supérieures à 500 Hz (Do 4), nous avons testé l'algorithme dans une configuration

supplémentaire – notée Tolonen-500 –, sur les notes comprises entre le Do 0 (33 Hz) et le Si 3 (494 Hz) seulement.

Résultats généraux

La figure 6.8 représente les résultats de l'évaluation objective lorsque la polyphonie est inconnue par les systèmes. La F-mesure donne une évaluation globale des résultats. De ce point de vue, notre système se démarque des autres en polyphonie 1 et 2, avec un score de 94% contre 89% (polyphonie 1) et 92% (polyphonie 2) pour le système de Klapuri. La tendance s'inverse ensuite entre les deux systèmes, les F-mesures obtenues pour le système de Klapuri et le nôtre valant respectivement 92% et 89% en polyphonie 3, et 73% et 65% en polyphonie 6. Pour l'ensemble des résultats, nous constatons que le système de Tolonen, même limité en tessiture, est moins performant.

Quelle que soit la polyphonie, le score de précision est élevé pour chaque système – entre 85 et 97% – et particulièrement pour le nôtre, alors que le rappel a tendance à diminuer quand la polyphonie augmente. Il faut par ailleurs noter qu'il est courant qu'un système d'estimation de fréquences fondamentales multiples soit moins performant en polyphonie 1 qu'en polyphonie 2 ou 3 lorsqu'il doit estimer le nombre de notes car le risque d'ajouter des notes lorsqu'il n'y en a qu'une présente est grand.

Estimation de la polyphonie

Les capacités de la méthode de Klapuri et de notre algorithme à détecter le bon nombre de notes sont représentées sur la figure 6.9. Jusqu'à la polyphonie 5 incluse, les systèmes parviennent à déterminer le bon nombre de notes présentes plus souvent que tout autre nombre. Notre système a en outre été testé dans des conditions plus défavorables que les autres étant donné que la polyphonie 0, c'est-à-dire le silence, peut être détectée. On constate qu'il est détecté dans un minimum de cas.

Estimation à polyphonie connue

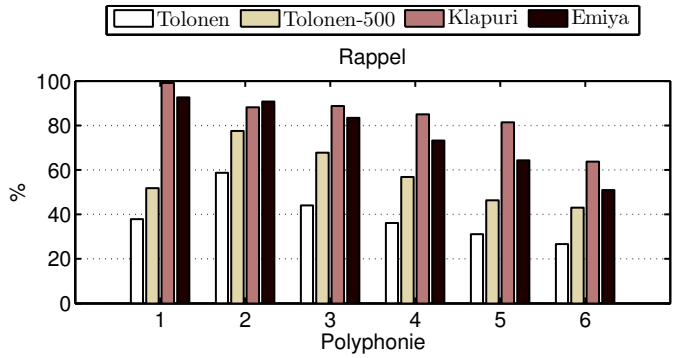
La figure 6.10 représente le taux de notes correctes lorsque la polyphonie est connue par les systèmes. Dans ce cas, le rappel, la précision et la F-mesure se valent et l'on parle de taux de notes correctes. On observe ici une petite dégradation des performances de notre système, qui arrive systématiquement en deuxième position après celui de Klapuri.

Accords aléatoires et accords usuels

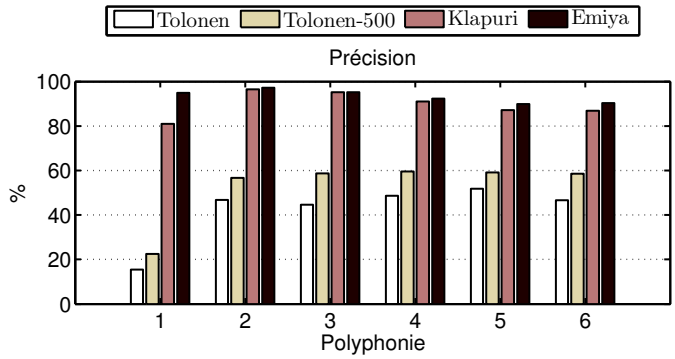
Les résultats obtenus sur les accords composés aléatoirement (classe RAND de la base MAPS) et sur les accords usuels (classe UCHO) sont représentés séparément sur la figure 6.11. Les résultats sont globalement meilleurs dans le cas d'accords usuels. Ils présentent certes un recouvrement spectral plus important que les accords aléatoires, mais ces derniers sont composés de notes qui peuvent être très éloignées sur le clavier, cause probable de la tendance observée.

Détection d'octave

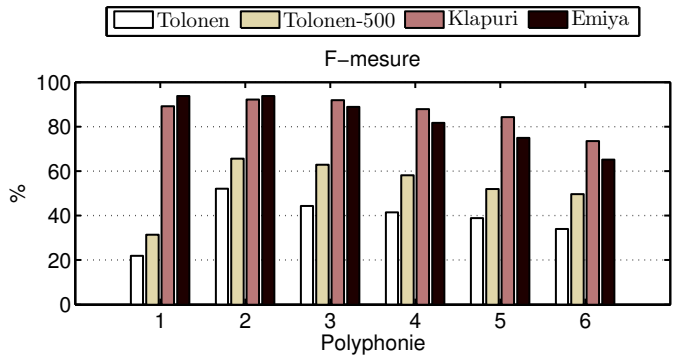
La figure 6.12 donne les résultats obtenus sur les sons de la base composés exclusivement d'une octave. Ce cas de figure fait partie des plus difficiles et nous voyons que les résultats sont moins élevés que ceux obtenus de manière générale en polyphonie 2. Notre système est ici le plus performant avec une F-mesure égale à 85%, contre 76% pour le système de Klapuri, et 77%/66% pour celui de Tolonen. Il semble donc que le modèle d'enveloppe spectrale et de recouvrement de spectre de notre algorithme soit particulièrement efficace.



(a)

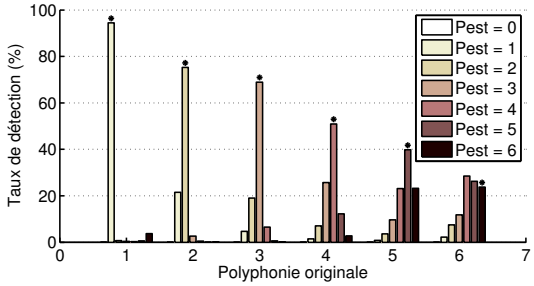


(b)

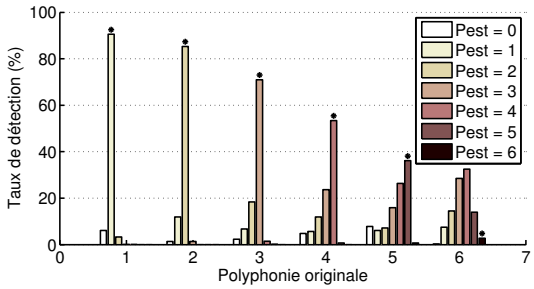


(c)

FIGURE 6.8 – Résultats de l'estimation de fréquences fondamentales multiples, la polyphonie étant inconnue. Le rappel (figure 6.8(a)), la précision (figure 6.8(b)) et la F-mesure (figure 6.8(c)) sont représentés pour chaque algorithme, en fonction de la polyphonie originale.



(a) Système de Klapuri.



(b) Système d'Emiya.

Figure 6.9 – Estimation de la polyphonie : la polyphonie estimée P_{est} est représentée en fonction de la polyphonie originale P, en proportion du nombre de sons de polyphonie P. Une astérisque (*) rappelle la polyphonie originale. Les résultats sont donnés pour le système de Klapuri (à gauche) et pour le nôtre (à droite).

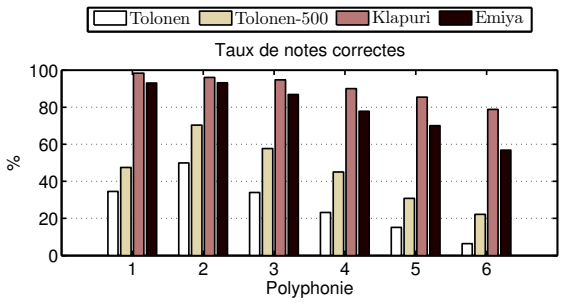
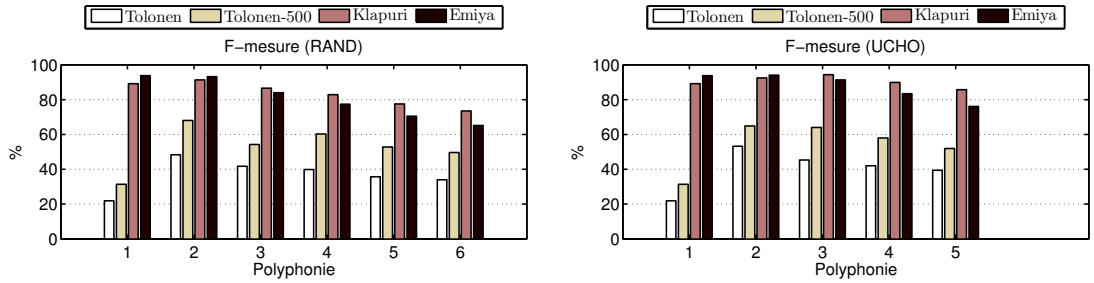


Figure 6.10 – Résultats de l'estimation de fréquences fondamentales multiples, la polyphonie étant connue. Le taux de notes correctes est représenté pour chaque algorithme, en fonction de la polyphonie originale.



(a) Accords aléatoires.

(b) Accords usuels.

FIGURE 6.11 – Performances en fonction de la consonance des accords : résultats pour des accords aléatoires (à gauche) et des accords usuels (à droite).

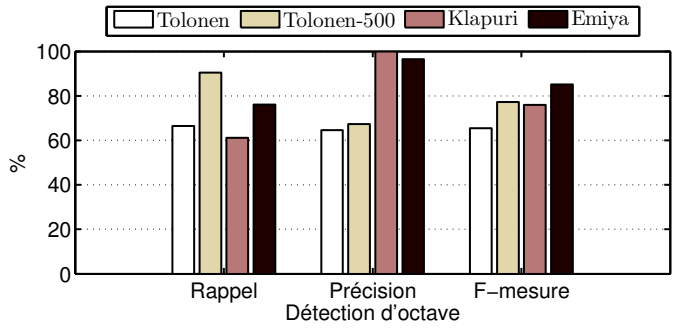


FIGURE 6.12 – Détection des octaves : détails des résultats pour les 90 sons d'octaves contenus dans la base de test (45 sons pour le système Tolonen-500).

Résultats en fonction des instruments et des conditions d'enregistrement

La figure 6.13 présente les résultats généraux par instrument, selon la nomenclature du tableau 6.4 (p. 139). Ces résultats sont intéressants sur plusieurs points. Tout d'abord, notre système a des performances relativement constantes en fonction de l'instrument et des conditions d'enregistrements utilisés, ce qui laisse penser que notre algorithme est robuste sur ces points. Par ailleurs, nous soulignons ici que les paramètres de notre algorithme ont été réglés sur une partie des sons provenant des pianos notés AkPnBsdf et AkPnCGdD. Les performances obtenues sur ces pianos et sur les autres étant similaires, nous avons pris le parti de les présenter ensemble, plutôt que de séparer la base de test et la base d'évaluation (démarche que nous aurions adoptée si les performances sur la base de test avaient été supérieures).

Par ailleurs, la principale variation visible sur ces figures est celle des performances du système de Klapuri en polyphonie 1. Nous avons constaté que la dégradation des résultats pour certains pianos était due à une baisse de la précision, le rappel étant stable. Cela voudrait indiquer que cet algorithme a parfois tendance à ajouter des notes.

Enfin, la différence de comportement des algorithmes face aux sons du piano réel (ENSTDk) par opposition aux sons synthétisés par des logiciels ne peut être vraiment établie. Certes, l'algorithme de Klapuri est un peu moins performant sur le piano réel, mais nous venons de voir que son comportement varie aussi parmi les différents pianos de synthèse. Par ailleurs, la qualité des sons de synthèse que nous avons utilisés nous laissent penser qu'ils sont très proches de sons réels, en particulier pour une application comme la transcription automatique.

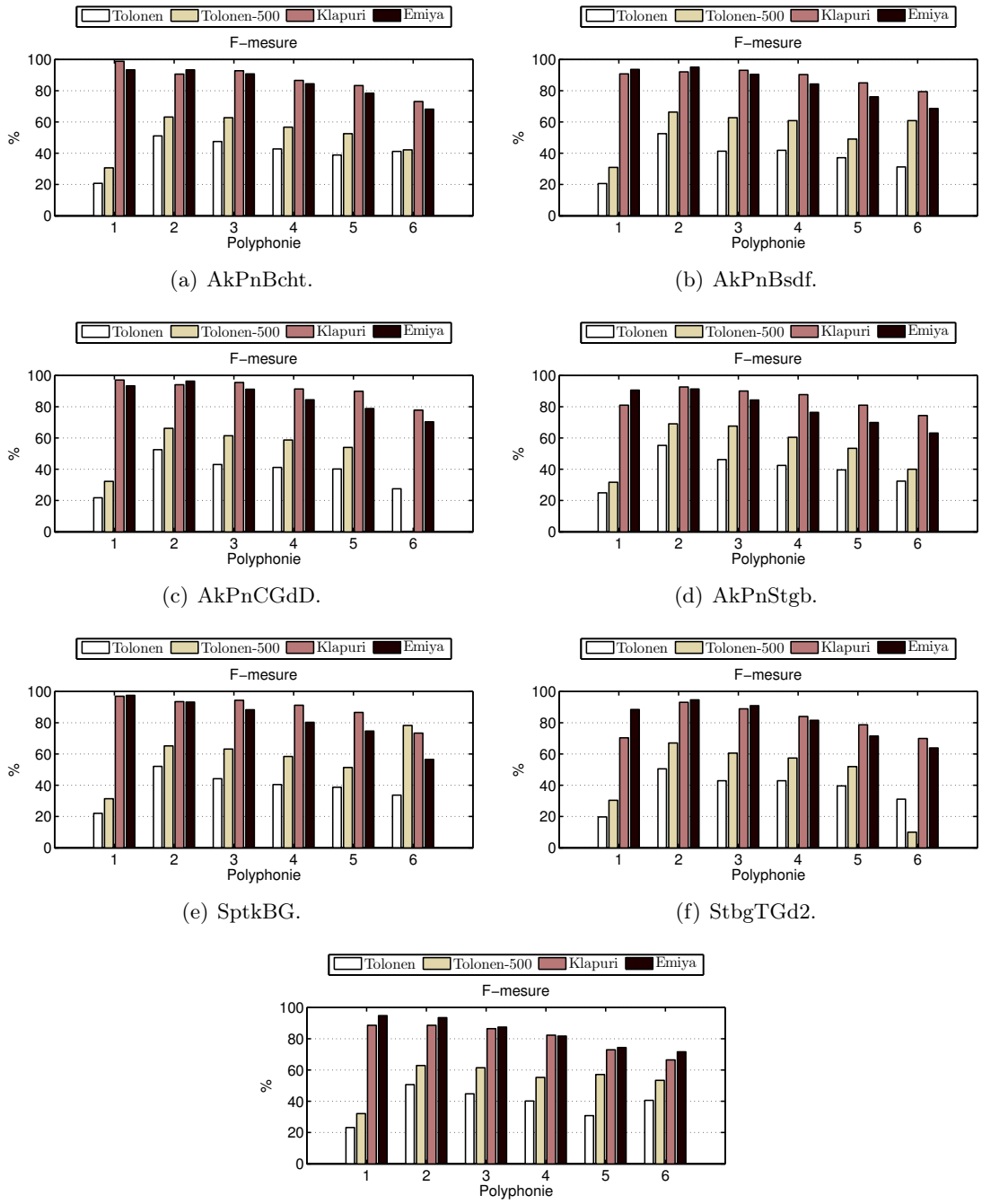
Performance de la sélection de notes candidates

Dans le cas spécifique de notre approche, il est important que la sélection de notes candidates soit performante puisque les notes qui sont éliminées à cette étape ne sont plus en mesure d'être choisies lors de l'estimation de fréquences fondamentales proprement dite. La figure 6.14 représente ces performances en fonction de la polyphonie et de la note originale. Seul le rappel y figure car l'objectif de la sélection de candidats est de ne pas oublier les bonnes notes, quitte à en ajouter des mauvaises.

Pour une polyphonie allant jusqu'à 3 ou 4, les performances sont très satisfaisantes : 99% des bonnes notes sont sélectionnées en polyphonie 1 et 2, 95% en polyphonie 3 et 87% en polyphonie 4. Les performances se dégradent un peu en polyphonie 5 et 6, avec un rappel valant respectivement 79% et 71%. Par ailleurs, les erreurs sont essentiellement commises dans le registre grave, en-deçà du La 1 (MIDI 45), et dans l'aigu au-delà du La 5 (MIDI 93). On peut ainsi supposer qu'une partie non négligeable des erreurs d'estimation en polyphonie élevée observée sur la figure 6.8(c) (p. 147) est due à l'étape de sélection de candidats, qu'il serait intéressant d'améliorer.

6.3.2 Système de transcription

Le système de transcription a été évalué sur les morceaux musicaux de la base MAPS (classe de sons MUS). Seules les trente premières secondes ont été extraites sur 215 morceaux. Plusieurs transcriptions de ces morceaux ont été obtenues, en utilisant, outre notre système, ceux de Bertin *et al.* [2007], les deux systèmes de Vincent *et al.* [2008] – une version B de base utilisant la décomposition en matrices non-négatives et une version H plus élaborée prenant en compte une contrainte d'harmonicité – et de Marolt [2004], les programmes ayant été fournis par leurs auteurs.



(a) AkPnBcht.

(b) AkPnBsdf.

(c) AkPnCGdD.

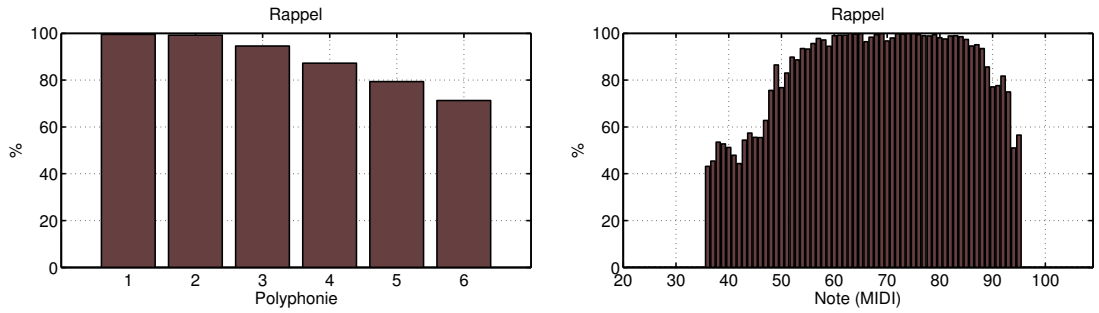
(d) AkPnStgb.

(e) SptkBG.

(f) StbgTGd2.

(g) ENSTDk.

FIGURE 6.13 – Performances en fonction des enregistrements : le détail de la nomenclature est donné dans le tableau 6.4 (p. 139).



(a) Rappel en fonction de la polyphonie.

(b) Rappel par note.

Figure 6.14 – Performances de la sélection de notes candidates : le rappel est tracé en fonction de la polyphonie (à gauche) et pour chaque note originale (à droite).

Les résultats sont représentés sur les figures 6.15 et 6.16(a). Pour chaque mesure utilisée, le score obtenu pour chaque morceau a été calculé, puis la moyenne des scores sur tous les morceaux est présentée.

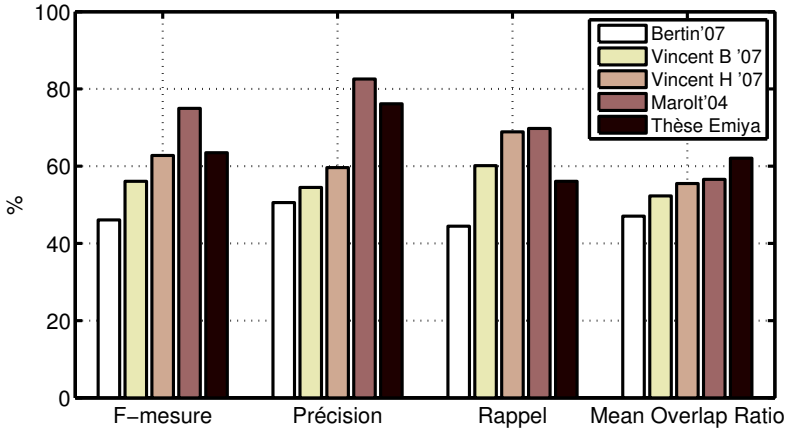


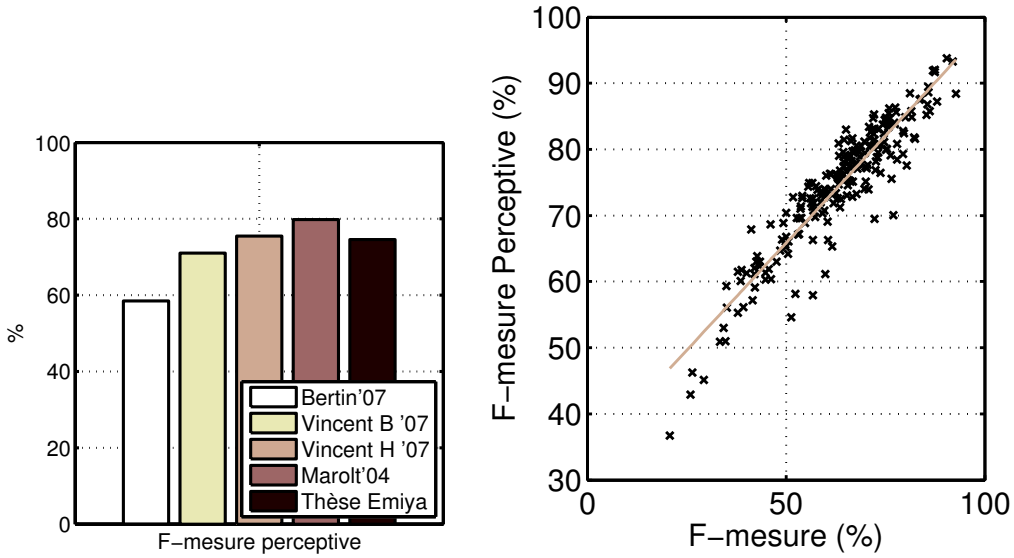
Figure 6.15 – Évaluation objective des transcriptions.

La F-mesure moyenne obtenue par notre système (figure 6.15) est égale à 63%. Elle arrive en deuxième position après celle de Marolt (75%), suivie de près par les systèmes de Vincent (56% et 63% respectivement pour les versions B et H), puis par celui de Bertin (46%). Notre système est plus perfomant en terme de précision (76%) que de rappel (56%), et suit de ce point de vue les tendances de la méthode d'estimation de fréquences fondamentales multiples. Il se distingue par ailleurs en arrivant en tête pour l'estimation de la durée des notes, le taux de recouvrement moyen (MOR) entre notes originales et transcrites atteignant 62%, contre 57%, 56%, 52% et 47% pour les autres systèmes. Cette performance illustre les qualités du système quant au suivi des mélanges de notes. L'architecture choisie, composée d'une segmentation selon les attaques et de HMM dans chaque segment, semble ici particulièrement efficace.

La F-mesure perceptive (figure 6.16(a)) donne des scores différents de la F-mesure, les valeurs ayant tendance à se rapprocher entre elles. Le système de Marolt reste en tête avec un score de 80%, suivi de la version H du système de Vincent et de notre système (75%), puis de la version B (71%) et du système de Bertin (58%). La figure 6.16(b) montre que

pour certains morceaux, la F-mesure perceptive et la F-mesure ne donnent pas du tout les mêmes résultats. On peut par exemple trouver un point de coordonnées $(41; 68)$ et un autre de coordonnées $(77; 70)$, qui sont relativement écartés de la courbe croissante moyenne.

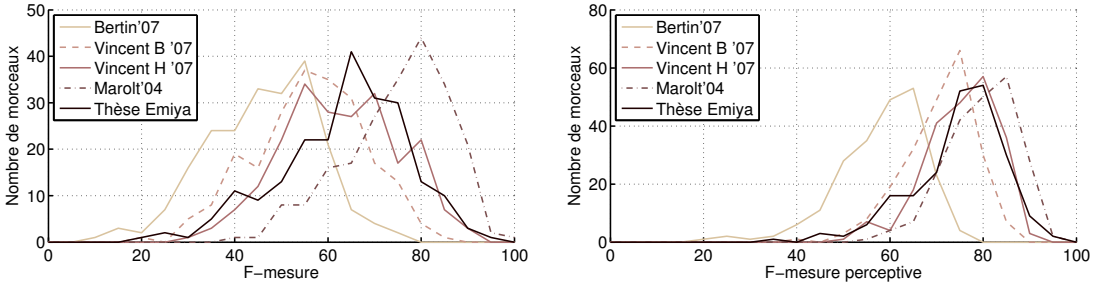
La figure 6.17 représente par ailleurs, pour chaque algorithme, la distribution des scores (F-mesure et F-mesure perceptive) selon les transcriptions. Ce sont ces valeurs qui ont été moyennées précédemment et l'on voit ici que les performances de chaque système dépend des morceaux originaux. Les différences obtenues pour un même système résultent le plus souvent des variations dans le niveau technique des morceaux, en particulier dans le niveau de polyphonie et suivant le tempo. Nous constatons par ailleurs de nouveau que la F-mesure perceptive (figure 6.17(b)) a davantage tendance à regrouper les scores que la F-mesure (figure 6.17(a)).



(a) F-mesure perceptive moyenne, par système.

(b) Correspondance entre F-mesure et F-mesure perceptive, pour chaque transcription obtenue (croix) par notre système. Le segment de droite est une régression linéaire entre les points, d'équation : F-mesure perceptive $= 0,65 \times$ F-mesure $+ 33$.

Figure 6.16 – Évaluation des transcriptions par la F-mesure perceptive.



(a) F-mesure.

(b) F-mesure perceptive.

Figure 6.17 – Distribution des résultats selon le morceau : en fonction de l'algorithme, le nombre de transcriptions est représentée en fonction de la F-mesure (à gauche) et la F-mesure perceptive (à droite), par pas de 5%.

6.4 Conclusion

Ce chapitre a été l'occasion d'aborder en détail plusieurs aspects liés à l'évaluation des transcriptions. Dans la première partie sur les méthodes d'évaluation, nous nous sommes intéressé aux critères d'évaluation, en introduisant une dimension perceptive qui n'a, à notre connaissance, jamais été utilisée dans le domaine de la transcription. Nous avons ainsi mis en évidence que d'un point de vue perceptif, les erreurs de transcription n'étaient pas toutes perçues avec la même sensibilité. Nous avons alors défini des mesures perceptives d'évaluation, qui généralisent les mesures utilisées habituellement. Enfin, nous avons introduit le cadre d'évaluation de la qualité utilisé en codage vidéo pour quantifier la qualité de ces métriques.

Dans la deuxième partie, nous avons décrit la base de données MAPS, construite de manière spécifique pendant cette thèse pour l'évaluation des systèmes de transcription automatique et d'estimation de fréquences fondamentales. Elle se compose de plusieurs classes de sons de piano entièrement et précisément annotés, que nous nous sommes efforcé de rendre les plus variés possible.

Dans la troisième et dernière partie, nous avons évalué notre algorithme d'estimation de fréquences fondamentales multiples et notre système de transcription en utilisant les outils précédents. Nous avons ainsi pu analyser sous plusieurs angles le comportement de ces méthodes, ainsi que d'approches de la littérature. Les résultats obtenus sont à la hauteur des systèmes les plus récents. Nous avons pu établir les qualités propres à chaque système. Ainsi, notre méthode d'estimation de fréquences fondamentales multiples est particulièrement efficace lorsque la polyphonie est inconnue, il est robuste aux changement de conditions d'enregistrement et présente des résultats très satisfaisant quant à la détection des octaves. Notre système de transcription offre des résultats satisfaisants d'une manière générale, avec une bonne estimation des durées des notes.

Conclusion et perspectives

Bilan de la thèse

Les travaux menés au cours de cette thèse ont apporté des éléments de réponse à des questions relatives à la transcription automatique de la musique de piano. Nous avons tout d'abord dégagé les enjeux liés à cette problématique en la situant par rapport aux domaines de recherche connexes – perception et estimation de la hauteur en général, estimation de fréquences fondamentales multiples, transcription automatique – et en soulignant les spécificités du piano. Il en est ressorti d'une part un besoin de caractérisation des sons de piano, et d'autre part des défis en matière de transcription automatique en général tels que la recherche de fréquences fondamentales sur une grande tessiture, la modélisation du recouvrement spectral, l'estimation du degré de polyphonie ou encore la question de l'évaluation de la qualité de la transcription. Suivant ces enjeux, la transcription automatique du piano a été abordée à plusieurs niveaux tels que la modélisation des sons, l'estimation des notes ou la manipulation des ensembles de notes obtenues.

Nous avons tout d'abord étudié la structure tonale des sons de piano et la paramétrisation associée (cf. chapitre 2). Après avoir décrit les outils appropriés d'analyse spectrale, nous avons mené une étude sur l'inharmonicité des sons et la localisation des fréquences des partiels d'une note grâce au couple de paramètres composé de la fréquence fondamentale et du coefficient d'inharmonicité. Nous avons en particulier proposé deux algorithmes d'estimation de ces paramètres et montré qu'ils parvenaient à localiser précisément les fréquences des partiels alors qu'une modélisation plus grossière – avec une inharmonicité moyenne ou nulle – ne le permettait pas au-delà d'un certain ordre de partiel. La quantification du gain d'une telle prise en compte de l'inharmonicité constitue le second résultat de cette étude. Nous avons ainsi montré que cette caractérisation fine de l'inharmonicité améliorait significativement la modélisation, sur la base d'un critère général tel que le rapport signal à bruit entre les sinusoïdes identifiées et le résiduel.

Autre aspect de la paramétrisation spectrale des sons de piano, la question des enveloppes spectrales et de leur modélisation pour la transcription a ensuite été abordée. Cette problématique nous a semblé particulièrement importante dans le cadre de l'estimation de fréquences fondamentales multiples en particulier pour lever les indéterminations telles que celle d'octave. Nous avons proposé un modèle autorégressif (AR) d'enveloppe spectrale qui reprend l'idée de *spectral smoothness* et lui confère un cadre plus formel que celui présenté dans la littérature. Nous avons utilisé le modèle de processus harmonique pour intégrer cette enveloppe spectrale dans un modèle de son. Par ailleurs, nous avons montré l'intérêt de modéliser le bruit résiduel par un processus à moyenne ajustée (MA).

Nous nous sommes ensuite intéressé (cf. chapitre 3) aux conditions difficiles d'estimation de fréquences fondamentales que constituent l'analyse de trames courtes et la contrainte d'une tessiture étendue. Nous avons montré que ces conditions faisaient chuter les perfor-

mances lorsqu'on utilise des techniques classiques pour l'estimation de hauteur telles que l'analyse de Fourier et les fonctions de détection élémentaires (autocorrélation, produit spectral). Nous avons proposé une solution paramétrique qui s'appuie sur une estimation sinusoïdale à haute résolution. La fonction de détection est ensuite construite de façon paramétrique. Elle offre des résultats très satisfaisants, en surpassant ceux de la littérature, en particulier aux extrêmités grave et aiguë de la tessiture.

Nous avons ensuite proposé une approche pour l'estimation de fréquences fondamentales multiples de sons de piano dans le chapitre 4. Elle intègre dans un cadre statistique le modèle de son et d'enveloppe spectrale présenté auparavant. L'étape d'estimation des paramètres aborde en particulier la question du recouvrement entre spectres de notes. Nous avons proposé un estimateur des amplitudes des partiels qui utilise l'information portée par les observations et les enveloppes spectrales pour estimer la contribution liée à chaque note. Les paramètres du modèle ayant été estimés, nous avons étudié leur intégration dans une fonction de détection de fréquences fondamentales multiples, et en particulier les difficultés liées à l'utilisation de la fonction de vraisemblance pour un modèle d'ordre variable. Nous avons alors proposé une solution approximative comme fonction d'estimation conjointe de fréquences fondamentales multiples et du degré de polyphonie. Les performances globales obtenues sont au niveau de l'état de l'art. Notre algorithme est particulièrement efficace lorsque la polyphonie est inconnue et il surpasse ses concurrents sur la question délicate de l'estimation d'octaves.

Ces résultats d'estimation de fréquences fondamentales multiples ont été intégrés dans un système de transcription automatique pour le piano (cf. chapitre 5). Nous avons considéré les spécificités de la problématique dans le cas du piano et avons proposé une solution adaptée, consistant à segmenter le signal selon les attaques et à suivre les mélanges de notes possibles dans les segments obtenus. Ce suivi prend la forme d'une estimation par modèles de Markov cachés (HMM) dont les états sont les mélanges potentiels. Le système est alors capable d'analyser un enregistrement monaural de piano et d'en estimer les notes jouées, avec des résultats à la hauteur de ceux de l'état de l'art. Le système s'est par ailleurs distingué par sa robustesse face aux variations de conditions d'enregistrement observées.

