Modélisation Sinusoïdale des Sons Polyphoniques

Mathieu Lagrange

N° ORDRE : 2917

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Mathieu LAGRANGE**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Modélisation sinusoïdale des sons polyphoniques

Soutenue le : 16 décembre 2004

Après avis de : MM. Philippe Depalle ... (McGill U., Montréal) Rapporteurs
Xavier Serra (U. Pompeu Fabra, Barcelone)

Devant la Commission d'examen formée de :

Mme.	Jenny Benois-Pineau	Professeur	Présidente
M.	Philippe Depalle	Professeur	Examinateurs
Mme	Myriam Desainte-Catherine	Professeur	
MM.	Sylvain Marchand	Maître de conférences	
	Jean-Bernard Rault	Docteur-Ingénieur ...	
	Xavier Serra	Professeur	

2004

This document was entirely written using LaTeX.

Remerciements

J'aimerais tout d'abord remercier Jenny Benois-Pineau qui m'a fait l'honneur d'être la présidente du jury de ma soutenance de thèse. Son intérêt pour les problématiques d'indexation développées à la fin de ma thèse m'encourage à poursuivre dans cette voie.

Je ne sais comment remercier mon directeur de thèse Sylvain Marchand. Son ouverture, sa rigueur et sa volonté d'excellence désintéressée sont pour moi des modèles de discipline de recherche. J'ai particulièrement apprécié sa manière toute personnelle d'ouvrir des horizons non explorés aux cours de nos discussions tout en me laissant toute latitude concernant les moyens à proposer pour y parvenir. Ensuite, il a su consacrer du temps pour me prodiguer conseils et encouragements, malgré la distance et les obligations diverses. Enfin, Je lui suis gré de m'avoir encouragé à publier des articles et participer à des conférences, ce qui m'a permis de confronter mes travaux à différentes communautés de chercheurs.

Je remercie également Myriam Desainte-Catherine, directrice du SCRIME pour avoir accepter de diriger ma thèse. Sa foi en l'informatique musicale permet le développement de synergies appréciables entre informaticiens, compositeurs et acousticiens.

Je tiens à témoigner toute ma reconnaissance à Philippe Depalle pour avoir rapporté ce document. Tout d'abord, ses nombreux conseils et remarques ont permis d'en améliorer la clarté et la qualité. Ensuite, sa vaste connaissance de domaines tels que le traitements du signal, l'acoustique et l'informatique ainsi que son intérêt particulier pour certaines problématiques développées dans ce document ont permis de soulever de nombreuses perspectives de recherches.

Toute ma gratitude va également à Xavier Serra qui a accepté de rapporter ce document malgré son emploi du temps chargé. Ses suggestions m'ont permis de m'interroger sur la portée de mon travail sur des domaines plus ouverts.

Je remercie tout les membres de l'équipe du traitement du son du site de Rennes de France Télécom Recherche et Développement. J'ai trouvé au sein de cette équipe un mélange subtil de professionnalisme, de camaraderie et d'humour toujours apprécié. En particulier, je remercie chaudement Jean-Bernard Rault qui, par son approche didactique, sa rigueur et son dévouement quotidien, a su me guider durant ces trois années. Merci à Pierrick Philippe pour ces discussions enflammées et ses points de vues originaux. Un grand merci également à Cathy Colomès pour m'avoir initiée aux méthodologies de tests subjectifs.

Un signe de la main à Mathias Rossignol ; les nombreuses discussions tardives dans notre salon ont rythmées agréablement ces trois années de cheminement intellectuel. Certaines idées soulevées sont devenues réalités, ce document en témoigne. D'autres, plus fantasques, sont encore dans les cartons. Elles ne sont pas les seules car le Vietnam, c'est loin !

Une pensée enfin à une jolie fleur qui accompagne mes jours au rythme des saisons depuis de nombreuses années déjà. Son épanouissement est mon soutien et ma joie.

Table des matières

Liste des tableaux

Table des figures

Introduction

Ce document traite de la modélisation sinusoïdale d'enregistrements audionumériques où plusieurs instruments sont enregistrés simultanément. En particulier, l'extraction de la partie quasi périodique d'un enregistrement musical en vue de meilleures possibilités d'interprétation constitue la problématique à laquelle on tentera de répondre.

Représentation du son

Avant le XXième siècle, la performance d'un musicien constituait un événement unique et éphémère. La notation musicale a alors été utilisée pour permettre de conserver l'intention du compositeur par le biais d'une partition. Comme tout langage, le solfège demande une interprétation d'un être humain pour obtenir un son proche de celui voulu par le compositeur. Cette représentation est donc peu fidèle. Elle est par contre très expressive, car le solfège permet de consigner en peu de symboles les contraintes nécessaires à l'exécution de l'intention de celui qui a écrit la partition.

Au début de ce siècle, il est devenu possible de fixer une réplique de l'onde sonore sur un support solide comme la cire ou le plastique vinyle. Plus tard, la possibilité de stocker des données chiffrées a permis de représenter le son non plus de manière analogique mais numérique. Ces découvertes ont profondément modifié notre perception de la musique car un phénomène unique est devenu reproductible à volonté. Malheureusement, il est ardu d'obtenir des informations structurelles d'une telle représentation. La forme d'onde est, à l'inverse de la partition, très fidèle mais en contrepartie très peu expressive.

La nécessité de stocker et de transmettre des données audionumériques a amené un effort de recherche conséquent en codage à réduction de débit. Le principe du codage est de représenter le son sous une forme plus compacte en exploitant les propriétés de l'oreille humaine [ZF90]. Il est donc important pour ces codeurs, dits perceptifs [Bra99], d'opérer la compression d'une représentation perceptive du son. Les transformées spectrales qui évaluent l'énergie du signal pour certaines fréquences sont la base de ces codeurs (appelés aussi *de facto* codeurs par transformée). Comme ces transformées sont inversibles, la représentation spectrale est aussi fidèle que la représentation temporelle, tout en étant plus expressive. Cette représentation est en effet privilégiée pour les applications de reconnaissance de phonèmes [DM80] et d'identification du locuteur [BW04].

Modélisation paramétrique

Pour effectuer des transformations pertinentes du son telles que la transposition ou l'étirement, il est utile d'avoir un modèle du son. Dans les années 70, les premiers vocodeurs de phase ont été implantés pour la modélisation de signaux harmoniques sous forme de sinusoïdes. Il a ensuite été proposé dans [SS87] de conserver ce modèle pour la modélisation de sons inharmoniques ou des sons harmoniques dont la hauteur varie au cours du temps. La volonté de modéliser les sons bruités de manière perceptuellement pertinente a amené Serra [Ser89] à proposer une décomposition du signal sonore en une partie pseudo périodique et une partie stochastique. Un niveau de décomposition supplémentaire a ensuite été proposé [VM98, BAM02] pour modéliser les parties transitoires du son, comme les sons percussifs.

Cette représentation est dite *paramétrique* car à chacune de ces trois composantes est associée une série de paramètres qui contrôlent un générateur d'un type particulier. Par exemple, la partie quasi périodique est synthétisée grâce à des oscillateurs sinusoïdaux (les *partiels*). Les paramètres de contrôle sont alors la fréquence, l'amplitude et la phase.

Dans le domaine de la vision par ordinateur, il est utile de décomposer une image en zones régulières, en contours et textures. Une telle décomposition permet de mieux analyser la scène visuelle. Par analogie, la notion de structure apportée par la représentation paramétrique permet de simplifier l'interprétation des signaux musicaux à des fins d'indexation. Par exemple, l'analyse de la partie transitoire est utile pour la détection du tempo et du rythme tandis que l'analyse des parties périodiques est utile pour la détection de la hauteur et de la mélodie.

Si la séparation d'un enregistrement audionumérique en ces trois composantes (sinusoïdes, bruit et transitoires) est correctement effectuée, on dispose alors d'une représentation fidèle et particulièrement expressive du son. Cependant, une telle décomposition demeure un défi majeur car les frontières entre ces trois parties sont souvent floues. On propose dans ce document des traitements informatiques qui permettront d'améliorer l'extraction de la partie quasi périodique des enregistrements musicaux polyphoniques.

Présentation générale

Le chapitre 1 étudie de manière succincte la structure des différents types d'instruments de musique les plus courants ainsi que certains travaux d'acousticiens. Cette étude motivera la représentation couramment utilisée pour représenter la partie quasi périodique du son. Certaines notions, comme composante court-terme (pic spectral) et composante long-terme (partiel) sont explicitées. Ensuite, l'étude des caractéristiques des parties dites transitoires et de bruit permettra de définir les requis de robustesse d'un algorithme d'extraction de la partie quasi périodique. Ces requis seront utiles lors de l'évaluation des différents algorithmes proposés par la suite.

Une extraction pertinente de la partie quasi périodique requiert tout d'abord l'estimation précise des paramètres du modèle (amplitude, fréquence et phase).

Le chapitre 2 étudie certains estimateurs de fréquence, d'amplitude et de phase, basés sur une transformée spectrale bien connue : la transformée de Fourier. En particulier, nous proposons une amélioration d'une technique d'estimation de la fréquence basée sur l'étude des dérivées successives du signal [DCM00]. Grâce à l'étude d'une interprétation purement trigonométrique de l'évolution au cours du temps de l'amplitude des composantes sinusoïdales du signal sonore, on améliore la précision de cet estimateur dans les hautes fréquences.

Les signaux sont rarement stationnaires durant l'intervalle d'observation considéré pour estimer leurs paramètres. Ceci amène une estimation incorrecte du paramètre de phase, en particulier lorsque la fréquence varie. On propose dans la deuxième partie de ce chapitre, deux estimateurs de phase robustes aux modulations linéaire de fréquence. La comparaison avec les estimateurs de phase classiques montrent une amélioration notable de la robustesse.

Après l'estimation locale des paramètres sinusoïdaux, la seconde étape dite de suivi de partiels consiste à déterminer quand un partiel débute, se termine et quelle est l'évolution de ses paramètres durant sa phase d'activité. Lors de l'analyse de signaux polyphoniques, l'utilisation d'heuristiques simples ne permet généralement pas d'obtenir une représentation interprétable. L'utilisation du caractère prédictible de l'évolution des paramètres de fréquence et d'amplitude permet d'améliorer l'identification des partiels dont les paramètres sont modulés (comme dans le cas d'une note avec un vibrato ou un trémolo) et d'éviter la majeure partie des composante dites de bruit [LMRR03, LMR04b]. Ensuite, l'absence théorique de haute fréquence dans l'évolution des paramètres des partiels permet de proposer un algorithme de suivi adapté aux contenus polyphoniques. Grâce à ce critère, de meilleurs résultats sont obtenus dans le cas des sinusoïdes dont les fréquences sont proches ou se croisent.

On montre dans le chapitre 4 que le caractère prédictible des évolutions des paramètres des partiels peut aussi être exploité pour interpoler la partie quasi périodique d'un son durant un intervalle de temps conséquent. Comparée aux différents méthodes proposées dans la littérature, la méthode proposée apporte un gain de qualité significatif.

Enfin, les nouvelles possibilités d'interprétation de la représentation à long terme offertes par les algorithmes présentés dans le chapitre 3 sont exploitées dans le chapitre 5 pour structurer cette représentation. Les partiels présentant certaines corrélations sont agrégés pour former des entités sonores, chaque entité étant perçue par le système auditif humain non plus comme plusieurs sons simples mais comme un unique son complexe. Plusieurs indices issus d'études psychoacoustiques [Bre90] tels que l'apparition simultanée de partiels, leur relation d'harmonicité, les évolutions corrélées de leurs paramètres permettent de mener à bien cette agrégation.

1 Éléments de modélisation paramétrique du signal sonore

La modélisation des sons quasi périodiques sous forme de sinusoïdes trouve ses racines dans de nombreux travaux d'acousticiens et psychoacousticiens, dont certains sont synthétisés dans une première partie. Sont ensuite présentés différents modèles paramétriques attachés aux trois composantes du son : pseudo périodique, bruit et transitoire. En particulier, les modèles sinusoïdaux à court terme et à long terme pour la représentation de la partie pseudo périodique sont détaillés. Pour tester la validité d'un modèle à court terme, la communauté dispose de nombreux outils. Il n'existe en revanche pas de méthodologie pour la validation d'une modélisation à long terme. Pour pallier ce manque, la fin de ce chapitre définit les requis d'une modélisation à long terme ainsi que plusieurs critères quantitatifs de validation qui seront utiles pour évaluer les méthodes d'extraction d'une représentation à long terme proposées dans ce document.

1.1 Introduction

L'homme a eu très tôt son attention attirée par les sons musicaux : ces sons, de timbre particulièrement agréable et suggestif, étaient primitivement fournis par des instruments à corde (lyre, harpe, etc.) et les instruments à vent (flûte, pipeau, cheng des Chinois, etc.).

Le son le plus simple est donné par un diapason électronique ; on dit que l'allure du son émis est rigoureusement sinusoïdale comme on le voit à gauche de la figure 1.1. La vitesse de la branche du diapason en mouvement qui produit le son augmente, passe par un maximum, décroît, change de sens et recommence à un rythme constant ; le diapason oscille régulièrement, périodiquement, autour de sa position d'équilibre. Une représentation très compacte de ce type de signal est la représentation fréquence/amplitude, dite représentation spectrale. Il suffit en effet de connaître la fréquence et l'amplitude de l'onde pour décrire et reproduire le son $s(t)$ produit par le diapason :

$$s(t) = a\cos(\omega t) \tag{1.1}$$

$$\omega = 2\pi f \tag{1.2}$$

où a, ω et f désignent respectivement l'amplitude, la pulsation exprimée en radians par seconde et la fréquence exprimée en Hertz. À droite de la figure 1.1 est représenté le spectre du signal émis par un diapason électronique.

Bien entendu, la plupart des instruments décrits plus haut produisent des sons plus riches. Prenons l'exemple d'une simple corde fixée à ses deux extrémités. Si l'on excite cette corde, elle va former un ventre dont le maximum de déformation est situé au milieu de la corde, deux ventres dont les maxima sont situés aux quarts supérieurs et inférieurs de la corde, trois ventres, etc. Les largeurs des ventres étant des sous-multiples de la longueur de la corde, chaque ventre va se comporter comme un diapason vibrant à une fréquence propre. Ceci mène à l'évidence que la plupart de sons musicaux "contiennent" d'autres sons [Sch66]. Ces sons différents provenant de la même corde sont dans des rapports simples (harmoniques) : leurs fréquences sont multiples d'une fréquence dite fréquence de fondamentale, celle du premier ventre. L'ajout de cordes de tailles et de tensions différentes les unes aux autres ou l'ajout de trous dans des tuyaux vibrants étend ce principe des ventres pour donner des instruments plus riches. L'agencement des tailles et tensions des cordes ou des trous sur les tuyaux d'instruments à vent, ouvre des possibilités de création de sons variés. En général, cet agencement est effectué de manière à conserver une perception "agréable".

Expliquer pourquoi un instrument "sonne" de façon agréable est ardu. Dans le présent document, on évitera volontairement les notions d'harmonie et de timbre qui permette de mieux appréhender ce problème, car ces notions ont des implications extrêmement complexes tenant autant de la physique que de la perception et des sciences humaines.

Une expérience menée par Helmholtz a permis de vérifier la pertinence d'une décomposition des sons complexes en sons simples [Hem63]. Des sphères ou

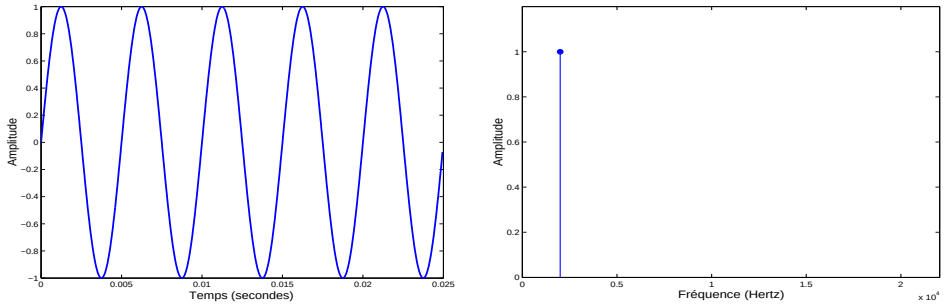


FIG. 1.1 – Représentation d'une onde sinusoïdale produite par un diapason sur le plan temps/amplitude (à gauche) et sur le plan fréquence/amplitude (à droite). On dit que ce son est très pauvre car l'énergie produite par le diapason est concentrée sur une très faible partie du spectre.

des cylindres sont percés d'un trou de faible diamètre destiné à recevoir le signal sonore et d'un autre trou de diamètre plus grand destiné à être approché par l'oreille de l'auditeur, voir figure 1.2. Ces sphères ont des dimensions bien déterminées de manière à entrer en résonance pour une fréquence donnée. Chaque sphère est alors façonnée pour isoler une composante particulière d'un signal composite. Considérons une série "harmonique" de sphères de tailles choisies pour entrer en résonance à des fréquences en relation harmonique d'une fréquence de fondamentale. Leur diamètre doit être en relation harmonique $(1, 1/2, 1/3, 1/4,$ etc.). L'expérience montre qu'il est possible de décomposer un signal musical d'une hauteur égale à la fréquence de résonance de la première sphère en plusieurs sons simples correspondants aux harmoniques de la fondamentale. En effet, si l'auditeur approche son oreille, il entendra parfaitement l'harmonique correspondante comme un son simple, proche de celui produit par un diapason. Chaque sphère agit comme un filtre simple, avec une seule fréquence où le signal est amplifié, voir figure 1.3.

Modélisation source/filtre

Il est donc possible d'imiter un son harmonique par la production de plusieurs sons purs avec des fréquences particulières. Cependant, si les amplitudes de ces sons purs sont choisies de manière arbitraire (tout les sons purs ont la même amplitude par exemple), le son obtenu sera très "synthétique", éloigné d'un son naturel.

Ceci s'explique par le fait que la plupart des instruments précités ne se composent pas seulement d'un système excitateur (corde ou sifflet) mais aussi d'un corps qui opère un filtrage (caisse d'une guitare ou tuyau d'une flûte). Ce corps exploite le phénomène de résonance pour amplifier et modifier le son émis par l'excitateur. On peut noter que ce phénomène d'amplification ne peut être obtenu que s'il existe une relation entre les dimensions du corps qui la subit et la fréquence de vibration du son émis par l'excitateur. En effet, un corps donné, de par sa structure, entre en résonance pour une même série de fréquences dites

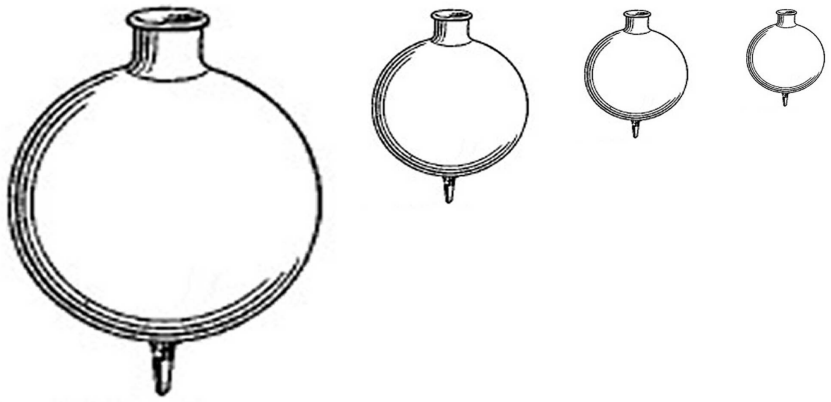


FIG. 1.2 – Résonateurs de Helmholtz. De par le diamètre de sa cavité, un résonateur va amplifier un signal dont la fréquence est proche de sa fréquence de résonance.

fréquences propres.

De ces considérations découlent une modélisation dite "source/filtre" du système de production sonore d'instruments simples comme ceux présentés précédemment. Prenons l'exemple d'une source qui émet un signal de base composé de sinusoïdes de fréquences multiples d'une fréquence de fondamentale. Le signal émis par la source est ensuite filtré par un dispositif physique qui va amplifier certaines fréquences et en atténuer d'autres, voir figure 1.4. Ce modèle est particulièrement utilisé pour la modélisation de la parole voisée (prononciation des voyelles, chant) où les cordes vocales et le conduit vocal forment le couple source/filtre. Ce modèle se généralise aux parties entretenues des notes produites par de nombreux instruments comme ceux à vent où l'anche ou le sifflet constituent la source et le tuyau le filtre. Pour les instruments à cordes, les cordes forment la source et la caisse constitue le filtre. Les amplitudes des harmoniques sont donc modifiées par la réponse en fréquence du filtre induit par le corps amplificateur.

Le système auditif humain

Bien avant des connaissances étendues sur le système auditif humain qui corroborent la validité d'une représentation des sons entretenus par des sommes de sinusoïdes mises en forme spectralement, les acousticiens ont supposé que ce type de décomposition était pertinente. Ohm se base sur la perception : " seul ce mouvement particulier de l'air que nous avons appelé *vibration simple*, dans lequel les particules se meuvent en avant et en arrière selon la loi du mouvement pendulaire, est susceptible de donner à l'oreille la sensation d'un son simple unique. Donc tout mouvement de l'air correspondant à un ensemble composite de sons musicaux peut, d'après la loi d'Ohm, être analysé en une somme de vibrations pendulaires simples, et à chacune de ces vibrations simples uniques

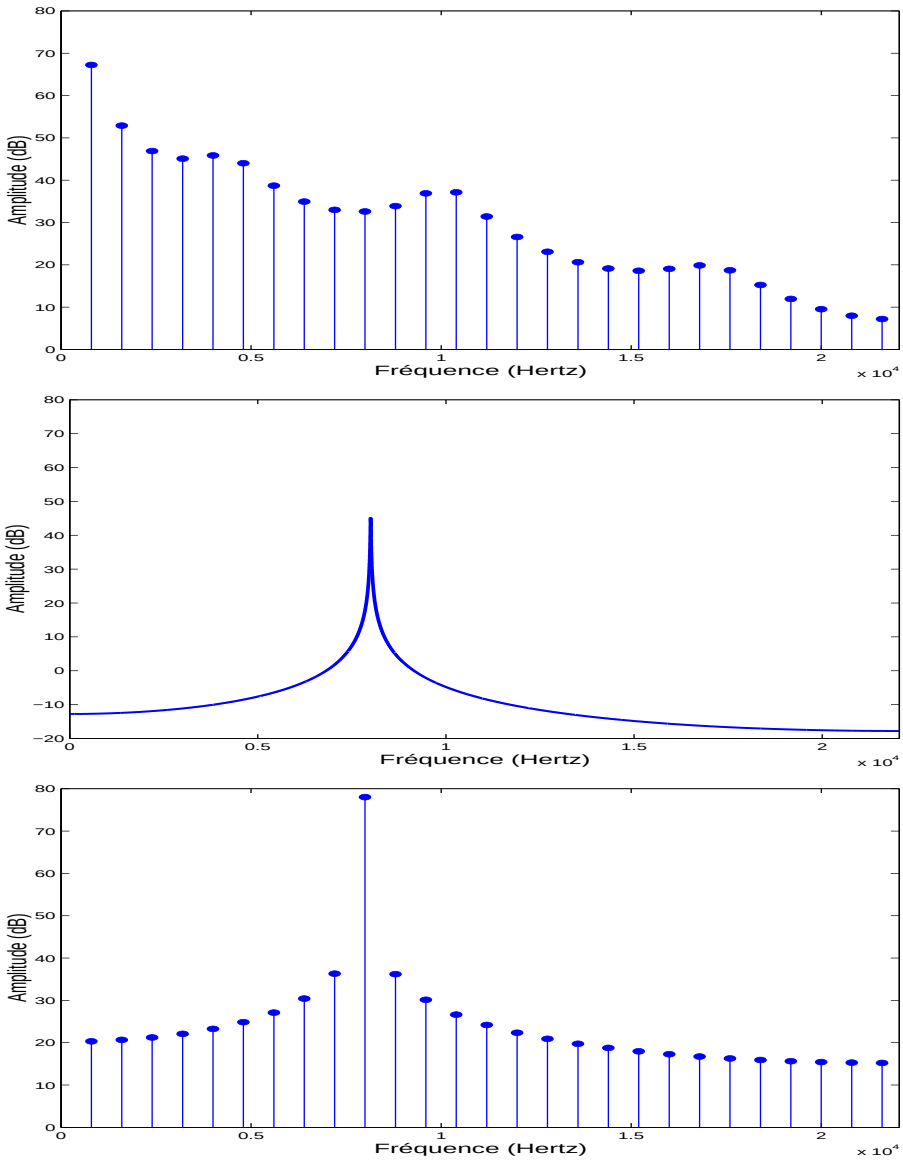


FIG. 1.3 – Principe de fonctionnement d'un résonateur de Helmholtz. Le signal composite (dont le spectre est représenté en haut) entre dans la sphère par l'ouverture inférieure. Si une harmonique de ce son complexe a une fréquence proche de cette fréquence de résonance (approximativement 8 kHz sur la figure du milieu), elle va être amplifiée tandis que les autres harmoniques vont être dissipées sous forme de chaleur lors de leur déplacement dans la sphère.

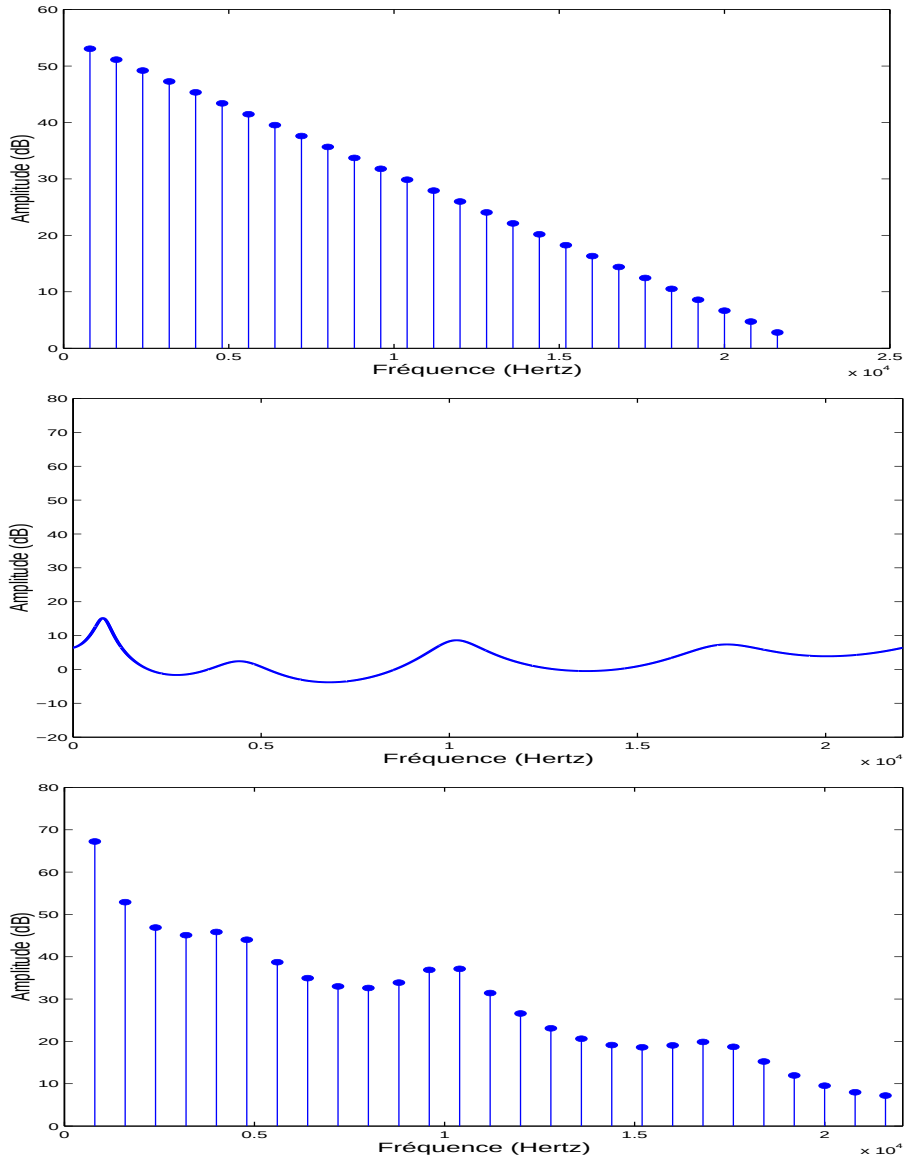


FIG. 1.4 – Modélisation source/filtre d'un signal voisé de parole. En haut, les cordes vocales (source) émettent une forme d'onde dont le spectre est composé d'une série de sinusoïdes dont les fréquences sont en relation harmonique et leur amplitude décroît régulièrement. Le conduit vocal, en fonction de sa forme, va amplifier certaines fréquences et en atténuer d'autres. Au milieu est représenté la réponse en fréquence du filtre induit par ce conduit vocal. On voit clairement se dégager la structure dite formantique. Les sinusoïdes de fréquence proche des maxima de l'enveloppe (formants) seront amplifiées tandis celles situées dans les creux seront atténuées. Le signal qui sort de la bouche du locuteur est alors le signal émis par les cordes vocales puis filtré par le conduit vocal.

correspond un son simple, repérable par l'oreille, dont la hauteur est déterminée par la durée de la période du mouvement de l'air correspondant.[1]"

L'hypothèse selon laquelle le système auditif humain fonctionne d'une manière similaire aux résonateurs de Helmholtz, c'est-à-dire que ce système est constitué d'éléments entrant en résonance pour différentes fréquences (un peu comme les cordes d'une harpe) a été énoncé par Helmholtz [Hem63]. Cette hypothèse est encore à ce jour sujet à discussion. On sait néanmoins que la cochlée (organe en forme d'escargot situé après le tympan et les osselets) est tapissée par une membrane dite basilaire composée de cellules cilliées qui émettent un signal électrique lorsqu'elles sont déformées. Békési [Bék60] a montré que plus la fréquence du signal perçu est élevée, plus la position du maximum de déformation de la membrane est éloigné de la base de la cochlée. L'oreille opère donc bien une décomposition spectrale.

Cette transformation est située très en amont de la chaîne de traitement du système auditif humain. En effet, même si notre oreille interne est capable de décomposer un signal composite en plusieurs sons simples, il est très difficile pour un auditeur de dissocier des composantes sinusoïdales simples avec des propriétés particulières comme une relation d'harmonicité entre leurs fréquences. De nombreux dispositifs "réflexes" relevant plus du domaine de la psychologie que de l'acoustique nous permettent d'analyser aisément une scène complexe, en agrégeant les sons simples en différents sons composites, considérés comme des entités perceptives. Ces dispositifs cognitifs sont particulièrement intéressants pour qui se préoccupe des capacités de hiérarchisation et d'interprétation du système auditif humain. L'étude des différentes théories qui tentent d'expliquer comment le système auditif humain interprète les informations données par l'oreille interne sort du sujet de ce chapitre mais sera repris dans le chapitre 5. Ce chapitre est en effet consacré à l'extraction d'entités perceptuelles où les différents dispositifs réflexes présentés dans [Bre90] permettront de regrouper plusieurs sons simples en entités perceptuelles. Avant cela, on doit être en mesure de décomposer de manière automatique un enregistrement sonore en sons simples.

La modélisation des sons musicaux comme une somme de sinusoïdes est développée dans une première section et les différents traitements associés à ce modèle seront largement développés dans les chapitres 2 et 3. Bien entendu, ce type de modèle est réservé à certaines classes de sons quasi périodiques et se prête bien à la modélisation des parties d'entretien et de relâchement d'une note, voir figure 1.5. En effet, ces sons diffus en temps, sont très localisés en fréquence comme on peut le constater sur la figure 1.1.

Les autres types de signaux sont aussi importants dans la perception des signaux musicaux. Citons Helmholtz : "De plus, le son de la plupart des instruments est d'habitude accompagné de bruits irréguliers caractéristiques comme le grattement ou frottement de l'archet dans le violon, le passage de l'air dans la flûte et dans les tuyaux d'orgue, le battement des anches, etc. Ces bruits, qui nous sont déjà familiers dans la mesure où ils caractérisent les instruments, facilitent matériellement notre pouvoir de les distinguer dans une masse com-

[1] Traduction de P. Schaeffer

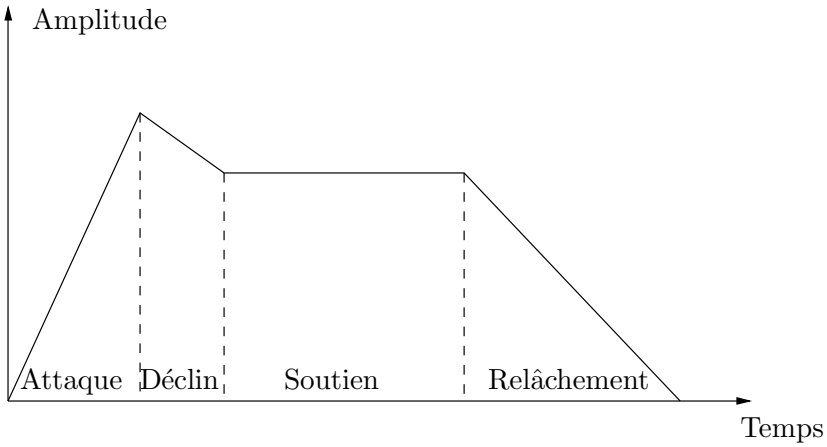


FIG. 1.5 – Représentation schématique de l'enveloppe d'amplitude d'une note. On peut distinguer quatre parties : la phase d'attaque (souvent très localisée en temps) la phase de déclin puis les phases de soutien et de relâchement.

posite de sons [2]". Ces signaux se prêtent mal à la décomposition en sons purs car la représentation sous forme de sinusoïdes s'éloigne notablement du système de production. Prenons l'exemple de la partie non voisée du signal de parole, comme la prononciation de consonnes. Le flux d'air produit par les poumons est contraint à sortir par une petite ouverture au niveau du conduit vocal, amenant des turbulences. Ces signaux dits bruités sont diffus en temps et en fréquence. Dans un signal de bruit blanc, toutes les fréquences sont équiprobables, le spectre est alors quasiment plat.

D'autres signaux comme ceux issus de percussions sont très peu résonants. Ces signaux dits transitoires sont très localisés en temps et souvent diffus en fréquence, à l'inverse des signaux quasi périodiques. Malgré l'aspect fugitif de l'événement sonore engendré par une transitoire comme un début de note, cet événement est aussi prépondérant dans la reconnaissance du type d'instrument. Cet aspect est abordé en profondeur dans [Sch66].

Les besoins de modélisation générique de larges classes de signaux musicaux ont donc amené à l'introduction de modèles hybrides où le signal à analyser n'appartient pas à une classe précise de sons. Idéalement, ce signal est décomposé en sous signaux, chacun appartenant à un des trois types de signaux présentés : sinusoïdal, bruit ou transitoires. Ces modèles associés à chacun de ces sous signaux sont présentés succinctement dans une deuxième section. Dans la troisième section, on propose des critères d'évaluation pour l'analyse sinusoïdale lorsqu'elle est confrontée à des signaux polyphoniques contenant des signaux bruités ou transitoires. Les critères proposés permettront d'évaluer la qualité des algorithmes proposés dans la suite de ce document.

[2] traduction de P. Schaeffer

1.2 Modèle sinusoïdal

Dans un modèle sinusoïdal, le signal est représenté sous forme d'une somme de sinusoïdes ayant chacune une fréquence, une phase et une amplitude propres. Ces techniques se basent sur le théorème de Fourier qui montre que toute fonction périodique peut être modélisée sous la forme d'une somme de sinusoïdes d'amplitudes données et de fréquences en relation harmonique. En citant Helmholtz : "La possibilité mathématique démontrée par Fourier de décomposer toutes les vibrations périodiques en vibrations simples ne nous autorise pas davantage à conclure qu'elle est la seule forme permissible d'analyse, si nous ne pouvons pas établir en sus que cette analyse a aussi une signification essentielle dans la nature. Que cela soit, de fait, le cas (que cette analyse a un sens dans la nature indépendamment de la théorie) est rendu probable par ce fait que l'oreille effectue précisément la même analyse, et aussi par cette circonstance, déjà mentionnée, que cette sorte d'analyse a de plus grands avantages pour l'investigation mathématique qu'aucune autre.[3]" Ces propos se vérifient encore aujourd'hui où les modèles alternatifs comme les transformées en ondelettes ou la modélisation fractale restent d'utilisation anecdotiques pour la modélisation des signaux sonores alors que les applications se basant sur la modélisation sinusoïdale sont très nombreuses. La proximité de ce modèle avec la perception et la pertinence perceptuelle de manipulations mathématiques simples en font, de fait, un modèle particulièrement riche et flexible.

1.2.1 Modélisation à court terme

Comme les caractéristiques des signaux sonores évoluent en fonction du temps, considérer les paramètres des composantes du modèle comme constants durant toute la durée du signal est peu pertinent. On peut segmenter le signal à analyser en petits fragments dans lesquels on considère que l'approximation faite par l'utilisation d'un modèle stationnaire est valide, voir figure 1.7. En effet, plus l'intervalle de temps considéré est petit, plus les variations des paramètres seront faibles. Chaque segment est alors représenté par un ensemble de sinusoïdes, voir figure 1.8. Ainsi, l'approximation faite par le modèle est considérée comme négligeable. Le signal dans une trame d'indice n est modélisé comme suit :

$$s_n(t) = \sum_{i=1}^{N} s_n^i(t) \tag{1.3}$$

$$s_n^i(t) = a_n^i \cos\left(2\pi f_n^i (t - n\,\Delta_T) + \phi_n^i\right) \text{ pour } n\,\Delta_T < t < n\,\Delta_T + T \tag{1.4}$$

où ϕ_n^i désigne la phase à $n\,\Delta_T$ et f_n^i et a_n^i désignent la fréquence et l'amplitude qui sont considérées comme constantes dans l'intervalle $[n\,\Delta_T, n\,\Delta_T + T[$. Pour chaque trame d'indice n et de durée T, un ensemble de paramètres sinusoïdaux à court terme $\mathcal{C}_n = \{p_n^0, \cdots, p_n^{N-1}\}$ est estimé. Les paramètres du système pour cette trame sont alors les N triplets $p_n^i = \{f_n^i, a_n^i, \phi_n^i\}$. Sur la figure 1.6, le pas

[3] traduction de P. Schaeffer

d'avancement Δ_T est égal à la moitié de la taille d'une trame, mais il peut être différent.

Le signal ainsi modélisé peut être synthétisé à partir de l'équation 1.4 sur l'intervalle de temps de la trame considérée. Ces différentes trames successives sont ensuite combinées pour obtenir le signal synthétisé, comme représenté sur la figure 1.6. Les différentes trames sont préalablement pondérées à l'aide d'une fenêtre de manière à réduire les discontinuités aux bornes :

$$s(t) = \sum_{n=-\infty}^{\infty} s_n(t) \cdot w(t - n\Delta_T) \tag{1.5}$$

où Δ_T est le pas d'avancement en secondes et $w(t)$ est une fenêtre de pondération ayant les propriétés suivantes (pour $\Delta_T = \frac{T}{2}$) :

$$w(t) = 0 \text{ pour } t < 0 \text{ ou } t \geq T \tag{1.6}$$

$$w(t) + w\left(t + \frac{T}{2}\right) = 1 \text{ pour } t \in [0, T/2] \tag{1.7}$$

Dans la figure 1.6, le signal $s_n(t)$ est pondéré par une fenêtre triangulaire mais l'utilisation d'autres fenêtres comme celles utilisées pour l'analyse sinusoïdale est possible. Ensuite, le pas de synthèse est égal au pas d'analyse. Pour des besoins spécifiques, il peut être plus faible, les valeurs entre $\mathcal{C}_n$ et $\mathcal{C}_{n+1}$ doivent alors être interpolées.

Cette modélisation dite à "court terme" est très populaire car pratiquement tous les signaux peuvent être modélisés grâce à cette approche, des sons stationnaires [Por76, LVS99] aux sons bruités [MC97, HDC01]. Des méthodes ont été proposées pour modéliser tout type de signaux grâce à un modèle sinusoïdal à court terme [GS97, VVHK99]. Pour certains signaux très localisés en temps et qui contiennent souvent de fortes variations d'amplitudes, des modèles plus élaborés ont été proposés comme les sinusoïdes amorties où les amplitudes des sinusoïdes peuvent décroître de façon exponentielle [NHD98]. Il peut être aussi utile de disposer non seulement de la fréquence mais aussi de la pente de la fréquence dans la trame d'analyse. Des modèles non stationnaires comportant des sinusoïdes dont la fréquence varie linéairement proposés dans la littérature [MA86, Mas96, ML03b] seront étudiés dans la section 2.3.

1.2.2 Modélisation à long terme

Pour des sons quasi périodiques, des corrélations entre les paramètres des sinusoïdes de trames successives peuvent être exploitées. Par exemple, on constate que pour les phases de soutien et de relâchement d'une note de piano (à partir de la trame d'indice 20 sur la figure 1.9), la fréquence de chaque composante sinusoïdale est quasi constante et l'amplitude décroît régulièrement. Il est alors utile de considérer un modèle sinusoïdal à long terme où les amplitudes et les fréquences des sinusoïdes évoluent lentement et de manière continue avec le temps, de manière à conserver une continuité de phase. Ces paramètres contrôlent un ensemble d'oscillateurs quasi sinusoïdaux communément appelés

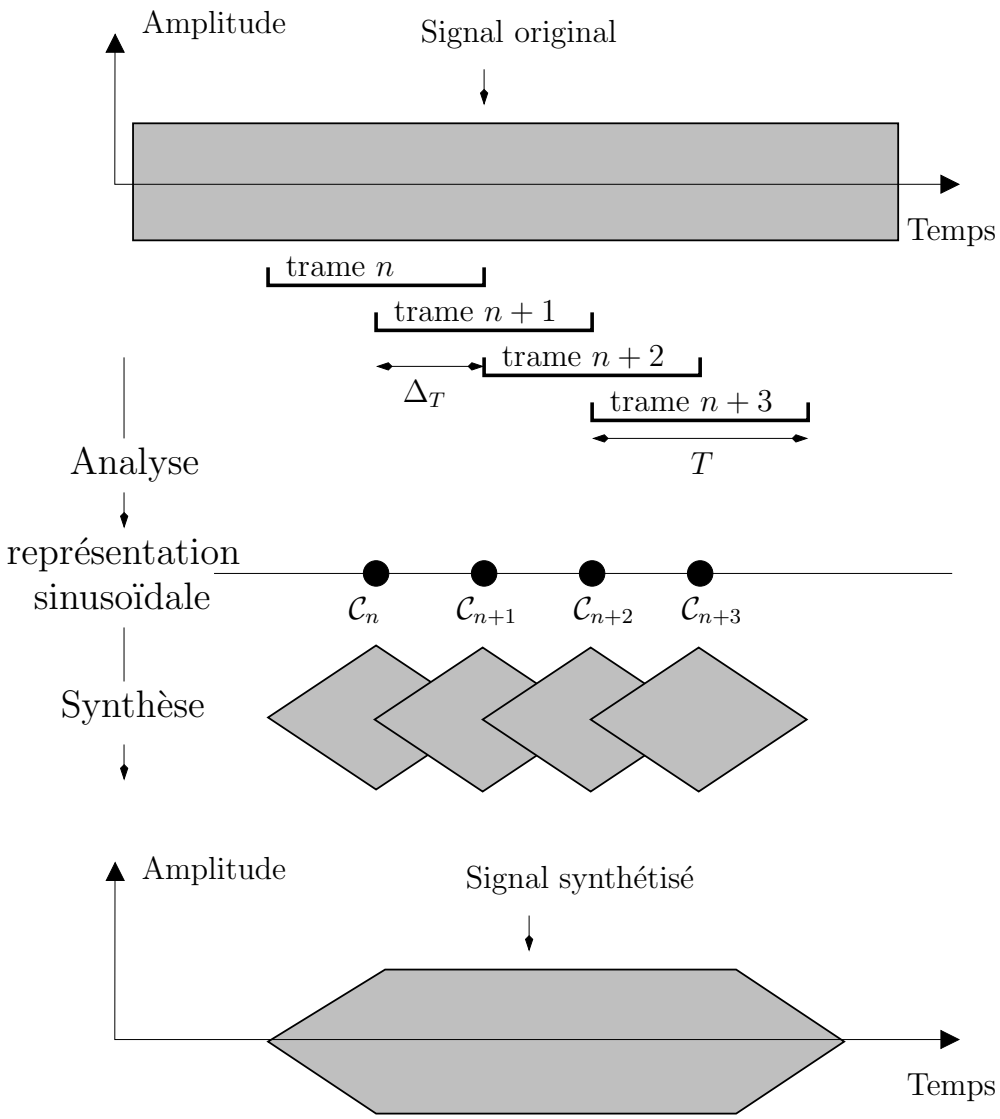


FIG. 1.6 – Schéma de principe d'un module d'analyse/synthèse à court terme. Pour une trame d'indice n, un ensemble de paramètres sinusoïdaux $\mathcal{C}_n$ est estimé. Ici, la largeur du pas d'analyse est égale à la moitié de la taille d'une trame, mais peut être différente. De chaque $\mathcal{C}_n$ est synthétisé un signal qui est ensuite mis en forme par une fenêtre de pondération de manière à réduire les discontinuités aux bornes. L'addition des contributions de tous les $\mathcal{C}_n$ donne le signal synthétisé.

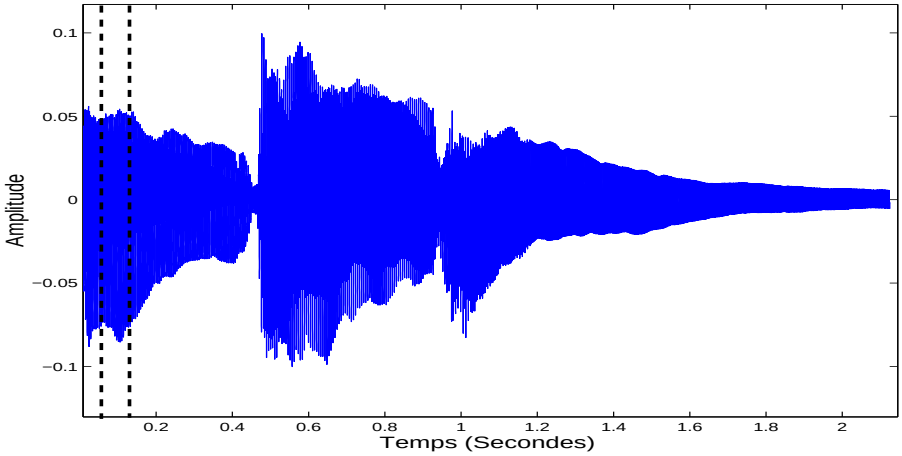


FIG. 1.7 – Forme d'onde associée à trois notes de piano. Dans un modèle sinusoïdal à court terme, on représente une petite partie de ce signal (délimitée par des tirets sur la figure) par une somme de sinusoïdes dont les paramètres sont constants.

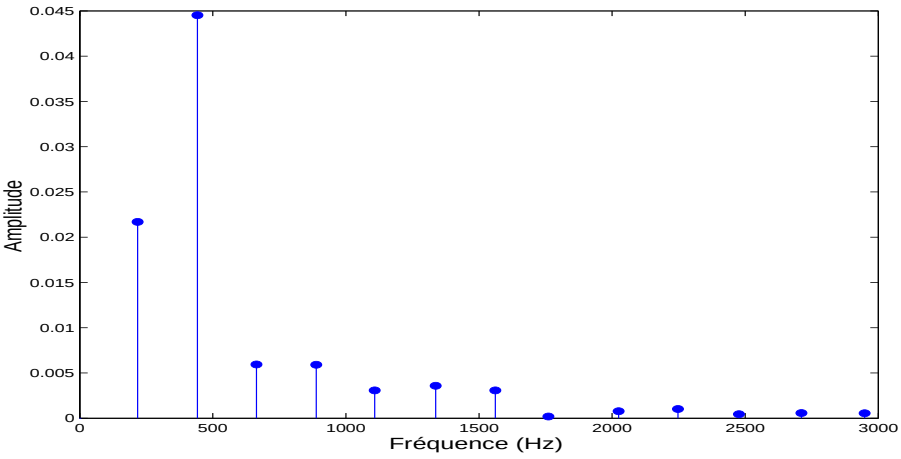


FIG. 1.8 – Modélisation sinusoïdale stationnaire d'un fragment de signal de piano, délimité par des lignes en tirets sur la figure 1.7.

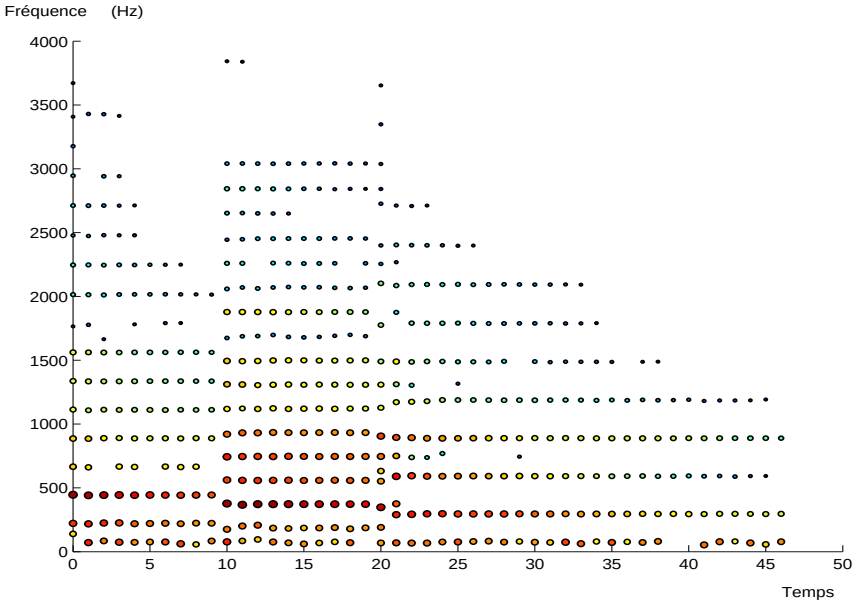


FIG. 1.9 – Représentation sinusoïdale à court terme du signal de la figure 1.7. À chaque trame, le signal est représenté par un ensemble de sinusoïdes (points). La taille du point est fonction de l'amplitude de la sinusoïde.

partiels. Le signal s peut être calculé à partir d'un ensemble de paramètres sinusoïdaux à long terme $\mathcal{S} = \{P_0, \cdots, P_{N-1}\}$ grâce aux équations suivantes :

$$s(t) = \sum_{k=1}^{N} A_k(t) \ \cos(\Phi_k(t)) \tag{1.8}$$

où,

$$\Phi_k(t) = \Phi_k(0) + 2\pi \int_0^t F_k(u) \ du \tag{1.9}$$

Les paramètres F_k, A_k et Φ_k sont respectivement les fréquences amplitudes et phases instantanées du partiel P_k. Les N triplets $(F_k(t), A_k(t), \Phi_k(t))$ au temps t sont les paramètres d'un modèle sinusoïdal à court terme valide à l'instant t. Une représentation de 3 notes de piano suivant ce modèle est donnée par la figure 1.10.

Il est difficile de définir la notion de variation "lente" des paramètres des partiels. Une définition perceptive est utilisée dans de document : les paramètres de contrôle d'un oscillateur sinusoïdal varient lentement s'il n'existe pas d'énergie notable pour des fréquences supérieures à 20 Hz dans le spectre de leur évolution. Dans le cas contraire, une telle modélisation n'est plus pertinente car elle ne correspond plus à la perception.

Les hypothèses de variation lente et de continuité temporelle qui sont la base de ce modèle font qu'il est plus contraint que le modèle à court terme. Cette contrainte peut être appréciée de manière négative car elle restreint le panel de sons correctement modélisés. Elle peut aussi être appréciée de

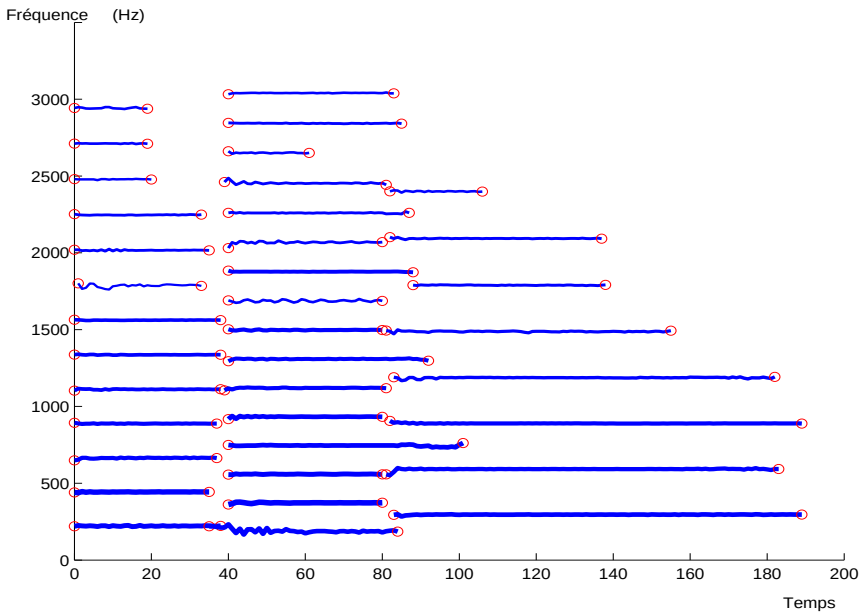


FIG. 1.10 – Représentation sinusoïdale à long terme du signal de la figure 1.7. Les cercles marquent le début et la fin de chaque sinusoïde représentée par un trait plein.

manière positive, car cette description est très concise et les manipulations y sont mathématiquement simples et perceptuellement pertinentes. Il est possible de reproduire, modifier des sons, car il existe des algorithmes d'analyse implantés dans des logiciels d'analyse/synthèse comme PARSHL [SS87], SMS (CLAM) [SS90], AudioSculpt [IRC96] LEMUR [FH96] et InSpect [MS99, Mar00a]. Des algorithmes de modification de hauteur et de durée de signaux monophoniques de haute qualité sont possibles. Ces algorithmes sont aussi utiles pour des applications de conversion de texte en signal de parole ou des applications de synchronisation de sons ainsi qu'en codage très bas débit des signaux de parole et de musique. En effet, les paramètres sinusoïdaux qui varient lentement peuvent être encodés efficacement. Le HILN pour "*Harmonic and Individual Lines plus Noise*" [PM00] et le SSC pour "*SinuSoid Codec*" [dBSO02] sont des codeurs musicaux qui se basent sur le modèle sinusoïdal à long terme. La possibilité d'obtenir une description de haut niveau est aussi particulièrement intéressante pour des applications comme la transcription musicale [FCQ98] et la séparation de sources [VK00].

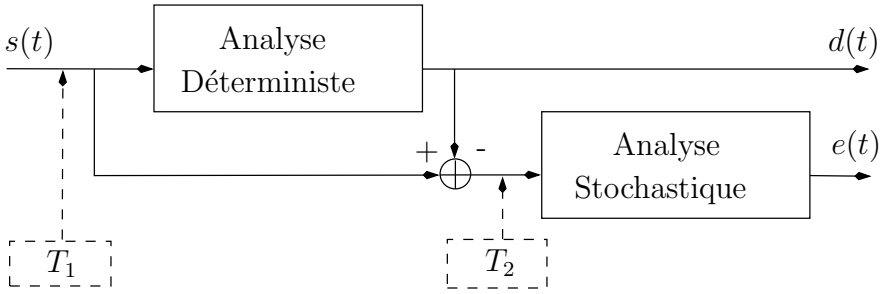


FIG. 1.11 – Procédure d'analyse hybride du signal sonore. La modélisation des transitoires peut être faite avant (cas T_1) ou après (cas T_2) l'analyse sinusoïdale qui extrait la partie déterministe du signal. La partie stochastique est estimée à partir du résidu, signal original auquel on a soustrait la partie déterministe et la partie transitoire.

1.3 Modèles hybrides

À proprement parler, les sons quasi périodiques sont produits parfaitement par peu d'instruments de musique. Pour une approche plus universelle, le modèle sinusoïdal est utile pour représenter les phases de soutien et de décroissance de notes produites par des instruments non percussifs et le signal de parole voisé. Les signaux bruités (fricatives de la voix parlée, turbulences) et les signaux transitoires (prononciation de consonnes, débuts abrupts de notes) doivent être modélisés séparément.

1.3.1 Modèle Sinusoïdes+Bruit

Le modèle Sinusoïdes+Bruit (SB) a été introduit dans [SS90] pour pallier les insuffisances du système PARSHL [SS87] qui se base uniquement sur une modélisation sinusoïdale à long terme. En effet, utiliser des sinusoïdes pour représenter des signaux de bruits est très coûteux en nombre de sinusoïdes car un spectre de bruit est très diffus. De plus, la modélisation du bruit avec des sinusoïdes ne permet pas des modifications pertinentes. Il est donc proposé de décomposer le signal temporel $s(t)$ en deux sous signaux :

$$s(t) = d(t) + e(t) \tag{1.10}$$

où $d(t)$ désigne la partie déterministe du signal et $e(t)$ le résiduel ou la partie stochastique, voir figure 1.11. La première partie $d(t)$ est représentée sous forme d'une somme de sinusoïdes avec cohérence de phase. Une fois cette partie $d(t)$ estimée, le signal résiduel $s(t) - d(t)$ peut alors être modélisé par un processus stochastique comme celui décrit par l'équation 1.11. Ce résiduel est donc souvent appelé la partie stochastique du signal. Dans [SS90], il est proposé de représenter ce résiduel par un bruit blanc u filtré par un filtre à réponse impulsionnelle h

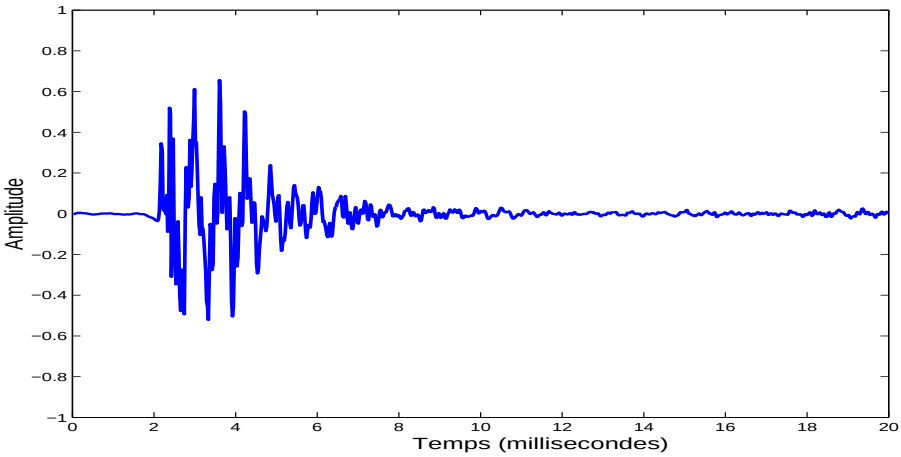


FIG. 1.12 – Signal transitoire émis par une paire de castagnettes.

qui représente l'enveloppe du résiduel au cours du temps :

$$e(t) = \int_0^t h(t,\tau)\, u(t-\tau) d\tau \tag{1.11}$$

où $u(t)$ est un bruit blanc et $h(t,\tau)$ est un filtre particulier dont les coefficients varient dans le temps. Ce modèle SB permet une description concise d'une plus grande gamme de sons, $h(t,\tau)$ étant un filtre contrôlé par peu de paramètres.

Toutefois, deux problèmes se posent lors de l'utilisation de ce type de modèle. Tout d'abord, la séparation déterministe/stochastique est rarement une discrimination claire. Certains partiels apparaissent comme des sinusoïdes bruitées ou très modulées. Des modèles utilisant des sinusoïdes sans cohérence de phase [MQ92] dont les paramètres de fréquence ou d'amplitude peuvent être modulés aléatoirement [Fit99] ont été proposés pour pouvoir modéliser ces partiels d'un type particulier. Ensuite, les transitoires (attaques et plosives dans le signal de parole) peuvent nécessiter un traitement particulier.

1.3.2 Modèle Sinusoïdes+Transitoires+Bruit

L'extension pour les transitoires dans les modèles Sinusoïdes + Transitoires + Bruit (STB) a été introduite pour modéliser les augmentations soudaines et brèves de l'énergie du signal de manière à prendre en compte les signaux très percussifs. Ces signaux sont très localisés en temps et on observe souvent une décroissance exponentielle de l'énergie après l'excitation, comme on peut le constater sur la figure 1.12. Cette atténuation dépend de nombreux facteurs comme la réponse de l'environnement acoustique qui varie en fonction de la fréquence. Des modèles sinusoïdaux non stationnaires ont été proposés pour modéliser l'atténuation exponentielle de ce type de signaux [NHD98].

Ce modèle est efficace si le début de l'attaque est proche du début de la fenêtre d'analyse, voir figure 1.14(a). Dans le cas contraire, la partie précédente (échantillons 0 à 500 sur la figure 1.14(c)) devra être modélisée par des sinusoïdes

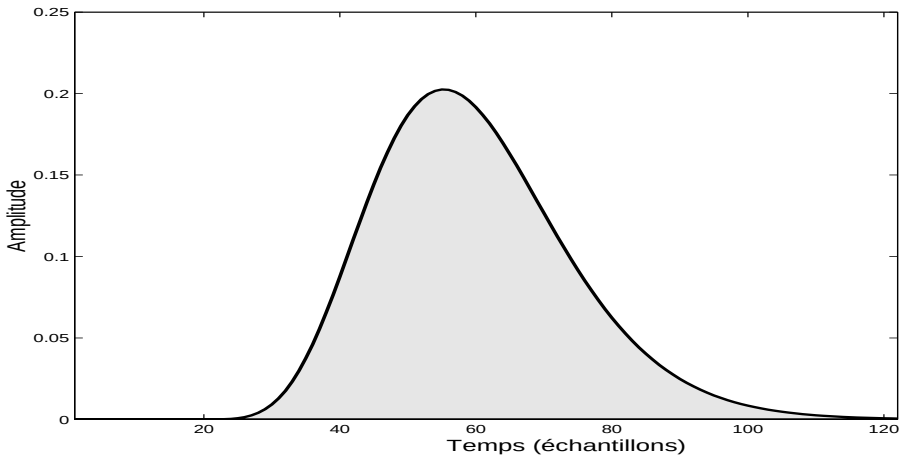


FIG. 1.13 – Fenêtre de Meixner pour $\beta = 20$, $\gamma = 0.8$ et un délai de 20 échantillons.

amorties en opposition de phases. La représentation est alors très inefficace et peu adéquate : de nombreuses sinusoïdes sont nécessaires pour modéliser du silence. Pour pallier ce problème, de nombreuses solutions ont été proposées. Une première consiste à placer les trames d'analyse au début des transitoires [PGV97], mais ceci amène à une segmentation irrégulière de l'axe temporel basée sur une détection préalable des attaques. Une seconde solution consiste à rajouter un délai pour chaque sinusoïde [Goo97, BAM02] pour obtenir un modèle à base de sinusoïdes amorties retardées (SAR). Une troisième solution considère que l'oreille n'est que peu sensible à un faible décalage temporel et propose donc d'aligner le début des transitoires sur le début des trames d'analyse [VHK01].

L'utilisation de ce type de modèle amène à considérer un grand nombre de paramètres. En considérant que la réponse acoustique est la même pour tout le spectre, on peut appliquer la même enveloppe temporelle à un ensemble de sinusoïdes stationnaires. Il est proposé dans [SOdBG02] d'utiliser une fenêtre dite de Meixner, représentée sur la figure 1.13. Cette enveloppe $g(n)$, calculée grâce à l'équation 1.12, a une attaque rapide contrôlée par le paramètre β suivie d'une décroissance exponentielle contrôlée par le paramètre γ :

$$g(n) = (1-\gamma^2)^{\beta/2}\sqrt{\frac{h(n)}{n!}}\,\gamma^n \tag{1.12}$$

$$h(n) = \beta\cdot(\beta+1)\cdot\cdots\cdot(\beta+n-1) \tag{1.13}$$

$$h(0) = 1 \tag{1.14}$$

avec $\beta > 0$, $0 < \gamma < 1$ et $n = 0, 1, 2, \ldots$. Cette approche est utilisé dans le codeur SSC [dBSO02] pour la modélisation des attaques et des transitoires.

Les modèles sinusoïdaux non stationnaires étant des généralisations du modèle stationnaire, l'utilisation de ce modèle permet de couvrir un large éventail de signaux. D'autres méthodes proposent des modèles explicitement dédiés aux signaux transitoires. Il est proposé dans [Dau00, Mol03] d'utiliser les arbres

dyadiques de coefficients d'ondelettes pour modéliser les transitoires. Le formalisme est alors complètement différent de celui utilisé pour les sinusoïdes et ne sera pas détaillé.

Dans [VM98, LS99], il est proposé d'utiliser un modèle sinusoïdal stationnaire pour modéliser les signaux transitoires dans le domaine fréquentiel grâce à la dualité temps/fréquence. Ceci permet de garder la flexibilité et les possibilités de manipulation des modèles sinusoïdaux tout en modélisant de façon explicite les signaux transitoires. Il est proposé d'utiliser la Transformée en Cosinus Discrète (TCD) [OS89] pour transformer le signal temporel. Comme on peut le constater sur la figure 1.14, le signal résultant d'une TCD d'un signal transitoire est compatible avec une modélisation sinusoïdale stationnaire. Certains paramètres, comme la position de l'attaque dans la trame d'analyse, peuvent être déduits simplement d'une analyse sinusoïdale de ce signal résultant.

À la lumière de ce bref historique, on peut constater que la modélisation sinusoïdale, avec certaines adaptations, peut être appropriée à la modélisation de signaux peu périodiques comme les sinusoïdes bruitées et les transitoires. Pour des applications musicales où la compacité de la représentation est un facteur négligeable en regard des potentialités de modifications expressives du son, il est même proposé dans [HDC03, Han03] de représenter un signal apériodique par des sinusoïdes de répartition fréquentielle aléatoire.

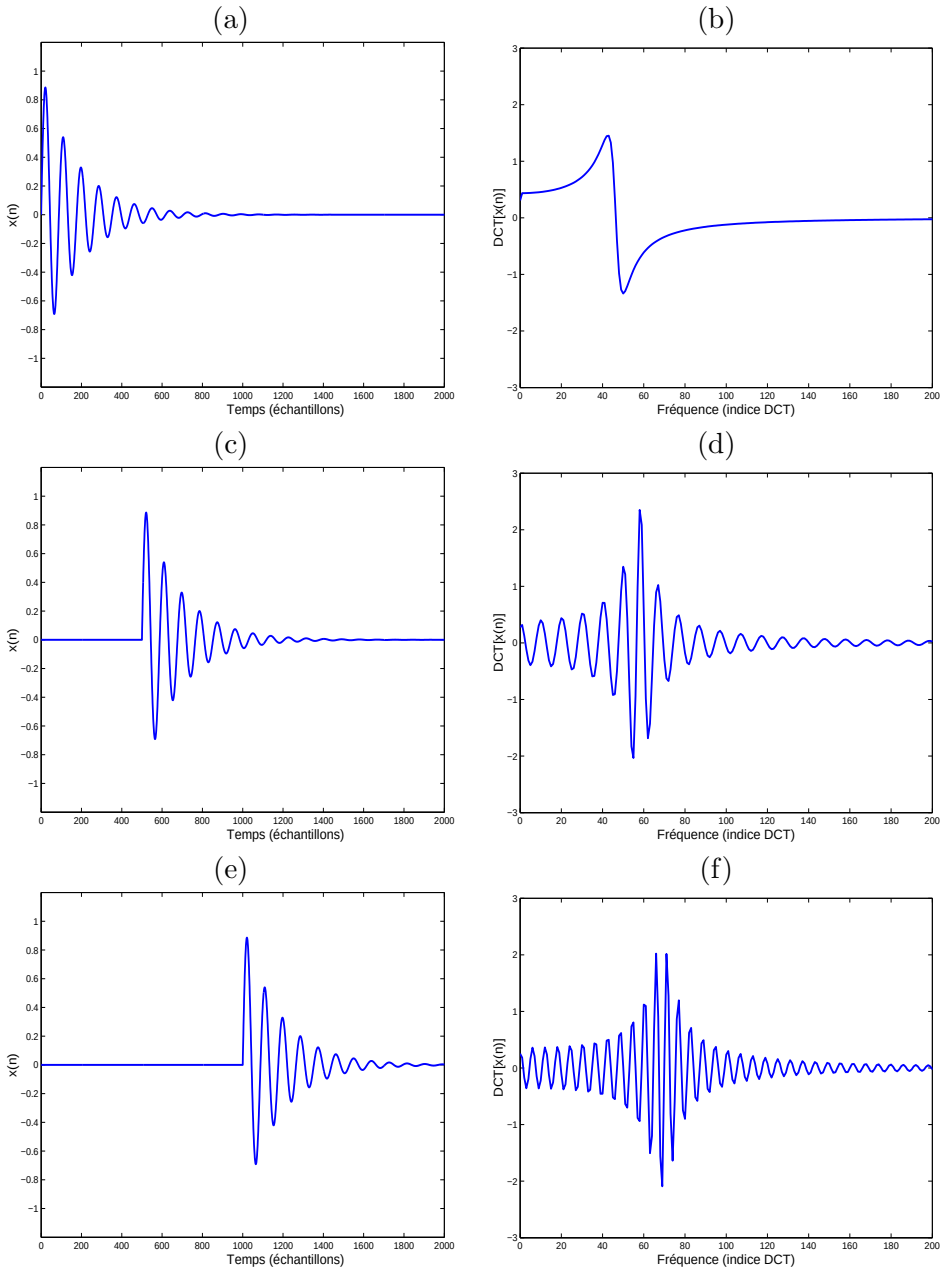


FIG. 1.14 – À gauche, le signal temporel d'une exponentielle amortie avec de haut en bas un délai de 0, 500 et 1000 échantillons. À droite, les transformées TCD correspondantes. Par la dualité temps/fréquence, l'utilisation d'un modèle sinusoïdal pour ce signal transformé est pertinent. Par exemple, un délai en temporel amène à une fréquence plus élevée du signal transformé.

1.4 Requis d'une modélisation à long terme

Comme on l'a vu dans la section 1.2.2, la modélisation long terme est d'un intérêt particulier. Or, les signaux de contrôle de l'amplitude, de la fréquence et de la phase de chaque partiel pour un signal particulier sont inconnus. En pratique, on estime les paramètres instantanés des partiels grâce à un modèle à court terme. Le principe des algorithmes dit de suivi de partiels est alors de déterminer quelles composantes court-terme (pics) doivent être allouées à une composante long-terme (partiel). Une fois la succession des pics identifiée pour un partiel, les signaux de contrôle peuvent alors être interpolés entre deux valeurs instantanées successives.

La complexité d'un signal polyphonique, l'ajout de bruit amènent des difficultés d'interprétation. Par exemple, dans une chaîne de traitement de type STB, la partie transitoire du signal peut être extraite directement du signal analysé comme dans le cas T_1 de la figure 1.11. La suppression d'une partie du signal amène à un manque d'informations spectrales nécessaires à l'analyse sinusoïdale pendant un intervalle de temps conséquent. Pendant quelques trames consécutives, aucun pic n'est disponible. La méthode de suivi de partiels doit donc être robuste à un manque temporaire d'informations spectrales. Dans le cas T_2 de la figure 1.11, la partie transitoire est extraite à partir du résiduel. La partie transitoire va donc corrompre pendant un bref instant les informations spectrales nécessaires à l'analyse sinusoïdale.

Les méthodes qui permettent le passage d'une représentation temporelle à une représentation sinusoïdale à court terme nécessitent peu de structures algorithmiques complexes. Pour chaque pic détecté à une trame donnée, les paramètres sinusoïdaux sont estimés selon des méthodes performantes détaillées dans le chapitre suivant. Il existe de plus de nombreux protocoles de validation de ces estimateurs comme les bornes de Cramér-Rao [Cra46].

En revanche, le passage d'une représentation sinusoïdale à court terme à une représentation à long terme n'est pas aussi immédiate. Des algorithmes plus complexes, comprenant de nombreuses prises de décision, doivent être mis en œuvre. Un algorithme dit de suivi de partiels doit déterminer, à une trame donnée, si un pic est la continuation d'un partiel existant, le début d'un nouveau ou un pic de bruit. À notre connaissance, il n'existe pas dans la littérature de méthodologie d'évaluation d'une méthode de suivi de partiels.

On introduit donc ici certains critères de validation qui seront utiles pour évaluer les algorithmes de suivi de partiels proposés dans le chapitre 3.

Formalisation

Nous proposons tout d'abord de formaliser l'opération de suivi de partiels de manière à ensuite proposer des critères d'évaluation. On adopte deux types de notations pour un pic selon son appartenance au modèle court-terme (p en minuscule) ou au modèle long-terme (P en majuscule). Soit $\mathcal{C}$, une représentation

à court terme composée de N_c trames :

$$\mathcal{C} = \bigcup_{n=1}^{N_c} \mathcal{C}_n \tag{1.15}$$

$$\mathcal{C}_n = \bigcup_{i=1}^{N_i} p_n^i \tag{1.16}$$

$$p_n^i = \{a_n^i, f_n^i, \phi_n^i\} \tag{1.17}$$

où $\mathcal{C}_n$ désigne l'ensemble des pics p_n^i à une trame d'indice n. Soit maintenant $\mathcal{S}$, une représentation à long terme :

$$\mathcal{S} = \bigcup_{k=1}^{N} P_k \tag{1.18}$$

où,

$$P_k = \{P_k(m), m = [n_k, \cdots, n_k + l_k - 1]\} \tag{1.19}$$

$$P_k(m) = \{A_k(m), F_k(m), \Phi_k(m)\} \tag{1.20}$$

où P_k est un partiel d'indice k né à la trame n_k et de longueur l_k. Ce partiel est un ensemble de pics dont les indices de trame sont successifs, où $P_k(m)$ représente le pic d'indice m du partiel k. Un pic ne peut être alloué qu'à un unique partiel. Cette représentation est donc soumise à l'allocation exclusive :

$$\forall i \neq j \; P_i(n) \neq P_j(n) \tag{1.21}$$

où n est un indice de trame.

