Segmentation et indexation des signaux sonores musicaux

Stéphane Rossignol

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

Spécialité :

Acoustique, traitement du signal et informatique appliqués à la musique

Présentée par

M. Stéphane ROSSIGNOL

pour obtenir le grade de DOCTEUR de l'UNIVERSITÉ PARIS 6

Sujet de la thèse :

SEGMENTATION ET INDEXATION DES SIGNAUX SONORES MUSICAUX

Soutenue le 12 07 2000 devant le jury composé de :

Madame	Régine	ANDRÉ-OBRECHT	RAPPORTEUR
Monsieur	Jean-Luc	COLLETTE	EXAMINATEUR
Monsieur	Nicolas	MOREAU	RAPPORTEUR
Monsieur	Pierrick	PHILIPPE	EXAMINATEUR
Monsieur	Jean-Dominique	POLACK	EXAMINATEUR
Monsieur	Xavier	RODET	DIRECTEUR DE THÈSE

TRAVAIL EFFECTUÉ À :

L'IRCAM – CENTRE GEORGES POMPIDOU 1, PLACE IGOR-STRAVINSKY, 75004 PARIS
SUPÉLEC – CAMPUS DE METZ 2, RUE ÉDOUARD BELIN, 57078 METZ

Remerciements

Je tiens à remercier tout d'abord Philippe Depalle et Xavier Rodet pour m'avoir offert l'opportunité de faire une partie de ma thèse au sein de l'équipe analyse-synthèse à l'ircam , et Joël Soumagne pour m'avoir offert l'opportunité de faire l'autre partie de ma thèse au sein de l'équipe Systèmes de Traitement du Signal (STS) de Supélec – Campus de Metz ; ainsi que pour leur accueil, leur disponibilité et la qualité de leur encadrement.

Cette thèse a été effectuée grâce à un financement du ccett. Je tiens à exprimer me profonde gratitude à Pierrick Philippe.

Je remercie également Jean-Luc Collette, avec qui j'ai eu le plaisir de travailler et dont j'ai pu apprécier la grande disponibilité et les ingénieuses idées.

J'ai eu l'occasion d'encadrer un stagiaire, Dragos Spataru. Son apport à ce travail concerne la *segmentation en sources*.

Je tiens à remercier ma famille, mes amis de l'ircam, mes amis de Metz, et ceux d'ailleurs, pour leur soutien.

Je remercie enfin les membres du jury, Madame Régine ANDRÉ-OBRECHT (Rapporteur), Monsieur Jean-Luc COLLETTE (Examinateur), Monsieur Nicolas MOREAU (Rapporteur), Monsieur Pierrick PHILIPPE (Examinateur), Monsieur Jean-Dominique POLACK (Examinateur), Monsieur Xavier RODET (Directeur de thèse), pour avoir accepté de participer à ce jury et pour s'être intéressé à ce travail. J'en suis très honoré.

Table des matières

Liste des tableaux

Table des figures

Première partie

Introduction

0.1 Enjeu

Le travail de thèse présenté dans cet exposé est le résultat de la collaboration de trois centres de recherche : le Centre Commun d'Études en Télécommunications et en Télédiffusion (CCETT) à RENNES, l'Institut de Recherche et Coordination Acoustique/Musique (IRCAM) à PARIS, et SUPÉLEC – CAMPUS DE METZ. Il concerne la segmentation, l'indexation et la manipulation des sons.

L'indexation et la segmentation des sons, quels qu'ils soient (parole, musique, bruit...), d'où qu'ils viennent (radios, bandes son de films, CD...), sont des domaines qui sont en plein essort, du fait notamment de l'émergence de MPEG-7. MPEG-7 est un standard en cours de développement. Sera défini un ensemble d'outils de description de « contenus » multimédia, pour en faciliter la recherche et l'identification. Ce standard international sera fixé dans les mois qui viennent, normalement en septembre 2001.

Dans cet exposé nous proposons et nous étudions des techniques pour segmenter, étiqueter et, dans une moindre mesure, manipuler les signaux sonores. Nous définissons des objets (un objet est un « segment étiqueté »). Ces objets sont extraits suivant des procédures hiérarchisées. Cette hiérarchisation se fait au sens d'une description de plus en plus précise des sons. Les étapes de l'extraction sont successives et complémentaires.

Nous nous sommes plus particulièrement intéressé aux signaux sonores musicaux. Relativement à ce qui existe en ce qui concerne la parole, très peu de travaux ont été consacrés à la caractérisation des signaux sonores musicaux. Le premier à s'être intéressé au problème (ou du moins à un problème connexe : la transcription automatique) est MOORER en 1975 dans sa thèse « On the segmentation and analysis of continuous musical sound by digital computer » (voir [Moo75]). L'objectif de MOORER était la transcription automatique de sons composés au plus de deux voix. Il s'imposa des restrictions importantes, restrictions dont nous voulons nous affranchir.

- Seuls les sons harmoniques, c'est-à-dire composés d'une succession de notes, sont considérés : nous nous intéressons aussi aux sons inharmoniques.
- Pour segmenter (en notes ici), n'est considéré que le trajet de la dérivée de l'énergie : nous nous intéressons aussi aux variations brusques de la hauteur et du contenu spectral. Il faut remarquer de plus que la procédure de segmentation n'est pas automatique : le seuillage de la dérivée de l'énergie est fait « à la main ».
- Seule la musique est traitée : dans une certaine mesure, nous traitons aussi la parole.
- Les sons où un vibrato, et/ou un trémolo, et/ou un glissando sont présents sont rejetés : nous traitons aussi ces sons.
- Une limite de deux voix (c'est-à-dire deux sons) harmoniques mixées est fixée.
- Il faut que la longueur des notes soit supérieure à 80 millisecondes : nous ne voulons pas faire d'hypothèse sur la longueur des notes.
- Il faut que la fondamentale des sons soit présente (ce n'est pas le cas ni pour la voix parlée téléphonique, ni, en théorie du moins, pour certains instruments, comme le basson) : les logiciels de l'IRCAM f_0, ADDITIVE et HMM s'affranchissent de ce problème.
- Il faut que le peigne harmonique soit dense, c'est-à-dire qu'il ne manque pas d'harmoniques (ce n'est pas le cas, en théorie du moins, pour la clarinette, pour laquelle les harmoniques pairs manquent, ou sont d'amplitude très faible) : f_0, ADDITIVE et HMM s'affranchissent de ce problème.

Notre objectif, dans cette thèse, est de traiter le plus de types de sons musicaux possible (harmoniques, percussifs...), le plus automatiquement possible. Des techniques sont développées et validées pour les sons « simples », et sont adaptées pour un certain nombre de sons plus « compliqués ». Les limitations de ces techniques sont discutées. Un ensemble d'outils informatiques, documentés et utilisables par les trois centres de recherche, a été bâti.

Il s'agit de segmenter les sons en « zones stables » ou « segments », et d'« étiqueter » chacun des segments obtenus. Une *zone (temporelle) « stable »* correspond à une portion du signal telle que certains paramètres du signal ne changent pas (ou ne varient que peu) sur toute cette zone. Ces paramètres peuvent être : la fréquence fondamentale, l'énergie, le contenu fréquentiel... *Segmenter*

veut dire poser des marques dans le domaine temporel aux moments où le signal « varie brusquement », pour passer d'une zone « stable » à une autre zone « stable » ; il s'agit de la détection d'« événements » et leur localisation temporelle. Une fois que le signal sonore est segmenté, il s'agit d'*étiqueter* chacun de ses segments afin de les caractériser, de les décrire. Nous obtenons ainsi des « objets ».

0.2 Cheminement

Dans un premier temps, nous avons considéré des sons que nous avons appelé « simples ». Ces sons « simples » sont *monophoniques*, c'est-à-dire composés d'une seule voix. Ils sont *harmoniques*, c'est-à-dire que chaque zone stable est composée d'une somme de sinusoïdes dont les fréquences sont des multiples de la fréquence fondamentale f_0 : ainsi, chaque zone stable correspond ici à une note ou à un phone. Et ils sont *non modulés*, c'est-à-dire que les sinusoïdes composant une zone stable ne sont ni modulées en fréquence (la modulation de fréquence correspond à un vibrato), ni modulées en amplitude (la modulation d'amplitude correspond à un trémolo [1]). Ainsi, en ce qui concerne la segmentation en zones stables, il s'est agi pour ces sons de les segmenter en *notes* ou en *phones*. Les traitements à appliquer dans le cas de ces sons « simples » ont été développés. Il faut remarquer que l'étiquetage pour ces sons simples consiste principalement à les transcrire, c'est-à-dire à retrouver les notes jouées (ou chantées).

Dans un second temps, nous avons fait la constatation que, pour des sons plus compliqués, c'est-à-dire :

- modulés en fréquence (vibrato), et/ou en amplitude (trémolo)
- et/ou inharmoniques,
- et/ou polyphoniques

les traitements utilisés pour les sons « simples » doivent être adaptés, ou même les traitements utilisés pour les sons plus compliqués doivent être différents de ceux appliqués pour les sons « simples ». Des classes de sons ont été définies : pour chaque classe, il faut redéfinir ce qu'est une « zone stable ».

La stabilité du signal, du point de vue de la musique, pour un son monophonique et harmonique, est définie ainsi : « nous ne changeons pas de note » (malgré la présence de vibrato, nous arrivons à perceptivement discerner les notes et à détecter les transitions entre elles). La stabilité du signal, du point de vue de la parole, est définie ainsi : « nous ne changeons pas de phone ». Il faut noter aussi que dans le cas de la voix chantée, les deux points de vue peuvent se rejoindre ou se confronter.

Donc, pour la musique instrumentale monophonique et harmonique, une zone stable est une *note*, alors que pour la voix chantée [2], une zone stable est une *note*, ou un *phone*.

Ainsi, la définition du terme « zone stable » dépend :

- De la classe à laquelle appartient le son considéré.
- Des désirs de l'utilisateur (par exemple, pour la voix chantée, il peut vouloir segmenter en notes, ou en phones...).

Chaque classe de sons pose des problèmes nouveaux. La détermination de la classe à laquelle appartient le son considéré a fait l'objet de ce que nous avons appelé la *segmentation en caractéristiques*.

Dans un troisième temps, nous avons constaté que la musique et la parole ne peuvent pas être traitées à l'identique, puisque la parole [3] est le plus souvent monophonique (une seule personne parle à la fois), non modulée, et constituée d'une rapide succession de zones stables – une

1. Ces définitions du vibrato et du trémolo sont utilisées dans [Mas96], page 82.

2. Nous considérons dans cet exposé que les signaux de voix chantée sont composés d'une suite de zones stables harmoniques : c'est-à-dire que nous considérons que les consonnes non voisées sont absentes de la voix chantée. À ce sujet, voir [DGR94] : « Due to the great predominance of vowels over consonants, we only processed the vowels. »

3. Ce que nous disons ici à propos de la *segmentation en zones stables* de la parole est purement informel : nous ne traiterons pas la *segmentation en zones stables* de la parole dans cet exposé.

zone stable ici est un phone[4] – harmoniques (les voyelles ou les consonnes voisées) et de zones stables inharmoniques (les consonnes non voisées), alors que la musique peut être monophonique ou polyphonique, modulée ou non, et constituée d'une succession rapide ou non de zones stables composées chacune d'une somme de sinusoïdes (pas forcément harmoniques) auxquelles s'ajoutent ou non de petits bruits (percussions). Donc, avant de *segmenter en zones stables*, il faut détecter si nous sommes en présence de parole ou de musique. Cette détection concerne ce que nous avons appelé la *segmentation en sources*.

Dans un quatrième temps, nous avons discuté, pour les sons polyphoniques, de la nécessité ou non de segmenter avant de séparer les sources présentes. Le succès de la séparation de sources, et donc de la transcription automatique de plusieurs sons monophoniques mixés, dépend de la qualité de la segmentation.

Toutes ces remarques nous ont conduit à complexifier la procédure de segmentation, qui a été divisée en trois niveaux de segmentation, hiérarchiquement organisés. Ces niveaux de segmentation échangent des informations, informations qui circulent des niveaux supérieurs vers le niveaux inférieurs. Vient finalement se greffer à ces modules de segmentation un module de séparation de sources. Nous donnons sur la figure 1 un résumé schématique de ce que nous venons d'indiquer, et nous donnons dans quelle(s) partie(s) de cet exposé chacun des modules sera plus spécifiquement explicité.

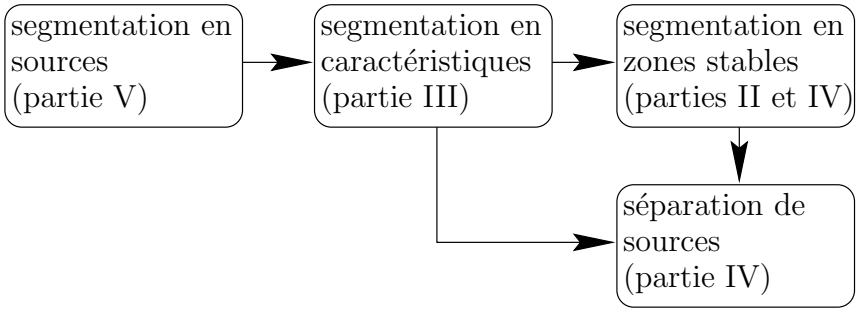


FIG. 1 – *Les modules de segmentation et de séparation : organisation hiérarchique*

L'objectif étant que les procédures pour segmenter et séparer décrites dans l'exposé soient le plus automatiques possible, au fur et à mesure que nous avons rencontré des paramètres libres (tailles de fenêtre d'analyse, etc.), nous les avons discutés et si possible réduits. Nous avons pour ce faire utilisé des bases de données sonores décrites dans l'exposé. Tout du moins, ces paramètres libres peuvent être contrôlés par l'utilisateur des programmes qui ont été développés au cours de cette thèse.

0.3 Plan détaillé de l'exposé

La **deuxième partie** de cet exposé traite de la *segmentation en zones stables* des sons monophoniques, harmoniques et non modulés. Pour de tels sons, une variation brusque du signal peut avoir lieu :

- En fréquence fondamentale. Le plus souvent, deux notes successives n'ont pas la même fréquence fondamentale. Cependant, deux phones successifs peuvent être chantés avec la même fréquence fondamentale. Ainsi, la détection des variations brusques de f_0 ne suffit pas.

4. Nous ne prenons pas en compte le phénomène de co-articulation. Ce phénomène est l'effet contextuel que produit un phonème sur ses voisins. Il est provoqué par le fait que, lors de la prononciation d'un phonème, l'appareil articulatoire se prépare pour la production du suivant.

- En énergie. Parfois, le saut de fréquence entre deux notes successives peut être petit ou même nul, et donc difficile à détecter. Lors de la transition entre deux notes, souvent l'énergie du signal diminue. Il peut parfois être plus facile de détecter cette variation d'énergie que de détecter la variation en fréquence fondamentale. Nous supposons qu'au cours de chaque note l'énergie ne change pas ou peu. Cependant, si deux notes successives sont liées (légato), l'énergie ne nous apporte que peu d'informations.
- En contenu fréquentiel. Par exemple, pour la voix chantée, nous pouvons avoir plusieurs phones consécutifs chantés avec la même fréquence fondamentale et avec la même énergie. Ici, ce sont les formants, c'est-à-dire le contenu fréquentiel, qui changent. Ainsi, pour la voix chantée, suivant ce que nous désirons, nous pouvons soit segmenter en phones, soit segmenter en notes. Si nous voulons segmenter en phones, il faut que nous nous intéressions au contenu fréquentiel.

Le plus souvent, bien sûr, ces trois types de variations brusques sont simultanément présents. Chacun de ces types de variations concerne l'un des trois attributs communément utilisés pour caractériser perceptivement un son [5] : la hauteur (f_0), l'énergie (intensité) et le timbre (ou contenu fréquentiel).

Diverses fonctions d'observation ont été mises en place qui permettent de mettre en évidence ces variations. Ces fonctions d'observation sont, pour la plupart, spécialisées : elles ne sont capables de réagir qu'à un seul type de variations. Alors, dans un second temps, nous devons fusionner les résultats obtenus avec les différentes fonctions d'observation. Cette fusion peut elle-même dépendre de ce que veut l'utilisateur (fonctions d'observation prises ou non en compte) : par exemple, pour la voix chantée, si l'utilisateur veut segmenter en phones, la détection des variations brusques de f_0 n'est pas suffisante.

L'utilisation d'un grand nombre de fonctions d'observation a pour objectif aussi de diminuer le nombre de fausses alarmes (nous faisons l'hypothèse qu'à un instant donné dans une zone stable peu de fonctions d'observation réagissent), et d'augmenter le nombre de bonnes détections (nous faisons l'hypothèse qu'à un instant donné correspondant à une transition peu de fonctions d'observation ne réagissent pas). Nous voulons ainsi améliorer la robustesse de la segmentation. Le but de la fusion de données est de prendre en compte les résultats obtenus avec plusieurs capteurs imparfaits pour aboutir à un meilleur résultat que celui obtenu avec un seul des capteurs.

Finalement, l'analyse se décompose en quatre étapes :

- La première étape est l'extraction de fonctions d'observation (appelées aussi, dans la littérature, « features » : voir [RRS$^+$]..., ou « caractéristiques »). Nous attendons d'une fonction d'observation qu'elle présente un pic aussi fin et grand que possible au moment de chaque transition entre deux zones stables, et que ses valeurs aient une moyenne et une variance petites pendant les zones stables. Nous cherchons aussi à ce que ces fonctions d'observation soient le plus possible indépendantes les unes des autres. Elles ne sont pas forcément reliées à des caractères perceptifs.
- La deuxième étape est la prise de décision pour chaque fonction d'observation. Il s'agit de détecter les pics correspondant à des transitions. Pour le moment, il s'agit de calculer un seuil et de l'appliquer. Ce seuil est périodiquement ou automatiquement adapté aux données. Ou la période d'adaptation du seuil est fixée à dix secondes ; ou, à partir des informations fournies par la *segmentation en caractéristiques*, nous déterminons automatiquement quand il faut le recalculer. Nous obtenons des fonctions de décision, c'est-à-dire des fonctions à valeurs dans $\{0, 1\}$.
- La troisième étape concerne la prise de décision finale. Elle est basée sur le traitement des prises de décision particulières obtenues pour chacune des fonctions d'observation, c'est-à-dire sur le traitement des fonctions de décision. Il s'agit donc ici d'une sorte de « fusion de données ».

5. Voir [ZF81], page 4 : « Quand nous entendons un son composé nous percevons en plus d'une sensation d'intensité et de hauteur, une autre sensation qui différencie ce son composé d'un son pur. Nous appellerons ceci le timbre du son composé. »

- Enfin, la dernière étape concerne l'étiquetage des segments obtenus. Pour les sons simples, cet étiquetage est équivalent à la recherche des notes jouées, c'est-à-dire à la transcription du son.

Cependant, la présence d'un vibrato gêne la détection des variations brusques de f_0. La présence d'un glissando en fréquence pose problème aussi. De la même façon, la présence d'un trémolo gêne la détection des variations brusques de l'énergie, et la présence d'un glissando en énergie pose problème.

De plus, le signal peut ne pas être harmonique : il peut être percussif, ou être composé de consonnes non voisées : c'est-à-dire que nous sommes en présence de bruit. Les variations brusques ont alors lieu en terme de variation des moments statistiques ($\hat{m} = E[X]$, $\hat{\sigma}^2 = E\left[(X-\hat{m})^2\right]$, $\hat{\mu}_3 = E\left[(X-\hat{m})^3\right]$...) du signal. Et nous constatons aussi que dans ce cas la détection des variations brusques de f_0 n'a plus aucun sens.

Ainsi, les traitements à appliquer pour *segmenter en zones stables* sont différents suivant que le signal est modulé ou non, suivant qu'il est harmonique ou non, etc. Ce sont des caractéristiques du signal qu'il faut déterminer avant d'opérer la *segmentation en zones stables*. Cette détermination fait l'objet de la **troisième partie** de l'exposé. Il s'agit d'une segmentation plus grossière, d'un niveau plus élevé que le précédent. Nous l'avons appelée *segmentation en caractéristiques*. La *segmentation en zones stables* de tous les signaux monophoniques (qu'ils soient harmoniques ou non, qu'ils soient modulés ou non) est dès lors possible.

Ce niveau de *segmentation en caractéristiques* a de plus pour but de nous indiquer quand les seuils, lors de la deuxième étape de l'analyse *segmentation en zones stables*, doivent être adaptés.

Cette partie traite aussi de la détection du vibrato, problème relevant du niveau de *segmentation en caractéristiques*. La détection, l'estimation des paramètres et la suppression sur le trajet de f_0 du vibrato sont des problèmes auxquels nous nous sommes particulièrement attachés. Pour cette raison, elles font l'objet de la plus grande part de cette partie. Quand un vibrato est détecté, il est intéressant (par exemple pour des modifications du son) et nécessaire (pour la *segmentation en zones stables*) de le supprimer du trajet de f_0. La plupart des considérations (détection, estimation, suppression) que nous ferons à propos du vibrato (modulation de la fréquence) seront aussi valables pour le trémolo (modulation de l'énergie, ou de l'amplitude).

La **quatrième partie** de cet exposé traite de la *segmentation en zones stables* des sons polyphoniques, c'est-à-dire composés d'une somme de voix, chacune étant elle-même harmonique ou inharmonique, modulée ou non modulée.

La plus grande partie des fonctions d'observation utilisées pour la *segmentation en zones stables* des sons monophoniques peuvent aussi bien s'appliquer pour des sons polyphoniques. Cependant, sont inutilisables toutes celles basées sur f_0.

L'étiquette monophonique/polyphonique est une des caractéristiques dont s'occupe le niveau de *segmentation en caractéristiques*. Ce niveau de segmentation a principalement pour but d'informer le niveau de *segmentation en zones stables* des fonctions d'observation qu'il peut utiliser : il est donc nécessaire de détecter la polyphonie avant de *segmenter en zones stables*.

La *séparation de sources* consiste à reconstruire les diverses voix d'un son. Elle est différente de la « blind separation » des antennistes, pour laquelle nous avons moins de sources que d'antennes (ou, à la limite, au moins autant d'antennes que de sources). Dans notre cas, nous n'avons qu'une seule « antenne » : nous considérons des sons enregistrés en monophonie. La *séparation de sources* fait aussi l'objet de la **quatrième partie** de l'exposé.

Nous avons vu que la définition de « zone stable » diffère suivant la nature du signal considéré :

- Pour les sons instrumentaux monophoniques et harmoniques, il s'agit d'une note.
- Pour la voix chantée, il s'agit d'une note ou d'un phone.
- Pour la voix parlée, il s'agit d'un phone. Un phone est harmonique quand il s'agit d'une voyelle ou d'une consonne voisée, ou inharmonique quand il s'agit d'une consonne non voisée.

Notons encore une fois que dans cet exposé nous ne discuterons pas de la segmentation en zones stables des signaux de parole.

- Pour les sons instrumentaux inharmoniques (par exemple pour un extrait de castagnettes), nous voulons détecter plutôt les instants d'attaque.

Il faut donc d'abord, avant de *segmenter en zones stables* et de *segmenter en caractéristiques*, déterminer la nature de signal considéré. Ceci fait l'objet d'un autre niveau de segmentation, plus grossier encore que le niveau de *segmentation en caractéristiques*. Nous l'appelons *segmentation en sources*. Ce niveau de segmentation fait l'objet de la **cinquième partie** de l'exposé.

Cette partie traite de la segmentation, par exemple des bandes son de film ou des enregistrements radiophoniques, suivant la nature du son analysé. Deux types de sons sont considérés pour le moment : voix parlée, et voix chantée et/ou musique instrumentale. Des sons d'autres natures devront être considérés : bruits de machines, bruits de rue...

La longueur des segments fournis par le niveau de *segmentation en sources* peut être de quelques minutes. La longueur des segments fournis par le niveau de *segmentation en caractéristiques* est communément plus petite : elle est disons de l'ordre de quelques dizaines de seconde. La longueur des segments fournis par le niveau de *segmentation en zones stables* est le plus souvent inférieure à une seconde. Chaque niveau de segmentation est concerné aussi par l'étiquetage des segments qu'il fournit.

Dans la **cinquième partie** de cet exposé, nous donnerons aussi les dépendances entre les trois niveaux de segmentation. Elles traitent des informations échangées par les trois niveaux de segmentation définis (*segmentation en parties stables*, *segmentation en caractéristiques* et *segmentation en sources*). Cette partie traite enfin des relations qui existent entre la *segmentation* et la *séparation de sources*.

Nous donnons une conclusion générale à cet exposé, ainsi que des perspectives, dans la **sixième partie**.

La **septième partie** rassemble les annexes.

Deuxième partie

Segmentation en notes et/ou en phones et indexation des sons monophoniques, harmoniques et non modulés

Chapitre 1

Présentation du problème

Nous considérons dans cette partie des sons monophoniques (c'est-à-dire où n'intervient qu'une voix : un orateur [1] seul ou un chanteur seul ou un instrument seul), harmoniques (c'est-à-dire que le signal pour chaque zone stable se décompose en une somme de sinusoïdes disposées sur un peigne harmonique), et non modulés (c'est-à-dire que les sinusoïdes présentes pour chaque zone stable ne sont modulées ni en fréquence ni en amplitude). Le but est de segmenter ces sons en zones stables, et d'étiqueter chacune d'elles. Une zone stable, dans cette partie, correspond à une note ou à un phone harmonique. La segmentation consiste à détecter les variations brusques du signal, à détecter les transitions entre deux zones stables successives. L'étiquetage consiste à retranscrire la partition jouée.

Un programme pour *segmenter en zones stables* a été développé en C au cours de cette thèse. Nous l'avons appelé *segmentation*.

L'analyse est composée de quatre étapes.

Nous avons tout d'abord défini un ensemble de **fonctions d'observation**. Elles doivent résumer et faire ressortir les caractéristiques des sons : des zones stables séparées par des transitions brusques. La mise en évidence des transitions a été faite d'un point de vue purement traitement du signal ou d'un point de vue plus perceptif. Les fonctions d'observation sont présentées dans le **deuxième chapitre** (chapitre 2) de cette partie. L'extraction de ces fonctions d'observation constitue la *première étape* de l'analyse *segmentation en zones stables*.

Ensuite nous avons défini un ensemble de critères permettant de déterminer, à partir des fonctions d'observation, à quels moments nous avons une variation importante du son, c'est-à-dire les moments où nous devons poser une marque de segmentation : il s'agit de la prise de décision. À partir de chaque fonction d'observation, nous obtenons une **fonction de décision**, c'est-à-dire une liste de marques de segmentation. Nous décrivons les procédures de prise de décision utilisées dans le **troisième chapitre** (chapitre 3) de cette partie. Ces prises de décision constituent la *deuxième étape* de l'analyse *segmentation en zones stables*.

Puis nous nous sommes intéressé dans le **quatrième chapitre** (chapitre 4) de cette partie à la fusion des résultats des prises de décisions obtenus pour chacune des fonctions d'observation. Nous traitons aussi dans ce chapitre d'un problème qui se rencontre aussi bien au cours de la *deuxième étape de l'analyse* (chapitre 3) qu'au cours de celle-ci : il s'agit du traitement des marques de segmentation trop proches les unes des autres. Ce problème est dû au fait que les transitions ne sont pas instantanées et que certaines fonctions d'observation réagissent plutôt au début des transitions, d'autres plutôt à la fin des transitions, et d'autres plusieurs fois au cours d'une même transition. Nous obtenons une **fonction de décision finale**. Cette prise de décision finale constitue la *troisième étape* de l'analyse *segmentation en zones stables*.

1. Voir la note de la page 3.

L'étiquetage des segments est discuté succintement dans le **cinquième chapitre** (chapitre 5) de cette partie. Cet étiquetage constitue la *quatrième étape* de l'analyse *segmentation en zones stables*.

Dans le **sixième chapitre** (chapitre 6) de cette partie, nous avons dans un premier temps observé le comportement des fonctions d'observation avec un son de flûte simple, c'est-à-dire parfaitement monophonique (pas de reverbération, etc.), harmonique, et quasi non modulé. Puis nous discutons les résultats obtenus avec le système complet sur un ensemble de signaux simulés ou réels.

Dans le **septième chapitre** (chapitre 7) de cette partie, les corrélations entre les fonctions d'observation sont étudiées, d'abord en vue d'éliminer celles qui n'apportent pas d'information supplémentaire et donc de réduire le nombre de fonctions d'observation à utiliser. Nous indiquons dans la section 2.6 (chapitre 2) que la présence de deux capteurs absolument corrélés (fonctions de décision absolument identiques) implique que nous n'atteignons pas les performances optimales lors de la fusion de données. Les corrélations sont aussi étudiées en vue de caractériser perceptivement chaque fonction d'observation (elle réagit plutôt aux transitions en f_0, aux transitions en énergie, aux transitions en contenu spectral, ou à une combinaison de ces trois types de transitions).

Dans le **huitième chapitre** (chapitre 8) de cette partie, nous faisons quelques remarques qui montrent que le système de segmentation tel qu'il est décrit dans cette partie n'est plus suffisant dès que le signal n'est plus monophonique, harmonique et/ou non modulé, et qu'il est nécessaire de mettre en place un niveau de segmentation supérieur à celui présenté dans cette partie. Nous appelons cet autre niveau (décrit dans la partie III) *segmentation en caractéristiques*. Cette segmentation, plus grossière en ce sens qu'elle nous donne des segments plus longs que la *segmentation en zones stables*, aide à la *segmentation en zones stables*.

Nous donnons une conclusion à cette partie dans le **neuvième chapitre** (chapitre 9) de cette partie.

Chapitre 2

Les fonctions d'observation

2.1 Introduction

L'extraction de fonctions d'observation est la *première étape* de l'analyse *segmentation en zones stables*.

Nous avons vu dans l'introduction générale à cet exposé que, pour les sons monophoniques, harmoniques et non modulés, une variation brusque du signal peut avoir lieu :

- En fréquence fondamentale. Les fonctions d'observation qui permettent de mettre en évidence ce type de variations brusques sont décrites dans la **deuxième section** (section 2.2) de ce chapitre.
- En énergie. Les fonctions d'observation qui permettent de mettre en évidence ce type de variations brusques sont présentées dans la **troisième section** (section 2.3) de ce chapitre.
- En contenu fréquentiel. Les fonctions d'observation qui permettent de mettre en évidence ce type de variations brusques sont présentées dans la **quatrième section** (section 2.4) de ce chapitre.

D'autres fonctions d'observation, n'entrant pas dans ces trois catégories, sont décrites dans la **cinquième section** (section 2.5) de ce chapitre. Elles tentent de détecter des variations brusques dans le signal lui-même, ou de mettre en évidence plusieurs types de variations brusques simultanément.

Les fonctions d'observation peuvent encore être séparées en deux groupes : celles qui sont unidimensionnelles et celles qui sont multidimensionnelles. Une fonction d'observation multidimensionnelle est une fonction d'observation qui nous donne par exemple un trajet pour chaque harmonique, ou pour chaque bande de fréquence considérée. Nous voulons réduire chaque fonction d'observation multidimensionnelle à une fonction d'observation unidimensionnelle. C'est-à-dire que nous voulons mixer les résultats obtenus pour chaque dimension : apparaissent les premiers problèmes de fusion de données. Leur discussion et leur résolution font l'objet de la **sixième section** (section 2.6) de ce chapitre.

Dans la **septième section** (section 2.7) de ce chapitre, un récapitulatif des fonctions d'observation étudiées est fait.

Nous concluons ce chapitre avec sa **huitième section** (section 2.8).

2.2 Fonctions d'observation basées sur les variations de f_0

2.2.1 Introduction

2.2.1.1 Détermination du trajet de f_0

Tout signal périodique peut être décomposé en une somme de sinusoïdes dont les fréquences sont toutes multiples d'une fréquence particulière, appelée fréquence fondamentale. Ces sinusoïdes sont appelées les harmoniques, ou les partiels harmoniques du signal. Et le signal est lui-même dit

harmonique. Les sons considérés dans cette partie sont harmoniques localement, c'est-à-dire sur chaque note, ou encore sur chaque zone stable. Au cours de chaque note, nous n'observons que quelques périodes du signal. Pour un la_3 ($f_0 = 440\ Hz$) de durée un quart de seconde, nous avons 110 périodes.

La fréquence fondamentale f_0 d'un signal sonore est obtenue grâce au logiciel f_0 développé à l'IRCAM (voir entre autres [Dov94] et [DGR93]).

2.2.1.2 Fonctionnement du logiciel f_0

Nous présentons succintement le principe simplifié de la méthode.

Il est fait l'hypothèse que le signal à analyser est harmonique. Il est découpé en portions (ou fenêtres d'analyse, ou encore trames) larges de quelques dizaines de millisecondes, et décalées de quelques millisecondes. Un spectre d'amplitude est calculé pour chacune de ces portions, d'instant central i, multipliée par une fenêtre de pondération. Tous les pics significatifs du spectre d'amplitude sont détectés : ils ont pour fréquences respectives $[f_1\ \dots\ f_N]$. Il s'agit ensuite de déterminer quelle fréquence fondamentale explique au mieux l'ensemble de ces pics. Un ensemble, fixe quelque soit i, de M fréquences fondamentales candidates est utilisé : $[f_0^{(1)}\ \dots\ f_0^{(M)}]$. Chaque pic gardé, de fréquence f_k, pour une candidate $f_0^{(j)}$, correspond à l'harmonique de numéro d'ordre l_k, avec : $l_k = \left(\frac{f_k}{f_0^{(j)}}\right)_{ent}$, où $(A)_{ent}$ est l'opérateur *plus proche entier de A*. M ensembles $[l_1\ \dots\ l_N]^{(j)}$, un par fréquence fondamentale candidate, sont obtenus : il s'agit de déterminer lequel est le plus vraisemblable, en prenant en compte la probabilité d'absence de chaque harmonique, la probabilité de dispersion de chaque harmonique l autour de sa fréquence théorique lf_0, etc.

Nous obtenons donc le trajet de f_0 dans le temps : $f_0(i)$. Les sauts de f_0 correspondent à un changement de note (de hauteur) et ce sont ces sauts que nous voulons détecter, dans cette section 2.2.

2.2.2 Dérivées de f_0

2.2.2.1 Les valeurs absolues des dérivées de f_0

La fonction d'observation « valeur absolue de la dérivée de f_0 », $|df_0(i)|$, se calcule, à l'instant i, à partir des $f_0(i)$ que nous avons définis dans la section 2.2.1.2, ainsi :

$$\left|df_0^I(i)\right| = |f_0(i+1) - f_0(i-1)| \text{ ou } \left|df_0^{II}(i)\right| = |f_0(i) - f_0(i-1)|$$

Nous définissons la fonction d'observation « valeur absolue de la dérivée relative de f_0 » ainsi :

$$\left|\delta f_0^I(i)\right| = \frac{\left|df_0^I(i)\right|}{f_0(i)} = \left|\frac{f_0(i+1) - f_0(i-1)}{f_0(i)}\right| \text{ ou } \left|\delta f_0^{II}(i)\right| = \frac{\left|df_0^{II}(i)\right|}{f_0(i)} = \left|\frac{f_0(i) - f_0(i-1)}{f_0(i)}\right|$$

Les fonctions d'observation $\left|df_0^I\right|$ et $\left|df_0^{II}\right|$ et les fonctions d'observation $\left|\delta f_0^I\right|$ et $\left|\delta f_0^I\right|$ sont implémentées dans le programme *segmentation*. Beaucoup d'autres dérivées « numériques » pourront être utilisées : voir entre autres celles données dans l'article [Boa92] [1]. Il s'agit d'une perspective.

2.2.2.2 L'apport des harmoniques de numéros d'ordre supérieurs

Nous calculons de la même façon les dérivées des harmoniques de numéros d'ordre supérieurs. Nous obtenons ainsi une fonction d'observation multidimensionnelle. Il faut alors fusionner les dimensions (voir la section 2.6). Ceci n'est pas fait dans le programme *segmentation*.

1. La dérivée $f(i+1)-f(i-1)$ y est appelée « CFD », pour « Central Finite Difference », et la dérivée $f(i)-f(i-1)$ y est appelée « BFD », pour « Backward Finite Difference ».

2.2.3 Indices de voisement – Première forme

2.2.3.1 Principe de la méthode

Nous calculons un indice de voisement pour chaque harmonique. Nous utilisons le trajet de f_0 déterminé par le logiciel f$_0$.

La procédure est la suivante :

- Nous connaissons la fréquence f_0 à l'instant i.
- Nous considérons une portion du signal centrée à cet instant i. La largeur T de cette portion est égale à quelques dizaines de millisecondes. Classiquement, elle doit couvrir trois ou quatre périodes du signal[2] : ainsi, pour un la$_3$ ($f_0 = 440\ Hz$), nous obtenons une dizaine de millisecondes. Nous multiplions cette portion du signal par une fenêtre de pondération w, de transformée de FOURIER $\hat{W}$ et de spectre d'amplitude $\left|\hat{W}\right|$. Nous calculons le spectre d'amplitude $\left|\hat{S}\right|$ de cette portion pondérée du signal.
- $\left|\hat{S}\right|$ est tronqué autour des fréquences harmoniques de f_0. $\left|\hat{W}\right|$ est tronqué autour de la fréquence nulle. Nous calculons la corrélation entre chaque tronçon normalisé en énergie de $\left|\hat{S}\right|$, et $\left|\hat{W}\right|$ tronqué et normalisé en énergie.

C'est-à-dire que nous calculons pour tous les l entre 1 et L, ou l correspond au numéro d'ordre de l'harmonique, l'indice de voisement :

$$V_l(i) = \sum_{f=(lf_0)_{ent}-N}^{(lf_0)_{ent}+N} \left|\hat{S}_{norm}(f)\right| \left|\hat{W}_{norm}(f-(lf_0)_{ent})\right|$$

où : L est tel que $Lf_0 \leq \frac{f_e}{2} \leq (L+1)f_0$ (notons que L varie dans le temps, c'est-à-dire avec i, suivant $f_0(i)$) ; f_e est la fréquence d'échantillonnage du signal sonore ; $(A)_{ent}$ est l'opérateur *plus proche entier de A* ; N est la largeur de la troncature sur laquelle nous calculons la corrélation (la taille des spectres d'amplitude tronqués est donc $2N+1$) ; et $(A)_{norm}$ est l'opérateur *normalisation de l'énergie de A à 1.* La normalisation en énergie d'un spectre d'amplitude $\left|\hat{T}\right|$ tronqué entre a et b se fait ainsi :

$$\left|\hat{T}_{norm}(f)\right| = \frac{\left|\hat{T}(f)\right|}{\sqrt{\sum_{f=a}^{b}\left|\hat{T}(f)\right|^2}}$$

pour $f \in [a\ \ldots\ b]$. Dans notre cas $a = (lf_0)_{ent} - N$ et $b = (lf_0)_{ent} + N$.

Nous donnons ici quelques notations, qui seront valables dans tout l'exposé. T est la largeur temporelle des fenêtres d'analyse (en seconde), t_{SIG} la même largeur en nombre d'échantillons. t_{FFT} est la taille des FFT. Les échantillons fréquentiels des spectres sont numérotés de $-\frac{t_{FFT}}{2}+1$ à $\frac{t_{FFT}}{2}$, et les fréquences correspondantes sont $-\frac{f_e}{2}+\Delta f$ et $\frac{f_e}{2}$. À l'échantillon fréquentiel 0, correspond une fréquence de 0. Δf vaut $\frac{f_e}{t_{FFT}}$.

Chacun des indices de voisement $V_l(i)$ vaut 1 si une sinusoïde de fréquence lf_0 est présente et moins de 1 sinon. Nous obtenons ainsi une fonction d'observation multidimensionnelle. Il faut alors fusionner les données (voir la section 2.6).

2. Communément, il est fait l'hypothèse que les signaux sonores sont stationnaires sur des périodes de 30 ou 40 millisecondes. Cette taille de fenêtre d'analyse permet d'assurer que les lobes principaux dus à deux harmoniques de numéros d'ordre successifs soient résolus, c'est-à-dire que leur somme ne forme pas qu'un seul lobe, repéré par un seul maximum local dans le spectre d'amplitude.

2.2.3.2 Relâcher la contrainte « Nous connaissons parfaitement $f_0(i)$ »

Si à l'instant i une petite erreur est faite sur f_0, quand nous calculons les indices de voisement comme nous l'indiquons ci-dessus dans la section 2.2.3.1, les effets de cette erreur se répercutent pour tous ces indices de voisement V_l. Il vaut donc mieux relâcher la contrainte $(lf_0)_{ent}$. Ainsi, nous calculons les indices de voisement obtenus pour A_l échantillons fréquentiels autour de cette valeur $(lf_0)_{ent}$, et, ensuite, n'est pris en compte que le plus grand. L'indice de voisement de numéro d'ordre l devient :

$$V_l^A(i) = \max_{j=-A_l \to A_l} \sum_{f=(lf_0)_{ent}-N+j}^{(lf_0)_{ent}+N+j} \left|\hat{S}_{norm}(f)\right| \left|\hat{W}_{norm}(f-(lf_0)_{ent}-j)\right|$$

Une petite erreur ϵ sur $f_0(i)$ nous donne une erreur $l\epsilon$ sur l'harmonique l. Donc A_l doit être une fonction linéaire du numéro d'ordre l de l'harmonique. Nous prenons : $A_l = lA_0$ [3].

2.2.3.3 Les paramètres libres

Les paramètres libres sont au nombre de 6. Ce sont : T, la largeur de chaque fenêtre d'analyse ; N, la demi-largeur de la troncature ; L_{pris}, le nombre d'harmoniques pris en compte ($L_{pris} \leq L$) ; A_0 ; t_{FFT}, la taille de la FFT ; la fenêtre de pondération. Nous avons vu que T est de l'ordre de quelques dizaines de millisecondes. N est choisie de telle façon que X % de l'énergie totale de $\left|\hat{W}\right|$ soit gardée. X est choisi de l'ordre de 90. L_{pris} n'est pas choisi très grand. Souvent l'amplitude des harmoniques diminue rapidement avec leur numéro d'ordre. Ainsi le lobe principal des harmoniques de numéro d'ordre élevé est trop déformé par le bruit pour que nous obtenions un indice de voisement proche de 1. Le plus souvent, $L_{pris} = 3$. A_0 est égal à $\frac{yf_0(i)}{\Delta f}$, où y est petit ($y = 0{,}01$ par exemple). Plus t_{FFT} est grand, plus Δf est petit, et donc plus les lf_0 sont proches d'un $\alpha\Delta f$, où α est entier. Ainsi, si nous nous plaçons dans le cas d'un signal harmonique, plus t_{FFT} est grand, plus les indices de voisement sont proches de 1. Nous n'avons pas de contrainte de temps de calcul : alors nous pouvons utiliser un taux de bourrage de zéros [4] grand. Le dernier paramètre libre est la fenêtre de pondération : souvent, nous utilisons celle de BLACKMAN.

L'indice de voisement première forme implémenté dans le programme *segmentation* est l'indice V_l^A. La fonction d'observation finale est la « fusion des valeurs absolues des dérivées des indices de voisement première forme ».

2.2.4 Indices d'inharmonicité

2.2.4.1 Fonctionnement d'ADDITIVE

Nous avons vu que le logiciel f_0 nous donne la fréquence fondamentale aux instants i. Dans le logiciel de l'IRCAM ADDITIVE, il est fait l'hypothèse que le signal sonore est une somme de sinusoïdes quasi harmoniques. Une fois que f_0 a été estimée à un instant donné i, ADDITIVE détermine une sinusoïde (sa fréquence, son amplitude et sa phase) pour chaque bande de fréquences centrée autour de lf_0, avec l variant de 2 à L. Il s'agit de détecter le maximum du spectre d'amplitude dans chaque bande de fréquence. La largeur des bandes de fréquence est au plus de f_0. Les positions de ces maximums nous donnent les fréquences f_l des harmoniques.

3. Il est un problème de notation. Dans tout cet exposé, nous considérons que le premier partiel harmonique ou premier harmonique correspond à la fréquence fondamentale, numérotée indifféremment 0... ou 1. Ainsi, en règle générale dans cet exposé : $f_0 = f_1$... et donc ici $A_0 = A_1$.

4. Ou « zero-padding ». Le bourrage de zéros consiste à ajouter t_{ZER} zéros à la fin du signal à analyser, de taille t_{SIG}. Le bourrage de zéros est utilisé pour deux raisons :

- afin que $t_{SIG} + t_{ZER}$ soit une puissance de 2 (il s'agit de profiter des algorithmes de FFT)
- afin d'obtenir un spectre d'amplitude lissé, interpolé (fort taux de bourrage de zéros)

2.2.4.2 Principe de la méthode

À chaque instant i et pour chaque harmonique de numéro d'ordre l (l variant de 2 à L) nous calculons l'indice d'inharmonicité $H_l(i)$ de la façon suivante :

$$H_l(i) = \left| \frac{f_l(i) - l f_0(i)}{l f_0(i)} \right|$$

où $f_l(i)$ est la fréquence du $l^{ème}$ harmonique obtenu avec ADDITIVE.

L'idée ici est qu'au cours des transitions, f_0 n'a plus de sens physique. Alors, l'inharmonicité devient importante pour tous les harmoniques.

Nous obtenons un critère de dimension $(L-1)$. Nous avons donc ici de nouveau un problème de fusion de données (voir la section 2.6).

2.2.4.3 Relâcher la contrainte « Nous connaissons parfaitement $f_0(i)$ »

Dans la méthode présentée dans la section 2.2.4.2 ci-dessus, nous accordons une grande confiance à $f_0(i)$. Les effets d'une petite erreur sur $f_0(i)$ se répercutent pour tous les indices d'inharmonicité. Il vaut donc mieux relâcher la contrainte $l f_0$. Nous calculons un nouvel $f_0(i)$, appelé $f_0^n(i)$, ainsi :

$$f_0^n(i) = \frac{1}{L_{pris}} \sum_{j=1}^{L_{pris}} \frac{f_j(i)}{j}$$

Bien sûr, dans les zones stables, $f_0^n(i)$ est très proche de $f_0(i)$.

Nous calculons alors l'indice d'inharmonicité $H_l^n(i)$ de la façon suivante :

$$H_l^n(i) = \left| \frac{f_l(i) - l f_0^n(i)}{l f_0^n(i)} \right|$$

pour chaque harmonique de numéro d'ordre l, avec l variant désormais de 1 à L_{pris}.

2.2.4.4 Paramètres libres

Le seul paramètre libre est L_{pris} le nombre d'harmoniques pris en compte ($L_{pris} \leq L$). Nous l'avons choisi petit, le plus souvent égal à trois. Ceci parce que l'amplitude de l'harmonique décroissant rapidement (en $\frac{1}{f}$ par exemple) avec son numéro d'ordre, le bruit détériore l'estimation de ses trois paramètres, c'est-à-dire notamment de sa fréquence.

L'indice d'inharmonicité implémenté dans le programme *segmentation* est l'indice H_l^n. La fonction d'observation finale est la « fusion des valeurs absolues des dérivées des indices d'inharmonicité ».

2.2.5 Analyse statistique sur le trajet de f_0

Nous avons adapté ici un programme implémenté en C par Laurent CERVEAU lors d'un stage de DEA effectué à l'IRCAM (voir [Cer94]).

2.2.5.1 Principe de la méthode

À l'instant i, nous observons le signal, c'est-à-dire ici le trajet de f_0, sur deux portions adjacentes de taille N échantillons (voir la figure 2.1) :

- O_1 : $O_1 = [x(i-N+1) \quad \ldots \quad x(i)]$.
- O_2 : $O_2 = [x(i) \quad \ldots \quad x(i+N-1)]$.

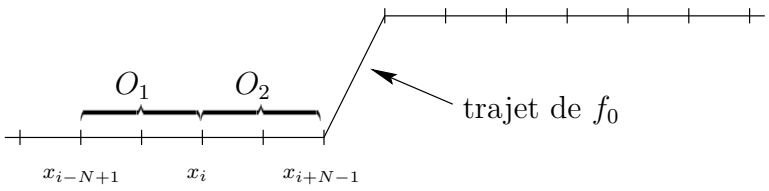


FIG. 2.1 – *Disposition des deux portions pour l'analyse statistique sur le trajet de* f_0

Nous faisons l'hypothèse que sur chaque portion nous obtenons N observations d'une variable aléatoire normale. Les densités de probabilité pour les deux variables aléatoires X_1 et X_2 sont normales et respectivement égales à $\mathcal{N}\left(m_1,\sigma_1^2\right)$ et $\mathcal{N}\left(m_2,\sigma_2^2\right)$. Il s'agit de détecter les moments où X_1 correspond à une note et X_2 à une autre. L'écart maximal entre les deux suites aléatoires est obtenu quand la portion 1 couvre la fin d'une note et la portion 2 le début de la note suivante.

Nous calculons alors la probabilité que les modèles correspondants aux deux suites aléatoires soient au plus éloignés d'une distance S, c'est-à-dire que nous calculons la probabilité à tout instant i :

$$p(i) = \mathrm{P_r}\left(|X_2(i) - X_1(i)| \leq S\right) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \int_{-S}^{+S} \exp\left[-\frac{1}{2}\left(\frac{y-(m_2-m_1)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)^2\right] dy$$

Nous donnons en annexe A les calculs qui nous permettent d'aboutir à cette intégrale. Nous n'obtenons pas une formule analytique : il reste cette intégrale, que nous calculons aisément numériquement. La taille des deux portions étant N, les moyennes sont estimées ainsi :

$$\hat{m}_1 = \frac{1}{N}\sum_{j=i-N+1}^{i} x(j) \text{ et } \hat{m}_2 = \frac{1}{N}\sum_{j=i}^{i+N-1} x(j)$$

et les variances ainsi (voir l'annexe F) :

$$\hat{\sigma}_1^2 = \frac{1}{N-1}\sum_{j=i-N+1}^{i} \left(x(j)-\hat{m}_1\right)^2 \text{ et } \hat{\sigma}_2^2 = \frac{1}{N-1}\sum_{j=i}^{i+N-1} \left(x(j)-\hat{m}_2\right)^2,$$

qui sont des estimateurs non biaisés des variances ; ou ainsi :

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{j=i-N+1}^{i} \left(x(j)-\hat{m}_1\right)^2 \text{ et } \hat{\sigma}_2^2 = \frac{1}{N}\sum_{j=i}^{i+N-1} \left(x(j)-\hat{m}_2\right)^2,$$

qui sont des estimateurs biaisés des variances.

N étant de l'ordre de quelques dizaines, c'est-à-dire étant petit, il est préférable d'utiliser l'estimateur non biaisé de la variance, quoique en variance il soit moins performant que l'estimateur biaisé.

2.2.5.2 Un nouvel ensemble de valeurs

Au lieu de considérer l'ensemble des valeurs $p(i)$, nous considérons l'ensemble des valeurs $g(i)$ telles que : $g(i) = p(i-1)p(i)p(i+1)$. Le passage d'une note à une autre se déroulant sur un certain laps de temps (la transition n'étant pas instantanée), nous espérons en utilisant $g(i)$ plutôt que $p(i)$ augmenter la robustesse de l'algorithme.

2.2.5.3 L'apport des harmoniques de numéros d'ordre supérieurs

Nous utilisons cet algorithme de la même façon sur les trajets des harmoniques de numéros d'ordre supérieurs. Il faut alors fusionner les résultats (voir la section 2.6). Ceci n'est pas fait dans le programme *segmentation*.

2.2.5.4 Les paramètres libres

Ils sont au nombre de 4. Ce sont : N, la taille des portions ; S, le seuil ; n, le numéro de l'ensemble considéré[5] ; L_{pris}, le nombre d'harmoniques pris en compte. N ne doit ni être trop grande : il ne faut pas qu'une portion couvre plus d'une note ou deux ; ni trop petite : il ne faut pas que les variances des estimées $\hat{m}_1$, $\hat{m}_2$, $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ soient trop grandes. Il faut faire ici une hypothèse sur la longueur des notes : supposons qu'elle soit de l'ordre du quart de seconde. Souvent, le trajet de f_0 est échantillonné à 100 Hz, donc la valeur de N choisie est d'environ 25. Dans cette méthode-ci il est un seuil : S. Cependant, puisque nous analysons le trajet de la fréquence fondamentale, S peut facilement être relié à une donnée musicale, en l'occurence à un écart de ton.

Finalement, la fonction d'observation considérée est ou bien $AS_{f_0}^{I}(i) = 1 - p(i)$, ou bien $AS_{f_0}^{II}(i) = 1 - g(i)$. Les deux sont disponibles dans le programme *segmentation*. Par défaut, $AS_{f_0}^{I}$ est utilisée.

2.2.5.5 Utilisation d'une fonction de score

Une fois les marques de segmentation posées[6] aux instants s_i, leur validité est testée en définissant les deux fenêtres d'analyse ainsi :

- $O_1 = [x(s_{i-1}) \quad \dots x(s_i) \quad]$
- $O_2 = [x(s_i) \quad \dots x(s_{i+1}) \quad]$

Et en calculant le score $sc(i)$ de la marque $s(i)$ de la façon suivante :

$$sc(i) = 1 - P_r\left(|X_2(s(i)) - X_1(s(i))| \leq S\right)$$

Si le score $sc(i)$ est trop faible, il convient de réexaminer et la position de la marque i et le nombre de marques présentes entre s_{i-1} et s_{i+1}. Ceci n'est pas implémenté dans le programme *segmentation*.

2.2.6 La rupture de modèles sur le trajet de f_0

Nous avons adapté ici un programme implémenté en C par Claude BARRAS au LAFORIA et par Jérôme DANIEL à l'IRCAM (voir [Dan95]). L'algorithme est dû à Régine ANDRÉ-OBRECHT (voir [LO95], [BB82], [Jeh97] et [BN93] page 318 : « Divergence algorithm »).

Le signal échantillonné $x(l)$ (ici le trajet de f_0), que nous devons segmenter, est considéré comme étant un processus auto-régressif.

2.2.6.1 Les processus auto-régressifs (ou AR)

Le signal $x(1) \dots x(N)$ représente la réalisation d'un processus auto-régressif d'ordre P s'il satisfait l'équation :

$$x(n) = -a_{P,1}x(n-1) - a_{P,2}x(n-2) - \dots a_{P,P}x(n-P) + e(n)$$

où les coefficients $a_{P,k}$ sont les paramètres AR et $e(n)$ un bruit normal $\mathcal{N}\left(0,\sigma^2\right)$.

La densité spectrale de puissance[7] est alors donnée par la relation :

$$S(f) = \frac{\sigma^2}{f_e \left|1 + \sum_{k=1}^{P} a_{P,k} \exp\left(-2j\pi k \frac{f}{f_e}\right)\right|^2}$$

5. Si nous considérons l'ensemble des $p(i)$, $n = 1$; si nous considérons l'ensemble des $g(i)$, $n = 0$: notons que ce sont les deux seules possibilités offertes par le programme *segmentation*. D'autres pourraient être envisagées : il s'agit d'une perspective.

6. C'est-à-dire après avoir seuillé la fonction d'observation AS_{f_0} : voir le chapitre 3. Initialement, dans le programme de Laurent CERVEAU, ce seuillage était effectué avec la deuxième méthode de seuillage (voir l'annexe B), qui n'est pas forcément la plus efficace.

7. Voir [Mar80] page 442 ou [KM81] page 1388, où $S(f)$ est appelée « power spectral density (PSD) ».

En écrivant l'auto-corrélation du signal ainsi : $R^{xx}(k) = E[x(n+k)x^*(n)]$, nous obtenons les équations de YULE-WALKER :

$$\begin{bmatrix} R^{xx}(0) & R^{xx}(-1) & \cdots & R^{xx}(-P) \\ R^{xx}(1) & R^{xx}(0) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ R^{xx}(P) & \cdots & \cdots & R^{xx}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_{P,1} \\ \vdots \\ a_{P,P} \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Nous avons à notre disposition diverses méthodes permettant de résoudre récursivement (ordre variant de 1 à P) ces équations de YULE-WALKER : DURBIN-LEVINSON, BURG, MARPLE (pour plus de détails, voir [Mar80] et [KM81]). Ces trois méthodes ont été implémentées dans le programme *segmentation*.

2.2.6.2 Notations et procédure

La décision de rupture à un instant i est basée sur l'observation de l'entropie croisée de deux modèles AR du signal. Le premier est calculé à partir d'une portion M_1 du signal, de taille fixe L et telle que $M_1 = [x(i-L+1)\cdots x(i)]$; et le second à partir d'une portion M_2 du signal dont l'origine est fixe et correspond à l'instant de détection de la rupture précédente, que nous notons j, avec bien sûr $j < i$. Ainsi : $M_2 = [x(j)\cdots x(i)]$ (voir la figure 2.2). Les moyennes de M_1 et M_2 sont normalisées à 0. L'ordre des deux modèles est P :

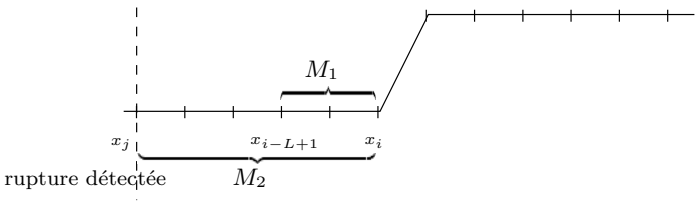


FIG. 2.2 – *Disposition des deux portions pour la rupture de modèles sur le trajet de* f_0

Nous adoptons les notations suivantes :

- $(A_1,\sigma_1) = (a_1^1 \cdots a_1^P,\ \sigma_1)$ les paramètres AR obtenus pour le premier modèle.
- $(A_2,\sigma_2) = (a_2^1 \cdots a_2^P,\ \sigma_2)$ les paramètres AR obtenus pour le second modèle.

De plus :

• $X_{k-1} = [x_1 \cdots x_{k-1}]$ correspond au passé de l'échantillon x_k.

et :

• $p_1^k(x/X_{k-1})$ et $p_2^k(x/X_{k-1})$ correspondent aux densités de probabilité conditionnelles associées respectivement à $M_1^k(A_1,\sigma_1)$ et à $M_2^k(A_2,\sigma_2)$.

2.2.6.3 L'entropie croisée

L'entropie croisée W_n est alors :

$$w_k = \int_{-\infty}^{+\infty} \left\{ p_2^k(x/X_{k-1}) \log_2 \left(\frac{p_1^k(x/X_{k-1})}{p_2^k(x/X_{k-1})} \right) \right\} dx - \log_2 \left(\frac{p_1^k(x_k/X_{k-1})}{p_2^k(x_k/X_{k-1})} \right)$$

Ce qui s'écrit dans le cas gaussien (voir [BB82]) :

$$w_k = \frac{1}{2} \left\{ 2\frac{e_2^k e_1^k}{\sigma_1^2} - \left[1 + \frac{\sigma_2^2}{\sigma_1^2} \right] \frac{e_2^k e_2^k}{\sigma_2^2} + \left[1 - \frac{\sigma_2^2}{\sigma_1^2} \right] \right\}$$

avec $e_\alpha^k = x_k - \sum_{m=1}^{P} a_\alpha^m x_{k-m}$ l'erreur de prédiction pour le modèle α, avec α valant 1 ou 2.

La statistique est définie comme étant la somme cumulée :

$$W_k = \sum_{l=1}^{k} w_l + \delta$$

où δ est un biais positif systématiquement ajouté, selon la règle de PAGE-HINKLEY (voir [BB82] page 26 et [LO95]).

2.2.6.4 Détection de rupture

Le problème est à présent de détecter un maximum local de W_k, moyennant un certain seuil λ. Cela se fait de la façon suivante :

si $(\max_{1 \leq l \leq k} W_l) - W_k > \lambda$ alors *rupture à l'instant* k, $k = k + 1$, et réinitialisation de W_k à 0
sinon $k = k + 1$

2.2.6.5 Le retour en arrière

Or, w_k a un comportement asymétrique par rapport aux modèles obtenus pour M_1 et M_2. Cela veut dire que certains sauts sont plus facilement détectables si nous nous déplaçons dans le sens temporel rétrograde que si nous nous déplaçons dans le sens normal. Le signal est donc traité dans le sens rétrograde dès qu'il est soupçonné qu'une omission a eu lieu. Nous décidons qu'il y a eu peut-être omission quand : $i - j > L_{min}$. Alors nous allons dans le sens rétrograde. Une fois que nous allons dans le sens rétrograde, si nous détectons une rupture à l'instant j^r et si $j^r - j > D_{min}$, la rupture à l'instant j^r est acceptée et nous reprenons l'analyse dans le sens temporel normal à partir de l'instant j^r.

2.2.6.6 L'apport des harmoniques de numéros d'ordre supérieurs

Nous utilisons cette algorithme de la même façon sur les harmoniques de numéros d'ordre supérieurs. Il faut alors fusionner les résultats (voir la section 2.6, en remarquant qu'ici nous obtenons des **fonctions de décision**, c'est-à-dire des listes de marques de segmentation, et non pas des **fonctions d'observation**). Ceci n'est pas fait dans le programme *segmentation*.

2.2.6.7 Les paramètres libres du programme

Ils sont au nombre de 7. Ce sont : P, l'ordre des modèles AR ; L, la taille de la fenêtre d'analyse de taille fixe ; δ, le biais systématiquement ajouté ; λ, le seuil de détection ; L_{min}, la longueur à partir de laquelle nous soupçonnons qu'une omission a eu lieu ; D_{min}, la longueur minimale entre une détection normale et une détection rétrograde ; L_{pris}, le nombre d'harmoniques pris en compte. Par défaut, nous avons : $P = 1$; L, telle que la largeur de la fenêtre de taille fixe soit de 200 millisecondes ; $\delta = 0{,}001$; $\lambda = 50$. Par défaut, l'analyse dans le sens rétrograde n'est pas effectuée. Si elle est utilisée, nous avons : $L_{min} = 6L$; $D_{min} = 3L$.

2.2.6.8 Conclusion

Contrairement aux autres fonctions d'observation, la prise de décision est incluse dans la « rupture de modèles AR ». Cette remarque, très importante, est valable aussi pour la « rupture de modèles AR » sur le trajet de l'énergie (voir la section 2.3.2) et pour la « rupture de modèles AR » directement sur le son (voir la section 2.5.2). En fait, les deux premières étapes de l'analyse *segmentation en zones stables* sont confondues. Cette remarque implique que ce que nous présentons dans le chapitre 3 ne s'applique pas ici.

Ce n'est pas le cas pour d'autres fonctions d'observation basées elles aussi sur la modélisation AR : voir les sections 2.4.3.3 et 2.5.3.

2.3 Fonctions d'observation basées sur les variations de l'énergie

2.3.1 Introduction : détermination du trajet de l'énergie

Nous entendons par « énergie », ici, l'énergie calculée sur une portion du son de taille t_{SIG} échantillons (t_{SIG} classiquement correspond à une portion de quelques dizaines de millisecondes), c'est-à-dire que cela correspond plutôt à une puissance.

Nous obtenons le trajet de l'énergie E du signal sonore dans le temps. Elle change pour chaque note et, de plus, pour certains signaux, aux instants de transition, elle passe par des minimums très faibles. Nous voulons ici détecter les sauts de la fonction temporelle $E(i)$.

Moorer ([Moo75]), Morris et Schwartz ([MSE93]), notamment, utilisent l'énergie pour segmenter.

Cette fois-ci, le critère n'a qu'une dimension, et nous n'aurons pas les problèmes de fusion des résultats rencontrés pour les fonctions d'observation basées sur les variations de f_0 et des harmoniques de numéros d'ordre supérieurs (voir la section 2.2.6.6).

2.3.2 Les fonctions d'observation

De la même façon que pour le trajet de f_0, nous utilisons, comme fonctions d'observation, pour le trajet de l'énergie :

- Les dérivées $\left|d^I E\right|$, $\left|d^{II} E\right|$, $\left|\delta^I E\right|$ et $\left|\delta^{II} E\right|$ du trajet de l'énergie. Elles sont implémentées dans le programme *segmentation*.
- L'analyse statistique AS_E^I ou AS_E^{II} sur le trajet de l'énergie. Elle n'est pas implémentée dans le programme *segmentation*.
- La rupture de modèles sur le trajet de l'énergie. Elle n'est pas implémentée dans le programme *segmentation*.

2.4 Fonctions d'observation basées sur les variations du contenu spectral

2.4.1 Indice de voisement – Deuxième forme, calculé avec le spectre d'amplitude

2.4.1.1 Introduction

Deux indices de voisement sont définis : le premier, qui est basé sur l'analyse des spectres **d'amplitude**, dans cette section 2.4.1 ; et le second, qui est basé sur l'analyse des spectres **complexes**, dans la suivante (section 2.4.2). Ils sont différents des indices de voisement définis dans la section 2.2.3 en ce que nous ne tenons pas compte de f_0. Nous ne faisons ici aucune hypothèse sur l'harmonicité du signal. Ces considérations nous amèneront, dans le chapitre 19 (partie IV), page 149, à faire une distinction, toute personnelle, entre « signal voisé » et « signal harmonique ».

Ces deux nouveaux indices de voisement ne sont pas encore implémentés en C, et donc ne sont pas inclus dans le programme *segmentation*.

2.4.1.2 Principe de la méthode (voir la section 2.2.3.1)

Une portion du son large de quelques dizaines de millisecondes et d'instant central i est considérée. Cette portion est multipliée par la fenêtre de pondération w. Le spectre d'amplitude $\left|\hat{S}\right|$ de cette portion pondérée du signal est calculé. Le spectre d'amplitude $\left|\hat{W}\right|$ de la fenêtre de pondération w est calculé. $\left|\hat{W}\right|$ est tronqué autour de la fréquence nulle. Ce tronçon normalisé en énergie, $\left|\hat{W}^{(2N+1)}\right|$, est large de $2N+1$ échantillons fréquentiels. Nous considérons des tronçons

normalisés en énergie $\left|\hat{S}_{(m)}^{(2N+1)}\right|$ glissants de $\left|\hat{S}\right|$. m est la position de l'échantillon fréquentiel central de chaque tronçon, et la largeur de chaque tronçon est de $2N+1$ échantillons fréquentiels. Nous corrélons $\left|\hat{W}^{(2N+1)}\right|$ et chaque $\left|\hat{S}_{(m)}^{(2N+1)}\right|$. Rappelons que les normalisations en énergie se font ainsi (voir page 13) :

$$\left|\hat{S}_{(m)}^{(2N+1)}(k)\right| = \frac{\left|\hat{S}(k)\right|}{\sqrt{\left(\sum_{j=m-N}^{m+N}\left|\hat{S}(j)\right|^2\right)}}, \text{ pour } k \in [m-N \ \dots \ m+N] \text{ ; et :}$$

$$\left|\hat{W}^{(2N+1)}(k)\right| = \frac{\left|\hat{W}(k)\right|}{\sqrt{\left(\sum_{j=-N}^{+N}\left|\hat{W}(j)\right|^2\right)}}, \text{ pour } k \in [-N \ \dots \ N]$$

Les notations sont telles que pour l'échantillon fréquentiel d'indice 0, la fréquence est 0, pour l'échantillon fréquentiel d'indice 1, $f = \frac{f_e}{t_{FFT}}$, pour l'échantillon fréquentiel d'indice -1, $f = -\frac{f_e}{t_{FFT}}$...

La corrélation pour le tronçon $\left|\hat{S}_{(m)}^{(2N+1)}\right|$ de position centrale m se calcule ainsi :

$$C(m) = \sum_{j=1}^{2N+1}\left|\hat{W}^{(2N+1)}(j)\right|\left|\hat{S}_{(m)}^{(2N+1)}(j)\right|$$

Le spectre est calculé pour M échantillons fréquentiels entre 0 et $\frac{f_e}{2}$. Il est donc composé de $M-N+1$ tronçons (m varie de 0 à $M-N$). Nous obtenons un vecteur de corrélation C composé de $M-N+1$ points.

Nous détectons les maximums locaux de ce vecteur C. Pour chacun de ceux qui dépassent un seuil s, nous décidons que nous avons une sinusoïde en cet endroit m du spectre d'amplitude $\left|\hat{S}\right|$. Il est nécessaire d'ajouter une « condition supplémentaire » : cependant, si la distance d entre deux maximums locaux supérieurs à s consécutifs ($d = m_{(2)} - m_{(1)}$) est inférieure à $2N+1$, nous ne gardons que la sinusoïde correspondant à la corrélation la plus grande [8]. Cette « condition supplémentaire » permet d'assurer qu'aucun échantillon fréquentiel n'est présent dans plusieurs des sinusoïdes gardées, que nous ne comptons pas plusieurs fois l'énergie d'un échantillon fréquentiel.

Nous calculons l'énergie de chaque sinusoïde gardée, c'est-à-dire l'énergie du tronçon concerné, de position centrale m :

$$e_m = \sum_{j=m-N}^{m+N}\left|\hat{S}(j)\right|^2$$

Nous additionnons les énergies de toutes les sinusoïdes gardées. Nous obtenons une énergie e. L'énergie totale du signal est $E = \sum_{j=0}^{M}\left|\hat{S}(j)\right|^2$. Si le rapport $R = \frac{e}{E}$ des deux énergies est supérieur à un seuil S à fixer, nous décidons que le signal est plutôt voisé, c'est-à-dire plutôt composé de sinusoïdes ; sinon, nous décidons qu'il n'est pas voisé [9]. Grâce à la condition supplémentaire, il est assuré que le rapport R soit forcément inférieur ou égal à 1.

8. Nous pourrions plutôt garder celle des deux sinusoïdes dont l'énergie est la plus grande ; ou mixer les deux tests (par exemple utiliser le produit de la corrélation par l'énergie pour chaque sinusoïde et ne garder que la sinusoïde correspondant au plus grand de ces produits). Il s'agit de perspectives.

9. Cette remarque concerne les parties III et IV, c'est-à-dire la *segmentation en caractéristiques* et la détection de la polyphonie.

Nous avons obtenu ainsi un indice de voisement proche de celui défini dans la section 2.2.3, mais, contrairement à celui-ci, indépendant de la détermination de f_0 (pour l'indice de voisement défini dans la section 2.2.3, nous corrélons avec les portions du spectre autour de f_0, $2f_0$, $3f_0$...).

La fonction d'observation implémentée sous MATLAB est la « valeur absolue de la dérivée de $1 - R$ », soit :

$$\left|dV^{(R)}(i)\right| = |1 - R(i) - 1 + R(i-1)| = |R(i-1) - R(i)|$$

2.4.1.3 Cas où le signal est un bruit normal

Il faut remarquer que quand le signal est du bruit, les corrélations ne sont pas nulles.

Préliminaires Si le signal est une suite aléatoire normale de loi $\mathcal{N}\left(0,\text{std}^2\right)$, nous prouvons (voir [Bri81]) que les parties réelles et imaginaires du spectre complexe sont indépendantes et qu'elles suivent des lois normales $\mathcal{N}\left(0,\dfrac{\text{std}^2}{2}\right)$. Le spectre d'amplitude est le module du complexe $Z = X + iY$ où X et Y sont respectivement les parties réelles et imaginaires du spectre complexe. Nous prouvons alors que $|Z|$ suit une loi de RAYLEIGH (voir [Cha96], pages 38 – 39). La densité de probabilité de $|Z|$ est :

$$p_{|Z|}(|z|) = \frac{|z|}{\sigma^2} \exp\left(-\frac{|z|^2}{2\sigma^2}\right)$$

avec $\sigma^2 = \dfrac{(\text{std})^2}{2}$.

Développements Nous prouvons (voir l'annexe G), que la moyenne de la variable aléatoire $v = \sqrt{\left(\sum_{j=m-N}^{m+N} \left|\hat{S}(j)\right|^2\right)}$ tend vers $\sqrt{2}\sqrt{2N+1}\sigma$ quand $2N+1$ tend vers l'infini. La valeur moyenne de l'amplitude de chaque échantillon fréquentiel est : $\sqrt{\dfrac{\pi}{2}}\sigma$. Ainsi :

$$E\left[C(m)\right] \simeq \frac{\sum_{j=1}^{2N+1} \left|\hat{W}^{(2N+1)}(j)\right|}{\sqrt{2N+1}} \frac{\sqrt{\pi}}{2} \qquad \text{formule (1)}$$

qui ne dépend pas des paramètres du bruit. Remarquons que puisque $\sqrt{2}\sqrt{2N+1}\sigma$ est supérieur à la valeur vraie de v, $E\left[C(m)\right]$ est supérieure à $\dfrac{\sum_{j=1}^{2N+1} \left|\hat{W}^{(2N+1)}(j)\right|}{\sqrt{2N+1}} \dfrac{\sqrt{\pi}}{2}$. Ceci est représenté sur les figures 2.3, 2.4 et 2.5. Pour chaque figure, la courbe en trait interrompu représente la valeur théorique et la courbe en trait plein la valeur estimée des moyennes de C en fonction de N, la largeur de chaque tronçon étant $2N + 1$. Les fenêtres de pondération utilisées ont été la fenêtre RECTANGULAIRE, la fenêtre de HANNING et la fenêtre de BLACKMAN. D'après ces courbes et la formule (1), nous constatons que plus N est grand plus la moyenne des corrélations avec le bruit est petite : ainsi, a priori, il faut prendre N le plus grand possible.

2.4.1.4 Influence du bourrage de zéros

Appliquer un bourrage de zéros ne change rien aux résultats en ce qui concerne la corrélation avec le bruit ou les lobes secondaires d'une sinusoïde. En fait, dans le cas d'un signal suivant une loi normale $\mathcal{N}\left(0,\sigma^2\right)$, les échantillons fréquentiels réels et imaginaires du spectre complexe suivent des lois normales $\mathcal{N}\left(0,\dfrac{\sigma^2}{2}\dfrac{t_{SIG}}{t_{FFT}}\right)$. Mais, puisque la moyenne des corrélations ne dépend pas de la variance des échantillons fréquentiels (voir la formule (1)), le bourrage de zéros n'a pas d'influence sur elles.

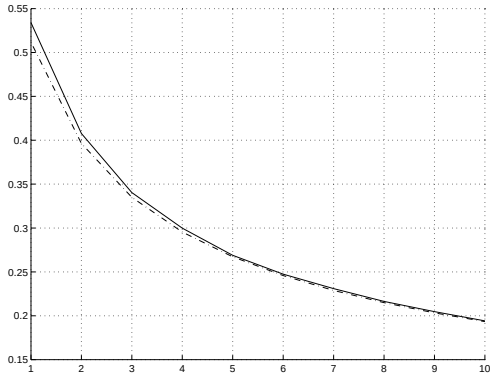


FIG. 2.3 – *Influence de N sur la corrélation. En abscisse : N ; en ordonnée : la corrélation. La fenêtre utilisée est la fenêtre* RECTANGULAIRE. *Le signal est normal. Trait interrompu : valeur théorique ; trait plein : valeur estimée de $E[C(m)]$*

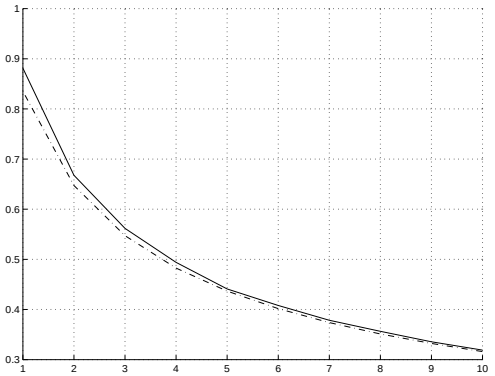


FIG. 2.4 – *Influence de N sur la corrélation. En abscisse : N ; en ordonnée : la corrélation. La fenêtre utilisée est la fenêtre de* HANNING. *Le signal est normal. Trait interrompu : valeur théorique ; trait plein : valeur estimée de $E[C(m)]$*

Les corrélations pratiques et théoriques en fonction de N sont montrées, sans bourrage de zéros sur la figure 2.5, et avec un taux de bourrage de zéros de 16 (la taille de la FFT est 16 fois la taille du signal) sur la figure 2.6. Bien sûr, dans le second cas, il faut, pour obtenir des résultats comparables au cas sans bourrage de zéros, prendre des tronçons plus larges.

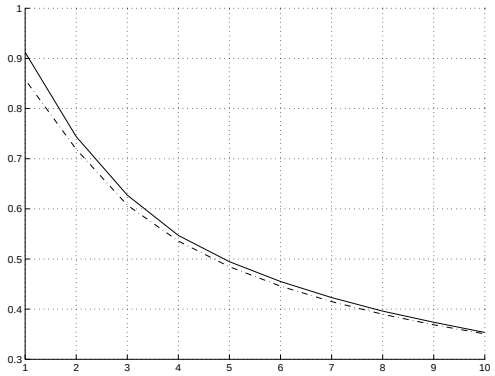


FIG. 2.5 – *Influence de N sur la corrélation. En abscisse : N ; en ordonnée : la corrélation. La fenêtre utilisée est la fenêtre de* BLACKMAN. *Le signal est normal. Trait interrompu : valeur théorique ; trait plein : valeur estimée de $E[C(m)]$. Aucun bourrage de zéros n'est appliqué*

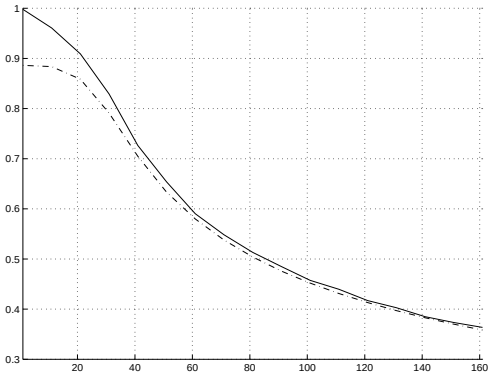


FIG. 2.6 – *Influence de N sur la corrélation. En abscisse : N ; en ordonnée : la corrélation. La fenêtre utilisée est la fenêtre de* BLACKMAN. *Le signal est normal. Trait interrompu : valeur théorique ; trait plein : valeur estimée de $E[C(m)]$. Un taux de bourrage de zéros de 16 est appliqué*

Par contre, le bourrage de zéros permet d'améliorer les résultats de la corrélation avec le lobe principal donné par une sinusoïde pure : ainsi, en moyenne, la fréquence de cette sinusoïde est d'autant plus proche de celle d'un échantillon fréquentiel du spectre que le taux de bourrage de zéros est important, et donc la corrélation s'approche d'autant plus de 1. En moyenne, car cela dépend bien sûr de la fréquence de la sinusoïde : il y a toujours le cas défavorable où la fréquence d'une sinusoïde tombe juste entre deux échantillons fréquentiels du spectre.

2.4.1.5 Cas d'une sinusoïde : corrélations avec les lobes secondaires

Dans ce cas, les échantillons fréquentiels du spectre d'amplitude ne suivent pas une loi de RAYLEIGH.

La fenêtre de pondération utilisée est la fenêtre BLACKMAN. Les autres paramètres sont :

- taille du signal (t_{SIG}) : 1024
- taille de la FFT (t_{FFT}) : 16384 (taux de bourrage de zéros : 16)
- amplitude de la sinusoïde quand elle est présente : 1
- largeur du demi-tronçon (N) : 80 (les deux premiers lobes secondaires de la fenêtre de pondération dans le domaine fréquentiel sont pris en compte)
- fréquence de la sinusoïde (f) : 400
- fréquence d'échantillonnage (f_e) : 4000

Ici, nous nous intéressons avant tout au cas où nous considérons un tronçon (de taille $2N+1$) du spectre d'amplitude dans les fréquences éloignées de f. Les résultats sont présentés sur les figures 2.7, 2.8, 2.9 et 2.10. Nous constatons que la moyenne des corrélations avec les lobes secondaires est égale à la moyenne des corrélations avec le bruit.

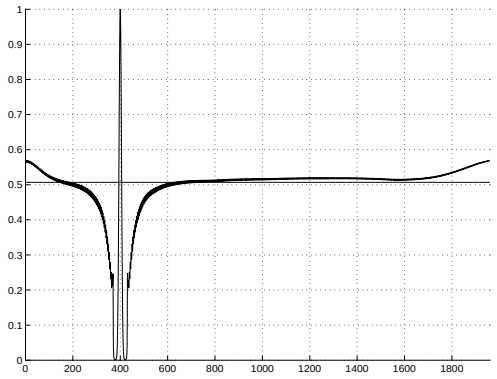


FIG. 2.7 – *Corrélations sur un spectre d'amplitude entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, est présente. Pas de bruit. La moyenne théorique $E[C(m)]$ des corrélations quand le signal est du bruit est le trait horizontal juste au-dessus de* 0,5

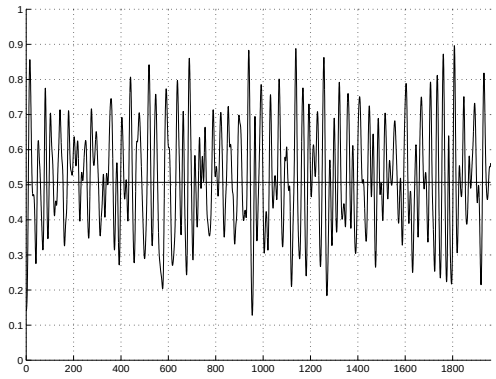


FIG. 2.8 – *Corrélations sur un spectre d'amplitude entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Seul un bruit normal ($m = 0$, $\sigma = 1$) est présent. La moyenne théorique $E[C(m)]$ des corrélations quand le signal est du bruit est le trait horizontal juste au-dessus de* 0,5

2.4.1.6 Les paramètres libres

Ils sont au nombre de 5. Ce sont : N, la demi-largeur [10] des tronçons pour la corrélation ; s, le seuil au-dessus duquel il est décidé qu'une sinusoïde est présente ; t_{FFT}, la taille de la FFT ; t_{SIG}, la taille des fenêtres d'analyse ; et la fenêtre de pondération utilisée. Pour les signaux réels sur lesquels cette fonction d'observation a été testée (voir les figures 6.15 page 63 et 8.2 page 80), nous avons choisi $N = 20$; $s = 0{,}8$; $t_{FFT} = 4096$; $T = 0{,}04$ s ($t_{SIG} = 1764$) ; et la fenêtre de pondération de BLACKMAN.

10. La taille du tronçon est en fait $2N + 1$.

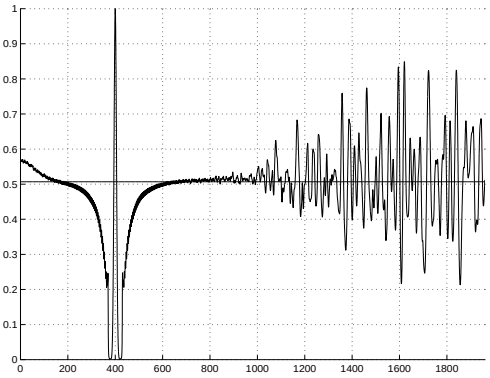


FIG. 2.9 – *Corrélations sur un spectre d'amplitude entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, et un bruit normal ($m = 0$, $\sigma = 10^{-7}$) sont présents. La moyenne théorique $E[C(m)]$ des corrélations quand le signal est du bruit est le trait horizontal juste au-dessus de* 0,5

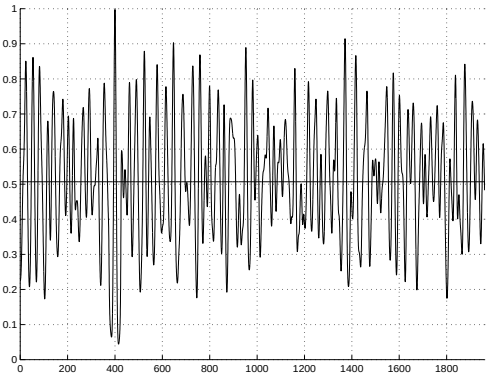


FIG. 2.10 – *Corrélations sur un spectre d'amplitude entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, et un bruit normal ($m = 0$, $\sigma = 0{,}5$) sont présents. La moyenne théorique $E[C(m)]$ des corrélations quand le signal est du bruit est le trait horizontal juste au-dessus de* 0,5

2.4.2 Indice de voisement – Deuxième forme, calculé avec le spectre complexe

2.4.2.1 Méthode

Cette fois nous avons :

$$\hat{S}^{(2N+1)}_{(m)}(k) = \frac{\hat{S}(k)}{\sqrt{\left(\sum_{j=m-N}^{m+N} \left|\hat{S}(j)\right|^2\right)}} \text{ pour } k \in [m-N \ \ldots \ m+N]$$

et :

$$\hat{W}^{(2N+1)}(k) = \frac{\hat{W}(k)}{\sqrt{\left(\sum_{j=-N}^{+N} \left|\hat{W}(j)\right|^2\right)}} \text{ pour } k \in [-N \ \ldots \ N]$$

La corrélation s'écrit alors :

$$C(m) = \left| \sum_{j=1}^{2N+1} \hat{W}^{(2N+1)}(j) \hat{P}^{(2N+1)}_{(m)}(j) \right|$$

Les corrélations avec les lobes secondaires et dans le cas où le signal est un bruit normal sont montrés sur les figures 2.11, 2.12, 2.13 et 2.14. Nous constatons que les corrélations avec les lobes secondaires sont nulles, mais que dès qu'un petit bruit est présent nous obtenons les mêmes résultats qu'avec l'indice de voisement deuxième forme calculé avec les spectres d'amplitude. Nous avons pris les mêmes valeurs pour les paramètres que ci-dessus (section 2.4.1.5). Les paramètres libres sont les mêmes que pour l'indice de voisement deuxième forme calculé avec les spectres d'amplitude (section 2.4.1.6).

La fonction d'observation implémentée sous MATLAB est la « valeur absolue de la dérivée de $1 - R$ », soit :

$$\left| dV^{(C)}(i) \right| = |R(i-1) - R(i)|.$$

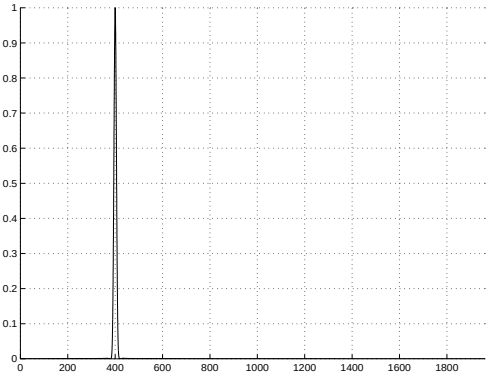


FIG. 2.11 – *Corrélations sur un spectre complexe entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, est présente. Pas de bruit*

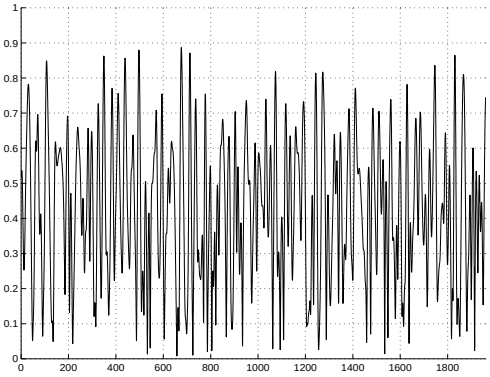


FIG. 2.12 – *Corrélations sur un spectre complexe entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Seul un bruit normal ($m = 0$, $\sigma = 1$) est présent*

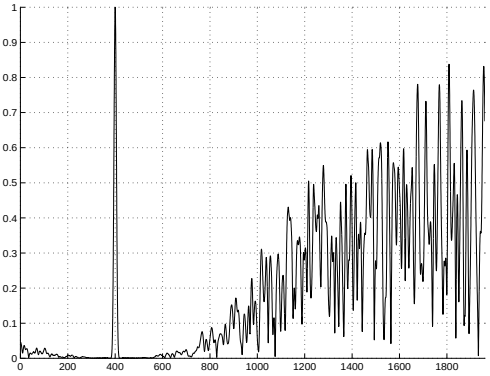


FIG. 2.13 – *Corrélations sur un spectre complexe entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, et un bruit normal ($m = 0$, $\sigma = 10^{-7}$) sont présents*

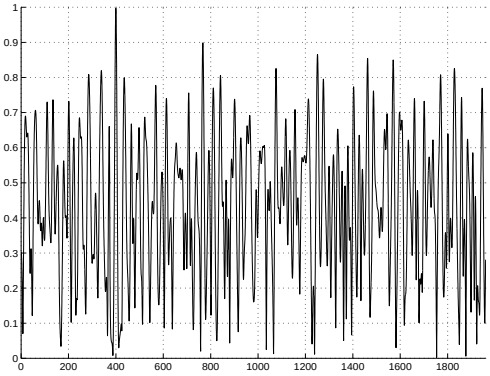


FIG. 2.14 – *Corrélations sur un spectre complexe entier. En abscisse : la fréquence ; en ordonnée : la corrélation. Une sinusoïde, de fréquence 400 Hz, et un bruit normal ($m = 0$, $\sigma = 0{,}5$) sont présents*

2.4.2.2 Perspectives

Premièrement, plus N est grande, plus les performances sont bonnes, c'est-à-dire plus les corrélations sont proches de 0 quand nous utilisons les spectres complexes et un signal non bruité et corrélons avec des portions de spectre loin de la zone d'influence de la sinusoïde. De la même façon, plus N est grande, plus les corrélations avec le bruit (et les lobes secondaires quand nous utilisons le spectre d'amplitude) sont petites.

Cependant, si une perturbation (c'est-à-dire une modulation d'amplitude ou de fréquence) vient déformer les lobes principaux, plus N est grande, plus la corrélation avec ces lobes principaux est petite.

La première solution serait de trouver un compromis.

Une autre solution serait d'étudier l'influence de ces perturbations sur les lobes principaux et d'essayer d'améliorer les indices de voisement en conséquence. Quelques travaux ont été faits en ce sens, qui ne sont pas présentés dans cet exposé.

Deuxièmement, cette fonction d'observation nous permet de déterminer dans quelles bandes de fréquences le signal, avec des sons « simples » (c'est-à-dire monophoniques, harmoniques et non modulés), est plutôt voisé et dans quelles autres il ne l'est plutôt pas. Nous allons voir que les « flux spectraux » (voir ci-dessous la section 2.4.3) peuvent être calculés sur toutes les fréquences ou par bandes de fréquences. Le voisement nous permet de régler automatiquement le paramètre libre « position de la césure » pour les « flux spectraux ».