Le travail sur l'évaluation (chapitre 6) a été motivé par le besoin de grands volumes de sons correctement annotés, par la recherche de consensus sur les modalités d'évaluation, par les limites des critères habituellement utilisés et par la complexité de la question dans le cas de la transcription. Nous avons donc porté une attention particulière à la problématique de l'évaluation et avons proposé deux contributions distinctes. D'une part, nous avons montré que les erreurs de transcription pouvaient être classées dans plusieurs catégories d'erreurs typiques plus ou moins gênantes perceptivement. Par conséquent, l'évaluation des systèmes de transcription ne peut se limiter qu'en première approximation à un dénombrement des erreurs. Dans une démarche nouvelle au sein du champ de la transcription automatique, nous avons proposé des moyens de faire évoluer les métriques en incluant ces critères qualitatifs sur les erreurs et les poids perceptifs associés. Nous avons en outre importé le cadre d'évaluation de la qualité des métriques utilisé en particulier dans le domaine du codage vidéo pour l'appliquer à la question de l'évaluation des transcriptions. D'autre part, nous avons créé une base de données de sons de piano adaptée à l'évaluation des systèmes de transcription et d'estimation de fréquences fondamentales. Elle est entièrement annotée et son contenu est varié, à la fois vis à vis du type de sons enregistrés et des conditions d'enregistrement. Enfin, en exploitant cette base de sons et les outils d'évaluation présentés, nous avons fourni les résultats d'une évaluation comparative de plusieurs algorithmes, dont les nôtres, dans laquelle nous avons mis en évidence les qualités et faiblesses propres à

chaque système.

Perspectives

Les perspectives que nous envisageons au terme de ces travaux de thèse concernent à la fois l'exploitation des résultats proposés et la poursuite de nouvelles thématiques à la lumière des travaux effectués.

La modélisation des enveloppes spectrales en général et du piano en particulier reste une problématique déterminante pour la transcription automatique. Nous espérons que notre discussion à ce sujet, ainsi que le modèle proposé, contribueront à mieux la cerner. Les nombreuses approches déjà proposées – lois *a priori* sur les amplitudes, modèles de mélange de gaussiennes, *pattern matching* avec apprentissage des enveloppes, enveloppes moyennées par *spectral smoothness* ou modèles autorégressifs d'enveloppes – montrent à la fois l'enjeu et la difficulté sous-jacente. Nous sommes convaincus que la thématique demeurera essentielle dans les préoccupations à venir.

La modélisation statistique des signaux audio pour la transcription est actuellement une direction de recherche majeure. Elle offre un cadre théorique solide, laissant entrevoir des résultats prometteurs liés à une grande capacité de modélisation. Cependant, elle pose également des difficultés théoriques importantes, en particulier quant à la phase d'inférence. Nous pensons que ce genre d'approche ne fournit pour le moment pas forcément les résultats escomptés et qu'elles donnent souvent lieu à une mise en œuvre assez lourde, mais que ces difficultés – qui ne nous ont pas épargné – sont à la hauteur des enjeux. C'est pour cette raison que nous avons choisi cette direction, et nous pensons que si notre méthode d'estimation de fréquences fondamentales multiples présente des points forts quant au modèle et à l'estimation des paramètres proposés, la fonction d'estimation proprement dite souffre néanmoins de quelques faiblesses et que les enjeux théoriques à ce sujet demeurent importants.

Après les nombreux travaux sur l'estimation de fréquences fondamentales simples proposés au cours des dernières décennies, il nous semble particulièrement important de nous intéresser aujourd'hui à la robustesse des méthodes. La robustesse est souvent assimilée à la sensibilité vis à vis du rapport signal à bruit et a été étudiée en ces termes dans la littérature, mais elle fait également référence aux facteurs que nous avons étudiés – taille de la fenêtre d'analyse, tessiture et compromis temps-fréquence –, ainsi qu'à d'autres paramètres tels que la qualité vocale et sa grande variabilité. Les performances de notre estimateur de fréquences fondamentales simples nous encouragent à poursuivre ces travaux dans cette direction et à essayer de les généraliser aux autres instruments, et surtout à la parole.

La question de l'évaluation suscite de nombreux efforts dans la communauté, notamment dans le cadre d'évaluations indépendantes telles que MIREX. Nos travaux sur les mesures d'évaluation nous ont montré qu'il n'était pas trivial de concevoir un système d'évaluation fidèle qui puisse donner des résultats proches d'un jugement humain. Seules quelques erreurs typiques ont pu être prises en compte, alors qu'une évaluation subjective fait appel à des notions de plus haut niveau telles que la tonalité. Aussi, nous pensons qu'il serait tout à fait profitable de développer les nombreuses pistes et travaux que nous avons pu aborder avec A. Daniel à ce sujet. Il conviendrait en particulier de mener de nouveaux travaux sur les critères de perception de la qualité d'une transcription, en prenant en compte des notions telles que la tonalité, le rythme ou la mélodie.

Notre système de transcription présente des qualités qu'il nous paraît important de souligner et laisse par ailleurs d'autres directions à approfondir. L'utilisation d'une étape

de détection d'attaque nous semble une bonne approche, que ce soit pour le piano ou pour d'autres instruments. D'une part, les algorithmes de détection d'attaques sont aujourd'hui relativement performants, alors qu'il n'est pas évident de détecter le début des notes à partir d'une seule analyse de fréquences fondamentales multiples sur des trames successives. D'autre part, la dimension rythmique nous semble sous-exploitée dans les systèmes de transcription – peut-être en raison de l'obsession d'une bonne estimation de fréquences fondamentales – alors qu'elle occupe une place de premier plan dans notre perception de la musique. L'intégration de modèles rythmiques élaborés laisse alors entrevoir des perspectives d'amélioration des systèmes de transcription. Nous avons également proposé un cadre de modèles de Markov cachés pour l'utilisation d'une méthode d'estimation jointe de fréquences fondamentales. Ce cadre nous a permis de modéliser l'évolution locale du mélange de notes présentes. Il pourrait être enrichi avec un modèle « musicologique » tel que ceux déjà proposés dans la littérature (cf. partie 1.4.5 (p. 39)). Nous suggérons alors que cette information sur le contenu tonal soit introduite au niveau des probabilités initiales des HMM, à la différence des approches déjà proposées dans lesquelles l'information se situe au niveau de la matrice de transition. De cette façon, l'aspect tonal est exploité au niveau des notes, c'est-à-dire sur une échelle de temps plus grande indépendante de la longueur de la trame d'analyse, alors que les transitions à l'intérieur des HMM continuent à prendre en charge l'enchaînement entre trames.

ANNEXES

Annexe A

Méthodes de traitement du signal numérique

A.1 Notes de probabilités

A.1.1 Variables et vecteurs gaussiens

Définition A.1.1 (Variable aléatoire gaussienne centrée réduite)**.** *Une variable aléatoire réelle X est dite* ***gaussienne centrée réduite*** *si elle admet pour densité de probabilité par rapport à la mesure de Lebesgue la fonction*

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{A.1}$$

Définition A.1.2 (Variable aléatoire gaussienne)**.** *Une variable aléatoire réelle X est dite* ***gaussienne*** *s'il existe μ et σ tels que $X = \sigma X_c + \mu$, où X_c est gaussienne centrée réduite. On le note alors $X \sim \mathcal{N}(\mu, \sigma^2)$ et l'on a*

$$\mathbb{E}[X] = \mu \tag{A.2}$$

$$Var[X] = \sigma^2 \tag{A.3}$$

Propriété A.1.3. *Si $\sigma \neq 0$, la densité de probabilité de X est*

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{A.4}$$

Définition A.1.4 (vecteur aléatoire gaussien)**.** *Le vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_n)$ est dit* ***gaussien*** *si toute combinaison linéaire $a^t X$ de ses composantes, où $a \in \mathbb{R}^n$, est une variable aléatoire gaussienne. On le note $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma)$, avec*

$$\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu} \tag{A.5}$$

$$\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^t\right] = \Gamma \tag{A.6}$$

Propriété A.1.5. *Si $\det \Gamma \neq 0$, $\boldsymbol{X}$ admet pour densité la fonction*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Gamma}} e^{-\frac{1}{2}(x-\boldsymbol{\mu})^t \Gamma^{-1} (x-\mu)} \tag{A.7}$$

Propriété A.1.6. *Soit $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Gamma)$ un vecteur gaussien de taille n, A une matrice $m \times n$ et $\boldsymbol{b} \in \mathbb{R}^m$. $\boldsymbol{Y} \triangleq A\boldsymbol{X} + \boldsymbol{b}$ est un vecteur gaussien et l'on a*

$$\boldsymbol{Y} \sim \mathcal{N}\left(A\boldsymbol{\mu} + \boldsymbol{b}, A\Gamma A^t\right) \tag{A.8}$$

Définition A.1.7 (Variable aléatoire gaussienne complexe)**.** *Une variable aléatoire complexe Z est dite **gaussienne complexe** si ses parties réelle X et imaginaire Y sont des variables aléatoires réelles gaussiennes indépendantes de même variance $\frac{\sigma^2}{2}$. On le note alors $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$ avec $\mu \triangleq \mathbb{E}[X] + i\mathbb{E}[Y]$.*

A.1.2 Changement de variable

Propriété A.1.8. *Soit X une variable aléatoire multivariée admettant une densité p_X et $Y = f(X)$, où f est une fonction bijective. Alors Y admet pour densité de probabilité la fonction*

$$p_Y(y) = p_X\left(f^{-1}(y)\right)\left|Jac[f]\left(f^{-1}(y)\right)\right|^{-1} \tag{A.9}$$

A.1.3 Autres lois de probabilité

A.1.3.1 Loi Gamma

Définition A.1.9 (Loi Gamma)**.** *Une variable aléatoire X réelle positive suit une **loi Gamma de paramètre de forme** $k \in \mathbb{R}_+$ **et de paramètre d'échelle** $E \in \mathbb{R}_+$ si elle admet une densité de probabilité de la forme*

$$p_X(x) = \frac{1}{E^\alpha \Gamma(k)} x^{k-1} e^{-\frac{x}{E}} \tag{A.10}$$

où $z \mapsto \Gamma(z)$ désigne la fonction Gamma d'Euler. On le note alors : $X \sim \Gamma(k, E)$. Les densités associées à plusieurs valeurs de paramètres sont représentées sur la figure A.1(a).

Propriété A.1.10 (Moyenne et variance)**.** *Si $X \sim \Gamma(k, E)$, alors*

$$\mathbb{E}[X] = kE \tag{A.11}$$

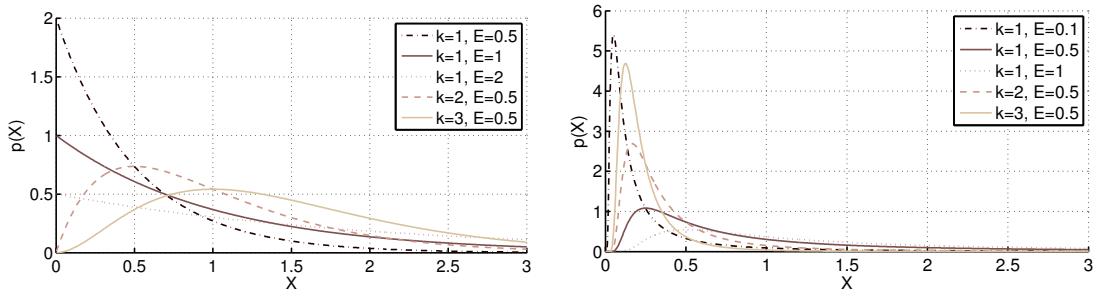
$$Var[X] = kE^2 \tag{A.12}$$

A.1.3.2 Loi Gamma Inverse

Définition A.1.11 (Loi Gamma Inverse)**.** *Une variable aléatoire X réelle positive suit une **loi Gamma Inverse de paramètre de forme** $k \in \mathbb{R}_+$ **et de paramètre d'échelle** $E \in \mathbb{R}_+$ si elle admet une densité de probabilité de la forme*

$$p_X(x) = \frac{E^k}{\Gamma(k)} x^{-k-1} e^{-\frac{E}{x}} \tag{A.13}$$

On le note alors : $X \sim \mathcal{IG}(k, E)$. Les densités associées à plusieurs valeurs de paramètres sont représentées sur la figure A.1.



(a) Famille de lois Gamma.

(b) Famille de lois Gamma Inverse.

FIGURE A.1 – Densités des lois Gamma (gauche) et Gamma Inverse (droite) pour quelques valeurs de paramètres de forme k et d'échelle e.

Propriété A.1.12 (Moyenne et variance)**.** *Soit* $X \sim \mathcal{IG}(k, e)$*. On a*

$$\mathbb{E}[X] = \frac{E}{k-1} \text{ si } k > 1 \tag{A.14}$$

$$Var[X] = \frac{E^2}{(k-1)^2 (k-2)} \text{ si } k > 2 \tag{A.15}$$

Propriété A.1.13 (Inverse)**.** *Si* $X \sim \mathcal{IG}(k, E)$*, alors* $\frac{1}{X} \sim \Gamma\left(k, \frac{1}{E}\right)$

A.2 Modélisation AR et MA

Nous rappelons ici quelques résultats sur les processus autorégressifs (AR) et à moyenne ajustée (MA). Le cas général du processus autorégressif à moyenne ajustée (ARMA) n'est pas traité car il n'intervient pas dans ces travaux de thèse.

A.2.1 Processus AR

Définition A.2.1 (processus AR)**.** $\{X_n\}_{n\in\mathbb{Z}}$ *est un* ***processus AR*** *d'ordre* p *s'il est stationnaire au second ordre et s'il est solution de l'équation*

$$X_n = \sum_{k=1}^{p} a_k X_{n-k} + W_n \tag{A.16}$$

où W_n *est un bruit blanc de variance* σ^2*. Même si ce n'est pas le cas le plus général, on supposera de plus que* W_n *est gaussien, en particulier pour considérer la densité de probabilité.*

On notera $A(z)$ la quantité [1]

$$A(z) \triangleq 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{A.17}$$

1. Dans la littérature consacrée à la modélisation AR et MA, différentes notations sont utilisées : ainsi $A(z)$ (équation (A.17)) a la forme d'une transformée en z, mais peut prendre celle d'un polynôme, auquel cas l'exposant de la variable z est k au lieu de $-k$; de même, l'équation (A.16) est présentée comme une formule de filtrage récursif, alors que l'on trouve également des présentations dans lesquelles les a_k ont des signes opposés, pour $k \geq 1$.

Propriété A.2.2 (Existence)**.** *Nous admettons le résultat suivant : l'équation* (A.16) *admet une solution stationnaire au second ordre si et seulement si $A(z)$ n'a pas de racine sur le cercle unité. La solution est alors unique et s'exprime en fonction des coefficients h_k du développement en série de Laurent de $\frac{1}{A(z)}$ au voisinage du cercle unité :*

$$X_n = \sum_{k=-\infty}^{+\infty} h_k W_{n-k} \tag{A.18}$$

On notera $H(z)$ la fonction

$$H(z) \triangleq \sum_{k=-\infty}^{+\infty} h_k z^{-k} \tag{A.19}$$

Propriété A.2.3 (Densité spectrale de puissance)**.** *Le processus X_n possède une densité spectrale de puissance dont l'expression est*

$$\Gamma(f) = \frac{\sigma^2}{\left|1 - \sum_{k=1}^{p} a_k e^{-i2\pi fk}\right|^2} \qquad = \frac{\sigma^2}{\left|A\left(e^{2i\pi f}\right)\right|^2} \tag{A.20}$$

Pour étudier la causalité d'un processus AR, nous introduisons le minimum et le maximum des modules des racines de $A(z)$:

$$\rho_m \triangleq \min\left\{|z| / A(z) = 0\right\} \tag{A.21}$$

$$\rho_M \triangleq \max\left\{|z| / A(z) = 0\right\} \tag{A.22}$$

Propriété A.2.4 (Causalité)**.** *Si $\rho_M < 1$, les pôles sont à l'intérieur du cercle unité et $H(z)$ est analytique sur la couronne ouverte $\{z/\,|z| > \rho_M\}$. En considérant la limite de $H\left(z^{-1}\right)$ en 0, on obtient alors $\forall k < 0, h_k = 0$: le processus est causal.*

De même, si $\rho_m > 1$, les racines sont à l'extérieur du cercle unité et $H(z)$ est analytique sur le disque ouvert $\{z/\,|z| < \rho_m\}$, on a alors $\forall k > 0, h_k = 0$, et le processus est anticausal.

Dans les autres cas, les racines sont de part et d'autre du cercle unité et le processus n'est ni causal, ni anticausal.

Propriété A.2.5 (Équations de Yule-Walker)**.** *On suppose que X_n est causal. Les équations de Yule-Walker relient la fonction d'autocovariance $\gamma(k) \triangleq \mathbb{E}\left[X_n X_{n-k}\right]$ de X_n et les paramètres $a_1, \ldots, a_P, \sigma^2$ du modèle AR sous la forme matricielle suivante :*

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(p) \\ \gamma(1) & \gamma(0) & \ddots & \gamma(p-1) \\ \vdots & \ddots & \ddots & \vdots \\ \gamma(p) & \gamma(p-1) & \ldots & \gamma(0) \end{pmatrix} \begin{pmatrix} 1 \\ -a_1 \\ \vdots \\ -a_P \end{pmatrix} = \begin{pmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{A.23}$$

Nombre de méthodes permettent d'estimer les paramètres d'un modèle AR à partir de l'observation d'une réalisation du processus. La plus connue consiste à substituer l'autocovariance empirique à la fonction d'autocovariance dans l'équation (A.23), puis à inverser le système linéaire obtenu. La matrice à inverser étant de Toeplitz, l'algorithme de Levinson-Durbin offre une résolution récursive avec une complexité en $\mathcal{O}\left(p^2\right)$ (contre $\mathcal{O}\left(p^3\right)$ pour l'inversion d'une matrice dans le cas général).

Densité de probabilité

D'après la définition A.2.1, le processus AR X_n s'exprime linéairement en fonction du processus gaussien W_n. C'est donc un processus gaussien et la log-densité du vecteur gaussien $X = (X_1, \ldots, X_n)^{\mathrm{t}}$ est de la forme

$$\ln p\,(X) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2}\ln\det\Gamma - \frac{1}{2}\frac{X^{\mathrm{t}}\Gamma^{-1}X}{\sigma^2} \tag{A.24}$$

où $\Gamma\sigma^2$ est la matrice de covariance de X. Il est cependant difficile de déterminer Γ car X_n s'exprime en fonction d'un nombre infini de termes W_m (cf. équation (A.18)), il n'est donc pas possible d'utiliser un simple changement de variable (cf. propriété A.1.6). Une première solution (proposition A.2.6) consiste à s'intéresser à la densité de probabilité conditionnelle de $(X_{p+1}, \ldots, X_n | X_1, \ldots, X_p)$. La deuxième (proposition A.2.8) consiste à modifier les hypothèses en supposant une relation de filtrage circulaire entre X_n et W_n. D'autres considérations sur le calcul de cette densité peuvent être trouvées dans les travaux d' Ansley [1979] par exemple.

Proposition A.2.6 (Densité conditionnelle)**.** *La log-densité de probabilité conditionnelle de* $(X_{p+1}, \ldots, X_n | X_1, \ldots, X_p)$ *est*

$$\ln p\,(\underline{X}|X_0) = -\frac{n-p}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|A\underline{X} + A_0X_0\right\|^2 \tag{A.25}$$

avec

$$\underline{X} \triangleq \begin{pmatrix} X_n \\ \vdots \\ X_{p+1} \end{pmatrix}, \qquad X_0 \triangleq \begin{pmatrix} X_p \\ \vdots \\ X_1 \end{pmatrix},$$

$$A \triangleq \begin{pmatrix} 1 & -a_1 & \ldots & -a_P & 0 & \ldots & 0 \\ 0 & 1 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & -a_P \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & -a_1 \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & 1 \end{pmatrix}, \qquad A_0 \triangleq \begin{pmatrix} 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & & \vdots \\ -a_P & \ddots & \vdots \\ \vdots & \ddots & 0 \\ -a_1 & \ldots & -a_P \end{pmatrix}$$

De plus, l'équation (A.25) *se réécrit sous la forme*

$$\ln p\,(\underline{X}|X_0) = -\frac{n-p}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|\underline{X} - \boldsymbol{X}\,\vec{a}\right\|^2 \tag{A.26}$$

avec

$$\boldsymbol{X} \triangleq \begin{pmatrix} X_{n-1} & \ldots & X_{n-p} \\ \vdots & & \vdots \\ X_p & & X_1 \end{pmatrix}, \qquad \vec{a} \triangleq \begin{pmatrix} a_1 \\ \vdots \\ a_P \end{pmatrix} \tag{A.27}$$

Démonstration. D'après (A.16) et en reprenant les notations précédentes, on a

$$W = A\underline{X} + A_0 X_0 \qquad \text{avec} \qquad W \triangleq \begin{pmatrix} W_n \\ \vdots \\ W_{p+1} \end{pmatrix}, \tag{A.28}$$

On a alors $\underline{X} = A^{-1}W - A^{-1}A_0X_0$. D'après A.1.6, $W \sim \mathcal{N}\left(0, \boldsymbol{I}\sigma^2\right)$ donc

$$\underline{X}|X_0 \sim \mathcal{N}\left(-A^{-1}A_0X_0, A^{-1}\left(A^{-1}\right)^{\mathrm{t}}\sigma^2\right) \tag{A.29}$$

Et

$$\ln p\left(X|X_0\right) = -\frac{n-p}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|AX + A_0X_0\right\|^2 \tag{A.30}$$

□

Corollaire A.2.7 (Estimation des paramètres du processus AR par maximum de vraisemblance)**.** *Au sens du maximum de vraisemblance, l'estimation des coefficients* $\vec{a}$ *a pour solution l'estimateur des moindres carrés* $\widehat{\vec{a}} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\underline{X}$ *et la variance estimée est* $\widehat{\sigma^2} = \frac{1}{n-p}\left\|\underline{X} - \boldsymbol{X}\,\widehat{\vec{a}}\right\|^2$.

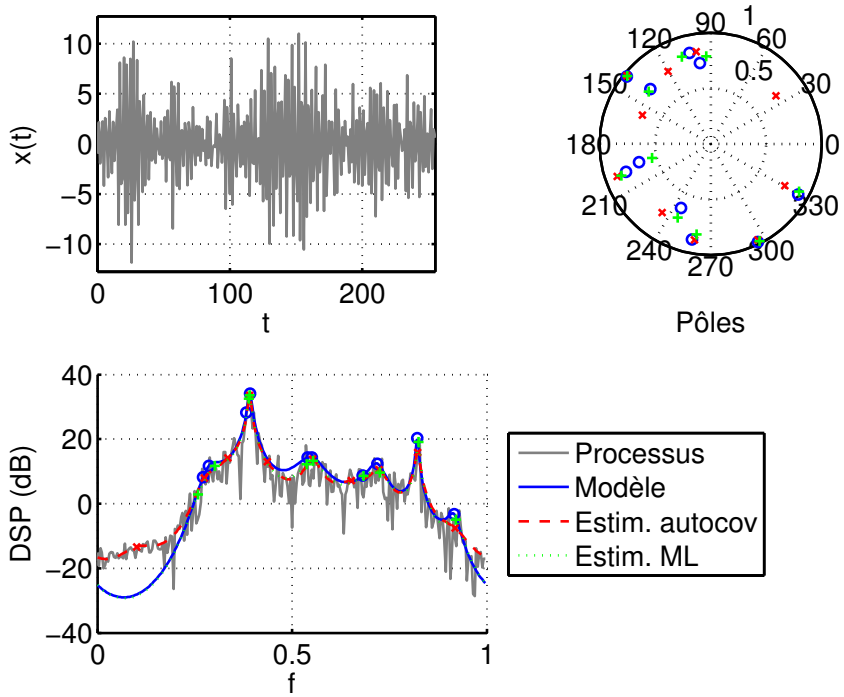


Figure A.2 – Estimation des paramètres AR : les paramètres (en bleu) sont estimés à partir du processus généré (en gris), par la fonction d'autocovariance (en rouge) et par maximum de vraisemblance (en vert). L'écart entre le modèle et la réalisation en bas à droite est dû au fenêtrage du signal et aux lobes secondaires que celui-ci provoque.

L'estimation par maximum de vraisemblance donne de meilleurs résultats que la méthode par la fonction d'autocovariance (cf. figure A.2). À noter qu'elle ne garantit pas que le modèle estimé est causal, et qu'il faut donc inverser le module des pôles situés à l'extérieur du cercle unité pour rendre la solution causale.

Proposition A.2.8 (Processus AR circulaire)**.** *On redéfinit ici le processus AR en supposant une relation de filtrage circulaire entre le vecteur* $W = (W_1, \ldots, W_n)^t$ *gaussien centré de covariance* $I_n\sigma^2$ *et le vecteur* $X = (X_1, \ldots, X_n)^t$ *:*

$$W = (1, -a_1, \ldots, -a_P) \otimes X \tag{A.31}$$

L'expression de la log-densité de X *est alors :*

$$\ln p\,(X) = -\frac{n}{2}\ln 2\pi\sigma^2 + \frac{1}{2}\sum_{k=0}^{n-1}\ln\left|A\left(e^{2i\pi\frac{k}{n}}\right)\right|^2 - \frac{\|(1, -a_1, \ldots, -a_P) \otimes X\|^2}{2\sigma^2} \tag{A.32}$$

ou de manière équivalente,

$$\ln p\,(X) = -\frac{n}{2}\ln 2\pi\sigma^2 + \frac{1}{2}\sum_{k=0}^{n-1}\ln\left|A\left(e^{2i\pi\frac{k}{n}}\right)\right|^2 - \frac{1}{2n\sigma^2}\sum_{k=0}^{n-1}\left|\frac{\mathcal{F}_X\left(\frac{k}{n}\right)}{1/A\left(e^{2i\pi\frac{k}{n}}\right)}\right|^2 \tag{A.33}$$

où $\mathcal{F}_X(f)$ *est la transformée de Fourier discrète de* X *en* f *et* $A(z)$ *est la transformée en* Z *du filtre de réponse impulsionnelle* $(1, -a_1, \ldots, -a_P)$*.*

Démonstration. La relation de filtrage s'écrit sous forme matricielle :

$$W = A_{\text{circ}}X \tag{A.34}$$

où A_{circ} est la matrice circulante

$$A_{\text{circ}} \triangleq \begin{pmatrix} 1 & 0 & \ldots & 0 & -a_P & \ldots & -a_1 \\ -a_1 & 1 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & -a_P \\ -a_P & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \ldots & 0 & -a_P & \ldots & -a_1 & 1 \end{pmatrix} \tag{A.35}$$

W étant un vecteur gaussien de loi $\mathcal{N}\left(0, \boldsymbol{I}\sigma^2\right)$, X est gaussien de loi $\mathcal{N}\left(0, A_{\text{circ}}^{-1}\left(A_{\text{circ}}^{-1}\right)^\dagger\sigma^2\right)$. A_{circ} étant circulante, on a

$$\det\left(A_{\text{circ}}A_{\text{circ}}^\dagger\right) = \prod_{k=0}^{n-1}\left|A\left(e^{2i\pi\frac{k}{n}}\right)\right|^2 \tag{A.36}$$

puis, en utilisant le fait que $X^\dagger\left(A_{\text{circ}}^{-1}\left(A_{\text{circ}}^{-1}\right)^\dagger\right)^{-1}X = \|A_{\text{circ}}X\|^2$, la densité de X s'écrit

$$p(X) = \frac{1}{\sqrt{(2\pi)^n \det\left(A_{\text{circ}}^{-1}\left(A_{\text{circ}}^{-1}\right)^\dagger \sigma^2\right)}} e^{-\frac{1}{2}X^\dagger\left(A_{\text{circ}}^{-1}\left(A_{\text{circ}}^{-1}\right)^\dagger\sigma^2\right)^{-1}X} \tag{A.37}$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \left(\prod_{k=0}^{n-1}\left|A\left(e^{2i\pi\frac{k}{n}}\right)\right|^2\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|A_{\text{circ}}X\|^2} \tag{A.38}$$

L'équation (A.33) s'obtient ensuite en utilisant l'identité de Parseval. □

Cette dernière méthode n'est pas standard. Nous l'avons employée dans [Emiya *et al.*, 2007a] et l'introduisons ici pour simplifier l'estimation par maximum de vraisemblance. Qualitativement, les deux densités (A.26) et (A.32) sont proches. Leur premier terme est identique à taille d'échantillon égale, ainsi que leur dernier terme si l'on considère qu'il correspond à $\frac{\|W_1,\ldots,W_n\|^2}{2\sigma^2}$. La seule différence qualitative est le deuxième terme de (A.32), absent dans (A.26).

A.2.2 Processus MA

Définition A.2.9 (processus MA)**.** $\{X_n\}_{n\in\mathbb{Z}}$ *est un* ***processus MA*** *d'ordre q s'il s'écrit*

$$X_n = \sum_{k=0}^{q} b_k W_{n-k} \tag{A.39}$$

où W_n est un bruit blanc de variance σ^2 et $b_0 = 1$. Comme dans le cas AR, on supposera de plus que W_n est gaussien.

On notera $B(z)$ la transformée en z du filtre unitaire associé

$$B(z) \triangleq \sum_{k=0}^{q} b_k z^{-k} \tag{A.40}$$

Propriété A.2.10 (Propriétés du second ordre)**.** *Le processus MA X_n est du second ordre, centré, et d'autocovariance*

$$\gamma(m) = \mathbb{E}\left[X_n X_{n-m}\right] \tag{A.41}$$

$$= \sigma^2 \sum_{k=0}^{n-|m|} b_k b_{k+|m|} \tag{A.42}$$

avec, en particulier, $\gamma(m) = 0$ pour $|m| > q$.

Propriété A.2.11 (Densité spectrale de puissance)**.** *Le processus X_n possède une densité spectrale de puissance dont l'expression est*

$$\Gamma(f) = \sigma^2 \left|\sum_{k=0}^{q} b_k e^{-i2\pi fk}\right|^2 \tag{A.43}$$

Propriété A.2.12 (Densité de probabilité)**.** *D'après la définition A.2.9, le processus MA X_n s'exprime linéairement en fonction du processus gaussien W_n. La log-densité du vecteur gaussien $X = (X_1, \ldots, X_n)^t$ est de la forme*

$$\ln p\,(X) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2}\ln\det\Gamma - \frac{1}{2\sigma^2}X^t\Gamma^{-1}X \tag{A.44}$$

où $\Gamma\sigma^2$ est la matrice de covariance de X, dont l'élément (i,j) est égal à $\gamma\,(|i-j|)$ (cf. propriété A.1.6).

Le calcul du déterminant et de l'inverse de Γ est cependant coûteux et l'on peut vouloir simplifier l'expression de ces calculs en introduisant une approximation. Deux solutions sont proposées : le calcul d'une vraisemblance approchée et, comme dans le cas AR, la redéfinition du processus MA par filtrage circulaire.

Proposition A.2.13 (Vraisemblance approchée)**.** *Le vecteur $X = (X_1, \ldots, X_n)^t$ s'écrit*

$$X = B_+ \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} + B_- \begin{pmatrix} W_0 \\ \vdots \\ W_{-q+1} \end{pmatrix} \tag{A.45}$$

avec

$$B_+ \triangleq \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ b_1 & 1 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ b_q & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \ldots & 0 & b_q & \ldots & b_1 & 1 \end{pmatrix} \quad \textit{et} \quad B_- \triangleq \begin{pmatrix} b_1 & \ldots & b_q \\ \vdots & \iddots & 0 \\ b_q & \iddots & \vdots \\ 0 & \iddots & \vdots \\ \vdots & \iddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix} \tag{A.46}$$

En négligeant le second terme qui fait intervenir les W_n pour $n \leq 0$ dans le membre de droite de (A.45)*, la densité de probabilité de X devient :*

$$\ln p\,(X) \approx -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}X^t\Theta X \tag{A.47}$$

avec

$$\Theta \triangleq \left(B_+^{-1}\right)^t B_+^{-1} \tag{A.48}$$

Proposition A.2.14 (Processus MA circulaire)**.** *On redéfinit ici le processus MA en supposant une relation de filtrage circulaire entre le vecteur $W = (W_1, \ldots, W_n)^t$ gaussien centrée de covariance $I_n\sigma^2$ et le vecteur $X = (X_1, \ldots, X_n)^t$:*

$$X = (1, b_1, \ldots, b_q) \otimes W \tag{A.49}$$

L'expression de la log-densité de X est alors :

$$\ln p(X) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\sum_{k=0}^{n-1}\ln\left|B\left(e^{2i\pi\frac{k}{n}}\right)\right|^2 - \frac{1}{2n\sigma^2}\sum_{k=0}^{n-1}\left|\frac{\mathcal{F}_X\left(\frac{k}{n}\right)}{B\left(e^{2i\pi\frac{k}{n}}\right)}\right|^2 \tag{A.50}$$

où $\mathcal{F}_X(f)$ est la transformée de Fourier discrète de X en f et B est la transformée en z du filtre de réponse impulsionnelle $(1, b_1, \ldots, b_q)$.