Critères d'évaluation

Soit une représentation long-terme $\mathcal{S}$ connue (cette représentation peut être estimée à partir de signaux naturels ou créée artificiellement). Si on supprime toute notion de continuité de cette représentation, on obtient une représentation court-terme $\mathcal{C}$. Un algorithme de suivi de partiels est performant lorsqu'il est à même de retrouver une représentation long terme $\hat{\mathcal{S}}$ avec le même indiçage que celui de $\mathcal{S}$ à une permutation d'indices près, à partir de $\mathcal{C}$ ou d'une version dégradée par l'ajout de pics parasites, la suppression de pics ou la corruption des paramètres des pics existants. Dans la suite, nous considérons que les pics de bruit sont indicés avec l'indice 0 et n'appartiennent donc à aucun partiel.

Formellement, la représentation obtenue doit être **précise**. Un partiel de $\hat{\mathcal{S}}$ ne doit contenir des pics que d'un seul partiel de $\mathcal{S}$:

$$\forall i, \exists j \, | \forall k \, \hat{P}_i(k) \subseteq P_j \tag{1.22}$$

Réciproquement, la représentation sinusoïdale doit être **efficace**, tous les pics d'un partiel de l'ensemble $\mathcal{S}$ doivent être inclus dans un seul et unique partiel de l'ensemble $\hat{\mathcal{S}}$:

$$\forall i, \exists j \, | \forall k \, P_i(k) \subseteq \hat{P}_j \tag{1.23}$$

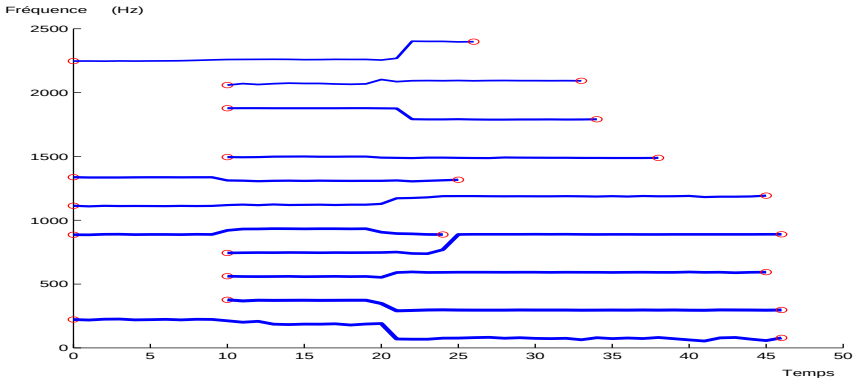


FIG. 1.15 – Représentation sinusoïdale à long terme possible des trois notes de piano de la figure 1.7. Cette représentation est peu précise (au sens de la notion de précision définie dans l'équation 1.22) car de nombreux partiels englobent des harmoniques de plusieurs notes différentes. Cette représentation est peu efficace (au sens de la notion d'efficacité définie dans l'équation 1.23) car de nombreuses harmoniques de la première note sont manquantes.

On peut définir des critères d'évaluation quantitatifs fonction de ces contraintes. Une représentation $\hat{\mathcal{S}}$ précise est une représentation qui comporte peu de partiels dont les pics appartiennent à deux partiels différents dans la représentation d'origine $\mathcal{S}$. Un critère quantitatif de précision C_p est alors :

$$C_p = 1 - \frac{\text{Card}\,\{\hat{P}_i \,|\, \hat{P}_i(n) \subseteq P_j \wedge \hat{P}_i(q) \subseteq P_k \wedge j \neq k\}}{\text{Card}\,\hat{\mathcal{S}}} \tag{1.24}$$

où i, j, et k sont des indices de partiels et n et q sont des indices de trames.

Réciproquement, une représentation $\hat{\mathcal{S}}$ est efficace si peu de partiels de $\mathcal{S}$ ont des pics alloués à plusieurs partiels de $\hat{\mathcal{S}}$ et peu des pics de $\mathcal{S}$ sont omis. Un critère quantitatif d'efficacité C_e est alors :

$$C_e = 1 - \frac{\text{Card}\,\{P_i \,|\, P_i(n) \subseteq \hat{P}_j \wedge P_i(q) \subseteq \hat{P}_k \wedge j \neq k \vee P_i(n) \not\subseteq \hat{P}_l\}}{\text{Card}\,\mathcal{S}} \tag{1.25}$$

où i, j, k et l sont des indices de partiels et n, q et r sont des indices de trames.

À titre d'exemple, une représentation long terme telle que chaque partiel ne contient qu'un seul pic comme celle de la figure 1.9 est une représentation précise mais en contrepartie très inefficace. La représentation à long terme de la figure 1.15 est plus efficace mais l'omission de certaines parties de la représentation d'origine font que l'efficacité n'est pas parfaite. De plus, cette représentation est très imprécise, car de nombreux partiels englobent des harmoniques de plusieurs notes différentes. La représentation de la figure 1.10 est une représentation précise car aucun partiel n'englobe deux harmoniques différentes. Cette représentation est efficace car aucun pic appartenant aux harmoniques du signal d'origine ne sont omis et aucune harmonique n'est séparée en plusieurs partiels.

2 Modélisation sinusoïdale à court terme

L'analyse sinusoïdale à court terme se compose de deux parties ; la première consiste à détecter la présence d'une composante sinusoïdale dans le signal analysé (pic dans le spectre de Fourier) et la seconde à estimer ses paramètres d'amplitude, de fréquence et de phase. Différentes méthodes d'estimation des paramètres stationnaires et non stationnaires sont étudiées dans ce chapitre. En particulier, une interprétation purement trigonométrique de l'estimateur de fréquence dit de la dérivée [DCM00] permet d'améliorer sa précision dans les hautes fréquences. On propose ensuite deux estimateurs de phase robustes aux modulations linéaires de fréquence. Lors de l'analyse de signaux bruités, de nombreux pics détectés ne sont pas pertinents. De manière à écarter ces pics non conformes au modèle sinusoïdal, on ne conserve que les pics dont le spectre est semblable au gabarit spectral d'une sinusoïde. Considérer les paramètres non stationnaires lors du calcul de ce gabarit permet de mieux distinguer les sinusoïdes modulées des pics de bruit et ainsi améliorer la sélection de pics par conformité au modèle sinusoïdal.

2.1 Analyse stationnaire

Dans un modèle stationnaire, on considère que les paramètres d'amplitude et de fréquence sont constants durant un court intervalle de temps, voir équation 1.4. Pour extraire les paramètres d'amplitude, de phase, et de fréquence des sinusoïdes, la majeure partie des implantations logicielles se basent sur une transformée spectrale bien connue, la transformée de Fourier. Une première partie présente d'abord cette transformée qui, grâce à des techniques de calcul efficaces, est encore aujourd'hui une estimation spectrale très largement utilisée. La suite traite de l'estimation des paramètres sinusoïdaux (fréquence, amplitude et phase) dans un spectre à court terme, calculé grâce à la transformée de Fourier.

2.1.1 Spectre de Fourier

Soit $x(t)$ une fonction de la variable t ; sous certaines conditions on démontre l'égalité suivante :

$$x_c(t) = \int_{-\infty}^{+\infty} X_c(f) e^{2j\pi ft} df \tag{2.1}$$

avec,

$$X_c(f) = \int_{-\infty}^{+\infty} x_c(t) e^{-2j\pi ft} dt \tag{2.2}$$

La fonction $X_c(f)$ est la transformée de Fourier de $x_c(t)$. Plus communément $X_c(f)$ est appelé spectre du signal $x_c(t)$. Prenons par exemple un cosinus :

$$s(t) = \cos(\omega t) \tag{2.3}$$

où ω désigne la pulsation en radians par secondes. Ce signal s'exprime sous forme d'exponentielles :

$$s(t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2} \tag{2.4}$$

Le spectre de $s(t)$ est composé de deux impulsions de Dirac, une située à la fréquence $-\omega$ et l'autre en la fréquence ω, comme représenté sur la figure 2.1.

Cas discret

Considérons maintenant la version discrète de ce signal :

$$x[n] \quad = \quad x_c(n/F_e) \tag{2.5}$$

où F_e désigne la fréquence d'échantillonnage. L'échantillonnage d'un signal peut être considéré comme une convolution du signal avec un train d'impulsions de Dirac. Comme la transformée de Fourier d'un train d'impulsions est un train d'impulsions, le spectre d'un signal échantillonné est le spectre du signal continu

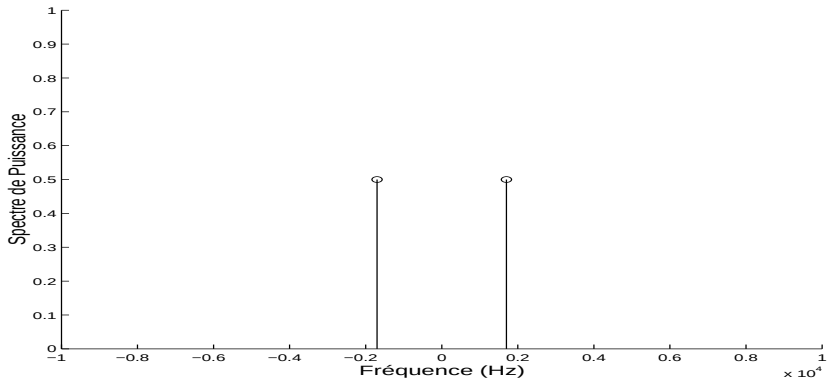


FIG. 2.1 – Transformée de Fourier continue d'une fonction cosinus.

multiplié par un train d'impulsions. Le spectre d'un signal échantillonné $X_e(f)$ est donc périodique de période F_e :

$$X_e(f) = F_e \sum_{k=-\infty}^{\infty} X_c(f - kF_e) \tag{2.6}$$

La version discrète de la transformée de Fourier (DFT) et de son inverse (IDFT) sont des opérations mathématiques définies comme suit :

$$X[m] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-2j\pi \frac{nm}{N}} \tag{2.7}$$

$$x[n] = \sum_{m=0}^{N-1} X[m] e^{2j\pi \frac{nm}{N}} \tag{2.8}$$

pour $m \in [0, ..., N-1]$. La DFT est une projection sur une base orthonormée d'exponentielles complexes de taille N. Les amplitudes de ces exponentielles forment le spectre de puissance du signal et les phases forment le spectre de phase.

Si le signal analysé est réel, les coefficients $X[0]$ et $X[N/2]$ sont respectivement la moyenne des échantillons temporels et la différence des échantillons pairs et impairs, et sont donc purement réels. Au-delà de la fréquence dite de Nyquist ($F_e/2$), les autres coefficients sont complexes conjugués, $X[k] = X^*[N-k]$. Dans la suite, on représentera donc un spectre discret à partir de l'indice 0 à $N/2$.

Des équations 2.7 et 2.8 découle la propriété de linéarité de la transformée de Fourier :

$$\mathrm{DFT}[a\, x_1[n] + b\, x_2[n]] = a\, X_1[m] + b\, X_2[m] \tag{2.9}$$

Le calcul d'une DFT a une complexité en $O(N^2)$ multiplications complexes. Dans le cas où le nombre d'échantillons est un multiple de deux, on peut réduire

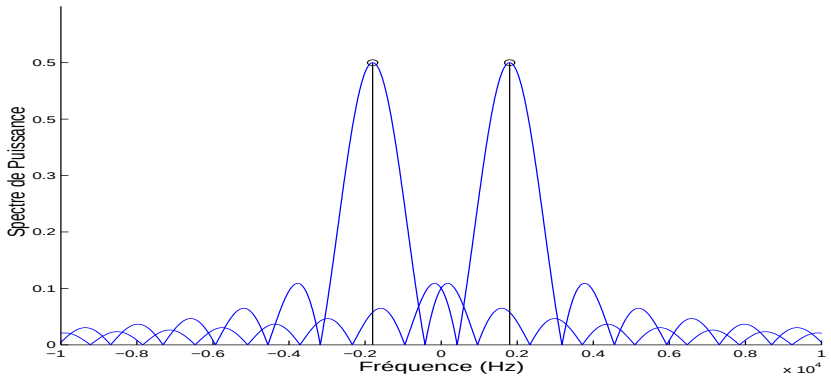


FIG. 2.2 – Convolution du spectre d'une fonction cosinus par le spectre d'une fenêtre rectangulaire.

ce coût de calcul en $O(N\log(N))$. En effet, le calcul d'une DFT d'ordre N revient au calcul de deux transformées d'ordre $N/2$ auquel on ajoute $N/2$ multiplications complexes. Par $\log(N/2)$ itérations de ce processus de simplification, on aboutit à des transformées simples qui ne demandent pas de multiplication. Ce type d'algorithme est appelé FFT de l'anglais "*Fast Fourier Transform*".

L'équation 2.7 effectue un fenêtrage implicite du signal car seulement N échantillons du signal sont exploités pour calculer le spectre de Fourier. D'une manière plus générale, on calcule :

$$X[m] = X\left(m\,\frac{F_e}{N}\right) = \frac{1}{N}\sum_{n=0}^{N-1} w[n]\,x[n]e^{-\frac{2j\pi}{N}nm} \tag{2.10}$$

pour $m \in [0, ..., N-1]$ et $w[n]$ étant une certaine fenêtre de pondération. On se ramène à l'équation 2.7 avec la fenêtre rectangulaire :

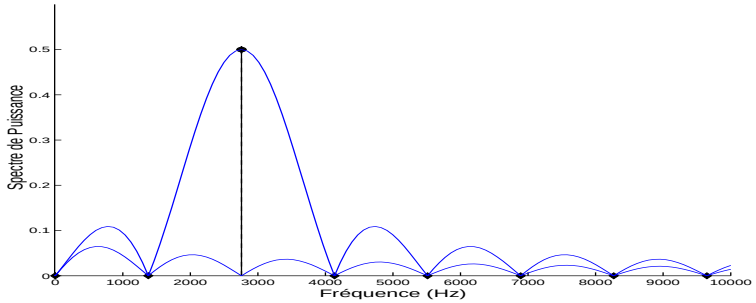
$$w_r[n] = \begin{cases} 1 \text{ pour } 0 \leq n < N \\ 0 \text{ sinon} \end{cases} \tag{2.11}$$

Une multiplication dans le domaine temporel équivaut à une convolution dans le domaine spectral. La DFT d'une fonction cosinus est donc la convolution de deux impulsions de Dirac avec le spectre de la fenêtre de pondération comme illustré dans la figure 2.2. On peut remarquer que les deux spectres se chevauchent. Ce phénomène s'observe aussi bien en proximité de 0 que de F_e car le spectre d'un signal échantillonné est périodique de période F_e, voir équation 2.6.

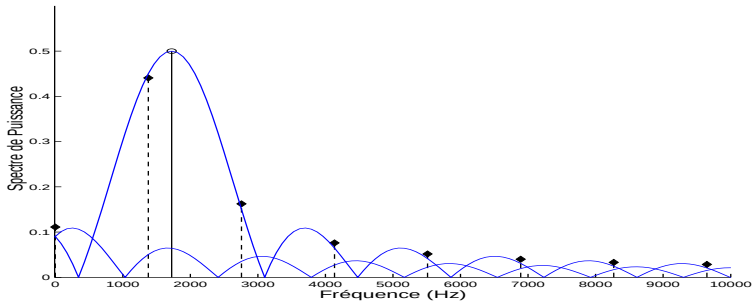
Si la fréquence du cosinus vérifie la relation suivante :

$$f = k\,\frac{F_e}{N} \tag{2.12}$$

k étant un entier compris entre 0 et $N/2$, une seule composante de la DFT du signal fenêtré par une fenêtre rectangulaire a une norme égale à la moitié de l'amplitude du cosinus analysé, tandis que toutes les autres ont une amplitude



(a)



(b)

FIG. 2.3 – En haut, cas de figure idéal où la fréquence du cosinus (trait plein) est multiple de F_e/N, l'énergie est alors concentrée sur une seule composante DFT. En bas, cas de figure général, où l'énergie est dispersée sur toutes les composantes DFT (tirets).

égale à zéro comme illustré dans la figure 2.3(a). Si la fréquence du cosinus ne vérifie pas cette relation, l'énergie du signal est dispersée sur toutes les composantes de la DFT comme illustré dans la figure 2.3(b). Ce phénomène peut être observé de manière équivalente en considérant la DFT comme un banc de filtres passe-bande. Chaque composante de la DFT est considérée comme un filtre qui a pour réponse fréquentielle le spectre de la fenêtre convoluée par une impulsion de Dirac de fréquence f vérifiant l'équation 2.12. La valeur de l'énergie récupérée par chaque composante de la DFT est donc celle de la réponse de son filtre à la fréquence du cosinus analysé comme illustré dans la figure 2.4.

Dans le cas général, la résolution fréquentielle dépend de la taille de la fenêtre utilisée pour le calcul de la DFT. Pour améliorer la résolution, une première solution consiste alors à augmenter le nombre d'observations. Cette augmentation de la précision fréquentielle implique un coût de calcul plus important et se fait au détriment de la précision temporelle. Les signaux sonores étant en général non stationnaires, le nombre d'échantillons temporels nécessaires à l'estimation spectrale doit être le plus faible possible. Une solution pour améliorer la précision de la DFT sans augmenter le nombre de points d'observation est d'augmenter la sélectivité des filtres passe-bande en utilisant une fenêtre de pondération plus appropriée.

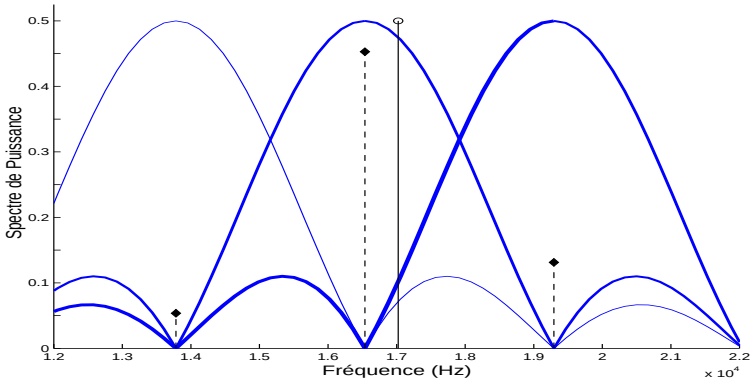


FIG. 2.4 – Représentation de la DFT sous forme de banc de filtres passe-bande. L'énergie récupérée par chaque composante de la DFT (tirets) est celle de la réponse de son filtre à la fréquence du cosinus analysé (trait plein).

2.1.2 Fenêtre de pondération

En utilisant une fenêtre de pondération particulière, on cherche généralement à accélérer la décroissance des lobes secondaires (ceux qui ne sont pas centrés en la fréquence de la composante de la DFT) tout en minimisant l'élargissement du lobe principal (lobe qui est centré en la fréquence de la composante de la DFT).

La fenêtre rectangulaire w_r définie par l'équation 2.11 a un spectre W_r défini par :

$$W_r(\omega) = e^{-j\omega \frac{N-1}{2}} \frac{\sin(\omega\, N/2)}{\sin(\omega/2)} \tag{2.13}$$

où ω est exprimé en radians par secondes et N est la taille de la fenêtre en nombre d'échantillons. Parmi les nombreuses autres fenêtres proposées dans la littérature, nous n'étudierons que les fenêtres dites trigonométriques :

$$w_\alpha(n) = \alpha - (1-\alpha)\cos\left(\frac{2\pi n}{N}\right) \tag{2.14}$$

Le spectre de cette fenêtre s'exprime sous la forme suivante :

$$W_\alpha(\omega) = \alpha\, W_r(\omega) + \frac{(1-\alpha)}{2} W_r\left(\omega - \frac{1}{N}\right) + \frac{(1-\alpha)}{2} W_r\left(\omega + \frac{1}{N}\right) \tag{2.15}$$

où ω est exprimé en radians et N est la taille de la fenêtre en nombre d'échantillons. L'équation 2.14 est une expression générale qui permet de parcourir l'ensemble des fenêtres trigonométriques. La fenêtre rectangulaire w_r se retrouve avec $\alpha = 1$. La fenêtre de Hamming w_m utilise un $\alpha = 0.54$, tandis que la fenêtre de Hann w_n utilise un $\alpha = 0.5$. On constate sur la figure 2.5 que la fenêtre rectangulaire présente un lobe principal très fin ($2F_e/N$ Hz) et en contrepartie une atténuation faible des lobes secondaires (-6 dB par octave). La fenêtre de Hamming présente une forte atténuation des lobes secondaires (-43 dB) au détriment de l'atténuation asymptotique des autres lobes. La fenêtre de Hann

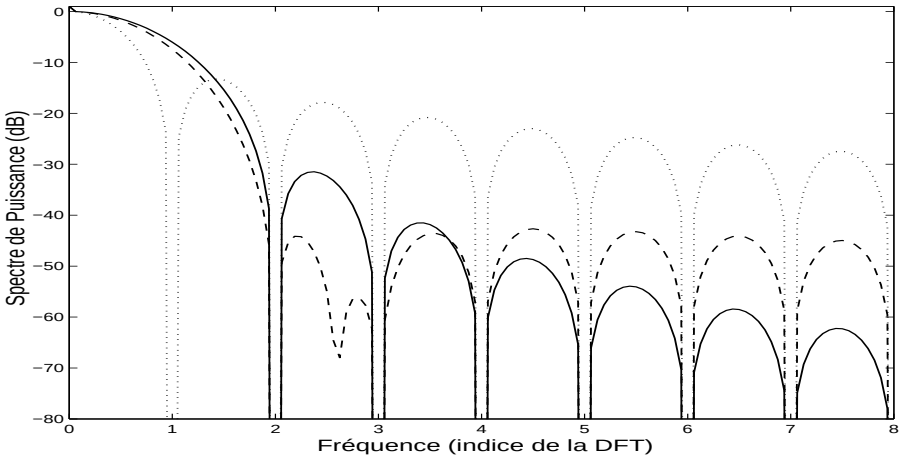


FIG. 2.5 – Spectres de trois fenêtres de pondération trigonométriques. En pointillés, la fenêtre rectangulaire présente un lobe principal très fin et en contrepartie une atténuation faible des lobes secondaires. En tirets, la fenêtre de Hamming présente une atténuation quasi optimale du second lobe mais l'atténuation asymptotique des autres lobes est faible. En trait plein, la fenêtre de Hann a un lobe principal plus large, mais présente une atténuation asymptotique des lobes secondaires particulièrement prononcée.

a un lobe principal un peu moins précis que celui de la fenêtre de Hamming et des lobes secondaires plus élevés, mais présente une atténuation asymptotique des lobes secondaires particulièrement prononcée (-18 dB par octave).

Comme le montre l'équation 2.15, le spectre d'une fenêtre trigonométrique est composé du spectre de la fenêtre rectangulaire et de deux autres composantes. Étudions ces deux dernières composantes pour la fenêtre de Hamming et celle de Hann. Elles élargissent le lobe principal et, en contrepartie, s'opposent aux lobes secondaires de la fenêtre rectangulaire, voir figure 2.6. On peut remarquer que avec le paramètre $\alpha = 0.54$, on obtient une corrélation presque parfaite entre le spectre de la fenêtre rectangulaire et celui des deux composantes au niveau du second lobe, ce qui explique l'atténuation très forte des lobes secondaires de la fenêtre de Hamming. Dans le cas de fenêtre de Hann, cette corrélation est moins prononcée au niveau du second lobe. Par contre, elle s'améliore asymptotiquement pour se rapprocher des lobes secondaires de la fenêtre rectangulaire et permettre une atténuation asymptotique prononcée.

La comparaison des performances de ces fenêtres sort du propos de ce document. On retiendra néanmoins que la fenêtre de Hann est un bon candidat pour l'estimation des paramètres sinusoïdaux [Mar00b].

2.1.3 Estimation de la fréquence

Déterminer la fréquence d'une sinusoïde dans le spectre de Fourier consiste à isoler un maximum local du spectre de puissance et à estimer sa fréquence.

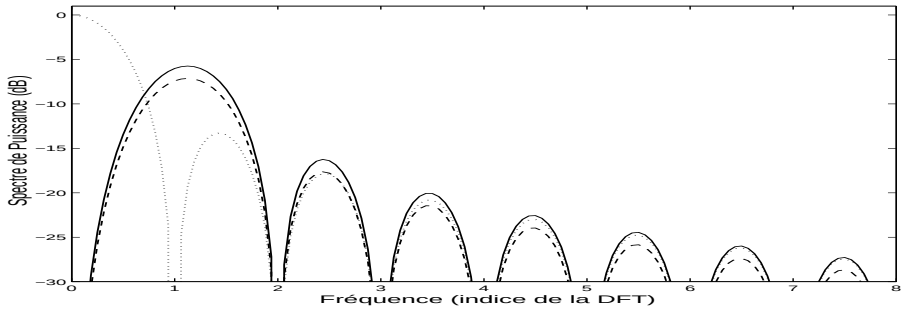


FIG. 2.6 – Influence des deux dernières composantes de l'équation 2.15. En pointillés, le spectre de la première composante (la fenêtre rectangulaire). En tirets et en trait plein, on représente le spectre des deux dernières composantes de cette équation pour la fenêtre de Hamming et de Hann.

Méthode des maxima locaux

Le spectre de Fourier d'une composante sinusoïdale de fréquence f a l'allure présentée figure 2.3(a), avec un maximum en f. Grâce à la DFT, on dispose des valeurs du spectre de Fourier aux points $k\,\frac{F_e}{N}$. Supposons un maximum en k, la fréquence estimée par la méthode des maxima locaux est :

$$\hat{f} = k\,\frac{F_e}{N} \tag{2.16}$$

L'erreur de cet estimateur est bornée par $\frac{F_e}{2N}$. De manière à réduire l'erreur possible, on peut augmenter N, l'ordre de la DFT. On souhaite dans le même temps conserver un nombre d'échantillons faible. De manière à calculer plus de points fréquentiels sans augmenter le nombre d'échantillons observés, on peut utiliser la méthode dite du "*zero-padding*", voir figure 2.7. Les $(Z-1)N$ zéros sont ajoutés aux échantillons temporels avant d'effectuer une transformée de Fourier de taille ZN. Le gain de précision apporté par cette méthode pour l'estimation de la fréquence est évalué dans la partie gauche de la table 2.1. Il est à noter que la résolution fréquentielle n'est en rien améliorée par ce type de manipulation, aussi on demeure incapable de distinguer deux cosinus de fréquence f_1 et f_2 tels que $|f_1 - f_2| < F_e/N$. On estime simplement le spectre sur plus de points. En particulier, l'utilisation de cette méthode sur des signaux bruités amène à la détection de plus de maxima locaux comme on peut le voir sur la figure 2.8.

L'ajout d'un nombre conséquent de zéro engendre un surcoût en temps de calcul. Certaines simplifications peuvent être apportées à l'algorithme FFT, car des multiplications complexes par des échantillons nuls peuvent être évitées [Mar71, Ski76, Hol87]. Le gain est malheureusement assez faible, car la complexité en $O(ZN\log(ZN))$ est réduite à une complexité en $O(ZN\log(N))$.

Une autre solution consiste à interpoler le spectre DFT à proximité du maximum local en utilisant une parabole passant par les points $|X[k-1]|$, $|X[k]|$ et $|X[k+1]|$ comme approximation du lobe principal sur l'échelle des dB [PTVF92]. Ce type d'interpolation est particulièrement adapté à la fenêtre

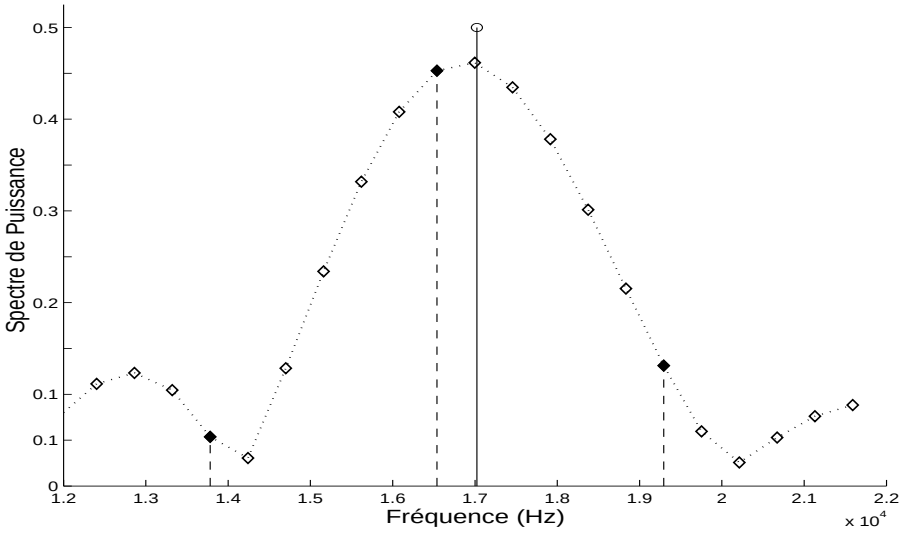


FIG. 2.7 – Sont représentées sur cette figure, la composante sinusoïdale analysée (trait plein), les composantes de la DFT (tirets) et les composantes de la DFT avec un facteur de *zero-padding* de 6 (diamants évidés).

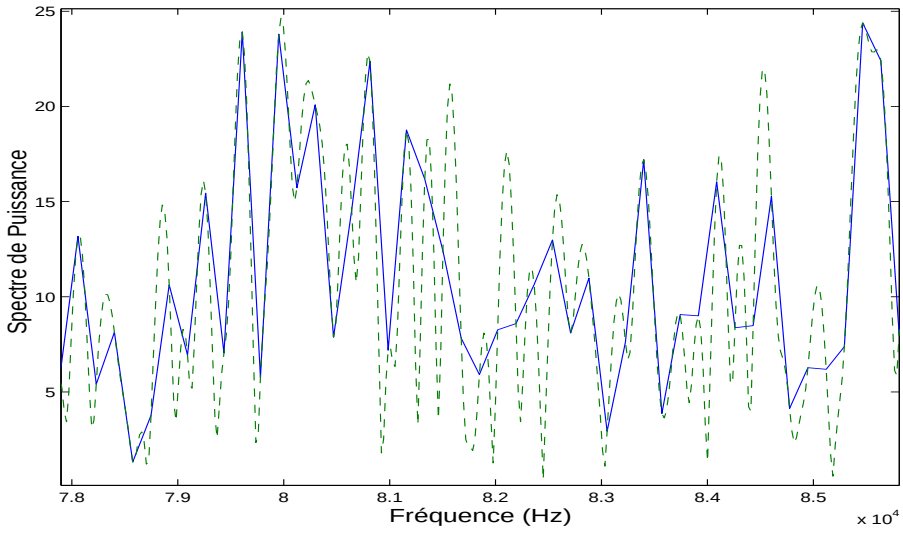


FIG. 2.8 – Spectre DFT d'un bruit blanc (trait plein) et spectre DFT du même signal avec un facteur de *zero-padding* de 8 (tirets).

gaussienne car son lobe principal est exactement une parabole sur l'échelle des déciBels. La méthode du *zero-padding* et l'interpolation parabolique peuvent être combinées pour un meilleur compromis complexité / précision comme décrit dans [Ser97].

Réassignement spectral

Auger et Flandrin proposent dans [AF95] d'utiliser la connaissance de la formule analytique de la dérivée de la fenêtre d'analyse de manière à ajuster la fréquence du maximum local du spectre DFT. Cette méthode est appelée le réassignement fréquentiel. La fréquence réassignée $\hat{f}_r$ du maximum local à l'indice k est donnée par l'équation suivante :

$$\hat{f}_r = k\frac{F_e}{N} - \Im\left(\frac{X_{w'}[k]}{X_w[k]}\right)\frac{F_e}{2\pi} \tag{2.17}$$

où $X_w(k)$ et $X_{w'}(k)$ sont les composantes d'index k des spectres calculés en utilisant la fenêtre $w(n)$ ou sa dérivée $w'(n)$ et $\Im(X)$ désigne la partie imaginaire de X. Cette méthode est utilisée par Fitz [Fit99] et Peeters [PR99].

Méthode de la dérivée

L'analyse de Fourier d'ordre m introduite dans [DCM00] montre qu'il est possible d'améliorer la précision de l'analyse de Fourier classique en considérant les m premières dérivées du signal. Pour $m = 1$, cette méthode est connue sous le nom d'algorithme de la dérivée. En pratique, la dérivée du signal continu est inconnue. On approxime donc cette dérivée par la différence du signal à un instant n avec ce signal à un instant $n-1$ divisée par l'intervalle de temps qui les sépare :

$$x^0[n] = s[n] \tag{2.18}$$
$$\tilde{x}^1[n] = F_e(x^0[n] - x^0[n-1]) \tag{2.19}$$

Cet estimateur exploite le fait que la dérivée d'un cosinus est un cosinus de même fréquence mais d'amplitude et de phase différente. Le rapport entre les amplitudes de ces deux cosinus permet d'estimer la fréquence du cosinus de manière précise. La fréquence corrigée $\hat{f}_d$ du maximum local à l'indice k est donnée par l'équation suivante :

$$\hat{f}_d = \frac{F_e}{\pi}\arcsin\left(\frac{1}{2F_e}\frac{|X^1[k]|}{|X^0[k]|}\right) \tag{2.20}$$

où $X^m(k)$ est la composante d'indice k du spectre de la dérivée d'ordre m du signal x. Il est à noter que l'amplitude du cosinus et la fenêtre de pondération utilisée influencent $X^0(k)$ et $X^1(k)$ d'un même facteur. Ces influences se compensent donc par la division de l'équation 2.20. Cet estimateur donne de très bonnes performances en utilisant la fenêtre de Hann.

Des études ont été menées pour comparer les performances des estimateurs utilisant le réassignement ou la méthode de la dérivée. Dans [KM02],

ces estimateurs sont comparés en terme d'erreur moyenne, de variance et d'erreur maximale, avec différents types de signaux synthétiques ou naturels. Ces expérimentations montrent que les deux méthodes d'estimation sont (au moins en pratique) équivalentes. Dans [HM03], ces estimateurs sont comparés en utilisant le formalisme des bornes de Cramér-Rao. La borne basse de Cramér-Rao est définie comme la limite de la meilleure performance atteignable pour un estimateur en fonction d'un jeu de données. Les résultats de ces expérimentations pointent un défaut de la méthode de la dérivée : sa précision se dégrade en se rapprochant de la fréquence de Nyquist.

L'étude d'une approche purement trigonométrique du problème de l'estimation de la fréquence d'un cosinus, détaillée dans l'appendice A, permet de proposer une amélioration de cet estimateur dans les hautes fréquences pour une complexité égale. Soit $s(n)$ un signal fenêtré, on calcule :

$$S_0[k] = \text{DFT}\ [s(n), \cdots, s(n+N-1)] \tag{2.21}$$

$$S_1[k] = \text{DFT}\ [s(n-1), \cdots, s(n+N)] \tag{2.22}$$

Pour chaque maximum local d'indice k du spectre S_0, la fréquence estimée est :

$$\hat{f} = \begin{cases} \frac{F_e}{\pi} \arcsin\left(\frac{|S_1[k]-S_0[k]|}{|S_0[k]|}\right) & \text{si } k < N/4 \\ \frac{F_e}{\pi} \arccos\left(\frac{|S_0[k]+S_1[k]|}{|S_0[k]|}\right) & \text{sinon} \end{cases} \tag{2.23}$$

De manière à apprécier le gain de précision apporté par cette méthode, on la compare à la méthode des maxima locaux pour une DFT avec un facteur de *zero-padding* variable. La méthodologie de test utilisée pour obtenir les résultats de la table 2.1 est utilisée pour tous les tests de ce chapitre. On utilise 2048 échantillons temporels, toutes les fréquences de 0 à $F_e/2$ par pas de 10 Hz sont testées dans un premier temps. Ensuite, toutes les fréquences de $F_e/4 - 2F_e/2048$ à $F_e/4 + 2F_e/2048$ par pas de 1 Hz sont testées.

Ces méthodes d'estimation de la fréquence impliquent le calcul de deux FFT et sont donc équivalentes en terme de complexité à une FFT avec un facteur de *zero-padding* de 2. Comme on peut le voir sur la table 2.1 le gain de précision est conséquent. De plus, on remarque que la méthode d'interpolation par *zero-padding* n'apporte que peu de gain de précision à la méthode.

2.1.4 Estimation de l'amplitude

Soit un maximum local du spectre de puissance à l'indice k, le module $|X(k)|$ est une approximation de la moitié de l'amplitude du cosinus :

$$\hat{a}_{\text{DFT}} = 2\,|X[k]| \tag{2.24}$$

Encore une fois, cette estimation est parfaite si la fréquence du cosinus vérifie $f = k\,\frac{F_e}{N}$, comme on le montre la figure 2.3(a). La méthode du *zero-padding* ou l'interpolation polynomiale citées précédemment sont utiles pour estimer de façon plus précise l'amplitude d'un cosinus. En effet, en interpolant le lobe principal, ces méthodes permettent d'estimer de façon plus précise la position du maximum en fréquence mais aussi en amplitude.

Z	f_{DFT}			$\hat{f}$		
	moyenne	variance	maximum	moyenne	variance	maximum
1	5.4004	9.8831	21.533	0.014301	0.13937	15.448
2	2.7169	2.6655	20	0.0159	0.15945	15.4486
3	1.815	1.2814	12.822	0.015658	0.12491	10.244
4	1.37	0.86601	14.616	0.01757	0.16017	11.978
5	1.0963	0.58606	11.533	0.014405	0.10113	8.878
6	0.91843	0.50238	12.822	0.015751	0.12496	10.244
7	0.791	0.45879	13.847	0.016452	0.14405	11.237
8	0.69335	0.37397	11.92	0.014856	0.10955	9.385
9	0.62003	0.36435	12.822	0.015728	0.12494	10.244
10	0.56083	0.36052	13.54	0.016273	0.13814	10.939

TAB. 2.1 – Erreur d'estimation de la fréquence en fonction du facteur de *zero-padding* pour deux types d'estimateurs. Le premier est la fréquence de la composante de la DFT d'amplitude la plus élevée et le second est l'estimateur proposé. Pour ces deux estimateurs, on considère l'erreur moyenne, la variance et l'erreur maximale. Le gain de précision de l'estimateur $\hat{f}$ est conséquent et ne s'améliore pas en fonction du facteur de *zero-padding*.

Dans le cas où l'on estime la fréquence $\hat{f}$ grâce aux méthodes décrites précédemment (méthode du réassignement, méthode de la dérivée), on peut corriger l'amplitude du maximum local. Comme le spectre d'un cosinus est le spectre de la fenêtre d'analyse centré en la fréquence $\hat{f}$, il suffit de connaître l'écart entre la fréquence du maximum local (deuxième composante DFT sur la figure 2.3(b)) et la fréquence $\hat{f}$ pour en déduire l'amplitude corrigée :

$$\hat{a}_w = 2\,\frac{|X[k]|}{|W(\hat{f} - k \cdot F_e/N)|} \tag{2.25}$$

où $W(f)$ est la réponse en fréquence (le spectre) de la fenêtre d'analyse. Cette réponse peut être calculée de manière analytique grâce à l'équation 2.15 dans le cas des fenêtres trigonométriques. La réponse en fréquence de certaines fenêtres n'étant pas connue de façon analytique, on peut pré-calculer celle-ci pour un ensemble fini de fréquences à l'aide par exemple d'une FFT avec un facteur de *zero-padding* fonction de la précision désirée.

Une autre méthode consiste à calculer directement le spectre du signal fenêtré pour la fréquence estimée $\hat{f}$ à l'aide d'une transformée de Fourier en un point :

$$X(\hat{f}) = \frac{1}{N}\sum_{n=0}^{N-1} w[n]x[n]\,e^{-2j\pi n\frac{\hat{f}}{F_e}} \tag{2.26}$$

où $w[n]$ est la fenêtre de pondération et $x[n]$ le signal considéré. L'amplitude estimée est alors :

$$\hat{a}_c = 2\,\frac{|X(\hat{f})|}{|W(0)|} \tag{2.27}$$

et

$$W(0) = \sum_{n=0}^{N-1} w(n) \tag{2.28}$$

La complexité de cette opération provient en grande partie du calcul des fonctions cosinus et sinus, parties réelles et imaginaires de l'exponentielle complexe. On peut réduire la complexité de cette opération en utilisant un algorithme de calcul récursif présenté dans la section 2.2.

2.1.5 Estimation de la phase

On considère généralement que l'oreille est peu sensible à un déphasage du signal [Ris91]. La synthèse d'un son perceptuellement proche de l'original peut donc être obtenue sans prendre en compte le paramètre de phase obtenu à l'analyse. En revanche, si le modèle sinusoïdal est intégré dans un modèle hybride, la forme d'onde du signal synthétisé doit être en phase avec le signal original pour permettre le calcul du signal résiduel par soustraction.

Les paramètres d'amplitude et de fréquence des composantes de la DFT sont une moyenne de l'évolution de ces paramètres dans la fenêtre d'observation, de n à $n+N-1$. On considère alors que ces paramètres désignent l'amplitude et la fréquence de la sinusoïde à l'échantillon $n+N/2$. Or, les phases des composantes DFT désignent les phases à l'origine, c'est-à-dire à l'échantillon n. Pour obtenir les phases à l'instant $n+N/2$, on doit opérer une rotation de phase. Une première méthode consiste à incrémenter la phase de $N/2$ fois la pulsation de la composante de la DFT :

$$\varphi\left(\frac{N}{2}\right) = \varphi(0) + 2\pi\frac{N}{2}\frac{f}{F_e} \tag{2.29}$$

Une autre méthode, plus efficace, consiste à effectuer une translation circulaire des échantillons temporels avant de calculer les composantes DFT. En effet, une translation circulaire des échantillons temporels entraîne une rotation des phases des composantes DFT :

$$\sum_{n=-\infty}^{\infty} x[n-n_0]e^{-2j\pi\frac{nk}{N}} = X[k]\, e^{-2j\pi\frac{n_0 k}{N}} \tag{2.30}$$

En pratique, soit un tableau d'échantillons temporels :

$$[x(n), \cdots, x(n+N-1)]$$

Ces échantillons sont fenêtrés par une fenêtre symétrique de longueur impaire $N-1$:

$$[0, x_w(n+1), \cdots, x_w(n+N/2), x_w(n+N/2+1), \cdots, x_w(n+N-1)]$$

Ces échantillons sont ensuite translatés circulairement pour obtenir la répartition suivante :

$$[x_w(n+N/2+1), \cdots, x_w(n+N-1), 0, x_w(n+1), \cdots, x_w(n+N/2)]$$

Si l'on opère cette translation à une fenêtre donnée (trigonométrique par exemple), le spectre obtenu par calcul de la DFT est purement réel. L'opération de fenêtrage n'introduit donc pas de déphasage supplémentaire. Le spectre est alors dit "à phase nulle".

L'estimateur basique de la phase d'une sinusoïde est la phase de la composante de la DFT la plus proche en fréquence. L'estimateur est alors :

$$\phi_{\mathrm{DFT}} = \angle X[k] \tag{2.31}$$

où k est l'index DFT du maximum local. L'utilisation d'un spectre à phase nulle rend cet estimateur plus robuste, car la phase d'un cosinus est constante dans le lobe principal, voir figure 2.9(d). Ainsi, une faible erreur d'estimation de la fréquence (inférieure à la moitié de la largeur du lobe) n'entraîne pas une estimation faussée de la phase.

Il est proposé dans [KAZ01] un estimateur de phase qui considère la moyenne des valeurs complexes des composantes d'une DFT autour du maximum local (la DFT est à phase nulle et comporte un facteur de *zero-padding* supérieur à 1) :

$$\overline{X_k} = \frac{1}{3} \sum_{l=k-1}^{k+1} \frac{X[l]}{|X[l]|} \tag{2.32}$$

où k est l'index DFT du maximum local. L'estimateur de phase ϕ_m est alors :

$$\phi_m = \angle \overline{X_m} \tag{2.33}$$

Lorsque la fréquence de la sinusoïde est connue avec une précision supérieure à la résolution de la DFT, la transformée de Fourier en $\hat{f}$ (voir équation 2.26) peut être utilisée pour estimer la phase de manière plus précise. L'estimateur est alors :

$$\phi_c = \angle X(\hat{f}) \tag{2.34}$$

où $\hat{f}$ désigne la fréquence estimée. Les performances de ces trois estimateurs sont très similaires avec un SNR supérieur à 0 dB, comme on peut le voir sur la table 2.2. Au-delà, ϕ_c présente une erreur moins prononcée. Au vu de ces résultats effectués sur des signaux synthétiques stationnaires, les deux derniers estimateurs n'apportent pas d'amélioration sensible de la précision.

Le problème de l'estimation de la phase lors de l'analyse de signaux non stationnaires sera étudié dans la section 2.4.3 et deux estimateurs qui améliorent sensiblement la précision de l'estimateur basique ϕ_{DFT} seront proposés.

2.1.6 Sélection de pics par conformité au modèle

De nombreux maxima locaux du spectre à court terme sont des pics dits de bruit, provenant de processus stochastiques ou transitoires. Il convient de les écarter pour ne modéliser que la partie pseudo périodique du son. Pour cela, on peut définir des critères qui permettent de classer les pics candidats en fonction de leur "proximité" avec une sinusoïde.

Le spectre d'une sinusoïde continue est une impulsion de Dirac. Il est donc pertinent de considérer les maxima locaux du spectre DFT pour identifier une

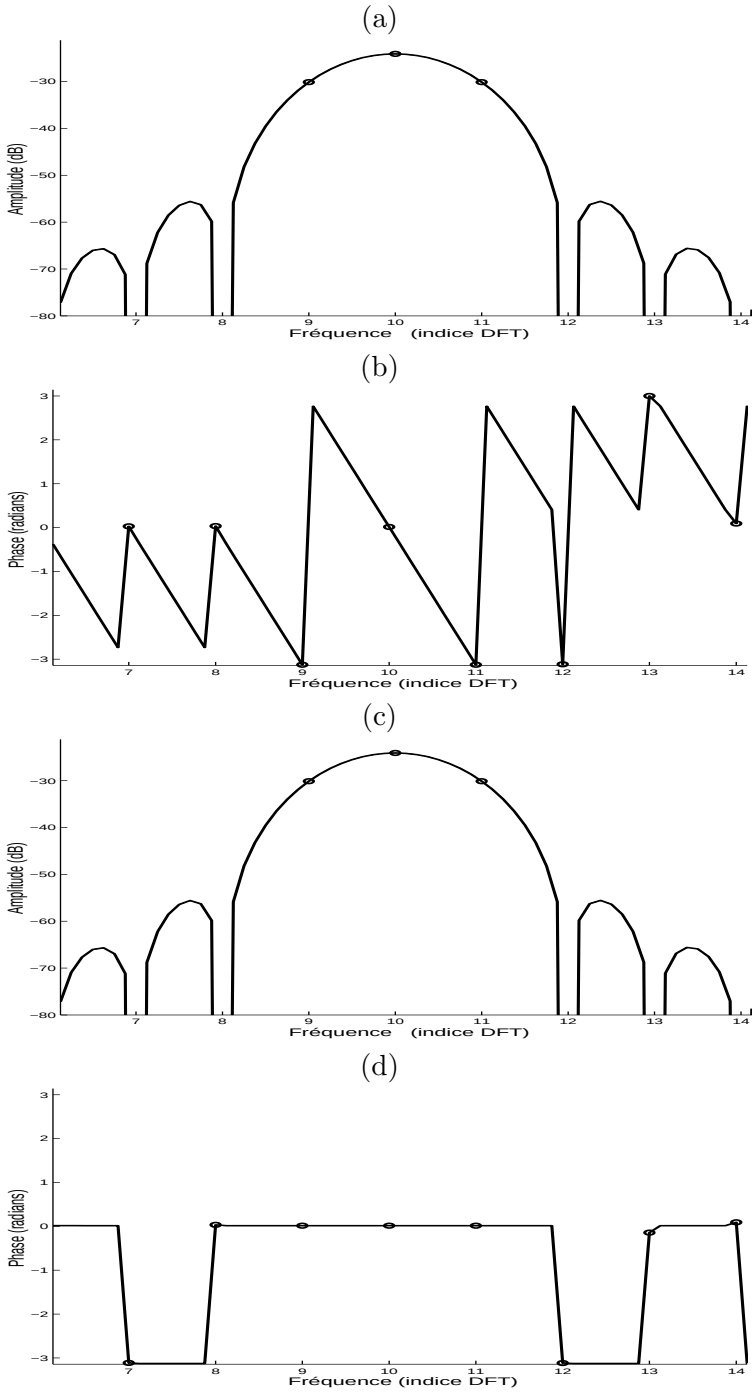


FIG. 2.9 – Spectres de puissance et de phase d'une sinusoïde utilisant un fenêtrage linéaire classique (a) et (b) et un fenêtrage à phase nulle (c) et (d). Les points représentent les composantes de la DFT sans *zero-padding*. La ligne représente le spectre de puissance ou de phase interpolé par *zero-padding*.

SNR	ϕ_{DFT}		ϕ_m		ϕ_c	
	moyenne	variance	moyenne	variance	moyenne	variance
90	9.90 E-07	4.98 E-13	9.93 E-07	5.00 E-13	9.92 E-07	5.00 E-13
80	3.04 E-06	5.96 E-12	3.05 E-06	5.99 E-12	3.03 E-06	5.92 E-12
70	9.22 E-06	5.51 E-11	9.25 E-06	5.54 E-11	9.20 E-06	5.59 E-11
60	3.32 E-05	6.64 E-10	3.33 E-05	6.67 E-10	3.32 E-05	6.68 E-10
50	9.84 E-05	5.18 E-09	9.87 E-05	5.22 E-09	9.80 E-05	5.14 E-09
40	3.14 E-04	5.41 E-08	3.15 E-04	5.44 E-08	3.15 E-04	5.32 E-08
30	1.02 E-03	5.09 E-07	1.02 E-03	5.12 E-07	1.01 E-03	5.13 E-07
20	3.02 E-03	5.74 E-06	3.02 E-03	5.77 E-06	3.00 E-03	5.67 E-06
10	1.07 E-02	5.86 E-05	1.07 E-02	5.92 E-05	1.07 E-02	5.84 E-05
0	3.14 E-02	5.30 E-04	3.14 E-02	5.33 E-04	3.14 E-02	5.34 E-04
-10	9.69 E-02	5.56 E-03	9.71 E-02	5.58 E-03	9.64 E-02	5.28 E-03
-20	1.29 E+00	8.65 E-01	1.29 E+00	8.65 E-01	3.11 E-01	6.39 E-02
-30	1.59 E+00	7.72 E-01	1.59 E+00	7.72 E-01	1.02 E+00	6.78 E-01

TAB. 2.2 – Erreur d'estimation de la phase en fonction du SNR pour les trois estimateurs ϕ_{DFT}, ϕ_m et ϕ_c.

sinusoïde. Plus formellement, une composante DFT $X[k]$ est un maximum local si :

$$|X[k-1]| < |X[k]| > |X[k+1]| \tag{2.35}$$

Ce type de critère de conformité à l'avantage de n'omettre aucun candidat si la résolution fréquentielle est suffisante. Ce critère est en contrepartie très permissif. En effet, le spectre de puissance d'un bruit blanc est composé d'approximativement $N/4$ maxima locaux, N étant la taille de la DFT. Pour réduire le nombre de candidats, de nombreux critères ont été proposés. La plupart de ces critères se basent sur les propriétés du spectre d'une sinusoïde pour définir le "gabarit" du candidat idéal. La similarité du spectre du maximum local candidat et de ce gabarit permet de déterminer si ce candidat doit être conservé ou non.

Le spectre d'un signal sinusoïdal pondéré par une fenêtre correspond au spectre de cette fenêtre centré en la fréquence de la sinusoïde. Une corrélation entre le voisinage du maximum local d'indice m du spectre DFT avec le spectre de la fenêtre centré en la fréquence estimée $\hat{f}$ donne une mesure de conformité du maximum local :

$$\Gamma_s = \left| \sum_{l \in [k-B, k+B]} \frac{X[l]}{|X[k]|} \cdot W\left(\left| l\frac{F_e}{N} - \hat{f} \right| \right) \right| \tag{2.36}$$

où B est un entier, k est l'indice du maximum local. Cette méthode a été utilisée comme indice de voisement dans le domaine du codage de la voix [GL85] et en modélisation sinusoïdale comme indice de conformité [PR99]. Le gabarit est celui d'une sinusoïde dont les paramètres de fréquence et d'amplitude sont constants durant l'intervalle d'observation. En conséquence, ce critère a tendance à identifier les sinusoïdes modulées comme des maxima locaux de bruit. Ce problème sera étudié dans la section 2.4.4 dédiée à la détection de sinusoïdes dans un modèle non stationnaire.

Dans le modèle psychoacoustique du codeur MPEG Layer I et II [MPE92], un autre critère sur l'amplitude est utilisé pour déterminer la conformité du maximum local, en utilisant une contrainte de relation d'amplitude entre les composantes de la DFT au voisinage du maximum local, cf. figure 2.9(a). Formellement, un maximum local est considéré comme une sinusoïde si la relation suivante est vérifiée :

$$||X[k]| - |X[k+l]|| \geq 7\text{dB} \tag{2.37}$$

où k et l sont les indices des composantes d'une DFT de 1024 points. Pour le MPEG Layer I, l est choisi suivant les indices suivants :

$$l = \begin{cases} -2, 2 & \text{pour} \quad 2 < k < 63 \\ -3, -2, 2, 3 & \text{pour} \quad 63 \leq k < 127 \\ -6, -3, -2, 2, 3, 6 & \text{pour} \quad 127 \leq k \leq 250 \end{cases} \tag{2.38}$$

Comme on peut le constater au bas de la figure 2.9(d), le spectre de phase est constant dans le lobe principal d'un spectre à phase nulle. Il est proposé dans [KAZ01], d'utiliser cette propriété pour discriminer les maxima locaux provenant de composantes sinusoïdales des maxima locaux provenant de composantes de bruits. Soit $\overline{X_k}$, la moyenne des termes complexes des composantes proches du maximum local d'indice k définie dans l'équation 2.32. Le critère de conformité est donné par la déviation relative des termes complexes des composantes DFT proches du maximum local à leur valeur moyenne :

$$E_\phi = \sum_{l=k-1}^{k+1} \left| \frac{X[l]}{|X[l]|} - \overline{X_k} \right| \tag{2.39}$$

Si pour un maximum local donné $E_\phi > 0.8$ avec les paramètres d'analyse à court terme donnés dans [KAZ01], le maximum local est rejeté.

Lors de l'estimation de la fréquence par la méthode du réassignement ou la méthode de la dérivée, la cohérence des estimations peut être mise à profit pour écarter certains maxima locaux. En premier lieu, si on analyse un signal sinusoïdal, la composante DFT qui est de plus forte énergie est la composante qui a la fréquence la plus proche de celle de la sinusoïde analysée. Si la fréquence estimée par la méthode de la dérivée est plus proche de celle d'une autre composante DFT, cette autre composante DFT aurait dû être celle d'amplitude maximale. Ceci mène à une incohérence. Par conséquent, si la distance entre la fréquence estimée et la fréquence du maximum local est plus grande que une demie fois la précision fréquentielle de la DFT, il est proposé dans [Mar00b] de rejeter le maximum local. Formellement, le maximum local est écarté si :

$$|\hat{f} - f_k| > \frac{F_e}{2N} \tag{2.40}$$

On a vu dans la section 2.1 que la DFT peut être interprétée comme un banc de filtres. La sélectivité imparfaite des filtres passe bande fait que l'estimation de la fréquence par la méthode de la dérivée est similaire pour les composantes DFT proches d'un maximum local, voir figure 2.10. On peut définir un critère de

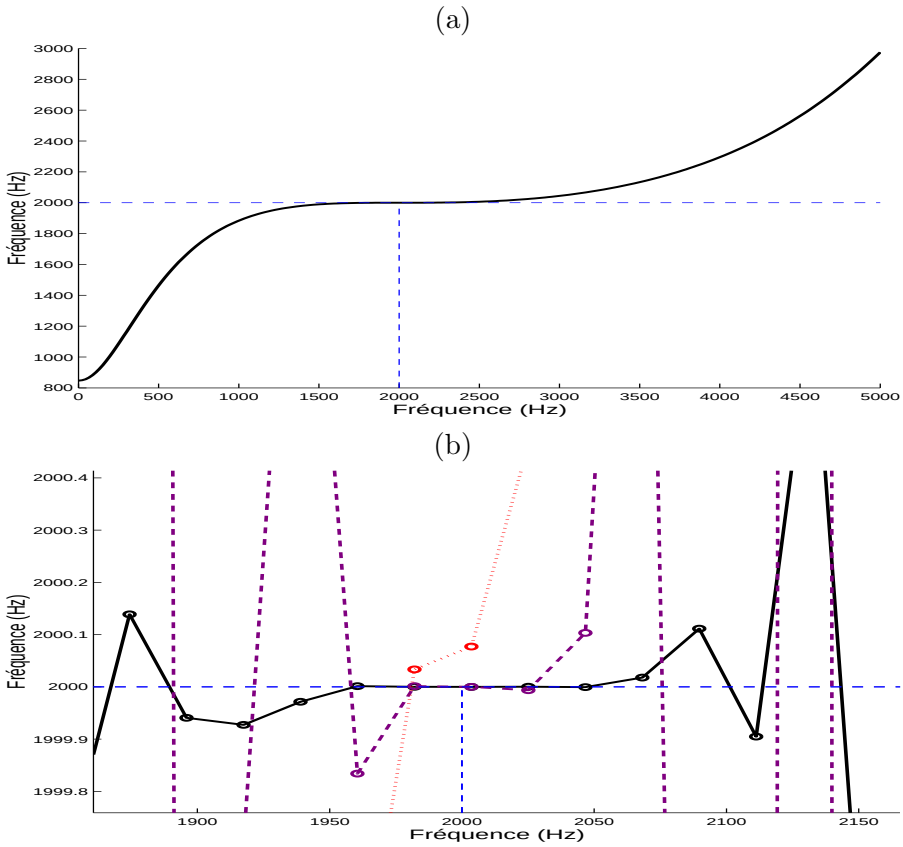


FIG. 2.10 – En haut, fréquence estimée grâce à la méthode de la dérivée en fonction de la fréquence des composantes d'une DFT de 2048 points. Le signal analysé est une sinusoïde de fréquence 2 kHz. En bas, un bruit a été ajouté de manière à obtenir un SNR de 90 dB en trait plein, de 50 dB en tirets et de 10 dB en pointillés.

conformité fonction de la déviation des fréquences estimées aux composantes DFT adjacentes $\hat{f}(l)$ par rapport à la fréquence estimée au maximum local $\hat{f}(k)$:

$$E_f = \frac{1}{3\,\hat{f}(k)} \sum_{l=k-1}^{k+1} \left| \hat{f}(l) - \hat{f}(k) \right| \tag{2.41}$$

De même, des études statistiques des valeurs de réassignement de la fréquence sont utilisées dans [HMW01] pour effectuer une discrimination sinusoïde/bruit.

2.2 Synthèse stationnaire

La synthèse de sinusoïdes dans un modèle stationnaire à court terme consiste à générer une somme de sinusoïdes dont les paramètres d'amplitude et de fréquence sont constants dans la trame de synthèse. Ces différentes trames de synthèse sont combinées entre elles comme expliqué dans la partie 1.2.1.

2.2.1 Approche temporelle

L'approche temporelle consiste à calculer les échantillons temporels pour chacune des sinusoïdes :

$$x[n] = a\,\cos(\Delta_\phi\, n + \phi) \text{ où } \Delta_\phi = \frac{2\pi f}{F_e} \tag{2.42}$$

pour $n = 0, 1, \cdots, N-1$.

Toutes les contributions des différentes sinusoïdes sont ensuite sommées pour obtenir la trame synthétisée. Comme la pulsation Δ_ϕ est constante au court du temps, un algorithme récursif de calcul de la fonction cosinus peut être mis en œuvre. Cet algorithme se base sur des propriétés trigonométriques de la fonction cosinus :

$$\cos(\alpha + \beta) = 2\cos\,\alpha\,\cos\,\beta - \cos(\alpha - \beta) \tag{2.43}$$

En posant $\alpha = \Delta_\phi(n-1)$ et $\beta = \Delta_\phi$, l'équation 2.42 peut se calculer de façon récursive par :

$$x[0] = A\,\cos(\phi) \tag{2.44}$$
$$x[1] = A\,\cos(\Delta_\phi + \phi) \tag{2.45}$$
$$\cdots$$
$$x[n] = 2\cos(\Delta_\phi)x[n-1] - x[n-2] \tag{2.46}$$

Cette opération ne nécessite qu'une addition et une multiplication flottante par échantillon. Cependant, chaque étape de la récursion est dépendante des deux précédentes, ce qui se prête mal à une parallélisation des opérations de calcul, dispositif commun aux microprocesseurs modernes. On trouvera dans [MP02] des algorithmes permettant de réduire cette dépendance, amenant à une réduction par deux des temps de calcul.

2.2.2 Approche spectrale

Le principe de l'approche spectrale est d'accumuler les contributions fréquentielles de chaque sinusoïde puis d'effectuer une IDFT pour obtenir les échantillons temporels [FRD92, FRD93]. L'efficacité de cette approche est basée sur le fait qu'une sinusoïde est très localisée en fréquence. Comme expliqué dans la section 2.1, la contribution fréquentielle d'une sinusoïde est le spectre de la

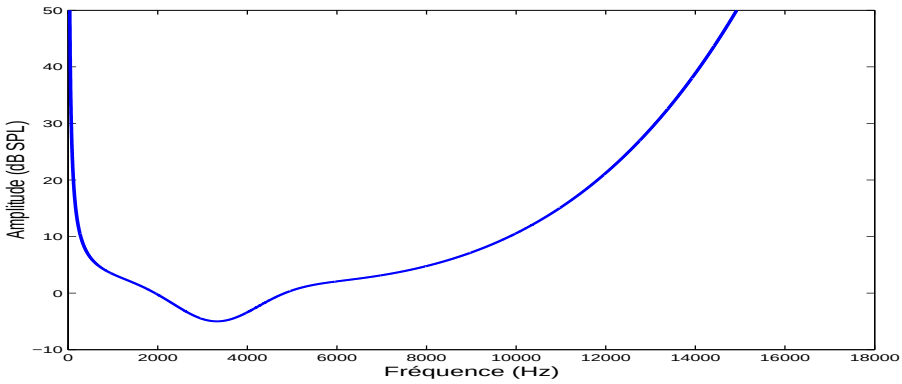


FIG. 2.11 – Seuil d'audibilité du système auditif humain.

fenêtre de synthèse centrée en la fréquence de la sinusoïde, déphasée par un facteur égal à la phase de la sinusoïde et mis à l'échelle par son amplitude. Ce spectre peut être obtenu grâce à l'équation 2.15.

Comme le calcul de chaque composante de la DFT pour chaque sinusoïde est coûteux, on procède plutôt au calcul préalable d'une table sur-échantillonnée du lobe principal à partir de la fréquence 0 jusqu'à une fréquence de troncature donnée [RD93]. De cette fréquence de troncature dépend le compromis efficacité/précision. Il est proposé dans [KAZ01] d'utiliser une fréquence de troncature égale à 0.24 radians. Comme la fréquence de la sinusoïde n'est généralement pas multiple du pas d'échantillonnage choisi pour générer la table du prototype, on utilise une interpolation linéaire pour déterminer l'amplitude et la phase des composantes de la DFT au voisinage de la fréquence de la sinusoïde que l'on veut synthétiser. Des études sur des implantations de ces deux approches sur des architectures récentes [MP02] montrent que l'approche fréquentielle est plus performante que l'approche temporelle dès que l'on a à synthétiser plus de 35 sinusoïdes par trames.

2.2.3 Réduction du nombre de sinusoïdes

Les algorithmes de synthèse décrits ci-dessus sont quasi optimaux en terme de complexité algorithmique. Pour réduire le temps nécessaire au calcul des échantillons temporels, il est nécessaire de ne synthétiser que les sinusoïdes perçues par le système auditif. En effet, certaines sinusoïdes ne sont pas perceptibles car le système auditif humain a des capacités de perception limitées. Par exemple, une sinusoïde de très faible amplitude ne sera pas perçue.

Le système auditif humain a un seuil d'audibilité qui dépend de la fréquence du signal. L'équation 2.47 donne une bonne approximation de ce seuil [PS00] en dB SPL, représenté sur la figure 2.11 :

$$\mathcal{A}(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \qquad (2.47)$$

Le principe est alors de déterminer si l'amplitude du pic considéré est inférieure

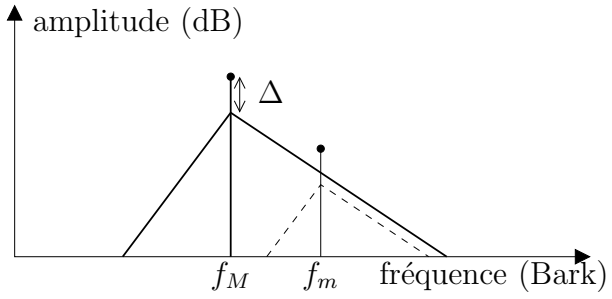


FIG. 2.12 – Masquage d'un pic de fréquence f_m par un autre pic de fréquence f_M. Le masquage est maximal quand f_m et f_M sont proches. On peut raisonnablement approximer ce masquage par un triangle dans l'échelle Bark/dB.

au seuil. Le cas échéant, la synthèse de ce pic peut être évitée. La fonction $\mathcal{A}(f)$ étant complexe à calculer, on évalue cette fonction pour un nombre limité de fréquences. La comparaison de l'amplitude du partiel avec le seuil pré-calculé se fait grâce à une interpolation linéaire.

Ensuite, chaque sinusoïde à un pouvoir masquant, c'est-à-dire qu'une deuxième sinusoïde relativement proche en fréquence de cette sinusoïde et d'amplitude sensiblement plus faible ne sera pas perçue. Ce masque peut être approximé [DGR93a] par un triangle sur une échelle Bark/dB avec les propriétés suivantes, voir figure 2.12 :

- La différence Δ entre l'amplitude du pic et le seuil du masque est de -10 dB.
- La pente de la courbe de masquage vers les basses fréquences est de 27 dB/Bark.
- La pente de la courbe de masquage vers les hautes fréquences est de -15 dB/Bark.

Cette propriété est d'un intérêt particulier pour notre problème. En effet, plus le nombre de pics devient grand, plus il y a de chance pour que de nombreux pics soient masqués et donc imperceptibles. Le nombre théorique de pics à synthétiser en fonction du nombre total de pics devient logarithmique. Le problème est donc d'être capable de décider efficacement si un pic est masqué ou non.

Comme le masque M varie au cours du temps au contraire du seuil d'audibilité, il doit être recalculé à chaque trame en ajoutant la contribution de chaque pic au masque. De manière à réduire autant que possible les temps de mises à jour du masque, les pics sont d'abord triés par ordre d'amplitude décroissante. Grâce à ce tri, la contribution du pic P_3 sur la figure 2.13 n'a pas à être ajoutée au masque M, car la contribution de P_2 (pic qui masque P_1) a déjà été ajoutée.

Deux représentations différentes du masque ont été testées dans [Lag01]. La première reprend le principe d'un échantillonnage discret du masque M utilisé pour le seuil d'audibilité. Dans les applications classiques de type analyse/synthèse, le nombre maximal de pics est borné par la résolution fréquentielle de l'analyse spectrale (on suppose ici qu'un unique signal est analysé, puis

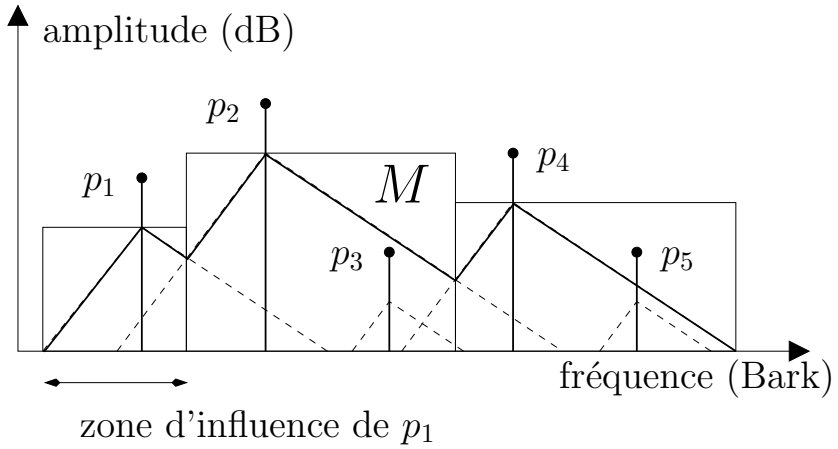


FIG. 2.13 – Cinq pics et le masque associé M. p_1, p_2, et p_4 sont des pics masquants et contribuent à M. Les zones d'influences sont représentées par des rectangles. p_5 n'est ni masqué ni masquant et p_3 est masqué.

synthétisé sans mixage préalable avec un autre signal). Dans le cas des estimateurs de Fourier présentés dans ce chapitre, le nombre de pics est au plus de $N/4$, N étant le nombre d'échantillons utilisés pour calculer la transformée de Fourier. On peut donc considérer que le nombre de pics moyen est de l'ordre d'une centaine, ce qui rend acceptable le temps nécessaire au calcul d'un masque échantillonné.

En revanche, pour le rendu temps/réel de scènes polyphoniques, impliquant plusieurs centaines voire plusieurs milliers d'oscillateurs, il est utile de disposer d'une structure de donnée plus efficace [LM01]. Le principe est de stocker pour chaque pic masquant sa zone d'influence (bande de fréquence où son masque est supérieur à tous les autres) par indice de fréquence croissant, voir figure 2.13. Pour chaque pic, l'algorithme de mise à jour du masque est le suivant :

- on recherche les voisins immédiats (pics de fréquence immédiatement inférieure et supérieure) ;
- si la fréquence du pic est comprise dans la zone d'influence du voisin gauche et son amplitude est inférieure au masque, le pic n'est pas synthétisé ;
- si la fréquence est comprise dans la zone d'influence du voisin droit et son amplitude inférieure au masque, le pic n'est pas synthétisé ;
- sinon, le pic est synthétisé et si le pic contribue au masque, on insère sa zone d'influence dans la structure de donnée et on met à jour les zones d'influences des pics voisins.

On remarque que la mise à jour du masque et la détection masqué/non masqué sont effectuées dans le même temps.

La structure de données doit donc être efficace en insertion d'éléments. On peut avoir recours à une structure en arbre binaire qui permet une recherche très rapide. Mais du fait de sa structure très rigide à maintenir, les insertions sont laborieuses.

On propose d'utiliser une Skip List [Pug90]. Cette structure de donnée est une liste simplement chaînée qui peut donc être manipulée comme telle, mais qui possède aussi une arborescence de pointeurs qui permet de "passer par dessus" – *skip* – des nœuds non pertinents.

La recherche du pic masquant numéro 6 dans une liste triée par ordre croissant est représentée sur la figure 2.14. On procède comme suit : le tableau de pointeur de la tête de liste (premier élément à gauche) est parcouru en premier, le parcours d'un tableau de pointeurs se faisant de haut en bas. Le premier pointeur du tableau associé à la tête de liste pointe sur le marqueur de fin (NIL sur la figure 2.14), on étudie alors le pointeur immédiatement inférieur. Le deuxième pointeur de ce même tableau pointe sur le pic numéro 2 qui est situé dans la liste avant le pic numéro 6. La recherche se poursuit alors à partir du tableau de pointeurs associé au pic numéro 2. Le même processus est répété jusqu'à trouver le pic recherché ou évaluer sans succès le pointeur le plus bas d'un tableau. Dans ce cas, l'élément n'est pas présent dans la liste. Dans cet exemple, le nombre de pointeurs évalués est réduit d'un facteur 1/3 par rapport au parcours d'une liste simplement chaînée.

Comme on peut le constater sur la figure 2.15, l'insertion dans une *skip-list* ne nécessite que des mises à jour de pointeurs existants. Pour gagner en complexité, la hauteur du tableau de pointeurs d'un élément à ajouter est tirée au hasard entre 0 et une taille maximale égale à la taille du tableau de pointeurs de la tête de liste et du marqueur de fin. Grâce à ce tirage aléatoire de la hauteur des tableaux de pointeurs, la probabilité de faire un parcours de liste complet lors de la recherche du plus grand élément décroît exponentiellement quand on augmente le nombre de nœuds. On obtient ainsi une structure souple, rapide en insertion et en consultation qui représente le masque sans perte de précision.

Des tests comparatifs de performances de ces deux représentations (présentés sur la figure 2.16) ont montré la supériorité de la représentation par *skip-list* mais aussi le coût prohibitif du tri en amplitude implémenté pour ces tests par un "*quick-sort*". Or la répartition des partiels variant peu au court du temps. Il serait judicieux d'étudier, à titre de perspective, une réduction de la fréquence des tris des partiels.

L'approche par *skip-list* a été intégrée dans un module de synthèse par oscillateurs. Les tests utilisant cinq sources (toutes clairement audibles) ont montré que le coût de calcul est pratiquement divisé par deux par rapport à la synthèse de l'intégralité des sinusoïdes composant les cinq sourcs.

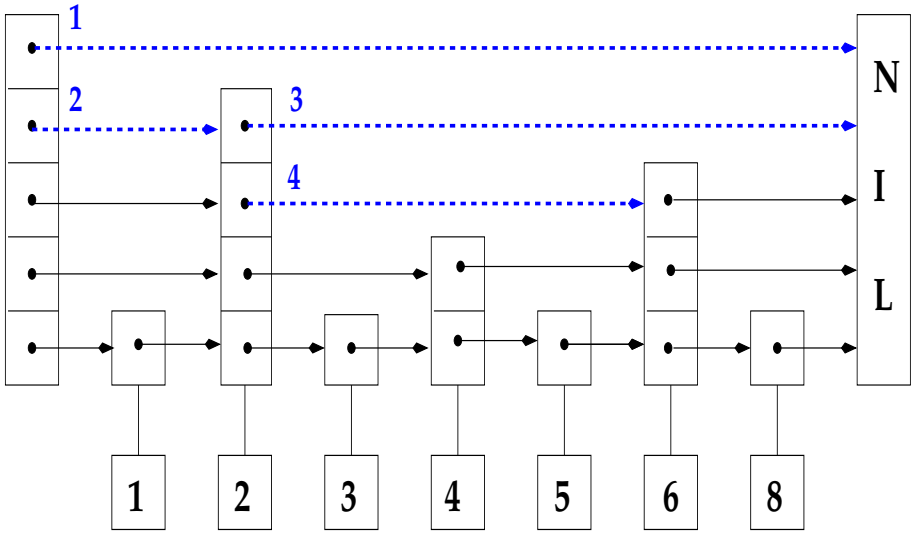


FIG. 2.14 – Recherche de l'élément 6 dans une skip-list où seuls les pointeurs en tirets sont utilisés.