2.4.2.3 Conclusion

Le comportement de la fonction d'observation est très différent en ce qui concerne les lobes secondaires suivant que nous la calculons avec le spectre d'amplitude ou le spectre complexe : le spectre complexe donne de meilleurs résultats. Cependant, en présence de bruit, même de variance faible, cet apport n'est plus visible. Utiliser l'une ou l'autre de cette version de cette fonction d'observation ne semble donc pas déterminant.

2.4.3 Les « flux spectraux »

2.4.3.1 Définition(s) du terme « flux spectral »

Un flux spectral est la somme échantillon fréquentiel par échantillon fréquentiel de la valeur absolue de la différence entre deux spectres d'amplitude, entre deux enveloppes spectrales, ou entre un spectre d'amplitude et une enveloppe spectrale.

Les spectres d'amplitude sont calculés par FFT (voir ci-dessous la section 2.4.3.2) ; les enveloppes spectrales à partir de la modélisation AR (voir ci-dessous la section 2.4.3.3), du cepstre (voir ci-dessous la section 2.4.3.4) ou des maximums locaux du spectre d'amplitude (voir la section 12.3.5, dans la partie III).

2.4.3.2 1°Avec les spectres d'amplitude

Nous calculons le spectre d'amplitude sur une portion pondérée de largeur T du signal (T vaut quelques dizaines de millisecondes : c'est-à-dire que t_{SIG} vaut quelques centaines d'échantillons), puis sur une portion pondérée du signal décalée de quelques millisecondes (c'est-à-dire de Q échantillons) par rapport à la première. Les spectres d'amplitude sont calculés pour t_{FFT} échantillons fréquentiels. Nous appelons le premier spectre d'amplitude $\left|\hat{S}_1\right|$ et le second $\left|\hat{S}_2\right|$. Nous avons donc $\left|\hat{S}_1(m)\right|$ et $\left|\hat{S}_2(m)\right|$ pour m variant de $\frac{-t_{FFT}}{2}+1$ à $\frac{t_{FFT}}{2}$.

Nous normalisons les spectres d'amplitude $\left|\hat{S}_1\right|$ et $\left|\hat{S}_2\right|$ pour qu'ils aient la même énergie comme nous l'indiquons page 13, avec $a=-\frac{t_{FFT}}{2}+1$ et $b=\frac{t_{FFT}}{2}$. Nous obtenons $\left|\hat{S}_1^{norm}\right|$ et $\left|\hat{S}_2^{norm}\right|$.

Il s'agit de calculer :

$$F = \sum_{m=0}^{\frac{t_{FFT}}{2}} \left| \left| \hat{S}_1^{norm}(m) \right| - \left| \hat{S}_2^{norm}(m) \right| \right|$$

Aux moments des transitions, c'est-à-dire aux changements de notes, la fonction d'observation F, ou « flux spectral », croît, puisque le signal n'est plus stationnaire.

2.4.3.3 2°Avec l'enveloppe spectrale basée sur la modélisation AR

Introduction L'enveloppe spectrale d'un spectre d'amplitude est une fonction lisse enveloppant ses pics (voir la figure 2.15), correspondant chacun à une sinusoïde. Une petite variation de l'amplitude ou de la fréquence (due à un trémolo ou à vibrato, entre autres) de ces sinusoïdes ne modifie pas énormément, d'une portion du signal à la suivante, l'enveloppe spectrale, alors que le spectre d'amplitude lui est très différent. Les effets du vibrato et du trémolo, ainsi que ceux du bruit, sont donc moins importants sur les enveloppes spectrales que sur les spectres d'amplitude. Nous espérons en utilisant les enveloppes spectrales plutôt que les spectres d'amplitude augmenter la robustesse de la fonction d'observation « flux spectral ».

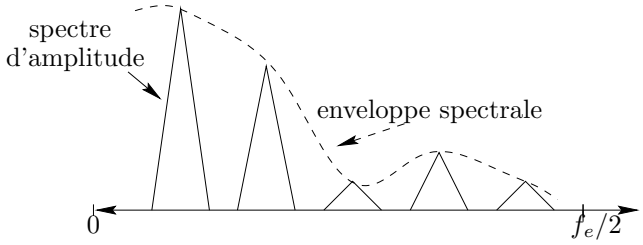


FIG. 2.15 – *Spectre d'amplitude et enveloppe spectrale lui correspondant*

Il existe plusieurs moyens permettant d'obtenir l'enveloppe spectrale d'un spectre d'amplitude. Tout d'abord il y a la modélisation auto-régressive (AR) : voir la section 2.2.6.1. Le choix de l'ordre des modèles pose un problème : il ne doit pas être trop élevé pour ne pas avoir une résonance trop importante pour les sinusoïdes présentes dans le signal. De plus, le temps de calcul devient rapidement très grand quand l'ordre augmente. Cette méthode est celle qui a été implémentée dans le programme *segmentation*.

Mais nous pouvons aussi considérer le cepstre (voir ci-dessous la section 2.4.3.4), ou, encore, interpoler, linéairement ou d'une autre façon, entre les pics les plus grands (le problème étant de choisir les pics correspondant réellement à des sinusoïdes : voir la section 12.3 de la partie III) d'un spectre d'amplitude obtenu par transformée de FOURIER ; etc.

Méthode Ici, nous appelons la densité spectrale de puissance définie dans la section 2.2.6.1, page 18, l'enveloppe spectrale.

Les deux formes de la fonction d'observation Il y a deux versions possibles de cette fonction d'observation :

- Nous calculons la valeur absolue de la dérivée de $F_1^{AR} = \sum_{m=0}^{\frac{t_{FFT}}{2}} \left| \left| \hat{S}(m) \right| - \hat{S}^{AR}(m) \right|$, où $\left| \hat{S} \right|$ est le spectre d'amplitude calculé pour une portion du signal large de quelques dizaines de millisecondes et où $\hat{S}^{AR}$ est l'enveloppe spectrale calculée pour la même portion. m est le numéro d'ordre des échantillons fréquentiels. Dans les zones où le signal est composé de sinusoïdes (signal voisé), la différence F_1^{AR} entre le spectre d'amplitude et l'enveloppe spectrale est plus grande que dans les parties où il est transitoire (signal non voisé).

- Cette version ressemble plus au flux spectral décrit ci-dessus (voir la section 2.4.3.2). Il s'agit de calculer $F_2^{AR} = \sum_{m=0}^{\frac{t_{FFT}}{2}} \left|\hat{S}_1^{AR}(m) - \hat{S}_2^{AR}(m)\right|$, où $\hat{S}_1^{AR}$ et $\hat{S}_2^{AR}$ sont deux enveloppes spectrales AR calculées pour deux portions successives du signal, décalées de Q échantillons.

2.4.3.4 3°Avec l'enveloppe spectrale basée sur le liftrage du cepstre

Cette méthode est similaire à la méthode décrite dans 2°(section 2.4.3.3). Elle diffère simplement par la technique utilisée pour calculer les enveloppes spectrales. Il y a deux versions possibles de la fonction d'observation :

- Nous calculons la dérivée de $F_1^{ceps} = \sum_{m=0}^{\frac{t_{FFT}}{2}} \left|\left|\hat{S}(m)\right| - \left|\hat{M}(m)\right|\right|$, où $\left|\hat{S}\right|$ est le spectre d'amplitude calculé sur une portion du signal large de t_{SIG} échantillons et où $\left|\hat{M}\right|$ est le spectre d'amplitude reconstruit après liftrage (voir la section 24.2.2.4 de la partie V pour une présentation du cepstre et du liftrage) calculé sur la même portion. m est le numéro d'ordre des échantillons fréquentiels.
- Cette version ressemble plus au flux spectral décrit dans 2.4.3.2. Il s'agit de calculer $F_2^{ceps} = \sum_{m=0}^{\frac{t_{FFT}}{2}} \left|\left|\hat{M}_1(m)\right| - \left|\hat{M}_2(m)\right|\right|$, où $\left|\hat{M}_1\right|$ et $\left|\hat{M}_2\right|$ sont deux spectres d'amplitude reconstruits après liftrage calculés pour deux portions successives du signal, décalées de Q échantillons.

2.4.3.5 Une fonction d'observation à plusieurs dimensions

Nous pouvons découper les spectres d'amplitude ou les enveloppes spectrales en plusieurs bandes de fréquence. Alors, nous normalisons en énergie chacune de ces bandes indépendamment des autres. Nous calculons le « flux spectral » pour chacune de ces bandes. Nous avons la possibilité :

- De découper les spectres d'amplitude ou les enveloppes spectrales en B bandes de tailles égales.
- De les découper en B bandes de tailles croissant linéairement. Les fréquences centrales des bandes de fréquence peuvent être disposées logarithmiquement, de telle manière que nous suivions une échelle Bark/Mels (voir [ZF81] par exemple).
- De les découper en deux bandes, la position de la césure c étant à entrer en paramètre, ou à déterminer automatiquement, le but étant de déterminer dans quelle bande de fréquences le signal est plutôt stable et dans quelle bande de fréquences il correspond plutôt à du bruit (voir la section 2.4.2.2).

Ceci n'a pas été implémenté dans le programme *segmentation*. Le nombre de paramètres libres augmenterait d'une unité. Il faudrait ajouter B, le nombre de bandes considérées, ou c la position de la césure : nous n'avons pas essayé au cours de cette thèse de déterminer la position optimale de la césure. Il s'agit de perspectives.

2.4.3.6 Les paramètres libres

Ils sont au nombre de 4. Ce sont :

- t_{SIG} : la taille des portions de signal.
- Suivant la méthode considérée pour calculer le « flux spectral » :
  1°la taille t_{FFT} de la FFT
  2°et/ou l'ordre P de la modélisation AR
  3°et/ou t_{FFT} et l'endroit où nous coupons le cepstre (voir la section 24.2.2.4, le seuil S_C)
- Q : l'écart temporel entre les deux portions successives s'il y a lieu. Pour les méthodes avec les enveloppes spectrales (2°et 3°), selon la version du « flux spectral » utilisée, il n'y a pas forcément besoin de Q.
- La fenêtre de pondération quand il y a lieu (pour 1°et 3°).

L'ordre des modèles AR est choisi petit, de l'ordre de 6. Q est tel que l'écart temporel entre les deux fenêtres soit de l'ordre de 5 millisecondes.

2.4.3.7 Conclusion

Cette fonction d'observation a été utilisée dans le but de mettre en évidence avant tout les variations du contenu fréquentiel. Plusieurs bandes de fréquence seraient à utiliser pour deux raisons. D'abord, une évolution lente de l'énergie, sans changement de hauteur, agit surtout sur les hautes fréquences : les harmoniques de numéros d'ordre élevés sortent du bruit ou disparaissent dans le bruit. Ensuite, à énergie et à hauteur restant constantes, le bruit, pour les harmoniques de numéros d'ordre élevés, d'amplitudes souvent petites, est important, ce qui rend le flux spectral inutilisable pour ces fréquences : c'est-à-dire que nous obtenons les mêmes valeurs pour le flux spectral dans les zones stables et aux moments des transitions. Cette raison explique que nous ayons considéré les enveloppes spectrales.

2.5 Autres fonctions d'observation

2.5.1 Dérivées du centroïde

Le centroïde est défini dans la section 2.5.5 de cette partie et dans la section 24.2.2.2 de la partie V. Le centroïde est utilisé aussi dans la *méthode de* MASRI (voir la section 2.5.5) et la *méthode de* HAJDA (voir la section 2.5.6).

La fonction d'observation est $|dC|$ la « valeur absolue de la dérivée du centroïde ». La taille des fenêtres d'analyse t_{SIG} et la fenêtre de pondération sont les paramètres libres. La méthode est implémentée seulement sous MATLAB.

2.5.2 La rupture de modèles directement sur le son

Il s'agit d'appliquer la même méthode que dans les sections 2.2.6 et 2.3.2, mais cette fois directement sur le signal sonore. Les paramètres libres sont les mêmes que ceux décrits dans la section 2.2.6. Cette méthode est utilisée pour la parole par [BB82] et [BN93], et pour la musique par [LO95]. Nous essayons ici de détecter directement les variations brusques du signal. La fonction d'observation est appelée RM_S.

2.5.3 Le test de BRANDT

2.5.3.1 Méthode

L'algorithme (voir [vB83] et [Jeh97]) est basé sur l'étude de trois modèles AR calculés pour trois portions F_1, F_2 et F_3 du signal sonore. Leurs tailles sont fixes. Elles sont simultanément glissantes. Le degré de dissemblance entre ces trois modèles est mesuré. Cette mesure, nous donnant des pics au moment des transitions et un bruit dans les zones stables, est une fonction d'observation. Les trois portions sont disposées comme il est indiqué sur la figure 2.16.

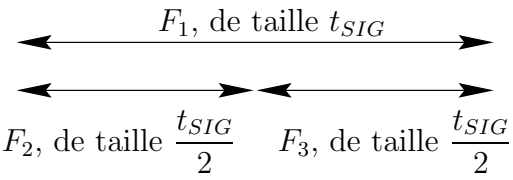


FIG. 2.16 – *Disposition des trois fenêtres d'analyse pour le test de* BRANDT

Un modèle AR d'ordre P est calculé pour chacune de ces portions. Nous obtenons les variances des trois erreurs de prédiction (des trois bruits blancs générateurs) : σ_1^2, σ_2^2 et σ_3^2. La fonction d'observation est alors le test de BRANDT :

$$BRANDT(t) = \left| t_{SIG} \log_e \sigma_1 - \frac{t_{SIG}}{2} \log_e \sigma_2 - \frac{t_{SIG}}{2} \log_e \sigma_3 \right|$$

t est l'instant central de la fenêtre F_1. Le moment de dissemblance maximale est obtenu quand la portion F_2 couvre la fin d'une note et la portion F_3 le début de la note suivante.

2.5.3.2 Paramètres libres

Ils sont au nombre de 2. Ce sont : t_{SIG}, la taille de la portion F_1 ; P, l'ordre des modèles AR. En général, P est pris petit (quelques unités).

2.5.4 L'analyse de la stationnarité

2.5.4.1 Introduction

D'un point de vue statistique, un signal « stable » peut être défini comme étant un signal stationnaire (au second ordre) et ergodique. Un processus aléatoire $X(t)$ (signal analogique) est stationnaire au second ordre s'il vérifie :

- $E[X(t)]$ indépendant de t (I)
- $E\left[X^2(t)\right] < +\infty$, et indépendant de t (II)
- $R^{xx}(t_1,t_2) = E[X(t_1)X(t_2)]$ fonction uniquement de $\tau = t_2 - t_1$ (III)

De la même façon, pour une suite aléatoire X_n (signal échantillonné) nous avons :

- $E[X_n]$ indépendant de n (I)
- $E\left[X_n^2\right] < +\infty$, et indépendant de n (II)
- $R^{xx}(n_1,n_2) = E[X_{n_1}X_{n_2}]$ fonction uniquement de $\eta = n_2 - n_1$ (III)

Et le processus ou la suite aléatoire est ergodique (à l'ordre i) si ses moyennes temporelles (jusqu'à l'ordre i) sont indépendantes du choix de la réalisation (IV).

Nous considérons deux portions du signal de taille t_{SIG} et décalées de η échantillons : ce sont nos deux observations. La somme de sinusoïdes et d'un bruit normal est un processus stationnaire au second ordre et ergodique au moins jusqu'à l'ordre 2. Soit le cas d'un signal constitué d'une somme de N sinusoïdes de fréquences f_i et d'amplitudes a_i (aucun bruit n'est présent) :

$$X_n = \sum_{i=1}^{N} a_i \cos\left(2\pi f_i \frac{n}{f_e} + \phi_i\right)$$

Nous obtenons alors :

$$R^{xx}(n_1,n_1+\eta) = R_{n_1}^{xx}(\eta) = E[X_{n_1}X_{n_1+\eta}] = \frac{1}{2}\sum_{i=1}^{N} a_i^2 \cos\left(2\pi f_i \frac{\eta}{f_e}\right) \quad \forall\eta$$

2.5.4.2 La fonction d'observation

La première idée est de considérer les coefficients d'auto-corrélation pour différents η et de fusionner les résultats. La fonction d'observation à la base est donc multidimensionnelle, avec tous les problèmes de fusion de données qui se posent alors (voir la section 2.6). Il est nécessaire de considérer plusieurs η. En effet, il peut fort bien arriver que pour certains η nous ayons :

$$\sum_{i=1}^{N} a_i^2 \cos\left(2\pi f_i^{(a)} \frac{\eta}{f_e}\right) = \sum_{i=1}^{M} b_i^2 \cos\left(2\pi f_i^{(b)} \frac{\eta}{f_e}\right)$$

où les $f_i^{(a)}$ et les a_i correspondent aux fréquences et aux amplitudes des N composantes sinusoïdales présentes dans une zone stable et les $f_i^{(b)}$ et les b_i aux fréquences et aux amplitudes des M composantes sinusoïdales présentes dans la zone stable juste adjacente à la première. Dans ce cas la transition n'est pas détectée.

La fonction d'observation choisie est la « moyenne arithmétique des valeurs absolues des dérivées numériques des coefficients d'auto-corrélation », c'est-à-dire :

$$CORR_{\Theta}(n) = \sum_{i=1}^{A} \left(\left|R_{n+1}^{xx}(\eta_i) - R_{n-p}^{xx}(\eta_i)\right|\right)$$

où $\Theta = [\eta_1 \ \ldots \ \eta_A]$ est un jeu de décalages temporels à définir, n le temps et $p \in 0, 1$. Pour $\Theta = [0]$, nous obtenons ce que nous appelons la « dérivée de l'énergie » dans la section 2.3.2.

2.5.4.3 Les paramètres libres

Ils sont au nombre de 2. Ce sont : t_{SIG}, la taille des portions du signal sur lesquelles nous calculons les coefficients d'auto-corrélation ; Θ, le jeu de décalages temporels. Nous prenons en compte les petits coefficients d'auto-corrélation.

2.5.4.4 Conclusion

Les performances en présence d'un signal « perturbé » (par exemple par la présence d'un vibrato) sont très dégradées. Il faut aussi voir qu'en pratique t_{SIG} est très loin de tendre vers l'infini : souvent, nous utilisons, pour estimer les coefficients d'autocorrélation, des fenêtres d'analyse de seulement quelques centaines d'échantillons (T vaut quelques dizaines de millisecondes).

2.5.5 La méthode de Masri

Masri, dans [Mas96] et [MB96], définit une fonction d'observation sensible aux attaques (« onset ») de notes. Soit $S(i)$ un signal de taille t_{SIG} et $\hat{X}(k)$ sa transformée de Fourier, de taille t_{FFT} (k varie de $-\frac{t_{FFT}}{2} + 1$ à $-\frac{t_{FFT}}{2}$). Nous avons :

$$E = \sum_{k=1}^{\frac{t_{FFT}}{2}} \left\{\left|\hat{X}(k)\right|^2\right\}$$

(nous commençons à sommer à partir de $k = 1$, c'est-à-dire que nous ne prenons pas en compte le continu) ; et (« HFC : High Frequency Content », ou centroïde non normalisé en énergie) :

$$HFC = \sum_{k=1}^{\frac{t_{FFT}}{2}} \left\{\left|\hat{X}(k)\right|^2 k\right\}$$

Un changement soudain dans un signal introduit des discontinuités de phase, donc de l'énergie apparaît dans les hautes fréquences. Voilà pourquoi nous les favorisons dans HFC.

Le signal $S(i)$ est une portion du signal sonore, portion multipliée par une fenêtre de pondération. Cette portion est glissante : le pas de déplacement est de Q échantillons. Nous indiçons chaque portion temporellement par r.

Ainsi, la fonction d'observation FOM (pour Fonction d'Observation de MASRI) s'écrit :

$$FOM(r) = \frac{HFC(r)}{HFC(r-1)} \frac{HFC(r)}{E(r)}$$

Il s'agit à présent de la seuiller. Le problème est le même que pour les autres fonctions d'observation : comment choisir automatiquement le seuil de détection ? MASRI ne répond pas à cette question : il seuille à la main. Nous donnons des solutions à ce problème dans le chapitre 3 et l'annexe B. Nous tentons ici de détecter simultanément les variations de l'énergie et les variations du contenu spectral. Cette fonction d'observation n'a pas été implémentée, ni sous MATLAB, ni en C. Les paramètres libres sont t_{SIG} et Q. t_{SIG} est telle que la taille de la fenêtre d'analyse vaut quelques dizaines de millisecondes ; et Q est tel que le pas de déplacement vaut quelques millisecondes.

2.5.6 Énergie et Centroïde – HAJDA

HAJDA, dans [Haj96], utilise lui aussi HFC (qu'il appelle le centroïde) et l'énergie (calculée dans le domaine temporel, et non pas comme pour la méthode de MASRI dans le domaine fréquentiel). Le but ici aussi est de simultanément détecter les variations de l'énergie et les variations du contenu spectral. Chaque note est découpée en quatre zones, qui se succèdent : l'attaque, la transition, l'état stable et la chute. Le comportement (croissance ou décroissance) de l'énergie et du centroïde est observé pour décider dans quelle zone nous sommes. Ainsi, HAJDA donne (et explique), le tableau 2.1.

	énergie	centroïde
attaque	croissante	décroissant
transition	croissante	croissant
état stable	indifférent	indifférent
chute	décroissante	décroissant (croissant)

TAB. 2.1 – *Utilisation du centroïde et de l'énergie dans la méthode de* HAJDA *pour segmenter « plus petit que la note »*

Les étiquettes qui sont données ici (attaque, transition, état stable, chute) complètent la liste d'étiquettes donnée dans le chapitre 5 de cette partie. Cette fonction d'observation n'a pas été implémentée, ni sous MATLAB, ni en C.

2.5.7 La méthode de SMITH

SMITH, dans l'article [Smi94], filtre un signal S, obtenu à partir du signal sonore s original en prenant en compte certaines caractéristiques de la cochlée (filtrage passe-bande, les fréquences centrales des filtres étant logarithmiquement placées (échelle Mels) ; redressement de chaque signal obtenu après ce filtrage passe-bande ; sommation des signaux redressés), par des filtres basés sur des idées appliquées en traitement de l'image (voir [MH80]).

Ces filtres sont de la forme :

$$f_E(x,k) = k \exp(-kx) \text{ } (DoE\text{: pour « Difference of Exponentials Filter »})$$

ou :

$$f_G(x,k) = \sqrt{k} \exp(-kx^2) \text{ } (HDoG\text{: pour « Half Difference of Gaussians »})$$

Et le signal à segmenter est alors (différence entre une moyenne sur un temps court et une moyenne sur un temps long) :

$$O(t,k,r) = \int_0^t \left[f_z(t-x,k) - f_z\left(t-x, \frac{k}{r}\right) \right] S(x)dx$$

z valant E ou G.

SMITH appelle les fonctions $O(t,k,r)$ les opérateurs « *onset/offset* ».

Le problème réside toujours dans la difficulté rencontrée pour seuiller et dans la détermination des paramètres libres k, r, t et z. Cette fonction d'observation n'a pas été implémentée, ni sous MATLAB, ni en C.

2.6 La fusion de données

2.6.1 Problématique de la fusion de données

Au sujet de la fusion de données, voir, pour une introduction, le livre de VARSHNEY [Var96]. Le problème peut se formuler ainsi : plusieurs capteurs imparfaits tentent de détecter le même phénomène ; il faut rassembler leurs résultats, ceci dans l'objectif d'obtenir de meilleurs performances que si un seul des capteurs était utilisé.

Deux types de relations entre les capteurs dont nous voulons fusionner les données sont considérées dans cet exposé :

- Les capteurs nous offrent des données homogènes. Ceci a lieu quand nous considérons les fonctions d'observation multidimensionnelles : ce cas concerne donc ce chapitre. Alors, nous pouvons simplement moyenner les données, c'est-à-dire ici les dimensions des fonctions d'observation.
  Faisons l'hypothèse que x_1 ... x_N soient N variables aléatoires normales indépendantes $\mathcal{N}\left(0,\sigma_x^2\right)$ de même moyenne et de même variance. L'écart-type de $y = \dfrac{1}{N}\sum_{i=1}^{N} x_i$ est $\sigma_y = \dfrac{\sigma_x}{\sqrt{N}}$.
  Soient pour une dimension d'une fonction d'observation particulière nos deux modes : le premier, correspondant au bruit, suit une loi normale $\mathcal{N}\left(0,\sigma_{(1)}^2\right)$; le second, correspondant aux pics à détecter, suit une loi normale $\mathcal{N}\left(m,\sigma_{(2)}^2\right)$. Ces densités de probabilité sont les mêmes pour les N dimensions de la fonction d'observation considérée. Après moyennage des dimensions, les densités de probabilité des deux modes sont respectivement $\mathcal{N}\left(0,\dfrac{\sigma_{(1)}^2}{N}\right)$ et $\mathcal{N}\left(m,\dfrac{\sigma_{(2)}^2}{N}\right)$. Ce moyennage a éloigné les deux modes l'un de l'autre : voir la section B.21.5, où nous montrons que plus la distance entre les deux modes est grande plus le seuillage automatique (ceci concerne la *deuxième étape* de l'analyse *segmentation en zones stables* : voir le chapitre 3) est efficace.
  La fusion de données homogènes correspond à la **fusion d'observations**.
- Les capteurs nous offrent des données hétérogènes.
  Si les dimensions de la fonction d'observation, ou si les fonctions d'observation, ne suivent pas les mêmes densités de probabilité, le raisonnement du précédent paragraphe ne tient plus : nous sommes en présence de données hétérogènes.
  Supposons que nous ayons trois variables aléatoires normales indépendantes a, b et c. Considérons trois cas.

  1° $p_a(a) = p_b(b) = p_c(c) = \mathcal{N}(0,1)$. La variance de la moyenne est $\dfrac{1}{3} = 0{,}33$.

  2° $p_a(a) = p_b(b) = p_c(c) = \mathcal{N}(0,7)$. La variance de la moyenne est $\dfrac{7}{3} = 2{,}33$.

  3° $p_a(a) = \mathcal{N}(0,1)$, $p_b(b) = \mathcal{N}(0,4)$ et $p_c(c) = \mathcal{N}(0,7)$. La variance de la moyenne est $\dfrac{12}{9} = 1{,}33$, c'est-à-dire supérieure à celle de la variable aléatoire a. Ainsi, ici, la fusion des observations dégrade les performances du système (utiliser un seul capteur, a, donne de meilleurs performances qu'en utiliser trois).

Aussi, nous devons d'abord prendre les décisions pour chaque dimension de la fonction d'observation ou chaque fonction d'observation, et fusionner les fonctions de décision obtenues. Ceci a lieu lors de la *troisième étape* de l'analyse *segmentation en zones stables* (voir le chapitre 4), c'est-à-dire quand nous confrontons les fonctions de décision, quand il y a compétition entre elles. La fusion de données hétérogènes correspond à la **fusion de décisions**.

Nous allons, dans la suite de l'exposé, faire l'hypothèse que les dimensions des fonctions d'observation multidimensionnelles correspondent à des données homogènes, alors que deux fonctions d'observation différentes correspondent à des données hétérogènes.

2.6.2 Fusion dans le cas de données homogènes – Fusion de données pour les fonctions d'observation multidimensionnelles

Nous avons présenté dans les sections précédentes des fonctions d'observation multidimensionnelles : les « valeurs absolues des dérivées relatives des trajets des harmoniques du signal », les « valeurs absolues des dérivées des indices d'inharmonicité »... Le problème est que nous voulons réduire chaque fonction d'observation multidimensionnelle à une fonction d'observation unidimensionnelle. Nous avons fait l'hypothèse que les données des dimensions de chaque fonction d'observation multidimensionnelle sont homogènes. La méthode la plus simple pour réduire le nombre de dimensions est de les moyenner. Les moyennes qui peuvent être considérées sont (il s'agit des moyennes classiques) :

- La moyenne arithmétique : $V_a(i) = \frac{1}{L}\sum_{l=1}^{L} v_l(i)$, où $v_l(i)$ est la fonction d'observation pour sa dimension l à l'instant i. Cette moyenne est celle qui est utilisée le plus communément.
- La moyenne géométrique : $V_g(i) = \left(\prod_{l=1}^{L} v_l(i)\right)^{\frac{1}{L}}$.
- La moyenne harmonique : $V_h(i) = \left(\frac{1}{L}\sum_{l=1}^{L} \frac{1}{v_l(i)}\right)^{-1}$.

La moyenne arithmétique est implémentée dans le programme *segmentation* pour toutes les fonctions d'observation multidimensionnelles. La moyenne géométrique (sans la racine d'ordre L) est implémentée dans le programme *segmentation* pour fusionner les valeurs absolues des dérivées des indices d'inharmonicité.

Dans les moyennes présentées ci-dessus, les dimensions ont toutes le même poids. Il pourrait être utile de moins prendre en compte celles qui concernent les hautes fréquences, de pondérer chacune d'elle par l'amplitude de l'harmonique qui la concerne, ou par l'énergie de la bande de fréquence qui la concerne... Il s'agit de perspectives.

2.6.3 Fusion dans le cas de données hétérogènes – Capteurs décorrélés

Cette section concerne la fusion des fonctions de décision.

2.6.3.1 Capteurs identiques

Les quatre décisions possibles sont O^O (le capteur a détecté quelque chose et il y avait quelque chose à détecter) ou N^O (le capteur n'a rien détecté alors qu'il y avait quelque chose à détecter) ; et O^N (le capteur a détecté quelque chose alors qu'il n'y avait rien à détecter) ou N^N (le capteur n'a rien détecté et il n'y avait rien à détecter).

Faisons l'hypothèse que nous ayons trois capteurs identiques (« identiques » veut dire qu'ils ont la même probabilité de bonne détection : p_{bd} et la même probabilité de fausse alarme : p_{fa}). Supposons que la probabilité de bonne détection p_{bd} (c'est-à-dire : le capteur détecte quelque chose et il y a quelque chose à détecter) pour chacun d'eux soit égale à 0,8 (cas 1). Supposons que la probabilité de fausse alarme (c'est-à-dire : le capteur détecte quelque chose alors qu'il n'y a rien à détecter) pour chacun d'eux soit égale à 0,3 (cas 2).

Remarquons que les deux cas sont semblables, si nous formulons le second ainsi : « supposons que la probabilité de bonne détection (c'est-à-dire : le capteur ne détecte rien et il n'y a rien à détecter) pour chacun d'eux soit égale à 0,7 ». Cela indique qu'il y a deux sortes de bonnes détections.

Considérons le premier cas : il y a quelque chose à détecter. Nous décidons qu'il y a quelque chose si au moins deux capteurs nous indiquent qu'ils ont détecté quelque chose (c'est-à-dire quand le poids du O^O est supérieur au poids du N^O : il s'agit de la règle de la majorité). Nous donnons dans le tableau 2.2 les diverses configurations possibles (les trois capteurs ont détectés quelque chose, deux des capteurs ont détecté quelque chose, etc.), la probabilité d'apparition de chacune de ces configurations, le poids du O^O et du N^O pour chaque configuration et la décision finale pour chaque configuration. Grâce à la fusion des fonctions de décisions de ces trois capteurs, la probabilité de bonne détection (c'est-à-dire : on détecte quelque chose et il y a quelque chose à détecter) passe de 0,8 à 0,896 (somme des probabilités des configurations I, II, III et V).

configuration	capt. 1	capt. 2	capt. 3	probabilité	poids O	poids N	décision finale
I	O^O	O^O	O^O	0,512	3	0	O
II	O^O	O^O	N^O	0,128	2	1	O
III	O^O	N^O	O^O	0,128	2	1	O
IV	O^O	N^O	N^O	0,032	1	2	N
V	N^O	O^O	O^O	0,128	2	1	O
VI	N^O	O^O	N^O	0,032	1	2	N
VII	N^O	N^O	O^O	0,032	1	2	N
VIII	N^O	N^O	N^O	0,008	0	3	N

TAB. 2.2 – *Fusion des décisions fournies par trois capteurs identiques et non corrélés*

Considérons le second cas : il n'y a rien à détecter. Nous travaillons toujours avec les trois mêmes capteurs. Grâce à la fusion des fonctions de décisions de ces trois capteurs, la probabilité de fausse alarme (c'est-à-dire : quelque chose est détecté alors qu'il n'y a rien à détecter) passe de 0,3 à 0,216.

2.6.3.2 Capteurs non identiques

Ci-dessus nous avons considéré trois capteurs identiques (mêmes probabilités de bonne détection et mêmes probabilités de fausse alarme). Clairement, la fusion des fonctions de décisions de ces trois capteurs permet d'obtenir une plus grande probabilité de bonne détection et une plus petite probabilité de fausse alarme.

Quand les capteurs ne sont pas identiques, l'apport de la fusion des fonctions de décisions devient incertain. Nous donnons quelques exemples dans le tableau 2.3. Nous constatons que la fusion de données n'est pas efficace pour tous les exemples (la fusion de données est inefficace quand l'un des capteurs a une p_{bd} supérieure à la p_{bd} finale).

	p_{bd} du capteur 1	p_{bd} du capteur 2	p_{bd} du capteur 3	p_{bd} finale	fusion efficace ?
cas 1	0,80	0,75	0,70	0,845	oui
cas 2	0,85	0,75	0,70	0,865	oui
cas 3	0,90	0,75	0,70	0,885	non
cas 4	0,80	0,80	0,70	0,864	oui
cas 5	0,80	0,80	0,60	0,832	oui
cas 6	0,80	0,80	0,50	0,800	–
cas 7	0,80	0,80	0,40	0,768	non

TAB. 2.3 – *Fusion des décisions fournies par trois capteurs non identiques (plusieurs cas sont considérés) et non corrélés*

Dans deux cas la fusion de données est inefficace. Pour le cas 3, ceci est dû à ce que l'un des capteurs a des performances très supérieures à celles des deux autres. Pour le cas 7, ceci est dû à ce que l'un des capteurs a des performances très inférieures à celles des deux autres.

Ces exemples nous montrent qu'il faut utiliser la fusion de décisions avec prudence. Il faut d'abord s'assurer que les fonctions de décision obtenues aient des performances sensiblement identiques.

Nous pourrions ajouter des poids sur les capteurs : nous faisons plus confiance en tel ou tel capteur. Ceci ne ferait qu'augmenter la difficulté du processus de fusion.

2.6.4 Fusion dans le cas de données hétérogènes – Deux capteurs sont corrélés

Nous considérons dans cette section que nous avons 5 capteurs. Deux cas sont étudiés : ou bien deux des capteurs sont absolument corrélés (fonctions de décision identiques), ou bien ils sont tous décorrélés. Nous donnons dans le tableau 2.4 la p_{bd} finale pour plusieurs configurations des p_{bd} des capteurs, ceci suivant que deux des capteurs sont absolument corrélés (dans ce cas, ce sont les capteurs 1 et 2 qui sont corrélés) ou décorrélés.

	2 capteurs corrélés ?	p_{bd} 1	p_{bd} 2	p_{bd} 3	p_{bd} 4	p_{bd} 5	p_{bd} finale
cas 1	non	0,80	0,80	0,80	0,80	0,80	0,942
	oui	0,80	0,80	0,80	0,80	0,80	0,896
cas 2	non	0,80	0,80	0,70	0,70	0,70	0,887
	oui	0,80	0,80	0,70	0,70	0,70	0,847
cas 3	non	0,70	0,70	0,80	0,80	0,80	0,908
	oui	0,70	0,70	0,80	0,80	0,80	0,848

TAB. 2.4 – *Fusion des décisions fournies par cinq capteurs identiques ou non et corrélés ou non*

Ces exemples nous montrent que l'utilisation de fonctions de décision égales (capteurs absolument corrélés) fait que les performances de la fusion de données sont moins bonnes qu'espéré. Il faut d'abord s'assurer que les capteurs soient décorrélés.

2.6.5 Fusion dans le cas de données hétérogènes – Capteurs corrélés

Nous considérons dans cette section que nous avons n capteurs. La probabilité de bonne détection pour chaque capteur est $p_{bd} = 0{,}8$ (capteurs identiques). Nous étudions la probabilité de bonne détection après la fusion de données. Plusieurs cas sont considérés :

1° Tous les capteurs sont décorrélés. Alors, bien sûr, la probabilité de bonne détection augmente avec n.

2° Deux capteurs sont absolument corrélés. Remarquons que la probabilité de bonne détection obtenue pour i capteurs dans ce cas est exactement égale à la probabilité de bonne détection obtenue pour $i-2$ capteurs dans le cas précédent (1°) .

3° Deux capteurs sont absolument corrélés et deux autres capteurs sont absolument corrélés.

4° Trois capteurs sont absolument corrélés.

La probabilité de bonne détection après la fusion de décisions, pour ces quatre cas, est donnée en fonction de n sur la figure 2.17.

Ainsi nous pouvons voir qu'il est important d'utiliser des fonctions d'observation décorrélées.

2.7 Recensement des fonctions d'observation étudiées

Le recensement des fonctions d'observation utilisables pour la *segmentation en zones stables* d'un signal monophonique, harmonique et non modulé (c'est-à-dire pour la *segmentation en notes*

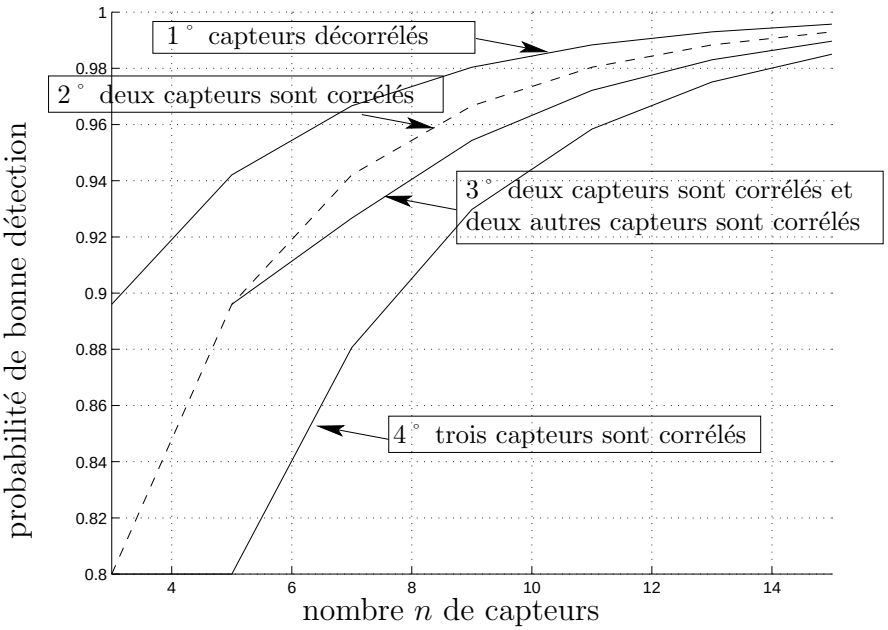


FIG. 2.17 – *Probabilité de bonne détection en fonction du nombre de capteurs et des corrélations entre eux*

et/ou en phones) est fait dans le tableau 2.5. Nous donnons dans ce tableau pour chaque fonction d'observation à quel endroit de cet exposé elle est décrite ; son nombre de dimensions ; son degré d'indépendance par rapport aux programmes de l'IRCAM ; si elle est implémentée dans le programme *segmentation* ; et son nom symbolique. Les performances de ces fonctions d'observation sont évaluées sur des bases de données qui sont présentées dans la section 6.2, ainsi que sur des signaux simulés.

2.8 Conclusion

Nous avons défini un processus de segmentation, décomposé en quatre étapes. Nous venons de décrire le première : l'extraction de fonctions d'observation. Tel est ce processus de segmentation que nous pouvons toujours implémenter d'autres fonctions d'observation et aisément les y intégrer. Ainsi, un certain nombre de fonctions d'observation ont été définies par différents auteurs, que nous n'avons pas présentées ici. La méthode d'ALKULAIBI (voir la référence [ASD97]) nous donne une fonction d'observation basée sur les moments d'ordres supérieurs (« Higher Order Moments : HOM »). La méthode de SOLBACH (voir la référence [SW96]) nous donne une fonction d'observation basée sur les ondelettes. Dans [FCM+96] et [JW88] sont définies des fonctions d'observation basées sur les coefficients cepstraux. Etc.

Seulement, quelques remarques sont à faire à propos de l'intégration d'éventuelles nouvelles fonctions d'observation :

- Il serait intéressant que, comme la grande majorité de celles qui ont déjà été étudiées, elles eussent pour principe de présenter des pics aussi fins et aussi grands que possible au moment des transitions et un bruit de moyenne et de variance aussi petites que possible pendant les parties stables du signal.
- Il existe énormément de fonctions d'observation, mais leurs performances sont inégales. Ces performances dépendent du signal à traiter. Les fonctions d'observation ne sont pas universelles : pour certains sons, elles échouent. Ou bien elles ne détectent pas certaines variations brusques du signal, ou bien certaines caractéristiques du signal les rendent inutilisables. D'où la nécessité d'un deuxième niveau de segmentation, qui doit nous permettre de détecter pour

principe	section	dimension	dépendante	implémentée	nom symbolique
dérivées de f_0	2.2.2	1 ou $\frac{f_e}{2f_0}$	oui	oui	$\lvert df_0^I \rvert$, $\lvert df_0^{II} \rvert$, $\lvert \delta f_0^I \rvert$, $\lvert \delta f_0^{II} \rvert$
dérivées du voisement (forme 1)	2.2.3	$\frac{f_e}{2f_0}$	oui	oui	$\sum_l \lvert dV_l^A \rvert$
dérivées des inharmonicités	2.2.4	$\frac{f_e}{2f_0}$	oui	oui	$\sum_l \lvert dH_l^n \rvert$, $\prod_l \lvert dH_l^n \rvert$
analyse statistique sur f_0	2.2.5	1 ou $\frac{f_e}{2f_0}$	oui	oui	$AS_{f_0}^I$, $AS_{f_0}^{II}$
rupture de modèles sur f_0	2.2.6	1 ou $\frac{f_e}{2f_0}$	oui	oui	RM_{f_0}
f_0 après filtrage de HILBERT	14.5	1 ou $\frac{f_e}{2f_0}$	oui	non	$\lvert X \rvert$
dérivées de l'énergie	2.3.2	1	non	oui	$\lvert dE^I \rvert$, $\lvert dE^{II} \rvert$, $\lvert \delta E^I \rvert$, $\lvert \delta E^{II} \rvert$
analyse statistique sur l'énergie	2.3.2	1	non	non	AS_E^I, AS_E^{II}
rupture de modèles sur l'énergie	2.3.2	1	non	non	RM_E
dérivées du voisement (formes 2)	2.4.1, 2.4.2	1 ou B bandes de fréquences	non	non	$\lvert dV^{(R)} \rvert$, $\lvert dV^{(C)} \rvert$
flux entre spectres	2.4.3.2	1 ou B bandes de fréquences	non	oui	F
flux enveloppe AR - spectre	2.4.3.3	1 ou B bandes de fréquences	non	non	$\lvert dF_1^{AR} \rvert$
flux entre enveloppes AR	2.4.3.3	1 ou B bandes de fréquences	non	oui	F_2^{AR}
flux enveloppe cepstre - spectre	2.4.3.4	1 ou B bandes de fréquences	non	non	$\lvert dF_1^{ceps} \rvert$
flux entre enveloppes cepstres	2.4.3.4	1 ou B bandes de fréquences	non	non	F_2^{ceps}
flux enveloppe maximums - spectre	12.3.9.1	1 ou B bandes de fréquences	non	non	$\lvert dF_1^{max} \rvert$
flux entre enveloppes maximums	12.3.9.1	1 ou B bandes de fréquences	non	non	F_2^{max}
flux entre enveloppes superposées	12.3.7.2, 12.3.9.2	1 ou B bandes de fréquences	non	non	F_3^{max}
dérivées du centroïde	2.5.1	1	non	non	$\lvert dC \rvert$
rupture de modèles sur le signal	2.5.2	1	non	non	RM_S
test de BRANDT	2.5.3	1	non	non	$BRANDT$
analyse de la stationnarité	2.5.4	A η considérés	non	non	$CORR_\Theta$
méthode de MASRI	2.5.5	1	non	non	FOM
méthode de HAJDA	2.5.6	1	non	non	C et E
méthode de SMITH	2.5.7	1	non	non	O

TAB. 2.5 – *Recensement des fonctions d'observation étudiées*

chaque son considéré quelles fonctions d'observation sont utilisables pour le *segmentation en zones stables*. Trois types de généralisation sont envisagés :

- Le signal n'est pas harmonique.
- Le signal est modulé en fréquence et/ou en amplitude.
- Le signal n'est pas monophonique.

Ceci sera discuté plus en détail dans les chapitres 6 et 8, puis dans les autres parties de l'exposé.

Lors du calcul de certaines fonctions d'observation, le programme *segmentation* offre la possibilité de tenir compte de la fonction d'audition de l'oreille (voir l'annexe E). Ainsi, ces fonctions d'observation peuvent aussi bien être considérées d'un point de vue purement traitement du signal, que d'un point de vue plus perceptif. Pour les flux spectraux, cette prise en compte de la fonction d'audition de l'oreille se traduit par le calcul du produit de la valeur de chaque échantillon fréquentiel des spectres d'amplitude ou des enveloppes spectrales par la valeur de la fonction d'audition de l'oreille à la fréquence pour cet échantillon fréquentiel. Pour les indices de voisement et pour les indices d'inharmonicité, cette prise en compte de la fonction d'audition de l'oreille se traduit par le calcul du produit de chaque indice par la valeur de la fonction d'audition de l'oreille à la fréquence du partiel qui nous donne cet indice.

Chapitre 3

Prise de décision pour chacune des fonctions d'observation

3.1 Introduction

Les fonctions d'observation implémentées dans le programme *segmentation* sont au nombre de dix. Il s'agit de :

- la « valeur absolue de la dérivée de la fréquence fondamentale »
- la « valeur absolue de la dérivée relative de la fréquence fondamentale »
- la « valeur absolue de la dérivée de l'énergie »
- la « valeur absolue de la dérivée relative de l'énergie »
- la « somme (moyenne arithmétique) des valeurs absolues des dérivées des indices d'inharmonicité »
- le « produit (moyenne géométrique) des valeurs absolues des dérivées des indices d'inharmonicité »
- la « somme des valeurs absolues des dérivées des indices de voisement première forme »
- le « flux spectral, calculé soit à partir de deux spectres d'amplitude, soit à partir de deux enveloppes spectrales du type AR, la fonction d'atténuation de l'oreille étant prise en compte ou non »
- l'« écart entre deux modèles statistiques, méthode utilisée sur le trajet de f_0 »
- la « détection de rupture de modèles utilisant la modélisation auto-régressive, méthode utilisée elle aussi sur le trajet de f_0 »

Seule la dernière fonction d'observation est basée sur un algorithme de décision. Pour les autres, il nous faut discriminer les pics qui ont un sens, c'est-à-dire ceux qui correspondent à une transition, de ceux qui correspondent à du bruit, c'est-à-dire qui sont présents dans les parties stables du signal. Ceci est la *deuxième étape* de l'analyse *segmentation en zones stables*.

Dans la **deuxième section** (section 3.2) de ce chapitre, est présentée la méthode de seuillage qui a été retenue.

Dans la **troisième section** (section 3.3) de ce chapitre, nous discutons de la normalisation des fonctions d'observation, cette normalisation ayant pour but d'éviter le seuillage.

Nous donnons une conclusion dans la **quatrième section** (section 3.4) de ce chapitre. Nous donnons aussi quelques perspectives.

3.2 Seuillages

3.2.1 Calcul automatique de la valeur des seuils

Chaque fonction d'observation correspond à une variable aléatoire dont les échantillons se répartissent en deux classes : une qui correspond à du bruit (dans les zones stables) ; l'autre aux pics à détecter (aux moments des transitions). Il s'agit de discriminer ces deux classes, c'est-à-dire de poser un seuil. Il existe un grand nombre de méthodes pour seuiller automatiquement. Elles sont présentées, non exhaustivement, dans l'annexe B. Leurs qualités et leurs défauts respectifs sont discutées dans cette même annexe. Elles viennent principalement du traitement des images : voir notamment [SSW88].

Dans notre cas, les pics dus aux transitions sont très rares et leur variance est grande. Nous montrons dans l'annexe B que les performances des méthodes de seuillage sont énormément dégradées quand la probabilité a priori d'une classe est très petite et quand la variance de cette même classe est très supérieure à la variance de l'autre classe.

La méthode implémentée dans le programme *segmentation* est la méthode des 3σ. Si nous considérons que le bruit suit une loi normale, nous pouvons estimer σ, l'écart-type, en retenant par exemple $n = 90\%$ (la méthode est relativement robuste à la valeur arbitrairement fixée n, comme il est indiqué dans l'annexe B, sur la figure B.56) des plus petits échantillons de la fonction d'observation. Alors, la valeur du seuil est 3σ. C'est-à-dire qu'il est décidé que les pics plus grands que 3σ sont des pics correspondant à des transitions.

Il y a une justification théorique à ce seuil, avec le « seuil universel » (« universal threshold »), défini par DONOHO dans [Don94] et [DJ94]. Ce seuil est égal à $\sigma\sqrt{2\log_e(M)}$, où M est la taille du signal en nombre d'échantillons. DONOHO prouve que quand M tend vers l'infini la probabilité qu'un échantillon de bruit dépasse ce seuil tend vers 0. Et $\sigma\sqrt{2\log_e(M)}$, pour un M de valeur « raisonnable » pour nous, c'est-à-dire de l'ordre de 1000 [1], par exemple, est très proche de 3 (3,72).

Le seuillage nous donne pour chaque *fonction d'observation* une *fonction de décision* qui vaut 0 (nous parcourons une zone stable) ou 1 (une transition a été détectée).

3.2.2 Quelques remarques

Actualisation des seuils Le seuillage est local : les seuils sont calculés, pour chacune des fonctions d'observation, sur des portions glissantes larges de n secondes. n est un paramètre libre. Dans l'avenir (il s'agit d'une perspective), nous actualiserons les seuils toutes les 10 secondes, ou nous déterminerons automatiquement quand il doit l'être (grâce au niveau de *segmentation en caractéristiques* : voir la partie III).

Cette actualisation automatique des seuils n'a pas été implémentée dans le programme *segmentation*. Ce programme a été testé pour des sons dont la longueur est de l'ordre de la dizaine ou, au plus, de la vingtaine de secondes (voir le chapitre 6).

Contrôle des seuils Dans le programme *segmentation*, l'utilisateur peut modifier la valeur des seuils à la main.

3.3 Normalisation des fonctions d'observation

Les « flux spectraux » (voir la section 2.4.3) sont calculés à partir de deux « spectres » (deux spectres d'amplitude, ou deux enveloppes spectrales) normalisés en énergie, c'est-à-dire que l'intégrale de chacun d'eux entre $f = 0$ et $f = \frac{f_e}{2}$ est égale à 1. Donc, la valeur maximale de la plupart des « flux spectraux » est de 2, valeur obtenue dans le cas extrême (et fort improbable) où les supports fréquentiels des deux « spectres » sont disjoints : voir la figure 3.1. Ainsi, la plupart

1. Les fonctions d'observation étant le plus souvent échantillonnées à 100 Hz, $M = 1000$ échantillons correspond à une réactualisation des seuils toutes les 10 secondes.

des « flux spectraux » sont déjà normalisés (entre 0 et 2). Ne le sont pas ceux basés sur un spectre d'amplitude et une enveloppe spectrale.

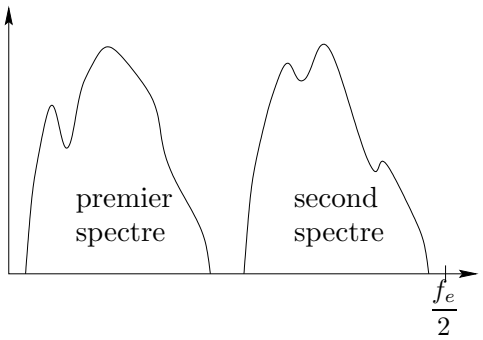


FIG. 3.1 – *Exemple de spectres d'amplitude à supports disjoints. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude des échantillons fréquentiels*

Le problème ne se pose pas pour les fonctions d'observation basées sur la rupture de modèles (voir la section 2.2.6), puisqu'en fait nous obtenons directement des fonctions de décision.

Les fonctions d'observations basées sur l'écart entre deux modèles statistiques (voir la section 2.2.5) sont elles aussi normalisées (entre 0 et 1), puisqu'elles correspondent à la mesure d'une probabilité.

Ce n'est pas le cas pour les autres fonctions d'observation, c'est-à-dire notamment pour toutes les valeurs absolues des dérivées et des dérivées relatives.

Par exemple, la valeur absolue de la dérivée à l'instant d'échantillonnage i de la fonction f est égale à :

$$|df(i)| = |f(i+a) - f(i-1)|$$

et la valeur absolue de la dérivée relative à :

$$|\delta f(i)| = \frac{|f(i+a) - f(i-1)|}{f(i)}$$

(a, ici, peut être égal à 0 ou à 1)

Nous définissons une autre dérivée, que nous appelons la dérivée normalisée. Elle est égale à :

$$|\Delta f(i)| = \frac{|f(i+a) - f(i-1)|}{max\,[f(i+a)\ f(i-1)]}$$

où l'opérateur max nous donne la plus grande valeur du tableau à deux éléments [. .].

Puisque les fonctions d'observations avant dérivation (trajet de f_0, trajet de l'énergie, indices d'inharmonicité, indices de voisement, centroïde, coefficients d'auto-corrélation...) sont toujours positives, nous sommes sûr que leur dérivée normalisée est toujours comprise entre 0 et 1. Le cas le plus défavorable est quand $f(i+a) = f(i-1) = 0$: alors, la valeur de $\Delta f(i)$ doit être imposée à 0. Donc, ces fonctions d'observation sont normalisables.

Le test de BRANDT ne donne pas une fonction d'observation normalisée. La fonction d'observation décrite dans la section 14.5 de la partie III n'est pas non plus normalisée.

3.4 Conclusion et perspectives

3.4.1 Perspectives

3.4.1.1 Seuiller ou ne pas seuiller

Pourquoi normaliser les fonctions d'observation ? Les méthodes de seuillage proposées dans l'annexe B sont adaptées à des variables aléatoires normales. Nous avons testé ces méthodes de

seuillage en supposant que nous étions en présence de telles variables aléatoires, ce qui ne correspond pas à la réalité. Le bruit suit plutôt une loi de RAYLEIGH et les pics dus aux transitions une loi normale. Ainsi, la prise décision automatique par seuillage automatique n'est pas forcément efficace pour notre cas. Il faudrait donc :

- ou bien adapter les méthodes de seuillage à nos données
- ou bien nous affranchir du seuillage automatique

Considérons la seconde solution :

Premièrement, cette normalisation nous donne la possibilité d'utiliser des classifieurs, comme les k plus proches voisins ($k$ppv ; ou kNN, pour « k Nearest Neighbours ») ou les réseaux de neurones (la normalisation est nécessaire aussi bien pour les $k$ppv que pour les réseaux de neurones : voir à ce sujet le rapport [Rap95]), pour prendre les décisions. Ceci pour chaque fonction d'observation, mais aussi pour la prise de décision finale (voir le chapitre 4), puisqu'elle travaille avec les *fonctions de décision*, qui par définition sont comprises entre 0 et 1. Il faut entraîner les classifieurs. Pour le faire correctement les réseaux de neurones, ainsi que les k plus proches voisins, il faut avant tout utiliser une grande base de sons.

Deuxièmement, si nous utilisons des fonctions d'observation normalisées, nous pouvons, pour chaque fonction d'observation, étudier la position optimale du seuil à appliquer. Cette position optimale est déterminée après avoir étudié le comportement de chaque fonction d'observation sur une base de sons conséquente. Ainsi, après cet entraînement, nous n'avons plus besoin de déterminer automatiquement un seuil. Cette méthode nous permettrait, après entraînement, de détecter les pics qui nous intéressent sans avoir à déterminer automatiquement la position d'un seuil et sans avoir à utiliser de classifieur.

Ainsi, d'autres méthodes que le seuillage automatique devront être testées (réseaux de neurones, k plus proches voisins...). Au sujet de la segmentation avec l'aide des réseaux de neurones, voir l'article [KHM96]. Nous avons vu qu'il existe beaucoup de fonctions d'observation. Avant d'intégrer d'autres fonctions d'observation dans le programme *segmentation*, nous nous attacherons à l'étude d'autres techniques de prises de décision automatiques. Indiquons-le de nouveau : il s'agit de perspectives.

Cependant, comme il a été dit, un réseau de neurones ou les $k$ppv doivent être entraînés, et l'un de nos objectifs est de construire un programme de segmentation le plus automatique possible. Aussi, le seuillage automatique est conservé.

3.4.1.2 Une fonction de coût

Nous présentons dans cette section une dernière perspective. JUNQUA et WAKITA, dans l'article [JW88], définissent une fonction de coût C pour chaque marque de segmentation trouvée, utilisable directement pour une grande partie de nos fonctions d'observation. Soient $val(i)$ la valeur de la fonction d'observation au moment i considéré, pour lequel nous avons détecté une marque de segmentation ; t_i la localisation temporelle de la marque ; t_{i-1} celle de la marque précédente ; $valmax$ la valeur maximale de cette fonction d'observation ; et min la valeur de la fonction d'observation dans le creux (minimum local le plus petit) précédent la marque en i. Nous avons alors :

$$C(i) = \frac{val}{valmax}\frac{val - min}{val + min}(t_i - t_{i-1})$$

Ainsi, les trop petits pics sont pénalisés (premier terme du produit) ; les pics pas assez prononcés sont rejetés (deuxième terme) ; et les pics trop proches sont pénalisés aussi (troisième terme). JUNQUA et WAKITA donnent une autre fonction de coût pour les « plateaux » (c'est-à-dire les « zones stables » pour nous), prenant en compte leurs longueurs : il ne faut pas qu'ils soient

trop longs. Mais, comme nous l'avons mentionné dans l'introduction, nous ne voulons pas faire d'hypothèse sur la longueur des notes. De plus, de nouveaux paramètres libres sont introduits, et il faudrait seuiller C.

3.4.2 Conclusion

Cependant, ce problème de prises de décisions automatiques est compliqué par le fait que les fonctions d'observation ne réagissent pas exactement aux mêmes moments, du fait que les transitions ne sont pas instantanées. Dans ce chapitre, nous avons considéré la prise de décisions d'un point de vue local : pour chaque instant d'échantillonnage i des fonctions d'observation, sans considérer ce qui a été obtenu aux instants voisins ; il faut relâcher cette contrainte. Ceci est discuté dans le chapitre suivant.

Chapitre 4

La fusion des résultats obtenus avec chaque fonction d'observation

4.1 Procédure pour fusionner : les objectifs

Un travail important, qui est fait ici, est lié encore à un problème de prise de décision : pour chaque fonction d'observation, nous avons vu que chaque critère de prise de décision nous donne un ensemble de marques : une *fonction de décision*. Il faut déterminer à présent, à partir de tous ces ensembles de marques, quelles marques sont à garder et en quels endroits il faut les positionner, avec une fonction de confiance pour chacune indiquant son degré de validité. Ceci constitue la *troisième étape* de l'analyse *segmentation en zones stables*.

Cette fusion de résultats a pour objectif d'améliorer la qualité de la segmentation :

- Chaque fonction d'observation est imparfaite. Elle présente des pics parasites (fausses alarmes), ou des pics sont absents (bonnes détections manquantes). Utiliser plusieurs fonctions d'observation permet de compenser les défaillances des unes par les qualités des autres.
- Le seuillage (prise de décision pour chaque fonction d'observation) n'est pas toujours efficace. Chaque fonction de décision est imparfaite. Il s'agit de compenser ce défaut.

Nous donnons dans la **deuxième section** (section 4.2) de ce chapitre la procédure de base que nous avons utilisée pour fusionner. Nous faisons entrer en concurrence les fonctions de décision.

Cette procédure de base, du fait que les transitions ne sont pas instantanées, n'est pas suffisante. Ce problème va nous amener à discuter des procédures mises en place pour rassembler les marques trop proches les unes des autres. Ces procédures sont appliquées lors de la deuxième étape de l'analyse *segmentation en zones stables*, c'est-à-dire après que chaque fonction d'observation a été seuillée, et lors de cette troisième étape de l'analyse *segmentation en zones stables*. Ceci fait l'objet de la **troisième section** (section 4.3) de ce chapitre. La règle de fusion utilisée est beaucoup plus compliquée que la règle de la majorité. Ceci implique notamment que ce qui est dit dans la section 2.6 à propos de l'effet de l'utilisation de capteurs corrélés n'est pas forcément directement valable ici.

Nous donnons une conclusion dans la **quatrième section** (section 4.4) de ce chapitre.

4.2 Première étape – Sommation

Dans l'exemple présenté sur la figure 4.1, nous supposons que nous extrayons trois fonctions d'observation et nous supposons qu'il y a trois transitions à détecter. Ces transitions surviennent environ aux moments indiqués par les trois traits interrompus verticaux du premier carré de la figure 4.1 (en haut à gauche).

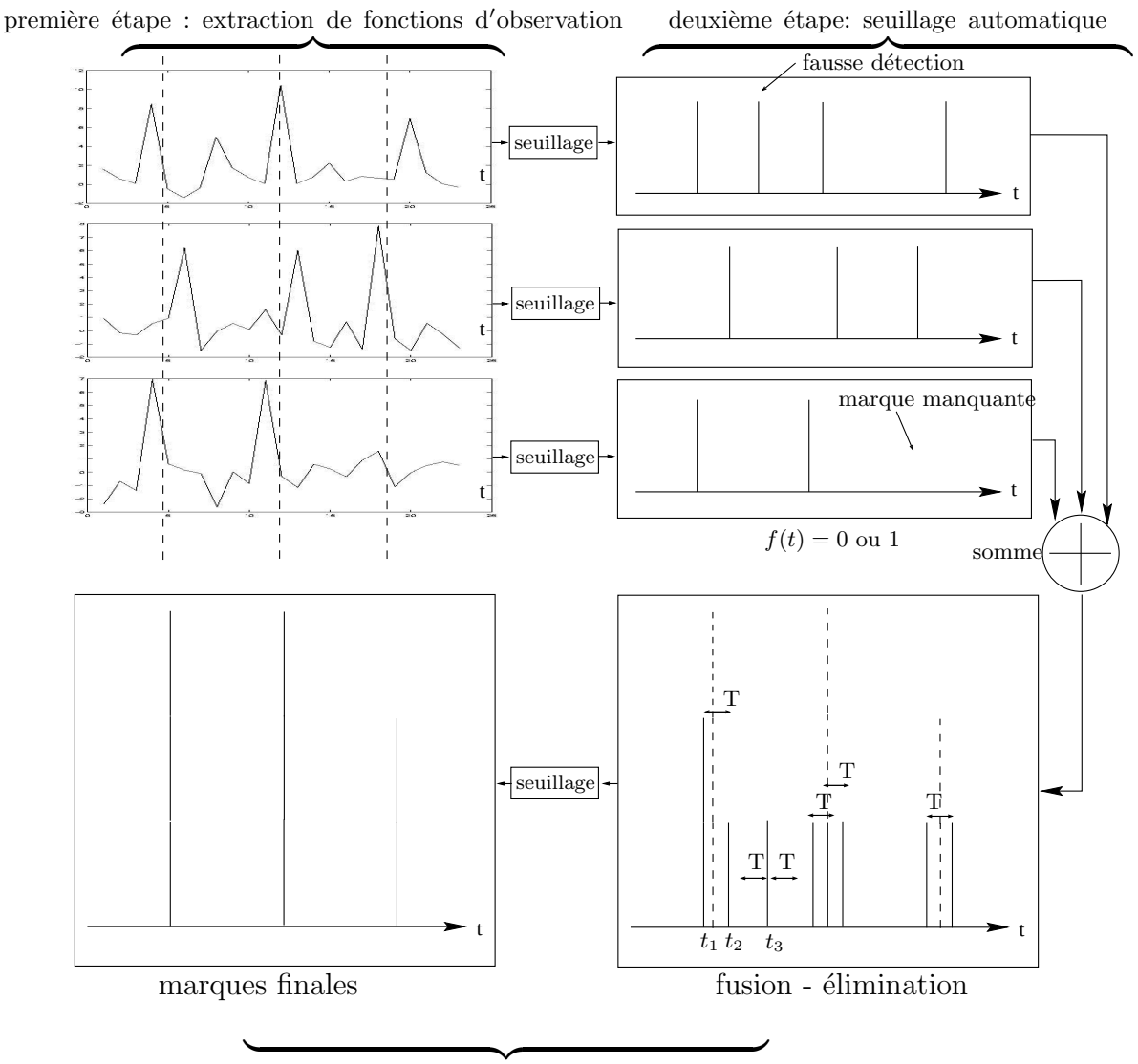


FIG. 4.1 – *Les trois premières étapes de la segmentation en zones stables*

Les fonctions d'observation présentent bien un pic à chaque transition, seulement ils ne surviennent pas au même moment pour toutes les fonctions d'observation. Ceci est dû à ce que les transitions ne sont pas instantanées.

Pour chaque fonction d'observation, un seuil automatique est calculé, et nous obtenons les marques présentées dans le carré en haut à droite de la figure 4.1. Nous obtenons trois fonctions de décision qui ont pour valeurs 0 (partie stable du signal) ou 1 (transition).

Nous constatons que pour la première fonction d'observation nous avons une fausse alarme (seuil trop bas ou pic de bruit trop grand) et que pour la troisième il nous manque une marque (seuil trop haut ou pic trop bas). Ainsi, nous voyons la nécessité d'utiliser plusieurs fonctions d'observation : il s'agit d'éviter les fausses alarmes (présentes par hypothèse sur peu de fonctions d'observation à un moment donné) et les marques manquantes (ce qui arrive par hypothèse pour peu de fonctions d'observation pour une transition donnée). En rassemblant les résultats obtenus pour chaque fonction d'observation, nous espérons mettre en évidence ceci (voir la section 2.6.3.1, les deux « sortes » de p_{bd}).

La **première étape** du mixage des résultats (fusion des données) consiste à additionner les fonctions de décision (à valeurs dans {0,1}). Nous obtenons, à chaque instant i, $Som(i)$. Quand plusieurs fonctions d'observation réagissent au même moment, nous obtenons une plus grande valeur (voir dans le carré en bas à droite de la figure 4.1, les segments de droite verticaux et en traits pleins) pour la fonction *Som(i)*.

Cependant, comme nous l'avons déjà mentionné, les transitions ne sont pas instantanées. Ainsi, pour une transition donnée, les fonctions d'observation réagissent diversement :

- plutôt au début des transitions pour certaines transitions
- plutôt à la fin pour d'autres
- plutôt au milieu pour d'autres
- et plusieurs fois pour certaines autres (cas non présenté sur la figure 4.1)

De plus, chaque fonction d'observation ne se comporte pas de la même façon d'une transition à la suivante. Donc, les marques obtenues après sommation des fonctions de décision sont éparpillées en groupes plus ou moins denses : il s'agit de sommer celles qui sont « trop » proches les unes des autres pour que réellement il y ait là la possibilité d'avoir deux marques à poser, et de les replacer dans le but de mieux estimer/approximer la position des transitions – ou de leur centre. Ceci est fait dans la section 4.3.1.

Les trois premiers points constituent un seul problème (appelons-le p_a), résolu dans les sections 4.3.1, 4.3.3.3 et 4.3.3.4. Le quatrième point constitue un problème (appelons-le p_b), résolu dans la section 4.3.3.2.

4.3 Deuxième étape – Traitement des marques trop rapprochées : éliminations

4.3.1 Algorithme de base

La **seconde étape** de la fusion consiste à rassembler les marques trop proches les unes des autres. L'algorithme à la base est le suivant :

- si à l'instant i_1 il y a une marque ($Som(i_1)$ différente de 0) [1], nous regardons s'il y a une autre marque dans l'intervalle $[i_1 \quad i_1 + T]$
  - s'il n'y en a pas, nous ne touchons pas à la marque en i_1 et nous passons à la marque suivante
  - s'il y en a une en i_2, nous regardons s'il y a une autre marque dans l'intervalle $[i_2 \quad i_2 + T]$, et nous gardons i_1 et i_2 en mémoire
  - nous regardons s'il y a une autre marque dans l'intervalle $[i_2 \quad i_2 + T]$, etc.

1. Implicitement, il n'y a pas de marque entre $i_1 - T$ et i_1.

- ensuite, il faut fusionner (c'est-à-dire les remplacer par une seule) les marques que nous avons en mémoire avant de passer à la marque suivante (qui forcément est éloignée de la dernière marque en mémoire de plus de T)

Si nous remplaçons les marques en mémoire par une marque placée en leur centre de gravité et de valeur leur somme, nous obtenons les marques en pointillés du carré en bas à droite de la figure 4.1. Les marques en mémoire sont éliminées.

Finalement, un seuil S_F (à fixer) est appliqué, qui élimine les plus petites marques. Nous faisons l'hypothèse qu'elles correspondent à des fausses alarmes. Nous obtenons les marques en traits pleins du carré en bas à gauche de la figure 4.1.

Ainsi, pour chaque marque obtenue, nous obtenons une sorte de « fonction de confiance ». Plus est elle grande, plus il est probable qu'une transition soit réellement présente. De plus, nous obtenons aussi une estimée de la position de chaque transition, c'est-à-dire en fait, puisque les transitions ne sont pas instantanées, de son centre.

Cependant, deux paramètres libres sont introduits : T et S_F.

4.3.2 Les problèmes

Nous discutons ici de quelques problèmes qui rendent la fusion des fonctions de décision, telle qu'elle a été décrite pour le moment, difficile.

Le premier problème (appelons-le p_c : il est résolu dans la section 4.3.3.5) vient de ce que, par exemple, dans un son communément harmonique par zones il est de temps en temps des segments non harmoniques relativement longs (un chanteur qui reprend son souffle : la longueur du segment non harmonique est de l'ordre d'une seconde). Donc, pour ce segment non harmonique, nous obtenons, pour les fonctions d'observation basées sur le trajet de f_0 par exemple, un grand nombre de marques très proches les unes des autres. Sommer ces marques comme nous le proposons dans la section 4.3.1 nous donne finalement une seule marque, située au centre du segment, ce qui ne correspond pas à ce que nous voulons. Dans ce cas, nous aimerions obtenir une marque au début de la zone non voisée et une marque à la fin.

Il s'agit de l'une des généralisations qu'il faudra effectuée : nous ne sommes pas toujours en présence d'un signal harmonique. Avant de *segmenter en zones stables*, il faut déterminer quand le signal est harmonique et quand il ne l'est pas. Toutes les fonctions d'observation ne sont pas toujours utilisables. À ce sujet, voir les autres parties de cet exposé.

Le deuxième problème (appelons-le p_d : il est en partie résolu dans la section 4.3.3.7) vient de ce que si nous traitons un signal constitué de notes (ou de phones) très courtes (moins de 20 millisecondes, par exemple), les groupes de marques obtenus pour deux transitions successives se mélangent. La méthode pour fusionner en deux étapes proposée dans la section 4.2 et la section 4.3.1 échoue : nous obtenons une marque quelque part au centre de l'une de ces très courtes notes. Le résultat là non plus ne correspond pas à ce que nous voulons. Nous souhaiterions plutôt (au moins) garder toutes les marques, même si elles sont excessivement abondantes.

Nous concevons difficilement que deux transitions puissent se superposer. Ainsi, considérons un son où deux transitions sont présentes. Nous faisons ici l'hypothèse qu'elles sont parfaitement instantanées. La fin (une fonction d'observation réagit à ce moment-là) de la première transition a toujours lieu avant le début de la seconde transition (une fonction d'observation, par exemple la même que ci-dessus, réagit à ce moment).

Cependant, les fonctions d'observation sont calculées sur des portions du signal sonore, portions larges de quelques dizaines de millisecondes, ce qui a pour effet d'étaler la zone d'influence de chaque transition : deux zones d'influence successives peuvent se recouvrir. D'où, déjà, le mélange possible des groupes de marques.

Mais le caractère instantané de la transition entre deux notes est ce qu'il faut principalement remettre en cause. Et, de la même façon :

- La réverbération étale les notes (fin d'une note mélangée avec le début de la suivante).

- Les notes, pour certains instruments, peuvent se chevaucher (fin d'une note mélangée avec le début de la suivante) : voir la harpe, par exemple.
- Les harmoniques naissent l'un après l'autre [2] (des plus basses fréquences au plus hautes, au fur et à mesure que l'énergie augmente), ce qui étale encore la transition.

Dans les deux premiers cas, le signal n'est plus parfaitement monophonique.

4.3.3 Méthode utilisant une « distance minimale entre deux marques »

4.3.3.1 Introduction

Reprenons l'algorithme de base décrit dans la section 4.3.1 ci-dessus. Dans un premier temps, nous observons des groupes de marques. Chaque groupe est constitué de marques consécutives séparées par moins de T millisecondes ; et la distance entre la dernière marque d'un groupe et la première marque du groupe suivant est supérieure à T. T a été fixé à 50 millisecondes.

Nous indiquons ci-dessous trois méthodes, aucune ne visant à être plus qu'un moindre mal, pour éliminer les marques trop proches. Un traitement supplémentaire doit être appliqué dans les cas pathologiques. De nouveaux paramètres libres sont introduits, dont il faut essayer de se débarrasser.

4.3.3.2 Élimination « simple »

Dans un groupe de marques, nous ne gardons que la marque pour laquelle la valeur correspondante de la fonction d'observation est la plus grande : voir la figure 4.2. Cette méthode d'« élimination » de marques est appliquée à chaque fonction d'observation après qu'elle a été seuillée, c'est-à-dire lors de la deuxième étape de l'analyse *segmentation en zones stables* sur chaque fonction de décision. Chaque marque obtenue a pour valeur 1. L'idée est de ne garder qu'une marque par transition, certaines fonctions d'observation réagissant plusieurs fois au cours de certaines d'entre elles.

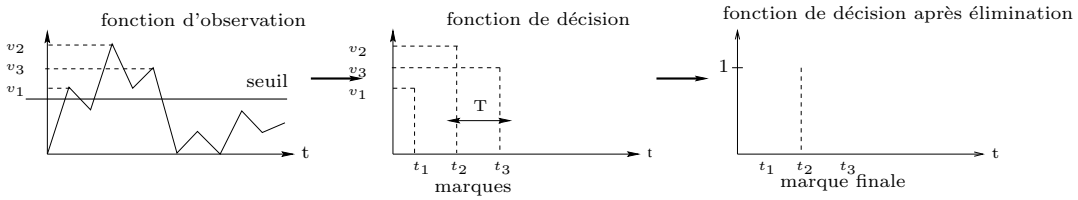


FIG. 4.2 – *Élimination « simple » : dans un groupe de marques, nous ne gardons que la marque pour laquelle la valeur correspondante de la fonction d'observation est la plus grande*

4.3.3.3 Élimination « somme »

Nous ne gardons que la marque pour laquelle la valeur correspondante de la fonction d'observation est la plus grande. Nous changeons la valeur de la marque que nous gardons, en lui ajoutant la valeur de chacune de celles qui ont été éliminées pondérée en fonction de sa distance à la marque gardée. Voir la figure 4.3.

Cette méthode a été utilisée à la fin de la troisième étape de l'analyse *segmentation en zones stables*, une fois que nous avions sommé toutes les fonctions de décision. Elle a été appliquée sur cette somme : cela veut dire avec des v_i ayant des valeurs entières. Mais, en fait, dans le programme *segmentation*, a été implémentée la méthode décrite dans la section 4.3.3.4.

2. Considérons que « les harmoniques naissent l'un après l'autre ». Alors, les résultats obtenus pour chaque harmonique avec les fonctions d'observation multidimensionnelles ne sont plus homogènes. Nous ne pouvons donc plus faire simplement les « moyennes » (voir la section 2.6). Chaque dimension devient une fonction d'observation à part entière, et la fusion doit être faite une fois que les décisions pour chacune ont été prises. C'est-à-dire que ce sont les fonctions de décision qui sont fusionnées. Il s'agit d'une perspective.

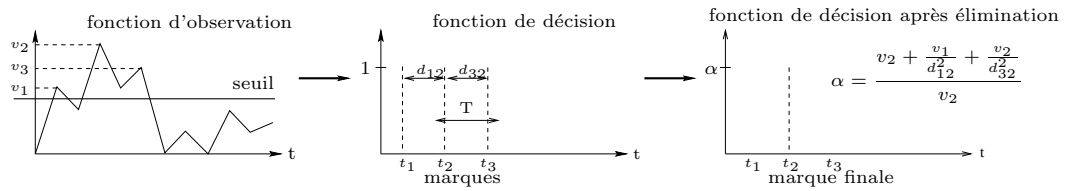


FIG. 4.3 – *Élimination « somme » : dans un groupe de marques, nous ne gardons que la marque pour laquelle la valeur correspondante de la fonction d'observation est la plus grande ; à la valeur de la marque que nous gardons nous ajoutons la valeur de chacune de celles qui ont été éliminées pondérée en fonction de sa distance à la marque gardée*

Dans le cas présenté sur la figure 4.3, la pondération se fait ainsi :

$$\alpha = \frac{v_2 + \frac{v_1}{d_{12}^2} + \frac{v_3}{d_{13}^2}}{v_2}$$

ou ainsi : $\alpha = \dfrac{v_2 + \frac{v_1}{d_{12}^2\beta} + \frac{v_3}{d_{13}^2\beta}}{v_2}$ ou encore ainsi : $\alpha = \dfrac{v_2 + v_1 \exp\left(-\frac{d_{12}^2}{\beta}\right) + v_3 \exp\left(-\frac{d_{13}^2}{\beta}\right)}{v_2}$
β étant un paramètre libre à fixer.

4.3.3.4 Élimination « somme et positionnement de la marque »

Nous calculons la position de la marque gardée à partir du calcul du centre de gravité du groupe de marques. Cette position est égale à l'instant d'échantillonnage de la fonction d'observation le plus proche du centre de gravité. Nous modifions la valeur de la marque que nous gardons en lui ajoutant la valeur de chacune de celles qui ont été éliminées, pondérée en fonction de sa distance à la marque gardée. Voir la figure 4.4.

Les pondérations se font comme dans la section 4.3.3.3.

Cette méthode est utilisée à la fin de la troisième étape de l'analyse, une fois que nous avons sommé toutes les fonctions de décision (elle est appliquée sur cette somme : cela veut dire que les v_i ont des valeurs entières).

L'idée est d'obtenir une estimation de la position centrale de la transition.

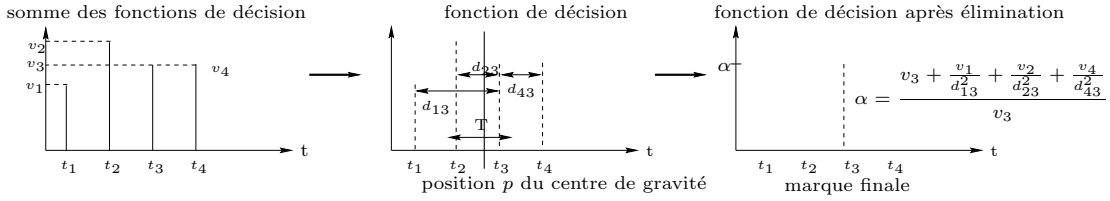


FIG. 4.4 – *Élimination « somme et positionnement de la marque » : dans un groupe de marques, nous ne gardons que la marque la plus proche du centre de gravité du groupe ; à la valeur de la marque que nous gardons nous ajoutons la valeur de chacune de celles qui ont été éliminées pondérée en fonction de sa distance à la marque gardée*

4.3.3.5 Traitement spécial

Si les marques à rassembler couvrent une zone supérieure à TG millisecondes (c'est-à-dire si la largeur du groupe de marques est supérieure à TG), cela veut dire :

- soit il faut poser une marque au début et une autre à la fin de la zone : cette zone est bruitée, cela veut dire que nous avons utilisé des fonctions d'observation pas aptes à déceler

le caractère stable de cette zone (par exemple, les « valeurs absolues des dérivées de f_0 » sont utilisées alors que nous sommes en présence d'un signal percussif)

- soit il y a plusieurs transitions très rapprochées : il faut donc plutôt garder toutes les marques.

En fait, cela indique que nous sommes en présence d'un cas pathologique, que nous ne savons pas traiter.

Ainsi, un deuxième fenêtrage est effectué. La longueur de cette fenêtre est cette fois de TG millisecondes, avec $TG = 500$ millisecondes par exemple.

Dans le programme *segmentation*, ce second fenêtrage est implémentée. Par défaut, la première solution est appliquée : nous posons une marque au début et une autre à la fin de la zone. Bien sûr, ni l'une ni l'autre des solutions données ici n'est satisfaisante.

4.3.3.6 Remarques

Nous considérons ici que les longueurs T et TG sont fixes : elles correspondent à deux paramètres libres, à fixer à un moment ou à un autre. Nous pouvons les rendre adaptatifs, mais cela ne résoudra pas tous les problèmes. Par exemple, quand deux groupes de marques correspondant à deux transitions sont mélangés, comment les séparer ? Dans la section suivante et surtout dans la section 4.3.4, un début de réponse au problème rencontré (problème p_d) est apporté. Il s'agit de perspectives.

4.3.3.7 Une idée pour nous affranchir de TG

Nous pouvons peut-être nous affranchir d'ors et déjà de TG, comme vont le montrer les remarques qui suivent.

Considérons une segmentation basée sur neuf fonctions d'observation et considérons les marques obtenues après sommation des fonctions de décision (nous considérons ici qu'*élimination simple* a été appliquée, avec T et TG, pour chaque fonction d'observation : ce n'est pas restrictif). Nous classons les groupes de marques en trois types.

1° – L'allure d'un groupe de marques concernant une transition seule est plutôt celle présentée sur la figure 4.5.

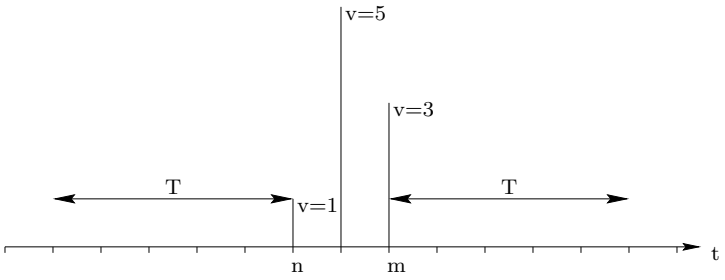


FIG. 4.5 – *Allure du groupe de marques quand il concerne une seule transition : il ne faut garder qu'une seule marque*

2° – L'allure d'un groupe de marques concernant une zone bruitée est plutôt celle présentée sur la figure 4.6.

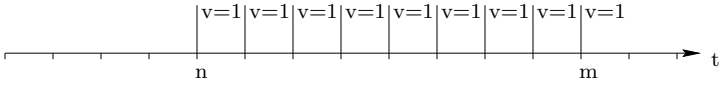


FIG. 4.6 – *Allure du groupe de marques quand il concerne une zone bruitée : il faut garder une marque au début et une autre à la fin*

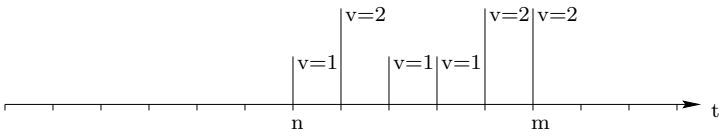


FIG. 4.7 – *Allure du groupe de marques quand il concerne deux transitions, c'est-à-dire quand deux groupes se mélangent : il faut les séparer*

3° – L'allure d'un groupe de marques concernant deux groupes mélangés (deux transitions très proches) est plutôt celle présentée sur la figure 4.7.

La position $t^{'}$ du centre de gravité de chaque groupe de marques se calcule ainsi :

$$t^{'} = \frac{\sum_{i=n}^{m} v_i t_i}{\sum_{i=n}^{m} v_i}$$

Supposons que nous remplacions toutes les marques du groupe par une marque en $t^{'}$ et que nous donnions à celle-ci pour valeur :

$$v^{'} = \sum_{i=n}^{m} v_i \exp\left(-\frac{(t^{'} - t_i)^2}{\beta}\right)$$

nous constatons que $v^{'}$ a un comportement différent suivant le type de groupe de marques considéré.

- Dans le cas 1°, $v^{'}$ est grande et supérieure à la plus grande des valeurs v_{max} dans l'ensemble $[v_n \ \ldots \ v_m]$.
- Dans le cas 2°, $v^{'}$ et v_{max} sont du même ordre de grandeur et petites (nous sommons beaucoup de petites marques très fortement pondérées à cause de leur grande distance au centre de gravité).
- Dans le cas 3°, $v^{'}$ est inférieure à v_{max} (le centre de gravité est très éloigné de la position de v_{max} : il tombe entre les deux classes, c'est-à-dire entre les deux transitions, au milieu d'une zone stable).

Pour les exemples données, nous obtenons (avec $\beta = 0{,}4$ et $t_i = i$) :

- Cas 1° : $v^{'} = 5{,}10$ et $v_{max} = 5$
- Cas 2° : $v^{'} = 1{,}16$ et $v_{max} = 1$
- Cas 3° : $v^{'} = 1{,}15$ et $v_{max} = 2$

Il reste à tester cette méthode sur des signaux réels. Il faudrait aussi nous affranchir de T.

4.3.4 Autre méthode

4.3.4.1 Présentation

Cette méthode est basée sur le groupement (ou « clusterisation » : voir [Buh95]) par les centres. Il s'agit d'une méthode de classification. Son principe est expliqué ci-dessous.

Principe Nous avons N données x_i à classifier dans K classes. Chaque $M_{i\nu}$ (avec i variant de 1 à N et ν variant de 1 à K) correspond au taux d'assignement du point x_i à la classe ν, soit encore à la probabilité que le point x_i appartienne à la classe ν. Ces taux valent donc 0 ou 1, et nous avons :

$$\sum_{\nu=1}^{K} M_{i\nu} = 1 \text{ (équation 1)}$$

Soit :

$$y_\nu = \frac{\sum_{i=1}^{N} M_{i\nu} x_i}{\sum_{i=1}^{N} M_{i\nu}} \text{ (équation 2)}$$

la position du centre de chaque classe. Soient $D_{i\nu} = (x_i - y_\nu)^2$ les distances quadratiques de chaque point à chaque centre (d'autres « distances » peuvent être considérées). Il s'agit de minimiser itérativement suivant les inconnues (c'est-à-dire les $M_{i\nu}$) la fonction de coût :

$$\mathcal{E} = \sum_{i=1}^{N} \sum_{\nu=1}^{K} M_{i\nu} D_{i\nu} \text{ (équation 3)}$$

Cette minimisation se fait ainsi :

étape 1 – initialisation des y_ν (par exemple avec les K premiers points à classifier)
étape 2 – calcul des taux d'assignement :

$$M_{i\alpha} = \begin{cases} 1 & \text{si } \| x_i - y_\alpha \| < \| x_i - y_\nu \| \quad \text{pour tout} \quad \nu \neq \alpha \\ 0 & \text{sinon} \end{cases}$$

étape 3 – recalcul des centres (voir l'équation 2)
étape 4 – les étapes 2 et 3 sont répétées jusqu'à la convergence (c'est-à-dire jusqu'à ce que plus aucun des points à classifier ne change de classe d'une itération à la suivante)

Extension Au lieu que les $M_{i\nu}$ soient strictement à valeurs dans $\{0, 1\}$, nous pouvons leur faire prendre toutes les valeurs entre 0 et 1. Alors, les taux d'assignement se calculent ainsi (ceci constitue la seule différence par rapport à l'algorithme donné ci-dessus) :

$$M_{i\alpha} = \frac{\exp\left(-\beta \left(x_i - y_\alpha\right)^2\right)}{\sum_{\nu=1}^{K} \exp\left(-\beta \left(x_i - y_\nu\right)^2\right)}$$

β est initialisé (étape 1) à une petite valeur : les K taux d'assignement pour chaque points sont proches. Puis β augmente à chaque itération. Un β égal à l'infini donne des taux d'assignements à valeur dans $\{0, 1\}$. Cette optimisation est reliée à la méthode de recherche de Monte Carlo, β correspondant ici à la température.

Quand nous ne connaissons pas le nombre de classes K, l'utilisation de cette méthode, comme de toutes les méthodes de classification, devient difficile. Telle est notre situation (groupe de marques de type 1° : $K = 1$; groupe de marques de type 3° : $K = n$, avec $n > 1$). Mais il existe des moyens pour résoudre cette difficulté (minimisations simultanées du coût $\mathcal{E}$ et d'un coût de complexité)... Nous nous attachons à les étudier : il s'agit d'une perspective.

Si nous travaillons avec les trois types de groupes de marques décrits dans la section précédente, nous constatons que les types 1° et 3° sont traités correctement par cette méthode, sans que nous ayons plus besoin de T. Il reste le problème du type 2°, où il s'agit de poser une marque de segmentation au début de la zone et une autre à la fin.

4.4 Conclusion

Ce problème du traitement automatique des marques trop rapprochées est la limitation principale que nous avons rencontrée pour la procédure de *segmentation en zones stables* toute automatique et sans apprentissage décrite dans cet exposé. Nous faisons quelques remarques :

- Les méthodes décrites dans ce chapitre ne sont que des pis-aller. Voir les figures 6.28 et 6.29, pour constater que cependant elles améliorent les performances de la segmentation.
- Trois paramètres libres « généraux » sont ajoutés dans le programme *segmentation* : T, TG et S_F. Notamment, avec l'utilisation du paramètre libre T, nous réintroduisons une hypothèse restrictive sur la longueur des notes : il faut qu'elle soit supérieure à T.

- Trois voies pour passer outre cette limitation sont proposées en tant que perspectives :
  - La mise en place de méthodes d'apprentissage devrait permettre de la résoudre.
  - Définir une fonction d'observation universelle, qui réagisse à tous les types de transition, permettrait de ne pas avoir à fusionner.
  - Pour cela, il faudrait changer de modèle de signal, c'est-à-dire prendre le problème d'un autre point de vue.