Démonstration. La démonstration est similaire à celle de la proposition A.2.8. □

Comme dans le cas AR, cette méthode n'est pas standard et a été utilisée dans [Emiya *et al.*, 2007a]. Alors que les approximations introduites sont asymptotiquement peu gênantes (lorsque $n >> q$), le gain de complexité est significatif. Dans le cas de la proposition A.2.13, l'inversion de matrice est moins coûteuse car la matrice est triangulaire, et pour le cas A.2.14, il n'y a plus de matrice à inverser mais uniquement une transformée de Fourier discrète à calculer. Par ailleurs, ces simplifications peuvent également être utiles pour estimer les paramètres MA au sens du maximum de vraisemblance, avec une vraisemblance approximative.

A.3 Approximation de Laplace

L'approximation de Laplace [Chickering et Heckerman, 1997] permet de calculer approximativement une intégrale en fonction des propriétés du second ordre de l'intégrande au niveau de son maximum, en approximant la fonction à intégrer par une gaussienne.

Soit une fonction f définie sur $\mathbb{R}^n$ à valeurs dans $\mathbb{R}_+^*$, admettant un maximum en $x_0 \in \mathbb{R}^n$ et dont le logarithme est deux fois différentiable en x_0. Le développement de Taylor de $\log f$ à l'ordre 2 en x_0 est

$$\log f(x) = \log f(x_0) - \frac{1}{2}(x - x_0)^{\mathrm{t}}\boldsymbol{H}(x - x_0) + o\left((x - x_0)^2\right) \tag{A.51}$$

avec

$$\boldsymbol{H} \triangleq -\boldsymbol{\nabla^2}\{\log f\}(x_0) \tag{A.52}$$

où $\boldsymbol{\nabla^2} f$ désigne la matrice hessienne de la fonction f.

On peut alors approximer f sous la forme $f \approx \alpha g$ où α est une constante multiplicative et g une gaussienne centrée en x_0 et de matrice de covariance $\boldsymbol{H}^{-1}$. On a

$$\int f(x)\,dx \approx \alpha \int g(x)\,dx \tag{A.53}$$

$$\approx \alpha \tag{A.54}$$

$$\approx \frac{f(x_0)}{g(x_0)} \tag{A.55}$$

$$\approx f(x_0)(2\pi)^{\frac{n}{2}}(\det \boldsymbol{H})^{-\frac{1}{2}} \tag{A.56}$$

La validité de l'approximation de Laplace dépend bien évidemment de la fonction à intégrer et de la précision que l'on souhaite obtenir. Dans bien des situations, l'approximation obtenue est grossière, mais peut s'avérer néanmoins utile (c'est le cas en particulier dans nos travaux). Pour plus de détails sur la validité de cette approximation et son usage, on pourra se référer aux nombreux travaux publiés sur le sujet, par exemple ceux de MacKay [1998].

Annexe B

Preuves mathématiques

B.1 État de l'art

Démonstration de l'équation (1.8) *p. 28.* La maximisation de la log-vraisemblance par rapport à σ_w^2 implique

$$\frac{d}{d\sigma_w^2} L\left(\mathbf{x}|\mathbf{a}, \boldsymbol{\varphi}, f_0, \sigma_w^2\right) = 0 \tag{B.1}$$

et donne la valeur optimale pour ce paramètre :

$$\widehat{\sigma_w^2} = \frac{1}{T} \sum_{t=0}^{T-1} \left(x\left(t\right) - s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t)\right)^2 \tag{B.2}$$

qui est insérée dans (1.7) :

$$L\left(\mathbf{x}|\mathbf{a}, \boldsymbol{\varphi}, f_0\right) \triangleq L\left(\mathbf{x}|\mathbf{a}, \boldsymbol{\varphi}, f_0, \widehat{\sigma_w^2}\right) \tag{B.3}$$

$$= -\frac{T}{2} \log \frac{2\pi e \sum_{t=0}^{T-1} \left(x\left(t\right) - s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t)\right)^2}{T} \tag{B.4}$$

Maximiser $L\left(\mathbf{x}|\mathbf{a}, \boldsymbol{\varphi}, f_0\right)$ par rapport à $(\mathbf{a}, \boldsymbol{\varphi}, f_0)$ revient alors à minimiser

$$\epsilon_X\left(\mathbf{a}, \boldsymbol{\varphi}, f_0\right) \triangleq \sum_{t=0}^{T-1} \left(x\left(t\right) - s_{f_0,\mathbf{a},\boldsymbol{\varphi}}(t)\right)^2 \tag{B.5}$$

qui devient, en utilisant l'identité de Parseval,

$$\epsilon_X\left(\mathbf{a}, \boldsymbol{\varphi}, f_0\right) = \frac{1}{T} \sum_{k=0}^{T-1} \left| X\left(\frac{k}{T}\right) - S_{f_0,\mathbf{a},\boldsymbol{\varphi}}\left(\frac{k}{T}\right) \right|^2 \tag{B.6}$$

où X et $S_{f_0,\mathbf{a},\boldsymbol{\varphi}}$ sont les transformées de Fourier discrètes de x et $s_{f_0,\mathbf{a},\boldsymbol{\varphi}}$. La minimisation par rapport à $(\mathbf{a}, \boldsymbol{\varphi})$ s'obtient en constatant que

$$S_{f_0,\mathbf{a},\boldsymbol{\varphi}}(f) = \begin{cases} a_h e^{2i\pi h f_0 + i\varphi_h} & \text{si } f = h f_0, h \in h \in [\![1; H]\!] \\ a_h e^{-2i\pi h f_0 - i\varphi_h} & \text{si } f = -h f_0, h \in [\![1; H]\!] \\ 0 & \text{sinon} \end{cases} \tag{B.7}$$

car $s_{f_0,\mathbf{a},\boldsymbol{\varphi}}$ est un signal de fréquence fondamentale f_0 et $f_0 T \in \mathbb{N}$. Les valeurs optimales $\widehat{\mathbf{a}} \triangleq (\widehat{a_1}, \ldots, \widehat{a_H})$ et $\widehat{\boldsymbol{\varphi}} \triangleq (\widehat{\varphi_h}, \ldots, \widehat{\varphi_h})$ de $(\mathbf{a}, \boldsymbol{\varphi})$ sont donc

$$\widehat{a_h} = |X(hf_0)| \tag{B.8}$$
$$\widehat{\varphi_h} = \angle X(hf_0) \tag{B.9}$$

pour $h \in [\![1; H]\!]$. L'expression à minimiser, par rapport à f_0, devient alors

$$\epsilon_X(\widehat{\mathbf{a}}, \widehat{\boldsymbol{\varphi}}, f_0) = \sum_{\substack{k=0 \\ \frac{k}{T} \neq h f_0}}^{T-1} \left| X\left(\frac{k}{T}\right) \right|^2 = \sum_{k=0}^{T-1} \left| X\left(\frac{k}{T}\right) \right|^2 - 2\sum_{h=1}^{H} |X(hf_0)|^2 \tag{B.10}$$

Le premier terme de cette somme étant constant par rapport à f_0, la solution est donnée en maximisant la somme spectrale présente dans le second terme. Il convient de noter que l'on a affaire à des modèles emboîtés, dans la mesure où toute fréquence sous-multiple de la vraie fréquence fondamentale est solution du problème, et qu'il faut donc choisir la plus grande. □

B.2 Estimation de fréquences fondamentales multiples

Démonstration des équations (4.44) *et* (4.45) *p. 95.* On cherche un estimateur de la forme

$$\hat{\alpha}_{k_0} \triangleq \eta X(f_{k_0}) \tag{B.11}$$

avec $\eta \in \mathbb{C}$. L'erreur quadratique moyenne est alors

$$\epsilon_{k_0}(\eta) = \mathbb{E}\left[|\alpha_{k_0} - \eta X(f_{k_0})|^2 \right] \tag{B.12}$$

La valeur optimale $\hat{\eta}$ vérifie la condition d'optimalité $\frac{d\epsilon_{k_0}}{d\eta}(\hat{\eta}) = 0$, qui est équivalente à la décorrélation entre l'erreur $(\alpha_{k_0} - \hat{\eta} X(f_{k_0}))$ et la donnée $X(f_{k_0})$, d'où

$$\begin{aligned} 0 &= \mathbb{E}\left[\left(\alpha_{k_0}^* - \hat{\eta}^* X^*(f_{k_0})\right) X(f_{k_0}) \right] \\ &= \mathbb{E}\left[\alpha_{k_0}^* X(f_{k_0})\right] - \hat{\eta}^* \mathbb{E}\left[|X(f_{k_0})|^2 \right] \end{aligned} \tag{B.13}$$

On a donc un résultat qui présente des analogies avec celui obtenu dans le cas du filtrage de Wiener :

$$\hat{\eta} = \frac{\mathbb{E}\left[\alpha_{k_0} X^*(f_{k_0})\right]}{\mathbb{E}\left[|X(f_{k_0})|^2 \right]} \tag{B.14}$$

Dans le cas présent, on a

$$\begin{aligned} \mathbb{E}\left[\alpha_{k_0} X^*(f_{k_0})\right] &= \sum_{k=1}^{K} W^*(f_{k_0} - f_k)\, \mathbb{E}\left[\alpha_{k_0} \alpha_k^*\right] \\ &= W^*(0)\, \sigma_{k_0}^2 \end{aligned} \tag{B.15}$$

et

$$\begin{aligned}\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right] &= \sum_{k=1}^{K}\sum_{l=1}^{K} W\left(f_{k_0}-f_k\right) W^*\left(f_{k_0}-f_l\right) \mathbb{E}\left[\alpha_k \alpha_l^*\right] \\ &= \sum_{k=1}^{K} \left|W\left(f_{k_0}-f_k\right)\right|^2 \sigma_k^2 \end{aligned} \tag{B.16}$$

On a donc prouvé (4.44). L'erreur associée est

$$\begin{aligned}\epsilon_{k_0}\left(\hat{\eta}\right) &= \mathbb{E}\left[\left|\alpha_{k_0}\right|^2\right] + \left|\hat{\eta}\right|^2 \mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right] - \hat{\eta}\mathbb{E}\left[\alpha_{k_0}^* X\left(f_{k_0}\right)\right] - \hat{\eta}^*\mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right] \\ &= \mathbb{E}\left[\left|\alpha_{k_0}\right|^2\right] - \frac{\left|\mathbb{E}\left[\alpha_{k_0} X^*\left(f_{k_0}\right)\right]\right|^2}{\mathbb{E}\left[\left|X\left(f_{k_0}\right)\right|^2\right]} \\ &= \sigma_{k_0}^2 - \frac{\left|W\left(0\right)\right|^2 \sigma_{k_0}^4}{\sum_{k=1}^{K}\left|W\left(f_{k_0}-f_k\right)\right|^2 \sigma_k^2} \end{aligned} \tag{B.17}$$

□

Annexe C

Correspondance entre notes, fréquences fondamentales et échelle MIDI

Le tableau C.1 donne la correspondance entre les notes (nom et numéro d'octave conventionnels), fréquences fondamentales selon un tempérament égal avec un La 3 à 440 Hz et codes MIDI associés (entre 0 et 127). Seules les notes de la tessiture standard du piano sont représentées. Les formules de conversion sont les suivantes :

$$\text{Code MIDI} = \left\lfloor 12 \log_2 \frac{F_0}{440} \right\rceil + 69 \tag{C.1}$$

$$F_0 = 2^{\frac{\text{Code MIDI}-69}{12}} \times 440 \tag{C.2}$$

$$\text{N}^\text{o}\text{ octave} = \left\lfloor \frac{\text{Code MIDI} - 60}{12} + 3 \right\rfloor \tag{C.3}$$

Remarque : ce tableau est donné à titre général. Dans le cas particulier du piano, il ne s'applique pas exactement, l'instrument n'étant habituellement pas accordé selon un tempérament égal. En raison de l'inharmonicité des cordes, l'accordeur doit en effet « étirer les octaves » pour qu'il n'y ait pas de battements entre les partiels de deux notes à l'octave. Il en résulte des fréquences fondamentales plus élevées (resp. plus basses) qu'avec le tempérament égal pour les notes aigues (resp. graves).

N° octave	**Do** (C)	**Do#** (C#)	**Ré** (D)	**Mi♭** (E♭)	**Mi** (E)	**Fa** (F)	**Fa#** (F#)	**Sol** (G)	**Sol#** (G#)	**La** (A)	**Si♭** (B♭)	**Si** (B)	
−1 (0)					20,6	21,8	23,1	24,5	26,0	27,5	29,1	30,9	F_0 (Hz)
					16	17	18	19	20	21	22	23	Code MIDI
0 (1)	32,7	34,6	36,7	38,9	41,2	43,7	46,2	49,0	51,9	55,0	58,3	61,7	F_0 (Hz)
	24	25	26	27	28	29	30	31	32	33	34	35	Code MIDI
1 (2)	65,4	69,3	73,4	77,8	82,4	87,3	92,5	98,0	104	110	117	123	F_0 (Hz)
	36	37	38	39	40	41	42	43	44	45	46	47	Code MIDI
2 (3)	131	139	147	156	165	175	185	196	208	220	233	247	F_0 (Hz)
	48	49	50	51	52	53	54	55	56	57	58	59	Code MIDI
3 (4)	262	277	294	311	330	349	370	392	415	440	466	494	F_0 (Hz)
	60	61	62	63	64	65	66	67	68	69	70	71	Code MIDI
4 (5)	523	554	587	622	659	698	740	784	831	880	932	988	F_0 (Hz)
	72	73	74	75	76	77	78	79	80	81	82	83	Code MIDI
5 (6)	1047	1109	1175	1245	1319	1397	1480	1568	1661	1760	1865	1976	F_0 (Hz)
	84	85	86	87	88	89	90	91	92	93	94	95	Code MIDI
6 (7)	2093	2217	2349	2489	2637	2794	2960	3136	3322	3520	3729	3951	F_0 (Hz)
	96	97	98	99	100	101	102	103	104	105	106	107	Code MIDI
7 (8)	4186												F_0 (Hz)
	108												Code MIDI

Table C.1 – Correspondance entre notes, fréquences fondamentales (F_0) et échelle MIDI, sur la tessiture standard du piano. Les noms de notes et la numérotation des octaves entre parenthèses correspondent à la notation anglo-saxone.

Bibliographie

M. Abe et J. Smith : AM/FM rate estimation for time-varying sinusoidal modeling. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 201–204, Philadelphia, PA, USA, Mars 2005.

M. Alonso, G. Richard et B. David : Extracting note onsets from musical recordings. *Proc. of the ICME*, Amsterdam, The Netherlands, Juil. 2005.

M. Alonso, G. Richard et B. David : Accurate tempo estimation based on harmonic plus noise decomposition. *EURASIP Journal on Advances in Signal Processing*, 2007 (1), p. 161–174, 2007.

C. Ansley : An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66 (1), p. 59–65, 1979.

B. Atal et L. Rabiner : A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24 (3), p. 201–212, 1976.

B. Atal et S. Hanauer : Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *The Journal of the Acoustical Society of America*, 50 (2B), p. 637–655, 1971.

R. Badeau : Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées. Application aux signaux de musique. Thèse de doctorat, *École Nationale Supérieure des Télécommunications*, France, 2005.

R. Badeau et B. David : Weighted maximum likelihood autoregressive and moving average spectrum modeling. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 3761–3764, Las Vegas, NV, USA, Mars-Avr. 2008.

R. Badeau, B. David et G. Richard : A new perturbation analysis for signal enumeration in rotational invariance techniques. *IEEE Transactions on Signal Processing*, 54 (2), p. 450–458, Fév. 2006.

B. Bank : Physics-based Sound Synthesis of the Piano. Master's thesis, Helsinki Univ. of Technology, 2000.

B. Bank et L. Sujbert : On the nonlinear commuted synthesis of the piano. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Hamburg, Germany, Sept. 2002.

B. Bank et L. Sujbert : Generation of longitudinal vibrations in piano strings : From physics to sound synthesis. *The Journal of the Acoustical Society of America*, 117 (4), p. 2268–2278, 2005.

I. Barbancho, A. Barbancho, A. Jurado et L. Tardón : Transcription of piano recordings. *Applied Acoustics*, 65 (12), p. 1261–1287, Déc. 2004.

J. Bello : Towards the automated analysis of simple polyphonic music : A knowledge-based approach. Thèse de doctorat, *Univ. of London*, England, 2003.

J. Bello, L. Daudet et M. Sandler : Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (6), p. 2242–2251, Nov. 2006.

J. Bensa : Analysis and Synthesis of Piano Sounds using Physical and Signal Models. Thèse de doctorat, *Univ. de la Mediterranée*, France, 2003.

N. Bertin, R. Badeau et G. Richard : Blind signal decompositions for automatic transcription of polyphonic music : NMF and K-SVD on the benchmark. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 65–68, Honolulu, HI, USA, Avr. 2007.

E. Bigand, F. Madurell, B. Tillmann et M. Pineau : Effect of global structure and temporal organization on chord processing. *Journal of experimental psychology : Human perception and performance*, 25 (1), p. 184–197, 1999.

R. Bradley : Some Statistical Methods in Taste Testing and Quality Evaluation. *Biometrics*, 9 (1), p. 22–38, 1953.

J. C. Brown : Musical fundamental frequency tracking using a pattern recognition method. *The Journal of the Acoustical Society of America*, 92 (3), p. 1394–1402, 1992.

A. Camacho : SWIPE : a sawtooth waveform inspired pitch estimator for speech and music. Thèse de doctorat, *Univ. of Florida*, USA, 2007.

A. Cemgil : Bayesian Music Transcription. Thèse de doctorat, *SNN, Radboud Univ. Nijmegen*, The Netherlands, 2004.

A. Cemgil, H. Kappen et D. Barber : A generative model for music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (2), p. 679–694, 2006.

Y. Cheng et D. O'Shaughnessy : Automatic and reliable estimation of glottal closure instant and period. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37 (12), p. 1805–1815, 1989.

D. M. Chickering et D. Heckerman : Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*, 29 (2), p. 181–212, 1997.

H. A. Conklin Jr. : Design and tone in the mechanoacoustic piano. Part I. Piano hammers and tonal effects. *The Journal of the Acoustical Society of America*, 99 (6), p. 3286–3296, 1996a.

H. A. Conklin Jr. : Design and tone in the mechanoacoustic piano. Part III. Piano strings and scale design. *The Journal of the Acoustical Society of America*, 100 (3), p. 1286–1298, 1996b.

H. A. Conklin Jr. : Piano strings and "phantom" partials. *The Journal of the Acoustical Society of America*, 102 (1), p. 659, 1997.

A. Cont, S. Dubnov et D. Wessel : Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Bordeaux, France, Sept. 2007.

C. Creusere, K. Kallakuri et R. Vanam : An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. *IEEE Transactions on Audio, Speech and Language Processing*, 16 (1), p. 129–136, Janv. 2008.

C. d'Alessandro, V. Darsinos et B. Yegnanarayana : Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, 6 (1), p. 12–23, Janv. 1998a.

C. d'Alessandro, S. Rosset et J. Rossi : The pitch of short-duration fundamental frequency glissandos. *The Journal of the Acoustical Society of America*, 104 (4), p. 2339–2348, 1998b.

A. Daniel, V. Emiya et B. David : Perceptually-based evaluation of the errors usually made when automatically transcribing music. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, Sept. 2008.

L. Daudet et B. Torrésani : Sparse adaptive representations for musical signals. A. Klapuri et M. Davy, éds : *Signal Processing Methods for Music Transcription*. Springer, 2006.

B. David, R. Badeau, N. Bertin, V. Emiya et G. Richard : Multipitch detection for piano music : Benchmarking a few approaches. *The Journal of the Acoustical Society of America*, 122 (5), p. 2962–2962, 2007.

B. David, V. Emiya, R. Badeau et Y. Grenier : Harmonic plus noise decomposition : Time-frequency reassignment versus a subspace-based method. *120th Convention of the Audio Engineering Society*, Paris, France, Mai 2006.

M. Davy, S. Godsill et J. Idier : Bayesian analysis of polyphonic western tonal music. *The Journal of the Acoustical Society of America*, 119 (4), p. 2498–2517, 2006.

M. Davy et S. Godsill : Bayesian harmonic models for musical signal analysis. *Proc. of Bayesian Statistics 7*, Valencia, Spain, Juin 2002. Oxford University Press.

A. de Cheveigné : Separation of concurrent harmonic sounds : Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *The Journal of the Acoustical Society of America*, 93 (6), p. 3271–3290, 1993.

A. de Cheveigné : Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, 103 (3), p. 1261–1271, 1998.

A. de Cheveigné et H. Kawahara : Multiple period estimation and pitch perception model. *Speech Communication*, 27 (3-4), p. 175–185, 1999.

A. de Cheveigné et H. Kawahara : YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111 (4), p. 1917–1930, 2002.

A. Dempster, N. Laird et D. Rubin : Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), p. 1–38, 1977.

S. Dixon : On the computer recognition of solo piano music. *Australasian Computer Music Conference*, Brisbane, Australia, 2000.

A. Doucet et X. Wang : Monte carlo methods for signal processing : a review in the statistical signal processing context. *IEEE Signal Processing Magazine*, 22 (6), p. 152–170, 2005.

B. Doval et X. Rodet : Estimation of fundamental frequency of musical sound signals. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 3657–3660, Toronto, Ontario, Canada, Avr. 1991.

Z. Duan, Y. Zhang, C. Zhang et Z. Shi : Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 16 (4), p. 766–778, 2008.

B. Efron et R. J. Tibshirani : *An Introduction to the Bootstrap*. Chapman & Hall, London, 1993.

A. El-Jaroudi et J. Makhoul : Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39 (2), p. 411–423, 1991.

D. Ellis et D. Rosenthal : Mid-level representations for computational auditory scene analysis. *International Joint Conference on Artificial Intelligence - Workshop on Computational Auditory Scene Analysis*, Montreal, Quebec, Août 1995.

D. Ellis et G. Poliner : Classification-based melody transcription. *Machine Learning*, 65 (2), p. 439–456, 2006.

V. Emiya, R. Badeau et B. David : Multipitch estimation of inharmonic sounds in colored noise. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, p. 93–98, Bordeaux, France, Sept. 2007a.

V. Emiya, R. Badeau et B. David : Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. *Proc. of the European Conference on Signal Processing (EUSIPCO)*, Lausanne, Switzerland, Août 2008.

V. Emiya, B. David et R. Badeau : A parametric method for pitch estimation of piano tones. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 249–252, Honolulu, HI, USA, Avr. 2007b.

A. Eronen et A. Klapuri : Musical instrument recognition using cepstral coefficients and temporal features. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 753–756, Istanbul, Turkey, Juin 2000.

S. Essid, G. Richard et B. David : Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (1), p. 68–80, 2006.

P. Flandrin : *Temps-Fréquence*. Hermès, 1993.

H. Fletcher, E. D. Blackham et R. Stratton : Quality of piano tones. *The Journal of the Acoustical Society of America*, 34 (6), p. 749–761, 1962.

N. H. Fletcher et T. D. Rossing : *The Physics of Musical Instruments*. Springer, 1998.

M. Gainza et E. Coyle : Automating ornamentation transcription. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 69–72, Honolulu, HI, USA, Avr. 2007.

A. Galembo et A. Askenfelt : Measuring inharmonicity through pitch. *Speech Transmission Lab. Quart. Rep., Dept. Speech Commun. Music Acoust., Royal Inst. Technol., Stockholm, Sweden,*, STL-QPSR 1/1994, p. 135–144, 1994.

A. Galembo et A. Askenfelt : Signal representation and estimation of spectral parameters by inharmonic comb filters with application to the piano. *IEEE Transactions on Speech and Audio Processing*, 7 (2), p. 197–203, 1999.

A. Gelman, H. Stern et D. Rubin : *Bayesian data analysis*. Chapman and Hall/CRC, 2004.

O. Gillet : Transcription des signaux percussifs. Application à l'analyse de scènes musicales audiovisuelles. Thèse de doctorat, *École Nationale Supérieure des Télécommunications*, France, 2007.

S. Godsill et M. Davy : Bayesian computational models for inharmonicity in musical instruments. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2005.

D. Godsmark : A computational model of the perceptual organisation of polyphonic music. Thèse de doctorat, *Univ. of Sheffield*, England, 1998.

E. Gómez : Tonal description of polyphonic audio for music content processing. *INFORMS J. on Computing*, 18 (3), p. 294–304, 2006.

M. Goto : A real-time music-scene-description system : predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43 (4), p. 311–329, 2004.

M. Goto, H. Hashiguchi, T. Nishimura et R. Oka : RWC music database : Music genre database and musical instrument sound database. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, Baltimore, MD, USA, Oct. 2003.

M. Goto et Y. Muraoka : A beat tracking system for acoustic signals of music. *Proc. of the ACM Int. Conf. on Multimedia*, San Francisco, CA, USA, Oct. 1994a. ACM Press New York, NY, USA.

M. Goto et Y. Muraoka : A Sound Source Separation System for Percussion Instruments. *IEICE Transactions*, 77, p. 901–911, 1994b.

D. E. Hall : Piano string excitation in the case of small hammer mass. *The Journal of the Acoustical Society of America*, 79 (1), p. 141–147, 1986.

D. E. Hall : Piano string excitation II : General solution for a hard narrow hammer. *The Journal of the Acoustical Society of America*, 81 (2), p. 535–546, 1987a.

D. E. Hall : Piano string excitation III : General solution for a soft narrow hammer. *The Journal of the Acoustical Society of America*, 81 (2), p. 547–555, 1987b.

D. E. Hall : Piano string excitation. VI : Nonlinear modeling. *The Journal of the Acoustical Society of America*, 92 (1), p. 95–105, 1992.

D. E. Hall et A. Askenfelt : Piano string excitation V : Spectra for real hammers and strings. *The Journal of the Acoustical Society of America*, 83 (4), p. 1627–1638, 1988.

D. E. Hall et P. Clark : Piano string excitation IV : The question of missing modes. *The Journal of the Acoustical Society of America*, 82 (6), p. 1913–1918, 1987.

P. Hanna et P. Ferraro : Polyphonic music retrieval by local edition of quotiented sequences. *Proc. Int. Work. on Content-Based Multimedia Indexing (CBMI)*, Bordeaux, France, Juin 2007.

W. Hess : *Pitch determination of speech signals.* Springer-Verlag New York, 1983.

A. J. M. Houtsma, T. D. Rossing et W. M. Wagenaars : Auditory demonstrations on compact disc. *The Journal of the Acoustical Society of America*, 83 (S1), p. 58, 1988.

R. Huber et B. Kollmeier : PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (6), p. 1902–1911, Nov. 2006.

H. Indefrey, W. Hess et G. Seeser : Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain-preliminary results. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 415–418, Tampa, FL, USA, Mars 1985.

International Music Information Retrieval Systems Evaluation Laboratory : Multiple fundamental frequency estimation & tracking. *Music Information Retrieval Evaluation eXchange (MIREX)*, Vienna, Austria, Sept. 2007.

H. Kameoka : Statistical Approach to Multipitch Analysis. Thèse de doctorat, *Univ. of Tokyo*, Japan, 2007.

H. Kameoka, T. Nishimoto et S. Sagayama : Harmonic temporal-structured clustering via deterministic annealing EM algorithm for audio feature extraction. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, Sept. 2005.

H. Kameoka, T. Nishimoto et S. Sagayama : A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering. *IEEE Transactions on Audio, Speech and Language Processing*, 15 (3), p. 982–994, 2007.

M. Karjalainen et T. Tolonen : Separation of Speech Signals Using Iterative Multi-Pitch Analysis and Prediction. *Proc. Eur. Conf. on Speech Communication and Technology (Eurospeech)*, p. 2187–2190, Budapest, Hungary, Sept. 1999. ISCA.

K. Kashino, K. Nakadai, T. Kinoshita et H. Tanaka : Application of the Bayesian probability network to music scene analysis. D. F. Rosenthal and H. S. Okuno eds., éd. : *Computational Auditory Scene Analysis*, p. 115–137. Lawrence Erlbaum Associates, Inc. Mahwah, NJ, USA, 1998.

H. Kawahara, Y. Atake et P. Zolfaghari : Accurate Vocal Event Detection Method Based on a Fixed-Point Analysis of Mapping from Time to Weighted Average Group Delay. *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, p. 664–667, Beijing, China, Oct. 2000. ISCA.

A. Klapuri : Number theoretical means of resolving a mixture of several harmonic sounds. *Proc. of the European Conference on Signal Processing (EUSIPCO)*, Island of Rhodes, Greece, Sept. 1998.

A. Klapuri : Pitch estimation using multiple independent time-frequency windows. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 1999a.

A. Klapuri : Wide-band pitch estimation for natural sound sources with inharmonicities. *Proc. of the 106th AES Convention*, Munich, Germany, Mai 1999b.

A. Klapuri : Signal processing methods for the automatic transcription of music. Thèse de doctorat, *Tampere Univ. of Technology*, Finland, 2004.

A. Klapuri : Multiple fundamental frequency estimation by summing harmonic amplitudes. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, p. 216–221, Victoria, Canada, Oct. 2006.

A. Klapuri : Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech and Language Processing*, 16 (2), p. 255–266, Fév. 2008.

A. Klapuri et M. Davy : *Signal Processing Methods for Music Transcription*. Springer, 2006.

A. Klapuri : Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11 (6), p. 804–816, 2003.

A. Klapuri : A perceptually motivated multiple-f0 estimation method. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2005.

A. Klapuri, A. Eronen et J. Astola : Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (1), p. 342–355, 2006.

A. Kobzantsev, D. Chazan et Y. Zeevi : Automatic transcription of piano polyphonic music. *Proc. Int. Symp. Image and Sig. Process. and Analysis (ISPA)*, p. 414–418, Zagreb, Croatia, Sept. 2005.

B. Krueger : Classical Piano MIDI files. http ://www.piano-midi.de, 2008.

O. Lartillot et P. Toiviainen : A Matlab toolbox for musical feature extraction from audio. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, p. 237–244, Bordeaux, France, Sept. 2007.

J. Lattard : Influence of inharmonicity on the tuning of a piano—measurements and mathematical simulation. *The Journal of the Acoustical Society of America*, 94 (1), p. 46–53, 1993.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné et S. Sagayama : Single and multiple F_0 contour estimation through parametric spectrogram modeling of speech in noisy environments. *IEEE Transactions on Audio, Speech and Language Processing*, 15 (4), p. 1135–1145, Mai 2007.

D. Lee et H. Seung : Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755), p. 788–791, 1999.

M. Lee, K. ; Slaney : Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *IEEE Transactions on Audio, Speech and Language Processing*, 16 (2), p. 291–301, Fév. 2008.

H. Lehtonen, H. Penttinen, J. Rauhala et V. Välimäki : Analysis and modeling of piano sustain-pedal effects. *The Journal of the Acoustical Society of America*, 122 (3), p. 1787–1797, 2007.

P. Leveau, E. Vincent, G. Richard et L. Daudet : Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech and Language Processing*, 16 (1), p. 116–128, Janv. 2008.

V. I. Levenshtein : Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1 (1), p. 8–17, 1965.

D. MacKay : Choice of Basis for Laplace Approximation. *Machine Learning*, 33 (1), p. 77–86, 1998.

R. Maher : A Approach for the Separation of Voices in Composite Musical Signals. Thèse de doctorat, *Univ. of Illinois at Urbana-Champaign*, USA, 1989.

J. Makhoul : Linear prediction : A tutorial review. *Proceedings of the IEEE*, 63 (4), p. 561–580, 1975.

M. Marolt : Transcription of polyphonic solo piano music. Thèse de doctorat, *Univ. of Ljubljana*, Slovenia, 2002.

M. Marolt : A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6 (3), p. 439–449, 2004.

D. W. Martin et W. D. Ward : Subjective evaluation of musical scale temperament in pianos. *The Journal of the Acoustical Society of America*, 26 (5), p. 932–932, 1954.

K. D. Martin : A blackboard system for automatic transcription of simple polyphonic music. Rap. tech. 385, M.I.T. Media Lab Perceptual Computing Technical Report, 1996.

R. McAulay et T. Quatieri : Speech analysis/Synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34 (4), p. 744–754, 1986.

R. McAulay et T. Quatieri : Pitch estimation and voicing detection based on a sinusoidal speech model. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 249–252, Albuquerque, NM, USA, Avr. 1990.

R. Meddis et M. J. Hewitt : Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I : Pitch identification. *The Journal of the Acoustical Society of America*, 89 (6), p. 2866–2882, 1991a.

R. Meddis et M. J. Hewitt : Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II : Phase sensitivity. *The Journal of the Acoustical Society of America*, 89 (6), p. 2883–2894, 1991b.

D. Mellinger : Event formation and separation in musical sound. Thèse de doctorat, *Stanford Univ.*, USA, 1991.

M. Mongeau et D. Sankoff : Comparison of musical sequences. *Computers and the Humanities*, 24 (3), p. 161–175, 1990.

G. Monti et M. Sandler : Automatic polyphonic piano note extraction using fuzzy logic in a blackboard system. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Hamburg, Germany, 2002.

J. Moorer : *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. Dept. of Music, Stanford Univ., 1975.

A. M. Noll : Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41 (2), p. 293–309, 1967.

L. I. Ortiz-Berenguer : Identificatión automática de acordes musicales. Thèse de doctorat, *Univ. Politécnica de Madrid*, Spain, 2002.