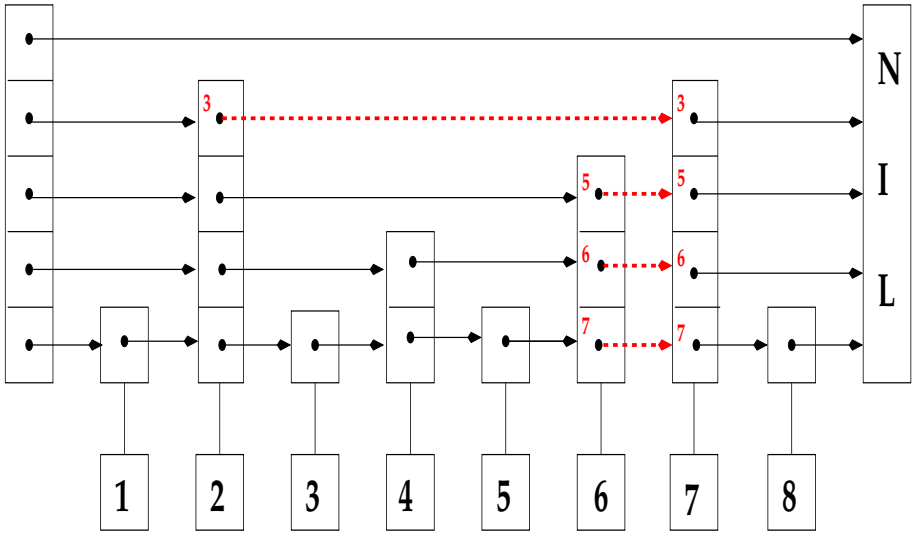


FIG. 2.15 – Insertion de l'élément 7 dans une skip-list. Seuls les pointeurs en tirets sont mis à jour.

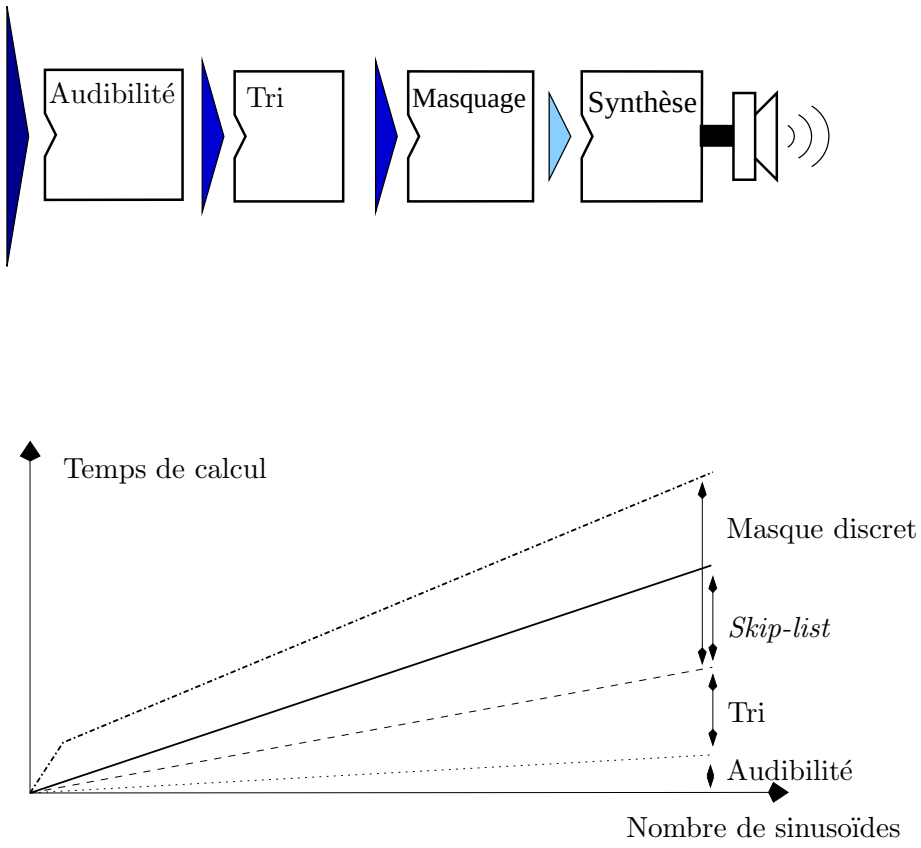


FIG. 2.16 – En haut, est schématisée l'architecture logicielle du module de synthèse. La taille des triangles verticaux symbolise le nombre de pics à synthétiser. Dans un premier temps, les pics d'amplitude trop faible sont écartés. Les pics restants sont ensuite triés par amplitude décroissante. La détection masqué/non masqué permet de réduire encore le nombre de sinusoïdes. En bas, le temps de calcul des différents composants de cette architecture est représenté en fonction du nombre de sinusoïdes.

2.3 Modélisation non stationnaire

Les signaux naturels ne sont pas parfaitement stationnaires à court terme. L'application du modèle stationnaire de l'équation 1.4 fait que les paramètres d'amplitude et de fréquence des composantes sinusoïdales sont les moyennes des évolutions de ces paramètres dans la trame d'analyse. L'utilisation de ce modèle requiert donc une bonne précision temporelle car le nombre d'échantillons nécessaires à l'estimation spectrale doit être le plus faible possible pour que les paramètres puissent être raisonnablement considérés comme constants.

Or, l'analyse spectrale de polyphonies implique une bonne résolution fréquentielle, le nombre d'échantillons doit donc être grand. Il est donc pertinent de considérer un modèle non stationnaire d'ordre 1 où la fréquence ou l'amplitude varient linéairement/exponentiellement dans l'intervalle considéré. Si les paramètres non stationnaires sont correctement estimés, l'approximation effectuée par ce nouveau modèle sera plus faible.

Comme on le verra dans le chapitre 3, la plupart des méthodes de suivi de partiels se basent sur la proximité des fréquences des pics de trames adjacentes qui permet d'assurer la continuité à l'ordre 1. Les informations d'ordre 1 peuvent aussi être utiles pour lier les pics de trames adjacentes entre eux pour former les partiels. Nous nous intéressons aussi à la modélisation non stationnaire pour affiner le critère utilisé pour former les partiels en ajoutant une contrainte sur la continuité à l'ordre supérieur.

Dans ce modèle, on représente le signal analysé $s(t)$ comme suit :

$$s(t) = \sum_{p=1}^{P} a(t)\ \cos(\phi_p^{ns}(t)) \tag{2.48}$$

$$\phi_p^{ns}(t) = \phi_p(0) + 2\pi \int_0^t \left(f_p(0) + \Delta_{f_p}\, u\right)\, du \tag{2.49}$$

$$= \phi_p(0) + 2\pi \left(f_p(0)\, t + \frac{\Delta_{fp}}{2}\, t^2\right) \tag{2.50}$$

$$a(t) = a_p(0) 10^{\frac{\Delta_{a_p}\, t}{20}} \tag{2.51}$$

où $a_p(0)$, $f_p(0)$ et $\phi_p(0)$ désignent respectivement l'amplitude, la fréquence et la phase à l'origine. Les paramètres de variation exponentielle d'amplitude Δ_{ap} et linéaire de fréquence Δ_{f_p} sont constants au cours de la trame d'analyse. Par mesure de clarté, l'indice p est omis dans la suite.

2.4 Analyse non stationnaire

L'utilisation d'une fenêtre de pondération avec de bonnes propriétés mathématiques comme la fenêtre gaussienne permet l'expression analytique de Δ_f en fonction du spectre du maximum local. L'utilisation d'autres fenêtres plus adaptées à l'estimation des paramètres sinusoïdaux telle que la fenêtre de Hann est problématique car on ne dispose pas de formule analytique de la transformée de Fourier d'une sinusoïde dont la fréquence varie linéairement :

$$S_{f_0}^{ns}(f) = \sum_{n=-N/2+1}^{N/2-1} \frac{(1-\cos(2\pi n/N))}{2} \cos(\phi^{ns}(n)) \cdot e^{\frac{-2j\pi fn}{N}} \tag{2.52}$$

où $\phi^{ns}(n)$ désigne la phase d'une sinusoïde modulée de fréquence f_0 calculée grâce à l'équation 2.50.

On étudie donc des observations empiriques ou des approximations en des conditions particulières proposées dans la littérature pour estimer les paramètres non stationnaires. On montre ensuite comment, grâce à la connaissance des paramètres non stationnaires, on peut corriger le biais en amplitude et en phase introduits par des modulations non stationnaires. Dans une deuxième section, le critère de conformité au modèle sinusoïdal de l'équation 2.36 est étendu au modèle non stationnaire et permet ainsi une meilleure distinction des maxima locaux issus du bruit de ceux issus de sinusoïdes modulées.

2.4.1 Estimation des paramètres non stationnaires

La présente section est dédiée aux estimateurs de paramètres non stationnaires se basant sur le spectre de Fourier. D'autres méthodes utilisant des transformées temps/fréquence ou des méthodes de minimisation explicitement dédiées existent [NHD98, BAM02], et sont souvent de complexité plus élevée.

Estimateur de Marques et Almeida

Marques et Almeida utilisent les propriétés mathématiques de la fenêtre gaussienne pour exprimer analytiquement le spectre d'une sinusoïde modulée linéairement en fréquence [MA86]. Soit $s_g(n)$ une sinusoïde modulée en fréquence (voir équation 2.48) fenêtrée par une fenêtre gaussienne de variance σ :

$$w_g(t) = e^{-\frac{t^2}{2\sigma^2}} \tag{2.53}$$

En posant :

$$\omega_0 = 2\pi \frac{f(0)}{F_e} \tag{2.54}$$

$$\Delta = \frac{\Delta_f}{2} \tag{2.55}$$

Le spectre de $s_g(n)$ s'exprime analytiquement :

$$S_g(\omega) = a\, e^{i\phi_0}\, B(\omega) e^{jC(\omega)} \tag{2.56}$$

$$B(\omega) = \sqrt{\frac{1+2j\Delta\,\sigma^2}{1+4\Delta^2\sigma^4}}\, e^{-\frac{\sigma^2}{2}\frac{(\omega-\omega_0)^2}{1+4\Delta_f\sigma^4}} \tag{2.57}$$

$$C(\omega) = \frac{\Delta\,\sigma^4}{1+4\Delta^2\sigma^4}(\omega-\omega_0)^2 \tag{2.58}$$

Le lobe principal du spectre de fréquence d'une sinusoïde fenêtrée par une fenêtre gaussienne est une parabole sur l'échelle des logarithmes d'amplitude. Si

$$\log S_g(\omega) = \alpha\omega^2 + \beta\omega + \zeta \tag{2.59}$$

alors

$$\omega_0 = -\frac{\beta}{\alpha} \tag{2.60}$$

$$|\Delta| = \pm\frac{1}{2\sigma^2}\sqrt{\frac{-\sigma^2}{2\alpha} - 1} \tag{2.61}$$

Le signe de Δ est déterminé par la concavité ou la convexité du spectre de phase autour de ω_0, voir figure 2.18. Cet estimateur est malheureusement réservé à la fenêtre gaussienne qui possède de mauvaises propriétés de précision spectrale, ce qui constitue un handicap sérieux lors de l'analyse de signaux polyphoniques.

Estimateurs de Masri

On a exposé dans la section 2.1 que la transformée de Fourier effectuait une moyenne de l'évolution des paramètres de fréquence et d'amplitude de la sinusoïde analysée sur l'intervalle d'observation. La transformée de Fourier ne paraît donc pas être un bon candidat pour l'analyse des variations de ces paramètres. Il n'en reste pas moins que cette transformée est inversible et à reconstruction parfaite. Comme le fait remarquer Masri [Mas96], les non stationnarités ne sont donc pas représentées explicitement mais représentées sous forme de distorsions.

On a vu qu'une sinusoïde dont les paramètres de fréquence et d'amplitude sont constants pendant l'intervalle d'observation a un spectre de phase constant dans le lobe principal pourvu que le spectre soit à phase nulle, voir figure 2.9(d). Considérons une sinusoïde dont l'amplitude augmente de 5 dB ou décroît de 5 dB dans l'intervalle d'observation. On constate sur la figure 2.17 que les distorsions du spectre de puissance sont peu perceptibles. Au contraire, la distorsion du spectre de phase est très informative. La distorsion est antisymétrique autour de l'indice DFT du maximum local et l'orientation de la pente est dépendante du signe de la modulation d'amplitude. Si l'amplitude croît, la phase décroît avec la fréquence et inversement.

Considérons maintenant une sinusoïde dont la fréquence croît de 40 Hz ou décroît de 40 Hz dans l'intervalle d'observation. On constate que ces modulations entraînent des distorsions du spectre de puissance et de phase, voir la figure 2.18. La bande de fréquence dans laquelle évolue la sinusoïde étant plus

large, le spectre de puissance est étalé. La distorsion du spectre de phase est symétrique autour de l'indice DFT du maximum local et l'orientation dépend du signe de la modulation de fréquence.

Masri utilise l'analyse de la distorsion de phase pour estimer les paramètres de variation d'amplitude et de fréquence [Mas96]. Il est proposé de considérer la différence $(\angle S[k+1] - \angle S[k-1])$ et la somme $(\angle S[k+1] + \angle S[k-1])$ des phases adjacentes au maxima local d'indice k dans un spectre à phase nulle pour estimer les modulations d'amplitude et de fréquence. Considérons l'évolution de la phase dans le lobe principal, voir figures 2.17 et 2.18. La dérivée première de la phase en fonction de la fréquence dépend de la modulation d'amplitude et la dérivée seconde de la phase en fonction de la fréquence dépend de la modulation de fréquence. On utilisera donc dans la suite Φ' et Φ'' comme estimateurs des modulations d'amplitude et de fréquence :

$$\Phi' = \angle S[k+1] - \angle S[k-1] \tag{2.62}$$

$$\Phi'' = -2\angle S[k] + \angle S[k+1] + \angle S[k-1] \tag{2.63}$$

De manière à obtenir des mesures $\angle S[k+1]$ et $\angle S[k-1]$ les plus proches possibles du maximum local, Masri propose d'utiliser la méthode du *zero-padding*. La figure 2.19 montre l'évolution de Φ' et Φ'' en fonction de Δ_a et de Δ_f.

Une relation quasi linéaire entre Δ_a et Φ' est constatée :

$$\Delta_a \simeq \alpha\, \Phi' \tag{2.64}$$

où α est une constante qui dépend du type de la fenêtre d'analyse, du nombre d'échantillons et du facteur de *zero-padding*. La relation entre Δ_f et Φ'' est plus complexe et non inversible car la courbe de Φ'' en fonction Δ_f possède un point d'inflexion pour $\Delta_f = 6$. On se restreint donc à des modulations de fréquences comprises entre 0 et $6\frac{F_e}{N}$ Hz par trame et on approxime la partie choisie par un polynôme d'ordre 3. Comme montré dans [Mas96], la forme générale des courbes affichées en haut de la figure 2.19 ne dépend pas des paramètres de l'analyse, du moins dans le cas des fenêtres trigonométriques. En revanche, les valeurs des coefficients des polynômes utilisés dépendent de la taille de la fenêtre, du type de fenêtre utilisé et du facteur de *zero-padding*.

D'autre part, l'indépendance entre les deux estimateurs n'est pas parfaite. Une modulation de fréquence influence Φ' et inversement une modulation d'amplitude influence Φ'' comme on peut le constater dans les figures 2.19(c) et 2.19(d). De manière à augmenter la robustesse de ces estimateurs, il est proposé dans [MC98] de considérer plusieurs points d'observations pour estimer les dérivées première et seconde de l'évolution de la phase.

Estimateurs de Master

Master propose d'estimer la variation linéaire de la fréquence d'une sinusoïde de fréquence et de phase nulles, modulée linéairement en fréquence et fenêtrée par la fenêtre de Hann. Pour de grandes valeurs de Δ_f, il est proposé dans [ML03b] d'utiliser le formalisme des intégrales de Fresnel pour obtenir une expression approchée du spectre de puissance et de phase [ML03a] autour du

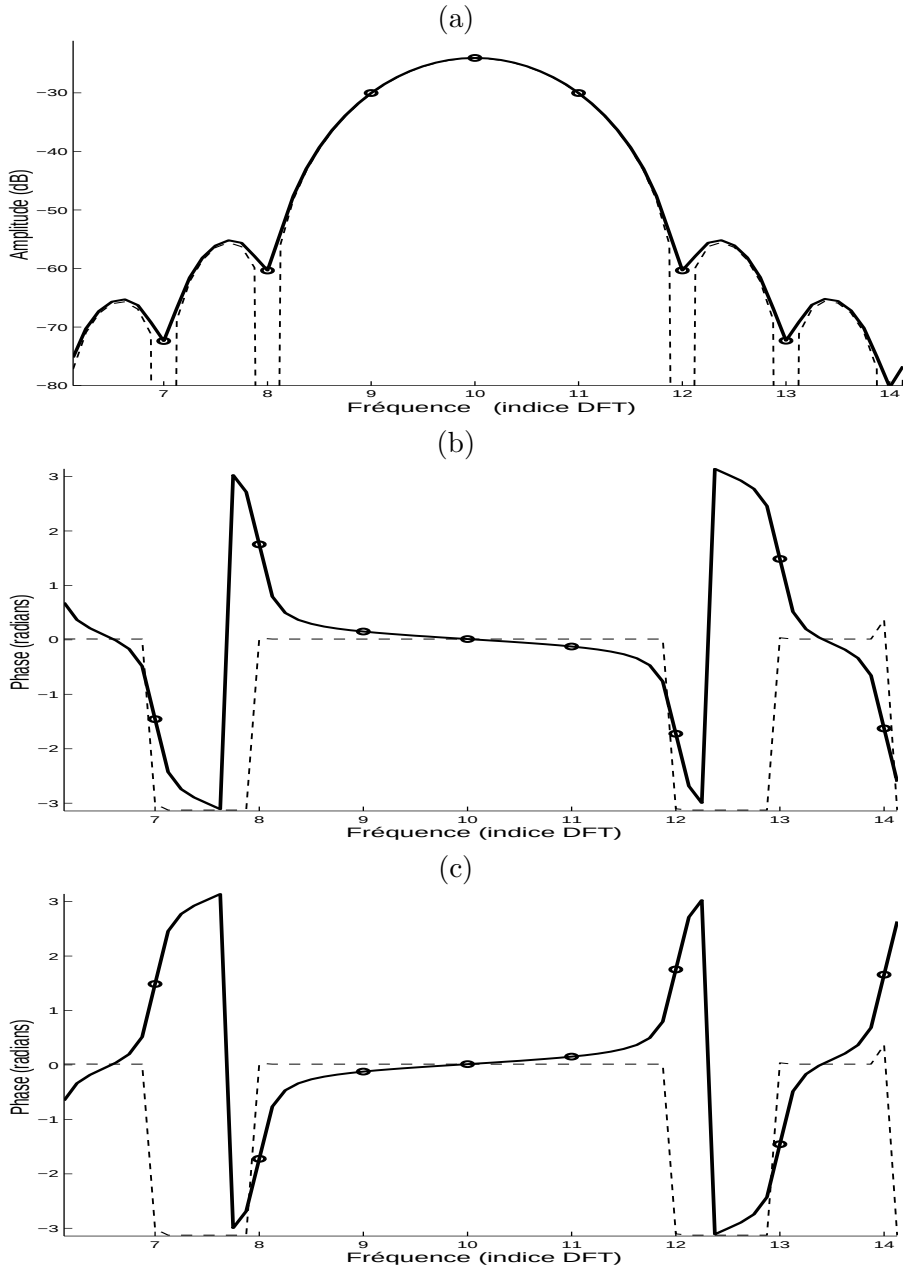


FIG. 2.17 – En haut, est représenté le spectre de puissance d'une sinusoïde avec une modulation exponentielle d'amplitude de plus ou moins 5 dB. Le trait en pointillés représente le spectre d'une sinusoïde non modulée et sert ainsi de référence. Le trait plein représente le spectre calculé avec un facteur de *zero-padding* de 8. Les points correspondent aux composantes d'une DFT de 2048 points calculée sans *zero-padding*. Au milieu, est représenté le spectre de phase d'une sinusoïde avec une modulation exponentielle d'amplitude de 5 dB. En bas, est représenté le spectre de phase d'une sinusoïde avec une modulation exponentielle d'amplitude de −5 dB.

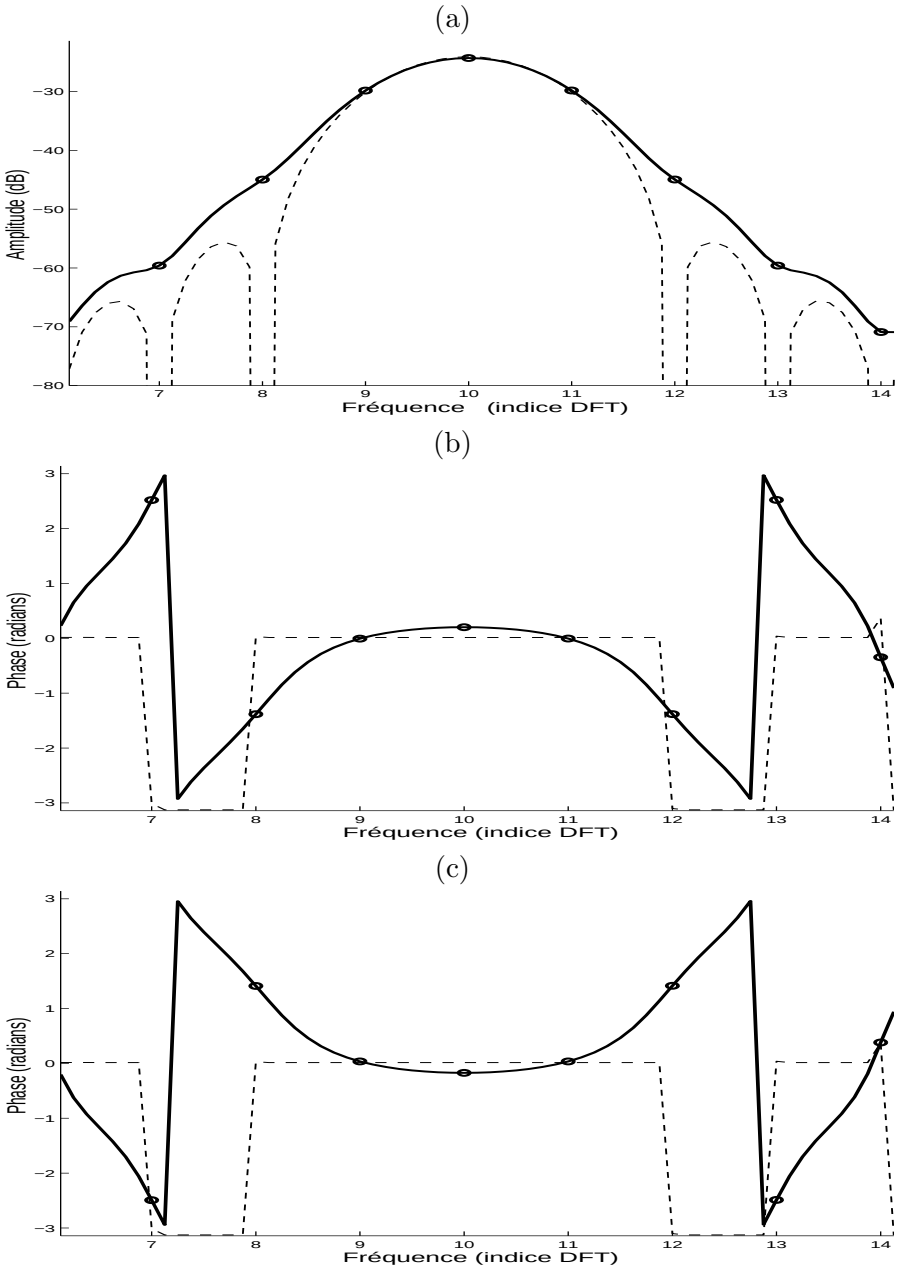


FIG. 2.18 – En haut, est représenté le spectre de puissance d'une sinusoïde avec une modulation linéaire de fréquence de plus ou moins 40 Hz. Le trait en pointillés représente le spectre d'une sinusoïde non modulée et sert ainsi de référence. Le trait plein représente le spectre calculé avec un facteur de *zero-padding* de 8. Les points correspondent aux composantes d'une DFT de 2048 points calculée sans *zero-padding*. Au milieu, est représenté le spectre de phase d'une sinusoïde avec une modulation linéaire de fréquence de 40 Hz. En bas, est représenté le spectre de phase d'une sinusoïde avec une modulation linéaire de fréquence de -40 Hz.

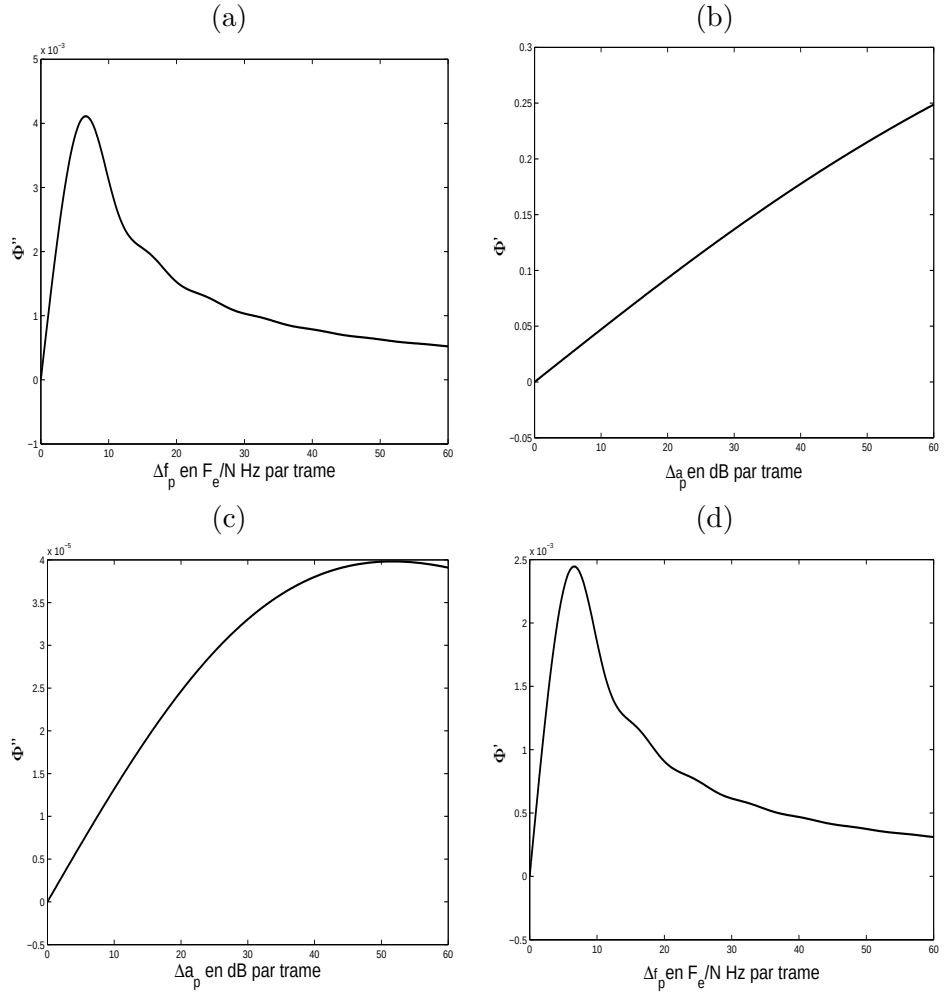


FIG. 2.19 – Les figures (a) et (d) représentent respectivement l'évolution de Φ' et de Φ'' en fonction d'une modulation linéaire de fréquence. Les figures (b) et (c) représentent respectivement les évolutions de Φ' et Φ'' en fonction d'une modulation exponentielle d'amplitude. Le signal test est échantillonné à $F_e = 44100$ Hz et analysé avec une DFT de taille $N = 2048$ points.

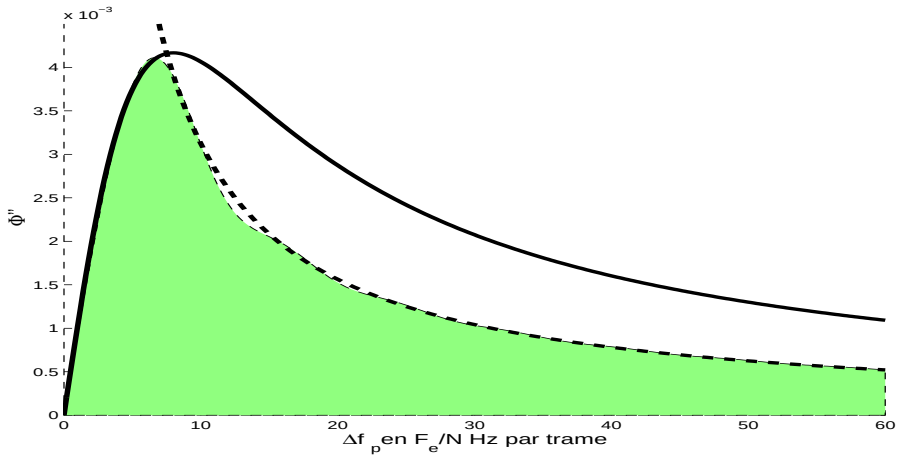


FIG. 2.20 – Approximation du degré de courbure de la phase par l'approche des intégrales de Fresnel (trait en pointillés) et de l'approximation des intégrales de Taylor (trait plein) en fonction d'une modulation linéaire de fréquence.

maximum local. On voit sur la figure 2.20(b) que ce type d'estimateur est limité à des Δ_f supérieurs à $6\frac{F_e}{N}$ Hz par trame. Dans le cas où Δ_f est proche de 0, il est proposé d'utiliser une approximation en séries de Taylor d'un signal exponentiel avec un argument faible. Comme l'estimateur proposé par Masri, ces deux estimateurs utilisent la courbure de la phase du lobe principal du spectre à phase nulle.

Estimateurs de Röbel

Röbel propose d'utiliser le formalisme du réassignement pour estimer la variation linéaire de la fréquence. Le calcul de la fréquence corrigée par la méthode du réassignement nécessite le spectre du signal fenêtré et le spectre du signal fenêtré par la dérivée de la fenêtre, voir équation 2.17. Il est proposé dans [Röb02] de considérer en plus le spectre du signal fenêtré par la dérivée seconde pour estimer Δ_f.

2.4.2 Estimation de l'amplitude

Comme on peut le voir sur les figures 2.17(a) et 2.18(a), les modulations de fréquence et d'amplitude modifient le spectre de puissance. On propose dans [LMR02] de corriger le biais introduit par des modulations d'ordre 1. On considère non plus le spectre de la fenêtre d'analyse comme détaillé dans la section 2.1, mais le spectre DFT d'un signal sinusoïdal modulé :

$$\hat{a} = 2\,\frac{|X[k]|}{S^{ns}_{\hat{f}}\left(\frac{k\,F_e}{N}\right)} \tag{2.65}$$

où k désigne l'indice du maximum local et $S^{ns}_{\hat{f}}(f_k)$ est calculé grâce à l'équation 2.52 avec des paramètres non stationnaires préalablement estimés.

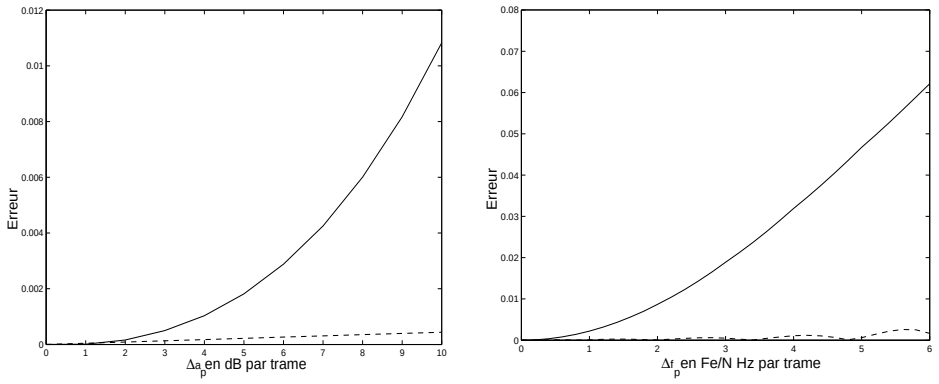


FIG. 2.21 – Erreur d'estimation de l'amplitude en utilisant la correction par le calcul du spectre de la fenêtre de pondération (en trait plein) et la méthode de correction par le calcul du spectre d'une sinusoïde modulée (en tiret). À gauche, on introduit une modulation exponentielle de l'amplitude et à droite, une modulation linéaire de fréquence.

La corrélation du signal avec un signal exponentiel modulé et fenêtré par la fenêtre de Hann est une opération du même ordre de complexité que dans le cas de l'équation 2.26. Cette opération permet d'estimer non seulement l'amplitude mais aussi la phase. Si le spectre modulé a une formule analytique, comme dans le cas de la fenêtre gaussienne [MA86], l'amplitude et la phase sont estimées par une méthode des moindres carrés moins coûteuse en temps de calcul.

2.4.3 Estimation de la phase

Comme on peut le constater en bas de la figure 2.18, une modulation de fréquence introduit un biais au niveau de l'estimation de la phase à l'indice du maximum local. Une étude empirique montre que ce biais est fonction du Δ_f mais aussi du nombre d'échantillons N utilisé pour le calcul de la DFT. Si l'on dispose d'une estimation de Δ_f, on peut corriger ce biais :

$$\hat{\Phi}_{\xi_N} = \angle S(\hat{f}) - \xi_N \frac{\Delta_f N}{F_e} \tag{2.66}$$

où ξ_N dépend de N, voir table 2.3.

Comme on peut le voir sur la figure 2.18(c), les phases d'un spectre à phase nulle situées en $\hat{f} \pm \frac{F_e}{N}$ suggèrent un estimateur robuste aux modulations linéaires de fréquence. De manière à être autant que possible robuste aux modulations d'amplitude qui amènent une distorsion antisymétrique de la phase, on considère la demie somme de ces deux valeurs de phases. Formellement :

$$\hat{\Phi}_\tau = \frac{1}{2}\angle S\left(\hat{f} + \frac{F_e}{N}\right) + \frac{1}{2}\angle S\left(\hat{f} - \frac{F_e}{N}\right) \tag{2.67}$$

Les phases $\angle S(\hat{f} \pm \frac{F_e}{N})$ peuvent être obtenues soit par calcul des transformées de Fourier en un point $S(\hat{f} \pm \frac{F_e}{N})$ soit approchées en utilisant une estimation

N	ξ_N
8	0.10659681
16	0.10473404
32	0.10472224
64	0.10472145
128	0.10471888
256	0.10470863
512	0.10466765
1024	0.10450403
2048	0.10385471
4096	0.10133763
8192	0.09242542

TAB. 2.3 – Valeurs du correctif ξ_N en fonction du nombre d'échantillons N utilisés pour le calcul de la DFT du signal.

du spectre avec un facteur de *zero-padding* suffisant (dans les simulations qui suivent ce facteur est égal à 8). Le principal avantage de cet estimateur est de ne pas être dépendant d'une estimation préalable de Δ_f.

De manière à apprécier les performances de ces deux estimateurs, on procède à l'expérience suivante : une sinusoïde de fréquence arbitraire est modulée soit en fréquence soit en amplitude. La table 2.4 présente la moyenne de la distance de ces estimateurs à la phase théorique dans le cas d'une modulation linéaire de fréquence. Les maxima d'erreurs sont approximativement égaux à deux fois l'erreur moyenne et la variance est faible. Ces mesures ne sont donc pas considérés dans la table.

Comme on peut le constater avec les résultats de la colonne 1, ϕ_{DFT} est un mauvais estimateur à cause du biais introduit par la modulation de fréquence. L'estimateur $\phi_{\mathrm{DFT}+\xi_N}$ qui compense le biais grâce à la valeur exacte du Δ_f est 100 fois plus précis que ϕ_{DFT}, comme le montre les résultats de la colonne 2. Sans connaissance du Δ_f, l'estimateur ϕ_τ donne des résultats 10 fois plus précis que ϕ_{DFT}, voir colonne 4. Pour chacune de ces deux méthodes, l'utilisation de transformées de Fourier en un point pour estimer les valeurs de phase apporte un faible gain de performance, voir colonnes 3 et 5.

La table 2.5 présente la moyenne de la distance de ces estimateurs à la phase théorique dans le cas d'une modulation exponentielle de l'amplitude. De même, les maxima d'erreurs et la variance ne sont pas considérés dans la table. La fréquence étant constante, les résultats de ϕ_{DFT} et de $\phi_{\mathrm{DFT}+\xi_N}$ sont égaux, voir colonnes 1 et 2. Une modulation d'amplitude introduit une distorsion antisymétrique de la phase du lobe principal, voir en bas de la figure 2.17. Si la fréquence de la sinusoïde n'est pas multiple de $\frac{F_e}{N}$, l'estimation de la phase est biaisée car le spectre de phase n'est plus constant dans le lobe principal. Le calcul du spectre de Fourier en la fréquence $\hat{f}$ pour estimer la phase en la fréquence de la sinusoïde amène donc de bien meilleurs résultats, comme le montre les résultats de la colonne 3 de la table 2.5. Pour la même raison que citée précédemment, l'indépendance de l'estimateur ϕ_τ vis-à-vis de

Δ_f	ϕ_{DFT}	$\phi_{\text{DFT}+\xi_N}$	$\phi_{c+\xi_N}$	ϕ_τ	$\phi_{c\tau}$
0	2.9131 E-17	2.9131 E-17	3.0310 E-18	3.4529 E-17	4.7922 E-17
5	2.3800 E-02	3.1532 E-04	2.8735 E-04	3.9611 E-03	3.8602 E-03
10	4.7561 E-02	6.6867 E-04	6.1280 E-04	7.7848 E-03	7.5857 E-03
15	7.1247 E-02	1.0981 E-03	1.0144 E-03	1.1337 E-02	1.1045 E-02
20	9.4819 E-02	1.6414 E-03	1.5302 E-03	1.4491 E-02	1.4113 E-02
25	1.1824 E-01	2.3366 E-03	2.1980 E-03	1.7130 E-02	1.6675 E-02
30	1.4147 E-01	3.2212 E-03	3.0556 E-03	1.9147 E-02	1.8626 E-02
35	1.6447 E-01	4.3328 E-03	4.1406 E-03	2.0453 E-02	1.9877 E-02
40	1.8721 E-01	5.7086 E-03	5.4902 E-03	2.0972 E-02	2.0351 E-02

TAB. 2.4 – Moyenne de l'erreur de cinq estimateurs de phase en fonction de modulations linéaire de fréquence. L'estimateur ϕ_{DFT} considère la phase du maximum local, $\phi_{\text{DFT}+\xi_N}$ considère la phase du maximum local et la corrige grâce à Δ_f. L'estimateur $\phi_{c+\xi_N}$ considère la phase du spectre de Fourier en la fréquence estimée et la corrige grâce à Δ_f. Les estimateurs ϕ_τ et $\phi_{c\tau}$ considèrent la demie somme entre les valeurs de phase à gauche et à droite du maximum local approchées par une DFT avec un facteur de *zero-padding* de 8 ou calculées en les fréquences $S(\hat{f} \pm \frac{F_e}{N})$. Sans connaissance de Δ_f, ϕ_τ est 10 fois plus précis que ϕ_{DFT}. Avec connaissance de Δ_f, $\phi_{\text{DFT}+\xi_N}$ est 100 fois plus précis que ϕ_{DFT}.

la modulation d'amplitude n'est pas parfaite, comme on peut le constater sur la colonne 4 de la table 2.5. Le calcul de deux transformées de Fourier en les fréquences $\hat{f} \pm \frac{F_e}{N}$ pour estimer les deux phases nécessaires à l'équation 2.67 permet une bien meilleure indépendance aux modulations d'amplitude, comme le montre les résultats de la colonne 5 de la table 2.5.

2.4.4 Sélection de pics par conformité au modèle

De manière à pouvoir distinguer les pics de bruit des sinusoïdes dont les paramètres sont modulés, la détection de sinusoïdes basée sur la forme du spectre de puissance requiert une estimation de Δ_f et Δ_a. En effet, ces modulations amènent des distorsions du spectre de puissance, comme on peut le constater sur la figure 2.18(a). Le critère de conformité valide pour des signaux localement stationnaires (voir équation 2.36) n'est alors plus approprié.

Si l'on dispose d'une estimation de Δ_f et Δ_a, la corrélation peut se faire non plus avec le spectre de puissance de la fenêtre mais avec le spectre d'une sinusoïde ayant ces paramètres non stationnaires. Sauf dans le cas où la fenêtre d'analyse est une gaussienne, le calcul de ce spectre se fait par corrélation grâce à l'équation 2.26. Le critère de conformité non stationnaire est donc :

$$\Gamma_{ns} = \left| \sum_{l \in [k-B, k+B]} \frac{X[l]}{|X[k]|} \cdot |S^{ns}_{\hat{f}}(f_l)| \right| \tag{2.68}$$

où B est un entier positif, k est l'indice DFT du maximum local et $S^{ns}_{\hat{f}}(f_l)$ désigne le spectre à la fréquence $f_l = l\frac{F_e}{N}$ d'une sinusoïde d'amplitude 1, de

Δ_a	ϕ_{DFT}	$\phi_{d+\xi_N}$	$\phi_{c+\xi_N}$	ϕ_τ	$\phi_{c\tau}$
0	2.9131 E-17	2.9131 E-17	3.0310 E-18	3.4529 E-17	4.7922 E-17
2	1.4676 E-03	1.4676 E-03	3.1789 E-07	2.3215 E-03	6.3697 E-07
4	2.9337 E-03	2.9337 E-03	6.3833 E-07	4.6330 E-03	1.2713 E-06
6	4.3968 E-03	4.3968 E-03	9.6387 E-07	6.9244 E-03	1.9004 E-06
8	5.8553 E-03	5.8553 E-03	1.2971 E-06	9.1866 E-03	2.5217 E-06
10	7.3080 E-03	7.3080 E-03	1.6405 E-06	1.1411 E-02	3.1326 E-06
12	8.7532 E-03	8.7532 E-03	1.9966 E-06	1.3589 E-02	3.7309 E-06
14	1.0190 E-02	1.0190 E-02	2.3681 E-06	1.5715 E-02	4.3139 E-06
16	1.1616 E-02	1.1616 E-02	2.7574 E-06	1.7782 E-02	4.8788 E-06

TAB. 2.5 – Moyenne de l'erreur de cinq estimateurs de phase en fonction d'une modulation exponentielle de l'amplitude. L'estimateur ϕ_{DFT} considère la phase du maximum local, $\phi_{\mathrm{DFT}+\xi_N}$ considère la phase du maximum local et la corrige grâce à Δ_f. L'estimateur $\phi_{c+\xi_N}$ considère la phase du spectre de Fourier en la fréquence estimée et la corrige grâce à Δ_f. Les estimateurs ϕ_τ et $\phi_{c\tau}$ considèrent la demie somme entre les valeurs de phase à gauche et à droite du maximum local obtenue respectivement par une DFT avec un facteur de *zero-padding* de 8 et deux calculs du spectre de Fourier en les fréquences $S(\hat{f} \pm \frac{F_e}{N})$.

SNR	ϕ_{DFT}		ϕ_τ		$\phi_{c\tau}$	
	moyenne	variance	moyenne	variance	moyenne	variance
90	4.75 E-02	2.51 E-09	7.78 E-03	3.20 E-08	7.58 E-03	9.24 E-12
80	4.75 E-02	2.52 E-09	7.78 E-03	3.20 E-08	7.58 E-03	3.78 E-11
70	4.75 E-02	2.56 E-09	7.78 E-03	3.17 E-08	7.58 E-03	3.14 E-10
60	4.75 E-02	4.15 E-09	7.78 E-03	3.55 E-08	7.58 E-03	3.35 E-09
50	4.75 E-02	1.69 E-08	7.78 E-03	6.41 E-08	7.58 E-03	3.62 E-08
40	4.75 E-02	1.31 E-07	7.81 E-03	3.34 E-07	7.61 E-03	2.87 E-07
30	4.75 E-02	1.71 E-06	7.73 E-03	3.98 E-06	7.53 E-03	3.87 E-06
20	4.71 E-02	1.48 E-05	8.81 E-03	2.98 E-05	8.67 E-03	2.93 E-05
10	4.63 E-02	1.45 E-04	1.67 E-02	1.40 E-04	1.66 E-02	1.35 E-04
0	5.56 E-02	1.15 E-03	4.60 E-02	1.41 E-03	4.58 E-02	1.42 E-03
-10	1.03 E-01	5.87 E-03	1.36 E-01	1.13 E-02	1.40 E-01	1.14 E-02
-20	1.26 E+00	8.87 E-01	1.27 E+00	8.66 E-01	5.03 E-01	2.03 E-01
-30	1.59 E+00	8.60 E-01	1.58 E+00	8.54 E-01	8.52 E-01	3.97 E-01

TAB. 2.6 – Moyenne de l'erreur de trois estimateurs de phase en fonction d'un SNR décroissant. La sinusoïde est modulée en fréquence ($\Delta_f = 10$). L'estimateur ϕ_{DFT} considère la phase du maximum local. Les estimateurs ϕ_τ et $\phi_{c\tau}$, considèrent la demie somme entre les valeurs de phase à gauche et à droite du maximum local obtenue respectivement par une DFT avec un facteur de *zero-padding* de 8 et deux calculs du spectre de Fourier en les fréquences $S(\hat{f} \pm \frac{F_e}{N})$.

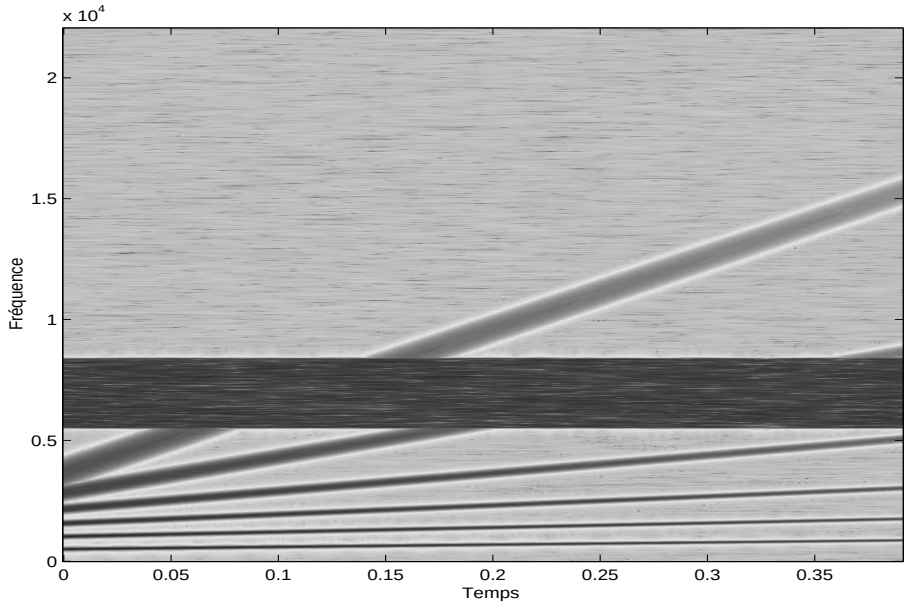


FIG. 2.22 – Spectrogramme du signal test. Ce signal est composé de 6 sinusoïdes dont la fréquence croît et l'amplitude décroît au cours du temps mixés avec un bruit blanc filtré passe bande.

fréquence $\hat{f}$ calculée grâce à l'équation 2.52 avec des paramètres non stationnaires préalablement estimés.

Pour évaluer le gain apporté par la prise en compte du caractère non stationnaire des évolutions de fréquence et d'amplitude dans la sélection des maxima locaux pertinents, on considère un signal test dont le spectrogramme est représenté dans la figure 2.22. Ce signal est composé de 6 sinusoïdes dont la fréquence croît et l'amplitude décroît au cours du temps mixées avec un bruit blanc filtré de manière à n'occuper qu'une région du spectre. À chaque trame, les maxima locaux sont triés selon trois critères : l'amplitude de la composante DFT, Γ_s défini par l'équation 2.36 ou Γ_{ns} défini par l'équation 2.68. Dans la figure 2.23, seuls les dix pour cent des maxima locaux ayant l'indice de tri le plus élevé sont conservés. On peut constater sur la figure 2.23(c) que la prise en compte des paramètres non stationnaires (ici calculés grâce aux estimateurs de Masri) permet de mieux distinguer les maxima locaux issus du bruit et ceux issus de sinusoïdes modulées.

Certains maxima locaux issus de sinusoïdes fortement modulées sont supprimés alors que l'évolution des paramètres de fréquence et d'amplitude au cours du temps est parfaitement déterministe.

Nous sommes convaincus que déterminer si un maximum local est effectivement issu d'une composante sinusoïdale ne doit donc pas être fait à un instant donné mais en combinant les résultats d'analyses effectuées en plusieurs instants successifs. Cet aspect sera détaillé dans le chapitre suivant, dédié au modèle sinusoïdal à long terme.

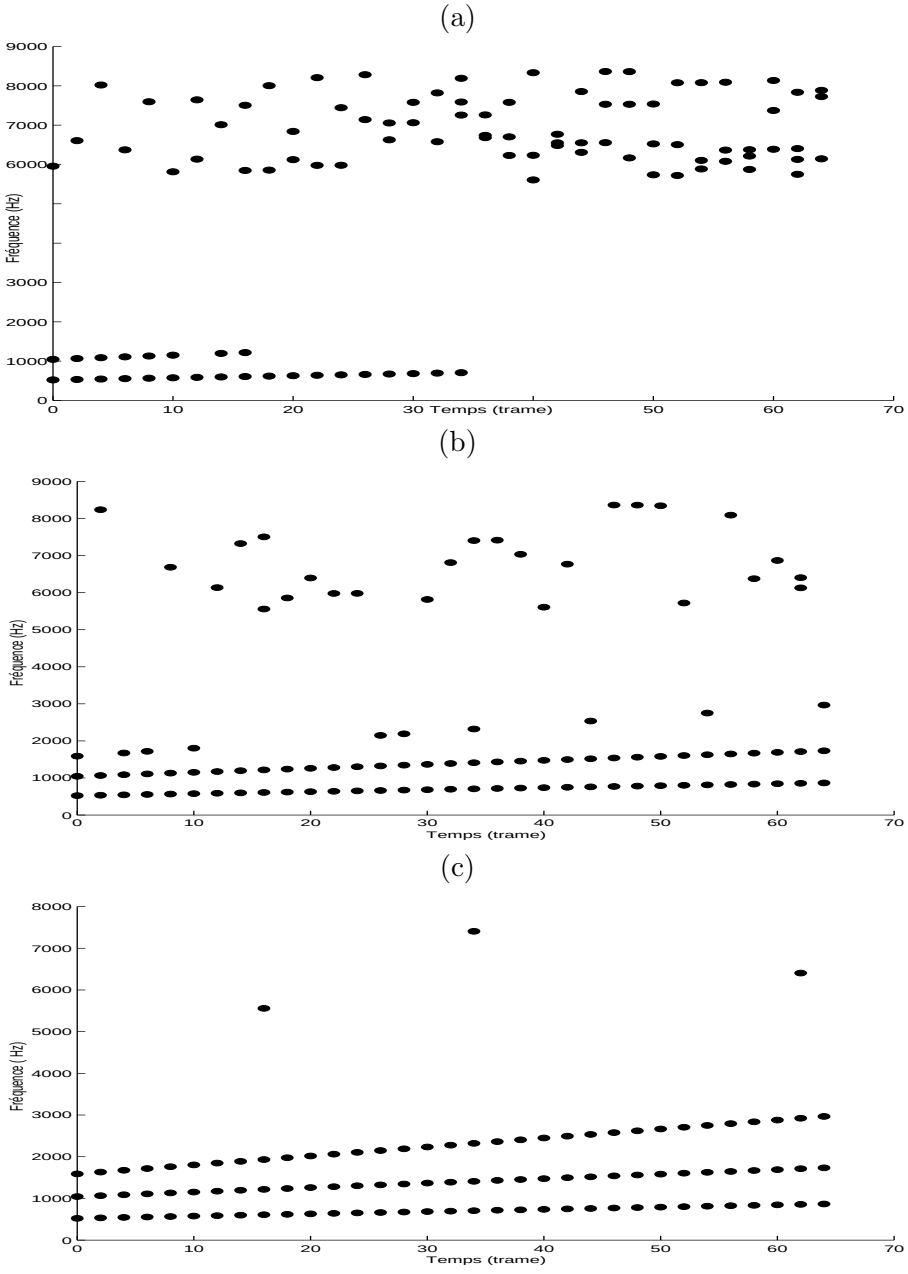


FIG. 2.23 – Maxima locaux conservés par une sélection sur les amplitudes des maxima locaux (en haut), en utilisant le critère de conformité stationnaire Γ_s (au milieu) et en utilisant le critère de conformité non stationnaire Γ_{ns} (en bas).

2.5 Synthèse non stationnaire

La synthèse de sinusoïdes dans un modèle non-stationnaire à court terme consiste à générer une somme de sinusoïdes dont les paramètres d'amplitude et de fréquence varient dans la fenêtre de synthèse.

2.5.1 Approche temporelle

Lors de la synthèse d'une sinusoïde dont la fréquence varie linéairement au cours du temps, l'incrément de phase entre deux échantillons n'est pas constant. En revanche, cet incrément de phase est augmenté d'un facteur constant à chaque échantillon. Ceci mène à un algorithme de calcul rapide d'une exponentielle complexe $x_c[n]$ dont la fréquence varie linéairement au cours du temps. Soient w_0 la fréquence réduite au début de la trame et Δ_ω, le delta de fréquence réduite dont ω augmente durant la trame de synthèse :

$$\omega_0 = 2\pi \frac{f(0)}{F_e} \tag{2.69}$$

$$\Delta_\omega = 2\pi N \frac{\Delta_f}{{F_e}^2} \tag{2.70}$$

La valeur du cosinus $x[n]$ est égale à la partie réelle de $x_c[n]$. La phase d'initialisation est la suivante :

$$x_c[0] = a\, e^{j\phi(0)} \tag{2.71}$$

$$D_1[0] = e^{2j\left(w_0 \frac{\Delta_\omega}{2N}\right)} \tag{2.72}$$

$$D_2 = e^{\left(j \frac{\Delta_\omega}{N}\right)} \tag{2.73}$$

où $\phi(0)$, $f(0)$ et Δ_f sont les paramètres de la sinusoïde à synthétiser, voir équation 2.48. La récursion se déroule comme suit :

$$x_c[n] = x_c[n-1] \cdot D_1[n-1] \tag{2.74}$$

$$D_1[n] = D_1[n-1] \cdot D_2 \tag{2.75}$$

$$x[n] = \Re\left(x_c[n]\right) \tag{2.76}$$

où $\cdot$ désigne une multiplication complexe et $\Re(x)$ désigne la partie réelle de x. Comme une multiplication complexe implique 6 opérations flottantes, le nombre total d'opération flottantes par échantillon est donc de 12.

2.5.2 Approche spectrale

L'utilisation de la IDFT pour la synthèse de sinusoïde dont la fréquence varie amène plusieurs difficultés. D'une part, la forme du spectre est dépendante des modulations. D'autre part, comme on l'a vu dans les sections précédentes, une modulation de fréquence étale le lobe principal, voir figure 2.18(a). Le nombre de points spectraux des contributions spectrales des sinusoïdes à synthétiser

doit donc être plus grand pour conserver une qualité de synthèse similaire. Une alternative pour conserver une complexité raisonnable est de considérer les paramètres de fréquences et d'amplitudes comme constants sur de plus petits intervalles de temps que celui de l'analyse et d'utiliser l'algorithme stationnaire de synthèse par approche fréquentielle décrit dans la section 2.2.

Conclusions

On a étudié dans ce chapitre des estimateurs de paramètres stationnaires qui améliorent sensiblement la précision des estimateurs basiques de la transformée de Fourier. Ensuite, différents estimateurs de paramètres non stationnaires comme la dérivée de la fréquence ou l'évolution exponentielle de l'amplitude existant dans la littérature ont été présentés. Seul l'estimateur de Almeida, basé sur la fenêtre gaussienne, possède une formulation analytique. L'utilisation de fenêtres trigonométriques dans un modèle non stationnaire amène des difficultés car l'expression analytique du spectre d'une sinusoïde dont la fréquence varie linéairement est à ce jour inconnue. Pour pallier ce problème, des approches empiriques ou des approximations sont proposées dans la littérature.

La connaissance de la dérivée de la fréquence est d'un intérêt particulier pour le suivi de partiels car elle permet de mieux déterminer si deux pics de trames adjacentes font partie du même partiel. Cependant, l'utilisation des estimateurs de Masri n'a pas amené à des résultats concluants. De plus, les signaux qui contrôlent les partiels ne sont pas simplement linéaires, il serait donc utile de considérer les dérivées d'ordre supérieures. Or, l'estimation des paramètres non stationnaires d'ordre 1 en utilisant une fenêtre trigonométrique est un domaine de recherche encore jeune et actif [Röb02, ML03a], des estimateurs des dérivées d'ordres supérieurs n'ont pas encore été étudiés à notre connaissance.

Alternativement, on propose dans la suite un modèle d'évolution des paramètres des partiels non polynômial. Ce modèle exploite l'estimation précise des paramètres stationnaires pour améliorer le suivi et l'interpolation de partiels comme cela sera détaillé dans les chapitres 3 et 4.

3 Modélisation sinusoïdale à long terme

On propose dans ce chapitre plusieurs algorithmes de suivi de partiels adaptés à l'analyse de signaux polyphoniques. Une modélisation explicite des transitions entre pics de trames non adjacentes permet de proposer un premier algorithme destiné à être robuste aux dégradations de la représentation à court terme (défauts de pics et pics de bruit). Cette formulation probabiliste d'un intérêt théorique certain se révèle en pratique peu efficace car trop éloigné des contraintes relatives au modèle sinusoïdal. On exploite alors dans un second algorithme le caractère prédictible des évolutions des paramètres des partiels. Une modélisation autorégressive de ces évolutions permet d'améliorer notablement l'extraction des partiels de sons modulés comme ceux présentant un vibrato ou un trémolo. Grâce à cette modélisation plus fine, de nombreux pics de bruit peuvent être évités. Cette approche ne peut en revanche éviter certaines erreurs lorsque les fréquences des partiels sont très proches. Or, ce cas est fréquent lors de l'analyse de signaux polyphoniques. Un nouveau critère basé sur l'absence théorique de hautes fréquences dans les évolutions des paramètres des partiels permet de pallier ce problème et d'obtenir une représentation adaptée à nos besoins.

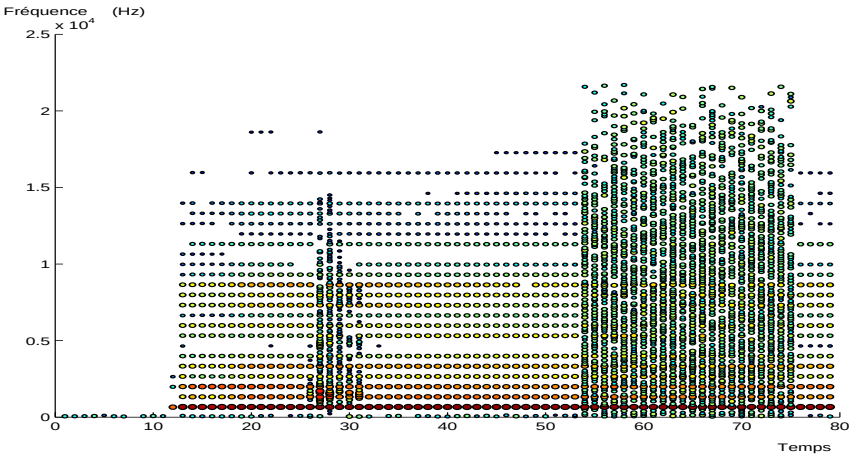


FIG. 3.1 – Représentation à court terme composée de pics (maxima locaux du spectre), d'un enregistrement d'une note tenue d'accordéon avec le claquement d'une paire de castagnettes puis un bruit blanc.

3.1 Introduction

Le modèle sinusoïdal long terme permet une représentation de haut niveau du signal sonore, utile pour de nombreuses applications comme la modification des sons, le codage et plus récemment l'indexation et la séparation de sources, voir section 1.2.2. Un modèle sinusoïdal à court terme (ensemble de pics) est utilisé pour identifier les paramètres instantanés des composantes sinusoïdales long-terme (les partiels) à un instant donné.

Une représentation composée de trames court-terme successives est par définition discrète. Le propos de l'extraction d'un modèle sinusoïdal à long terme est donc de restaurer la continuité de cette représentation du contenu spectral. Plus précisément, il est nécessaire d'identifier le début et la fin d'une composante long-terme et de définir quelles composantes court-terme la compose. C'est le propos des algorithmes dits de *suivi de partiels* : certains pics sont sélectionnés et reliés de trame en trame pour former des partiels.

Les premiers algorithmes de suivi de partiels dédiés à la modélisation sinusoïdale des signaux audionumériques ont été proposés dans les années 80. L'un a été proposé par Mc Aulay et Quatieri [MQ86] pour la modélisation des signaux de paroles voisés. L'autre a été proposé par Smith et Serra [SS87] pour la modélisation des sons inharmoniques et de sons harmoniques de fréquence de fondamentale variable. Ces algorithmes de faible complexité sont présentés dans la section 3.2.

Lors de l'analyse de sons non purement sinusoïdaux, certains pics dit de "bruit" ne doivent pas être retenus dans une représentation sinusoïdale long terme. Ces pics sont issus de processus transitoires ou stochastiques (castagnettes ou bruit blanc sur la figure 3.1). De plus, certains pics peuvent être

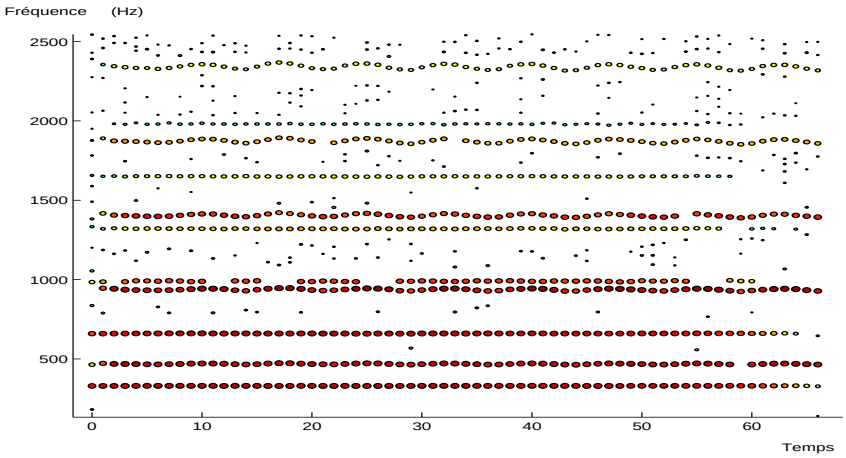


FIG. 3.2 – Pics spectraux (maxima locaux du spectre) d'un duo de flûtes analysé avec une fenêtre d'analyse de 2048 échantillons. Certaines harmoniques proches ne sont pas différenciées par le module d'extraction de pics.

manquant, voir trame 48 à 10 kHz sur la figure 3.1. Ceci peut être dû à une amplitude trop faible, à une modulation trop forte ou à une corruption du spectre. En effet, lors de l'analyse de signaux polyphoniques, les composantes sinusoïdales du signal analysé peuvent être arbitrairement proches. Le module d'analyse à court terme n'identifie qu'un seul pic alors que deux composantes sinusoïdales sont présentes dans le signal analysé, voir trame 10 à 1 kHz sur la figure 3.2. Une solution consiste à augmenter la résolution de la transformée de Fourier en augmentant le nombre d'échantillons temporels considérés tout en conservant le même pas d'avancement. Cette méthode a ses limites car la perte de précision temporelle peut être très gênante lors de l'analyse de signaux modulés comme la voix. Ce phénomène appelé "étalement spectral" peut être observé sur la figure 3.3.

Pour résumer, un algorithme de suivi "idéal" doit être capable de n'utiliser que des pics pertinents pour la formation des partiels, de rejeter ainsi les pics de bruit ou les pics corrompus et d'être robuste aux pics manquants. Les sections 3.4, 3.6 et 3.7 décrivent les trois méthodes originales développées lors de ce travail de thèse autant du point de vue théorique qu'algorithmique. Ces méthodes sont ensuite évaluées dans la section 3.8. Tout d'abord, les qualités intrinsèques d'un module de suivi sont évaluées selon un protocole original. Ses qualités sont ensuite évaluées selon des critères objectifs et subjectifs lors d'une intégration de ce module dans une chaîne complète d'analyse/synthèse.

Une fois les paramètres d'un modèle sinusoïdal long terme estimés, des algorithmes de synthèse présentés dans la section 3.9 permettent d'obtenir un signal temporel de cette représentation spectrale.

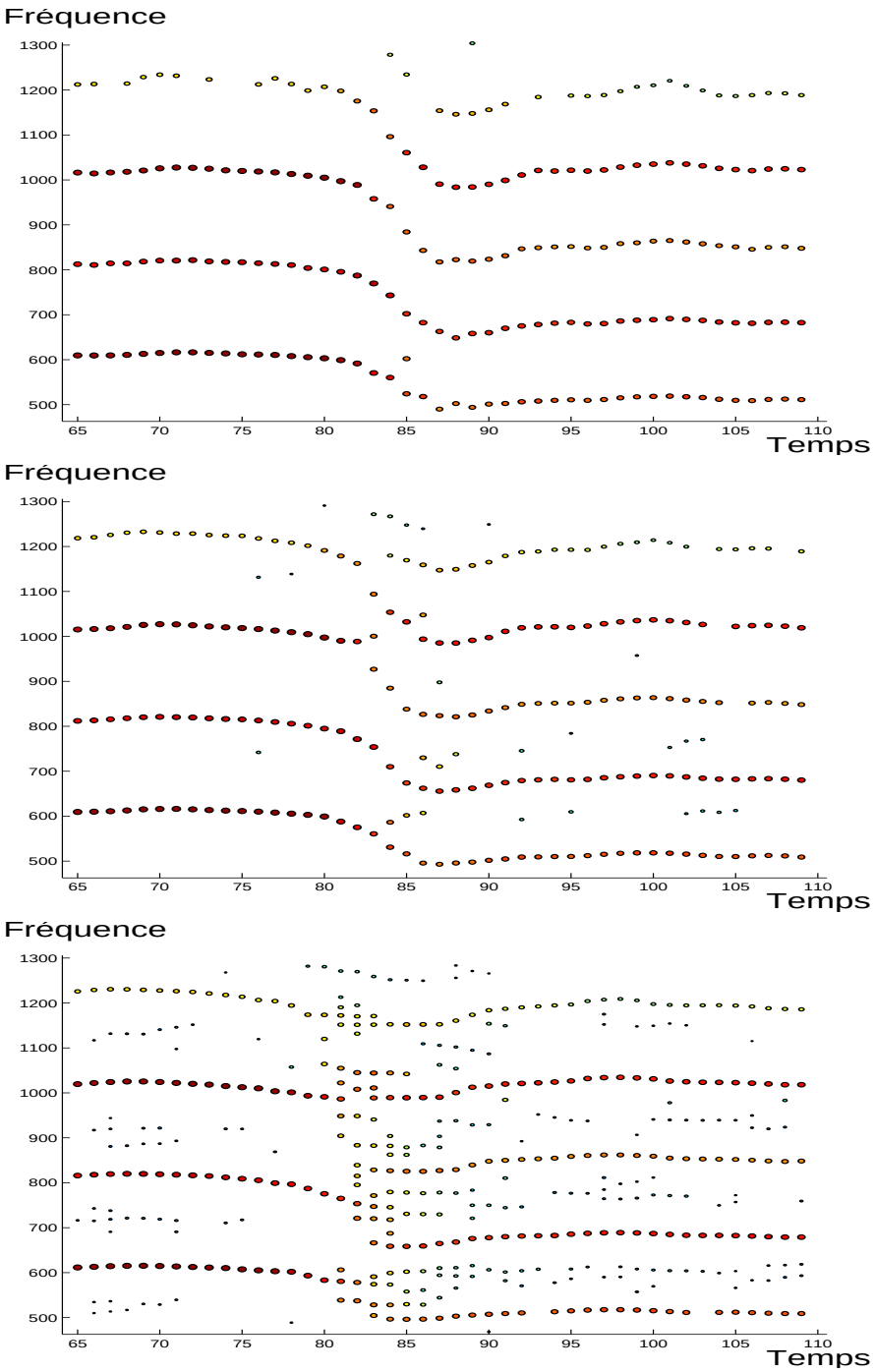


FIG. 3.3 – Pics spectraux (maxima locaux du spectre) d'une voix chantée analysée avec un pas d'avancement de 512 échantillons et une fenêtre d'analyse de taille 1024, 2048 et 4096 échantillons respectivement. Plus la taille de fenêtre est grande, plus l'étalement temporel est prononcé.

3.2 Algorithmes de suivi de faible complexité

L'algorithme dit de Mc Aulay et Quatieri (noté MAQ dans la suite) est proposé dans [MQ86] pour la modélisation sinusoïdale à long terme des signaux de parole voisés. On cherche à lier les pics de la trame k avec ceux de la trame $k+1$. Si un pic de la trame k ne trouve pas de successeur dans la trame $k+1$, le partiel auquel il appartient est déclaré "mort" et se termine donc à la trame k. Si un pic de la trame $k+1$ n'est pas lié à un pic de la trame k, un partiel "naît". Pour déterminer un successeur, on procède comme suit. On définit un seuil maximal Δ_f de variation de fréquence (voir figure 3.4) entre les deux pics consécutifs d'un partiel :

$$|f_k^i - f_{k+1}^j| < \Delta_f \tag{3.1}$$

où f_k^i désigne la fréquence en Hz du pic d'indice i de la trame k. Le déroulement de l'algorithme se fait trame par trame et par fréquence croissante. Supposons que l'on ait relié chaque pic de rang inférieur à i pour une trame d'indice k. Au pic d'indice i de la trame k noté p_k^i, on associe un pic de la trame $k+1$ non encore relié tel que la différence de fréquence entre ces deux pics est minimale. Si cette différence est supérieure à Δ_f, ce partiel est déclaré mort et écarté de tout autre considération. Dans le cas contraire, l'équation suivante est vérifiée :

$$|f_k^i - f_{k+1}^j| < |f_k^i - f_{k+1}^m| < \Delta_f \ \forall m \neq j \tag{3.2}$$

Si le pic candidat p_{k+1}^j n'a pas une fréquence plus proche du pic de fréquence supérieure p_k^{i+1} non encore lié :

$$|f_k^i - f_{k+1}^j| < |f_k^{i+1} - f_{k+1}^j| < \Delta_f \tag{3.3}$$

alors on relie définitivement p_k^i et p_{k+1}^j (cas du partiel P_1 de la figure 3.4). Si ce n'est pas le cas et qu'il n'existe pas d'alternative pour le pic p_k^i, alors le partiel auquel il appartient est déclaré mort (cas du partiel P_2 de la figure 3.4). Lorsque toutes les liaisons entre les pics de la trame k et ceux de la trame $k+1$ sont faites, pour tout pic non encore relié de la trame $k+1$, on fait naître un nouveau partiel.

Pour éviter des discontinuités du signal lors de la synthèse, il est proposé d'ajouter au début et à la fin des partiels des pics d'amplitude nulle, les fréquences et les phases étant extrapolées. Ceci est particulièrement utile dans la modélisation de la voix, mais peut entraîner des phénomènes désagréables de lissage des attaques dans le cas de la modélisation des signaux musicaux lors de l'utilisation de fenêtre d'analyse de taille élevée.

État "zombie"

Pour pouvoir extraire une représentation long-terme des harmoniques de rang élevé ou des harmoniques de signaux polyphoniques, le suivi de partiels doit être robuste aux défauts de pics spectraux, voir figures 3.1 à la trame 48 et

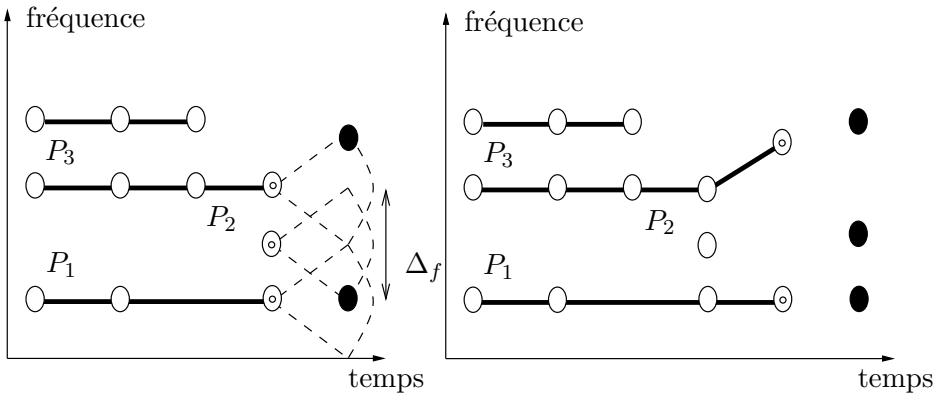


FIG. 3.4 – Déroulement et résultat de l'algorithme proposé par Mc Aulay et Quatieri. Les pics évidés ne sont plus actifs, tandis que les pics pleins sont des pics non encore reliés. Les pics évidés ayant un cercle représente les "queues" des partiels actifs.