Chapitre 5

Étiquetage des segments obtenus

5.1 Introduction

Il s'agit de la *quatrième et dernière étape* de la *segmentation en zones stables*.

Dans cette partie, puisque nous ne considérons délibérément que les sons monophoniques et harmoniques, l'étiquetage consiste principalement à déterminer la note jouée, c'est-à-dire à opérer la transcription automatique de l'extrait segmenté. Ceci fait l'objet de la **deuxième section** (section 5.2) de ce chapitre.

D'autres étiquettes, qui vaudront pour des sons plus compliqués, sont présentées dans la **troisième section** (section 5.3) de ce chapitre.

5.2 Transcription automatique

5.2.1 Les notes jouées

Dans un **premier temps**, il s'agit d'effectuer la transcription automatique. Pour trouver la note jouée pour un segment particulier, ces trois mesures sont utilisées :

1° m_1 : la moyenne de f_0 sur le segment
2° m_2 : la médiane de f_0 sur le segment
3° m_3 : la moyenne de f_0 sur le segment, sans le premier et le quatrième quart de ce segment

Les distances entre les fréquences vraies FV (voir le tableau 6.2) des notes et chaque mesure m_i sont calculées. La plus petite distance FV_i nous donne la note jouée NV_i. Nous obtenons ainsi les décisions NV_1, NV_2 et NV_3 pour les trois mesures.

Donnons à présent les avantages et les inconvénients de chaque mesure. Si la marque de début du segment considéré est placée légèrement trop tôt, ou si la marque de fin du segment considéré est placée légèrement trop tard, m_1 peut être suffisamment faussée pour qu'à la prise de décision nous nous trompions d'une note (voire de plusieurs notes). m_3, ne prenant pas en compte le début et la fin du segment, pallie ce problème. Cependant, puisque m_3 est calculée avec moins de points que m_1, la variance de m_3 est plus grande que la variance de m_1 : m_3 étant calculée avec deux fois moins de points que m_1, si nous admettons que tous les points obéissent à la même loi de probabilité, sa variance est deux fois plus grande que celle de m_1. Les segments étant courts, ils couvrent peu d'échantillons du trajet de f_0. Alors, cette différence de variance a une influence notable. Et m_3 à son tour peut être suffisamment faussée pour qu'à la prise de décision nous nous trompions d'une note (voire de plusieurs notes). m_2 semble être un bon compromis.

Il s'agit de décider quelle note a été jouée à partir des trois décisions précédentes. Nous avons de nouveau un problème de fusion des données. La décision finale est prise ainsi (cinq cas ont été considérés) :

- Deux ou trois des décisions sont identiques : elles nous donnent la note NV. La décision finale est NV.

- Les trois décisions sont différentes : les trois sont proches (elles correspondent à trois notes consécutives). La décision finale est $NV = NV_2$.
- Les trois décisions sont différentes : deux, parmi lesquelles NV_2 celle obtenue à partir de m_2, sont proches (une note les sépare) ; la dernière est éloignée des deux autres. La décision finale est $NV = NV_2$.
- Les trois décisions sont différentes : deux sont proches (une note les sépare) ; la dernière, qui est celle obtenue à partir de m_2, est éloignée des deux autres. La décision finale est, d'une façon indéterminée, NV_1 ou NV_3.
- Les trois décisions sont différentes : les trois sont éloignées les unes des autres. Cela veut dire que la segmentation n'est pas correcte.

Remarquons que si un vibrato d'amplitude importante est présent, ces trois mesures ne peuvent pas être utilisées.

Dans un **deuxième temps**, il s'agira de déterminer le rythme, le tempo, à partir de la durée des notes. Ceci constitue une perspective.

5.2.2 Plus petit qu'une note

Dans un **troisième temps**, les étiquettes données par HAJDA dans [Haj96] (attaque, transition, état stable, chute : voir la section 2.5.6) sont à considérer. Il s'agit de segmenter plus petit que la note. Ceci constitue une perspective.

5.3 Les autres étiquettes

Comme nous allons par la suite considérer des sons plus compliqués que les sons monophoniques, harmoniques et non modulés, il s'agira aussi de déterminer, segment par segment, par exemple :

1. si le segment correspond à du silence ou à du signal
2. si le segment est constitué de bruit ou est composé d'une somme de sinusoïdes [1]
3. si le son est harmonique ou inharmonique
4. la note jouée s'il y a lieu
5. s'il y a du vibrato ou pas
6. s'il y a du trémolo ou pas
7. si le son est monophonique ou polyphonique

Nous pouvons classer les critères de cette liste en trois groupes :

- Il y a ceux qui sont plutôt absolus : ou bien le signal est monophonique, ou bien il est polyphonique. Il s'agit des critères 1 et 7.
- Il y a des critères plus flous : il y a plus ou moins de vibrato. Il s'agit des critères 2, 3, 5 et 6.

Dans ces deux cas, il faut prendre une décision, même si elle est plus ou moins incertaine.

- Il y a ceux qui consistent à estimer une valeur. Il s'agit du critère 4.

1. Dans le chapitre 19 de la partie IV, nous faisons une distinction entre « voisé » et « harmonique », distinction toute personnelle. En résumé, pour nous, un signal harmonique est un signal composé d'une somme de sinusoïdes, disposées sur un peigne harmonique plutôt plein ; un signal voisé est un signal composé d'une somme de sinusoïdes, disposées sur un peigne harmonique plutôt vide et aux « dents » très rapprochées.

Chapitre 6

Quelques performances

6.1 Introduction

Une série de tests a été effectuée sur l'ensemble de signaux d'étude que nous décrivons dans la section 6.2 ci-dessous. Ces sons peuvent être trouvés ici :

`http://www.ircam.fr/equipes/analyse-synthese/rossigno/these/segm.html`

Nous décrivons dans la **deuxième section** (section 6.2) de ce chapitre la base de données de sons utilisée.

Dans la **troisième section** (section 6.3) de ce chapitre, nous donnons les performances du système avec un extrait de flûte monophonique, harmonique et quasi non modulé. Nous donnons pour ce son les trajets des fonctions d'observation ; les trois fenêtres de l'interface graphique du programme *segmentation*, correspondant aux trois premières étapes de la *segmentation en zones stables* ; et les résultats de la transcription automatique.

Dans la **quatrième section** (section 6.4) de ce chapitre, nous résumons les performances du système obtenues avec les sons de la base de données. Les performances du système complet : fusion des résultats obtenus pour plusieurs fonctions d'observation, etc., sont donnés dans cette même section.

Dans la **cinquième section** (section 6.5) de ce chapitre, nous donnons une conclusion à ce chapitre.

6.2 Les sons réels utilisés

Une base de données composée de sons monophoniques a été construite. Elle contient :

- Les sons de l'IRCAM présentés dans la section 6.2.1.
- Les sons du CD Sqam présentés dans la section 6.2.2. Le terme « Sqam » est mis pour : « Sound Quality Assessment Material ». Ce CD a été édité par : « European Broadcasting Union ».

6.2.1 Sons monophoniques – IRCAM

Les *signaux d'étude* suivants ont été considérés :

- **flute.sf** : extrait de flûte.
- **brahms2.sf** : extrait de clarinette.
- **Violon2.sf** : extrait de violon.
- **voiceP.sf** : extrait de voix chantée.
- **piano2.sf** : une note de piano.
- **casta.sf** : extrait de castagnettes.

6.2.2 Sons monophoniques – CD Sqam

Les *signaux d'étude* suivants ont été considérés :

- **3.sf** : extrait d'un gong électronique artificiel.
- **6.sf** : extrait d'un autre gong électronique artificiel.
- **8a.sf** : extrait d'un violon.
- **11a.sf** : extrait d'une contrebasse.
- **14a.sf** : extrait d'un hautbois.
- **16a.sf** : extrait d'une clarinette.
- **17a.sf** : extrait d'une clarinette basse.
- **19a.sf** : extrait d'un contre-basson.
- **20a.sf** : extrait d'un saxophone.
- **20b.sf** : autre extrait d'un saxophone.
- **29.sf** : extrait d'une grosse caisse.
- **30.sf** : extrait d'une timbale.
- **39a.sf** : extrait d'un piano.
- **42a.sf** : extrait d'un accordéon.
- **43a.sf** : extrait d'un orgue.

6.3 Performances pour l'extrait de « flûte »

6.3.1 Description du son considéré

Il est monophonique et parfaitement harmonique. Un léger vibrato est de temps en temps présent (par exemple sur la plus longue note, entre la $4^{ème}$ et la $6^{ème}$ seconde : voir la figure 6.1), mais il est suffisamment petit pour ne pas nous gêner et pour que nous considérions que le signal n'est pas modulé. Il a été enregistré en salle anéchoïque, donc la réverbération est quasi absente : la fin d'une note ne se superpose jamais au début de celle qui la suit. Les transitions sont donc relativement rapides.

Remarquons que, la fréquence d'échantillonnage du signal étant 32000 Hz, pour une fondamentale de 440 Hz (la_3) la période du signal couvre environ 73 échantillons.

6.3.2 Les fonctions d'observation

Nous donnons ci-dessous les trajets des fonctions d'observation obtenues pour l'extrait de flûte.

6.3.2.1 Fonctions d'observation basées sur f_0

Sur la figure 6.1 nous présentons le trajet de f_0. Sur la figure 6.2 nous donnons le trajet de la « valeur absolue de la dérivée du trajet de f_0 » (voir la section 2.2.2).

Sur la figure 6.3, nous présentons le trajet de la « somme des valeurs absolues des dérivées des indices de voisement première forme » (voir la section 2.2.3).

Nous donnons les trajets des 11 premiers harmoniques, des inharmonicités calculées du 1^{er} au $11^{ème}$ harmonique, et des deux fonctions d'observation (voir la section 2.2.4) « somme (moyenne arithmétique) », et « produit (moyenne géométrique), calculées avec les trois premiers indices d'inharmonicité », respectivement sur les figures 6.4, 6.5, 6.6 et 6.7.

Nous présentons sur la figure 6.8 le trajet de $1 - p(i)$: « analyse statistique appliquée au trajet de la fréquence fondamentale » (voir la section 2.2.5).

Nous donnons sur la figure 6.9 les marques obtenues sur le trajet de la fondamentale avec la « rupture de modèles AR » (voir la section 2.2.6). Nous avons posé à la main, sur la figure 6.10, les marques « vraies » sur le trajet de f_0. Pour ce son, les transitions sont très aisément décelables à l'oreille.

Voir la section 14.5 en ce qui concerne la fonction d'observation basée sur le filtrage de HILBERT.

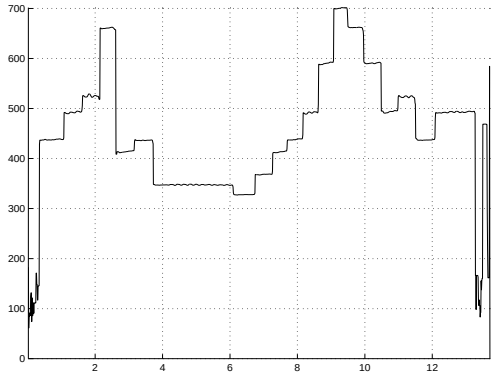


FIG. 6.1 – *Trajet de la fréquence fondamentale f_0. En abscisse : le temps ; en ordonnée : la fréquence en Hz*

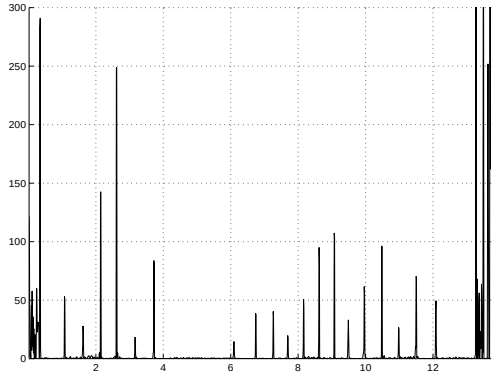


FIG. 6.2 – *Trajet de la « valeur absolue de la dérivée de la fréquence fondamentale f_0 ». En abscisse : le temps*

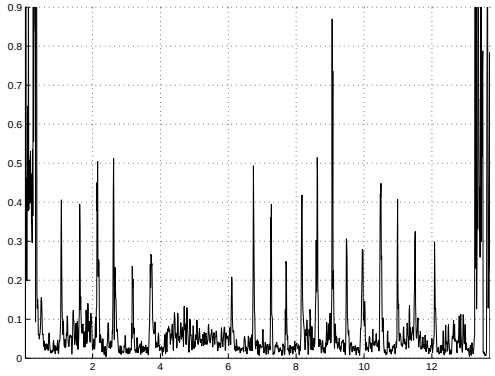


FIG. 6.3 – *Trajet de la « somme des valeurs absolues des dérivées des dix premiers indices de voisement première forme ». En abscisse : le temps*

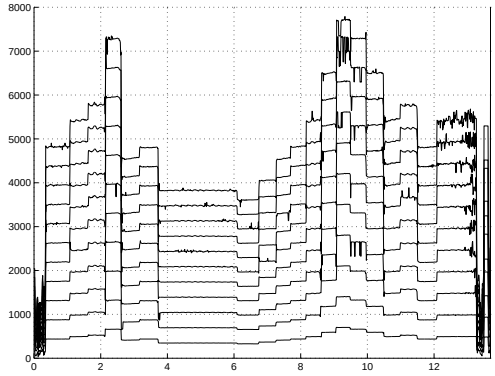


FIG. 6.4 – *Trajets des onze premiers harmoniques. En abscisse : le temps ; en ordonnée : la fréquence en Hz*

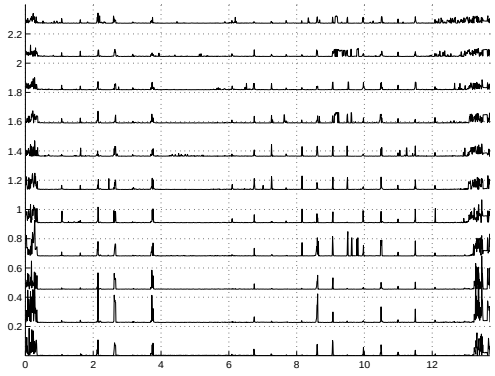


FIG. 6.5 – *Trajets des indices d'inharmonicités du 1^{er} au $11^{ème}$ harmonique. En abscisse : le temps ; en ordonnée : une échelle arbitraire*

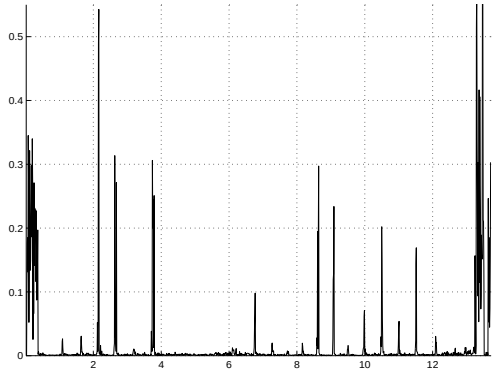


FIG. 6.6 – *Trajet de la « somme des valeurs absolues des dérivées des trois premiers indices d'inharmonicité ». En abscisse : le temps*

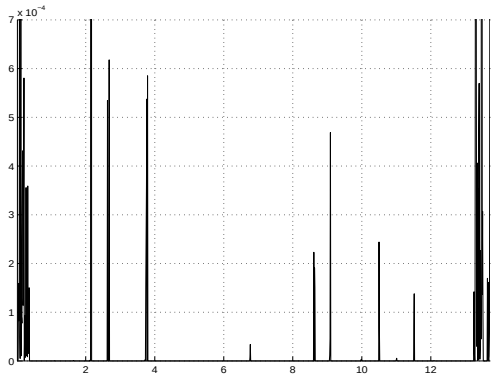


FIG. 6.7 – *Trajet du « produit des valeurs absolues des dérivées des trois premiers indices d'inharmonicité ». En abscisse : le temps*

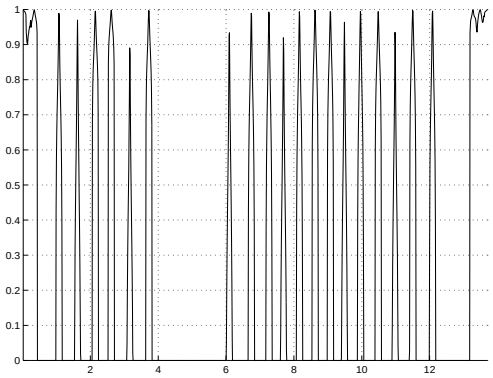


FIG. 6.8 – *Trajet de l'« analyse statistique appliquée au trajet de f_0 ». En abscisse : le temps ; en ordonnée : la probabilité $1 - p(i)$*

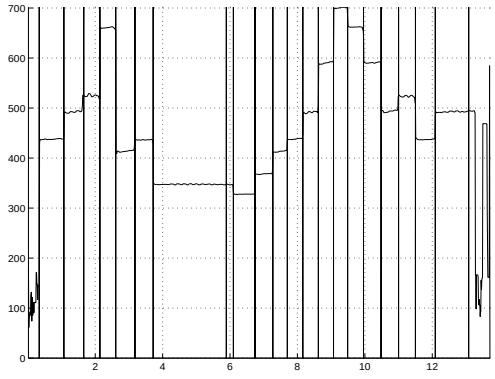


FIG. 6.9 – *Marques de segmentation données par la « rupture de modèles » sur le trajet de la fondamentale. En abscisse : le temps ; en ordonnée : la fréquence en Hz. Les traits verticaux sont les marques*

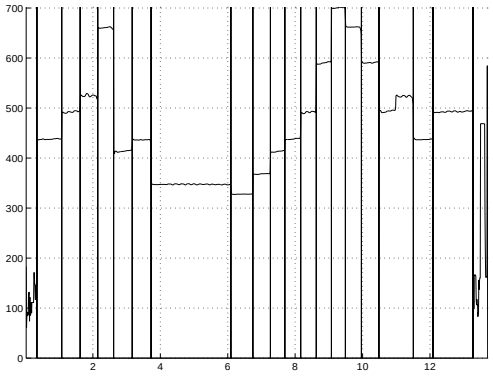


FIG. 6.10 – *Marques de segmentation posées à la main sur le trajet de la fréquence fondamentale. En abscisse : le temps ; en ordonnée : la fréquence en Hz. Les traits verticaux sont les marques*

6.3.2.2 Fonctions d'observation basées sur l'énergie

Sur la figure 6.11 nous présentons le trajet de l'énergie. Sur la figure 6.12 nous donnons la « valeur absolue de la dérivée du trajet de l'énergie » (voir la section 2.3.2).

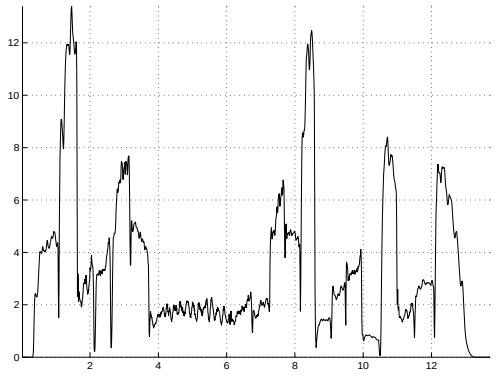


FIG. 6.11 – *Trajet de l'énergie. En abscisse : le temps ; en ordonnée : l'énergie*

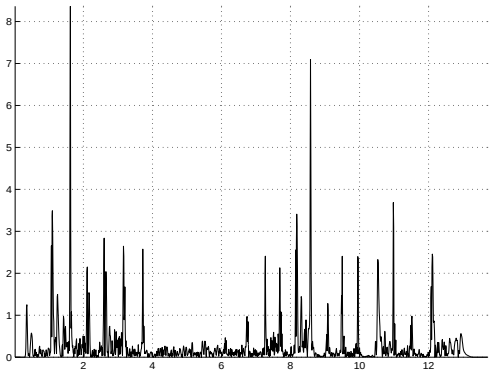


FIG. 6.12 – *Trajet de la « valeur absolue de la dérivée de l'énergie ». En abscisse : le temps*

Nous présentons sur la figure 6.13 le trajet de $1-p(i)$: « analyse statistique appliquée au trajet de l'énergie » (voir la section 2.3.2).

Nous donnons sur la figure 6.14 les marques obtenues sur le trajet de l'énergie avec la « rupture de modèles AR » (voir la section 2.3.2).

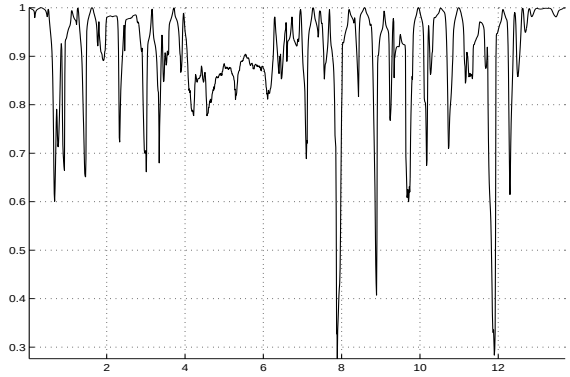


FIG. 6.13 – *Trajet de l'« analyse statistique appliquée au trajet de l'énergie ». En abscisse : le temps ; en ordonnée : la probabilité* $1-p(i)$

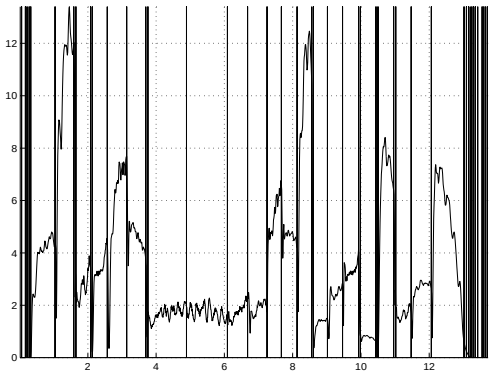


FIG. 6.14 – *Marques de segmentation données par la « rupture de modèles » sur le trajet de l'énergie. En abscisse : le temps ; en ordonnée : l'énergie. Les traits verticaux sont les marques*

6.3.2.3 Fonctions d'observation basées sur le contenu spectral

Sur la figure 6.15, nous présentons le trajet de la « valeur absolue de la dérivée de l'indice de voisement deuxième forme calculé avec le spectre d'amplitude » (voir la section 2.4.1, et aussi la section 2.4.2).

Nous donnons ensuite les trajets des divers « flux spectraux ». Nous présentons tout d'abord le trajet du « flux spectral calculé avec les spectres d'amplitude » (voir la section 2.4.3.2). Nous donnons le trajet du flux spectral calculé sur toutes les fréquences, puis seulement sur les basses fréquences et enfin seulement sur les hautes fréquences. La césure est posée au centre des spectres d'amplitude. Voir les figures 6.16, 6.17 et 6.18.

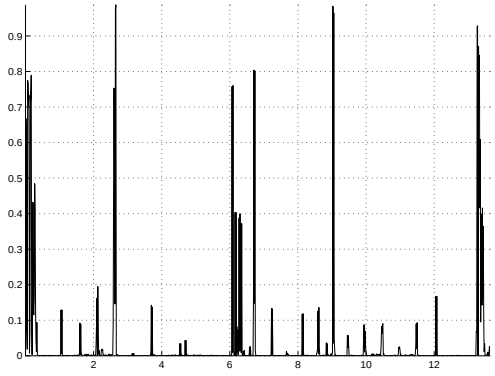


FIG. 6.15 – *Trajet de la « valeur absolue de la dérivée de l'indice de voisement deuxième forme calculé avec le spectre d'amplitude ». En abscisse : le temps*

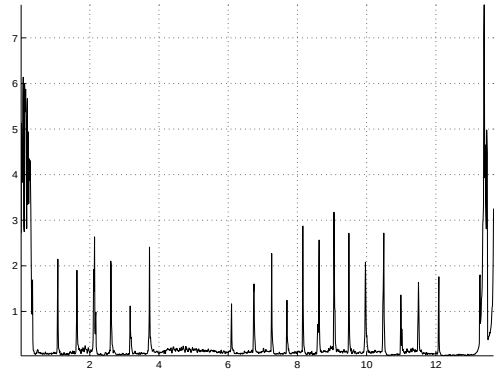


FIG. 6.16 – *Trajet du « flux spectral calculé avec toutes les fréquences des spectres d'amplitude ». En abscisse : le temps*

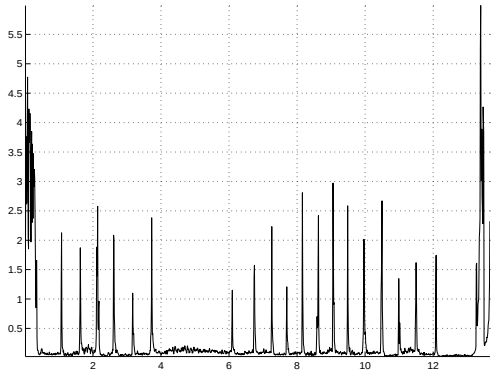


FIG. 6.17 – *Trajet du « flux spectral calculé avec les basses fréquences des spectres d'amplitude ». En abscisse : le temps*

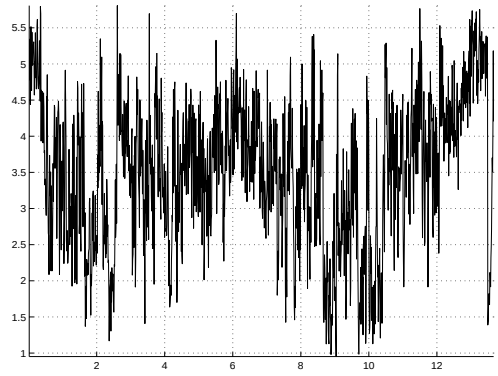


FIG. 6.18 – *Trajet du « flux spectral calculé avec les hautes fréquences des spectres d'amplitude ». En abscisse : le temps*

Avec des modèles AR d'ordre 6 et pour les deux formes du « flux spectral » qui existent quand nous le calculons avec les « enveloppes spectrales basées sur la modélisation AR » (voir la section 2.4.3.3), nous obtenons les trajets des figures 6.19 et 6.20.

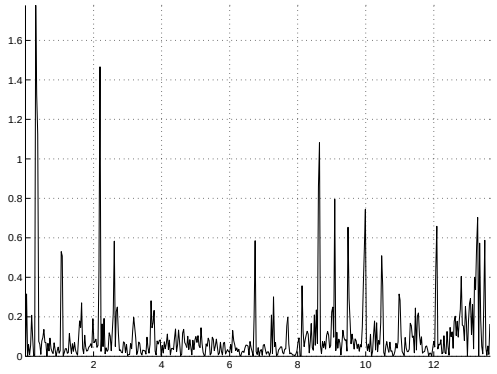


FIG. 6.19 – *Trajet de la « valeur absolue de la dérivée du flux calculé entre l'enveloppe spectrale* AR *(modèles d'ordre 6) et le spectre d'amplitude ». En abscisse : le temps en seconde*

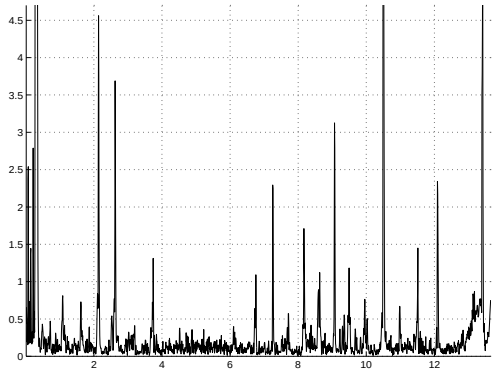


FIG. 6.20 – *Trajet du « flux spectral calculé sur toutes les fréquences de deux enveloppes* AR *(modèles d'ordre 6) successives ». En abscisse : le temps en seconde*

Nous présentons respectivement sur les figures 6.21 et 6.22 les trajets obtenus avec les deux formes de « flux spectral » qui existent quand nous utilisons les « enveloppes spectrales basées sur le cepstre » (voir la section 2.4.3.4).

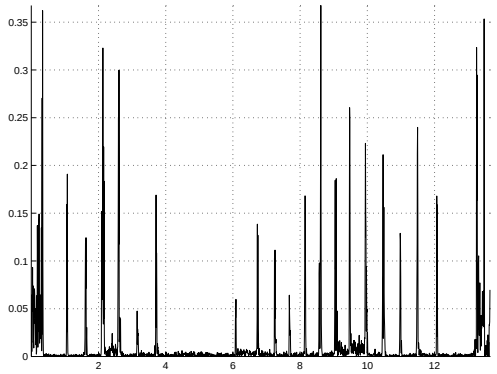


FIG. 6.21 – *Trajet de la « valeur absolue de la dérivée du flux calculé entre le spectre d'amplitude reconstruit après liftrage et le spectre d'amplitude ». En abscisse : le temps en seconde*

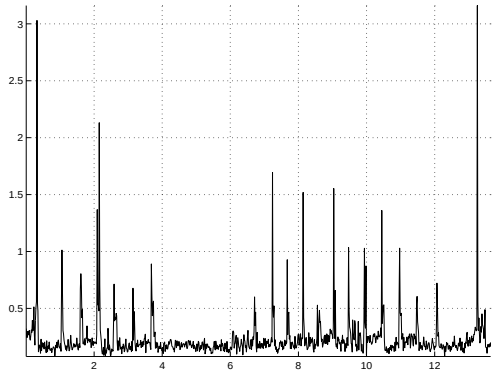


FIG. 6.22 – *Trajet du « flux spectral calculé entre deux spectres d'amplitude reconstruits après liftrage ». En abscisse : le temps en seconde*

Pour les trois « flux » calculés avec les « enveloppes spectrales basées sur les maximums », voir les sections 12.3.7.2 et 12.3.9.

6.3.2.4 Autres fonctions d'observation

Le trajet de la « valeur absolue de la dérivée du centroïde » (voir la section 2.5.1) est donné sur la figure 6.23.

Le trajet obtenu pour le « test de BRANDT » (voir la section 2.5.3) est présenté sur la figure 6.24.

Pour l'« analyse de la stationnarité » (voir la section 2.5.4), avec $\Theta = [0]$ nous obtenons le trajet donné sur la figure 6.11 ; et avec :

$$\Theta = [0\ 1\ 3\ 6\ 10\ 15\ 21\ 28\ 36\ 45\ 55\ 66\ 78\ 91\ 105]$$

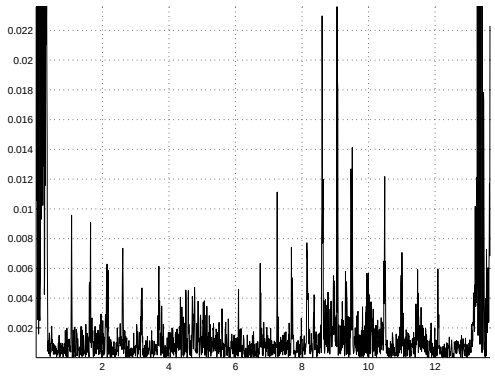


FIG. 6.23 – *Trajet de la « valeur absolue de la dérivée du centroïde ». En abscisse : le temps en seconde*

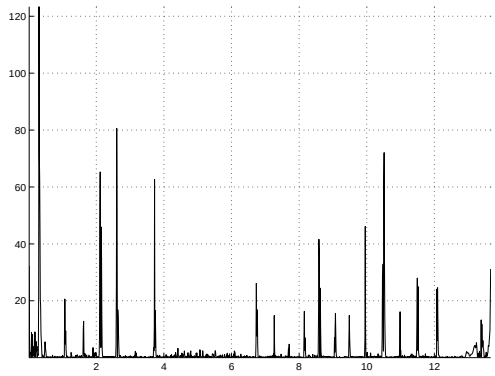


FIG. 6.24 – *Trajet obtenu avec le « test de* BRANDT *(modèles d'ordre 1) ». En abscisse : le temps en seconde*

nous obtenons le trajet présenté sur la figure 6.25.

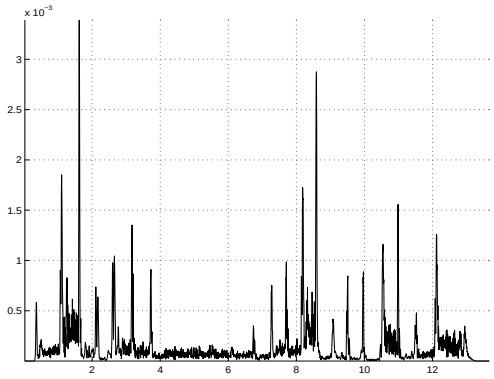


FIG. 6.25 – *Trajet de la « somme des valeurs absolues des dérivées de certains coefficients d'autocorrélation ». En abscisse : le temps en seconde*

6.3.3 L'interface graphique du programme *segmentation*

Les trois premières étapes de l'analyse *segmentation en zones stables* sont représentées chacune dans une « fenêtre ». Ces fenêtres sont données respectivement sur les figures 6.26 (fonctions d'observation), 6.27 (fonctions de décision) et 6.28 (fonction de décision finale). Nous donnons de plus, sur la figure 6.29, la fonction de décision finale obtenue si aucun traitement des marques trop proches n'est effectué (voir la section 4.3).

6.3.4 Transcription automatique

Les résultats de la transcription automatique pour l'extrait de flûte sont donnés sur la figure 6.30. Sont représentés sur cette figure le trajet de f_0, les marques de segmentation trouvées automatiquement et les notes trouvées automatiquement.

Nous donnons dans le tableau 6.1 les fréquences et les notes trouvées pour chacun des segments avec chacune des mesures. Ces trois mesures sont toujours d'accord, sauf pour le premier et le dernier segment, qui correspondent à du bruit.

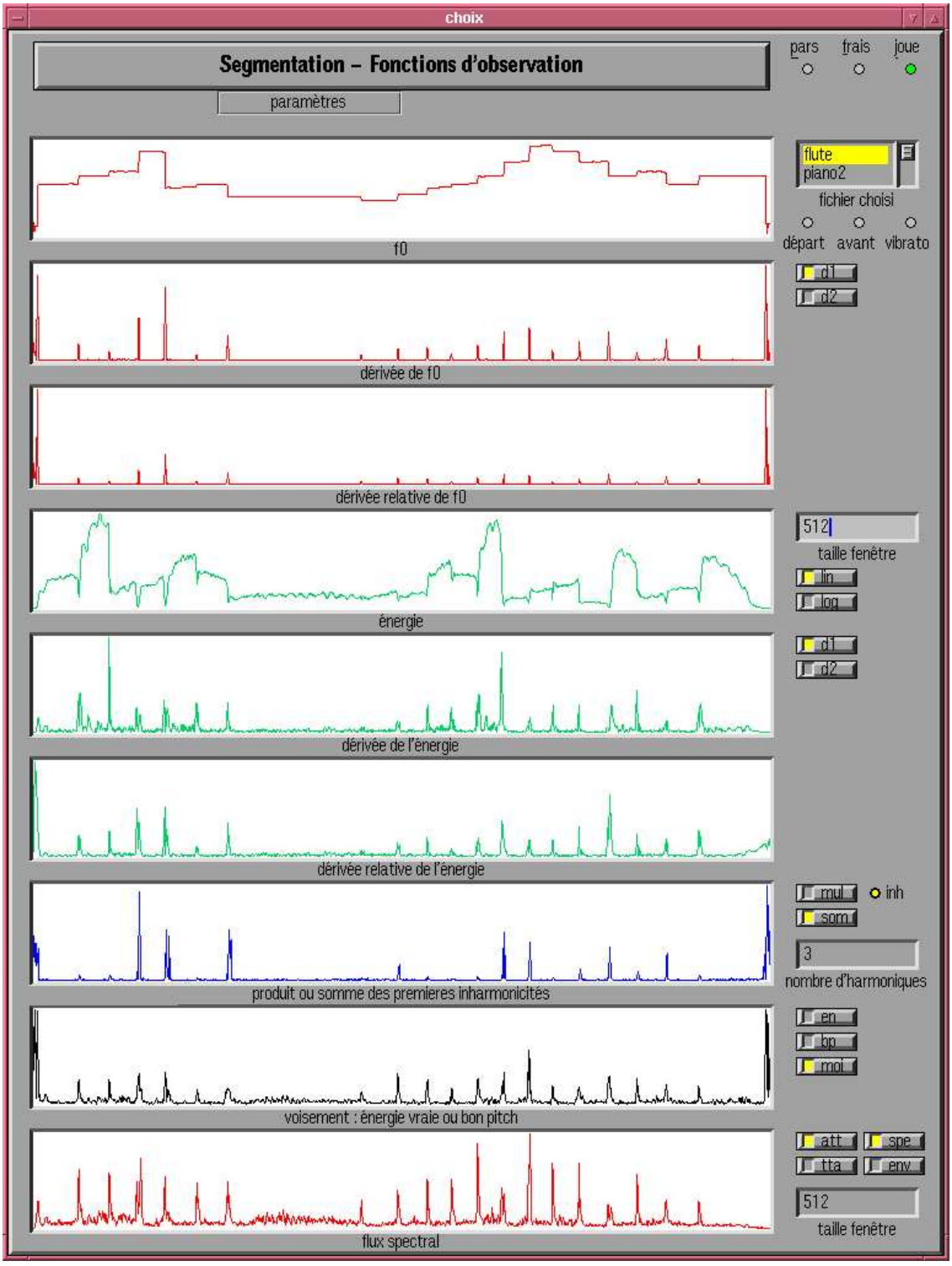


FIG. 6.26 – *Les fonctions d'observation utilisées par le programme segmentation sont présentées. L'extraction de fonctions d'observation est la première étape de l'analyse « segmentation en zones stables ». Du haut en bas, nous avons les trajets : de f_0, de la valeur absolue de la dérivée de f_0, de la valeur absolue de la dérivée relative de f_0, de l'énergie, de la valeur absolue de la dérivée de l'énergie, de la valeur absolue de la dérivée relative de l'énergie, de la somme des valeurs absolues des dérivées des indices d'inharmonicité, de la somme des valeurs absolues des dérivées des indices de voisement première forme, et du flux spectral calculé avec les spectres d'amplitude*

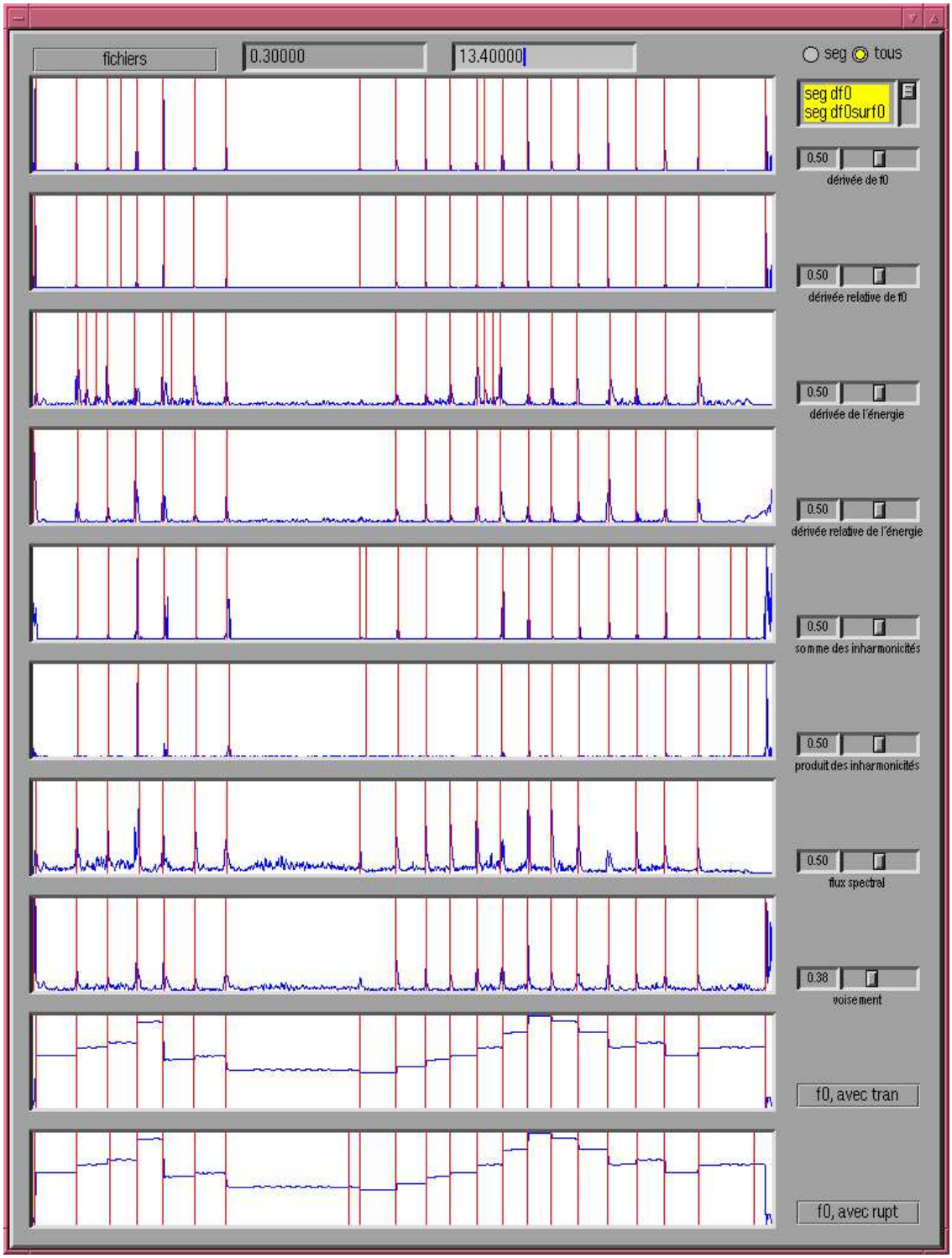


FIG. 6.27 – *Les fonctions de décision obtenues par seuillage automatique des fonctions d'observation sont présentées. Ces prises de décision sont la deuxième étape de l'analyse « segmentation en zones stables ». Les trajets des fonctions d'observation sont donnés. Les lignes verticales sont les marques de segmentation trouvées. Du haut en bas, nous avons les résultats pour : la valeur absolue de la dérivée de f_0, la valeur absolue de la dérivée relative de f_0, la valeur absolue de la dérivée de l'énergie, la valeur absolue de la dérivée relative de l'énergie, la somme des valeurs absolues des dérivées des indices d'inharmonicité, le produit des valeurs absolues des dérivées des indices d'inharmonicité, la somme des valeurs absolues des dérivées des indices de voisement première forme, le flux spectral calculé avec les spectres d'amplitude, l'analyse statistique sur f_0 et la rupture de modèles sur f_0. Pour chaque fonction d'observation, des fausses alarmes et des marques manquantes sont observées. Nous constatons, qu'en ce qui concerne ces fausses alarmes et ces marques manquantes, les fonctions d'observation ne réagissent pas de la même manière*

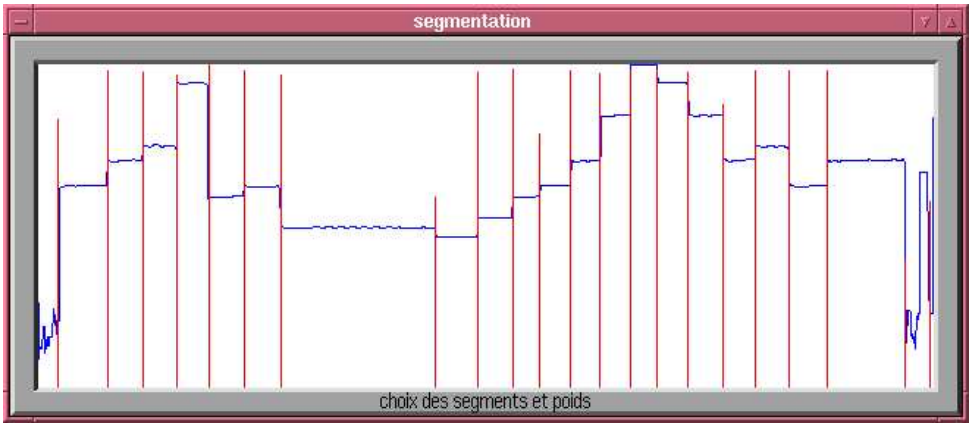


FIG. 6.28 – *La fonction de décision finale est présentée. Cette prise de décision finale est la troisième étape de l'analyse « segmentation en zones stables ». Le trajet de f_0 est donné. Les lignes verticales sont les marques de segmentation trouvées. La hauteur des lignes représente la confiance accordée à chaque marque. Ici : $T = 0{,}1$, $TG = 1{,}0$ et le seuil final $S_F = 0{,}4$ (voir le chapitre 4)*

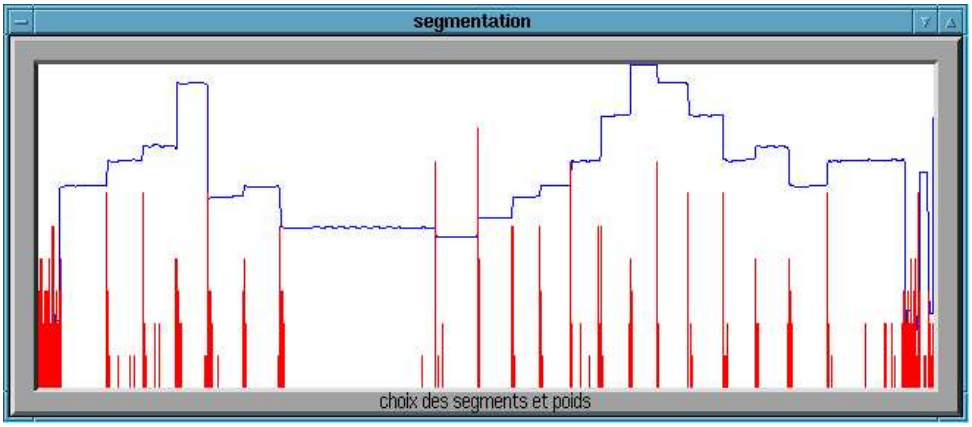


FIG. 6.29 – *La fonction de décision finale est présentée. Cette fois, $T = 0$, $TG = 100$ (c'est-à-dire qu'aucun traitement des marques trop rapprochées n'est effectué : toutes les marques sont gardées) et le seuil final vaut 1 (c'est-à-dire que toutes les marques sont gardées)*

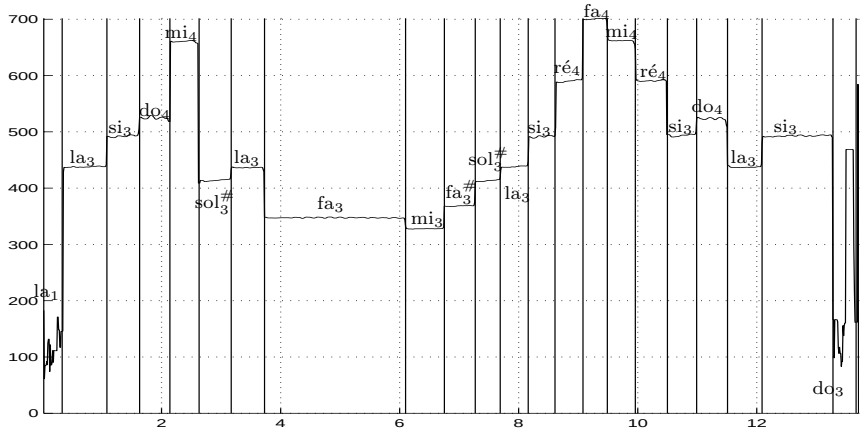


FIG. 6.30 – *Résultats de la transcription automatique pour l'extrait de flûte*

moyenne		médiane		moyenne du centre	
115,65	$la_1^{\#}$	111,35	la_1	110,10	la_1
430,66	la_3	437,52	la_3	437,58	la_3
493,08	si_3	492,35	si_3	492,43	si_3
524,45	do_4	524,44	do_4	525,47	do_4
647,50	mi_4	660,50	mi_4	660,99	mi_4
414,01	$sol_3^{\#}$	413,59	$sol_3^{\#}$	413,48	$sol_3^{\#}$
434,62	la_3	436,29	la_3	436,26	la_3
347,31	fa_3	347,28	fa_3	347,57	fa_3
328,47	mi_3	327,79	mi_3	327,78	mi_3
369,13	$fa_3^{\#}$	368,53	$fa_3^{\#}$	368,28	$fa_3^{\#}$
413,18	$sol_3^{\#}$	413,55	$sol_3^{\#}$	413,03	$sol_3^{\#}$
437,83	la_3	437,74	la_3	437,89	la_3
490,61	si_3	491,72	si_3	491,70	si_3
588,78	$ré_4$	590,28	$ré_4$	590,03	$ré_4$
699,86	fa_4	700,28	fa_4	700,68	fa_4
661,75	mi_4	661,89	mi_4	661,92	mi_4
588,23	$ré_4$	590,46	$ré_4$	590,39	$ré_4$
494,30	si_3	494,07	si_3	493,21	si_3
521,37	do_4	523,09	do_4	523,56	do_4
438,01	la_3	437,01	la_3	436,80	la_3
489,87	si_3	492,41	si_3	492,81	si_3
252,63	si_2	166,28	mi_2	261,16	do_3

TAB. 6.1 – *Fréquence pour chaque segment trouvé automatiquement avec le programme segmentation et note jouée correspondante*

		do_2	130,81	do_3	261,62	do_4	523,24
		$do_2^{\#}$	138,59	$do_3^{\#}$	277,18	$do_4^{\#}$	554,36
		$ré_2$	146,83	$ré_3$	293,66	$ré_4$	587,32
		$ré_2^{\#}$	155,565	$ré_3^{\#}$	311,13	$ré_4^{\#}$	622,26
		mi_2	164,815	mi_3	329,63	mi_4	659,26
		fa_2	174,615	fa_3	349,23	fa_4	698,46
		$fa_2^{\#}$	184,995	$fa_3^{\#}$	369,99	$fa_4^{\#}$	739,98
		sol_2	195,995	sol_3	391,99		
$sol_1^{\#}$	103,8275	$sol_2^{\#}$	207,655	$sol_3^{\#}$	415,31		
la_1	110	la_2	220	la_3	440		
$la_1^{\#}$	116,54	$la_2^{\#}$	233,08	$la_3^{\#}$	466,16		
si_1	123,4725	si_2	246,945	si_3	493,89		

TAB. 6.2 – *Fréquence fondamentale et note correspondante*

	type	durée	nbre de transitions	nbre de bd	nbre de fa
flûte	harmonique	13,685 s	21	21	3
clarinette	harmonique	20 s	31	30	3
violon	harmonique	15 s	16	16	15
voiceP	voix chantée	25,572607 s	21	21	9
piano	harmonique	4,458219 s	2	2	3
castagnettes	percussion	7,72512 s	43	42	0

TAB. 6.3 – *Performances de la segmentation avec une seule fonction d'observation (valeur absolue de la dérivée de f_0 pour les sons harmoniques et la voix chantée ; valeur absolue de la dérivée de l'énergie pour les percussions) – Sons de l'*IRCAM

6.4 Résumé des performances du système complet pour les sons des bases de données

6.4.1 Sons de l'IRCAM

Voir les tableaux 6.3 et 6.4. Dans la troisième colonne, nous donnons la durée du son considéré ; dans la quatrième, le nombre de transitions à détecter ; dans la cinquième, le nombre de transitions détectées, appelé dans le tableau « nbre de bd » (bd pour « bonnes détections ») ; et dans la sixième, le nombre de fausses alarmes, appelé dans le tableau « nbre de fa » (fa pour « fausses alarmes »). Faisons quelques remarques :

- L'extrait de flûte est un son « simple ». Ce son a été enregistré en salle anéchoïque, donc la réverbération est très réduite. De plus, ce son est parfaitement harmonique et monophonique, et il est quasi non modulé. Utiliser plusieurs fonctions d'observation permet de diminuer le nombre de fausses alarmes.
- L'extrait de clarinette est un son « simple ». Il comprend des notes extrêmement courtes.
- L'extrait de violon ayant été enregistré à un niveau très faible, il est très bruité. Un vibrato est présent. De plus, nous entendons le bruit des feuilles de partition quand elles sont tournées (le signal n'est plus tout à fait monophonique). Le nombre de fausses alarmes passe de 15 à 2. L'apport de la fusion des résultats obtenus avec plusieurs fonctions d'observation est montré.
- Pour l'extrait de voix chantée, un vibrato très important est présent. La même voyelle est chantée tout du long de l'extrait. Pour obtenir les résultats du tableau 6.3, nous utilisons la valeur absolue de la dérivée du trajet de f_0 sans suppression du vibrato (la « suppression du vibrato » est présentée dans la partie III, chapitre 15). Pour obtenir les résultats du tableau 6.4, nous segmentons avec les fonctions d'observation basées sur f_0 une fois que le vibrato a été supprimé. L'apport de la suppression du vibrato sur le trajet de f_0 est montré : nous passons de 9 fausses alarmes à 0.
- La note de piano n'est pas tout à fait harmonique. Le bruit de l'étouffoir quand il se referme sur la corde à la fin de la note est audible. Remarquons que les dérivées de l'énergie nous permettent de détecter cet instant.
- L'extrait de castagnettes n'est pas harmonique. Pour segmenter, nous n'avons pas utilisé les fonctions d'observation basées sur le trajet de f_0. Les dérivées de l'énergie et le flux spectral donnent de bons résultats.

6.4.2 Sons du CD Sqam

Voir les tableaux 6.5 et 6.6. Dans la troisième colonne, nous donnons la durée du son considéré ; dans la quatrième, le nombre de transitions à détecter ; dans la cinquième, le nombre de transitions détectées ; et dans la sixième, le nombre de fausses alarmes. Faisons quelques remarques :

- Ce son de gong artificiel est formé de douze fois la même note. Chaque note est composée d'une seule sinusoïde, amortie exponentiellement avec le temps. La fréquence de cette sinusoïde est $f_0 = 100\ Hz$. Les notes sont rassemblées en quatre groupes de trois notes très rapprochées.

	type	durée	nbre de transitions	nbre de bd	nbre de fa
flûte	harmonique	13,685 s	21	21	0
clarinette	harmonique	20 s	31	31	3
violon	harmonique	15 s	16	16	2
voiceP	voix chantée	25,572607 s	21	20	0
piano	harmonique	4,458219 s	2	2	0
castagnettes	percussion	7,72512 s	43	42	0

TAB. 6.4 – *Performances de la segmentation avec le système complet – Sons de l'*IRCAM

Les groupes sont séparés par des silences de durée 1 s. Pour chaque groupe, nous avons 4 transitions à détecter. La durée de chaque note est 1,3 s. Les indices d'inharmonicités ne peuvent pas être utilisés, puisque chaque note n'est formée que d'une seule sinusoïde.

- Ce son de gong artificiel est formé de huit fois la même note. Chaque note est composée d'une seule sinusoïde, amortie exponentiellement avec le temps. La fréquence de cette sinusoïde est $f_0 = 475\ Hz$. Un vibrato est présent. L'amplitude du vibrato est 20 Hz. La durée de chaque note est 1,3 s. Un silence de 0,5 s sépare deux notes successives. Ainsi, nous avons 16 transitions à détecter. Les indices d'inharmonicités ne peuvent pas être utilisés, puisque chaque note n'est formée que d'une seule sinusoïde.
- Pour cet extrait de violon, un vibrato, très petit, est présent. Les notes jouées (arpèges) vont du sol_2 ($f_0 = 196\ Hz$) au sol_5 ($f_0 = 1568\ Hz$).
- Les notes jouées par la contrebasse sont très graves : les fréquences fondamentales sont comprises entre 61,7 Hz (si_0) et 392 Hz (sol_3).
- Les notes jouées par le hautbois sont plutôt aiguës (les fréquences fondamentales sont comprises entre 293,66 Hz (ré_3) et 1174,64 Hz (ré_5).
- L'extrait de clarinette (arpèges) est un son simple. Les notes sont longues. Aucun problème.
- L'extrait de clarinette basse (arpèges) est un son simple. Les notes sont longues. Aucun problème. La marque manquante quand nous utilisons le système complet est la dernière, c'est-à-dire celle correspondant à la fin de la dernière note, dont la chute est lente.
- L'extrait de contre-basson (arpèges) est un son simple. Il est composé de notes très graves : entre 32,7 Hz (do_0) et 130,8 Hz (do_2). La taille t_{SIG} des fenêtres d'analyse doit être choisie plus grande : pour le do_0, t_{SIG} doit être de l'ordre de 80 millisecondes.
- L'extrait de saxophone (arpèges) est un son simple. Les deux fausses alarmes qui apparaissent quand nous segmentons seulement avec la « valeur absolue de la dérivée de f_0 » ont lieu lors de la dernière note, où un léger vibrato est présent.
- L'extrait de saxophone (mélodie) est un son simple. Pour les deux extraits de saxophone, la chute de la dernière note est très lente, il est donc difficile de déterminer à quel moment elle finit. Les résulats donnés par le logiciel f_0 deviennent de plus en plus chahutés. Pour cet extrait, l'une des fausses alarmes qui apparaît quand nous segmentons seulement avec la « valeur absolue de la dérivée de f_0 » est due à cette décroissance lente.
- L'extrait de grosse caisse est composé de six coups séparés par environ trois secondes. Seules les « valeurs absolues des deux dérivées de l'énergie » sont utilisables. Chacune des deux fonctions d'observation nous donne un grand nombre de fausses alarmes. Une fois la fusion de données faite, elles sont éliminées.
- L'extrait de piano est composé de dix notes. Les indices d'inharmonicité sont inutilisables.
- L'extrait de timbale est composé de dix coups séparés par au moins une seconde et demie. Nous entendons une hauteur, mais aussi des battements : certaines sinusoïdes ont donc des fréquences très proches. Il y a des partiels perturbateurs. Le programme f_0 ne parvient pas à nous donner une fréquence fondamentale. Seules les valeurs absolues des deux dérivées de l'énergie sont utilisables. Chacune des deux fonctions d'observation nous donne un grand nombre de fausses alarmes. Une fois la fusion de données faite, elles sont éliminées.
- L'extrait d'accordéon est composé de vingt-quatre notes, dont certaines sont très courtes (moins de 100 millisecondes). Les « valeurs absolues des dérivées de l'énergie » et le « flux spectral » sont inutilisables.

	type	durée	nbre de transitions	nbre de bd	nbre de fa
gong électronique 1	harmonique	27 s	16	16	0
gong électronique 2	harmonique	23 s	16	16	0
violon (arpèges)	harmonique	13 s	11	11	10
contrebasse (arpèges)	harmonique	14 s	11	11	4
hautbois (arpèges)	harmonique	9,7 s	8	8	9
clarinette (arpèges)	harmonique	8 s	8	8	0
clarinette basse (arpèges)	harmonique	9 s	8	8	5
contre-basson (arpèges)	harmonique	8,5 s	8	8	4
saxophone (arpèges)	harmonique	6,5 s	10	10	2
saxophone (mélodie)	harmonique	11 s	8	8	4
grosse caisse	percussif	24 s	6	6	6
timbale	percussif	26 s	10	10	5
piano (arpèges)	quasi harmonique	14,5 s	11	11	20
accordéon (mélodie)	harmonique	9,1 s	25	23	3
orgue (arpèges)	harmonique	10 s	8	8	3

TAB. 6.5 – *Performances de la segmentation avec une seule fonction d'observation (valeur absolue de la dérivée de f_0 pour les sons harmoniques et la voix chantée ; valeur absolue de la dérivée de l'énergie pour les percussions) – Sons du* CD *Sqam*

- L'extrait d'orgue (arpèges) est composé de sept notes. Seules les fonctions d'observation basées sur le trajet de f_0 (« valeurs absolues des dérivées de f_0 », « somme des valeurs absolues des indices de voisement première forme ») sont utilisables.

6.5 Conclusion

En premier lieu, les performances de ce système pour *segmenter en zones stables* sont bonnes, mais devront être étudiées pour plus de sons, et surtout pour plus de types de sons. Ensuite, tel qu'il a été construit, ce système ne peut pas nous donner la position de chaque transition – ou de chaque centre de transition – avec une précision temporelle supérieure à 10 millisecondes (ceci correspond à la période d'échantillonnage des fonctions d'observation). Pour améliorer cette précision, il faudrait soit ne plus travailler avec des portions décalées, c'est-à-dire changer de modèle de signal (utiliser la rupture de modèles, etc.) ; soit ajouter un « post-traitement », c'est-à-dire une cinquième étape à l'analyse *segmentation en zones stables*. Enfin, en troisième perspective, il s'agirait de calculer une barre d'erreur sur la position de chaque transition.

	type	durée	nbre de transitions	nbre de bd	nbre de fa
gong électronique 1	harmonique	27 s	16	16	0
gong électronique 2	harmonique	23 s	16	16	0
violon (arpèges)	harmonique	13 s	11	11	1
contrebasse (arpèges)	harmonique	14 s	11	11	0
hautbois (arpèges)	harmonique	9,7 s	8	8	3
clarinette (arpèges)	harmonique	8 s	8	8	0
clarinette basse (arpèges)	harmonique	9 s	8	7	0
contre-basson (arpèges)	harmonique	8,5 s	8	8	0
saxophone (arpèges)	harmonique	6,5 s	10	10	0
saxophone (mélodie)	harmonique	11 s	8	8	0
grosse caisse	percussif	24 s	6	6	0
timbale	percussif	26 s	10	9	1
piano (arpèges)	quasi harmonique	14,5 s	11	11	0
accordéon (mélodie)	harmonique	9,1 s	25	25	1
orgue (arpèges)	harmonique	10 s	8	8	0

TAB. 6.6 – *Performances de la segmentation avec le système complet – Sons du* CD *Sqam*

Chapitre 7

Corrélations entre les fonctions d'observation

7.1 Introduction

L'un des objectifs est de réduire le nombre de fonctions d'observation utilisées. Ceci est valable aussi bien pour la *segmentation en zones stables* (dans cette partie : *en notes et/ou en phones*) que pour la *segmentation en sources* (voir la partie V). En effet, deux fonctions d'observation très corrélées ne nous apportent pas plus d'information que l'une des deux, seule : voir la section 2.6.5.

Il s'agit aussi d'étudier en soi les liens entre les fonctions d'observation, qui correspondent elles-mêmes à des attributs perceptifs (la hauteur, l'intensité ou le timbre) ou psychoacoustiques des sons (par exemple, le centroïde est relié à la « brillance » du son).

De plus, nous avons vu qu'il existe trois types de variations brusques (en f_0, en énergie et en contenu spectral). Dans le cas général de la *segmentation en zones stables*, il n'y a pas de raison pour que nous favorisions l'un de ces types de transition. Ainsi, il faudrait utiliser sensiblement autant de fonctions d'observation, décorrélées, pour chacun de ces types de variations brusques.

Pour étudier la « corrélation » (dépendance statistique) entre deux fonctions d'observation, trois mesures sont couramment utilisées :

- Le coefficient de corrélation. Il est défini dans la **deuxième section** (section 7.2) de ce chapitre.
- L'information mutuelle. Elle est définie dans la **troisième section** (section 7.3) de ce chapitre.
- Le test du χ^2. Il est défini dans la **quatrième section** (section 7.4) de ce chapitre.

Dans la **cinquième section** (section 7.5) de ce chapitre, nous donnons quelques particularités de chacune de ces mesures.

Dans la **sixième section** (section 7.6) de ce chapitre, nous donnons des résultats pour des signaux réels. Des résultats pour des signaux simulés sont présentés dans l'annexe H.

7.2 Le coefficient de corrélation

Le coefficient de corrélation entre deux variables aléatoires X et Y (ce sont deux des fonctions d'observation : nous avons pour chacune d'elles M observations) est égal à :

$$\rho = \frac{E\left[X^c Y^c\right]}{\sigma_x \sigma_y}$$

avec :

$$X^c = X - E[X],\ Y^c = Y - E[Y]$$

et :

$$\sigma_x^2 = E\left[|X - E[X]|^2\right] \text{ et } \sigma_y^2 = E\left[|Y - E[Y]|^2\right]$$

où l'opérateur E représente l'espérance de la variable aléatoire.

7.3 L'information mutuelle

7.3.1 Définition

L'information mutuelle de SHANNON $IM(X,Y)$ entre deux variables aléatoires X et Y est par définition égale à :

$$IM(X,Y) = H(X) + H(Y) - H(X,Y)$$

où $H(Z)$ représente l'entropie de la variable aléatoire Z. L'entropie est elle-même égale à :

$$H(Z) = -\int_{-\infty}^{+\infty} p_z(z) \log_2 (p_z(z))\, dz$$

où $p_z(z)$ est la densité de probabilité de Z.

Si $p_x(x)$, $p_y(y)$ et la densité de probabilité conjointe $p_{xy}(x,y)$ existent, $IM(x,y)$ est égale à la distance de KULLBACK-LEIBLER, c'est-à-dire à :

$$IM(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p_{xy}(x,y) \log_2 \left(\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}\right) dxdy$$

Dans notre cas, nous n'avons pas accès aux densités de probabilité analytiques. Aussi, nous les estimons à partir des histogrammes normalisés h_x, h_y et h_{xy} des variables aléatoires, et nous avons :

$$IM(X,Y) \simeq \sum_{i=1}^{N}\sum_{j=1}^{N} h_{xy}(i,j) \log_2 \left(\frac{h_{xy}(i,j)}{h_x(i)h_y(j)}\right)$$

où N est le nombre de cases des histogrammes (h_{xy} compte $N \times N$ cases). La normalisation est effectuée de telle façon que nous ayons : $\sum_I h_Z(I) = 1$.

Dans le cas où $Y = X$, nous obtenons $IM(X,Y) = H(X)$, qui ne vaut pas forcément 1.

7.3.2 Normalisation dans le cas de variables aléatoires uniformément réparties

Le coefficient de corrélation a pour ensemble de définition $[0 \ldots 1]$. Pour pouvoir comparer ses performances avec celles de l'information mutuelle, il faudrait que celle-ci fût à valeurs dans cet intervalle aussi.

$$IM(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \leq H(X) \leq \log_2(N)$$

Le maximum de l'entropie $H(X)$ a lieu quand la variable aléatoire est uniformément répartie. Ainsi, dans notre cas (estimation de la densité de probabilité avec l'histogramme), nous obtenons : $0 \leq H(X) \leq \log_2(N)$. Et, comme $H(X|Y) \leq H(X)$ (cela implique aussi que $IM(X,Y)$ est toujours > 0), le coefficient :

$$im(X,Y) = \frac{1}{\log_2(N)} \sum_{i=1}^{N}\sum_{j=1}^{N} h_{xy}(i,j) \log_2 \left(\frac{h_{xy}(i,j)}{h_x(i)h_y(j)}\right)$$

est à valeurs dans [0 ... 1] (quoique ce ne soit pas encore tout à fait sûr, puisque nous n'avons qu'une estimée de $IM(X,Y)$).

Dans le cas général, c'est-à-dire quand nous ne sommes pas dans le cas de variables aléatoires uniformément réparties, cette normalisation n'est plus efficace (c'est-à-dire que l'information mutuelle normalisée $im(X,X)$ peut être dans des cas extrêmes très petite).

7.4 Le test du χ^2

Soient les histogrammes h_x et h_y à N cases et h_{xy} à N^2 cases des deux variables aléatoires X et Y. Le test du χ^2 est égal à (nous utilisons les histogrammes normalisés : c'est-à-dire une estimée de la densité de probabilité) :

$$\chi^2(X,Y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(h_{xy}(i,j) - h_x(i)h_y(j))^2}{h_x(i)h_y(j)}$$

7.5 Quelques particularités de ces mesures de la corrélation

- Le coefficient de corrélation mesure plutôt des dépendances linéaires entre les variables : voir l'annexe H, où sont donnés des tests faits avec les signaux simulés. Il s'agit plutôt d'un inconvénient.
- L'information mutuelle $im(X,X)$ ne vaut pas forcément 1. Il s'agit plutôt d'un inconvénient.

7.6 Résultats pour quelques fonctions d'observation

Dans le tableau 7.1, fo_1 correspond à la « valeur absolue de la dérivée de la fréquence fondamentale », fo_2 à la « valeur absolue de la dérivée relative de la fréquence fondamentale », fo_3 à la « valeur absolue de la dérivée de l'énergie », fo_4 à la « valeur absolue de la dérivée relative de l'énergie », fo_5 à la « somme des valeurs absolues des dérivées des indices d'inharmonicité », fo_6 au « produit des valeurs absolues des dérivées des indices d'inharmonicité », fo_7 à la « somme des valeurs absolues des dérivées des indices de voisement première forme », fo_8 au « flux spectral calculé à partir des spectres d'amplitude la fonction d'atténuation de l'oreille étant prise en compte », fo_9 à l'« analyse statistique appliquée au trajet de f_0 ». Chaque case renferme trois chiffres : le premier correspond au coefficient de corrélation, le second à l'information mutuelle et le troisième au test du χ^2. Nous obtenons, pour l'extrait de flûte **flute.sf**, le tableau 7.1.

Faisons remarquer que comme l'information mutuelle n'est pas normalisée, elle ne peut se regarder que relativement, c'est-à-dire ligne par ligne dans le tableau 7.1, la référence pour chaque ligne i étant $im(fo_i,fo_i)$. Normaliser $im(fo_i,fo_j)$ par $im(fo_i,fo_i)$ ($im_{(n)}(fo_i,fo_j) = \frac{im(fo_i,fo_j)}{im(fo_i,fo_i)}$) assurerait que la mesure fût à valeurs dans [0 ... 1]. Le problème est que normaliser ainsi rendrait la mesure non symétrique, c'est-à-dire que nous aurions $im_{(n)}(fo_i,fo_j) \neq im_{(n)}(fo_j,fo_i)$!

Nous commentons ci-dessous les résultats du tableau 7.1. Le coefficient de corrélation, pour certaines cases, est écrit en plus gros et mis en gras. Il s'agit des cas où il est supérieur à 0,7 : nous disons alors que les deux variables aléatoires sont plutôt corrélées. Pour d'autres cases, le coefficient de corrélation, est écrit en plus gros et mis en italique. Il s'agit des cas où il est inférieur à 0,20 : nous disons alors que les deux variables aléatoires sont plutôt décorrélées.

- Les trois mesures sont le plus souvent en accord. Par exemple, considérons la ligne fo_6. Le coefficient de corrélation indique que fo_6 n'est pas corrélée avec fo_7, un peu moins encore avec fo_9 et encore moins avec fo_8. Nous obtenons la même hiérarchie avec les deux autres mesures. Ceci est vérifié presque partout pour toutes les lignes du tableau. Cela indique que les dépendances sont plutôt linéaires, c'est-à-dire que le coefficient de corrélation suffit pour juger du degré de corrélation entre les fonctions d'observation.

	fo_1	fo_2	fo_3	fo_4	fo_5	fo_6	fo_7	fo_8	fo_9
fo_1	0,999 0,059 1,000	**0,851** 0,017 0,201	*0,141* 0,000 0,000	0,390 0,011 0,061	0,211 0,011 0,061	*0,095* 0,008 0,166	**0,730** 0,031 0,216	*0,195* 0,009 0,020	0,314 0,019 0,023
fo_2		0,999 0,025 1,000	*0,068* 0,000 0,000	0,328 0,004 0,030	0,208 0,006 0,037	*0,072* 0,005 0,083	**0,724** 0,019 0,188	*0,147* 0,002 0,005	0,238 0,007 0,009
fo_3			0,999 0,107 1,000	0,442 0,021 0,059	*0,071* 0,001 0,000	*0,041* 0,000 0,000	0,239 0,005 0,010	*0,192* 0,009 0,016	0,430 0,037 0,040
fo_4				0,999 0,100 1,000	0,453 0,015 0,071	*0,198* 0,003 0,012	0,587 0,024 0,079	0,407 0,021 0,052	0,520 0,039 0,047
fo_5					0,999 0,091 1,000	**0,758** 0,015 0,230	0,468 0,018 0,105	0,300 0,017 0,066	0,374 0,032 0,039
fo_6						0,999 0,018 1,000	0,201 0,007 0,079	*0,091* 0,001 0,002	*0,185* 0,005 0,006
fo_7							0,999 0,108 1,000	0,376 0,027 0,057	0,488 0,043 0,052
fo_8								0,999 0,348 1,000	0,331 0,025 0,026
fo_9									0,999 0,564 1,000

TAB. 7.1 – *Mesures de la corrélation pour la flûte. Case par case, nous trouvons : le coefficient de corrélation en haut, l'information mutuelle au milieu et le test du χ^2 en bas*

- fo_1 et fo_2 sont très corrélées : il s'agit des deux dérivées du trajet de f_0. fo_1 est corrélée avec fo_7, qui correspond au voisement première forme. fo_7 est aussi une fonction d'observation basée sur le trajet de f_0. Par contre, fo_1 n'est corrélée ni avec fo_3 ni avec fo_8. fo_3 est basée sur l'énergie, fo_8 sur le contenu spectral. fo_1 n'est pas corrélée avec fo_6, qui est pourtant une fonction d'observation basée sur le trajet de f_0.

- Les mêmes résultats sont obtenus pour fo_2 et fo_1.

- fo_3 n'est pas corrélée avec les trois fonctions d'observation fo_5, fo_6 et fo_8. Les deux premières fonctions d'observation sont basées sur le trajet de f_0 et la troisième sur le contenu spectral, alors que fo_3 est basée sur le trajet de l'énergie.

- fo_4, qui est basée sur le trajet de l'énergie, et fo_6, qui est basée sur le trajet de f_0, ne sont pas corrélées.

- fo_5 et fo_6 sont très corrélées : il s'agit des deux fonctions d'observation basées sur les indices d'inharmonicité.

- fo_6, qui est basée sur le trajet de f_0, n'est pas corrélée avec fo_8, qui est basée sur le contenu spectral. fo_6 n'est pas non plus corrélée avec fo_9, qui est pourtant basée aussi sur le trajet de f_0.

Ainsi, pour conclure, nous pouvons dire que deux fonctions d'observation ayant pour but de réagir à deux types de transitions différents sont plutôt décorrélées. Deux fonctions d'observation qui ont pour mission de mettre en évidence le même type de transitions sont le plus souvent corrélées, mais ce n'est pas toujours le cas.

Chapitre 8

Remarque : nécessité de traitements particuliers (pour le vibrato par exemple)

8.1 Introduction

Nous donnons ici quelques résultats obtenus avec un son réel qui n'est plus « simple ». Ici, en l'occurence, les modulations sont importantes. Afin de constater les effets d'un vibrato sur la fonction d'observation « flux spectral calculé avec les spectres d'amplitude », voir la section 11.2.4 page 90, où ils sont montrés pour un signal simulé. Les effets du vibrato sur la fonction d'observation « valeur absolue de la dérivée de f_0 » pour un signal simulé sont donnés dans la section 15.1 page 137.

8.2 Performances de quelques fonctions d'observation pour l'extrait de voix chantée voiceP.sf

8.2.1 Description du son considéré

Il s'agit d'un extrait de voix chantée. Le vibrato est très important. Le trémolo lui aussi est important : pour la voix, la présence de vibrato implique la présence de trémolo, puisque les harmoniques suivent à peu près, en amplitude, les formants. Ainsi, la fréquence de modulation du trémolo est la même que la fréquence de modulation du vibrato. Tout du long de l'extrait, la chanteuse chante le même phone : la voyelle « a ». Elle utilise parfois le vibrato pour passer d'une note à l'autre. Au centre de l'extrait, la chanteuse reprend son souffle, durant environ une seconde. Les dernières notes sont chantées beaucoup plus fort que les premières. Le trajet de f_0 est donné sur la figure 8.1.

8.2.2 Figures

Les résultats, pour cet extrait de voix chantée, de la fonction d'observation « valeur absolue de la dérivée du trajet de f_0 » sont donnés sur la figure 15.4. Nous constatons que la présence d'un vibrato rend cette fonction d'observation inutilisable pour la *segmentation en zones stables*.

Les résultats de la fonction d'observation « indice de voisement deuxième forme calculé en utilisant le spectre d'amplitude », sont donnés sur la figure 8.2. Nous constatons que la présence d'un vibrato rend cette fonction d'observation inutilisable pour la *segmentation en zones stables*.

Les résultats de la fonction d'observation « flux spectral calculé avec les spectres d'amplitude » sont donnés sur la figure 8.3. Les résultats de la fonction d'observation « valeur absolue de la dérivée de l'énergie » sont donnés sur la figure 8.4.

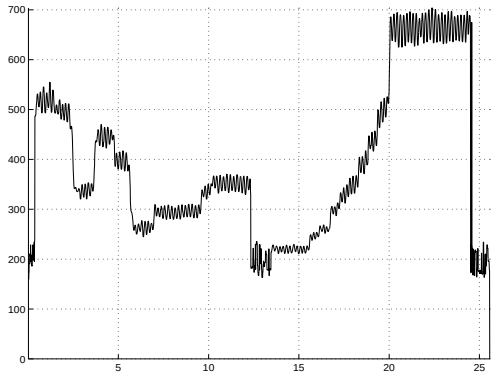


FIG. 8.1 – *Trajet de la fréquence fondamentale pour l'extrait de voix chantée* **voiceP.sf**. *En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

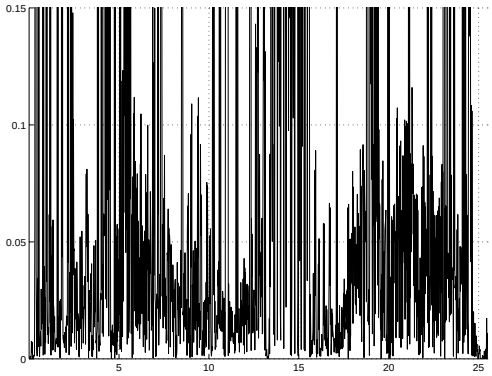


FIG. 8.2 – *Trajet de la valeur absolue de la « dérivée de l'indice de voisement deuxième forme calculé avec les spectres d'amplitude » pour l'extrait de voix chantée* **voiceP.sf**. *En abscisse : le temps en seconde*

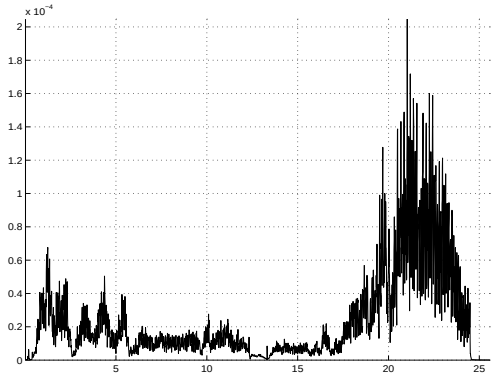


FIG. 8.3 – *Trajet du « flux spectral calculé avec les spectres d'amplitude » pour l'extrait de voix chantée* **voiceP.sf**. *En abscisse : le temps en seconde*

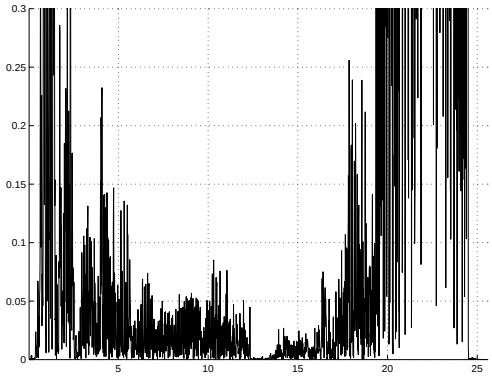


FIG. 8.4 – *Trajet de la « valeur absolue de la dérivée de l'énergie » (grossie entre 0 et 0,3) pour l'extrait de voix chantée* **voiceP.sf**. *En abscisse : le temps en seconde*

Chapitre 9

Conclusion de la deuxième partie

9.1 Bilan de la deuxième partie

Nous avons explicité les étapes que nous avons définies en ce qui concerne le processus de la *segmentation en zones stables*. Nous avons mis en place :

- Des **fonctions d'observation**. Nous demandons de chacune d'elles qu'elle présente un pic aussi grand et fin que possible lors des transitions, et un bruit de moyenne et de variance aussi petites que possible dans les zones stables. Les transitions sont classées en trois catégories, chacune de ces catégories correspondant à l'un des phénomènes de la perception communément utilisés pour caractériser psychoacoustiquement un son :
  - Les transitions en fréquence fondamentale.
  - Les transitions en énergie.
  - Les transitions en contenu spectral.

  Certaines de ces fonctions d'observation se décomposent en plusieurs dimensions. Il s'agit ensuite de fusionner ces dimensions. Des techniques de fusion de données sont utilisées. Elles consistent à calculer une « moyenne » de ces dimensions, de telle façon que la variance du bruit et la variance des pics à détecter soient les plus petites possibles.
  Quelques fonctions d'observation présentées dans la littérature tentent de mettre en évidence plusieurs de ces catégories de transitions simultanément. Nous avons choisi de les détecter les unes indépendamment des autres. Nous avons préféré cette procédure d'abord parce que l'utilisateur ne veut pas forcément détecter toutes les transitions : pour la voix chantée, il peut être soit intéressé par la segmentation en notes (transition en fréquence fondamentale), soit par la segmentation en phones (transition en contenu spectral). Il s'agit ensuite de rassembler les informations obtenues pour chaque catégorie de transition.
- Des techniques de prise de décision sont appliquées à ces fonctions d'observation pour obtenir des **fonctions de décision**. Ces fonctions de décision correspondent à des listes de marques de segmentation. La prise de décision pour chaque fonction d'observation se décompose en deux points :
  - Le premier point est la détermination automatique d'un seuil.
  - Le second point consiste à rassembler les marques trop proches les unes des autres. Les transitions n'étant pas instantanées, certaines fonctions d'observation présentent plusieurs pics au cours d'une même transition. En résumé, il s'agit, dans un groupe de marques à rassembler, de ne garder que la marque pour laquelle la fonction d'observation est la plus grande.
- Un **processus de prise de décision final** qui, à partir de l'ensemble des fonctions de décision particulières, détermine une fonction de décision unique. Là encore, ce processus se décompose en deux points :
  - Le premier point consiste à fusionner les fonctions de décision obtenues à l'étape de l'analyse *segmentation en zones stables* précédente. La règle de fusion utilisée est plus

compliquée que la règle de la majorité. Notons que pour cette règle de la majorité il est nécessaire d'utiliser un nombre impair de capteurs, à moins de déterminer un poids pour chaque capteur, poids correspondant à la confiance qui lui est accordée : ceci complique encore la fusion de données. Les effets des corrélations entre les fonctions d'observation sur la procédure de fusion des fonctions de décision sont présentés pour des signaux simulés.

- Le second point consiste encore une fois à rassembler les marques trop proches les unes des autres. En résumé, il s'agit de ne garder qu'une marque que nous plaçons au centre de gravité du groupe de marques à rassembler.

- Un **processus d'étiquetage**. Il s'agit de décrire le contenu de chaque segment. Ce processus a consisté ici, pour les sons monophoniques, harmoniques et non modulés, principalement en la transcription automatique. Notons que si les marques de segmentation sont mal placées, la transcription est nécessairement erronée. Des techniques pour éviter ce problème d'accumulation des erreurs sont proposées.

Le processus de *segmentation en zones stables* est, pour le moment, c'est-à-dire pour un son monophonique, harmonique et non modulé, celui présenté sur la figure 9.1.

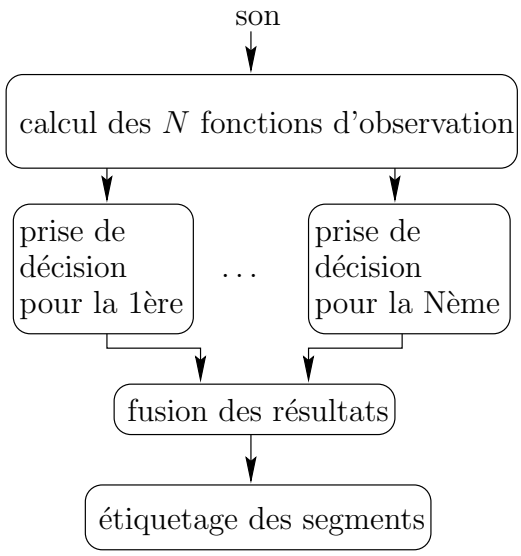


FIG. 9.1 – *Segmentation en zones stables d'un son monophonique, harmonique et non modulé : algorithme de base*

9.2 Perspectives pour les parties suivantes

Nous avons vu (chapitre 8) que nous rencontrons des problèmes lors de la *segmentation en zones stables* dès que le signal n'est plus monophonique, harmonique et/ou non modulé.

Les sons sont classées en plusieurs catégories. Pour chacune des catégories de son (son instrumental harmonique sans vibrato, voix chantée avec un vibrato et un trémolo importants, son polyphonique, son percussif...) :

- Certaines des fonctions d'observation décrites dans cette partie sont, pour la *segmentation en zones stables*, utilisables et d'autres non (ainsi, s'il est un vibrato important par exemple, les « fusions des valeurs absolues des dérivées des indices de voisements », et plusieurs autres fonctions d'observation, ne peuvent plus être utilisées).
- Certaines fonctions d'observation doivent être adaptées (par exemple, le vibrato doit être supprimé du trajet de f_0 avant que les « valeurs absolues des dérivées de f_0 » ne soient calculées : voir la partie III, chapitre 15).

L'un des objectifs sera, dans la suite de cet exposé, de déterminer la configuration de fonctions d'observation adéquate pour chaque catégorie de sons. Ceci nécessitera des traitements particuliers. Un autre objectif sera de résoudre les problèmes posés par certaines catégories de sons à certaines fonctions d'observation. Aussi, il s'agira d'abord de déterminer, avant de *segmenter en zones stables*, à quelle catégorie appartient le son considéré. Ceci nous amènera à définir deux autres niveaux de segmentation qui, hiérarchiquement, seront exécutés avant le niveau de *segmentation en zones stables* : le niveau de *segmentation en caractéristiques* (parties III et IV) et le niveau de *segmentation en sources* (partie V).

Pour déterminer la catégorie de sons à laquelle appartient un son donné, une série de questions est posée. Dans un premier temps, nous organisons cette série de question ainsi :

q_1 Parole (remarquons que souvent la parole est monophonique : le plus souvent une seule personne parle à la fois ; et que la parole est une succession de zones plutôt harmoniques et de zones plutôt de bruit)
q_1 ou Musique ?

q_2 *si Musique :* Monophonique

q_3 *si Monophonique :* Voisé (c'est-à-dire composé d'une somme de sinusoïdes [1])

q_4 *si Voisé :* Harmonique (sinusoïdes appartenant à un peigne harmonique empli d'une manière dense)
q_5 *si Harmonique :* Avec vibrato (et/ou trémolo)
q_5 ou Sans vibrato (ni trémolo) ?
q_4 ou Inharmonique (sinusoïdes n'appartenant pas à un peigne harmonique empli d'une manière dense) ?

q_3 ou Non voisé (c'est-à-dire pas composé de sinusoïdes : son percussif, bruit...) ?

q_2 ou Polyphonique ?

La première question (q_1) fait l'objet de la *segmentation en sources* (voir la partie V). Les autres (q_2, q_3, q_4 et q_5) font l'objet de la *segmentation en caractéristiques* (voir la partie III). Cette série de questions est schématique : elle a pour fonction de présenter très rapidement le contenu du reste de l'exposé.

Indiquons-le de nouveau : la procédure utilisée pour *segmenter en zones stables* doit être adaptée à chacune des catégories de sons, puisque la définition du terme « zone stable » est différente pour chacune d'elle. Avec les techniques qui ont été présentées pour le moment, nous pouvons segmenter les sons musicaux monophoniques, voisés harmoniques et non modulés.

1. Au sujet de la distinction que nous faisons entre « voisé » et « harmonique », voir la note de la page 57 et le chapitre 19 de la partie IV. Remarquons que dans cette partie IV nous décrirons une procédure pour détecter la polyphonie (q_2) basée sur l'étude simultanée d'un indice de voisement (q_3) et d'un indice d'inharmonicité (q_4). Ainsi, l'ordre des questions tel qu'il est proposé ici ne sera pas respecté. Mais ceci est dû à l'une des hypothèses restrictives que nous faisons dans le chapitre 18 : mixage de voix harmoniques aux peignes non superposés. Dans le cas général de la polyphonie, l'ordre sera rétabli (le traitement du cas polyphonique général est une perspective).

Troisième partie

Le problème du vibrato (Segmentation en caractéristiques)

Chapitre 10

Introduction

10.1 Segmentation en caractéristiques

L'objectif de la *segmentation en caractéristiques* est d'indexer des segments de son – ou encore : de poser des étiquettes sur des segments de son – dont la taille soit de l'ordre de la seconde, ceci principalement dans le but d'aider à la *segmentation en notes ou en phones (ou, plus généralement, en zones stables)* (voir la partie précédente). D'abord, il s'agit de déterminer quelles fonctions d'observation sont utilisables pour la *segmentation en zones stables*. Ensuite, pour certaines catégories de sons, il faut calculer certaines fonctions d'observation (l'extraction des fonctions d'observation est la *première étape* de la *segmentation en zones stables*) d'une façon différente de la façon utilisée pour les sons « simples », c'est-à-dire monophoniques, voisés harmoniques et non modulés.

Nous avons considéré les caractéristiques suivantes :

- Présence ou absence de vibrato. Rappelons que le vibrato correspond à une modulation de la fréquence. La détection du vibrato est basée soit directement sur l'analyse du signal sonore, soit sur l'analyse du trajet de f_0. Si un vibrato est détecté, nous déterminons ses paramètres (fréquence, amplitude, phase). Voir les chapitres 11 à 16.
- Présence ou absence de trémolo. Rappelons que le trémolo correspond à une modulation de l'amplitude. Si un trémolo est détecté, nous déterminons ses paramètres (fréquence, amplitude, phase). Il s'agit d'une perspective.
- Silence ou présence de son. Il s'agit d'une perspective. Cette détection est basée sur l'analyse du trajet de l'énergie. Mais, jugeant cette analyse insuffisante, nous la couplons à l'analyse du taux de passage par zéro. L'idée vient des considérations suivantes :

  1. Pendant un silence, l'énergie est très petite : en effet, est toujours présent un bruit ; et, si nous faisons l'hypothèse que ce bruit est blanc, le taux de passage par zéro est très grand.
  2. Si nous sommes en présence d'un signal qui n'est pas du bruit et dont l'amplitude est très faible (fin de la chute d'une note par exemple), l'énergie est très petite, mais le taux de passage par zéro lui aussi est petit.