L. I. Ortiz-Berenguer et F. J. Casajús-Quirós : Polyphonic transcription using piano modeling for spectral pattern recognition. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Hamburg, Germany, 2002.

G. Peeters : Music pitch representation by periodicity measures based on combined temporal and spectral representations. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 53–56, Toulouse, France, Mai 2006.

M. Plumbley, S. Abdallah, T. Blumensath et M. Davies : Sparse representations of polyphonic music. *Signal Processing*, 86 (3), p. 417–431, 2006.

G. Poliner et D. Ellis : A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 8, p. 1–9, 2007.

L. Rabiner : On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25 (1), p. 24–33, 1977.

L. Rabiner, M. Cheng, A. Rosenberg et C. McGonegal : A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24 (5), p. 399–418, 1976.

L. Rabiner : A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2), p. 257–286, 1989.

C. Raphael : Automatic transcription of piano music. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, p. 15–19, Paris, France, Oct. 2002.

J. Rauhala : The beating equalizer and its application to the synthesis and modification of piano tones. *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Bordeaux, France, Sept. 2007.

J. Rauhala, H.-M. Lehtonen et V. Välimäki : Fast automatic inharmonicity estimation algorithm. *The Journal of the Acoustical Society of America*, 121 (5), p. 184–189, 2007.

J. Rauhala et V. Valimaki : Tunable dispersion filter design for piano synthesis. *IEEE Signal Processing Letters*, 13 (5), p. 253–256, Mai 2006.

G. Richard et C. d'Alessandro : Analysis/synthesis and modification of the speech aperiodic component. *Speech Communication*, 19 (3), p. 221–244, 1996.

M. Ross, H. Shaffer, A. Cohen, R. Freudberg et H. Manley : Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22 (5), p. 353–362, 1974.

L. Rossi : Identification de sons polyphoniques de piano. Thèse de doctorat, *Univ. de Corse*, France, 1998.

J. Rouat, Y. Liu et D. Morissette : A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21 (3), p. 191–207, 1997.

R. Roy, A. Paulraj et T. Kailath : ESPRIT–a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34 (5), p. 1340–1342, 1986.

M. Ryynänen et A. Klapuri : Polyphonic music transcription using note event modeling. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2005.

M. P. Ryynänen et A. P. Klapuri : Modelling of note events for singing transcription. *Proc. Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Jeju, Korea, Oct. 2004. ISCA.

M. P. Ryynänen et A. P. Klapuri : Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32 (3), p. 72–86, 2008.

M. R. Schroeder : Period histogram and product spectrum : New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43 (4), p. 829–834, 1968.

O. H. Schuck et R. W. Young : Observations on the vibrations of piano strings. *The Journal of the Acoustical Society of America*, 15 (1), p. 1–11, 1943.

X. Serra et J. Smith : Spectral Modeling Synthesis : A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition. *Computer Music Journal*, 14 (4), p. 12–24, 1990.

R. N. Shepard : Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36 (12), p. 2346–2353, 1964.

P. Smaragdis et J. Brown : Non-negative matrix factorization for polyphonic music transcription. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 177–180, New Paltz, NY, USA, Oct. 2003.

J. Smith et S. Van Duyne : Commuted piano synthesis. *Proc. Int. Computer Music Conf. (ICMC)*, Banff, Canada, Sept. 1995.

A. Sterian : Model-based Segmentation of Time-frequency Images for Musical Transcription. Thèse de doctorat, *Univ. of Michigan*, USA, 1999.

P. Stoica et Y. Selen : Model-order selection : a review of information criterion rules. *IEEE Signal Processing Magazine*, 21 (4), p. 36–47, 2004.

H. W. Strube : Determination of the instant of glottal closure from the speech wave. *The Journal of the Acoustical Society of America*, 56 (5), p. 1625–1629, 1974.

D. Thomson et R. Chengalvarayan : Use of voicing features in HMM-based speech recognition. *Speech Communication*, 37 (3-4), p. 197–211, 2002.

T. Tolonen et M. Karjalainen : A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8 (6), p. 708–716, 2000.

V. Tuan et C. d'Alessandro : Robust Glottal Closure Detection Using the Wavelet Transform. *Proc. Eur. Conf. on Speech Communication and Technology (Eurospeech)*, p. 2805–2808, Budapest, Hungary, Sept. 1999. ISCA.

R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering et R. van Oostrum : Using transportation distances for measuring melodic similarity. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, Baltimore, MD, USA, Oct. 2003.

R. Typke, F. Wiering et R. Veltkamp : Transportation Distances and Human Perception of Melodic Similarity. *Musicae Scientiae*, 4A, p. 153–181, 2007.

P. P. Vaidyanathan : *Multirate systems and filter banks*. Englewoods Cliffs, NJ, USA : Prentice Hall, 1993.

C. J. Van Rijsbergen : *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979. ISBN 0408709294.

T. Viitaniemi, A. Klapuri et A. Eronen : A probabilistic model for the transcription of single-voice melodies. *Proc. of Finnish Signal Processing Symposium*, Tampere, Finland, Mai 2003.

E. Vincent : Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux. Thèse de doctorat, *Univ. Paris 6 - UPMC*, France, 2004.

E. Vincent, N. Bertin et R. Badeau : Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 109–112, Las Vegas, NV, USA, Mars - Avr. 2008.

E. Vincent, R. Gribonval et C. Fevotte : Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (4), p. 1462–1469, Juil. 2006.

T. Virtanen : Sound Source Separation in Monaural Music Signals. Thèse de doctorat, *Tampere Univ. of Technology*, Finland, 2006.

T. Virtanen : Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15 (3), p. 1066–1074, Mars 2007.

A. Viterbi : Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2), p. 260–269, 1967.

P. Walmsley, S. Godsill et P. Rayner : Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 119–122, New Paltz, NY, USA, Oct. 1999.

D. Wang et G. Brown : *Computational auditory scene analysis : principles, algorithms, and applications*. IEEE Press/Wiley-Interscience, 2006.

G. Weinreich : Coupled piano strings. *The Journal of the Acoustical Society of America*, 62 (6), p. 1474–1484, 1977.

X. Wen et M. Sandler : A partial searching algorithm and its application for polyphonic music transcription. *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, Sept. 2005a.

X. Wen et M. Sandler : Transcribing piano recordings using signal novelty. *Proceedings of the 118th AES Convention*, Barcelona, Spain, Mai 2005b.

N. Wiener : *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

S. Winkler : *Digital Video Quality : Vision Models and Metrics*. Wiley, 2005.

J. Wise, J. Caprio et T. Parks : Maximum likelihood pitch estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24 (5), p. 418–423, Oct. 1976.

M. Wu, D. Wang et G. Brown : A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11 (3), p. 229–241, 2003.

C. Yeh : Multiple fundamental frequency estimation of polyphonic recordings. Thèse de doctorat, *Univ. Paris 6 - UPMC*, France, 2008.

C. Yeh, A. Robel et X. Rodet : Multiple fundamental frequency estimation of polyphonic music signals. *Proc. of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, p. 225–228, Philadelphia, PA, USA, Mars 2005.

R. W. Young : Inharmonicity of plain wire piano strings. *The Journal of the Acoustical Society of America*, 24 (3), p. 267–273, 1952.

Overview

In this thesis, automatic transcription of music will refer to the process of analyzing a music recording for extracting information related to notes. Primarily, pitches, onset times, durations, loudnesses are targeted but sometimes higher-level features like rhythm patterns, key and time signatures. Shortly, it consists in converting a stream of raw audio data into a symbolic representation, as in audio-to-score or audio-to-MIDI applications.

Automatic transcription of music is one of the major topics in the field of Music Information Retrieval (MIR), and is strongly related to several MIREX [1] tasks as Onset Detection and Multiple Fundamental Frequency Estimation and Tracking. In the MIR context, automatic transcription can also serve as a basis for further applications such as indexing tasks, query by humming (QbH) and more generally symbolic audio similarity analysis, or score alignment and following.

This PhD dissertation focuses on automatic transcription of piano music and its related tasks. Our choice to limit the study to this single instrument is motivated by both the large ratio of piano solo recordings and the scientific challenge specific to the instrument. Some papers point out that the piano automatic transcription remains one of the most difficult compared to the case of other musical instruments. The main issues that we have to cope with include:

- the large fundamental frequency (F_0) range;
- the fast and compact groups of notes caused by the virtuosity of pieces for piano;
- the high polyphony levels;
- the typical characteristics like the deviation from exact harmonicity or the beats occurring in its sounds.

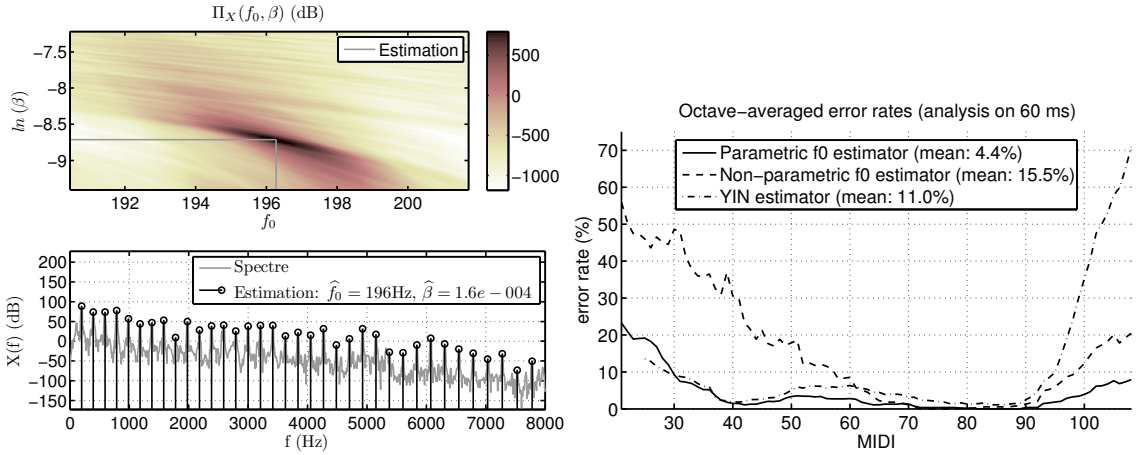
In addition, this is the opportunity to wonder whether a general topic like automatic transcription should be investigated through generic approaches, as it has been done for several decades, or by dividing the overall problem into more specific tasks, such as melody/bass line extraction, source separation and instrument-specific transcription, which has been the object of more recent studies.

The above motivations are developed in **Chapter 1**, where we review the state-of-the-art advances in four chosen directions: an introduction to the main principles of pitch estimation, including its relation to perception; a description of the approaches for multi-pitch estimation, which is often a key point in transcription systems; an overview of the numerous automatic transcription systems proposed for about thirty years; finally, in the specific context of the transcription of piano music, some insights into the physics of this instrument and a review of the existing transcription systems. The chapter ends with a set of questions in order to describe the thesis issues: which strategy could be adopted to

1. Music Information Retrieval Evaluation eXchange

transcribe piano music with reasonable chances of success? How to take into account the spectral overlap between simultaneous, harmonically-related notes for multipitch estimation? How does the inharmonicity of piano tones impact the transcription results? How to estimate the number of simultaneous notes? And what makes a good transcription from a perceptual point of view?

Chapter 2 addresses the sound models for automatic transcription. First, we present the general framework of the harmonic process, which will be used further. Two specific aspects of piano sounds are then investigated: the inharmonic distribution of the frequencies of the partials (cf. figure C.1(a)) and the modeling of the spectral envelope. In both cases, we propose some models and algorithms adapted to the piano and to the transcription task. Finally, we focus on the noise modeling for which we choose a Moving-Average model.



(a) Inharmonicity/F_0 fine estimation.

(b) Pitch estimation results.

Figure C.1: overview (1)

Chapter 3 is dedicated to an approach of pitch estimation in the case of a short analysis window and a large F_0 search range. From a general point of view, these adverse conditions are a challenge when analyzing audio – either speech or music –, because of the constraints related to the pseudo-stationarity of signals and to the time-frequency trade-off to cope with. In the case of piano tones, pitch estimation algorithms must often deal with such conditions, since the virtuosity of piano pieces may be high, the notes being very short, while the F_0 search range spreads over one of the largest existing tessituras. Within this framework, we propose a pitch estimation method for piano tones with satisfying results on the whole piano range – *i.e.* $7^1/_4$ octaves, the fundamental frequencies being distributed from $27,5$ to 4200 Hz – using a 60ms analysis window – while 93ms windows are commonly used (cf. figure C.1(b)). The method is based on a parametric approach using a high-resolution analysis to extract the parameters of the sinusoidal components. The F_0 estimator then includes a spectral function and a temporal function, jointly combined as described by Peeters [2006]. Thanks to the parametric approach, the inharmonicity of sounds is taken into account both in the time and in the frequency domains and the precision of the F_0 numeric estimation is enhanced. For more details, one may refer to the related article [Emiya *et al.*, 2007b].

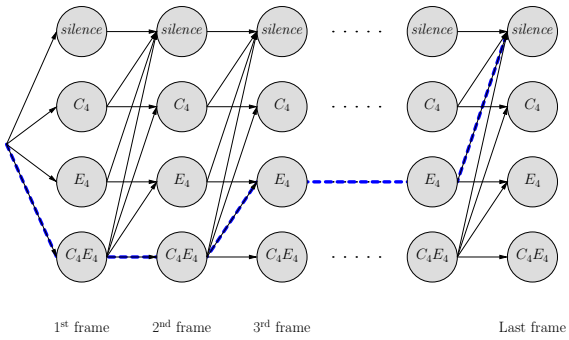
In **Chapter 4**, we propose another approach to address the multipitch estimation task. It is based on a statistical framework and a maximum likelihood resolution. The main idea

consists in finding parametric models for the spectral envelopes of notes and for the noise. By using a low-order autoregressive (AR) model, we propose a formalization of the idea of the spectral smoothness [Klapuri, 2003], allowing to deal with the variability of piano spectra. Besides, the parametric aspect makes it possible to derive an estimator for the amplitudes of partials in the case of overlapping spectra. The noise is modeled by a moving-average (MA) process, which is a model more suitable for audio signals than the commonly-chosen white noise. In the case of a sinusoids+noise mixture, using a MA noise model is an advantage with respect to an AR noise model: the latter may consider a residual sinusoid as a pole, whereas the former cannot model it well, thus enhancing the discrimination between the sinusoidal part and the noise part. The resulting multipitch estimation technique is a joint estimation approach, including the estimation of the polyphony level (*i.e.* the number of simultaneous notes) and an F_0-candidate selection stage aiming at reducing the intrinsic complexity of joint approaches. An early version of these works was published [Emiya *et al.*, 2007a].

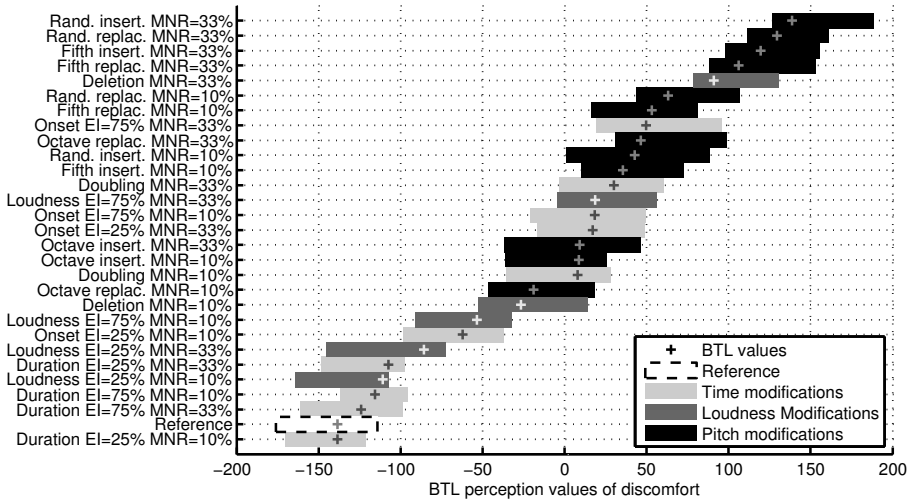
In **chapter 5**, the whole transcription system is described. The transcription strategy is based on some features of the piano and of piano pieces. It results in a framework in which the signal is segmented according to an onset detection stage. Each segment is then analyzed by means of a hidden Markov model (HMM, cf. figure C.2(a)) embedding the multipitch estimation method detailed in Chapter 4. The transcription system is able to analyze any piece of piano solo music from any style, recorded in ordinary conditions, with fair limits in terms of maximum polyphony, of speed of the played notes, and of F_0 range. An early version of these works was published [Emiya *et al.*, 2008].

Chapter 6 deals with the evaluation of automatic transcriptions. The topic has been studied in two directions. In the first part, the choice and design of the evaluation method is questioned, leading to an enhancement of the usual metrics. The limit of common evaluation systems are pointed out, showing the need for identifying the nature of errors and for taking it into account in the evaluation. Some perceptually-based versions of the original metrics are then designed, using the results of a perception test in which typical transcription errors are sorted and scored (cf. figure C.2(b)). In the second part of the chapter, we introduce a database specifically developed for multipitch estimation and automatic transcription of piano music. It is composed of recordings of isolated notes, random and usual chords and pieces of music. A large number of parameters are varying from one file to the other, such as loudness, durations or sustain pedal activation. Recordings are obtained either using a "midified" piano (Disklavier) or from high quality piano synthesis, associating an accurate ground truth to the audio files. The third section of the chapter is a detailed evalution of our multipitch algorithm and of our transcription system. Part of the works on evaluation was published [Daniel *et al.*, 2008].

Finally, conclusions are drawn, including a summary of our contributions and some perspectives.



(a) HMM for automatic transcription.



(b) Perception of typical errors.

Figure C.2: overview (2)

Liste de publications

Les publications dont la référence figure en gras sont jointes au manuscript de thèse.

Articles de conférences

[EUSIPCO'08]
Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches, V. Emiya, R. Badeau et B. David, European Conference on Signal Processing, Lausanne, Suisse, août 2008 (accepté).

[ISMIR'08]
Perceptually-based evaluation of the errors usually made when automatically transcribing music, A. Daniel, V. Emiya et B. David, International Conference on Music Information Retrieval, Philadelphie, États-Unis, septembre 2008 (accepté).

[DAFx'07]
Multipitch estimation of inharmonic sounds in colored noise, V. Emiya, R. Badeau, et B. David, 10th International Conference on Digital Audio Effects, Bordeaux, France, 10-15 septembre 2007.

[ICASSP'07]
A parametric method for pitch estimation of piano tones, V. Emiya, B. David et R. Badeau, 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, Hawaii, Etats-Unis, 15-20 Avril 2007.

[120th AES]
Harmonic plus noise decomposition : time-frequency reassignment versus a subspace based method, B. David, V. Emiya, R. Badeau et Y. Grenier, 120th Audio Engineering Society Convention, Paris, 20-23 mai 2006.

[Forum Acusticum 2005]
Two representation tools to analyse non-stationary sounds in a perceptual context, V. Emiya, B. David et V. Gibiat, Forum Acusticum 2005, Budapest, Hongrie, 29 août - 2 septembre 2005.

Conférences invitées

[ASA 2007]
Multipitch detection for piano music : Benchmarking a few approaches, B. David, R. Badeau, N. Bertin, V. Emiya et G. Richard, The Journal of the Acoustical Society of America, 122 (5) p. 2962, novembre 2007.

[ASA 2005]
Phase characterization of soundscapes,V. Gibiat, A. Padilla, V. Emiya et L. Cros, The Journal of the Acoustical Society of America, 117 (4) p. 2550, avril 2005.

Posters

[JJCAAS 2005]
Utilisation de la phase pour l'amélioration de la localisation temporelle et fréquentielle de l'analyse spectrographique, V. Emiya, Deuxièmes Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio (JJCAAS), Laboratoire de Mécanique et d'Acoustique, Marseille, 9-11 mars 2005.

Brevets

[PATENT 2002]
Tone detector and method therefor, L.F.C. Pessoa, V. Emiya, D. Melles, and D. Vallot, Freescale Semiconductor Inc., United States Patent 20040047370, European Patent EP1395065, 2002.

Séminaires internes

[SEMINAIRE 2006]
L'estimation de pitch : présentation de quelques systèmes de référence, V. Emiya, Séminaire audio ENST, juillet 2006.

[SEMINAIRE 2005]
La phase de la TFCT - Utilisation pour l'amélioration de la localisation temporelle et fréquentielle de l'énergie & Application à l'analyse spectrographique, V. Emiya, Séminaire audio ENST, janvier 2005.

Divers

[ATIAM 2004]
Spectrogramme d'Amplitude et de Fréquence Instantanées (SAFI), V. Emiya, Rapport de DEA ATIAM, juillet 2004.

[ENST 2003]
Amélioration et mise à jour des GRM Tools sur le système Digidesign HD, V. Emiya, rapport de stage de fin d'études, janvier 2003.

Sélection de publications

AUTOMATIC TRANSCRIPTION OF PIANO MUSIC BASED ON HMM TRACKING OF JOINTLY-ESTIMATED PITCHES

Valentin Emiya, Roland Badeau, Bertrand David

TELECOM ParisTech (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France
valentin.emiya@enst.fr

ABSTRACT

This work deals with the automatic transcription of piano recordings into a MIDI symbolic file. The system consists of subsequent stages of onset detection and multipitch estimation and tracking. The latter is based on a Hidden Markov Model framework, embedding a spectral maximum likelihood method for joint pitch estimation. The complexity issue of joint estimation techniques is solved by selecting subsets of simultaneously played notes within a pre-estimated set of candidates. Tests on a large database and comparisons to state-of-the-art methods show promising results.

1. INTRODUCTION

In this work, we consider the music transcription task of analyzing a recorded piece and estimating a symbolic version from it, as a MIDI file for instance. Accomplishing this task either humanly or with an automatic system often proves to be a complex issue that requires a multilevel analysis. Important subtasks include the estimation of pitch, loudness, onset and offset of notes, which can be completed by higher level analysis, such as extracting rhythm information, recognizing the played instruments or looking into the tonal context.

Many transcription systems have been developed in the recent decade, using different approaches and often mixing theoretical framework, empirical considerations and fine tuning. For instance, Davy et al. [2] and Kameoka et al. [3] use two different Bayesian approaches while the technique proposed by Marolt [4] relies on oscillators and neural networks. Blind decomposition techniques are utilized by Bertin et al. [5] and Vincent et al. [6]. The task is also accomplished by detecting temporal [1] or spectral [7] patterns with online or offline learning techniques. The system proposed by Ryynänen and Klapuri [8] takes into account the psychoacoustic knowledge of the human peripheric auditory perception and involves some musicological considerations in the postprocessing for enhancing the results.

This paper, while presenting a system whose output is a MIDI-transcribed file, focuses more on a Hidden Markov Model (HMM) framework in which a joint multiple fundamental frequency (F_0) estimation method [9] is embedded. As with most joint estimation approaches, a high number of chord combinations must be tested. Consequently, strategies are required in order to prune this search space [10]. Here, the use of an onset detector and a selector of F_0 candidates solves this computational issue by picking a reduced set of likely combinations to test.

The research leading to this paper was supported by the European Commission under contract FP6-027026-K-SPACE and by the French GIP ANR under contract ANR-06-JCJC-0027-01, *Décomposition en Éléments Sonores et Applications Musicales - DESAM*. The authors would also like to thank M. Marolt, N. Bertin, E. Vincent and M. Alonso for sharing their programs, and L. Daudet for its piano database. [1]

This paper is structured as follows. Section 2 gives an overview of the system and details each of its stages. Section 3 then reports test results and compares them with some state-of-the-art systems. Finally, conclusions are presented in section 4.

2. TRANSCRIPTION SYSTEM

The whole processing system is represented by Algorithm 1. The input is the recording of a piece of piano music, denoted by $x(t)$, the corresponding discrete sequence being sampled at $F_s = 22050$ Hz. Firstly, an onset detection is applied. The signal portion between two successive onsets will be herein referred to as a **segment**. For each of these variable-length segments, we select a presumably oversized set of F_0 candidates. The segment is then split into overlapping **frames** of constant length N, which are processed using a HMM. In this framework, the likelihood related to each possible set of simultaneous notes (or local chord) is derived using a recent spectral Maximum Likelihood (ML) estimation technique [9]. A MIDI file is generated at the end.

Algorithm 1 (System overview)

Input: Waveform
  Detect onsets
  for each segment between subsequent onsets **do**
    Select note candidates
    Track the most likely combination of notes within the HMM framework
  end for
  Detect repetitions of notes from one segment to the next one
Output: MIDI file

2.1 Onset detection

Onset detection is performed by an existing algorithm [11] based on the so-called spectral energy flux. An adaptive threshold is used to extract the local maxima of a detection function that measures phenomenal accents. The temporal accuracy of the onset detection is about 12 ms.

2.2 Selection of F_0 candidates

In order to select the set of note candidates, a normalized product spectrum is derived as a function of the fundamental frequency. The function is defined by:

$$S(f_0) = \frac{1}{H(f_0)^{\nu}} \ln \prod_{h=1}^{H(f_0)} |X(f_h)|^2 \qquad (1)$$

where $H(f_0)$ is the number of overtones below a predefined cut-off frequency for fundamental frequency f_0, $X(f)$ is a whitened version of the Discrete Fourier Transform (DFT) of the current frame, ν is a parameter empirically

set to .38 to adjust the normalization term $H(f_0)^\nu$ and f_h is the frequency of the overtone with order h defined by $f_h = hf_0\sqrt{1+\beta h^2}$, β being the so-called inharmonicity coefficient of the piano tone [12]. The whitening process reduces the spectral dynamics. It is performed by modeling the background noise with an autoregressive process and by applying the inverse filter to the original signal frame. The normalization by $H(f_0)^\nu$ aims at correcting the slope due to the variation of the number of overtones with respect to f_0. In the special case $\nu = 0$, the function in (1) equals the product spectrum [13], which has been designed for a constant number of overtones. For each note in the search range, the function is maximized in a 2-dimensional plane in the neighborhood of $(f_0, \beta(f_0))$ where f_0 is the well-tempered scale value of the fundamental frequency and $\beta(f_0)$ is a typical value of the inharmonicity coefficient for this note [12, pp. 365]. The N_c notes with the highest values are then selected as candidates. In practice, the function is computed and averaged over the first three frames of each segment in order to be robust to the temporal spreading of onsets before the steady state of notes is reached. N_c is set to 9, the frame length is 93 ms, with a 50%-overlap. The cut-off frequency related to $H(f_0)$ varies from 1200 Hz in the bass range ($f_0 = 65$ Hz) to $F_s/2$ in the treble range ($f_0 = 1000$ Hz and above).

The candidate selection stage has two purposes. Firstly, as with all joint estimation techniques, the approach described in this paper faces a high combinatory issue. Selecting a reduced set of candidates is a way to reduce the number of combinations to be tested, and to reach a realistic computational cost. For instance, if the maximum polyphony is 5 and the note search range spreads over 5 octaves (*i.e.* 60 notes), around $5 \cdot 10^6$ combinations have to be tested, whereas after selecting $N_c = 9$ candidates among the 60 notes, the size of the set of possible chords decreases to only 382 elements. Secondly, the efficiency of the transcription system depends on the selection of optimized values for fundamental frequencies and inharmonicity. These values can be obtained by either maximizing the candidate selection function defined above, which is a criterion on the energy of overtones, or maximizing the likelihood described in the next section. The advantage of the first solution is that it reduces the overall computational cost: the optimization task is performed on a lower number of candidates (*e.g.* with the figures mentioned above, 60 note candidates instead of 382 chords) and a call to the candidate selection function does not require as much computation time as a call to the likelihood function.

2.3 HMM tracking of most likely notes

This section describes how to track the possible combinations of note candidates, along frames, by means of one HMM [14] per segment. From now on, each possible combination of notes in a frame is referred to as a *chord*.

Let us consider a segment obtained by the onset detection stage, *i.e.* delimited by two consecutive onsets. It is composed of U frames, numbered from 1 to U. In frame u, $1 \leq u \leq U$, the observed spectrum X_u is a random variable that depends on the underlying chord, denoted by c_u. c_u is also a random variable. When one or several notes are played, they spread over a number of consecutive frames and may be extinguished at any moment. Thus, in a short-term context like the duration of a segment, the polyphonic content c_u of frame u strongly depends on the short-term past. Consequently, a first-order Markovian process is assumed for the chord sequence $c_1 \dots c_U$: for $u \geq 2$, chord c_u only depends on chord c_{u-1} and does not depend on u, resulting in

$$p(c_{u+1}|c_1 \dots c_u) = p(c_{u+1}|c_u) \tag{2}$$

Thus the transcription of the segment consists in finding the best sequence of hidden chords $\hat{c}_1\hat{c}_2 \dots \hat{c}_U$ given the sequence of observations $X_1X_2 \dots X_U$.

In statistical terms, $\hat{c}_1 \dots \hat{c}_U$ is obtained by solving:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U|X_1 \dots X_U) \tag{3}$$

which is rewritten using the Bayes rule as:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}} \frac{p(c_1 \dots c_U)\, p(X_1 \dots X_U|c_1 \dots c_U)}{p(X_1 \dots X_U)} \tag{4}$$

The denominator is removed without changing the argmax result:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U)\, p(X_1 \dots X_U|c_1 \dots c_U) \tag{5}$$

Given that the observed spectrum X_u only depends on the underlying chord c_u, we obtain:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U) \prod_{u=1}^{U} p(X_u|c_u) \tag{6}$$

Using eq. (2), this finally reduces to:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1) \prod_{u=2}^{U} p(c_u|c_{u-1}) \prod_{u=1}^{U} p(X_u|c_u) \tag{7}$$

In a HMM context, c_u is the so-called hidden state in frame u, X_u is the observation, $p(c_1)$ is the initial-state probability, $\lambda_{c_{u-1},c_u} \triangleq p(c_u|c_{u-1})$ is the state-transition probability from c_{u-1} to c_u and $p(X_u|c_u)$ is the observation probability in state c_u. Thanks to the selection of N_c note candidates in the current segment and to the polyphony limit set to $P_{\max}$, possible values for c_u are restricted to the sets composed of 0 to $P_{\max}$ notes among the candidates. Thus, the number of states equals $\sum_{P=0}^{P_{\max}} \binom{N_c}{P}$. Transition probabilities are defined as follows:

- for $2 \leq u \leq U$, the birth of a note is not allowed in frame u since a segment is defined as being delimited by two consecutive onsets. Thus $\lambda_{c,c'} \triangleq 0$ if c' is not a subset of notes from c.
- as a result, the only transition from the "silence" state is toward itself: $\lambda_{\emptyset,\emptyset} \triangleq 1$.
- transitions from c are allowed toward c itself or toward a subset c' of c. It only depends on the number of notes in chord c and in chord c'. Probability transitions are learnt from musical pieces, as described below.

The initial-state probabilities are also learnt as a function of the number of notes in the state. The learning process is performed on MIDI music files. The initial-state probabilities are learnt from the polyphony observed at each onset time and the transition probabilities from the polyphony evolution between consecutive onset times. Finally, $p(X_u|c_u)$ is the likelihood detailed in the next section and given in a logarithmic version in eq. (13): $\ln p(X_u|c_u) = \tilde{L}_{X_u}(c_u)$.

The Viterbi algorithm [15] is applied to extract the best sequence of states that explains the observations, as illustrated in Table 1 and Figure 1. The obtained chord sequence corresponds to the set of notes present at the beginning of the segment and to the possible terminations within the segment.

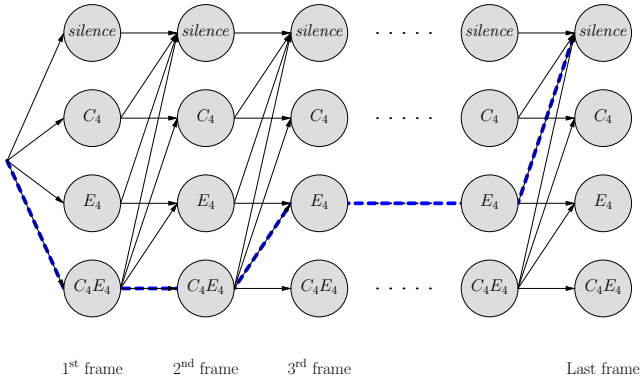


Figure 1: Chord network corresponding to the transition matrix of Table 1. Due to the sparsity of the transition matrix, transitions are allowed toward the same chord or toward a "subchord", when note endings occur. The thick, dashed line shows a possible Viberbi path: chord $\{C_4, E_4\}$ is detected, C_4 dies at frame 3 while E_4 lasts until next to last frame of the segment.

c \ c'	$\emptyset$	C_4	E_4	C_4 E_4
$\emptyset$	1	0	0	0
C_4	.17	.83	0	0
E_4	.17	0	.83	0
C_4E_4	.07	.06	.06	.80

Table 1: Example of transition matrix, *i.e.* the probability of going from chord c at time t to chord c' at time $t+1$. For graphical convenience, only $N_c = 2$ candidates are selected (notes C_4 and E_4). The transition probability is learnt as a function of the number of notes in c and c'.