3.2 à la trame 12. Pour cela, il est proposé dans [SS87] de permettre à un partiel qui ne trouve pas de pic satisfaisant à une trame donnée d'être mis dans un état "zombie" pour un nombre donné de trames. Ce partiel passe de l'état zombie à l'état vivant s'il trouve un pic vérifiant les contraintes fixées avant d'avoir épuisé un nombre donné d'utilisation de l'état zombie. Dans le cas contraire, le partiel passe de l'état zombie à l'état mort et n'est plus considéré, voir figure 3.5. Ce partiel se termine alors à l'indice de la trame correspondant au dernier pic inséré. La durée pendant laquelle un partiel peut rester dans l'état zombie peut être le même pour tout l'ensemble des partiels ou bien être spécifique à chaque partiel.

Grâce à cette méthode, un pic n'est pas utilisé à une trame donnée parce qu'il ne satisfait pas une contrainte locale et non parce qu'un pic appartenant aux trames suivantes est plus approprié. La transition entre deux pics de trames non adjacentes est ainsi rendu possible de manière implicite.

Paramétrage

Le seuil Δ_f peut être fixé suivant plusieurs critères. Dans le cas de l'analyse de signaux monophoniques et harmoniques comme la partie voisée du signal de parole, on peut utiliser une transformée de Fourier avec une résolution fréquentielle choisie en fonction d'une estimation de la fréquence de fondamentale. On peut alors fixer Δ_f en fonction de la résolution spectrale.

Dans le cas contraire, on peut considérer les propriétés physiques des signaux que l'on souhaite modéliser. Par exemple, l'harmonique de rang n d'une note comportant un vibrato a une excursion en fréquence qui est n fois l'excursion en fréquence de la fondamentale. Si l'excursion maximale pour la fréquence de fondamentale est égale à 4 Hz entre 2 trames, l'excursion maximale pour la dixième harmonique sera de 40 Hz. Cette valeur est un compromis acceptable en pratique mais ne permet pas d'identifier les harmoniques de rang d'harmonicité

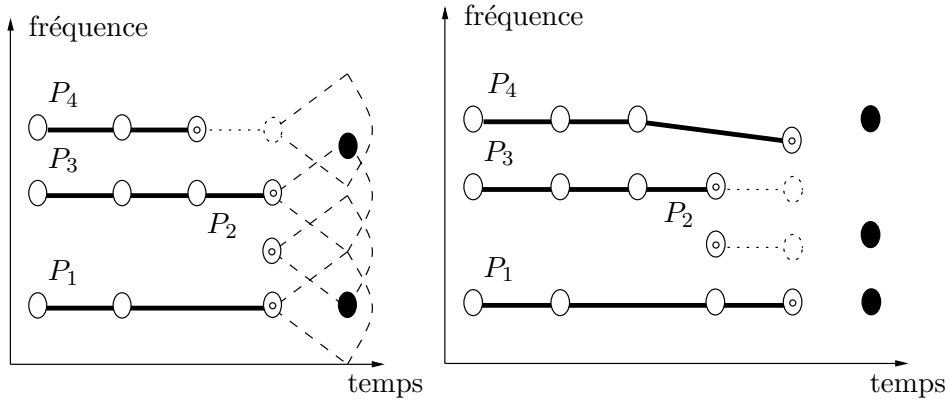


FIG. 3.5 – Utilisation du mode "zombie" avec l'algorithme de Mc Aulay et Quatieri. Un partiel n'ayant pas relié un pic lors d'une étape de l'algorithme (le partiel P_4) reste actif et se prolonge (pic en pointillés).

élevé dans le cas d'une forte modulation.

De plus, il est souhaitable que Δ_f soit suffisamment faible pour que les partiels en relation d'harmonicité n'entrent pas en concurrence. Ce seuil doit donc être deux fois plus petit que la plus petite hauteur de fondamentale que l'on est supposé analyser. Là encore, un Δ_f de 40 Hz est acceptable.

De part sa simplicité et son efficacité, l'algorithme MAQ est devenu particulièrement populaire. Lorsque la résolution fréquentielle est adaptée à la hauteur du signal monophonique et harmonique, la représentation à court terme est de très bonne qualité : peu de pics sont manquants et peu de pics de bruit sont présents. En ce cas, l'application de ce type d'algorithme permet l'extraction d'une représentation à long terme interprétable et permet une synthèse très fidèle.

Lors de l'analyse de signaux polyphoniques ou non stationnaires et/ou bruités, la représentation à court terme est de moins bonne qualité. La représentation à long terme extraite par l'algorithme MAQ conserve alors le plus souvent une fidélité satisfaisante. En revanche, les possibilités d'interprétation sont très réduites comme on peut le constater sur la figure 3.32(a). En particulier, les modulations de fréquences (vibrato) ne sont pas clairement identifiées et de nombreux partiels modélisent plusieurs harmoniques de notes différentes.

De plus, l'algorithme MAQ propose de représenter tous les pics sous forme de partiels, or certains pics sont des maxima locaux issus de composantes de bruit et doivent être évités. Pour être à même d'identifier une ou plusieurs composantes sinusoïdales dans un signal bruité, des algorithmes utilisant des formalismes statistiques ont été proposés dans la littérature et sont étudiés dans la section suivante.

3.3 Suivi par modèle de Markov caché

Les algorithmes de suivi par modèle de Markov caché (HMM), de l'anglais "*Hidden Markov Model*", sont présentés dans cette section. La littérature abondante nous amène à faire un court historique. Un algorithme dédié au suivi de partiels dans des signaux musicaux [DGR93b] utilisant certains de ces aspects est ensuite présenté.

3.3.1 Historique

Du fait des applications essentiellement militaires dans un contexte de fin de guerre froide, le suivi de lignes fréquentielles (équivalent des partiels pour l'imagerie radar ou sonar) par application du modèle des HMM a reçu un effort conséquent de la part de la communauté du traitement du signal [SB90, XE91] et est toujours d'actualité [PJ03].

Une ou plusieurs lignes de fréquence sont identifiées dans des équivalents de spectrogrammes issus de dispositifs radar ou sonar. Ces lignes de fréquences sont la représentation spectrale des vibrations induites par des pièces mécaniques vibrant à un certain régime (moteur de propulsion d'un sous-marin par exemple). Dans ce type d'applications, le nombres de composantes à suivre est très faible (unique dans la plupart des cas) et peu modulées (le moteur d'un sous-marin a un régime assez stable). En revanche, le rapport de signal à bruit est extrêmement faible (inférieur à -10 dB).

Dans un domaine plus proche de cette thèse, ce formalisme a été appliqué au suivi de formants [Kop86] dans les signaux de paroles. Une présentation complète du formalisme HMM sortant du cadre de ce manuscrit, le lecteur est invité à se référer à [RJ86] pour de plus amples informations. Brièvement, une chaîne de Markov à nombre fini d'états est caractérisée par une matrice de probabilité de transition $\mathcal{A}$. Les éléments de cette matrice sont les probabilités d'effectuer une transition d'un état à un autre de la chaîne. Dans le cas des HMM, on considère une autre matrice, la matrice de probabilité d'observations $\mathcal{B}$. Les éléments de $\mathcal{B}$ définissent la probabilité d'être dans un état donné en fonction d'une observation particulière.

Dans les algorithmes de suivi présentés ci-dessus, les différentes bandes de fréquences que peut occuper la ligne de fréquence ou le formant représentent les états. Accessoirement, on peut ajouter un état 0, qui représente le cas où aucune ligne ou formant n'est présent à un instant donné. Des spectres à court-terme quantifiés constituent les observations. La matrice $\mathcal{A}$ est supposée connue car elle représente une connaissance *a priori*. En pratique, cette matrice est obtenue par apprentissage sur une large base de tests. Conditionnellement à une série finie d'observations, on trouve grâce à l'algorithme de Viterbi [GDF73] la succession d'états optimale. Ce type de modèle limite à 1 le nombre d'éléments que l'on cherche à identifier.

Il est proposé dans [Kop86] d'étendre ce modèle au suivi de trois formants. Il considère comme état de la chaîne de Markov un triplet de fréquence de formants. L'utilisation de ce modèle entraîne 10^3 états et demande l'exploration

de 10^6 transitions à chaque étape de l'algorithme de Viterbi. Cette solution fut donc rejetée pour des raisons de complexité. Il est proposé de réduire le nombre de transitions considérées en mettant à zéro des éléments de la matrice $\mathcal{B}$ pendant l'apprentissage. Soient des observations (x, y, z) notant trois détections aux fréquences x, y et z et des états $[x, y, z]$ notant la présence de trois formants de fréquences x, y et z. Il est par exemple hautement improbable que le premier formant ait une fréquence plus élevée que le deuxième ou troisième formant. Par conséquent, l'élément de la matrice $\mathcal{B}$ qui associe une observation du type $(200, 500, 1000)$ à l'état $[500, 200, 1000]$ peut être mis à 0. Malheureusement, lors du suivi de partiels de signaux polyphoniques, ces considérations permettant de réduire la complexité ne peuvent plus être utilisées car les partiels peuvent se croiser au contraire des formants. De plus, le nombre d'éléments que l'on cherche à identifier est inconnu.

3.3.2 Application aux signaux de musique

Des recherches menées à l'IRCAM [DGR93a, DGR93b] ont permis de proposer un algorithme de suivi de partiels dédié aux signaux musicaux. Dans cet algorithme, les informations des trames court-terme (fréquences et amplitudes des pics) ne sont pas utilisées comme observations, mais sont utilisées directement pour calculer les probabilités de transition entre différents états (éléments de la matrice $\mathcal{A}$). L'algorithme de Viterbi est alors utilisé pour déterminer la séquence d'états qui minimise la fonction de coût globale (comprenant tout les partiels) sur plusieurs trames. Une fonction de coût prenant en compte la dérivée seconde de la fréquence et la dérivée de l'amplitude entre trois pics de trames successives est proposée pour identifier la séquence d'états optimale en fonction d'un nombre donné de trames court-terme.

Soit une représentation à court terme d'un signal comprenant cinq pics, un état est une combinaison de deux trames court-terme (ordonnées en fréquence) annotées de cette manière : $[0, 0, 1, 3, 2]$. Les deux premiers pics sont des pics de bruit (indice 0), les trois autres appartiennent aux partiels 1, 2 et 3. Les deux derniers partiels ont effectué un croisement.

L'algorithme identifie les séquences d'états pour un nombre fini de trames d'observations (fixé ici à 2). Une fenêtre glissante avec recouvrement est utilisée. Supposons que l'algorithme lors d'une première étape ait identifié une séquence d'états ($[1, 0, 2]$, $[1, 0, 2]$) aux trames d'indices 1 et 2. Lors de la seconde étape, la séquence d'états identifiée aux temps 2 et 3 est : $[1, 0, 0]$ et $[1, 0, 0]$. On constate que le partiel d'indice 2 est mort au temps 2.

Les points forts de cette méthode sont : d'une part de considérer une minimisation globale d'une fonction de coût, et d'autre part d'effectuer cette minimisation sur un certain nombre de trames, ce qui permet d'être plus robuste à certains artefacts du modèle à court terme comme les pics de bruit. La complexité de ce type d'algorithme est toutefois élevée [Gar92]. Pour une fenêtre de 5 trames contenant 10 pics, le nombre total d'états est de 10^6 et le nombre de probabilités de transition est de 10^9.

3.4 Suivi par exploration des trajectoires futures

On introduit ici un nouvel algorithme de suivi qui conserve le principe d'une minimisation d'une fonction de coût sur plusieurs trames pour être à même d'éviter les pics de bruit. En revanche, la minimisation globale sur tout l'ensemble des partiels n'est pas conservée car cette minimisation entraîne un surcoût de calcul important. On préfère déterminer la prolongation "optimale" de chaque partiel au sens d'une certaine fonction de coût indépendamment de la présence des autres partiels. La gestion de la concurrence entre partiels est effectuée après coup.

L'objectif de cet algorithme est d'être résistant à des dégradations diverses du modèle à court terme. Ces dégradations peuvent être : des pics corrompus, des pics de bruit ou des pics manquants. Pour cela, on propose dans [LMR04a] de modéliser de manière explicite une transition entre deux pics de trames non adjacentes, au contraire du mode zombie introduit dans la partie 3.2 où ce type de transition est rendue possible de manière implicite. Cette modélisation originale permet d'être robuste à l'absence de pic mais aussi d'éviter un pic corrompu ou de bruit à une trame donnée en choisissant de manière explicite une transition vers un pic de la trame suivante qui amène une liaison de distorsion plus faible.

La probabilité que nous proposons d'associer à ce type de transition est présentée dans la première partie. À partir d'un sous-ensemble de pics non encore utilisés par des partiels (noté Γ) de petites trajectoires sont construites sur la base de cette probabilité de transition, voir figure 3.6. Cette construction, détaillée dans la deuxième partie, se fait dans le sens inverse du temps. Ces trajectoires se terminent à la queue d'un partiel et indiquent ainsi l'extension optimale du dit partiel. Un pic de l'ensemble Γ peut être utilisé dans plusieurs trajectoires alors que les partiels sont soumis aux règles de l'allocation exclusive. Pour résoudre ce problème, un algorithme de gestion de la concurrence est introduit dans une troisième et dernière partie.

3.4.1 Probabilité de transition

On introduit ici une probabilité de transition entre pics issus de trames non adjacentes. Le modèle choisi pour la probabilité d'une transition entre deux pics de trames adjacentes est tout d'abord explicité. Une expression analytique de cette probabilité valide pour toutes les longueurs de transition est ensuite obtenue par récurrence.

Une n-transition est une transition directe (sans pic intermédiaire) d'un pic de la trame k à un pic de la trame $k+n$. La probabilité λ_n est la probabilité d'effectuer une n-transition d'un pic p_k^i de fréquence f_k^i à un pic p_{k+n}^j de fréquence f_{k+n}^j. Nous proposons de modéliser la probabilité d'effectuer une 1-transition entre deux pics p_k^i et p_{k+1}^j par une gaussienne de variance σ, prenant comme

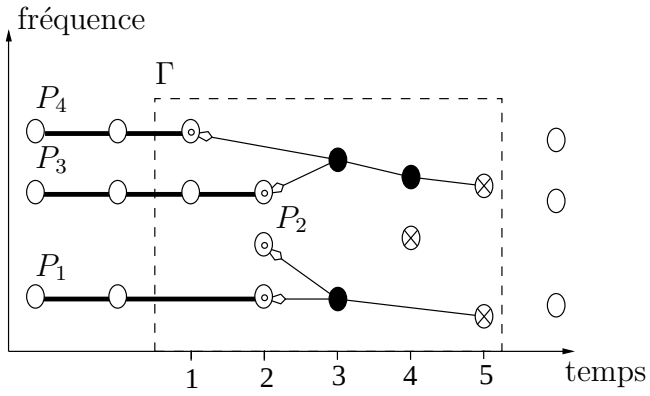


FIG. 3.6 – Construction des trajectoires (flèches) dans Γ en sens inverse du temps. Ces trajectoires peuvent partager des pics et effectuer des transitions entre pics de trames non adjacentes. Les pics blancs sont inactifs, les pics avec un double cercle sont les queues des partiels (en trait épais).

paramètre la différence de leurs fréquences respectives :

$$\lambda_1(p_k^i, p_{k+1}^j) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\left(\frac{(f_k^i - f_{k+1}^j)^2}{2\sigma^2}\right)} \tag{3.4}$$

f_k^i étant la fréquence de p_k^i, f_{k+1}^j la fréquence de p_{k+1}^j et σ, la variance de la gaussienne, étant un paramètre à déterminer. Par mesure de clarté, l'indice du pic dans la trame est omis dans la suite et le pic de départ est situé à la trame d'indice $k = 0$.

Lors d'une 2-transition entre deux pics p_0 et p_2, on considère que l'on est passé par un pic "virtuel" $\tilde{p}_1$ uniquement déterminé par sa fréquence dont on ne connaît pas la valeur précise mais dont les valeurs possibles $\tilde{f}_1$ suivent une loi de probabilité gaussienne g de même variance σ, centrée en la valeur moyenne $f_g = (f_0 + f_2)/2$:

$$g(f_g, f) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\left(\frac{(f_g - f)^2}{2\sigma^2}\right)} \tag{3.5}$$

La probabilité d'effectuer une transition de taille 2 s'exprime donc par :

$$\lambda_2(p_0, p_2) = \int_{-\infty}^{\infty} g(f_g, \tilde{f}_1)\lambda_1(p_0, \tilde{p}_1)\lambda_1(\tilde{p}_1, p_2)\, d\tilde{f}_1 \tag{3.6}$$

où $\lambda_1(p_0, \tilde{p}_1)\lambda_1(\tilde{p}_1, p_2)$ représente la composition de deux 1-transitions passant par un pic virtuel $\tilde{p}_1$ ayant pour fréquence $\tilde{f}_1$. En développant on trouve :

$$\lambda_2(p_0, p_2) = \frac{1}{\sqrt{12}(\sigma\sqrt{\pi})^2}\, e^{-\left(\frac{(f_0 - f_2)^2}{4\sigma^2}\right)} \tag{3.7}$$

Le principe d'extension est de considérer qu'une $n+1$-transition est la composition linéaire en fréquence (de distorsion minimale) d'une 1-transition et d'une

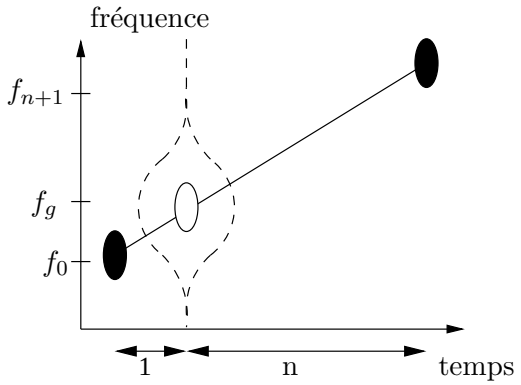


FIG. 3.7 – Principe d'extension de la mesure de probabilité. Les pics réels sont en noirs, le pic virtuel est en blanc. La fréquence de ce pic est distribuée selon une distribution gaussienne et ayant une probabilité de présence maximale en f_g.

n-transition (cf. figure 3.7). Sur ce principe, pour une transition de taille 3, la valeur intermédiaire f_g qui maximise la probabilité d'effectuer une composition linéaire de transitions est le barycentre de f_0 et f_3, f_0 ayant un poids de 2 et f_3 un poids de 1, à savoir $f_g = \frac{2f_0+f_3}{3}$. La probabilité d'effectuer une 3-transition s'exprime donc par :

$$\lambda_3(p_0, p_3) = \int_{-\infty}^{\infty} g(f_g, \tilde{f}_1)\lambda_1(p_0, \tilde{p}_1)\lambda_2(\tilde{p}_1, p_3)\, d\tilde{f}_1 \tag{3.8}$$

et en développant :

$$\lambda_3(p_0, p_3) = \frac{1}{\sqrt{60}(\sigma\sqrt{\pi})^3}\, e^{-\left(\frac{(f_0-f_3)^2}{6\sigma^2}\right)} \tag{3.9}$$

On voit se dégager une expression générale de λ_n de la forme :

$$\lambda_n(p_0, p_n) = \frac{1}{K_n(\sigma\sqrt{\pi})^n}\, e^{-\left(\frac{(f_0-f_n)^2}{2n\sigma^2}\right)} \tag{3.10}$$

K_n étant une suite à déterminer. Exprimons maintenant λ_{n+1} en fonction de λ_n et λ_1 :

$$\lambda_{n+1}(p_0, p_{n+1}) = \int_{-\infty}^{\infty} g(f_g, \tilde{f}_1)\lambda_1(p_0, \tilde{p}_1)\, \lambda_n(\tilde{p}_1, p_{n+1})\, d\tilde{f}_1 \tag{3.11}$$

où $f_g = \frac{f_0 n+f_{n+1}}{n+1}$, f_g étant la fréquence du pic virtuel qui minimise la distorsion en fréquence. Le produit $\lambda_1(p_0, \tilde{p}_1)\lambda_n(\tilde{p}_1, p_{n+1})$ représente la probabilité d'effectuer une $n+1$-transition de p_0 vers p_{n+1} en passant par un pic virtuel $\tilde{p}_1$ de fréquence $\tilde{f}_1$. En calculant l'intégrale, on trouve :

$$\lambda_{n+1}(p_0, p_{n+1}) = \frac{1}{\sqrt{\frac{2(2n+1)}{n}}K_n(\sigma\sqrt{\pi})^{n+1}}\, e^{-\left(\frac{(f_0-f_{n+1})^2}{2(n+1)\sigma^2}\right)} \tag{3.12}$$

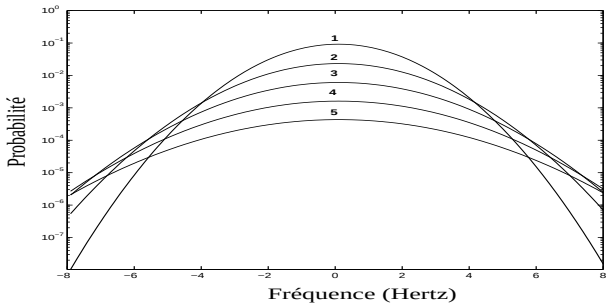


FIG. 3.8 – Représentation de λ_n, la probabilité associée à une n-transition entre deux pics distants d'un certain écart de fréquence, pour $n \in [1, 5]$. À partir d'un certain écart de fréquence (dépendant de σ), il devient plus probable d'effectuer une n-transition qu'une $n-1$-transition.

K_n est définie de manière récurrente par :

$$K_1 = \sqrt{2} \tag{3.13}$$

$$K_n = K_{n-1}\sqrt{2\,\frac{2n-1}{n-1}} \tag{3.14}$$

En développant cette relation de récurrence, on a :

$$K_n = K_1 \prod_{i=2}^{n} \sqrt{2\,\frac{2n-1}{n-1}} \tag{3.15}$$

$$K_n = \sqrt{2\,\frac{(2n-1)!}{(n-1)!(n-1)!}} \tag{3.16}$$

On reconnaît une forme en C_p^n :

$$K_n = \sqrt{2n\,C_n^{2n-1}} \tag{3.17}$$

On montre ainsi que la probabilité d'une n-transition entre deux pics de fréquence f_1 et f_{n+1} est :

$$\lambda_n(p_0, p_n) = \frac{1}{\sqrt{2n\,C_n^{2n-1}}\,(\sigma\sqrt{\pi})^n}\,e^{-\left(\frac{(f_0-f_n)^2}{2n\sigma^2}\right)} \tag{3.18}$$

On obtient alors une mesure de probabilité de transition dont le seul paramètre est σ, la variance de la gaussienne représentée sur la figure 3.8 pour plusieurs tailles de transition. Cette probabilité nous permet de construire les prolongements optimaux des partiels dans les trames futures en modélisant de manière explicite des transitions entre pics de trames non adjacentes.

3.4.2 Trajectoires optimales

À chaque trajectoire est associée une probabilité égale au produit des probabilités associées à chaque n-transition composant la trajectoire. Une trajectoire

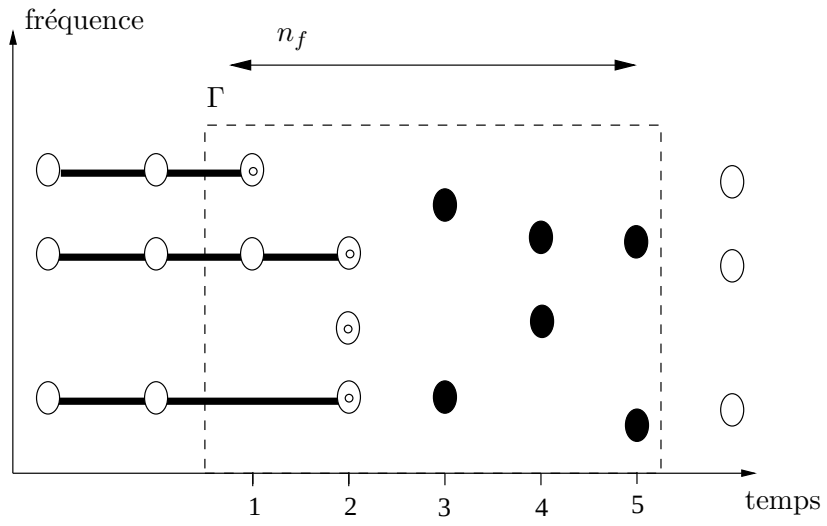


FIG. 3.9 – Positionnement de Γ par rapport aux queues des partiels.

optimale de p_k^i vers p_{k+m}^j est l'ensemble des transitions successives partant de p_k^i et qui mènent à p_{k+m}^j en maximisant la probabilité associée.

Une propriété très importante est exploitée dans l'algorithme de Viterbi [GDF73] : il n'existe qu'une seule trajectoire optimale de p_k^i vers p_{k+n}^j. Dans notre modèle, cette propriété n'est pas immédiatement vérifiée, car rien ne nous permet d'assurer que $\lambda_{n+m}(a,c) \neq \lambda_n(a,b)\,\lambda_m(b,c), \forall$ (a, b, c). Si un tel cas se présente, on choisit la trajectoire qui contient le plus de pics. De cette propriété découle un algorithme récursif détaillé ci-après.

Construction des trajectoires

On considère deux paramètres, le premier est le nombre de trames futures que nous considérons pour la formation des trajectoires optimales noté n_f. On ne permet que des n-transitions de taille bornée. Le second paramètre est la taille maximale d'une n-transition, noté n_m (dans la suite, on considère $n_m = 3$). L'ensemble de pics Γ représente la zone prospective des partiels. Elle doit donc être au moins constituée de toutes les trames contenant les derniers pics attribués aux partiels jusqu'aux trames immédiatement atteignables. On a donc : $n_f \geq 2n_m - 1$ (cf. figure 3.9).

De manière à réduire le coût de calcul induit par la construction des trajectoires, on procède d'abord au pré-calcul des probabilités de transition pour chaque couple de pics contenu dans Γ qui engendrent une n-transition de taille inférieure à n_m.

En l'absence d'informations sur les différentes trajectoires possibles contenues dans Γ, chaque pic est placé dans un état initial en fonction de sa position. Un pic qui peut atteindre un pic de la trame d'indice supérieur au dernier indice de trame de Γ est dans un état de continuation (⊕). En effet, s'il ne trouve aucun pic avec lequel effectuer une liaison dans Γ, le partiel qui se prolonge sui-

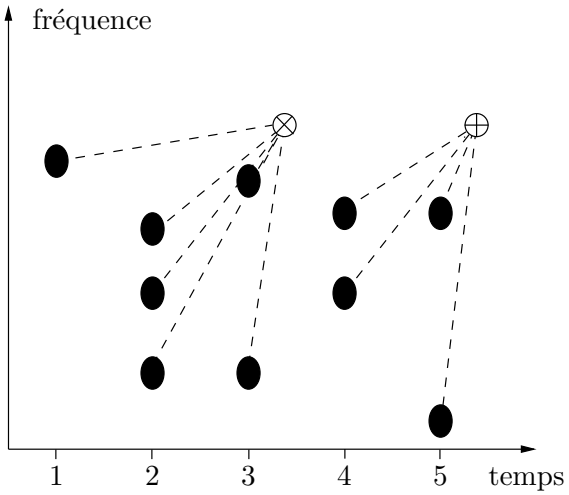


FIG. 3.10 – Initialisation des trajectoires de Γ vers un état de continuation $\oplus$ si le pic peut effectuer une n-transition vers une trame d'indice relatif supérieur à n_f ou vers un état de non continuation $\otimes$ dans le cas contraire ($n_m = 3$).

vant une telle trajectoire sera en mesure de se prolonger à l'extérieur de Γ, voir figure 3.10. Une probabilité initiale égale à celle d'une n-transition de variation de fréquence nulle est attribuée à ces pics, n étant la distance entre la trame du pic et la trame située immédiatement après Γ.

A contrario, les pics issus de trames d'indices inférieurs ne peuvent pas faire partie d'une trajectoire prospective menant à l'extérieur de Γ sans au préalable passer par des pics intermédiaires d'indices supérieurs contenu dans Γ. Sans information sur les trajectoires possibles, on les met dans un état de non continuation ($\otimes$) et on leur attribue une probabilité initiale très faible.

La construction des trajectoires proprement dite se déroule en remontant les indices de trames, voir figure 3.11. Pour tout pic p_3^i de la trame 3, on recherche de manière exhaustive une n-transition vers un pic d'une trame atteignable engendrant un poids maximal défini comme le produit de la probabilité de la 1-transition de p_3^i vers p_4^j et le poids associé à p_4^j. On met à jour le poids associé à p_3^i si cette transition est d'un poids supérieur à celui apporté par l'initialisation de p_3^i. L'algorithme se déroule ensuite de manière récursive jusqu'à la trame 1.

Certaines trajectoires passent par le même pic. Si deux partiels réservent ces trajectoires, on doit pouvoir décider laquelle des deux trajectoires est la plus probable pour déterminer quel partiel doit pouvoir réserver sa trajectoire en premier. La section suivante adresse ce problème en proposant une méthode de comparaison des trajectoires optimales.

Comparaison

Les trajectoires optimales n'ont pas forcément la même longueur et peuvent être décalées, c'est-à-dire ne pas avoir le même indice de trame de début ou de fin. On ne compare donc que leur partie commune. Si la partie commune est

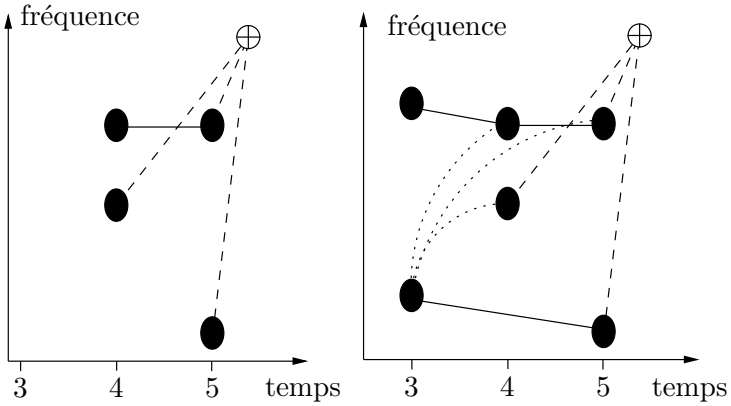


FIG. 3.11 – Construction récursive des trajectoires optimales de Γ. À gauche, premier pas de construction. À droite, deuxième pas de construction. Les arcs en pointillés représentent les trajectoires explorées avant de choisir celle de poids maximal.

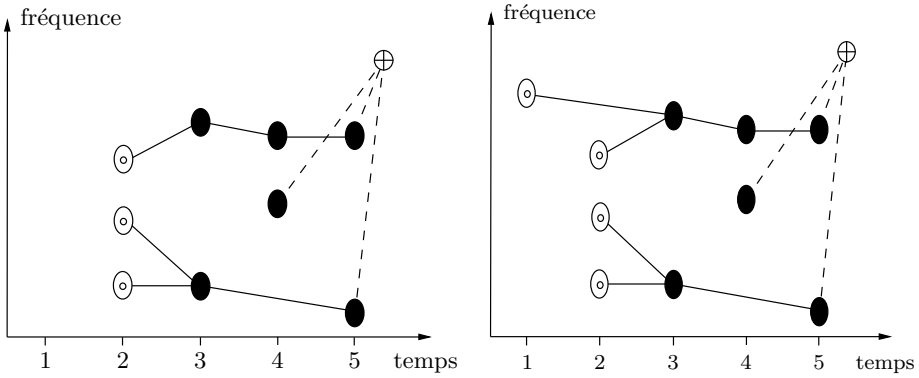


FIG. 3.12 – Construction récursive des trajectoires optimales de Γ. À gauche, troisième pas de construction. À droite, quatrième pas de construction.

identique, on choisit alors de façon arbitraire la trajectoire la plus longue, de même si la partie commune est vide.

Il reste à déterminer comment extraire la partie commune pour chacune des deux trajectoires. Si les positions de départ et de fin de la partie commune sont les mêmes pour les deux trajectoires, il suffit de comparer les produits de probabilités respectifs. Dans le cas contraire (voir figure 3.11), supposons que la trajectoire 1 ait pour première transition une m-transition et que la trajectoire 2 soit la plus longue et effectue une n-transition ($n > m$). Pour pouvoir comparer la partie commune des deux trajectoires, on doit réduire la taille de la n-transition de la trajectoire 2. On remplace dans le produit de probabilité associé au trajectoire $\lambda_n(a, b)$ par $\lambda_m(a, a + \frac{(b-a)m}{n})$, $\frac{bm}{n}$ étant la fréquence d'un pic interpolé ayant même index de trame que le premier pic de la trajectoire 1.

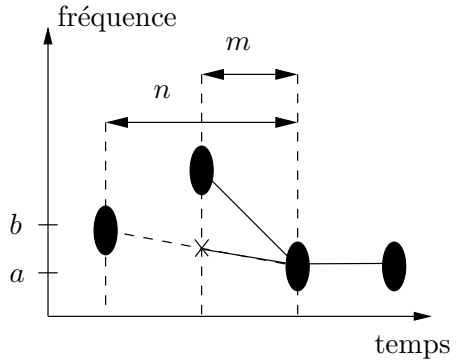


FIG. 3.13 – De manière à comparer les deux trajectoires, on doit calculer de la probabilité associée à la partie commune des deux trajectoires comparées. La croix montre la position du pic interpolé.

3.4.3 Algorithme

À chaque trame, l'algorithme de suivi de partiels consiste en l'itération de la procédure suivante :

- calcul des trajectoires optimales dans Γ ;
- tri des partiels qui sont prolongés par une trajectoire (ce tri par valeur décroissante est effectué en fonction du poids de la trajectoire qui prolonge le partiel) ;
- pour chaque partiel trié : si tous les pics de la trajectoire qui le prolonge sont disponibles, les pics de la trajectoire sont rendus indisponibles et le partiel est prolongé avec le premier pic de la trajectoire si ce pic est dans la trame courante.

On itère ce processus tant que des partiels ne sont pas encore prolongés et qu'il existe des trajectoires qui se terminent avec les queues de partiels.

Exemple

Soit Γ dans l'état présenté dans la figure 3.9. Après construction des trajectoires optimales, on obtient l'ensemble des trajectoires présenté à droite de la figure 3.11. La probabilité associée à chaque trajectoire étant d'autant plus élevée que la distorsion en fréquence est faible, les partiels P_1 et P_3 vont réserver leurs trajectoires comme le montre la figure 3.14(a). Comme les premiers pics de ces trajectoires sont dans la trame courante ici d'indice 3, ces pics font maintenant partie des partiels.

Les partiels P_2 et P_4 ayant élu des trajectoires passant par des pics déjà réservés dont le poids est plus faible , une nouvelle segmentation de Γ est effectuée. Le partiel P_4 ayant le meilleur poids associé réserve sa trajectoire (voir figure 3.14(b)) mais une liaison définitive n'est pas effectuée car le pic élu n'est pas dans la trame courante. Le partiel P_2 n'ayant pas pu réserver la trajectoire qu'il avait élu et n'ayant aucune alternative est déclaré mort, voir figure 3.14(c). Il ne reste aucun pic actif, on peut déplacer Γ d'une trame. alors dans la situation

Commentaires

On a présenté dans cette partie une approche probabiliste appliquée au suivi de partiel qui permet, par une probabilité de transition originale, la modélisation explicite des transitions entre pics de trames non adjacentes. Cette approche peut aisément s'adapter à des problématiques similaires dans des domaines différents où la gestion de l'absence sporadique d'informations est d'intérêt car la probabilité proposée se base sur peu de caractéristiques propres au modèle sinusoïdal à long terme. Cette généralité est à la fois une force et une faiblesse. En effet, cette méthode s'est montrée en pratique difficile à paramétrer. En particulier, l'influence d'une réduction de la complexité sur la qualité du suivi est très difficile à évaluer.

Nous nous sommes donc orientés dans la suite de nos recherches vers des méthodes qui se basent sur l'utilisation des contraintes relatives au modèle sinusoïdal, comme le caractère prédictible des évolutions des paramètres des partiels et l'absence théorique de hautes fréquences dans ces évolutions. Ces nouvelles méthodes conservent de plus la souplesse et l'efficacité de l'algorithme MAQ grâce à une structure générique présentée dans la section suivante.

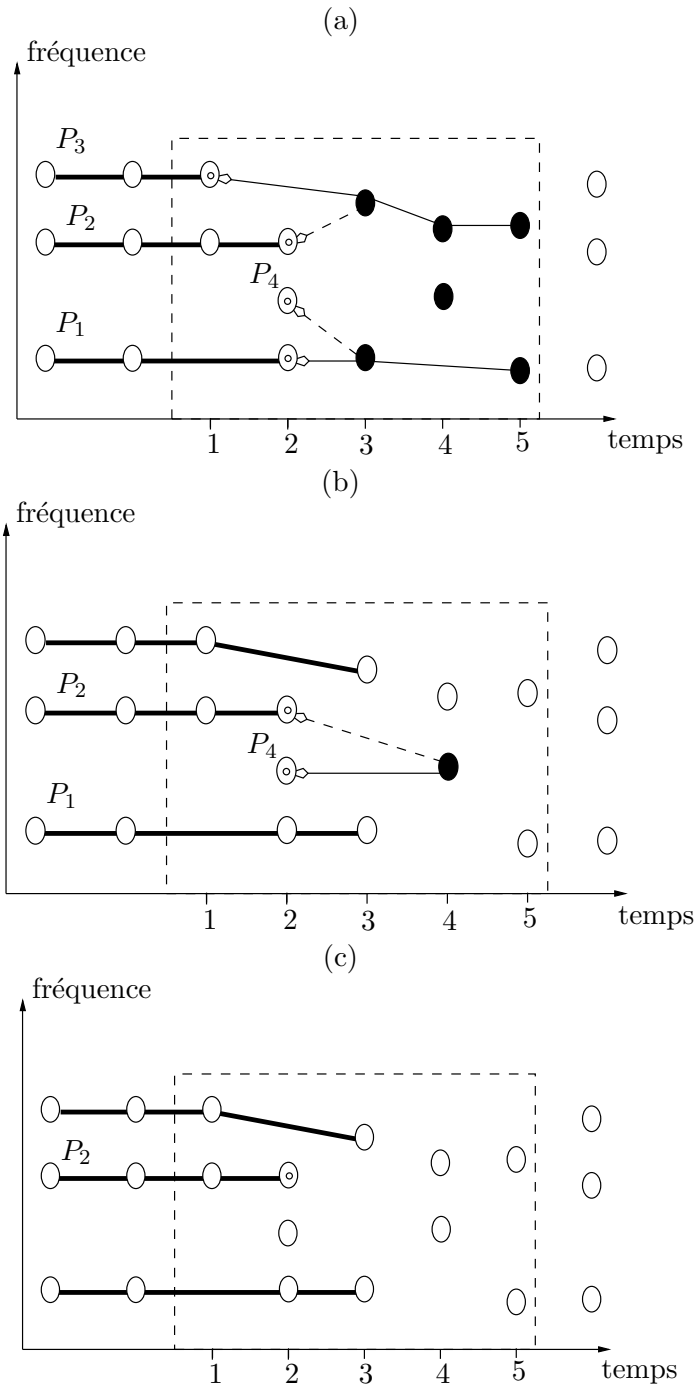


FIG. 3.14 – Trois phases du processus d'élection/confirmation. Seul les pics en noir sont des éléments actifs des trajectoires optimales, les pics en blanc étant soit déjà inclus dans des partiels ou des trajectoires réservés, soit à l'extérieur de Γ et donc non atteignables.

3.5 Algorithme de suivi de partiels générique

Les algorithmes présentés dans ce chapitre ont en commun plusieurs contraintes qui amènent une structure similaire. Tout d'abord, ces algorithmes sont destinés à pouvoir être utilisés en *streaming*. Les partiels sont traités de manière indépendante, aucune considération d'harmonicité ou de similarité d'évolution entre partiels n'est exploitée pour effectuer le suivi de partiels. Le lecteur est invité à se référer à [MB94, Ser97] pour des détails sur l'utilisation des contraintes d'harmonicité pour le suivi de partiels. Enfin, le principe d'allocation exclusive est appliqué, c'est-à-dire que un pic ne peut être alloué qu'à un seul partiel.

Mêmes si certains algorithmes sont amenés à considérer plusieurs trames dans le futur, l'allocation des pics aux partiels se fait trame par trame. Cette allocation peut se segmenter en trois parties :

- l'élection, où le meilleur prolongement du partiel est choisi ;
- l'ordonnancement, où les partiels sont triés par ordre de priorité ;
- la confirmation, où chaque partiel, par ordre de priorité, se prolonge avec le pic élu s'il est disponible.

3.5.1 Élection

L'élection consiste à déterminer quel pic est le meilleur prolongement d'un partiel donné. La fonction de coût utilisée pour répondre à ce problème est déterminante pour la qualité du suivi. Dans l'algorithme MAQ, la fonction de coût est la distance entre la fréquence du dernier pic du partiel et le pic considéré. D'autres fonctions plus élaborées sont proposées dans les sections suivantes. On se contentera ici d'étudier brièvement différentes caractéristiques des pics spectraux qui peuvent être utiles pour une fonction de coût pertinente à faible complexité.

La différence d'amplitude entre deux pics peut être utilisée au même titre que celle en fréquence même si les évolutions des partiels en amplitude sont plus chaotiques et moins pertinentes perceptivement.

Des distances provenant de différents plans peut être combinées. Comme les distances n'ont pas la même grandeur physique, il est proposé dans [DGR93b] de considérer ces distances comme des variables aléatoires gaussiennes de moyenne nulle et d'écarts types pouvant être estimés sur une large base de tests.

Lors d'un vibrato, les variations d'amplitude et de fréquence sont corrélées et peuvent être déphasées comme on peut le voir sur la figure 3.15. On peut exploiter cette corrélation pour proposer une distance d_e :

$$d_e(p_k^i, p_{k+1}^j) = \sqrt{(f_k^i - f_{k+1}^j)^2\,(a_k^i - a_{k+1}^j)^2} \tag{3.19}$$

où f_k^i et a_k^i sont la fréquence et l'amplitude du pic p_k^i. La distance qui en découle n'est pas une distance euclidienne et doit donc être utilisée dans un voisinage en fréquence restreint. Cette mesure, utilisée dans [LMR04a], est particulièrement performante lors de l'analyse de signaux modulés et bruités comme le violon, voir figure 3.16.

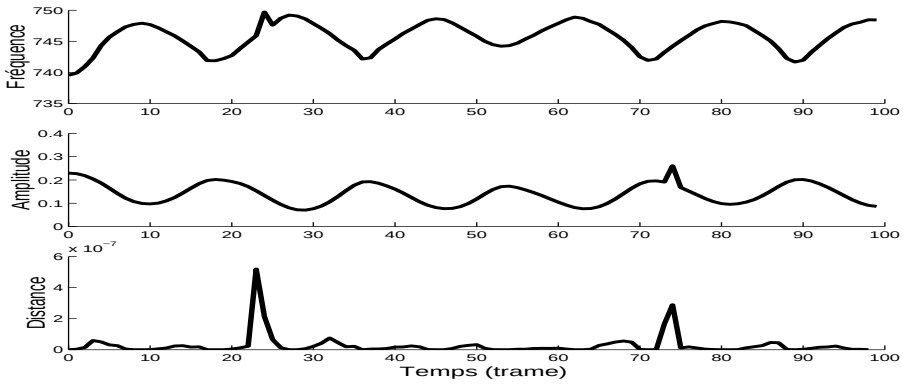


FIG. 3.15 – Fréquence (en haut), amplitude (au milieu) et la distance d_e correspondante (en bas) pour la première harmonique d'une note de saxophone modulée par un vibrato. La fréquence et l'amplitude sont déphasées. Les valeurs de fréquence et d'amplitude sont corrompues de manière artificielle respectivement aux trames d'index 25 et 75.

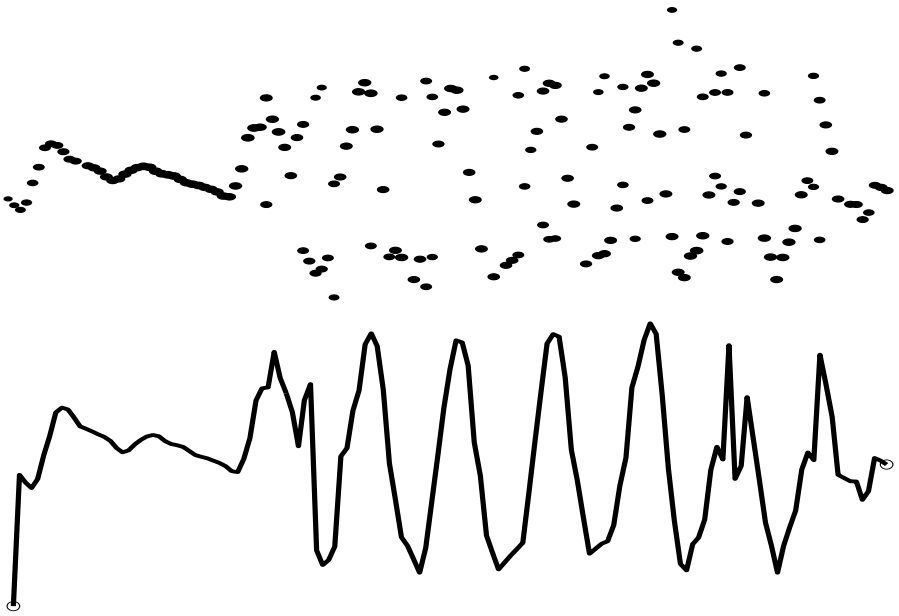


FIG. 3.16 – En haut, représentation à court terme d'un vibrato de violon. Les modulations et le frottement de l'archet génèrent de nombreux pics de bruit. En bas, partiel obtenu en minimisant la distance d_e.

3.5.2 Confirmation

Dans la phase de confirmation, le partiel se prolonge dans la trame active en utilisant le pic qu'il a élu lors de la phase d'élection si ce pic est toujours disponible. Par application du principe d'allocation exclusive, le pic ne peut alors plus être utilisé pour la prolongation d'un autre partiel.

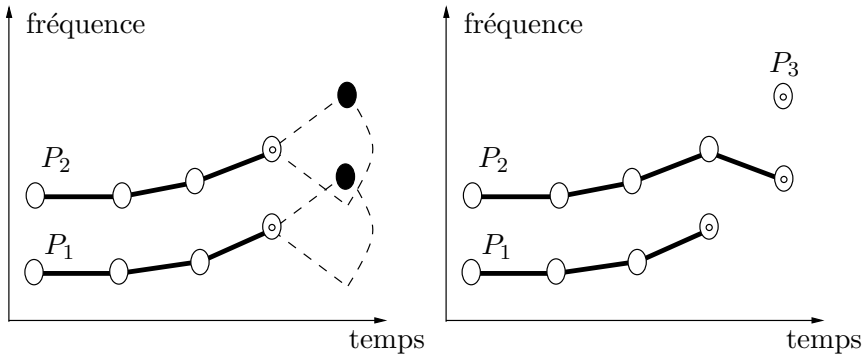


FIG. 3.17 – Cas particulier où l'algorithme MAQ est mis en défaut. Le pic disponible de fréquence la plus faible est plus proche de P_2 que de P_1. Le partiel P_1 ne peut donc se prolonger grâce à ce pic et meurt. Lorsque le partiel P_2 est actif, ce pic est toujours disponible. Ce partiel se prolonge donc en utilisant ce pic.

3.5.3 Ordonnancement

L'ordonnancement a pour but de déterminer dans quel ordre les partiels vont sélectionner les pics de la trame active. Comme chaque pic n'est alloué qu'à un seul partiel, cet ordonnancement permet de gérer de manière implicite la concurrence entre les partiels.

Comme les pics spectraux sont issus d'une DFT, l'algorithme MAQ utilise ce tri des pics par fréquence croissante implicite et est de ce fait particulièrement efficace en terme de complexité. Ceci est à propos lors de l'analyse de signaux monophoniques de voix car les partiels les plus audibles (de plus forte amplitude) sont situés dans les basses fréquences. Lors de l'analyse de signaux polyphoniques, il est souhaitable que l'ordonnancement soit plus flexible et puisse prendre en compte des fonctions complexes combinant par exemple l'amplitude du partiel, sa longueur (nombre de pics déjà insérés), etc.

De plus, la gestion de la concurrence dans l'algorithme MAQ est minimale. On doit s'assurer que le pic élu par le partiel actif (celui qui est considéré à une étape donnée de l'algorithme) ne peut pas être mieux utilisé par le partiel de fréquence supérieure. Dans le cas d'une modulation de fréquence, cet algorithme peut être mis en défaut. Dans la figure 3.17, le pic de fréquence la plus faible a sa fréquence plus proche de la fréquence du dernier pic inséré dans P_2 que celui de P_1. Le partiel P_1 ne peut donc se prolonger grâce à ce pic et meurt. Lorsque le partiel P_2 devient le partiel actif, ce pic est toujours disponible, P_2 se prolonge donc en utilisant ce pic.

Pour gagner en flexibilité et pallier ces deux problèmes, on propose d'effectuer un tri des partiels en fonction d'un indice de priorité qui peut être défini soit en utilisant uniquement des caractéristiques propres au partiel comme sa longueur où la fréquence du dernier pic utilisé, soit en utilisant des caractéristiques du couple (partiel, pic élu) comme la distance entre la fréquence du dernier pic utilisé et la fréquence du pic candidat.

Dans le premier cas, l'algorithme se déroule linéairement comme expliqué sur la figure 3.18(a). Tous les partiels sont triés par ordre de priorité décroissante. Puis chaque partiel élit un pic, par ordre de priorité. Le cas échéant, ce pic est utilisé pour prolonger le partiel, le partiel n'est plus actif et le pic n'est plus disponible. Dans le cas contraire, le partiel n'est pas actif pour cette trame et entre dans le mode zombie. Ce processus se répète jusqu'à ce qu'il n'y ait plus de partiels actifs. Les pics restant sont alors utilisés pour faire naître de nouveaux partiels.

Dans le deuxième cas, tous les partiels procèdent d'abord à une élection. Les partiels ayant élu un pic sont considérés comme actifs. Ces partiels sont triés en fonction de caractéristiques du couple formé par le partiel actif et le pic élu, par ordre décroissant. À une étape donnée du processus, si le pic est disponible, on procède alors à la confirmation et le pic n'est plus actif. Dans le cas contraire, une nouvelle phase d'élection est effectuée pour ce partiel. Si un pic est élu, le couple (partiel, pic) est replacé dans la liste de couples en fonction de sa priorité. Dans le cas contraire, le partiel n'est plus actif. Ce processus se répète jusqu'à ce qu'il n'y ait plus de partiels actifs. Les pics restant sont utilisés pour faire naître de nouveaux partiels.

3.5.4 Sélection des partiels par conformité au modèle

À cause de la présence de perturbations dans le signal, voir figure 3.1, certains partiels extraits ne correspondent pas à une composante sinusoïdale du signal analysé. On introduit ici des critères simples de sélection qui sont utilisés pour toutes les méthodes évaluées dans la section 3.8. Tout d'abord, un partiel est une composante à long terme et par conséquent ne peut avoir une durée de vie inférieure à un seuil donné. On considère en pratique que cette durée est comprise entre 50 ms et 200 ms en fonction du degré de détail désiré. Ensuite, son amplitude doit être suffisante. Pour cela, on s'assure qu'il existe un rapport suffisant entre l'énergie de ce partiel et l'énergie de tous les partiels présents durant sa phase d'activité. Formellement, soient n_1 et n_2 les indices de trames de la naissance et de la mort du partiel P_i, le critère C_i se définit comme suit :

$$A_i = \sum_{n=n_1}^{n_2} A_i(n) \tag{3.20}$$

$$B = \sum_{n=n_1}^{n_2} \sum_{j=1}^{m} A_j(n)\, e_j(n) \tag{3.21}$$

$$C_i = \frac{A_i}{B} \tag{3.22}$$

où m est le nombre de partiels et $e_j(n)$ vaut 1 si le partiel P_j est présent à la trame d'indice n et 0 sinon. Empiriquement, on considère que si C_i est inférieur à 0.005, le partiel P_i est rejeté.

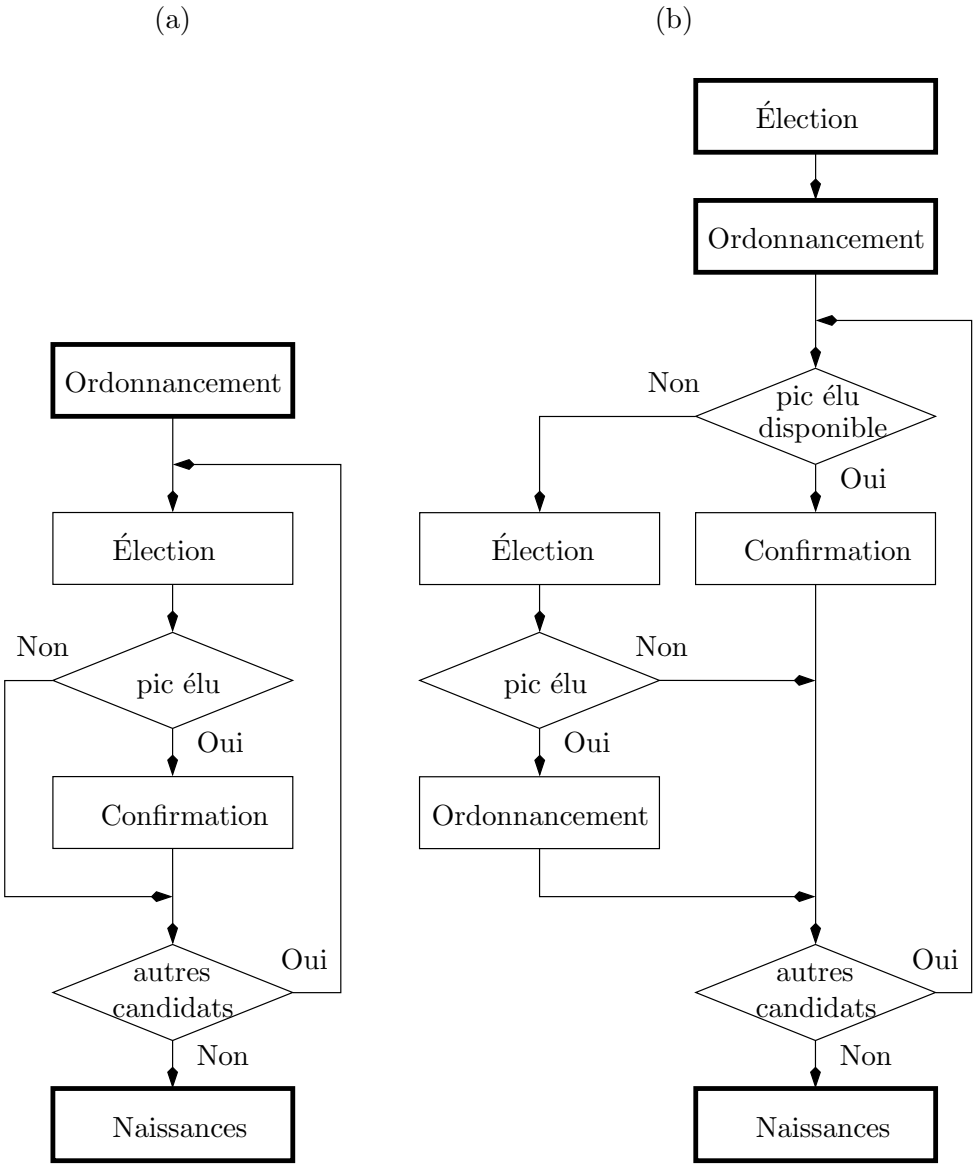


FIG. 3.18 – Schémas de principe d'une étape d'un algorithme de suivi de partiels générique en fonction du type d'ordonnancement. Les rectangles épais correspondent à des opérations globales sur l'ensemble des partiels. À gauche, l'ordonnancement des partiels est effectué en fonction de paramètres concernant seulement le partiel, comme sa longueur où la fréquence du dernier pic utilisé, etc. À droite, l'ordonnancement des partiels est effectué en fonction de caractéristiques concernant le couple (partiel, pic élu) comme la distance entre la fréquence du dernier pic utilisé et la fréquence du pic élu.

3.6 Suivi par prédiction linéaire

Il est raisonnable de penser que les paramètres d'un partiel ont une évolution prévisible dans le plan temps/fréquence et qu'une rupture brusque de cette propriété équivaut à la mort de ce partiel et à la naissance d'un nouveau partiel. En effet, un signal est prévisible lorsque l'on peut décrire avec précision un échantillon de ce signal en ne considérant que les échantillons qui l'ont précédé.

Ce modèle de description peut être contraint. On fixe alors certains paramètres du modèle par des études statistiques sur une base de partiels supposés bien suivis [SW98]. On peut au contraire n'avoir aucun *a priori* sur les caractéristiques de l'évolution de la fréquence des partiels que l'on considère. Les paramètres du modèle sont dans ce cas entièrement libres et déduits de l'évolution passée de la fréquence du partiel. C'est le cas de l'algorithme introduit dans cette section.

La phase d'élection consiste alors à prédire la fréquence mais aussi l'amplitude du pic suivant en fonction des paramètres des pics déjà insérés dans un partiel. On propose d'utiliser la méthode de la prédiction linéaire (LP) présentée dans une première partie pour effectuer cette prédiction. Un pic dont les paramètres sont proches de ces prédictions est ensuite choisi grâce à une méthode explicitée dans la suite.

3.6.1 Prédiction linéaire

Dans le formalisme de la prédiction linéaire, on prédit l'évolution d'un système que l'on suppose autorégressif (AR) grâce à un certain nombre d'observations. L'échantillon courant $x(n)$ est approximé par une combinaison linéaire des observations (les échantillons passés produits par le système). L'échantillon prédit $\hat{x}(n)$ est calculé par filtrage à réponse impulsionnelle finie (FIR) des d dernières observations :

$$\hat{x}(n) = \sum_{i=1}^{d} a(i)\, x(n-i) \tag{3.23}$$

Soient N observations consécutives, on veut calculer les coefficients $a(i)$ qui amènent à la meilleure approximation du signal $x(n)$ au sens des moindres carrés :

$$E = \sum_{n=0}^{N-1} (x(n) - \hat{x}(n))^2 \tag{3.24}$$

L'estimation des coefficients $a(i)$ nécessite quelques connaissances *a priori* du système que l'on souhaite approximer. Tout d'abord, il est utile d'avoir un ordre de grandeur de l'ordre du modèle d. Ensuite, l'ordre de grandeur de N et surtout le rapport entre N et d sont importants pour le choix de la bonne méthode de minimisation. Dans cette section, trois méthodes d'estimation des coefficients $a(i)$ sont présentées :

- la méthode d'autocorrélation ;

- la méthode de covariance ;
- la méthode de Burg.

Les deux premières méthodes sont détaillées dans [Mak75]. La dernière, dite de Burg [Kay88, KAZ00], est présentée autant du point de vue théorique que du point de vue de l'implantation.

Méthode d'autocorrélation

La méthode d'autocorrélation minimise l'erreur quadratique de prédiction avant sur un support supposé infini, c'est-à-dire :

$$\epsilon_{\infty}^{f} = \frac{1}{N} \sum_{n=-\infty}^{\infty} |x(n) - \hat{x}(n)|^2 \tag{3.25}$$

En pratique, le nombre d'observations est fini. Les échantillons nécessaires à la minimisation qui ne sont pas observés sont fixés à une valeur nulle car cette méthode est définie pour des signaux centrés. Les observations sont ensuite fenêtrées pour une fenêtre de type trigonométrique calculée grâce à l'équation 2.14 pour minimiser les discontinuités aux bornes de l'intervalle d'observation. Le calcul les coefficients $a(i)$ en minimisant ϵ_{∞}^{f} conduit à la résolution d'un système très régulier d'équations normales mettant en jeu une matrice dite de Toeplitz. Cette régularité permet d'utiliser l'algorithme très efficace dit de Levinson-Durbin [Kay88] pour effectuer cette résolution.

Méthode de covariance

La méthode de covariance effectue la minimisation de l'erreur quadratique de prédiction avant sur un support fini :

$$\epsilon^{f} = \frac{1}{(N-d)} \sum_{n=d}^{N-1} |x(n) - \hat{x}(n)|^2 \tag{3.26}$$

où $\hat{x}(n)$ est l'estimation calculée grâce à l'équation 3.23. Comme le support est fini, les observations ne sont pas fenêtrées. Cette méthode est donc utile pour l'estimation des coefficients $a(i)$ avec un nombre d'observations faible. Malheureusement, cette méthode peut mener à des filtres qui ne sont pas à phase minimale (les pôles estimés ne sont pas garantis d'être dans le cercle unité : $a(i) < 1 \ \forall i$). Ce problème est négligeable pour des applications de type estimation spectrale où cette méthode se trouve être particulièrement appropriée [KKZS03]. Ce phénomène est par contre particulièrement gênant pour la prédiction car si le filtre induit par les coefficients $a(i)$ est instable, l'intervalle de variation de la prédiction n'est plus borné.

Méthode de Burg

Notons respectivement $e_k^f(n)$ et $e_k^b(n)$ les erreurs de prédiction avant et arrière pour un ordre donné k :

$$e_k^f(n) = x(n) - \sum_{i=1}^{k} a_k(i)x(n-i) \tag{3.27}$$

$$e_k^b(n) = x(n-k) + \sum_{i=1}^{k} a_k(i)x(n-k+i) \tag{3.28}$$

La méthode de Burg [Bur75] minimise la moyenne des erreurs quadratiques avant et arrière sur un support fini et ceci de manière récursive. Pour obtenir $a_k(i)$, on minimise :

$$\epsilon_k = \frac{1}{2}(\epsilon_k^f + \epsilon_k^p) \tag{3.29}$$

où

$$\epsilon_k^f = \frac{1}{(N-k)} \sum_{n=k}^{N-1} |e_k^f(n)|^2 \tag{3.30}$$

$$\epsilon_k^b = \frac{1}{(N-k)} \sum_{n=k}^{N-1} |e_k^b(n)|^2 \tag{3.31}$$

et

$$a_k(i) = \left\{ \begin{matrix} a_{k-1}(i) + r_k a_{k-1}(k-i) \text{ pour } i = 1, 2, ..., k-1 \\ r_k \text{ pour } i = k \end{matrix} \right\} \tag{3.32}$$

où r_k est appelé le coefficient de réflexion d'ordre k. En substituant l'équation 3.32 dans les équations 3.30 et 3.31, on trouve une formulation récursive des erreurs avant et arrière :

$$e_k^f(n) = e_{k-1}^f(n) + r_k e_{k-1}^b(n-1) \tag{3.33}$$

$$e_k^b(n) = e_{k-1}^b(n-1) + r_k e_{k-1}^f(n) \tag{3.34}$$

où

$$e_0^f(n) = e_0^b(n) = x(n) \tag{3.35}$$

Pour trouver r_k, on dérive l'énergie de l'erreur de prédiction d'ordre k en fonction de r_k. En fixant cette dérivée à zéro (pour une erreur minimale), on obtient :

$$r_k = \frac{-2\sum_{n=k}^{N-1} e_{k-1}^f(n) e_{k-1}^b(n-1)}{\sum_{n=k}^{N-1} |e_{k-1}^f(n)|^2 + |e_{k-1}^b(n-1)|^2} \tag{3.36}$$

La méthode de Burg combine les avantages des deux méthodes précédentes. Comme la méthode d'autocorrélation, la méthode de Burg est à phase minimale ($\forall i, a(i) < 1$). En effet, l'expression des r_k est de la forme :

$$r_k = \frac{2bc}{|b|^2 + |c|^2} \tag{3.37}$$

où b et c sont des vecteurs. Par l'inégalité de Schwarz, il est vérifié que les r_k ont une amplitude plus faible que 1. Et comme la méthode de covariance, la

méthode de Burg estime les coefficients $a(i)$ sur un support fini. L'algorithme suivant calcule le vecteur a en utilisant la méthode de Burg à un ordre d, en utilisant le vecteur x comme vecteur d'observations :

ef $\leftarrow x$
eb $\leftarrow x$
$a \leftarrow 1$
pour m de 0 à $d-1$ faire
  efp $\leftarrow$ ef sans son premier élément
  ebp $\leftarrow$ eb sans son premier élément
  $k \leftarrow -2\text{ebp} \cdot \text{efp}/(\text{ebp} \cdot \text{ebp} + \text{efp} \cdot \text{efp})$
  ef $\leftarrow$ efp $+ k$ ebp
  eb $\leftarrow$ ebp $+ k$ efp
  $a \leftarrow (a[0], a[1], \cdots, a[m], 0) + k(0, a[m], a[m-1], \cdots, a[0])$
fin pour

3.6.2 Prédiction de l'évolution des paramètres des partiels

L'algorithme MAQ utilise un prédicteur constant en fréquence :

$$\hat{f}(n+k) = f(n) \tag{3.38}$$

où k est la distance entre la valeur prédite et la dernière observation. Ce type de prédicteur pose des problèmes lors de l'analyse de signaux modulés en fréquence et de signaux polyphoniques dans lesquels les partiels se croisent. Il est proposé dans [DGR93b] d'utiliser un prédicteur qui considère le moment de la fréquence entre les deux derniers pics insérés du partiel :

$$\hat{f}(n+k) = f(n) + (f(n) - f(n-1)) \cdot k \tag{3.39}$$

Dans ces deux algorithmes, l'état zombie n'étant pas utilisé, k est toujours égal à 1.

Comme le montre une littérature conséquente [JVV86, Vas88a, Vas88b, Vas92, Val91, Ett96, KKS01, KR02], la prédiction linéaire peut être utilisée avec succès pour extrapoler les signaux audionumériques. On montre ici que la prédiction linéaire peut aussi être utilisée pour prédire les évolutions des paramètres de fréquence et d'amplitude des partiels, qui sont des signaux temporels, avec néanmoins une fréquence d'échantillonnage beaucoup plus faible. Ces prédictions de l'évolution des paramètres de fréquence et d'amplitude des partiels dans les trames futures permettent ainsi de mieux sélectionner les pics à insérer.

À une étape donnée du processus de suivi, l'évolution future des paramètres d'un partiel est déduite des paramètres des N derniers pics insérés dans le partiel. La méthode de Burg présentée dans la section 3.6.1 est utilisée pour estimer les coefficients AR en fonction des N observations de fréquence ou d'amplitude. Ces coefficients sont alors utilisés pour obtenir une prédiction de l'évolution du partiel en fréquence ou en amplitude grâce à l'équation 3.23. Si le nombre de pics est inférieur à N (au début du partiel), la méthode de la réflexion

décrite ci-après, est utilisée pour obtenir un nombre suffisant d'observations. Soit un vecteur x de taille M, les valeurs réfléchies à gauche sont :

$$x(-i) = 2\,x(0) - x(i) \tag{3.40}$$

Symétriquement les valeurs réfléchies à droite sont :

$$x(N-1+i) = 2\,x(N-1) - x(i) \tag{3.41}$$

Cette méthode conserve la continuité d'ordre zéro et un.

Choix de la méthode

Des trois méthodes décrites dans la section 3.6.1, la méthode de Burg apparaît comme le meilleur compromis. En effet, du fait du faible taux d'échantillonnage des paramètres des partiels, une minimisation de l'erreur sur un support fini apparaît indispensable. En raison de l'instabilité potentielle des vecteurs de prédiction calculés par la méthode de la covariance, cette méthode doit être évitée. Le choix se porte donc naturellement vers la méthode de Burg. Il est à noter que c'est précisément cette méthode qui a été retenue par Kauppinen pour l'extrapolation de signaux audionumériques.

L'expérience suivante confirme ces remarques. Les trois méthodes sont utilisées pour prédire l'échantillon $x(n)$ (trait plein sur la figure 3.19) considèrent $[x(n-13), \cdots, x(n-1)]$ comme observations. Le signal original est un signal quasi périodique avec un échantillon déplacé d'un facteur 1.01. La partie quasi périodique permet d'évaluer la précision du prédicteur et la partie non stationnaire permet d'apprécier la résistance de la méthode au bruit. Pour la méthode de l'autocorrélation, on a retranché aux observations la moyenne de ces observations avant d'effectuer la minimisation, voir section 3.6.1.

Comme on le voit sur la figure 3.19, la prédiction obtenue en utilisant la méthode de l'autocorrélation n'est pas satisfaisante avec un nombre réduit d'observations. La prédiction obtenue grâce à la méthode de la covariance est très précise dans les parties stationnaires mais diverge en cas de non stationnarité, voir trame 32 de la figure 3.19. Lors du suivi de partiels, cette instabilité peut amener à une mauvaise identification du pic successeur. La méthode de Burg semble être un bon compromis en terme de précision et de résistance au bruit. Cette méthode est donc celle qui sera retenue dans la suite pour le suivi de partiels par prédiction linéaire.

Choix des paramètres

L'ordre du modèle d et le nombre d'observations utilisées N sont d'une importance majeure pour la qualité de la prédiction. En premier lieu, des considérations théoriques permettent d'obtenir un ordre de grandeur pour ces deux paramètres en fonction des contraintes associées au suivi de partiels. Ces considérations sont ensuite affinées par une évaluation des performances d'un module de prédiction de l'évolution de la fréquence des partiels de signaux naturels en fonction de différentes valeurs de d et N. Ces performances sont

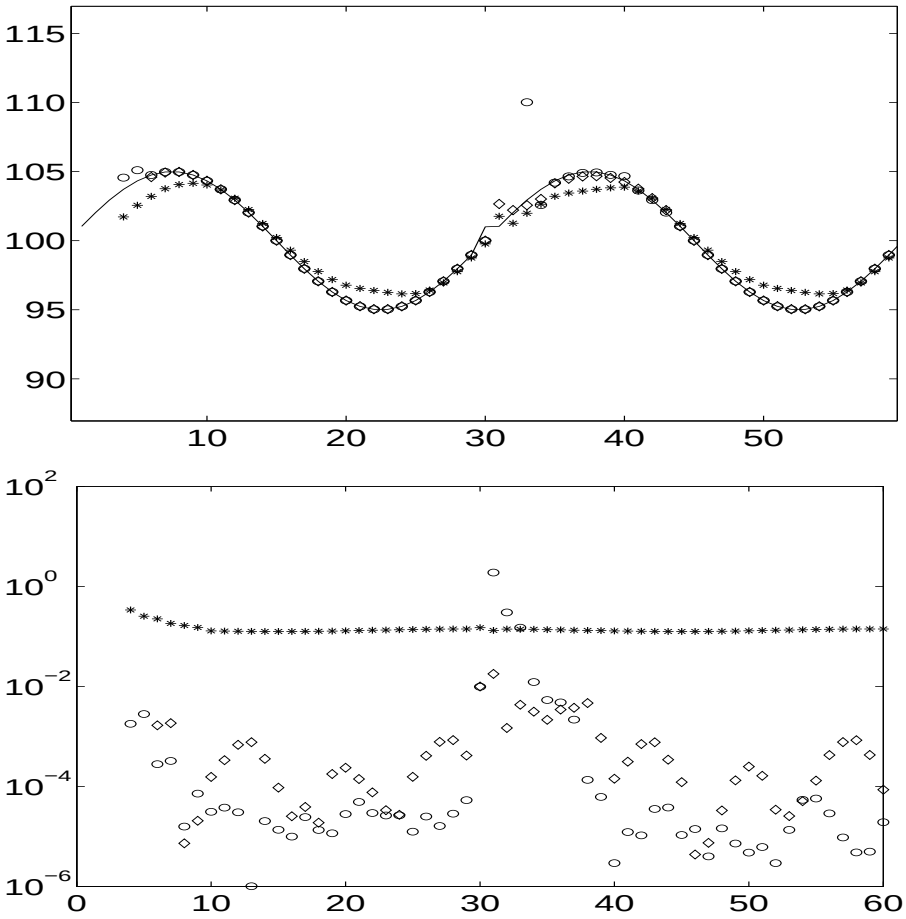


FIG. 3.19 – En haut, prédiction de l'échantillon $x(n)$ (trait plein) en fonction des observations $[x(n-13), \cdots, x(n-1)]$ pour les trois méthodes LP : autocorrélation ($*$), covariance ($\circ$) et Burg ($\diamond$). Le signal original est un signal quasi périodique avec un échantillon déplacé d'un facteur 1.01 (pour des raisons de clarté de la figure, des prédictions trop divergentes de la méthode de la covariance ne sont pas affichées). En bas, les erreurs de prédiction associées sont affichées sur une échelle logarithmique.

comparées à celles de prédicteurs simples présentés dans les équations 3.38, 3.39 et 3.41.

En ce qui concerne le paramètre de fréquence, les types d'évolutions classiques sont : une fréquence quasi constante, une croissance ou la décroissance exponentielle (dans le cas d'un portamento) et une évolution de type sinusoïdale (vibrato). L'ordre ne doit donc pas être inférieur à 2. Les tests expérimentaux détaillés dans la suite montrent qu'un ordre compris entre 2 et 8 est satisfaisant.

Le nombre d'échantillons doit être suffisamment large pour permettre l'identification de la périodicité du signal. Inversement, il doit être suffisamment faible pour ne pas être trop contraint par l'évolution passée du partiel et ainsi être plus réactif à un changement de dynamique. Le module d'analyse à court terme (voir section 2) utilise une transformée en fréquence glissante avec un pas d'avancement de 512 échantillons pour des signaux audionumériques échantillonnés à une qualité CD (44.1 kHz). Les fréquences et les amplitudes des partiels sont donc échantillonnées à $\approx$ 86 Hz. Comme un vibrato naturel a une fréquence proche de 4 Hz, au moins 20 échantillons sont nécessaires pour obtenir une période du vibrato. Des tests expérimentaux détaillés dans la suite montrent qu'un nombre d'observations compris entre 4 et 32 est satisfaisant.

Pour l'amplitude, les types d'évolutions attendus sont similaires à ceux de la fréquence. On observe néanmoins des variations beaucoup plus rapides. Il convient donc de considérer un nombre d'observations assez réduit. Des tests expérimentaux montrent qu'un nombre d'observations compris en 4 et 20 est satisfaisant. L'ordre étant fixé à la moitié du nombre d'observations.