  Dans [MC98], une technique basée seulement sur l'énergie est proposée. Dans [JEA99], une sorte d'inventaire (« survey ») des méthodes de détection des silences est donnée.
- Son harmonique ou inharmonique : cette détection est basée sur l'analyse des indices d'inharmonicité que nous avons définis dans la partie précédente.
- Son voisé ou non voisé : cette détection est basée sur l'analyse des indices de voisement que nous avons définis dans la partie précédente.

Ce niveau de segmentation concerne donc l'étiquetage de segments de sons avec ces caractéristiques. Cet étiquetage a pour base l'extraction de fonctions d'observation. Certaines de ces fonctions d'observation sont utilisées pour la *segmentation en zones stables*.

10.2 Le vibrato

Le cas du vibrato est développé en détails dans le reste de cette partie. Dans les chapitres suivants, nous nous attacherons aux problèmes de la **détection** du vibrato, de l'**estimation** de ses paramètres, et de sa **suppression** sur le trajet de f_0. L'absence ou la présence de vibrato est l'une des étiquettes définies dans la section 10.1. Le vibrato posant de grands problèmes quand il s'agit de *segmenter en zones stables*, nous nous sommes particulièrement intéressé à la résolution de ces problèmes.

10.2.1 Limite entre « signal avec vibrato » et « signal sans vibrato »

Dans [ZF81], pages 63 – 66, une limite psychoacoustique est donnée. Quand la valeur de l'amplitude A_{vib} de la modulation est plus petite qu'un certain seuil, elle n'est pas entendue. Il s'agit de ce qui est appelé le seuil différentiel de fréquence. Ce seuil dépend de la fréquence f_0 du son pur (« pur » veut dire parfaitement sinusoïdal). Au-dessous de 500 Hz, il est à peu près indépendant de la fréquence et vaut 1,8 Hz. Au-delà de 500 Hz, il croît à peu près linéairement avec la fréquence f_0 et vaut approximativement 3,5 $10^{-3} f_0$. Mais ce seuil dépend aussi de la fréquence de modulation f_{vib} : les mesures que nous venons d'indiquer sont valables pour $f_{vib} = 4\ Hz$; il dépend aussi du niveau sonore ; et il vaut pour les sons purs, pas pour les sons complexes (« complexe » veut dire composé d'une somme de sinusoïdes, de partiels harmoniques ou non). Nous n'avons donc pas utilisé ce critère pour la prise de décision.

10.2.2 Plan de la partie

Nous présentons dans le **deuxième chapitre** (chapitre 11) de cette partie le problème du vibrato.

Dans le **troisième chapitre** (chapitre 12) de cette partie, les trois méthodes basées sur l'analyse directe des signaux sonores que nous avons implémentées pour détecter le vibrato sont explicitées. Certaines d'entre elles nous permettent d'estimer les paramètres du vibrato.

Dans le **quatrième chapitre** (chapitre 13) de cette partie, les trois méthodes basées sur l'analyse des trajets de f_0 que nous avons implémentées pour détecter le vibrato sont explicitées. Certaines d'entre elles nous permettent d'estimer les paramètres du vibrato.

Dans le **cinquième chapitre** (chapitre 14) de cette partie, nous discutons et comparons les performances de ces méthodes sur des signaux réels.

Nous présentons dans le **sixième chapitre** (chapitre 15) de cette partie la méthode de suppression du vibrato sur le trajet de f_0 (pour l'aide à la *segmentation en zones stables* notamment) que nous avons utilisée.

Dans le **septième chapitre** (chapitre 16) de cette partie les techniques de fusion de données utilisées dans le cas du vibrato sont introduites.

Le **huitième chapitre** (chapitre 17) de cette partie constitue une conclusion à cette partie.

Chapitre 11

Le vibrato : présentation du problème

11.1 Introduction

Nous nous intéressons au vibrato d'abord dans le but de le **détecter**, puis d'**estimer** ses paramètres, et enfin de le **supprimer** sur le trajet de f_0. Ainsi, il est nécessaire de détecter le vibrato, d'estimer ses paramètres et de le supprimer sur le trajet de f_0 :

- Pour caractériser – étiqueter – le son (« il y a du vibrato », ou « il n'y a pas de vibrato » ; ou : « il y a plutôt du vibrato », ou « il n'y en a plutôt pas »), dans le cadre de MPEG-7 par exemple : ceci concerne la détection
- Pour aider à la *segmentation en zones stables* (dérivées du trajet de f_0) : ceci concerne la détection, et la suppression
- Pour des traitements ultérieurs (analyse du son ; modification des paramètres du vibrato ; resynthèse du son avec le nouveau vibrato) : ceci concerne la détection, l'estimation, et la suppression

Le vibrato détériore en fait les performances des algorithmes de *segmentation en zones stables* décrits dans la partie II. La nécessité de la suppression du vibrato sur le trajet de f_0 réside en ce que son amplitude peut être supérieure à celle du saut de fréquence entre deux notes. Nous donnons sur la figure 11.1 un exemple de ce que nous pourrions avoir à l'extrême. Le trajet de f_0 d'un son simulé est présenté. Nous avons modélisé le vibrato par une sinusoïde pure de fréquence 5 Hz et de phase 1,9 *rad*. Les deux notes jouées sont un la_3 ($f_0^c = 440$ Hz) suivi d'un $\text{la}_3^{\#}$ ($f_0^c = 466{,}16$ Hz). L'amplitude du vibrato est 30 Hz. Pour le modèle de transition en fréquence entre les deux notes utilisé, voir la section 11.2.3.1. Nous avons pris $a = 0{,}75$ et $b = 0{,}1$.

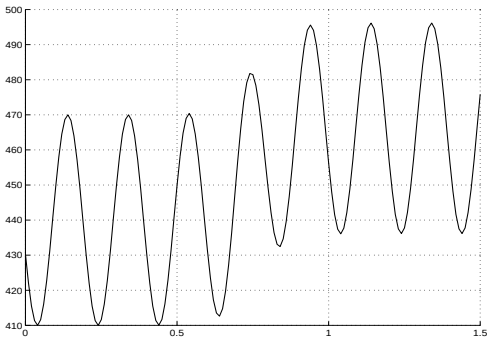


FIG. 11.1 – *Trajet de f_0 simulé. Un vibrato est présent. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

Quand un vibrato est présent la détection de la transition entre les deux notes (avec la fonction d'observation « valeur absolue de la dérivée de f_0 » par exemple) devient problématique (voir la figure 15.1, page 138). Nous devons supprimer ce vibrato sur le trajet de f_0 avant de calculer la fonction d'observation.

Mais il est intéressant aussi de pouvoir, pour un extrait sonore, supprimer le vibrato sur le trajet de f_0, ajouter un autre vibrato sur le trajet de f_0 une fois le vibrato initial supprimé, puis resynthétiser le son avec ce nouveau trajet de f_0.

Auparavant, il faut déterminer si un vibrato est présent ou non.

Pour une note donnée, le vibrato correspond à une variation quasi périodique de f_0 autour de sa valeur centrale. Cette variation n'est pas forcément sinusoïdale : il faut modéliser le vibrato comme une somme de sinusoïdes (d'harmoniques du vibrato !) dont les amplitudes et les fréquences instantanées varient dans le temps. La fréquence fondamentale du vibrato est en général comprise entre 3 Hz et 11 Hz.

Le trémolo correspond à une modulation de l'amplitude de même que le vibrato correspond à une modulation de la fréquence. Il faut le supprimer, lui, sur le trajet de l'énergie. La plupart des méthodes décrites dans cette partie pour résoudre le problème du vibrato sont adaptables au cas du trémolo.

Mais, auparavant, nous allons présenter, dans la section suivante (section 11.2), les modèles du trajet de f_0 quand un vibrato est présent mis en place.

11.2 Modèles utilisés

11.2.1 Modèle complet du vibrato sur une note

Nous nous intéressons dans cette section à la modélisation du trajet de la fréquence fondamentale f_0 quand un vibrato est présent. La fréquence fondamentale $f_0(t)$ instantanée à l'instant t s'écrit :

$$f_0(t) = f_0^c(t) + \sum_{k=1}^{p(t)} A_{vib(k)}(t) \cos(\phi_{(k)}(t)) + b(t)$$

où :

- $f_0^c(t)$, est la composante continue, représentant la hauteur du son, la note jouée : il s'agit de ce que nous voulons retrouver après suppression du vibrato sur le trajet de f_0 (voir le chapitre 15).
- $p(t)$, est le nombre d'harmoniques du vibrato pris en compte.
- $A_{vib(k)}(t)$, est l'amplitude instantanée de l'harmonique du vibrato de numéro d'ordre k.
- $\phi_{(k)}(t)$, est la phase instantanée de l'harmonique du vibrato de numéro d'ordre k. Nous avons :

$$\phi_{(k)}(t) = \phi_{(k)}(t - \Delta t) + 2\pi k \int_{t-\Delta t}^{t} f_{vib}(t)dt$$

- Δt, est la période d'échantillonnage du trajet de f_0.
- $f_{vib}(t)$, est la fréquence instantanée du premier harmonique du vibrato.
- $b(t)$, est un résidu de modélisation.

11.2.2 Modèle simplifié du vibrato sur une note

Dans cet exposé, nous avons essayé d'extraire seulement la « fréquence fondamentale » f_{vib} du vibrato[1]. De plus, si nous faisons l'hypothèse que la hauteur du son est stable sur une note, que les paramètres du vibrato (fréquence, amplitude, phase) ne changent pas sur une note, et qu'il n'y a pas de résidu de modélisation, nous avons, pour chaque note :

$$f_0(t) = f_0^c + A_{vib}\cos(2\pi f_{vib}t + \varphi_{vib})$$

où φ_{vib} est la phase à l'origine du vibrato ; et l'évolution de la phase du signal $s(t)$ s'écrit :

$$\phi(t) = 2\pi\int_0^t f_0(t)dt = 2\pi\left[f_0^c t + \frac{A_{vib}}{2\pi f_{vib}}\sin\left(2\pi f_{vib}t + \varphi_{vib}\right)\right]$$

Alors : $s(t) = \cos\left(\phi(t)\right)$

11.2.3 Modélisation de la transition entre deux notes harmoniques

11.2.3.1 Pour la fréquence

Nous modélisons le saut en fréquence entre deux sinusoïdes de fréquences respectives $f_0^{(1)}$ et $f_0^{(2)}$ par :

$$f_0(t) = f_0^{(1)} + \left[\tanh\left(\frac{t-a}{b}\right) + 1\right]c$$

où a représente le moment où le saut en fréquence a lieu, b la rapidité à laquelle ce saut se fait, et c l'amplitude du saut. Ainsi, $c = \dfrac{f_0^{(2)} - f_0^{(1)}}{2}$. Quand x tend vers $-\infty$, $\tanh(x)$ tend vers -1 ; et quand x tend vers $+\infty$, $\tanh(x)$ tend vers $+1$: ainsi, $f_0(t)$ passe de $f_0^{(1)}$ à $f_0^{(2)}$. La phase s'écrit :

$$\phi(t) = 2\pi\int_0^t f_0(t)dt + \varphi_1 = 2\pi f_0^{(1)}t + 2\pi ct + 2\pi cb\left\{\log_e\left[\cosh\left(\frac{t-a}{b}\right)\right] - \log_e\left[\cosh\left(-\frac{a}{b}\right)\right]\right\} + \varphi_1$$

Pour une somme $s(t)$ de L sinusoïdes harmoniques, la fréquence fondamentale variant comme indiqué ci-dessus, nous avons :

$$s(t) = \sum_{l=1}^{L} B_l(t)\cos\left(2\pi f_0^{(1)}lt + 2\pi clt + 2\pi cbl\left\{\log_e\left[\cosh\left(\frac{t-a}{b}\right)\right] - \log_e\left[\cosh\left(-\frac{a}{b}\right)\right]\right\} + \varphi_l\right)$$

Si, en plus, un vibrato est présent, de fréquence et d'amplitude fixes, nous obtenons :

$$\begin{aligned} s(t) \quad = \quad & \\ & \sum_{l=1}^{L} B_l(t)\cos\left(2\pi\left(f_0^{(1)}lt + clt + cbl\left\{\log_e\left[\cosh\left(\frac{t-a}{b}\right)\right] - \log_e\left[\cosh\left(-\frac{a}{b}\right)\right]\right\}\right) + \right. \\ & \left. \frac{lA_{vib}}{f_{vib}}\sin\left(2\pi f_{vib}t + \varphi_{vib}\right) + \varphi_l\right) \end{aligned}$$

Avec $f_0^{(1)} = 440\ Hz$ (la_3), $f_0^{(2)} = 493{,}89\ Hz$ (si_3), $a = 1{,}07$ et $b = 0{,}005$, nous obtenons le trajet de la fondamentale présenté sur la figure 11.2.

Les variations d'amplitude (modulation, naissance ou mort d'un partiel) sont modélisées dans les $B_l(t)$: voir la section 11.2.3.2.

1. Voir la section 13.5.2, où nous montrons pour un signal sonore réel que les amplitudes des harmoniques du vibrato de numéros d'ordre supérieurs sont très petites, tellement petites que nous pouvons considérer que ces harmoniques du vibrato sont absents.

11.2.3.2 Pour l'amplitude

Nous faisons l'hypothèse que l'amplitude de chaque partiel passe de c_1 (amplitude au cours de la première note) à c_2 (amplitude au cours de la seconde note), en s'approchant de 0 au « moment » de la transition. c_1 et c_2 peuvent être modulés (trémolo). Nous modélisons l'amplitude $B(t)$ de chaque partiel ainsi :

$$B(t) = \left[1 - \tanh\left(\frac{t - a_1}{b_1}\right)\right]\frac{c_1}{2} + \left[1 + \tanh\left(\frac{t - a_2}{b_2}\right)\right]\frac{c_2}{2}$$

L'amplitude du partiel passe par son minimum entre a_1 et a_2 (donc a_2 est plus grand que a_1). b_1 représente la rapidité de la chute de la première note, et b_2 la rapidité de l'attaque de la seconde note. Avec $a_1 = 1{,}065$, $b_1 = 0{,}005$, $c_1 = 0{,}022$ et $a_2 = 1{,}085$, $b_2 = 0{,}025$, $c_2 = 0{,}04$, nous obtenons le trajet de l'amplitude présenté sur la figure 11.3.

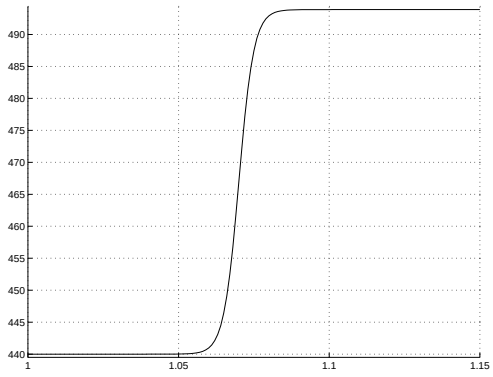


FIG. 11.2 – *Modèle du trajet de la fréquence d'un partiel lors d'un changement de note. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

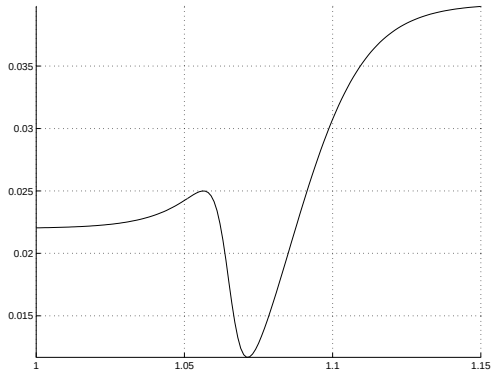


FIG. 11.3 – *Modèle du trajet de l'amplitude d'un partiel lors d'un changement de note. En abscisse : le temps en seconde ; en ordonnée : l'amplitude*

Considérons une portion de l'extrait de flûte **flute.sf**. Ce signal a été enregistré en salle anéchoïque, donc la réverbération est nulle : ainsi, la fin d'une note ne se superpose pas au début de la note qui suit. La portion de signal considérée couvre le premier changement de note : nous passons d'un la_3 ($f_0^c = 440\ Hz$) à un si_3 ($f_0^c = 493{,}89\ Hz$). Elle est donnée sur la figure 11.4. Les figures 11.5 et 11.6 présentent respectivement le trajet en fréquence et le trajet en amplitude de la fondamentale pour cette portion du signal. La fréquence de la fondamentale et son amplitude ont été déterminées à partir du spectre d'amplitude calculé sur des fenêtres d'analyse larges de 6 millisecondes, ce qui représente à peu près 2,6 périodes pour la première note et 2,9 pour la seconde. Puisque $f_e = 32000\ Hz$, chaque fenêtre d'analyse est large de 192 échantillons. L'échantillon fréquentiel p pour lequel nous avons le maximum du spectre d'amplitude entre 0 et 750 Hz est une première estimation de la fréquence fondamentale. La valeur x_p du spectre d'amplitude pour l'échantillon fréquentiel p est une première estimation de l'amplitude de la fondamentale. Nous faisons passer un polynôme d'ordre 2 par les échantillons fréquentiels de numéros d'ordre $p-1$, p et $p+1$. L'endroit où sa dérivée s'annule nous donne une estimation plus précise de la fréquence fondamentale, et la valeur de ce polynôme à cet endroit nous donne une estimation plus précise de l'amplitude de la fondamentale.

11.2.4 Un exemple : influence du vibrato sur les performances du flux spectral pour un son simulé

Nous avons simulé un signal, avec ou sans vibrato, formé de 30 partiels harmoniques dont les amplitudes décroissent en $\frac{1}{l^2}$, où l est le numéro d'ordre des harmoniques. Deux notes se succèdent. Lors de la transition, la fréquence fondamentale passe de 440 Hz à 480 Hz. Nous présentons les

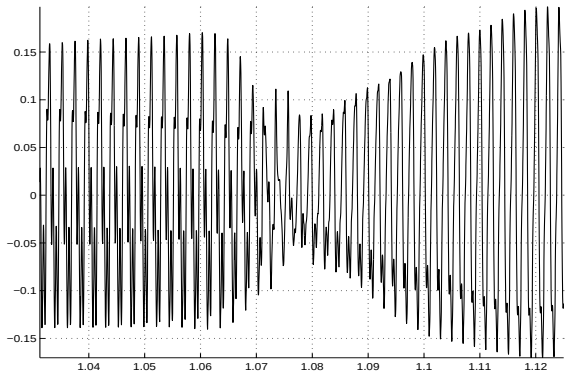


FIG. 11.4 – *Signal de flûte réel (***flute.sf***) lors du premier changement de note. En abscisse : le temps en seconde ; en ordonnée : l'amplitude des échantillons*

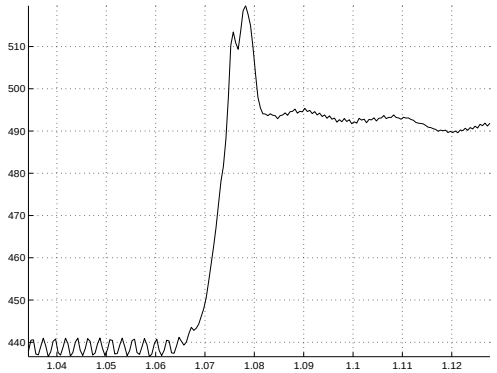


FIG. 11.5 – *Trajet de la fréquence de la fondamentale pour le signal de flûte lors du premier changement de note. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

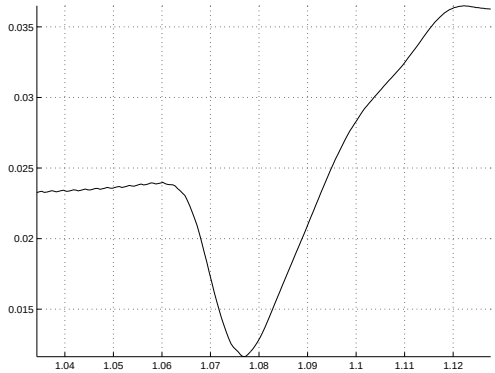


FIG. 11.6 – *Trajet de l'amplitude de la fondamentale pour le signal de flûte lors du premier changement de note. En abscisse : le temps en seconde ; en ordonnée : l'amplitude*

résultats pour la fonction d'observation « flux spectral calculé avec les spectres d'amplitude » (voir la partie II, sections 2.4.3.1 et 2.4.3.2, page 27) sur les figures 11.7 (trajet de f_0) et 11.8 (trajet du « flux spectral ») quand aucun vibrato n'est présent, et sur les figures 11.9 (trajet de f_0) et 11.10 (trajet du « flux spectral ») quand un vibrato est présent. Nous constatons que la présence d'un vibrato rend le « flux spectral » inutilisable pour la *segmentation en zones stables*.

Les paramètres libres pour le « flux spectral » ont été fixés à $t_{SIG} = 1764$ ($T = 0{,}04$ seconde) ; $Q = 220$ (0,005 seconde) ; $t_{FFT} = 4096$; et la fenêtre de pondération utilisée est celle de Blackman. La fréquence du vibrato est $f_{vib} = 5\ Hz$, son amplitude $A_{vib} = 30\ Hz$ et sa phase à l'origine $\varphi_{vib} = 2{,}4125\ rad$. Pour le modèle de transition en fréquence utilisé, voir la section 11.2.3.1. Nous avons choisi $a = 0{,}75$ et $b = 0{,}05$. Le modèle de transition en amplitude décrit dans la section 11.2.3.2 n'a pas été utilisé.

Dans le premier cas (pas de vibrato), la fonction d'observation réagit nettement lors de la transition ; dans le second cas (vibrato présent), la fonction d'observation ne réagit pas de façon significative lors de la transition.

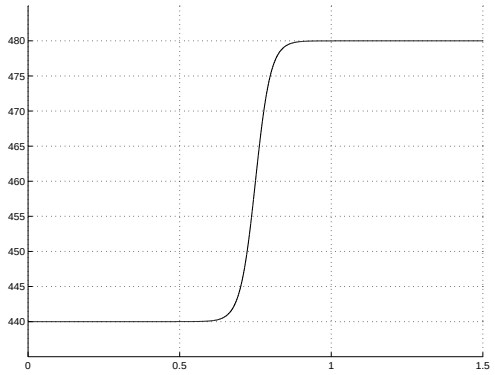


Fig. 11.7 – *Trajet de la fréquence fondamentale pour un son simulé. Présence d'une transition. Pas de vibrato. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

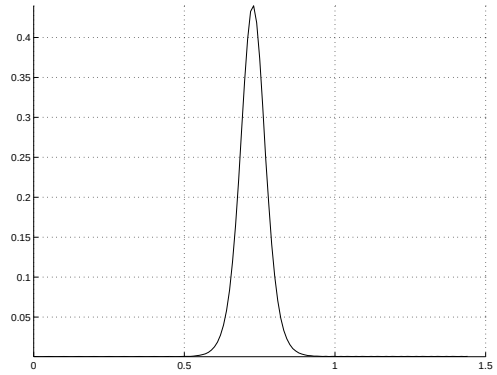


Fig. 11.8 – *Trajet du « flux spectral calculé avec les spectres d'amplitude » pour le son simulé (pas de vibrato). En abscisse : le temps en seconde*

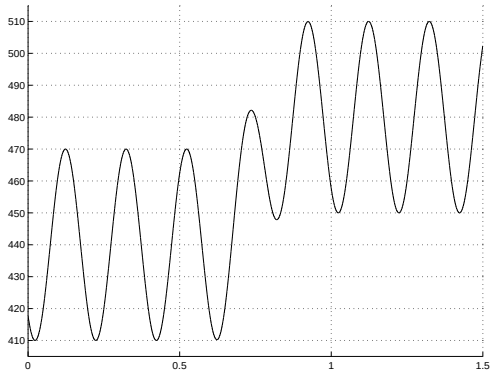


Fig. 11.9 – *Trajet de la fréquence fondamentale pour un son simulé. Présence d'une transition. Vibrato présent. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

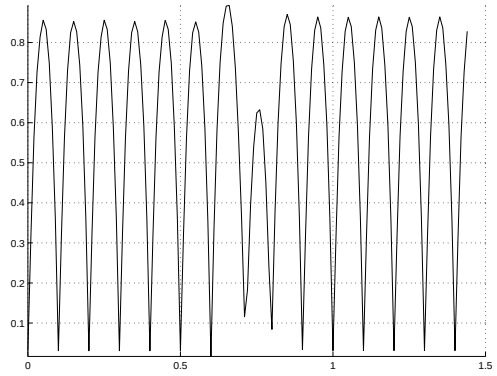


Fig. 11.10 – *Trajet du « flux spectral calculé avec les spectres d'amplitude » pour le son simulé (vibrato présent). En abscisse : le temps en seconde*

Chapitre 12

Méthodes de détection du vibrato à partir du son

12.1 Préambule

Il est intéressant de détecter le vibrato et d'estimer ses paramètres sans passer par le trajet de f_0. Ceci fait l'objet des méthodes proposées dans ce chapitre. La méthode décrite dans la **deuxième section** (section 12.2) de ce chapitre ne permet pas de détecter le vibrato sans extraire ses paramètres. Il est intéressant de détecter la présence de vibrato, sans passer par le trajet de f_0, mais aussi sans estimer les paramètres du vibrato. Ceci fait l'objet de la méthode présentée dans la **troisième section** (section 12.3) de ce chapitre. La méthode décrite dans la **quatrième section** (section 12.4) de ce chapitre nécessite d'être adaptée pour être utilisée sur des sons réels. La **cinquième section** (section 12.5) de ce chapitre constitue une conclusion à ce chapitre.

12.2 Méthode basée sur la modélisation du spectre complexe

12.2.1 Introduction

La méthode consiste à minimiser par les moindres carrés la distance – l'erreur – entre un spectre complexe calculé par transformée de Fourier pour une portion d'un signal sonore et un spectre complexe calculé en utilisant les estimations des paramètres inconnus de ce signal. Il s'agit de déterminer les valeurs des paramètres inconnus du signal telles que cette distance soit minimale. Ces paramètres, pour le modèle que nous utilisons, sont la partie réelle a_l de l'amplitude complexe de chaque harmonique (dont le numéro d'ordre est l), la partie imaginaire b_l de l'amplitude complexe de chaque harmonique, l'excursion A_{vib} en Hz du vibrato, la fréquence f_{vib} en Hz du vibrato, la phase à l'origine φ_{vib} du vibrato et la fréquence fondamentale f_0 en Hz.

12.2.2 Influence du vibrato sur le spectre d'amplitude

Nous avons simulé le signal suivant :

$$s(t) = \sum_{l=1}^{60} \cos\left(\varphi_l + 2\pi(l f_0^c)t + \frac{l A_{vib}}{f_{vib}} \sin\left(2\pi f_{vib} t\right)\right)$$

Il s'agit donc de la somme d'harmoniques (de numéro d'ordre l, l variant de 1 à 60 ; de fréquences $l f_0^c$, avec $f_0^c = 300\ Hz$; d'amplitudes 1 et de phases aléatoires uniformément distribuées entre 0 et 2π) pour lesquels un vibrato de fréquence $f_{vib} = 5\ Hz$, de phase à l'origine nulle et d'amplitude $l A_{vib}\ Hz$, avec $A_{vib} = 15$, est présent. Ce signal est échantillonné à $f_e = 44100\ Hz$. Il faut remarquer que l'amplitude du vibrato étant $l A_{vib}\ Hz$ pour l'harmonique de numéro d'ordre l, le signal est à tout moment harmonique.

La longueur du signal est de 0,75 seconde. Nous considérons des portions de ce signal larges de 13230 échantillons, soit larges de 0,3 seconde, ce qui est beaucoup eu égard à ce qui est communément utilisé comme taille de fenêtre d'analyse [1], mais ce qui correspond à l'ordre de grandeur des tailles des fenêtres d'analyse utilisées dans le chapitre 13 pour la détection du vibrato à partir des trajets des harmoniques du signal. Le pas d'avancement entre deux portions successives est de 100 échantillons. Chacune d'elles est multipliée par la fenêtre de pondération W. Nous calculons le spectre d'amplitude pour chaque portion pondérée. L'instant courant pour chaque fenêtre d'analyse est l'instant t_j de son centre. Le trajet théorique de la fréquence fondamentale instantanée, f_0^{th}, est obtenu en considérant que $f_0^{th}(t_j) = f_0^c + A_{vib}\cos(2\pi f_{vib} t_j)$.

La méthode sommaire utilisée pour estimer la fréquence et l'amplitude du $l^{\text{ème}}$ harmonique est la suivante. Le maximum du spectre d'amplitude entre $lf_0^{th} - \frac{f_0^{th}}{2}$ et $lf_0^{th} + \frac{f_0^{th}}{2}$ est détecté. L'amplitude de ce $l^{\text{ème}}$ harmonique est égale à l'amplitude de ce maximum. Nous estimons ainsi les amplitudes $a_l(t_j)$ des neuf premiers harmoniques. Le produit $p(t_j) = \prod_{l=1}^{9} a_l(t_j)$ (il s'agit d'une version de la moyenne géométrique : voir la section 2.6.2) est calculé. Nous présentons sur la figure 12.1 le trajet de f_0^{th} (courbe du haut) et le trajet du produit p (courbe du bas : l'échelle, pour ce produit, n'est pas respectée) obtenus.

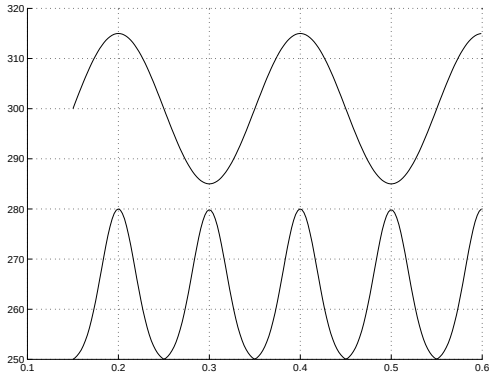


FIG. 12.1 – *Courbe du bas : trajet du produit de l'amplitude des neuf premiers harmoniques ; courbe du haut : trajet de la fréquence fondamentale. En abscisse : le temps en seconde ; en ordonnée : pour la courbe du haut, la fréquence en Hz, et pour la courbe du bas, échelle arbitraire*

Nous constatons que le produit p des amplitudes passe par des maximums quand la fréquence instantanée du premier harmonique (et donc celle de chaque harmonique) passe par des maximums ou des minimums ($f_0^{th} = f_0^c \pm A_{vib}$), et par des minimums quand sa valeur est voisine de la fréquence fondamentale quand il n'y a pas de vibrato ($f_0^{th} = f_0^c$).

En observant les spectres d'amplitude (voir les figures 12.2 et 12.3), nous constatons que ce sont les harmoniques de numéros d'ordre élevés qui sont les plus distordus : plus leur fréquence augmente, plus leur amplitude décroît et plus la largeur de leur lobe principal s'accroît.

Nous considérons deux cas.

Pour le premier cas, à t_j, la fréquence fondamentale passe par un maximum ($f_0^{th} = f_0^c + A_{vib}$: l'évolution de la fréquence au centre de la portion est faible). Nous donnons sur la figure 12.2 le spectre d'amplitude obtenu.

Pour le second cas, à t_j, la valeur absolue de la dérivée de la fréquence fondamentale passe par un maximum ($f_0^{th} = f_0^c$: l'évolution de la fréquence au centre de la portion est importante). Nous donnons sur la figure 12.3 le spectre d'amplitude obtenu.

1. Voir la note de la page 13.

Sont représentés dans les deux cas seulement les 9 premiers harmoniques. Le spectre d'amplitude du signal quand il n'y a pas de vibrato est superposé en trait interrompu dans le second cas (figure 12.3).

Dans le premier cas, le pic obtenu pour le premier harmonique est 3,5 dB plus petit que le pic obtenu quand il n'y a pas de vibrato. Pour le neuvième harmonique, la différence est de 9 dB.

Dans le second cas, pour le premier harmonique, la différence est déjà de 6 dB. Pour le neuvième harmonique, elle est de 13 dB si nous considérons le maximum du spectre d'amplitude dans la bande de fréquences $\left]9f_0^{th} - \frac{f_0^{th}}{2} \quad 9f_0^{th} + \frac{f_0^{th}}{2}\right[$ et de 16 dB si nous considérons la valeur du spectre d'amplitude au centre de cette bande (voir la figure 12.3 : les lobes sont symétriques autour de lf_0^{th} et creux au centre).

La différence entre ces deux cas est très nette. Cependant, dans les deux cas, plus nous considérons un harmonique de numéro d'ordre élevé, plus son amplitude est détériorée par le vibrato. De plus, la largeur du pic augmente avec le numéro d'ordre de l'harmonique.

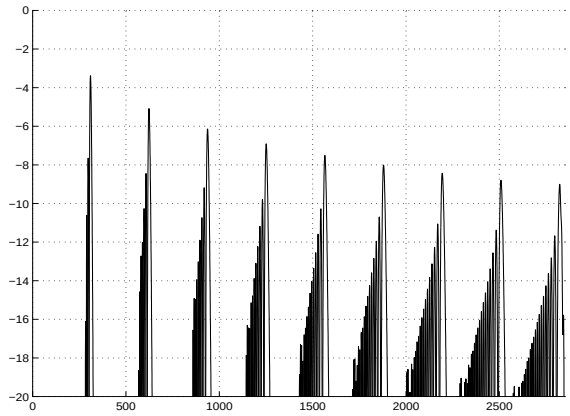


FIG. 12.2 – *Spectre d'amplitude du signal – Premier cas : l'évolution de la fréquence au centre de la fenêtre d'analyse est faible. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude en dB*

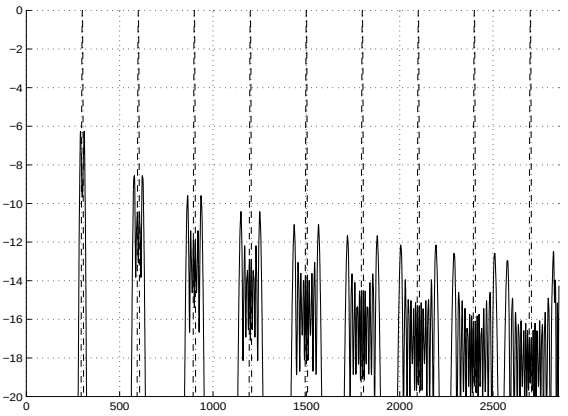


FIG. 12.3 – *Spectre d'amplitude du signal – Second cas : l'évolution de la fréquence au centre de la fenêtre d'analyse est importante. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude en dB*

Nous tenterons d'expliquer plus en détail ces phénomènes par la suite. Quand nous prenons des fenêtres d'analyse plus petites les résultats sont les mêmes, seulement le spectre d'amplitude du signal quand il y a du vibrato est moins détérioré globalement : ces phénomènes ne sont plus nettement visibles que sur les harmoniques de numéros d'ordre très élevés : voir la section 12.3.

12.2.3 Explication

12.2.3.1 Formalisation du problème

Nous avons modélisé le trajet de la fréquence fondamentale ainsi (voir la section 11.2.2) :

$$f_0(t) = f_0^c + A_{vib}\cos\left(2\pi f_{vib}t\right)$$

Ceci correspond à une sinusoïde pure modulée en fréquence. L'évolution de la phase s'écrit :

$$\phi(t) = 2\pi\int_0^t f_0(t)dt = 2\pi\left[f_0^c t + \frac{A_{vib}}{2\pi f_{vib}}\sin(2\pi f_{vib}t)\right]$$

Nous faisons les hypothèses que la phase du vibrato et que la phase du signal à l'origine ($t = 0$) sont nulles : ces hypothèse ne sont pas restrictives. Donc le signal s'écrit :

$$\begin{aligned} M(t) &= \cos(\phi(t)) \\ &= \cos\left(2\pi f_0^c t + \frac{A_{vib}}{f_{vib}}\sin(2\pi f_{vib} t)\right) \\ &= \cos\left(2\pi f_0^c t\right)\cos\left(\frac{A_{vib}}{f_{vib}}\sin(2\pi f_{vib} t)\right) - \sin\left(2\pi f_0^c t\right)\sin\left(\frac{A_{vib}}{f_{vib}}\sin(2\pi f_{vib} t)\right) \end{aligned}$$

Or, nous avons :

$$\cos(a\sin(kx)) = J_0(a) + 2\sum_{n=1}^{+\infty} J_{2n}(a)\cos(2nkx)$$
$$\sin(a\sin(kx)) = 2\sum_{n=1}^{+\infty} J_{2n-1}(a)\sin((2n-1)kx)$$

où $J_n(x)$ est la fonction de Bessel d'ordre n entier.

En identifiant : $a = \frac{A_{vib}}{f_{vib}}$ et $kx = 2\pi f_{vib} t$, nous pouvons écrire :

$$\begin{aligned} M(t) = & \\ & \cos\left(2\pi f_0^c t\right)\left[J_0\left(\frac{A_{vib}}{f_{vib}}\right) + 2\sum_{n=1}^{+\infty} J_{2n}\left(\frac{A_{vib}}{f_{vib}}\right)\cos(2n2\pi f_{vib} t)\right] - \\ & \sin\left(2\pi f_0^c t\right) 2\sum_{n=1}^{+\infty} J_{2n-1}\left(\frac{A_{vib}}{f_{vib}}\right)\sin((2n-1)2\pi f_{vib} t) \end{aligned}$$

Ce qui, en sachant que $J_{-n}(x) = (-1)^n J_n(x)$, s'écrit, après quelques calculs :

$$M(t) = \sum_{n=-\infty}^{+\infty} J_n\left(\frac{A_{vib}}{f_{vib}}\right)\cos(2\pi(f_0^c + n f_{vib})t)$$

La transformée de Fourier $\hat{M}(f)$ de cette expression se calcule aisément :

$$\hat{M}(f) = \frac{1}{2}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{A_{vib}}{f_{vib}}\right)\left[\delta\left(f - (f_0^c + n f_{vib})\right) + \delta\left(f + (f_0^c + n f_{vib})\right)\right]$$

Nous considérons une portion du signal, portion multipliée par la fenêtre de pondération $w(t)$. La transformée de Fourier de la fenêtre de pondération est $\hat{W}(f)$.

La transformée de Fourier de cette portion pondérée du signal est alors :

$$\hat{M}_{(W)}(f) = \frac{1}{2}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{A_{vib}}{f_{vib}}\right)\left[\hat{W}\left(f - (f_0^c + n f_{vib})\right) + \hat{W}\left(f + (f_0^c + n f_{vib})\right)\right]$$

Nous obtenons donc, au lieu d'un pic d'amplitude $\frac{1}{2}$ ayant la forme de la transformée de Fourier de la fenêtre de pondération, la somme d'une infinité de pics d'amplitude $\frac{1}{2}J_n\left(\frac{A_{vib}}{f_{vib}}\right)$ (il faut cependant noter que les coefficients de Bessel, à x constant, décroissent rapidement avec $|n|$ croissant) ayant chacun la forme de la transformée de Fourier de la fenêtre de pondération. Ces pics sont séparés de $f_{vib}\ Hz$, et ils sont constructifs ou destructifs les uns entre les autres suivant les phases (c'est-à-dire suivant le signe des coefficients de Bessel).

12.2.3.2 Conséquences

Mais cela n'explique pas pourquoi ce sont les harmoniques de fréquences élevées qui sont les plus distordus par le vibrato. Nous donnons ci-dessous un élément de réponse.

La fonction génératrice des fonctions de BESSEL d'ordre entier n s'écrit :

$$\exp\left(\frac{x}{2}\left(t-\frac{1}{t}\right)\right) = \sum_{n=-\infty}^{+\infty} t^n J_n(x)$$

Donc, si nous posons $t = 1$, nous avons :

$$1 = \sum_{n=-\infty}^{+\infty} J_n(x) \qquad \text{(formule 1)}$$

Or, nous avons, dans notre cas, $x = \dfrac{lA_{vib}}{f_{vib}}$, où l est le numéro d'ordre de l'harmonique. A_{vib} et f_{vib} sont des constantes. Ainsi x croît linéairement avec le numéro d'ordre de l'harmonique. Et, pour $n = 0$, la valeur absolue du coefficient de BESSEL décroît rapidement avec x.

Ainsi, plus le numéro d'ordre de l'harmonique est élevé, plus il faut additionner de coefficients de BESSEL (ordre 0, puis ordre -1 et 1, puis ordre -2 et 2, etc.) pour atteindre, avec la somme précédente (formule 1), une valeur proche de 1. Nous avons calculé cette somme (formule 1) pour différentes valeurs de x. Nous obtenons la courbe 12.4. En abscisse, nous avons le nombre de coefficients de BESSEL pris en compte. La courbe la plus à gauche correspond à $x = 2$ et la plus à droite à $x = 20$ (le pas d'avancement est 2). Donc plus le numéro d'ordre de l'harmonique est élevé, plus son énergie se disperse sur un nombre élevé de pics : ainsi, plus son lobe sera étalé autour de la fréquence centrale, et plus l'amplitude de ce lobe sera faible.

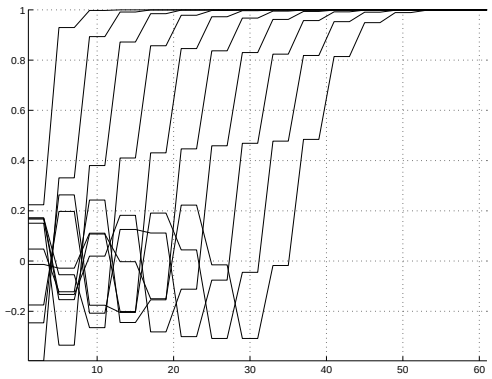


FIG. 12.4 – *Croissance de la fonction génératrice en fonction de x : $x \in [2 \ \ldots \ 20]$. En abscisse : le nombre de coefficients pris en compte $(2n + 1)$; en ordonnée : valeur de la fonction génératrice $(t = 1)$*

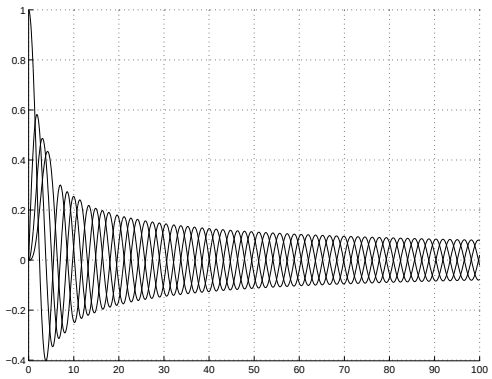


FIG. 12.5 – *Fonctions de* BESSEL *d'ordre n, pour n valant 0, 1, 2 et 3. En abscisse : x ; en ordonnée : valeur des fonctions de* BESSEL *$J_n(x)$*

Nous présentons les fonctions de BESSEL sur la figure 12.5.

Bien sûr, cette explication n'est pas suffisante puisque nous n'avons pas tenu compte des phases : ce sont elles qui expliquent les différences qui existent entre les spectres d'amplitude des figures 12.2 et 12.3. Des détails supplémentaires et importants sont donnés dans la section 12.3.

12.2.4 Modélisation plus complète du signal quand un vibrato est présent

Pour le moment dans ce chapitre, nous avons modélisé l'évolution de la fréquence d'une sinusoïde modulée en fréquence ainsi :

$$f_0(t) = f_0^c + A_{vib}\cos\left(2\pi f_{vib} t\right)$$

où f_0^c est la fréquence fondamentale « centrale », A_{vib} l'amplitude en Hz du vibrato, et f_{vib} la fréquence du vibrato.

Ajoutons à présent à ce modèle une phase à l'origine φ_1 pour la fondamentale[2], et une phase à l'origine φ_{vib} pour le vibrato. $f_0(t)$ s'écrit alors :

$$f_0(t) = f_0^c + A_{vib}\cos\left(2\pi f_{vib}t + \varphi_{vib}\right)$$

et l'évolution de la phase du signal s'écrit :

$$\begin{aligned}\phi(t) &= 2\pi\int_0^t f_0(t)dt + \varphi_1 \\ &= 2\pi f_0^c t + \frac{A_{vib}}{f_{vib}}\sin(2\pi f_{vib}t + \varphi_{vib}) + \varphi_1\end{aligned}$$

Nous obtenons :

$$\begin{aligned}M(t) &= \cos\left[2\pi f_0^c t + \frac{A_{vib}}{f_{vib}}\sin(2\pi f_{vib}t + \varphi_{vib}) + \varphi_1\right] \\ &= \sum_{n=-\infty}^{+\infty} J_n\left(\frac{A_{vib}}{f_{vib}}\right)\cos\left[(2\pi f_{vib}t + \varphi_{vib})n + 2\pi f_0^c t + \varphi_1\right]\end{aligned}$$

Si nous considérons un signal formé de L partiels harmoniques modulés par un vibrato, d'amplitudes respectives B_l et de phases à l'origine respectives φ_l, l variant de 1 à L, nous avons :

$$\begin{aligned}M(t) &= \sum_{l=1}^{L} B_l\cos\left[2\pi l f_0^c t + \frac{lA_{vib}}{f_{vib}}\sin(2\pi f_{vib}t + \varphi_{vib}) + \varphi_l\right] \\ &= \sum_{l=1}^{L}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right)B_l\cos\left[(2\pi f_{vib}t + \varphi_{vib})n + 2\pi l f_0^c t + \varphi_l\right]\end{aligned}$$

Nous multiplions le signal par une fenêtre de pondération $w(t)$. La transformée de Fourier de la fenêtre de pondération s'écrit $\hat{W}(f)$. La transformée de Fourier de :

$$\cos[(2\pi f_{vib}t + \varphi_{vib})n + 2\pi l f_0^c t + \varphi_l]w(t)$$

est :

$$\frac{1}{2}\hat{W}\left(f - lf_0^c - nf_{vib}\right)\exp[i(n\varphi_{vib} + \varphi_l)] + \frac{1}{2}\hat{W}\left(f + lf_0^c + nf_{vib}\right)\exp[-i(n\varphi_{vib} + \varphi_l)]$$

La transformée de Fourier du signal $M(t)$ multiplié par la fenêtre de pondération $w(t)$ est alors :

$$\begin{aligned}\hat{M}_{(W)}(f) &= \\ &\sum_{l=1}^{L}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right) \\ &\frac{B_l}{2}\left\{\hat{W}\left(f - lf_0^c - nf_{vib}\right)\exp\left[i(n\varphi_{vib} + \varphi_l)\right] + \hat{W}\left(f + lf_0^c + nf_{vib}\right)\exp\left[-i(n\varphi_{vib} + \varphi_l)\right]\right\}\end{aligned}$$

2. Dans cet exposé, la fréquence fondamentale est appelée indifféremment f_0 ou f_1. La phase de la fondamentale est appelée φ_0 ou φ_1. Voir la note de la page 14.

En fait, dans la suite nous allons considérer :

$$\begin{cases} a_l &= \dfrac{B_l}{2}\cos(\varphi_l) \\ b_l &= \dfrac{B_l}{2}\sin(\varphi_l) \end{cases}$$

Donc :

$$\hat{M}_{(W)}(f) = \sum_{l=1}^{L}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right) \left\{\hat{W}(f - lf_0^c - nf_{vib})(a_l + ib_l)\exp(in\varphi_{vib}) + \hat{W}(f + lf_0^c + nf_{vib})(a_l - ib_l)\exp(-in\varphi_{vib})\right\}$$

12.2.5 Les moindres carrés pour le vibrato

Nous allons ici essayer de retrouver les paramètres inconnus du signal en minimisant par les moindres carrés l'erreur quadratique entre un spectre complexe connu et un spectre complexe estimé.

12.2.5.1 Les moindres carrés

Nous voulons minimiser l'erreur quadratique ϵ entre un vecteur colonne connu $\underline{S}$, de taille N, et un vecteur colonne estimé $\underline{M}$, contrôlé par P paramètres x_p. Si les dérivées d'ordre supérieur à 1 de $\underline{M}$ par rapport aux x_p ne sont pas nulles, le problème n'est pas linéaire et ϵ est itérativement minimisée. Nous allons considéré ici ce cas général. Soit (k) représentant le numéro de l'itération courante. Nous avons :

$$\epsilon_{(k)} = \left(\underline{M}_{(k)}^H - \underline{S}^H\right)\left(\underline{M}_{(k)} - \underline{S}\right)$$

où l'exposant H représente la transposition-conjugaison. Les autres notations utilisées dans ce chapitre sont : l'exposant t, qui représente la transposition ; l'exposant -1, qui représente l'inversion ; l'exposant $*$, qui représente la conjugaison ; et le soulignement, qui indique que nous sommes en présence d'un vecteur ou d'une matrice.

Nous avons :

$$\underline{M}_{(k)} = \underline{M}_{(k-1)} + \alpha_{1(k-1)}\frac{\partial \underline{M}_{(k-1)}}{\partial x_1} + \cdots + \alpha_{P(k-1)}\frac{\partial \underline{M}_{(k-1)}}{\partial x_P}$$

Nous posons :

$$\underline{D}_{(k-1)} = \underbrace{\left[\frac{\partial \underline{M}_{(k-1)}}{\partial x_1} \;\cdots\; \frac{\partial \underline{M}_{(k-1)}}{\partial x_P}\right]}_{P \text{ colonnes}} \Bigg\} N \text{ lignes}$$

Et :

$$\underline{\alpha}_{(k-1)} = \left[\begin{array}{c} \alpha_{1(k-1)} \\ \vdots \\ \alpha_{P(k-1)} \end{array}\right] \Bigg\} P \text{ lignes}$$

Alors $\underline{\alpha}$, qui est l'inconnue, correspond au pas de descente optimal. Nous le montrons sur la figure 12.6 dans le cas où M et S sont des scalaires et $P = 1$.

Nous avons donc :

$$\underline{M}_{(k)} = \underline{M}_{(k-1)} + \underline{D}_{(k-1)}\underline{\alpha}_{(k-1)}$$

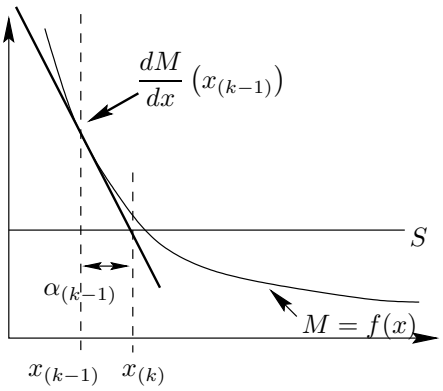


FIG. 12.6 – *Pas de descente optimal pour les moindres carrés*

Et :

$$\epsilon_{(k)} = \left(\underline{M}^H_{(k-1)} + \underline{\alpha}^H_{(k-1)}\underline{D}^H_{(k-1)} - \underline{S}^H\right)\left(\underline{M}_{(k-1)} + \underline{D}_{(k-1)}\underline{\alpha}_{(k-1)} - \underline{S}\right)$$

Nous voulons minimiser ϵ, c'est-à-dire annuler sa dérivée : $\dfrac{\partial \epsilon}{\partial \underline{\alpha}_{(k-1)}} = 0$. Or :

$$\frac{\partial \epsilon}{\partial \underline{\alpha}_{(k-1)}} = 2\underline{D}^H_{(k-1)}\left(\underline{M}_{(k-1)} + \underline{D}_{(k-1)}\underline{\alpha}_{(k-1)} - \underline{S}\right)$$

Donc, nous avons :

$$\underline{\alpha}_{(k-1)} = \left(\underline{D}^H_{(k-1)}\underline{D}_{(k-1)}\right)^{-1}\underline{D}^H_{(k-1)}\left(\underline{S} - \underline{M}_{(k-1)}\right)$$

Nous avons finalement :

$$\underline{x}_{(k)} = \underline{x}_{(k-1)} + \underline{\alpha}_{(k-1)}$$

Ce processus continue jusqu'à sa convergence (il faut définir un critère de convergence), s'il converge ; ou jusqu'à ce que $k = K_{max}$, où K_{max} est un paramètre libre fixé. Dans cet exposé, nous avons utilisé la seconde condition d'arrêt, avec K_{max} de l'ordre de 20.

12.2.5.2 Notre cas

Ici, $\underline{\hat{S}}(f)$ est le spectre complexe obtenu avec la FFT, rangé dans un vecteur colonne ; $\underline{\hat{M}}(f)$ est le spectre complexe estimé ; et les paramètres x_p sont :

$$\begin{cases} a_l & \text{la partie réelle de l'amplitude complexe de l'harmonique } l \\ b_l & \text{la partie imaginaire de l'amplitude complexe de l'harmonique } l \\ A_{vib} & \text{l'excursion en } Hz \text{ du vibrato} \\ f_{vib} & \text{la fréquence du vibrato} \\ f_0^c & \text{la fréquence fondamentale « centrale »} \\ \varphi_{vib} & \text{la phase du vibrato} \end{cases}$$

Nous avons donc $2L + 4$ paramètres à déterminer (il est fait l'hypothèse que L est connu).

Il ne nous reste plus qu'à dériver $\underline{\hat{M}}(f)$ par rapport à ces paramètres, à donner des formules analytiques pour ces dérivées. Il n'existe pas de formule analytique, ni pour les fonctions de BESSEL ni pour leurs dérivées. Cependant, elles sont aisément estimables, avec une grande précision, numériquement.

12.2.5.3 Dérivation par rapport à a_l

$$\frac{\partial \underline{\hat{M}}(f)}{\partial a_l} = \sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right)\left\{\underline{\hat{W}}\left(f - lf_0^c - nf_{vib}\right)\exp\left(in\varphi_{vib}\right) + \underline{\hat{W}}\left(f + lf_0^c + nf_{vib}\right)\exp\left(-in\varphi_{vib}\right)\right\}$$

12.2.5.4 Dérivation par rapport à b_l

$$\frac{\partial \underline{\hat{M}}(f)}{\partial b_l} = \sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right)\left\{i\underline{\hat{W}}\left(f - lf_0^c - nf_{vib}\right)\exp\left(in\varphi_{vib}\right) - i\underline{\hat{W}}\left(f + lf_0^c + nf_{vib}\right)\exp\left(-in\varphi_{vib}\right)\right\}$$

12.2.5.5 Dérivation par rapport à A_{vib}

$$\frac{\partial \underline{\hat{M}}(f)}{\partial A_{vib}} = \sum_{l=1}^{L}\sum_{n=-\infty}^{+\infty} \frac{\partial J_n\left(\frac{lA_{vib}}{f_{vib}}\right)}{\partial A_{vib}}$$

$$\left\{\underline{\hat{W}}\left(f - lf_0^c - nf_{vib}\right)\left(a_l + ib_l\right)\exp\left(in\varphi_{vib}\right) + \underline{\hat{W}}\left(f + lf_0^c + nf_{vib}\right)\left(a_l - ib_l\right)\exp\left(-in\varphi_{vib}\right)\right\}$$

avec :

$$\frac{\partial J_n(x)}{\partial x} = \frac{n}{x}J_n(x) - J_{n+1}(x)$$

et :

$$\frac{\partial f(g(x))}{\partial x}(x_0) = \frac{\partial f}{\partial g(x)}(g(x_0))\frac{\partial g}{\partial x}(x_0)$$

Donc :

$$\frac{\partial J_n\left(\frac{lA_{vib}}{f_{vib}}\right)}{\partial A_{vib}} = \frac{l}{f_{vib}}\left[\frac{n}{\left(\frac{lA_{vib}}{f_{vib}}\right)}J_n\left(\frac{lA_{vib}}{f_{vib}}\right) - J_{n+1}\left(\frac{lA_{vib}}{f_{vib}}\right)\right]$$

12.2.5.6 Dérivation par rapport à f_{vib}

$$\frac{\partial \underline{\hat{M}}(f)}{\partial f_{vib}} = \sum_{l=1}^{L}\left[\sum_{n=-\infty}^{+\infty} \frac{\partial J_n\left(\frac{lA_{vib}}{f_{vib}}\right)}{\partial f_{vib}}\left\{\underline{\hat{W}}(f - lf_0^c - nf_{vib})(a_l + ib_l)\exp(in\varphi_{vib}) + \right.\right.$$

$$\left.\underline{\hat{W}}(f + lf_0^c + nf_{vib})(a_l - ib_l)\exp(-in\varphi_{vib})\right\} +$$

$$\sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right)\left\{-n\frac{\partial \underline{\hat{W}}}{\partial f}(f - lf_0^c - nf_{vib})(a_l + ib_l)\exp(in\varphi_{vib}) + \right.$$

$$\left.\left. n\frac{\partial \underline{\hat{W}}}{\partial f}(f + lf_0^c + nf_{vib})(a_l - ib_l)\exp(-in\varphi_{vib})\right\}\right]$$

avec :

$$\frac{\partial J_n\left(\frac{lA_{vib}}{f_{vib}}\right)}{\partial f_{vib}} = \left(\frac{-lA_{vib}}{f_{vib}^2}\right)\left[\frac{n}{\left(\frac{lA_{vib}}{f_{vib}}\right)}J_n\left(\frac{lA_{vib}}{f_{vib}}\right) - J_{n+1}\left(\frac{lA_{vib}}{f_{vib}}\right)\right]$$

Nous donnons dans l'annexe C les dérivées de certaines fenêtres de pondération dans le domaine fréquentiel.

12.2.5.7 Dérivation par rapport à f_0^c

$$\frac{\partial \underline{\hat{M}}(f)}{\partial f_0^c} = \sum_{l=1}^{L} \sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right) \left\{ -l\frac{\partial \underline{\hat{W}}}{\partial f}(f - lf_0^c - nf_{vib})\,(a_l + ib_l)\exp(in\varphi_{vib}) + l\frac{\partial \underline{\hat{W}}}{\partial f}(f + lf_0^c + nf_{vib})\,(a_l - ib_l)\exp(-in\varphi_{vib}) \right\}$$

12.2.5.8 Dérivation par rapport à φ_{vib}

$$\frac{\partial \underline{\hat{M}}(f)}{\partial \varphi_{vib}} = \sum_{l=1}^{L} \sum_{n=-\infty}^{+\infty} J_n\left(\frac{lA_{vib}}{f_{vib}}\right) \left\{ in\underline{\hat{W}}\,(f - lf_0^c - nf_{vib})\,(a_l + ib_l)\exp(in\varphi_{vib}) - in\underline{\hat{W}}\,(f + lf_0^c + nf_{vib})\,(a_l - ib_l)\exp(-in\varphi_{vib}) \right\}$$

12.2.6 Simulations

Les simulations ont été faites sous MATLAB.

Tout d'abord, dans l'exemple qui suit, nous supposons que la fréquence fondamentale f_0^c est connue. Nous avons constaté que ce paramètre ralentissait la convergence ou, au pire, la rendait problématique. Dans l'exemple, nous avons pris $L = 3$. Les valeurs des paramètres à trouver sont rangées dans la première colonne du tableau 12.1 et les valeurs initiales (itération $k = 1$) de ces paramètres sont rangées dans la deuxième colonne de ce même tableau. f_0^c est égale à 400 Hz.

$A_{vib} = 20$	$A_{vib(1)} = 19{,}0$
$f_{vib} = 5$	$f_{vib(1)} = 5{,}25$
$\varphi_{vib} = 1$	$\varphi_{vib(1)} = 0{,}95$
$B_1 = 1$	$B_{1(1)} = 1{,}05$
$B_2 = 0{,}5$	$B_{2(1)} = 0{,}475$
$B_3 = 1/3$	$B_{3(1)} = 0{,}2375$
$\varphi_1 = 1{,}0$	$\varphi_{1(1)} = 1{,}05$
$\varphi_2 = 0{,}7$	$\varphi_{2(1)} = 0{,}665$
$\varphi_3 = 1{,}6$	$\varphi_{3(1)} = 1{,}68$

TAB. 12.1 – *Paramètres à trouver et conditions initiales*

Nous faisons initialement 5 % d'erreur sur chaque paramètre. La fenêtre de pondération utilisée est celle de BLACKMAN. Nous prenons en compte 41 coefficients de BESSEL pour chaque sinusoïde ($-20 \leq n \leq 20$). Nous avons $T = 0{,}04$ seconde ($t_{SIG} = 1765$), $f_e = 44100\ Hz$, le nombre de points pour la FFT égal à $t_{FFT} = 4096$ et nous travaillons avec les $N = 201$ premiers points du spectre complexe des fréquences positives.

Nous calculons l'erreur ε sur le paramètre $param$ à l'itération k ainsi :

$$\varepsilon(k) = 10\log_{10}\left\{\left|\frac{param(k) - param}{param}\right|\right\}$$

Donc 5 % d'erreur correspond à $-13\ dB$, 1 % à $-20\ dB$, puis 0,1 % à $-30\ dB$, etc.

Nous donnons sur la figure 12.7 le spectre d'amplitude réel en dB sur lequel nous travaillons (l'énergie de la fenêtre de pondération n'étant pas normalisée, nous n'avons pas 0 dB pour la première des sinusoïdes, d'amplitude 1), et sur les figures 12.8 et 12.9 l'erreur ε en fonction de k respectivement pour $A_{vib(k)}$, $f_{vib(k)}$ et $\varphi_{vib(k)}$, puis pour $B_{1(k)}$, $B_{2(k)}$, $B_{3(k)}$, $\varphi_{1(k)}$, $\varphi_{2(k)}$ et $\varphi_{3(k)}$ ($k = 1$ correspond aux conditions initiales). Nous constatons que l'algorithme converge vers les bonnes valeurs des paramètres.

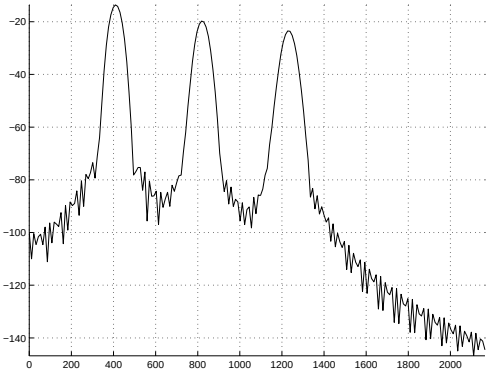


FIG. 12.7 – *Spectre d'amplitude obtenu avec le signal dont il faut retrouver les paramètres. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude en dB*

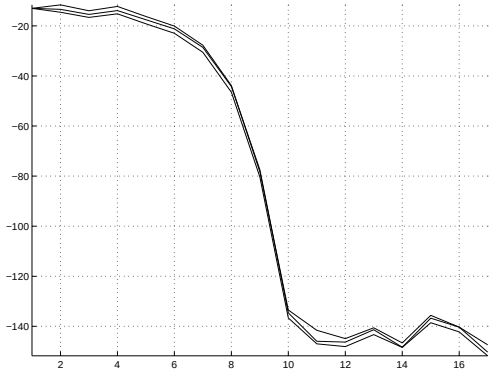


FIG. 12.8 – *Erreur sur $A_{vib(k)}$, $f_{vib(k)}$ et $\varphi_{vib(k)}$. En abscisse : le numéro de l'itération ; en ordonnée : les erreurs en dB*

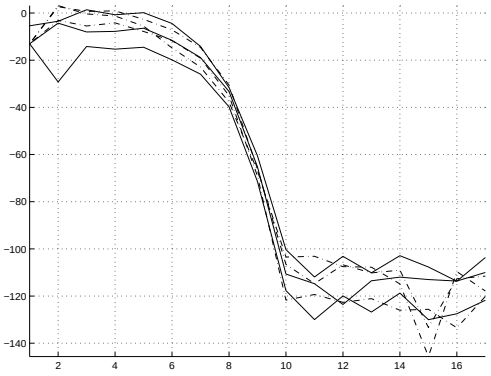


FIG. 12.9 – *Erreur sur les amplitudes $B_{1(k)}$, $B_{2(k)}$, $B_{3(k)}$ (traits pleins) des harmoniques ; et sur les phases $\varphi_{1(k)}$, $\varphi_{2(k)}$ et $\varphi_{3(k)}$ (traits interrompus) des harmoniques. En abscisse : le numéro de l'itération ; en ordonnée : les erreurs en dB*

12.2.7 Conclusion

En ce qui concerne la présence ou l'absence de vibrato, la **décision** est prise ainsi :

- Si f_{vib} est comprise entre 3 Hz and 11 Hz et A_{vib} est grand, il y a du vibrato.

Les **estimations** de f_{vib}, A_{vib} et φ_{vib} sont directement obtenues.

12.3 Méthode basée sur la distorsion des enveloppes spectrales

12.3.1 Introduction

La méthode est basée sur l'extraction de caractéristiques (fonctions d'observation) pour des fenêtres d'analyse larges de quelques dizaines de millisecondes, et leur traitement. Cela peut paraître étrange d'utiliser des portions du signal aussi petites, qui sont loin de couvrir une période du vibrato (classiquement, la fréquence du vibrato est d'environ 5 Hz, donc sa période est de quelques centaines de millisecondes). Cependant, il ne s'agit pas ici de déterminer l'amplitude, la fréquence et la phase du vibrato : nous voulons seulement le détecter, et non déterminer ces paramètres. Notamment, le manque d'information dû au fait que nous utilisons des petites fenêtres d'analyse va se traduire par l'obtention d'une équation à deux variables inconnues de la forme : $A = \frac{A_{vib}}{f_{vib}}$, où A_{vib} est l'amplitude du vibrato (modélisé par une sinusoïde : voir la section 11.2.2) et f_{vib} sa fréquence (voir la section 12.2.3 pour les notations).

En ce qui concerne la modulation de fréquence, une quantité, appelée « bande de Carson », est, pratiquement, définie. Lors de la modulation en fréquence d'une porteuse, l'énergie de cette porteuse s'éparpille sur toutes les fréquences du spectre : la transformée de Fourier du signal est la somme d'une infinité de pics décalés de f_{vib} Hz, correspondant chacun à la transformée de Fourier de la fenêtre de pondération utilisée, et d'amplitudes les coefficients de Bessel (voir la section 12.2.3.1). Cependant, la plus grande partie de l'énergie du spectre (98 %) est comprise dans la bande de Carson, qui est centrée sur la fréquence de la porteuse et est égale à : $B = 2(\Delta f + F_{max})$, où F_{max} est l'excursion maximale en fréquence de la modulation et Δf la fréquence maximale de la modulation. Ainsi, avec nos notations, nous avons, pour le $l^{ème}$ harmonique : $B_l = 2(f_{vib} + lA_{vib})$, qui augmente linéairement avec le numéro d'ordre l de l'harmonique. Avec $f_0^c = 220$ Hz, $f_{vib} = 5$ Hz et $A_{vib} = 11$ Hz (ce qui correspond à 5 % de f_0^c), nous avons $B \geq 2f_0^c$ à partir de l'harmonique de numéro d'ordre $l = 20$, donc de fréquence $f_{20} = 20f_0^c = 4400$ Hz. Cela veut dire qu'à partir du vingtième harmonique, les lobes dus à deux harmoniques de numéros d'ordre successifs se recouvrent. Notons de plus que dans ce cas $f_{20} = 20f_0^c = 4400$ Hz tombe juste dans la bande de fréquence où l'oreille humaine entend le mieux (voir l'annexe E). Nous allons tenter de mettre en évidence cette croissance linéaire de la bande de Carson : la méthode décrite dans cette section 12.3 est basée sur ce principe.

12.3.2 La méthode – Présentation

Nous avons vu dans la section 12.2 que la présence d'un vibrato détériore la forme de chaque lobe principal (notamment son amplitude) correspondant à un partiel harmonique du signal. Nous avons vu que cette détérioration est d'autant plus importante que le numéro d'ordre de cet harmonique est élevé. Nous allons de même montrer que, pour un harmonique de numéro d'ordre donné, plus la taille de la fenêtre d'analyse est grande plus cette détérioration est importante. Nous allons essayer de mettre cette détérioration en évidence et de déterminer certaines de ses caractéristiques.

Nous considérons deux fenêtres d'analyse, toutes les deux centrées sur le même instant d'analyse i. La taille de l'une (portion F_1) a été fixée à 60 millisecondes et la taille de l'autre (portion F_2) à 20 millisecondes : voir la figure 12.10. Le spectre d'amplitude est calculé pour chacune de ces portions (ces spectres d'amplitudes sont appelés respectivement s_1 et s_2). Notons que la fenêtre de pondération que nous utilisons est la fenêtre de Blackman.

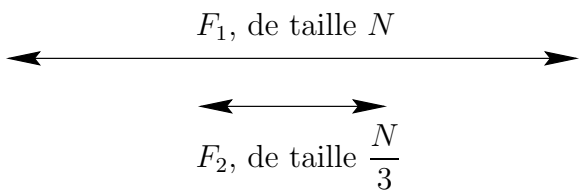


FIG. 12.10 – *Disposition des deux fenêtres d'analyse pour la détection du vibrato à partir de l'étude des enveloppes spectrales*

D'après ce qui a été mentionné ci-dessus, nous pouvons dire que les lobes principaux du spectre d'amplitude s_1 (grande portion) sont plus petits, à cause du vibrato, que les lobes principaux correspondant du spectre d'amplitude s_2 (petite portion), et ce d'autant plus qu'ils correspondront à des harmoniques de numéros d'ordre élevés. Nous avons indiqué qu'il ne s'agit ici ni de déterminer la fréquence fondamentale du signal ni de déterminer les paramètres du vibrato. Aussi, nous allons plutôt raisonner en terme d'enveloppe spectrale. Ainsi : plus la fréquence considérée sera grande, plus la différence entre les deux enveloppes spectrales sera grande.

Nous montrons dans la section suivante (section 12.3.3) que, si les harmoniques ont tous la même amplitude (cela veut dire que l'enveloppe spectrale « vraie », c'est-à-dire en l'absence de vibrato, est horizontale), cette différence croît linéairement avec la fréquence : cette *croissance linéaire* (ou du moins *quasi linéaire*) est ce que nous voulons détecter. Elle nous indiquera sûrement si un vibrato est présent ou non.

Nous pouvons remarquer aussi que les pics correspondant aux lobes dus aux harmoniques non seulement sont plus petits pour le spectre d'amplitude s_1 que pour le spectre d'amplitude s_2, mais sont en plus décalés en fréquence, et ce d'autant plus que le numéro d'ordre de l'harmonique considéré est grand.

12.3.3 Forme de la croissance de la différence entre les deux enveloppes spectrales

Nous avons tenté dans la section 12.2.3 d'expliquer pourquoi ce sont les harmoniques de numéros d'ordre supérieurs qui sont les plus détériorés par le vibrato. Nous dîmes que ceci est dû à ce que plus $x = \frac{lA_{vib}}{f_{vib}}$ (l est le numéro d'ordre de l'harmonique) est grand moins les coefficients $J_n(x)$ d'ordre $|n|$ grand sont négligeables. Ceci est illustré sur la figure 12.11 (avec $A_{vib} = 15$ et $f_{vib} = 5$).

Les amplitudes des harmoniques ont été fixées à la même valeur, soit ici 1. Nous avons pris : $f_0^c = 440\ Hz$, $f_{vib} = 5\ Hz$, $A_{vib} = 15\ Hz$, $\varphi_{vib} = 0$ (nous nous plaçons dans le cas le plus défavorable : c'est-à-dire le cas où les lobes sont le plus détériorés) et des phases aléatoires, uniformément réparties entre 0 et 2π, pour chacun des harmoniques. La fenêtre de pondération utilisée a été la fenêtre de BLACKMAN (des résultats similaires sont obtenus avec les fenêtres de HANNING et de HAMMING). La taille de la FFT est $t_{FFT} = 16384$.

Nous vérifions que l'amplitude du lobe principal de chaque harmonique (qui est la somme d'une infinité de fenêtres de pondération dans le domaine fréquentiel d'amplitudes $J_n(x)$ décalées de $f_{vib}\ Hz$) décroît avec le numéro d'ordre de l'harmonique. Il s'agit d'essayer de déterminer une loi pour cette décroissance. Sur la figure 12.12 nous donnons l'amplitude du maximum du lobe principal des quinze premiers harmoniques en fonction de leur position, et ceci pour plusieurs tailles de fenêtres d'analyse. À cause du vibrato la position du maximum du lobe principal n'est pas tout à fait lf_0. À la limite, si nous prenons une très grande fenêtre d'analyse, le maximum pour le premier harmonique se situe à $n = 2$ ou $n = -2$ (voir la figure 12.11), c'est-à-dire à $f_0^c \pm 2f_{vib}$; pour le deuxième harmonique autour de $2f_0^c \pm 5f_{vib}$; pour le troisième harmonique autour de $3f_0^c \pm 7f_{vib}$; etc. Et le lobe principal pour chaque harmonique se décompose en plusieurs petits lobes symétriquement disposés (voir la figure 12.3). Ici, nous utilisons des fenêtres d'analyse de tailles « raisonnables », c'est-à-dire de quelques dizaines de millisecondes, de telle façon que la somme des « petits lobes » nous donne un grand et large lobe, unique, tout du moins pour les premiers harmoniques : dans l'exemple utilisé ici, ceci est vérifié au moins pour les quinze premiers

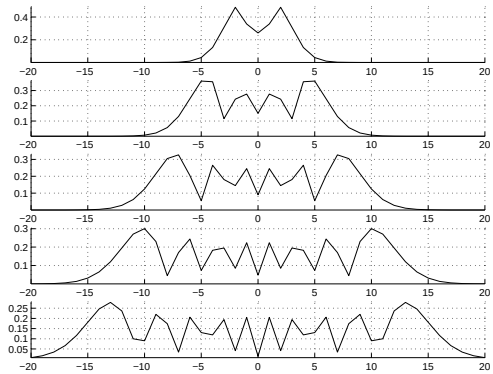


FIG. 12.11 – *Coefficients de* BESSEL *suivant x et n. Pour chacune des cinq courbes nous avons en abscisse : n ; en ordonnée : $|J_n(x)|$. Le premier harmonique ($x = A_{vib}/f_{vib}$) en représenté en haut ; le cinquième ($x = 5A_{vib}/f_{vib}$) en bas. Notons que le plus grand coefficient de* BESSEL *pour le premier harmonique vaut 0,486 ; pour le deuxième 0,362... ; et pour le cinquième 0,279*

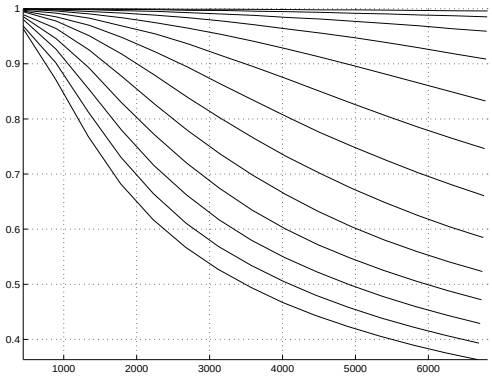


FIG. 12.12 – *Décroissance de l'amplitude des lobes en fonction de la taille de la fenêtre d'analyse. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude linéaire. La taille de la plus petite fenêtre est* 0,02 *seconde (courbe du haut) ; la taille de la plus grande* 0,08 *seconde (courbe du bas). Le pas entre deux tailles de fenêtre d'analyse successives est de* 0,005 *seconde*

harmoniques.

Nous constatons que, pour une petite fenêtre d'analyse, de taille 20 millisecondes, les amplitudes de tous les harmoniques sont proches de 1 : ainsi, pour la petite fenêtre d'analyse, le vibrato ne détériore pas le spectre d'amplitude. Nous constatons aussi que, pour la grande fenêtre d'analyse, nous obtenons une détérioration à la fois importante et presque linéaire, et les lobes des harmoniques de numéros d'ordre supérieurs ne se décomposent pas encore en plusieurs lobes. D'où le choix des tailles des deux fenêtres d'analyse qui a été fait (voir la section 12.3.2)

12.3.4 Influence des amplitudes respectives des harmoniques

Le fait que, dans le cas général, l'enveloppe vraie ne soit pas plate et horizontale (ce qui a lieu seulement quand les partiels ont la même amplitude) n'est pas un problème : nous pouvons en effet, dans le cas général, retrouver une *croissance quasi linéaire* de la différence entre les enveloppes spectrales es_1 et es_2. Considérons tout d'abord a_l^1, l'amplitude de l'harmonique de numéro d'ordre l du spectre d'amplitude s_1 ; a_l^2, celle du spectre d'amplitude s_2 ; et a_l, l'amplitude vraie de l'harmonique. Alors $c_l = \dfrac{a_l - a_l^1}{a_l}$ croît quasi linéairement avec l . Donc, si nous faisons l'hypothèse que a_l^2 est très proche de a_l, ce qui est presque vérifié puisque plus la fenêtre d'analyse est petite moins l'influence du vibrato se fait sentir (voir la figure 12.12), le problème est résolu. Ceci est montré sur la figure 12.14, où nous donnons, sur les courbes en traits pleins, $\dfrac{a_l^1}{a_l^{est}}$ [3]. Ici, nous avons un « formant » situé autour du cinquième harmonique, formant de forme gaussienne. L'amplitude de l'harmonique de numéro d'ordre l est : $\exp\left(-\dfrac{(l-5)^2}{50}\right)$, soit : 0,7 0,8 0,9 0,98 1,0 0,98 ... 0,1. Les amplitudes a_l sont estimées en utilisant une fenêtre d'analyse de 20 millisecondes prise au centre de la fenêtre d'analyse de 60 millisecondes qui nous donne les a_l^1 (voir la figure 12.10). En trait interrompu, nous avons les amplitudes vraies des harmoniques, et les amplitudes trouvées avec la petite fenêtre d'analyse : nous constatons que les deux courbes sont quasi superposées. Pour obtenir les courbes en traits pleins de cette figure 12.14, nous avons directement divisé la valeur du maximum de chacun des lobes obtenus pour la grande fenêtre par la valeur du maximum du lobe correspondant obtenu pour la petite fenêtre d'analyse, sans tenir compte des décalages de

3. *est* pour « estimée » : $a_l^{est} = a_l^2$

positions en fréquence de ces maximums selon la fenêtre d'analyse. Dans la suite de l'exposé, nous interpolons linéairement entre ces maximums, pour obtenir les enveloppes spectrales, et ce sont ces enveloppes spectrales que nous comparons, échantillon fréquentiel par échantillon fréquentiel, pour obtenir la courbe c. Ces enveloppes spectrales sont données sur la figure 12.13 pour les différentes fenêtres d'analyse. La fenêtre de pondération utilisée est celle de BLACKMAN. La phase (qui est liée à la position des fenêtres d'analyse) du vibrato ici est 0, mais nous obtenons des résultats similaires quelle que soit cette phase : seulement, dans certains cas, les amplitudes a_l sont un peu moins bien estimées (en fait, au pire, les points de la vraie enveloppe spectrale sont décalés en fréquence : le point en f se retrouve en $f \pm \frac{f}{f_0} f_{vib}$) : mais nous retrouvons tout de même une croissance de la différence entre les deux enveloppes spectrales d'allure quasi linéaire. La pente de cette croissance dépend de A_{vib} et de f_{vib}.

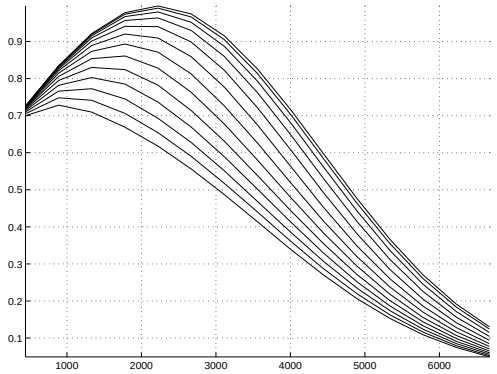


FIG. 12.13 – *Enveloppes spectrales. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude linéaire. La taille de la plus petite fenêtre est* 0,02 *seconde (courbe du haut) ; la taille de la plus grande* 0,08 *seconde (courbe du bas). Le pas entre deux tailles de fenêtre d'analyse successives est de* 0,005 *seconde*

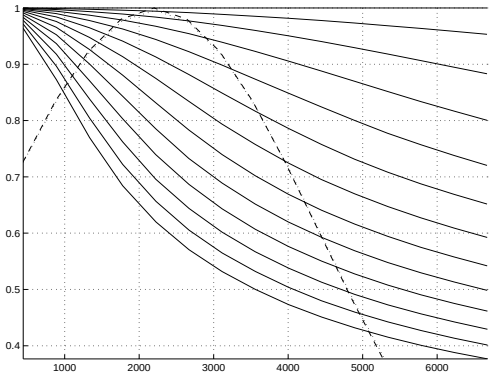


FIG. 12.14 – *Traits pleins : décroissance de l'amplitude de chaque pic divisée par l'amplitude du pic correspondant pour la plus petite fenêtre (*0,02 *seconde). En trait interrompu : les amplitudes vraies des harmoniques et les amplitudes trouvées avec la petite fenêtre d'analyse. En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude linéaire*

12.3.5 Calcul de l'enveloppe spectrale – signal « réel »

La technique présentée dans cette section vient de la thèse de Xavier SERRA : [Ser89], pages 42 – 47.

Le problème de l'obtention de l'enveloppe spectrale est un problème compliqué, et ce d'autant plus qu'ici nous nous basons sur les spectres d'amplitude. En effet, puisque nous voulons obtenir la vraie enveloppe spectrale, c'est-à-dire celle qui passe exactement pas les sommets des lobes du spectre d'amplitude correspondant à des sinusoïdes, nous n'utilisons pas la modélisation AR (voir la partie II, section 2.4.3.3) ou le cepstre (voir la partie II, section 2.4.3.4) pour obtenir l'enveloppe spectrale. L'idée est de détecter les maximums locaux du spectre d'amplitude (correspondant à des lobes dus chacun à la présence d'un partiel) et d'interpoler (linéairement) entre eux. Bien sûr, les pics parasites (lobes secondaires, bruit...) détériorent énormément les performances de la méthode.

L'algorithme présenté ci-dessous est loin d'être parfait :

- Calcul des deux spectres d'amplitude.
- Lissage des « creux » de chaque spectre d'amplitude $\hat{S}$. Un creux est un minimum local :

$$\text{Si } (\hat{S}_{i-1} \geq \hat{S}_i \;\&\&\; \hat{S}_{i+1} > \hat{S}_i) \;||\; (\hat{S}_{i-1} > \hat{S}_i \;\&\&\; \hat{S}_{i+1} \geq \hat{S}_i),$$
$$\text{il y a un minimum local en l'échantillon fréquentiel de numéro d'ordre } i$$

avec $\geq$ est l'opérateur SUPÉRIEUR OU ÉGAL, && est l'opérateur ET, $>$ l'opérateur SUPÉRIEUR, et || l'opérateur OU.

Le lissage se fait ainsi : $\hat{S}_i = \frac{\hat{S}_{i-1} + \hat{S}_{i+1}}{2}$

Cette opération est répétée plusieurs fois sur tout le spectre d'amplitude, jusqu'à la convergence. Cette opération a pour effet d'éliminer certains des plus petits maximums locaux sans toucher aux plus grands. Comme il a déjà été dit, pour les besoins de la méthode il est important de ne pas modifier les plus grands maximums locaux.

- Nous détectons les maximums locaux de chaque spectre d'amplitude :

$$(\hat{S}_{i-1} \leq \hat{S}_i \;\&\&\; \hat{S}_{i+1} < \hat{S}_i) \;||\; (\hat{S}_{i-1} < \hat{S}_i \;\&\&\; \hat{S}_{i+1} \leq \hat{S}_i)$$

  et les minimums locaux. Remarquons que le fait de prendre les conditions données ci-dessus plutôt que plus simplement ($\hat{S}_{i-1} < \hat{S}_i \;\&\&\; \hat{S}_{i+1} < \hat{S}_i$), n'assure cependant pas encore que nous ne puissions pas avoir deux maximums locaux successifs (ou deux minimums) sans minimum local (ou maximum) intercalé entre eux !
- Certains maximums locaux sont éliminés : ceux pour lesquels le contraste avec les deux minimums locaux voisins (celui qui précède juste et celui qui vient juste après) est trop petit. Ainsi, nous espérons éliminer les maximums locaux correspondant à des lobes secondaires, au bruit, etc. Il faut que la moyenne des deux contrastes soit inférieure à 50 dB pour que le pic ne soit pas pris en compte.
- Interpolation linéaire entre les maximums locaux gardés. Nous obtenons alors les enveloppes spectrales.

D'autres algorithmes pour obtenir la vraie enveloppe spectrale sont succintement présentés dans la section 12.3.6.

12.3.6 Amélioration du calcul de l'enveloppe spectrale

Nous avons vu dans la section 12.3.5 qu'il n'est pas aisé de déterminer les enveloppes spectrales à partir de la détection des pics du spectre d'amplitude. Les fausses alarmes (pics de bruit ou des lobes secondaires détectés) aussi bien que les détections manquées (pics dus à des sinusoïdes non détectés) détériorent énormément les performances de la méthode. Faisons la remarque qu'ici, dans cette section 12.3, en aucun cas, nous ne voulons considérer la forme des lobes obtenus et la comparer à celle du lobe théorique (voir les sections 2.2.3, 2.4.1 et 2.4.2), puisque nous considérons que ces lobes sont déformés par les modulations d'amplitude et de fréquence.

12.3.6.1 Nouvelle méthode 1

Deux spectres d'amplitude sont calculés : l'un pour une « petite » portion du signal, l'autre pour une « grande » portion, les deux portions étant centrées au même instant. Les tailles des deux fenêtres d'analyse sont relativement proches. L'idée est que pour les deux spectres d'amplitude :

- les lobes dus à des sinusoïdes tombent aux mêmes endroits, c'est-à-dire aux mêmes échantillons fréquentiels,
- les lobes dus à des sinusoïdes ont sensiblement la même amplitude.

Et ceci même quand il y a du vibrato, puisque les fenêtres d'analyse ont quasi la même taille. Alors qu'au contraire :

- les pics dus aux lobes secondaires ou au bruit ne tombent pas forcément aux mêmes endroits,
- les pics dus aux lobes secondaires n'ont pas forcément la même amplitude.

Il s'agit de sélectionner les maximums locaux qui nous intéressent, c'est-à-dire ceux qui sont quasi identiques, en position et en amplitude, pour les deux spectres d'amplitude. Voir la figure 12.15. Une sinusoïde de fréquence 440 Hz est présente. Les maximums locaux dans le premier cas (portion large de 0,04 seconde) sont repérés par des + et des 1 ; dans le second cas (portion large de 0,05 seconde), par des O et des 2. Un seul pic est commun : celui qui a lieu vers 440 Hz.

12.3.6.2 Nouvelle méthode 2

Ici, nous utilisons exactement la même portion du signal, mais nous la multiplions par deux fenêtres de pondération différentes : l'une étant la fenêtre de BLACKMAN, et l'autre étant la fenêtre de HANNING, par exemple, ou l'une étant une fenêtre sans lobes secondaires (comme la fenêtre de HANNING-POISSON : voir l'annexe D). Ainsi, les lobes dus à des sinusoïdes tombent aux mêmes endroits (aux mêmes échantillons fréquentiels) pour les deux spectres d'amplitude et ils ont sensiblement la même amplitude (le vibrato ne changeant ici rien à l'affaire, puisque les fenêtres d'analyse ont la même taille), alors que pas forcément les pics dus aux lobes secondaires ou au bruit. Voir la figure 12.16. Une sinusoïde de fréquence 440 Hz est présente. Les maximums locaux dans le premier cas (fenêtre de pondération de BLACKMAN) sont repérés par des + et des 1 ; dans le second cas (fenêtre de pondération de HANNING), par des O et des 2. Un seul pic est commun : celui qui a lieu vers 440 Hz.

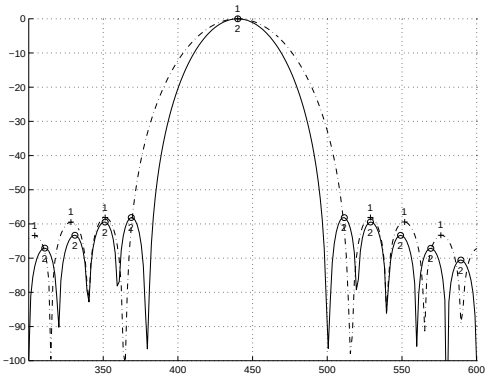


FIG. 12.15 – *En trait interrompu, le spectre d'amplitude obtenu pour une portion du signal large de* 0,04 *seconde ; en trait plein, le spectre d'amplitude obtenu pour une portion du signal centrée au même instant que la précédente mais large de* 0,05 *seconde. Dans les deux cas, la fenêtre de pondération utilisée est celle de* BLACKMAN. *En abscisse : la fréquence en* Hz *; en ordonnée : l'amplitude en dB*

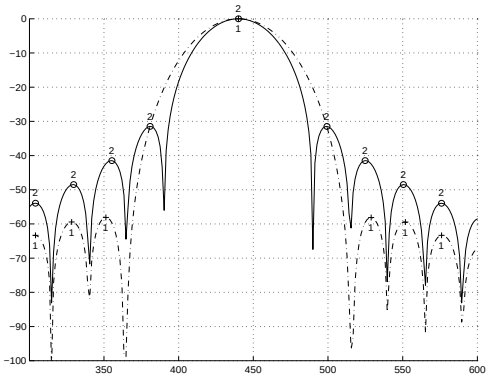


FIG. 12.16 – *En trait interrompu, le spectre d'amplitude obtenu pour une portion du signal multipliée par la fenêtre de pondération de* BLACKMAN *; en trait plein, le spectre d'amplitude obtenu pour la même portion du signal, mais multipliée par la fenêtre de pondération de* HANNING. *La fenêtre d'analyse est large de* 0,04 *seconde. En abscisse : la fréquence en* Hz *; en ordonnée : l'amplitude en dB*

12.3.6.3 Nouvelle méthode 3

Pour obtenir de meilleures enveloppes spectrales, des méthodes, notamment basées sur la « Ligne de Partage des Eaux (LPE) », méthodes venues de la segmentation des images et de la « morphologie mathématique » pourraient être utilisées. Ceci n'a pas été testé.