Note that this HMM-based approach shares some similarities with another transcription system [10] in which chords are also modeled by HMM states. However, major differences exists between the two systems, especially in the choice of the observations, of their likelihood estimation, and in the way the HMM networks are significantly simplified, thanks to the onset-based segmentation, the selection of note candidates and the use of one state per chord in the current paper.

2.4 Maximum likelihood estimation of simultaneous pitches

The core of the pitch estimation stage is performed at the frame level by estimating the likelihood of the observed DFT X given a possible chord. In 2.4.1 and 2.4.2, we present a summary of this method that has been developed in a recent work [9]. A normalization stage is then introduced, with a number of purposes: obtaining a homogeneous detection function with respect to low/high pitched notes, making it robust to polyphony variations, and adjusting dynamics between the note and the noise likelihoods. A silence detection mechanism is finally described.

2.4.1 Maximum likelihood principle

Let us consider a mixture of M notes and of an additive colored noise. We observe one frame from this mixture. For $1 \leq m \leq M$, note m is modeled by a set of sinusoidal components. The set of related frequency bins is denoted by $\mathcal{H}^{(m)}$. In the case of a piano note, $\mathcal{H}^{(m)}$ is defined by:

$$\mathcal{H}^{(m)} = \left\{ f_h \,\middle|\, f_h = h f_0^{(m)} \sqrt{1+\beta^{(m)}h^2} < \frac{F_s}{2} \right\} \quad (8)$$

where $f_0^{(m)}$ is the fundamental frequency of note m and $\beta^{(m)}$ is its inharmonicity coefficient.

We then assume that the noise spectrum is observed in any frequency bin that is not located within the primary spectral lobe of a note component. The set $\mathcal{N}$ of frequency bins related to noise observations is inferred by $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}$ and is thus defined by:

$$\mathcal{N} = \left\{ f \in \mathcal{F} \,\middle|\, \forall f' \in \bigcup_{m=1}^{M} \mathcal{H}^{(m)}, |f - f'| > \Delta f/2 \right\} \quad (9)$$

where $\mathcal{F}$ is the whole set of frequency bins and Δf is the width of the primary spectral lobe ($\Delta f = 4/N$ for a Hann window).

We now consider a set $\mathcal{S} \in \left\{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right\}$, *i.e.* the set of frequency bins related to any of the elements (either a note or the noise) in the mixture. We model the set of selected spectral observations $X(\mathcal{S})$ by a transfer function in a family of parametric functions (*e.g.* all-pole functions)[1]. We showed in [9] that the normalized log-likelihood of $X(\mathcal{S})$ can be analytically written as:

$$L_{\mathcal{S}}(R) = c - 1 + \ln\left(\rho_{\mathcal{S}}(R)\right) \quad (10)$$

where R is the considered parametric transfer function, $c = -\frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \ln(\pi|X(k)|^2)$ is a constant w.r.t. R ($\#\mathcal{S}$ denotes the number of elements in $\mathcal{S}$) and

$$\rho_{\mathcal{S}}(R) = \frac{\left(\prod_{k\in\mathcal{S}} \left|\frac{X(k)}{R(k)}\right|^2\right)^{\frac{1}{\#\mathcal{S}}}}{\frac{1}{\#\mathcal{S}} \sum_{k\in\mathcal{S}} \left|\frac{X(k)}{R(k)}\right|^2} \quad (11)$$

is equal to the ratio between the geometrical mean and the arithmetical mean of the set $\left\{\left|\frac{X(k)}{R(k)}\right|^2\right\}_{k\in\mathcal{S}}$. Such a ratio is maximal and equal to 1 when $|X(k)/R(k)|$ is constant, independent of k, which means that $\rho_{\mathcal{S}}(R)$ measures the

[1] Note that in the case of frequency overlap between two notes, this modelisation does not hold for the overlapped frequency bins. This phenomenon is ignored here.

whiteness, or the *flatness* of $\left\{\left|\frac{X(k)}{R(k)}\right|^2\right\}_{k\in\mathcal{S}}$. Thus the application of the Maximum Likelihood principle, *i.e.* the estimation of the parameterized function that maximizes $\rho_{\mathcal{S}}(R)$ results in whitening the spectral envelope $|X(\mathcal{S})|$. In our system, the note $m \in [\![1, M]\!]$ is modeled by an autoregressive (AR) filter whereas noise is modeled as a finite impulse response (FIR) filter of length $p \ll N$. A discussion on the choice of these models is presented in [9]. Approximate solutions $\widehat{R}_{\mathcal{H}_1}, \ldots, \widehat{R}_{\mathcal{H}_M}, \widehat{R}_{\mathcal{N}}$ to the optimization problem are obtained by means of estimation techniques in the case of partially observed spectra [16].

2.4.2 Joint estimation function

The M simultaneous notes are parameterized by $2M$ coefficients, which are the M fundamental frequencies $f_0^{(1)}, \ldots, f_0^{(M)}$ and M related inharmonicity coefficients $\beta^{(1)}, \ldots, \beta^{(M)}$. Using equations (8) and (9), a chord is thus characterized by the set $\mathcal{C} = \left\{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right\}$, *i.e.* by the way how the frequency sets of the various elements of the chord are built.

Our pitch estimator relies on a weighted maximum likelihood (WML) method: for a chord $\mathcal{C}$, we calculate the weighted log-likelihood of the observed spectrum

$$\begin{aligned} L_X(\mathcal{C}) &= L_X\left(\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right) \\ &= \frac{1}{2M}\sum_{m=1}^{M} \ln \rho_{\mathcal{H}^{(m)}}\left(\widehat{R}_{\mathcal{H}^{(m)}}\right) + \frac{1}{2}\ln \rho_{\mathcal{N}}\left(\widehat{R}_{\mathcal{N}}\right) \end{aligned} \tag{12}$$

2.4.3 Log-likelihood normalization

For any given set of bins $\mathcal{S}$ and transfer function R, the statistical properties of the flatness $\rho_{\mathcal{S}}(R)$ depend both on the statistical properties of the whitened data X/R and on the number of observations $\#\mathcal{S}$. The latter influence tends to lower $\rho_{\mathcal{S}}(R)$ when $\#\mathcal{S}$ increases, which raises a double issue when using expression (12):

- since $\#\mathcal{N} \gg \#\mathcal{H}^{(m)}$, the various terms of the sum have values and variations that cannot be compared
- the lack of homogeneity also appears when comparing the log-likelihoods $L_X(\mathcal{C}_1)$ and $L_X(\mathcal{C}_2)$ of two chords $\mathcal{C}_1$ and $\mathcal{C}_2$ since bins are not grouped in comparable subsets.

Thus expression (12) is normalized as:

$$\begin{aligned} \tilde{L}_X(\mathcal{C}) = &\frac{1}{2M}\sum_{m=1}^{M} \frac{\ln \rho_{\mathcal{H}^{(m)}}\left(\widehat{R}_{\mathcal{H}^{(m)}}\right) - \mu_{\#\mathcal{H}^{(m)}}}{\sigma_{\mathcal{H}}} \\ &+ \frac{1}{2}\frac{\ln \rho_{\mathcal{N}}\left(\widehat{R}_{\mathcal{N}}\right) - \mu_{\#\mathcal{N}}}{\sigma_{\mathcal{N}}} \end{aligned} \tag{13}$$

where $\mu_{\#\mathcal{S}}$ is the empirical median, depending on the observation vector size $\#\mathcal{S}$ and $\sigma_{\mathcal{S}}$ is the (constant) empirical standard deviation of $\ln \rho_{\mathcal{S}}$, both being adaptively estimated from the frame. $\tilde{L}_X(\mathcal{C})$ thus represents a normalized log-likelihood, with mean and variance around 0 and 1 respectively.

2.4.4 Silence model

When a given chord $\mathcal{C}$ is played, a peak appears such that $\tilde{L}_X(\mathcal{C}) \gg 1$, whereas for $\mathcal{C}' \neq \mathcal{C}$, $\mathcal{C}' \mapsto \tilde{L}_X(\mathcal{C}')$ is a centered process with variance 1. However, when no chord is played, no peak is generated in function $\tilde{L}_X(\mathcal{C})$ at the expected "empty chord" $\mathcal{C}_0$ = silence. A simple solution to detect silences is to set a constant value $\tilde{L}_X(\mathcal{C}_0) = \tilde{L}_0 \geq 1$ in order to obtain $\tilde{L}_X(\mathcal{C}) \gg \tilde{L}_0 = \tilde{L}_X(\mathcal{C}_0)$ when chord $\mathcal{C}$ is played, and to guarantee $\tilde{L}_X(\mathcal{C}_0) \geq \tilde{L}_X(\mathcal{C})$ for possible $\mathcal{C} \neq \mathcal{C}_0$ when no chord is played, since $\tilde{L}_X(\mathcal{C}) \leq 1$ in this case. In our implementation, $\tilde{L}_0$ is empirically set to 2.

2.5 Detection of repeated notes and MIDI file generation

The sets of notes estimated in frames of successive segments are assembled to obtain an overall discrete-time piano roll of the piece. When a note is detected both at the end of a segment and at the beginning of the following one, it may be either a repeated note, or a single one overlapping both segments. The variation of the note loudness is estimated by using the function introduced in eq. (1) for candidate selection, derived on the non-whitened version of the DFT. A note is then considered as repeated when its loudness variation is greater than a threshold empirically set to 3 dB. A MIDI file is finally generated with a constant loudness for each note.

3. EXPERIMENTAL RESULTS

Our transcription system is evaluated on a set of 30s-truncated versions of 90 piano pieces randomly chosen from B. Krueger's MIDI file database[2]. Tests are carried out using a 93-ms frame length with a note search range composed of 60 notes between C2 (MIDI note 36) and B6 (MIDI note 95). $N_c = 9$ notes are used for F_0 candidate selection, maximum polyphony is set to $P = 5$ and λ_0 is empirically set to 0.9. In order to compare the transcription with a reliable ground truth, audio files are first generated from the original MIDI files and the synthesized audio files then constitute the input of the transcription system. The audio generation is performed by virtual pianos (Steinberg The Grand 2, Native Instrument Akoustik Piano and Sampletekk Black Grand) that use a large amount of sampled sounds and provide a high quality piano synthesis. A note-based evaluation is drawn: a MIDI file comparison is performed between the original file and the transcription by classifying notes into True Positives (TP), False Positives (FP) and False Negatives (FN). TP are defined as notes with a correct pitch (with a quartertone tolerance) and an onset time error lower than 50 ms (commonly used threshold), FP are the other transcribed notes and FN are the non-transcribed notes. Performance is then evaluated through four rates: *recall* $\frac{\#\mathrm{TP}}{\#\mathrm{TP}+\#\mathrm{FN}}$, *precision* $\frac{\#\mathrm{TP}}{\#\mathrm{TP}+\#\mathrm{FP}}$, *F-measure* $\frac{2\ precision\times recall}{precision+recall}$ and *mean overlap ratio* $\frac{1}{\#\mathrm{TP}}\sum_{i\in\mathrm{TP}}\frac{\min(\mathrm{offsets}_i)-\max(\mathrm{onsets}_i)}{\max(\mathrm{offsets}_i)-\min(\mathrm{onsets}_i)}$, where onsets_i (or $\mathrm{offsets}_i$) denote the pair of onset (or offset) times for TP note i in the reference and in the transcription. The mean overlap ratio is an averaged ratio between the length of the intersection of the temporal supports of an original note and its transcription, and of the length of their union. Note that offset times are not taken into account when classifying notes between TP, FP and FN sets. This choice is due to two main reasons. First, since piano notes are damped sounds, endings are difficult to determine and depend on the instrument and on the recording conditions, even for a given original MIDI file. Secondly, the evaluation system used here enables to take offset times into account by means of the mean overlap ratio. Hence F-measure, precision and recall rates focus on the evaluation of pitches and onset times.

The results[3] of our system are reported in Figure 2, together with the performance of state-of-the-art methods [4, 5, 6]. All transcription systems are tested on the same database as our system, using their authors' original

[2] `http://www.piano-midi.de/`

[3] See also the audio examples available on the authors' web site at: `http://perso.enst.fr/~emiya/EUSIPCO08/`

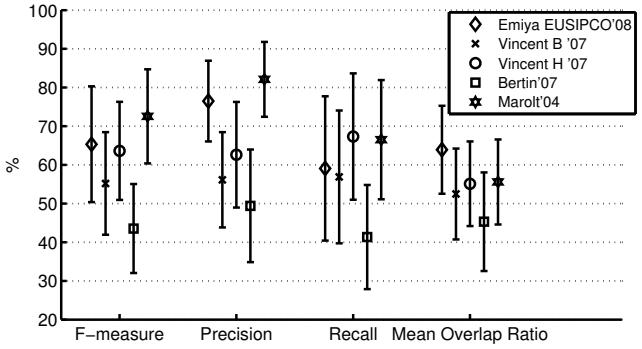


Figure 2: Evaluation results: F-measure, recall, precision and mean overlap ratio for several transcription systems, averaged on transcriptions of several pieces. *Vincent B '07* and *Vincent H '07* stand for the "Baseline" and "Harmonic" methods presented in [6]. Vertical lines show standard deviations of indicators w.r.t. all pieces, showing how results depend on music excerpts.

code. Performance rates have been averaged with respect to all test pieces, and the related standard deviation is also represented. Results dramatically depend on musical excerpts, with a number of consequences. Thus, high performance (*e.g.* F-measure greater than 90%) is reached for pieces with slow tempo or low polyphony while fast pieces are generally difficult to transcribe. This explains the large standard deviations, and means that such absolute figures should not be compared to results from other publications since they drastically depend on the database. Besides, the confidence in the results is not directly related to these standard deviations: it has been assessed by an ANOVA test on F-measure data, passed using a 0.01 test level.

Our approach is comparable to state-of-the-art systems in terms of global performance. The obtained F-measure (65%) is satisfying. We obtain a better mean overlap ratio than the competing methods, which suggests that the proposed HMM framework for note tracking is efficient both for selecting pitches among candidates, and for detecting their possible endings. It results in an efficient transcription of durations, enhancing the phrasing similarity with the original piece and thus participating in the subjective quality when hearing the correct transcribed notes. Similar results were obtained when the system was tested with real piano recordings, using 30s-excerpts of pieces used in [1]. In this case, the average F-measure reaches 79%, the musical content of the excerpts appearing quite easy to transcribe (low polyphony, medium tempi).

Generally, the main errors of our system are: polyphony limitation, missed notes in the onset detection and candidate selection stages, errors at the bass and treble bounds and harmonically related confusions (octave and others). Our system is implemented in Matlab and C, and its running time is less than 200 times realtime on a recent PC.

4. CONCLUSIONS

In this paper, we have put forward a new approach to the automatic transcription of piano music, applying some recent advances in pitch estimation and including an original structure to deal with joint estimation. The system has been tested on a database of musical pieces, compared with competing systems and has reached a satisfying performance.

Future work will deal with the estimation of loudness of notes, the question of overlap between overtones and the development of a more accurate silence model.

REFERENCES

[1] J.P. Bello, L. Daudet, and M.B. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2242 – 2251, Nov. 2006.

[2] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acous. Soc. Amer.*, vol. 119, no. 4, pp. 2498–2517, 2006.

[3] H. Kameoka, T. Nishimoto, and S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[4] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[5] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of ICASSP 2007*, Honolulu, Hawaii, USA, Apr.15–20 2007, vol. I, pp. 65–68.

[6] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Las Vegas, Nevada, USA, Mar. 30 – Apr. 4 2008.

[7] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, vol. 8, pp. 1–9, 2007.

[8] M. Ryynänen and A.P. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, New Paltz, NY, USA, Oct. 2005, pp. 319–322.

[9] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of inharmonic sounds in colored noise," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Bordeaux, France, Sept.10–15 2007, pp. 93–98.

[10] C. Raphael, "Automatic transcription of piano music," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[11] M. Alonso, G. Richard, and B. David, "Extracting note onsets from musical recordings," in *Proc. of the ICME*, Amsterdam, The Netherlands, July6–8 2005, pp. 1–4.

[12] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[13] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acous. Soc. Amer.*, vol. 43, no. 4, pp. 829–834, 1968.

[14] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[16] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Las Vegas, Nevada, USA, Mar.30 – Apr.4 2008.

PERCEPTUALLY-BASED EVALUATION OF THE ERRORS USUALLY MADE WHEN AUTOMATICALLY TRANSCRIBING MUSIC

Adrien DANIEL, Valentin EMIYA, Bertrand DAVID
TELECOM ParisTech (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France

ABSTRACT

This paper investigates the perceptual importance of typical errors occurring when transcribing polyphonic music excerpts into a symbolic form. The case of the automatic transcription of piano music is taken as the target application and two subjective tests are designed. The main test aims at understanding how human subjects rank typical transcription errors such as note insertion, deletion or replacement, note doubling, incorrect note onset or duration, and so forth. The Bradley-Terry-Luce (BTL) analysis framework is used and the results show that pitch errors are more clearly perceived than incorrect loudness estimations or temporal deviations from the original recording. A second test presents a first attempt to include this information in more perceptually motivated measures for evaluating transcription systems.

1 INTRODUCTION

In the benchmarking of Information Retrieval systems, performance is often evaluated by counting and classifying errors. Classically the ratio of relevant items that are returned out of the full set of original ones, referred to as *recall*, measures the completeness of the system performance whereas the proportion of relevant items that are retrieved, or *precision*, indicates the correctness of the answer. The F-measure, combining precision and recall, offers a single score to assess the performance. When music processing systems are involved, the question arises as to how to complement such a quantitative assessment by incorporating a certain amount of perceptually motivated criteria or weights.

This paper investigates the perceptual importance of typical errors occurring when transcribing polyphonic music excerpts into a symbolic form, *e.g.* converting a piece recorded in a PCM (.wav) format into a MIDI file. This particular Music Information Retrieval (MIR) task and its related subtasks (onset detection, multipitch estimation and tracking) have received a lot of attention [9] from the MIR community since the early works of Moorer [14] in the mid 70s. The approaches used to accomplish the goal are very diverse [4, 5, 14, 15, 16] and the evaluation of the performance for such systems is almost as varied. Some papers [4, 14] focus on a couple of sound examples, to probe typical errors such as octave errors, or deviations from ground truth such as duration differences, and so forth. However, the most widely used criteria for assessing automatic transcription are quantitative, even if the evaluation framework is not always similar (frame-based [15], note-based [16] or both [1]).

In the practical context of piano music for instance, the evaluation task is often handled by generating the PCM format piece from an original MIDI file which makes it possible to compare the input (ground truth) and output MIDI files. For that particular case, in this study, a perception test has been designed for subjectively rating a list of typical transcription errors (note insertions, deletions, incorrect onsets or duration...). The test is based on pairwise comparisons of sounds holding such targeted errors. The results are then analyzed by means of the Bradley-Terry-Luce (BTL) method [3].

In a second step, the question emerged of finding a way to take into account the perceptual ranking of the discomfort levels we obtained. Another test was designed to subjectively compare transcriptions resulting from different systems. It aimed at deriving more perceptually relevant metrics from the preceding BTL results by synthetically combining their main findings, and at checking their compliance with the test results. We worked in two directions: perceptually weighting typical errors, countable by comparing the input and output MIDI files, and adaptating similarity metrics [17].

2 THE EVALUATION MEASURES

The commonly-used F-measure is defined by:

$$f \triangleq 2\frac{rp}{r+p} = \frac{\#\text{TP}}{\#\text{TP} + \frac{1}{2}\#\text{FN} + \frac{1}{2}\#\text{FP}} \qquad (1)$$

The authors thank all the subjects involved in the perceptive test for their participation. They also thank M. Castellengo, A. de Cheveigné, D. Pressnitzer and J. Benenson for their useful remarks, and M. Marolt, N. Bertin and P. Leveau for sharing their programs. The research leading to this paper was supported by Institut TELECOM under the *Automatic Transcription of Music: Advanced Processing and Implementation - TAMTAM* project and by the French GIP ANR under contract ANR-06-JCJC-0027-01, *Décomposition en Éléments Sonores et Applications Musicales - DESAM*.

where r denotes the recall, p the precision, $\#$TP the number of true positives (TP), $\#$FN the number of false negatives (FN) and $\#$FP the number of false positives (FP). f is equivalent to the quantity a, that is referred to as either accuracy or score [5], since $f = \frac{2}{\frac{1}{a}+1}$. The F-measure is useful to obtain the error rate for individually counted errors, but does not consider aspects like sequentiality, chords, harmonic or tonal relationships, etc.

Another evaluation approach comes from the problem of finding the similarity between two (musical) sequences. At the moment, these methods are commonly used to search for similar melodies in large databases, rather than in the field of the evaluation of transcriptions.

Let us assume that one must compare two sequences of symbols, A and B. The Levenshtein's distance, or edit distance [11], is a metric that counts the minimal number of operations necessary to transform A to B. The possible operations on symbols are: deletion from A, insertion into B, or replacement of a symbol in A by another one in B.

Mongeau and Sankoff [13] proposed adapting this distance to the case of monophonic musical sequences, in order to define a similarity metric between two melodies. The two sequences of notes are ordered according to the onset of each note. Each note is characterized by its pitch and duration, which are used to compute the cost of the following possible operations: insertion, deletion, replacement, with costs depending on tonal criteria, fragmentation and consolidation of several notes with the same pitch. These operations reflect typical mistakes in transcriptions. The minimum distance between the sets of notes is then estimated using the edit distance framework.

This melodic edit distance being applicable only to monophonic sequences, an extension to the polyphonic case has been recently proposed [8]. In order to represent the polyphonic nature of musical pieces, quotiented sequences are used. So far, this representation has only been applied to chord sequences, which constitute a restricted class of musical pieces: the notes within a chord must have the same onset and duration.

Another way to compute the similarity between two musical sequences [17] consists in considering each set of notes as points in a multidimensional space, *e.g.* the pitch/time domain. The algorithm is based on two choices. First, each point must be assigned a weight, *e.g.* the note duration. Second, a distance between a point in the first set and a point in the second one is defined, *e.g.* the euclidian distance in the time/pitch space. Then, the overall distance can be computed with the *Earth Movers Distance* (EMD) or the *Proportional Transportation Distance* (PTD). It is related to the minimum amount of work necessary to transform one set of weighted points to the other using the previously-defined distance, making it possible to transfer the weight of a source note towards several targets.

In all of these methods, the setting of the parameters is a crucial point. Indeed, the weighting between the time and the pitch dimensions, for instance, depends on music perception. The tests presented in this paper aim at assessing the trends of the perceptive impact of typical errors and the distribution of their related weights.

3 EXPERIMENTAL SETUP

3.1 Overview

The perception test consists of two tasks, which are detailed below. It was available on the Internet in the spring of 2007 for two weeks and was announced by e-mail. Before accessing the tasks, the subject is given instructions and information on the recommended audio device (high-quality headphones or loudspeakers, and a quiet environment) and on the estimated duration of the test. He or she is then invited to complete the tasks. Both of them consist in hearing a musical excerpt and several transcriptions of it, and in focusing on the discomfort caused by the transcriptions, with respect to the original. Task 1 uses artificial transcriptions, *i.e.* some copies of the original piece into which errors were inserted whereas task 2 uses transcriptions obtained by automatic transcription systems. In both cases, the transcriptions are resynthesized in the same recording conditions as the original piece in order to be heard and compared by the subject. At the end, the subject was asked to describe the criteria he used to compare files and to add any comments. Due to the total duration of the test a subject can possibly endure (about 40' here), we limited the scope of the study to pieces of classical piano music, from different periods, with different tempi and harmonic/melodic content.

3.2 Test 1: Subjective Evaluation of Typical Transcription Errors

3.2.1 Principle

Test 1 aims at obtaining a specific score for typical transcription errors. In order to achieve this, the transcriptions to be evaluated are made by inserting one and only one kind of error into an original excerpt. The error is chosen among the following list of typical errors: note deletion, random-pitched note insertion (1 to 11 half-tones), random-pitched note replacement (1 to 11 half-tones), octave insertion, octave replacement, fifth insertion, fifth replacement, note doubling, onset displacement, duration change (offset modification) and loudness modification (MIDI velocity).

These errors are inserted into three excerpts from *Studies, op 10 / Study 1 in C Major* by Chopin (8 seconds), *Suite Bergamasque / III. Clair de Lune* by C. Debussy (20 seconds), and *Sonata in D Major KV 311 / I. Allegro con Spirito* by W.A. Mozart (13 seconds).

Ideally, we would like to obtain a ranking of the typical errors. Due to the large number of files, asking the subjects

Figure 1. Test 1: for each pair of audio files, the subject selects the one causing more discomfort.

to give a score to each of them is not feasible. We preferred to set up a pairwise comparison task, as shown in Figure 1 and derived the full scale as described in the next section.

3.2.2 Protocol and Settings

For each kind of error, several test files are created with various error rates. The number of modified notes is parametered by the Modified Note Rate (MNR), which is set to either 10%, or 33%. For some kinds of error, the error intensity (EI) is also parametrized. This is quantified as a ratio of the note duration for duration changes and onset changes, and as a ratio of the MIDI velocity for loudness modifications. The EI is set to either 25%, or 75%. Modified notes are randomly chosen using the MNR. Intensity changes are made randomly, uniformly in the range centered on the true value and with the EI as radius.

To derive a ranking scale from pairwise comparisons, we choose the BTL method which uses hidden, "true" values associated to the transcriptions, along a given dimension (here, the discomfort). For a given pair of transcriptions, the subject's answer is a comparison of a noisy version of the two true values, the noise modeling the subjectivity and the variable skill of subjects. Thanks to this statistical framework, the full subjective scale is then obtained by processing all the pairwise comparisons. For this test, 20 pairs out of 812 are randomly chosen and presented to each subject for each musical excerpt. This number has been chosen in order to adjust the test duration and is not critical for the results, as long as the number of subjects is high enough.

3.3 Test 2: Subjective Evaluation of Transcriptions of Musical Pieces

Test 2 aims at obtaining a perceptive score for a series of transcriptions from several pieces of music. Three original excerpts from *Prelude in C minor BWV 847* by J.S. Bach (13 seconds), *Suite Bergamasque / III. Clair de Lune* by C. Debussy (20 seconds), and *Sonata in D Major KV 311 / I. Allegro con Spirito* by W.A. Mozart (13 seconds) were chosen. For each excerpt, five transcriptions are presented, as shown in Figure 2. The subject has to assign a non-negative value to each transcription. These values express the discomfort caused by transcription errors in comparison with its reference. The subject can listen as many times as needed to each transcription and reference.

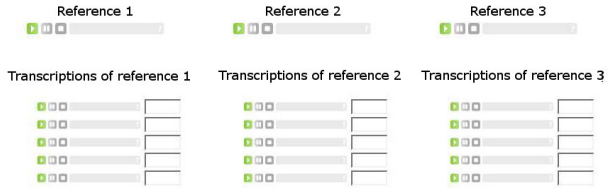


Figure 2. Test 2: the subject scores transcriptions with non-negative values.

In this test, all subjects are presented exactly the same audio files, in random order for each subject. One of the five transcriptions is the original piece in order to check whether the answers are consistent. The other four were obtained by automatic transcription systems, namely SONIC [12], available on the author's website, Bertin's system [2], a homemade system by P. Leveau based on [10] and an early version of [7]. The error rates and kinds of error thus depend on the specific behaviors of the transcription systems.

4 RESULTS

Thirty-seven subjects (24 musicians and 13 non-musicians) took part in this test. The results of Tests 1 and 2 are detailed here. The subjects' comments show that the instructions were understood correctly. They pointed out tone errors as a major cause of discomfort, while they seldom mentioned loudness and duration errors in an explicit way.

4.1 Test 1

Results of Test 1 are given in Figure 3. The BTL method makes it possible to obtain, from the pairwise comparisons of all the subjects, a subjective scale of discomfort for typical errors. A BTL perception value is thus assigned to each modification, which can be ordered according to this scale.

Different forms of evidence show the consistency of the obtained scale. First, increasing scores are obtained with increasing error rates, either MNR or EI, and decreasing harmonicity (octave, fifth, random pitches). Second, a minimum discomfort is obtained for the reference (taking into account its confidence interval). Third, as described in [6], the above 90% confidence intervals are related to a 5% risk. Thus, they are narrow enough to distinguish error types and to assert that the answers make sense, although adjacent error types should be considered perceptually equivalent.

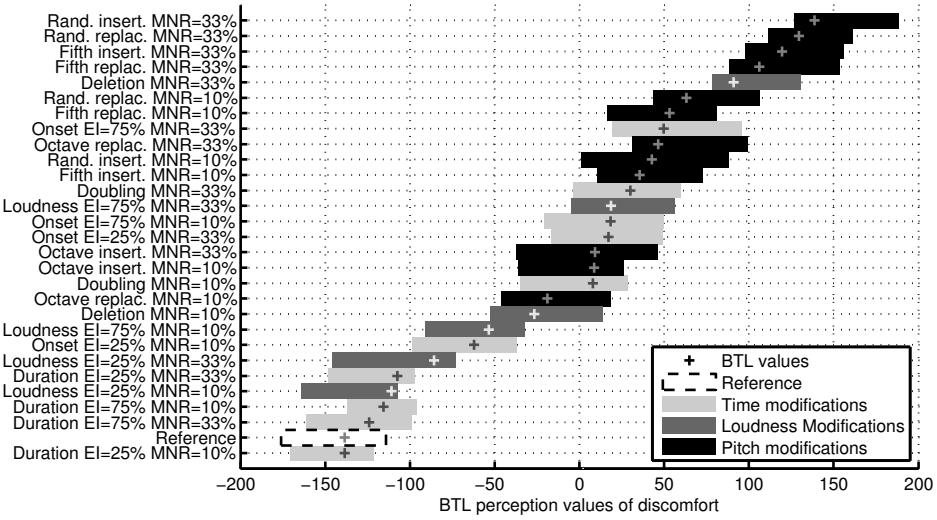


Figure 3. Test 1 : perceptive scale for typical errors. Crosses account for the related BTL value. Horizontal bars depict the 90% confidence intervals, obtained by a bootstrap method [6] using 100 resamplings of the data (because the data is not gaussian, confidence intervals may not be centered on BTL values).

Globally, as expected from the subjects' comments, the highest discomfort values are obtained with pitch modifications; loudness and time modifications cause low to medium discomfort. Regarding pitch changes, octave errors are judged much less serious than fifth changes, which cause a slightly lower discomfort than random changes. In each case, replacements and insertions are judged as equivalent, which would indicate that the discomfort is more induced by the added note than by the deleted note of a replacement. The lower values obtained for deletions confirm this hypothesis, which is commonly observed when working on transcription systems: a false negative usually sounds better than a false positive.

Results with time errors show that the modified onsets cause much more discomfort than duration changes. While one can expect that moving the beginning of an event causes a significant subjective change, it seems that subjects just did not perceive most of the modifications of duration. This can be explained by a specific feature of the piano: the ends of sounds generated from its freely-vibrating strings are less perceptible than for a musical instrument with forced vibrations. Thus current results for perception of note duration cannot be generalized to all instruments.

Finally, additional analysis of the results should be reported, which are not represented in Figure 3. First, similar results were obtained from subjects that were musicians and from non-musicians. Second, the three scales obtained for the three excerpts are also similar, with a little difference for the excerpt by Debussy in which deletions have a lower score and duration changes cause higher discomfort, probably because of the long durations in this slow piece of music.

4.2 Test 2

For each subject, scores of Test 2 are normalized by the maximum score he/she gave. Six subjects were removed since they scored a discomfort greater than 20% for the reference. Average scores and variances were then computed, with respect to the results from all the remaining subjects.

Results are represented in Figure 4. As the test is not a contest between existing algorithms, the systems were made anonymous, numbered from 1 to 4. The confidence in the results is assessed thanks to a 3 (composers) $\times$ 5 (algorithms) factorial ANOVA test, passed for each dimension and for interactions using a $p = 0.01$ test level. Thereby, the scores and the ranking of the algorithms are very dependent on the piece of music. This confirms that the performance of a transcription system is related to the musical content of pieces and thus depends on the test database. Large standard deviations indicate that the evaluation of a musical transcription depends greatly on proper subjective criteria. An important overlap between the answers makes it impossible to obtain a definitive ranking among the different algorithms even if for each excerpt, systems 2 and 3 are judged as the worst and the best one respectively.

5 EVALUATING THE TRANSCRIPTIONS

When comparing the results given by one of the objective evaluation methods proposed in Section 2 to the perceptive results of Test 1, several aspects are differentiated in the former case while they are not in the latter case, and vice versa. For instance, octave, fifth and random pitch changes have

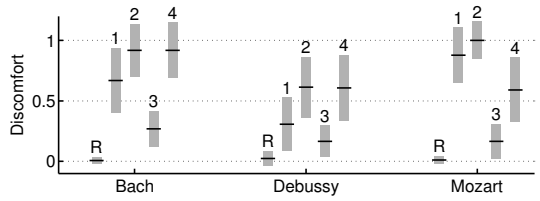


Figure 4. Results of Test 2: perceptive evaluation of the reference (R) and of transcriptions from four systems (1-4), with standard deviations (gray bars).

similar F-measures but cause an increasing discomfort. On the contrary, perceptive results are equivalent for replacements and insertions, whereas the F-measure is higher for insertions. Besides, perceptive results provide a balance between time and pitch influences, which is not taken into account in objective evaluation methods.