Les tests expérimentaux présentés dans la suite utilisent les évolutions en fréquence de partiels extraits de signaux naturels de types différents, comme une note de saxophone avec vibrato, une guitare et des voix chantées. Sur les figures 3.1 et 3.2 sont présentées les erreurs moyennes et entre parenthèses les erreurs maximales pour différents prédicteurs. Dans la partie gauche de ces figures, les erreurs de prédiction sont calculées pour plusieurs prédicteurs simples et dans la partie droite, les erreurs de prédiction du prédicteur LP. Ces erreurs sont calculées pour différentes valeurs de k (distance en indices de trames entre la dernière observation et la valeur prédite). En ce qui concerne la partie dédiée au prédicteur LP, l'ordre du modèle d augmente de la gauche vers la droite. Pour chaque valeur de k, le nombre d'observations N considérées est [4, 8, 16, 32].

Le formalisme LP est efficace pour la prédiction de processus stationnnaires, sinusoïdaux et exponentiellement croissants ou décroissants. En ce qui concerne les évolutions constantes (comme celles rencontrées pour la fréquences des harmoniques d'une guitare), le prédicteur constant est suffisant. La prédiction de processus sinusoïdaux (comme le vibrato sur le plan temps/fréquence, voir figure 3.20) est très précise dans le cas du prédicteur LP à condition que le nombre d'observations soit suffisant. Les résultats sont alors très bons, tant en termes d'erreur moyenne que maximale, surtout quand k grandit. En ce qui concerne les évolutions exponentielles (portamento sur le plan temps/fréquence, voir figure 3.21), l'erreur moyenne est bonne en comparaison de celle des prédicteurs simples. L'erreur maximale reste en revanche semblable, à cause de la transition non stationnaire (voir figure 3.21 à la trame 85).

k	Constant	Linéaire	Réflexion	ordre : 2	4	6	8
1	0.35 (0.9)	0.16 (0.8)	0.16 (0.8)	0.20 (0.8)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	0.17 (0.6)	0.17 (0.6)	0.18 (0.7)	- (-)
	- (-)	- (-)	- (-)	0.16 (0.6)	0.14 (0.6)	0.14 (0.6)	0.14 (0.6)
	- (-)	- (-)	- (-)	0.16 (0.7)	0.13 (0.6)	0.13 (0.7)	0.12 (0.6)
2	0.69 (1.8)	0.42 (1.4)	0.51 (1.6)	0.48 (1.4)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	0.43 (1.3)	0.42 (1.3)	0.45 (1.6)	- (-)
	- (-)	- (-)	- (-)	0.41 (1.3)	0.35 (1.3)	0.34 (1.4)	0.34 (1.3)
	- (-)	- (-)	- (-)	0.41 (1.3)	0.32 (1.2)	0.29 (1.1)	0.28 (1.1)
3	1.01 (2.5)	0.76 (2.4)	1.00 (3.1)	0.85 (2.3)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	0.75 (2.3)	0.77 (2.3)	0.80 (2.7)	- (-)
	- (-)	- (-)	- (-)	0.72 (2.2)	0.62 (2.0)	0.57 (2.2)	0.57 (2.1)
	- (-)	- (-)	- (-)	0.71 (2.1)	0.52 (1.7)	0.46 (1.7)	0.43 (1.4)
4	1.31 (3.1)	1.18 (3.7)	1.63 (4.6)	1.29 (3.5)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	1.13 (3.5)	1.18 (3.6)	1.20 (4.1)	- (-)
	- (-)	- (-)	- (-)	1.09 (3.5)	0.92 (3.3)	0.83 (3.1)	0.81 (3.0)
	- (-)	- (-)	- (-)	1.06 (3.3)	0.77 (2.5)	0.65 (2.3)	0.58 (1.9)

TAB. 3.1 – Erreur moyenne (et maximale) pour différents prédicteurs pour la prédiction de l'évolution de la fréquence de partiels d'une note de **saxophone avec vibrato**. Les erreurs de prédiction sont calculées pour plusieurs prédicteurs simples (partie gauche) et le prédicteur LP (partie droite) pour différentes valeurs de k (distance entre la dernière observation et la valeur prédite). En ce qui concerne la partie dédiée au prédicteur LP, l'ordre du modèle d augmente de la gauche vers la droite. Pour chaque valeur de k, le nombre d'observations N considérées est [4, 8, 16, 32]. L'erreur de prédiction moyenne est inférieure à celle du meilleur prédicteur simple et cette amélioration est de plus en plus sensible quand k grandit.

Note sur le filtrage adaptatif

Dans l'algorithme proposé dans cette partie, de nouveaux coefficients AR nécessaires à la prédiction sont calculés pour chaque partiel et à chaque trame. Dans le formalisme du filtrage adaptatif [WS85, Hay91], à chaque insertion d'une nouvelle observation, l'erreur de prédiction entre cette nouvelle observation et la valeur prédite est exploitée pour mettre à jour les coefficients AR. Du point de vue de la complexité, ce formalisme est mieux adapté. Toutefois, toutes les méthodes testées ont un délai de convergence qui les rend inutilisables pour notre problème.

Des méthodes d'estimations récursives des coefficients LP existent [MC81, Bar81]. Cependant, les coefficients calculés sont différents de ceux estimés grâce aux méthodes non adaptatives du fait du fenêtrage utilisé pour minimiser l'influence des anciennes observations. L'implantation et la comparaison de ces méthodes récursives sont envisagées dans un proche avenir.

3.6.3 Algorithme

L'algorithme de suivi par prédiction linéaire des paramètres de fréquence et d'amplitude [LMRR03, LMR04b] utilise la structure algorithmique présentée dans la section précédente. La procédure d'élection se déroule comme suit. Soit un partiel actif à la trame n, les valeurs prédites en fréquence $\hat{f}(n+1)$ et en amplitude $\hat{a}(n+1)$ sont calculées par la méthode LP. La prédiction en fréquence permet d'effectuer une première sélection. Tous les pics dont la fréquence est distante de moins de Δ_f de $\hat{f}(n+1)$ sont sélectionnés, voir figure 3.22. Parmi ces candidats, le pic ayant l'amplitude la plus proche de celle prédite est élu.

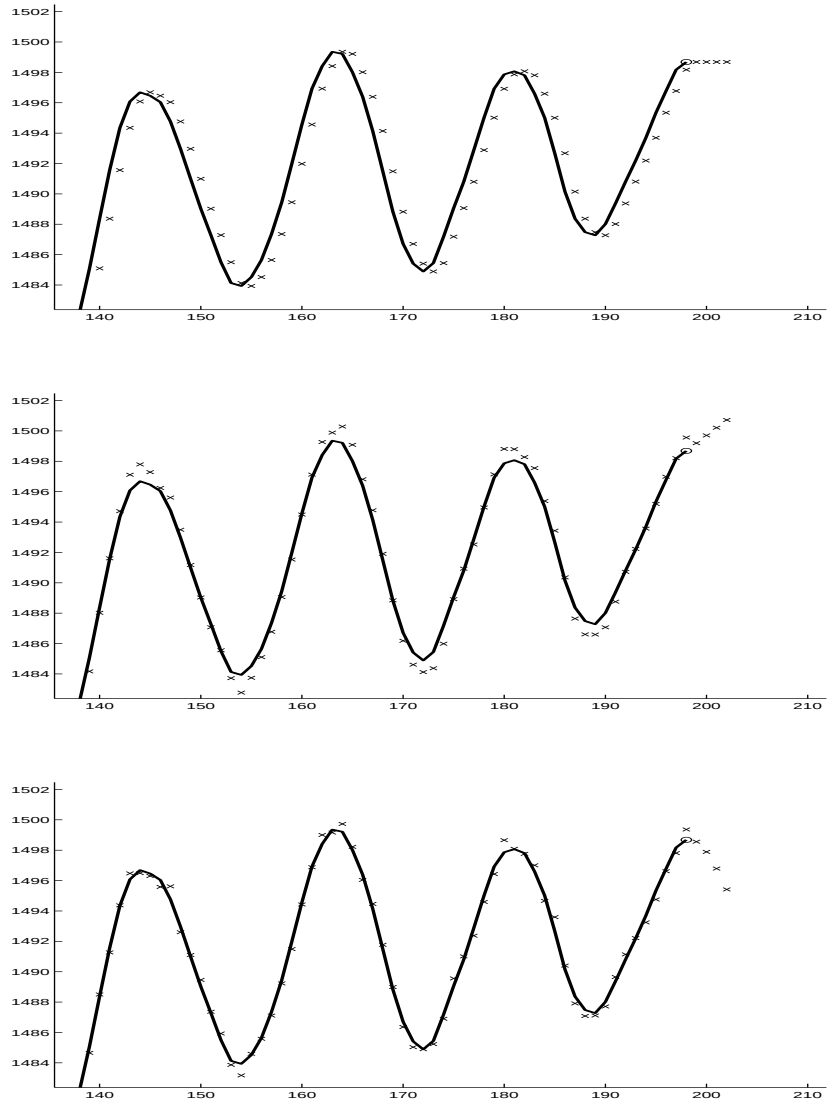


FIG. 3.20 – Évolutions de différents prédicteurs : constant (en haut), linéaire (au milieu), LP (en bas) pour un partiel de **saxophone**. L'ordre du prédicteur LP est fixé à 6 et 20 observations ont été utilisées pour estimer les coefficients de ce prédicteur. Les fréquences des partiels sont affichées avec des lignes et la fin d'un partiel est représentée par un cercle. Les fréquences prédites, marquées par des croix, sont calculées en fonction des dernières fréquences observées. Pour montrer la capacité d'un prédicteur à prédire à long terme l'évolution de la fréquence du partiel (situation commune par exemple dans le cas de pics manquants), les valeurs prédites sont affichées après la mort du partiel.

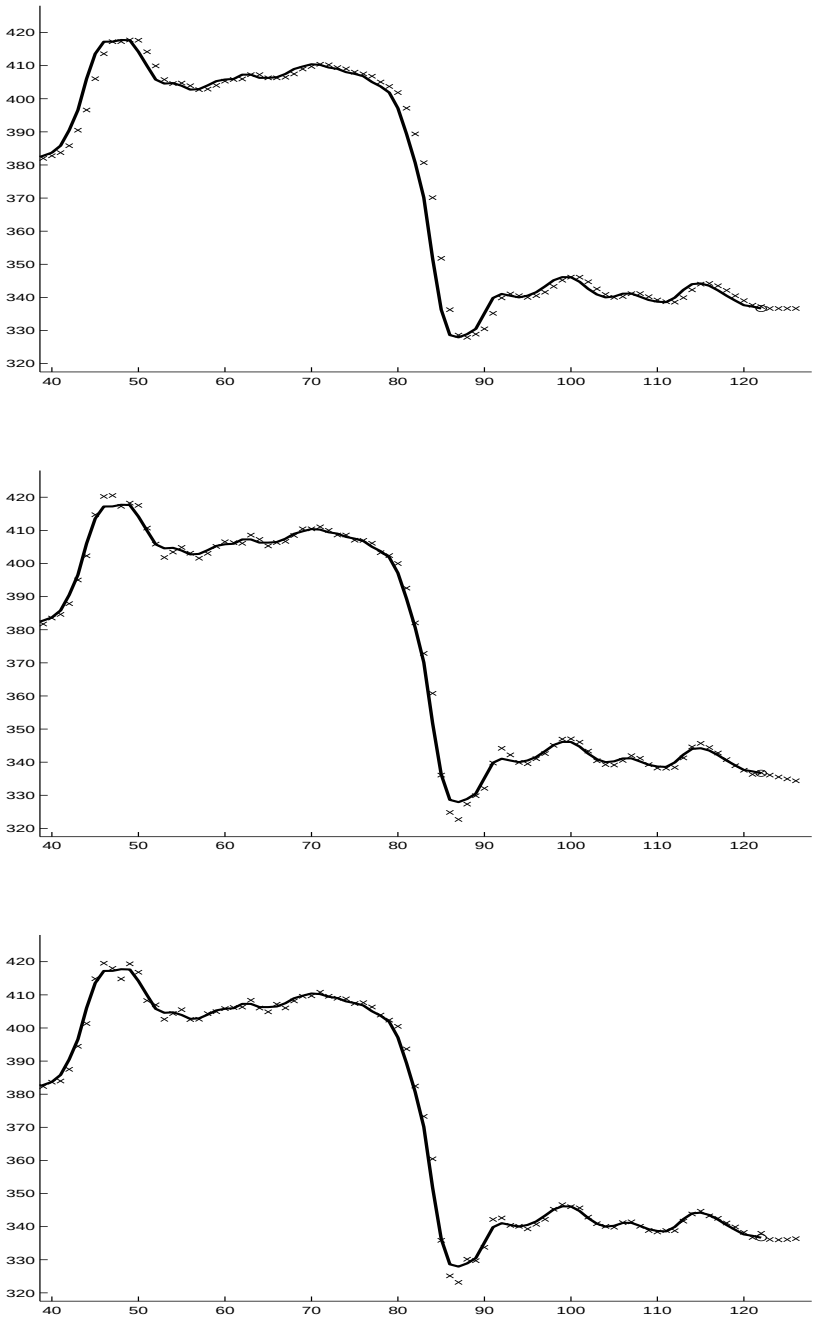


FIG. 3.21 – Évolutions de différents prédicteurs : constant (en haut), linéaire (au milieu), LP (en bas) pour un partiel de **voix chantée**. L'ordre du prédicteur LP est fixé à 6 et 20 observations ont été utilisées pour estimer ce prédicteur.

k	Constant	Linéaire	Réflexion	ordre : 2	4	6	8
1	2.69 (26.3)	1.16 (10.6)	1.16 (10.6)	1.39 (10.1)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	1.22 (10.0)	0.97 (12.2)	1.09 (11.5)	- (-)
	- (-)	- (-)	- (-)	1.19 (9.8)	0.86 (12.4)	0.89 (11.4)	0.92 (11.3)
	- (-)	- (-)	- (-)	1.17 (9.6)	0.83 (12.4)	0.84 (11.1)	0.86 (11.2)
2	5.32 (44.9)	3.19 (30.6)	3.99 (36.8)	3.66 (29.8)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	3.24 (28.7)	2.67 (35.1)	2.98 (33.1)	- (-)
	- (-)	- (-)	- (-)	3.17 (27.7)	2.35 (35.2)	2.35 (33.3)	2.43 (33.9)
	- (-)	- (-)	- (-)	3.12 (26.7)	2.24 (34.4)	2.24 (32.5)	2.26 (33.0)
3	7.81 (61.2)	6.01 (56.8)	8.06 (56.8)	6.61 (55.2)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	5.88 (53.0)	4.96 (62.2)	5.39 (61.3)	- (-)
	- (-)	- (-)	- (-)	5.76 (51.0)	4.33 (61.8)	4.21 (60.8)	4.32 (61.6)
	- (-)	- (-)	- (-)	5.68 (49.1)	4.10 (59.2)	4.00 (57.9)	3.99 (58.9)
4	10.17 (74.9)	9.45 (84.5)	12.76 (72.7)	10.02 (81.6)	- (-)	- (-)	- (-)
	- (-)	- (-)	- (-)	8.91 (78.0)	7.60 (88.6)	8.18 (86.1)	- (-)
	- (-)	- (-)	- (-)	8.77 (74.9)	6.64 (86.8)	6.40 (85.1)	6.51 (85.5)
	- (-)	- (-)	- (-)	8.67 (71.9)	6.29 (81.5)	6.07 (79.6)	6.00 (80.3)

TAB. 3.2 – Erreur moyenne (et maximale) pour différents prédicteurs pour la prédiction de l'évolution de la fréquence de partiels d'une **voix chantée**. Les erreurs de prédiction sont calculées pour plusieurs prédicteurs simples (partie gauche) et le prédicteur LP (partie droite) pour différentes valeurs de k (distance entre la dernière observation et la valeur prédite). En ce qui concerne la partie dédiée au prédicteur LP, l'ordre du modèle d augmente de la gauche vers la droite. Pour chaque valeur de k, le nombre d'observations N considérées est [4, 8, 16, 32]. L'erreur de prédiction moyenne est inférieure à celle du meilleur prédicteur simple et cette amélioration est de plus en plus sensible quand k grandit. L'erreur maximale reste néanmoins comparable à cause d'une transition non-stationnaire (voir 3.21 à la trame 85).

Le seuil Δ_f peut être le même pour tous les partiels. Une prédiction affinée permet alors de réduire ce seuil de tolérance par rapport à celui utilisé dans l'algorithme MAQ (une valeur convenable est approximativement 25 Hz.). Ce seuil peut aussi être fixé en fonction du gain de prédiction obtenu pour la trame précédente, comme proposé dans [RMC96].

Les partiels sont ordonnés selon leur amplitude et le mode zombie ne peut être utilisé que pour un nombre donné de trames consécutives. Ce paramètre dépend du pas d'avancement. Pour la configuration décrite précédemment, 4 est une valeur appropriée. En effet, le partiel ne peut se prolonger artificiellement au-delà de 20 ms sans engendrer d'artefacts audibles. Dans le cas où ce mode est utilisé, les valeurs prédites par la méthode LP peuvent être utilisées pour interpoler les valeurs de fréquence et d'amplitude des pics manquants. Dans le cas où les paramètres de phases sont nécessaires, l'algorithme d'interpolation de phase présenté dans la section 4.4 est utilisé.

Conclusion

On a proposé dans cette section d'utiliser la prédiction linéaire pour prédire et modéliser les évolutions de la fréquence et de l'amplitude des partiels. Ce formalisme est ici utilisé pour le suivi de partiels mais est aussi d'intérêt pour l'interpolation de partiels comme cela sera montré dans le chapitre 4.

L'utilisation de la prédiction linéaire permet de prédire l'évolution des partiels dans les trames futures. Comme les évolutions ne sont pas parfaitement prédictibles, on tolère une déviation par rapport à la valeur prédite en utilisant

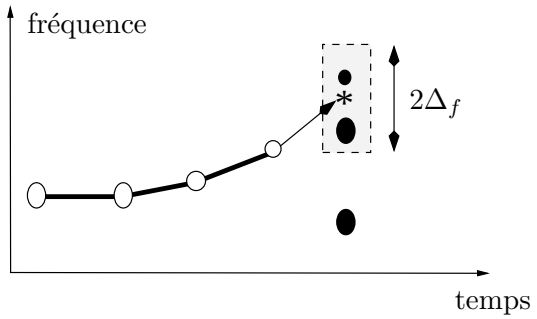


FIG. 3.22 – Étape d'élection pour un partiel en utilisant la prédiction linéaire des paramètres de fréquence et d'amplitude. Tous les pics dont la fréquence est distante de moins de Δ_f de la fréquence prédite (étoile) sont sélectionnés. On choisit parmi ces pics celui qui a l'amplitude la plus proche de celle prédite (pic de fréquence la plus élevée sur la figure).

un seuil Δ_f. Cette tolérance est plus faible que celle de l'algorithme MAQ grâce à une prédiction plus fine.

On conserve donc le principe d'un seuil de tolérance. Malheureusement, la diversité des signaux, les relations d'harmonicité qui amènent à des variation faibles dans les basses fréquences et des fortes variations en hautes fréquences font qu'il est très difficile de fixer un seuil de variation aussi faible que désiré. La section suivante propose un nouvel algorithme de suivi de partiel qui exploite la prédiction des paramètres et introduit un critère additionnel qui permet de s'assurer que le prolongement possible pour un partiel donné n'engendre pas de hautes fréquences dans les vecteurs de fréquence et d'amplitude du partiel.

3.7 Suivi par analyse fréquentielle des évolutions des partiels

L'algorithme de suivi introduit dans cette section utilise une mesure de distorsion originale qui nous permet de juger si un prolongement possible du partiel dans les trames futures n'engendre pas de discontinuités audibles. Cette mesure de distorsion qui exploite une contrainte inhérente au modèle sinusoïdal à long terme permet de s'affranchir de tout seuil de variation tout en apportant une garantie que les paramètres des partiels évoluent lentement.

3.7.1 Contraintes sur l'évolution des partiels

Un modèle sinusoïdal à long terme représente un signal par une somme d'oscillateurs quasi sinusoïdaux dont les paramètres doivent évoluer lentement au cours du temps, comme défini dans la section 1.2.2. En effet, un changement brusque des paramètres engendre une rupture de la continuité du signal. Le cas échéant, la distorsion peut devenir audible, la modélisation choisie n'est alors plus pertinente car elle ne correspond plus à la perception.

Comme cela a été évoqué précédemment, il est difficile d'établir dans le domaine temporel un procédé général qui permet d'identifier les variations brusques du signal. En revanche, la mise en place d'un tel procédé est plus aisé dans le domaine spectral, car les discontinuités du signal se manifestent par une énergie conséquente dans les hautes fréquences (HF). Pour que les paramètres évoluent lentement, on considère que le spectre de ces paramètres de contrôle ne doit pas présenter d'énergie notable pour une fréquence supérieure à 20 Hz, comme exposé dans la section 1.2.2.

L'expérience suivante montre que l'analyse des propriétés spectrales des paramètres permet de déterminer si un partiel a des paramètres de contrôle acceptables du point de vue du modèle. En haut de la figure 3.23 sont affichées les évolutions en fréquence de trois partiels. Le premier est une harmonique d'une note de saxophone. Le second est une harmonique de la même note avec une discontinuité et le troisième est un partiel extrait par erreur d'un signal de bruit blanc. Les spectres correspondants à ces évolutions en fréquence sont calculés grâce à une DFT et affichés en bas de la figure 3.23. Les signaux analysés sont préalablement fenêtrés par une fenêtre de Hann. À cause du vibrato, l'évolution en fréquence de l'harmonique correctement suivie a un spectre avec un niveau d'énergie élevé en basse fréquence. Par contre, au-dessus de 20 Hz, l'énergie est inférieure à -30 dB. Ce n'est pas le cas pour l'harmonique avec une discontinuité locale pour laquelle le niveau d'énergie dans les hautes fréquences est très élevé (-10 dB). L'évolution en fréquence du partiel extrait par erreur du bruit a un spectre avec un niveau d'énergie en haute fréquence plus faible (-20 dB), mais reste nettement plus élevé que pour le cas de l'harmonique correctement suivie. Ces deux derniers cas sont des exemples de ce qu'un module de suivi doit éviter. Le premier cas apparaît lorsque qu'un partiel suit une composante sinusoïdale puis change brusquement de composante ou se prolonge localement

avec un pic de bruit. Le second cas se produit lorsque le module de suivi relie des pics spectraux appartenant au bruit.

Il est donc possible de discriminer les évolutions lentes des autres (partiels de bruit, partiels avec une discontinuité locale) en considérant une estimation spectrale de l'évolution de la fréquence ou de l'amplitude du partiel. Toutefois, la suppression des "mauvais" partiels *a posteriori* peut amener à une représentation à long terme incomplète. Les partiels de bruit seront supprimés mais aussi les partiels avec une ou plusieurs discontinuités locales. Pour extraire des partiels qui sont conformes au modèle, l'estimation du contenu HF doit être opérée durant le suivi.

Le principe de l'algorithme présenté dans cette section est de considérer plusieurs petites trajectoires vraisemblables qui prolongent le partiel dans les trames futures. Pour chacune de ces trajectoires, l'analyse du contenu HF engendré par l'ajout de cette trajectoire au partiel permet alors de déterminer si un tel prolongement permet au partiel de conserver des paramètres de contrôles valides.

3.7.2 Estimation du contenu haute fréquence

Durant l'élection, on estime le contenu HF induit par l'insertion d'un pic candidat. Le taux d'échantillonnage faible ($\approx$ 100 Hz) des paramètres des partiels apporte une forte contrainte. Pour des raisons de localité et de complexité, l'estimation du contenu HF doit être effectuée avec un nombre très faible d'observations ($\approx$ 10). On propose d'utiliser un filtre passe-haut elliptique de faible délai pour estimer le contenu HF [PB87]. Les paramètres du filtre comme la fréquence de coupure et l'ordre dépendent du taux d'échantillonnage. Pour une fréquence d'échantillonnage proche de 100 Hz, un filtre d'ordre 4 avec une fréquence de coupure normalisée de 0.5 est satisfaisant. Une implantation efficace de ce filtre peut être faite grâce à des cellules à réponse impulsionnelle infinie (IIR) d'ordre 2 avec les coefficients suivants :

$$\begin{pmatrix} 1 & 0.2274 & 0 & 1 & -0.2346 & 0 \\ 1 & 0.1673 & -0.0137 & 1 & -1.2898 & 0.4076 \\ 1 & -0.3951 & 0.0201 & 1 & -2.0762 & 1.0762 \end{pmatrix}$$

À la naissance de chaque partiel, deux filtres sont initialisés, l'un destiné au filtrage du contenu HF des fréquences du partiel et l'autre est destiné au filtrage du contenu HF des amplitudes de ce même partiel. À chaque prolongement du partiel, on utilise pour mettre à jour les mémoires de ces filtres les paramètres du pic inséré auxquels on a soustrait la valeur de fréquence/amplitude du premier pic inséré. Lors de l'évaluation des trajectoires possibles, les mémoires des deux filtres sont alors utilisées en l'état, sans être mises à jour.

3.7.3 Étape d'élection

L'étape d'élection consiste à déterminer si un pic de la trame active est une prolongation satisfaisante pour un partiel donné. Au lieu de considérer les pics d'une seule trame future, on considère des trajectoires dans un nombre donné

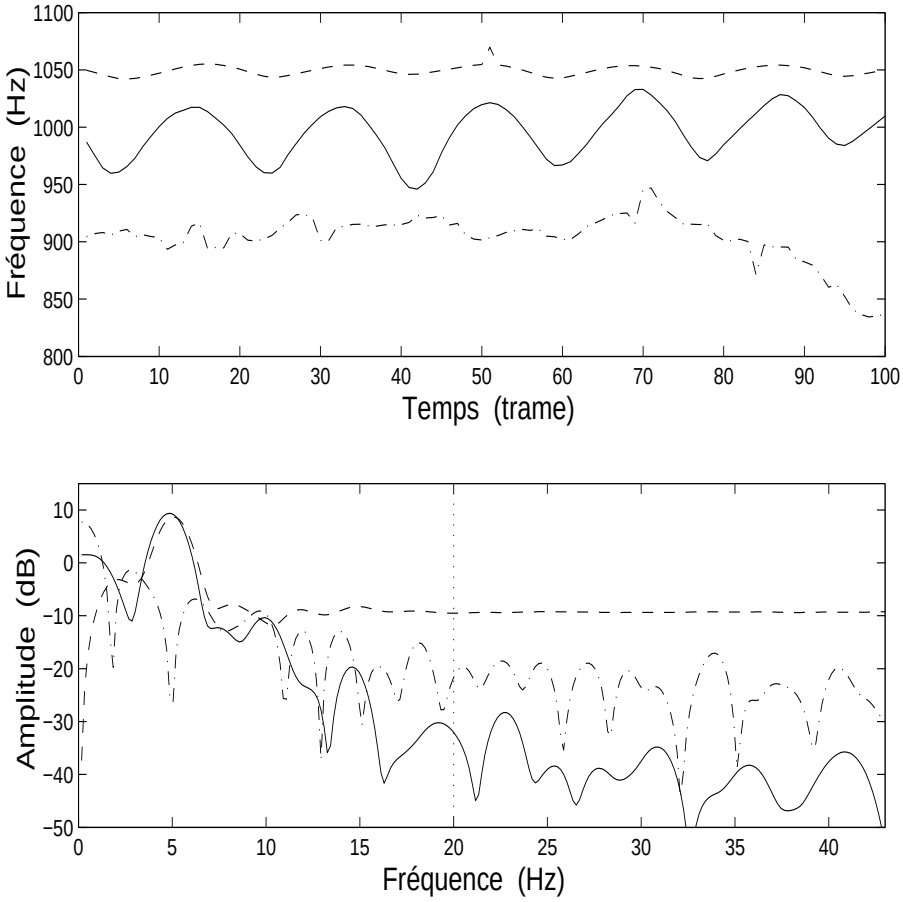


FIG. 3.23 – Trois évolutions en fréquence de partiels (en haut), extraits en utilisant l'algorithme MAQ et leurs spectres correspondants (en bas). De haut en bas : une harmonique d'une note de saxophone avec une discontinuité artificielle à la trame 50 ; une harmonique d'une note de saxophone correctement suivie ; un partiel extrait (par erreur) d'un signal de bruit blanc. Seul l'évolution en fréquence du partiel correctement suivi a un spectre concentré dans les basses fréquences avec un niveau d'énergie faible (-30 dB) lorsque la fréquence est supérieure à 20 Hz, et ce malgré des modulations importantes.

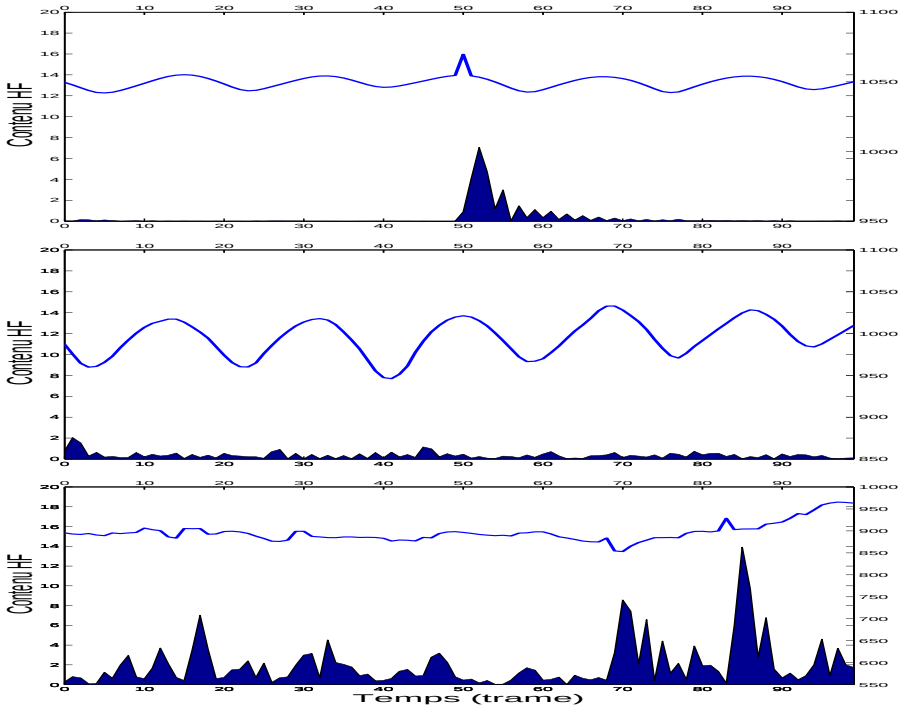


FIG. 3.24 – Sorties des filtres passe-haut (aire) en fonction des évolutions du paramètre de fréquence de trois partiels (trait).

n_f de trames futures. Ces trajectoires doivent être de longueur suffisante car les filtres destinés à l'estimation du contenu HF introduisent un délai. Une longueur égale à l'ordre du filtre (ici $n_f = 4$) est satisfaisante. On procède à une élection en trois étapes :

- les prédictions en fréquence et en amplitude sont calculées, voir figure 3.25(a) ;
- un sous-ensemble de pics dont les paramètres sont proches de ceux prédits est sélectionné dans les trames futures, voir figure 3.25(b) ;
- le contenu HF des trajectoires passant par ces pics extraits et des pics interpolés est évalué pour déterminer la meilleure prolongation.

Choix des pics candidats

Si n_c pics candidats sont sélectionnés dans n_f trames consécutives, le nombre de trajectoires possibles à tester est : $(n_c + 1)^{n_f}$, ce pour chaque partiel et à chaque trame. On réduit la complexité de l'algorithme en sélectionnant un sous-ensemble de pics candidats. En premier lieu, tous les pics dont la différence entre leur fréquence et la fréquence prédite est inférieure à un seuil Δ_f (ici, assez élevé) sont sélectionnés. Parmi ces pics, les n_c pics dont l'amplitude est la plus proche de l'amplitude prédite sont candidats. Par trames, on a donc $n_c + 1$ pics possibles, car on doit ajouter aux pics extraits la possibilité d'utiliser un pic interpolé.

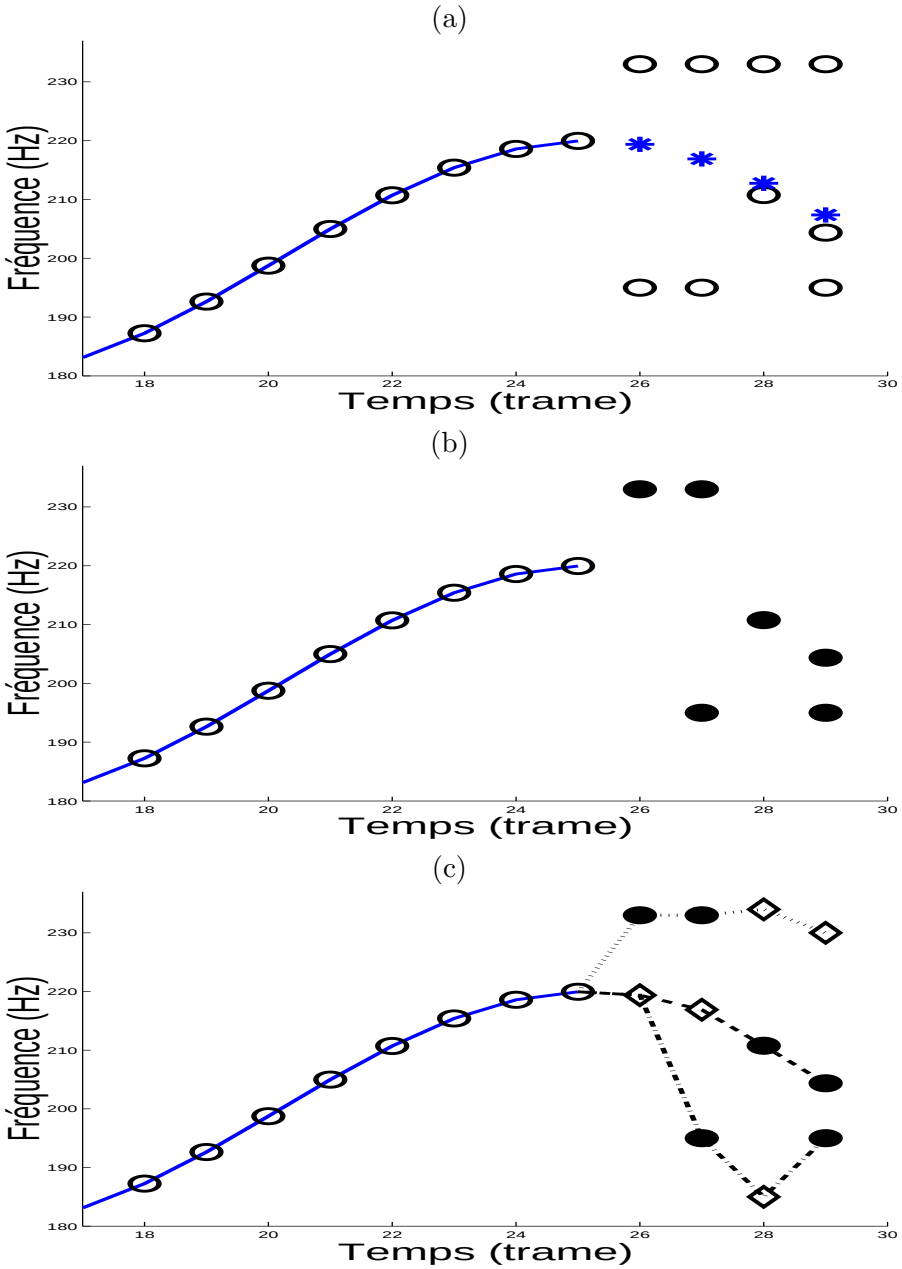


FIG. 3.25 – Sélection des pics candidats dans les trames futures. En haut, les fréquences prédites sont marquées par des étoiles. On choisit 2 pics par trame dont la différence entre la fréquence et la fréquence prédite est inférieure à un seuil Δ_f. On obtient alors les pics candidats représentés par des points noirs (au milieu). En bas sont représentées trois trajectoires possibles (traits non continus) parmi les 81 possibles. Ces trajectoires passent par des pics extraits par le module court-terme et des pics interpolés (diamants) dont les paramètres sont déterminés grâce aux coefficients AR précédemment calculés.

Exploration du treillis de trajectoires

Toutes les trajectoires passant parmi ces pics extraits par le module court-terme ou par un pic interpolé sont testées. Sur la figure 3.25(c) sont représentées 3 trajectoires parmi les 81 possibles. Chacune fait appel à des pics interpolés (diamants sur la figure 3.25(c)), dont les paramètres sont à déterminer. Lors du choix des pics candidats, les coefficients AR calculés grâce à l'évolution passée du partiel sont utilisés pour prédire l'évolution probable du partiel dans les trames futures. En supposant que ces évolutions sont localement stationnaires, ces coefficients sont directement utilisés pour le calcul des paramètres des pics interpolés utilisés dans les trajectoires.

Fonction de coût associé à une trajectoire

En supposant que le partiel ait été correctement extrait, c'est-à-dire que son évolution passée n'engendre pas de discontinuité, la prédiction dans les trames futures consiste en une évolution très stationnaire, au contenu HF faible. Au contraire, les erreurs d'estimation du module court-terme fait que les trajectoires passant par les pics extraits induisent toujours un contenu HF plus élevé. L'utilisation d'un pic interpolé doit donc être pénalisé.

Une fonction de coût c_t est calculée pour chaque trajectoire du treillis. Cette fonction est d'autant plus faible que le contenu HF en amplitude et en fréquence est faible et que le nombre de pics interpolés utilisés est réduit :

$$c_t = \left(\frac{1}{\xi}\right)^{n_t} \cdot \frac{\sum_{k=0}^{n_f} |\tilde{a}_k^{i_t}|^2}{K_a} \cdot \frac{\sum_{k=0}^{n_f} |\tilde{f}_k^{i_t}|^2}{K_f} \tag{3.42}$$

où $\tilde{a}_k^{i_t}$ et $\tilde{f}_k^{i_t}$ sont respectivement les amplitudes et les fréquences filtrées HF du pic d'indice i_t de la trame k. L'indice i_t dépend de la trajectoire considérée. En considérant qu'un pic interpolé a un indice égal à -1, la trajectoire de fréquence la plus faible sur la figure 3.25(c) a cette succession d'indices : $[-1, 0, -1, 0]$. Le paramètre n_t est le nombre de pics interpolés dans la trajectoire (2 dans la trajectoire présentée ci-dessus). K_a et K_f sont des constantes de normalisation. Comme le choix de la meilleure trajectoire mène à des contraintes sur l'ordre des fonctions de coût et non sur leurs valeurs dans l'absolu, K_a et K_f peuvent donc être mises à 1. Le paramètre $\xi \in]0, 1]$ permet de pénaliser plus ou moins l'utilisation de pics interpolés.

Le pic élu est alors celui qui débute la trajectoire dont la fonction de coût est la plus faible. Si ce pic est un pic interpolé, le partiel ne sera donc pas actif pour cette trame et utilisera ce pic interpolé pour se prolonger.

3.7.4 Algorithme

L'algorithme de suivi présenté dans cette section utilise la structure algorithmique présentée dans la section 3.5. Pour des raisons spécifiques à cette méthode, deux types de partiels sont introduits. Les filtres nécessaires à l'estimation HF associés à chaque partiel nécessitent quelques échantillons pour s'initialiser. Durant cette période longue de N_j trames, le partiel est considéré

comme “juvénile”. La méthode d’élection d’un partiel juvénile est similaire à celle de l’algorithme MAQ. Une fois que le partiel est devenu “mature” (par opposition avec juvénile), les sorties des filtres HF sont considérées comme pertinentes car N_j pics ont été insérés. L’élection peut alors se faire grâce à la méthode décrite précédemment. Lors de la confirmation, le partiel se prolonge par le pic élu s’il est disponible et rend indisponible tous les pics de la trajectoire.

Ordonnancement

Les partiels sont ordonnés selon le critère o_p calculé grâce à l’équation 3.43, de manière à ce que les partiels matures ayant l’amplitude la plus forte soient les premiers à sélectionner leur pic élu. Les partiels juvéniles sont considérés ensuite, par ordre de stationnarité. Soit n l’indice de trame courant, o_p est :

$$o_p = \begin{cases} A_p(n) & \text{si } p \text{ est mature} \\ -|f_{n+1} - F_p(n)| & \text{sinon} \end{cases} \tag{3.43}$$

où $A_p(n)$ et $F_p(n)$ sont l’amplitude et la fréquence du partiel P à la trame n, et f_{n+1} est la fréquence du pic élu.

Gestion du mode zombie

Un partiel juvénile à un nombre maximal de pics interpolés consécutifs limité à N_j. Si le nombre de pics interpolés consécutif atteint cette borne N_j, le partiel est déclaré mort. Les partiels matures utilisent une gestion du mode zombie différente. Chaque partiel mature à nombre maximal de pics interpolés consécutifs n_m qui peut changer de trame en trame avec une borne supérieure N_m. Lorsque le partiel devient mature, n_m est initialisé à N_y. Le paramètre n_m est incrémenté si le pic est extrait par le module court-terme et décrémenté si le pic est interpolé. Si le nombre de pics interpolés consécutifs atteint n_m, le partiel est déclaré mort. Ce type de gestion permet d’améliorer grandement la discrimination sinusoïde/bruit comme on le verra dans la section 3.8.

Conclusion

Cet algorithme présente de nombreux avantages : sa fonction de coût est proche d’une contrainte du modèle et l’utilisation de filtres HF implantés sous forme de cellules IIR d’ordre 2 permet de conserver une complexité très raisonnable. Cet algorithme permet de plus une adaptation aisée à des contenus sonores variés, ce qui faisait défaut aux deux méthodes présentées dans les sections 3.4 et 3.6. Comme on le verra dans la section suivante dédiée à l’évaluation, cet algorithme est performant autant du point de vue de la séparation déterministe/stochastique que de la gestion de signaux polyphoniques. L’analyse du contenu HF permet en outre de mieux modéliser les modulations des paramètres et de mieux identifier les indices de début et de fin des partiels. La représentation long-terme obtenue est ainsi plus interprétable.

3.8 Évaluation des algorithmes de suivi

Dans cette section, certaines méthodes présentées dans ce chapitre sont évaluées selon plusieurs critères. On introduit dans une première partie, une méthode d'évaluation des méthodes de suivi par leurs capacités intrinsèques. L'algorithme évalué est testé seul (sans avoir recours à un module d'analyse à court terme ni un module de synthèse) en partant d'une représentation court-terme $\mathcal{C}$ dont on connaît la représentation long-terme sous-jacente $\mathcal{S}$. On évalue alors la représentation long-terme obtenue $\hat{\mathcal{S}}$ en fonction de cette représentation sous-jacente.

Les parties suivantes évaluent ces méthodes de suivi lorsqu'elles sont insérées dans une chaîne complète d'analyse/synthèse sinusoïdale. Un ensemble de signaux temporels forment le jeu de tests, et le signal synthétisé permet une évaluation de la méthode de suivi. Pour chacune des méthodes comparées, le module d'analyse à court-terme et le module de synthèse sont identiques. Le module d'analyse à court-terme est paramétré comme suit : la taille de trame est de 2048 échantillons et le pas d'avancement de 360 à une fréquence d'échantillonnage de 44.1 kHz. L'algorithme de synthèse est celui présenté dans [MQ86], basé sur une interpolation linéaire de l'amplitude et une interpolation cubique de la phase, voir section 3.9.

Trois méthodes de suivi sont comparées. La première dite de MAQ est présentée dans la section 3.2 et utilise pour ces tests un Δ_f de 80 Hz et permet 4 utilisations successives du mode zombie. La seconde, basée sur la prédiction linéaire présentée dans la section 3.6, utilise un Δ_f de 40 Hz. Celle basée sur l'estimation du contenu HF présentée dans la section 3.7 utilise un ξ de 0.9. Concernant le mode zombie, cette dernière méthode utilise les paramètres suivant : $N_s = 20$, $N_y = 4$ et $N_m = 20$. Tous les partiels extraits ayant moins de 10 pics issus du module court-terme sont écartés.

L'algorithme de suivi par HMM, présenté dans la section 3.3, n'est pas évalué dans cette section. Les paramètres tels que les facteurs de normalisation des distances (variance des gaussiennes utilisée dans la mesure, voir section 3.5), aussi bien que les heuristiques utilisées pour réduire l'espace d'états n'étant pas disponibles, une ré-implantation aurait eu un comportement différent de celle présentée dans [DGR93a, DGR93b]. Les méthodologies de tests dont les résultats sont détaillés dans les trois première parties de cette section requièrent un nombre conséquent d'exécutions de l'algorithme pour atteindre une certaine validité statistique. La complexité de l'algorithme de suivi par exploration des trajectoires futures présenté dans la section 3.4 nous a amené à l'écarter de ces évaluations.

3.8.1 Évaluation des capacités intrinsèques

Lors de l'évaluation d'une chaîne d'analyse/synthèse sinusoïdale, on souhaite idéalement pouvoir évaluer chacun des éléments indépendamment des autres. En ce qui concerne le module d'analyse, l'estimation des paramètres des pics peut être évaluée grâce au formalisme des bornes de Cramér-Rao. La

qualité du module de synthèse peut être évaluée grâce à des mesures de SNR de reconstruction (R-SNR, voir équation 3.49) qui mesure objectivement la dégradation induite par une représentation particulière. Des tests d'écoute subjectifs où l'oreille humaine est juge de la qualité obtenue peuvent aussi être utilisés. En revanche, il n'existe pas dans la littérature de méthodologie de tests permettant d'évaluer les performances d'un algorithme de suivi. Dans cette partie, on introduit une méthode d'évaluation des algorithmes de suivi de partiels ne faisant aucune supposition sur les propriétés des modules d'analyse et de synthèse. Cette méthode se base sur les principes énoncées dans la section 1.4. Les éléments du jeu de tests sont constitués d'une représentation à court terme d'un partiel générée par une expression mathématique ou estimée en utilisant un algorithme de suivi de partiels. Un seul pic est donc présent par trame.

On souhaite ici pouvoir évaluer les capacités de robustesse aux dégradations classiques d'une représentation court-terme lors de l'analyse de signaux polyphoniques. Cette représentation est dégradée selon trois méthodes. Dans la première, certains pics sont enlevés. Dans la seconde, certains pics voient leurs paramètres modifiés. Dans la troisième, des pics supplémentaires sont ajoutés. La force de la dégradation est le nombre de pics dégradés ou ajoutés sur le nombre de pics initiaux. La trame où a lieu la dégradation est choisie au hasard.

La fréquence des pics modifiés est dégradée par l'ajout d'une valeur aléatoire comprise entre $[-\Delta_f, \Delta_f]$. La fréquence des pics ajoutés est choisie au hasard entre la fréquence moyenne des pics initiaux $\pm\Delta_f$. L'amplitude des pics modifiés ou ajoutés est choisie au hasard de 0 à l'amplitude maximale des pics initiaux.

La qualité de la représentation à long terme obtenue est évaluée selon quatre mesures :

- Q_e une mesure d'efficacité ;
- Q_c une mesure de complétude ;
- E_f une mesure d'erreur d'identification de la fréquence du partiel P ;
- E_a une mesure d'erreur d'identification de l'amplitude du partiel P.

Soient $A(n)$ et $F(n)$ les amplitudes et les fréquences du partiel d'origine et $\hat{\mathcal{S}}$, l'ensemble des partiels extrait par l'algorithme, ces quatre mesures sont :

$$Q_e = \begin{cases} 0 & \text{si Card } \hat{\mathcal{S}} = 0 \\ \frac{1}{\text{Card } \hat{\mathcal{S}}} & \text{sinon} \end{cases} \tag{3.44}$$

$$Q_c = \frac{1}{N_T} \text{Card } \{T_i | \exists \hat{P}_k \in \hat{\mathcal{S}} \wedge e_k(i) = 1\} \tag{3.45}$$

$$E_f = \frac{1}{2\,\Delta_f\,N_T} \sum_{i=0}^{N_T-1} \sum_{k=1}^{\text{Card } \hat{\mathcal{S}}} |F(i) - \hat{F}_k(i)|\, e_k(i) \tag{3.46}$$

$$E_a = \frac{1}{a_{\max}\,N_T} \sum_{i=0}^{N_T-1} \sum_{k=1}^{\text{Card } \hat{\mathcal{S}}} |A(i) - \hat{A}_k(i)|\, e_k(i) \tag{3.47}$$

où N_T est le nombre de trame T_i, $A_k(n)$ et $F_k(n)$ sont l'amplitude et la fréquence à la trame n du partiel d'indice k de $\hat{\mathcal{S}}$ et $e_k(i)$ vaut 1 si le partiel k est présent à la trame d'indice i et 0 sinon. Le paramètre $a_{\max}$ désigne l'amplitude maximale

du partiel d'origine.

Les représentations à long terme de tests sont composées de partiels synthétiques comme un vibrato, un trémolo, une sinusoïde dont les paramètres de fréquence et d'amplitude sont constants et une sinusoïde dont la fréquence est linéairement croissante et l'amplitude est constante. Des représentations long-terme de signaux naturels sont aussi utilisées, comme une harmonique d'un saxophone et d'une voix chantée. Pour chaque force de dégradation, les résultats obtenus pour chacun de ces signaux sont très similaires. Par conséquent, les résultats affichés sur les figures 3.26, 3.27 et 3.28 sont les résultats moyens de 100 simulations utilisant chaque élément de la base de tests.

La figure 3.26 montre les résultats obtenus par les trois méthodes testées lors de l'ajout de pics. Ce test montre la capacité d'un algorithme à éviter les pics de bruit. La gestion de l'état zombie est particulièrement importante dans ce cas. L'algorithme MAQ extrait de nombreux partiels, son efficacité est donc faible tandis que les méthodes proposées se comportent nettement mieux. La prise en compte du contenu HF notamment permet de rejeter des partiels composés de pics de bruit. On obtient alors une erreur plus faible et une efficacité nettement supérieure avec une complétude similaire.

La figure 3.27 montre les résultats lorsque les pics initiaux ont leur paramètres dégradés. Ce test a pour but de montrer comment une méthode de suivi est capable d'extraire une représentation qui n'engendre pas d'artefacts en partant d'une représentation court-terme dégradée. Même si l'algorithme LP est plus efficace que celui de MAQ, l'amélioration est faible. En revanche, l'estimation du contenu HF est d'un grand intérêt. Cette méthode de suivi obtient une erreur plus faible et une efficacité supérieure avec une complétude similaire.

La figure 3.28 montre les résultats dans le cas où certains pics sont supprimés. Ce test montre comment une méthode est capable de prédire sa continuation probable à travers plusieurs trames. Les deux méthodes proposées gardent une meilleure efficacité avec une complétude similaire. Les erreurs en fréquence et en amplitude sont très faibles pour les trois méthodes.

3.8.2 Séparation déterministe/stochastique

Les capacités de discrimination et d'efficacité des trois algorithmes sont évaluées en considérant maintenant ces algorithmes inclus dans une chaîne complète d'analyse/synthèse sinusoïdale. On considère une sinusoïde d'amplitude constante et de fréquence variant de manière sinusoïdale. La modulation appliquée à la fréquence du partiel est une sinusoïde de fréquence 4 Hz et d'amplitude 50 Hz. La fréquence oscille autour de 2 kHz. De manière à tester le comportement de ces trois algorithmes face au bruit, on ajoute à ce signal un niveau croissant de bruit blanc. La qualité de l'algorithme de suivi est alors évalué grâce au SNR de reconstruction (R-SNR) en fonction du SNR de dégradation (D-SNR).

Le D-SNR est le rapport d'énergie entre le perturbateur (le bruit blanc) et le signal test (le vibrato). Le R-SNR est le rapport d'énergie entre l'erreur de modélisation (la différence entre le signal original et le signal synthétisé) et le

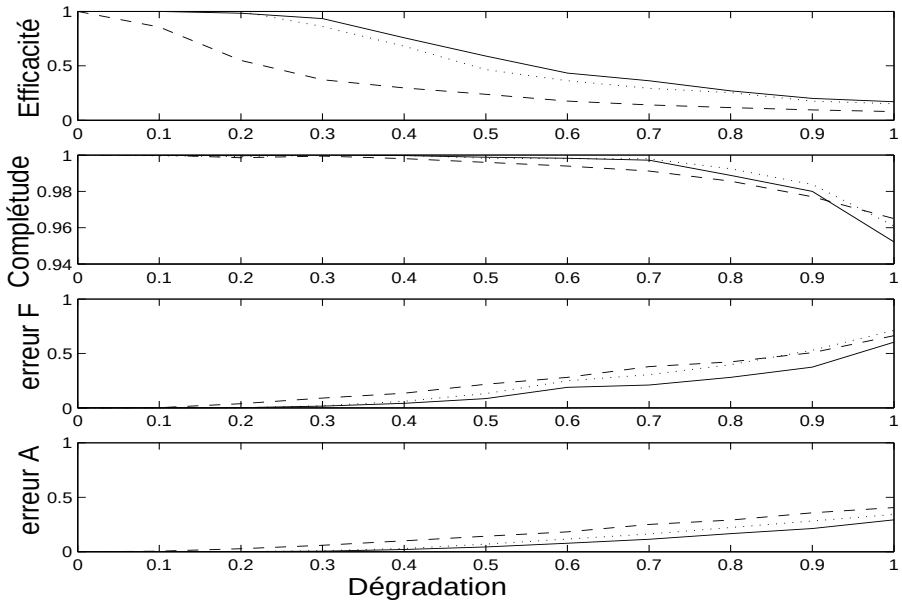


FIG. 3.26 – Évaluation de la résistance à l'ajout de pics pour les algorithmes suivants : l'algorithme de MAQ (en tirets), de l'algorithme LP (en pointillés) et de l'algorithme HF (en trait plein). Les résultats sont donnés en fonction du taux de pics ajoutés.

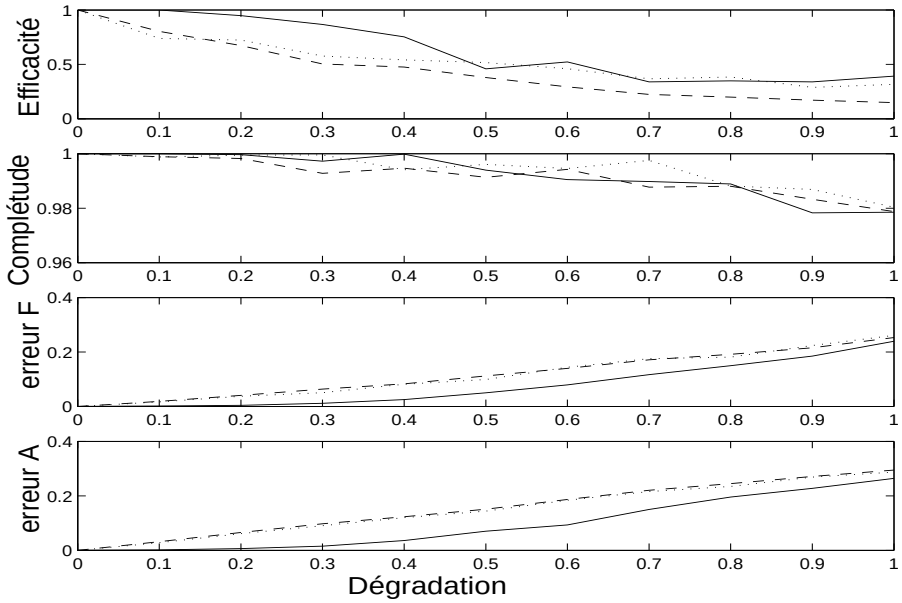


FIG. 3.27 – Évaluation de la résistance à la modification de la fréquence des pics pour les algorithmes suivants : l'algorithme de MAQ (en tirets), de l'algorithme LP (en pointillés) et de l'algorithme HF (en trait plein). Les résultats sont donnés en fonction du taux de pics modifiés.

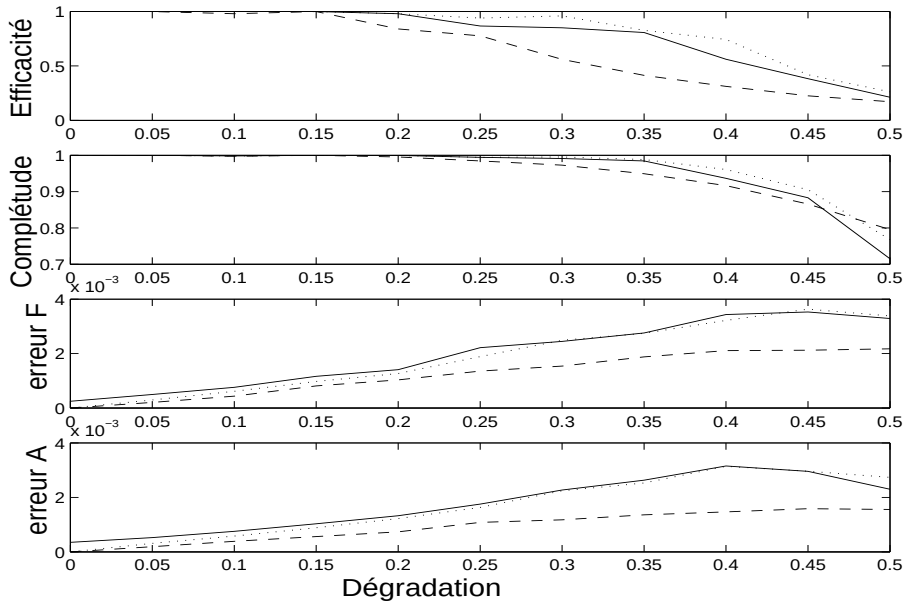


FIG. 3.28 – Évaluation de la résistance à la suppression de pics pour les algorithmes suivants : l'algorithme de MAQ (en tirets), de l'algorithme LP (en pointillés) et de l'algorithme HF (en trait plein). Les résultats sont donnés en fonction du taux de pics manquants.

signal test. Ces deux mesures se calculent grâce aux équations suivantes :

$$\text{D-SNR} = 10 \log 10 \left(\frac{\sum_{n=0}^{N-1} b^2(n)}{\sum_{n=0}^{N-1} x^2(n)} \right) \tag{3.48}$$

$$\text{R-SNR} = 10 \log 10 \left(\frac{\sum_{n=0}^{N-1} (x(n) - \hat{x}(n))^2}{\sum_{n=0}^{N-1} x^2(n)} \right) \tag{3.49}$$

où $x(n)$ est le signal original, $b(n)$ le signal perturbateur et $\hat{x}(n)$ le signal synthétisé. Dans une première expérience, l'efficacité est évaluée en ne conservant des partiels extraits que le partiel avec la plus grande amplitude moyenne. Seul ce partiel est utilisé pour synthétiser $\hat{x}(n)$.

À des niveaux de D-SNR inférieurs à -7 dB, l'algorithme MAQ extrait des partiels qui sont un mélange de pics issus de la composante sinusoïdale et de pics de bruit, voir figure 3.29(a). Les deux méthodes proposées ont un D-SNR qui reste stable, en partie grâce à une meilleure sélectivité. La lente dégradation du D-SNR est due aux erreurs d'estimation des paramètres du module court-terme.

Dans la seconde expérience, on évalue les capacités de discrimination entre processus déterministe et processus stochastique en conservant tous les partiels dont la fréquence est située entre 1.9 et 2.1 kHz. L'intégralité de ces partiels est utilisée pour la synthèse de $\hat{x}(n)$. L'utilisation de la prédiction et la recherche de partiels avec des paramètres qui varient lentement permet d'améliorer de manière conséquente la discrimination, voir figure 3.29(b).

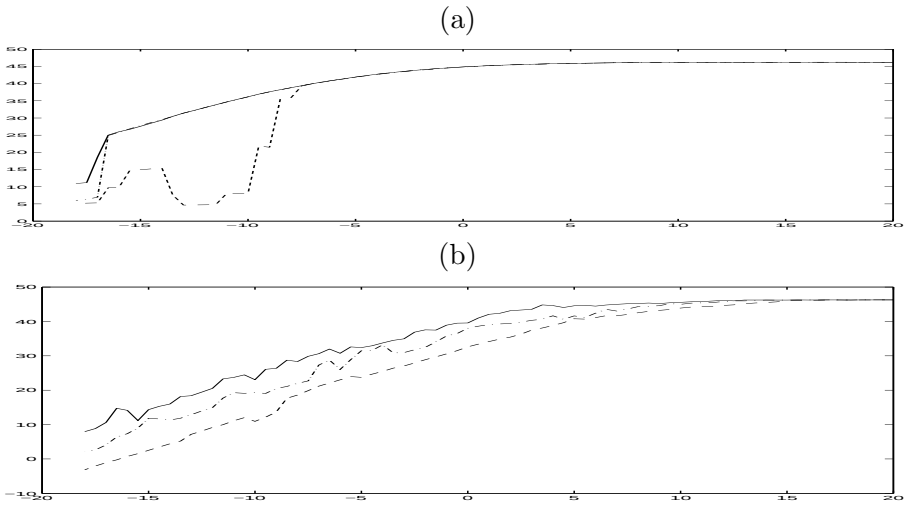


FIG. 3.29 – Évaluation de l'efficacité (a) et de la capacité de discrimination (b) entre processus déterministe et processus stochastique pour l'algorithme de MAQ (en tirets), l'algorithme LP (en pointillés) et l'algorithme HF (en trait plein). Les figures montrent le SNR de reconstruction et fonction du SNR de dégradation. Pour évaluer l'efficacité (en haut), seul le partiel de plus haute amplitude est synthétisé. Pour évaluer la capacité de discrimination (en bas), tout les partiels sont synthétisés.

3.8.3 Gestion des polyphonies

Pour une modélisation à long terme de signaux polyphoniques, le module de suivi de partiels doit pouvoir tolérer les composantes sinusoïdales qui se croisent ou qui ont des fréquences proches. Pour cela, ce module doit être capable d'identifier à l'avance l'évolution probable des partiels et interpoler les composantes court-terme manquantes, voir figure 3.30 pour le croisement et figure 3.2 pour les composantes de fréquences proches.

Pour évaluer le comportement des différentes méthodes face au croisement, on ajoute à une note de saxophone une sinusoïde synthétique d'amplitude constante et de fréquence augmentant linéairement de 200 Hz à 4 kHz. Cette sinusoïde qui croise toutes les harmoniques de la note de saxophone commence 20 trames après le début de cette note. Seuls les partiels dont la naissance est antérieure à la trame 20 sont synthétisés. L'algorithme MAQ utilise un Δ_f de 80 Hz, l'algorithme LP un Δ_f de 25 Hz, et l'algorithme HF un Δ_f de 25 Hz et un ξ de 0.9. Les résultats sont affichés sur la figure 3.31(a).

Avoir un modèle d'évolution permet une gestion plus efficace des croisements. Les algorithmes LP ou HF sont ainsi plus sélectifs et ont une meilleure capacité d'interpolation. De plus, ces algorithmes ordonnent les partiels par amplitude croissante. Avec un tel ordonnancement, les partiels avec la plus faible dégradation sont prolongés les premiers, ceci réduit la probabilité de gestion incorrecte. Par exemple, la composante sinusoïdale de la figure 3.30 dont la fréquence croît linéairement est prolongée en premier.

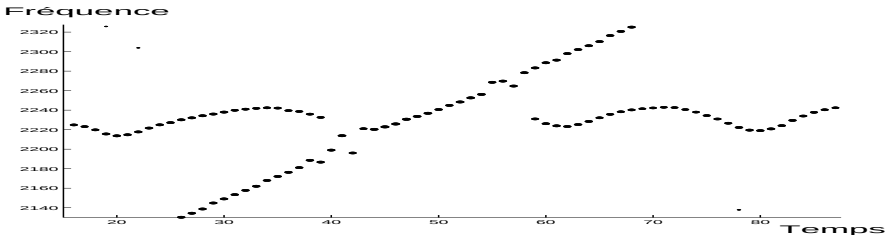


FIG. 3.30 – Représentation à court terme d'une harmonique d'une note de saxophone et d'une sinusoïde synthétique dont les fréquences respectives se croisent. Pour extraire une bonne représentation long terme de croisements, le module doit pouvoir prédire et interpoler les évolutions des paramètres de partiels dans la partie corrompue.

L'analyse sinusoïdale de signaux polyphoniques requiert une résolution fréquentielle arbitrairement haute. Dans le même temps, le compromis entre résolutions fréquentielle et temporelle fait que le module d'analyse à court terme doit utiliser une taille de fenêtre raisonnable. Les relations entre les hauteurs des différents notes de la gamme amènent des contaminations des composantes de la DFT ainsi que des sinusoïdes dont les fréquences sont arbitrairement proches.

Pour évaluer la gestion des sinusoïdes dont les fréquences sont proches, on ajoute à une note de saxophone de hauteur 440 Hz un signal composé d'un ensemble de sinusoïdes synthétiques de fréquences et d'amplitudes constantes et qui débutent 20 trames après le début de la note. Leurs fréquences sont : $F(h) = 370 + 440\,(h-1)$, où F_h désigne la fréquence de la sinusoïde de rang h. Seuls les partiels dont la naissance est antérieure à la trame 20 sont synthétisés pour calculer le R-SNR.

La figure 3.31(b) montre les avantages de l'analyse du contenu HF des évolutions des partiels sur plusieurs trames lorsque les partiels sont proches. Quand la note synthétique commence, la représentation à court terme contient beaucoup de pics de bruit entre les composantes sinusoïdales. La méthode LP est alors incapable d'éviter ces pics de bruit et donne de mauvais résultats. En revanche, l'algorithme HF donne de bons résultats même à haut D-SNR.

3.8.4 Interprétabilité de la représentation

Pour des applications comme l'indexation ou la séparation de sources de signaux quasi périodiques, une bonne représentation à long terme donne un haut niveau de description, utile pour extraire de multiples informations sur le signal analysé comme le début et la fin des notes, la hauteur, le type d'instrument, etc. De manière à détecter aisément le début et la fin des notes, une bonne séparation temporelle est nécessaire. Autrement dit, une bonne précision est requise (un partiel ne doit appartenir qu'à une note, voir section 1.2.2).

Pour pouvoir détecter la hauteur d'une note, le nombre de partiels dits "de bruit" doit être le plus réduit possible. Pour identifier quel partiel appartient à quelle source, les paramètres des partiels doivent avoir une évolution "claire"

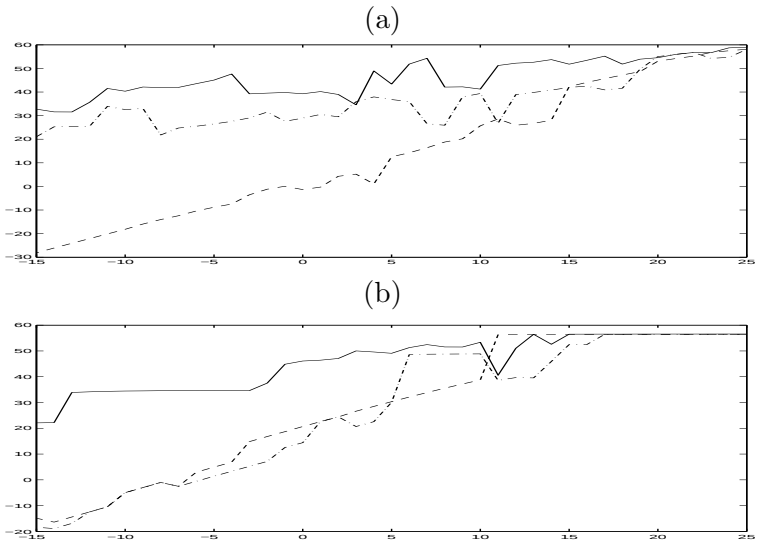


FIG. 3.31 – Évaluation de la gestion des sinusoïdes dont les fréquences se croisent (a) et des sinusoïdes proches (b) pour les algorithmes de MAQ (pointillés), LP (tirets) et HF (trait plein). Les figures montrent le R-SNR en fonction du D-SNR.

de manière à pouvoir les regrouper de manière robuste. Comme on peut le constater sur la figure 3.32(a), la représentation long-terme de l'algorithme de MAQ n'est pas satisfaisante. De nombreux partiels appartiennent à deux ou trois notes et il serait ardu de détecter la fréquence du vibrato de la deuxième note. L'algorithme LP identifie mieux le vibrato, mais la représentation n'est pas très précise car de nombreux partiels appartiennent à plus d'une note. L'algorithme HF donne de bien meilleurs résultats en terme de précision et le vibrato de la seconde note est plus prononcé, comme on peut le constater sur la figure 3.32(c).

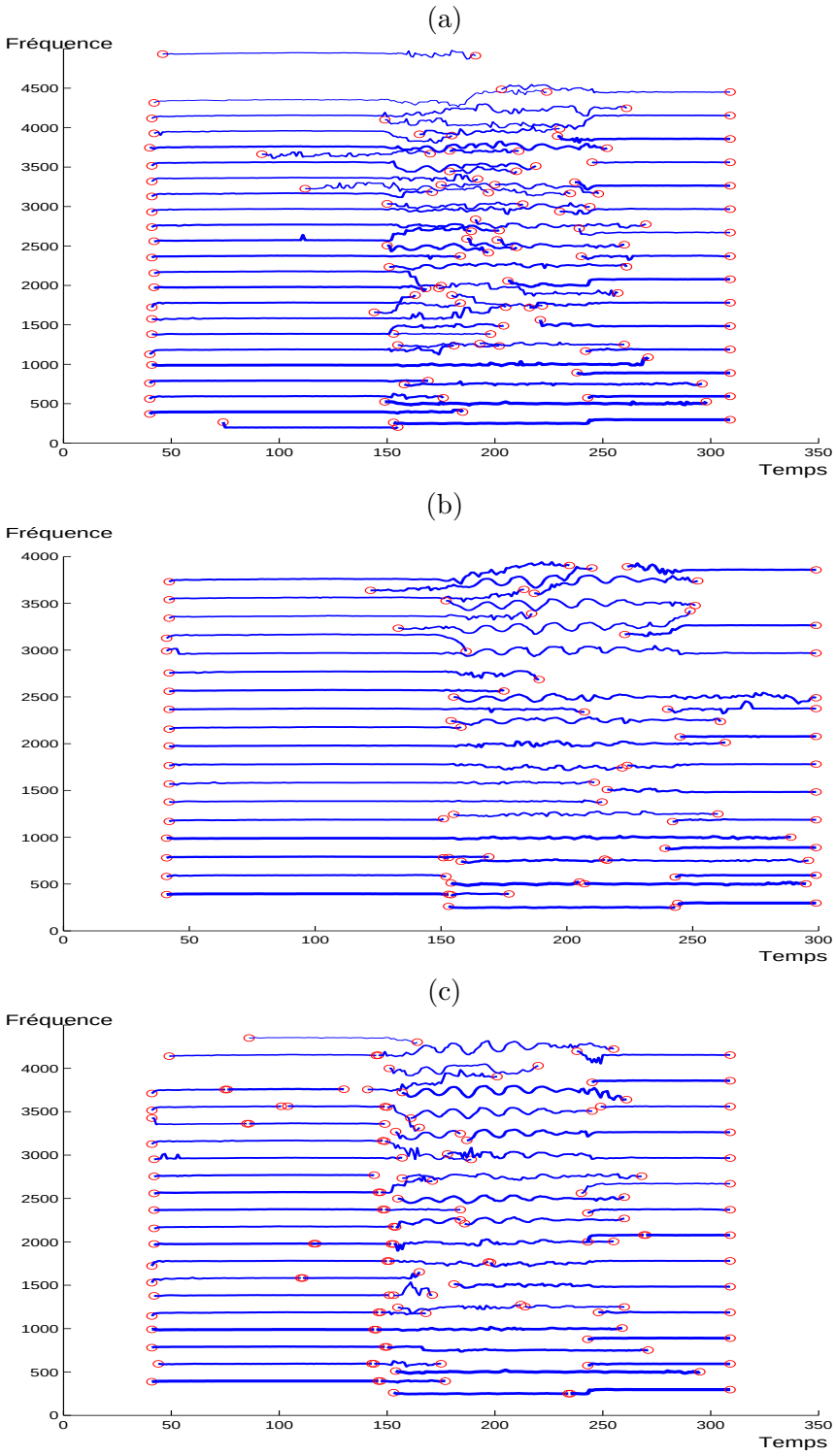


FIG. 3.32 – Représentations à long terme de trois notes de violon. Les représentations affichées de haut en bas sont le résultat des algorithmes de MAQ, LP et HF. Les partiels sont représentés par des lignes pleines, débutant en finissant par des cercles.

3.9 Synthèse

On présente trois approches de synthèse des sinusoïdes long-terme. La première approche utilise un algorithme de synthèse rapide tel que celui présenté dans la section 2.5.1. Le dernier nombre complexe calculé est ensuite stocké pour servir de référence de départ pour la synthèse de la trame suivante. Comme on n'exploite pas les informations de phase du partiel issues de l'analyse à court terme, cette méthode n'est utilisable qu'à des fins d'écoute car le signal synthétisé n'est pas en phase avec le signal analysé.

Dans un modèle Sinusoïdes+Bruit (SB), il est indispensable que ces deux signaux soient en phase pour pouvoir calculer le signal résiduel par soustraction. En utilisant les paramètres de fréquence et de phase donnés par le module d'analyse, les algorithmes rapides présentés dans les sections 2.2.1 et 2.5.1 peuvent être employés à condition que soient utilisés des trames de synthèse avec recouvrement. De cette manière, on se ramène en quelque sorte à un modèle sinusoïdal à court terme, voir figure 1.6. Ceci amène cependant à négliger toutes les informations de continuité apportées par le modèle à long terme.

Soient deux pics p_i et p_{i+1} consécutifs d'un partiel donné. Dans la suite, trois méthodes qui assurent la continuité de la phase entre les deux pics sont présentées. Leur principe est d'interpoler plus ou moins finement la phase du partiel dans la trame de synthèse de manière à éviter des discontinuités entre deux trames. Dans toutes ces méthodes, l'amplitude est interpolée linéairement.