12.3.7 Les fonctions d'observation prises en compte

Les premiers tests ont été effectués sur ce signal simulé :

- $f_e = 44100\ Hz$: fréquence d'échantillonnage du signal sonore.
- $f_0^c = 440\ Hz$: fréquence fondamentale « centrale ».
- $f_{vib} = 5\ Hz$: fréquence du vibrato.
- $A_{vib} =$ de 0 à 25 Hz : amplitude du vibrato.
- $duree = 1\ seconde$: durée du son.
- $nb_{har} = 30$: nombre d'harmoniques.
- $amp = \exp\left[-\left(\frac{l-2}{2}\right)^2\right] + 0{,}1\exp\left[-\left(\frac{l-9}{3}\right)^2\right]$: amplitudes des harmoniques, soit :

$$(0{,}78 \quad 1{,}00 \quad 0{,}78 \quad 0{,}37 \quad 0{,}12 \quad 0{,}06 \quad 0{,}07 \quad 0{,}09 \quad 0{,}10 \quad 0{,}09 \quad 0{,}06\ldots).$$

  Nous avons donc deux « formants » : le premier autour du deuxième harmonique et le second autour du neuvième harmonique.
- phases des harmoniques : aléatoires, uniformément réparties entre 0 et 2π.

La taille des FFT est de $t_{FFT} = 4096$ points et les enveloppes spectrales sont calculées entre le 30ème et le 512ème échantillon fréquentiel (c'est-à-dire entre 322 Hz et 5512 Hz). Les tailles des deux fenêtres d'analyse sont 60 et 20 millisecondes. Pour calculer les enveloppes spectrales, nous n'avons pas utilisé les techniques que nous avons présentées dans la section 12.3.6, mais celle donnée dans la section 12.3.5.

12.3.7.1 Pente de la droite

Nous modélisons la différence : $diff = \frac{es_2 - es_1}{es_2}$ par une droite (polynôme d'ordre 1) $poly = \alpha f + \beta$. S'il y a du vibrato, la pente α doit être positive et grande. Ceci pour plusieurs pas d'analyse consécutifs : nous lissons la valeur de la pente obtenue sur une largeur de 0,3 seconde. S'il n'y a pas de vibrato, nous obtenons du bruit : le lissage nous donne une valeur proche de 0. Voir les figures 12.17 (trajet de α pour un signal modulé en fréquence) et 12.18 (moyenne de α en fonction de A_{vib}).

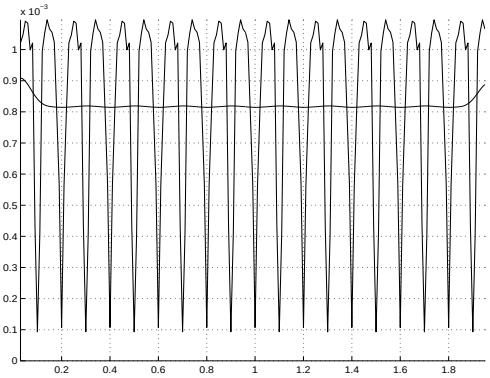


FIG. 12.17 – *Trajet de la « pente de la droite ». Un vibrato d'amplitude* 20 Hz *est présent. En abscisse : le temps en seconde ; en ordonnée : la valeur de la pente*

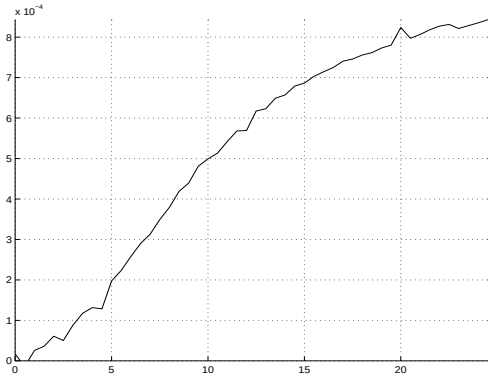


FIG. 12.18 – *Moyenne de la « pente* α *de la droite » en fonction de l'amplitude* A_{vib} *du vibrato. En abscisse :* A_{vib} *; en ordonnée : la valeur de la pente*

12.3.7.2 « Flux » entre les deux enveloppes spectrales

Le flux entre les deux enveloppes spectrales (normalisé par celle obtenue pour la petite fenêtre d'analyse) doit être *grand* s'il y a du vibrato, et *petit* s'il n'y en a pas. La valeur du flux est lissée

sur une largeur de 0,3 seconde (soit, ici, une période et demie du vibrato). Le lissage est effectué en filtrant passe-bas. Voir les figures 12.19 (trajet du « flux » pour un signal modulé en fréquence) et 12.20 (moyenne du « flux » en fonction de A_{vib}).

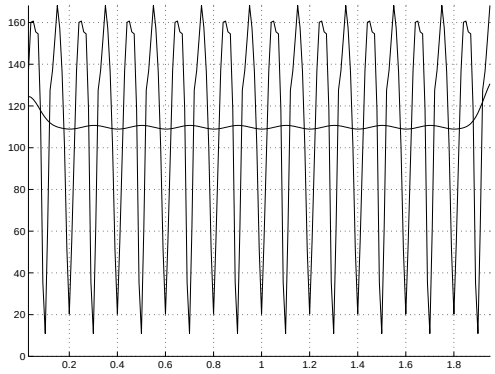


FIG. 12.19 – *Trajet du « flux ». Un vibrato d'amplitude* 20 Hz *est présent. La courbe non lissée et la courbe lissée sont données. En abscisse : le temps en seconde ; en ordonnée : la valeur moyenne du « flux »*

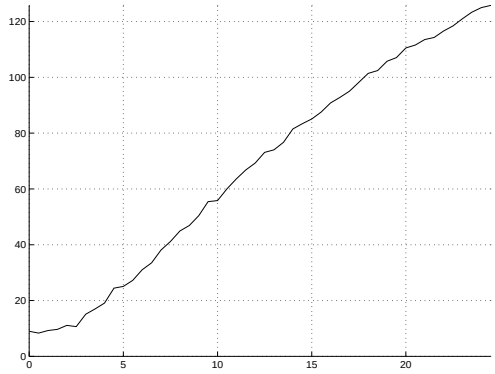


FIG. 12.20 – *Moyenne du « flux » en fonction de l'amplitude A_{vib} du vibrato. En abscisse : A_{vib} ; en ordonnée : la valeur du « flux »*

12.3.8 Une première remarque : le trémolo, son influence

L'influence du trémolo sur les enveloppes spectrales est très différente de l'influence du vibrato. Ainsi, un signal harmonique, composé de L sinusoïdes, un trémolo étant présent, s'écrit :

$$s(t) = \sum_{l=1}^{L} B_l \left[1 + m^{(l)} \cos\left(2\pi f_{tré}^{(l)} + \varphi_{tré}^{(l)}\right)\right] \cos\left(2\pi l f_0^c t + \varphi_l\right)$$

Dans le cas de la voix chantée, le trémolo dépendant du vibrato, toutes les fréquences $f_{tré}^{(l)}$ sont égales, et valent la fréquence du vibrato f_{vib}. Dans le cas général, ce n'est pas forcément vrai.

Ainsi, le lobe de chaque harmonique se décompose en trois lobes : un en lf_0^c d'amplitude $\frac{B_l}{2}$, un autre en $lf_0^c - f_{tré}^{(l)}$ d'amplitude $\frac{B_l m^{(l)}}{2}$ et le dernier en $lf_0^c + f_{tré}^{(l)}$ d'amplitude $\frac{B_l m^{(l)}}{2}$. Le lobe principal de chaque harmonique est détérioré indépendamment de ceux des autres harmoniques, et tous les lobes principaux sont élargis d'autant, contrairement à ce qui se passe pour la modulation de fréquence. Si nous considérons les deux fenêtres d'analyse précédentes, de tailles respectives 60 millisecondes et 20 millisecondes :

- Le flux entre les enveloppes spectrales est plus grand quand il y a du trémolo que quand il n'y en a pas.
- Nous ne pouvons pas mettre en évidence une croissance quasi linéaire en fonction de la fréquence de la différence entre les deux enveloppes spectrales.

Ainsi, les influences respectives du vibrato et du trémolo sur les enveloppes spectrales sont différentes.

12.3.9 Une seconde remarque : application de la méthode à la *segmentation en notes et/ou en phones (ou plus généralement en zones stables)*

Pour la *segmentation en zones stables* (voir le chapitre 2 de la partie II), nous obtenons de nouvelles fonctions d'observation. Voir les « flux spectraux » décrits dans la section 2.4.3.

12.3.9.1 Méthodes du type « flux spectral »

Pour avoir la définition du terme « flux spectral », voir la section 2.4.3.1 de la partie II, page 27. Il est deux versions à cette fonction d'observation (voir la section 2.4.3.3 de la partie II, page 28) :

- Nous calculons $|dF_1^{max}|$ la « valeur absolue de la dérivée du flux entre le spectre d'amplitude et l'enveloppe spectrale calculés sur une même portion du signal ».
- Nous calculons F_2^{max} le « flux entre deux enveloppes spectrales calculées pour deux portions successives, décalées de Q échantillons ».

12.3.9.2 Nouvelle méthode

Voir la section 12.3.7.2 ci-dessus. Nous utilisons directement le trajet F_3^{max} de ce « flux », mais avant le lissage.

12.3.10 Conclusion

Quand il y a du vibrato, la différence relative $diff = \frac{es_2 - es_1}{es_2}$ croît quasi linéairement avec f : $diff \simeq poly = \alpha f + \beta$, donc α est grand et positif. La **décision** est prise ainsi :

- Si α et le « flux » sont grands, il y a du vibrato.

Il est difficile d'obtenir les **estimations** de f_{vib}, A_{vib} et φ_{vib}.

Nous constatons sur les figures 14.3, 14.4, 14.5 et 14.6, obtenues pour des signaux réels, que les deux tests nous donnent les résultats attendus. Cependant, la qualité des enveloppes spectrales est à mettre en cause : « quelques » pics parasites de temps en temps ne sont pas éliminés. Nous obtenons alors des résultats aberrants, que le lissage tend à éliminer, sans y parvenir forcément.

12.4 Méthode de Laroche

12.4.1 Présentation

La méthode est décrite dans la thèse de Laroche (voir [Lar89]).

Nous modélisons le signal x comme la somme de L sinusoïdes complexes dont les amplitudes instantanées sont représentées par des polynômes d'ordre q aux coefficients b complexes. Ainsi, les modulations de fréquence et d'amplitude sont modélisées dans ces coefficients. Nous avons (avec $\overline{x}$ le signal estimé) :

$$\overline{x}_n = \sum_{l=1}^{L} \sum_{m=0}^{q} b_{l,m} n^m z_l^n$$

Avec : $z_l = \exp\left(2\pi i \frac{f_l}{f_e}\right)$ et n le numéro de l'échantillon (la taille du signal est $N+1$: $n \in [0 \dots N]$). Nous pouvons réécrire ce signal ainsi :

$$\underline{\mathbf{\Phi}}\ \underline{B} = \underline{\overline{x}}$$

Nous utilisons les mêmes notations que dans la section 12.2.5. Pour indiquer que nous sommes en présence d'une matrice et non d'un vecteur, nous utilisons les caractères gras (voir $\underline{\mathbf{\Phi}}$). Avec (en considérant que $0^0 = 1$) :

$$\underline{\overline{x}} = [\overline{x}_0 \dots \overline{x}_N]^t$$

$$\underline{B} = [\underline{b}_1 \dots \underline{b}_L]^t$$

$$\underline{b}_l = [b_{l,1} \dots b_{l,q}]$$

$$\underline{\mathbf{\Phi}} = [\ \underline{\mathbf{\Phi_1}}\ \dots\ \underline{\mathbf{\Phi_L}}\ ]$$

$$\underline{\mathbf{\Phi_l}} = \underbrace{\left.\begin{bmatrix} 1 & 0 & \dots & 0 \\ z_l & z_l & \dots & z_l \\ z_l^2 & 2z_l^2 & \dots & 2^q z_l^2 \\ \vdots & \vdots & \ddots & \vdots \\ z_l^N & N z_l^N & \dots & N^q z_l^N \end{bmatrix}\right\} N+1 \text{ lignes}}_{q \text{ colonnes}}$$

Pour estimer $\underline{B}$, nous utilisons les moindres carrés (voir la section 12.2.5). Ici, le problème est linéaire. Nous voulons minimiser l'erreur quadratique :

$$\epsilon = ((\ \underline{\mathbf{\Phi}}\ \underline{B})^H - \underline{x}^H)(\ \underline{\mathbf{\Phi}}\ \underline{B} - \underline{x})$$

Nous obtenons finalement :

$$\underline{B} = \left[\underline{\mathbf{\Phi}}^{\boldsymbol{H}}\underline{\mathbf{\Phi}}\right]^{-1}\ \underline{\mathbf{\Phi}}^{\boldsymbol{H}}\ \underline{x}$$

Faisons la remarque que la matrice $\underline{\mathbf{\Phi}}$ comprend $N+1$ lignes et qL colonnes, et que le vecteur colonne $\underline{B}$ comprend qL éléments.

12.4.2 Quelques exemples

Faisons l'hypothèse que $L = 1 : \overline{x}_n = \sum_{m=0}^{q} b_{1,m} n^m z_1^n$

- Si la fréquence et l'amplitude sont fixes :

$$x_n = \exp\left(2\pi i \frac{f_0}{f_e} n\right) = \left[\exp\left(2\pi i \frac{f_0}{f_e}\right)\right]^n = z_1^n$$

  et donc : $\underline{b}_1 = [1 \quad 0 \quad \dots \quad 0]$.
- Si la fréquence est modulée sinusoïdalement (vibrato) :

$$f_0 = A_{vib} \cos\left(2\pi \frac{f_{vib}}{f_e} n\right) + f_0^c$$

$$x_n = \exp\left[i \frac{A_{vib}}{f_{vib}} \sin\left(2\pi \frac{f_{vib}}{f_e} n\right) + 2\pi i \frac{f_0^c}{f_e} n\right] = \exp\left[i \frac{A_{vib}}{f_{vib}} \sin\left(2\pi \frac{f_{vib}}{f_e} n\right)\right] z_1^n$$

  Or, $f_{vib} \ll f_e$ et nous considérons que n reste petit, donc $\sin\left(2\pi \frac{f_{vib}}{f_e} n\right) \simeq 2\pi \frac{f_{vib}}{f_e} n$. Ainsi :

$$x_n \simeq \exp\left[i 2\pi \frac{A_{vib}}{f_e} n\right] z_1^n$$

  Or, $A_{vib} \ll f_e$ et nous considérons que n reste petit, donc $\exp\left[i 2\pi \frac{A_{vib}}{f_e} n\right] \simeq 1 + i 2\pi \frac{A_{vib}}{f_e} n$. Ainsi :

$$x_n \simeq \left[1 + i 2\pi \frac{A_{vib}}{f_e} n\right] z_1^n$$

  et donc : $\underline{b}_1 = \left[1 \quad i 2\pi \frac{A_{vib}}{f_e} \quad \dots\right]$.
- Si l'amplitude est modulée sinusoïdalement (trémolo) :

$$x_n = \left[1 + A_{tré} \cos\left(2\pi \frac{f_{tré}}{f_e} n\right)\right] z_1^n$$

Or, $f_{tré} \ll f_e$ et nous considérons que n reste petit, donc $\cos\left(2\pi\frac{f_{tré}}{f_e}n\right) \simeq 1-\frac{1}{2}\left(2\pi\frac{f_{tré}}{f_e}n\right)^2$.

Et donc : $\underline{b}_1 = \left[1 \quad 0 \quad -\frac{1}{2}\left(2\pi\frac{f_{tré}}{f_e}\right)^2 \quad \dots\right]$.

Nous donnons dans la section qui suit quelques résultats de simulations effectuées sous MATLAB.

12.4.3 Simulations

Nous avons pris $L = 1$. Le signal que nous avons simulé est le suivant :

$$x_n = A_n \exp(i\phi_n)$$

Avec :

$$\phi_n = 2\pi f_0^{c(v)}\frac{n}{f_e} + \frac{A_{vib}^{(v)}}{f_{vib}^{(v)}}\sin\left(2\pi f_{vib}^{(v)}\frac{n}{f_e} + \varphi_{vib}^{(v)}\right) + \varphi_0^{(v)}$$

et :

$$A_n = 1 + a_0^{(v)}\left(\frac{n}{f_e}\right) + a_1^{(v)}\left(\frac{n}{f_e}\right)^2 + a_2^{(v)}\left(\frac{n}{f_e}\right)^3$$

Les valeurs de $f_0^{c(v)}$, $A_{vib}^{(v)}$, $f_{vib}^{(v)}$, $\varphi_{vib}^{(v)}$, $\varphi_0^{(v)}$, $a_0^{(v)}$, $a_1^{(v)}$, $a_2^{(v)}$, f_e et $N+1$ sont données dans le tableau 12.2.

$f_0^{c(v)}$	=	100 Hz
$A_{vib}^{(v)}$	=	20 Hz
$f_{vib}^{(v)}$	=	10 Hz
$\varphi_{vib}^{(v)}$	=	1 rad
$\varphi_0^{(v)}$	=	2 rad
$a_0^{(v)}$	=	1
$a_1^{(v)}$	=	20
$a_2^{(v)}$	=	4000
f_e	=	44100 Hz
$N+1$	=	2206

TAB. 12.2 – *Vraies valeurs des paramètres pour la méthode de* LAROCHE

La taille de la fenêtre d'analyse est $T = 0{,}05$ seconde ($n \in [0 \ \dots \ 2205]$). La fréquence instantanée du signal passe de 110,8 Hz pour $t = 0$ seconde à 89,2 Hz pour $t = 0{,}05$ seconde. La partie réelle du signal simulé est présentée sur la figure 12.21.

Nous donnons sur les figures 12.22 et 12.23 les trajets de $|x|$ et de $|\overline{x}|$ (en haut à gauche), de $|x| - |\overline{x}|$ (en bas à gauche), de $Arg(x)$ et de $Arg(\overline{x})$ (en haut à droite), de $Arg(x) - Arg(\overline{x})$ (en bas à droite) ; ce respectivement pour $q = 3$ et $q = 10$, avec $f_1 = 85$ Hz : ceci veut dire que :

$$\overline{x}_n = \sum_{m=0}^{q} b_{1,m} n^m \exp\left(2\pi i\frac{f_1}{f_e}n\right)$$

L'argument Arg de $\overline{x}_n$ est choisi parmi ces trois mesures $Arg_n^{(1)}$, $Arg_n^{(2)}$ et $Arg_n^{(3)}$:

- $Arg_n^{(1)} = Arg(\overline{x}_n) - 2\pi$
- $Arg_n^{(2)} = Arg(\overline{x}_n)$
- $Arg_n^{(3)} = Arg(\overline{x}_n) + 2\pi$

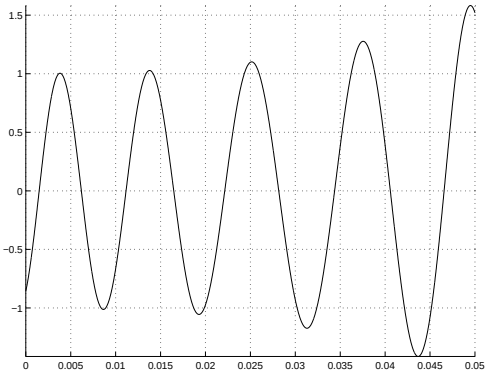


FIG. 12.21 – *Partie réelle du signal simulé. En abscisse : le temps en seconde ; en ordonnée : l'amplitude des échantillons*

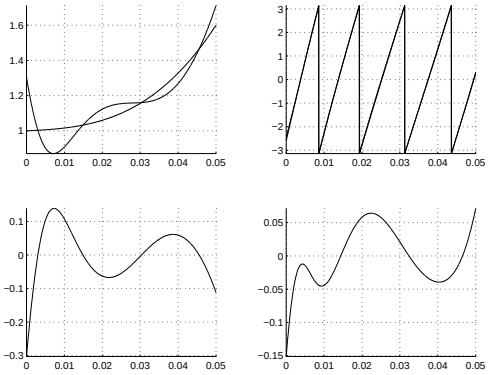


FIG. 12.22 – *À gauche, en haut : amplitude vraie et amplitude estimée ; en bas : différence. À droite, en haut : phase vraie et phase estimée ; en bas : différence. Ordre du modèle : 3*

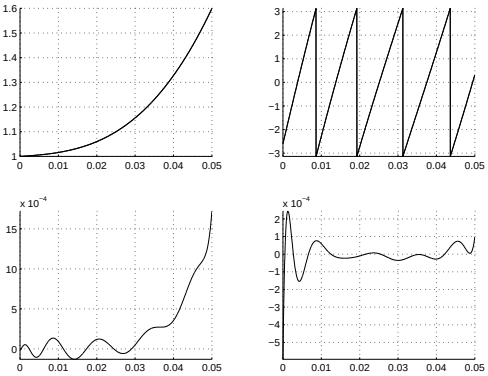


FIG. 12.23 – *À gauche, en haut : amplitude vraie et amplitude estimée ; en bas : différence. À droite, en haut : phase vraie et phase estimée ; en bas : différence. Ordre du modèle : 10*

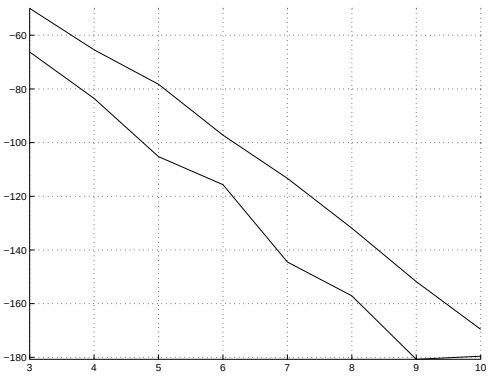


FIG. 12.24 – *Erreur moyenne E suivant l'ordre q et la fréquence f_1. Courbe supérieure : $f_1 = f_0^{c(v)}$; courbe inférieure : $f_1 = 85\ Hz$*

Les trois différences $\left|Arg_n^{(i)} - Arg(x_n)\right|$ sont calculées. Si la plus petite a lieu pour l'indice j, nous décidons que $Arg(\overline{x}_n) = Arg_n^{(j)}$.

Nous calculons l'erreur moyenne ainsi :

$$E = \frac{\sum_{n=0}^{N} ||x_n| - |\overline{x}_n||}{\sum_{n=0}^{N} |x_n|} + \frac{\sum_{n=0}^{N} |Arg(x_n) - Arg(\overline{x}_n)|}{\sum_{n=0}^{N} |Arg(x_n)|}$$

Nous donnons sur la figure 12.24 cette erreur moyenne en dB ($E_{dB} = 20\log_{10}(E)$), selon l'ordre q du polynôme utilisé (abscisse) et suivant la fréquence f_1 utilisée pour calculer $\mathbf{\Phi}$. Ainsi, la courbe du haut correspond à $f_1 = f_0^{c(v)}$ et celle du bas à $f_1 = 85\ Hz$. Nous voyons qu'environ 20 dB les séparent.

Le problème principal qui se pose quand nous augmentons l'ordre du modèle est que le conditionnement de la matrice $\mathbf{\Phi^H \Phi}$ (de taille $q \times q$, ou dans le cas général $qL \times qL$) à inverser devient de plus en plus mauvais, du fait que certaines de ses valeurs sont beaucoup plus grandes que d'autres (comparer 1 à N^q). Ainsi l'inversion de cette matrice est plus difficile. Cependant, Laroche s'était limité à l'ordre 3 (page 87 de [Lar89]) : nous pouvons dire que la méthode est efficace même pour des ordres supérieurs. Il suffit pour cela de constater que, pour construire $\mathbf{\Phi}$, à la place d'utiliser n, les numéros d'ordre des échantillons, nous pouvons utiliser nt_e, où t_e est la période d'échantillonnage du signal en seconde, ou toute autre valeur $n\alpha$, où α est une constante. Ceci nous permet de résoudre en partie les problèmes de conditionnement : par tâtonnement, nous choisissons un α qui nous permet d'inverser correctement la matrice $\mathbf{\Phi^H \Phi}$.

Le problème de la détection du vibrato (modulation de fréquence) et le problème de la détection du trémolo (modulation d'amplitude) peuvent être traités par cette méthode. Ces informations sont contenues dans le vecteur $\underline{B}$. Mais nous n'avons pas essayé de « séparer » les deux influences, de repérer ce qui dans $\underline{B}$ est dû à un vibrato et ce qui est dû à un trémolo. D'autres influences sont à considérer, comme les glissandos, en amplitude et en fréquence. Il s'agit de construire un critère de décision spécifique pour chaque influence, basé sur l'étude de la « forme » de $\underline{B}$. De plus, au lieu de considérer des signaux complexes, il serait intéressant de plutôt tout réécrire pour des signaux réels ; et il serait intéressant aussi d'étudier la sensibilité de cette méthode au bruit. Il s'agit de perspectives.

12.5 Conclusion

Pour détecter le vibrato et estimer ses paramètres à partir de l'analyse directe du son, il existe d'autres méthodes, comme celle de Masri (voir [Mas96], pages 74 - 85). Masri prend en compte la distorsion qu'induit dans la phase du lobe principal une modulation d'amplitude exponentielle, et celle qu'induit une modulation linéaire de fréquence. Ces deux distorsions sont selon Masri séparables.

En ce qui concerne le trémolo, il faudrait réécrire les modèles de signal utilisés par la première et la troisième méthodes décrites dans ce chapitre. La deuxième méthode ne peut pas être adaptée au cas du trémolo.

Les trois méthodes utilisent l'apport des harmoniques de numéros d'ordre supérieurs du signal pour détecter le vibrato.

En ce qui concerne la détection des harmoniques du vibrato, il faut remarquer que pour la première méthode nous aboutissons à un modèle avec des doubles sommes infinies si nous considérons deux harmoniques du vibrato ; qu'il semble difficile d'adapter la deuxième méthode pour ce faire ; et qu'il faudrait réecrire le modèle de signal pour la troisième.

Chapitre 13

Méthodes de détection du vibrato à partir du trajet de f_0

13.1 Préambule

Le signal traité ici est le trajet de f_0, trajet déterminé par le logiciel f_0 (voir la section 2.2.1.2, page 12). Nous travaillons dans toute la suite avec une fréquence d'échantillonnage pour f_0 de 100 Hz. Cette fréquence d'échantillonnage est celle qui est utilisée par défaut par les logiciels f_0 et ADDITIVE.

Les méthodes décrites dans ce chapitre ont été seulement appliquées sur le trajet de la fréquence fondamentale f_0. L'apport de la prise en compte du vibrato présent sur les harmoniques du signal de numéros d'ordre supérieurs n'a pas été étudié. Il s'agit de perspectives.

Pour améliorer l'estimation des paramètres du vibrato, il faudrait tenir compte de la position des transitions entre notes : il ne faudrait pas que les portions de f_0 analysées chevauchent l'une de ces transitions. Ainsi, il faut noter que le problème de la *segmentation en zones stables* et celui de l'*analyse du vibrato* sont indissociables : l'étude du vibrato a pour but d'améliorer les performances de la segmentation ; mais l'étude du vibrato serait facilitée si elle était précédée d'une segmentation efficace.

La méthode décrite dans la **deuxième section** (section 13.2) de ce chapitre est basée sur des techniques classiques d'analyse spectrale, adaptée pour nos besoins. La méthode, simple et robuste, décrite dans la **troisième section** (section 13.3) de ce chapitre est la seule qui ait été implémentée dans le programme *segmentation*. La méthode décrite dans la **quatrième section** (section 13.4) vient du traitement de la parole. La **cinquième section** (section 13.5) de ce chapitre constitue une conclusion à ce chapitre.

13.2 Méthode basée sur la prédiction linéaire

13.2.1 Principe de la méthode

Cette méthode repose sur des procédés classiques de l'analyse spectrale : la transformée de FOURIER et la modélisation AR.

La valeur de la fréquence fondamentale du vibrato est en général comprise entre 3 Hz et 11 Hz. Quand nous voulons détecter les sinusoïdes d'un signal périodique à partir du spectre d'amplitude calculé avec la FFT, la taille de la fenêtre d'analyse doit être d'environ trois fois la période fondamentale de ce signal. Ainsi, dans notre cas, où le signal est le trajet de f_0, cette taille doit être d'environ 1 seconde, dans le cas le plus défavorable, c'est-à-dire quand f_{vib} vaut 3 Hz. Cette taille est trop importante : le signal n'est pas stationnaire sur 1 seconde, ne serait-ce que parce que généralement les notes durent moins d'une seconde. Mais elle est trop petite aussi,

puisque cela correspond à une fenêtre d'analyse large de $N = 100$ échantillons, donnant donc après FFT une précision fréquentielle de $\frac{f_e}{N} = 1\ Hz$.

L'idée ici est d'utiliser les coefficients de la modélisation AR pour prédire les valeurs passées et futures de la portion de signal à partir de laquelle ils ont été calculés. Ainsi, nous pouvons et/ou réduire la taille de la fenêtre d'analyse effective (c'est-à-dire dont les échantillons sont du signal utile, c'est-à-dire des échantillons du trajet de f_0 original) et/ou augmenter la précision fréquentielle, ceci tout en ayant encore accès aux informations de phase et d'amplitude instantanées, qui sont des paramètres du vibrato que nous voulons estimer. Sur la figure 13.1, le synoptique de la méthode est donné.

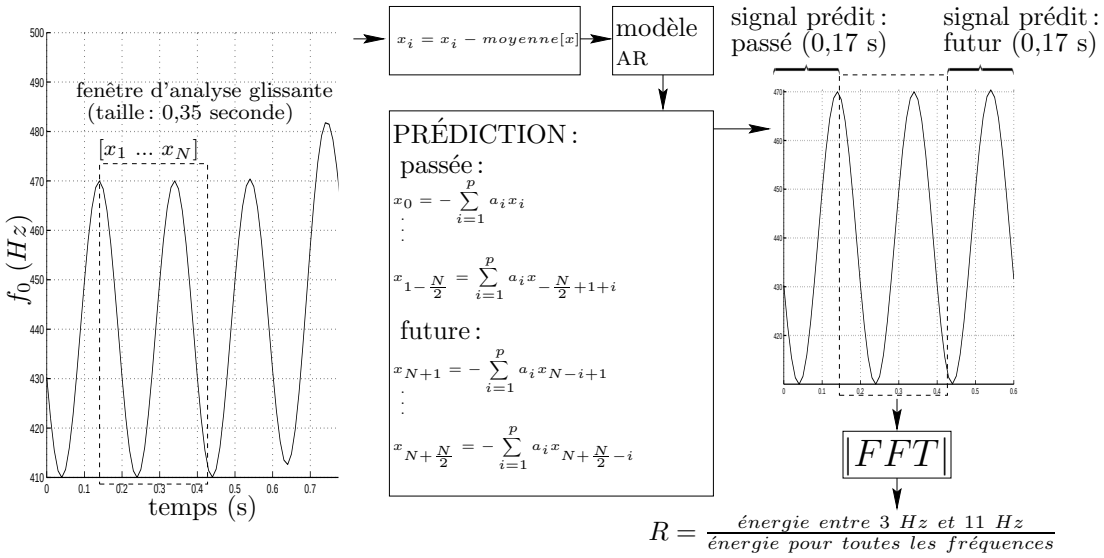


FIG. 13.1 – *Détection et estimation des paramètres du vibrato grâce à la prédiction linéaire*

13.2.2 Quelques simulations

13.2.2.1 Première simulation

D'abord, nous avons considéré un son formé de trois notes sur lesquelles un vibrato est présent. Nous avons simulé le trajet de f_0 suivant :

- première note : durée $t^{(1)} = 0{,}5$ seconde, $f_0^{(1)} = 200\ Hz$, $A_{vib}^{(1)} = 5$ et $f_{vib}^{(1)} = 5\ Hz$
- deuxième note : durée $t^{(2)} = t^{(1)}$ seconde, $f_0^{(2)} = 100\ Hz$, $A_{vib}^{(2)} = A_{vib}^{(1)}$ et $f_{vib}^{(2)} = 8\ Hz$
- troisième note : durée $t^{(3)} = t^{(1)}$ seconde, $f_0^{(3)} = 150\ Hz$, $A_{vib}^{(3)} = A_{vib}^{(1)}$ et $f_{vib}^{(3)} = 6\ Hz$

De plus, six harmoniques du vibrato sont ajoutés, ceci pour chaque note. L'amplitude de l'harmonique du vibrato de numéro d'ordre i est égale à $\frac{1}{i}$. L'amplitude du vibrato est modulée par un cosinus relevé de période la durée de chaque note. Un bruit normal de variance faible ($\sigma^2 = 0{,}0001$) est finalement ajouté.

Pour modéliser les transitions, nous utilisons le modèle décrit dans la section 11.2.3.1. Nous avons :

$$f_0(t) = f_0^{(1)} + \left[\tanh\left(\frac{t - a^{(12)}}{b}\right) + 1\right] c^{(12)} + \left[\tanh\left(\frac{t - a^{(23)}}{b}\right) + 1\right] c^{(23)} + \sum_{i=1}^{7} \frac{A_{vib}(t)}{i} \cos\left(\phi_i(t)\right)$$

où le terme $\sum_{i=1}^{7} A_{vib}(t)\cos(\phi_i(t))$ modélise les transitions en fréquence et en amplitude du vibrato ; et où : $a^{(12)} = t^{(1)}$, $a^{(23)} = t^{(1)} + t^{(2)}$, $c^{(12)} = \frac{f_0^{(2)} - f_0^{(1)}}{2}$, $c^{(23)} = \frac{f_0^{(3)} - f_0^{(2)}}{2}$, $b = 0{,}1$ et $t \in \begin{bmatrix} 0 & 3t^{(1)} \end{bmatrix}$. Nous avons :

$$A_{vib}(t) = A_{vib}^{(1)}\left[\cos\left(2\pi\frac{t}{t^{(1)}} - \pi\right) + 1\right]$$

et :

$$\begin{aligned}\phi_i(t) \quad &= \phi_i^{(0)} + 2\pi f_{vib}^{(1)} it + \\ &2\pi i\frac{f_{vib}^{(2)} - f_{vib}^{(1)}}{2}t + 2\pi i\frac{f_{vib}^{(2)} - f_{vib}^{(1)}}{2}b\left[\log_e\left(\cosh\left(\frac{t - t^{(1)}}{b}\right)\right) - \log_e\left(\cosh\left(-\frac{t^{(1)}}{b}\right)\right)\right] + \\ &2\pi i\frac{f_{vib}^{(3)} - f_{vib}^{(2)}}{2}t + 2\pi i\frac{f_{vib}^{(3)} - f_{vib}^{(2)}}{2}b\left[\log_e\left(\cosh\left(\frac{t - t^{(1)} - t^{(2)}}{b}\right)\right) - \log_e\left(\cosh\left(-\frac{t^{(1)} + t^{(2)}}{b}\right)\right)\right]\end{aligned}$$

Les phases $\phi_i^{(0)}$ sont des variables aléatoires uniformément réparties entre 0 et 2π.

Nous considérons des portions larges de 0,35 seconde (35 échantillons) du signal $f_0(t)$. La moyenne de chaque portion est normalisée à 0. Nous modélisons chaque portion moyennée par un processus auto-régressif d'ordre 16 (la méthode de BURG est utilisée : voir [Mar80]) et nous prédisons son futur, ainsi que son passé, sur 17 échantillons. Nous obtenons donc un portion large de 69 (35+17+17) échantillons.

Nous présentons sur les figures 13.2, 13.3, 13.4 et 13.5 respectivement le signal simulé, le trajet de la fréquence du vibrato obtenue à partir du spectre d'amplitude calculé sur chaque portion moyennée et fenêtrée (la fenêtre de pondération utilisée étant celle de BLACKMAN) large de 35 échantillons (la fréquence du vibrato est égale à la fréquence de l'echantillon fréquentiel pour lequel ce spectre d'amplitude passe par son maximum), le trajet de la fréquence du vibrato obtenue à partir de la densité spectrale de puissance calculée grâce aux coefficients de la modélisation AR (pour la définition de la densité spectrale de puissance, voir la section 2.2.6.1, partie II, page 17) pour cette portion, et enfin le trajet de la fréquence du vibrato obtenue à partir du spectre d'amplitude calculé sur chaque portion moyennée et fenêtrée (la fenêtre de pondération utilisée étant celle de BLACKMAN) large de 69 échantillons (c'est-à-dire les prédictions passées et futures étant prises en compte).

Nous voyons que les résultats sont corrects, sauf au moment des transitions. Les moyennes de R sur tout le signal sont respectivement, pour le spectre d'amplitude sans prédiction, la densité spectrale de puissance et le spectre d'amplitude avec prédiction 0,6168, 0,6479 et 0,6522 (elles valent 0,7337, 0,78 et 0,7835 si seulement deux harmoniques du vibrato sont considérés).

13.2.2.2 Seconde simulation

Nous donnons ici des résultats similaires à ceux que nous avons donnés dans la section précédente. Cependant, le signal n'a pas été simulé par nous : entre autres, nous ne connaissons ni la fréquence du vibrato, ni le modèle de transition utilisés. Les fichiers proviennent de Peter DESAIN et de Henkjan HONING, de l'« Institute for Logic, Language and Computation (ILLC) », d'AMSTERDAM (voir [Hon94]). Le trajet de f_0 est échantillonné à 333 Hz et nous utilisons des portions larges de 0,3 seconde. Donc nous calculons la FFT avec 100 échantillons utiles : ceci correspond à à peine plus d'une période du vibrato. Nous utilisons les résultats de la modélisation AR et de la prédiction. L'ordre des modèles est 6. Nous contatons sur les figures 13.6 et 13.7 que l'algorithme simple que nous utilisons est robuste, c'est-à-dire que le vibrato mesuré n'est pas absurde en comparaison de celui que nous observons « à l'œil », sauf aux moments des transitions, qui posent un problème non résolu : nous voulons extraire le vibrato pour améliorer la segmentation, mais une bonne segmentation améliorerait l'extraction du vibrato.

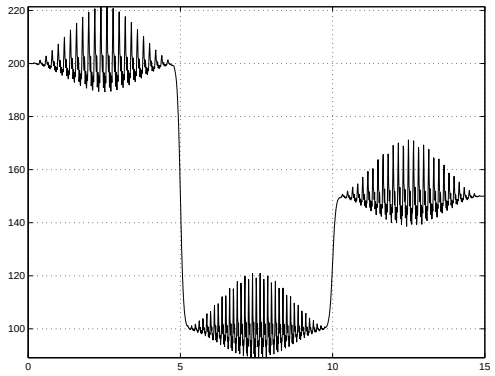


FIG. 13.2 – *Trajet simulé de la fréquence fondamentale f_0. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

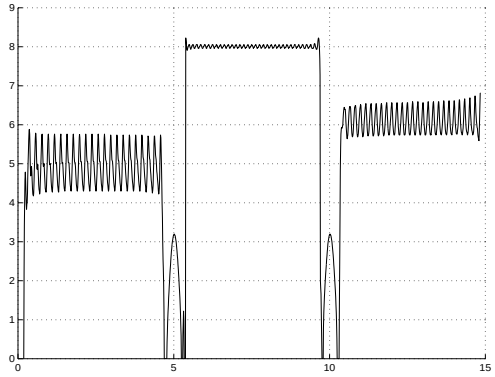


FIG. 13.3 – *Fréquence du vibrato obtenue grâce aux spectres d'amplitude du signal $f_0(t)$. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

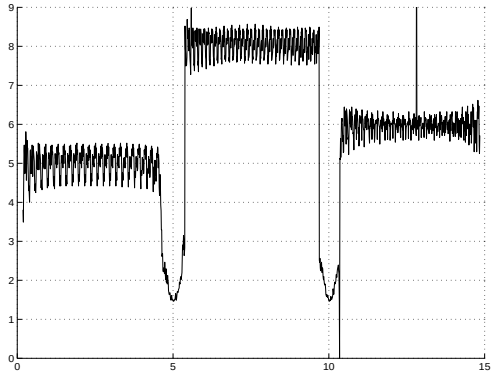


FIG. 13.4 – *Fréquence du vibrato obtenue grâce à la densité spectrale de puissance calculée avec les coefficients* AR *pour le signal $f_0(t)$. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

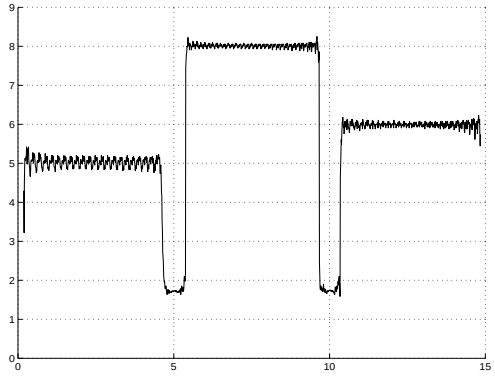


FIG. 13.5 – *Fréquence du vibrato obtenue grâce aux spectres d'amplitude du signal $f_0(t)$ avec prédiction. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

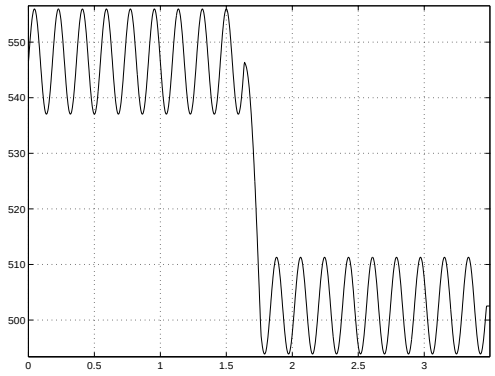


FIG. 13.6 – *Trajet simulé de la fréquence fondamentale f_0. Modèle utilisé non connu. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

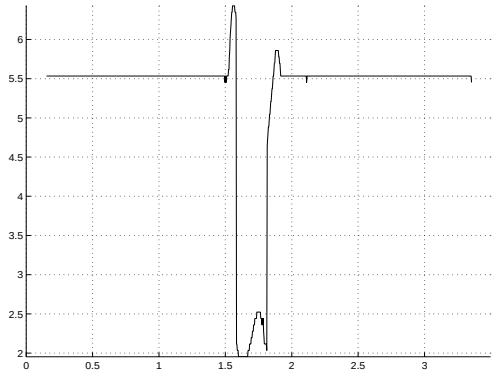


FIG. 13.7 – *Fréquence du vibrato obtenue grâce aux spectres d'amplitude du signal $f_0(t)$ avec prédiction. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

13.2.3 Conclusion

La **décision** est prise ainsi :

- Si R est grand, il y a du vibrato.

Les **estimations** des paramètres du vibrato sont faites ainsi :

- $A_{vib}(n)$ = **valeur** du maximum de $|FFT|(n)$ entre 3 Hz et 11 Hz
- $f_{vib}(n)$ = **position** du maximum de $|FFT|(n)$ entre 3 Hz et 11 Hz

13.3 Méthode basée sur les minimums et les maximums locaux

13.3.1 La méthode

13.3.1.1 Principe

Un algorithme pour détecter, estimer et aussi supprimer le vibrato est présenté ici. Cet algorithme a été implémenté dans le programme *segmentation* (voir la partie II).

Son principe est le suivant. Le signal est $f_0(t_i)$, avec $i \in [0 \ \ldots \ N]$. Nous détectons les M maximums locaux $M(k)$, survenant aux instants tM_k. Un maximum local a lieu en l'instant i si :

$$f_0(t_{i-1}) < f_0(t_i) \ \&\& \ f_0(t_i) > f_0(t_{i+1})$$

D'autres détecteurs de maximums locaux peuvent être utilisés. Par exemple celui-ci (voir la remarque faite dans la section 12.3.5, page 108, à propos de la possibilité d'avoir deux maximums locaux successifs sans minimum local intermédiaire) :

$$(f_0(t_{i-1}) \leq f_0(t_i) \ \&\& \ f_0(t_i) > f_0(t_{i+1})) \ \ || \ \ (f_0(t_{i-1}) < f_0(t_i) \ \&\& \ f_0(t_i) \geq f_0(t_{i+1}))$$

De la même façon, nous obtenons les m minimums locaux $m(k)$ aux instants tm_k.

Nous interpolons linéairement entre ces maximums pour obtenir une enveloppe supérieure du signal original échantillonnée aux instants t_i. Nous appelons cette enveloppe $E(t_i)$. Nous obtenons l'enveloppe inférieure de la même façon avec les minimums locaux. Nous l'appelons $e(t_i)$. Finalement, le signal « trajet de f_0 une fois le vibrato supprimé », appelé SV, est, à tout instant d'échantillonnage t_i de f_0, calculé ainsi :

$$SV(t_i) = 0{,}5\,[E(t_i) + e(t_i)]$$

13.3.1.2 Effets du vibrato – point de vue global

Nous considérons dans cette section le signal « trajet de f_0 » dans sa globalité, c'est-à-dire sur tout le son, soit quelques secondes ou dizaines de secondes. Donc, dans un premier temps, tous les maximums locaux du trajet de f_0 sont détectés. La liste LISTMAX de ces maximums est dressée. Le trajet MAXINTERP est obtenu en interpolant entre ces points. Toutes les distances temporelles entre deux maximums locaux successifs sont calculées, ce qui nous donne la liste DISTMAX.

Dans un deuxième temps, les mêmes traitements sont effectués avec les minimums locaux du trajet de f_0. Nous obtenons les listes LISTMIN et DISTMIN et le trajet MININTERP.

Puis :

- la variance des distances DISTMAX est calculée (nous obtenons VMAX)
- la variance des distances DISTMIN est calculée (nous obtenons VMIN)
- Le nombre de DISTMAX comprises entre 0,15 seconde et 0,25 seconde est estimé (nous obtenons le pourcentage PMAX)
- Le nombre de DISTMIN comprises entre 0,15 seconde et 0,25 seconde est estimé (nous obtenons le pourcentage PMIN)

Une période de vibrato comprise entre 0,15 seconde et 0,25 seconde correspond à une fréquence du vibrato comprise entre 4 Hz et 6,7 Hz. Quand un vibrato est présent, la plupart des périodes du vibrato sont comprises dans ce petit intervalle, ce qui veut dire que les valeurs des deux variances sont petites et que les valeurs des deux pourcentages sont grandes. Nous pouvons le constater dans le tableau 13.1. Les résultats présentés ont été obtenus pour l'extrait de flûte **flute.sf**, pour lequel il y a quasi absence de vibrato, et pour l'extrait de voix chantée **voiceP.sf**, pour lequel il y a présence d'un fort vibrato. Ces deux sons sont ceux utilisés dans le chapitre 14 pour tester les performances des techniques décrites dans cette partie.

Finalement, si un vibrato a été détecté, les distances relatives à tous les instants d'échantillonnage entre le trajet MAXINTERP et le trajet MININTERP sont calculées.

La liste DISTFREQ est formée. Si la moyenne $\text{MFREQ}^{(G)}$ de ces distances relatives est grande, le vibrato est significatif et il est nécessaire de le supprimer.

- $\text{MFREQ}^{(G)}$: moyenne de (MAXINTERP - MININTERP)/(NOUVEAU TRAJET DE f_0)

	VMAX	VMIN	PMAX	PMIN	$\text{MFREQ}^{(G)}$
voix chantée	0,0010	0,0012	85 %	84 %	0,087
flûte	0,0090	0,0100	50 %	46 %	0,032

TAB. 13.1 – *Quelques résultats de la méthode pour la détection du vibrato d'un point de vue global*

La moyenne à chaque instant t_i du trajet MAXINTERP et du trajet MININTERP nous donne le NOUVEAU TRAJET DE f_0, c'est-à-dire le « trajet de f_0 sur lequel le vibrato est supprimé ».

L'amplitude A_{vib} du vibrato est obtenue en calculant la moyenne de MAXINTERP - MININTERP sur tout le son. La fréquence f_{vib} du vibrato est obtenue en faisant la moyenne des DISTMAX et des DISTMIN.

L'algorithme général est présenté sur la figure 15.6, page 139.

13.3.1.3 Effets du vibrato – point de vue local

Nous avons décrit ci-dessus une méthode pour décider de la présence d'un vibrato et estimer ses paramètres d'un point de vue global, c'est-à-dire en considérant tout le son. Cependant, nous sommes plutôt intéressé par la détection du vibrato d'un point de vue local, c'est-à-dire pour des portions du signal larges de quelques dixièmes de seconde. Nous nous alignons sur l'ordre de grandeur utilisé pour les autres méthodes de détection du vibrato : ainsi, nous prenons des portions larges de 0,35 seconde.

Nous comptons le nombre NB_M de maximums locaux par portion.

- Avec $f_{vib} \in \left[3 \quad \frac{2}{0{,}35} \simeq 5{,}7\right[$ Hz, pour une portion donnée, le nombre NB_M de maximums locaux présents est égal à 1 ou 2.
- Avec $f_{vib} \in \left[\frac{2}{0{,}35} \quad \frac{3}{0{,}35} \simeq 8{,}6\right[$ Hz, pour une portion donnée, le nombre NB_M de maximums locaux présents est égal à 2 ou 3.
- Avec $f_{vib} \in \left[\frac{3}{0{,}35} \quad 11\right]$ Hz, pour une portion donnée, le nombre NB_M de maximums locaux présents est égal à 3 ou 4.

Nous ne pouvons utiliser ni les variances VMAX et VMIN ni les pourcentages PMAX et PMIN calculés avec aussi peu de points. Soient : $NB_M(n)$ et $NB_m(n)$ respectivement le nombre de maximums et de minimums locaux pour la portion considérée, d'instant central n ; et $\text{MFREQ}^{(l)}(n)$, la moyenne des distances relatives pour cette même portion. L'idée est de construire une fonction d'observation basée sur ces trois mesures. Nous avons utilisé, comme fonction d'observation :

$$pb(n) = \exp\left(-\frac{(NB_{max}(n)-2)^2}{3}\right)\exp\left(-\frac{(NB_{min}(n)-2)^2}{3}\right)\text{MFREQ}(n)$$

$\exp\left(-\frac{(NB-2)^2}{3}\right)$ vaut respectivement pour $NB = [0\ \ 1\ \ 2\ \ 3\ \ 4\ \ 5]$:

[0,264 0,716 1,000 0,716 0,264 0,05]

13.3.2 Conclusion

La **décision** est prise ainsi :

- Si *pb* est grand, il y a un vibrato.

Les **estimations** sont faites ainsi :

- $A_{vib}(n) = \text{DISTFREQ}(n)$, ou $A_{vib}(n) = \text{MFREQ}^{(l)}(n)$.
- $f_{vib}(n)$ est estimée à partir des positions $[P_1\ \ldots\ P_{NB_M}]$ des NB_M maximums locaux et des positions $[p_1\ \ldots\ p_{NB_m}]$ des NB_m minimums locaux pour la portion centrée sur n :

$$\hat{f}_{vib}^{(M)}(n) = \frac{1}{NB_M - 1} \sum_{j=1}^{NB_M-1} \frac{f_e}{P_{j+1} - P_j} \text{ et } \hat{f}_{vib}^{(m)}(n) = \frac{1}{NB_m - 1} \sum_{j=1}^{NB_m-1} \frac{f_e}{p_{j+1} - p_j}$$

et, finalement : $\hat{f}_{vib}(n) = \frac{\hat{f}_{vib}^{(M)}(n) + \hat{f}_{vib}^{(m)}(n)}{2}$

13.4 Méthode basée sur le signal analytique

13.4.1 Introduction

Une troisième méthode pour détecter le vibrato et estimer ses paramètres une fois que le trajet de la fondamentale a été obtenu est donnée dans cette section. Cette méthode est basée sur la détermination de la fréquence instantanée par filtrage de HILBERT. Des variantes de cette méthode sont classiquement utilisées pour la détermination de la fondamentale d'un signal de parole et sa segmentation (segmentation en phones) simultanée : voir [Hes83]. Elle a été adaptée au cas du vibrato, le signal considéré ici étant le trajet de f_0. Dans la méthode originale, le signal est filtré en $\frac{1}{f^K}$, K étant entier. Ce filtrage assure que la fondamentale soit prédominante. Dans notre cas, ce filtre n'est pas nécessaire, et nous nous contentons d'un filtre passe-bande, avec : $f_1 = 4\ Hz$ et $f_2 = 9\ Hz$. Ainsi, nous sommes sûr que le continu (c'est-à-dire, ici, f_0^c) est éliminé (ce qui est nécessaire pour le calcul du signal analytique) et que les harmoniques du vibrato sont eux aussi éliminés. Le synoptique de la méthode est donné sur la figure 13.8. Le module des complexes $X(n)$ nous donne des indications sur l'amplitude du vibrato. Seuiller le module relatif $\frac{|X(n)|}{\hat{f}_0^c(n)}$ nous permet de déterminer les parties du signal où un vibrato est présent et celles où il n'y a pas de vibrato : nous posons des marques sur le trajet de f_0.

13.4.2 Filtrage passe-bande

La qualité de la réjection du continu (c'est-à-dire de f_0 « moyen », ou f_0^c) est primordiale pour le bon déroulement des traitements qui suivent (pour le filtrage de HILBERT principalement). De plus, l'ordre des filtres doit être petit. En effet, le signal est stationnaire sur de courts segments : il n'y a que quelques périodes de vibrato par note. Le filtre utilisé n'obéit pas tout à fait au gabarit idéal défini, mais ce défaut ne prête pas à conséquence d'une manière trop importante. Ceci est simplement ennuyeux pour la détermination du module $|X|$, c'est-à-dire de l'amplitude du vibrato, et pour le calcul de l'estimation $\hat{f}_0^c$ de f_0^c : dans ce cas, l'amplitude du vibrato étant sous-estimée (gain du filtre passe-bande inférieur à 1), il n'est pas complètement supprimé sur le trajet de f_0 (c'est-à-dire $\hat{f}_0^c \neq f_0^c$). Avec un ordre de 35, l'amplitude de la réponse en fréquence du filtre passe-bande est donnée sur la figure 13.9.

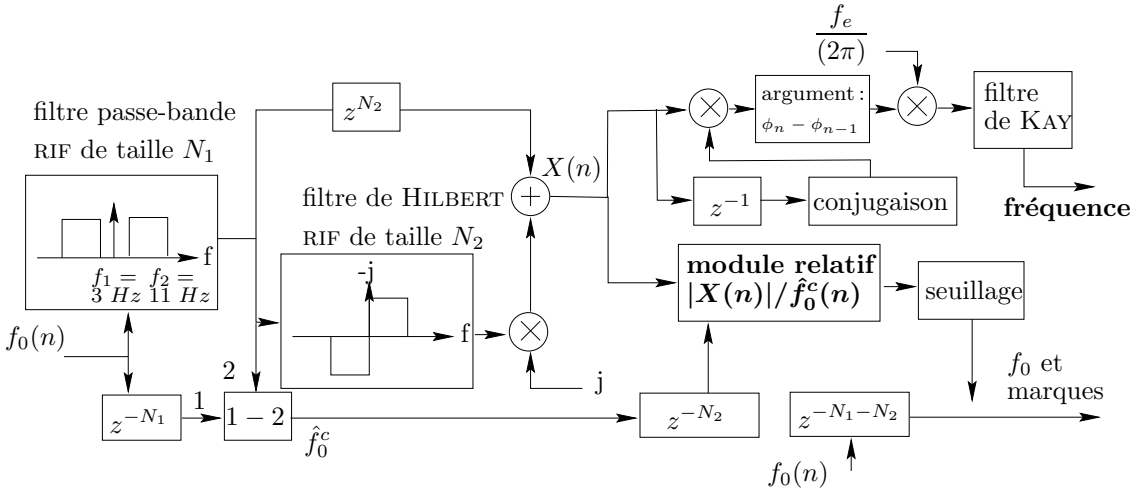


FIG. 13.8 – *Synoptique de la méthode de détection du vibrato basée sur le signal analytique*

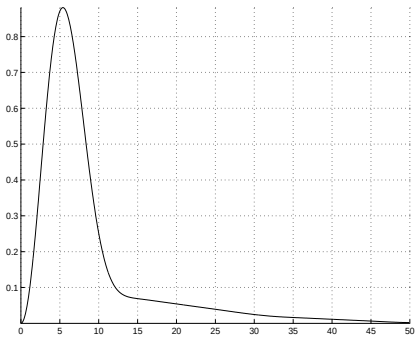


FIG. 13.9 – *Amplitude de la réponse en fréquence du filtre « passe-bande ». En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude de la réponse en fréquence*

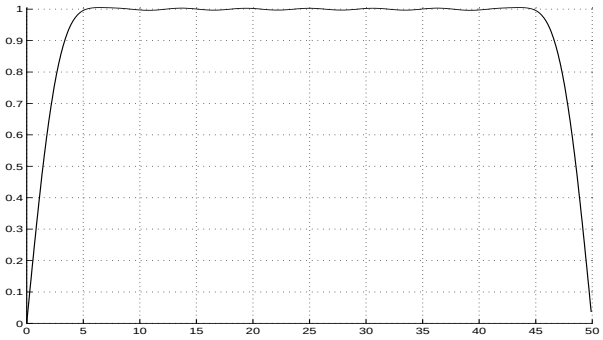


FIG. 13.10 – *Amplitude de la réponse en fréquence du filtre de* HILBERT. *En abscisse : la fréquence en Hz ; en ordonnée : l'amplitude de la réponse en fréquence*

13.4.3 Filtrage de HILBERT

Après le filtrage passe-bande décrit dans la section précédente, nous obtenons, si un vibrato sinusoïdal est présent, idéalement, une sinusoïde pure, de fréquence f_{vib} variant lentement et d'amplitude A_{vib} variant lentement. Donc, sur un court horizon, nous pouvons considérer que nous avons : $s_n = \cos\left(2\pi f_{vib}\frac{n}{f_e}\right)$. Le signal s_n est en tout cas à bande étroite. Le signal analytique associé à ce signal réel s'écrit :

$$C_n = s_n + js'_n = s_n + jH\left(s_n\right) = \cos\left(2\pi f_{vib}\frac{n}{f_e}\right) + j\sin\left(2\pi f_{vib}\frac{n}{f_e}\right) = \exp\left(2\pi j f_{vib}\frac{n}{f_e}\right)$$

où l'opérateur H correspond au filtrage de HILBERT.

En fait, si $\hat{S}$ et $\hat{C}$ sont respectivement les transformées de FOURIER de s et C, nous voulons :

$$\begin{aligned}\hat{C}(f) &= 2\hat{S}(f) \quad \text{pour } f > 0\\ &= 0 \quad \text{pour } f < 0\end{aligned}$$

Soit :

$$\hat{C}(f) = \left[1 + j(-j\text{signe}(f))\right]\hat{S}(f) = \hat{S}(f) + j\hat{H}(f)\hat{S}(f)$$

La réponse en fréquence $\hat{H}(f)$ d'un filtre de HILBERT est donc :

$$\hat{H}(f) = -j\text{signe}(f)$$

Ainsi, sa réponse impulsionnelle est :

$$\begin{aligned}q(k) &= \frac{1}{f_e}\int_{-f_e/2}^{f_e/2} -j\text{signe}(f)\exp\left(2j\pi\frac{k}{f_e}f\right)df\\ &= \frac{2}{\pi k}\sin^2\left(\frac{\pi k}{2}\right) \text{ (pour } k \neq 0 \text{ ; pour } k = 0 \text{ nous avons } q(0) = 0)\end{aligned}$$

Nous obtenons un filtre non causal et non réalisable physiquement. En tronquant la réponse impulsionnelle, nous arrivons à obtenir une approximation convenable dans une certaine bande de fréquences. Les basses fréquences et les hautes fréquences sont éliminées.

Dans notre cas, la taille du filtre utilisé est, comme pour le filtre passe-bande, 35. L'amplitude de la réponse en fréquence est donnée sur la figure 13.10. Les coefficients du filtre sont égaux à :

$$\begin{aligned}h(k) &= \frac{2}{\pi k}\text{ Hamming}(k) \quad \text{pour } k \in [-17\ -15\ \ldots\ 15\ 17]\\ &= 0 \quad \text{pour } k \in [-16\ -14\ \ldots\ 14\ 16]\end{aligned}$$

L'important est d'obtenir une réponse sensiblement linéaire et horizontale entre 3 et 11 Hz. Ceci est à peu près vérifié ici (tout du moins entre 4 et 11 Hz). La multiplication par la fenêtre de HAMMING permet d'obtenir cette réponse plate, mais réduit la bande-passante du filtre. Si nous utilisons la fenêtre RECTANGULAIRE, la réponse est plus large mais elle ondule dans la bande-passante. La solution serait peut-être en fait de d'abord sous-échantillonner le trajet de f_0. Cependant, si nous voulions garder une taille pour les fenêtres d'analyse de 0,35 seconde, nous serions obligé de diminuer le nombre de coefficients du filtre ! et, ainsi, il n'est pas évident que le sous-échantillonnage apporterait quelque amélioration.

13.4.4 Filtrage de KAY

La valeur de la fréquence instantanée $\hat{f}_{vib}(n) = \frac{f_e}{2\pi}(\phi_n - \phi_{n-1})$ est lissée sur un horizon de $N-1$ points. La première idée serait de calculer simplement la moyenne :

$$\hat{f}_{vib}(n) = \frac{1}{N-1}\sum_{i=0}^{N-2} \hat{f}_{vib}(n-i)$$

soit : $\hat{f}_{vib}(n) = \frac{f_e}{2\pi(N-1)}(\phi_n - \phi_{n-N+1})$. Ainsi, $N-2$ données ϕ ne sont pas utilisées, ce qui est en soi absurde. KAY (voir [Boa92] page 543, [Kay88]) a prouvé qu'en terme de variance de l'estimée de la fréquence, l'optimal est de prendre :

$$\hat{f}_{vib}(n) = \frac{f_e}{2\pi}\sum_{i=0}^{N-2} h_i\,(\phi_{n-i} - \phi_{n-i-1})$$

avec :

$$h_i = \frac{1{,}5N}{N^2-1}\left(1 - \left[\frac{i-(N/2-1)}{N/2}\right]^2\right) \text{ avec } i \in [0 \ldots N-2]$$

Nous avons choisi un N égal à 35, qui nous donne un lissage sur une fenêtre d'analyse large de 0,35 seconde. Cette taille est celle des portions de signal que nous avions utilisées pour les autres filtres et pour les autres méthodes de détermination des paramètres du vibrato à partir du trajet de f_0 (voir les sections 13.2 et 13.3).

13.4.5 Réjection du continu

La méthode présentée dans la section 13.3 nous permettait d'obtenir la composante continue f_0^c du trajet de f_0. Ainsi, nous pouvons appliquer l'algorithme décrit ci-dessus, mais sans le filtrage passe-bande, directement sur $f_0 - f_0^c$. Mais même si nous considérons que les amplitudes des harmoniques du vibrato de numéros d'ordre supérieurs sont très petites (voir la section 13.5.2), l'expérience nous a montré qu'un filtrage passe-bas est nécessaire. Il est cependant plus facile à réaliser, à taille de filtre constante, que le filtre passe-bande initial. L'amplitude de la réponse en fréquence du filtre utilisé est donné sur la figure 13.11. Les résultats de l'algorithme appliqué à $f_0 - f_0^c$ sont donnés, pour l'extrait de flûte et l'extrait de voix chantée (voir la section 14.1), sur les figures 13.12 et 13.13. Nous voyons que nous obtenons de bons résultats, similaires à ceux obtenus avec le filtrage passe-bande (ces résultats sont présentés sur les figures 14.17 et 14.18).

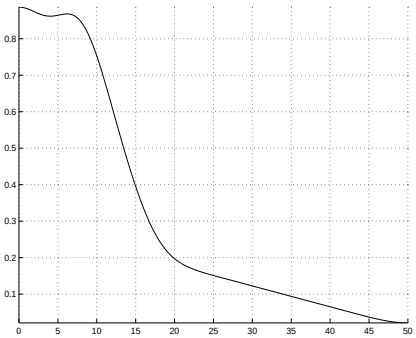


FIG. 13.11 – *Amplitude de la réponse en fréquence du filtre « passe-bas »*

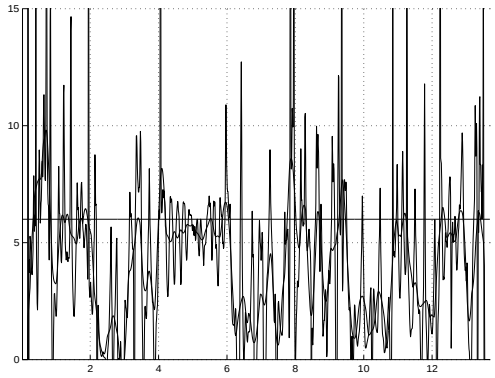


FIG. 13.12 – *Trajet non lissé et trajet lissé de la fréquence du vibrato pour la flûte. En abscisse : le temps ; en ordonnée : la fréquence en Hz. Une barre horizontale à 6 Hz est ajoutée*

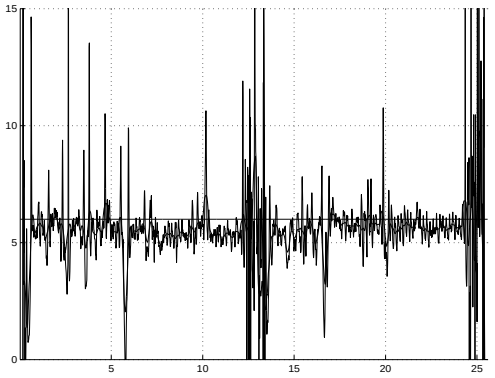


FIG. 13.13 – *Trajet non lissé et trajet lissé de la fréquence du vibrato pour la voix chantée. En abscisse : le temps ; en ordonnée : la fréquence en Hz. Une barre horizontale à 6 Hz est ajoutée*

13.4.6 Conclusion

La **décision** est prise ainsi :

- Si $\frac{|X(n)|}{\hat{f}_0^c(n)}$ est grand, il y a du vibrato.

Les **estimations** des paramètres du vibrato sont faites ainsi :

- $A_{vib}(n) \simeq |X(n)|$
- $f_{vib}(n) =$ fréquence

13.5 Conclusion

Les trois méthodes décrites dans ce chapitre pourront nous servir pour le cas du trémolo. Il s'agit d'une perspective.

13.5.1 Une première remarque : prise en compte du vibrato présent sur les harmoniques du signal de numéros d'ordre supérieurs

Le vibrato bien sûr affecte tous les harmoniques du signal, et il les affecte de la même façon qu'il affecte la fréquence fondamentale : c'est-à-dire que, pour une note, les vibratos présents sur les harmoniques du signal sont tous en phase, que la fréquence du vibrato est pour tous les harmoniques du signal f_{vib} et que l'amplitude du vibrato (nous faisons ici l'hypothèse qu'il est modélisé par une sinusoïde pure) présent sur l'harmonique de numéro d'ordre l (donc dont la fréquence est lf_0) du signal est égale à lA_{vib}. Ainsi, il est possible d'utiliser les informations disponibles sur les harmoniques de numéros d'ordre supérieurs à 1 pour améliorer la robustesse des algorithmes de détection du vibrato que nous avons décrits dans ce chapitre et celle de l'algorithme de suppression du vibrato sur le trajet de f_0 que nous allons décrire dans le chapitre 15, ne serait-ce qu'en travaillant avec le signal :

$$f_{moy}(i) = \frac{1}{L_{pris}} \sum_{l=1}^{L_{pris}} \frac{f_l(i)}{l}$$

plutôt qu'avec le signal $f_0(i)$ [1]. Nous avons : L_{pris} le nombre d'harmoniques pris en compte et $f_l(i)$ le trajet du $l^{ème}$ harmonique. Il s'agit d'une perspective.

1. Avec toujours $f_1 = f_0$: voir la note de la page 14.

13.5.2 Une seconde remarque : détection des harmoniques du vibrato sur des signaux réels

Il s'agit ici d'extraire les harmoniques du vibrato pour un signal réel. Ce signal réel est **voiceP.sf**. Nous considérons le mi_4 ($f_0 = 659{,}26\ Hz$) final, qui est tenu et pour lequel un vibrato important est présent. Nous n'utilisons pas les résultats obtenus par la modélisation AR et la prédiction (voir la section 13.2). Nous travaillons avec des fenêtres d'analyse larges de 0,9 seconde, c'est-à-dire avec des portions plutôt plus larges que les portions habituelles : la note étant tenue, nous pouvons nous le permettre. Nous pouvons alors calculer la FFT sur plus de points utiles : ainsi, la résolution de la transformée de FOURIER est augmentée. Nous donnons sur les figures 13.14 et 13.15 respectivement le signal sur lequel nous travaillons et les résultats obtenus, en nous limitant à cinq harmoniques du vibrato.

Nous constatons que les harmoniques que nous obtenons suivent assez bien les trajets qu'ils devraient suivre théoriquement, c'est-à-dire kf_{vib}, avec $k = 2 \ldots 5$, et avec f_{vib} le premier harmonique du vibrato obtenu : ce sont les courbes en pointillés de la figure 13.15. Nous constatons que les amplitudes des harmoniques du vibrato de numéros d'ordre supérieurs à 1 sont faibles, relativement à celle du premier, qui vaut, en moyenne, 33,7. En effet, les autres valent, en moyenne, respectivement 2,34, 1,24, 0,95 et 0,49.

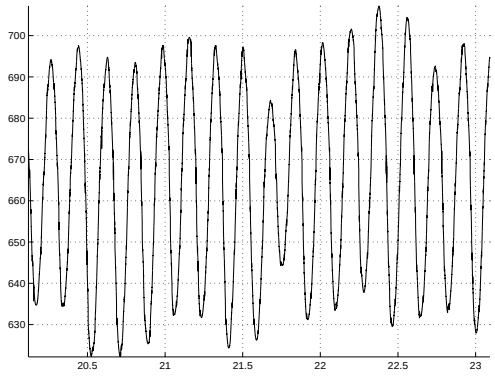


FIG. 13.14 – *Trajet de la fréquence fondamentale f_0 pour une note chantée : il s'agit d'une partie du* mi_4 *de* **voiceP.sf**. *Un vibrato important est présent. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

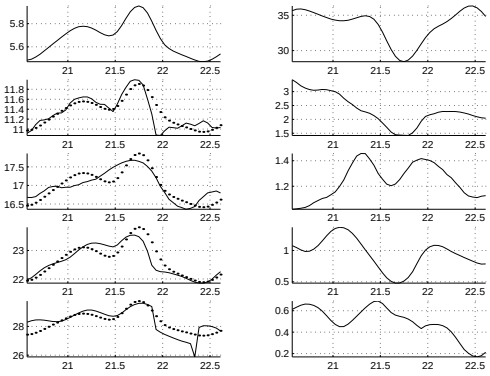


FIG. 13.15 – *À droite, les fréquences des harmoniques du vibrato ; en abscisse : le temps en seconde, en ordonnée : la fréquence en Hz. À gauche, les amplitudes des harmoniques du vibrato ; en abscisse : le temps en seconde, en ordonnée : l'amplitude. De haut en bas : du premier harmonique du vibrato au cinquième*

13.5.3 Perspective : découpage en sous-bandes

Nous considérons que la fréquence fondamentale du vibrato est comprise dans la bande [3 Hz − 11 Hz]. Nous découpons cette bande en deux sous-bandes. Nous souhaitons que, quelque soit la fréquence fondamentale du vibrato, elle se retrouve seule dans l'une des deux sous-bandes, et que ses harmoniques se retrouvent soit dans l'autre sous-bande, soit au-delà de 11 Hz. Aussi, le découpage doit être fait de la façon suivante : première sous-bande [3 Hz − 5,75 Hz], et seconde sous-bande [5,75 Hz − 11 Hz].

Ainsi, idéalement, pour les fréquences fondamentales du vibrato comprises dans la bande [3 Hz − 3,66 Hz], nous retrouvons la fréquence fondamentale dans la première sous-bande et deux harmoniques dans la seconde (car $3{,}66 \times 3 = 11$) ; pour celles comprises dans la bande [3,66 Hz − 5,5 Hz], nous retrouvons la fréquence fondamentale dans la première sous-bande et un harmonique dans la seconde ; pour celles comprises dans la bande [5,5 Hz − 5,75 Hz], nous retrouvons la fréquence fondamentale dans la première sous-bande et rien dans la seconde ; pour celles comprises dans la bande [5,75 Hz − 11 Hz], nous retrouvons la fréquence fondamentale dans

la seconde sous-bande et rien de plus. Nous considérons que l'énergie du signal se concentre dans l'une de ces deux sous-bandes, même dans les cas où il y a un ou deux harmoniques du vibrato dans l'autre puisque leur amplitude décroît rapidement (voir la section 13.5.2) avec leur numéro d'ordre. Il suffit alors de calculer les spectres définis dans la section 13.2.1 sur la sous-bande où l'énergie est la plus grande.

Cette technique a pour but d'améliorer l'extraction de la fréquence fondamentale du vibrato. À la rigueur, elle pourrait nous permettre d'extraire les harmoniques du vibrato de numéros d'ordre supérieurs (cependant, quand deux harmoniques du vibrato sont présents ensemble dans une sous-bande, la méthode basée sur le signal analytique ne peut pas être utilisée). Elle n'a pas été implémentée, ni sous MATLAB ni en C. Il s'agit de perspectives.

Chapitre 14

Performances comparées des méthodes de détection du vibrato et d'estimation de ses paramètres

14.1 Les sons réels considérés

Les deux signaux réels utilisés sont l'extrait de flûte **flute.sf**, pour lequel le vibrato est très petit, ou inexistant, selon la note ; et l'extrait de voix chantée **voiceP.sf**, pour lequel le vibrato est très grand, et pour lequel nous avons la présence d'un trémolo qui perturbe les algorithmes décrits dans les deux chapitres précédents. Le trajet de f_0 pour la flûte et pour la voix chantée sont redonnés, respectivement sur les figures 14.1 et 14.2. Les résultats présentés dans ce chapitre ont fait l'objet d'une communication : [RRD+99].

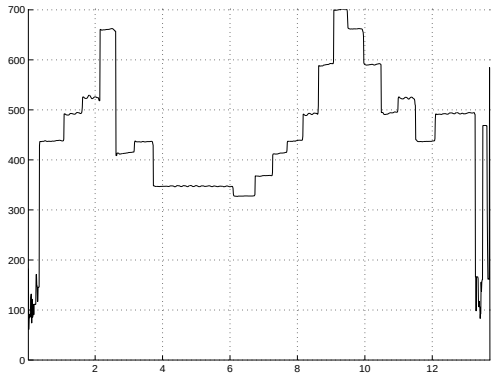


FIG. 14.1 – *Trajet de f_0 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

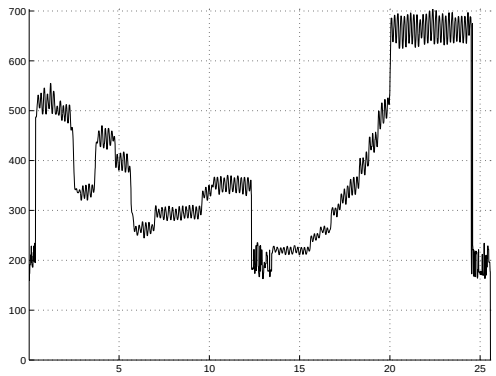


FIG. 14.2 – *Trajet de f_0 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

14.2 Méthode 2 : « distorsion des enveloppes spectrales »

Pour la **détection** du vibrato, l'évolution temporelle de α et l'évolution du « flux » sont étudiées. Voir les figures 14.3, 14.4, 14.5 et 14.6 pour les deux sons considérés.

Cette méthode ne nous permet pas d'estimer les paramètres du vibrato.

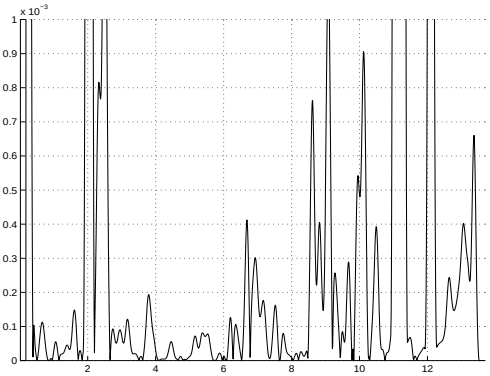


FIG. 14.3 – **Détection** *du vibrato. Résultats de la méthode 2 pour la flûte. En abscisse : le temps en seconde ; en ordonnée :* α

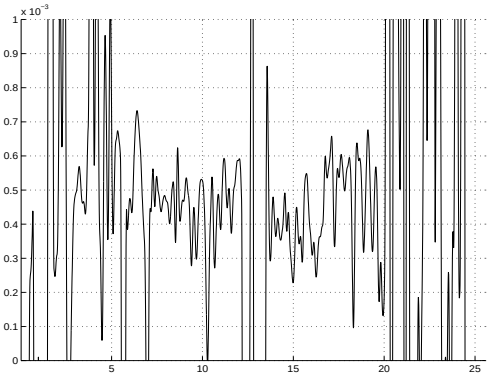


FIG. 14.4 – **Détection** *du vibrato. Résultats de la méthode 2 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée :* α

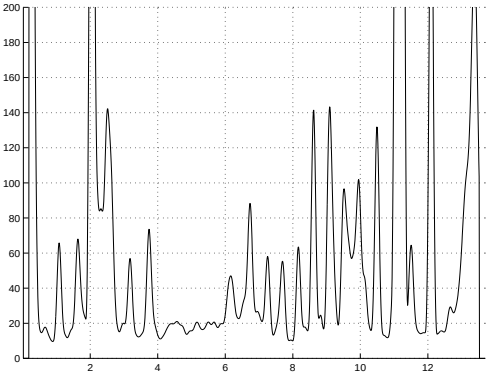


FIG. 14.5 – **Détection** *du vibrato. Résultats de la méthode 2 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : le « flux »*

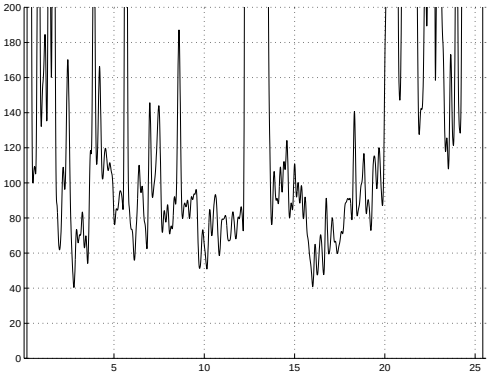


FIG. 14.6 – **Détection** *du vibrato. Résultats de la méthode 2 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : le « flux »*

14.3 Méthode 4 : « prédiction AR »

Pour la **détection**, l'évolution temporelle de R est étudiée. Voir les figures 14.7 et 14.8 pour les deux sons considérés.

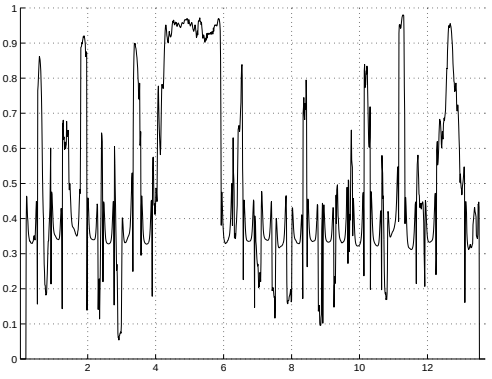


FIG. 14.7 – **Détection** *du vibrato. Résultats de la méthode 4 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : R*

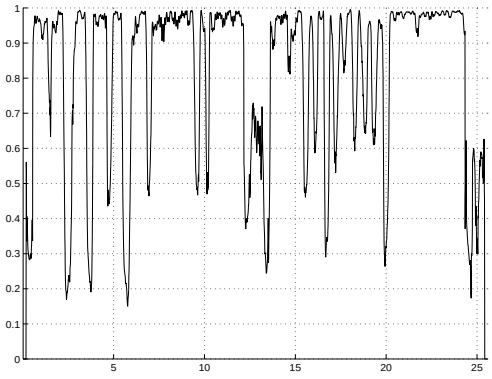


FIG. 14.8 – **Détection** *du vibrato. Résultats de la méthode 4 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : R*

Pour l'**estimation** des paramètres du vibrato, voir les figures 14.9 et 14.10, où nous donnons les trajets de f_{vib} trouvés pour les deux sons considérés (les droites horizontales représentent la fréquence du vibrato mesurée « à l'œil » : barre horizontale à 6 Hz). Nous donnons les résultats pour la flûte bien que ce son ne soit quasi pas modulé en fréquence.

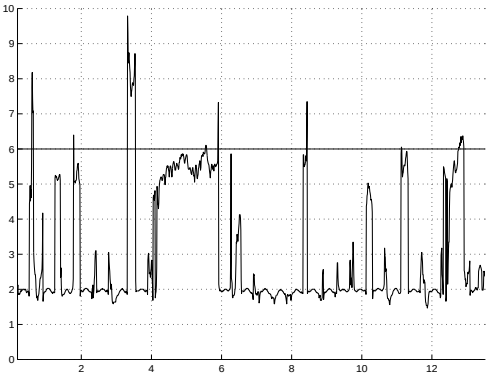


FIG. 14.9 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 4 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

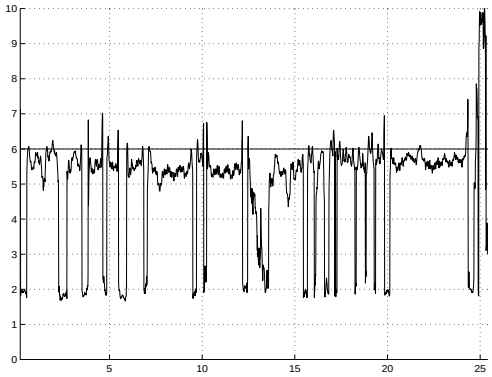


FIG. 14.10 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 4 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

14.4 Méthode 5 : « détection des minimums et maximums »

Pour la **détection**, voir le tableau 13.1 page 122. Mais voir surtout les figures 14.11 et 14.12 pour les deux sons considérés.

Pour l'**estimation** des paramètres du vibrato, voir les figures 14.13 et 14.14, où nous donnons les trajets de f_{vib} trouvés pour les deux sons considérés (les droites horizontales représentent la fréquence du vibrato mesurée « à l'œil » : barre horizontale à 6 Hz). Nous donnons les résultats pour la flûte bien que ce son ne soit quasi pas modulé en fréquence.

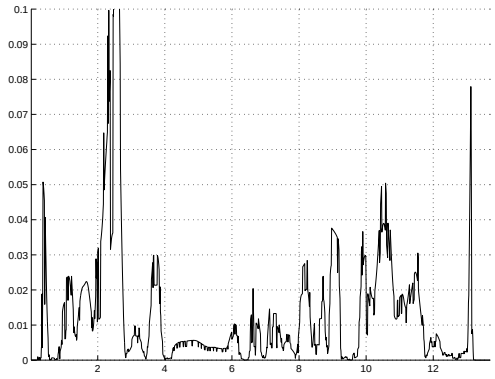


FIG. 14.11 – **Détection** *du vibrato. Résultats de la méthode 5 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : pb*

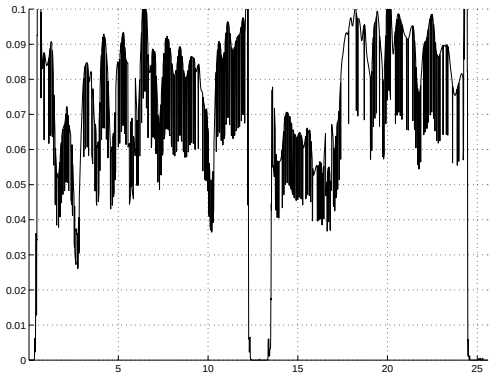


FIG. 14.12 – **Détection** *du vibrato. Résultats de la méthode 5 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : pb*

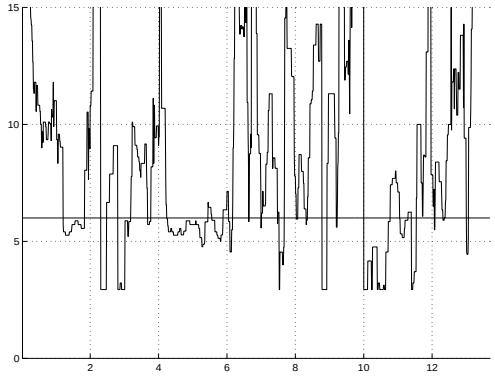


FIG. 14.13 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 5 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

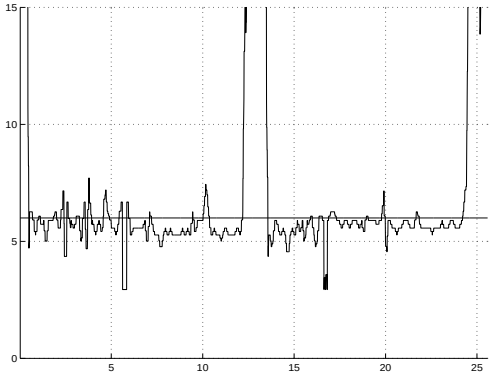


FIG. 14.14 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 5 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

14.5 Méthode 6 : Méthode « signal analytique »

Pour la **détection**, l'évolution temporelle de $mod = \frac{|X|}{\hat{f}_0^c}$ est étudiée. Voir les figures 14.15 et 14.16 pour les deux sons considérés.

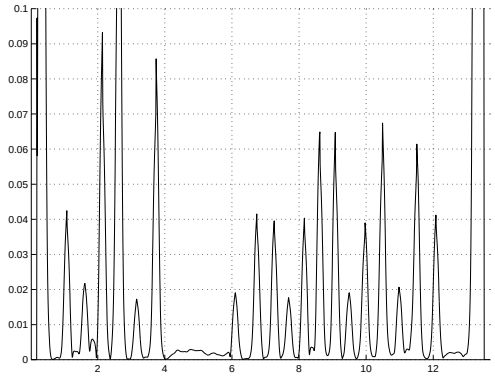


FIG. 14.15 – **Détection** *du vibrato. Résultats de la méthode 6 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : mod*

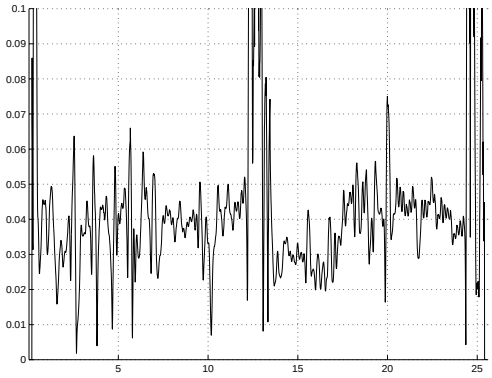


FIG. 14.16 – **Détection** *du vibrato. Résultats de la méthode 6 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : mod*

Pour l'**estimation** des paramètres du vibrato, voir les figures 14.17 et 14.18, où nous donnons les trajets de f_{vib} trouvés pour les deux sons considérés (les droites horizontales représentent la fréquence du vibrato mesurée « à l'œil » : barre horizontale à 6 Hz). Nous donnons les résultats pour la flûte bien que ce son ne soit quasi pas modulé en fréquence.

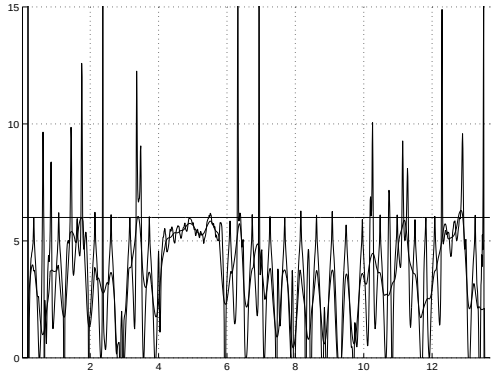


FIG. 14.17 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 6 pour la flûte. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

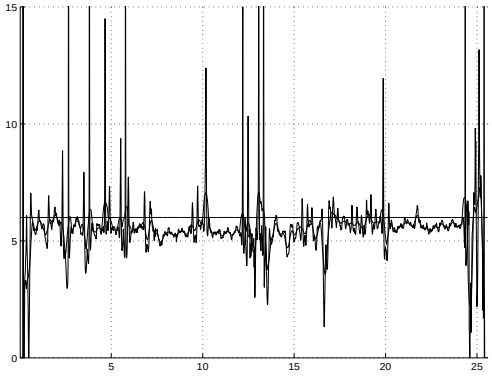


FIG. 14.18 – **Estimation** *de la fréquence du vibrato. Résultats de la méthode 6 pour la voix chantée. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

Pour l'extrait de voix chantée, nous obtenons les résultats intermédiaires (trajets de la fondamentale filtrée et du module) présentés sur les figures 14.19 et 14.20.

Les mêmes courbes (voir 14.21 et 14.22) sont données dans le cas de la flûte, où il n'y a du vibrato, du reste très petit, que sur peu de notes. Nous voyons que le trajet du vibrato est beaucoup plus chahuté pour la flûte (ceci, sauf pour la note de plus de 2 secondes à $f_0^c = 347\ Hz$: le vibrato quoique petit est visible « à l'œil », et il est aussi parfaitement estimé avec cet algorithme) que pour la voix chantée : nous obtenons du bruit. De la même façon, le module et le module relatif sont plus petits pour la flûte que pour la voix chantée.