In this part, we estimate weighting coefficients of typical errors from Test 1 results, and we then apply them to adapt two existing metrics: the F-measure and the PTD. These modified methods are validated by applying them to the excerpts used in Test 2 and by comparing the results with the discomfort expressed by the subjects.

5.1 Extraction of the Weighting Coefficients

To extract the weighting coefficients, the results of Test 1 are normalized between 0 and 1. We only used results with MNR 33%, and the results were averaged for pitch modifications, insertions and replacements. Six criteria [1] are obtained, to be integrated into metrics. Their related weighting coefficients are given in the following table:

Octave	Fifth	Other intervals	Deletion	Duration	Onset
$\alpha_1 =$	$\alpha_2 =$	$\alpha_3 =$	$\alpha_4 =$	$\alpha_5 =$	$\alpha_6 =$
0.1794	0.2712	0.2941	0.2475	0.0355	0.4687

The coefficients are normalized so that $\frac{1}{3}\sum_{i=1}^{3}\alpha_i + \sum_{i=4}^{6}\alpha_i = 1$, since octave, fifth and random pitch represents alternative false positive errors.

5.2 Perceptive F-measure

In eq.(1), errors are the number of FP and FN, with an equal weight ($\frac{1}{2}$). We thus define the perceptive F-measure by:

$$f_{\text{percept}} \triangleq \frac{\#\text{TP}}{\#\text{TP} + \sum_{i=1}^{6}\alpha_i w_i \#E_i} \qquad (2)$$

[1] Loudness was not considered since the results were not satisfying, probably due to the difficulty of having a trustworthy loudness scale. Doubled notes were not used either because they could not be integrated into metrics.

where $\#E_i$ is the number of errors of type i (see below), $w_1 = w_2 = w_3 = w_4 = 1$, w_5 is the average duration error in the transcription, and w_6 is the average onset error. (The average errors are computed as the square root of the mean square error.) Note that a similar perceptive accuracy could be defined by using the equivalence mentioned in Section 2 and that the expression (2) equals the F-measure in the case $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{2}$ and $\alpha_5 = \alpha_6 = 0$.

Errors from MIDI files are extracted as follows:

1. TP are estimated as notes with correct pitch (rounded to the nearest semitone) and onset deviation lower than 150 ms. For each TP, the relative onset deviation and the relative duration deviation (both with respect to the original note parameters) are extracted. Then, let $\#E_5 = \#E_6 = \#\text{TP}$.

2. FP are transcribed notes which are not TP. The set of FP is split as follows: for each FP, (a) if there is an original note at the octave or sub-octave, at the same time (*i.e.* with any overlap of both time supports), the FP is added to the set E_1 of octave FP; (b) otherwise, if there is an original note at the upper or lower fifth at the same time, the FP is added to the set E_2 of fifth FP; (c) otherwise, the FP is added to the set E_3 of other pitch FP.

3. FN are the set E_4 of original notes that are not associated with one TP.

5.3 Perceptive PTD

The PTD is originally used to evaluate melodic similarities (see Section 2). In this context, note duration as weights to transfer and the euclidian distance in the time/pitch space seem to be appropriate choices. Nevertheless, when comparing generic musical pieces, both of these choices should be changed. PTD weights should be defined in a musical sense but this is beyond the scope of the current work and we thus chose to assign an equal and unitary PTD weight to each note. Using the perceptual coefficients introduced in Section 5.1, the distance between two notes is then defined in the multidimensionnal space of criteria composed of pitch (octave, fifth or others), duration and onset modifications.

5.4 Results

Figure 5 shows the results of the application of the original two objective measures and of their perceptive versions to the musical excerpts from Test 2. In order to compare them to the discomfort results, F-measures were changed by applying the function $x \mapsto 1 - x$. In order to best fit the discomfort scale, all results were scaled by a multiplicative coefficient obtained by minimizing the mean square error.

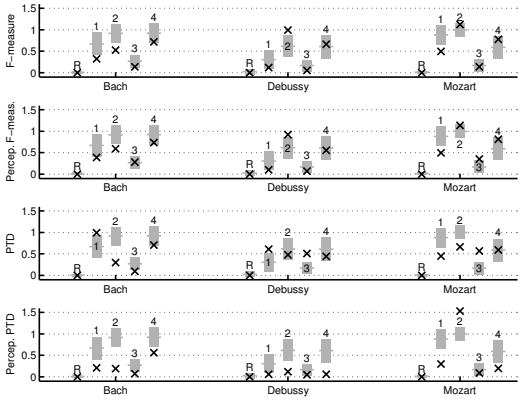


Figure 5. Transcription evaluation results with several objective and perceptive measures: in each case, crosses show the normalized error related to a measure, and the gray bars indicate the discomfort obtained in Test 2.

Results with the perceptive F-measure are slightly closer to the discomfort values than the original F-measure. Moreover, the ranking of the 15 excerpts is also closer to the discomfort-based ranking. Results of the perceptive PTD do not look better than the original, due to a high isolated value for the excerpt with highest discomfort (Mozart, System 2), that makes it difficult to scale the results adequately. However, the achieved ranking is dramatically better than the ranking by the original PTD, and also slightly better than the ranking by the perceptive F-measure. Thus, even if the relation between the discomfort and the perceptive PTD may be non-linear, the latter is appropriate in a ranking task.

6 CONCLUSIONS

The main idea of these tests was to get a ranking of the typical automatic transcription errors, to extract perception weights, and to integrate them into several musical sequence distance metrics. These primary results are consistent and the proposed perceptive metrics give satisfying results.

However further investigations should focus on a number of aspects, such as non-linear relations between specific error rates and discomfort, musical-based typical errors (taking into account tonality, melody, chords, etc.), and more specific algorithms to identify them.

References

[1] Multiple fundamental frequency estimation & tracking. In *Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.

[2] N. Bertin, R. Badeau, and G. Richard. Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. of ICASSP*, Honolulu, Hawaii, USA, April 2007.

[3] R.A. Bradley. Some statistical methods in taste testing and quality evaluation. *Biometrics*, 9(1):22–38, 1953.

[4] A.T. Cemgil, H.J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Trans. Audio, Speech and Lang. Proces.*, 14(2):679–694, 2006.

[5] S. Dixon. On the computer recognition of solo piano music. *Australasian Computer Music Conf.*, 2000.

[6] B. Efron and R. J. Tibshirani. An introduction to the bootstrap. In *London: Chapman & Hall*, 1993.

[7] V. Emiya, R. Badeau, and B. David. Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. In *Proc. of EUSIPCO*, Lausanne, Switzerland, August 2008.

[8] P. Hanna and P. Ferraro. Polyphonic music retrieval by local edition of quotiented sequences. In *Proc. of CBMI*, Bordeaux, France, June 2007.

[9] A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer, 2006.

[10] P. Leveau, E. Vincent, G. Richard, and L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Trans. Audio, Speech and Lang. Proces.*, 16(1):116–128, January 2008.

[11] V. I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1):8–17, 1965.

[12] M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Trans. on Multimedia*, 6(3):439–449, 2004.

[13] M. Mongeau and D. Sankoff. Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175, 1990.

[14] J.A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. Dept. of Music, Stanford University, 1975.

[15] G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 8:1–9, 2007.

[16] M. Ryynänen and A.P. Klapuri. Polyphonic music transcription using note event modeling. In *Proc. of WASPAA*, pages 319–322, New Paltz, NY, USA, 2005.

[17] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. van Oostrum. Using transportation distances for measuring melodic similarity. In *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003.

MULTIPITCH ESTIMATION OF QUASI-HARMONIC SOUNDS IN COLORED NOISE

Valentin Emiya, Roland Badeau, Bertrand David

GET - Télécom Paris (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France
valentin.emiya@enst.fr

ABSTRACT

This paper proposes a new multipitch estimator based on a likelihood maximization principle. For each tone, a sinusoidal model is assumed with a colored, Moving-Average, background noise and an autoregressive spectral envelope for the overtones. A monopitch estimator is derived following a Weighted Maximum Likelihood principle and leads to find the fundamental frequency (F_0) which jointly maximally flattens the noise spectrum and the sinusoidal spectrum. The multipitch estimator is obtained by extending the method for jointly estimating multiple F_0's. An application to piano tones is presented, which takes into account the inharmonicity of the overtone series for this instrument.

1. INTRODUCTION

Multipitch estimation is a critical topic for many applications, both in the field of speech processing (*e.g.* prosody analysis) [1] and in the context of musical signal analysis (*e.g.* automatic transcription) [2, 3]. The challenge offered by the spectral interference of the overtones of simultaneous notes has been taken up by various methods, some aiming at detecting a periodicity in the signal [4] or in its spectrum [5] while others use a combination of both spectral and temporal cues [6, 7]. Recent trends in the task include estimation in a bayesian framework [8] or in a perceptually compliant context [7]. The technique proposed in this paper is based on a Weighted Maximum Likelihood (WML) principle and belongs to the spectral estimators category.

This paper is organized as follows. Section 2 introduces the Maximum Likelihood principle applied to the proposed signal model. Section 3 then details the adaptation of the theoretical method to the multipitch estimation task in the case of piano sounds. Experimental results are given in section 4. Finally, conclusions are presented in section 5.

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales - DESAM, and by the French Ministry of Education and Research under the Music Discover project of the ACI-Masse de données

2. WEIGHTED MAXIMUM LIKELIHOOD PITCH ESTIMATOR

2.1. Main idea

This work focuses on signals which can be decomposed into a sum of sinusoidal components and a colored noise. In the following, a moving average process is assumed for the latter, with a corresponding FIR filter of transfer function $B(z)$. The spectral envelope of the partials is modeled by an autoregressive filter of transfer function $\frac{1}{A(z)}$. The technique presented herein is based on the decomposition of the set of DFT frequencies into two subsets: the subset $\mathcal{N}$ owing to the background noise properties and the other, $\mathcal{H}$, associated with the sinusoidal part. Once both $1/A(z)$ and $B(z)$ have been estimated, the constructed likelihood is maximized for the true value of F_0 since it simultaneously whitens the noise sub-spectrum and the sinusoidal sub-spectrum. In the case where a bad F_0 candidate is selected, the choice of a FIR $\mathcal{N}$-support sub-spectrum and an AR $\mathcal{H}$-support sub-spectrum ensures that such a flatness of both sub-spectra is not achieved.

2.2. Statistical framework

Let $\boldsymbol{x}$ denote the N-dimensional vector containing N successive samples of data, $\boldsymbol{X}$ the N-dimensional vector of its Digital Fourier Transform (DFT) and $\boldsymbol{F}$ the $N \times N$ orthonormal DFT matrix ($F_{(p,q)} = \frac{1}{\sqrt{N}} e^{-2i\pi \frac{pq}{N}}$). We assume that $\boldsymbol{x}$ results from the circular filtering of a centered white complex Gaussian random vector $\boldsymbol{w}$ of variance σ^2. Let $\boldsymbol{h}$ be the corresponding impulse response vector, and $\boldsymbol{H}$ its N-dimensional DFT vector. Since $\boldsymbol{X} = \mathrm{diag}\{\boldsymbol{H}\}\boldsymbol{F}\boldsymbol{w}$, $\boldsymbol{X}$ is a centered Gaussian random vector of covariance matrix $\sigma^2 \,\mathrm{diag}\{|\boldsymbol{H}|^2\}$.

Below, we consider that the observed data consist of a subset $\mathcal{S}$ of the DFT coefficients in vector $\boldsymbol{X}$. Then the previous discussion shows that the probability law of the

observed data is

$$p(X_{\mathcal{S}}) = \prod_{k\in\mathcal{S}} \frac{1}{\pi\sigma^2|H(k)|^2} e^{-\frac{|X(k)|^2}{\sigma^2|H(k)|^2}}.$$

Thus the normalized log-likelihood $L_{\mathcal{S}}(\sigma,\boldsymbol{h}) = \frac{1}{\#\mathcal{S}}\ln p(X_{\mathcal{S}})$ can be written in the form

$$L_{\mathcal{S}}(\sigma,\boldsymbol{h}) = C + \frac{1}{\#\mathcal{S}}\sum_{k\in\mathcal{S}}\left[\ln\left(\frac{|X(k)|^2}{\sigma^2|H(k)|^2}\right) - \frac{|X(k)|^2}{\sigma^2|H(k)|^2}\right] \quad (1)$$

where $C = -\frac{1}{\#\mathcal{S}}\sum_{k\in\mathcal{S}}\ln(\pi|X(k)|^2)$ is a constant with respect to σ and $\boldsymbol{h}$, and $\#\mathcal{S}$ denotes the number of elements in $\mathcal{S}$. Normalizing the likelihood by factor $1/\#\mathcal{S}$ aims at obtaining comparable, homogeneous values when $\#\mathcal{S}$ varies. Maximizing $L_{\mathcal{S}}$ with respect to σ yields the estimate

$$\widehat{\sigma}^2 = \frac{1}{\#\mathcal{S}}\sum_{k\in\mathcal{S}}\left|\frac{X(k)}{H(k)}\right|^2. \quad (2)$$

Then substituting equation (2) into equation (1) yields

$$L_{\mathcal{S}}(\boldsymbol{h}) \triangleq L_{\mathcal{S}}(\widehat{\sigma}^2,\boldsymbol{h}) = C - 1 + \ln\left(\rho_{\mathcal{S}}(\boldsymbol{h})\right) \quad (3)$$

where

$$\rho_{\mathcal{S}}(\boldsymbol{h}) = \frac{\left(\prod_{k\in\mathcal{S}}\left|\frac{X(k)}{H(k)}\right|^2\right)^{\frac{1}{\#\mathcal{S}}}}{\frac{1}{\#\mathcal{S}}\sum_{k\in\mathcal{S}}\left|\frac{X(k)}{H(k)}\right|^2} \quad (4)$$

is equal to the ratio between the geometrical mean and the arithmetical mean of the set $\left\{\left|\frac{X(k)}{H(k)}\right|^2\right\}_{k\in\mathcal{S}}$. Such a ratio is maximal and equal to 1 when $|X(k)/H(k)|$ is constant, independant of k, which means that $\rho_{\mathcal{S}}(\boldsymbol{h})$ measures the *whiteness*, or the *flatness* of $\left\{\frac{X(k)}{H(k)}\right\}_{k\in\mathcal{S}}$. The next step consists in choosing a parametric model for $\boldsymbol{h}$, and maximizing $L_{\mathcal{S}}$ with respect to the filter parameters. This optimization results in maximizing $\rho_{\mathcal{S}}(\boldsymbol{h})$. For instance, if $\boldsymbol{h}$ is modeled as an autoregressive (AR) filter, an approximate solution $\widehat{\boldsymbol{h}}$ to the optimization problem can be obtained by means of linear prediction techniques [9]. If $\boldsymbol{h}$ is modeled as a finite impulse response (FIR) filter of length $p \ll N$, an approximate solution $\widehat{\boldsymbol{h}}$ can be obtained by windowing a biased estimate of the autocovariance function.

2.3. Application to pitch estimation

Our pitch estimator relies on a weighted maximum likelihood (WML) method: for all subsets $\mathcal{H}$, *i.e.* for all possible F_0's, we calculate the weighted likelihood

$$L_{\mathcal{H}} = \alpha\ln\hat{\rho_{\mathcal{H}}} + (1-\alpha)\ln\hat{\rho_{\mathcal{N}}} \quad (5)$$

$$\text{with}\begin{cases}\hat{\rho_{\mathcal{H}}} &= \max\limits_{A}\rho_{\mathcal{H}}\left(\frac{1}{A(z)}\right)\\ \hat{\rho_{\mathcal{N}}} &= \max\limits_{B}\rho_{\mathcal{N}}\left(B(z)\right)\end{cases}$$

where $\mathcal{N} = \overline{\mathcal{H}}$ is the complement set of $\mathcal{H}$ and $0 < \alpha < 1$ (in practice we choose $\alpha = 1/2$). The pitch estimate is given by the set $\hat{\mathcal{H}}$ which maximizes $L_{\mathcal{H}}$. This maximum depends on the sum of the two $\mathcal{H}$-dependent terms in (5): $\ln\hat{\rho_{\mathcal{H}}}$ and $\ln\hat{\rho_{\mathcal{N}}}$. The flatness $\hat{\rho_{\mathcal{H}}}$ of the whitened components has a local maximum for a smooth spectral envelope, obtained when analyzing the true F_0 (see figure 1) or one of its multiples (*i.e.* $\mathcal{H}$ is a subset of the right set of overtones, see figure 2), or when $\mathcal{H}$ only contains noisy components. Low values of $\hat{\rho_{\mathcal{H}}}$ are obtained when amplitudes at the frequencies of $\hat{\mathcal{H}}$ are alternately low and high since AR filters have no zero, which means that they cannot fit a spectrum where some sinusoidal components in $\mathcal{H}$ are missing. This particularly happens for a sub-harmonic of the true F_0 (see figure 3). In other respects, when considering the spectral envelope of the noisy part of the sound, FIR filters have no pole, which means that they cannot fit any sinusoidal component: the spectral flatness $\hat{\rho_{\mathcal{N}}}$ of the whitened residual part reaches high values when the frequencies of overtones have been selected in $\mathcal{H}$, *i.e.* when analyzing any sub-harmonic frequency of the true F_0 (see figure 3). As illustrated in figure 4, by combining both spectral flatnesses $\hat{\rho_{\mathcal{H}}}$ and $\hat{\rho_{\mathcal{N}}}$, a global maximum is found for the true F_0 while any other local maximum in $\hat{\rho_{\mathcal{H}}}$ (or $\hat{\rho_{\mathcal{N}}}$) is attenuated by $\hat{\rho_{\mathcal{N}}}$ (or $\hat{\rho_{\mathcal{H}}}$), particularly harmonics and sub-harmonics.

3. APPLICATION TO MULTI-PITCH ESTIMATION OF PIANO TONES

3.1. Inharmonicity in piano tones

In a piano note, the stiffness of strings causes the frequencies of overtones to slightly differ from a perfect harmonic distribution. We are focussing one these quasi-harmonic sounds and exclude from this study other inharmonic tones like bell tones. The frequency of the overtone of order n is thus given by the inharmonicity law [10]:

$$f_n^{(f_0,\beta)} = nf_0\sqrt{1+\beta\left(n^2-1\right)} \quad (6)$$

where f_0 is the fundamental frequency and β is the inharmonicity coefficient. Note that β varies along the range of the piano keyboard and from one instrument to the other. Thus, the set $\mathcal{H}$, characterized by these two parameters, is defined as:

$$\mathcal{H}^{(f_0,\beta)} = \left\{f_n^{(f_0,\beta)}/n \in \mathbb{N}, f_n^{(f_0,\beta)} < F_s/2\right\} \quad (7)$$

Analysis of a synthetic signal with fundamental frequency 1076.6602 Hz.

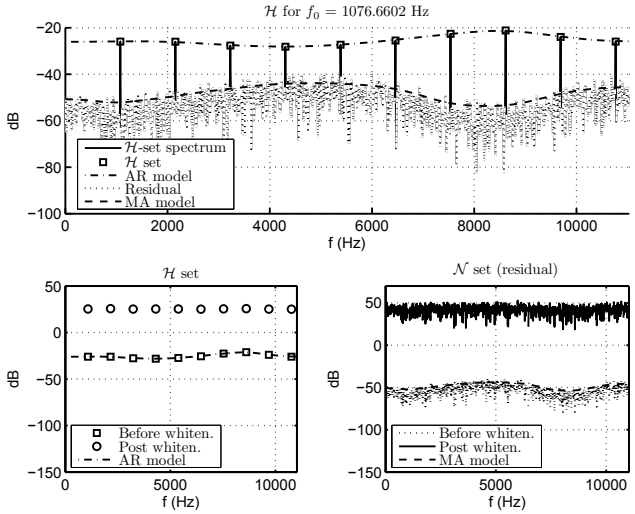


Figure 1: $L_{\mathcal{H}}$ estimation for $\mathcal{H} = \hat{\mathcal{H}}$ (true F_0). Overtones are selected in the spectrum (top), amplitudes of components fit the AR model (bottom left) and the residual spectrum is well whitened by the MA model (bottom right). In order to avoid overlapping between curves in the graphical representation, an constant offset is added to post-whitening dB-curves.

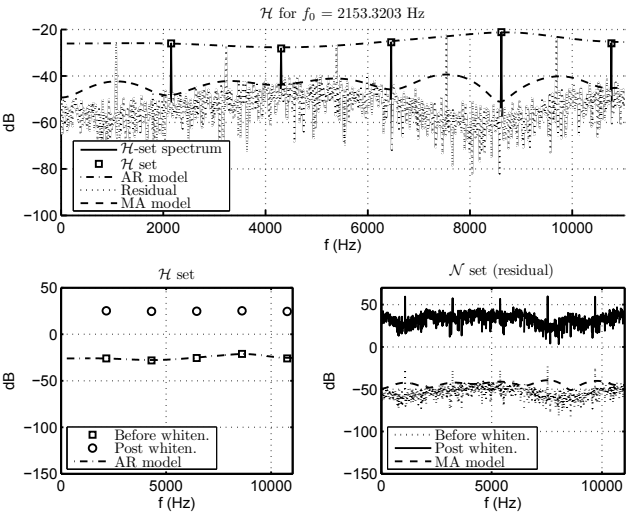


Figure 2: $L_{\mathcal{H}}$ estimation at twice the true F_0. Amplitudes of components fit the AR model whereas the residual spectrum is not perfectly whitened by the MA model, due to remaining components.

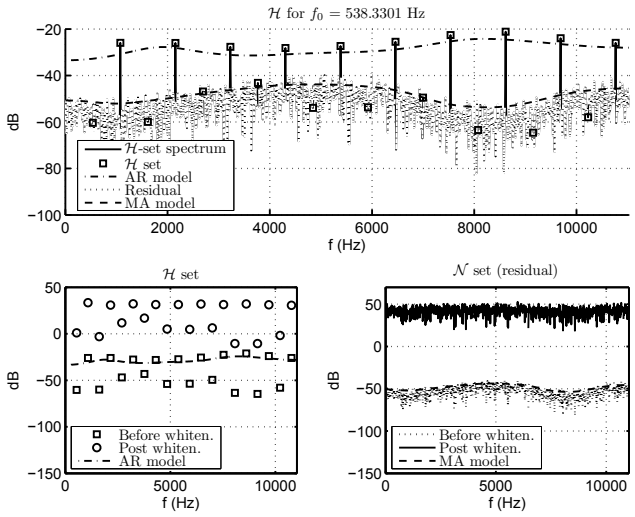


Figure 3: $L_{\mathcal{H}}$ estimation at half the true F_0. While residual spectrum is well whitened by the MA model, amplitudes of components do not fit the AR model, resulting in a low flatness of whitened amplitudes (bottom left, circles).

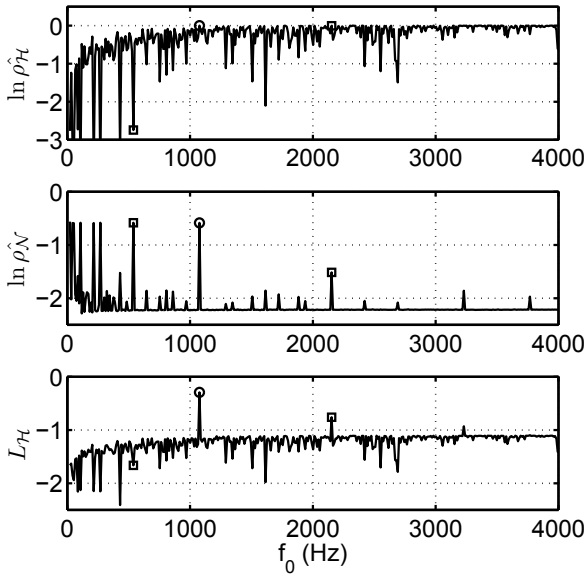


Figure 4: $\mathcal{H}$-dependent terms $\ln \hat{\rho_{\mathcal{H}}}$ (top) and $\ln \hat{\rho_{\mathcal{N}}}$ (middle), and weighted likelihood $L_{\mathcal{H}}$ (bottom), computed for all possible F_0's (*i.e.* all possible $\mathcal{H}$'s).

where F_s is the sampling frequency. Optimizing the log-likelihood $L\left(\mathcal{H}^{(f_0,\beta)}\right)$ with respect to $\mathcal{H}^{(f_0,\beta)}$ then consists in maximizing it with respect to f_0 and β.

3.2. From the theoretical model to real sounds

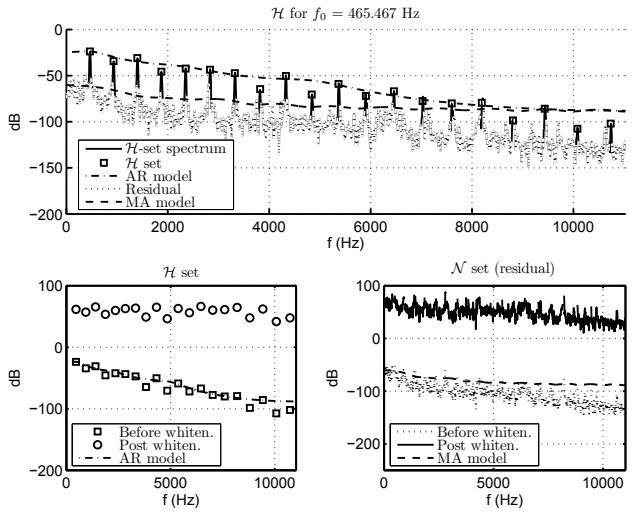


Figure 5: Real piano tone: separation between note components and residual part, and related MA and AR models

How do real piano tones fit the signal model described above? The AR model for the sinusoidal component, the MA noise model and the inharmonicity distribution of frequencies seem to be robust hypotheses. Conversely, the practical application of the method has to cope with two deviations from the theoretical point of view:

1. the assumption that f_n lies in the exact center of a frequency bin (multiple of $1/N$) is usually false, and spectral leakage thus influences the $\mathcal{N}$-support sub-spectrum.

2. the amplitude of the overtone may vary within the analysis frame, reflecting various effects as the energy loss of the sound and the beating between close adjacent components. This can affect the spectral envelope of the $\mathcal{H}$-support sub-spectrum.

The windowing of the analyzed waveform by a Hann window has proved to be a robust trade-off to overcome these issues. It prevents the spectral leakage associated with high energy components from masking weak overtones. Amplitudes of every overtone k are estimated by performing a parabolic interpolation of the spectrum (in decibels) based on the values in the nearest Fourier bins. The resulting (linear) value is used when computing the sinusoidal-part spectral flatness $\hat{\rho_{\mathcal{H}}}$, *i.e.* in place of $X(k)$ in equation (4). In order to minimize the effects described above in $\hat{\rho_{\mathcal{N}}}$ (see equation (4)), primary lobes of the frequencies selected in $\mathcal{H}$ are removed from $\mathcal{N}$, which is redefined as:

$$\mathcal{N} = \{k'/\forall f \in \mathcal{H}, |k'/N - f| > \Delta f/2\} \quad (8)$$

where Δf is the width of the primary lobe ($\Delta f = \frac{4}{N}$ for a Hann window). Note that the question of removing a set of components is a key step in the implementation of our algorithm. As shown in figure 5, the proposed method performs an approximate removal that offers a satisfying trade-off between efficiency and computational cost. Other techniques based on amplitude estimation and adapted filter design have been tested without bringing major improvements. The non-stationary nature of signals seems to be responsible for this limitation. It should be taken into account for enhancing the separation between a set of components and the residual signal.

3.3. Extension to polyphonic sounds

We now consider that the deterministic signal $s(n)$ is a sum of M inharmonic sounds: $s(n) = \sum_{m=1}^{M} s^{(m)}(n)$ and $\forall m \in \{1 \dots M\}$, $f_n^{(m)} = n f_0^{(m)} \sqrt{1 + \beta^{(m)}(n^2 - 1)}$, where $f_0^{(m)}$ is the pitch and $\beta^{(m)} > 0$ is the inharmonicity coefficient of the m^{th} tone. Each note is associated with one individual AR model, and weights in the likelihood are uniformly distributed among notes. Thus the WML principle consists in maximizing the log-likelihood:

$$L(\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(M)}) = \frac{1}{2M} \sum_{m=1}^{M} \ln \rho_{\mathcal{H}^{(m)}} \left(\frac{1}{A^{(m)}(z)} \right) + \frac{1}{2} \ln \hat{\rho_{\mathcal{N}}} \quad (9)$$

where $\mathcal{H}^{(m)} = \mathcal{H}^{\left(f_0^{(m)}, \beta^{(m)}\right)}$ and $\mathcal{N}$ is the set of bins outside primary lobes of frequencies of any $\mathcal{H}^{(m)}$. The optimization is performed with respect to each of the sets $\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(M)}$. Each set $\mathcal{H}^{(m)}$ is defined by the parameters $\{(f_0^{(m)}, \beta^{(m)})\}_{m \in \{1 \dots M\}}$ and $1/A^{(m)}(z)$ is the AR filter related to note m. Two distinct sets $\mathcal{H}^{(m_1)}$ and $\mathcal{H}^{(m_2)}$ may intersect, allowing overlap between spectra of notes m_1 and m_2. The algorithm presented in section 2.3 can be applied straightforwardly.

3.4. Multi-pitch estimator implementation

Multi-pitch estimation is often performed either in an iterative or in a joint process. The proposed method belongs to the joint estimation category. While iterative methods consist in successively estimating and removing a predominant F_0, joint estimation simultaneously extracts the set of

F_0's. Thus, a direct implementation of the algorithm described above would require to compute the ML of all possible combinations of notes, leading to a high-order combinatory task. For instance, more than 2.10^6 different chords exist for a 4-note polyphony in the full piano range, each of these candidates requiring several calls to the likelihood function since the exact F_0 and β values are unknown.

In order to reduce the cost of the ML estimation, a two-step algorithm is proposed. First, each possible chord is evaluated on a reduced number of points N_p in the $(f_0^{(m)}, \beta^{(m)})$ region around F_0 values from the well-tempered scale and approximate β values. N_{cand} chord candidates are extracted among all combinations by selecting the N_{cand} greatest likelihood values. Then, the likelihood of each selected candidate is locally maximized with respect to coefficients $f_0^{(m)}$ and $\beta^{(m)}$. A simplex method is used to perform this optimization, which is initialized with the $f_0^{(m)}$ and $\beta^{(m)}$ values selected during the first step. Finally, the chord with maximum accurately-computed likelihood is selected as the chord estimate.

4. EXPERIMENTAL RESULTS

The algorithm has been tested on a database composed of about 540 isolated piano tones of the RWC database [11] and random chords generated by several virtual piano softwares based on sampled sounds. About 600 two-note chords and 600 three-note chords were evaluated. In each case, the polyphony is known a priori by the algorithm and the estimation results from the analysis of one 93 ms frame, beginning 10 ms after the onset. F_0 estimates are rounded to the nearest half-tone in the well-tempered scale in order to determine if an estimated note is correct. This approximation on F_0 is carried out in order to evaluate the pitch estimation at a note level rather than at a frequency level. The note search range spreads over 5 octaves, from MIDI note 36 ($f_0 = 65$ Hz) to MIDI note 95 ($f_0 = 1976$ Hz). These test conditions are similar to the ones used in competitor systems [4, 5, 7] in terms of frame length, F_0 search range and error rate definition.

The parameters of the system have been adjusted as follows. Sounds are sampled at 22050 Hz. DFT are computed on 4096 points after zero-padding the 2048-point frame. The AR model order is set to 8, the MA model order to 20. In the first step of the implementation described in section 3.4, all chord combinations are evaluated, each one with $N_p = 10$ (polyphony ≤ 2) or $N_p = 5$ (polyphony three) different $(f_0^{(m)}, \beta^{(m)})$ values. Then $N_{cand} = 75$ (monophony) or $N_{cand} = 150$ (polyphony ≥ 2) chord candidates are selected for the second step.