La phase étant connue modulo 2π, il convient d'estimer un coefficient de déroulement. Ce coefficient est le nombre de fois où la phase modulo 2π s'est repliée pendant la durée de la trame de synthèse. Le nombre de valeurs possibles de ce coefficient étant infini, ce coefficient est choisi tel que l'évolution de la phase soit la plus "douce" possible. Dans un modèle sinusoïdal stationnaire, ce coefficient est [MQ86] :

$$M_{\mathrm{s}} = \left[\frac{1}{2\pi}\left((\phi_i - \phi_{i+1}) + (f_i + f_{i+1})\frac{N}{2F_e}\right)\right] \tag{3.50}$$

Où $[x]$ dénote la partie entière de x. Dans un modèle sinusoïdal non stationnaire, ce coefficient est plus complexe [GMdM+03] :

$$M_{\mathrm{ns}} = \left[\frac{1}{2\pi}\left((\phi_i - \phi_{i+1}) + (f_i + f_{i+1})\frac{N}{2F_e} + (\Delta^f_{i+1} - \Delta^f_i)\frac{N^2}{40F_e}\right)\right] \tag{3.51}$$

Si on veut conserver uniquement la continuité de la phase, un modèle à phase linéaire peut être utilisé :

$$\phi(n) = \phi_i + \frac{\phi_{i+1} - \phi_i + 2\pi M_{\mathrm{s}}}{N} n \tag{3.52}$$

où M_s est calculé grâce à l'équation 3.50. Ce modèle de phase simple permet l'utilisation d'algorithmes rapides tel que celui présenté dans la section 2.2.1.

Si on veut conserver la continuité de la phase et de la fréquence, on obtient quatre contraintes aux extrémités de la trame de synthèse. Le polynôme de phase doit être de degré trois :

$$\phi(n) = \phi_i + \omega\, n + \alpha\, n^2 + \beta\, n^3 \tag{3.53}$$

C'est la méthode couramment appelée méthode de Mc Aulay et Quatieri du nom de ses auteurs [MQ86]. Une fois ces valeurs de phase calculées, la valeur de l'échantillon temporel correspondante est obtenue par calcul du cosinus.

Dans un modèle non stationnaire, si on veut conserver la continuité de la phase, de la fréquence et de la dérivée de la fréquence, on obtient six contraintes aux bornes de la trame de synthèse. Le polynôme de phase doit donc être de degré cinq :

$$\phi(n) = \phi_i + \omega\, n + \frac{\delta_f}{2} n^2 + \alpha\, n^3 + \beta\, n^4 + \gamma\, n^5 \tag{3.54}$$

Les résultats publiés dans [GMdM+03] montrent que dans le cas où la taille de fenêtre d'analyse est fixe (ce qui est le cas pour l'analyse de signaux polyphoniques), les trois modèles de phase ont des performances proches. L'amélioration du SNR du modèle de phase linéaire au modèle cubique est inférieur à 1 dB. L'amélioration obtenue du modèle de phase cubique au modèle d'ordre cinq est approximativement de 0.2 dB.

4 Restauration à long terme

Dans de nombreux cas de figures, certains paramètres des partiels peuvent être détériorés ou indisponibles. Lors de l'analyse sinusoïdale à long terme, un événement transitoire peut perturber suffisamment le spectre pour que les informations de fréquence, d'amplitude et de phase extraites soient inexploitables. Lors de transmission de la partie sinusoïdale dans une application de codage en *streaming*, des défauts de transmission peuvent engendrer des pertes d'informations nécessaires à la synthèse. Il convient alors d'interpoler une représentation sinusoïdale à long terme durant des intervalles de temps conséquents. Or, les résultats obtenus par les méthodes existantes sont jugés artificiels par les auditeurs lors de l'interpolation d'une partie manquante de durée supérieure à 50 ms. On propose une méthode basée sur la modélisation autorégressive des paramètres de fréquence et d'amplitude des partiels pour interpoler une partie manquante. Cette méthode originale conserve les micro-modulations de ces paramètres (phénomène important pour la perception) et permet ainsi une meilleure restauration.

4.1 Introduction

On peut distinguer deux cas de dégradation d'une représentation à long terme. Une perte d'informations peut toucher les partiels durant un même intervalle de temps. C'est un cas de perte synchrone comme le montre la figure 4.2(a). On a alors une partie manquante débutant à la trame n_2 et se terminant à la trame n_3 qui corrompt un ensemble de sinusoïdes $\mathcal{S}$. De manière à restaurer cette partie manquante, deux ensembles de partiels $\mathcal{G}$ et $\mathcal{D}$ sont utilisés. Le premier ensemble $\mathcal{G}$ est constitué des partiels à *gauche* de la dégradation sur un axe temporel se terminant à la trame n_2. L'autre ensemble $\mathcal{D}$ est constitué des partiels à *droite* de la dégradation commençant à la trame n_3.

Bien que les applications potentielles soient nombreuses, peu de travaux ont été spécifiquement dédiés à ce problème. Quatieri et Danisewicz [QD90] proposent une méthode pour interpoler certaines harmoniques qui se chevauchent pour permettre la séparation de voyelles de deux voix dans un enregistrement mono-voie. Cet algorithme est basé sur le principe utilisé pour la synthèse de partiels, voir section 3.9. L'amplitude $\hat{a}(t)$ est interpolée linéairement entre $a(n_2)$ et $a(n_3)$. La phase $\hat{\phi}(t)$ est interpolée grâce à un polynôme cubique entre $\phi(n_2)$ et $\phi(n_3)$. La fréquence $\hat{f}(t)$ est alors obtenue par dérivation de la phase interpolée. Bien que cette méthode, nommée méthode d'interpolation polynomiale dans la suite, ait été à l'origine dédiée à l'interpolation entre trames des paramètres sinusoïdaux pour des applications de synthèse, elle donne de bons résultats pour des parties manquantes de durées comprises entre 20 et 100 ms.

Une telle interpolation préserve globalement l'énergie du partiel et assure la continuité de phase. En revanche, les modulations des paramètres des partiels ne sont pas pris en compte. Par exemple, la fréquence d'un partiel avec un vibrato naturel est une sinusoïde sur le plan temps/fréquence d'approximativement 4 Hz additionnée d'une constante. Comme le polynôme interpolateur de phase est cubique, le polynôme résultant pour la fréquence est quadratique. Or, une sinusoïde est correctement approximée par un polynôme de degré 2 pendant moins d'un quart de période. Si on veut être à même d'interpoler des partiels présentant un vibrato naturel, l'utilisation d'une telle méthode est limitée à des parties manquantes de longueur d'au plus 60 ms. De manière similaire, si on veut pouvoir interpoler un trémolo naturel, l'utilisation de cette méthode est limitée à des parties manquantes de longueur d'au plus 20 ms. Or, les modulations des paramètres de fréquence ou d'amplitude sont importants pour la perception comme cela est détaillé dans [Bre90]. On a donc besoin d'une méthode qui conserve les modulations sur des intervalles de temps conséquent.

On propose dans ce chapitre un algorithme de restauration composé de 4 étapes. Le schéma de principe de cet algorithme est donné par la figure 4.1. Cet algorithme se base sur la prédiction des évolutions des paramètres des partiels des deux ensembles dans la partie manquante. En fonction de ces prédictions, un algorithme développé dans une première section, décide de l'appariement de certains partiels de l'ensemble $\mathcal{G}$ avec certains partiels de l'ensemble $\mathcal{D}$. Ces partiels appariés forment alors un unique partiel avec une partie manquante,

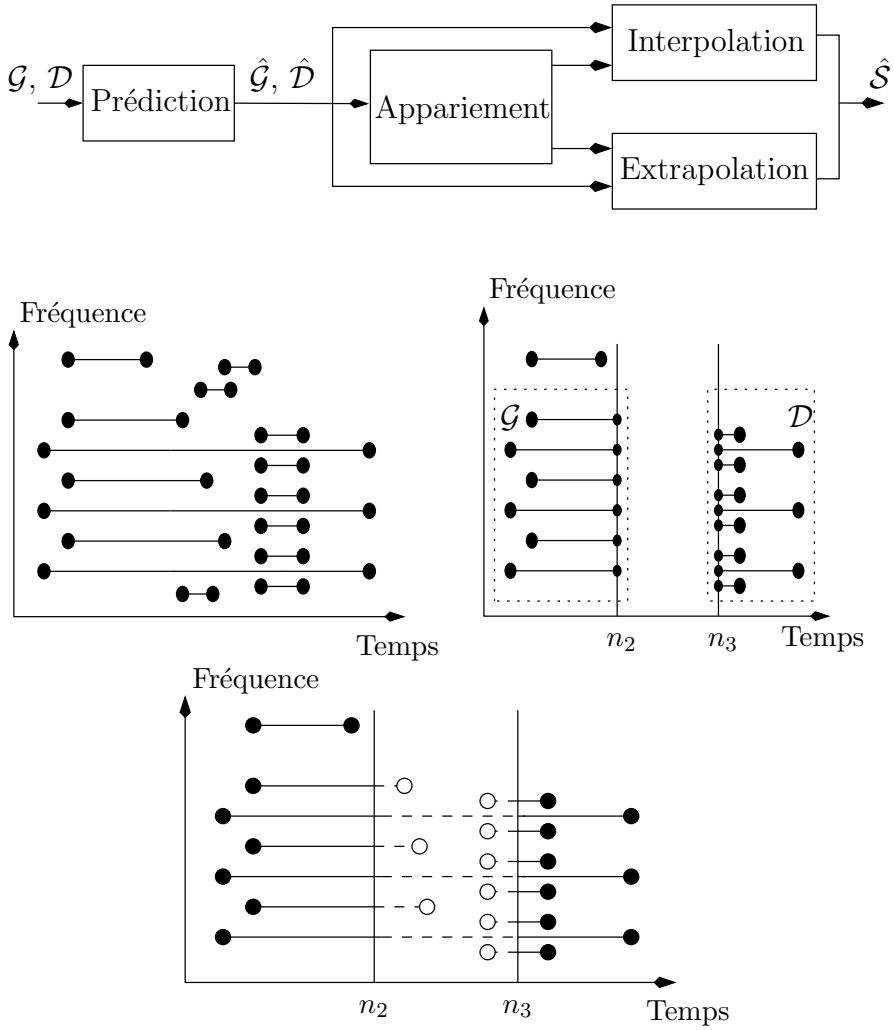


FIG. 4.1 – En haut, schéma de principe de l'algorithme de restauration de la représentation sinusoïdale à long terme. La représentation originale $\mathcal{S}$ est corrompue par une partie manquante débutant au temps n_2 et se terminant au temps n_3. Deux ensembles de partiels $\mathcal{G}$ et $\mathcal{D}$ sont utilisés. L'ensemble $\mathcal{G}$ est constitué des partiels à *gauche* de la dégradation se terminant à la trame n_2. L'ensemble $\mathcal{D}$ est constitué des partiels à *droite* de la dégradation commençant à la trame n_3. Grâce aux prédictions de l'évolution des fréquences et des amplitudes de ces partiels dans la partie manquante ($\hat{\mathcal{G}}$ et $\hat{\mathcal{D}}$), des partiels de $\mathcal{G}$ sont appariés avec des partiels de $\mathcal{D}$. La partie manquante (tirets) de chacune de ces paires est ensuite interpolée selon l'algorithme présenté dans la section 4.4. Les partiels non appariés (se terminant avec des points vides) sont extrapolés dans la partie manquante.

interpolée grâce à une méthode originale introduite dans une deuxième section. Les partiels non appariés sont eux extrapolés dans la partie manquante grâce à une méthode présentée dans la troisième section.

Lors de l'analyse de signaux polyphoniques, les dégradations éventuelles

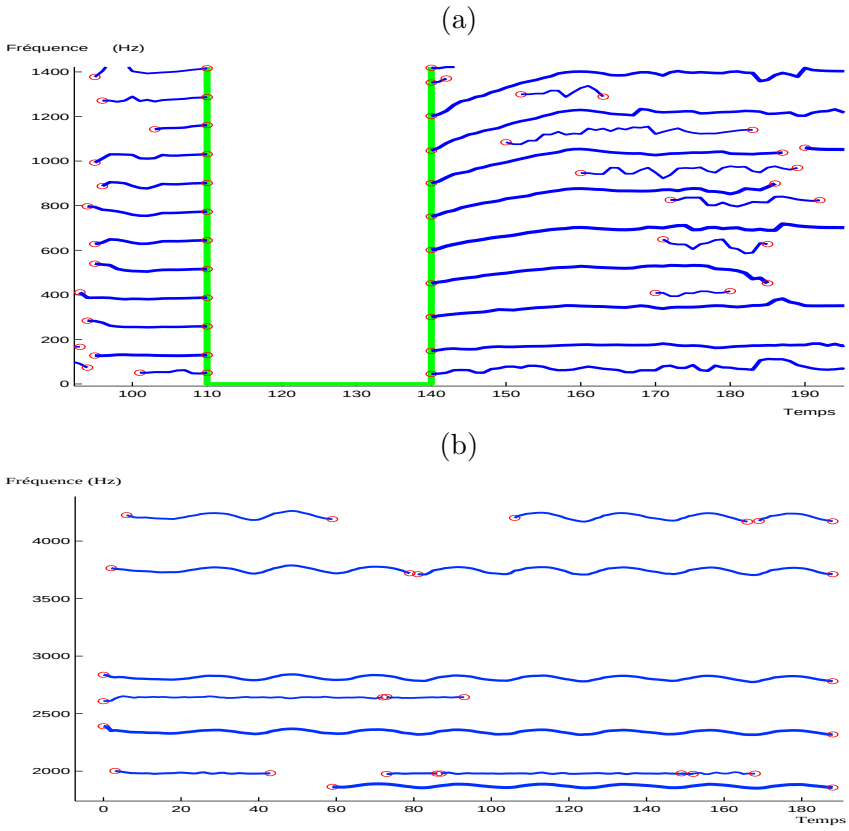


FIG. 4.2 – Deux cas de dégradation d'une représentation à long terme. En haut, toutes les informations relatives aux partiels sont indisponibles pendant un intervalle de temps commun à tous les partiels, la dégradation est dite synchrone. En bas, la représentation à long terme est morcelée en des intervalles de temps particuliers à chaque partiel ; la dégradation est dite asynchrone. L'algorithme de restauration proposé dans ce chapitre est suffisamment générique pour être appliqué dans ces deux cas.

sont asynchrones : les partiels peuvent être dégradés durant des intervalles de temps différents, comme le montre la figure 4.2(b). Dans une troisième partie, on présente comment l'algorithme introduit dans ce chapitre peut être adapté simplement à ce nouveau problème. La restauration d'enregistrements musicaux dégradés par des événements impulsifs est une problématique qui reçoit une attention particulière dans la littérature [Vas88b, JVV86, Ett96, KKS01, Mah94]. Dans une dernière partie, l'algorithme proposé est appliqué à ce problème et comparé à deux méthodes connues. Au vu des tests subjectifs effectués, la méthode proposée apporte un gain de qualité significatif. Grâce à la modélisation AR des paramètres de partiels, une interpolation réaliste est opérée pour des parties manquantes de durée élevée.

4.2 Prédiction des paramètres

De manière à clarifier l'exposé des parties suivantes, on introduit quelques notations. Soient P_i et P_j, des partiels des ensembles $\mathcal{G}$ et $\mathcal{D}$:

$$\mathcal{G} = \{P_i, i = 1, \cdots, N\} \tag{4.1}$$
$$\mathcal{D} = \{P_j, j = 1, \cdots, M\} \tag{4.2}$$
$$P_i = \{P_i(n), n = n_2 - l_i, \cdots, n_2\} \tag{4.3}$$
$$P_j = \{P_j(n), n = n_3, \cdots, n_3 + l_j\} \tag{4.4}$$
$$P_i(n) = \{F_i(n), A_i(n), \Phi_i(n)\} \tag{4.5}$$

où l_i et l_j sont respectivement les longueurs de P_i et P_j.

On a montré dans la section 3.6 que la modélisation AR des paramètres de fréquence et d'amplitude est utile pour améliorer le suivi de partiels car cette modélisation prend en compte les modulations prédictibles. Pour chaque partiel de $\mathcal{G}$ et $\mathcal{D}$, on calcule l'évolution probable de la fréquence et de l'amplitude dans l'intervalle manquant. Formellement :

$$\hat{\mathcal{G}} = \{\hat{P}_i, i = 1, \cdots, N\} \tag{4.6}$$
$$\hat{\mathcal{D}} = \{\hat{P}_j, j = 1, \cdots, M\} \tag{4.7}$$
$$\hat{P}_i = \{\hat{P}_i(n_2 + k), k = 1, \cdots, n_3 - n_2 - 1\} \tag{4.8}$$
$$\hat{P}_j = \{\hat{P}_j(n_3 - k'), k' = 1, \cdots, n_3 - n_2 - 1\} \tag{4.9}$$
$$\hat{P}_i(n) = \{\hat{F}_i(n), \hat{A}_i(n)\} \tag{4.10}$$

où $\hat{P}_i(n)$ est un couple de paramètres instantanés prédits grâce à la méthode de la prédiction linéaire décrite dans la section 3.6. Le calcul de $\hat{F}_i(n)$ se fait grâce à $k = \min(20, \frac{n_2 - n_1}{2})$ coefficients AR estimés avec les observations $[F_i(n_2 - 2k), \cdots, F_i(n_2)]$. Le nombre de coefficients utilisé est le même pour $\hat{A}_i(n)$. Pour $\hat{F}_j(n)$ et $\hat{A}_j(n)$, la même méthode est appliquée à ceci près que l'extrapolation est faite en arrière, voir figure 4.6.

4.3 Appariement de partiels

La première étape de l'interpolation d'une représentation à long terme est de décider quel partiel de l'ensemble $\mathcal{G}$ doit être apparié avec un partiel de l'ensemble $\mathcal{D}$ pour former un unique partiel. On propose de guider cette étape de décision par les informations de prédiction $\hat{\mathcal{G}}$ et $\hat{\mathcal{D}}$.

C'est une problématique très similaire au suivi de partiels. On doit faire correspondre, appariés, des éléments de part et d'autre d'un intervalle de temps où la continuité est perdue. Une première approche consiste à adapter l'algorithme de suivi de partiels de Mc Aulay et Quatieri décrit dans la section 3.2. Les partiels de couples (P_i, P_j) tels que la distance entre la dernière fréquence de P_i et la première fréquence de P_j est inférieure à un certain seuil Δ_f sont appariés :

$$F_i(n_2) - F_j(n_3) < \Delta_f \tag{4.11}$$

où $f_i(n_2)$ est la dernière fréquence de P_i et $f_j(n_3)$ la première fréquence P_j et Δ_f est un paramètre. Comme il est remarqué dans [Mah94], si le contenu de la représentation à long terme varie au court du temps, l'utilisation d'un tel algorithme amène à des résultats non satisfaisants. L'intervalle de temps entre les éléments à apparier est bien plus long que dans le cas du suivi de partiels. L'hypothèse de stationnarité en fréquence supposé par l'algorithme n'est alors plus vérifiée, voir figure 4.3(a).

Une prédiction plus fine dans la partie manquante permet d'apparier de façon plus robuste les partiels de deux ensembles $\mathcal{G}$ et $\mathcal{D}$. Notons $d_f(P_i, P_j)$ la distance euclidienne entre les fréquences prédites $\hat{f}_i$ et $\hat{f}_j$:

$$d_f(P_i, P_j) = \sum_{t=n_2}^{n_3} \left(\hat{F}_i(n) - \hat{F}_j(n)\right)^2 \tag{4.12}$$

La distance euclidienne $d_a(P_i, P_j)$ entre les amplitudes prédites $\hat{a}_i$ et $\hat{a}_j$ est définie de manière similaire. La première phase de l'algorithme consiste à détecter chaque couple (P_i, P_j) tel que $d(f_i, f_j)$ est inférieur à un seuil Δ_f. Ces couples sont alors candidats à l'appariement.

Les couples candidats sont évalués par ordre de distance décroissante grâce à deux critères mettant en jeu les prédictions en fréquences d'une part et les prédictions en amplitude d'autre part. L'application d'un seuil sur les distances $d_f(P_i, P_j)$ et $d_a(P_i, P_j)$ pour décider si les partiels P_i et P_j doivent être appariés peut générer des difficultés. En effet, si les prédictions varient de manière conséquente, l'opération de décision par application d'un seuil doit être plus tolérante que dans le cas où les prédictions sont quasi constantes, voir figure 4.4.

Pour résoudre ce problème, on propose de normaliser la distance euclidienne entre les deux prédictions par la somme des écarts types de ces deux prédictions.

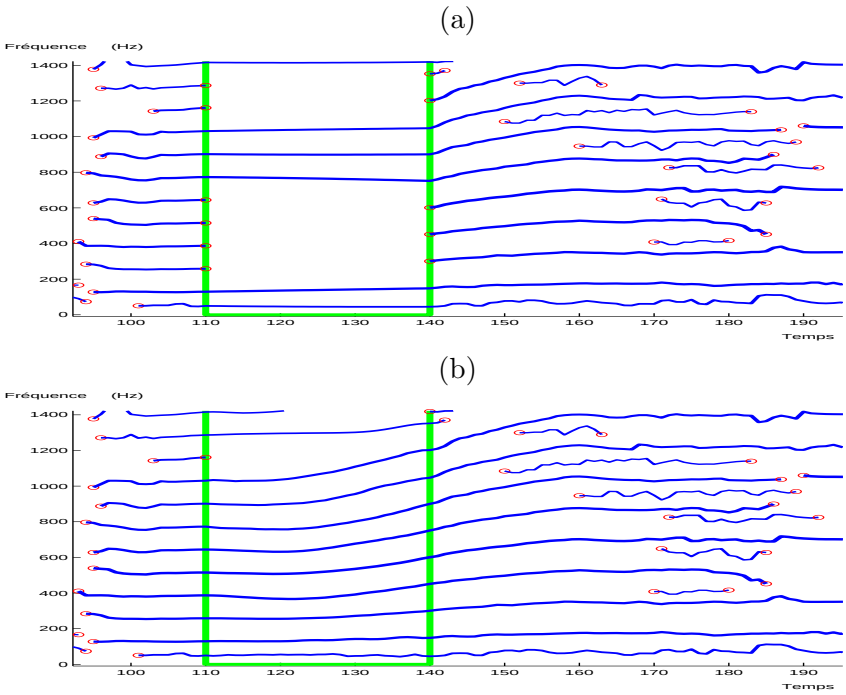


FIG. 4.3 – Résultat du processus d'appariement dans le cas d'un glissando de trombone en utilisant la méthode de référence MAQ (en haut) et la méthode proposée (en bas).

Ces critères C_f et C_a sont définis dans les équations suivantes :

$$C_f = \frac{\sqrt{d_f(P_i, P_j)/(n_3 - n_2)}}{1 + \sigma(\hat{F}_i) + \sigma(\hat{F}_j)} < T_f \tag{4.13}$$

$$C_a = \frac{\sqrt{d_a(P_i, P_j)/(n_3 - n_2)}}{1 + \sigma(\hat{A}_i) + \sigma(\hat{A}_j)} < T_a \tag{4.14}$$

où $\sigma(x)$ est l'écart type du vecteur x et T_f et T_a sont respectivement des seuils en fréquence et en amplitude. Si ces deux critères sont satisfaits pour un couple (P_i, P_j), chaque autre couple candidat dont un élément est P_i ou P_j est enlevé de la liste triée et ces deux partiels sont appariés. La partie manquante du partiel résultant est alors interpolée grâce à la méthode décrite dans la section 4.4. Ce processus est répété jusqu'à ce qu'il n'existe plus de couple satisfaisant les critères pré-cités, voir figure 4.5. Grâce à cet algorithme de décision, l'appariement est correctement opéré dans le cas de modulations, sans appariement non désiré dans les cas stationnaires. Des exemples d'appariement utilisant cette méthode avec le même jeu de seuils T_f et T_a sont donnés par les figures 4.3 et 4.4(b).

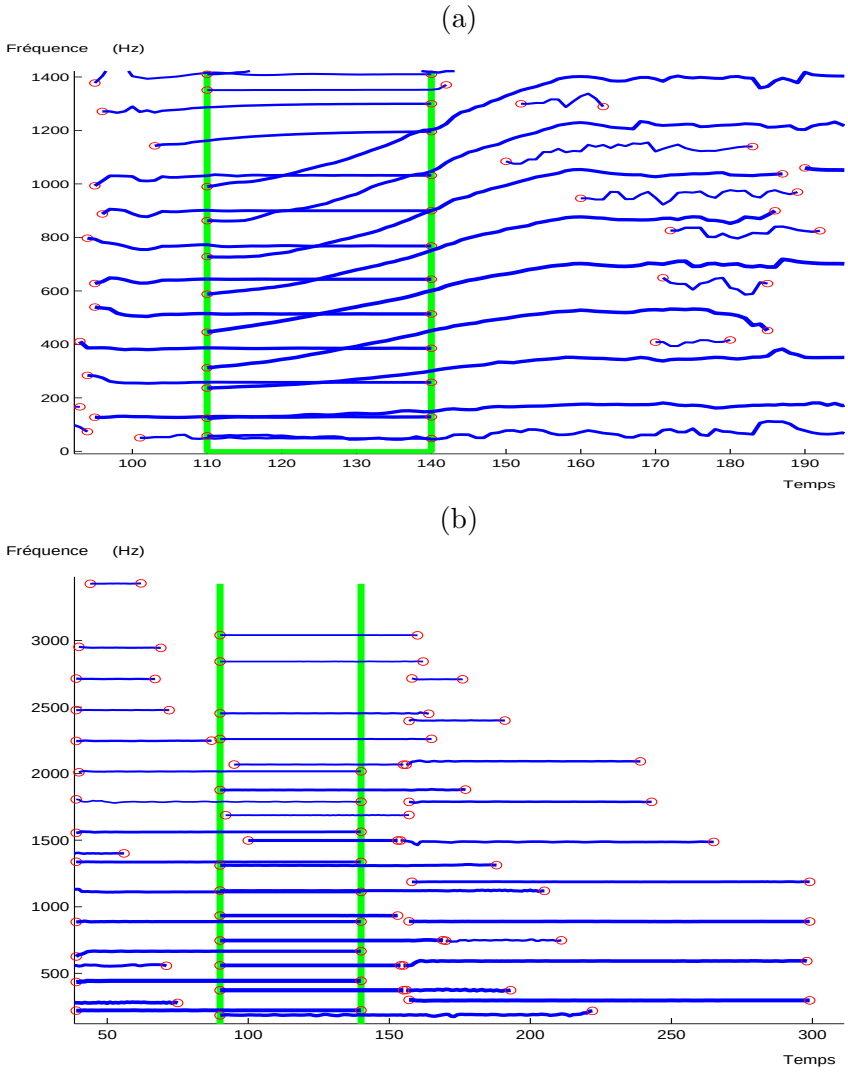


FIG. 4.4 – Prédictions des partiels des deux ensembles pour une note de trombone avec un glissando (en haut) et une transition entre deux notes de piano (en bas). D'une part, la distance euclidienne entre les prédictions des partiels de la note de trombone est élevée et les partiels de $\mathcal{G}$ et $\mathcal{D}$ doivent être appariés. D'autre part, les prédictions des deux notes de piano sont très proches et les partiels ne doivent pas être appariés.

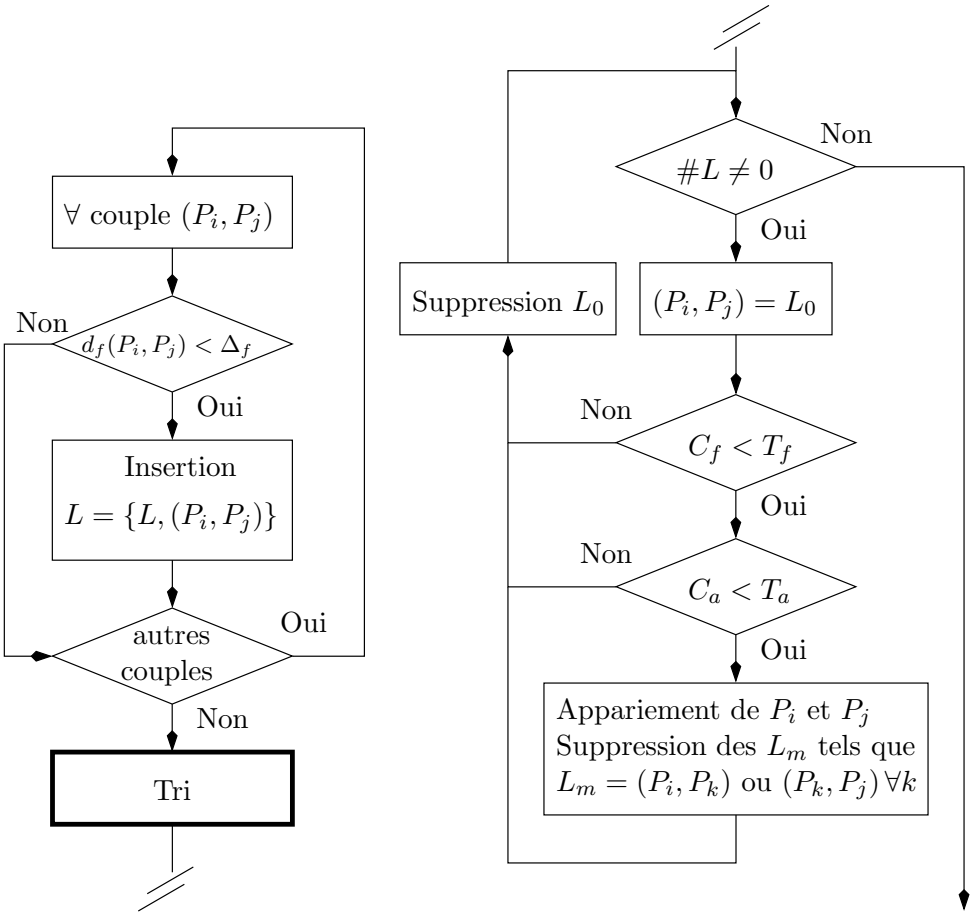


FIG. 4.5 – Schéma de principe de l'algorithme d'appariement. Le schéma de gauche décrit la construction de la liste de couples de partiels et celui de droite, la sélection des couples à apparier parmi les éléments de cette liste.

4.4 Interpolation de partiels

Soit un partiel P né à la trame d'indice n_1 et mort à la trame n_4 résultat de l'appariement des partiels P_i et P_j. Comme P_i se termine à la trame n_2 et P_j débute à la trame n_3, les paramètres de P sont inconnus entre n_2 et n_3, avec $n_1 < n_2 < n_3 < n_4$. On cherche à interpoler ces paramètres manquants $\hat{P}$ en exploitant les paramètres connus de P_i et de P_j :

$$P_i = \{P_i(n), t = n_1, ..., n_2\} \tag{4.15}$$
$$P_j = \{P_j(n), n = n_3, ..., n_4\} \tag{4.16}$$
$$P(n) = \{F(n), A(n), \Phi(n)\} \tag{4.17}$$
$$\hat{P} = \{\hat{P}(n), n = n_2 + 1, ..., n_3 - 1\} \tag{4.18}$$
$$\hat{P}(n) = \{\hat{F}(n), \hat{A}(n), \hat{\Phi}(n)\} \tag{4.19}$$

De manière à interpoler les paramètres de fréquence et d'amplitude dans la partie manquante entre deux partiels P_i et P_j que l'on a appariés, on utilise une combinaison des prédictions $\hat{P}_i$ et $\hat{P}_j$ calculées à partir des deux parties connues. On introduit dans une première partie une méthode originale de combinaison qui permet de favoriser la meilleure prédiction. Dans la deuxième partie, on décrit une méthode qui prend en compte le caractère non stationnaire du paramètre d'amplitude. Enfin, la troisième section introduit une méthode d'interpolation originale de la phase qui assure la continuité aux bornes de la partie manquante. L'application de ces trois méthodes est appelée dans la suite interpolation AR.

4.4.1 Interpolation de la fréquence

Pour calculer les paramètres de fréquence du partiel dans la partie manquante $\hat{F}$ en fonction des deux vecteurs de prédiction $\hat{F}_i$ et $\hat{F}_j$, une combinaison de ces deux vecteurs est opéré en multipliant $\hat{F}_i$ par une fenêtre de pondération $w(t)$ calculée grâce à l'équation 4.24 et $\hat{F}_j$ par $1 - w(t)$:

$$\hat{F}(n) = w\left(\frac{n - n_2}{n_3 - n_2}\right)\hat{F}_i(n) + \left(1 - w\left(\frac{n - n_2}{n_3 - n_2}\right)\right)\hat{F}_j(n) \tag{4.20}$$

Dans le cas de la figure 4.6, l'extrapolation de la partie gauche est moins atténuée que celle de droite, car P_i est plus long que P_j. La fenêtre utilisée doit donc être asymétrique de manière à favoriser l'extrapolation obtenue grâce à des coefficients AR calculés avec le plus de données. La fenêtre calculée grâce à l'équation 4.22 est symétrique. Cette fonction est donc égale à 0.5 au milieu de la partie manquante. La combinaison symétrique obtenue grâce à cette fenêtre de pondération n'est donc appropriée que dans le cas où P_i et P_j sont de même longueur. Si P_i est trois fois plus long que P_j, la fenêtre de pondération $w(t)$ doit

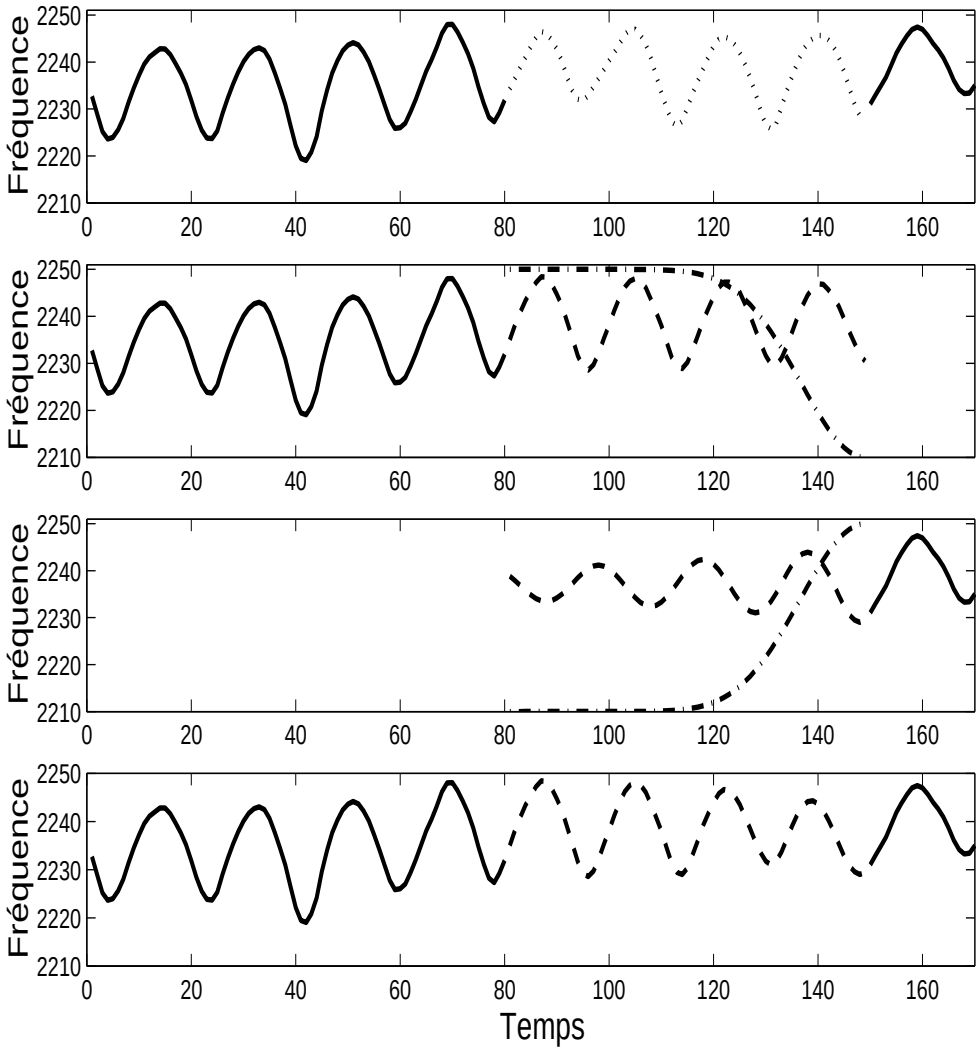


FIG. 4.6 – Interpolation de fréquences manquantes (représentées en haut par des pointillés). On utilise une modélisation AR des fréquences d'un partiel d'une note de saxophone avec vibrato. La partie gauche P_i du partiel est extrapolée en avant et la partie droite P_j en arrière. Ces deux extrapolations sont ensuite combinées en utilisant une fenêtre asymétrique (en trait mixte sur les deux figures du milieu), de manière à favoriser l'extrapolation calculée avec le plus de points (ici, $\hat{F}_i$).

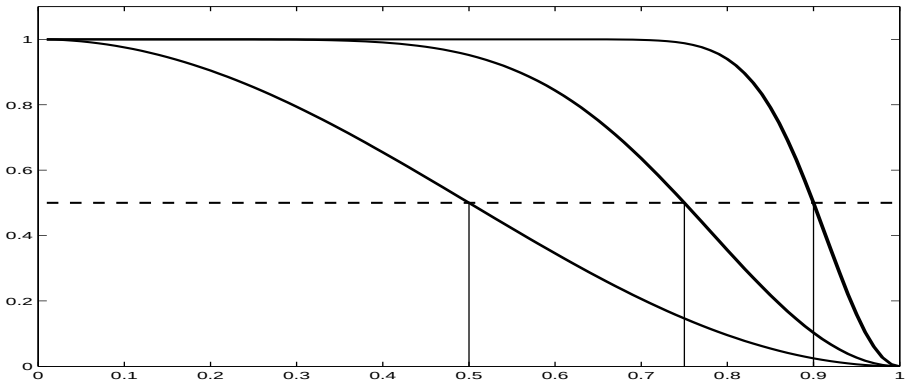


FIG. 4.7 – Trois fenêtres de pondération obtenues grâce à l'équation 4.24 avec, de gauche à droite, $l_i/l_j = [1/2, 1/3, 1/9]$.

atteindre 0.5 aux 3/4 de la partie manquante, voir figure 4.7. D'une manière générale, on souhaite que :

$$w\left(\frac{l_g}{l_g + l_d}\right) = \frac{1}{2} \tag{4.21}$$

Une telle fenêtre asymétrique peut être calculée grâce à l'équation 4.24 avec un facteur d'asymétrie calculé grâce à l'équation 4.23 en fonction des longueurs respectives de P_i et P_j :

$$c(t) = \frac{1 + \cos(\pi.(1+t))}{2} \tag{4.22}$$

$$r(a,b) = \frac{\log(1/2)}{\log(c(\frac{a}{a+b}))} \tag{4.23}$$

$$w(t) = \begin{cases} c\,(t)^{r(l_i,l_j)} & \text{si } l_i > l_j \\ 1 - [1 - c\,(t)]^{r(l_j,l_i)} & \text{sinon} \end{cases} \tag{4.24}$$

où t est compris entre 0 et 1, l_i est la longueur de P_i, l_j est la longueur de P_j et log désigne le logarithme népérien.

4.4.2 Interpolation de l'amplitude

L'amplitude d'un partiel est souvent plus modulée que la fréquence, par exemple dans des signaux de parole. Ceci à pour conséquence que les prédictions ne sont souvent pas utilisables en l'état. En effet, même si les micro-modulations sont modélisées correctement, les prédictions à long terme sont peu satisfaisantes.

L'extrapolation en amplitude du partiel P_i est contrainte à atteindre à la trame d'indice n_3 une amplitude égale à l'amplitude moyenne du partiel P_j calculée de la trame d'indice n_3 à celle d'indice $\min(n_3 + M, n_4)$. Le paramètre M doit être choisi de manière à obtenir une estimation de l'énergie du début

du partiel P_j. Dans la configuration présentée dans la section 3.6.2, M peut être fixé à 30. Cette contrainte peut être obtenue par l'ajout d'un incrément $\delta_i(t)$ calculé grâce à l'équation 4.25. La même contrainte est appliquée à $\hat{a}_j$ par l'ajout de $\delta_j(t)$ calculé grâce à l'équation 4.26.

$$\delta_i(n) = \frac{n-n_2}{n_3-n_2}\left(\frac{\sum_{\tau=n_3}^{\min(n_3+M,n_4)} A_j(\tau)}{\min(n_3+M,n_4)-n_3} - \hat{A}_i(n_3)\right) \tag{4.25}$$

$$\delta_j(n) = \frac{n_3-n}{n_3-n_2}\left(\frac{\sum_{\tau=\min(n_2-M,n_1)}^{n_2} A_i(\tau)}{n_2-\min(n_2-M,n_1)} - \hat{A}_j(n_2)\right) \tag{4.26}$$

La combinaison de ces amplitude corrigées est ensuite opérée comme suit :

$$\hat{A}(n) = w\left(\frac{n-n_2}{n_3-n_2}\right)\left(\hat{A}_i(n)+\delta_i(n)\right) + \left(1-w\left(\frac{n-n_2}{n_3-n_2}\right)\right)\left(\hat{A}_j(n)+\delta_j(n)\right) \tag{4.27}$$

4.4.3 Interpolation de la phase

En utilisant la méthode d'interpolation polynomiale décrite dans [QD90], la phase est interpolée grâce au polynôme d'ordre 3 le plus lisse possible qui satisfait les quatre contraintes aux bornes : $F_i(n_2)$, $\Phi_i(n_2)$ et $F_j(n_3)$, $\Phi_j(n_3)$. La fréquence est alors contrainte par l'évolution de la phase.

Inversement, on dispose d'informations de fréquences que l'on souhaite conserver tout en obtenant une phase cohérente et qui n'engendre pas de discontinuités aux bornes. On propose d'intégrer la phase par la méthode des trapèzes, un incrément ρ est ajouté aux phases calculées pour assurer la continuité de phase aux bornes. On note $\varphi(n)$ la phase déroulée à la trame d'indice n, ($\Phi(n) = \varphi(n) \bmod 2\pi$). Considérant $\Phi(n_2)$, $\Phi(n_3)$ et $F(n_2), \hat{F}(n_2+1)\cdots\hat{F}(n_3-1), F(n_3)$ les fréquences soit connues soit interpolées, les phases interpolées peuvent être en première approximation calculées comme suit :

$$\hat{\varphi}(n) = \varphi(n_2) + 2\pi\,\Delta_T \sum_{\tau=n_2}^{n} \frac{F(\tau-1)+F(\tau)}{2} \tag{4.28}$$

où $n \in]n_2, n_3]$ et Δ_T est le pas d'échantillonnage du modèle long-terme exprimé en secondes. Toutefois, une discontinuité de phase peut apparaître à la fin de la partie manquante : $\hat{\Phi}(n_3) \neq \Phi(n_3)$. Un incrément de phase ρ, calculé grâce à l'équation 4.29, est alors ajouté de manière à satisfaire la contrainte de continuité de phase à la fin de la partie manquante. Les phases sont alors calculées en utilisant l'équation 4.30.

$$\rho = \begin{cases} e_\Phi + 2\pi & \text{si } |e_\Phi| > |e_\Phi + 2\pi| \\ e_\Phi - 2\pi & \text{si } |\hat{e}_\Phi| > |e_\Phi - 2\pi| \\ e_\Phi & \text{sinon} \end{cases} \tag{4.29}$$

où $e_\Phi = \hat{\Phi}(n_3) - \Phi(n_3)$.

$$\hat{\varphi}(n) = \varphi(n_2) + \rho \frac{n - n_2}{n_3 - n_2} + 2\pi \, \Delta_T \sum_{\tau=1}^{n} \frac{F(\tau - 1) + F(\tau)}{2} \tag{4.30}$$

où $n \in]n_2, n_3]$ et Δ_T est la taille du pas d'échantillonnage du modèle long-terme en secondes.

4.4.4 Évaluation

L'utilisation de modèles hybrides de type SN ou STN (voir section 1.3) requiert une synthèse du modèle sinusoïdal à long terme proche du signal original de manière à obtenir un signal résiduel d'énergie minimale. Dans une première partie, on évalue de manière objective les capacités des deux algorithmes d'interpolation présentés dans la section précédente par calcul de la dégradation induite par l'interpolation. Dans des applications de type restauration d'enregistrements ou dans la transmission numérique en "*streaming*", la qualité subjective (jugée par l'auditeur) est déterminante. La seconde partie décrit la méthodologie et exploite les résultats de tests d'écoute subjectifs effectués à France Télécom R&D.

Évaluation objective

Pour ce test, une partie manquante dans une représentation à long terme est simulée en supprimant des pics d'indices consécutifs aux partiels qui sont nés avant et morts après la partie manquante. Les autres partiels sont laissés en l'état, voir figure 4.8. Les paramètres de ces pics supprimés sont alors interpolés selon les deux méthodes. La dégradation induite par l'interpolation est évaluée grâce au SNR de Reconstruction (R-SNR) calculé grâce à l'équation 3.49, où $x(n)$ est le signal temporel synthétisé à partir de la représentation à long terme originale et $\hat{x}(n)$ celui synthétisé à partir de la représentation à long terme reconstruite. Pour chaque taille de partie manquante, le résultat présenté sur la figure 4.9 est le R-SNR moyen pour toutes les positions possibles de la partie manquante.

En utilisant une modélisation AR des paramètres de fréquence et d'amplitude, les modulations musicales sont préservées. Les résultats sont donc meilleurs que ceux de l'interpolation polynomiale pour les deux premiers signaux qui comportent respectivement un vibrato ou un trémolo, voir figure 4.9. Les résultats sont équivalents dans le cas stationnaire, par exemple dans le cas d'une note de clavecin.

Évaluation subjective

Les deux méthodes sont ici évaluées grâce à un test de qualité subjective. Quatre signaux sont utilisés : une note de saxophone, une note de vibraphone, un enregistrement d'une voix chantée de soprano et une pièce d'orchestre. Les parties manquantes sont de tailles variables, de 80 à 820 ms. Pour chaque signal et pour chaque taille de partie manquante, il est demandé à 10 experts d'écouter

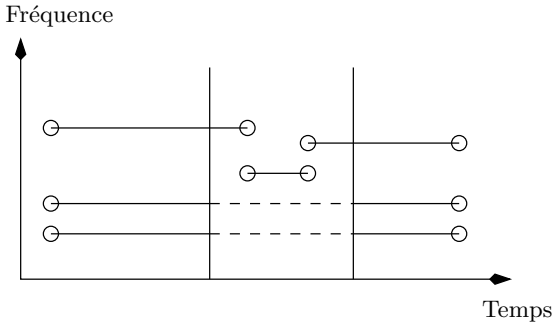


FIG. 4.8 – De manière à ne tester que la capacité d'interpolation des paramètres des partiels, seuls les partiels nés avant et morts après la partie manquante ont leurs paramètres interpolés durant la partie manquante. Les autres sont laissés en l'état.

le signal temporel issu de la synthèse de la représentation à long terme originale. Ce signal original constitue une référence explicite. Après cette première écoute, il est ensuite demandé aux experts de juger la qualité de quatre versions, une version dégradée sans interpolation, une version restaurée par une interpolation polynomiale, une version restaurée avec une interpolation AR et l'original constituant ainsi une référence cachée. L'échelle de notation à 100 points de Mushra est utilisée. Les notes obtenues sont affichées sur la figure 4.10.

Comme on peut le voir en haut de la figure 4.10, l'utilisation d'une modélisation AR permet une interpolation de haute qualité dans le cas de signaux musicaux modulés (saxophone et vibrato) avec des parties manquantes longues de près de 1 seconde. Les signaux modulés de manière plus complexe comme la voix chantée ou les enregistrements polyphoniques sont plus difficiles à interpoler de manière réaliste. Néanmoins, la méthode proposée apporte un gain significatif par rapport à la méthode polynomiale comme on peut le constater en bas de la figure 4.10.

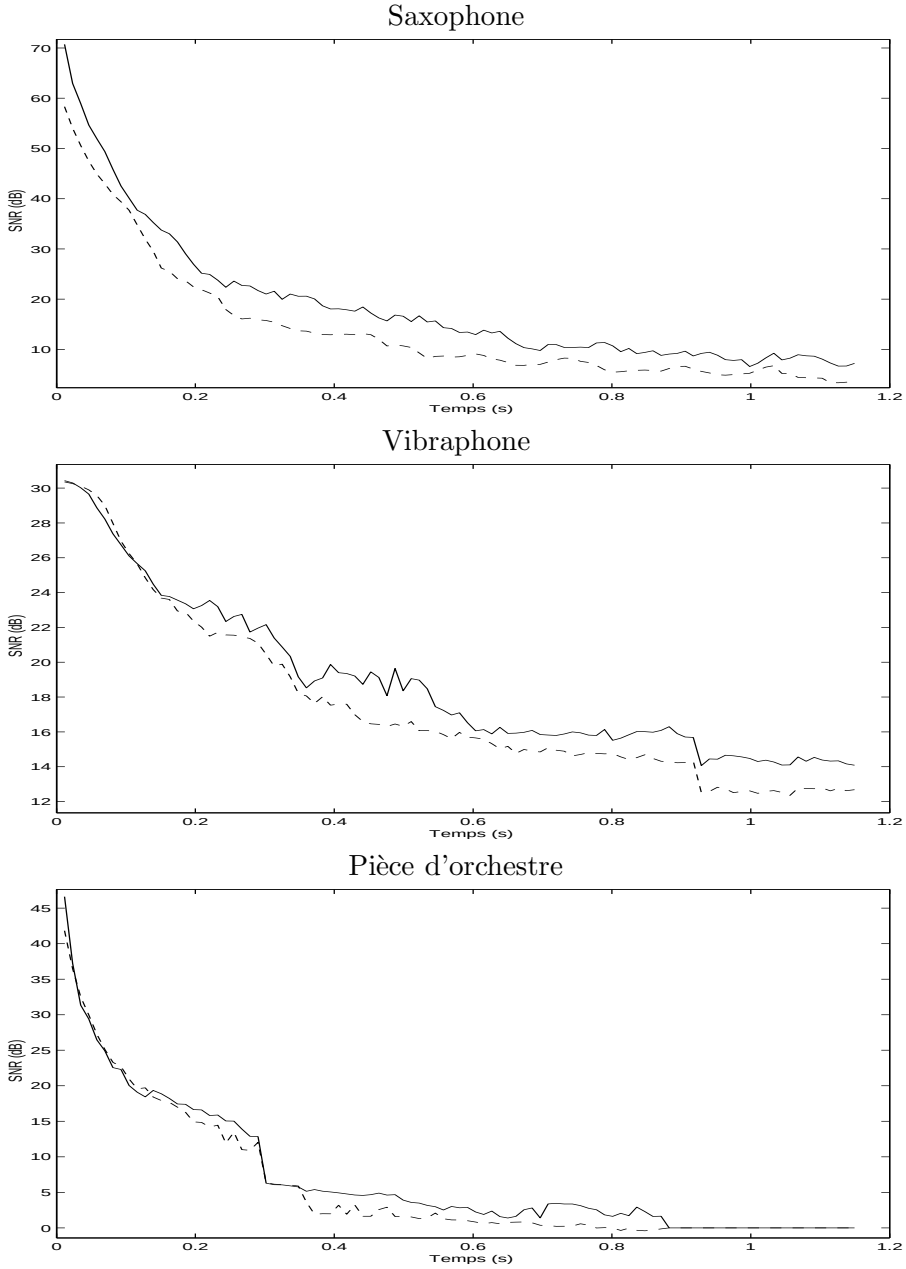


FIG. 4.9 – Comparaison objective de l'interpolation AR (trait plein) et l'interpolation polynomiale (tirets). Trois signaux sont utilisés : une note de saxophone avec vibrato (en haut), une note de vibraphone (au milieu) et une note de clavecin (en bas). La méthode proposée atteint un meilleur R-SNR dans le cas de modulations (saxophone et vibraphone) et un R-SNR équivalent dans un cas stationnaire (clavecin).

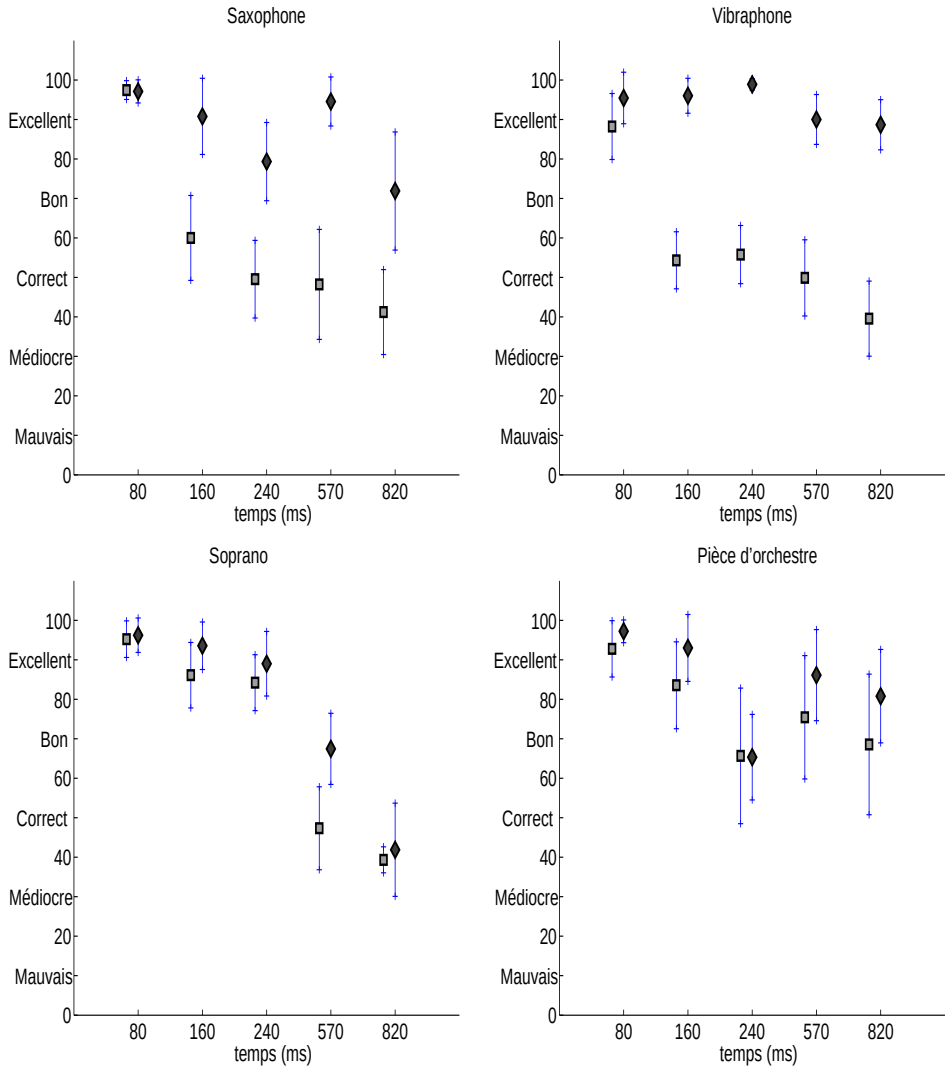


FIG. 4.10 – Résultats moyens des tests d'écoute comparant la méthode d'interpolation polynomiale (carrés) et la méthode d'interpolation AR (diamants) pour cinq tailles de partie manquante. Pour chaque méthode, le symbole désigne la moyenne des scores et les lignes les intervalles de confiance.

4.5 Extrapolation des partiels non appariés

Si les partiels de l'ensemble $\mathcal{G}$ et ceux de l'ensemble $\mathcal{D}$ sont correctement appairés, les partiels non appairés de $\mathcal{G}$ appartiennent à une note qui se termine alors que ceux de $\mathcal{D}$ appartiennent à une note qui a commencé pendant la partie manquante, voir figure 4.1.

On propose d'extrapoler ces partiels non appariés en considérant les prédictions des paramètres d'amplitude et de fréquence. Les fréquences prédites sont utilisées en l'état, et les phases correspondantes sont calculées grâce à l'équation 4.28. On définit deux paramètres : $l_{\mathcal{G}}$ la longueur maximale d'une extrapolation d'un partiel de l'ensemble $\mathcal{G}$ et $l_{\mathcal{D}}$ la longueur maximale d'une extrapolation d'un partiel de l'ensemble $\mathcal{D}$. L'amplitude extrapolée $\tilde{A}_i(n)$ est définie comme l'amplitude prédite $\hat{A}_i(n)$ pondérée grâce à l'équation 4.33 :

$$n_e = \min(n_2 + l_{\mathcal{G}}, n_3) \tag{4.31}$$

$$\gamma_i(n) = \frac{n - n_2}{n_3 - n_2} \max(\hat{A}_i(n_e), 0) \tag{4.32}$$

$$\tilde{A}_i(n) = \hat{A}_i(n) - \gamma_i(n) \tag{4.33}$$

pour n de n_2 à n_3. Si $\tilde{A}_i(n+1) < 0$ avec $t+1 < n_e$, le partiel P_i se termine à la trame d'indice t. De même, $\tilde{A}_j(n)$ est définie comme l'amplitude prédite $\hat{A}_j(n)$ pondérée grâce à l'équation 4.36 :

$$n_b = \max(n_3 - l_{\mathcal{D}}, n_2) \tag{4.34}$$

$$\gamma_j(n) = \frac{n_3 - n}{n_3 - n_2} A_j(n_b) \tag{4.35}$$

$$\tilde{A}_j(n) = \hat{A}_j(n) - \gamma_j(n) \tag{4.36}$$

pour n de n_2 à n_3. Le partiel P_j extrapolé débute au dernier indice de trame t tel que $\tilde{A}_j(n-1) \leq 0$.

Obtenir une extrapolation réaliste de l'amplitude d'un partiel est un problème difficile. En particulier, le système doit être capable de décider à qu'elle trame un partiel se termine s'il appartient à $\mathcal{G}$ et à quelle trame commence le partiel s'il appartient à $\mathcal{D}$.

En général, les amplitudes des partiels ont une évolution assez prédictible à la dernière partie de la note (phase de soutien ou de relâchement). Les amplitudes prédites peuvent donc être exploitées pour détecter quand le partiel doit se terminer. C'est pourquoi γ_i est fixé à la valeur maximale entre $\hat{A}_i(n_e)$ et 0, voir équation 4.32. En conséquence, les partiels extrapolés peuvent se terminer avant n_e, comme illustré sur la figure 4.1.

Au contraire, les amplitudes des partiels durant une attaque abrupte (celle du piano par exemple) sont très riches en informations. Ces évolutions ne peuvent être déduites des évolutions des partiels durant les parties tenues (soutien et relâchement) de la note. De manière à simuler une attaque, quasiment tous les partiels non appariés de l'ensemble $\mathcal{D}$ et appartenant à la même entité sonore, doivent commencer en même temps. C'est pourquoi γ_j est fixé en fonction de $A_j(n_b)$, voir équation 4.35.

Les paramètres $l_{\mathcal{G}}$ et $l_{\mathcal{D}}$ doivent être choisis en fonction de l'application. Pour des applications de type interpolation de perte d'informations d'une représentation sinusoïdale à long terme due à une défaillance de transmission sur un réseau (*streaming* temps réel), la taille maximale d'une partie manquante est généralement faible à cause des capacités limités de la mémoire tampon du décodeur. Dans cette optique, l'extrapolation doit être paramétrée de manière à être tolérante aux erreurs potentielles d'appariement. Ceci peut être fait en fixant $l_{\mathcal{G}} = l_{\mathcal{D}} = n_3 - n_2$, de manière à assurer un fondu des extrapolations des partiels non appariés. Au contraire, durant la phase d'analyse ou dans une application en temps différé de type restauration d'enregistrement audionumériques, des informations auxiliaires peuvent être exploitées pour estimer l'indice de trame auquel les partiels non appariés de l'ensemble $\mathcal{D}$ doivent débuter. Le paramètre $l_{\mathcal{D}}$ doit alors être fixé à une valeur particulière pour chaque partie manquante, en fonction d'une estimation de la position de l'attaque.

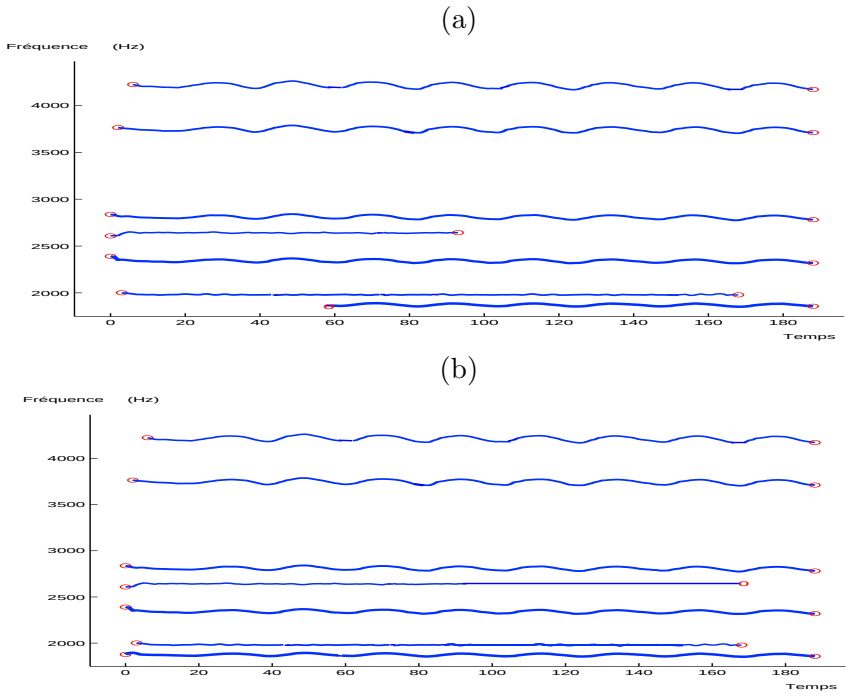


FIG. 4.11 – Application des algorithmes de restauration des sections 4.3 et 4.5 à la restauration d'entités sonores. En haut, résultat de la phase d'appariement des partiels de la figure 4.2(b). En bas, résultat de la phase d'extrapolation.

4.6 Application à la restauration d'entités sonores

Lors de l'extraction d'entités sonores détaillée dans le chapitre 5, certains partiels composant ces entités peuvent être segmentés. Nous sommes dans le cas asynchrone car les parties manquantes ne sont pas de même longueur et sont placées à des positions différentes, comme le montre la figure 4.2(b). Certaines informations structurelles peuvent être exploitées, car si l'entité sonore considérée est harmonique, cet indice peut être exploité pour le problème de l'appariement comme cela est proposé dans [RR04]. Néanmoins, une application directe des méthodes présentées dans les sections précédentes donne des résultats satisfaisants.

La phase d'appariement considère tous les couples possibles de partiels présents. L'utilisation de l'équation 4.12 lors du tri favorise implicitement les parties manquantes de taille faible. Le résultat de cette phase d'appariement des partiels de la figure 4.2(b) est présenté sur la figure 4.11(a). Lors de l'extraction d'entités sonores, on dispose des indices de début et de fin de l'entité. Ces indices sont exploités pour extrapoler les partiels qui commencent après ou se termine avant ces indices de début et de fin, voir figure 4.11(b).

4.7 Application à la restauration d'enregistrements musicaux

La restauration d'enregistrements musicaux consiste généralement à interpoler des échantillons temporels consécutifs. On compare ici trois méthodes de restauration d'enregistrements musicaux grâce à des tests subjectifs utilisant le même protocole que celui utilisé dans la section 4.4. La première méthode est dite temporelle. Les 2000 échantillons temporels des deux côtés de la partie manquante sont utilisés pour calculer pour chacune des parties 1000 coefficients AR en utilisant la méthode de Burg. Les extrapolations obtenues par filtrage IIR sont combinées grâce à la fenêtre de pondération de l'équation 4.22.

Les deux autres sont basées sur un modèle sinusoïdal à long terme. La première est une adaptation de la méthode proposée par Mc Aulay et Quatieri [MQ86]. L'appariement de partiels est effectué grâce à l'équation 4.11 avec $\Delta_f =$ 40 Hz. Les paramètres de phase et de fréquence durant la partie manquante sont obtenus grâce à un polynôme interpolateur d'ordre 3 tandis que l'amplitude est interpolée linéairement. La dernière est la méthode proposée (notée méthode AR). L'appariement est effectué grâce à l'algorithme décrit dans la section 4.3 avec $T_f = 0.5$ et $T_a = 0.1$. L'interpolation des paramètres manquants est opérée grâce à la méthode décrite dans 4.4. L'extrapolation est effectuée comme décrit dans la section 4.5 avec $l_{\mathcal{G}} = n_3 - n_2$ et $l_{\mathcal{D}} = n_2 + (n_3 - n_2)/2$.

Cinq signaux audionumériques sont utilisés pour ces tests. Une note de violon avec un vibrato, une note de piano, une pièce d'orchestre, une note de gong et l'enregistrement de deux sopranos. Pour les résultats affichés sur la figure 4.13, une partie d'une note est supprimée, tandis que pour ceux présentés sur la figure 4.12, une phase de transition entre deux notes est supprimée.

L'interpolation temporelle par modélisation AR est adéquate pour l'interpolation de segments de centaines d'échantillons de signaux audionumériques de qualité CD [KKS01, KR02]. Pour des parties manquantes plus grandes, la qualité de l'interpolation dépend notamment de la complexité et de la stationnarité du signal. Si le signal est composé de partiels très stationnaires comme une note de piano par exemple, les erreurs de prédiction associées aux deux vecteurs de coefficients AR sont assez faibles. Par conséquent, l'effet d'atténuation de la prédiction est assez faible. En revanche, si le signal interpolé consiste en le même nombre d'harmoniques (aux environs de 10) mais cette fois-ci modulées par un vibrato, l'atténuation est extrêmement prononcée, car le signal ne correspond plus au modèle, voir figure 4.15(a). Ce phénomène d'atténuation explique pourquoi les scores obtenus par cette méthode sont mauvais dans le cas de modulations prononcées (scores compris entre 30 et 50). L'utilisation d'une représentation sinusoïdale permet de répondre à ce problème d'atténuation, comme on peut le constater sur les figures 4.15(b) et 4.15(c).

La méthode polynomiale se comporte mieux que la méthode temporelle pour de segments supprimés de taille supérieure à 320 ms, voir figure 4.13. En contrepartie, toute forme de modulation disparaît. Le son est perçu par les auditeurs comme "figé" durant l'interpolation. Pour des parties manquantes plus

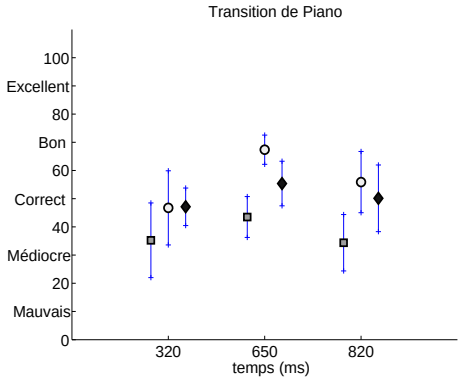


FIG. 4.12 – Résultats des tests d'écoute subjectifs comparant la méthode polynomiale (carrés), la méthode LP (diamants) et la méthode temporelle (cercles) pour trois tailles de partie manquante placée pendant une transition entre deux notes de piano. Pour chaque méthode, le symbole désigne la moyenne des votes et les lignes les intervalles de confiance.

grandes, cette méthode est jugée artificielle par les auditeurs. Les scores obtenus peuvent être alors moins bons que ceux obtenus par la méthode temporelle. Ce cas apparaît lors de l'interpolation d'une partie supprimée de 820 ms de la note de violon.

La méthode proposée tire partie des avantages des deux méthodes. L'utilisation d'une représentation sinusoïdale permet d'éviter tout problème d'atténuation. L'interpolation de partie manquante de taille conséquente est donc possible. De plus, la modélisation AR des paramètres de fréquence et d'amplitude permet de préserver les modulations durant l'interpolation. Les partiels de la note de gong et des deux sopranos ont des modulations qui sont assez faibles tandis que ceux de la note de violon sont modulés de manière conséquente. Pour ces signaux de tests, la méthode obtient des scores de 90 à 70 avec une décroissance régulière pour des tailles de 320 à 820 ms. Les voix de soprano peuvent même être interpolées durant 1.6 secondes avec un score *bon*. Les partiels de la pièce d'orchestre ont des modulations complexes à cause de la présence de bruit et de nombreuses harmoniques. Les capacités d'interpolation sont donc plus faibles que dans les cas précédents, mais un score *correct* est obtenu pour des tailles allant jusqu'à 450 ms.

Si la partie manquante est placée durant une phase de transition, des informations importantes sont perdues, en particulier l'attaque de la nouvelle note, qui est très difficile à simuler. La qualité de l'interpolation est donc moins bonne. Dans cette situation, la méthode temporelle est la mieux notée, probablement à cause de l'effet d'atténuation évoqué précédemment qui simule un effet de fondu centré au milieu de la partie manquante. En ce qui concerne les méthodes sinusoïdales, l'utilisation d'un algorithme d'appariement robuste (dans ce cas, les partiels des deux notes sont pas appairés) apporte un gain non négligeable, voir figure 4.12.

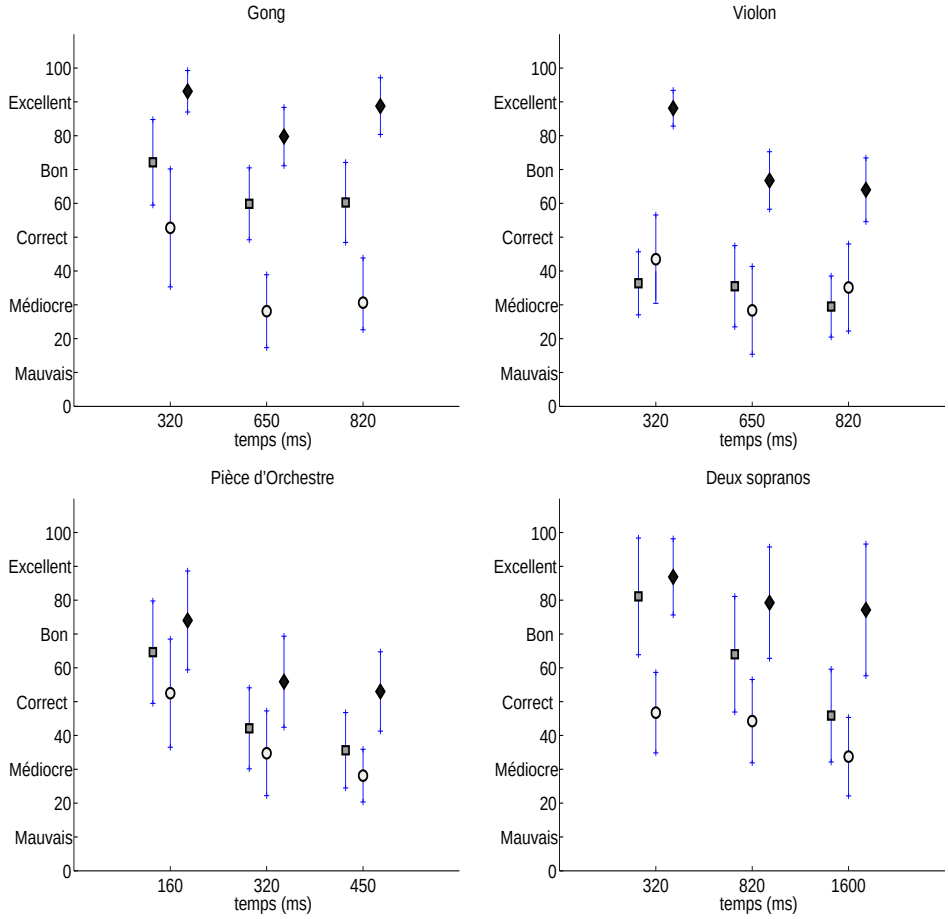


FIG. 4.13 – Résultats des tests d'écoute subjectifs comparant la méthode polynomiale (carrés), la méthode LP (diamants) et la méthode temporelle (cercles) pour trois parties manquantes de tailles différentes. Pour chaque méthode, le symbole désigne la moyenne des votes et les lignes les intervalles de confiance.

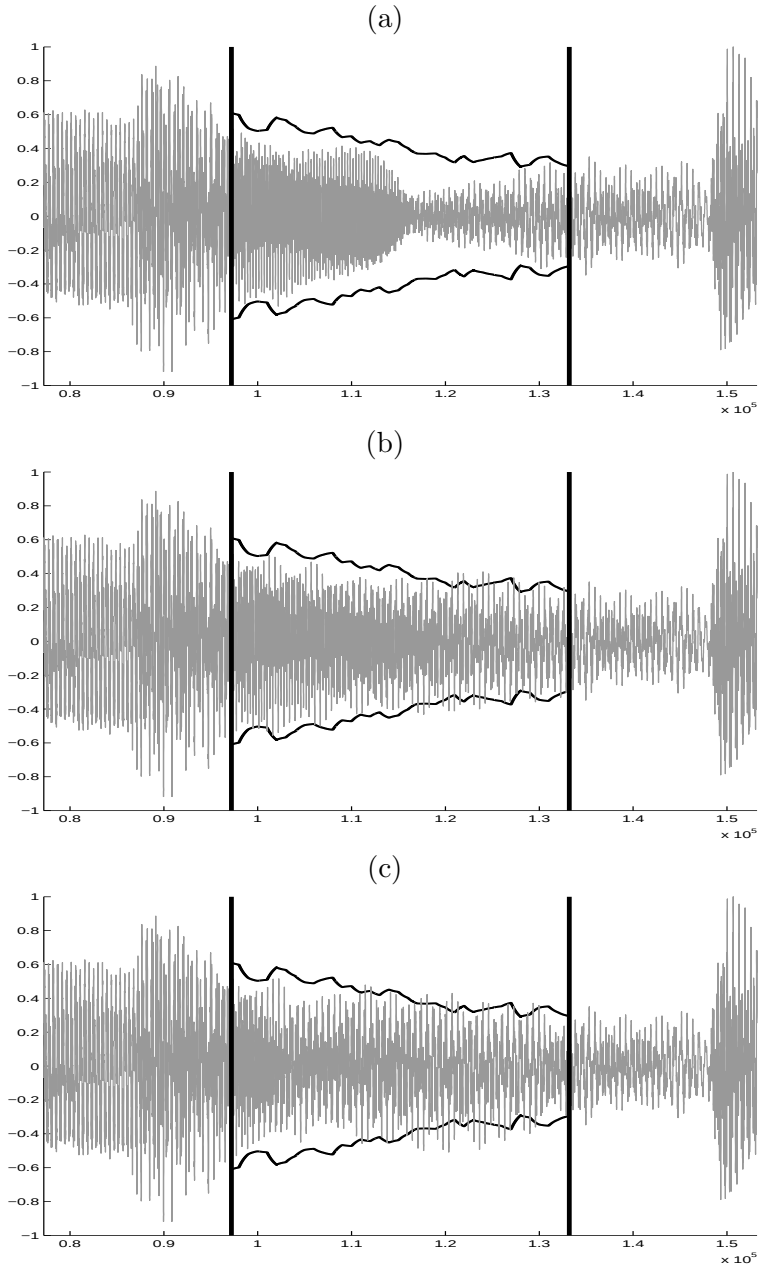


FIG. 4.14 – Une note de piano est interpolée durant 820 ms grâce aux trois méthodes présentées. Les deux lignes verticales marquent les limites de la zone interpolée. L'enveloppe du signal original est symbolisée par les deux lignes brisées symétriques. Les méthodes temporelle (a), polynomiale (b) et AR (c) sont utilisées pour restaurer la partie manquante.