Pour l'exemple de la flûte, nous constatons sur la figure 14.22 que le trajet du module peut être utilisé pour détecter les transitions, lors de la *segmentation en zones stables*. Remarquons que

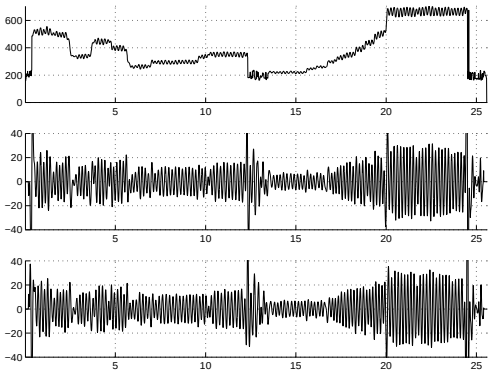


FIG. 14.19 – *Trajets de f_0, de f_0 filtrée passe-bande et de f_0 filtrée passe-bande puis filtrée par* HILBERT. *Extrait de voix chantée. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

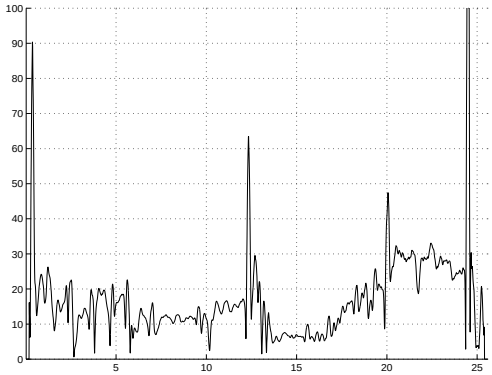


FIG. 14.20 – *Trajet du module $|X|$. Extrait de voix chantée. En abscisse : le temps en seconde*

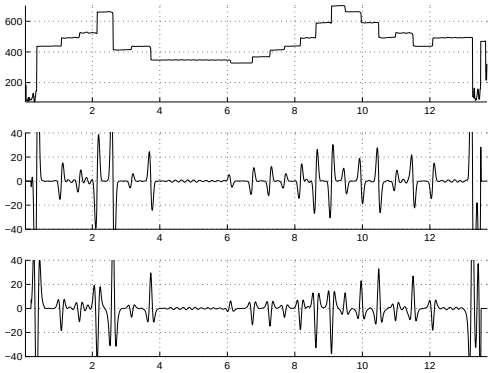


FIG. 14.21 – *Trajets de f_0, de f_0 filtrée passe-bande et de f_0 filtrée passe-bande puis filtrée par* HILBERT. *Extrait de flûte. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

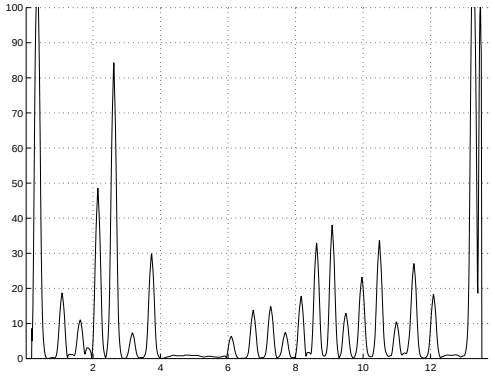


FIG. 14.22 – *Trajet du module $|X|$. Extrait de flûte. En abscisse : le temps en seconde*

le module, contrairement à f_{vib}, n'est pas lissé sur une fenêtre d'analyse large de 0,35 seconde (filtrage de KAY). La fonction d'observation pour la *segmentation en zones stables* (voir le chapitre 2, page 11) est alors $|X|$.

Chapitre 15

Méthode de suppression du vibrato sur le trajet de f_0 (pour l'aide à la *segmentation en zones stables*)

15.1 Introduction

Nous montrons sur la figure 15.1 pourquoi au cas où le vibrato est significatif les fonctions d'observation basées sur le trajet de f_0 échouent quand il s'agit de *segmenter en zones stables*, et, donc, pourquoi il est nécessaire de le supprimer sur ce trajet. Le signal utilisé ici est simulé. Pour le modèle, voir la section 11.2.3.1 : nous avons utilisé $a = 1$, $b = 0{,}1$, $f_0^{(1)} = 246{,}945\ Hz$ (si_2), $f_0^{(2)} = 261{,}62\ Hz$ (do_3) et $\varphi_{vib} = 0{,}1234\ rad$.

À gauche, le trajet de f_0 original est donné ainsi que la valeur absolue de sa dérivée. Nous pouvons voir que le pic de transition est mêlé aux pics dus au vibrato.

À droite, le « trajet de f_0 une fois le vibrato supprimé » (obtenu avec la méthode présentée dans la section 13.3) et la valeur absolue de sa dérivée sont donnés. La fonction d'observation présente un pic significatif au moment de la transition. Cependant la largeur de cette dernière a augmenté : nous n'avons pas un pic « aussi fin et grand » que possible au moment de la transition.

La méthode présentée dans la section 13.3 est la seule des six méthodes présentés dans les chapitres 12 et 13 qui permette de supprimer le vibrato sur le trajet de f_0.

15.2 Performances de la méthode pour un signal réel

Nous avons analysé ci-dessous l'extrait de voix chantée **voiceP.sf**, pour lequel un vibrato d'amplitude très importante est présent. Nous ne considérons que la seconde moitié du son (après que la chanteuse eut repris son souffle). Nous avons une succession de dix notes de plus en plus hautes, la chanteuse se servant du vibrato pour passer de l'une à l'autre. Les neuf transitions ont lieu dans un intervalle de temps inférieur à cinq secondes.

Nous donnons sur les figures 15.2, 15.3, 15.4 et 15.5 respectivement le trajet de f_0 original ; le trajet de f_0 sur lequel il n'y a plus de vibrato ; le trajet de la « valeur absolue de la dérivée de f_0 original » ; et le trajet de la « valeur absolue de la dérivée de f_0 sur lequel il n'y a plus de vibrato ». Pour les quatre figures, les notes chantées et les marques de segmentation posées à la main sont ajoutées.

Sur la figure 15.5, est aussi donné la position du seuil (seuil posé à la main). Si nous posons une marque à chaque maximum local plus grand que le seuil, nous constatons que les instants de segmentation sont tous détectés. Ainsi, la méthode est efficace.

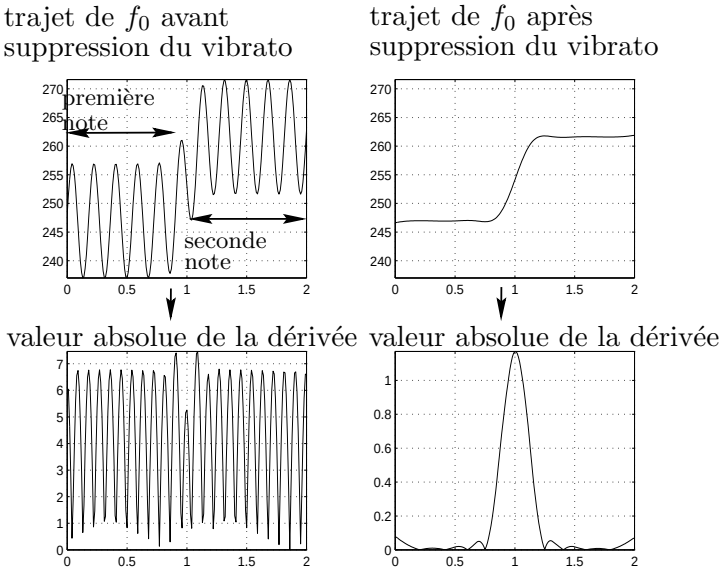


FIG. 15.1 – *Pourquoi nous supprimons le vibrato sur le trajet de f_0*

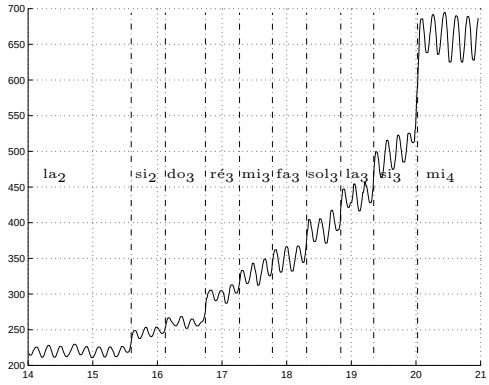


FIG. 15.2 – *Trajet original de f_0, notes chantées et marques de segmentation posées à la main. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

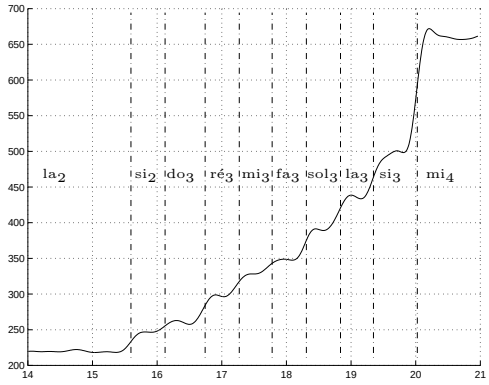


FIG. 15.3 – *Trajet de f_0 une fois le vibrato supprimé, notes chantées et marques de segmentation posées à la main. En abscisse : le temps en seconde ; en ordonnée : la fréquence en Hz*

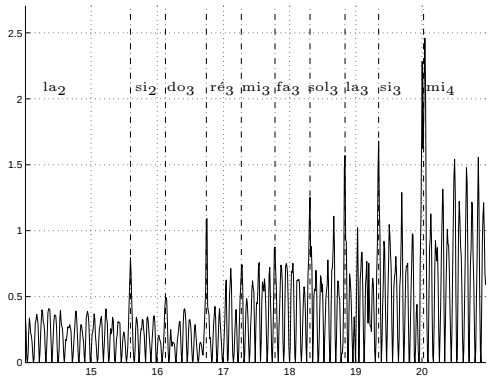


FIG. 15.4 – *Valeur absolue de la dérivée du trajet de f_0 original, notes chantées et marques de segmentation posées à la main. En abscisse : le temps en seconde*

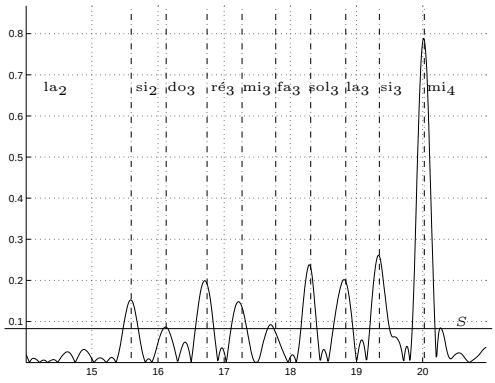


FIG. 15.5 – *Valeur absolue de la dérivée du trajet de f_0 une fois le vibrato supprimé, notes chantées et marques de segmentation posées à la main. En abscisse : le temps en seconde*

Le principal défaut de cette méthode est qu'elle perd de l'information disponible : ainsi, pour calculer le trajet de f_0 sur lequel il n'y a plus de vibrato, nous n'utilisons qu'environ dix pour cent des points du trajet de f_0 original. En effet, la fréquence d'échantillonnage de f_0 est de 100 Hz et la fréquence du vibrato d'environ 5 Hz, donc 1 échantillon sur 20 correspond à un maximum local, et 1 sur 20 à un minimum local. À cause de cela, les durées des transitions sont allongées. Et la dérivée est lissée en conséquence. Ceci se traduit parfois par un léger décalage des maximums locaux de la fonction d'observation « valeur absolue de la dérivée de f_0 une fois le vibrato supprimé », décalage visible sur la figure 15.5. Malgré ce décalage, les résultats sont très intéressants. D'autres exemples nous ont montré que les grands sauts de f_0 ne sont pas catastrophiquement détériorés par cette méthode.

15.3 Conclusion

L'algorithme général de détection du vibrato, d'estimation de ses paramètres et de sa suppression sur le trajet de f_0 à partir des minimums locaux et des maximums locaux du trajet de f_0 est schématisé sur la figure 15.6 (voir la section 13.3). Rappelons que cette méthode a été intégrée dans le programme *segmentation*.

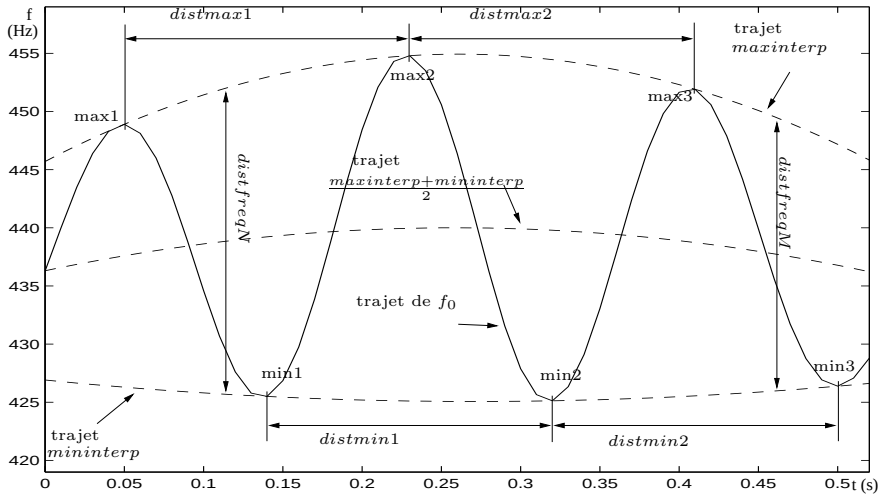


FIG. 15.6 – *Détection et suppression du vibrato sur le trajet de f_0*

Chapitre 16

La fusion de données dans le cas du vibrato – Introduction

En ce qui concerne le problème du vibrato, les deux types de fusion de données (voir la section 2.6.1 page 34) sont à considérer.

- D'abord, nous fusionnons des fonctions de décision : elles nous disent s'il y a du vibrato ou s'il n'y a pas de vibrato. Nous sommes dans le cas de la fusion de données hétérogènes (voir la section 2.6.1).

  Pour savoir s'il est nécessaire ou non de supprimer le vibrato du trajet de f_0, il faut une décision ferme : ou bien « il y a du vibrato », ou bien « il n'y en a pas ». Nous ne pouvons pas nous contenter d'une décision floue du type « le vibrato est plutôt petit » ou « il y a plutôt un vibrato important ». La règle de fusion de la majorité est utilisée. Les résultats sont donnés dans le tableau 16.1. Les décisions sont prises d'un point de vue global, c'est-à-dire sur tout le son. fd_1 correspond à la fonction de décision obtenue à partir du trajet de α ; fd_2 à celle obtenue à partir du trajet du « flux » ; fd_3 à celle obtenue à partir du trajet de R ; fd_4 à celle obtenue à partir du trajet de pb ; et fd_5 à celle obtenue à partir du trajet de mod.

  Les seuils pour les cinq fonctions d'observation ont été fixés respectivement à 0,0002, 67, 0,5, 0,02 et 0,025. Les fonctions de décisions sont présentées sur les figures 16.1 et 16.2. Nous considérons que l'extrait de flûte n'est pas modulé et que l'extrait de voix chantée l'est. Nous constatons que la fusion est efficace : pour l'extrait de flûte, la probabilité de présence d'un vibrato après la fusion de données est plus petite que la plus petite des probabilités de présence d'un vibrato obtenues pour chaque fonction d'observation ; et pour l'extrait de voix chantée, la probabilité de présence d'un vibrato après la fusion de données est plus grande que la plus grande des probabilités de présence d'un vibrato obtenues pour chaque fonction d'observation.

  Cependant, nous avons vu qu'un très petit vibrato est présent pour l'extrait de flûte sur la plus longue note. Seule la méthode 4 pourrait nous permettre de le détecter. Et nous avons vu que la chanteuse, pour l'extrait de voix chantée, reprend son souffle pendant quelques dixièmes de seconde vers le milieu de l'extrait : ainsi, un vibrato n'est pas toujours présent pour cet extrait. La fonction d'observation « flux » (méthode 2) et la méthode 6, telles qu'elles sont utilisées, ne permettent pas de mettre cela en évidence.
- Ensuite, nous fusionnons des fonctions d'estimation : les valeurs des paramètres du vibrato. Nous sommes dans le cas de la fusion de données homogènes (voir la section 2.6.1).

	fd_1	fd_2	fd_3	fd_4	fd_5	**fusion**
Probabilité de présence de vibrato pour la flûte	0,284	0,238	0,324	0,207	0,225	**0,188**
Probabilité de présence de vibrato pour la voix	0,751	0,88	0,839	0,892	0,887	**0,923**

TAB. 16.1 – *Détection du vibrato – fusion de données*

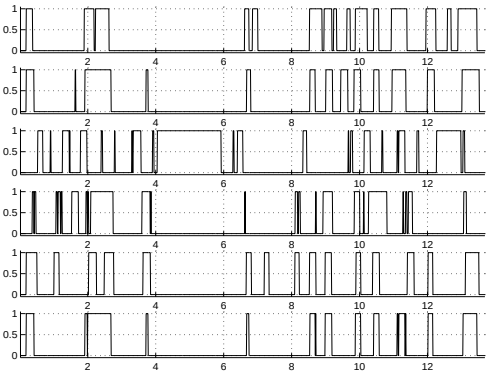


FIG. 16.1 – *Fonctions de décision pour la flûte. De haut en bas : méthode 2, pente ; méthode 2, flux ; méthode 4 ; méthode 5 ; méthode 6 ; et résultats de la fusion. En abscisse : le temps ; en ordonnée : il y a du vibrato (1) ou il n'y a pas de vibrato (0)*

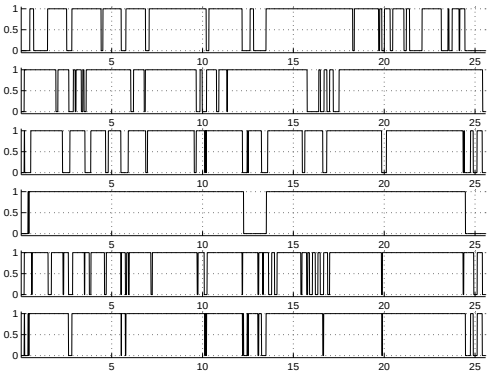


FIG. 16.2 – *Fonctions de décision pour la voix chantée. De haut en bas : méthode 2, pente ; méthode 2, flux ; méthode 4 ; méthode 5 ; méthode 6 ; et résultats de la fusion. En abscisse : le temps ; en ordonnée : il y a du vibrato (1) ou il n'y a pas de vibrato (0)*

Nous donnons la procédure globale, en nous limitant aux quatre méthodes qui ont été testées efficacement sur des signaux réels, sur la figure 16.3.

FIG. 16.3 – *La fusion de données dans le cas du vibrato*

Chapitre 17

Conclusion de la troisième partie

17.1 En ce qui concerne le vibrato

L'analyse du problème du vibrato se décompose en trois étapes : **I.**, la détection du vibrato ; **II.**, l'estimation des paramètres du vibrato ; **III.**, la suppression du vibrato sur le trajet de f_0.

I. En ce qui concerne la détection du vibrato, six méthodes ont été étudiées. Trois sont basées sur l'analyse directe du signal :

1. Méthode basée sur la « modélisation du spectre ».
   Cette méthode, inexistante dans la littérature, n'a pas été testée sur des signaux réels. Telle qu'elle a été présentée dans cet exposé, elle souffre de limitations qui ne le permettent pas :
   - Nous supposons que f_0^c ne varie pas sur la fenêtre d'analyse
   - Nous supposons que les amplitudes des harmoniques ne varient pas sur la fenêtre d'analyse
   - Nous supposons que les paramètres du vibrato ne varient pas sur la fenêtre d'analyse

   Des modèles du signal plus complexes devront être mis en place. Il s'agit de perspectives.
2. Méthode basée sur la « détection de la distorsion des enveloppes spectrales ».
   Cette méthode, inexistante dans la littérature, a été utilisée avec succès sur des signaux réels.
3. Méthode basée sur l'approche de Laroche.
   Cette méthode n'a pas été testée sur des signaux réels. Elle a été légèrement améliorée.

Trois autres méthodes sont basées sur l'analyse du trajet de f_0 :

4. Méthode basée sur la « prédiction ar ».
   Les radaristes utilisent la prédiction ar pour améliorer la résolution fréquentielle obtenue après transformée de Fourier (voir [Bro95]) : les radaristes travaillent eux aussi avec des signaux stationnaires sur très peu de périodes. Elle a été testée avec succès sur des signaux réels.
5. Méthode basée sur la « détection des minimums et des maximums ».
   Cette méthode, malgré sa simplicité, a été testée avec succès sur des signaux réels. Elle est implémentée dans le programme *segmentation*.
6. Méthode basée sur le « signal analytique ».
   Cette méthode a été adaptée d'une méthode d'estimation de f_0 utilisée en traitement de la parole. Elle a été testée avec succès sur des signaux réels.

II. En ce qui concerne l'estimation des paramètres du vibrato, les méthodes présentées ci-dessus nous donnent :

- Méthode 1 : les estimations de f_{vib}, A_{vib} et φ_{vib} sont directement obtenues

- Méthode 2 : cette méthode ne peut pas être utilisée pour estimer f_{vib}, A_{vib} et φ_{vib}
- Méthode 3 : les estimations de f_{vib}, A_{vib} et φ_{vib} ne sont pas directes
- Méthode 4 : les estimations de $f_{vib}(n)$ et $A_{vib}(n)$ sont obtenues
- Méthode 5 : les estimations de $f_{vib}(n)$ et $A_{vib}(n)$ sont obtenues
- Méthode 6 : les estimations de $f_{vib}(n)$ et $A_{vib}(n)$ sont obtenues

III. La suppression du vibrato sur le trajet de f_0 n'est possible qu'avec la cinquième méthode.

L'intérêt d'avoir implémenté et testé plusieurs méthodes, d'origines très diverses, pour détecter le vibrato, puis pour déterminer ses paramètres et le traiter, réside en ce que nous pouvons confronter leurs résultats, qui, grâce à cette diversité, sont plutôt décorrélés. Il s'agit donc ici encore d'un problème de fusion de données, le but étant toujours d'améliorer la robustesse du système. Aussi bien lors de la détection du vibrato que lors de l'estimation de ses paramètres, des techniques de fusion des données fournies par les différentes techniques mises en places sont utilisées.

De nombreuses autres méthodes non dédiées au problème du vibrato (de sa détection, de l'estimation de ses paramètres et de sa suppression du trajet de f_0) peuvent être adaptées pour lui, notamment parmi celles qui concernent la détection et l'estimation de f_0 : voir [Hes83] à ce sujet.

La détection, l'estimation des paramètres et la suppression sur le trajet de f_0 du vibrato sont intéressantes en elles-mêmes, pour les applications qu'elles ont en synthèse par exemple ; mais nous avons vu aussi qu'elles sont nécessaires pour obtenir une *segmentation en zones stables* efficace.

Aussi, il est nécessaire de détecter le vibrato, d'estimer ses paramètres (fréquence, amplitude, phase), et de l'extraire du trajet de f_0 :

- Pour modifier le son. Nous considérons l'extrait de voix chantée **voiceP.sf**.
  Le son peut être resynthétisé avec un logiciel de l'IRCAM, appelé DIPHONE [1]. Pour ce faire, nous avons besoin :
  - Du trajet de la fondamentale.
  - Du trajet de l'énergie : ici, nous avons pris la même énergie tout le long du signal, sauf au moment où la chanteuse reprend son souffle où nous l'avons posée à 0. Il s'agit là de la première transformation effectuée.
  - Des phonèmes chantés : ici, la chanteuse chante tout du long de l'extrait un « a ».
  - De la voix du chanteur : est-ce un chanteur ou une chanteuse ? un ténor ou un soprano ou etc. ? Il s'agit de placer les formants où il faut.

  Nous avons resynthétisé les deux sons :
  - Dans un premier temps, avec le trajet de f_0 original.
  - Dans un second temps, avec le trajet de f_0 sur lequel il n'y a plus de vibrato (voir la figure 15.3 page 138).

  Ces sons peuvent être trouvés à cette adresse :

  `http://www.ircam.fr/equipes/analyse-synthese/rossigno/dafx99/article6.html`

  Faisons simplement remarquer que pour les deux sons resynthétisés, en ce qui concerne la seconde partie du son, nous entendons bien les dix notes présentes sur le son original.
- Pour caractériser les sons (MPEG-7) : étiquette *il y a du vibrato* ou *il n'y a pas de vibrato*, ou « il y a un vibrato plutôt petit » ou « il y a un vibrato plutôt important »
- Pour aider à améliorer la *segmentation en notes (zones stables)* des extraits de voix chantée

Les techniques utilisées pour la détection, l'estimation des paramètres et la suppression sur le trajet de f_0 du vibrato sont robustes. Il reste à extraire les paramètres du vibrato pour des modèles du vibrato plus complexes.

1. Voir `http://www.ircam.fr/produits/logiciels/log-forum/` pour une documentation en ligne de DIPHONE.

17.2 En ce qui concerne la *segmentation en caractéristiques*

Nous nous sommes dans cette partie avant tout intéressé au problème du vibrato. Pour la plupart des caractéristiques données dans l'introduction à cette partie, les travaux en sont encore à l'état de perspectives. Indiquons seulement que la majorité des techniques que nous venons de décrire dans le cas du vibrato sont adaptables au cas du trémolo.

Le nouvel algorithme pour la segmentation est présenté sur la figure 17.1. Le niveau de *segmentation en caractéristiques* est ajouté. Il a d'abord pour fonction d'indiquer au niveau de *segmentation en zones stables* quelles fonctions d'observation il peut utiliser. Il a aussi pour fonction de fournir au niveau de *segmentation en zones stables* des fonctions d'observation traitées de telle manière qu'elles soient utilisables (par exemple, le vibrato est supprimé sur le trajet de f_0, et les fonctions d'observation basées sur f_0 sont calculées avec ce nouveau trajet de f_0 : voir le chapitre 15).

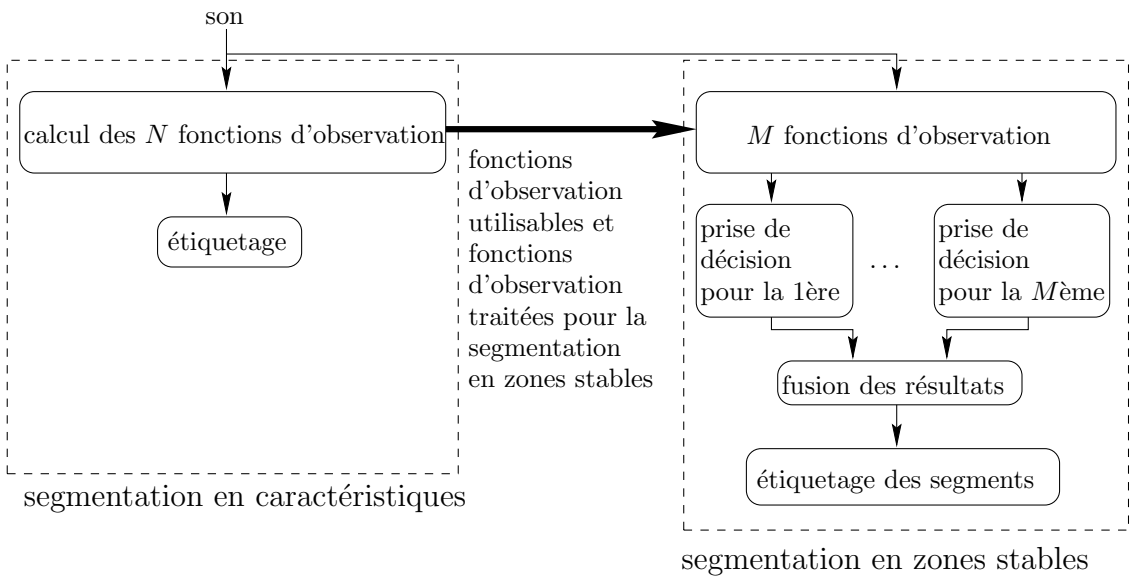


FIG. 17.1 – *Segmentation en zones stables : algorithme plus complet que celui de la figure 9.1*

Un exemple avec l'inharmonicité est développé ici. Un signal de castagnettes, par exemple, est toujours inharmonique. Donc, la fonction d'observation « fusion des valeurs absolues des dérivées des indices d'inharmonicité » ne présente pas de pics fins et grands quand les transitions ont lieu (dans ce cas, les transitions sont des transitoires d'énergie). Donc, cette fonction d'observation ne peut pas nous aider à poser les marques de segmentation pour la *segmentation en zones stables*. De la même façon, les autres fonctions d'observation basées sur le trajet de f_0 (« valeurs absolues des dérivées de f_0 », « fusion des valeurs absolues des dérivées des indices de voisement première forme »), les fonctions d'observation basées sur le contenu spectral (« valeur absolue de la dérivée de l'indice de voisement deuxième forme » et « flux spectraux ») ne peuvent pas être utilisées. Dans le cas extrême du signal de castagnettes, seule les fonctions d'observation basées sur l'énergie nous donnent des résultats exploitables.

Ainsi, lors de la *segmentation en zones stables*, il s'agit de classer les échantillons de la fonction d'observation en deux classes. Lors de la *segmentation en caractéristiques*, il s'agit de décider si les échantillons de la fonction d'observation sont utilisables (moyenne et variance petites sur un segment d'une seconde, calculées avec les n % plus petits : voir la méthode de seuillage décrite dans la section 3.2.1, page 42) ou non (moyenne et variances grandes).

Quatrième partie

Segmentation et indexation des sons polyphoniques – Introduction à la séparation de sources

Chapitre 18

Présentation des problèmes

Nous nous intéressons dans cette partie au cas des sons polyphoniques. Un son polyphonique est un son composé de plusieurs voix. Un son polyphonique est le mixage de plusieurs sons monophoniques : une voix chantée avec un ou plusieurs instruments, ou deux ou plusieurs instruments, etc. Il s'agit de définir ce qu'est la stabilité d'un signal sonore dans le cas polyphonique.

Dans le cas monophonique, nous avons vu que la stabilité correspond à :

- Une fréquence fondamentale qui varie peu. Ce critère n'est pas valide dans le cas polyphonique :
  - Puisque plusieurs fréquences fondamentales sont présentes simultanément si les voix mixées sont harmoniques
  - Ou puisque une partie importante de l'énergie correspond à du bruit si l'une des voix n'est pas harmonique (percussion)
- Une énergie qui varie peu. Ce critère reste valide.
- Ou un contenu spectral qui varie peu. Ce critère est encore valide ici.

Dans le cas polyphonique, la stabilité correspond à une énergie et à un contenu spectral qui varient peu. Ainsi, certaines des fonctions d'observation décrites dans la partie II sont encore utilisables pour segmenter en zones stables les sons polyphoniques.

Considérons un mixage de deux ou plusieurs voix parfaitement harmoniques. Dans un premier temps, nous voulons détecter la polyphonie sans passer par le trajet des partiels et la détection de plusieurs fréquences fondamentales par le groupement de ces partiels en plusieurs peignes harmoniques emplis d'une manière dense. C'est-à-dire que dans un premier temps nous ne nous intéressons pas à la séparation de sources, que nous ne nous intéressons pas au nombre de sources présentes, etc. Avant d'essayer de séparer les sources, il faut d'abord avoir détecté que nous sommes en présence d'un son polyphonique. Nous partons des considérations suivantes :

- Un signal polyphonique n'est pas harmonique (comme l'est un son dû à une flûte seule par exemple), ou alors la fréquence fondamentale est très petite (le Plus Grand Commun Diviseur de tous les partiels présents est petit) et il manque beaucoup d'harmoniques à cette fréquence fondamentale. Ainsi, sur une zone stable, nous n'avons pas même forcément une période de la fréquence fondamentale.
- Un signal polyphonique n'est pas « presque » harmonique (comme l'est une note de piano seule par exemple). C'est-à-dire que nous ne considérons pas dans cette partie d'instruments qui comme le piano ne soient que « presque » harmoniques. Il s'agit de la première hypothèse restrictive faite.
- La proportion d'énergie due à du bruit est petite devant la proportion d'énergie due à des sinusoïdes : en fait, cela veut dire que dans cet exposé nous nous intéressons au mixage de plusieurs voix parfaitement harmoniques. Il s'agit de la deuxième hypothèse restrictive faite. Dans un second temps (il s'agit de perspectives), nous nous intéresserons au mixage de « bruits » (percussions) et d'instruments harmoniques.

- Une troisième hypothèse restrictive est faite : le nombre de voix mixées est deux.
- Une quatrième et dernière hypothèse restrictive est faite : nous supposons que les deux peignes harmoniques ne se chevauchent pas complètement mais au plus en partie. C'est-à-dire que nous supposons qu'aucune des fréquences fondamentales ne correspond à un harmonique de l'autre son présent.

Notre premier objectif, dans cette partie, est de construire une « fonction d'observation » qui nous permette de détecter la polyphonie. Cette fonction d'observation intervient bien sûr lors de la *segmentation en caractéristiques*. Au sujet de la détection de la polyphonie, voir le **deuxième chapitre** (chapitre 19) de cette partie.

Le **troisième chapitre** (chapitre 20) de cette partie concerne la *segmentation en zones stables* des sons polyphoniques. Nous recensons les fonctions d'observation utilisables dans ce cas. Quelques performances sont données.

Le **quatrième chapitre** (chapitre 21) de cette partie concerne la *séparation de sources*, c'est-à-dire l'extraction des voix présentes dans le son.

Nous donnons une conclusion à cette partie dans le **cinquième chapitre** (chapitre 22).

Chapitre 19

Détection de la polyphonie

19.1 Introduction

Pour la catégorie restreinte de sons polyphoniques que nous considérons dans cet exposé (deux voix harmoniques mixées n'appartenant pas à un peigne harmonique commun), la stratégie choisie pour détecter la polyphonie passe par l'étude simultanée d'un indice d'inharmonicité (voir la section 2.2.4 de la partie II) et d'un indice de voisement (voir les sections 2.2.3, et surtout 2.4.1 et 2.4.2 de la partie II).

Il s'agit ici de caractériser automatiquement un signal avec cette **première étiquette** (cette étiquette concerne le niveau de *segmentation en caractéristiques* : voir la partie III et le chapitre 25 de la partie V) :

- Ou bien : le signal est plutôt bruité, c'est-à-dire qu'il est non voisé (voir la page 21 : $R = \frac{e}{E}$ est petit).
- Ou bien : le signal est plutôt composé d'une somme de sinusoïdes, c'est-à-dire qu'il est voisé ($R = \frac{e}{E}$ est grand).
  Tout signal stationnaire, de durée infinie et composé d'une somme de sinusoïdes de périodes commensurables, est harmonique, c'est-à-dire périodique. Seulement, les sons ne sont pas de durée infinie : ils ne sont stationnaires (stables) que par zones. Considérons un son composé de plusieurs voix mixées. Si deux sinusoïdes (l'une appartenant à une voix et l'autre à une autre), au cours d'une zone stable, ont des fréquences très proches, la fréquence fondamentale résultante est très petite (la période fondamentale est très grande), il manque beaucoup d'harmoniques de cette fondamentale, et la durée de cette zone stable n'est pas forcément plus grande que la période fondamentale. À ce propos, dans l'introduction à cette partie, nous dîmes abusivement qu'un signal polyphonique n'est pas harmonique, du fait des constatations précédentes.

La **seconde étiquette**, valable seulement pour les sons voisés (selon la définition de « voisé » que nous venons de donner, soit : un son voisé est un son composé de sinusoïdes), est :

- Ou bien : le signal est monophonique. Il est composé d'une somme de sinusoïdes appartenant toutes à un peigne harmonique commun, ce peigne étant empli d'une manière dense (c'est-à-dire que peu d'harmoniques sont absents).
- Ou bien : le signal est polyphonique. Selon les considérations précédentes : si le signal est voisé, si la fréquence fondamentale est petite, s'il manque beaucoup d'harmoniques et s'il y a moins d'une période du signal dans une zone stable, le signal est polyphonique.

Ainsi, nous disons qu'un signal voisé et harmonique est un signal monophonique, alors qu'un signal voisé mais pas harmonique est un signal polyphonique. Voir le tableau 19.1, qui synthétise ces remarques.

	signal voisé (composé de sinusoïdes)	signal non voisé (pas composé de sinusoïdes)
signal harmonique (périodique par zones stables)	son monophonique	–
signal inharmonique (non périodique par zones stables)	son polyphonique	bruit

TAB. 19.1 – *Essai de classification des sons en « sons monophoniques » et « sons polyphoniques »*

19.2 Détection du voisement, de l'harmonicité et de la polyphonie

19.2.1 Introduction

Nous calculons le voisement avec l'un des indices de voisement décrit dans la section 2.4 de la partie II. Supposons que nous ayons détecté que le signal est voisé (c'est-à-dire composé de sinusoïdes). Il faut mesurer la périodicité – leur plus grand commun diviseur – des positions des maximums locaux de la corrélation gardés. Si cette mesure est grande le signal est monophonique, sinon, il est polyphonique.

Mais, en fait, bien sûr, la fréquence f_l du $l^{\text{ème}}$ partiel n'est pas tout à fait égale à lf_0. Ce pour plusieurs raisons : nous faisons une petite erreur sur f_0 (due à du bruit par exemple), ou, en général, nous faisons une petite erreur sur les f_l ; de plus, lors du calcul du voisement, nous obtenons les corrélations pour des fréquences $j\Delta f$, où j est un entier et Δf égal à $\frac{f_e}{t_{FFT}}$, et non pas continuement pour toutes les fréquences. Ainsi, le plus grand commun diviseur n'existe pas, ou est, absurdement, très petit, même quand le son est parfaitement harmonique. À ce sujet voir la section 2.2.1.2 de cet exposé et la thèse de Boris DOVAL [Dov94]. DOVAL préconise de prendre comme densité de probabilité pour f_l une gaussienne dont la moyenne m est la fréquence f_l trouvée et dont la variance σ^2 dépend de f_0. De la même façon, est définie une probabilité d'absence de l'harmonique de numéro d'ordre l, et la régularité de l'enveloppe spectrale est prise en compte.

Cependant, ici, nous ne connaissons pas f_0 et nous ne sommes pas intéressé par sa détermination. Nous voulons seulement une mesure simple et rapide de l'harmonicité, ou de la densité du peigne harmonique où sont rangés au mieux les partiels détectés. Deux méthodes pour mesurer le degré d'harmonicité d'un ensemble de pics sont données dans les sections suivantes (sections 19.2.3 et 19.2.4).

19.2.2 Principe des mesures de l'harmonicité

Soient $[f_1 \ \ldots \ f_N]$ les positions, en Hz, des N pics de corrélation gardés, ordonnées de la plus petite à la plus grande. Les configurations possibles sont :

- Les N fréquences sont presque harmoniques
- M fréquences sont presque harmoniques et m correspondent à des fausses alarmes (dues au bruit ou aux lobes secondaires) : nous avons $M \gg m$
- M_1 fréquences correspondent au premier son présent et M_2 au second son présent : nous avons $M_1 \simeq M_2$ ($\simeq$ voulant dire ici du même ordre de grandeur)
- M_1 fréquences correspondent au premier son présent, M_2 au second son présent, et m à des fausses alarmes : nous avons $M_1 \simeq M_2$, $M_1 \gg m$ et $M_2 \gg m$

En pratique, seules les deuxième et quatrième configurations sont obtenues. Il s'agit de mettre en évidence leurs différences, en se rendant compte que les m fausses alarmes vont avoir une influence néfaste sur les mesures de l'harmonicité. Il faut réduire l'influence des fausses alarmes sans détruire celle du deuxième son harmonique.

19.2.3 Première mesure

Le signal est présumé harmonique. Il s'agit de prouver qu'il ne l'est pas.
Trois cas sont possibles :

- Si nous faisons l'hypothèse que f_1 est égale à f_0, f_i correspond à l'harmonique de numéro d'ordre $l = \text{round}\left(\frac{f_i}{f_1}\right)$, où « round » est l'opérateur d'arrondi (round(a,b) est égal à a si $b \in [0 \ \ldots \ 5]$ et à $a+1$ si $b \in]5 \ \ldots \ 10[$). Remarquons que plusieurs f_i peuvent correspondre au même harmonique l ! Par exemple, f_i et f_{i+1} peuvent être très proches : l'une est due à un partiel réel et l'autre est une fausse alarme, ou les deux sont des fausses alarmes. L'indice d'inharmonicité h_i pour l'harmonique i est alors : $h_i = \left|\frac{f_i}{f_1} - l\right|$. L'indice d'inharmonicité global est égal à : $H_1 = \frac{1}{N}\sum_{i=1}^{N} h_i$.
- Si nous faisons l'hypothèse que f_1 est égale à jf_0, où j est un entier (c'est-à-dire qu'il manque les harmoniques $[1 \ \ldots \ j-1]$), f_i correspond à l'harmonique de numéro d'ordre $l = \text{round}\left(\frac{jf_i}{f_1}\right)$. L'indice d'inharmonicité global est égal à : $H_j = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{jf_i}{f_1} - l\right|$.
  Nous calculons l'indice d'inhamonicité global H_j pour plusieurs j, en nous limitant aux J plus petits j puisque la probabilité d'absence des premiers harmoniques est faible (voir [Dov94]). Si $\alpha = \min[H_1 \ \ldots \ H_J]$ (l'opérateur « min » nous donne la plus petite valeur du tableau à J éléments [. . .]) est suffisamment petite, le signal est harmonique, et donc monophonique. Sinon, il est polyphonique. Il existe malheureusement un cas pathologique, décrit dans le point suivant.
- Si f_1 est une fausse alarme, la méthode s'effondre tout à fait. En fait, ce défaut rend à lui seul cette méthode, trop simple, peu fiable.

Quelques résultats de la méthode sont donnés dans la section 19.2.5, sur la figure 19.1.

19.2.4 Deuxième mesure

Si le son est monophonique, l'écart e_i entre f_{i+1} et f_i ($e_i = f_{i+1} - f_i$) est, à peu près, un multiple de f_0 (pour la clarinette, par exemple, les amplitudes des harmoniques de numéros d'ordre pairs étant très faibles, souvent ces harmoniques ne sont pas détectés : dans ce cas, e_i est proche de $2f_0$). Nous dressons la liste des écarts $[e_0 \ \ e_1 \ \ \ldots \ \ e_{N-1}]$, e_0 valant f_1, c'est-à-dire dans le cas monophonique sans fausses alarmes presque f_0. Ainsi, la valeur médiane ME de ces écarts correspond à peu près à jf_0, où j est un entier. La suite de la procédure est itérative :

Étape 1 Nous déterminons l'écart e_I (correspondant par exemple à la différence entre les fréquences f_1 et f_2) qui est « le plus éloigné » de cette valeur médiane.

Pour ce faire, le tableau $\left[\underbrace{\left\{\frac{ME}{e_0} \ \ldots \ \frac{ME}{e_{N-1}}\right\}}_{\text{partie 1}}\underbrace{\left\{\frac{e_0}{ME} \ \ldots \ \frac{e_{N-1}}{ME}\right\}}_{\text{partie 2}}\right]$ est formé. Nous détectons le maximum de ce tableau.

Étape 2 Nous cherchons par quel entier d il faut multiplier (si le maximum est trouvé dans la première partie de la liste) ou diviser (si le maximum est trouvé dans la seconde partie de la liste) cet écart pour qu'il soit le plus près de cette valeur médiane.

Étape 3 Si d est différent de 1 : nous éliminons cet écart de la liste, nous calculons la nouvelle valeur médiane avec cette nouvelle liste, et nous retournons en 1 (avec $N = N - 1$).

Étape 4 Sinon, nous allons en 5.

Étape 5 Nous divisons la taille de la liste initiale par la taille de la liste finale : si ce rapport v est grand, le son n'est pas monophonique ; s'il est petit, le son est monophonique.

Quelques résultats de la méthode sont donnés dans la section 19.2.5, sur la figure 19.2.

19.2.5 Quelques tests

Pour chacune des deux méthodes, nous donnons les quatre histogrammes des mesures (α pour la première, v pour la deuxième) calculés chacun pour 10000 observations. Voir respectivement les figures 19.1 et 19.2. Les quatre configurations sont :

- signal monophonique : N fréquences presque harmoniques (histogramme en haut à gauche)
- signal monophonique : N fréquences presque harmoniques, et n fausses alarmes (histogramme en haut à droite)
- signal polyphonique : N_1 fréquences presque harmoniques, et N_2 fréquences presque harmoniques (histogramme en bas à gauche)
- signal polyphonique : N_1 fréquences presque harmoniques, N_2 fréquences presque harmoniques, et n fausses alarmes (histogramme en bas à droite)

La fréquence fondamentale f est choisie aléatoirement pour chaque observation et dans le cas polyphonique pour chaque son entre 0,125 et 1 (cet écart permet de couvrir trois gammes). La demi-fréquence d'échantillonnage est 20. Ainsi, N (cas monophonique), et N_1 et N_2 (cas polyphonique) sont compris entre 20 et 160. La fréquence de l'harmonique de numéro d'ordre l est égale à $lf + \mathcal{N}\left(0{,}10^{-8}\right)$. Le nombre de fausses alarmes est soit 0, soit $\dfrac{N}{4}$ (cas monophonique) ou $\dfrac{N_1 + N_2}{4}$ (cas polyphonique). La position des fausses alarmes est une variable aléatoire uniformément répartie entre 0 et 20.

La première méthode nous permet de détecter la polyphonie, malgré un cas pathologique gênant.

La deuxième méthode nous permet de détecter efficacement la polyphonie.

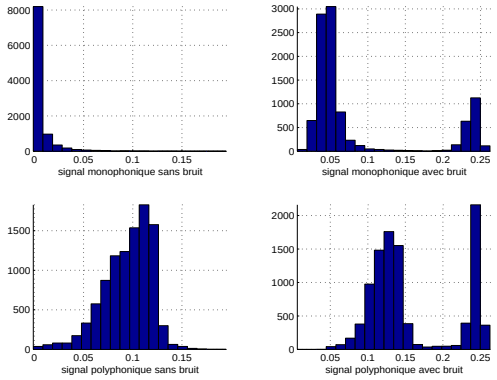


FIG. 19.1 – *Résultats pour la première mesure. Histogrammes de α*

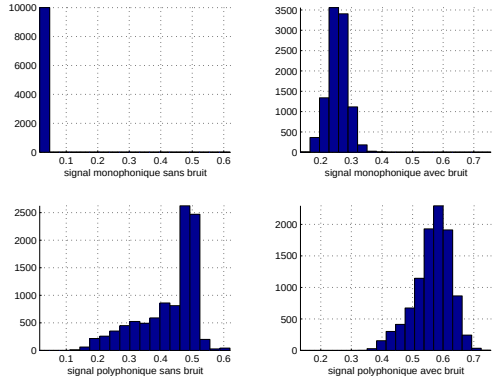


FIG. 19.2 – *Résultats pour la deuxième mesure. Histogrammes de v*

19.2.6 Conclusion

Quelques performances ont été données. Il s'agit de résultats préliminaires : les perspectives seront d'améliorer les techniques présentées ici et de les tester avec des signaux réels.

Chapitre 20

Segmentation des sons polyphoniques

20.1 Fonctions d'observation utilisables dans le cas polyphonique

Dans le tableau 20.1, nous indiquons lesquelles des fonctions d'observation décrites dans la partie II (voir le tableau 2.5) sont utilisables dans le cas polyphonique restreint considéré dans cet exposé.

principe	son polyphonique admis
dérivées de f_0	non
dérivées du voisement (forme 1)	non
dérivées des inharmonicités	non
analyse statistique sur f_0	non
rupture de modèles sur f_0	non
f_0 après filtrage de HILBERT	non
dérivées de l'énergie	oui
analyse statistique sur l'énergie	oui
rupture de modèles sur l'énergie	oui
dérivées du voisement (formes 2)	oui
flux entre les spectres	oui
flux enveloppe AR - spectre	oui
flux entre enveloppes AR	oui
flux enveloppe cepstre - spectre	oui
flux entre enveloppes cepstres	oui
flux enveloppe maximums - spectre	oui
flux entre enveloppes maximums	oui
flux entre enveloppes superposées	oui
dérivées du centroïde	oui
rupture de modèles sur le signal	oui
test de BRANDT	oui
analyse de la stationnarité	oui
méthode de MASRI	oui
méthode de HAJDA	oui
méthode de SMITH	oui

TAB. 20.1 – *Fonctions d'observation pour la polyphonie*

Nous supposons que les sons étudiés dans cette partie sont composés de deux voix et que chacune d'elle est harmonique par zone stable (c'est-à-dire par note). Les transitions peuvent être de deux types :

- Nous avons un changement de note pour les deux voix simultanément.
- Nous avons un changement de note pour une seule des voix : pour l'autre voix, nous restons sur la même note.

20.2 Quelques performances avec un signal synthétique

20.2.1 Le signal sonore polyphonique synthétique

Le signal synthétisé est la somme de deux sons de deux secondes.

- Le premier son est formé de deux notes successives d'une seconde chacune. Les fréquences fondamentales respectives des deux notes sont $f_0^{(1)} = 300\ Hz$ et $f_0^{(2)} = 710\ Hz$. Chaque note est composée des 20 premiers harmoniques. L'amplitude de chaque harmonique est égale à $\frac{1}{l^2}$, où l est le numéro d'ordre de l'harmonique.
  Le modèle de transition entre les deux notes harmoniques utilisé est celui décrit dans la section 11.2.3, page 89. Pour la transition en fréquence, nous avons pris : $a = 1$ et $b = 0{,}002$. Pour la transition en amplitude, nous avons pris : $a_1 = 1 - 0{,}001$, $a_2 = 1 + 0{,}001$, $b_1 = b_2 = 0{,}001$ et $c_1 = c_2 = \frac{1}{l^2}$. Ceci correspond à une transition très brutale, comme nous le constatons sur la figure 20.1 où est représenté ce son au moment de la transition.

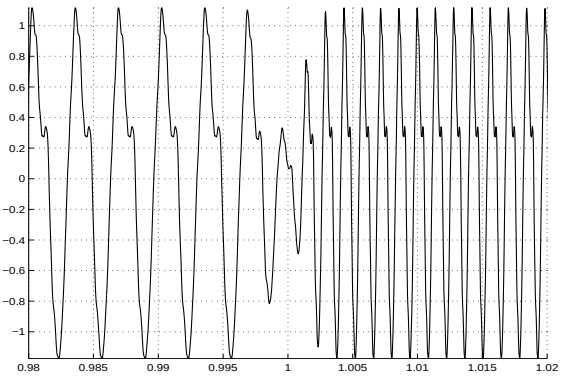


FIG. 20.1 – *Signal sonore simulé. Le signal lors de la brusque transition entre les deux notes est représenté. En abscisse : le temps en seconde ; en ordonnée : l'amplitude du signal*

- Le second son est formé d'une note fixe, de fréquence fondamentale $f_0^{(3)} = 490\ Hz$. Elle est composée des 20 premiers harmoniques, dont les amplitudes respectives sont égales à $\frac{2}{l^2}$.

Il s'agit de détecter la transition entre les notes du premier son.

20.2.2 Performances avec la fonction d'observation « analyse de la stationnarité »

La définition de la fonction d'observation est donnée dans la section 2.5.4, page 31. Nous avons utilisé $\Theta = [0\ 2\ 4\ \ldots\ 200]$ et une fenêtre d'analyse large de 45 millisecondes. Nous obtenons les trajets présentés sur la figure 20.2.

Pour la première zone stable, nous avons :

$$E\left[X_n X_{n+\eta}\right] = \frac{1}{2}\sum_{l=1}^{20}\frac{1}{l^2}\cos\left(2\pi l f_0^{(1)}\frac{\eta}{f_e}\right) + \frac{1}{2}\sum_{i=l}^{20}\frac{2}{l^2}\cos\left(2\pi l f_0^{(3)}\frac{\eta}{f_e}\right)$$

Et pour la seconde :

$$E\left[X_n X_{n+\eta}\right] = \frac{1}{2}\sum_{l=1}^{20}\frac{1}{l^2}\cos\left(2\pi l f_0^{(2)}\frac{\eta}{f_e}\right) + \frac{1}{2}\sum_{l=1}^{20}\frac{2}{l^2}\cos\left(2\pi l f_0^{(3)}\frac{\eta}{f_e}\right)$$

Pour $\eta = 0$, l'influence du son « continu » (du son qui ne change pas) est deux fois plus importante que l'influence du son « variable ». Nous constatons qu'en effet la dérivée de $E\left[X_n X_n\right]$ ne réagit pas ou presque pas au moment de la transition. Mais, pour certains η, l'influence du son « continu » et celle du son « variable » s'équilibrent, et à la transition nous obtenons un pic. La « moyenne des valeurs absolues des dérivées obtenues pour un grand nombre de η » nous permet de détecter la transition.

20.2.3 Performances avec la fonction d'observation « test de Brandt appliqué au signal »

La définition de la fonction d'observation est donnée dans la section 2.5.3. Nous avons utilisé une fenêtre d'analyse large de 30 millisecondes. Nous obtenons le trajet présenté sur la figure 20.3.

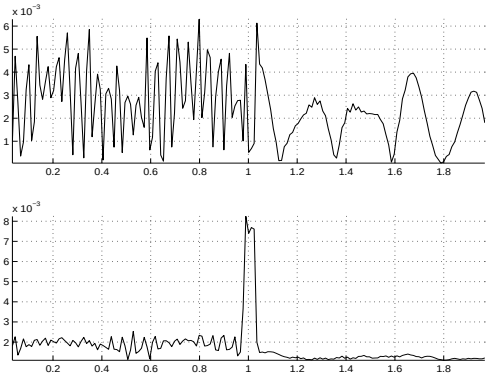


Fig. 20.2 – *Trajets obtenus avec l'analyse de la stationnarité sur le signal polyphonique synthétique. En haut : la « valeur absolue de la dérivée de l'énergie ». En bas : la « moyenne des valeurs absolues des dérivées des coefficients d'autocorrélation ». En abscisse : le temps en seconde*

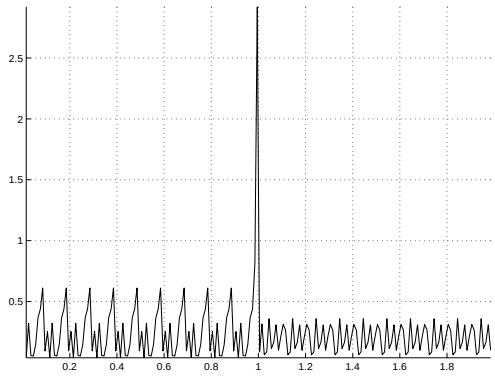


Fig. 20.3 – *Trajet obtenu avec le « test de Brandt » sur le signal polyphonique synthétique. En abscisse : le temps en seconde*

20.2.4 Performances avec la fonction d'observation « flux spectral calculé avec les spectres d'amplitude »

La définition de la fonction d'observation est donnée dans la section 2.4.3.2, page 27. Nous avons pris $t_{SIG} = 1280$, $t_{FFT} = 2048$ et $Q = 220$. La fenêtre de pondération de Blackman est utilisée. Nous obtenons le trajet présenté sur la figure 20.4.

20.2.5 Performances avec la fonction d'observation « indice de voisement deuxième forme calculé avec le spectre d'amplitude »

La définition de la fonction d'observation est donnée dans la section 2.4.1, page 20. Nous avons pris $N = 72$, $t_{FFT} = 16384$, $s = 0{,}8$ et $t_{SIG} = 1280$. La fenêtre de pondération de Blackman est utilisée. Nous obtenons le trajet présenté sur la figure 20.4.

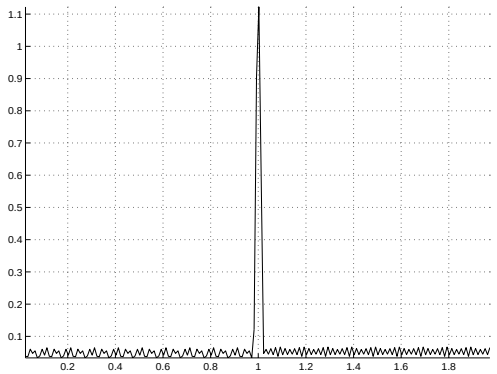


FIG. 20.4 – *Trajet obtenu avec le « flux spectral calculé avec les spectres d'amplitude » sur le signal polyphonique synthétique. En abscisse : le temps en seconde*

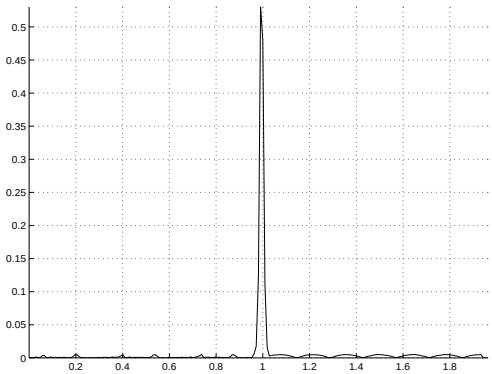


FIG. 20.5 – *Trajet obtenu avec l'« indice de voisement deuxième forme calculé avec le spectre d'amplitude » sur le signal polyphonique synthétique. En abscisse : le temps en seconde*

Chapitre 21

Introduction à la séparation de sources

21.1 Introduction

Les sons enregistrés à la radio que nous utilisons pour la *segmentation en sources* (voir la partie V, chapitre 24) sont communément polyphoniques en ce qui concerne la musique (voix chantée plus instrument ou instruments), contrairement à ce qui se passe en ce qui concerne la voix parlée (rarement plusieurs personne parlent en même temps).

En ce qui concerne la séparation de sources, un certain nombre de travaux existent (articles, thèses et livres) : voir entre autres [Moo75], [CJ86], [Mah90], [Mel91], [Wei84], [Ell96], [Bre90], [Wan94], etc. Une bibliographie et un état de l'art sont donnés dans [Kla97]. Il s'agit d'obtenir les trajets des partiels, de segmenter, d'apparier les partiels de chaque segment en groupes harmoniques, et enfin d'apparier les groupes d'un segment avec ceux des autres segments.

Le principal problème vient sans doute du traçage difficile des trajets des partiels dans le cas polyphonique. Dans le cas monophonique, des algorithmes pour détecter f_0 robustes existent, et, une fois que la fréquence f_0 a été déterminée, il est aisé de déterminer les trajets des harmoniques de numéros d'ordre supérieurs : il est fait l'hypothèse que les partiels appartiennent à un peigne harmonique. En ce qui nous concerne, les trajets des harmoniques sont obtenus grâce au logiciel ADDITIVE, logiciel développé à l'IRCAM.

Dans le cas polyphonique, nous ne pouvons plus utiliser ce trajet de f_0, et nous devons déterminer les trajets des partiels l'un après l'autre. D'autres problèmes existent :

- Il arrive que la fréquence d'un harmonique d'une voix et la fréquence d'un harmonique d'une autre voix soient identiques : alors, de l'information est perdue. Dans le cas extrême, les deux sinusoïdes sont en opposition de phase, et ainsi disparaissent toutes les deux complètement.
- Un harmonique d'une voix peut croiser un harmonique d'une autre voix : par exemple, la fréquence fondamentale d'une voix reste fixe alors que pour l'autre nous avons un glissando. Alors, il est difficile de faire les raccordements après le croisement pour obtenir deux trajets continus.

Le problème principal demeure que les partiels ne sont plus répartis sur un peigne harmonique empli d'une manière dense, et que donc nous devons déterminer les trajets des partiels un à un. Aussi, nous utilisons le logiciel HMM (HMM pour « Hidden Markov Model »), qui a été développé à l'IRCAM, par Guillermo GARCIA notamment (voir l'article [DGR93]), et qui permet de pallier ces inconvénients. HMM utilise des modèles de MARKOV pour déterminer le chemin des partiels.

Dans la **deuxième section** (section 21.2) de ce chapitre, nous présentons très succintement le logiciel HMM.

Dans la **troisième section** (section 21.3) de ce chapitre, nous présentons la procédure de base suivie pour la séparation de sources.

Dans la **quatrième section** (section 21.4) de ce chapitre, nous présentons les problèmes que pose l'appariement des partiels.

Dans la **cinquième section** (section 21.5) de ce chapitre, nous discutons les performances obtenues avec notre système simple.

21.2 Fonctionnement abrégé du logiciel HMM

Cette fois, contrairement à ce qui se passe pour le logiciel f_0, il n'est pas fait d'hypothèse sur l'harmonicité du signal. Les spectres d'amplitude sont toujours calculés pour des portions glissantes du signal, larges de quelques dizaines de millisecondes. Les maximums locaux de ces spectres d'amplitude sont détectés. Puis nous essayons de relier les pics détecter pour un trame à ceux détectés pour les $N-1$ trames précédentes. Le cœur du logiciel se situe dans la phase de suivi des partiels, qui utilise un algorithme inspiré des modèles de MARKOV cachés (HMM). Pour chaque séquence de N pics possible, une fonction de score est calculée qui prend en compte les différences de fréquence, d'amplitude et de phase entre deux pics de deux trames successives.

Soit $p^{(i)}$, avec $i = 0 \ \ldots \ N-1$, une séquence de pics extraits de N trames successives. Le score S de cette séquence vaut : $S = S^{(1)} \ S^{(2)} \ \ldots \ S^{(N-2)}$ avec (f est mis pour fréquence, a pour amplitude et φ pour phase) :

$$S^{(i)} = S_f^{(i)} S_a^{(i)} S_\varphi^{(i)}$$

et :

$$S_f^{(i)} = \exp\left(-\frac{\mathrm{G}_f \left(p_f^{(i-1)} - 2p_f^{(i)} + p_f^{(i+1)} \right)^2}{\sigma_{\mathrm{f}}^2} \right)$$

$$S_a^{(i)} = \exp\left(-\frac{\mathrm{G}_a \left(p_a^{(i-1)} - 2p_a^{(i)} + p_a^{(i+1)} \right)^2}{{p_a^{(i)}}^2 \sigma_{\mathrm{a}}^2} \right)$$

$$S_\varphi^{(i)} = \exp\left(-\frac{\mathrm{G}_\varphi \left(p_\varphi^{(i-1)} - 2p_\varphi^{(i)} + p_\varphi^{(i+1)} \right)^2}{{\sigma^2}_\varphi} \right)$$

Ainsi, les évolutions linéaires des caractéristiques fréquence, amplitude et phase sont favorisées. Les variances σ_f^2, σ_a^2 et σ_φ^2 permettent de contrôler l'écart à la linéarité de chaque caractéristique. Les gains G_f, G_a et G_φ permettent de contrôler le poids relatif de chaque caractéristique. Ce sont des paramètres libres contrôlables par l'utilisateur de HMM. Ne sont gardées que les séquences ayant un score suffisamment élevé.

21.3 Procédures

21.3.1 Hypothèses faites dans cet exposé

Dans un premier temps, nous faisons quelques hypothèse simplificatrices :

- Le son est le mixage de deux sons monophoniques.
- Chacun de ces deux sons est harmonique.
- HMM nous donne des trajets de partiels continus et numérotés, échantillonnés par défaut à 100 Hz. Nous supposons que nous avons obtenu le trajet de tous les partiels à tout instant d'échantillonnage (sauf aux moments des transitions). C'est-à-dire que nous avons obtenu même le trajet des partiels qui sont très proches et des partiels qui se croisent.
- Nous supposons que nous avons déjà obtenu la segmentation du son polyphonique en zones stables.

21.3.2 Première étape : appariement des partiels pour un segment

Nous admettons que les problèmes de traçage des partiels ont été résolus. Alors, des axes de recherche se dessinent. Notamment, le plus communément, après avoir obtenu les trajets des partiels, des critères sont développés qui permettent de les séparer en n groupes (si n sons sont entremêlés : dans notre cas, nous nous limiterons à $n = 2$). Par exemple, Ellis, dans [Ell96] (Mellinger dans [MMR91] et [Mel91] définit sensiblement les mêmes) donne les critères suivants :

- La quasi simultanéité des instants de début et de fin d'existence pour les harmoniques d'une même voix.
- Des relations d'harmonicité entre les harmoniques d'une même voix.
- Les mêmes modulations d'amplitude pour les harmoniques d'une même voix.
- Les mêmes modulations de fréquence pour les harmoniques d'une même voix.

21.3.3 Deuxième étape : appariement des segments

Il faut à présent apparier les segments.
Nous divisons les sons mixés en trois catégories :

- Les deux sons ne sont pas formés des mêmes notes et les instants de ruptures entre elles ne tombent pas aux mêmes instants pour les deux sons.
- Les deux sons ne sont pas formés des mêmes notes mais les instants de ruptures entre elles tombent aux mêmes instants pour les deux sons. Alors, pour séparer les sons, nous considérons le fait qu'ils ont par exemple des vibratos et/ou des trémolos différents.
- Les deux sons sont formés des mêmes notes ou constituent un accord consonant, et les instants de ruptures entre elles tombent aux mêmes instants pour les deux sons. Alors, pour les séparer, nous considérons le fait qu'ils ont par exemple des vibratos et/ou des trémolos différents.

21.3.4 Procédure suivie dans cet exposé

Pour les cas (plutôt idéals) présentés dans cet exposé, la procédure suivie pour séparer deux sons est schématisée sur la figure 21.1.

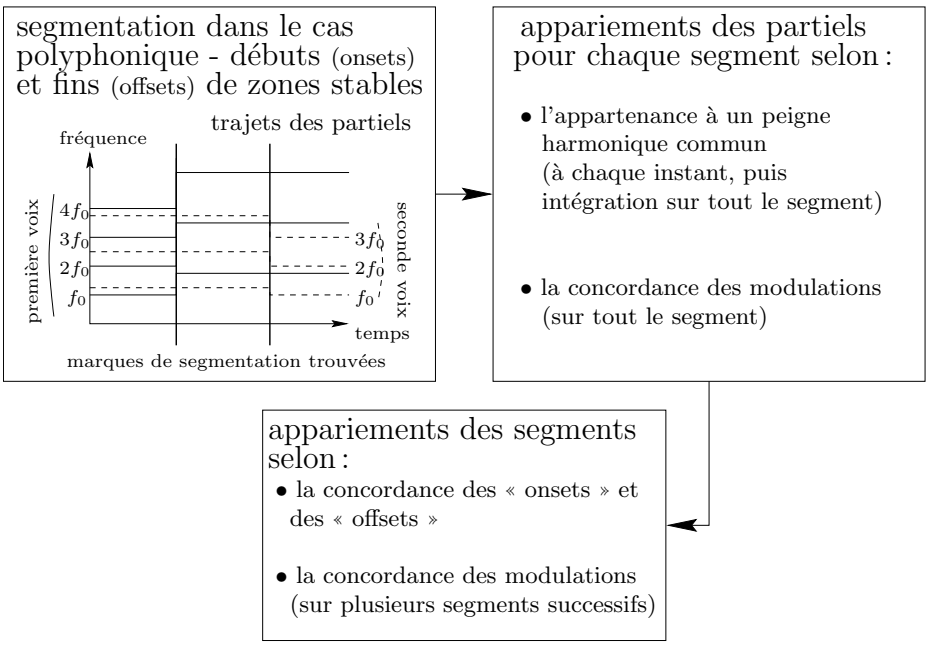


Fig. 21.1 – *Algorithme de base pour la séparation de sources*

- Dans un premier temps, nous considérons chaque segment i séparément.
  1° D'abord, à chaque instant d'échantillonnage des trajets des partiels, nous apparions les partiels en deux groupes : l'un rassemble les partiels de la note jouée par la première voix, et l'autre les partiels de la note jouée par la seconde voix.
  2° Ensuite, nous intégrons ces résultats sur tout le segment.
  3° Alors, nous comparons les modulations (de fréquence et d'amplitude) présentes sur les partiels de chaque voix, et nous vérifions qu'elles concordent. Finalement, pour le segment i, nous obtenons les deux groupes de partiels A_i et B_i.
- Dans un second temps, il s'agit de regrouper les segments. Il faut déterminer quelle configuration est la bonne : A_{i+1} est la suite de A_i et donc B_{i+1} celle de B_i ; ou B_{i+1} est la suite de A_i et donc A_{i+1} celle de B_i.

  1° En détectant s'il y a continuité en fréquence entre un groupe du premier segment et un groupe du deuxième segment. Il y a trois cas :

  - Une telle continuité existe en effet pour deux groupes, et pas pour les deux autres. Cela veut dire que nous avons eu un changement de note pour une des voix et pas pour l'autre. Les liens entre les groupes sont aisés à faire. Et nous savons à quelle voix est due la transition.
  - Une telle continuité n'existe pas. La transition est due aux deux voix : pour les deux voix, nous avons eu un changement de note. Pour lier les groupes, il faut utiliser des traitements supplémentaires. Il faut étudier les modulations. Par exemple, une voix peut être modulée en fréquence (vibrato) et pas l'autre : alors nous pouvons lier les groupes. Nous pourrions étudier aussi la continuité des enveloppes spectrales, etc. : il s'agit d'une perspective.
  - Une telle continuité existe en effet pour deux groupes, mais aussi pour les deux autres. La transition est due à un transitoire d'énergie, ou alors il y a un trou dans le trajet d'un des partiels, dû à des difficultés rencontrées par HMM (l'amplitude de ce partiel peut être petite, et donc il est noyé dans du bruit). Cette configuration est pathologique.

  2° En comparant les modulations des quatre groupes A_i, A_{i+1}, B_i et B_{i+1} de partiels.
  3° En comparant les enveloppes spectrales. Ce n'est pas fait dans cet exposé : il s'agit d'une perspective.

Pour l'exemple schématique présenté sur la figure 21.1, nous voyons que les deux transitions sont chacune due à une seule voix.

21.4 Le problème de l'appariement des partiels

21.4.1 Appariement à chaque instant d'échantillonnage du trajet des partiels

21.4.1.1 Répartition des partiels en « groupes harmoniques »

Si, à un instant donné, nous avons détecté N partiels, nous prouvons qu'il y a :

$$S_N = \sum_{j=1}^{N-1} 2^{j-1} \text{ pour } N \geq 2$$

possibilités de les ranger en deux groupes, chacun de ces groupes étant constitué d'au moins 1 partiel.

En effet, supposons que nous ayons $N = 2$, le nombre de partiels. Les deux groupes sont appelés A et B. Nous n'avons qu'une possibilité : $A_1 = \{x_1\}$ et $B_1 = \{x_2\}$. Donc, si S_i est le nombre de possibilités pour $N = i$, nous avons $S_2 = 1$.

Si nous ajoutons une troisième donnée ($N = 3$) x_3, elle peut être classée soit avec x_1 (dans A_1), soit avec x_2 (dans B_1), soit toute seule les deux autres étant ensemble. Nous obtenons ainsi les configurations possibles :

$$\begin{array}{llll} A_1 = \{x_1x_3\} & \text{et} & B_1 = \{x_2\} \\ A_2 = \{x_1\} & \text{et} & B_2 = \{x_2x_3\} \\ A_3 = \{x_3\} & \text{et} & B_3 = \{x_1x_2\} \end{array}$$

Il n'y a pas d'autre possibilité. Nous avons donc $S_3 = 3$.

Si nous ajoutons encore une autre donnée ($N = 4$) x_4, elle peut être classée dans l'un des 6 groupes donné ci-dessus, ou rester toute seule. Nous obtenons ainsi :

$$\begin{array}{llll} A_1 = \{x_1x_3x_4\} & \text{et} & B_1 = \{x_2\} \\ A_2 = \{x_1x_3\} & \text{et} & B_2 = \{x_2x_4\} \\ A_3 = \{x_1x_4\} & \text{et} & B_3 = \{x_2x_3\} \\ A_4 = \{x_1\} & \text{et} & B_4 = \{x_2x_3x_4\} \\ A_5 = \{x_3x_4\} & \text{et} & B_5 = \{x_1x_2\} \\ A_6 = \{x_3\} & \text{et} & B_6 = \{x_1x_2x_4\} \\ A_7 = \{x_4\} & \text{et} & B_7 = \{x_1x_2x_3\} \end{array}$$

Il n'y a pas d'autre possibilité. Nous avons donc $S_4 = 7$.

Etc.

Nous voyons donc que $S_{i+1} = 2S_i + 1$. Et, puisque $S_2 = 1$, nous avons :

$$S_N = \sum_{j=1}^{N-1} 2^{j-1} \text{ pour } N \geq 2$$

Si 10 partiels sont présents par voix et si nous les détectons tous, nous avons S_{20}, le nombre de configurations possibles, qui vaut 524287. Si nous détectons 20 partiels par voix, nous avons S_{40} qui vaut à peu près 550 milliards !

De plus, en fait, il faudrait sans doute mieux considérer trois groupes : l'un pour les partiels de la première voix, un autre pour ceux de la seconde, et un dernier pour les « faux » pics (fausses alarmes, bruit...). Dans ce cas, nous obtenons T_{20} qui vaut près de 600 millions et T_{40} qui vaut deux milliards de milliards !

En effet, supposons que nous ayons $N = 3$. Les trois groupes sont appelés A, B et C. Nous n'avons qu'une possibilité : $A_1 = \{x_1\}$, $B_1 = \{x_2\}$ et $C_1 = \{x_3\}$. Donc, si T_i est le nombre de possibilités pour $N = i$, nous avons $T_3 = 1$.

Si nous ajoutons une quatrième donnée ($N = 4$), elle peut être classée soit avec x_1 (dans A_1), soit avec x_2 (dans B_1), soit avec x_3 (dans C_1), soit toute seule les trois autres étant réparties en 2 groupes. Nous obtenons ainsi :

$$\begin{array}{llllll} A_1 = \{x_1x_4\} & \text{et} & B_1 = \{x_2\} & \text{et} & C_1 = \{x_3\} \\ A_2 = \{x_1\} & \text{et} & B_2 = \{x_2x_4\} & \text{et} & C_2 = \{x_3\} \\ A_3 = \{x_1\} & \text{et} & B_3 = \{x_2\} & \text{et} & C_3 = \{x_3x_4\} \\ A_4 = \{x_4\} & \text{et} & B_4 = \{x_1x_3\} & \text{et} & C_4 = \{x_2\} \\ A_5 = \{x_4\} & \text{et} & B_5 = \{x_1\} & \text{et} & C_5 = \{x_2x_3\} \\ A_6 = \{x_4\} & \text{et} & B_6 = \{x_3\} & \text{et} & C_6 = \{x_1x_2\} \end{array}$$

Il n'y a pas d'autre possibilité. Nous avons donc $T_4 = 6$.

Et la règle générale s'écrit : $T_{i+1} = 3T_i + S_i$. Nous avons finalement :

$$T_N = 3^{N-3} + \sum_{j=3}^{N-1} \left(\sum_{k=1}^{j-1} 2^{k-1} 3^{N-j-1} \right) \text{ pour } N \geq 3$$

qui se simplifie peut-être (?).

Nous n'allons dans la suite ne considérer que deux groupes, mais le problème n'est pas résolu pour autant. Nous voyons que le problème combinatoire, du fait du grand nombre de répartitions/configurations

possibles, devient rapidement très coûteux en temps de calcul, si ce n'est insoluble. Le fait de ne pas considérer les configurations où il n'y aurait qu'un ou deux partiels dans l'un des groupes ne résoudrait qu'imparfaitement le problème.

Dans la suite, même pour les signaux réels, nous n'allons considérer que les premiers partiels, c'est-à-dire que ceux dont la fréquence est inférieure à quelques milliers de Hz (entre 1500 Hz et 3300 Hz pour les exemples donnés dans la section 21.5).

21.4.1.2 Mesure de la qualité de chaque répartition

Il s'agit à présent de tester la validité de chacune des configurations, et de ne garder que celle qui soit la plus valide, la plus probable.

Il s'agit de trouver le f_0 le plus probable pour chacun des groupes. Pour cela, nous utilisons les résultats de la thèse de DOVAL, notamment sur la probabilité d'absence de chaque partiel et la probabilité de dispersion de chaque partiel autour de sa position théorique.

21.4.2 Intégration sur un segment

HMM nous donne les trajets des partiels dans le temps. Ainsi, nous pouvons couper le signal en segments. L'instant limite entre deux segments successifs est donné par la naissance ou la mort d'un partiel. Ainsi, sur un segment, nous avons à tous les instants exactement le même nombre de partiels, et chacun de ces partiels est continu. Ces segments ne correspondent pas à la segmentation en zones stables : par exemple, pendant quelques échantillons, un partiel peut manquer, pour une raison ou une autre.

Il reste donc à vérifier sur un segment que les deux groupes trouvés à un instant sont cohérents avec ceux trouvés aux autres instants. Supposons que nous ayons trois partiels numérotés 1, 2 et 3. Si à l'instant i nous retenons les groupes $A = \{1 \quad 2\}$ et $B = \{3\}$, à l'instant $i+1$ les groupes $A = \{3\}$ et $B = \{1 \quad 2\}$ et à l'instant $i+2$ les groupes $A = \{1 \quad 2\}$ et $B = \{3\}$, il y a cohérence et l'intégration se fait aisément (note : les instants i, $i+1$ et $i+2$ sont directement consécutifs). Par contre, si à l'instant i nous trouvons que les groupes les plus probables sont $A = \{1 \quad 2\}$ et $B = \{3\}$, à l'instant $i+1$ $A = \{1\}$ et $B = \{2 \quad 3\}$ et à l'instant $i+2$ $A = \{1 \quad 3\}$ et $B = \{2\}$, il n'y a pas cohérence.

21.5 Les sons utilisés

21.5.1 Introduction

Des exemples avec des sons réels sont disponibles via internet. Ils correspondent aux sons sur lesquels ont travaillé par exemple WANG ([Wan94]) ou COOKE et BROWN ([CBCG94]). Nous ne les avons pas utilisés. Nous avons préféré travailler avec des sons parfaitement calibrés (fréquences fondamentales connues, etc.).

Dans un premier temps, nous simulons des sons monophoniques que nous mixons.

Nous avons pris comme sons monophoniques :

- Premier son : durée 2 secondes ; première note, formée de 4 partiels : $f_0 = 440\ Hz$, de $t = 0$ à $t = 1$ seconde ; deuxième note, formée de 4 partiels : $f_0 = 494\ Hz$, de $t = 1$ à $t = 2$ secondes.
- Second son : durée 2 secondes ; note formée de 3 partiels : $f_0 = 554\ Hz$, de $t = 0$ à $t = 2$ secondes.

Nous mixons ces deux sons. Pour nous essayons de les extraire (de les démixer).

D'autres sons ont été utilisés et traités. Les résultats sonores (sons avant mixage, sons mixés et sons démixés) peuvent être trouvés ici :

`http://www.ircam.fr/equipes/analyse-synthese/rossigno/these/separ.html`

L'exemple décrit ci-dessus est le deuxième exemple présenté.

21.5.2 Quelques performances

Le premier son extrait a été reconstruit avec 4 partiels pour la première note puis avec 4 partiels pour la seconde ; le second son a été reconstruit avec 3 partiels. Nous donnons sur la figure 21.2 les trajets des partiels du son résultant du mixage. À la main, nous avons indiqué à quel son appartient chaque trajet trouvé. Sur les figures 21.3 et 21.4 nous donnons le résultat de la séparation de sources. Nous voyons que les deux groupes de partiels trouvés sont corrects.

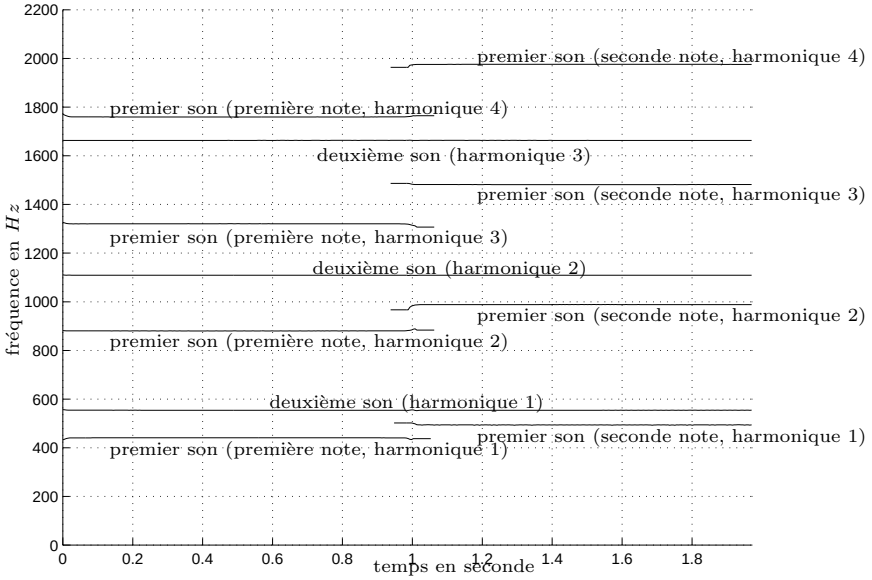


FIG. 21.2 – *Trajets des partiels pour le son résultant du mixage de deux sons harmoniques*

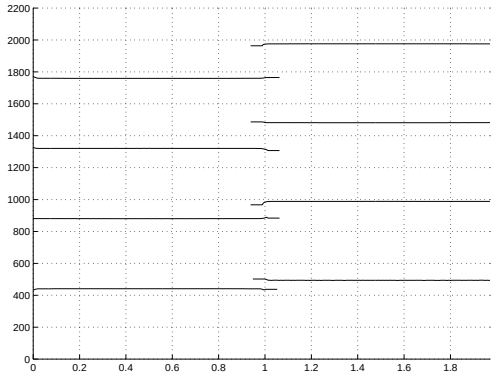


FIG. 21.3 – *Trajets des partiels pour le premier son extrait. En abscisse : le temps ; en ordonnée : la fréquence*

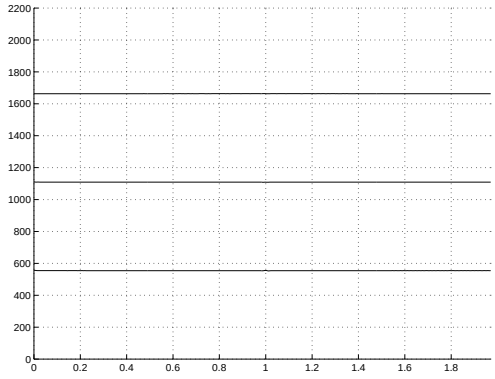


FIG. 21.4 – *Trajets des partiels pour le second son extrait. En abscisse : le temps ; en ordonnée : la fréquence*

Pour les autres exemples, nous décrivons succintement les sons :

Exemple 1 Sons mixés :

- Premier son : $f_0 = 440\ Hz$, durée : 2 secondes, nombre d'harmoniques : 12
- Second son : $f_0 = 554\ Hz$, durée : 2 secondes, nombre d'harmoniques : 4

Sons démixés :

- Premier son extrait : reconstruit avec 4 partiels

– Second son extrait : reconstruit avec 3 partiels

Exemple 3 Sons mixés :

– Premier son : $f_0 = 440\ Hz$, durée : 2 secondes, nombre d'harmoniques : 3 ; avec un vibrato ($f_{vib} = 5\ Hz$ et $A_{vib} = 20\ Hz$) et un trémolo ($f_{tré} = 5\ Hz$)
– Second son : $f_0 = 523\ Hz$, durée : 2 secondes, nombre d'harmoniques : 4

Sons démixés :

– Premier son extrait : reconstruit avec 3 partiels
– Second son extrait : reconstruit avec 4 partiels

Exemple 4 Sons mixés :

– Premier son : $f_0 = 554\ Hz$, durée : 2 secondes, nombre d'harmoniques : 5
– Second son : $f_0 = 440\ Hz$, durée : 0,7 seconde, nombre d'harmoniques : 4 ; puis $f_0 = 494\ Hz$, durée : 1,3 seconde, nombre d'harmoniques : 4

Sons démixés :

– Premier son extrait : reconstruit avec 5 partiels
– Second son extrait : reconstruit avec 4 partiels ; puis avec 4 partiels

Exemple 5 Sons mixés :

– Premier son : f_0 varie de $440\ Hz$ à $491\ Hz$ (loi polynômiale, polynôme d'ordre 2 : $f_0 = 440 + 20t + 20t^2$), l'amplitude des harmoniques varie (de 0,26 à 0,32 pour le premier, la variation étant linéaire), durée : 2 secondes, nombre d'harmoniques : 3
– Second son : $f_0 = 554\ Hz$, durée : 2 secondes, nombre d'harmoniques : 4

Sons démixés :

– Premier son extrait : reconstruit avec 3 partiels
– Second son extrait : reconstruit avec 3 partiels

Exemple 6 Sons mixés :

– Premier son : flûte : $f_0 = 277\ Hz$, de 0,4 à 2 secondes
– Second son : hautbois : $f_0 = 349\ Hz$, de 0 à 2 secondes

Sons démixés :

– Premier son extrait : reconstruit avec 7 partiels
– Second son extrait : reconstruit avec 6 partiels

Exemple 7 Sons mixés :

– Premier son : clarinette : $f_0 = 262\ Hz$, de 0,35 à 2 secondes
– Second son : hautbois : $f_0 = 349\ Hz$, de 0 à 2 secondes

Sons démixés :

– Premier son extrait : reconstruit avec 6 partiels
– Second son extrait : reconstruit avec 9 partiels

Exemple 8 Sons mixés :

- Premier son : violon : $f_0 = 554\ Hz$, de 0 à 1,3 seconde
- Second son : flûte : $f_0 = 438\ Hz$, de 0 à 0,72 seconde ; puis $f_0 = 492\ Hz$, de 0,72 à 1,26 seconde

Sons démixés :

- Premier son extrait : reconstruit avec 6 partiels
- Second son extrait : reconstruit avec 6 partiels ; puis avec 6 partiels

Faisons la remarque que les sons de clarinette, de hautbois, de violon et de flûte sont réels. Ils ont été enregistrés en salle anéchoïque. Ces sons purs ont été récupérés dans la base de sons du projet SOL (pour « StudioOnLine »). L'adresse du site est :

`http://www.ircam.fr/produits/techno/sol/intro/`

Chapitre 22

Conclusion de la quatrième partie

La **détection de la polyphonie** et la **segmentation en zones stables des sons polyphoniques** ont pour objectif d'aider à la **séparation de sources**. Quelques résultats ont été présentés dans cette partie concernant ces trois thèmes. Nous les discutons ci-dessous.

- La détection de la polyphonie en soi [1] n'existait pas.
  Les fonctions d'observation mises en place pour détecter la polyphonie peuvent être utilisées pour la *segmentation en sources* (voir la partie V, le chapitre 24) : en effet, le signal de parole est plus souvent monophonique que le signal de musique.
- La séparation de sources se faisant classiquement en suivant les partiels, c'est-à-dire en déterminant quels partiels varient ensemble dans le temps (changement de notes, vibratos, trémolos...), il est nécessaire et de parvenir à bien extraire vibratos et trémolos, et de segmenter en zones stables efficacement.
- Quelques-unes des limitations des techniques utilisées dans la littérature pour séparer en sources sont montrées.
  Les problèmes existent en très grand nombre, et certains sont particulièrement difficiles à résoudre. Nous en présentons, succintement, quatre :

  – **Problème 1** : Deux partiels, se chevauchant, sont abîmés ou même se détruisent. Il faut alors utiliser des informations relevant du timbre pour reconstituer les trajets de ces partiels. Des critères basés sur la régularité des enveloppes spectrales doivent être développés. Nous aboutissons alors à des problèmes de reconnaissance du timbre – c'est-à-dire de l'instrument –, qui ne sont pas du tout abordés dans cet exposé. Ainsi : la segmentation, l'indexation (la reconnaissance du timbre faisant partie intégrante de l'indexation), et la séparation de sources sont liées.
  – **Problème 2** : Un accord musical sonne d'autant mieux que les sons « fusionnent » (perceptivement) correctement. L'analyse d'un accord se fait en superposant la représentation fréquentielle des voix le constituant, et en observant comment se superposent les partiels des sons.
    Considérons un accord formé de deux notes. Dans le cas d'un accord d'octave, un harmonique sur deux de la note la plus aiguë se superpose à un harmonique de la note la plus grave. Nous obtenons l'accord le plus consonant. Dans le cas d'un accord de quinte, un harmonique sur trois de la note la plus aiguë se superpose à un harmonique de la note la plus grave. Il s'agit d'un des accords les plus consonants après l'accord d'octave.
    Quand deux partiels se superposent mal, disons avec un écart de 10 Hz, ils produisent des battements, c'est-à-dire une désagréable modulation d'amplitude.
    Ainsi, les accords consonants sont plutôt plus courants que les accords dissonants. Donc, le **problème 1** se rencontre très souvent. Dans cet exposé, nous n'avons pas du tout tenu compte de ces considérations, nous avons même fait plutôt l'hypothèse que les peignes harmoniques des voix mixées ne se chevauchent plutôt pas.

1. C'est-à-dire : « la détection de la polyphonie pour la détection de la polyphonie ».

– **Problème 3** : Nous avons vu les difficultés combinatoires que pose l'appariement des partiels en n groupes. Dans la section 2.2.1.2, nous présentons rapidement le fonctionnement du logiciel f_0. Nous mentionnons qu'un ensemble fixe de M fréquences fondamentales candidates est utilisé. Nous pourrions étendre la méthode au cas de la détection de deux fréquences fondamentales simultanément présentes. Si le cardinal de l'ensemble fixe de fréquences fondamentales candidates est M dans le cas monophonique, il est de M^2 quand deux voix harmoniques sont mixées, et de M^n dans un cas polyphonique plus général : n voix harmoniques mixées. Faisons deux remarques :

  * M^n croît très rapidement avec n, donc cette méthode ne résoud que partiellement les problèmes combinatoires.
  * Pour utiliser cette méthode, il faut connaître le nombre n, ou le déterminer automatiquement.

– **Problème 4** : Comme nous l'avons déjà mentionné, le problème majeur qui se pose aux techniques proposées dans cet exposé est l'extraction des trajets des partiels. Des techniques du type haute résolution (MUSIC [2] : voir par exemple [Fri90] ; etc.) ou du type de celle décrite dans [Wan94] ou [Cor99] doivent être utilisées.

Ces problèmes nous ouvrent de très nombreuses et très intéressantes perspectives.

2. Mais MUSIC ne nous donne accès ni aux amplitudes ni aux phases des partiels. Ainsi, il faudrait utiliser en parallèle une technique haute résolution, qui nous indiquerait les positions fréquentielles des partiels, et la transformée de FOURIER, qui nous donnerait les autres paramètres des partiels.

Cinquième partie

Le système complet

Chapitre 23

Introduction

Dans les parties précédentes ont été présentés les divers modules d'un système pour *segmenter en zones stables* les sons musicaux. Comme nous l'avons déjà mentionné, les techniques présentées dans cet exposé ne s'appliquent pas au cas de la parole.

Aussi, dans cette partie, dans un premier temps, nous décrivons le troisième niveau de segmentation, qui a pour but de nous indiquer quand nous sommes en présence de musique et quand nous sommes en présence de parole. Ceci fait l'objet du **deuxième chapitre** (chapitre 24) de cette partie. Il s'agit de ce que nous appelons la *segmentation en sources*.