Error rates are 2.0% in monophony, 7.5% in polyphony two and 23.9% in polyphony three. They are reported in

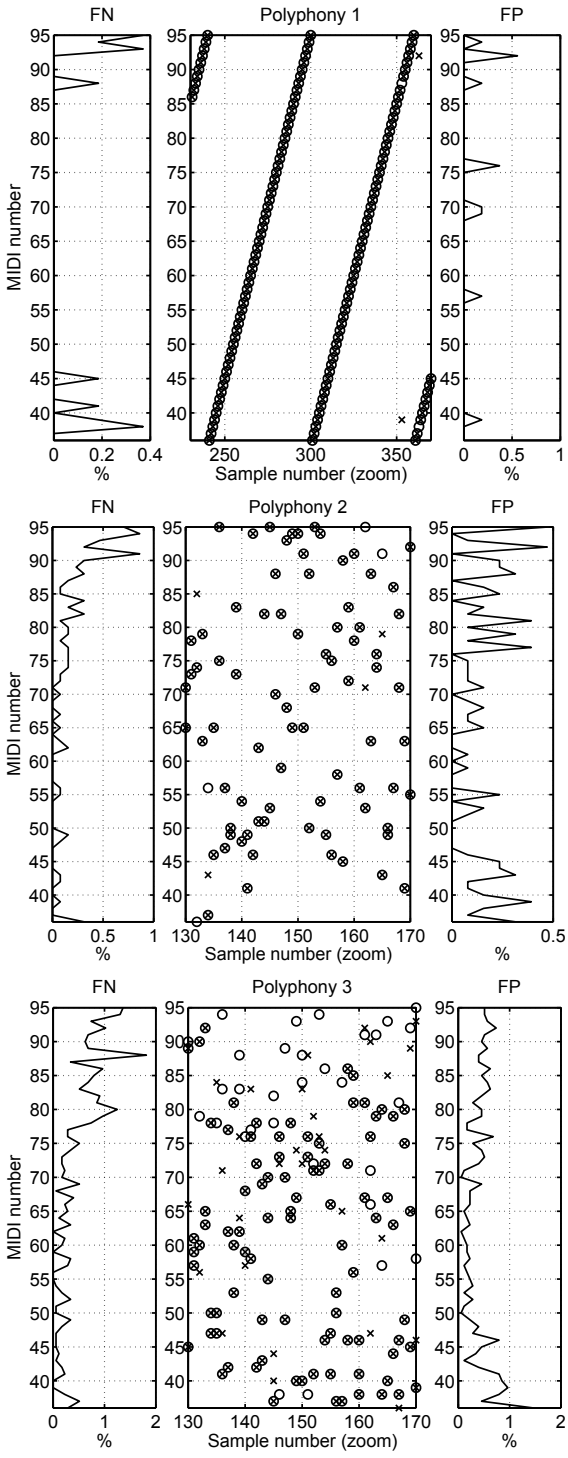


Figure 6: Estimation results: for a given polyphony (1 to 3 from top to bottom), random chords are generated (circles) and estimated (crosses). For visual representation clarity, only 50 samples of them are shown (center). Distribution of false negatives is displayed on the left. Distribution of false positives is displayed on the right.

Polyphony	1	2	3
Error rate	2.0% $\pm 0.6\%$	7.5% $\pm 1.1\%$	23.9% $\pm 2.2\%$
Octave error rate	0%	1.6%	5.2%
State of the art	$2 \sim 11\%$	$7 \sim 25\%$	$\approx 10 \sim 35\%$

Table 1: Error rates with respect to polyphony. Lower and upper bounds of state-of-the-art performances are also reported. Confidence interval is derived as the standard deviation of the error rate estimator.

table 1 and can be compared to the three competitor systems previously mentionned. Their performances have been established in [7] for polyphony one, two, four and six: error rates vary from 2 to 11% in monophony, from 7 to 25% in polyphony two and from 14 to 41% in polyphony four. Error rates in polyphony three are not given, but could be figured out as intermediate values between results in polyphony two and four, which would lead to approximate error rates between 10 and 35%. The proposed pitch estimator is comparable to competitor systems in terms of performance. Error rates are particularly competitive in polyphonies one and two.

The evaluation task has been performed using randomly uniformly-distributed notes in order to provide experimental results from an objective point of view rather than from musical considerations. The distribution of errors is reported in figure 6. The few errors in polyphony one occur in the lowest and highest pitch regions. In polyphony two and three, most of missed notes (or false negatives, FN) are located in the treble part of the piano range whereas the false-alarm notes (or false positives, FP) estimated in place of them tend to be distributed in a more uniform manner along the piano range. Closely-spaced chords in the medium range seem easier to detect than widely-spaced chords. Octave error are scarce – around one fifth of all errors for each polyphony number –, which can be explained by the complementary contributions of note and noise likelihoods. On the contrary, high-pitched FN and large-interval errors often occur, in spite of the likelihood normalization stage, due to the sensitivity of the ML approach to the variable number of frequency parameters that depends on F_0 candidates.

5. CONCLUSIONS

The multipitch estimation task has been performed here through a Maximum Likelihood approach. It consists in modeling notes and residual noise by AR and MA models, and results in a criterion on their spectral flatness after a whitening process based on the models. The method has been validated by satisfying experimental results for polyphony one to three.

Future works will deal with managing the overlap between notes spectra, with improving the model for the spectral envelope of notes and with making the computational cost decrease in order to both benefit from the efficiency of the estimator and avoid the inherent complexity of joint estimation of multiple F_0's.

6. REFERENCES

[1] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[2] M. Ryynänen and A.P. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 319–322.

[3] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[4] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[5] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, November 2003.

[6] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. of ICASSP 2006*, Toulouse, France, May 14-29 2006, IEEE, vol. 5, pp. 53–56.

[7] A.P. Klapuri, "A perceptually motivated multiple-f0 estimation method," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 291– 294.

[8] Manuel Davy, Simon Godsill, and Jerome Idier, "Bayesian analysis of polyphonic western tonal music," *JASA*, vol. 119, no. 4, pp. 2498–2517, 2006.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[10] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[11] T. Nishimura M. Goto, H. Hashiguchi and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003, pp. 229–230.

A PARAMETRIC METHOD FOR PITCH ESTIMATION OF PIANO TONES

Valentin EMIYA, Bertrand DAVID, Roland BADEAU

Ecole Nationale Supérieure des Télécommunications - Département TSI
46 rue Barrault, 75634 PARIS cedex 13 - France

ABSTRACT

The efficiency of most pitch estimation methods declines when the analyzed frame is shortened and/or when a wide fundamental frequency (F_0) range is targeted. The technique proposed herein jointly uses a periodicity analysis and a spectral matching process to improve the F_0 estimation performance in such an adverse context: a 60ms-long data frame together with the whole, $7^1/_4$-octaves, piano tessitura. The enhancements are obtained thanks to a parametric approach which, among other things, models the inharmonicity of piano tones. The performance of the algorithm is assessed, is compared to the results obtained from other estimators and is discussed in order to characterize their behavior and typical misestimations.

Index Terms— audio processing, pitch estimation

1. INTRODUCTION

Numerous methods dedicated to fundamental frequency (F_0) estimation of periodic signals try to extract the signal self-similarities by maximizing a function of time or frequency. In this manner, they measure a degree of internal resemblance in the waveform (ACF [1, 2], AMDF [3, 4], cepstrum [5]) or in the spectrum [6]. When processing real world musical sounds, these techniques are confronted to deviations from the theoretical model, such as the presence of noise, which can be both stationary and non stationary, or the possibly non-uniform distribution of the harmonics.

The development and applications of the quoted methods often deal with an extension to subband processing [2, 7], to an optimization of the main function [4, 7] or to the joint use of both time and frequency domains [8]. Typical errors that usually occur give a general idea of the difficulties the F_0 estimation task must cope with. Temporal or spectral methods tend to make sub-octave or octave errors respectively. Both of them come up against difficulties like a large F_0 search range (*e.g.* 27-4200 Hz for the piano), non-regular spectral envelopes and inharmonic deviations of the frequency components [6, 9]. In addition, a short analysis frame prevents spectral methods from resolving components for low F_0 values whereas the uniformely-distributed discrete time scale used by temporal methods makes the estimation fail above some F_0 limit.

The new F_0 estimation algorithm we describe aims at enhancing F_0 estimation results in the case of a short analysis window and a large F_0 search range. We will focus on piano sounds since they present all the listed difficulties and usually cause one of the worst estimation error rates per instrument (*e.g.* see [8]). The pitch of a harmonic or quasi-harmonic sound is an attribute that only depends on the sinusoidal components of the signal. Thus a F_0 estimator only requires the parameters of components such as frequency, amplitude, damping factor and initial phase. So far, the other part of the sound, including the ambient noise, transients, etc. has not been used in the F_0 estimation task, as far as the authors know. Therefore, the preliminary task in the F_0 estimation method we present consists in extracting the parameters of components. The F_0 estimator then includes a spectral function and a temporal function. The parametric approach enables to take into account the inharmonicity of sounds both in time and frequency domains and to optimize the precision of the F_0 numeric estimation.

The F_0 estimation system is described in section 2. Evaluation results and comparisons with other algorithms are then detailed in section 3 and conclusions are finally presented in section 4.

The research leading to this paper was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décomposition en Éléments Sonores et Applications Musicales - DESAM

2. PITCH ESTIMATION SYSTEM

2.1. High Resolution analysis

The N_a-length analyzed waveform is modeled by:

$$s(t) \quad = \quad \sum_{k=1}^{K} \alpha_k z_k^t + w(t) \qquad (1)$$

defined for $t \in [\![0, N_a - 1]\!]$ and composed of a sum of K exponentially-modulated sinusoids $\alpha_k z_k^t, k \in [\![1, K]\!]$ with complex amplitudes $\alpha_k = A_k e^{i\Phi_k} \in \mathbb{C}^*$, ($A_k$ being the real, positive amplitude and Φ_k the initial phase), and distinct poles $z_k = e^{d_k + i2\pi f_k}$ (f_k being the frequency and d_k the damping factor), plus an additive colored noise $w(t)$. This section details how the signal is preprocessed, how poles z_k are then estimated via the ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) algorithm [10], and how amplitudes α_k are finally extracted.

Preprocessing. A two-step preprocessing stage is applied to the signal sampled at 32 kHz:

1. The cubic computational cost of the ESPRIT algorithm is reduced when the number of poles to be estimated is low. This is achieved by using a filter bank. The signal is splitted into $D = 32$ subbands with width 500-Hz by using cosine-modulated filters [11]. The order of magnitude of the computational cost drops from N_a^3 to N_a^3/D^2 (N_a^3/D^3 per band) leading to a satisfactory processing time for the analysis bloc.
2. Components of piano sounds are particularly well represented by the exponential sinusoidal plus noise model introduced in (1). However, the ESPRIT algorithm only applies to the restrictive case of white noise. Thus, the second preprocessing step consists in whitening the noise in each subband thanks to an AR filter estimated on the smoothed spectrum of the signal.

ESPRIT algorithm. The signal in each preprocessed subband is a sum of exponentially-modulated sinusoids plus white noise. Assuming the number of poles is known, the ESPRIT algorithm [10]

gives an estimation of those poles. The method is based on a subspace projection on the so-called signal subspace and benefits from the rotational invariance property of this signal subspace.

Estimation of the number of poles. In the current application, the number of poles in each subband is not known a priori and thus must be estimated. The ESTER [12] algorithm establishes a criterion $J(p)$ that provides an estimation of the number of poles as $\text{argmax}_{p\in P}\,(J(p) > \delta_J)$, P being the set of candidates for the number of poles and δ_J a threshold tuned to $\delta_J = 10$ in the current study. The result obtained by this method is either correctly estimated, or slightly over-estimated. As shown in [12], the latter case is not disturbing for the ESPRIT analysis, and weak amplitudes are estimated for the spurious poles.

Estimation of amplitudes. Once the poles extracted, amplitudes are estimated by a least squares algorithm applied to the subband signal. The effects of the preprocessing stage on the amplitudes in each subband are corrected by applying the inverse filters of the various preprocessing steps – whitening, filter bank and preemphasis filter series –, leading to the estimation of the amplitudes $\alpha_k, k \in [\![1, K]\!]$.

2.2. Pitch estimation

A temporal method and a spectral method are first introduced. Although each one could account for a F_0 estimator, they are jointly used in the same manner as in [8] to obtain the whole, more efficient estimator detailed in the last part.

2.2.1. Temporal method

Periodicity is often analyzed by assuming the signal is an observation of a real, wide-sense stationary (WSS) process y and by estimating its autocovariance function $R_y(\tau) = \mathbb{E}\left[y(t)y(t+\tau)\right]$. When the signal is periodic, the maxima of $R_y(\tau)$ are located at $\tau = 0$ and at every multiple of the period. Let us consider a real, WSS process y composed of K undamped sinusoids with frequencies ν_k, real amplitudes $2a_k$, initial phases φ_k, which are assumed to be independant and uniformely distributed along $[0, 2\pi[$. The autocovariance function of y is $R_y(\tau) = \sum_{k=1}^{K} 2a_k^2 \cos(2\pi\nu_k\tau) + \delta(\tau)\sigma_{w_y}^2$. Therefore we can define a temporal function $R(\tau)$ for F_0 estimation from the parameters estimated by the high resolution analysis:

$$R(\tau) = \sum_{k=1}^{K} p_k \cos(2\pi f_k \tau) \quad (2)$$

$$p_k = \begin{cases} |\alpha_k|^2 & \text{if } |z_k| = 1 \\ \frac{|\alpha_k|^2}{N_a}\frac{1-|z_k|^{2N_a}}{1-|z_k|^2} & \text{otherwise} \end{cases} \quad (3)$$

where $\tau > 0$, $f_k = \frac{\arg(z_k)}{2\pi}$ is the normalized frequency of component k, and the instantaneous power p_k is an estimate of coefficient $2a_k^2$ over the analysis frame.

In the case of a slightly inharmonic sound, the frequency deviation weakens or even removes the maxima of $R(\tau)$ at the multiples of the period. The inharmonicity law [13] for a piano tone of fundamental frequency f_0 causes partial h not to be located at frequency hf_0 but at $hf_0\sqrt{1+\beta(h^2-1)}$, β being the inharmonicity coefficient of the note. As illustrated in fig. 1, this frequency stretching may be inversed by remapping the set of estimated frequencies $\{f_k, k \in [\![1, K]\!]\}$ to a set of frequencies $\{g_{f_0,k}, k \in [\![1, K]\!]\}$:

$$g_{f_0,k} = \frac{f_k}{\sqrt{1+\beta(f_0)\left(h^2(f_0, f_k)-1\right)}} \quad (4)$$

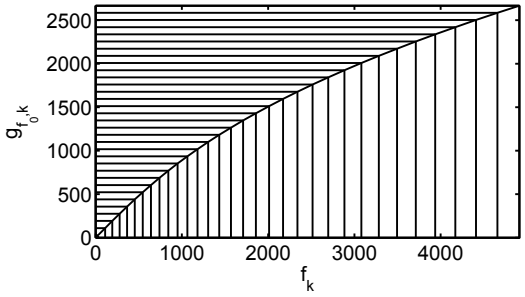


Fig. 1. At any given F_0, the frequencies f_k are remapped to $g_{f_0,k}$, leading to a harmonic distribution for the actual F_0. One theoretical partial over 5 is represented with $f_0 = 27.5$Hz and $\beta = 2.54e-4$.

where $\beta(f_0)$ is an approximative inharmonicity coefficient for fundamental frequency f_0 averaged from the results presented in [13, pp. 365]. The assumed partial order $h(f_0, f_k)$ associated to frequency f_k is extracted from the inharmonicity law:

$$h^2(f_0, f_k) = \frac{\sqrt{(1-\beta(f_0))^2 + 4\beta(f_0)\frac{f_k^2}{f_0^2}} - 1 + \beta(f_0)}{2\beta(f_0)} \quad (5)$$

As the remapping process causes the remapped frequencies $g_{f_0,k}$ of the partials to be perfect multiples of the actual fundamental frequency f_0, we replace f_k with $g_{\frac{1}{\tau},k}$ in (2) to obtain a temporal function $R_{\text{inh}}(\tau)$ for piano tones which is maximum for $\tau = \frac{1}{f_0}$:

$$R_{\text{inh}}(\tau) = \sum_{k=1}^{K} p_k \cos\left(2\pi g_{\frac{1}{\tau},k}\tau\right) \quad (6)$$

2.2.2. Spectral method

A parametric amplitude spectrum is designed from the estimates of frequencies f_k and energies E_k of components $k \in [\![1, K]\!]$. It is composed of a sum of K gaussian curves centered in f_k with constant standard deviation σ, weighted by the square root of the component energies as average amplitudes:

$$S(f) = \sum_{k=1}^{K} \frac{\sqrt{E_k}}{\sqrt{2\pi}\sigma} e^{-\frac{(f-f_k)^2}{2\sigma^2}} \quad (7)$$

σ is set to $f_{0\text{min}}/4$ where $f_{0\text{min}}$ is the lower bound of the F_0 search range in order to prevent overlap between successive partials.

Our spectral estimator $U(f)$ relies on maximizing a scalar product between the parametric amplitude spectrum and theoretical harmonic unitary patterns of F_0 candidates:

$$U(f) = \sum_{h=1}^{H_f} w_{f,h} S(hf) \quad (8)$$

where H_f is the maximum number of partials possible for fundamental frequency f and $\{w_{f,h}, h \in [\![1, H_f]\!]\}$ is the pattern associated to f. The choice of the pattern is based on an approximative logarithmic spectral decrease of components. The slope p of a linear

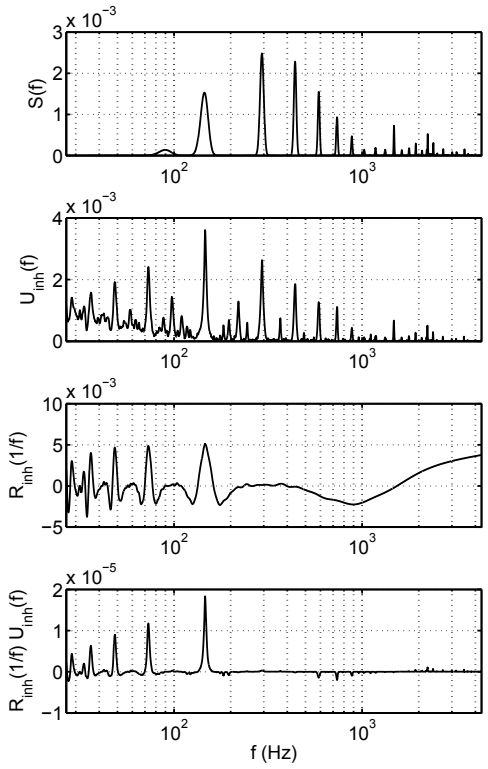


Fig. 2. From top to bottom, on a logarithmic frequency scale: parametric spectrum, spectral estimation function $U_{\text{inh}}(f)$, remapped temporal estimation function $R_{\text{inh}}\left(\frac{1}{f}\right)$, joint F_0 estimation function. Functions result from the 60 ms analysis of a D3 piano note.

regression between $\log(\sqrt{E_k})$ and f_k is extracted and weights $w_{f,h}$ are then defined as:

$$w_{f,h} \quad = \quad w_0 e^{phf} \tag{9}$$

where $w_0 = \left(\sum_{h=1}^{H_f} e^{2phf}\right)^{-\frac{1}{2}}$ is a normalizing term such that $\sum_{h=1}^{H_f} w_{f,h}^2 = 1$.

The spectral estimator is then adapted to piano tones by selecting the values of the spectrum on an inharmonic stretched scale instead of a harmonic scale:

$$U_{\text{inh}}(f) \quad = \quad \sum_{h=1}^{H_f} w_{f,h} S\left(hf\sqrt{1+\beta(f)\left(h^2-1\right)}\right) \tag{10}$$

Finally, the estimator efficiency can be improved by ignoring all frequencies and weights below a cut-off frequency of 100 Hz since the impedance at the piano bridge [13] causes a significant deviation of low frequencies from the inharmonicity law and the highest weigths $w_{f,h}$ of patterns are allocated to those frequencies.

2.2.3. Pitch estimator

As mentioned in the introduction, sub-harmonic and harmonic error trends are opposed in temporal and spectral methods. A way to benefit from this phenomenon is described in [8]. It consists in multiplying a temporal and a spectral function on a common F_0 scale in order to preserve common peaks from both functions and to remove or attenuate other peaks (see fig. 2). Thus, the pitch is estimated by maximizing the product of the methods $R_{\text{inh}}\left(\frac{1}{f}\right)$ and $U_{\text{inh}}(f)$:

$$\hat{f}_0 \quad = \quad \text{argmax}_f \left(R_{\text{inh}}\left(\frac{1}{f}\right) U_{\text{inh}}\left(f\right)\right) \tag{11}$$

Thanks to the analytic expressions (6) and (10), $R_{\text{inh}}\left(\frac{1}{f}\right)$ and $U_{\text{inh}}(f)$ can be directly evaluated for any f value. As the F_0 distribution of an equal-tempered musical scale is logarithmic, the F_0-scale support is set to N_f points logarithmically spaced in the F_0-search range. This unconstrained choice is a key advantage of the method since the logarithmic F_0 distribution is not offered by many methods (see [4, 8]). Actually, temporal methods have a linearly distributed time scale, which results in a lack of precision in high frequency and too much resolution in low frequency, whereas Fourier-based spectral methods have a linear F_0 distribution. In those cases, the estimation function must often be interpolated to achieve the required precision and may still suffer from this.

In a Matlab implementation on a 2.4GHz-CPU, the overall processing of a 60ms frame averages 6.5s. About 1s is necessary for the analysis. About 95% of the remaining time is required by the spectral F_0 estimator and may be optimized and written in C for a computationally-efficient implementation.

3. EVALUATION

The algorithm has been evaluated on isolated piano tones from various sources: 3168 notes from three pianos of RWC database [14], 270 notes from five pianos of a PROSONUS database and 264 notes from a Yamaha upright piano of a private database. All recordings include several takes of all the 88 notes of piano range (except PROSONUS in which notes are spaced by fourth) with a varying loudness. RWC samples also offer various play modes (normal, staccato, with pedal). The F_0 search scale is composed of $N_f = 8192$ values logarithmically distributed between $f_{0\text{min}} = 26.73$ Hz and $f_{0\text{max}} = 4310$ Hz. The estimation is performed after the analysis of a single 60 ms or 93 ms frame: 60 ms is quite a challenging frame length since it is below twice the period of lowest notes while 93 ms is a well spread duration for this kind of evaluation. Each estimated F_0 is associated to the closest note in the equal tempered scale with A4 tuned to 440 Hz. Errors are then defined as incorrect note estimations. The method is compared to two estimators. The first one is as similar to our estimator as possible, replacing the ESPRIT analysis stage with a classical analysis: the ACF is estimated from the signal by the formula $r(\tau) = \frac{N_a}{N_a-\tau}\text{DFT}^{-1}\left[\left|\text{DFT}\left[s\right]\right|^2\right]$, the factor $\frac{N_a}{N_a-\tau}$ being a correction of the bias; the spectral estimator $U_{\text{inh}}(f_0)$ is computed by replacing the parametric spectrum with the modulus of the DFT of the signal, using a zero-padding on $8N_f$ points; $r(\tau)$ is mapped to the frequency scale by interpolation as described in [8]; the pitch is finally estimated by maximizing the product between the spectral function and the remapped $r(\tau)$. The second method is the YIN algorithm [4] which is considered as a very efficient monopitch estimator. We used the code available on the authors' website.

Evaluation results are reported in fig. 3. At the target window length of 60 ms, the global error rate of our estimator is around 4.4% which is at least twice better than the other estimators. This is due to a low error rate on a large F_0 range (1.1% in the F_0 range $65-2000$ Hz) and slowly increasing values at the very bass and treble limits. In comparison, the non-ESPRIT based estimator achieves

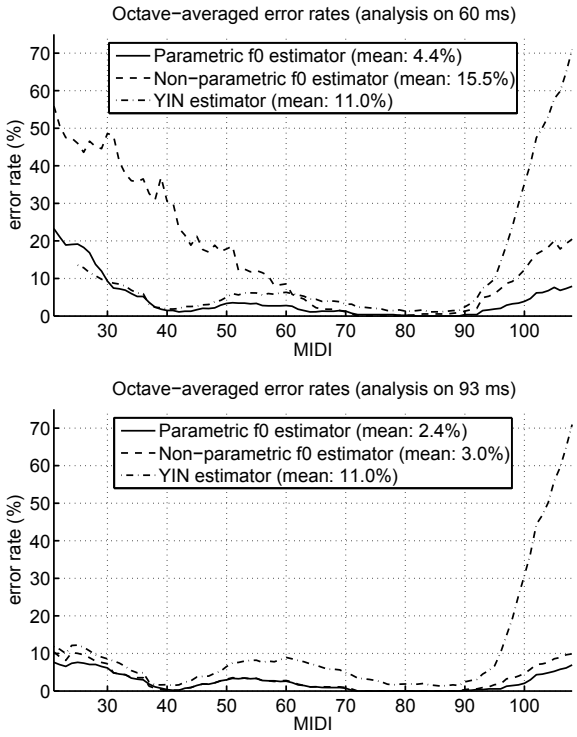


Fig. 3. Octave-averaged error rates per note with two different frame lengths, for the parametric F_0 estimator and two other methods: a similar but non-parametric algorithm and the YIN estimator

a 1.1% error rate in the range $240 - 2000$ Hz. Its low efficiency outside this range shows how the F_0 estimation is improved by both the high resolution analysis and the handling of parametric, analytic formulas. The YIN algorithm is slightly less efficient in the medium range than our estimator and has similar results in the bass range (for the first octave both curves should be at the same level, but our estimator results seem to be worse since they include the lowest four note error rates that cannot be estimated by the YIN algorithm with a 60 ms window length). In the high range, it shows a quite high error rate, which is a typical behavior of temporal methods. Global results are improved with a 93 ms frame length. Nevertheless, the high resolution analysis does not enhance significatively the F_0 estimation even if its error rate remains the lowest.

Typical errors are now discussed, in the 60 ms analysis case. As expected, usual errors are under-estimations of high f0s and over-estimations of low f0s. Around 18% of errors made by each algorithm are octave and suboctave errors. In the case of our algorithm, the remaining error intervals are of all kinds, with only 5% that are half-tone errors, whereas this rate reaches 10% for the other two algorithms. The YIN algorithm makes a high proportion of sub-harmonic errors (13% are sub-octaves, 8% are sub-nineteenth). Thus, even if our algorithm makes a reduced number of harmonic/subharmonic errors, those errors remain difficult to avoid. Half-tone error rates show the efficiency of our method while the other algorithms suffer from a lack of precision of temporal estimators for high F_0. Finally, the inharmonicity management contributes to lower the global error rate, from 4.9 to 4.4% in the 60-ms frame case. As expected, the improvement is localized in the lowest F_0 range: the error rate in the MIDI range $[\![21, 37]\!]$ decreases from 16.6 to 14.1%.

4. CONCLUSIONS

The F_0 estimator designed in this paper enables to address typical error trends in a short frame analysis and a wide F_0-range context. It is based on a preliminary extraction of the parameters of components and on the design of temporal and spectral parametric function. Satisfying performances have been obtained and a large part was allocated to the discussion on typical errors and the way to avoid them.

5. REFERENCES

[1] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[2] Ray Meddis and Michael J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *JASA*, vol. 89, no. 6, pp. 2866–2882, 1991.

[3] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[4] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] A. Michael Noll, "Cepstrum pitch determination," *JASA*, vol. 41, no. 2, pp. 293–309, 1967.

[6] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, November 2003.

[7] A.P. Klapuri, "A perceptually motivated multiple-f0 estimation method," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 291– 294.

[8] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. of ICASSP 2006*, Paris, France, May 14-29 2006, IEEE, vol. 5, pp. 53–56.

[9] S. Godsill and M. Davy, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2005, IEEE, pp. 283– 286.

[10] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT–a subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1340–1342, 1986.

[11] P. P. Vaidyanathan, *Multirate systems and filter banks*, Englewoods Cliffs, NJ, USA: Prentice Hall, 1993.

[12] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. on Signal Processing*, vol. 54, no. 2, pp. 450–458, February 2006.

[13] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[14] T. Nishimura M. Goto, H. Hashiguchi and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR*, Baltimore, Maryland, USA, 2003, pp. 229–230.

Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Harmonic plus noise decomposition: time-frequency reassignment versus a subspace based method

Bertrand David[1], Valentin Emiya[1], Roland Badeau[1] and Yves Grenier[1]

[1] *ENST, dept TSI, 46 rue Barrault, 75634 Paris Cedex 13, France*

Correspondence should be addressed to Bertrand David (bertrand.david@enst.fr)

ABSTRACT

This work deals with the Harmonic+Noise decomposition and, as targeted application, to extract transient background noise surrounded by a signal having a strong harmonic content (speech for instance). In that perspective, a method based on the reassigned spectrum and a High Resolution subspace tracker are compared, both on simulations and in a more realistic manner. The reassignment re-localizes the time-frequency energy around a given pair (analysis time index, analysis frequency bin) while the High Resolution method benefits from a characterization of the signal in terms of a space spanned by the harmonic content and a space spanned by the stochastic content. Both methods are adaptive and the estimations are updated from a sample to the next.

1. INTRODUCTION

In the context of musical signal processing [1], or audio coding (*cf.* MPEG4-HILN coder), or in the case of some specific forensic application where extracting weak audio transients buried in a sinusoidal foreground [2] is intended, one may need to efficiently decompose the signal into a sinusoidal part (also denominated as the harmonic part, or the deterministic part) and a noisy part (also denominated as the stochastic part or the residual).

More precisely, the model is that of M slowly varying complex exponentials, hence encompasses the case of real data, summed with a stochastic process [3], written down as:

$$s(t) = \sum_{k=1}^{M} b_k(t) \exp(j\Phi_k(t)) + w(t), \qquad (1)$$

where $t \in \mathbb{Z}$ denotes the discrete time index, M the order of the model —*e.g.* the number of complex exponentials, being even, $M = 2P$, when data is

composed of P real sinusoids —, $b_k(t) \geq 0$ the modulation law relative to the kth component magnitude (real). $\Phi_k(t)$ is the instantaneous phase of the component and is bound up to its instantaneous frequency $f_k(t)$ by differentiation:

$$\Phi_k(t)' = 2\pi f_k(t). \quad (2)$$

Note that the frequencies are *not* assumed to be multiples of some fundamental. The stochastic process $w(t)$ may describe several kinds of physical signals : background measurement noise, turbulence noise imputable to air friction when dealing with wind instruments or voice, impulse-shaped, transient noise when processing for instance the onset of a piano or a percussion sound.

Estimation of the model. Since the instantaneous amplitudes and frequencies b_k's and f_k's are expected to be varying, both the parameters of the sinusoidal part and the statistical properties of the stochastic process may be considered as *non-stationary*. To overcome this difficulty, most methods (*cf.* [3, 4]) tend to use a sequential estimation technique applied on overlapping segments of finite length along which a definite sinusoidal model is estimated. For instance, the phase is often taken as a polynomial of low degree (*typ.* 1 or 2) in the variable t. The process $w(t)$ is obtained as a residual, by subtracting the estimated deterministic part. A broad bulk of existing algorithms relies on a time-frequency analysis of the signal, facing the challenging trade-off of shortening the segments for more adequation to the assumption of stationarity while loosing frequency resolution and hence, leading to poor estimates.

In this paper, both answers to this issue concerning harmonic plus noise decomposition are compared: one is based on the reassigned spectrum [5] and the second one is an adaptive subspace based analysis. The methods are described separately in the following sections while the results are demonstrated afterward. More specifically, this work focuses on the ability of each method to extract the noise part while preserving its spectro-temporal shape. Clues on frequency estimation performance can be found in [6, 7, 8] and are not in the scope of this work.

2. HARMONIC+NOISE DECOMPOSITION WITH REASSIGNED SPECTRUM

2.1. Principles

Reassignment operators [5] . The derivation of the so-called reassignment operators in time and frequency relies on the continuous time definition of the Short Time Fourier Transform (STFT). Let be $s_a(t),\ t \in \mathbb{R}$, the analyzed signal, the associated STFT is formulated as:

$$\tilde{S_a}(\tau, f) = \int_{t\in\mathbb{R}} s_a(t)h(t-\tau)e^{-j2\pi ft}dt \quad (3)$$

When facing the problem of localizing amplitude and frequency-modulated sinusoids, the performance limitation is mainly due to the window length and its spectral width (of referred to as the time-frequency box). The reassignment tries to overcome this Fourier-related constraint using the STFT phase information. The STFT is now rewritten in terms of magnitude and phase:

$$\tilde{S_a}(\tau, f) = M(\tau, f)e^{j\varphi(\tau,f)}. \quad (4)$$

The reassignment operators are derived from the partial derivatives of $\varphi(t, f)$ with respect to each of its variables, leading respectively to the instantaneous frequency

$$F_i(\tau, f) = \frac{1}{2\pi}\frac{\partial\varphi(\tau, f)}{\partial\tau}, \quad (5)$$

and to the group delay

$$T_g(\tau, f) = -\frac{1}{2\pi}\frac{\partial\varphi(\tau, f)}{\partial f}. \quad (6)$$

These equations are often interpreted as follows. When considering the energy $M(\tau_0, f_0)^2$ spread around a given point (τ_0, f_0) of the time-frequency plane, its centroid is the point of normalized frequency $F_i(\tau_0, f_0)$ and discrete time $\tau_0 + T_g(\tau_0, f_0)$. Each energy coefficient is said to be *reassigned* to this centroid. The time-frequency content of the signal is then re-mapped on the plane.

2.2. Discrete-time implementation

The Short Time Fourier Transform (STFT) of the sampled data sequence $s(t),\ t \in \mathbb{Z}$ is defined as

$$\tilde{S}(\tau, \nu_k) = \sum_{t=\tau}^{\tau+N-1} s(t)h(t-\tau)e^{-j2\pi\nu_k t}, \quad (7)$$

where $\tau \in \mathbb{Z}$ is the analysis time lag, $\nu_k = k/K$ the frequency bin and $h(t)$ the window applied, assumed to be of finite length N. The order K of the transform has to be greater or equal to N, and is chosen as $K = 2N$ in our practical implementations. The STFT is then rewritten in its polar form as

$$\tilde{S}(\tau, \nu_k) = M(\tau, \nu_k) e^{j\varphi(\tau, \nu_k)}. \quad (8)$$

To approximate the continuous variable derivatives needed in equations (5) and (6) in the context of numerical processing, a numerical filter is used. This filter can be for instance designed with the help of a Remez-Parks-McLellan algorithm for linear phase Finite Impulse Response (FIR) filter. In this work, a different technique is employed, the starting point of which is a polynomial fitting of the sequence [9].