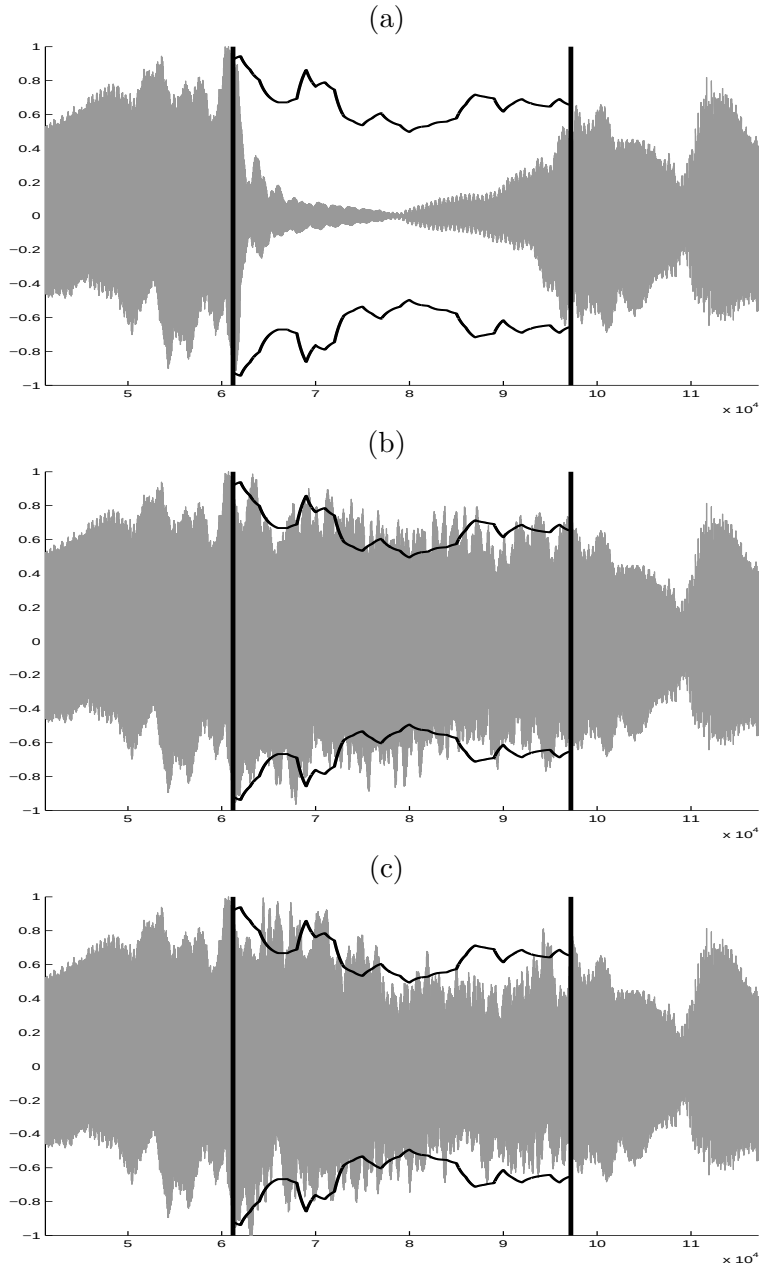


FIG. 4.15 – Une note de violon avec vibrato est interpolée durant 820 ms grâce aux trois méthodes présentées. Les deux lignes verticales marquent les limites de la zone interpolée. Les deux lignes brisées désigne une estimation de l'enveloppe du signal original. De haut en bas, les méthodes temporelle, polynomiale et AR sont utilisées. Les deux dernières méthodes, basées sur un modèle sinusoïdal à long terme, ne sont pas soumises au phénomène d'atténuation.

5 Extraction d'entités sonores

Les nouvelles possibilités d'interprétation de la représentation à long terme offertes par les algorithmes présentés dans le chapitre 3 sont exploitées dans ce chapitre. On étudie d'abord plusieurs indices issus d'études psychoacoustiques qui permettront de détecter quels partiels, une fois regroupés, seront perçus par l'auditeur comme une entité perceptuelle. On montre ensuite pourquoi la notion de continuité de la représentation à long terme permet de simplifier le processus de séparation et de mieux détecter les débuts de notes dont l'attaque est douce. Un premier niveau de structuration est apporté par le regroupement des partiels apparus simultanément. Les partiels de ces groupes sont ensuite classés en entités selon deux critères différents. On propose un premier critère basé sur la détection préalable de la fréquence de fondamentale associée à une entité dont les partiels ont des fréquences en relation harmonique. On propose ensuite un algorithme de classification basé sur un critère plus générique : l'évolution corrélée des paramètres des partiels. On montre alors que la modélisation autorégressive des paramètres est appropriée à la définition d'une mesure de similarité exploitant ce critère.

5.1 Introduction

Une entité sonore est une entité perceptuelle, c'est-à-dire qu'une telle entité est perçue par le système auditif humain comme un son simple ou complexe mais unique. En effet, l'auditeur n'est pas capable dans des conditions normales d'isoler les différentes composantes de cette entité.

Dans un modèle sinusoïdal, une entité sonore est un ensemble de partiels présentant certaines corrélations. À un niveau perceptif, ce sont ces différentes corrélations qui amènent le système auditif humain à percevoir non plus un ensemble de sons simples mais un unique son complexe. Ces indices de corrélations ont fait l'objet de nombreuses études psychoacoustiques, synthétisées dans la thèse de Mellinger [Mel91] que l'on exploitera au cours de ce chapitre. On fera aussi amplement référence à l'ouvrage de Bregman [Bre90] auquel le lecteur est invité à se référer pour approfondir ce domaine très riche.

Indices pour le regroupement de partiels

On peut distinguer plusieurs indices qui permettent de regrouper des partiels en entités sonores :

- l'apparition simultanée ;
- la relation d'harmonicité ;
- les évolutions corrélés des paramètres ;
- la position spatiale.

L'apparition simultanée de différents partiels est un indice fort du fait que ces partiels appartiennent à la même entité sonore. Cet indice fera donc l'objet d'une attention particulière dans la suite. Les relations d'harmonicité entre les partiels sont aussi un indice important car ces relations proviennent des propriétés physiques de certains instruments étudiées dans le chapitre 1. Malheureusement, de nombreux instruments sont inharmoniques comme le triangle, voir figure 5.3.

Grâce au modèle à long terme, des indices plus génériques peuvent être utilisés. Les partiels d'une même entité sonore ont des paramètres de fréquence et d'amplitude qui varient de manière corrélée même en cas d'inharmonicité. On appellera dans la suite variation (sans plus de précision), une variation forte du paramètre comme un vibrato ou un portamento pour la fréquence. Par contre, si les variations sont faibles, elles seront appelées "micro-modulations" [McA84]. Comme il est expliqué dans [McA89], ces deux types d'évolutions sont très importantes pour la perception. Des expériences ont montré que plus ces variations ou micro-modulations sont prononcées, plus les partiels soumis à ces modulations se combinent facilement en une entité sonore.

Les partiels d'une même entité sonore ont été émis par le même instrument. La position spatiale des partiels dans la scène sonore est donc un indice pertinent pour le regroupement de partiels en entités sonores. Toutefois, ces informations de position ne font pas partie du modèle sinusoïdal à long terme utilisé dans cette thèse. En revanche, la notion de continuité du modèle à long

terme rend son utilisation particulièrement appropriée à notre problématique, comme détaillé dans la suite.

Heuristique "ancien et nouveau"

Le système auditif humain est capable d'isoler des entités sonores dans des mélanges sonores complexes. Ce système utilise vraisemblablement les indices présentés ci-dessus mais aussi une capacité à interpréter une partie du son comme la continuation des sons passés. Ce système dispose probablement d'une mémoire des événements passés qui permet d'écarter tout ce qui ressemble ou qui peut se déduire de ce passé et de se concentrer sur la partie nouvelle (non expliquée) du son. Cette heuristique est formulée par Bregman [Bre90] : " Si vous pouvez interpréter de manière plausible une partie d'un groupe de composantes acoustiques comme une continuation d'un son qui vient de débuter, faites le et enlevez la du mélange. Ensuite, prenez la différence entre le son d'origine et l'extrapolation à partir du son passé comme le nouveau groupe à analyser.[1]"

Cette heuristique permet de comprendre pourquoi il est aisé de dissocier deux entités sonores qui n'ont pas débuté en même temps. Par application de cette heuristique, l'analyse de ce signal polyphonique revient en quelque sorte à l'analyse décalée de deux signaux monophoniques. On isole la première entité et on la supprime du signal original. Le signal résultant ne contient donc qu'une seule entité. Ce signal est donc monophonique et par conséquent plus simple à analyser.

L'utilisation d'un modèle à long terme est alors particulièrement pertinente, car ce modèle intègre dans son formalisme la notion de continuité et permet au fur et à mesure de l'identification de nouvelles entités sonores, de supprimer les partiels appartenant à ces entités et de simplifier ainsi grandement l'interprétation de l'ensemble des partiels restants.

Algorithme d'extraction d'entités sonores

L'utilisation d'une représentation à long terme du contenu spectral et l'application de l'heuristique "ancien et nouveau" permettent de proposer un algorithme itératif d'extraction d'entités sonores, dont le schéma de principe est représenté sur la figure 5.2. On cherche, dans l'ensemble des partiels $\mathcal{S}$, une apparition simultanée de partiels à une trame donnée. On regroupe alors dans $\mathcal{G}_1$ les partiels qui sont apparus ensemble. De ce groupe sont ensuite extraits les entités sonores $\mathcal{E}_1^1, ..., \mathcal{E}_1^n$, n étant le nombre d'entités trouvées dans le groupe $\mathcal{G}_1$, par recherche des partiels en relation harmonique ou ayant des paramètres évoluant de manière corrélée.

Ces entités sonores sont extraites d'une représentation long-terme qui n'est pas exempte de défauts. Ces entités sont donc reconstruites grâce à un algorithme de restauration d'entités sonores présenté dans le chapitre 4. Les partiels de $\mathcal{S}$ appartenant à ces entités restaurées sont ensuite enlevés et une nouvelle apparition simultanée est recherchée pour former $\mathcal{G}_2$.

[1] traduction de l'auteur

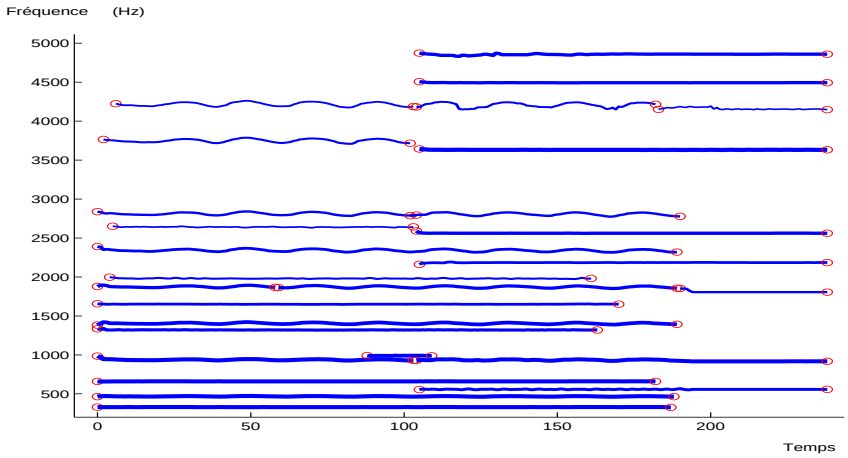


FIG. 5.1 – Représentation à long terme d'un mélange de deux flûtes jouant simultanément (l'une avec un vibrato et l'autre tenue) et d'un triangle.

La figure 5.3 présente un exemple d'application de cet algorithme sur un mélange de trois entités sonores : deux flûtes jouant simultanément (l'une avec un vibrato et l'autre tenue) et un triangle. Du premier groupe $\mathcal{G}_1$ (en trait plein sur la figure 5.2(a)) sont extraites deux entités sonores $\mathcal{E}_1^1$ et $\mathcal{E}_1^2$ (en trait plein sur la figure 5.2(c)) grâce à un critère d'harmonicité. Les partiels appartenant à ces deux entités sont supprimés de $\mathcal{S}$ avant d'obtenir un nouveau groupe $\mathcal{G}_2$ (en trait plein sur la figure 5.2(c)). De ce groupe est extrait l'entité $\mathcal{E}_2^1$ (en trait plein sur la figure 5.2(d)) par des critères de similarité de d'évolution.

L'apparition simultanée de partiels est étudiée dans la section 5.2. On propose un algorithme itératif qui détecte l'apparition d'un groupe d'une ou plusieurs entités sonores et rassemble dans ce groupe les partiels qui sont nés dans un même intervalle de temps. De ces groupes de partiels sont ensuite extraites des entités sonores grâce à deux algorithmes exploitant des indices de corrélation différents.

Dans la section 5.3, une méthode d'estimation de la fréquence de fondamentale dominante dans un groupe de partiels est proposée. Grâce à cette méthode, un algorithme extrait des entités sonores dont les fréquences des partiels sont en relation harmonique. Le second algorithme, introduit dans la section 5.4, exploite la notion de continuité du modèle sinusoïdal à long terme. Plusieurs mesures de corrélation entre les évolutions des paramètres de fréquence et d'amplitude des différents partiels sont étudiées. Les partiels d'un même groupe sont classés en fonction de ces mesures de corrélation de l'évolution de la fréquence ou de l'amplitude des partiels. L'utilisation d'une méthode de classification statistique permet de proposer un algorithme d'extraction d'entités sonores à partir d'un groupe de partiels où toutes les entités sonores sont identifiées simultanément.

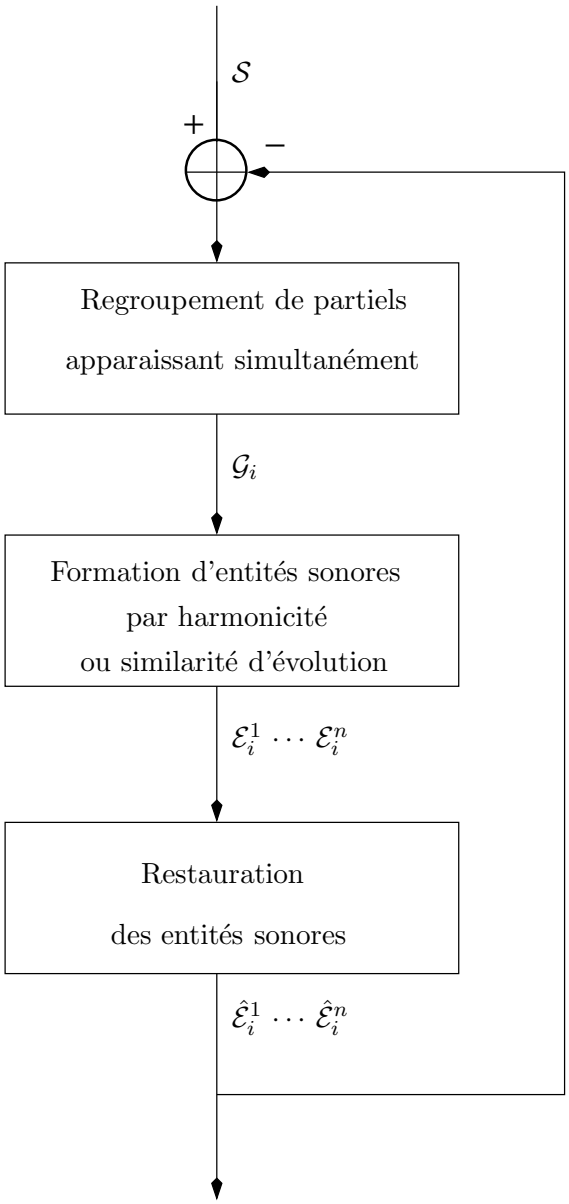


FIG. 5.2 – Schéma de principe de l'algorithme d'extraction d'entités sonores. On cherche dans l'ensemble des partiels $\mathcal{S}$ une apparition simultanée de partiels à une trame donnée. On regroupe alors dans $\mathcal{G}_1$ les partiels qui sont apparus ensemble. De ce groupe sont extraites les entités sonores $\mathcal{E}_1^{1\cdots n}$, n étant le nombre d'entités trouvées dans le groupe $\mathcal{G}_1$. Ces entités sonores sont ensuite restaurées. Les partiels de $\mathcal{S}$ appartenant à ces entités restaurées sont ensuite enlevés et une nouvelle apparition simultanée est recherchée pour former $\mathcal{G}_2$.

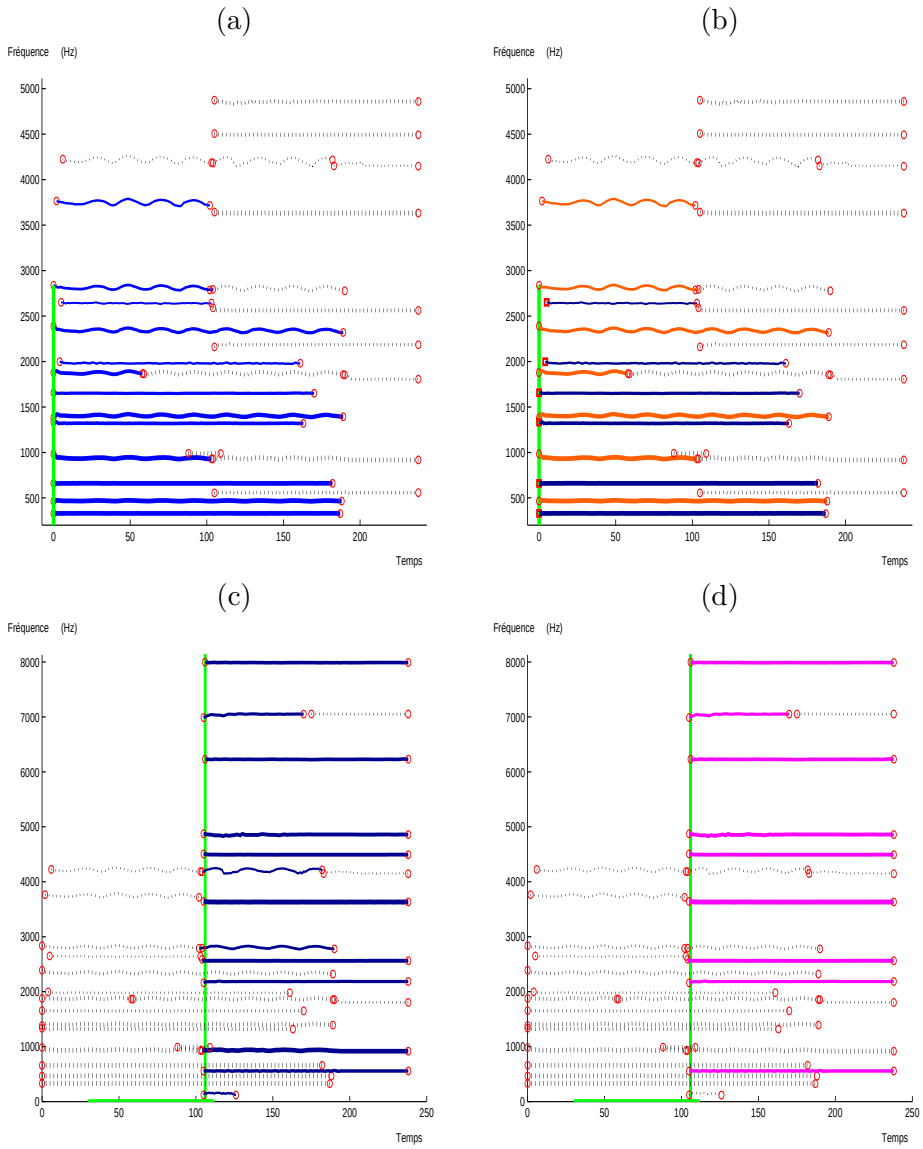


FIG. 5.3 – Exemple d'application de l'algorithme d'extraction d'entités sonores (les partiels considérés à une étape donnée sont représentés en trait plein). Du premier groupe $\mathcal{G}_1$ (a) sont extraites grâce à un critère d'harmonicité, deux entités sonores $\mathcal{E}_1^1$ et $\mathcal{E}_1^2$ (b). De manière à dissocier ces deux entités, les partiels de $\mathcal{E}_1^2$ débutent par un carré. Les partiels appartenant à ces deux entités sont supprimés de $\mathcal{S}$ avant d'obtenir un nouveau groupe $\mathcal{G}_2$ (c). De ce groupe est extraite $\mathcal{E}_2^1$ par des critères de similarité d'évolution (d).

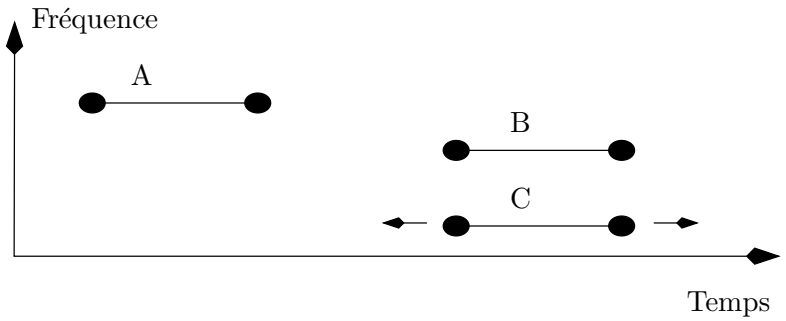


FIG. 5.4 – Répartition des partiels pour l'expérience de Bregman/Pinker. La position relative de B et C est modifiée en déplaçant C sur l'axe temporel.

5.2 Apparition simultanée

La détection des partiels apparaissant simultanément constitue la base de notre algorithme d'extraction d'entités sonores. Les considérations psychoacoustiques motivant cette place prépondérante sont détaillées dans une première partie. La seconde introduit un critère adapté à cette tâche et décrit un algorithme simple de détection de l'apparition simultanée de partiels.

5.2.1 Motivations

Un des indices les plus prépondérants pour la perception de plusieurs partiels comme une entité sonore est leur apparition simultanée. En citant Hartmann [Har88] : " Il est approprié que la position du début de la note prenne la première place dans la liste [des indices de séparation], car c'est probablement le premier indice par lequel les sons de différentes sources sont séparés dans l'écoute quotidienne de la voix et la musique.[2]"

Bregman a procédé à plusieurs expérimentations qui confirment l'influence prépondérante de l'apparition simultanée pour la perception groupée de plusieurs partiels. Une d'elle, effectuée avec Pinker [BP78], se propose de faire varier les caractéristiques de trois sinusoïdes pour changer la manière dont les partiels sont regroupés en entités sonores par le système auditif humain. Les trois sinusoïdes A, B et C (voir figure 5.4) sont présentées de manière répétitive à l'auditeur durant 150 ms chacune. Les résultats montrent que le degré de synchronisation entre le début de B et de C affecte fortement la perception de B comme une tonale isolée ou comme une composante d'un ensemble plus complexe BC. Une différence de 30 ms entre les débuts de B et C amène à une réduction significative du degré de combinaison. De plus, une étude publiée dans [Ras78] tend à montrer que l'asynchronie des attaques des différents instruments dans un ensemble musical se situe entre 27 et 49 ms. Les travaux de

[2] traduction de l'auteur

Gordon [Gor84] sur les attaques tendent à montrer que la durée perceptive des attaques d'un instrument est approximativement de 30 ms.

Cette durée, au-delà de laquelle deux sinusoïdes sont perçues comme deux éléments séparés (et qui représente aussi la durée maximale d'une attaque), est utile pour paramétrer l'algorithme que nous proposons. Notons Υ le nombre de trames correspondant à cette durée. Dans notre algorithme, Υ est multiplié par 2 à cause des imprécisions de la représentation à long terme extraite par l'algorithme de suivi proposé dans la section 3.7.

5.2.2 Algorithme

La détection de débuts de note est une problématique d'importance car elle constitue l'élément de base des algorithmes de segmentation de fichiers audionumériques. Une étude complète des multiples méthodes existantes sort du contexte de cette thèse et le lecteur est donc invité à se référer à [Kla99, DDS01, DBDS03, BS03] pour une description et des références plus complètes. De manière schématique, la plupart de ces méthodes utilisent l'évolution à court terme de certaines propriétés comme l'amplitude ou la phase du signal. Si la caractéristique choisie dépasse un certain seuil, un début de note est détecté. Sur ce principe, on peut proposer un premier critère d_A, qui considère la différence entre l'amplitude des partiels à la trame n et l'amplitude des partiels à la trame $n-1$:

$$d_A(n) = \sum_{i=1}^{m} A_i(n)\, e_i(n) - \sum_{j=1}^{m} A_j(n-1)\, e_j(n) \tag{5.1}$$

où m désigne le nombre de partiels, $A_i(n)$ l'amplitude du partiel d'indice i à la trame n et $e_i(n)$ vaut 1 si le partiel est présent à la trame d'indice n.

Ce critère pose deux problèmes. Tout d'abord, l'amplitude de certaines notes croît lentement comme pour le violon. Par son approche à court terme (seules deux trames consécutives sont exploitées), ce critère ne permet pas de détecter ce type de début de note, voir figure 5.5(a). De plus, d_A n'est pas normalisé et peut varier considérablement en fonction de l'énergie du signal. Il est alors difficile de fixer un seuil de détection de début de note.

En exploitant les propriétés du modèle à long terme, on peut pallier ces deux problèmes et proposer un nouveau critère $D_A(n)$ qui se base sur l'apparition simultanée de partiels. Les amplitudes moyennes de tous les partiels étant nés dans un certain intervalle autour de n sont accumulées, voir équation 5.2. $B(n)$, la somme des amplitudes moyennes des partiels dans cet intervalle est utilisée pour normaliser ce critère :

$$B(n) = \sum_{i=1}^{m} b_i(n)\,\bar{A}_i \tag{5.2}$$

$$C(n) = \sum_{j=1}^{m} \frac{1}{2\,\Upsilon + 1} \sum_{k=-\Upsilon}^{\Upsilon} A_j(n+k) \tag{5.3}$$

$$D_A(n) = \frac{B(n)}{C(n)} \tag{5.4}$$

où b_i est égal à 1 si le partiel P_i est né entre $n - \Upsilon$ et $n + \Upsilon$ et 0 sinon. $\bar{A}_i$ désigne l'amplitude moyenne du partiel P_i.

Ce nouveau critère permet la détection des débuts de notes dont l'amplitude croît lentement, voir figure 5.5(b) à la trame 317. La normalisation par $C(n)$ permet l'utilisation d'un unique seuil pour un nombre conséquent de signaux de dynamiques différentes. Comme D_A considère les partiels nés dans un intervalle de temps réduit, son évolution est composée soit d'intervalles nuls soit de pics proéminents, ce qui facilite l'utilisation d'un seuil pour la détection du début de note.

Ce critère est utilisé dans un algorithme de détection simple dont le schéma de principe est décrit dans la figure 5.8. À partir de l'indice courant n, le critère $D_A(n)$ est calculé. S'il est supérieur à une valeur seuil D_S, un maximum est détecté (max := 1) à la trame n ($n_m := n$). Cette position peut être affinée ensuite si $D_A(n)$ augmente. La recherche d'une apparition simultanée de partiels se termine lorsque l'indice courant est distant de Υ de l'indice du maximum ($n - n_m \leq \Upsilon$) et si $D_A(n)$ a chuté d'au moins la moitié de sa valeur maximale ($D_A(n) < D_M/2$). Ces deux derniers critères permettent de mettre en œuvre un masque, empêchant toute détection en deçà de ce masque, voir figure 5.7.

L'indice courant n est alors stocké pour être utilisé comme position de départ pour une détection ultérieure. Tous les partiels tels que $b_i(n_M) = 1$ sont regroupés dans un groupe $\mathcal{G}_i$. Un groupe peut être composé d'une ou plusieurs entités qu'il convient d'identifier. Les deux parties suivantes présentent deux méthodes d'extraction d'entités sonores sur des critères particuliers.

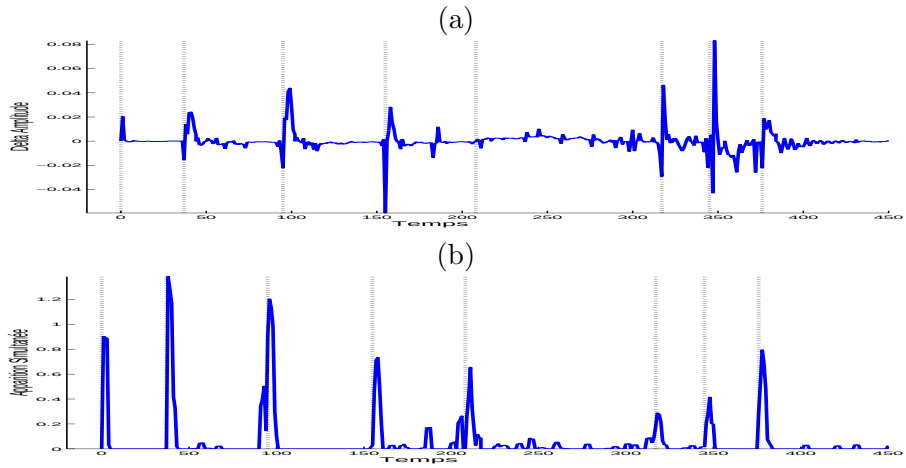


FIG. 5.5 – Deux mesures de détection de début de note. Les lignes verticales en pointillés représentent les débuts de notes relevés manuellement. En haut est représentée l'évolution du critère court-terme d_A en fonction du temps pour la représentation long terme de la figure 5.6. On note que le début de la note de violon n'est pas détectable (trame 210). En bas est représentée l'évolution du critère long-terme $D_A(n)$ en fonction du temps pour la même représentation à long terme. Les pics sont clairement marqués, même pour le début de la note de violon.

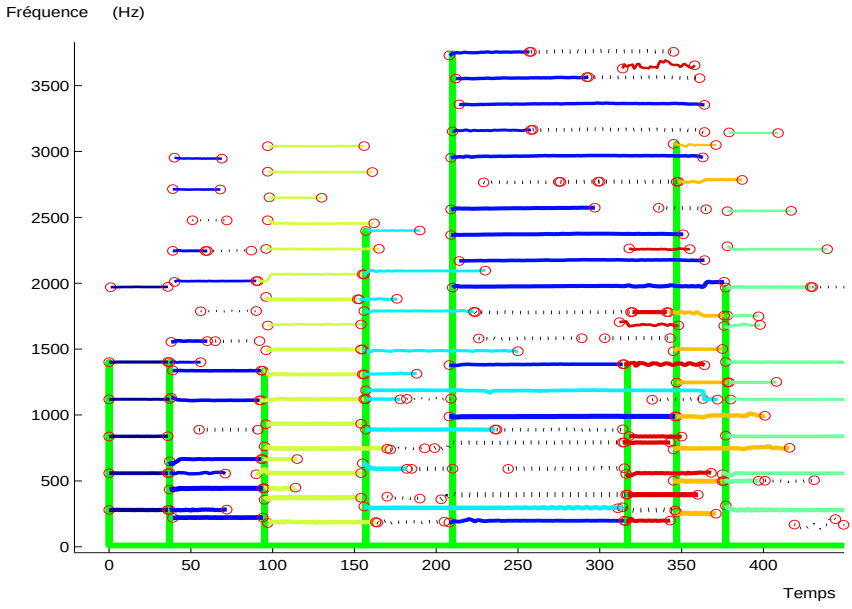


FIG. 5.6 – Représentation long terme de six notes de piano et d'une note de violon. Les lignes verticales représentent les apparitions simultanées de partiels.

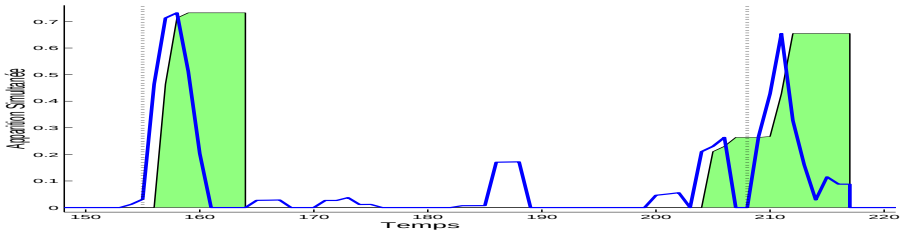


FIG. 5.7 – Évolution au cours du temps du masque D_M (aire) en fonction de D_A (trait).

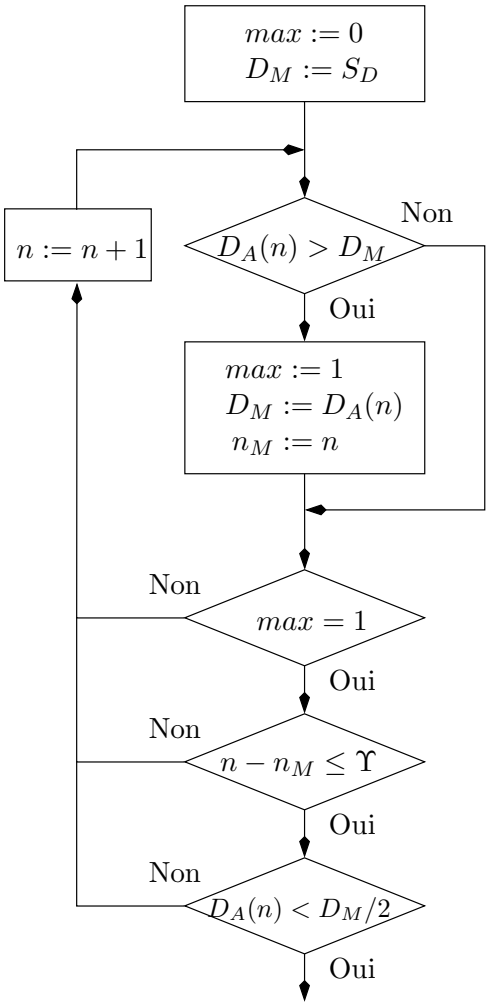


FIG. 5.8 – Schéma de principe de l'algorithme de détection d'une apparition simultanée de partiels.

5.3 Relation d'harmonicité

Comme cela est évoqué dans le chapitre 1, De nombreux instruments (en tout premier lieu la voix), de par leur structure physique, émettent des sons harmoniques. Même s'il n'existe pas à l'heure actuelle de démonstration neurophysiologique du fait que le système auditif humain effectue un regroupement des partiels par harmonicité, on peut noter que ce système est le pendant naturel du système de production de la voix. On peut donc supposer que ce système est capable, d'une manière ou d'une autre, d'agréger les partiels avec des fréquences en relation harmonique. Il serait en effet particulièrement désagréable d'entendre distinctement plusieurs composantes sinusoïdales à la place d'une voyelle.

La relation d'harmonicité est donc un indice prépondérant pour le regroupement de partiels ayant débuté simultanément. Il existe plusieurs manières de déterminer si des partiels sont en relation harmonique. Certaines méthodes se basent sur un critère de concordance harmonique des fréquences de deux partiels [RG04, VK00]. Par exemple, soient deux partiels P_i et P_k appartenant à la même entité harmonique. Le rapport de ces deux fréquences est proche d'un rapport de deux entiers positif a et b :

$$\frac{F_i(n)}{F_k(n)} \approx \frac{a}{b} \tag{5.5}$$

où $F_i(n)$ $F_k(n)$ sont les fréquences des partiels P_i et P_k à la trame d'indice n. En fixant une borne inférieure à la fréquence de fondamentale de cette entité harmonique F_{min}, il est possible de mesurer la dissimilarité harmonique des ceux partiels :

$$d_h(P_i, P_k) = \min_{a,b} \left| \log \left(\frac{F_i/F_k}{a/b} \right) \right| \tag{5.6}$$

pour a de 1 à $\lfloor F_i/F_{min} \rfloor$ et b de 1 à $\lfloor F_k/F_{min} \rfloor$.

Une autre approche se base sur une estimation préalable de la fréquence de fondamentale pour ensuite regrouper les partiels dont la fréquence est proche d'un multiple de cette fréquence de fondamentale.

On peut citer trois théories d'analyse de la hauteur par le système auditif humain. Rappelons que la hauteur ("*pitch*" en anglais) désigne une mesure perceptive, tandis que la fréquence de fondamentale désigne une mesure physique. La première théorie se base sur une mise en correspondance de la répartition des partiels sur le spectre avec un gabarit de série harmonique. Une seconde théorie exploite l'espacement régulier entre les fréquences des partiels d'une série harmonique pour estimer la fréquence de fondamentale. Cette théorie permet de mieux expliquer pourquoi l'oreille est capable d'attribuer une hauteur à un signal harmonique dont la fondamentale est manquante. Une troisième, plus complexe, exploite les phénomènes de battements entre les harmoniques adjacentes, en particulier dans les hautes fréquences [Bre90]. Plusieurs estimations de la hauteur sont effectuées dans différentes bandes de fréquences. Les résultats sont ensuite intégrés pour donner une estimation globale de la hauteur comme exploité dans [Kla03]. Cette théorie permet de mieux expliquer la capacité de

l'oreille à attribuer une hauteur à un signal légèrement inharmonique, comme celui produit par le piano.

5.3.1 Estimation de la fréquence de fondamentale

Des trois modèles présentés ci-dessus, le modèle par mise en correspondance de motifs a été à la base de nombreuses implantations de détecteurs de fréquence de fondamentale [DR91, Bro92, MB94]. Le principe de l'algorithme décrit dans [Bro92] est d'utiliser une transformée spectrale particulière, qui amène à une répartition logarithmique des fréquences des composantes harmoniques. L'écartement en fréquence des composantes harmoniques est alors indépendant de la fréquence de fondamentale. La corrélation maximale entre ce spectre et un gabarit particulier (spécifique à chaque instrument) permet alors d'estimer la fondamentale. Dans cet algorithme, la représentation spectrale est discrète, de même que le gabarit de comparaison.

Dans notre cas, les fréquences des partiels sont des nombres réels, et le gabarit doit donc être une fonction à support continu. De plus, les instruments considérés peuvent être de types variés. Il est donc indispensable pour notre application d'avoir un gabarit générique qui puisse ensuite être spécialisé grâce à un nombre réduit de paramètres. Tout d'abord, ce gabarit doit être quasi périodique de période la fréquence de fondamentale :

$$g_h(f) = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi f}{h}\right)\right) \tag{5.7}$$

où h est la fréquence de fondamentale et f est la fréquence à laquelle on évalue la fonction. La largeur des pics est dépendante de la fréquence de fondamentale utilisée, voir figure 5.9(a). Pour remédier à ce défaut et permettre de paramétrer la sélectivité des pics, il est proposé dans [Car02] de pondérer cette fonction :

$$g'_h(f) = g_h(f)^{\frac{-s}{\log(g_h(1))}} \tag{5.8}$$

où $s \in]0, 1]$ est un paramètre du modèle qui permet de régler la sélectivité des pics, voir figure 5.9(b). De manière à modéliser l'inharmonicité, on peut réduire la sélectivité des pics plus leur fréquence est élevée :

$$g''_h(f) = g'_h(f)^{u\,\frac{f-h}{Fe/2-h}\,(\frac{1}{s-1})} \tag{5.9}$$

où F_e est la fréquence d'échantillonnage et $u \in]0, 1]$ est un paramètre du modèle qui permet de régler l'étalement des pics en fonction de la fréquence, voir figure 5.9(c). On remarque que le gabarit proposé par la fonction est éloigné des enveloppes spectrales des signaux musicaux car la majeure partie de ces signaux ont une enveloppe exponentiellement décroissante. Il semble naturel d'appliquer une pondération similaire :

$$g'''_h(f) = g''_h(f)\, 10^{-(d\,\frac{f-h}{Fe})} \tag{5.10}$$

où d est un paramètre du modèle exprimé en dB dont dépend la pondération appliquée, voir figure 5.9(d). Cette pondération permet de lever les ambiguïtés

de détection des multiples ou sous-multiples de la fréquence de fondamentale. Grâce à cette fonction, on dispose d'un gabarit $\mathcal{P}_h(f)$ défini par :

$$\mathcal{P}_h(f) = \begin{cases} g_h'''(f) & \text{si } f > h/2 \text{ et } f < F_e - h/2 \\ 0 & \text{sinon.} \end{cases} \tag{5.11}$$

L'indice de vraisemblance V_h d'une fréquence de fondamentale h en fonction d'un ensemble de partiels à une trame n est alors défini par :

$$V_h = \sum_{i=0}^{m} A_i(n)\, \mathcal{P}_h(F_i(n)) \tag{5.12}$$

où m désigne le nombre de partiels et $A_i(n)$ et $F_i(n)$ désignent les paramètres d'amplitude et de fréquence du partiel P_i à la trame n. L'amplitude $A_i(n)$ vaut 0 si le partiel P_i n'est pas présent à la trame n.

5.3.2 Algorithme

L'algorithme proposé est un algorithme itératif. Tout d'abord, la fréquence de fondamentale dominante du groupe de partiels débutant ensemble est estimée. Si l'entité sonore associée à cette fréquence de fondamentale satisfait certains critères définis dans les équations 5.13 et 5.14, alors les partiels de cette entité sont soustraits du groupe de partiels. Les partiels restants sont alors utilisés pour détecter une autre fréquence de fondamentale.

À chaque partiel du groupe est associée une fréquence et une amplitude moyenne $\bar{F}$ et $\bar{A}$, calculée sur le même support temporel. L'exploration de toutes les fréquences pour trouver la fréquence h qui maximise V_h est très coûteux. On propose de ne considérer que la fréquence du partiel d'amplitude $\bar{A}$ la plus élevée et ses sous-multiples supérieurs à 50 Hz.

L'ensemble des partiels associés à cette fréquence de fondamentale est ensuite formé. Pour chaque multiple de cette fréquence de fondamentale $r_h\, h$, le partiel de fréquence moyenne $\bar{F}_i$ la plus proche est recherché. Ce partiel est associé si son rang harmonique est bien r_h et si une estimation de son inharmonicité est inférieure à un seuil donné :

$$\left[\frac{\bar{F}}{h}\right] = r_h \tag{5.13}$$

$$g_h''(\bar{F}_i) < S_h \tag{5.14}$$

où $[x]$ désigne l'entier le plus proche de x.

Ces partiels sont définitivement regroupés au sein d'une même entité si la vraisemblance et la moyenne des inharmonicités (comme calculées par l'équation 5.14) de ces partiels sont inférieures à des seuils donnés.

Cet algorithme d'extraction d'entités harmoniques, même s'il donne des résultats acceptables en pratique, souffre de deux handicaps. D'une part, le nombre de paramètres et de seuils est conséquent et nécessite une phase d'apprentissage à partir d'une base de signaux tests. D'autre part, cet algorithme

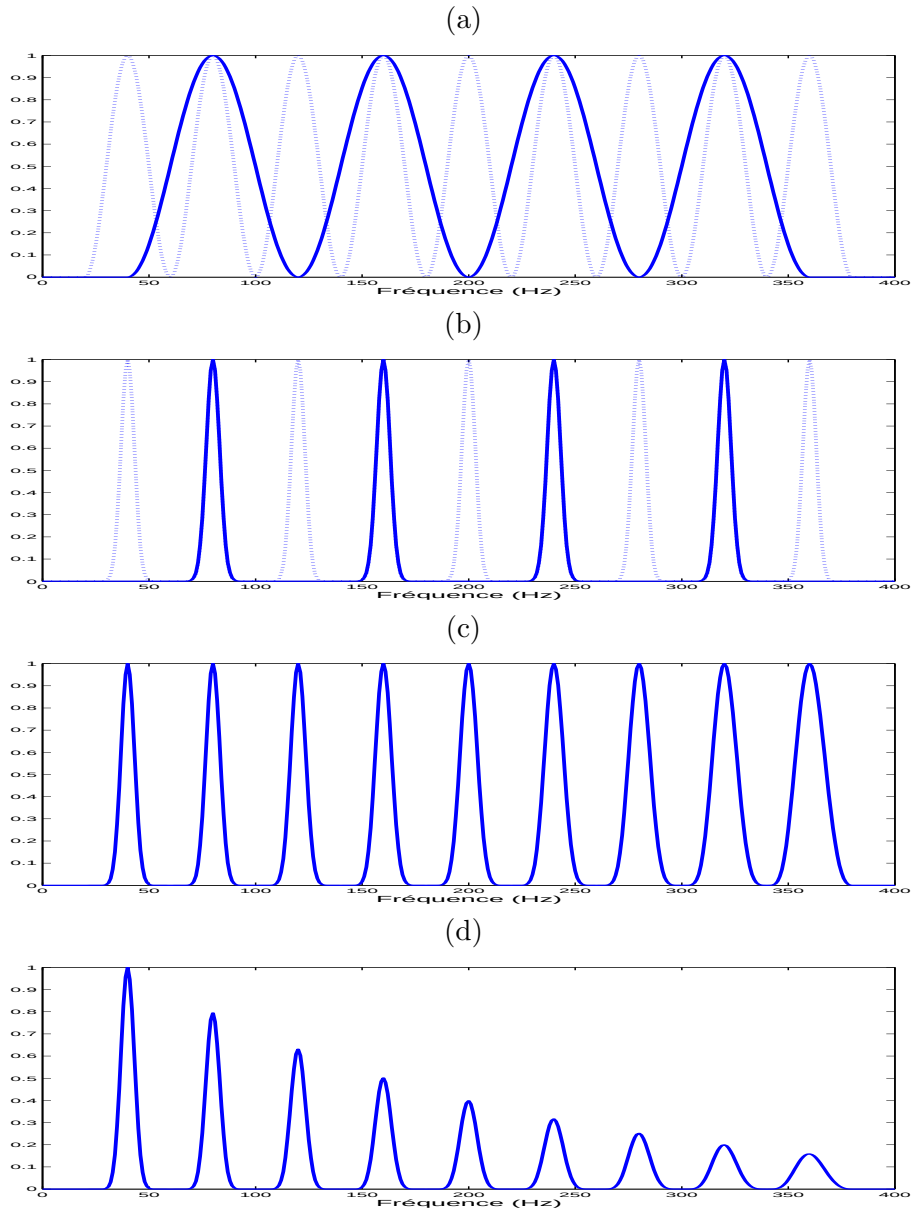


FIG. 5.9 – Sur la figure (a) est représentée g_h pour deux fréquences de fondamentale différentes (40 et 80 Hz). Sur la figure (b) est représentée g'_h pour ces deux fréquences avec un facteur de sélectivité $s = 0.05$. Sur la figure (c) est représentée g''_h avec un facteur d'étalement $e = 1$. Enfin, sur la figure (d) est représentée g'''_h avec une atténuation $d = 2$ dB.

suppose que chaque partiel appartient à une et une seule entité. Or, des harmoniques d'entités différentes peuvent se chevaucher et être représentées par un seul partiel. Ensuite, on peut remarquer que de nombreux instruments de musique sont, par nature, inharmoniques. Il est donc nécessaire d'utiliser un critère de regroupement de partiels qui soit plus générique.

5.4 Similarité d'évolution

On peut considérer que les évolutions des paramètres de fréquence et d'amplitude des partiels d'une même entité sonore sont corrélées. Une première justification provient des instruments harmoniques. En effet, les partiels de rang m et n auront des fréquences liées par un rapport égal à m/n.

De plus, de nombreuses études psychoacoustiques ont montré que les variations ou micro-modulations des paramètres influent grandement sur la perception. En citant Bregman : "De faibles fluctuations de la fréquence apparaissent naturellement dans la voix humaine et dans les instruments musicaux. Ces fluctuations ne sont souvent pas très prononcées, comprises entre moins d'un pour cent pour une note de clarinette à environ un pour cent pour une voix essayant de conserver une hauteur constante, avec des fluctuations plus fortes, de l'ordre de 20 pour cent pour le vibrato d'un chanteur. Même un très faible facteur de fluctuation de la fréquence peut avoir des effets notables sur le regroupement perceptif de composantes harmoniques.[3]"

Cette caractéristique a notamment été utilisée dans de nombreux synthétiseurs. Pour rendre le signal synthétisé plus naturel, on fait varier faiblement les fréquences des sinusoïdes. Au vu des résultats des expériences de McAdams [McA84], il semble important que les variations soient corrélées pour que les partiels soient effectivement perçus comme une entité sonore.

Dans le cas d'entités inharmoniques, cette propriété est particulièrement intéressante pour regrouper les partiels car les informations d'harmonicité sont par définition inexploitables. De plus, le fait que les évolutions des paramètres des partiels soient corrélées ne semble pas lié uniquement à l'harmonicité. En citant Moore [Moo03] : "Si on présente à l'auditeur un son complexe contenant plusieurs partiels avec des fréquences tirées au hasard, un seul son complexe s'approchant d'un signal bruité sera perçu avec un certain timbre. Un sous-ensemble de partiels est maintenant mis en exergue du reste des partiels en les faisant varier de manière cohérente en fréquence, en amplitude ou les deux. Ce groupe sera alors perçu comme une figure proéminente se dégageant d'un fond sonore et ces deux entités, figure et fond, auront un timbre différent du son original.[4]"

5.4.1 Dissimilarité entre partiels

De manière à pouvoir regrouper les partiels dont la fréquence et/ou l'amplitude varient de façon corrélées, on doit disposer d'une mesure de dissimilarité qui soit pertinente. La similarité entre objets est une mesure de la vraisemblance du fait que ces deux objets appartiennent à une même classe. La notion inverse de dissimilarité est préférée dans cet exposé car elle correspond mieux à l'application recherchée qui est bien de séparer, de distinguer des objets différents.

[3] traduction de l'auteur
[4] traduction de l'auteur

De manière à évaluer plusieurs dissimilarités existantes et à apprécier le gain apporté par celle proposée dans cette partie, on considère un ensemble de partiels issus d'ensembles de partiels de sons d'instruments variés : une note de saxophone avec un vibrato, une voix chantée modulée, une note de piano, et une note de triangle. Les cinq partiels d'amplitudes les plus fortes sont sélectionnés de chacun de ces ensembles. À ces partiels sont ajoutés cinq partiels extraits par erreur d'un bruit blanc (on attribue à ces partiels l'indice de classe 0). Les évolutions en fréquence de ces partiels sont représentées sur la figure 5.10. Tout ces partiels sont tronqués pour être de même durée (approximativement 1 seconde).

À l'exception de la dissimilarité d_v proposée par Virtanen dans [VK00] où les vecteurs de fréquence et d'amplitude sont utilisés en l'état, la moyenne est soustraite aux vecteurs préalablement à tout calcul. Les vecteurs d'amplitude sont normalisés par leur variance avant d'être utilisés car les différences d'amplitude sont très élevées entre les différentes harmoniques d'une même note. Comme les évolutions du paramètre de fréquence sont apparues lors de nos expérimentations comme les plus révélatrices, on utilise ce paramètre pour évaluer les différentes dissimilarités envisagées. Concernant l'amplitude, les performances relatives des différentes dissimilarités sont apparues lors des tests comme équivalentes, en étant plus faible comparativement à celles obtenues avec la fréquence.

Dissimilarités classiques

La distance euclidienne d_e entre deux vecteurs est définie par :

$$d_e(X,Y) = \sqrt{\sum_{i=0}^{N-1} (X(i) - Y(i))^2} \tag{5.15}$$

où X et Y sont des vecteurs de taille N. Cette dissimilarité n'est pas invariante par changement d'échelle, ce qui est problématique pour regrouper des partiels d'une entité harmonique et modulée, par exemple par un vibrato.

La dissimilarité cosinus (ou corrélation) d_c est définie par :

$$d_c(X,Y) = 1 - \frac{c(X,Y)}{\sqrt{c(X,X)c(Y,Y)}} \tag{5.16}$$

$$c(X,Y) = \sum_{i=0}^{N-1} X(i)\,Y(i) \tag{5.17}$$

où X et Y sont des vecteurs de taille N. La normalisation par la racine carrée du produit des moyennes rend cette dissimilarité plus appropriée mais, dans ce cas, on est incapable de séparer les fréquences des partiels de bruit et les fréquences des partiels du piano qui sont quasi constantes avec une faible perturbation due à l'imprécision de l'analyse à court terme, comme on le verra dans la suite.

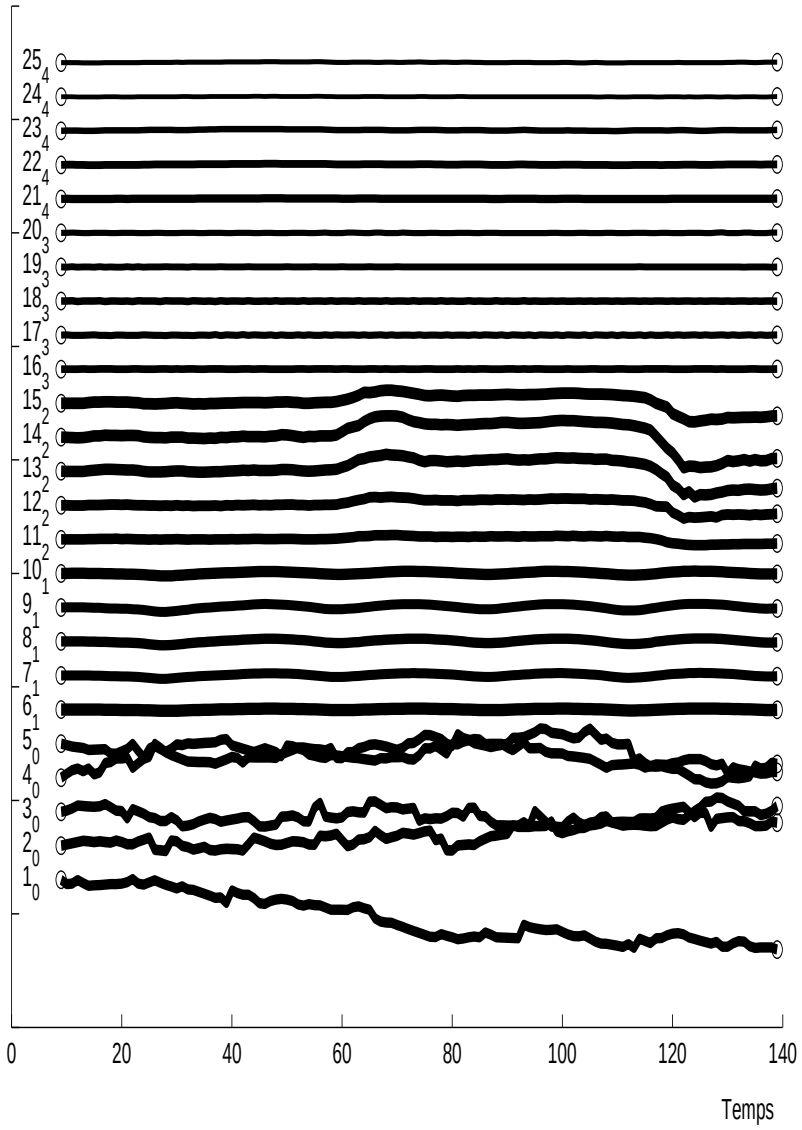


FIG. 5.10 – Évolution en fréquence de cinq partiels extraits d'un bruit blanc (entité $\mathcal{C}_0$), des partiels d'une note de saxophone avec un vibrato (entité $\mathcal{C}_1$), une voix chantée modulée (entité $\mathcal{C}_2$), une note de piano (entité $\mathcal{C}_3$) et une note de triangle (entité $\mathcal{C}_4$). Ces partiels sont indicés par indice croissant et en sous indice par numéro d'entités.

Dissimilarité avec normalisation par les moyennes

Virtanen propose [VK00] une dissimilarité qui consiste en une distance euclidienne au carré des vecteurs préalablement normalisés par leur moyenne :

$$d_v(X,Y) = \frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{X(i)}{\bar{X}} - \frac{Y(i)}{\bar{Y}}\right)^2 \tag{5.18}$$

La normalisation par la moyenne permet à la dissimilarité d'être invariante par facteur d'échelle. Pour les vecteurs de fréquences de différents partiels d'une entité harmonique, cette normalisation est pertinente et amène à des performances appréciables comme il sera montré dans la partie dédiée à l'évaluation. Ces bons résultats se généralisent malheureusement mal aux autres types de partiels comme ceux de bruit ou de piano.

Si on considère l'évolution des fréquences d'un partiel comme un signal, on peut le décomposer en deux parties, une partie qui est conforme au modèle sinusoïdal (qui varie lentement et continuement au cours du temps) et une autre partie dite de "bruit d'observation" qui provient des erreurs d'estimation de la fréquence du module d'analyse à court terme. Idéalement, cette indépendance vis-à-vis du facteur d'échelle devrait être appliquée uniquement à la forme recherchée et non pas au bruit dû aux erreurs d'estimation de l'analyse spectrale. Or, la normalisation effectuée dans l'équation 5.18 change le facteur d'échelle de la partie pertinente du signal, mais aussi celui du bruit. L'utilisation de la modélisation AR pour la modélisation des paramètres des partiels introduite dans 3.6 permet de proposer une méthode qui apporte une solution à ce problème.

Erreurs de prédiction croisées

On a montré dans la section 3.6 que la modélisation AR des paramètres de fréquence et d'amplitude des partiels était pertinente pour améliorer le suivi de partiels. Les intérêts théoriques d'une telle modélisation pour la définition d'une dissimilarité sont nombreux. Tout d'abord, on peut discriminer les évolutions prédictibles des évolutions non prédictibles. Ce faisant, on peut alors discriminer facilement les partiels de bruit des autres. On modélise donc un vecteur X_l (le vecteur de fréquence ou d'amplitude du partiel l) comme :

$$X_l(n) = \sum_{i=1}^{k} K_l(i)X_l(n-i) + E_l(n) \tag{5.19}$$

où E_l est l'erreur de prédiction directe. C'est donc le signal résiduel du filtrage du signal X_l en utilisant les coefficients de filtre K_l. On peut montrer que les coefficients $K_l(i)$ sont invariants par facteur d'échelle. La modélisation AR permet donc théoriquement de pouvoir comparer des vecteurs en fonction d'une partie prédictible indépendante du facteur d'échelle et d'une partie non prédictible qui reste à la même échelle pour tous les vecteurs.

Ceci est particulièrement approprié dans le cas des vecteurs de fréquences et d'amplitudes des partiels. En effet, dans le cas de vecteurs de fréquences

assez prédictibles comme ceux du piano ou du saxophone, l'erreur de prédiction représente l'imprécision de l'estimateur de fréquence du modèle à court terme. Cette erreur est uniformément répartie sur l'axe des fréquences dans le cas d'estimateurs comme ceux abordés dans la section 2 basés sur la transformée de Fourier.

Pour chaque vecteur de fréquence F_l, on calcule donc un vecteur K_l de 4 coefficients AR par la méthode de Burg décrite dans la section 3.6. La comparaison de ces deux vecteurs permet théoriquement d'identifier des signaux proches. Cette approche a été utilisée avec succès dans des applications de reconnaissance de formes [DG86] :

$$d_{\mathrm{AR}}(X_1, X_2) = d_e(K_1, K_2) \tag{5.20}$$

où X_1 et X_2 sont deux vecteurs que l'on souhaite comparer. Une telle dissimilarité donne de résultats inexploitables pour notre application comme on le verra dans la partie suivante.

L'erreur de prédiction croisée (noté E_1^2) se définit comme le signal résiduel du filtrage d'un vecteur X_1 en utilisant les coefficients de filtre K_2 :

$$E_1^2(n) = X_1(n) - \sum_{i=1}^{k} K_2(i) X_1(n-i) \tag{5.21}$$

Soient deux vecteurs de fréquence F_1 et F_2 à comparer. Les vecteurs AR K_1 et K_2 sont calculés de manière à minimiser respectivement l'énergie des erreurs de prédiction directes E_1 et E_2. Si les deux vecteurs F_1 et F_2 ont des propriétés similaires (la composante prédictible est sensiblement la même à un facteur d'échelle près), les énergies des erreurs de prédiction croisées E_1^2 et E_2^1 sont faibles. Le principe de la dissimilarité d_σ proposée est de cumuler ces deux dissimilarités antisymétriques pour obtenir une dissimilarité symétrique :

$$d_\sigma(X_1, X_2) = \frac{1}{2}\left(|E_1^2| + |E_2^1|\right) \tag{5.22}$$

où X_1 et X_2 sont deux vecteurs que l'on souhaite comparer. On peut considérer une autre dissimilarité d'_σ, qui consiste en le rapport entre les erreurs de prédiction croisées et les erreurs de prédiction directes :

$$d'_\sigma(X_1, X_2) = \frac{|E_1^2| + |E_2^1|}{1 + |E_1| + |E_2|} \tag{5.23}$$

où X_1 et X_2 sont deux vecteurs que l'on souhaite comparer. Les performances de ces deux dissimilarités sont discutées dans la partie suivante.

Évaluation de la capacité de discrimination

Une dissimilarité pertinente est une dissimilarité qui est faible entre deux éléments (partiels) d'une même classe (entité sonore) et élevée entre deux éléments de classes différentes. Autrement dit, la dissimilarité intra-classe doit être minimale et la dissimilarité inter-classe maximale de manière à pouvoir

Q_d	d_e	d_c	d_v	d_{AR}	d_σ	d'_σ
$\mathcal{C}_0$	3.9	4.3	0	6.6	8.9	16.8
$\mathcal{C}_1$	10.5	806.6	47634	2.5	97.3	46.9
$\mathcal{C}_2$	3.1	1586.6	37.4	5.7	23.1	29.5
$\mathcal{C}_3$	147.9	4.8	57.1	4.7	251.8	81.3
$\mathcal{C}_4$	49.5	43.9	83866	2.1	72.1	22.6
U	5.5	10.6	5.1	3.1	21.2	36.2

TAB. 5.1 – Estimation de la qualité en fonction du critère Q_d défini dans les équations 5.26 et 5.27 de différentes dissimilarités : d_e la distance euclidienne, d_c la distance cosinus, d_{AR} la distance euclidienne entre les coefficients AR calculés sur les deux vecteurs considérés, d_σ la dissimilarité par erreurs de prédiction croisées et d'_σ la dissimilarité par erreurs de prédiction croisées normalisée par les erreurs de prédiction directes. Les vecteurs comparés sont les fréquences des partiels représentés sur la figure 5.10.

séparer le mieux possible les différentes classes. Soit U l'ensemble des éléments de cardinal N et $\mathcal{C}_i$ la classe d'indice i parmi N_c classes, une estimation de la qualité d'une dissimilarité donnée $d(x, y)$ pour cette classe est :

$$\mathrm{intra}(\mathcal{C}_i) = \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} d(\mathcal{C}_i(j), \mathcal{C}_i(k)) \tag{5.24}$$

$$\mathrm{inter}(\mathcal{C}_i) = \sum_{j=0}^{n-1}\sum_{l=0}^{N-n+1} d(\mathcal{C}_i(j), \mathcal{F}_i(l)) \tag{5.25}$$

$$Q_d(\mathcal{C}_i) = \frac{\mathrm{inter}(\mathcal{C}_i)}{\mathrm{intra}(\mathcal{C}_i)} \tag{5.26}$$

où n est le nombre d'éléments de la classe $\mathcal{C}_i$ et $\mathcal{F}_i = U - \mathcal{C}_i$. La qualité globale Q_d est alors :

$$Q_d(U) = \frac{\sum_{i=0}^{N_c-1} \mathrm{inter}(\mathcal{C}_i)}{N_c \sum_{i=0}^{N_c-1} \mathrm{intra}(\mathcal{C}_i)} \tag{5.27}$$

À titre d'exemple, soient deux classes $\mathcal{A}_1 = \{\{1.1, 5.1\}, \{1.0, 5.2\}, \{1.0, 5.3\}\}$ et $\mathcal{A}_2 = \{\{1.0, 1.1\}, \{1.1, 0.9\}\}$, composées de points $e_i = \{x_i, y_i\}$ dans un espace à deux dimensions. Soient deux dissimilarités, l'une considérant les ordonnées $d_x(e_i, e_j) = |x_i - x_j|$ et l'autre considérant les abscisses $d_y(e_i, e_j) = |y_i - y_j|$. Au vu des données, la dissimilarité la plus discriminante est d_y. Cette observation est vérifiée par la mesure de qualité : $Q_{d_x} = 0.75 < Q_{d_y} = 32$.

Un autre indicateur ζ, utilise un critère plus proche d'un critère de classification et est ainsi plus indépendant de l'échelle de la dissimilarité choisie. Soit un ensemble d'objets X, $\zeta(X)$ est la proportion de couples (a, b) tels que b est l'élément le plus proche de a au sens de la dissimilarité choisie et a et b appartiennent à la même classe.

Soit une fonction classe définie comme :

$$\begin{array}{rccc} \text{classe} : & X & \rightarrow & \mathbb{N} \\ & a & \mapsto & i \end{array}$$

ζ	d_e	d_c	d_v	d_{AR}	d_σ	d'_σ
$\mathcal{C}_0$	0.6	0.6	0.6	0	0	1
$\mathcal{C}_1$	0	1	1	0	0.8	0.8
$\mathcal{C}_2$	0	1	1	0	0.2	1
$\mathcal{C}_3$	1	0.6	0.4	0	1	1
$\mathcal{C}_4$	0	1	1	0	1	1
U	0.06	0.168	0.16	0	0.12	0.192

TAB. 5.2 – Estimation de la qualité en fonction du critère ζ défini dans l'équation 5.28 de différentes dissimilarités : d_e la distance euclidienne, d_c la distance cosinus, d_{AR} la distance euclidienne entre les coefficients AR calculés sur les deux vecteurs considérés, d_σ la dissimilarité par erreurs de prédiction croisées et d'_σ la dissimilarité par erreurs de prédiction croisées normalisée par les erreurs de prédiction directes. Les vecteurs comparés sont les fréquences des partiels représentés sur la figure 5.10.

où i est l'indice de classe auquel appartient a.
On a :

$$\zeta(X) = \frac{\mathrm{Card}\ \{(a,b)\,|\,d(a,b) = \min_{c\in X} d(a,c) \wedge \mathrm{classe}(a) = \mathrm{classe}(b)\}}{\mathrm{Card}\ X} \quad (5.28)$$

X pouvant être soit une classe $\mathcal{C}_i$ soit l'ensemble des objets U.

Comme on peut le voir à la première colonne des tables 5.1 et 5.2, la dissimilarité d_e donne de mauvais résultats pour les vecteurs de fréquences des partiels de la note de saxophone et ceux de la voix modulée. Grâce à la normalisation par les écarts types, la dissimilarité par corrélation d_c se comporte mieux dans le cas de modulations, voir à la deuxième colonne des tables 5.1 et 5.2.

La dissimilarité par distance euclidienne normalisée par les moyennes montre des résultats disparates, voir troisième colonne de la table 5.1. Certaines classes sont aisément discriminées comme celles comprenant les partiels des notes de saxophone $\mathcal{C}_1$ et du triangle $\mathcal{C}_4$ au contraire d'autres comme les partiels de bruit où la normalisation par la moyenne n'est pas pertinente. L'évaluation par le critère ζ montre en revanche de bons résultats sauf pour les classes $\mathcal{C}_0$ et $\mathcal{C}_3$.

La dissimilarité par distance euclidienne des coefficients AR d_{AR} offre des résultats médiocres, comme on peut le constater à la quatrième colonne des tables 5.1 et 5.2.

La dissimilarité d_σ offre de bonnes performances pour les signaux prédictibles, voir l'avant-dernière colonne des tables 5.1 et 5.2. Par contre, les corrélations des partiels de voix ou de bruit avec des évolutions complexes sont mal représentées.

La dissimilarité d'_σ offre des résultats très homogènes pour les classes de partiels que l'on souhaite modéliser, voir la dernière colonne des tables 5.1 et 5.2. Cette homogénéité est particulièrement utile pour la méthode de classification qui sera étudiée dans la section suivante.

5.4.2 Classification ascendante hiérarchique

La classification non supervisée se propose d'identifier les classes dans lesquelles se regroupent les objets, et ceci en considérant uniquement les attributs des objets. On peut aborder ce problème de plusieurs manières ; nous présentons ici deux méthodes classiques [CM02].

La méthode des k-moyennes se propose d'effectuer cette classification en fixant le nombre de classes *a priori*. On part de k objets moyens (dont les attributs sont choisis au hasard), en agrégeant les objets les plus proches, on partitionne l'ensemble des objets en k classes. À chaque étape, on agrège à chacun de ces objets moyens l'objet le plus proche. On partitionne ainsi l'ensemble des objets en k classes. Pour chaque étape, les attributs d'objets moyens sont recalculés en fonction des nouvelles insertions. C'est pour cela que la méthode s'appelle aussi méthode des *moyennes mobiles*. Dans le cas d'un ensemble d'objets de cardinal faible (c'est le cas dans notre application), l'étape d'initialisation est contraignante car elle nécessite plusieurs exécutions de l'algorithme pour obtenir des résultats satisfaisants [RG04]. De plus, dans notre problème, le nombre d'entités contenues dans un groupe est inconnu.

La seconde méthode, dite de classification hiérarchique, consiste à construire une hiérarchie de classes que l'on interprète ensuite pour identifier les classes pertinentes de cette hiérarchie. Quelques définitions sont présentées dans une première partie. Puis un algorithme dit de classification ascendante hiérarchique est présenté. En utilisant cet algorithme avec la dissimilarité par erreur de prédiction croisées introduite dans la partie précédente, on obtient une classification pertinente pour l'agrégation de partiels en fonction des corrélations des évolutions, tant au niveau du paramètre de fréquence qu'à celui d'amplitude.

Définitions

Soit U l'ensemble de cardinal N des objets à classer. Une partition π de U est un ensemble de parties de U, notés h_i, non vides et disjointes deux à deux, dont l'union est U :

- $h_i \neq \emptyset \, \forall i$
- $h_i \cap h_j = \emptyset \, \forall i \neq j$
- $\cup h_i = U$

Une partition π_i est dite plus fine qu'une partition π_j si et seulement si tout élément de π_j est soit un élément de π_i soit l'union de plusieurs parties de π_i, soit l'union de plusieurs éléments de π_i. On définit π_1 la partition telle que tous ses éléments sont des singletons et π_N la partition telle que tous les objets sont dans la même classe.

Une hiérarchie H sur U est un ensemble de partitions $\{\pi_1, ..., \pi_N\}$ tel que π_{i-1} est plus fine que π_i et π_1 est la partition la plus fine de U. Autrement dit, une hiérarchie H sur U est un sous-ensemble des parties de U tel que :

- pour tout élément x de U, $\{x\} \in H$;
- pour tout couple d'éléments h_i et h_j de H avec $h_i \neq h_j$, on a :
  soit $h_i \cap h_j = \emptyset$,
  soit $h_i \cap h_j \neq \emptyset$, alors soit $h_i \subset h_j$, soit $h_j \subset h_i$.

Par exemple, considérons 6 points sur un plan à 2 dimensions dont les attributs sont leurs coordonnées spatiales, voir figure 5.11(a). Une hiérarchie possible de ces objets est représentée par le dendrogramme (arbre pondéré) de la figure 5.11(b). Les premiers éléments de la hiérarchie $h(1, ..., 6)$ sont constitués par les objets eux-mêmes (les feuilles du dendrogramme). La partition la plus fine π_1 est donc $\{h_1, ..., h_6\}$. L'élément h_7 (le premier nœud du dendrogramme) est l'union de h_4 et h_5, cette agrégation se poursuit jusqu'à obtenir $\pi_6 = h_{11}$ avec $h_{11} = U$.

Une hiérarchie indicée monotone est une hiérarchie sur un ensemble fini à laquelle on associe une suite de nombres réels r_i tels que si π_i est plus fine que π_{i-1} alors $r_i \geq r_{i-1}$. On peut exploiter cette information pour estimer le nombre de classes en trouvant les "points de ruptures" dans l'évolution de r_i [RL98]. Sur la figure 5.11(c), l'écartement entre deux éléments de la hiérarchie est fonction de r_i. On voit clairement se dégager deux classes, l'une rassemblant les objets dont l'abscisse est proche de 0 et l'autre rassemblant ceux dont l'abscisse est proche de 1.

On peut associer à une hiérarchie une distance entre individus telle que la distance d_h entre a et b est le plus petit niveau de partition qui contient les deux éléments. Par exemple, les points $(0.86, 0.92)$ et $(0.88, 0.75)$ de la figure 5.11 ont une distance $d_h = 1$ tandis que les points $(0.86, 0.92)$ et $(0.24, 0.91)$ ont une distance $d_h = 6$.