Dans un second temps, nous décrivons le système complet de segmentation et d'indexation que nous avons défini. Il rassemble les diverses techniques présentées tout du long de cet exposé. Les dépendances qui existent entre les trois modules de segmentation, puis celles qui existent entre la *segmentation* et le module de *séparation de sources* se concrétisent par des informations échangées. Nous donnons le type d'informations qui circulent entre les quatre modules. En ce qui concerne un certain nombre d'entre ces liens, il s'agit de perspectives. Ces descriptions font l'objet du **troisième chapitre** (chapitre 25) de cette partie.

Une conclusion est donnée dans le **quatrième chapitre** (chapitre 26) de cette partie.

Chapitre 24

Segmentation en sources

24.1 Présentation du problème

24.1.1 Introduction

Le but ici est de segmenter des bandes son de films, ou des programmes radiophoniques enregistrés, en parties où nous sommes en présence de voix parlée et en parties où nous sommes en présence de musique (voix chantée et/ou musique instrumentale). Ceci notamment pour aider au codage (choix du codeur le plus adapté à la parole ou de celui le plus adapté à la musique : pour le codage perceptif, voir [Pai92], [Phi95], [BS94], [PMMS92], [Col94]...), mais aussi parce que les techniques présentées dans les parties précédentes pour *segmenter en notes ou en phones ou plus généralement en parties stables* ne s'appliquent qu'à la musique.

D'autres catégories de sons seront considérées par la suite (il s'agit de perspectives) : voix chantée seule, bruits de machines, bruits de rue...

Pour le moment, nous nous limitons à ces deux catégories :

- voix parlée
- musique, c'est-à-dire musique instrumentale ou voix chantée seule (nous classons dans cet exposé la voix chantée seule, c'est-à-dire la voix a cappella, avec les instruments de musique) ou les deux ensemble

Un programme, appelé *sources*, a été développé en C, sous UNIX. Qui accompagne ce programme, une interface graphique a été développée. Elle a pour but de nous aider à l'interprétation.

L'analyse se décompose en deux étapes :

- Des caractéristiques de base sont extraites qui essaient de mettre en évidence les propriétés spécifiques de chaque classe. Des fonctions d'observation sont calculées à partir de ces caractéristiques de base afin de faire ressortir ces différences. Elles sont présentées dans la **deuxième section** (section 24.2) de ce chapitre.
- Nous classifions les échantillons des fonctions d'observation : les méthodes de classification utilisées sont présentées dans la **troisième section** (section 24.3) de ce chapitre. La classification se fait ici après un entraînement. Remarquons que les techniques utilisées jusqu'à présent, notamment lors de la *segmentation en zones stables* telle qu'elle est présentée dans cet exposé (voir les parties précédentes), ne nécessitent pas d'entraînement.

Quelques performances obtenues avec des signaux réels sont donnés dans la **quatrième section** (section 24.4) de ce chapitre. Elles sont commentées.

Les corrélations entre les caractéristiques de base sont étudiées dans la **cinquième section** (section 24.5) de ce chapitre. La plupart des techniques décrites dans ce chapitre jusqu'à cette section 24.5 sont tirées de la littérature concernant la *segmentation en sources* (principalement [SS97]). À partir de l'étude des histogrammes des caractéristiques de base (histogrammes nécessaires au

calcul des corrélations), de nouvelles fonctions d'observation sont proposées, visant à mieux mettre en lumière les différences entre la classe « parole » et la classe « musique ».

Les corrélations entre les fonctions d'observation sont données dans la **sixième section** (section 24.6) de ce chapitre.

L'interface graphique du programme *sources* est très rapidement présentée dans la **septième section** (section 24.7) de ce chapitre. Nous indiquons dans quelle mesure elle peut nous aider à l'interprétation.

24.1.2 Les sons utilisés

Pour les performances montrées dans cet exposé, nous avons utilisé quatre fichiers sons :

- **musique1wav.sf** : 625 secondes de musique enregistrée à la radio (quelques secondes d'accordéon seul, puis les Rita MITSOUKO, puis Léo FERRÉ, puis ZAZIE)
- **parole1wav.sf** : 656 secondes de voix parlée enregistrée à la radio (informations, discussions « sages », toux, rires, poèmes scandés ; 4 voix d'hommes, 1 voix d'homme au téléphone, 4 voix de femmes)
- **musique2wav.sf** : 761 secondes de musique enregistrée à la radio (Ophélie WINTER, puis quelques secondes de jazz, puis Michel DELPECH, puis Susan VEGA : voix a cappella parfois, puis Michel SARDOU)
- **parole2wav.sf** : 599 secondes de voix parlée enregistrée à la radio (discussions moins « sages » : personnes qui parlent ensemble, rires, publicité sans musique : homme ; 5 femmes, 1 homme)

Un « jeu de sons » correspond à à peu près 20 minutes de son enregistré à la radio, dont environ 10 sont de la musique et environ 10 de la voix parlée. Chaque jeu de sons est formé de la réunion d'un fichier de voix parlée et d'un fichier de musique. Les deux jeux de sons sont : **musique1wav.sf** + **parole1wav.sf**, et **musique2wav.sf** + **parole2wav.sf**.

24.2 Les fonctions d'observation

24.2.1 Introduction

Les fonctions d'observation utilisées sont basées sur l'extraction de caractéristiques « de base ». Ces caractéristiques de base sont le flux spectral, le centroïde, le taux de passage par 0, le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage et le « spectral rolloff point ». Dans la littérature, d'autres caractéristiques de base ont été mises en place : voir les articles [Sau96], [SS97], [SBZD99], [WE99] et [ZWG99] (ce dernier article a été écrit par des gens qui viennent du traitement de la parole).

Il sera nécessaire d'en développer de nouvelles quand nous segmenterons en plus de deux classes.

Les caractéristiques de base sont calculées pour des trames temporelles larges de quelques dizaines de millisecondes. Chaque caractéristique de base nous donne une valeur (un scalaire) pour chaque trame. Ensuite, les fonctions d'observation sont calculées pour des segments d'une seconde, à partir des valeurs des caractéristiques de base obtenues pour les trames de ce segment.

24.2.2 Les caractéristiques de base

24.2.2.1 Le flux spectral

La première caractéristique de base est le flux spectral calculé avec les spectres d'amplitude. Le « flux spectral » a déjà été décrit, par exemple dans la section 2.4.3 de la partie II. Les flux spectraux calculés avec les enveloppes spectrales n'ont pas été utilisés ici.

Le flux spectral est plus grand dans les parties non voisées de la voix que dans les parties voisées. En effet, les spectres d'amplitude de deux trames successives de bruit peuvent varier énormément. Pour la musique il vaut toujours sensiblement la même chose. Tous les flux spectraux devraient réagir ainsi.

24.2.2.2 Le centroïde

Le centroïde est le centre de gravité g du spectre d'amplitude (calculé avec la FFT pour chaque trame), c'est-à-dire :

$$g = \frac{\sum_{i=0}^{t_{FFT}/2} i \left|\hat{S}(i)\right|}{\sum_{i=0}^{t_{FFT}/2} \left|\hat{S}(i)\right|}$$

où $\left|\hat{S}\right|$ est le spectre d'amplitude et t_{FFT} la taille de la FFT (à comparer avec « HFC », défini dans la section 2.5.5).

Le centroïde est plus grand pour les trames non voisées de la voix que pour les trames voisées. Dans le premier cas, en effet, le spectre d'amplitude contient plus d'énergie dans les hautes fréquences (le signal est du bruit, blanc ou coloré) que dans le second (l'énergie est concentrée principalement dans les premiers harmoniques, c'est-à-dire dans les basses fréquences). Pour la musique il vaut toujours sensiblement la même chose.

24.2.2.3 Le taux de passage par 0 (TPPZ)

Le taux de passage par 0 est le nombre de fois que le signal dans le domaine temporel franchit 0 au cours d'une trame. Les N échantillons de la trame courante s'écrivent : $[x_1 \dots x_N]$. La détection d'un passage par 0 s'effectue ainsi :

$si\ x_i == 0\ ||\ (x_{i+1} > 0\ \&\&\ x_i < 0)\ ||\ (x_{i+1} < 0\ \&\&\ x_i > 0)$
alors il y a passage par 0

Dans une trame de bruit (correspondant à une portion d'une consonne non voisée), le nombre de passage par 0 est plus grand (présence de hautes fréquences) que dans une trame de voix voisée. Pour la musique il vaut toujours sensiblement la même chose.

24.2.2.4 Le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage

Cette étude a fait l'objet d'un stage d'un mois, effectué en septembre 1998 par un étudiant roumain, Dragos SPATARU, qui est actuellement en quatrième année à l'Université Polytechnique de BUCAREST (voir [Spa98]).

Le principe de la méthode est le suivant :

- Nous calculons le cepstre $\hat{C}$ pour une trame x :

$$\hat{C} = \text{Partie\_réelle}\left(\text{FFT}^{-1}\left(\log_e\left(\underbrace{|\text{FFT}(x)|}_{|\hat{S}|}\right)\right)\right)$$

  (FFT^{-1} correspond à la FFT inverse).

  Dans [Tem96] (page 268), nous trouvons une justification de certains termes utilisés dans cet exposé. Nous donnons le passage qui nous concerne : « ...the authors of the first paper devoted to this method [cepstrum], (Bogert et al. 1963) introduced a number of new terms: the spectrum-of-the-(logarithm)-spectrum is called the *cepstrum*, the variable along the horizontal axis is designated by the word *quefrency*,... while the word filter is replaced by *lifter*... ».

  Indiquons de plus que le cepstre réel $\hat{C}$ est symétrique par rapport à la quéfrence 0. Nous les numérotons ainsi : $\left[-\frac{t_{FFT}}{2}+1 \quad \frac{t_{FFT}}{2}\right]$ (voir la page 13, où nous donnons des notations similaires pour les spectres).

- Nous ne gardons que les n coefficients de $\hat{C}$ correspondant aux plus petites quéfrences positives, le coefficient correspondant à la quéfrence 0, et les n coefficients du cepstre correspondant aux plus petites quéfrences négatives en valeur absolue. La valeur des autres est mise à 0. Ceci correspond à un « liftrage ».
  Nous obtenons $\hat{C}'$.
  Dans le cas d'un son voisé, les coefficients de numéro d'ordre élevé (entre $n+1$ et $\frac{t_{FFT}}{2}$ et entre $-n-1$ et $-\frac{t_{FFT}}{2}+1$) nous permettent de remonter, après transformée inverse, au train d'impulsions qui est émis par la source. La périodicité de ces impulsions correspond à la période fondamentale.
  Les plus petites quéfrences en valeur absolue (de numéros d'ordre $-n$ à n) nous permettent de remonter, après transformée inverse, à la réponse impulsionnelle du filtre qui filtre ce train d'impulsions (passage dans la gorge, entre les lèvres...). Il s'agit de ce que nous voulons obtenir ici.
  En ce qui nous concerne, nous gardons 99 (les 50 premiers coefficients positifs et les 49 premiers coefficients négatifs) coefficients pour $t_{FFT} = 2048$. Dans le programme *sources*, la position de ce seuil S_C est fixe. Elle a été choisie complètement empiriquement. Il s'agit d'un paramètre libre à fixer.
  Des améliorations sont envisageables : il faudrait utiliser une fenêtre de pondération avant de calculer le cepstre ; et une autre fenêtre de pondération, moins sévère que la fenêtre RECTANGULAIRE utilisée, pour sélectionner les coefficients qui servent à la reconstruction de la réponse impulsionnelle du filtre. Il s'agit de perspectives.
- La transformée inverse est calculée, mais pas jusqu'à obtenir le signal dans le domaine temporel y : nous nous arrêtons au spectre d'amplitude reconstruit après liftrage $\left|\hat{S}'\right|$:

$$\left|\hat{S}'\right| = \left|\exp\left(\mathrm{FFT}\left(\hat{C}'\right)\right)\right|$$

- La caractéristique « de base » est alors :

$$FO = \sum_{i=0}^{\frac{t_{FFT}}{2}} \left|\left|\hat{S}(i)'\right| - \left|\hat{S}(i)\right|\right|$$

Ainsi, le spectre d'amplitude d'un bruit (voix non voisée) est mieux approximé que le spectre d'amplitude d'un signal harmonique (voix voisée) : voir pour s'en convaincre l'exemple donné page 520 du livre d'OPPENHEIM et SCHAFER, « Digital Signal Processing » ([OS75]). Pour la musique il vaut toujours sensiblement la même chose.

En fait, pour un son voisé, nous obtenons une sorte d'enveloppe spectrale, mais surbaissée. Nous pouvons remonter à l'enveloppe spectrale à partir de ce spectre d'amplitude reconstruit après liftrage : voir la thèse de HALLÉ ([Hal85]).

24.2.2.5 Le « Spectral Rolloff Point »

Le « Spectral Rolloff Point » est la position p de l'échantillon fréquentiel tel que 95 % (cas général : x %, x étant un paramètre libre à fixer) de l'énergie du spectre d'amplitude soit comprise entre le premier échantillon fréquentiel (pour lequel $f = 0$) et cet échantillon fréquentiel p (voir la figure 24.1). Nous distinguons ainsi encore une fois les parties voisées (pour lesquelles l'énergie est concentrée dans les basses fréquences : p est petite) des parties non voisées (pour lesquelles l'énergie est plus uniformément répartie sur tout le spectre d'amplitude : p est plus grande). Pour la musique il vaut toujours sensiblement la même chose.

Cette caractéristique de base n'est pas implémentée dans le programme *sources*.

24.2.3 Fonctions d'observation

Ces caractéristiques « de base » sont calculées pour des trames (des portions) temporelles larges de quelques dizaines de millisecondes (disons, communément, 20, ce qui, nos signaux étant

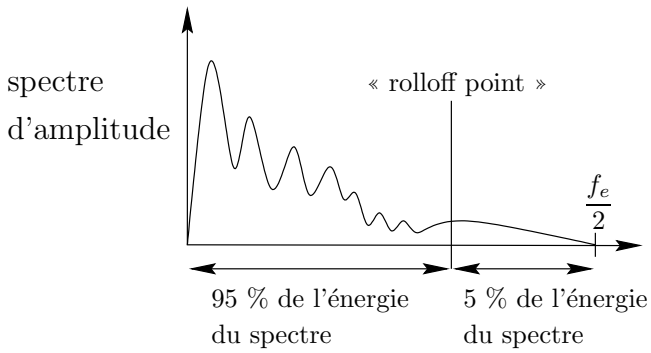


FIG. 24.1 – *Définition du « Spectral Rolloff Point »*

échantillonnés à 44100 Hz, correspond à $t_{SIG} = 882$ échantillons). Ces trames peuvent se chevaucher.

Les huit fonctions d'observation implémentées dans le programme *sources* sont :

- La MOYENNE et le LOGARITHME DÉCIMAL DE LA VARIANCE du « flux spectral » calculés sur un « segment ».
- La MOYENNE et le LOGARITHME DÉCIMAL DE LA VARIANCE du « centroïde » calculés sur un « segment ».
- La MOYENNE et le LOGARITHME DÉCIMAL DE LA VARIANCE du « taux de passage par 0 » calculés sur un « segment ».
- La MOYENNE et le LOGARITHME DÉCIMAL DE LA VARIANCE du « flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage » calculés sur un « segment ».

Dans cet exposé, les moyennes et les variances sont calculées sur des segments d'une seconde, c'est-à-dire à partir des valeurs des caractéristiques « de base » obtenues pour chaque **trame** comprise dans ce **segment**. Les segments peuvent se chevaucher.

24.2.4 Quelques interprétations

Nous avons vu que les caractéristiques « de base » nous donnent des valeurs très différentes pour les parties non voisées et pour les parties voisées du signal.

Pour la voix chantée, les voyelles (qui sont voisées) sont souvent plus longues que pour la voix parlée, et les consonnes non voisées sont très peu présentes [1]. La musique peut être considérée comme une succession de périodes de stabilité relative, malgré la présence de petits signaux (comme les percussions, ou les consonnes non voisées) qui ajoutent du bruit dans les hautes fréquences. Ces périodes de stabilité relative correspondent aux notes, ou aux phones (voix chantée), ou à la superposition de notes et de phones, ou à la superposition de plusieurs notes, et de phones... Ainsi, les caractéristiques « de base » restent relativement constantes sur un segment de musique. Au contraire, la parole [2] est plutôt une rapide succession de périodes de bruit (comme les consonnes non voisées) et de périodes de relative stabilité (comme les voyelles) : ainsi, les caractéristiques « de base » varient beaucoup sur un segment.

Comme nous pouvons le voir sur la figure 24.2, les variances sont plus grandes pour la parole que pour la musique, alors que les moyennes tendent à être plus grandes pour la musique que pour la parole.

1. Voir [DGR94] et la note de la page 3
2. Voir la note de la page 4.

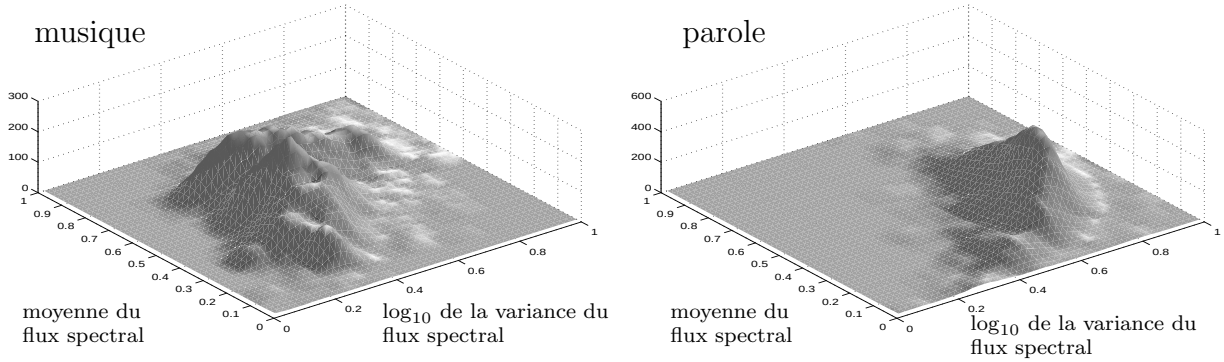


FIG. 24.2 – *Histogramme à 3 dimensions (premier jeu de sons)*

24.3 La classification

24.3.1 Introduction

Pour chacun des quatre sons utilisés pour valider la classification, les fonctions d'observation ont été calculées et rangées dans deux fichiers séparés : quatre segments sur cinq ont été utilisés pour l'entraînement, et un sur cinq a été utilisé pour les tests. Voir la section 24.4.1, où le protocole d'évaluation utilisé est explicité.

Il existe une multitude de classifieurs. Cette section ne vise pas à en donner un panorama complet : nous nous sommes focalisés plus spécialement sur trois d'entre eux, sans même chercher à approfondir l'étude de leur fonctionnement. Ce que nous faisons dans cette section en matière de classification est de toute façon « simple » : l'existant est plus compliqué, avancé.

24.3.2 Le mélange de gaussiennes (MG)

24.3.2.1 Cas général

Considérons une classe. Nous modélisons les données de cette classe par une somme de P densités de probabilité $p(x|\theta_\nu)$ multi-dimensionnelles gaussiennes. La classe considérée est formée de P composantes. La somme des densités de probabilité $p(x|\theta_\nu)$ de chaque composante nous donne la densité de probabilité $P(x|\Theta)$ du mélange pour la classe considérée :

$$P(x|\Theta) = \sum_{\nu=1}^{P} p(x|\theta_\nu)\pi_\nu$$

avec : x le vecteur des données (de taille N), π_ν les probabilités a priori de chaque composante, $\Theta = (\theta_1 \ \ldots \ \theta_P)$ et $\theta_\nu = (y_\nu, \Sigma_\nu)$, où y_ν est le vecteur des moyennes (sa taille est le nombre n de dimensions, c'est-à-dire le nombre de fonctions d'observation prises en compte) pour la composante ν et Σ_ν est la matrice de covariance (de taille $n \times n$) pour la même composante.

Il s'agit de déterminer les paramètres Θ du mélange.

La méthode utilisée ici est la méthode du maximum de vraisemblance. Plus communément, nous maximisons suivant Θ ceci :

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \log_e P(x_i|\Theta) = \sum_{i=1}^{N} \log_e \left(\sum_{\nu=1}^{P} p(x_i|\theta_\nu)\pi_\nu \right)$$

Cette maximisation se fait en utilisant l'algorithme EM (« Expectation – Maximization ») :

- **Étape E.** Les taux d'assignement $M_{i\alpha}$ [3] de chaque point i à chaque composante α sont estimés (chaque x_i est lui-même un vecteur de taille n) ainsi :

$$\begin{aligned} M_{i\alpha} &= \frac{p\left(x_i \mid y_\alpha, \Sigma_\alpha\right)\pi_\alpha}{\sum_{\nu=1}^{P} p\left(x_i \mid y_\nu, \Sigma_\nu\right)\pi_\nu} \\ &= \frac{\left|\Sigma_\alpha\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x_i - y_\alpha\right)^T \left(\Sigma_\alpha\right)^{-1}\left(x_i - y_\alpha\right)\right)}{\sum_{\nu=1}^{P}\left|\Sigma_\nu\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x_i - y_\nu\right)^T \left(\Sigma_\nu\right)^{-1}\left(x_i - y_\nu\right)\right)} \end{aligned}$$

- **Étape M.** Les paramètres du mélange sont réévalués ainsi :

$$y_\alpha = \frac{\sum_{i=1}^{N} M_{i\alpha} x_i}{\sum_{i=1}^{N} M_{i\alpha}}$$

$$\Sigma_\alpha = \frac{1}{\sum_{i=1}^{N} M_{i\alpha}} \sum_{i=1}^{N} M_{i\alpha}(x_i - y_\alpha)(x_i - y_\alpha)^T$$

$$\pi_\alpha = \frac{1}{N}\sum_{i=1}^{N} M_{i\alpha}$$

Faisons la remarque que ^{T} représente la transposition.

Les paramètres sont dans un premier temps initialisés, puis les deux étapes E et M sont répétées jusqu'à la convergence.

Supposons que nous considérions deux classes. La première est composée de $P^{(1)}$ composantes et la seconde de $P^{(2)}$. L'entraînement consiste à estimer les **paramètres** $\Theta^{(1)} = \left(\theta_1^{(1)} \ \ldots \ \theta_{P^{(1)}}^{(1)}\right)$, $\pi^{(1)} = \left(\pi_1^{(1)} \ \ldots \ \pi_{P^{(1)}}^{(1)}\right)$ et $\Theta^{(2)} = \left(\theta_1^{(2)} \ \ldots \ \theta_{P^{(2)}}^{(2)}\right)$, $\pi^{(2)} = \left(\pi_1^{(2)} \ \ldots \ \pi_{P^{(2)}}^{(2)}\right)$ des deux mélanges.

24.3.2.2 Notre cas

Dans notre cas, chacune des deux classes a été modélisée par une seule gaussienne multi-dimensionnelle. Les paramètres (vecteur moyenne et matrice de covariance) de chacune de ces deux gaussiennes peuvent être estimés indépendamment de ceux de l'autre gaussienne, et nous n'avons pas besoin d'utiliser l'agorithme « Expectation – Maximization ».

L'algorithme EM n'est nécessaire qu'au cas où nous modélisons chacune des deux classes par plusieurs gaussiennes. Ceci devra être effectué pour la classe « musique ». Il s'agit d'une perspective.

Pour classifier un nouveau point x_j (un point permettant de mesurer les performances du **test 1** ou du **test 2** : voir la section 24.4.1), il faut prendre en compte les probabilités a priori $\Pi^{(1)}$ et $\Pi^{(2)}$ de chaque classe, c'est-à-dire les nombres $N^{(1)}$ et $N^{(2)}$ de points de chaque classe utilisés pour entraîner. Nous avons : $\Pi^{(1)} = \dfrac{N^{(1)}}{N^{(1)} + N^{(2)}}$ et $\Pi^{(2)} = \dfrac{N^{(2)}}{N^{(1)} + N^{(2)}}$. Et, finalement, chaque classe étant modélisée par une seule gaussienne, le point x_j appartient à la classe 1 si (il s'agit du critère de Bayes : voir [Cha96] page 139) :

$$\begin{aligned} &\frac{\Pi^{(1)}}{\left|\Sigma^{(1)}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x_j - y^{(1)}\right)^T \Sigma^{(1)-1}\left(x_j - y^{(1)}\right)\right) > \\ &\frac{\Pi^{(2)}}{\left|\Sigma^{(2)}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x_j - y^{(2)}\right)^T \Sigma^{(2)-1}\left(x_j - y^{(2)}\right)\right) \end{aligned}$$

Dans notre cas, pour chaque jeu de sons, nous avons presque autant de points de parole que de points de musique, donc $\Pi^{(1)} \simeq \Pi^{(2)} \simeq 0{,}5$.

3. Voir le chapitre 4.3.4 de la partie II.

24.3.3 Les k plus proches voisins ($k$ppv)

Pour chaque point à classifier, nous cherchons ses k plus proches voisins (la distance mise en place est la distance euclidienne : d'autres sont possibles) dans la base d'entraînement. Si la majorité de ces k points appartient à la classe i, il est décidé que le point à classifier appartient à la classe i. Nous prenons k impair. k est un paramètre libre à fixer.

Cette fois, nous n'obtenons pas de **paramètres** décrivant (modélisant, compressant) la base de données d'entraînement, paramètres que nous supposons applicables ensuite pour la base de test (**test 1** : voir la section 24.4.1), puis pour tout nouveau son à classifier (**test 2** : voir la section 24.4.1) : les données d'entraînement constituent un dictionnaire.

24.3.4 Les réseaux de neurones (RN)

Nous avons utilisé un logiciel libre de droits (« freeware ») fourni par l'Université de STUTTGART, logiciel qui s'appelle SNNS (pour « Stuttgart Neural Network Simulator ») et qui nous permet de dessiner le réseau de neurones voulu : choix du nombre de couches, du nombre d'unités (ou de neurones) par couches, de la méthode d'entraînement, du nombre d'itérations pour l'entraînement... ; de visualiser le réseau ; de visualiser l'erreur obtenue à chaque itération sur les données d'entraînement et sur les données de test ; de choisir les fichiers d'entraînement, de test (**test 1** : voir la section 24.4.1) et de « cross validation » (**test 2** : voir la section 24.4.1) ; etc.

Le logiciel et sa documentation peuvent être obtenus ici :

`ftp.informatik.uni-stuttgart.de(129.69.211.2)`

dans le répertoire : `/pub/SNNS`

Voir aussi, pour une documentation en ligne :

`http//www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html`

Dans cet exposé, nous avons travaillé avec un réseau de neurones composé de trois couches, complètement connecté. C'est-à-dire que tous les neurones d'une couche sont reliés à tous les neurones de la couche précédente.

La couche cachée comprend 6 unités quand nous utilisons seulement deux fonctions d'observation, et 8 quand nous utilisons les six premières fonctions d'observation.

Nous utilisons la méthode d'entraînement par défaut, c'est-à-dire la « backpropagation » (voir la page 21 de la documentation de SNNS pour plus de détails).

Des tests plus poussés doivent être effectués. L'un des problèmes est que le choix des paramètres (taille du réseau...) demeure en grande partie empirique.

Les performances des réseaux de neurones dépendent beaucoup du processus d'entraînement des réseaux de neurones utilisé. Dans un premier temps présentons un échantillon de la classe « musique », puis un échantillon de la classe « parole », et ainsi de suite, en respectant l'ordre temporel : ainsi, deux segments de musique consécutifs dans le temps ont pour numéro d'ordre de présentation au réseau de neurones i et $i+2$. Dans un second temps, présentons un échantillon de musique, puis un échantillon de parole, et ainsi de suite, mais en ne respectant pas l'ordre temporel : les échantillons sont choisis aléatoirement. Ainsi, les résultats sont énormément améliorés et la convergence est accélérée. Nous passons de 9,1 % d'erreur pour le **test 2** (voir la section 24.4.1) quand nous respectons l'ordre temporel des segments à 4,7 % (voir le tableau 24.2) quand nous le rompons. Il s'agit d'une des règles empiriques utiles pour les réseaux de neurones : il faut présenter les points d'entraînement aussi aléatoirement que possible. Le simple entrelacement (un échantillon d'une classe puis un échantillon d'une autre) n'est pas suffisant !

Entraînons M fois le même réseau de neurones avec les mêmes fichiers d'entraînement. Les échantillons d'entraînement sont présentés dans un ordre aléatoire. Alors, le problème qui apparaît est que le réseau de neurones est entraîné légèrement différemment d'un entraînement à l'autre, et donne des performances légèrement différentes. Pour le mélange de gaussiennes, les paramètres obtenus sont M les mêmes, puisque les fichiers d'entraînement ne changent pas : donc nous obtenons aussi M fois les mêmes perforamances ; la même remarque peut être faite pour les k plus proches voisins.

24.3.5 L'entraînement « problématique » du programme *sources*

Dans le programme *sources*, nous avons accès aux tailles des fenêtres d'analyse (ou trames), à la taille des segments, à la fenêtre de pondération utilisée, etc. Mais modifier ces paramètres libres conduit à modifier, plus ou moins radicalement, l'allure des nuages de points (les valeurs des fonctions d'observation), et donc à rendre la classification à partir de fichiers d'entraînement qui seraient fournis par nous aux utilisateurs du programme *sources*, obtenus eux avec les valeurs par défaut des paramètres libres, très difficile ou absurde : en tout cas, à aboutir à de mauvais résultats.

Donc l'utilisateur doit faire ses propres fichiers d'entraînement, avec ses sons, ce qui est pénible pour lui puisqu'il a d'abord à étiqueter à la main ces fichiers sons. Ainsi, le caractère « automatisé » du programme *sources* se trouve en partie compromis. Or, nous voulons que le programme *sources* soit le plus automatique possible, au moins du point de vue de l'utilisateur : notamment, celui-ci ne doit pas avoir à s'occuper de l'entraînement. Pour lui, tout doit être transparent : il donne un son et il récupère un fichier de segmentation.

Ainsi, nous donnons des fichiers sons représentatifs de chaque classe, pour lesquels les fonctions d'observation sont calculées de nouveau à chaque fois qu'un nouveau son à segmenter est présenté ou à chaque fois qu'un paramètre libre est modifié. Le problème est que, notamment pour la musique, si nous voulons un fichier le plus représentatif possible, il risque d'être gros (ceci se traduit par le fait que la classe « musique » est plus étendue que la classe « parole »), d'où des problèmes de temps de calcul.

24.4 Les performances de la classification

24.4.1 Protocole d'évaluation

Trois types d'erreurs sont considérés :

- Les erreurs d'**entraînement** correspondent aux erreurs faites sur les données d'entraînement en utilisant les paramètres obtenus lors de cet entraînement. Pour le mélange de gaussiennes (MG), ces paramètres sont les vecteurs de moyennes et les matrices de covariance. Pour les réseaux de neurones, ce sont les poids sur les entrées des unités. Bien sûr, il n'y en a pas pour les k plus proches voisins.
- Les erreurs de **test 1** correspondent aux erreurs faites sur les données de test du jeu de sons i en utilisant les paramètres obtenus lors de l'entraînement avec les données d'entraînement du jeu de sons i. Nous disons alors que les données d'entraînement et de test sont homogènes : les données utilisées pour tester ont été calculées à partir des mêmes fichiers sonores que les données qui ont servi à entraîner.
- Les erreurs de **test 2** correspondent aux erreurs faites sur les données de test du jeu de sons i en utilisant les paramètres obtenus lors de l'entraînement avec les données d'entraînement du jeu de sons j ($j \neq i$). Nous disons alors que les données d'entraînement et de test sont hétérogènes : les données utilisées pour tester n'ont pas été calculées à partir des mêmes fichiers sonores que les données qui ont servi à entraîner. En fait, ce sont ces erreurs qui sont le plus intéressantes, car elles correspondent au fonctionnement du programme *sources* du point de vue d'un utilisateur quelconque : une boite noire à laquelle il est donné un son inconnu, boite noire qui segmente ce son.

Nous n'avons pas fait la distinction entre les erreurs qui correspondent à : un segment de parole est pris pour un segment de musique, et celles qui correspondent à : un segment de musique est pris pour un segment de parole.

24.4.2 Performances

24.4.2.1 Du programme *sources*

Quelques performances des méthodes de classification sont données dans les tableaux 24.1 et 24.2. Pour chaque jeu de sons n, nous obtenons un pourcentage d'erreur e_n. Dans chaque case des

tableaux nous donnons la moyenne $\bar{e} = \frac{1}{2}\sum_{n=1}^{2} e_n$ obtenue à partir des mesures faites sur les deux jeux de sons.

	entraînement	**test 1**	**test 2**
flux : *moyenne* et $\log_{10}(variance)$	4,7 %	5,4 %	9,1 %
centroïde : *moyenne* et $\log_{10}(variance)$	7,3 %	8,7 %	12,7 %
TPPZ : *moyenne* et $\log_{10}(variance)$	7,4 %	8,3 %	13,2 %
cepstre : *moyenne* et $\log_{10}(variance)$	8,2 %	10,6 %	13,8 %

TAB. 24.1 – *Pourcentage de segments mal classés pour chaque couple de fonctions d'observation en utilisant le classifieur kppv ($k = 7$)*

	entraînement	**test 1**	**test 2**
MG	4,0 %	3,6 %	9,3 %
$k$ppv ($k = 7$)	1,3 %	1,5 %	5,9 %
RN	1,9 %	1,7 %	4,7 %

TAB. 24.2 – *Pourcentage de segments mal classés en utilisant les six premières fonctions d'observation ensemble*

24.4.2.2 Avec le « Spectral Rolloff Point »

Cette caractéristique de base est disponible seulement sous MATLAB.

Avec le classifieur des k plus proches voisins ($k = 7$), nous obtenons 10 % d'erreur à l'**entraînement** pour le premier jeu de sons.

En fait, toutes les techniques qui permettent de mesurer le voisement donnent des résultats similaires. Ainsi, pour le « voisement deuxième forme » (voir les sections 2.4.1 et 2.4.2), avec le classifieur des k plus proches voisins ($k = 7$) et le premier jeu de sons, nous obtenons 11 % d'erreur à l'**entraînement**.

24.4.3 Conclusion

Nous constatons qu'utiliser plus de fonctions d'observation tend à améliorer les résultats, notablement même dans certains cas. Ceci pour les k plus proches voisins ($k$ppv) et les réseaux de neurones (RN) pour les trois types d'erreur, et pour le mélange de gaussiennes (MG) pour les deux premiers types d'erreur.

Par contre, en utilisant les six premières fonctions d'observation et en classifiant avec le mélange de gaussiennes (MG), nous nous rendons compte que les résultats sont mauvais. Ceci est dû au fait que les deux classes pour le deuxième jeu de sons sont beaucoup plus proches l'une de l'autre que pour le premier jeu de sons (ceci encore plus vrai pour le centroïde et le taux de passage par 0 que pour le flux spectral). En fait, les segments de parole pour les quatre jeux de sons sont situés au même endroit ; ce sont les segments de musique qui se sont approchés de ceux de la parole pour le deuxième jeu de sons. Les segments de musique s'étalent sur une plus grande zone que les segments de parole. Ceci est dû à la grande variété de genres qui existe pour la musique.

Une solution serait de modéliser chacune des classes en utilisant plusieurs gaussiennes par classe (notamment pour la classe « musique »), et d'utiliser beaucoup plus de segments de musique pour entraîner les classifieurs.

Nous pouvons dire que la segmentation en deux classes : parole/musique est efficace. Il faudra découper la classe musique en plusieurs sous-classes : musique instrumentale/voix chantée/bruits.

24.5 Corrélations entre les caractéristiques de base

24.5.1 Densités de probabilité

Les densités de probabilité des caractéristiques de base ont été estimées pour les deux fichiers d'environ dix minutes de musique et pour les deux fichiers d'environ dix minutes de parole. Les densités de probabilité estimées sont les histogrammes normalisés. Nous donnons les densités de probabilité obtenues pour chaque fichier (traits pleins). Et nous donnons les densités de probabilité obtenues pour les deux fichiers représentant une source pris ensemble (traits interrompus).

Pour le flux spectral, voir les figures 24.3 et 24.4. Les paramètres libres sont :

- la taille des fenêtres d'analyse (trames) : nous prenons 0,04 seconde, soit $t_{SIG} = 1764$
- la taille de la FFT : nous prenons $t_{FFT} = 4096$
- le pas d'avancement d'une trame à la suivante : nous prenons 0,01 seconde
- le décalage entre les deux fenêtres d'analyse : nous prenons 0,005 seconde
- la fenêtre de pondération utilisée : nous prenons BLACKMAN

Ainsi, avec le pas d'avancement choisi, les densités de probabilité obtenues pour chaque fichier sont chacune estimées à partir d'environ 60000 points.

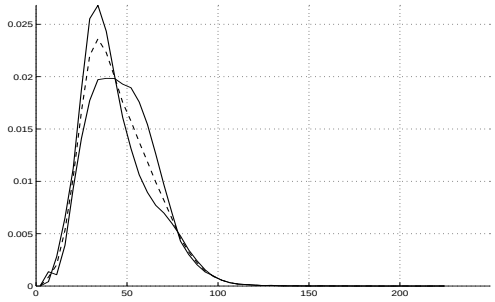


FIG. 24.3 – *Densités de probabilité (ddp) pour le flux spectral. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de musique. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

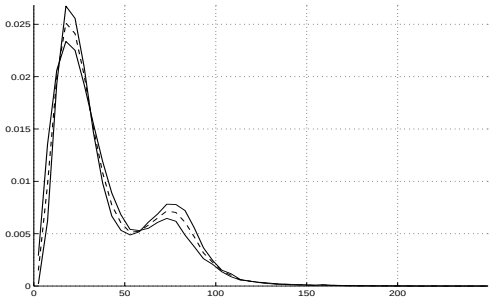


FIG. 24.4 – *Densités de probabilité (ddp) pour le flux spectral. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de parole. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

Pour le centroïde, voir les figures 24.5 et 24.6. Les paramètres libres sont les mêmes que pour le flux spectral. Pour le taux de passage par 0, voir les figures 24.7 et 24.8. Les paramètres libres sont toujours les mêmes. Pour le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage, voir les figures 24.9 et 24.10. Les paramètres libres sont toujours les mêmes.

Voir les remarques faites précédemment (section 24.2.4) et ci-dessous (section 24.5.2) à propos des différences entre les moyennes et les variances des caractéristiques de base pour la parole et la musique. Elles apparaissent ici clairement.

24.5.2 De nouvelles fonctions d'observation

Les densités de probabilité des caractéristiques de base nous indiquent que la classe « parole » est plutôt formée de deux modes. Tout du moins, ceci est nettement visible en ce qui concerne le flux spectral, le centroïde et le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage. Celui des modes dont la moyenne est la plus petite correspond aux zones voisées du signal de parole, alors que l'autre correspond à ses zones non voisées. Au contraire, la classe « musique » n'est plutôt formée que d'un seul mode. Ainsi, il s'agirait de mettre en évidence cette différence. La moyenne et la variance ne sont pas forcément de bonnes idées. Aussi nous avons développé d'autres mesures, dont nous avons testé les performances.

- Il y a plus de trames au-dessous de la moyenne pour la parole que pour la musique. La fonction d'observation est donc le nombre de trames au-dessous de la moyenne sur un segment.

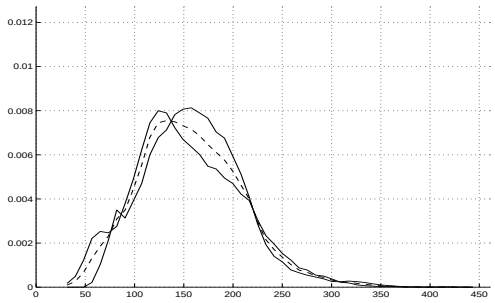


FIG. 24.5 – *Densités de probabilité pour le centroïde. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les densités de probabilité obtenues pour chaque fichier de musique. La courbe en trait interrompu est la densité de probabilité obtenue quand les 2 fichiers sont considérés ensemble*

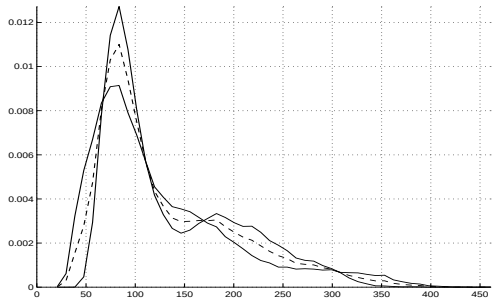


FIG. 24.6 – *Densités de probabilité (ddp) pour le centroïde. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de parole. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

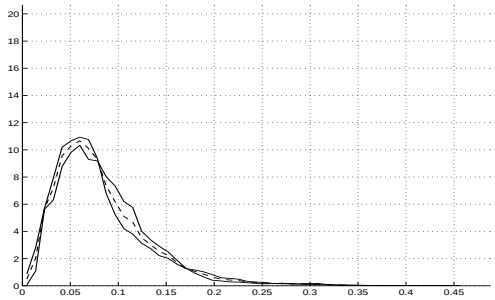


FIG. 24.7 – *Densités de probabilité (ddp) pour le taux de passage par 0. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de musique. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

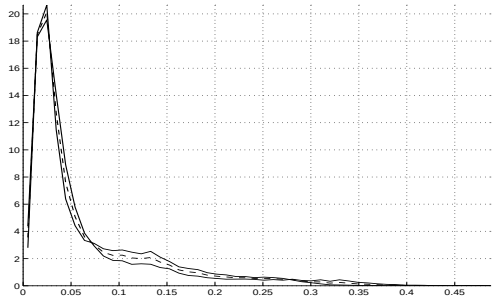


FIG. 24.8 – *Densités de probabilité (ddp) pour le taux de passage par 0. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de parole. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

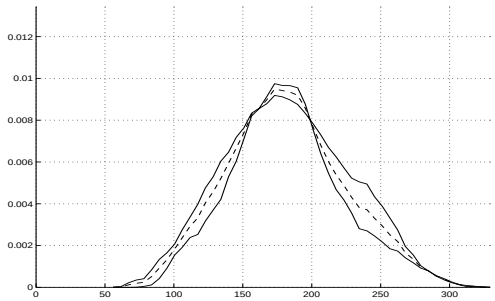


FIG. 24.9 – *Densités de probabilité (ddp) pour le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de musique. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

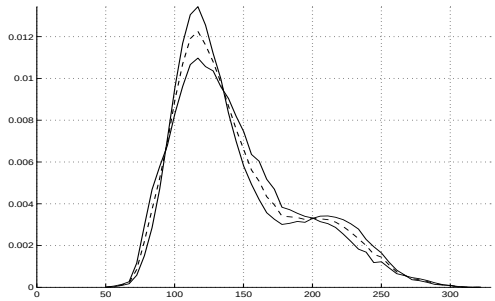


FIG. 24.10 – *Densités de probabilité (ddp) pour le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage. En abscisse x ; en ordonnée : $p_x(x)$. Les 2 courbes en trait plein sont les ddp obtenues pour chaque fichier de parole. La courbe en trait interrompu est la ddp obtenue quand les 2 fichiers sont considérés ensemble*

fonction d'observation	pourcentage d'erreur
moyenne	37,4 %
$\log_{10}$ de la variance	10,5 %
nombre de trames au-dessous de la moyenne	26,5 %
position $x(i)$ du maximum de l'histogramme	18,2 %
produit $x(i)$ par $\log_{10}$ de la variance du mode détecté	26,8 %

TAB. 24.3 – *Performances des fonctions d'observation avec le flux spectral*

- La position x du maximum de l'histogramme est plus petite pour la parole que pour la musique. Si cette position a lieu pour la case i de l'histogramme, $x = x_i$. La fonction d'observation est cette position.
- La position x du maximum de l'histogramme est plus petite pour la parole que pour la musique, et la variance du mode détecté ainsi est plus petite pour la parole que pour la musique. Si cette position a lieu pour la case i de l'histogramme, le mode détecté rassemble toutes les trames comprises dans les cases 1 à $2i-1$ de l'histogramme. La fonction d'observation est le produit de cette position et du logarithme décimal de la variance du mode détecté.

Pour évaluer les performances de ces mesures, nous utilisons le classifieur des k plus proches voisins. Seuls les résultats pour le **test 1** (voir la section 24.4.1) sont donnés. La caractéristique de base utilisée est le flux spectral. Pour chaque jeu de sons n, nous obtenons un pourcentage d'erreur e_n. La moyenne $\bar{e} = \frac{1}{2}\sum_{i=n}^{2} e_n$ est obtenue à partir des mesures faites sur les deux jeux de sons.

Les nouvelles mesures sont moins efficaces que le logarithme décimal de la variance, mais elles sont plus efficaces que la moyenne.

24.5.3 Mesures de la corrélation entre deux variables aléatoires

Trois mesures de la corrélation nous avaient servi lors de l'étude sur la corrélation entre les fonctions d'observation utilisées pour la *segmentation en zones stables* (voir la partie II, chapitre 7). Voir la section 7.2 pour avoir la définition du coefficient de corrélation. Voir la section 7.3 pour avoir la définition de l'information mutuelle. Voir la section 7.4 pour avoir la définition du test du χ^2. Nous avons utilisé ces trois mesures pour étudier la corrélation entre les caractéristiques de base, puis entre les fonctions d'observation (voir la section 24.6). Nous ne donnons que les résultats obtenus avec le coefficient de corrélation. Les deux autres mesures sont en accord avec le coefficient de corrélation.

24.5.4 Corrélations entre les caractéristiques de base

Les corrélations sont présentées dans les tableaux 24.4 et 24.5. ca_1 est mis pour le flux spectral, ca_2 pour le centroïde, ca_3 pour le taux de passage par 0 et ca_4 pour le flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage. Pour chaque fichier son de 10 minutes, numéroté i, un coefficient de corrélation c_i est obtenu. La moyenne est donnée.

	ca_1	ca_2	ca_3	ca_4
ca_1	1,000	0,758	0,796	*0,717*
ca_2		1,000	**0,857**	*0,715*
ca_3			1,000	0,818
ca_4				1,000

TAB. 24.4 – *Coefficients de corrélation pour la musique avec les 2 fichiers de 10 minutes*

Quelques remarques sont données ci-dessous :

- Les caractéristiques de base sont plutôt corrélées.

	ca_1	ca_2	ca_3	ca_4
ca_1	1,000	0,816	*0,794*	*0,789*
ca_2		1,000	**0,898**	0,801
ca_3			1,000	0,843
ca_4				1,000

TAB. 24.5 – *Coefficients de corrélation pour la parole avec les 2 fichiers de 10 minutes*

- Considérons le plus grand coefficient de corrélation obtenu pour chacune des deux classes. Nous pouvons dire que le centroïde (ca_2) et le taux de passage par 0 (ca_3) sont corrélés, ceci aussi bien pour la musique que pour la parole.
- ca_1 et ca_4, les deux flux, sont décorrélés, et ce pour les deux classes. Ceci est normal. En effet, le flux spectral calculé avec les spectres d'amplitude est petit dans les zones voisées et grand dans les zones non voisées, alors que l'inverse est obtenu pour le flux calculé à partir d'un spectre d'amplitude et d'un spectre d'amplitude reconstruit après liftrage.
- Considérons les plus petits coefficients de corrélation obtenus pour chacune des deux classes (non compris la corrélation entre ca_1 et ca_4). Pour la musique, les deux fonctions d'observation les moins corrélées sont ca_2 (centroïde) et ca_4 (flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage). Pour la parole, les deux fonctions d'observation les moins corrélées sont ca_1 (flux spectral) et ca_3 (taux de passage par 0).

24.6 Corrélations entre les fonctions d'observation

Bien sûr, les nuages de points obtenus avec les fonctions d'observation ne sont pas aussi compliqués que certains de ceux (cercle, sinus...) donnés dans l'annexe H. Ils ont plutôt la forme de « haricots » (arcs de cercle courts) : voir les figures 24.2, 24.11 et 24.12. Aussi, le coefficient de corrélation suffit, et nous ne donnons que les résultats obtenus avec lui.

Pour le premier fichier de musique nous obtenons le tableau 24.6.

- fo_1 : moyenne du flux spectral
- fo_2 : logarithme décimal de la variance du flux spectral
- fo_3 : moyenne du centroïde
- fo_4 : logarithme décimal de la variance du centroïde
- fo_5 : moyenne du taux de passage par zéro
- fo_6 : logarithme décimal de la variance du taux de passage par zéro
- fo_7 : moyenne du flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage
- fo_8 : logarithme décimal de la variance du flux entre le spectre d'amplitude et le spectre d'amplitude reconstruit après liftrage

	fo_1	fo_2	fo_3	fo_4	fo_5	fo_6	fo_7	fo_8
fo_1	1,000	0,601	0,835	0,461	0,909	0,688	0,862	0,649
fo_2		1,000	0,751	0,804	0,644	0,885	0,493	0,843
fo_3			1,000	0,609	0,912	0,804	0,779	0,728
fo_4				1,000	0,460	0,867	0,294	0,716
fo_5					1,000	0,737	0,893	0,657
fo_6						1,000	0,585	0,858
fo_7							1,000	0,532
fo_8								1,000

TAB. 24.6 – *Corrélations entre les fonctions d'observation*

Nous constatons que les variances sont plutôt corrélées entre elles et que les moyennes sont plutôt corrélées entre elles, mais que les variances et les moyennes sont plutôt décorrélées. Nous obtenons la même chose pour la parole.

24.7 Quelques interprétations

Nous pouvons, dans le cas de la musique, assez bien séparer chaque chanson dans l'espace des fonctions d'observation. Ainsi, avec la moyenne et le logarithme décimal de la variance du flux spectral, nous obtenons, pour les segments de musique du premier jeu de sons, la figure 24.11. Le programme utilisé ici est la version avec interface graphique du programme *sources*. Les courbes fermées représentant chacune l'un des extraits présents dans le son ont été ajoutées à la main. Ce sont des contours subjectifs, tracés après avoir écouté les segments [4] un à un.

L'extrait d'accordéon est purement monophonique, et parfois les notes sont longues. Ainsi, le flux spectral est petit, ainsi que sa variance.

L'extrait des Rita Mitsouko comporte plus de percussions que celui de Zazie, qui lui-même en comporte plus que l'extrait de Léo Ferré.

Pour le second jeu de sons (musique et parole), nous obtenons la figure 24.12. Les points les plus clairs (rouges) correspondent à la musique et les plus sombres (bleus) à la parole.

L'extrait de Susan Vega comporte des parties de voix chantée a cappella. Les segments de cet extrait de musique sont très proches des segments de parole.

Pour les deux jeux de sons, dans le cas de la parole, les points les plus à gauche et en bas, correspondent souvent à de longs « euh ! ».

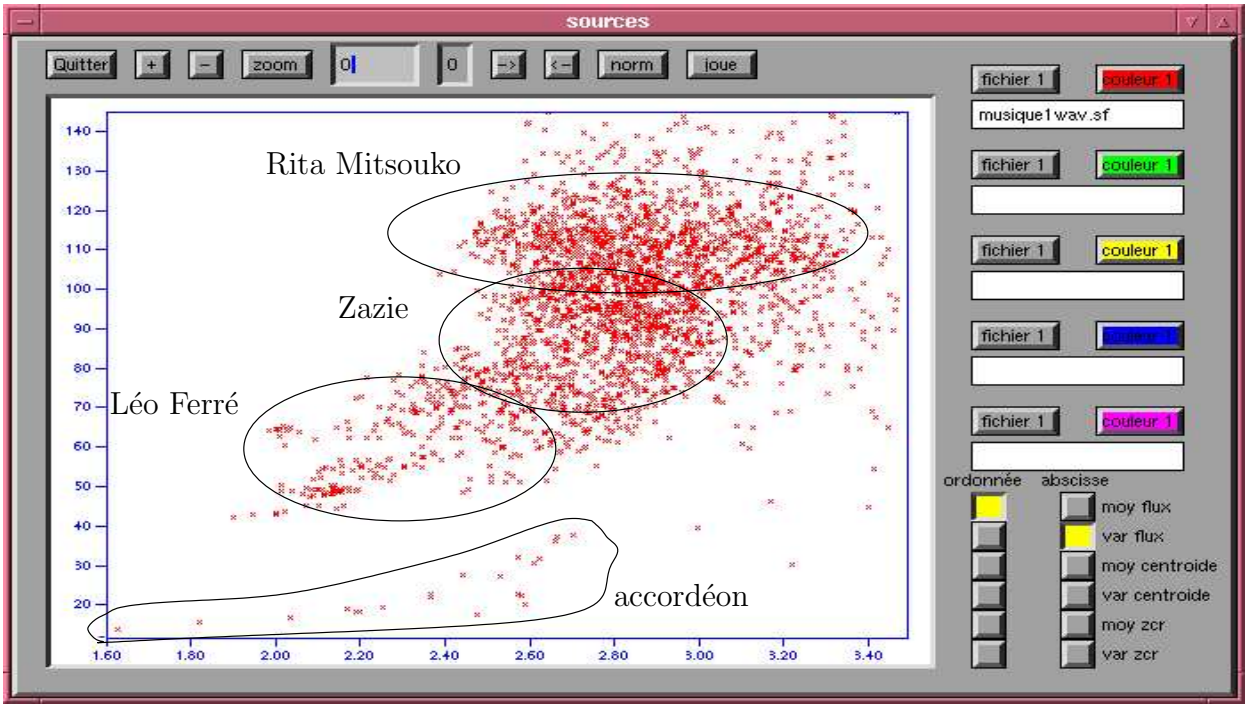


FIG. 24.11 – *L'interface graphique du programme « sources ». Les segments pour le fichier de musique du premier jeu de sons sont représentés. En ordonnée : la moyenne du flux spectral ; en abscisse : le* $\log_{10}$ *de la variance du flux spectral*

4. Chaque segment d'une seconde est représenté par une étoile. En cliquant sur une étoile, la seconde de son correspondante est jouée.

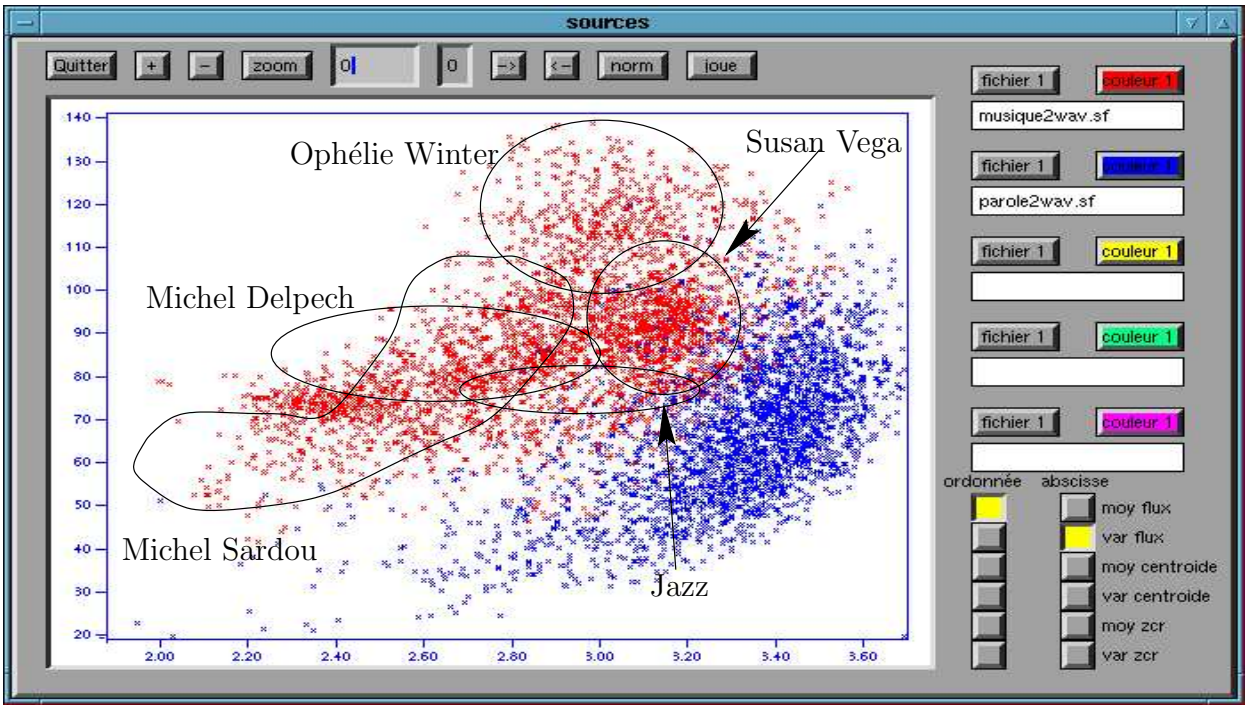


FIG. 24.12 – *L'interface graphique du programme « sources ». Les segments du deuxième jeu de sons sont représentés. En ordonnée : la moyenne du flux spectral ; en abscisse : le* $\log_{10}$ *de la variance du flux spectral*

Chapitre 25

Dépendances entre les trois niveaux de segmentation et la séparation de sources

25.1 Dépendances entre les niveaux de segmentation

Nous avons vu qu'en fait l'analyse est hiérarchisée ainsi : d'abord, la *segmentation en sources* (chapitre 24 de cette partie V) est effectuée, puis la *segmentation en caractéristiques* (partie III), et enfin la *segmentation en notes ou en phones, ou plus généralement en zones stables* (partie II principalement). Les informations entre niveaux circulent du premier de ces niveaux de segmentation au troisième. Quelques dépendances, et un schéma général pour la segmentation, sont donnés sur la figure 25.1.

Nous donnons dans les sections qui suivent quelques-uns des cas possibles. Les cas mentionnés ici ne représentent pas de façon exhaustive tous les cas possibles. L'une des perspectives est de les recenser.

Dans la conclusion à la deuxième partie (chapitre 9, page 83), nous avions explicité une série de questions que le système a à se poser tout du long de son analyse. Nous reportons pour chacun des cas traités ci-dessous lesquelles de ces questions sont résolues par le système.

25.1.1 Le cas du vibrato

Pour le vibrato, si nous nous plaçons délibérément dans le cas d'un son monophonique (q_2) et harmonique (q_4), l'algorithme est le suivant.

Quand de la parole (q_1) est identifiée au premier niveau de segmentation, il n'y a pas besoin d'essayer de détecter le vibrato. Donc, la variable VIBRATO est fixée à OFF. Et, quand les fonctions d'observation sont extraites pour le troisième niveau de segmentation, l'outil « suppression du vibrato » n'est pas utilisé.

Quand de la musique (q_1) est identifiée, l'outil « détection du vibrato » (q_5), inclus dans le niveau de *segmentation en caractéristiques*, est utilisé. Si un vibrato est détecté, la variable VIBRATO est fixée à ON. Et, dans ce cas, le vibrato est supprimé sur le trajet de f_0 (voir le chapitre 15, dans la partie III) et les quatre fonctions d'observation basées sur ce trajet (dérivées, analyse statistique, rupture de modèles) sont calculées à partir du nouveau trajet de f_0.

25.1.2 Le cas de l'inharmonicité

Pour l'indice d'inharmonicité, si nous nous plaçons délibérément dans le cas d'un son monophonique (q_2), l'algorithme est le suivant.

Un signal de castagnettes, par exemple, est toujours inharmonique (q_4). Donc, la fonction d'observation basée sur l'indice d'inharmonicité ne présente pas de pics fins et grands quand les transitions ont lieu (dans ce cas, les transitions sont des transitoires d'énergie). Cette fonction

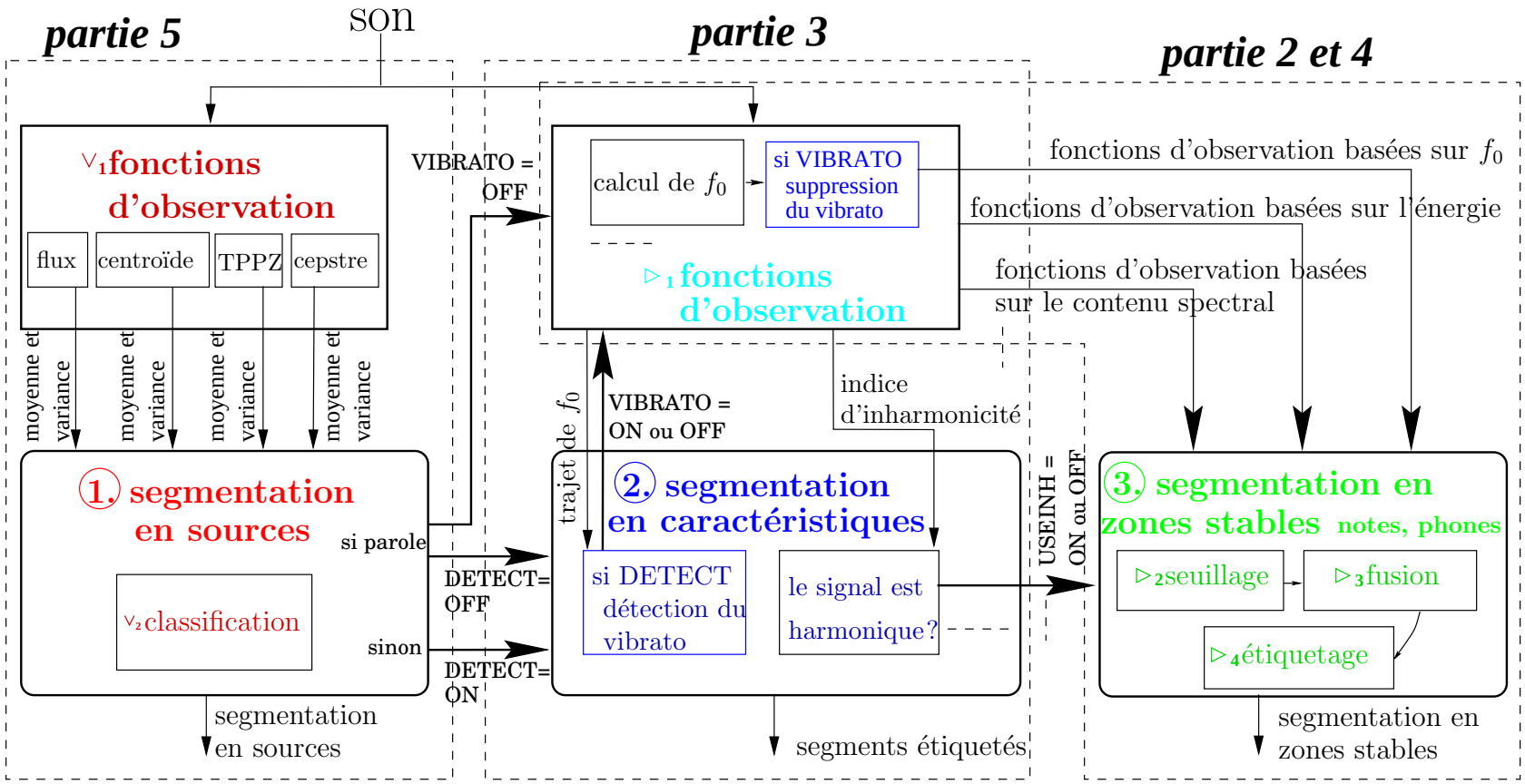


FIG. 25.1 – *Dépendances entre les niveaux de segmentation : algorithme*

d'observation ne peut pas nous aider à poser les marques pour le troisième niveau de segmentation (*segmentation en zones stables*). Ainsi, quand le signal est inharmonique, la variable USEINH est fixée à OFF.

25.1.3 Le cas de la polyphonie

Pour la polyphonie, l'algorithme est le suivant.

Quand de la musique (q_1) est identifiée au premier niveau de segmentation, le détecteur de la polyphonie (q_2), dans le niveau de *segmentation en caractéristiques*, est utilisé. Si la polyphonie est détectée, les fonctions d'observation basées sur le trajet de f_0 ne sont pas envoyées au niveau de *segmentation en zones stables*.

25.2 Dépendances entre la segmentation et la séparation

La séparation de sources et la segmentation sont liées. La détection de la polyphonie (q_2) est effectuée lors de la *segmentation en caractéristiques*. La séparation de sources a besoin d'informations qui sont obtenues lors de la segmentation : notamment, elle a besoin que le son soit segmenté en zones stables (voir la partie II et la partie IV). Ces relations sont schématisées sur la figure 25.2.

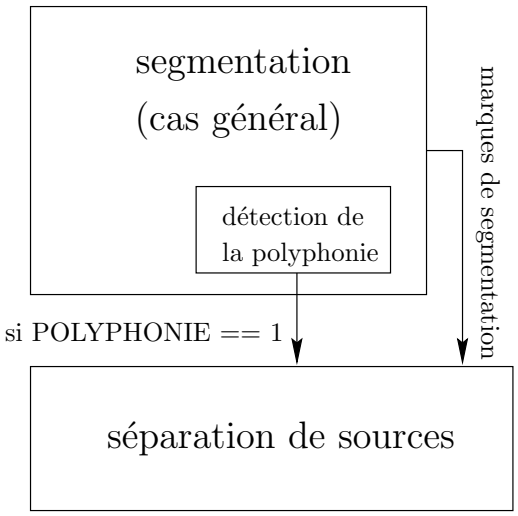


FIG. 25.2 – *Liens entre la segmentation et la séparation de sources : algorithme*

La séparation de sources n'a lieu que si la polyphonie a été détectée (q_2) au moment de la *segmentation en caractéristiques*. L'algorithme pour le *cas général* de la segmentation (*segmentation en sources*, *segmentation en caractéristiques*, *segmentation en zones stables*) est celui décrit dans la section précédente.

Chapitre 26

Conclusion de la cinquième partie

Cette conclusion se décompose en deux paragraphes, chacun concernant l'un des chapitres traités dans cette partie :

- En ce qui concerne la *segmentation en sources*, l'apport par rapport aux techniques présentées dans [SS97] concerne les deux étapes de l'analyse :

  Étape 1 : L'extraction des caractéristiques de base et des fonctions d'observation. Nous nous sommes principalement intéressé à ce problème.

  - En ce qui concerne les caractéristiques de base, nous savons que tous les critères mesurant le flux spectral (voir le tableau 2.5, page 39, pour en avoir la liste) permettent de *segmenter en sources*. Une première famille de caractéristiques de base est ainsi déterminée. De la même façon, nous savons que tous les critères mesurant le degré de voisement (voir le tableau 2.5) permettent de *segmenter en sources*. Une autre famille de caractéristiques de base est déterminée. Nous savons qu'il s'agit de mettre en évidence le fait que la parole est constituée d'une succession de zones stables plutôt non voisées (ou de bruit) et de zones stables plutôt voisées, alors qu'au contraire la musique est constitué d'une succession de zones stables du même type.
  - En ce qui concerne les fonctions d'observation, qui sont calculées à partir des caractéristiques de base, nous avons défini d'autres mesures que les deux mesures (c'est-à-dire la moyenne et la variance) utilisées dans la littérature. Les densités de probabilité des caractéristiques de base sont pour la parole constituées de deux modes (l'un correspondant aux zones voisées et l'autre aux zones non voisées) alors que les densités de probabilité des caractéristiques de base ne sont pour la musique constituées que d'un seul mode. Il s'est agi de mettre en évidence cette différence, plus efficacement qu'avec la moyenne et la variance.

  Étape 2 : La classification. Nous avons utilisé les réseaux de neurones, d'une manière simple, sans chercher à approfondir. Dans [SS97], ils ne sont pas utilisés. Les résultats de la classification ont été légèrement améliorés.
- En ce qui concerne le système complet, nous considérons que la généralisation de l'idée de « zone stable » est faite pour une partie des sons musicaux (sons monophoniques, voisés harmoniques et non modulés ; sons monophoniques, voisés harmoniques et modulés ; sons monophoniques percussifs ; certains sons polyphoniques simples). Pour ces cas, les problèmes aussi bien physiques qu'informatiques que posent la détection des fins et des débuts de zones stables ont été résolus.

Tout du long de cette partie, de nombreuses perspectives ont été indiquées.

Sixième partie

Conclusion générale et perspectives

Chapitre 27

Conclusion générale

Objectifs

Le premier objectif de cette thèse était de proposer et de valider des techniques automatiques de segmentation et d'étiquetage des signaux sonores musicaux.

Beaucoup de gens à l'IRCAM sont intéressés par la segmentation et l'étiquetage : pour la synthèse, pour la structuration d'événements sonores à des niveaux supérieurs, pour l'exploration du *phrasé*, du *geste musical*, pour des traitements du son du type PSOLA, etc. Le programme *segmentation* rend l'utilisation des divers outils de segmentation présentés conviviale, et permet qu'en retour ses utilisateurs nous indiquent les défauts et les bogues présents et nous ouvrent par leurs remarques de nouvelles voies d'exploration. Sur station de travail sous UNIX, une interface graphique a été développée : elle nous a aidé tout du long de la thèse à trouver des voies pour améliorer nos résultats.

Il s'agissait aussi de donner des étiquettes décrivant les sons, ceci entre autres dans l'optique de MPEG-7. Le CCETT aussi bien que l'IRCAM à travers CUIDAD prennent part aux discussions à propos de l'élaboration de ce standard.

Le second objectif de cette thèse était de fournir aux trois centres de recherche collaborant à ce projet des programmes suffisamment finalisés et documentés pour être utilisables.

Deux programmes ont été développés en C au cours de la thèse. Le premier, *segmentation*, qui a pour but de *segmenter en zones stables* (voir principalement la partie II), rassemble plus de 23000 lignes de code et est disponible en mode ligne de commande à l'IRCAM pour tous les utilisateurs potentiels. Il a été porté pour l'IRCAM sous UNIX SGI, UNIX ALPHA et LINUX ; pour SUPÉLEC – CAMPUS DE METZ sous UNIX SUN et sous VISUAL C ; et pour le CCETT sous VISUAL C. Le second, *sources*, qui a pour but de *segmenter en sources* (voir la partie V), rassemble 3000 lignes de code. Il est porté sur UNIX SUN et UNIX SGI.

Bilan

La difficulté majeure rencontrée pendant la thèse vient de la grande diversité des sons musicaux. Il ne nous a pas semblé possible de définir un modèle de signal unique pour tous les sons musicaux qui soit plus mathématiquement ou statistiquement précis que : les sons musicaux sont composés de « zones stables » séparées par des « transitions ». Il s'est agi ensuite de définir ce que nous entendions par « stable » et « transition ». Le système pour segmenter et indexer les sons (principalement les sons musicaux) présenté dans cet exposé répond à ces questions pour un certain nombre de types de sons.

1. Dans la **première partie** de l'exposé, nous avons décidé de nous intéresser aux sons musicaux les plus « simples » possibles : les sons monophoniques, harmoniques et non modulés. Dans ce cas, nous considérons que chaque zone « stable » est une note (ou un phone). En ce qui concerne le terme « transition », sa définition est moins aisée. Il est apparu en cours de thèse que trois types de transitions, quoique coexistant le plus souvent, peuvent être distingués : les

transitions en fréquence fondamentale, les transitions en énergie et les transitions en contenu spectral. Cette distinction en trois types de transitions correspond à une distinction existant réellement physiquement ou psychophysiquement : en effet, la hauteur, le niveau et le timbre [1] sont les trois paramètres qui définissent psychoacoustiquement un son. Dès lors, l'analyse s'est décomposée, presque naturellement, en quatre étapes :

- La première est l'extraction de fonctions d'observation.
  Nous demandons à ces fonctions d'observation de réagir aux moments des transitions. Trois familles de fonctions d'observation, correspondant aux trois types de transition, ont été définies. La première (pages 11 à 19) rassemble les dérivées numériques de f_0, les indices d'inharmonicité, une première espèce d'indices de voisement, l'analyse statistique de f_0 et l'analyse des changements brusques (en moyenne principalement) de f_0 ; la deuxième (page 20) les dérivées numériques de l'énergie, l'analyse statistique de l'énergie et l'analyse des changements brusques de l'énergie ; la troisième (pages 20 à 30) une seconde espèce d'indices de voisement et les flux spectraux. D'autres fonctions d'observation sont mentionnées (pages 30 à 34).
- La deuxième est la prise de décision pour chaque fonction d'observation.
  Il s'agit ici de seuiller les fonctions d'observation pour obtenir des fonctions de décision. Un grand nombre de méthodes de seuillage ont été testées et évaluées (ceci fait l'objet d'une annexe). La méthode classique des 3σ a été retenue (page 42).
- La troisième est la fusion des fonctions de décision.
  Pour qu'à un instant donné il soit décidé qu'il y a une transition, il faut qu'un « certain nombre » de fonctions de décision ait indiqué qu'il y a une transition à cet instant (page 46).
  Mais cette règle de fusion n'est pas suffisante, du fait que les transitions ne sont pas instantanées. Ceci est le principal problème qui a été rencontré dans cette partie. Ce problème rend le mixage des résultats particulièrement difficile. Quelques méthodes, imparfaites, ont été proposées pour résoudre cette difficulté (page 48).
- La quatrième et dernière est l'étiquetage des segments obtenus.
  Chaque segment est encadré par deux marques de segmentation, caractérisées chacune par sa position et la confiance que nous pouvons lui accorder. Une fois que les marques de segmentation sont obtenues, il reste à décrire le contenu de chacun des segments. Pour les sons simples, il s'est agi de retranscrire la partition (page 56).

Puis nous avons étendu – généralisé – ces résultats à des sons plus compliqués, les procédures utilisées pour les sons simples servant de base. Nous mentionnons sur la figure 27.1 pour quel types de sons des extensions ont été apportées : elles concernent les sons monophoniques percussifs (ces sons sont non voisés : il s'agit de détecter les transitoires d'énergie, c'est-à-dire de ne prendre en compte que les fonctions d'observation basées sur l'énergie), les sons monophoniques modulés et une catégorie restreinte de sons polyphoniques. En ce qui concerne ces derniers, des extensions possibles pour le futur sont indiquées. Comme l'indique la figure 27.1, de la même façon que les techniques utilisées pour les sons simples sont adaptées aux sons plus compliqués, les techniques utilisées pour les sons polyphoniques « simples » seront adaptées aux sons polyphoniques plus compliqués.

2. La première généralisation effectuée a fait l'objet de la **deuxième partie** de l'exposé. Elle a concerné les sons modulés. Les deux types de modulation considérés sont la modulation de fréquence (vibrato) et la modulation d'amplitude (trémolo). Nous nous sommes intéressé dans cette partie à la détection du vibrato, à l'estimation de ses paramètres et à sa suppression du trajet de f_0. Deux familles de méthodes ont été décrites.

   - La première rassemble les méthodes basées sur l'analyse directe du son. Deux méthodes originales ont été présentées pour détecter le vibrato.
     - L'une a pour base le fait que moduler sinusoïdalement en fréquence une porteuse

1. Cependant, la définition de « contenu spectral » telle qu'elle a été donnée dans cet exposé ne recouvre pas complètement celle de « timbre » : voir [ZF81]. Et, de plus, à ces trois premiers paramètres, il faut en ajouter un quatrième, qui concerne la perception de l'espace.

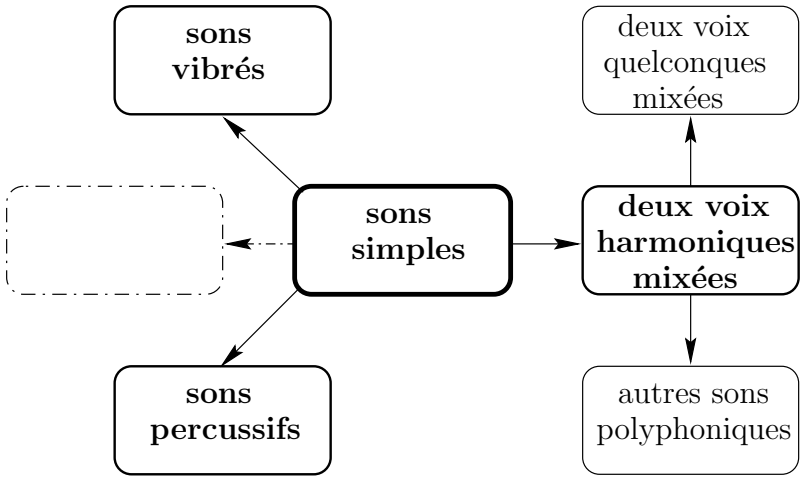


FIG. 27.1 – *La segmentation et l'étiquetage des sons musicaux. Des sons « simples » aux sons « compliqués » ; et des sons compliqués aux sons encore plus compliqués*

décompose dans le domaine fréquentiel son lobe en une somme infinie de lobes espacés de f_{vib} Hz et d'amplitudes les coefficients de BESSEL (page 93). Un modèle du spectre est obtenu. Il s'agit de déterminer les paramètres du modèle pour lesquels l'erreur entre un spectre théorique et un spectre vrai soit minimale.

- L'autre vient de considérations sur l'atténuation des lobes harmoniques induite par la modulation de fréquence (page 104). Nous avons montré que cette atténuation varie quasi linéairement avec le numéro d'ordre de l'harmonique. Il est possible de détecter cette atténuation quasi linéaire, et donc de déterminer sûrement si un vibrato est présent ou non. Nous avons montré la validité de cette méthode sur des signaux musicaux réels.
- La troisième méthode (page 112) vient de la thèse de LAROCHE (voir [Lar89]).

- La seconde rassemble les méthodes basées sur l'analyse du trajet de f_0. Le problème rencontré ici est que le signal (le trajet de f_0) ne peut être considéré comme stationnaire que sur très peu de périodes du vibrato. La méthode classique d'analyse spectrale qu'est la transformée de FOURIER ne peut pas alors être utilisée telle quelle.

  - La première idée (page 117) a été de prédire le passé et le futur du signal grâce à la modélisation auto-régressive afin de rendre la transformée de FOURIER utilisable.
  - La deuxième idée (page 121) a été de détecter les maximums et les minimums du signal (traitement non linéaire), pour déterminer un trajet de f_0 sur lequel le vibrato est supprimé.
  - La troisième idée (page 123) a été de calculer la transformée de HILBERT du signal pour obtenir sa phase – et donc sa fréquence – instantanée.

  Nous avons montré la validité des trois méthodes sur des signaux réels.

Puisqu'un grand nombre de techniques ont été mises en place pour détecter le vibrato et estimer ses paramètres, des méthodes de fusion de données ont été mises en place pour améliorer la robustesse des résultats. Pour le problème du vibrato, les limitations rencontrées dans la première partie de l'exposé lors de la fusion de données n'ont pas cours. Par exemple, en ce qui concerne la prise de décision, la règle de fusion de la majorité peut être utilisée.

3. La deuxième généralisation effectuée a concerné les sons polyphoniques. Pour des sons polyphoniques simples, des méthodes pour détecter la polyphonie ont été décrites (pages 149 à 152). Il a été indiqué que l'utilisation d'une des trois familles de fonctions d'observation pour *segmenter en zones stables* décrites dans la partie II pour les sons simples devient

problématique (page 153) pour les sons polyphoniques : il s'agit de celle rassemblant les fonctions d'observation basées sur le trajet de f_0. Dans cette partie, certaines limitations des techniques existantes pour séparer les sources ont été montrées (page 157). Ces études ont donné lieu à un grand nombre d'interrogations et ainsi ouvert tout un champ de perspectives.

4. Un système complet a été défini. Ce système a pour mission première de détecter quel est le type du son qui lui est donné à traiter : il doit déterminer ce qu'est une zone stable et ce qu'est une transition pour ce son particulier. Puis il doit adapter son analyse à ce son. Pour chaque type de sons, la procédure suivie est différente, mais la structure de base – le système – reste la même.

   - Tout d'abord, il s'est agi de déterminer la nature du son à traiter. Les parties précédentes ont été consacrées à résoudre les problèmes de segmentation et d'indexation pour certains sons musicaux. Ici, nous nous attachons à déterminer si nous sommes en présence d'un son de cette nature, c'est-à-dire si nous sommes en présence d'un son musical. Nous avons pour ce problème bénéficié de l'expérience de nos prédécesseurs (notamment de [SS97]). Plus qu'au problème de la classification, nous nous sommes intéressé au problème de la définition de caractéristiques de base (page 171) réagissant différemment suivant la nature du son, puis à la définition de fonctions d'observation permettant de mettre efficacement en évidence ces différences (page 173). En ce qui concerne les caractéristiques de base, nous avons montré que tous les flux spectraux et que tous les indices du voisement sont à même de nous aider à résoudre le problème. En ce qui concerne les fonctions d'observation, nous avons mis en place de nouvelles mesures, au moins autant efficaces que les mesures utilisées dans la littérature.
   - Ensuite, il s'est agi de rassembler les diverses techniques présentées au cours de l'exposé en un ensemble cohérent. Un « système » (page 187) a été élaboré. Il résume algorithmiquement le contenu de cet exposé. Informatiquement, il n'existe que par morceaux.

Les études menées pendant les trois ans de thèse ont donné lieu à la publication de plusieurs articles : trois communications référencées [RRS$^+$98], [RRS$^+$99] et [RRD$^+$99] ; et une publication avec comité de lecture référencée [RRS$^+$].

Trois axes de recherche pour des travaux futurs se sont dessinés au cours de la thèse :

point 1° Étendre le système tel qu'il a été défini à de plus en plus de types de sons musicaux.

point 2° Dépasser certaines limitations de ce système en redéfinissant sa structure interne : notamment pour éviter les problèmes de fusion de données (troisième étape de l'analyse *segmentation en zones stables* : voir le chapitre 4, page 46).

point 3° Prendre le problème d'un autre point de vue (modèle de signal différent).

Nous détaillons ces perspectives dans le chapitre suivant.

Chapitre 28

Perspectives

En ce qui concerne la poursuite des travaux présentés dans cet exposé, au cours de la thèse deux voies ont été envisagées. Ou bien nous gardons notre modèle de signal : des « zones stables », séparées par des transitions plus ou moins brusques relevant de trois types : transition en fréquence fondamentale, transition en énergie et transition en contenu spectral : ceci concerne les points 1° et 2° donnés à la fin du chapitre précédent. Ou bien nous adoptons un modèle de signal du type statistique (HMM [1], rupture de modèles AR...), que nous adaptons pour les signaux sonores musicaux : ceci concerne le point 3° donné à la fin du chapitre précédent. Ce modèle nous permettrait d'éviter entre autres tous les problèmes de fusion de données que nous avons rencontrés.

En ce qui concerne le point 1° , les perspectives sont les suivantes :

- Pour la segmentation en sources

  Considérer plus de deux types de sources Il serait intéressant de considérer plus de deux classes de sons. Ces classes pourraient être :

  – parole
  – voix chantée et musique instrumentale
  – musique instrumentale seule
  – voix chantée seule
  – bruits de machines, bruits de rue...

  Il faut pour cela développer des fonctions d'observation propre à séparer ces classes. Quelques travaux, basés sur une stratégie similaire à celle que nous avons adoptée dans cet exposé (extraction de fonctions d'observation, puis classification), existent en ce qui concerne la segmentation en trois classes ([SS97], [SBZD99]), mais le taux d'erreur obtenu est important (de l'ordre de 20 %).

  Les styles musicaux Il serait intéressant de séparer dans la classe musique les genres musicaux.

- Pour la segmentation en caractéristiques
  Nous avons vu que nous avons développé un critère qui nous permet de décider quand il y a plutôt du vibrato et quand il n'y en a plutôt pas. Des critères existent aussi pour décider quand le signal est plutôt harmonique ou inharmonique, plutôt voisé ou non voisé, etc. Il reste que la plupart des liens informatiques entre le premier niveau de segmentation et le troisième n'existent pas encore. C'est-à-dire qu'informatiquement le deuxième niveau de segmentation n'existe pas complètement. Ceci concerne le second objectif de la thèse.
- Pour la segmentation en zones stables
  Nous avons déjà mentionné que l'ensemble des critères de classification (mélange de gaussiennes, réseaux de neurones, k plus proches voisins) et que l'ensemble des fonctions d'observation sont extensibles (nous pourrions par exemple y inclure des analyses par ondelettes

1. Ces techniques sont utilisées en traitement de la parole : voir par exemple [ABF+94], [BFO93], [Fal95]...

du signal, des coefficients cepstraux, des moments d'ordre supérieurs, etc.). Il s'agit d'utiliser les fonctions d'observation les moins corrélées possible. Cependant, nous avons montré certaines limitations de cette approche : le fait que les transitions ne sont pas instantanées pose des problèmes partiellement résolus de fusion des données. Des comparaisons critiques des fonctions d'observation devront être effectuées plus systématiquement.

- Pour l'indexation
  Il faudrait la compléter. Avec le système présenté dans cet exposé, nous extrayons plutôt des descripteurs de bas niveau : le son est monophonique ou polyphonique, etc. D'autres descripteurs plutôt de bas niveau manquent, comme par exemple le nombre de sources présentes dans un son polyphonique. De plus, dans l'avenir, des descripteurs de plus haut niveau, relevant par exemple de la reconnaissance du timbre, de l'instrument, du sexe du chanteur, de la voix du chanteur, etc., devront être extraits.
- Pour la séparation de sources
  Dans les méthodes utilisées par les antennistes pour la séparation de sources, il est nécessaire que le nombre d'antennes soit au moins égal au nombre de sources présentes (voir [Car95]). Aussi, pour qu'elles commencent à devenir utilisables dans notre cas, il faut que nous considérions des sons enregistrés en stéréophonie. Cependant, le problème des antennistes n'est pas le même problème que celui présenté dans la partie IV de cet exposé : le modèle de signal est différent.
  De plus, les dépendances entre la segmentation et la séparation de sources ne doivent sans doute pas être aussi unilatérales que nous l'avons indiqué : les deux analyses devraient plutôt être menées de concert.

En ce qui concerne le point 2° , nous nous intéressons à la définition d'une « fonction d'observation » qui détecte toutes les transitions, qui soit universelle : qui soit insensible aux « petites variations » de la hauteur (vibrato, glissando) et aux « petites » variations de l'énergie (trémolo, mort lente d'une note, glissando). Cela peut paraître étrange de vouloir construire une fonction d'observation universelle, puisque, comme nous l'avons dit, suivant ce que désire l'utilisateur, nous voulons pouvoir obtenir plusieurs *segmentations en zones stables* : mais, alors, l'utilisateur aurait accès à certains « paramètres de contrôle » de cette fonction d'observation universelle. Ici, nous relâcherions la contrainte sur la différenciation des types de transitions brusques. Le modèle de signal dirait simplement : le son est composé de « zones stables », séparées par des transitions brusques.

En ce qui concerne le modèle de signal à utiliser pour la segmentation et l'indexation, il existe en fait au moins trois approches :

- Utiliser plusieurs fonctions d'observation, spécialisées, et mixer (fusionner) les résultats que chacune nous donne. Nous faisons ceci pour deux raisons : d'abord pour détecter tous les types de variations ; ensuite pour améliorer la robustesse du système. Cette approche est celle qui a été considérée dans cet exposé. Certaines de ses limites ont été montrées.
  De plus, la structure du système de segmentation présenté dans cet exposé est telle que plus nous intégrerons de nouveaux types de sons à segmenter, plus nous aurons de problèmes à résoudre simultanément en ce qui concerne les traitements à appliquer à chaque type de sons (point 1°) .
- Utiliser une seule fonction d'observation, qui réagit à toute « variation brusque » que nous voulons détecter (point 2°) .
- Changer de modèle de signal (point 3°) . Il s'agirait ici d'utiliser et d'adapter les modèles de signal utilisés en traitement de la parole : HMM, rupture de modèles AR.

La deuxième et la troisième approche ne remettent pas forcément en cause l'existence des deux niveaux de segmentation supérieurs (*segmentation en caractéristiques* et *segmentation en sources*). Par exemple, il est nécessaire de faire la distinction parole/musique avant de *segmenter en zones*

stables :

- Car la parole est plutôt monophonique, alors que la musique est souvent polyphonique (d'où la nécessité de traitements supplémentaires, relevant de la séparation de sources, à opérer, sans doute, en parallèle avec la segmentation).
- Car les signaux de parole et de musique, déjà dans le cas monophonique, sont très différents : importance relative des parties voisées et des parties non voisées, nombre de transitions par unité de temps, modulations, etc.

Aussi, le modèle de signal doit être différent pour la parole et la musique.

Un autre travail a été entrepris qui n'a pas été présenté dans cet exposé. Il s'agit de l'amélioration des indices de voisement à partir de la prise en compte des déformations que les lobes principaux dus aux sinusoïdes subissent dès que celles-ci ne sont plus stationnaires sur la fenêtre d'analyse, c'est-à-dire dès qu'une perturbation, soit un vibrato, et/ou un trémolo, et/ou un glissando en fréquence et/ou en énergie, est présente. Ces déformations suivant la perturbation considérée sont bien différenciées. Quand plusieurs perturbations sont présentes il s'agit de séparer leur influence respective. Le but de ce travail était entre autres de construire la « fonction d'observation universelle » dont il a été question dans les paragraphes précédents.

Septième partie

Annexes

Annexe A

Rappels sur le calcul de la probabilité $P_r\left(|X_2\left(i\right)-X_1\left(i\right)|\leq S\right)$

A.1 Le problème

Nous voulons calculer la probabilité $p(i)=P_r\left(|X_2\left(i\right)-X_1\left(i\right)|\leq S\right)$ quand X_1 et X_2 sont deux suites aléatoires dont les densités de probabilité sont des gaussiennes $\mathcal{N}\left(m_1,\sigma_1^2\right)$ et $\mathcal{N}\left(m_2,\sigma_2^2\right)$, et où S est un seuil fixé.

A.2 Méthode « classique »

Nous avons :

$$p(i)=P_r\left(|X_2\left(i\right)-X_1\left(i\right)|\leq S\right)=P_r\left(-S\leq X_2\left(i\right)-X_1\left(i\right)\leq S\right)$$

Or, nous savons (voir [Pap65], page 222) que la loi de probabilité de $Y=\sum_{i=1}^{N}\pi_i X_i$, où les X_i obéissent chacun à une loi normale $\mathcal{N}\left(m_i,\sigma_i^2\right)$ et où les π_i valent -1 ou 1, est normale et vaut :

$$\mathcal{N}\left(\sum_{i=1}^{N}\pi_i m_i,\sum_{i=1}^{N}\sigma_i^2\right)$$

Donc ce que nous voulions démontré est démontré.