$\varphi(\tau, \nu_k)$ is then extracted and unwrapped for each channel k and derivated to obtain the instantaneous frequency $F_i(\tau, k)$. The same procedure is applied along the frequency axis, yielding the group delay $T_g(\tau, k)$.

Adaptive computation. The algorithm is intended to work with a hop size of only one sample ($(N-1)-$samples overlap). To lower the complexity from the well-known $O(N \log(N))$ cost per sample to a linear one ($O(N)$) the STFT derivation is made adaptive [10, 11]. This gain benefits from the fact that a number of common windows are built with sines and thus, can be written as a sum of geometric sequences of the complex exponential form.

Let for instance the window $h(t)$ be the Hann window:

$$h(t) = \frac{1}{2}(1 - \cos(\frac{2\pi}{N} t)). \quad (9)$$

This is rewritten as

$$h(t) = \frac{1}{2}(1 - \frac{1}{2}(W_N^t + W_N^{-t})),\ t \in [0,\ N-1],$$

where $W_N = e^{j2\pi/N}$, which leads to a decomposition of the STFT:

$$\tilde{S}(\tau, \nu_k) = 0.5\tilde{S}_0(\tau, \nu_k) - 0.25(\tilde{S}_1(\tau, \nu_k) + \tilde{S}_2(\tau, \nu_k)), \quad (10)$$

where $\tilde{S}_0(\tau, \nu_k)$ is the STFT using the rectangular window $u_N(t) = 1,\ t \in [0,\ N-1]$ and $u_N(t) = 0$ otherwise, and where $\tilde{S}_1(\tau, \nu_k)$ and $\tilde{S}_2(\tau, \nu_k)$ are the STFT respectively windowed by $W_N^t u_N(t)$ and $W_N^{-t} u_N(t)$.

Defining the simple increment

$$\Delta s(\tau, k) = e^{-j2\pi\nu_k\tau} ((-1)^k s(\tau + N) - x(\tau)), \quad (11)$$

an update of each STFT is readily obtained as

$$\begin{cases} \tilde{S}_0(\tau+1, \nu_k) = \tilde{S}_0(\tau, \nu_k) + \Delta s(\tau, k) \\ \tilde{S}_1(\tau+1, \nu_k) = \tilde{S}_1(\tau, \nu_k) + W_N^{-1} \Delta s(\tau, k) \\ \tilde{S}_2(\tau+1, \nu_k) = \tilde{S}_2(\tau, \nu_k) + W_N \Delta s(\tau, k) \end{cases} \quad (12)$$

The update of the whole STFT then results from the equation (10).

Harmonic+noise decomposition

The reassignment principles have been applied for enhancing the time-frequency representation, for frequency estimation [5, 12] and also for source/filter modeling in speech processing [13]. As the formulae 5 and 6 cited above result in the precise localization of the frequency estimates, a reconstruction technique is to be determined to extract the harmonic part on one side and the noise on the other. As in many other works, the former is obtained at first and subtracted afterward from the original to get the latter.

For a given segment of analyzed data located in the interval $[\tau,\ \tau + N - 1]$, the *Harmonic part* of the signal is computed following the steps:

1. STFT computation and peak-picking of its magnitude,
2. derivation of $F_i(\tau, k)$ and $T_g(\tau, k)$ for each peak k,
3. selection among this collection of peaks of the bins l where the instantaneous frequency and the bin frequency match, *i.e* F_i must lie in the vicinity of the frequency center of the channel, for instance:

$$|F_i(\tau, l) - l/K| < \frac{3}{2}(1/2K). \quad (13)$$

This stage can be post-processed by a median filtering to remove isolated points,

4. for each selected bin l, a complex exponential at the frequency $F_i(\tau, l)$ is computed with an amplitude taking into account the phase and amplitude distorsion due to windowing at the frequency $F_i(\tau, l)$,
5. the synthetized component is added to the output segment, windowed by a Hann window centered on the time-instant $\tau + N/2 + T_g(\tau, l)$.

It is worth making mention here that the Hann window utilized for the synthesis is not of constant length, since it depends on the reallocation time in the analyzed interval. Let $L_h(\tau, l)$ be this length, this is expressed as:

$$L_h(\tau, l) = N - 2|T_g(\tau, l)|. \tag{14}$$

In addition, for approaching perfect reconstruction, the synthesis window $h_s(t)$ is weighted by the factor $(\sum_{t=0}^{L_h-1} h_s(t))^{-1}$ to be made unitary.

Once the steps 1-5 have been repeated all along the analyzed signal, the harmonic part $s_h(t)$ is derived. The noise part is then deducted as:

$$s_n(t) = s(t) - s_h(t) \tag{15}$$

3. ADAPTIVE HIGH RESOLUTION HNM DECOMPOSITION

Since the end of the 18th century [14, 15], Fourier analysis and High Resolution (HR) methods have been both complementary and competitors. While the former developed into the prominent tool in the field of the spectral analysis, the latter has revealed himself in the two last decades to be one of the most valuable estimation technique in the so-called Direction Of Arrival problem [16]. Notwithstanding its remarkable resolution properties, its use remains marginal in audio processing tasks, even though the underlying model is well adapted for tracking slow varying line spectra [17].

3.1. Theoretical background

Subspace analysis. Subspace decomposition is the theoretical foundation of a number of methods (Pisarenko [18], MUSIC [19], Matrix Pencil [7], ESPRIT [20]). The subspace analysis relies on the following remark. Let $x(t)$, $t \in \mathbb{Z}$ be a complex signal, linear combination of M complex exponentials:

$$x(t) = b_0 z_0^t + b_1 z_1^t + \ldots + b_{M-1} z_{M-1}^t, \tag{16}$$

where the z_k's, $k = 0, 1, \ldots, M-1$, are the complex poles of the signal and b_k's the associated complex amplitudes. More precisely, $z_k = \exp(\delta_k + j2\pi\nu_k)$ where $\delta_k \in \mathbb{R}$ is the damping or growing factor and $\nu_k \in [-0.5\ 0.5]$ is the normalized frequency. Expanding this definition to the vector of the n ($n \geq M$) subsequent samples $\mathbf{x} = \begin{bmatrix} x(0) & x(1) & \ldots & x(n-1) \end{bmatrix}^T$ leads to the matrix expression :

$$\mathbf{x} = \mathbf{V}\mathbf{b}, \tag{17}$$

where $\mathbf{b} = \begin{bmatrix} b_0 & b_1 & \ldots & b_{M-1} \end{bmatrix}^T$ and $\mathbf{V}$ is the Vandermonde matrix defined as:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ z_0 & z_1 & \ldots & z_{M-1} \\ z_0^2 & z_1^2 & \ldots & z_{M-1}^2 \\ \vdots & \vdots & \vdots & \vdots \\ z_0^{n-1} & z_1^{n-1} & \ldots & z_{M-1}^{n-1} \end{bmatrix} \tag{18}$$

For M distinct poles, the M vectors $\{\mathbf{v}(z_k)\}_{k=0,1,\ldots,M-1}$, defined as the column vectors of the matrix $\mathbf{V}$, $\mathbf{v}(z_k) = \begin{bmatrix} 1 & z_k & \ldots & z_k^{n-1} \end{bmatrix}^T$, are linearly independant. Thus the range space of $\mathbf{V}$ is of dimension M. In short, a vector of n subsequent samples of a signal combining linearly M complex exponential belongs to a M dimensional subspace, the so-called *signal subspace*. When dealing with a noisy signal model : $s(t) = x(t) + w(t)$, the vector $\mathbf{s} = \begin{bmatrix} s(0) & s(1) & \ldots & s(n-1) \end{bmatrix}^T$ belongs to a n-dimensional subspace. Under the hypothesis of a Wide Sense Stationary (WSS) white noise, this subspace can be decomposed as the direct sum of the M-dimensional signal subspace and its orthogonal complementary, of dimension $n - M$, referred to as *the noise subspace.*

Harmonic+noise decomposition. Let $\mathbf{W}$ be a $n \times M$ matrix, conveniently chosen as orthonormal, whose range space is the signal subspace. The projection matrices onto the signal subspace and onto the noise subspace are thus respectively $\mathbf{P}_s = \mathbf{W}\mathbf{W}^H$ and $\mathbf{P}_n = \mathbf{I} - \mathbf{P}_s$, where the subscript H denotes the hermitian transpose. For a given vector of data $\mathbf{s}$, the harmonic part is then obtained by:

$$\mathbf{s}_h = \mathbf{P}_s \mathbf{s} \tag{19}$$

while the noise part is the reminder:

$$\mathbf{s}_n = \mathbf{P}_n \mathbf{s} \tag{20}$$

These expressions need two remarks:

- even in the ideal case of stable signal components (neither amplitude nor frequency modulation) and WSS white noise, this decomposition *does not* lead to $s_h(t) = x(t)$, simply because considering a noise vector of n subsequent samples $\mathbf{w}$, this vector usually belongs to a n-dimensional space in which the noise subspace as defined above is included;
- neither estimation of the parameters (frequencies, damping factors, amplitudes) has to be made explicitly.

Tracking of W. In a number of methods, a matrix $\mathbf{W}$, the columns of which form a basis of the signal subspace, is derived by means of a Singular Value Decomposition of the covariance matrix $\mathbf{C}_{ss}$ of the data. Conversely, the subspace method used in this work is adaptive, referred to as the Fast Approximated Power Iteration in the literature [21]. Starting from a rank one update of the covariance matrix,

$$\mathbf{C}_{ss}(t) = \beta \mathbf{C}_{ss}(t-1) + \mathbf{s}(t)\mathbf{s}(t)^T, \qquad (21)$$

where $\beta < 1$ is a real positive forgetting factor and $\mathbf{s}(t) = \begin{bmatrix} s(t) & s(t+1) & \dots & s(t+n-1) \end{bmatrix}^T$, it conduces in $3nM + O(M^2)$ operations[1] to a rank one update of the form:

$$\mathbf{W}(t) = \mathbf{W}(t-1) + \mathbf{e}(t)\mathbf{g}(\mathbf{t})^H \qquad (22)$$

where $\mathbf{e}(t)$ and $\mathbf{g}(t)$ are column vectors. The whole description of the algorithm is beyond the scope of this paper and can be found in [21].

3.2. Preprocessing

As the method relies on a model comprehending an additive white stationary noise process, its performances lower when dealing with real signals the stochastic part of which is usually not white. In addition of the coloration of the noise, it is not rare in the audio field to encounter large dynamics. The estimation of the weak harmonic components, often settled in the upper part of the spectrum as low as 40 or 60 dB under the maximum, is then made dubious. The preprocessing designed for applying successfully the subspace tracker described above includes 3 steps:

1. pre-emphasis of the entire signal,
2. subband decomposition,
3. whitening in each subband.

Pre-emphasis and whitening. The first and third preprocessing steps are based on the same principle. The Power Spectral Density (PSD) of the considered sequence is estimated (for instance by means of a Welch-averaged periodogram) and an estimator of the noise PSD is derived as the non-linear median filtering of it. The corresponding AR coefficients are computed for a pre-defined model order K. As the aim of the pre-emphasis is a spectrum detrending, a low order K is chosen for the first step. In each subband, on the contrary, K will be of order 10 to 20, for the noise coloration must be drastically reduced. An exemple of the pre-emphasis of voice segment is given in figure 1. The original signal has then been filtered by a Finite Impulse Response (FIR) filter of length 5 to obtain the pre-emphasized signal.

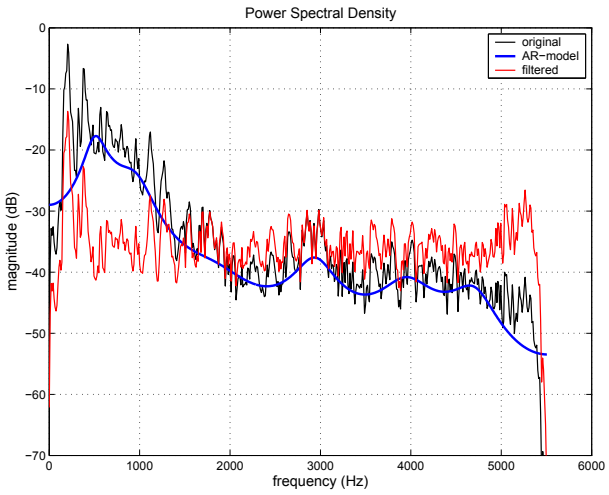


Fig. 1: Pre-emphasis of a voice segment of one second length. The model order is $K = 12$.

Filter bank. The subband decomposition is completed using a quasi-perfect reconstruction cosine-modulated filter bank [22]. Each subband signal is maximally decimated. The adjustment of the subband number depends on the sampling frequency and on the density of harmonic components in the resulting subband. Usual values vary from 4 to 16.

[1] an operation being defined as a Multiply and ACcumulate operation, MAC.

It can be noticed that even if in this work, only uniform filter banks are considered, an extension to non-uniform ones is readily obtained by dyadic iteration. An exemple of uniform 4-subbands decomposition is displayed on the figure 2.

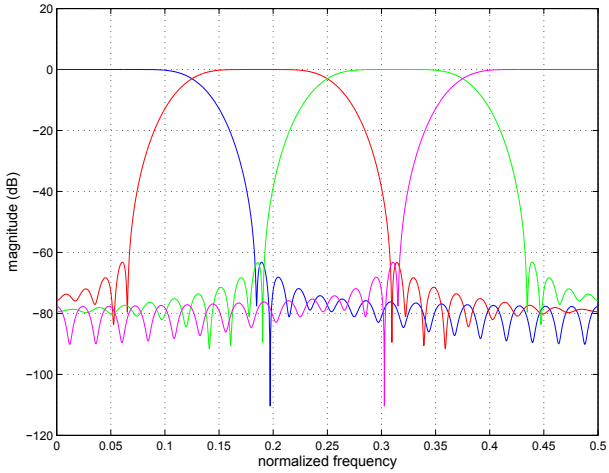


Fig. 2: Cosine modulated analysis filter bank, with 4 subbands.

4. EXPERIMENTS

The aim of this section is to demonstrate the abilities of both algorithms (Reassigned Spectrum-based or Subspace-based) in the task of extracting a background stochastic process and especially a transient (highly non-stationary process) from a signal including a strong harmonic content, speech being the target example of such a kind of signal. To be able to assess the results, a procedure for bringing into existence a non-stationary, impulsive-like process with known characteristics has been defined. The algorithms are then applied to various simulations to give some clues on the parameters tuning according to the context.

4.1. Creating a synthetic non stationary noise

The time-frequency profile of the process is defined as follows.

1. the spectrum at $t = 0$, the initial time instant is defined with the help of a set of poles, leading to an AutoRegressive (AR) spectrum,

2. a low f_l and a high f_h spectral limits are set, and a damping factor $\alpha(f_l)$ is defined for the low limit, owing to which the decreasing of the process around the frequency f_l is of the form $d(t) \propto \exp(-\alpha(f_l)t)$,

3. a damping law is given, as a power function of frequency, *i.e.* $\alpha(f) = \alpha(f_l)(\frac{f}{f_l})^p$

The whole operation is implemented by FFT-filtering of a white stationary noise. An example is drawn on figure 3, obtained at a sampling frequency of 8kHz with the following parameters: a solely pole of 0.99 magnitude at the frequency of 500 Hz, $f_l = 150$ Hz and $f_h = 3500$ Hz, $\alpha(f_l) = 4\,\mathrm{s}^{-1}$ and $p = 1$

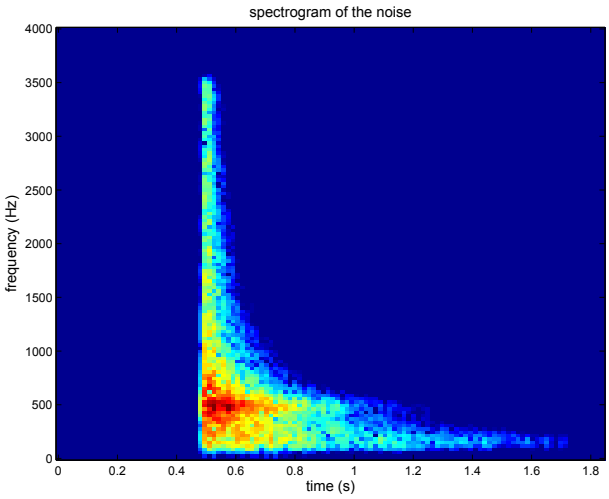


Fig. 3: Time-frequency (256 pts-FFT, Hann windowed) representation of the non-stationary noise.

4.2. Illustrative simulations

All the simulations of this section include a transient noise generated as described above in the section 4.1, and a white background stationary noise around 50 dB below the maximal signal power (this corresponds to an overall Signal To Noise Ratio around -25dB for the whole observation window). In the following, the Fourier-based method is referred to as RF-HND (Reassigned Fourier-Harmonic+Noise Decomposition), while the Subspace analysis-based method is abbreviated as HR-HND.

For each case, the time-frequency representation of the results are given, derived with a Hann windowed 256-points-FFT and jointly scaled to be comparable.

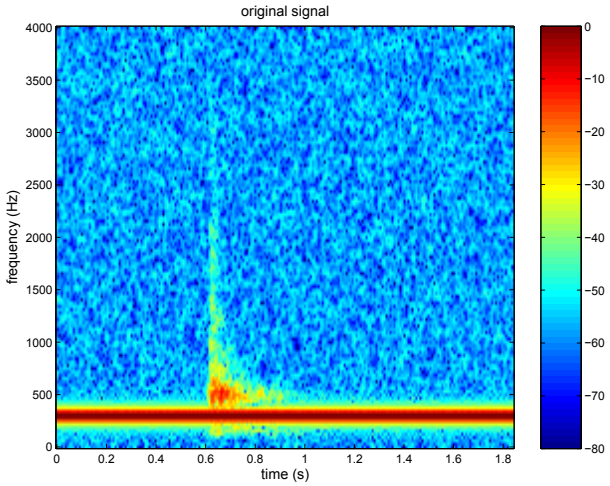


Fig. 4: Time-frequency representation of the original signal.

Pure Sine + noise. In this example, a 300 Hz-sinusoid is added to the noise, leading to a signal whose time-frequency representation is given in the figure 4.

Analysis parameters.

The HF-HND is applied with a window length $N = 256$ samples (32ms) and an order (number of frequency bins) $K = 512$.

The HR-HND is applied with the parameters

preprocessing AR-order	filter bank	analysis (length P, order M)
order 12	no	$P = 256$ (32ms) $M = 1$

Results and interpretation. The representation of the noise part respectively extracted by the RF-HND and the HR-HND methods is displayed in the figures 5 and 6.

In both cases, a satisfying extraction of the transient noise is performed. This results from the steadiness of the sinusoidal component which matches exactly the estimated model in both cases. Nevertheless the window length cannot be shortened without increasing the variance of the HR-HND estimator or lessening the resolution capability of the RF-HND estimator. Both influences imply a notching effect on the whole extracted stochastic part, around the sinusoid frequency. A manner of this effect can be observed on the RF-noise as a "hole" in the transform around 300 Hz.

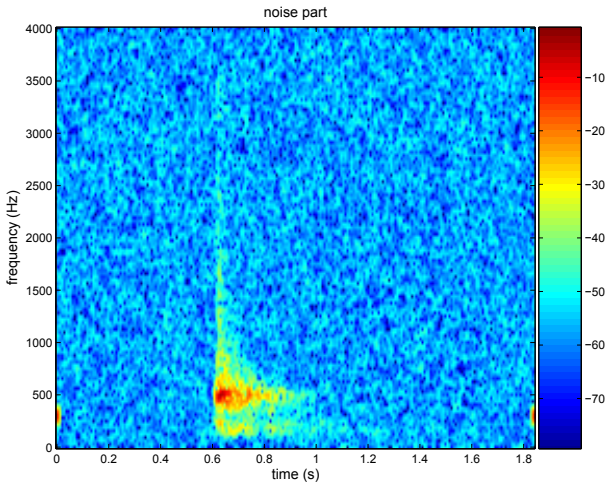


Fig. 5: Time-frequency representation of the noise part obtained by the RF-HND method.

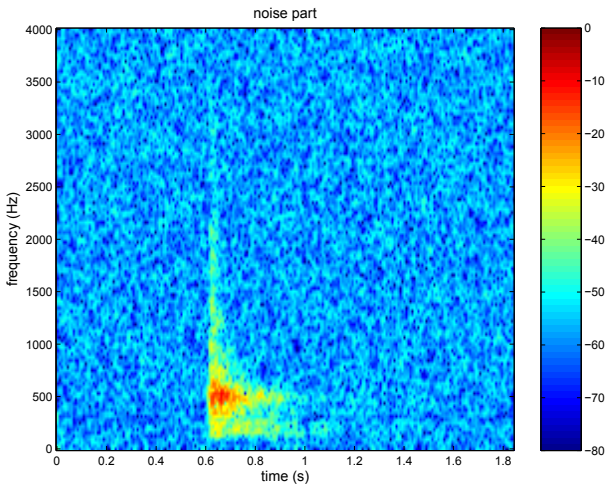


Fig. 6: Time-frequency representation of the noise part obtained by the HR-HND decomposition.

FM-modulated sine + noise. The sinusoid of this example is now modulated, leading to a 4 Hz vibrato of a semi-tone frequency deviation. Its rep-

resentation is given in figure 7.

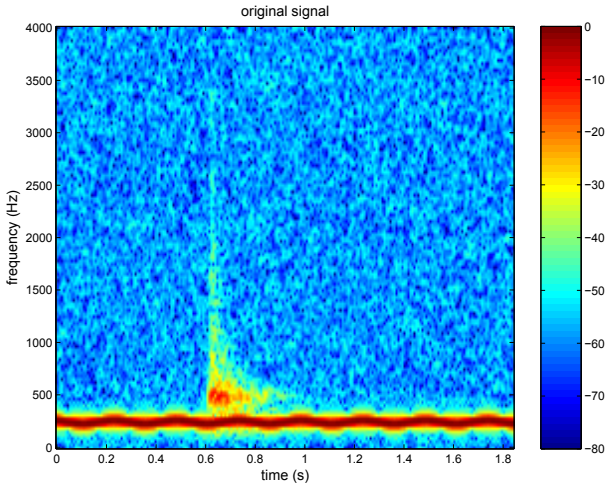


Fig. 7: Time-frequency representation of the original signal.

Analysis parameters.

The HF-HND is applied with a window length $N = 256$ samples and an order (number of frequency bins) $K = 512$.

The HR-HND is applied with the parameters

preprocessing AR-order	filter bank	analysis (length P, order M)
order 12	no	$P = 512$ (64ms) $M = 6$

Results and interpretation. The representation of the noise part respectively extracted by the RF-HND and the HR-HND methods is displayed in figures 8 and 9. This example illustrates the different tuning sensibility of both methods. For the RF-HND estimator, the window length must satisfy the trade-off between the intended resolution and the ability to track the frequency modulation. The HR-HND, in the contrary, uses a longer window and represents the modulation with the help of a higher model order. Similarly to the previous case, the spectral shape of the transient noise is roughly preserved and the RF-method causes a more pronounced notching effect.

4.3. Toward a real world application

To approach our targeted application, this last simulation is generated by mixing a real male speech utterance of the vowel 'a' (french) with the preceding transient noise.

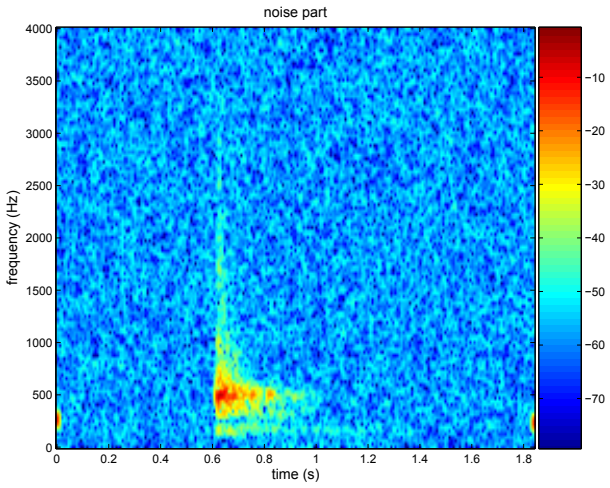


Fig. 8: Time-frequency representation of the noise part obtained by the RF-HND method.

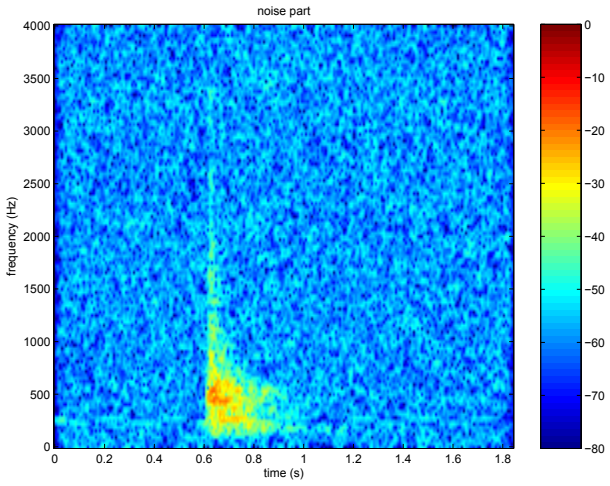


Fig. 9: Time-frequency representation of the noise part obtained by the HR-HND decomposition.

Analysis parameters.

The HF-HND is applied with a window length $N = 512$ samples and an order (number of frequency bins) $K = 1024$.

The HR-HND is applied with different parameters of the analysis for each subband.

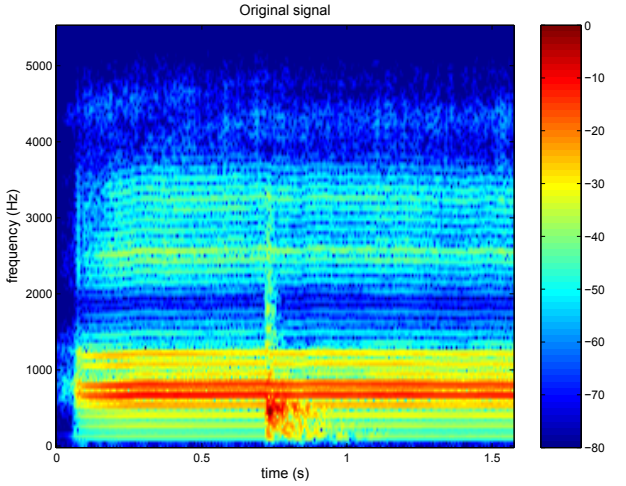


Fig. 10: Time-frequency representation of the original signal.

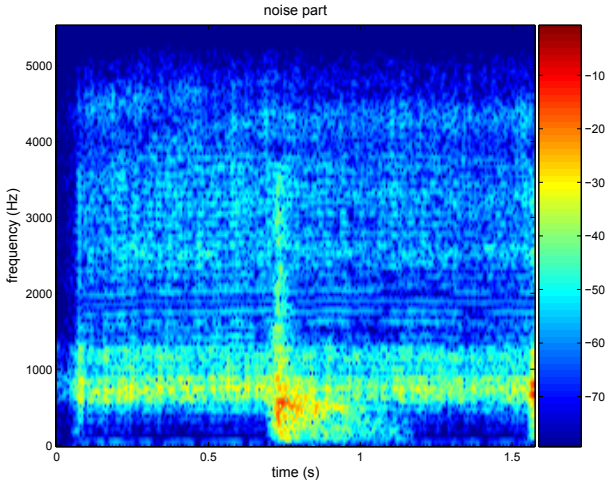


Fig. 11: Time-frequency representation of the noise part obtained by the RF-HND method.

preprocessing AR-order	filter bank	analysis (length P, order M)
order 12	4 bands AR12 whit.	$P = 200, 150, 50, 40$ $M = 40, 20, 25, 10$

Results and interpretation. The noise part extracted by the RF-HND estimator (figure 11), has a lower spectral density than that extracted by the HR-HND estimator(figure 12), especially in the upper part of the spectrum. Indeed, the subband decomposition used as preprocessing of the latter allows to process apart each subband: the window length can be adjusted differently in the lower range and in the upper range of the spectrum, leading to a kind of multiresolution processing. This might explain a more prominent notching effect for the RF-HND method while the overall formantic structure of the voice friction noise is better preserved by the HR-HND method. Conversely, the latter is more sensitive to the parameter set fine tuning,

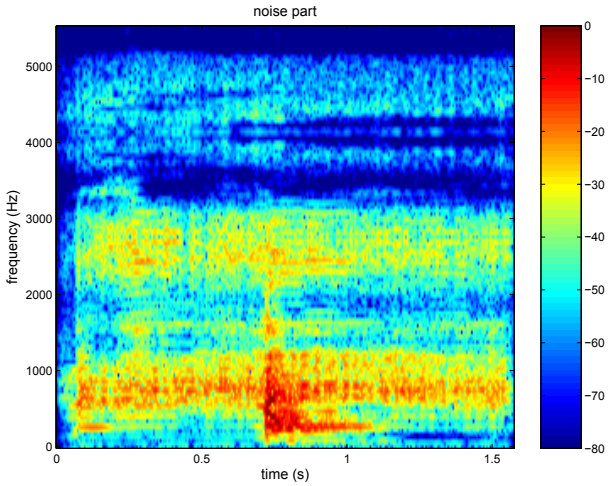


Fig. 12: Time-frequency representation of the noise part obtained by the HR-HND method.

5. CONCLUSIONS

This preliminary work on the extraction of a transient background noise surrounded by a signal with a strong harmonic content enlights the main differences and abilities of both methods: one being based on the reassigned STFT and the other being an adaptive subspace-based estimator. Both are trying to cope with the limitations related to the well-known time-frequency trade-off. Both are capable of extracting transient noise when it is reasonably strong and when the modulations of the harmonic content remain of low extent.

Future work may include the tracking of the line spectra as a preliminary of the resynthesis of the harmonic part and the test of non-uniform filter banks or multiresolution representations.

6. REFERENCES

[1] G. Peeters and X. Rodet, "SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum," in *Proc. of ICMCs*, Beijing, China, 1999.

[2] Y. Grenier and B. David, "Extraction of weak background transients from audio signals," in *Proc. of 114th Convention of the Audio Engineering Society (AES)*, Amsterdam, Netherlands, Mar. 2003.

[3] X. Serra and J. Smith, "Spectral modeling synthesis : a sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, 1990.

[4] L. S. Marques and L. B. Almeida, "Frequency-varying sinusoidal modeling of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 5, pp. 763–765, May 1989.

[5] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[6] C. R. Rao and L. C. Zhao, "Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals," *IEEE Trans. Signal Processing*, vol. 41, no. 3, pp. 1461–1464, Mar. 1993.

[7] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 5, pp. 814–824, May 1990.

[8] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*. Cambridge, UK: Cambridge University Press, 2001.

[9] M. Dvornikov, "Formulae of numerical differentiation," in *arXiv.org e-Print archive*. arXiv.org, 2003, pp. 1–14.

[10] C. Richard and R. Lengellé, "Joint recursive implementation of time–frequency representations and their modified version by the reassignment method," *Signal Processing*, vol. 60, no. 2, pp. 163–179, 1997.

[11] V. Gibiat, P. Jardin, and F. Wu, "Analyse spectrale différentielle - Application aux signaux de Myotis Mystacinus," *Acustica*, vol. 63, pp. 90–99, 1987, in French.

[12] V. Emiya, B. David, and V. Gibiat, "Two representation tools to analyse non-stationary sounds in a perceptual context," in *Proceedings of Forum Acusticum 2005*, Budapest, Hungary, Aug. 2005.

[13] D. J. Nelson, "Cross-spectral methods for processing speech," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2575–2592, 2001.

[14] J. Fourier, "Mémoire sur la propagation de la Chaleur dans les corps solides," *Nouveau Bulletin des sciences par la Société philomathique de Paris*, vol. 6, pp. 112–116, 1808, in French.

[15] G. M. Riche de Prony, "Essai expérimental et analytique: sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool différentes températures," *Journal de l'école polytechnique*, vol. 1, no. 22, pp. 24–76, 1795, in French.

[16] S. Chandran, *Advances in Direction-Of-Arrival Estimation*. Cambridge, UK: Artech House Publishers, Jan. 2006.

[17] B. David, R. Badeau, and G. Richard, "HRHATRAC Algorithm for Spectral Line Tracking of Musical Signals," in *Proc. of ICASSP'06*. Toulouse, France: IEEE, May 2006, (to be published).

[18] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical J. Royal Astron. Soc.*, vol. 33, pp. 347–366, 1973.

[19] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[20] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT–A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 5, pp. 1340–1342, Oct. 1986.

[21] R. Badeau, B. David, and G. Richard, "Fast Approximated Power Iteration Subspace Tracking," *IEEE Trans. Signal Processing*, vol. 53, no. 8, Aug. 2005.

[22] P. P. Vaidyanathan, *Multirate systems and filter banks*. Englewoods Cliffs, NJ, USA: Prentice Hall, 1993.