Cette distance entre objets définie grâce à une hiérarchie indicée est une distance ultramétrique. Une distance d_h est dite ultramétrique si :

$$d_h(a, a) = 0 \tag{5.29}$$
$$d_h(a, b) = d_h(b, a) \tag{5.30}$$
$$d_h(a, b) \leq \max(d_h((a, c), d(b, c)) \tag{5.31}$$

Algorithme de classification ascendante hiérarchique

L'algorithme de Classification Ascendante Hiérarchique (CAH) se propose de construire une hiérarchie indicée monotone en partant de la partition la plus fine de l'ensemble des objets à classer. Tout d'abord, on calcule la demi-matrice D de distance entre les objets de U. On cherche alors le couple d'indices (i, j) tel que $D(i, j)$ est minimal. On fusionne les deux éléments de la hiérarchie h_s et h_t associés à ces indices et on indice ce nouvel élément $h_u = h_s \cup h_t$ par $r_u = D(i, j)$. On calcule ensuite les distances entre ce nouvel élément et les autres éléments de la hiérarchie selon une méthode particulière, comme celle définie par l'équation 5.33. L'algorithme est itéré jusqu'à obtenir la partition la moins fine.

Pour des raisons d'efficacité, on souhaite éviter de revenir aux objets eux-mêmes pour calculer la distance entre l'union de deux éléments h_i et h_j de la hiérarchie et les éléments restants $h_k, \forall k \neq (i, j)$. Pour ce faire, une méthode consiste à prendre la distance la plus faible entre $d(h_i, h_k)$ et $d(h_j, h_k)$. La distance entre classes est alors celle des objets des deux classes les plus proches.

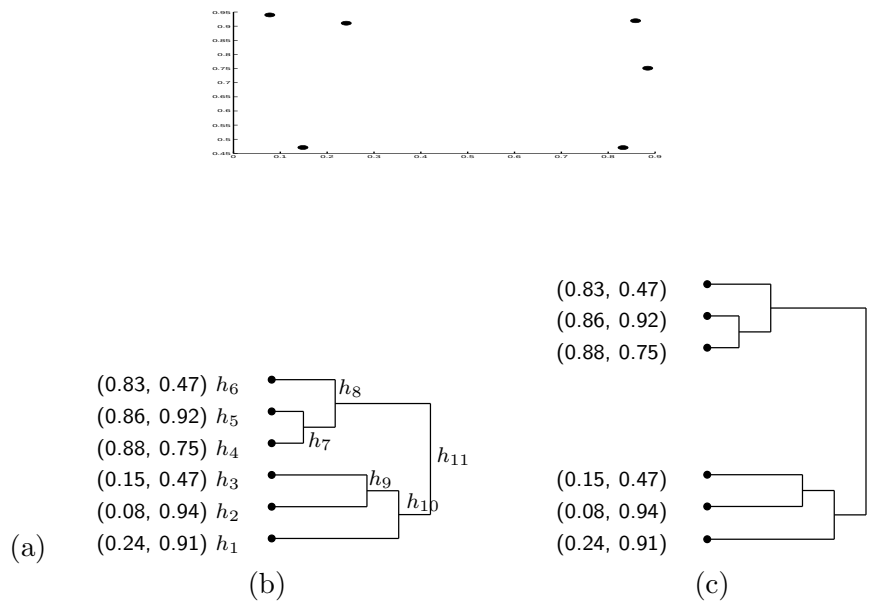


FIG. 5.11 – Dendrogrammes représentant une hiérarchie monotone (à gauche) et une hiérarchie monotone indicée (à droite) des points représentés en haut sur un plan. L'indice est exploité pour écarter plus ou moins les deux éléments à fusionner. On voit clairement se dégager deux classes, l'une avec des objets avec l'abscisse proche de 0 et l'autre avec l'abscisse proche de 1.

Une autre méthode, dite de Ward [War63], se propose de minimiser pendant l'agrégation l'inertie intra-classe. Soit une partition de K classes, l'inertie intra-classe mesure son homogénéité :

$$I = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(\mathcal{C}_k(i), \overline{\mathcal{C}_k}) \tag{5.32}$$

où $\mathcal{C}_k$ est une classe de n_k élément et $\overline{\mathcal{C}_k}$ est son barycentre, qui a pour attributs les attributs moyens des éléments dans la classe.

On calcule la distance entre l'union de deux éléments h_i et h_j de la hiérarchie et un autre élément h_k, respectivement de cardinaux n_i, n_j et n_k, comme suit :

$$d(h_i \cup h_j, h_k) = \frac{n_k + n_i}{n_k + n_j + n_i} d(h_i, h_k) + \frac{n_k + n_j}{n_k + n_j + n_i} d(h_j, h_k) + \frac{n_i + n_j}{n_k + n_j + n_i} d(h_i, h_j) \tag{5.33}$$

Cette méthode est adaptée à la classification de vecteurs de données scalaires et donne de bons résultats pour notre application. À titre d'exemple, considérons 12 points sur un plan dont les attributs sont leurs coordonnées, voir figure 5.12(a). Les figures 5.12(b) et 5.12(c) sont respectivement les dendrogrammes obtenus avec la méthode du minimum et la méthode de Ward. On note un effet de chaînage pour la méthode du minimum, rendant la hiérarchie difficile

à interpréter. La méthode de Ward génère un dendrogramme plus équilibré, ce qui permet de détecter plus facilement les classes pertinentes.

5.4.3 Algorithme

L'algorithme présenté dans cette section se propose de partitionner un groupe de partiels $\mathcal{G}_i$ en entités. Pour cela, on utilise la dissimilarité entre partiels utilisant les erreurs de prédiction croisées d'_σ calculée à partir des vecteurs de fréquence ou d'amplitude des partiels du groupe $\mathcal{G}_i$. Le support temporel de ces vecteurs doit être identique. Une hiérarchie de classes est alors générée grâce à l'algorithme de CAH. Pour déterminer le nombre de classes, de multiples critères sont envisageables. Dans notre algorithme, si la dérivée de l'indice de hiérarchie $r_{i+1} - r_i$ est supérieure à un seuil (égal à 0.1 sur la figure 5.13), une nouvelle classe est identifiée. Les partiels dont les vecteurs de données sont dans la même classe sont regroupés dans une même entité.

On représente sur les figures 5.13 et 5.14 les dendrogrammes résultant de classifications à partir des vecteurs de fréquence ou des vecteurs d'amplitude de l'ensemble de partiels représenté sur la figure 5.10. Le dendrogramme généré à partir des vecteurs de fréquence est très satisfaisant, les partiels sont correctement regroupés et les partiels de bruit insérés dans la hiérarchie en dernier, avec un indice r_i très élevé, ce qui permet de les éliminer facilement. Le dendrogramme généré à partir des vecteurs d'amplitude est moins bon, sans être tout à fait inexploitable. Une combinaison des deux classifications permettrait d'obtenir une répartition des partiels plus robuste.

La figure 5.15 montre les résultats d'une expérience mettant en jeu trois notes de hauteurs différentes d'un même piano, jouées à une intensité similaire. Les dendrogrammes obtenus à partir des vecteurs de fréquence et d'amplitude ne sont pas parfaits, mais présentent tout de même des résultats intéressants. Certaines corrélations sont clairement visibles, notamment pour les partiels de la note la plus aiguë (classe 3).

Conclusion

On a montré dans ce chapitre l'intérêt d'une modélisation sinusoïdale à long terme pour l'extraction d'informations de haut niveau qui permettent de structurer la représentation en agrégeant les partiels selon des critères pertinents. En particulier, cette modélisation à long terme permet l'utilisation de l'heuristique "ancien et nouveau" pour réduire le nombre d'entités à détecter simultanément. Ensuite, les évolutions à long terme des paramètres de fréquence ou d'amplitude peuvent être exploités. Ces évolutions se révèlent utiles pour agréger des partiels sans aucun *a priori* harmonique. L'analyse des variations (vibrato ou trémolo) et des micro-modulations permet donc de retirer des informations pertinentes. Ces informations sont utiles pour l'agrégation de partiels mais peuvent sans doute être d'intérêt pour la description d'entités sonores pour des applications de reconnaissance d'instruments ou de locuteur.

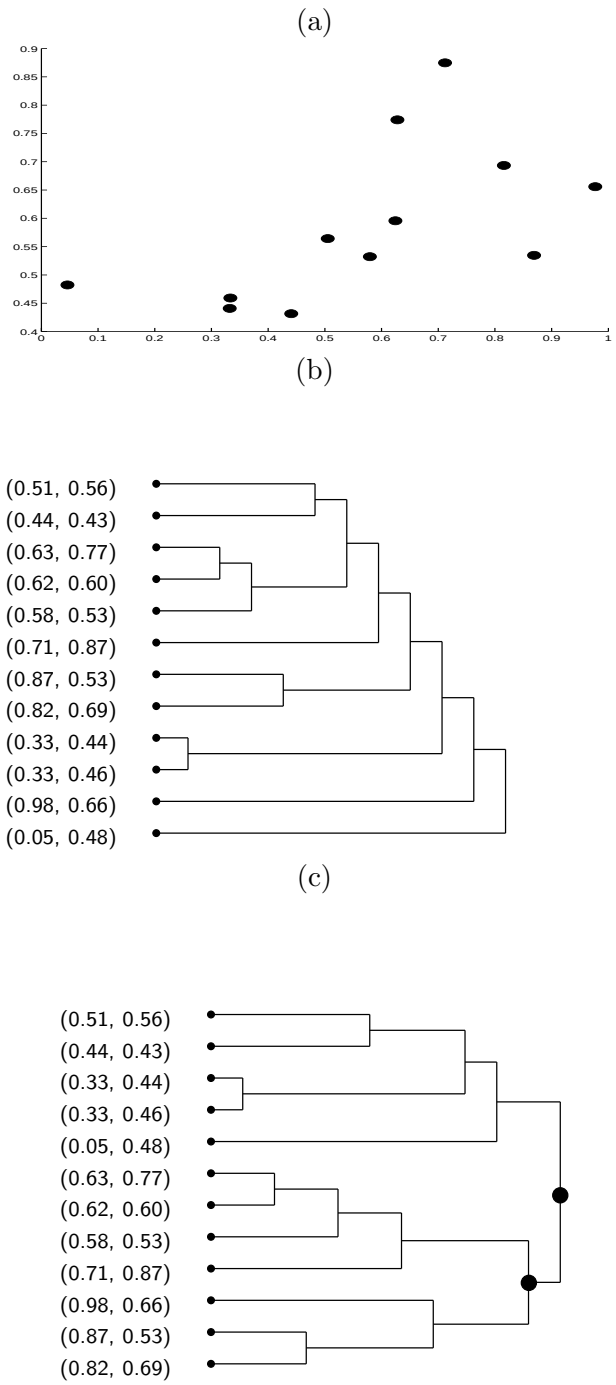


FIG. 5.12 – Dendrogrammes de deux hiérarchies obtenus avec le lien minimal (au milieu) et la méthode de Ward (en bas). La première présente un effet de chaîne qui la rend peu exploitable. La seconde, amène un arbre plus équilibré et permet une classification des objets en deux ou trois classes.

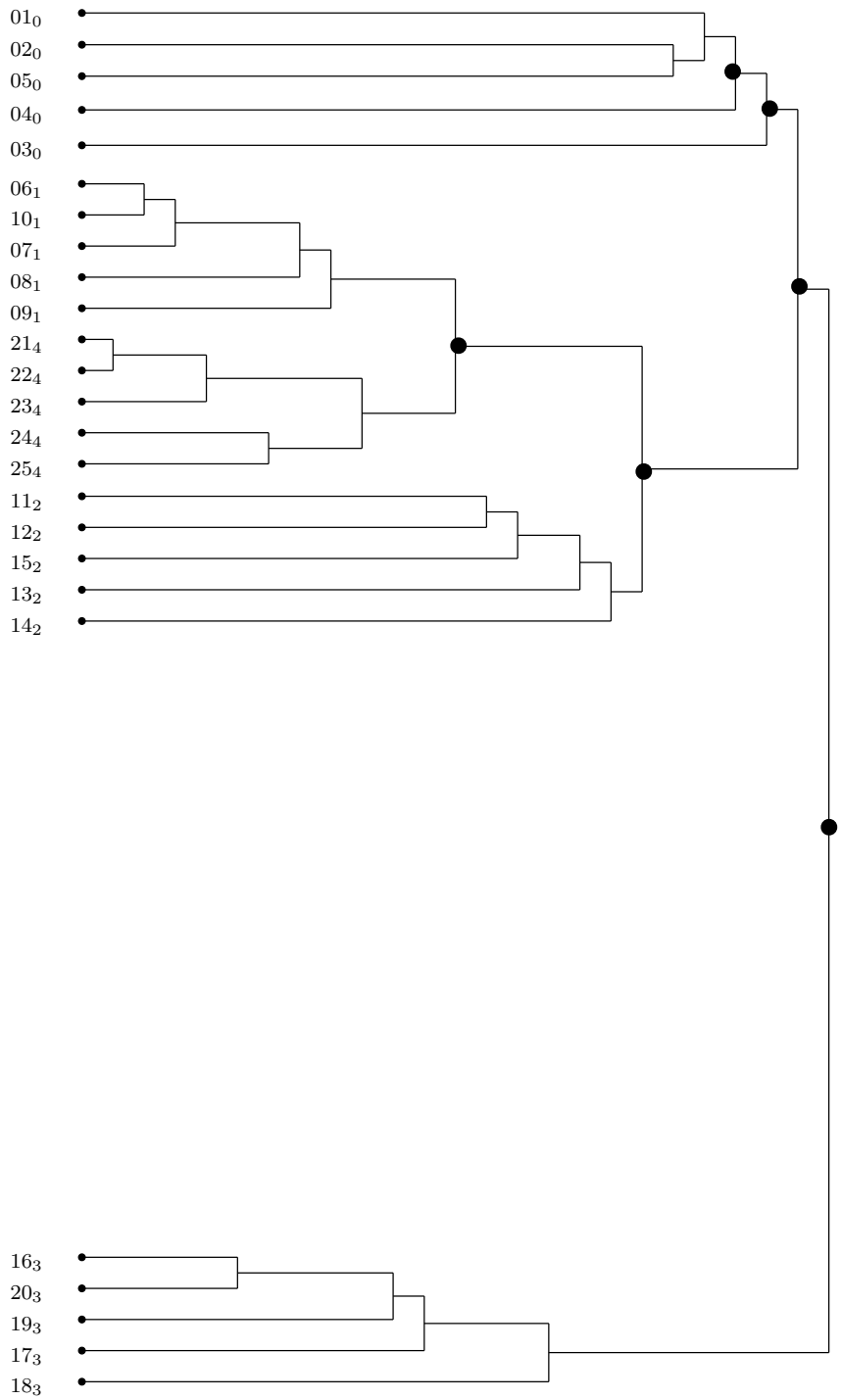


FIG. 5.13 – Dendrogramme obtenu à partir des distances entre vecteurs de fréquence pour les partiels de notes d'instruments différents (représentés sur la figure 5.10) avec la dissimilarité d'_σ. Ce dendrogramme permet de distinguer les classes de partiels pertinentes (les coupures sont représentées par des points) et de rejeter les partiels de bruit (classe d'indice 0).

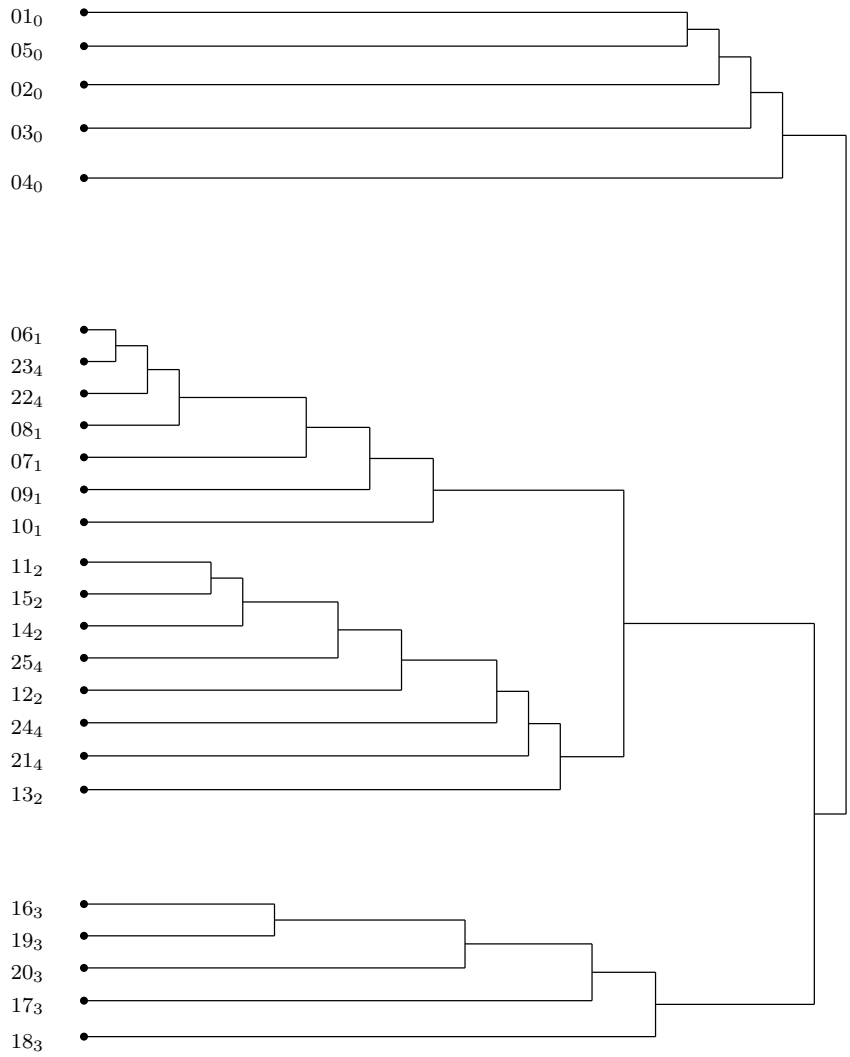


FIG. 5.14 – Dendrogramme obtenu à partir des distances entre vecteurs d'amplitude pour les partiels de notes d'instruments différents (représentés sur la figure 5.10) avec la dissimilarité d'_σ. Ce dendrogramme n'est pas directement exploitable car les partiels des classes $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_4$ sont mélangées, mais certaines corrélations sont visibles.

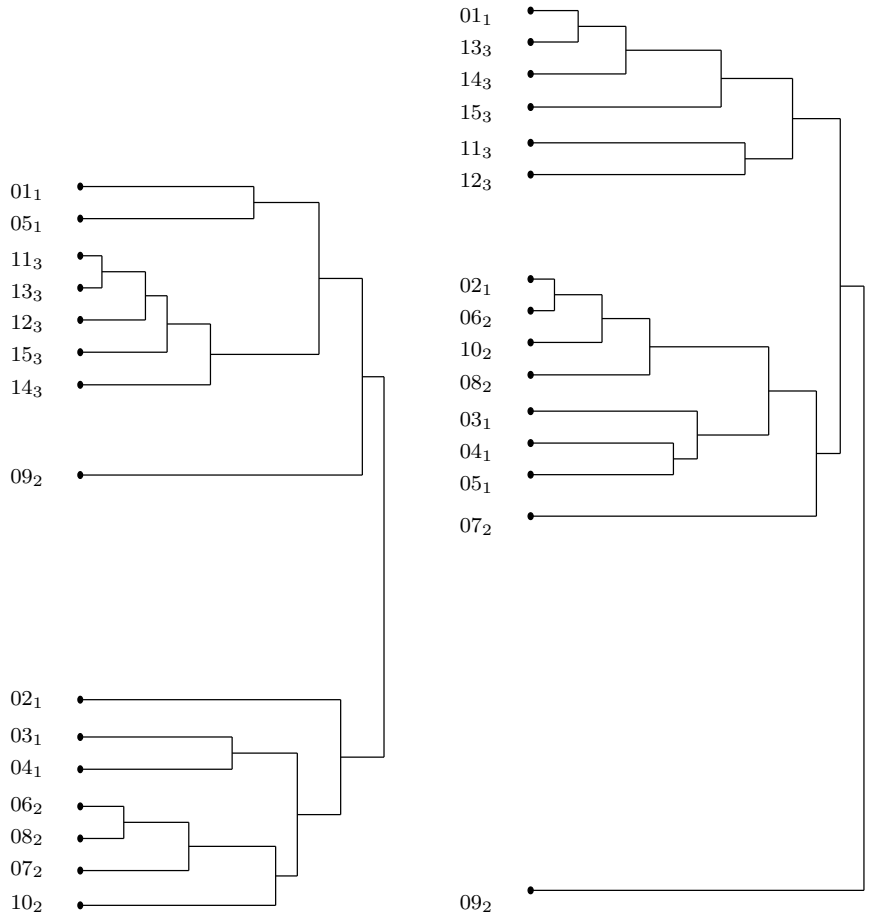


FIG. 5.15 – Dendrogrammes obtenus à partir des distances entre vecteurs de fréquence (à gauche) et d'amplitude (à droite) pour cinq partiels de trois notes de piano de hauteurs différentes avec la dissimilarité d'_σ.

Conclusions et Perspectives

Dans ce document, nous nous sommes efforcés d'améliorer les possibilités d'interprétation d'une modélisation à long terme de la partie quasi périodique d'enregistrements polyphoniques. Une bonne interprétation nécessite avant tout une estimation précise des paramètres. Dans le chapitre 2, nous avons présenté une interprétation trigonométrique d'un estimateur de la fréquence d'une sinusoïde basé sur la dérivée du signal [DCM00]. Cette nouvelle formulation a permis d'en améliorer la précision dans les hautes fréquences. En outre, l'estimation du paramètre de phase est importante pour être à même de retrancher la partie quasi périodique du signal original et ainsi obtenir un résidu d'énergie minimale. Malheureusement, l'estimation de la phase se révèle être sensible aux modulations de fréquence. Un premier estimateur qui nécessite une connaissance préalable de la variation de la fréquence dans la trame d'analyse est, dans le cas de signaux synthétiques, cent fois plus précis que l'estimateur classique. Un second estimateur qui présente l'avantage d'être indépendant de toute connaissance concernant les variations d'amplitude et de fréquence est ensuite proposé. Des tests similaires montrent que cet estimateur est dix fois plus précis que l'estimateur classique.

La présence de perturbations telles que des composantes transitoires ou des composantes de bruit amène la présence de nombreux pics qui n'appartiennent pas à la partie quasi périodique du signal analysé. On propose à la fin du chapitre 2 une extension au cas non stationnaire d'un critère de sélection de pics de de manière à ne pas rejeter les pics issus de composantes sinusoïdales modulées comme celle rencontrées lors de vibrato ou de trémolo.

Néanmoins, nous sommes convaincus que la décision d'allouer un pic particulier à une composante sinusoïdale ne doit pas se faire à un instant donné mais en combinant les résultats d'analyses effectuées en plusieurs instants successifs. Nous nous sommes donc concentrés dans le chapitre 3 sur une approche long-terme pour obtenir une représentation de la partie quasi périodique qui soit fidèle et interprétable. Malheureusement, l'application d'heuristiques simples pour le suivi de partiels se révèle être souvent inopérante dans le cas de signaux polyphoniques. L'utilisation de contraintes inhérentes au modèle comme le caractère prédictible des évolutions des paramètres des partiels de même que l'absence théorique de hautes fréquences dans ces évolutions permettent de proposer de nouveaux algorithmes de suivi adaptés au problème posé. Les tests des capacités intrinsèques ainsi que ceux effectués dans une chaîne complète d'analyse/synthèse montrent la pertinence de ces approches. La meilleure sélectivité apportée par l'utilisation de la prédiction linéaire permet de mieux rejeter les

composantes de bruit et l'analyse du contenu haute fréquence permet d'obtenir de meilleurs résultats dans le cas de partiels dont les fréquences sont proches ou se croisent, phénomènes fréquents lors de l'analyse de signaux polyphoniques. La représentation long-terme obtenue est alors plus interprétable et permet l'extraction d'informations de haut niveau. En particulier, les informations de début et de fin ainsi que les modulations des paramètres des partiels sont mieux identifiées.

À titre de perspective, nous pensons que l'exploitation d'enregistrements stéréophoniques pour extraire des informations spatiales associées aux composantes DFT peut permettre d'améliorer le problème de la contamination des composantes DFT par la présence de plusieurs harmoniques de fréquence proche. De plus, ces informations spatiales peuvent être utiles lors du regroupement de partiels en entités sonores car la proximité spatiale est un indice important.

Nous avons vu dans le chapitre 4 que la modélisation autorégressive permet de conserver les modulations lors de l'interpolation des paramètres d'amplitude et de fréquence. Ensuite, une méthode pour répartir l'erreur de phase permet de garantir une interpolation du signal synthétisé sans discontinuités aux bornes. Cette modélisation est intégrée dans un module d'interpolation générique qui s'applique aussi bien à l'interpolation d'une représentation sinusoïdale à long terme qu'à la restauration d'entités sonores et d'enregistrements musicaux. Les tests d'écoute ont montré que cette approche permet d'améliorer sensiblement la qualité subjective de l'interpolation des paramètres sinusoïdaux notamment grâce à la conservation des micro-modulations, phénomène important pour la perception.

Les algorithmes proposés dans le chapitre 5 exploitent les bonnes propriétés de la représentation à long terme pour apporter un niveau de structuration supplémentaire en agrégeant certains partiels de manière à former des entités sonores. L'utilisation d'une telle représentation permet notamment d'appliquer l'heuristique "ancien et nouveau" formulée dans [Bre90] et ainsi concilier efficacité et robustesse. Cette représentation permet aussi de détecter des débuts de notes avec des attaques très douces.

Pour former une entité perceptuelle, les partiels qui ont débuté en même temps doivent présenter des paramètres de fréquence qui évoluent de manière corrélée. On a montré dans la dernière partie du chapitre 5 que la mesure par erreurs de prédiction croisées permet de détecter de manière robuste la similarité entre deux vecteurs de fréquence ou d'amplitude. En effet, l'utilisation d'une telle mesure de dissimilarité permet de comparer la partie prédictible des deux vecteurs de manière indépendante de l'échelle tout en conservant la partie imprédictible à la même échelle.

L'utilisation de la classification ascendante hiérarchique (CAH) permet ensuite de classer ces partiels en fonction de cette dissimilarité. Cependant, la CAH opère une allocation exclusive. Un partiel ne peut donc être alloué qu'à une seule entité sonore. Or, certains partiels extraits peuvent être la représentation d'harmoniques de plusieurs entités du fait du manque de précision fréquentielle de l'analyse spectrale. Alternativement, l'utilisation de $2-3$ hiérarchies [Ber02] qui permettent une allocation non exclusive des éléments aux classes est envisagée

pour répondre à ce problème.

L'évolution corrélée des fréquences des partiels est une condition nécessaire à la perception d'une unique entité sonore comprenant ces partiels. Néanmoins, comme détaillé dans [Bre90], cette condition n'est pas suffisante car de nombreux autres facteurs peuvent être pris en compte. Il serait donc judicieux de combiner différentes mesures de similarité entres partiels comme l'harmonicité, l'évolution corrélée en fréquence ou en amplitude. Malheureusement, ces mesures étant d'unités différentes, leur combinaison est problématique.

À chaque hiérarchie obtenue en classant les partiels selon un critère particulier est associée une ultramétrique indépendante de l'unité de la mesure utilisée pour obtenir cette hiérarchie. La combinaison de différentes ultramétriques obtenues grâce aux classifications des même partiels selon des critères différents constitue une piste de recherche privilégiée. En effet, la classification multicritères est non seulement une problématique d'intérêt pour la formation d'entités sonores mais aussi pour regrouper certaines entités sonores sous forme de voix et ainsi apporter un niveau de structuration supplémentaire à la représentation de la partie quasi périodique d'un enregistrement musical.

Bibliographie

[AF95] François Auger and Patrick Flandrin. Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method. *IEEE Transactions on Signal Processing*, 43 :1068–1089, May 1995.

[BAM02] Rémi Boyer and Karim Abed-Meraim. Audio Transients Modeling by Damped and Delayed Sinusoids (DDS). In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, July 2002.

[Bar81] Thomas P. Barnwell. Recursive Windowing for Generating Autocorrelation Coefficients for LPC Analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(5) :1062–1066, October 1981.

[Ber02] P. Bertrand. Set Systems for Which Each Set Properly Intersects at Most One Other Set - Application to Pyramidal Clustering. In *IFCS2002, Classification, Clustering, and Data Analysis*, pages 38–39, July 2002.

[BP78] Albert S. Bregman and Steven Pinker. Auditory Streaming and the Building of Timbre. *Canadian Journal of Psychology*, 32(1) :19–31, 1978.

[Bra99] Karlheinz Brandenburg. MP3 and AAC explained. In *Proc. of AES 17th International Conference on High Quality Audio Coding (Florence)*, september 1999.

[Bre90] Albert S. Bregman. *Auditory Scene Analysis : The Perceptual Organization of Sound*. The MIT Press, 1990.

[Bro92] Judith C. Brown. Musical Fundamental Frequency Tracking Using a Pattern Recognition Method. *Journal of the Audio Engineering Society*, 92(3) :1394–1400, 1992.

[BS03] Juan Pablo Bello and Mark Sandler. Phase-Based Note Onset Detection for Music Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 441–444, April 2003.

[Bur75] John P. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, 1975.

[BW04] Mark A. Bartsch and Gregory H. Wakefield. Singing Voice Identification Using Spectral Envelope Estimation. *IEEE Transactions on Speech and Audio Processing*, 12(2) :100 – 109, 2004.

[Bék60] G. Békési. *Experiments in Hearing*. New-York : Mc Graw Hill, 1960.

[Car02] Grégory Cartier. Séparation de Sources Harmoniques. Master's thesis, Université Bordeaux 1, F-33405 Talence cedex, France, June 2002. In french.

[CM02] Antoine Cornuéjols and Laurent Miclet. *Apprentissage Artificiel*. Eyrolles, 2002.

[Cra46] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[Dau00] Laurent Daudet. *Représentation structurelle de signaux audiophoniques - Méthodes hybrides pour des applications à la compression*. PhD thesis, Université Aix-Marseille I, 2000. In french.

[DBDS03] Christopher Duxbury, Juan Pablo Bello, Mike Davies, and Mark Sandler. Complex Domain Onset Detection for Musical Signals. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 90–94. University of Limerick and COST (European Cooperation in the Field of Scientific and Technical Research), September 2003.

[dBSO02] Bert den Brinker, Erik Schuijers, and Werner Oomen. Parametric Coding for High-Quality Audio. In *112th Convention of the Audio Engineering Society*. Audio Engineering Society (AES), May 2002.

[DCM00] Myriam Desainte-Catherine and Sylvain Marchand. High Precision Fourier Analysis of Sounds Using Signal Derivatives. *Journal of the Audio Engineering Society*, 48(7/8) :654–667, July/August 2000.

[DDS01] Christopher Duxbury, Mike Davies, and Mark Sandler. Separation of Transient Information in Musical Audio using Multiresolution Analysis Techniques. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 1–4. University of Limerick and COST (European Cooperation in the Field of Scientific and Technical Research), December 2001.

[DG86] Susan R. Dubois and Filson H. Glanz. An Autoregressive Model Approach to Two-Dimensional Shape Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1) :55 – 66, 1986.

[DGR93a] Philippe Depalle, Guillermo Garcia, and Xavier Rodet. Analysis of Sound for Additive Synthesis : Tracking of Partials Using Hidden Markov Models. In *Proceedings of the International Computer Music Conference (ICMC)*, San Francisco, 1993. International Computer Music Association (ICMA).

[DGR93b] Philippe Depalle, Guillermo Garcia, and Xavier Rodet. Tracking of Partials for Additive Sound Synthesis Using Hidden Markov

Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 225–228, April 1993.

[DM80] Steven Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4) :357 – 366, 1980.

[DR91] Boris Doval and Xavier Rodet. Estimation of Fundamental Frequency of Musical Sound Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 3657 – 3660, April 1991.

[Ett96] Wilfried Etter. Restoration of a Discrete-Time Signal Segment by Interpolation Based on the Left-Sided and Right-Sided Autoregressive Parameters. *IEEE Transactions on Signal Processing*, 44(5) :1124–1135, 1996.

[FCQ98] Paulo Fernandez and Javier Casajus-Quiros. Multi-Pitch Estimation for Polyphonic Musical Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3565–3568, April 1998.

[FH96] Kelly Fitz and Lippold Haken. Sinusoidal Modeling and Manipulation Using Lemur. *Computer Music Journal*, 20(4) :44–59, Winter 1996.

[Fit99] Kelly R. Fitz. *The Reassigned Bandwith-Enhanced Method of Additive Synthesis*. PhD thesis, University of Illinois, 1999.

[FRD92] Adrian Freed, Xavier Rodet, and Philippe Depalle. Synthesis and Control of Hundreds of Sinusoidal Partials on a Desktop Computer without Custom Hardware. In *Proceedings of the ICSPAT'92 Conference*, 1992.

[FRD93] Adrian Freed, Xavier Rodet, and Philippe Depalle. Performance, Synthesis and Control of Additive Synthesis on a Desktop Computer Using FFT^{-1}. In *Proceedings of the International Computer Music Conference (ICMC)*, Tokyo, Japan, 1993. International Computer Music Association (ICMA).

[Gar92] Guillermo Garcia. Analyse des signaux sonores en termes de partiels et de bruit. Extraction automatique des trajets fréquentiels par des modèles de Markov cachés. Master's thesis, Orsay, France, 1992. In french.

[GDF73] Jr. G. David Forney. The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3) :268–278, March 1973.

[GL85] Daniel W. Griffin and Jae S. Lim. A New Model-Based Speech Analysis/Synthesis System. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tampa, 1985.

[GMdM+03] Laurent Girin, Sylvain Marchand, Jospeh di Martino, Axel Röbel, and Geoffroy Peeters. Conparing the order of a Polynomial Phase

Model for the Synthesis of Quasi-Harmonic Audio Signals. In *WASPAA*, New Paltz, NY, USA, October 2003. IEEE.

[Goo97] Michael Goodwin. Matching pursuit with damped sinusoids. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 2037 – 2040, April 1997.

[Gor84] John William Gordon. *Perception of Attack Transients in Muscial Tones*. PhD thesis, Stanford University, 1984.

[GS97] Bryan E. George and Mark J. T. Smith. Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model. *IEEE Transactions on Speech and Audio Processing*, 5(5) :389–406, September 1997.

[Han03] Pierre Hanna. *Modélisation statistique de sons bruités : étude de la densité spectrale, analyse, transformation musicale et synthèse*. PhD thesis, Bordeaux 1 University, LaBRI, 2003. in french.

[Har88] William Morris Hartmann. *Pitch Perception and the Segregation and Integration of Auditory Entities*. Gerald M. Edelman, W. Einar Gall, and W. Maxwell Cowan, editors, Auditory Function : Neurobiological Bases of Hearing, 1988.

[Hay91] Simon Haykin. *Adaptive Filter Theory*. Prentice Hall, 1991.

[HDC01] Pierre Hanna and Myriam Desainte-Catherine. Influence of Frequency Distribution on Intensity Fluctuation of Noise. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 120–124. University of Limerick and COST (European Cooperation in the Field of Scientific and Technical Research), December 2001.

[HDC03] Pierre Hanna and Myriam Desainte-Catherine. Time Scale modification of Noises Using a Spectral and Statistical Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2003.

[Hem63] Hermann Hemholtz. *Die Lehre von der Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Brunswick, Germany : Vieweg-Verlag, 1863.

[HM03] Stephen W. Hainsworth and Malcom D. Macleod. On Sinusoidal Parameter Estimation. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 151–156. Queen Mary, University of London, September 2003.

[HMW01] Stephen W. Hainsworth, Malcom D. Macleod, and Patrick J. Wolfe. Analysis of Reassigned Spectrograms for Musical Transcription. In *WASPAA*, pages 23–26, October 2001.

[Hol87] Sverre Holm. FFT Pruning Aplied to Time Domain Interpolation and Peak Localization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(12) :1776–1777, 1987.

[IRC96] IRCAM, Paris. *AudioSculpt User's Manual*, second edition, April 1996.

[JVV86] Augustus J. E. M. Janssen, Raymond N. J. Veldhuis, and Lodewijk B. Vries. Adaptive Interpolation of Discrete-Time Signals that can be Modeled as Autoregressive Processes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(2) :317–330, 1986.

[Kay88] Steven M. Kay. *Modern Spectral Estimation*, chapter Autoregressive Spectral Estimation : Methods, pages 228–231. Signal Processing Series. Prentice Hall, 1988.

[KAZ00] Florian Keiler, Daniel Arfib, and Udo Zölzer. Efficient Linear Prediction for Digital Audio Effects. In *Proceedings of the Digital Audio Effects (DAFx) Conference*. Università degli Studi di Verona and COST (European Cooperation in the Field of Scientific and Technical Research), December 2000.

[KAZ01] Florian Keiler, Daniel Arfib, and Udo Zölzer. Extraction Sinusoids from Harmonics Signals. *Journal of New Music Research*, 30(3) :243–258, 2001.

[KKS01] Ismo Kauppinen, Jyrki Kauppinen, and Pekka Saarinen. A Method for Long Extrapolation of Audio Signals. *Journal of the Audio Engineering Society*, 49(12) :1167–1180, December 2001.

[KKZS03] Florian Keiler, Can Karadogan, Udo Zölzer, and Albrecht Schneider. Analysis of Transient Musical Sounds by Auto-Regressive Modeling. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 301–304. Queen Mary, University of London, December 2003.

[Kla99] Anssi Klapuri. Sound Onset Detection by Applying Psychoacoustic Knowledge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3089 – 3092, 1999.

[Kla03] Anssi Klapuri. Multiple Fundamental Frequency Estimation by Harmonicity and Spectral Smoothness. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 6(11) :804–816, 2003.

[KM02] Florian Keiler and Sylvain Marchand. Survey on Extraction of Sinusoids in Stationary Sounds. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 51–58. University of the Federal Armed Forces - Hamburg, Germany, September 2002.

[Kop86] Gary Kopec. Formant Tracking Using Hidden Markov Models and Vector Quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4) :709 – 729, 1986.

[KR02] Ismo Kauppinen and Kari Roth. Audio Signal Extrapolation – Theory and Applications. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 105–110. University of the Federal Armed Forces - Hamburg, Germany, September 2002.

[Lag01] Mathieu Lagrange. Accélération de la Synthèse Sonore. Master's thesis, Université Bordeaux 1, F-33405 Talence cedex, France, June 2001. In french.

[LM01] Mathieu Lagrange and Sylvain Marchand. Real-Time Additive Synthesis of Sound by Taking Advantage of Psychoacoustics. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 5–9. University of Limerick and COST (European Cooperation in the Field of Scientific and Technical Research), December 2001.

[LMR02] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault. Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 59–64. University of the Federal Armed Forces - Hamburg, Germany, September 2002.

[LMR04a] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault. Partial Tracking Based on Future Trajectories Exploration. In *116th Convention of the Audio Engineering Society*, Berlin, May 2004. Audio Engineering Society (AES). Preprint 6046 (10 pages).

[LMR04b] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault. Using Linear Prediction to Enhance the Tracking of Partials. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 241–244, may 2004.

[LMRR03] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault. Enhanced Partial Tracking Using Linear Prediction. In *Proceedings of the Digital Audio Effects (DAFx) Conference*, pages 141–146. Queen Mary, University of London, September 2003.

[LS99] Scot Nathan Levine and Julius O. Smith. A Switched Parametric and Transform Audio Coder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 985–988, March 1999.

[LVS99] Scot Nathan Levine, Tony S. Verma, and Julius O. Smith. Multiresolution Sinusoidal Modeling for Wideband Audio with Modifications. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3585–3588, March 1999.

[MA86] Joachim Marques and Luis Almeida. A Background for Sinusoid Based Representation of the Voiced Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1233–1236, Tokyo, 1986.

[Mah94] Robert C. Maher. A Method for Extrapolation of Missing Digital Audio Data. *Journal of the Audio Engineering Society*, 42(5) :350–357, May 1994.

[Mak75] John Makhoul. Linear Prediction : A Tutorial Review. *Proceedings of the IEEE*, 63(4) :561–580, November 1975.

[Mar71] John D. Markel. FFT Pruning. *IEEE Transactions on Audio and Electroacoustics*, 19(4) :305–311, 1971.

[Mar00a] Sylvain Marchand. InSpect+ProSpect+ReSpect Software Packages. Online. URL : `http://www.scrime.u-bordeaux.fr`, 2000.

[Mar00b] Sylvain Marchand. *Sound Models for Computer Music (analysis, transformation, synthesis)*. PhD thesis, University of Bordeaux 1, LaBRI, December 2000.

[Mas96] Paul Masri. *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, 1996.

[MB94] Robert C. Maher and James W. Beauchamp. Fundamental Frequency Estimation of Musical Signals Using a Two-Way Mismatch Procedure. *Journal of the Audio Engineering Society*, (4) :2254 – 2263, 1994.

[MC81] John Makhoul and Lynn K. Cosell. Adaptive Lattice Analysis of Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3) :654–658, June 1981.

[MC97] Michael W. Macon and Mark A. Clements. Sinusoidal Modeling and Modification of Unvoiced Speech. *IEEE Transactions on Speech and Audio Processing*, 5(6) :557–560, 1997.

[MC98] Paul Masri and Nishan Canagarajah. Extracting More Detail from the Spectrum with Phase Distortion Analysis. In *Proceedings of the Digital Audio Effects (DAFx) Conference*. Audiovisual Institute, Pompeu Fabra University and COST (European Cooperation in the Field of Scientific and Technical Research), November 98.

[McA84] Stephen McAdams. *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*. PhD thesis, Stanford University, 1984.

[McA89] Stephen McAdams. Segregation of Concurrrents Sounds : Effects of Frequency Modulation Coherence. *Journal of the Audio Engineering Society*, 86(6) :2148–2159, 1989.

[Mel91] David K. Mellinger. *Event Formation and Separation in Musical Sound*. PhD thesis, Stanford University, 1991.

[ML03a] Aaron S. Master and Yi-Wen Liu. Nonstationary Sinusoidal Modeling with Efficient Estimation of Linear Frequency Chirp Parameters. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.

[ML03b] Aaron S. Master and Yi-Wen Liu. Robust Chirp Parameter Estimation for Hann Windowed Signals. In *IEEE International Conference on Multimedia and Exposition (ICME)*, 2003.

[Mol03] Stéphane Molla. *Audiophonic Signals : Hybrid Modelling and Coding Scheme*. PhD thesis, Université Aix-Marseille I, 2003. In french.

[Moo03] Brian C.J. Moore. *Introduction to the Psychology of Hearing.* Academic Press, 2003.

[MP02] Nikolaus Meine and Heiko Purnagen. Fast Sinusoid Synthesis for MPEG-4 HILN Parametric Audio Decoding. In *Proceedings of the Digital Audio Effects (DAFx) Conference.* University of the Federal Armed Forces - Hamburg, Germany, September 2002.

[MPE92] ISO MPEG2. ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5Mbit/s, standard n°11172, alias 'MPEG-1' ISO-MPEG, November 1992.

[MQ86] Robert J. McAulay and Thomas F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4) :744–754, 1986.

[MQ92] Robert J. McAulay and Thomas F. Quatieri. Shape Invariant Time-Scale and Pitch Modification of Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 40(3) :497–510, 1992.

[MS99] Sylvain Marchand and Robert Strandh. InSpect and ReSpect : Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 341–344, Beijing, China, October 1999. International Computer Music Association (ICMA).

[NHD98] Joost Nieuwenhuijse, Richard Heusdens, and Ed F. Deprettere. Robust Exponential Modeling of Audio Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3581–3584, May 1998.

[OS89] Alan V. Oppenheim and Ronald W. Schaffer. *Discrete time signal processing.* Prentice Hall, 1989.

[PB87] T.W. Parks and C.S. Burrus. *Digital Filter Design.* John Wiley & Sons, 1987.

[PGV97] Paolo Prandom, Mickael Goodwin, and Martin Vetterli. Optimal Time Segmentation for Signal Modeling and Compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 2029 – 2032, April 1997.

[PJ03] Sébastien Paris and Claude Jauffret. Frequency Line Tracking using HMM-based Schemes [passive sonar]. *IEEE Transactions on Aerospace and Electronic Systems*, 39(2) :439 – 449, 2003.

[PM00] Heiko Purnhagen and Nikolaus Meine. HILN - The MPEG-4 Parametric Audio Coding Tools. In *IEEE International Symposium on Circuits and Systems (ISCAS 2000)*, volume 3, pages 201–204, May 2000.

[Por76] Michael R. Portnoff. Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3) :243–248, 1976.

[PR99] Geoffroy Peeters and Xavier Rodet. SINOLA : A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum. In *Proceedings of the International Computer Music Conference (ICMC)*, Beijing, China, October 1999. International Computer Music Association (ICMA).

[PS00] Ted Painter and Andreas Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4) :451–515, 2000.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C (The Art of Scientific Computing)*, chapter 10 : Minimization or Maximization of Functions, pages 402–405. Cambridge University Press, USA, 2nd edition, 1992.

[Pug90] William Pugh. Skip Lists : A Probabilistic Alternative to Balanced Trees. In *Communications of the ACM*, volume 33, pages 668–676, June 1990.

[QD90] Thomas F. Quatieri and Ronald G. Danisewicz. An Approach to Co-Channel Talker Interference Suppression using a Sinusoidal Model for Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(1) :56–69, January 1990.

[Ras78] R. A. Rash. Synchronization in Performed Ensemble Music. *Acustica*, 43 :121–131, 1978.

[RD93] Xavier Rodet and Philippe Depalle. Spectral Envelope and Inverse FFT Synthesis. In *93rd Convention of the Audio Engineering Society*, San Francisco, 1993. Audio Engineering Society (AES).

[RG04] Julie Rosier and Yves Grenier. Unsupervised Classification Techniques for Multipitch Estimation. In *116th Convention of the Audio Engineering Society*. Audio Engineering Society (AES), May 2004.

[Ris91] Jean-Claude Risset. *Representation of Musical Signals*, chapter 1 : Timbre Analysis by Synthesis : Representations, Imitations, and Variants for Musical Composition, pages 7–43. MIT Press, Cambridge, Massachusetts, 1991.

[RJ86] Lawrence Rabiner and Biing-Hwang Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 1986.

[RL98] Valérie Rouat and Israel César Lerman. Problématique de la coupure dans la résolution de #sat par sériation. In *Actes de JNPC'98 (Résolution Pratique de Problèmes NP-Complets)*, École des Mines de Nantes, France, 1998.

[RMC96] Wilfried Roguet, Nadine Martin, and Alain Chehikian. Tracking of frequency in a time-frequency representation. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 341 – 344, June 1996.

[RR04] Sang-Uk Ryu and Kenneth Rose. Advances in Sinusoidal Analysis/Synthesis-Based Error Concealment in Audio Networ-

king. In *116th Convention of the Audio Engineering Society*. Audio Engineering Society (AES), May 2004.

[Röb02] Axel Röbel. Estimating Partial Frequency and Frequency Slope Using Reassignment Operators. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 122 – 125, 2002.

[SB90] Ros L. Streit and Ross F. Barrett. Frequency Line Tracking using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(4) :586–598, 1990.

[Sch66] Pierre Schaeffer. *Traité des Objets Musicaux*. Seuil, Paris, 1966. In French.

[Ser89] Xavier Serra. *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[Ser97] Xavier Serra. *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pages 91–122. Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[Ski76] David P. Skinner. Pruning the Decimation In-Time FFT Algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 193–194, 1976.

[SOdBG02] Erik G.P. Schuijers, Werner Oomen, Bert den Brinker, and Andy J. Gerrits. Advances in Parametric Coding for High-Quality Audio. In *IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA)*, November 2002.

[SS87] Julius O. Smith and Xavier Serra. An Analysis/Synthesis Program for Non-Harmonic Sounds based on a Sinusoidal Representation. In *Proceedings of the International Computer Music Conference*, San Francisco, 1987. Computer Music Association.

[SS90] Xavier Serra and Julius O. Smith. Spectral Modeling Synthesis : A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, 14(4) :12–24, 1990.

[SW98] Andrew Sterian and Gregory H. Wakefield. A Model-Based Approach to Partial Tracking for Musical Transcription. SPIE annual meeting, San Diego, California, 1998.

[Val91] Jean Christophe Valière. *La Restauration des Enregistrements Anciens par Traitement Numérique - Contribution à l'Étude de Quelques Techniques Récentes*. PhD thesis, Maine University, le Mans, 1991. In french.

[Vas88a] Saeed Vaseghi. A New Application of Adaptive Filters for Restoration of Archived Gramophone Recordings . In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 2548–2551, April 1988.

[Vas88b] Saeed Vaseghi. *Algorithm for Restoration of Archived Gramophone Recordings*. PhD thesis, University of Cambridge, Department of Engineering, 1988.

[Vas92] Saeed Vaseghi. Restoration of old gramophone recordings. *Journal of the Audio Engineering Society*, 40(10) :791–801, 1992.

[VHK01] Renat Vafin, Richard Heusden, and Bastiaan Kleijn. Modifying Transients for Efficient Coding of Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 3285 – 3288, May 2001.

[VK00] Tuomas Virtanen and Anssi Klapuri. Separation of Harmonic Sound Sources Using Sinusoidal Modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 765–768, April 2000.

[VM98] Tony S. Verma and Teresa H.Y. Meng. An Analysis/Synthesis Tool for Transient Signals that Allows a Flexible Sines+Transients+Noise Model for Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3573–3576, May 1998.

[VVHK99] Koen Vos, Renat Vafin, Richard Heusdens, and Bastiaan Kleijn. High-Quality Consistent Analysis-Synthesis in Sinusoidal Coding. In *AES 17th International Conference on High-Quality Audio Coding*, September 1999.

[War63] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58 :238 – 244, 1963.

[WS85] Bernard Widrow and Salmuel D. Stearns. *Adaptive Signal Processing Algortihms*. Prentice Hall, 1985.

[XE91] Xianya Xie and Robin J. Evans. Multiple Target Tracking and Multiple Frequency Line Tracking using Hidden Markov Models. *IEEE Transactions on Signal Processing*, 39(12) :2659–2676, 1991.

[ZF90] Eberhard Zwicker and Richard Feldtkeller. *Psychoacoustics Facts and Models*. Springer Verlag, 1990.

(132 références)

Annexe A

Estimateur de fréquence trigonométrique

On propose une approche purement trigonométrique du problème de l'estimation de la fréquence d'un cosinus. Soit un cosinus d'amplitude a et de pulsation ω exprimée en radians par secondes à deux instants différents t et $t - \Delta$:

$$s(t) = a\ \cos(\omega\, t + \phi) \tag{A.1}$$
$$s(t - \Delta) = a\ \cos(\omega\, (t - \Delta) + \phi) \tag{A.2}$$

La différence et la somme de ce signal aux deux instants t et $t - \Delta$ peuvent s'exprimer sous la forme :

$$s(t) - s(t - \Delta) = a'\ \cos(\omega\, t + \phi') \tag{A.3}$$
$$s(t) + s(t - \Delta) = a''\ \cos(\omega\, t + \phi'') \tag{A.4}$$

avec

$$a' = 2a \sin\left(\frac{\omega\, \Delta}{2}\right) \tag{A.5}$$
$$a'' = 2a \cos\left(\frac{\omega\, \Delta}{2}\right) \tag{A.6}$$

et ϕ' et ϕ'' deux phases particulières :

$$\phi' = \phi + (\omega\Delta)/2 \tag{A.7}$$
$$\phi'' = \phi - (\omega\Delta)/2 + \pi \tag{A.8}$$

On en déduit que :

$$\frac{a'}{a} = 2 \sin\left(\frac{\omega\, \Delta}{2}\right) \tag{A.9}$$
$$\frac{a''}{a} = 2 \cos\left(\frac{\omega\, \Delta}{2}\right) \tag{A.10}$$

On peut déduire des équations A.9 et A.10 deux estimateurs de fréquence f^+ et f^- respectivement basés sur l'addition et la soustraction du signal à un instant t et ce même signal à un instant $t - \Delta$:

$$f^- = \frac{1}{\pi\,\Delta} \arcsin\left(\frac{a'}{a}\right) \tag{A.11}$$

$$f^+ = \frac{1}{\pi\,\Delta} \arccos\left(\frac{a''}{a}\right) \tag{A.12}$$

Ces équations montrent que la fréquence d'un signal sinusoïdal peut être estimée à partir des amplitudes du signal et des signaux "différence" et "somme" entre le signal et sa version retardée.

Nous allons maintenant étudier les performances de ces deux estimateurs. Pour cela, on utilise des signaux échantillonnés à $F_e = 44100$ Hz et on fixe $\Delta = 1/F_e$. Les signaux tests sont des sinusoïdes de fréquence variant entre 0 et $F_e/2$ par pas de 10 Hz auxquelles on ajoute un bruit blanc $b(n)$ d'énergie variable, soit :

$$s(n) = a\,\cos\left(\frac{2\pi}{F_e} f\right) + b(n) \tag{A.13}$$

On calcule $s^+(n)$ et $s^-(n)$, les signaux somme et différence :

$$s^+(n) = s(n) + s(n-1) \tag{A.14}$$

$$s^-(n) = s(n) - s(n-1) \tag{A.15}$$

Les amplitudes des trois signaux sont estimées par le calcul de l'énergie sur 2048 points, soit :

$$\overline{s(n)} = \sqrt{2\sum_{n=0}^{2047} s^2(n)} \tag{A.16}$$

$$\overline{s^-(n)} = \sqrt{2\sum_{n=0}^{2047} s^-(n)^2} \tag{A.17}$$

$$\overline{s^+(n)} = \sqrt{2\sum_{n=0}^{2047} s^+(n)^2} \tag{A.18}$$

$$\tag{A.19}$$

En effet, l'énergie d'une sinusoïde d'amplitude a est égale à $e = a^2/2$. Les estimateurs f^- et f^+ sont alors :

$$f^- = \frac{F_e}{\pi} \arcsin\left(\frac{\overline{s^-(n)}}{2\overline{s(n)}}\right) \tag{A.20}$$

$$f^+ = \frac{F_e}{\pi} \arccos\left(\frac{\overline{s^+(n)}}{2\overline{s(n)}}\right) \tag{A.21}$$

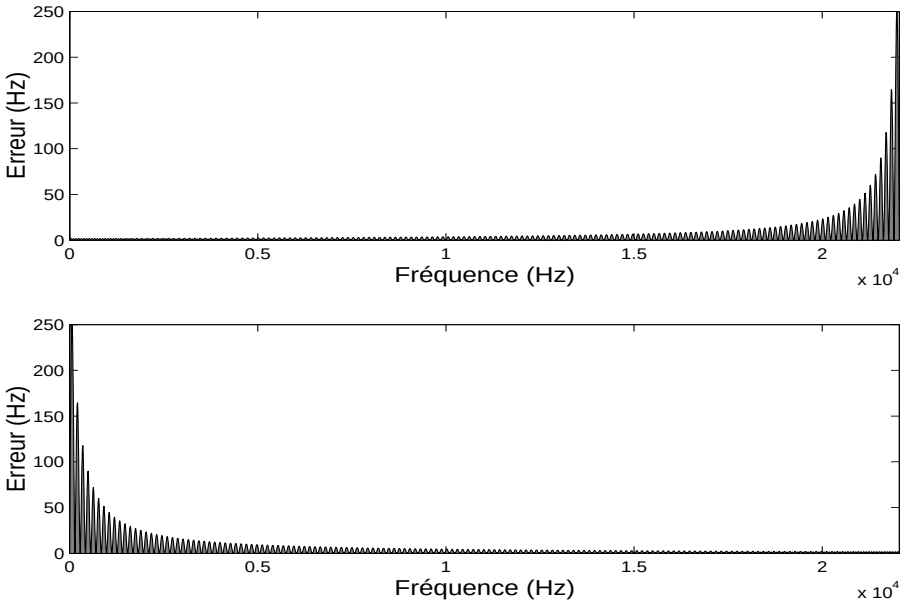


FIG. A.1 – En haut, erreur de l'estimateur f^- en fonction de la fréquence du cosinus analysé. En bas, erreur de l'estimateur f^+. Ces deux estimateurs ont des erreurs réparties symétriquement sur l'axe des fréquences.

Les résultats sont représentés dans la figure A.1. On observe que f^- a une erreur qui grandit au fur et à mesure que l'on se rapproche de la fréquence de Nyquist, à l'inverse de f^+ qui a une erreur qui grandit au fur et à mesure que l'on se rapproche de la fréquence 0.

Les performances de ces deux estimateurs sont liées aux fonctions arcsin et arccos utilisées dans les équations A.11 et A.12. Ces fonctions ne sont pas des fonctions de transfert linéaires. Si l'argument est proche de 1, une erreur faible sur l'argument amène une erreur conséquente sur l'estimation de la fréquence et inversement, voir figure A.2(b). Or, l'argument de la fonction arcsin dans l'équation A.11 tend vers 1 en la fréquence de Nyquist et l'argument de la fonction arcsin dans l'équation A.12 tend vers 1 en la fréquence 0 comme on peut le voir sur la figure A.2(a). Il est donc utile de considérer l'estimateur f^- lorsque la fréquence recherchée est inférieure à $F_e/4$ et de considérer l'estimateur f^+ lorsque la fréquence recherchée est supérieure à $F_e/4$ pour être plus résistant aux erreurs d'estimation des amplitudes.

De manière à tester la résistance au bruit de ces estimateurs, on ajoute un bruit gaussien à un cosinus de fréquence constante pour différents rapports de signal à bruit (SNR) exprimé en dB. L'estimateur testé est le suivant :

$$\hat{f}_{\pm} = \begin{cases} f^+ & \text{si } f^- > F_e/4 \text{ et } f^+ > F_e/4 \\ f^- & \text{sinon} \end{cases} \tag{A.22}$$

Comme on le voit sur la figure A.3, cet estimateur donne des résultats acceptables pour un SNR supérieur à 40 dB. Pour un niveau de SNR inférieur, un biais peut être noté. Selon la nature du bruit ajouté, l'estimateur va subir une

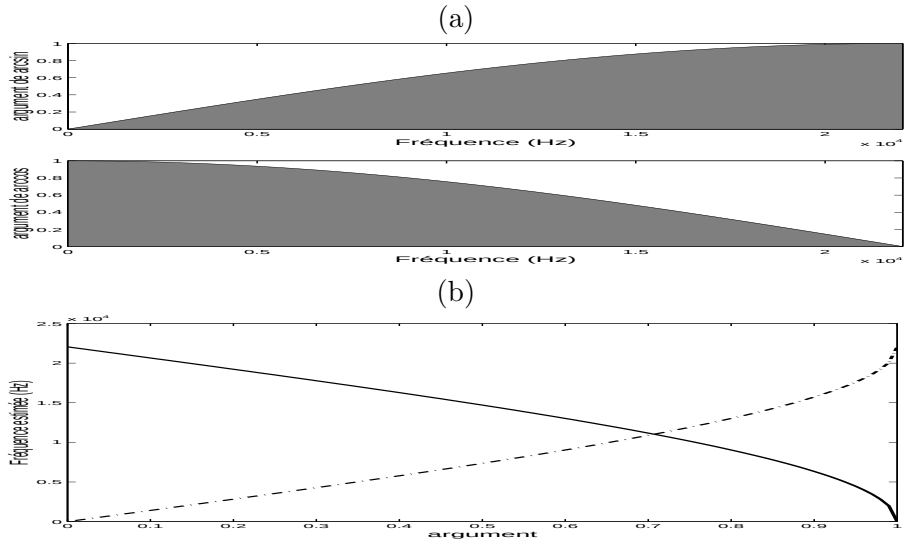


FIG. A.2 – En haut, évolutions des arguments des fonctions arcsin et arccos des équations A.11 et A.12 en fonction de la fréquence. En bas, fonctions de transferts arcsin et arccos utilisées dans les équations A.11 et A.12. Si l'argument (a) est proche de 1, une petite erreur sur l'argument amènera une grande erreur sur l'estimation de la fréquence et inversement.

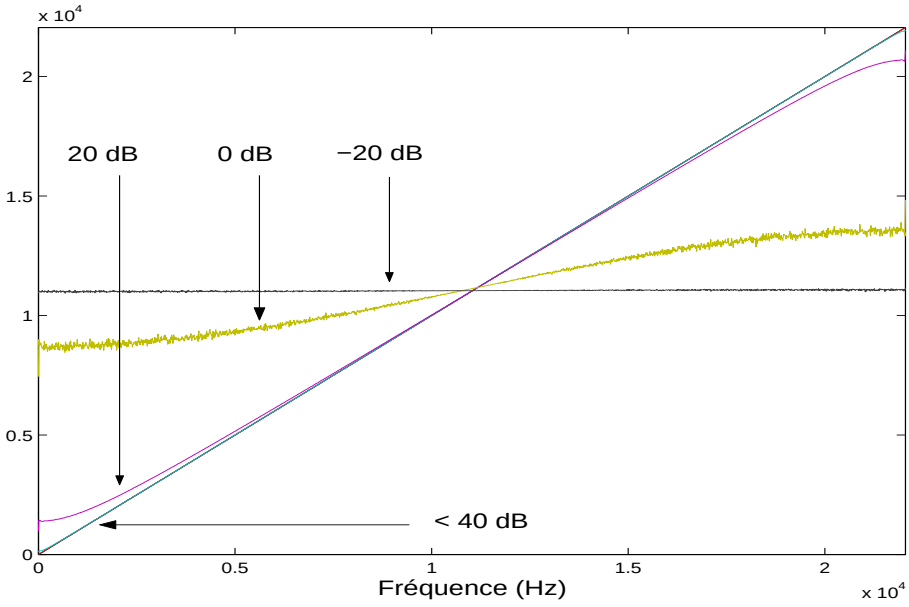


FIG. A.3 – Fréquence estimée en utilisant l'équation A.22 pour un signal composé d'un cosinus et d'un bruit gaussien de SNR donné, en fonction de la fréquence du cosinus analysé. La résistance au bruit de cet estimateur est assez faible.

attraction vers une fréquence donnée. Dans le cas d'un bruit blanc, la fréquence d'attraction est $F_e/4$. En effet, lorsque le rapport signal à bruit est très faible, on a : $\overline{s(n)}^2 = \overline{b(n)}^2$ et $\overline{s^+(n)}^2 = \overline{s^-(n)}^2 = 2\,\overline{b(n)}^2$. Les fréquences estimées sont alors :

$$f^+ = \frac{F_e}{\pi} \arccos\left(\frac{1}{\sqrt{2}}\right) = \frac{F_e}{4} \tag{A.23}$$

$$f^- = \frac{F_e}{\pi} \arcsin\left(\frac{1}{\sqrt{2}}\right) = \frac{F_e}{4} \tag{A.24}$$

Application aux composantes sinusoïdales multiples

Ces estimateurs sont limités à l'estimation de la fréquence d'un seul cosinus et le niveau de bruit toléré est assez faible. Or, les signaux musicaux comportent souvent de multiples composantes sinusoïdales. Pour estimer leurs fréquences, on doit disposer pour chaque composante, d'une estimation de l'amplitude de cette composante, de l'amplitude du signal "somme" de cette composante et de l'amplitude du signal "différence" de cette composante.

Soit un signal composé de N composantes sinusoïdales :

$$s(n) = \sum_{i=1}^{N} s_i(n) \tag{A.25}$$

$$s_i(n) = a_i \cos\left(\frac{2\pi}{F_e} f_i\, n + \phi_i\right) \tag{A.26}$$

On veut estimer les fréquences f_i. On pose :

$$s^-(n) = s(n) - s(n-1) = \sum_{i=1}^{N} s_i^-(n) \tag{A.27}$$

$$s^+(n) = s(n) + s(n-1) = \sum_{i=1}^{N} s_i^+(n) \tag{A.28}$$

où $s_i^-(n)$ et $s_i^+(n)$ sont des composantes sinusoïdales de fréquence f_i. On peut montrer que chaque fréquence f_i peut être estimée par :

$$f_i^- = \frac{F_e}{\pi} \arcsin\left(\frac{\overline{s_i^-(n)}}{2\overline{s_i(n)}}\right) \tag{A.29}$$

et

$$f_i^+ = \frac{F_e}{\pi} \arccos\left(\frac{\overline{s_i^+(n)}}{2\overline{s_i(n)}}\right) \tag{A.30}$$

où $\overline{s_i(n)}$, $\overline{s_i^-(n)}$ et $\overline{s_i^+(n)}$ sont respectivement les amplitudes des composantes de $s_i(n)$, $s_i^-(n)$ et $s_i^+(n)$. Pour estimer ces amplitudes, on fenêtre le signal par une fenêtre $w(n)$ et on utilise une DFT de taille N, choisie de manière à ce que les fréquences de 2 composantes du signal $s(n)$ soient éloignées d'au moins F_e/N.

Dans ce cas, chaque composante $s_i(n)$, $s_i^-(n)$ et $s_i^+(n)$ donnera un maximum local en l'indice DFT k_i tel que :

$$\frac{(k_i - 0.5)\,F_e}{N} \leq f_i \leq \frac{(k_i + 0.5)\,F_e}{N} \tag{A.31}$$

On peut montrer que :

$$|S[k_i]| = \overline{s_i(n)}\,K \tag{A.32}$$
$$|S^-[k_i]| = \overline{s_i^-(n)}\,K \tag{A.33}$$
$$|S^+[k_i]| = \overline{s_i^+(n)}\,K \tag{A.34}$$

avec

$$K = \frac{1}{2}\left|W\left(f_i - \frac{k_i\,F_e}{N}\right)\right| \tag{A.35}$$

où $W(f)$ est le spectre de la fenêtre d'analyse utilisée. Les estimateurs de f_i deviennent alors :

$$f_i^- = \frac{F_e}{\pi}\arcsin\left(\frac{\overline{S^-[k_i]}}{2\overline{S[k_i]}}\right) \tag{A.36}$$
$$f_i^+ = \frac{F_e}{\pi}\arccos\left(\frac{\overline{S^+[k_i]}}{2\overline{S[k_i]}}\right) \tag{A.37}$$

En posant,

$$S^-[k] = \text{DFT}\,[s(n) - s(n-1)] = \frac{1}{F_e}X^1[k] \tag{A.38}$$
$$S[k] = \text{DFT}\,[s(n)] = X^0[k] \tag{A.39}$$

on trouve :

$$f_i^- = \frac{F_e}{\pi}\arcsin\left(\frac{1}{2F_e}\frac{|X^1[k_i]|}{|X^0[k_i]|}\right) \tag{A.40}$$

soit l'estimateur correspondant à la méthode de la dérivée [DCM00]. L'estimateur f_i^+ est donc une amélioration de la méthode de la dérivée pour les hautes fréquences, c'est-à-dire supérieures à $F_e/4$, voir équation A.24.

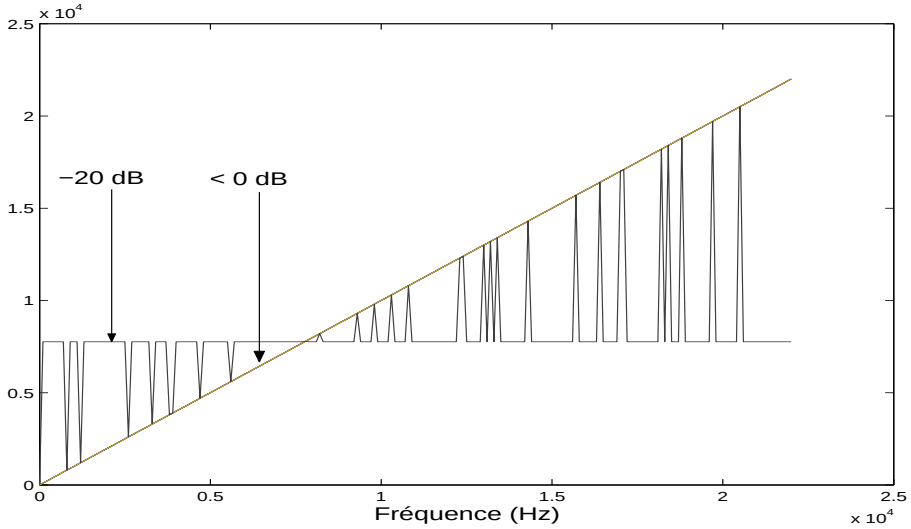


FIG. A.4 – Fréquence estimée en utilisant l'estimateur $\hat{f}$ calculé grâce à l'équation A.46 pour un signal composé d'un cosinus et d'un bruit gaussien en fonction de la fréquence du cosinus. La résistance au bruit est grandement améliorée par l'utilisation de la DFT.

Mise en œuvre

Les équations donnant f_i^- et f_i^+ montrent qu'il faut calculer la transformée de Fourier du signal à analyser, du signal "somme" et du signal "différence". En remarquant que,

$$S^-[k] = \text{DFT}\,[s^-(n)] = \text{DFT}\,[s(n)] - \text{DFT}\,[s(n-1)] \quad \text{(A.41)}$$

$$S^+[k] = \text{DFT}\,[s^+(n)] = \text{DFT}\,[s(n)] + \text{DFT}\,[s(n-1)] \quad \text{(A.42)}$$

$$S_0[k] = \text{DFT}\,[s(n)] \quad \text{(A.43)}$$

il apparaît que seules 2 transformées de Fourier sont nécessaires :

$$S_0[k] = \text{DFT}\,[s(n), \cdots, s(n+N-1)] \quad \text{(A.44)}$$

et

$$S_1[k] = \text{DFT}[s(n-1), \cdots, s(n+N)] \quad \text{(A.45)}$$

Ainsi, pour chaque maximum local d'indice k_i du spectre de puissance $|S_0|$, la fréquence estimée est :

$$\hat{f}_i = \begin{cases} \frac{F_e}{\pi} \arcsin\left(\frac{|S_1[k_i]-S_0[k_i]|}{|S_0[k_i]|}\right) & \text{si } k_i < N/4 \\ \frac{F_e}{\pi} \arccos\left(\frac{|S_0[k_i]+S_1[k_i]|}{|S_0[k_i]|}\right) & \text{sinon} \end{cases} \quad \text{(A.46)}$$

Grâce au filtrage passe-bande opéré par la DFT, la résistance au bruit de cet estimateur est sensiblement améliorée comme on peut le constater sur la figure A.4.

Sinusoidal Modeling of Polyphonic Sounds

Abstract

The aim of this thesis is to study a structured representation of polyphonic sounds.

Some peaks are selected from successive short-time spectra. An algorithm called *partial tracking* links some of those peaks from frame to frame. This algorithm forms partials : quasi-sinusoidal oscillators with parameters evolving slowly and continuously with time. Next, partials are clustered into acoustical entities on the basis of correlation cues so that each entity is no longer perceived by the human auditory system as multiple simple tones but as a unique complex tone.

Several constraints of the analysis of polyphonic signals lead to a spectral representation with artefacts, i.e. some peaks may be corrupted or missing. The tracking of partials across this corrupted spectral representation requires new tracking methods that use characteristics of the sinusoidal model. The predictability of the evolutions of the parameters of the partials as well as the theoretical lack of high frequencies in these evolutions are exploited to propose new algorithms useful for our purposes.

A set of tests simulating realistic degradations is proposed to evaluate tracking algorithms according to these criteria. The algorithms are tested in "stand alone" mode and in a complete analysis / synthesis module.

Several psychoacoustic cues may then be used to cluster partials like their common onset, their harmonic relation, and the correlated evolutions of their parameters. Initially, onsets are detected using the precision property of partials ; partials having simultaneous onsets are grouped. The partials of these groups are then clustered into acoustical entities using either an iterative algorithm based on harmonicity estimation or a simultaneous algorithm based on the similarity of the evolution of the parameters. Clustering methods gain efficiency and robustness because the use of a long-term representation of the spectrum allows a partial to be clustered only once and to be avoided during the detection of new entities.

After this clustering process, some partials may be incomplete due to the presence of several acoustical entities in the same frequency band. These partials are then completed by a novel interpolation algorithm based on the linear prediction of the frequency and amplitude parameters while avoiding phase discontinuities at the boundaries of the missing region.

Discipline : Computer Science

Keywords :
sound modeling,
audio material indexing,
audio segmentation,
source separation,
real-time additive synthesis.

Modélisation Sinusoïdale des Sons Polyphoniques

Résumé

L'objet de cette thèse est l'étude d'une représentation structurée pour les sons polyphoniques.

Certains maxima locaux du spectre de puissance, communément appelés *pics*, sont sélectionnés dans une succession de spectres à court terme. Un algorithme dit de *suivi de partiels* relie ensuite de trame en trame certains de ces pics pour former des partiels : oscillateurs quasi-sinusoïdaux dont les paramètres de fréquence et d'amplitude évoluent lentement et de façon continue au cours du temps.

Les contraintes liées à l'analyse de signaux polyphoniques amènent des artefacts dans la représentation spectrale, des pics sont manquants ou corrompus. Ces dégradations rendent le plus souvent inopérantes les techniques de suivi de partiels basées sur des heuristiques simples. L'utilisation des contraintes relatives au modèle sinusoïdal, comme le caractère prédictible des évolutions des paramètres des partiels de même que l'absence théorique de hautes fréquences dans ces évolutions permettent de proposer de nouveaux algorithmes de suivi adaptés au problème posé.

L'ensemble des partiels extraits se doit d'avoir de bonnes propriétés pour permettre ensuite une agrégation sans ambiguïté. Une première série de tests évalue un algorithme pour ses propriétés intrinsèques, tandis que la seconde série évalue les propriétés de cet algorithme lorsqu'il est inclus dans un module complet d'analyse / synthèse.

À la suite d'un défaut de transmission ou d'une insuffisance de l'analyse, certains partiels peuvent être incomplets ou dégradés. Ces partiels sont alors reconstruits selon un algorithme original d'interpolation basé sur la prédiction linéaire des paramètres de fréquences et d'amplitude évitant les discontinuités de phase aux bornes de la zone interpolée.

Enfin, les partiels présentant certaines corrélations sont ensuite agrégés pour former des entités sonores, chaque entité étant perçue par le système auditif humain non plus comme plusieurs sons simples mais comme un unique son complexe. On utilise pour cela plusieurs indices issus d'études psychoacoustiques tels que l'apparition simultanée de partiels, leur relation d'harmonicité et les évolutions corrélées de leurs paramètres.

Discipline : Informatique

Mots-clés :
modélisation du signal sonore,
indexation et segmentation de contenus sonores,
séparation de sources,
synthèse en temps réel.

U.F.R. de Mathématiques et d'Informatique

LaBRI
Université Bordeaux 1
351, cours de la Libération
F-33405 Talence cedex, FRANCE