A.3 Méthode « à la main »

La démonstration est beaucoup plus laborieuse, mais, pour emporter la conviction, il faut l'avoir faite une fois à la main.

La formule des probabilités totales nous donne :

$$p(i)=\int\limits_{\mathcal{D}X_1}P_r\left(|X_2\left(i\right)-X_1\left(i\right)|\leq S/X_1\left(i\right)\right)P_r\left(X_1\left(i\right)=x_1\right)dx_1$$

où $\mathcal{D}X_1$ est le domaine de définition de X_1. Ici $\mathcal{D}X_1=\mathcal{D}X_2=]-\infty,+\infty[$.

Nous obtenons donc :

$$p(i)=\frac{1}{2\pi\sigma_1\sigma_2}\int\limits_{-\infty}^{+\infty}\left\{\int\limits_{x_1-S}^{x_1+S}\exp\left[-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2\right]dx_2\right\}\exp\left[-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2\right]dx_1$$

Nous posons :

$y = x_2 - x_1$

donc :

$dy = dx_2$

$x_2 = y + x_1$

$x_2 = x_1 - S \rightarrow y = -S$

$x_2 = x_1 + S \rightarrow y = +S$

et :

$$\begin{aligned} p(i) &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \left\{ \int_{-S}^{+S} \exp\left[-\frac{1}{2} \left(\frac{y + x_1 - m_2}{\sigma_2} \right)^2 \right] dy \right\} \exp\left[-\frac{1}{2} \left(\frac{x_1 - m_1}{\sigma_1} \right)^2 \right] dx_1 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \left\{ \int_{-S}^{+S} \exp\left[-\frac{1}{2} \left(\frac{y^2}{\sigma_2^2} + \frac{x_1^2}{\sigma_2^2} + \frac{m_2^2}{\sigma_2^2} + \right. \right. \right. \\ &\quad \left. \left. \left. \frac{2yx_1}{\sigma_2^2} - \frac{2ym_2}{\sigma_2^2} - \frac{2x_1m_2}{\sigma_2^2} + \frac{x_1^2}{\sigma_1^2} + \frac{m_1^2}{\sigma_1^2} - \frac{2x_1m_1}{\sigma_1^2} \right) \right] dy \right\} dx_1 \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \left\{ \int_{-S}^{+S} \exp\left[-\frac{1}{2}\mathcal{T} \right] dy \right\} dx_1 \end{aligned}$$

avec :

$$\mathcal{T} = \frac{x_1^2}{\sigma_2^2} + \frac{x_1^2}{\sigma_1^2} + 2x_1 \left(\frac{y}{\sigma_2^2} - \frac{m_2}{\sigma_2^2} - \frac{m_1}{\sigma_1^2} \right) + \frac{m_2^2}{\sigma_2^2} + \frac{m_1^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - \frac{2ym_2}{\sigma_2^2}$$

Nous voulons écrire les trois premiers termes de $\mathcal{T}$ sous la forme :

$$\left(\frac{x_1 - m}{\sigma} \right)^2 + \mathcal{F} = \frac{x_1^2}{\sigma^2} + \frac{m^2}{\sigma^2} - \frac{2x_1m}{\sigma^2}$$

où $\mathcal{F}$ peut être fonction de tout sauf de x_1.
Nous obtenons :

$$\left\{ \begin{array}{lcll} \frac{1}{\sigma^2} & = & \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} & \rightarrow \text{ des 2 premiers termes} \\ \frac{m}{\sigma^2} & = & -\frac{y}{\sigma_2^2} + \frac{m_2}{\sigma_2^2} + \frac{m_1}{\sigma_1^2} & \rightarrow \text{ du troisième terme} \\ \frac{m^2}{\sigma^2} & = & \frac{1}{\sigma^2} \left[\sigma^2 \left(-\frac{y}{\sigma_2^2} + \frac{m_2}{\sigma_2^2} + \frac{m_1}{\sigma_1^2} \right) \right]^2 = -\mathcal{F} & \end{array} \right.$$

Nous avons donc :

$$\mathcal{T} = \left(\frac{x_1 - m}{\sigma} \right)^2 + \underbrace{\mathcal{F} + \underbrace{\frac{m_2^2}{\sigma_2^2}}_{I} + \underbrace{\frac{m_1^2}{\sigma_1^2}}_{II} + \underbrace{\frac{y^2}{\sigma_2^2}}_{III} - \underbrace{\frac{2ym_2}{\sigma_2^2}}_{IV}}_{\mathcal{A}}$$

Décomposons $\mathcal{F}$:

$$\mathcal{F} = -\sigma^2 \left(\frac{y^2}{\sigma_2^4} + \frac{m_2^2}{\sigma_2^4} + \frac{m_1^2}{\sigma_1^4} - \frac{2ym_2}{\sigma_2^4} - \frac{2ym_1}{\sigma_1^2\sigma_2^2} + \frac{2m_1m_2}{\sigma_1^2\sigma_2^2} \right)$$

Or, $\frac{1}{\sigma^2}=\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}$. Donc $\sigma^2=\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$, et :

$$\begin{aligned}\mathcal{F}=&\\ &\underbrace{-y^2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}}_{1}\underbrace{-m_2^2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}}_{2}\underbrace{-m_1^2\frac{\sigma_2^2}{\sigma_1^2}\frac{1}{\sigma_1^2+\sigma_2^2}}_{3}+\\ &\underbrace{2ym_2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}}_{4}+\underbrace{2ym_1\frac{1}{\sigma_1^2+\sigma_2^2}}_{5}\underbrace{-2m_1m_2\frac{1}{\sigma_1^2+\sigma_2^2}}_{6}\end{aligned}$$

Si nous combinons I et 2 nous obtenons :

$$\frac{m_2^2}{\sigma_2^2}-m_2^2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}=\frac{m_2^2}{\sigma_2^2}\left(\frac{\sigma_1^2+\sigma_2^2-\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)=\frac{m_2^2}{\sigma_1^2+\sigma_2^2}$$

De même si nous combinons II et 3 :

$$\frac{m_1^2}{\sigma_1^2}-m_1^2\frac{\sigma_2^2}{\sigma_1^2}\frac{1}{\sigma_1^2+\sigma_2^2}=\frac{m_1^2}{\sigma_1^2}\left(\frac{\sigma_1^2+\sigma_2^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)=\frac{m_1^2}{\sigma_1^2+\sigma_2^2}$$

Puis III et 1 :

$$\frac{y^2}{\sigma_2^2}-y^2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}=\frac{y^2}{\sigma_2^2}\left(\frac{\sigma_1^2+\sigma_2^2-\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)=\frac{y^2}{\sigma_1^2+\sigma_2^2}$$

Et, enfin, IV et 4 :

$$-\frac{2ym_2}{\sigma_2^2}+2ym_2\frac{\sigma_1^2}{\sigma_2^2}\frac{1}{\sigma_1^2+\sigma_2^2}=\frac{2ym_2}{\sigma_2^2}\left(\frac{-\sigma_1^2-\sigma_2^2+\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)=-\frac{2ym_2}{\sigma_1^2+\sigma_2^2}$$

$\mathcal{A}$ s'écrit alors :

$$\mathcal{A}=\frac{m_2^2}{\sigma_1^2+\sigma_2^2}+\frac{m_1^2}{\sigma_1^2+\sigma_2^2}+\frac{y^2}{\sigma_1^2+\sigma_2^2}-\frac{2ym_2}{\sigma_1^2+\sigma_2^2}+\frac{2ym_1}{\sigma_1^2+\sigma_2^2}-\frac{2m_1m_2}{\sigma_1^2+\sigma_2^2}$$

Or :

$$\left(y+m_1-m_2\right)^2=m_2^2+m_1^2+y^2-2ym_2+2ym_1-2m_1m_2$$

Donc :

$$\mathcal{A}=\frac{\left(y+m_1-m_2\right)^2}{\sigma_1^2+\sigma_2^2}$$

Nous avons ainsi :

$$\mathcal{T}=\left(\frac{x_1-m}{\sigma}\right)^2+\frac{1}{\sigma_1^2+\sigma_2^2}\left(y+m_1-m_2\right)^2$$

Alors :

$$\begin{aligned}p(i) &= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{+\infty}\left\{\int_{-S}^{+S}\exp\left[-\frac{1}{2}\left(\frac{x_1-m}{\sigma}\right)^2-\frac{1}{2}\frac{1}{\sigma_1^2+\sigma_2^2}\left(y+m_1-m_2\right)^2\right]dy\right\}dx_1\\ &= \frac{1}{2\pi\sigma_1\sigma_2}\underbrace{\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{2}\left(\frac{x_1-m}{\sigma}\right)^2\right]dx_1}_{=\sqrt{2\pi}\sigma}\int_{-S}^{+S}\exp\left[-\frac{1}{2}\frac{1}{\sigma_1^2+\sigma_2^2}\left(y+m_1-m_2\right)^2\right]dy\end{aligned}$$

Soit, finalement :

$$p(i) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \int_{-S}^{+S} \exp\left[-\frac{1}{2}\left(\frac{y - (m_2 - m_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)^2\right] dy$$

Puisque les bornes d'intégration sont finies et puisque la fonction à intégrer ne présente pas de singularités, cette intégrale se calcule aisément numériquement.

Annexe B

Le seuillage

B.1 Introduction

Dix-neuf méthodes de seuillage sont présentées dans cette annexe. Les dix premières sont explicitées dans un article de synthèse (voir [SSW88]) écrit par des traiteurs d'images ; la onzième a été construite par nous à partir de considérations faites à propos de la méthode d'analyse spectrale MUSIC ; et les deux suivantes sont utilisées entre autres pour le débruitage de signaux à partir de leur décomposition en paquets d'ondelettes : voir [Don94] et [DJ94]. Nous trouvons ces références aux adresses http suivantes :

`http://www.mathsoft.com/wavelets.html` et
`http://stat.Stanford.EDU/reports/donoho`

La quatorzième est celle proposée par SAHOO, SOLTANI et WONG dans [SSW88] ; et la dix-neuvième est une amélioration de cette méthode, due à quelques constatations que nous avons faites. Les autres ont été glanées ici et là. Bien sûr, nous ne prétendons pas avoir été exhaustif : certainement, un grand nombre d'autres méthodes de seuillage existent.

Nous considérons chacune de nos fonctions d'observation. Ses échantillons se répartissent en deux classes. Il s'agit de discriminer ces classes. Si nous supposons que les observations correspondant à chacune de ces deux classes obéissent à des suites aléatoires gaussiennes, la première, correspondant aux parties stables du signal, a une moyenne et une variance petites, alors que la seconde, correspondant aux pics à détecter, a une moyenne plus grande et une variance elle aussi plus grande. De plus, la première classe est représentée par la très grande majorité des échantillons : la seconde classe est représentée par très peu de points.

B.2 1 – Méthode du saut, ou du contraste

B.2.1 Principe

Nous classons par ordre de grandeur les échantillons, du plus petit au plus grand. Le tableau $[Y(1)\ \ldots\ Y(N_n)]$ est obtenu. Nous déterminons les valeurs $saut(i) = Y(i) - Y(i-1)$. Nous faisons l'hypothèse que les deux modes sont bien séparés, c'est-à-dire que nous supposons qu'il y a, une fois que nous avons classé les échantillons de la fonction d'observation par ordre de grandeur, un saut brusque quand nous passons d'un mode à l'autre. Nous supposons que le contraste est important. Le moment de la cassure a lieu dès que le saut $saut(i)$ dépasse un certaine valeur S, fixée. Le problème est de déterminer la valeur S du saut limite.

B.2.2 Remarque

En fait, nous avons utilisé plutôt cet algorithme :

- nous classons les échantillons du signal y de taille N_n en ordre croissant (nous obtenons Y)

- nous calculons tous les sauts $saut(i) = Y(i+1) - Y(i)$ (pour i variant de 1 à $N_n - 1$)
- $[sautmax\ imax] = max\left[saut\left(\frac{N_n}{\alpha}+1\right)\ \ldots\ saut\left(N_n - \frac{N_n}{\alpha}\right)\right]$ (nous ne prenons pas en compte les sauts concernant les plus petits échantillons et ceux concernant les plus grands) nous donne le plus grand saut (contraste) : $sautmax$ et son indice $imax$ dans le tableau des sauts
- la valeur du seuil est : $seuil = Y\left(imax + \frac{N_n}{\alpha}\right)$ (au lieu de $seuil = Y(imax)$ comme dans la section B.2.1)

qui, avec le signal du premier test effectué ci-dessous (voir la section B.21, pour plus de détails au sujet des tests), nous donne les meilleurs résultats pour des valeurs de α de l'ordre de 25 – 30.

α est un paramètre libre, que nous avons fixé : dans la suite, nous avons travaillé avec $\alpha = 25$.

Voir les figures B.6, B.24 et B.41 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.3 2 – Méthode des deux modes de l'histogramme

B.3.1 Exploitation analytique de l'ensemble des probabilités

Il s'agit de la méthode utilisée dans le programme *tran* original de Laurent CERVEAU : voir la section 2.2.5 de la partie II. Appelons $s(i)$ la fonction d'observation : le programme *tran* nous fournit $s(i) = 1 - p(i)$.

Nous calculons l'histogramme $H = [h_1 \ldots h_L]$ de l'ensemble des $s(i)$. Celui-ci présente deux modes : l'un, proche de 1 (ce mode correspond à une dissemblance maximale entre les deux modèles : l'une des fenêtres d'analyse couvre une partie d'une note et l'autre une partie d'une autre note), correspondant à l'état instable ; et l'autre, proche de 0 (ce mode correspond à une dissemblance minimale : les fenêtres d'analyse couvrent deux parties de la même note), correspondant à l'état stable. Il s'agit donc de déterminer la position $0 \leq c \leq 1$ de la césure séparant ces deux modes. Les marques de segmentation sont posées aux instants i tels que : $s(i) \leq c$.

Nous déterminons les K minimums locaux de l'histogramme , c'est-à-dire tous les points k tels que $h(k) < h(k-1)$ && $h(k) < h(k+1)$, pour k variant de 2 à $L-2$, L étant le nombre de cases de l'histogramme, numérotées de 1 à L.

Pour chacun de ces K triplets de points $k-1$, k, $k+1$, nous déterminons le polynôme d'ordre 2 qui passe par ses trois points. La position de césure possible $P(k)$ correspond à la valeur à laquelle s'annule la dérivée du polynôme, et le minimum correspondant est $M(k)$.

La valeur minimale de l'ensemble M à K éléments est déterminée. La position optimale T de la césure est la position $P(i)$, avec i tel que $M(i)$ est la valeur minimale de l'ensemble M.

Des discussions sur les justifications de cet algorithme, sur le nombre de classes à prendre pour l'histogramme, etc. sont disponibles dans [Cer94].

Voir les figures B.7, B.25 et B.42 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.3.2 Extension – Perspective

Cette extension n'est pas implémentée dans le programme *segmentation*. Elle correspond à une approche probabiliste de l'exploitation de l'ensemble des $s(i)$.

Nous supposons que les densités de probabilité des variables aléatoires X_1 et X_2 sont des gaussiennes. Il s'agit d'approximer au mieux, par les moindres carrés par exemple, l'histogramme par la somme de deux gaussiennes de moyennes m_1 et m_2, de variances σ_1 et σ_2, et de probabilités a priori π_1 et π_2 inconnues.

Une fois que nous avons estimé m_1, m_2, σ_1, σ_2, π_1 et π_2, nous pouvons utiliser la stratégie classique permettant de déterminer T, c'est-à-dire celle de BAYES.

B.3.2.1 Critère de Bayes

Le critère de Bayes consiste à minimiser une fonction de risque, qui évalue le « coût » d'une mauvaise décision. En général, la probabilité d'erreur est choisie comme fonction à minimiser.

Il y a deux configurations :

- Si $\sigma_1 = \sigma_2 = \sigma$:

$$T = \frac{2\sigma^2 \log_e\left(\frac{\pi_1}{\pi_2}\right) + m_2^2 - m_1^2}{2\,(m_2 - m_1)}$$

  Si en plus $\pi_1 = \pi_2 = 0{,}5$, $T = \dfrac{m_1 + m_2}{2}$.

- Dans le cas général ($\sigma_1 \neq \sigma_2$) :

$$T = \frac{m_1\sigma_2^2 - m_2\sigma_1^2 \pm \sqrt{m_1^2\sigma_2^4 + m_2^2\sigma_1^4 - 2m_1m_2\sigma_1^2\sigma_2^2 - (\sigma_2^2 - \sigma_1^2)\left[-2\log_e\left(\frac{\sigma_1}{\pi_1}\frac{\pi_2}{\sigma_2}\right)\sigma_1^2\sigma_2^2 + m_1^2\sigma_2^2 - m_2^2\sigma_1^2\right]}}{\sigma_2^2 - \sigma_1^2}$$

B.4 3 – Méthode d'Otsu

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ normalisé $\left(\sum_{i=0}^{L-1} h_i = 1\right)$ de la fonction d'observation.

Il s'agit de maximiser la variance inter-classes σ_B^2 suivant t le numéro de la case de l'histogramme courante. Cette variance inter-classe est égale à :

$$\sigma_B^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2$$

Avec :

$$\omega_0 = \sum_{i=0}^{t} h_i \text{ et } \omega_1 = 1 - \omega_0$$

Et, si nous posons :

$$\mu_t = \sum_{i=0}^{t} ih_i \text{ et } \mu_T = \sum_{i=0}^{L-1} ih_i$$

avec :

$$\mu_0 = \frac{\mu_t}{\omega_0} \text{ et } \mu_1 = \frac{\mu_T - \mu_t}{\omega_1}$$

Soit x_i l'abscisse de la case i de l'histogramme. Le seuil est égal à x_{t_M}, où t_M est le t tel que σ_B^2 soit maximale.

Voir les figures B.8, B.26 et B.43 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.5 4 – Première méthode entropique de Pun

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ normalisé $\left(\sum_{i=0}^{L-1} h_i = 1\right)$ de la fonction d'observation.

Les entropies a posteriori des deux modes (mesures de l'information a posteriori associée à chaque mode) sont respectivement (avec t le numéro de la case de l'histogramme courante) :

$$H_1 = -\sum_{i=0}^{t} h_i \log_e h_i \text{ et } H_2 = -\sum_{i=t+1}^{L-1} h_i \log_e h_i$$

Il s'agit de maximiser suivant t le terme $H = H_1 + H_2$.

Pun prouve qu'il est équivalent de maximiser $f(t)$ suivant t le numéro de la case de l'histogramme courante, avec :

$$f(t) = \frac{H_t}{H_T} \frac{\log_e P_t}{\log_e max(h_0 \ \dots \ h_t)} + \left(1 - \frac{H_t}{H_T}\right) \frac{\log_e(1 - P_t)}{\log_e max(h_{t+1} \ \dots \ h_{L-1})}$$

où :

$$H_t = -\sum_{i=0}^{t} h_i \log_e h_i \text{ et } H_T = -\sum_{i=0}^{L-1} h_i \log_e h_i \text{ et } P_t = \sum_{i=0}^{t} h_i$$

Voir la figure B.9 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.6 5 – Nombre de marques connu, ou méthode du pourcentile

Nous supposons que nous connaissons le nombre de marques à détecter. En traitement de l'image, cela correspond à : nous savons que cet objet occupe 20% des pixels.

Dans notre cas, si nous supposons qu'un échantillon sur dix correspond à une transition, il s'agit simplement de ne garder que les 10% plus grands.

B.7 6 – Deuxième méthode entropique de Pun

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ normalisé $\left(\sum_{i=0}^{L-1} h_i = 1\right)$ de la fonction d'observation.

Cette fois, nous considérons le coefficient d'anisotropie α :

$$\alpha = \frac{\sum_{i=0}^{m} h_i \log_e h_i}{\sum_{i=0}^{L-1} h_i \log_e h_i}$$

où m est le plus petit entier tel que :

$$\sum_{i=0}^{m} h_i \geq 0{,}5$$

et où t (numéro de la case de l'histogramme courante : le x_t correspondant est le seuil) est tel que :

$$\sum_{i=0}^{t} h_i = \begin{cases} 1 - \alpha & \text{si} \quad \alpha \leq 0{,}5 \\ \alpha & \text{si} \quad \alpha > 0{,}5 \end{cases}$$

Voir la figure B.10 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.8 7 – Méthode entropique de Kapur, Sahoo et Wong

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ (non normalisé) de la fonction d'observation. Nous avons :

$$H_1 = -\sum_{i=0}^{t} \frac{h_i}{P_t} \log_e \frac{h_i}{P_t} \text{ et } H_2 = -\sum_{i=t+1}^{L-1} \frac{h_i}{1 - P_t} \log_e \frac{h_i}{1 - P_t}$$

avec : $P_t = \sum_{i=0}^{t} h_i$

Il faut maximiser suivant t le terme $H_1 + H_2$.

Voir les figures B.11, B.27 et B.44 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.9 8 – Méthode entropique de Johannsen et Bille

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ normalisé $\left(\sum_{i=0}^{L-1} h_i = 1\right)$ de la fonction d'observation.

Elle est utilisée en traitement de l'image : elle est basée sur la minimisation de l'interdépendance, au sens de la théorie de l'information, entre deux groupes de pixels, classés suivant leurs niveaux de gris.

Le seuil x_{seuil} est égal au x_p (les x_p correspondent aux abscisses des cases de l'histogramme) qui minimise la fonction :

$$x_{seuil} = x_p \text{ tel que } \min_p(S_p + \bar{S}_p)$$

avec :

$$\begin{aligned} S_p &= \log_e \sum_{i=0}^{p} h_i - \left(h_p \log_e h_p + \left(\sum_{i=0}^{p-1} h_i \right) \log_e \sum_{i=0}^{p-1} h_i \right) / \sum_{i=0}^{p} h_i \\ \bar{S}_p &= \log_e \sum_{i=p}^{L-1} h_i - \left(h_p \log_e h_p + \left(\sum_{p+1}^{L-1} h_i \right) \log_e \sum_{p+1}^{L-1} h_i \right) / \sum_{i=p}^{L-1} h_i \end{aligned}$$

Voir les figures B.12, B.28 et B.45 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.10 9 – Méthode de la préservation des moments

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ (non normalisé) de la fonction d'observation.

La valeur du seuil est calculée de telle manière que les moments du signal à seuiller et les moments du signal seuillé (les échantillons de bruit sont mis à 0 et les échantillons correspondant à des transitions sont mis à 1) soient préservés.

Les moments m_α d'ordre α sont égaux à :

$$m_\alpha = \frac{1}{N_n} \sum_{i=0}^{L-1} i^\alpha h_i$$

où N_n est la taille du signal.

Alors, le t qui nous donne le seuil x_t est égal à :

$$t = \frac{z - m_1}{(c_1^2 - 4c_0)^{1/2}}$$

Avec :

$$c_0 = \frac{m_1 m_3 - m_2^2}{m_2 - m_1^2},\ c_1 = \frac{m_1 m_2 - m_3}{m_2 - m_1^2} \text{ et } z = \frac{1}{2} \left\{ (c_1^2 - 4c_0)^{\frac{1}{2}} - c_1 \right\}$$

Voir les figures B.13 et B.29 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.11 10 – Méthode de la superposition de deux gaussiennes

Cette méthode est basée sur l'histogramme $H = [h_0 \dots h_{L-1}]$ normalisé $\left(\sum_{i=0}^{L-1} h_i = 1\right)$ de la fonction d'observation.

Nous supposons que l'histogramme du signal est la somme de deux gaussiennes de densités de probabilité $N\left(\mu_1,\sigma_1^2\right)$ et $N\left(\mu_2,\sigma_2^2\right)$ et de probabilité a priori P_1 et P_2. Pour chaque case i de l'histogramme nous avons donc :

$$h_i = P_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right) + P_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)$$

La valeur du seuil est égale au x tel que :

$$\frac{(x-\mu_1)^2}{\sigma_1^2} + \log_e \sigma_1^2 - 2\log_e P_1 = \frac{(x-\mu_2)^2}{\sigma_2^2} + \log_e \sigma_2^2 - 2\log_e P_2$$

Hélas, nous ne connaissons pas μ_1, σ_1, P_1, μ_2, σ_2 et P_2.

Pour les déterminer, Kittler et Illingworth introduisent le critère :

$$J(t) = 1 + 2\left\{P_1(t)\log_e \sigma_1(t) + P_2(t)\log_e \sigma_2(t)\right\} - 2\left\{P_1(t)\log_e P_1(t) + P_2(t)\log_e P_2(t)\right\}$$

avec :

$$P_1(t) = \sum_{i=0}^{t} h_i \text{ et } P_2(t) = \sum_{i=t+1}^{L-1} h_i$$

$$\mu_1(t) = \left\{\sum_{i=0}^{t} i h_i\right\} / P_1(t) \text{ et } \mu_2(t) = \left\{\sum_{i=t+1}^{L-1} i h_i\right\} / P_2(t)$$

$$\sigma_1^2(t) = \left\{\sum_{i=0}^{t} (i-\mu_1(t))^2 h_i\right\} / P_1(t) \text{ et } \sigma_2^2(t) = \left\{\sum_{i=t+1}^{L-1} (i-\mu_2(t))^2 h_i\right\} / P_2(t)$$

Il faut minimiser ce critère suivant t. Alors, x_t est le seuil.

Voir les figures B.14, B.30 et B.46 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.12 11 – Méthode des deux segments de droites

Nous sommes parti de considérations sur la séparation en vecteurs propres de bruit et en vecteurs propres de signal d'une matrice d'autocorrélation. Les valeurs propres correspondantes, rangées dans l'ordre décroissant, sont représentées sur la figure B.1 pour laquelle nous avons 6 sources présentes. Voir [Ros94] et [FW88].

Nous voyons que la courbe obtenue peut être approximée par deux droites dont le point d'intersection correspond à la limite entre la partie signal et la partie bruit.

Nous appliquons dans notre cas une procédure similaire. Nous classons les échantillons du plus grand au plus petit, en conservant en index leur position temporelle. Nous appelons la courbe obtenue C. Puis nous cherchons à déterminer en quel endroit cette courbe se casse en deux : les plus petits échantillons correspondent aux zones stables du signal et les plus grands à des zones instables. La méthode pose les mêmes problèmes ici qu'aux techniques de séparation en un espace bruit et en un espace signal d'où elle vient : l'estimation du point de césure n'est pas évidente. Pour résoudre ce problème, nous modélisons C par deux segments de droites. L'un, s_1, partant du premier point du classement et s'arrêtant au point m. L'autre, s_2, partant de m et s'arrêtant au dernier point du classement. Si N_n est le nombre d'échantillons, nous faisons varier m de 1 à N_n.

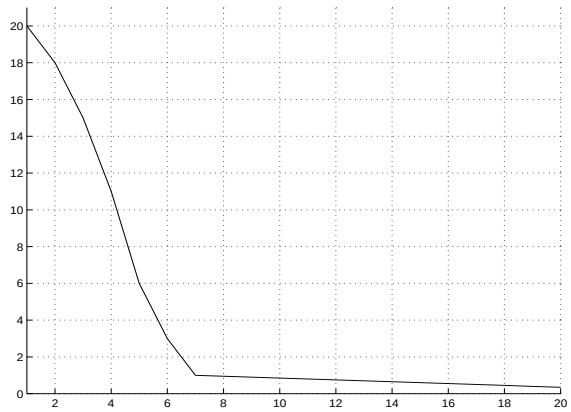


FIG. B.1 – *Séparation entre un espace signal et un espace bruit. En abscisse : le numéro de la valeur propre ; en ordonnée : la valeur de la valeur propre*

Pour effectuer cette modélisation, nous minimisons par les moindres carrés chacune des erreurs quadratiques :

$$E_1(m) = \sum_{i=1}^{m} [s_1(i) - C(i)]^2 \text{ et } E_2(m) = \sum_{i=m}^{N_n} [s_2(i) - C(i)]^2$$

Nous obtenons l'erreur quadratique totale : $E(m) = E_1(m) + E_2(m)$.

Nous constatons qu'il existe bien un minimum global pour $E(m)$ (et qu'il est même souvent l'unique minimum). Ce minimum survient pour un certain indice k. Nous positionnons alors le seuil à $C(k)$. Grâce aux index, nous pouvons poser les marques de segmentation.

La « valeur absolue de la dérivée de f_0 », entre 0,4 et 13,2 *secondes*, pour l'extrait de flûte, est représentée sur la figure B.2. La courbe $E(m)$ qui correspond à cette fonction d'observation est donnée sur la figure B.3. Le seuil trouvé vaut alors 27,8.

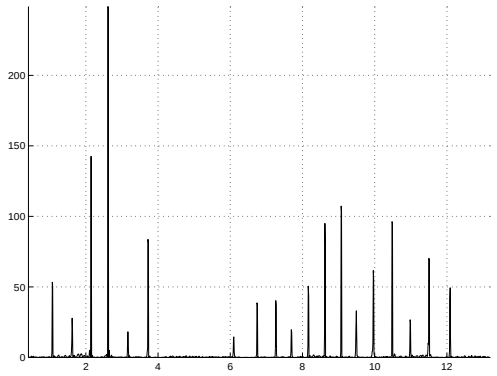


FIG. B.2 – *Trajet de la « valeur absolue de la dérivée de f_0 » pour l'extrait de flûte* **flute.sf**. *En abscisse : le temps en seconde*

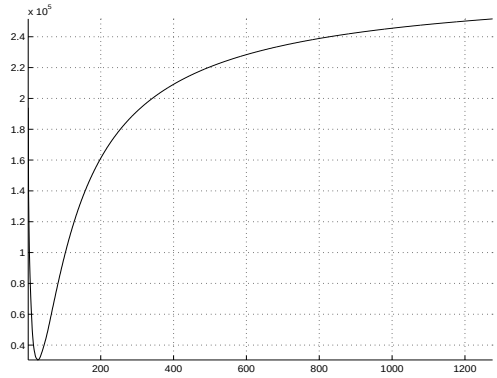


FIG. B.3 – *$E(m)$ pour la « valeur absolue de la dérivée de f_0 » pour l'extrait de flûte* **flute.sf**. *En abscisse : la position k ; en ordonnée : $E(m)$*

L'une des limitations de cette méthode est qu'elle est terriblement gourmande en temps de calcul.

Voir les figures B.15, B.31 et B.47 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.13 12 – Méthode SURE

Cette méthode est une méthode d'ondelettistes, utilisée pour le débruitage. Les ondelettistes ont en fait le même problème que nous. Les plus petits coefficients obtenus par une décomposition en paquets d'ondelettes correspondent à du bruit et les plus grands, qui sont beaucoup moins nombreux, à du signal. Voir [Don94] et [DJ94].

Le signal est : $x = [x_1 \dots x_{N_n}]$. Sa variance est σ^2.

Le seuil x_{seuil} est égal au λ qui minimise :

$$x_{seuil} = \min_{\lambda} \left[\sigma^2 \left(N_n - 2 \sum_{i=1}^{N_n} 1_{x_i<\lambda} \right) + \sum_{i=1}^{N_n} \min \left(x_i^2, \lambda^2 \right) \right]$$

où la fonction $1_{x_i<\lambda}$ renvoie 1 si x_i est inférieur à λ et 0 sinon.

Voir les figures B.16, B.32 et B.48 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.14 13 – Méthode des 3σ

Nous décidons que tous les pics supérieurs à 3 σ, où σ est l'écart-type du « bruit » (observé dans les zones stables), sont des pics ne correspondant pas à du « bruit » (correspondant donc à des transitions).

Il y a une justification théorique à ce seuil, avec le « seuil universel » (« universal threshold »), défini par Donoho dans [Don94] et [DJ94]. Ce seuil est égal à $\sigma\sqrt{2\log_e(M)}$, où M est la taille du signal en nombre d'échantillons. Donoho prouve que quand M tend vers l'infini la probabilité qu'un échantillon de bruit dépasse ce seuil tend vers 0. Et $\sigma\sqrt{2\log_e(M)}$, pour un M de valeur « raisonnable » pour nous, c'est-à-dire de l'ordre de 1000, par exemple, est très proche de 3 (3,72).

Le problème est de parvenir à déterminer σ correctement. Pour ce faire, nous classons les échantillons en ordre croissant et nous calculons σ avec les n % plus petits (avec n égale à 90, par exemple : comme il est montré dans la section B.21 sur la figure B.56, nous obtenons des résultats consistants, c'est-à-dire que la méthode est relativement robuste à n). Cependant, le seuil dépend de M, la taille du signal, ce qui n'est pas très intéressant.

Voir les figures B.17, B.33 et B.49 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.15 14 – Méthode de Sahoo

Dans leur article de revue (« survey ») des techniques de seuillage, Sahoo, Soltani et Wong donnent encore une autre méthode de seuillage. Elle est basée sur la *mesure de l'uniformité* :

$$U(t) = 1 - \frac{\sigma_1^2(t) + \sigma_2^2(t)}{C}$$

où t est la position du seuil (variant entre 1 et $N_n - 1$, N_n étant la taille du signal) ; C un « facteur de normalisation » ; $\sigma_1^2(t)$ est la variance du premier mode, supposé comprenant tous les échantillons inférieurs ou égaux à $x(t)$; et $\sigma_2^2(t)$ est la variance du second mode, supposé comprenant tous les échantillons strictement supérieurs à $x(t)$. Il s'agit de maximiser cette mesure. En fait, il est tout à fait équivalent de minimiser suivant t cette fonction :

$$V(t) = \sigma_1^2(t) + \sigma_2^2(t)$$

Voir les figures B.18, B.34 et B.50 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.16 15 – Méthode isodata

Cette méthode est itérative. À l'origine, elle est utilisée avec l'histogramme, mais elle peut être adaptée pour être utilisée directement avec les échantillons ordrés. La taille du signal est N_n. Les échantillons sont : $[x(1) \quad \dots \quad x(N_n)]$. L'algorithme est le suivant :

1. Nous avons, au départ (itération $i = 0$), $\text{seuil}_0 = x\left(\frac{N_n}{2}\right)$. C'est-à-dire que n_0, la position du seuil initial, est posée à $\frac{N_n}{2}$.
2. $i = i + 1$
3. Nous calculons les moyennes :

   $m_i^{(\text{bas})} = \text{moyenne de } [x(1) \quad \dots \quad x(n_{i-1})]$ et $m_i^{(\text{haut})} = \text{moyenne de } [x(n_{i-1}+1) \quad \dots \quad x(N_n)]$
4. La nouvelle valeur du seuil est alors : $\text{seuil}_i = \frac{\left(m_i^{(\text{bas})} + m_i^{(\text{haut})}\right)}{2}$.
5. Puis nous cherchons la valeur n_i telle que $x(n_i) \leq \text{seuil}_i \leq x(n_i + 1)$.
6. Si $n_i \neq n_{i-1}$, retour à l'étape 2
   Sinon, nous nous arrêtons et la valeur du seuil est $\text{seuil} = x(n_i)$

Voir les figures B.19, B.35 et B.51 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.17 16 – Méthode « symétrie » : forme 1

Cette méthode et la suivante sont basées sur l'étude de l'histogramme. Ces deux méthodes, basées sur la « symétrie », font l'hypothèse que l'un des deux modes produit un pic dominant : c'est-à-dire que la probabilité a priori de ce mode est plus grande que celle de l'autre, et/ou sa variance est plus petite que celle de l'autre. Ceci correspond à la plupart des cas étudiés par les tests définis dans la section B.21 : ce n'est pas vrai quand les deux modes ont autant de points. Cependant, le maximum qui nous intéresse appartenant toujours à la première partie de l'histogramme, même ce cas peut être traité par la méthode.

Le maximum de l'histogramme est détecté. Il a lieu pour la case numérotée i. Nous cherchons la case j, avec $j < i$ telle qu'au moins 95 % (dans le cas général : n %) des points compris entre la case 1 et la case i (y compris la case 1 et la case i) soient compris entre la case j et la case i (y compris la case j et la case i). La position du seuil est alors $p = i + (i - j)$ et la valeur du seuil est x_p. Ceci est illustré sur la figure B.4. En fait, nous avons constaté que le seuil trouvé est systématiquement trop petit, alors nous lui ajoutons une valeur, arbitraire : pour obtenir les résultats présentés dans cet exposé, nous l'avons fixée à 0,6. Nous devrions faire dépendre cette valeur du nombre de points dans le second mode (voir la figure B.20), de n, etc. : il s'agit de perspectives.

Voir les figures B.20, B.36 et B.52 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.18 17 – Méthode « symétrie » : forme 2

Cette fois, la position p nous permet d'obtenir une estimée de la variance du « bruit » : $\sigma^2 = \frac{1}{N_n}\sum_{i=1}^{p} n_i(x_i - m)^2$, avec N_n le nombre de points compris dans les cases de 1 à p, $m = \frac{1}{N_n}\sum_{i=1}^{p} n_i x_i$, n_i le nombre de points dans la $i^{\text{ème}}$ case de l'histogramme et x_i la coordonnée de cette case. Nous pouvons donc relier cette méthode à la règle des 3σ. Nous estimons σ, puis le seuil est fixé à 3σ (ou mieux à $3\sigma + m$, où m est la moyenne du bruit). Ainsi, nous nous affranchissons de ce n (voir la section B.14) qui correspondait à un seuil fixe, fixé arbitrairement.

Une perspective est envisagée : nous pourrions évaluer la variance du bruit en utilisant directement les échantillons du signal inférieurs à x_p.

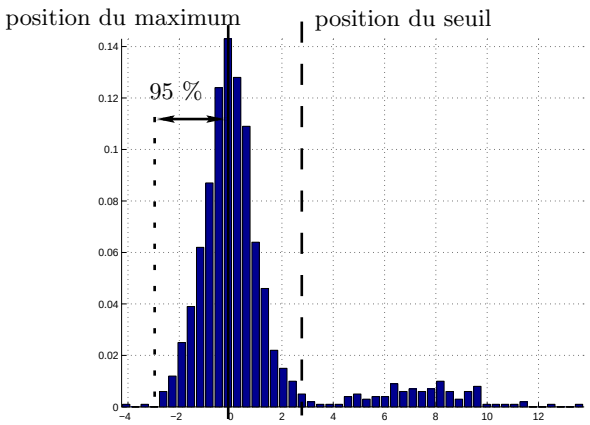


FIG. B.4 – *Histogramme – méthode « symétrie »*

Voir les figures B.21, B.37 et B.53 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.19 18 – Méthode du triangle

Cette méthode est basée elle aussi sur l'étude de l'histogramme (composé de N cases). Nous détectons le maximum de l'histogramme. Sa position est m, et sa valeur val_m : $[m, \text{val}_m]$. Nous relions par un segment de droite ce maximum au dernier point de l'histogramme ($[N, \text{val}_N]$). Pour chaque case i ($[i, \text{val}_i]$) tel que $i > m$ nous calculons sa distance euclidienne minimale d_i au segment de droite. Le plus grand des d_i, d_{max}, est obtenu pour la case p. Alors la position du seuil est val_p.

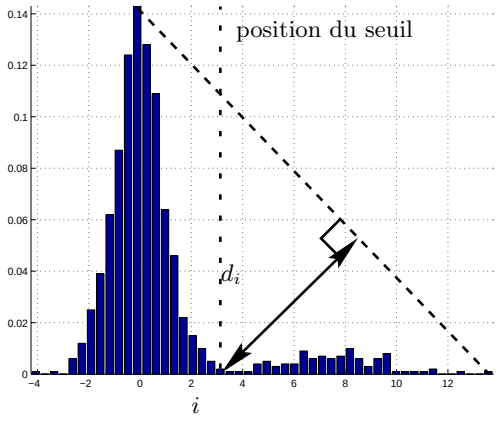


FIG. B.5 – *Histogramme – méthode du triangle*

Voir les figures B.22 (nous remarquons que quand les deux modes ont à peu près le même nombre de points, les performances de la méthode se dégradent : en fait, cette méthode est particulièrement adaptée au cas où l'un des modes est rare et a une variance grande, c'est-à-dire à nos fonctions d'observation), B.38 et B.54 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.20 19 – Quelques remarques, qui mènent à la méthode de SAHOO améliorée

B.20.1 Méthode

Nous prouvons (voir l'annexe F) que la variance de l'estimée à partir de N_n échantillons ($[x_1 \ \dots \ x_{N_n}]$) de la variance σ^2 d'un bruit gaussien $\mathcal{N}\left(m,\sigma^2\right)$ est égale à :

$$\mathrm{Var}\left[\hat{\sigma}^2_{(1)}\right] = 2\sigma^4\frac{N-1}{N^2}$$

dans le cas biaisé, c'est-à-dire quand nous estimons la variance ainsi :

$$\hat{\sigma}^2_{(1)} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \sum_{j=1}^{N}x_j\right)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}x_ix_j$$

et qu'elle est égale à :

$$\mathrm{Var}\left[\hat{\sigma}^2_{(1)}\right] = 2\sigma^4\frac{1}{N-1}$$

dans le cas non biaisé, c'est-à-dire quand nous estimons la variance ainsi :

$$\hat{\sigma}^2_{(2)} = \frac{1}{N_n-1}\sum_{i=1}^{N_n}\left(x_i - \sum_{j=1}^{N_n}x_j\right)^2 = \frac{1}{N_n-1}\sum_{i=1}^{N_n}x_i^2 - \frac{1}{N_n(N_n-1)}\sum_{i=1}^{N_n}\sum_{j=1}^{N_n}x_ix_j.$$

Elle dépend donc d'un terme en $\frac{1}{N_n}$. L'idée est d'ajouter un terme qui dépend du nombre de points utilisés pour calculer chacune des deux variances utilisées dans la *mesure de l'uniformité* (voir la section B.15) : plus le nombre de points est grand, plus nous pouvons faire confiance en l'estimée de la variance. Ainsi, il s'agit à présent de minimiser suivant t l'expression :

$$W(t) = t^\alpha\hat{\sigma}_1^2(t) + (N_n-t)^\alpha\hat{\sigma}_2^2(t)$$

Nous avons choisi $\alpha = 0{,}7$.

Voir les figures B.23, B.39 et B.55 pour les résultats de la méthode avec les tests définis dans la section B.21.

B.20.2 Perspective

Nous constatons que pour le premier test (voir la section B.21.3) cet estimateur n'est pas biaisé, ni quand il estime le nombre de bonnes détections, ni quand il estime le nombre de fausses alarmes. Par rapport à la méthode de SAHOO originale, l'amélioration, pour les trois tests, est nette. Nous pourrions aussi tenir compte du terme en σ^4, en prenant des puissances inférieures à 2 pour les écarts-types. Nous aurions, dans le cas général :

$$W_{\alpha,\ \beta}(t) = t^\alpha\hat{\sigma}_1^\beta(t) + (N_n-t)^\alpha\hat{\sigma}_2^\beta(t)$$

B.21 Les tests

B.21.1 Introduction

Trois problèmes se posent :

- L'un des modes est pauvre en points : le premier test vise à étudier la résistance des méthodes de seuillage à ce problème.
- La variance de ce même mode est plus grande que la variance de l'autre : le second test vise à étudier la résistance des méthodes de seuillage à ce problème.
- La distance (soit encore, à variances constantes, la différence entre la moyenne du premier mode et la moyenne du second mode) entre les deux modes varie (d'une fonction d'observation à l'autre) : le troisième test vise à étudier la résistance des méthodes de seuillage à ce problème.

B.21.2 Technique de comparaison

Nous avons simulé ce signal :

- Taille du signal $N_n = N_b + N_s$.
- Nombre de points du signal correspondant à des pics à détecter N_s. Ils suivent une loi normale $\mathcal{N}\left(m_1,\sigma_1^2\right)$.
- Nombre de points du signal correspondant à du bruit $N_b \gg N_s$. Ils suivent une loi normale $\mathcal{N}\left(m_2,\sigma_2^2\right)$.

Puis nous avons utilisé chacune des méthodes de seuillage pour seuiller ce signal. Nous décidons que tous les échantillons supérieurs à ce seuil correspondent à des transitions. Puisque nous savons les positions des échantillons qu'il faut détecter (disons, les N_s premiers sur N_n) et les positions des points réellement détectés, nous pouvons calculer le nombre de fausses alarmes (un échantillon de bruit a été donné comme étant un échantillon de transition) et le nombre de bonnes détections obtenus.

B.21.3 Premier test – L'un des modes est pauvre en points

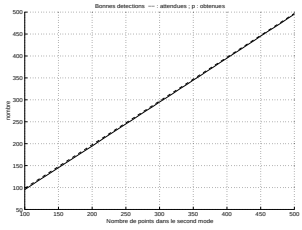
B.21.3.1 Définition des paramètres du signal

Nous avons pris :

- $N_n = 1000$ et N_s variant de 100 à 500
- $m_1 = 5$ et $\sigma_1^2 = 1$
- $m_2 = 0$ et $\sigma_2^2 = 1$

B.21.3.2 Résultats du premier test

Les résultats sont présentés sur les figures B.6, B.7, B.8, B.9, B.10, B.11, B.12, B.13, B.14, B.15, B.16, B.17, B.18, B.19, B.20, B.21, B.22 et B.23.



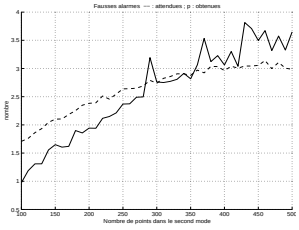
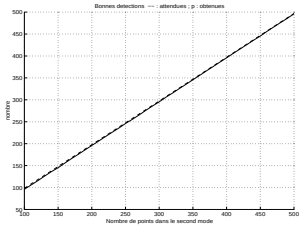


FIG. B.6 – *Méthode du saut – Premier test*



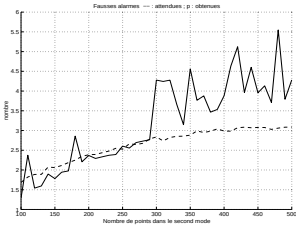
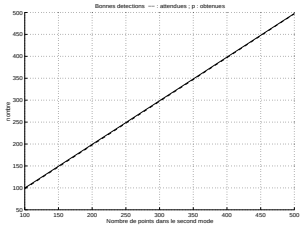


FIG. B.7 – *Détection des deux modes de l'histogramme – Premier test*



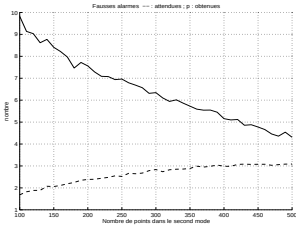
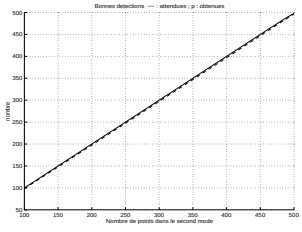


FIG. B.8 – *Méthode d'*OTSU – *Premier test*



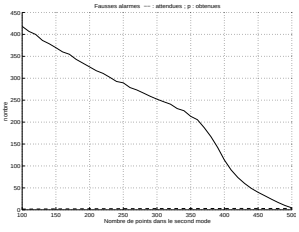


FIG. B.9 – *Première méthode de* PUN – *Premier test*

Les courbes de droite correspondent au nombre de bonnes détections obtenu (trait plein) en fonction de N_s. Les courbes de gauche représentent le nombre de fausses alarmes obtenu (trait plein) en fonction de N_s.

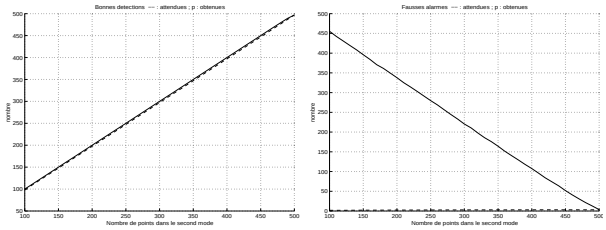


FIG. B.10 – *Deuxième méthode de* PUN – *Premier test*

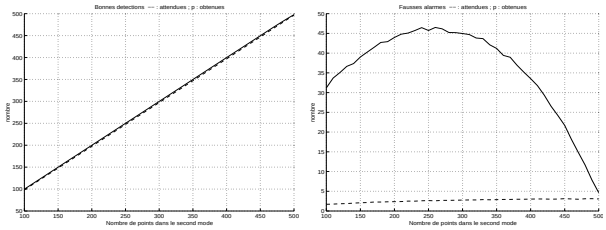


FIG. B.11 – *Méthode de* KAPUR, SAHOO *et* WONG – *Premier test*

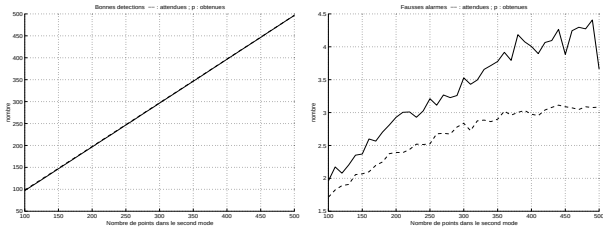


FIG. B.12 – *Méthode de* JOHANNSEN *et* BILLE – *Premier test*

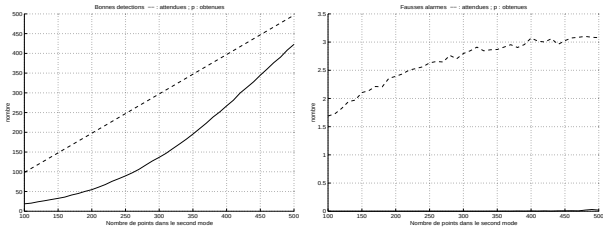


FIG. B.13 – *Méthode des moments – Premier test*

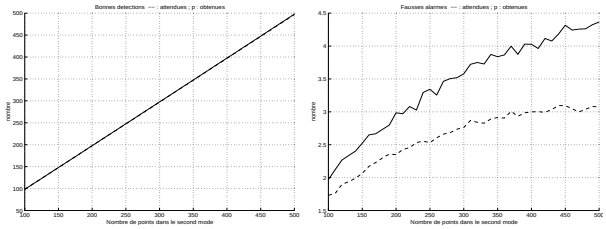


FIG. B.14 – *Méthode des deux gaussiennes – Premier test*

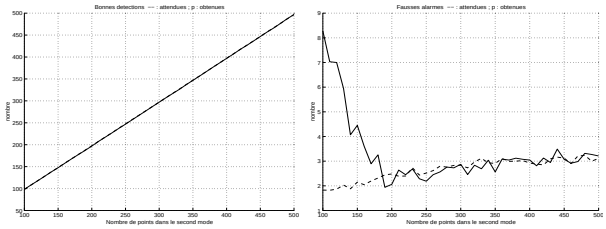


FIG. B.15 – *Méthode des deux droites – Premier test*

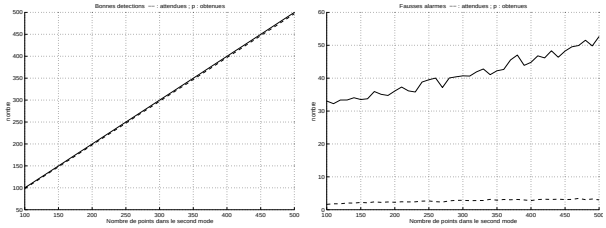


FIG. B.16 – *Méthode* SURE – *Premier test*

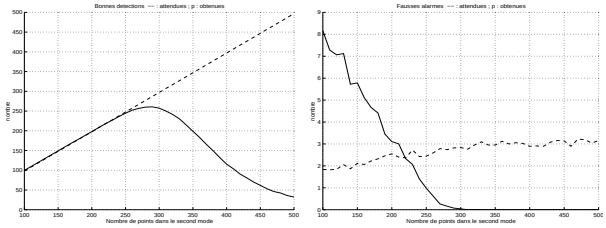


FIG. B.17 – *Règle des* 3σ – *Premier test*

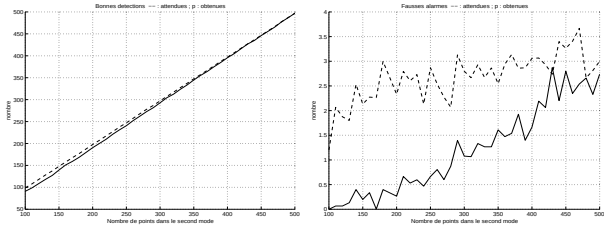


FIG. B.18 – *Méthode de* SAHOO – *Premier test*

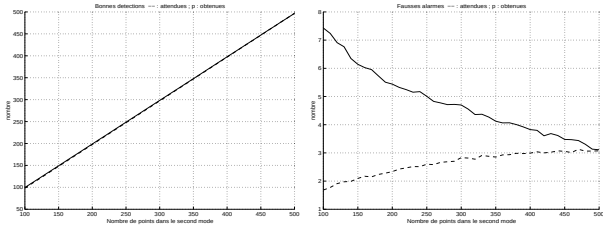


FIG. B.19 – *Méthode isodata – Premier test*

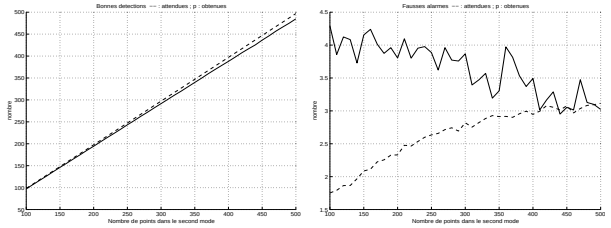


FIG. B.20 – *Méthode symétrie 1 – Premier test*

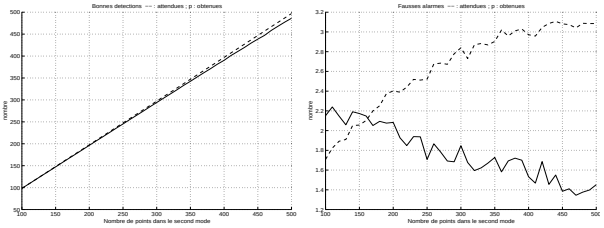


FIG. B.21 – *Méthode symétrie 2 – Premier test*

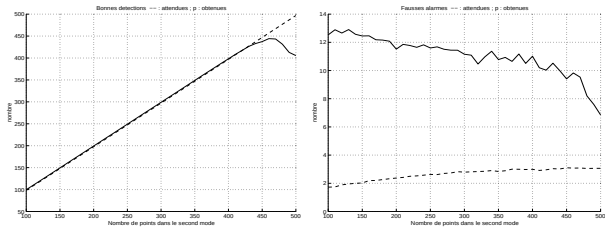


FIG. B.22 – *Méthode du triangle – Premier test*

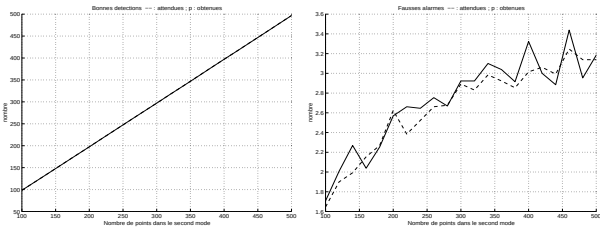


FIG. B.23 – *Méthode de* SAHOO *améliorée – Premier test*

Les courbes en pointillés représentent le nombre de bonnes détections attendu (à droite) et le nombre de fausses alarmes attendu (à gauche). Pour obtenir ces nombres, nous considérons que nous connaissons le nombre de marques à trouver et nous comptons, pour avoir le nombre de fausses alarmes, le nombre de pics de bruit plus grands que le plus petit pic de signal. En fait, toutes ces courbes en pointillés correspondent aux résultats que nous obtiendrions idéalement [1] avec la cinquième méthode de seuillage (« Nombre de marques connu, ou méthode du pourcentile » : voir la section B.6).

Les résultats donnés sont les moyennes obtenus pour un grand nombre de coups (1000 signaux de N_n échantillons).

Il est normal que la méthode des 3σ échoue, puisque l'importance relative des deux modes varie énormément. Les résultats donnés sont ceux obtenus pour $n = 75$ %. Nous conservons cette méthode dans la suite.

D'autres méthodes ont échoué. Il s'agit des deux méthodes entropiques de PUN (voir les sections B.5 et B.7). Ces méthodes sont adaptées aux cas où les deux modes sont équiprobables ou presque équiprobables. Nous ne les utiliserons plus dans la suite.

B.21.4 Deuxième test – La variance du mode rare est plus grande que la variance de l'autre

B.21.4.1 Définition des paramètres du signal

Nous avons pris :

- $N_b = 900$ et $N_s = 100$
- $m_1 = 5$ et σ_1^2 varie entre 1 et 4
- $m_2 = 0$ et $\sigma_2^2 = 1$

B.21.4.2 Résultats du deuxième test

Les résultats sont présentés sur les figures B.24, B.25, B.26, B.27, B.28, B.29, B.30, B.31, B.32, B.33, B.34, B.35, B.36, B.37, B.38, B.39.

Une méthode a échoué. Il s'agit de la méthode « Préservation des moments » (voir la section B.10). Nous ne l'utiliserons plus dans la suite.

1. C'est-à-dire si nous connaissions exactement le nombre de marques à trouver !

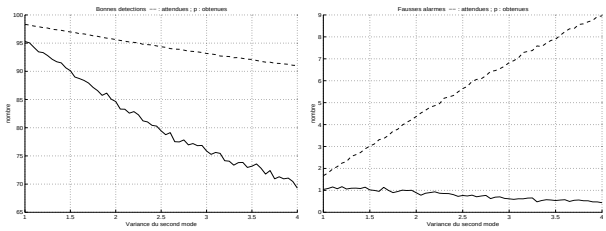


FIG. B.24 – *Méthode du saut – Deuxième test*

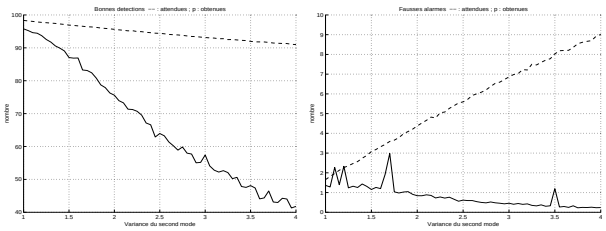


FIG. B.25 – *Détection des deux modes de l'histogramme – Deuxième test*

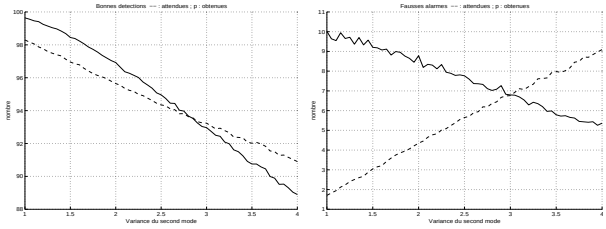


FIG. B.26 – *Méthode d'*OTSU *– Deuxième test*

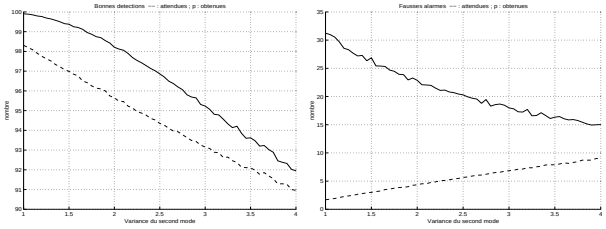


FIG. B.27 – *Méthode de* KAPUR, SAHOO *et* WONG *– Deuxième test*

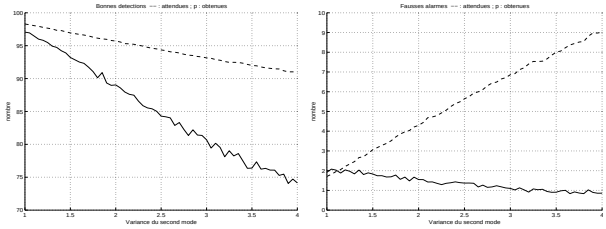


FIG. B.28 – *Méthode de* JOHANNSEN *et* BILLE *– Deuxième test*

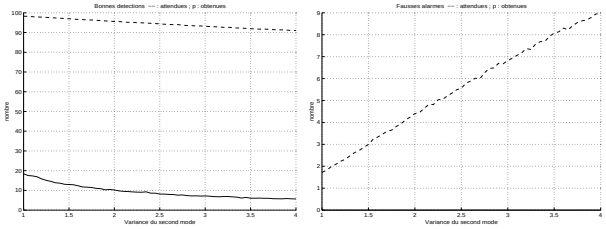


FIG. B.29 – *Méthode des moments – Deuxième test*

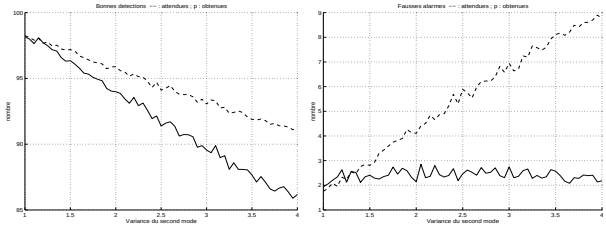


FIG. B.30 – *Méthodes des deux gaussiennes – Deuxième test*

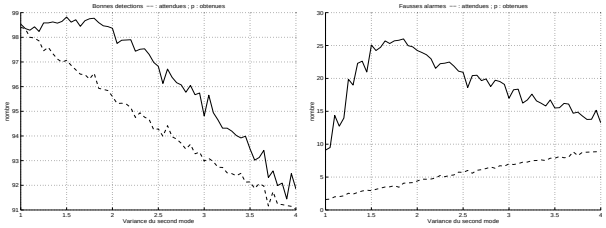


FIG. B.31 – *Méthode des deux droites – Deuxième test*

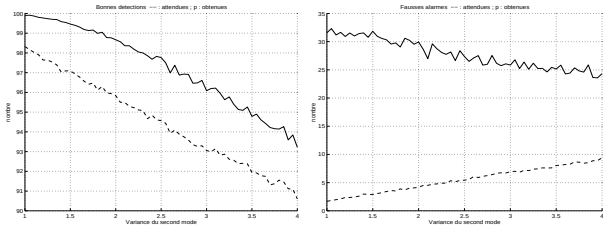


FIG. B.32 – *Méthode* SURE *– Deuxième test*

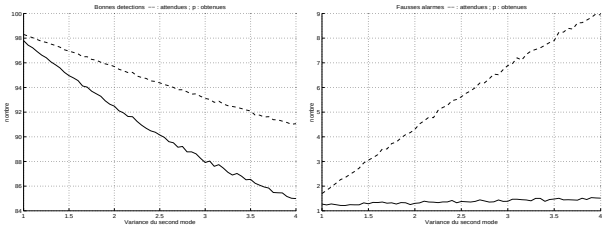


FIG. B.33 – *Règle des* 3σ *– Deuxième test*

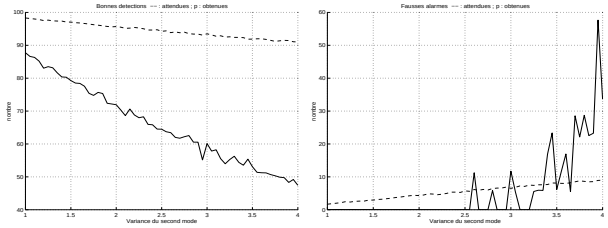


FIG. B.34 – *Méthode de* SAHOO *– Deuxième test*

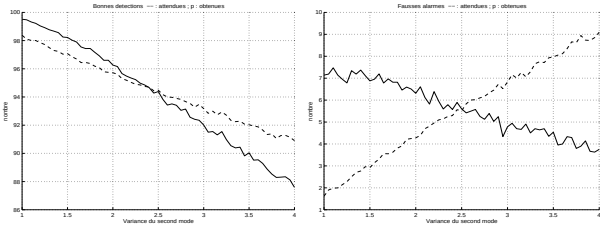


FIG. B.35 – *Méthode isodata – Deuxième test*

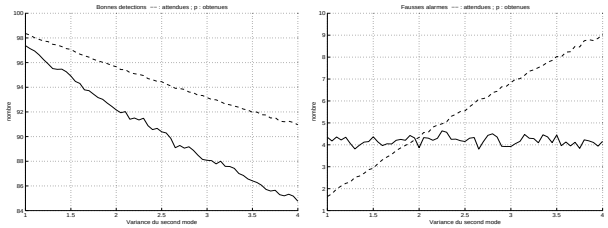


FIG. B.36 – *Méthode symétrie 1 – Deuxième test*

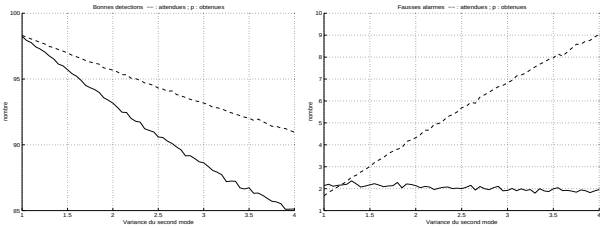


FIG. B.37 – *Méthode symétrie 2 – Deuxième test*

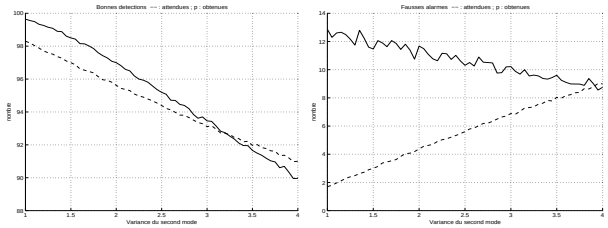


FIG. B.38 – *Méthode du triangle – Deuxième test*

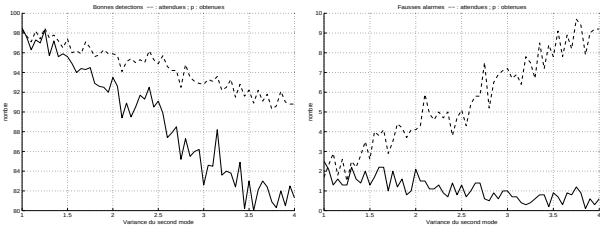


FIG. B.39 – *Méthode de* SAHOO *améliorée – Deuxième test*

B.21.5 Troisième test – La distance entre les deux modes varie

B.21.5.1 Définition des paramètres du signal

Nous avons pris :

- $N_b = 900$ et $N_s = 100$
- $m_1 = m$ variable (entre 3 et 10) et $\sigma_1^2 = 4$
- $m_2 = 0$ et $\sigma_2^2 = 1$

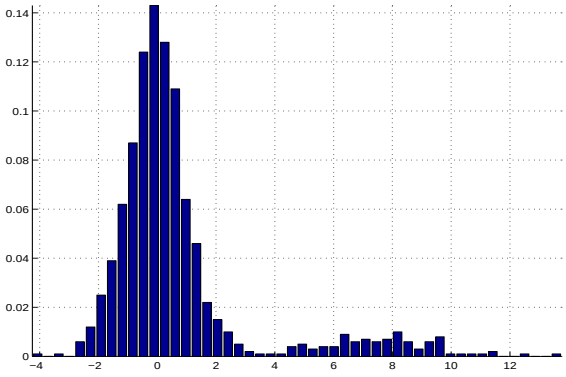


FIG. B.40 – *Exemple d'histogramme*

B.21.5.2 Résultats du troisième test

Les résultats sont présentés sur les figures B.41, B.42, B.43, B.44, B.45, B.46, B.47, B.48, B.49, B.50, B.51, B.52, B.53, B.54, B.55 et sur la figure B.56.

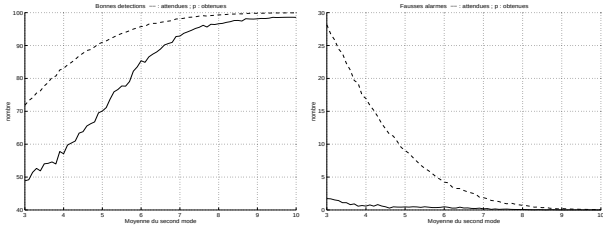


FIG. B.41 – *Méthode du saut – Troisième test*

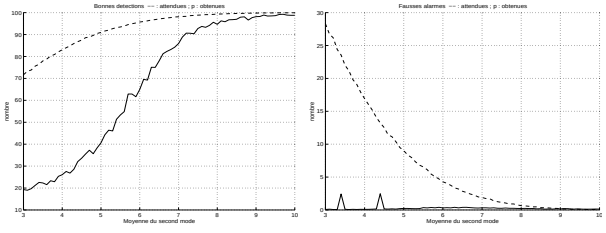


FIG. B.42 – *Détection des deux modes de l'histogramme – Troisième test*

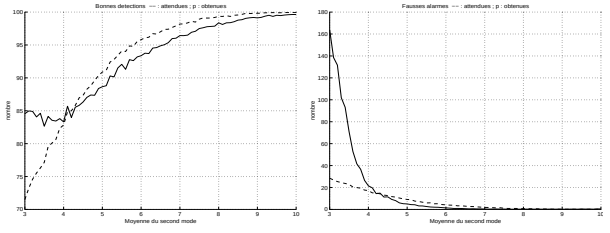


FIG. B.43 – *Méthode d'*OTSU *– Troisième test*

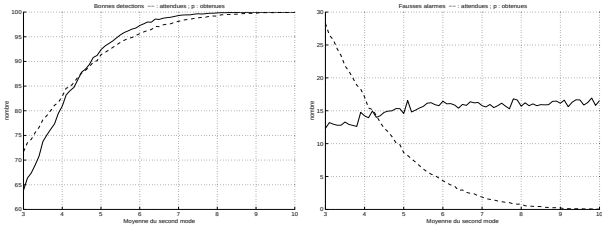


FIG. B.44 – *Méthode de* KAPUR*,* SAHOO *et* WONG *– Troisième test*

Considérons la figure B.56. À gauche, la courbe en pointillés donne le nombre de bonnes détections que nous pouvons espérer. Plus m est petit plus ce nombre est petit. En effet, plus

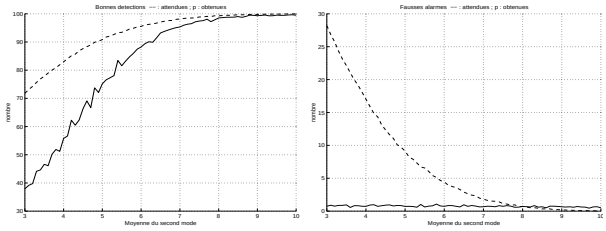


FIG. B.45 – *Méthode de* JOHANNSEN *et* BILLE *– Troisième test*

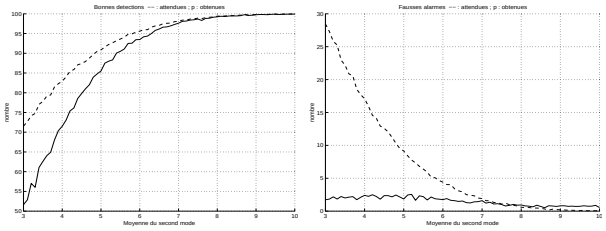


FIG. B.46 – *Méthode des deux gaussiennes – Troisième test*

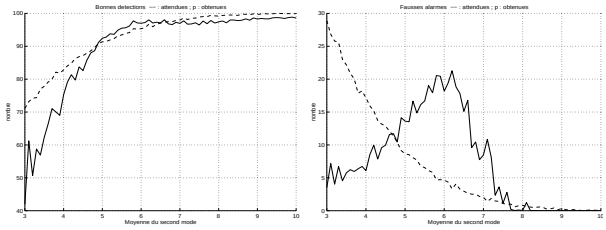


FIG. B.47 – *Méthode des deux droites – Troisième test*

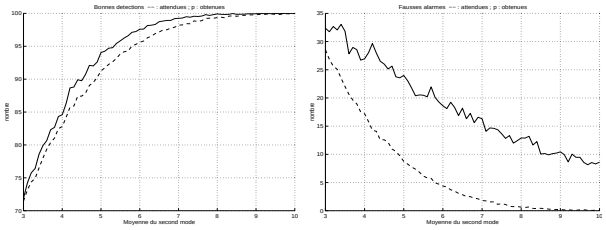


FIG. B.48 – *Méthode* SURE *– Troisième test*

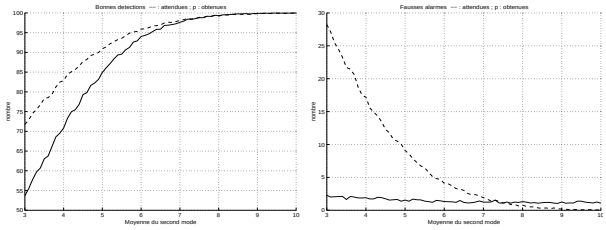


FIG. B.49 – *Règle des* 3σ *– Troisième test*

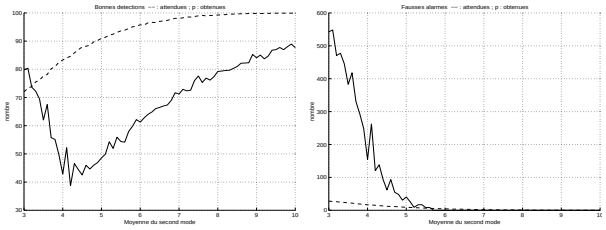


FIG. B.50 – *Méthode de* SAHOO *– Troisième test*

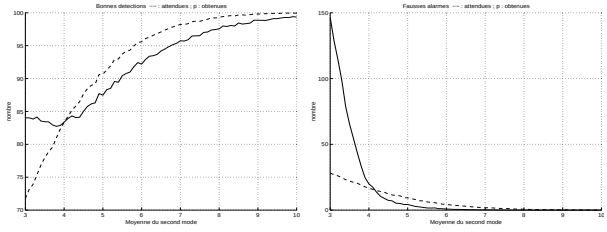


FIG. B.51 – *Méthode isodata – Troisième test*

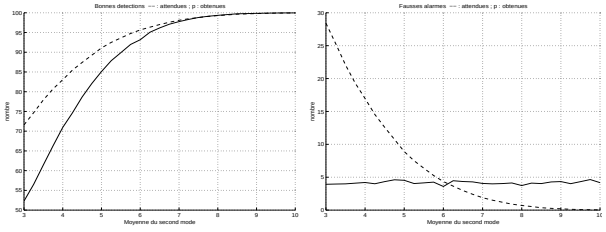


FIG. B.52 – *Méthode symétrie 1 – Troisième test*

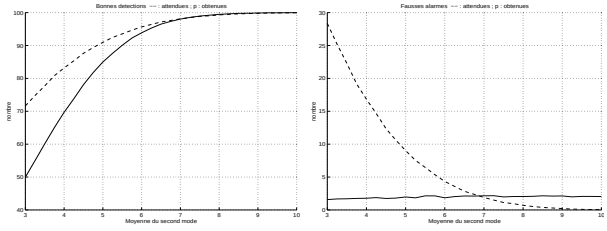


FIG. B.53 – *Méthode symétrie 2 – Troisième test*

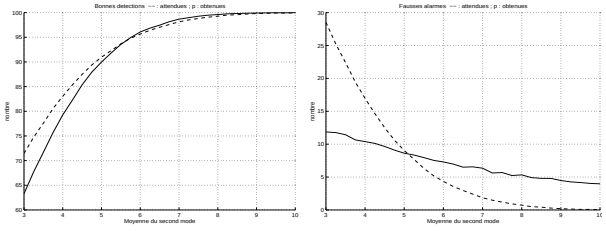


FIG. B.54 – *Méthode du triangle – Troisième test*

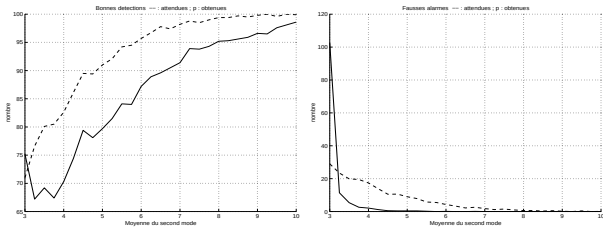


FIG. B.55 – *Méthode de* SAHOO *améliorée – Troisième test*

m est petit plus il y a de pics de bruit plus grands que les plus petits pics de signal. Pour $m = 4$, nous sommes à 84 environ. À droite, la courbe en pointillés donne le nombre de fausses alarmes obtenues si nous connaissions le nombre de pics à détecter : ce nombre est donc 100 moins le nombre de bonnes détections.

En haut, nous présentons les résultats obtenus avec la règle des 3σ, pour différents n. Plus n est petit, plus σ est sous-estimé, et donc plus le nombre de fausses alarmes augmente. Nous voyons que la méthode reste robuste, même quand nous faisons une grande erreur sur n.

Et en bas, nous présentons les résultats obtenus avec huit autres méthodes de seuillage. Parmi celles-ci, la plus classique, c'est-à-dire celle qui est basée sur la détection des deux modes de l'histogramme (voir la section B.3), est la moins fiable.

B.22 Conclusion

Dans le programme *segmentation*, seul le seuillage à 3σ a été implémenté (avec $n = 90\%$).

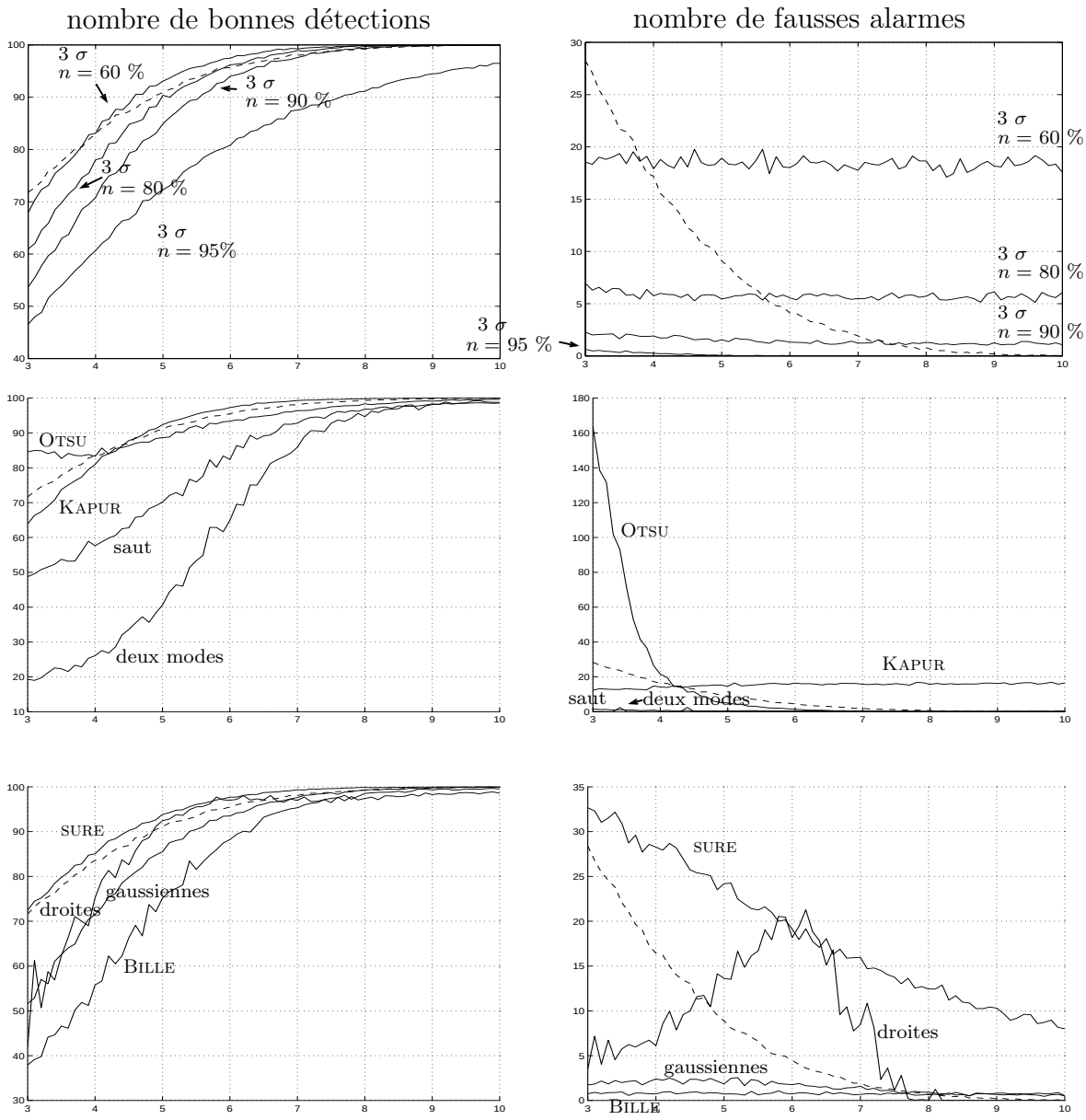


FIG. B.56 – *Performances comparées de différentes méthodes de seuillage*

Annexe C

Dérivée première des fenêtres de pondération

C.1 La fenêtre de pondération RECTANGULAIRE

C.1.1 Fenêtre dans le domaine temporel

$$W_R(t) = \mathcal{J}\left[-\frac{T}{2}, \frac{T}{2}\right]$$

C.1.2 Fenêtre de pondération dans le domaine fréquentiel

$$\hat{W}_R(f) = T\frac{\sin(\pi f T)}{\pi f T}$$

C.1.3 Dérivée de la fenêtre de pondération dans le domaine fréquentiel

$$\frac{\partial \hat{W}_R}{\partial f}(f) = T\left[\frac{\cos(\pi f T)}{f} - \frac{\sin(\pi f T)}{\pi f T}\frac{1}{f}\right]$$

C.2 Sommes de N cosinus

Ceci concerne par exemple les fenêtres de pondération de BLACKMAN, HAMMING, HANNING... (voir [Har78]).

C.2.1 Fenêtre de pondération dans le domaine temporel

$$W_S(t) = \mathcal{J}\left[-\frac{T}{2}, \frac{T}{2}\right]\sum_{i=0}^{N} a_i \cos\left(2\pi\frac{t}{T}i\right)$$

C.2.2 Fenêtre de pondération dans le domaine fréquentiel

$$\hat{W}_S(f) = \sum_{i=0}^{N}\frac{a_i}{2}\left\{\hat{W}_R\left(f - \frac{i}{T}\right) + \hat{W}_R\left(f + \frac{i}{T}\right)\right\}$$

C.2.3 Dérivée de la fenêtre de pondération dans le domaine fréquentiel

Comme $\frac{\partial}{\partial f}(h \star g) = h \star \frac{\partial g}{\partial f}$, où $\star$ représente la convolution, nous avons :

$$\frac{\partial \hat{W}_S}{\partial f}(f) = \sum_{i=0}^{N} \frac{a_i}{2} \left\{ \frac{\partial \hat{W}_R}{\partial f}\left(f - \frac{i}{T}\right) + \frac{\partial \hat{W}_R}{\partial f}\left(f + \frac{i}{T}\right) \right\}$$

C.3 La fenêtre de pondération de POISSON

C.3.1 Fenêtre de pondération dans le domaine temporel

$$W_P(t) = \exp\left(-\alpha \frac{|t|}{T/2}\right) \mathcal{J}\left[-\frac{T}{2}, \frac{T}{2}\right]$$

C.3.2 Fenêtre de pondération dans le domaine fréquentiel

Nous donnons le calcul plutôt compliqué qui nous permet d'aboutir à la transformée de FOURIER de la fenêtre de pondération de POISSON dans l'annexe D.

$$\hat{W}_P(f) = \frac{\alpha T}{\alpha^2 + (\pi f T)^2} \left\{ 1 + \left[\frac{\pi f T}{\alpha} \sin(\pi f T) - \cos(\pi f T) \right] \exp(-\alpha) \right\}$$

C.3.3 Dérivée de la fenêtre de pondération dans le domaine fréquentiel

$$\begin{aligned} \frac{\partial \hat{W}_P}{\partial f}(f) \quad = \quad & \\ & -\frac{2\pi^2 T^2 f}{\alpha^2 + (\pi f T)^2} \hat{W}_P + \frac{\alpha T}{\alpha^2 + (\pi f T)^2} \left\{ \frac{\pi T}{\alpha} \sin(\pi f T) + \frac{\pi^2 f T^2}{\alpha} \cos(\pi f T) + \pi T \sin(\pi f T) \right\} \exp(-\alpha) \end{aligned}$$

C.4 La fenêtre de pondération de HANNING-POISSON

C.4.1 Fenêtre de pondération dans le domaine temporel

$$W_{HP}(t) = \exp\left(-\alpha \frac{|t|}{T/2}\right) \left[0{,}5 + 0{,}5 \cos\left(\pi \frac{t}{T/2}\right)\right] \mathcal{J}\left[-\frac{T}{2}, \frac{T}{2}\right]$$

C.4.2 Fenêtre de pondération dans le domaine fréquentiel

$$\hat{W}_{HP}(f) = 0{,}5 \hat{W}_P(f) + 0{,}25 \hat{W}_P\left(f - \frac{1}{T}\right) + 0{,}25 \hat{W}_P\left(f + \frac{1}{T}\right)$$

C.4.3 Dérivée de la fenêtre de pondération dans le domaine fréquentiel

$$\frac{\partial \hat{W}_{HP}}{\partial f}(f) = 0{,}5 \frac{\partial \hat{W}_P}{\partial f}(f) + 0{,}25 \frac{\partial \hat{W}_P}{\partial f}\left(f - \frac{1}{T}\right) + 0{,}25 \frac{\partial \hat{W}_P}{\partial f}\left(f + \frac{1}{T}\right)$$

Annexe D

La fenêtre de pondération de Poisson

La fenêtre de pondération de Poisson dans le domaine temporel a pour expression :

$$F(n) = \exp\left(-\alpha\frac{|n|}{T/2}\right)\mathcal{J}\left[-\frac{T}{2},\frac{T}{2}\right]$$

où $\mathcal{J}$ est la fenêtre rectangulaire, de largeur T, centrée en 0, et valant 1 entre $-\frac{T}{2}$ et $\frac{T}{2}$.

La transformée de Fourier du premier terme, qui est la fonction de Poisson, est :

$$TF_P(f) = \frac{2a}{a^2 + (2\pi f)^2}$$

avec $a = \frac{2\alpha}{T}$; et celle du second terme est :

$$TF_R(f) = T\frac{\sin(\pi fT)}{\pi fT}$$

La transformée de Fourier TF_F du signal total est :

$$TF_F(f) = T\frac{\sin(\pi fT)}{\pi fT} \star \frac{2a}{a^2 + (2\pi f)^2}$$

où $\star$ représente la convolution.

Soit :

$$TF_F(f) = \int_{-\infty}^{+\infty} \underbrace{T\frac{\sin(\pi gT)}{\pi gT}\frac{2a}{a^2 + [2\pi(f-g)]^2}}_{E} dg = I$$

Nous allons dans cette annexe résoudre cette intégrale.

Nous décomposons, avec l'aide de Maple (voir [Red93]), l'expression E sous l'intégrale, en éléments simples. Nous obtenons :

$$E = \underbrace{\underbrace{\frac{2aT}{a^2 + (2\pi f)^2}}_{A}\frac{\sin(\pi gT)}{\pi gT}}_{E_1} \underbrace{-\frac{8a\pi}{a^2 + (2\pi f)^2}\frac{\sin(\pi gT)(-2f+g)}{a^2 + [2\pi(f-g)]^2}}_{E_2}$$

Nous avons $I = I_1 + I_2 = \int_{-\infty}^{+\infty} E_1 dg + \int_{-\infty}^{+\infty} E_2 dg$.

D.1 Intégrons E_1

Nous avons : $\int_{-\infty}^{+\infty} \frac{\sin(\pi g T)}{\pi g T} = \frac{1}{T}$

Donc $I_1 = \frac{A}{T}$.

D.2 Intégrons E_2

$$E_2 = \underbrace{\frac{16a\pi f}{a^2+(2\pi f)^2}}_{A\frac{8\pi f}{T}} \underbrace{\frac{\sin(\pi g T)}{a^2+[2\pi(f-g)]^2}}_{E_{21}} \underbrace{-\frac{8a\pi}{a^2+(2\pi f)^2}}_{-A\frac{4\pi}{T}} \underbrace{\frac{g\sin(\pi g T)}{a^2+[2\pi(f-g)]^2}}_{E_{22}}$$

D.2.1 Intégrons E_{21}

$$I_{21} = \int_{-\infty}^{+\infty} \frac{\sin(\pi g T)}{a^2+[2\pi(f-g)]^2} dg$$

Changeons de variable :

$$\begin{cases} h = f-g & g = f-h \\ dh = -dg & dg = -dh \end{cases}$$

$$\begin{aligned} I_{21} &= -\int_{+\infty}^{-\infty} \frac{\sin(\pi f T - \pi h T)}{a^2+(2\pi h)^2} dh \\ &= \underbrace{-\int_{+\infty}^{-\infty} \frac{\sin(\pi f T)\cos(\pi h T)}{a^2+(2\pi h)^2} dh}_{I_{31}} + \underbrace{\int_{+\infty}^{-\infty} \frac{\cos(\pi f T)\sin(\pi h T)}{a^2+(2\pi h)^2} dh}_{I_{32}} \end{aligned}$$

La fonction à intégrer dans I_{32} étant impaire, nous avons $I_{32} = 0$.

D.2.2 Intégrons E_{22}

$$I_{22} = \int_{-\infty}^{+\infty} \frac{g\sin(\pi g T)}{a^2+[2\pi(f-g)]^2} dg$$

Changeons de variable :

$$\begin{cases} h = f - g & g = f - h \\ dh = -dg & dg = -dh \end{cases}$$

$$\begin{aligned} I_{22} &= -\int_{+\infty}^{-\infty} \frac{(f-h)\sin(\pi fT - \pi hT)}{a^2 + (2\pi h)^2} dh \\ &= \underbrace{-\int_{+\infty}^{-\infty} \frac{f\sin(\pi fT)\cos(\pi hT)}{a^2 + (2\pi h)^2} dh}_{I_{41}} \quad \underbrace{+\int_{+\infty}^{-\infty} \frac{f\cos(\pi fT)\sin(\pi hT)}{a^2 + (2\pi h)^2} dh}_{I_{42}} \\ &\quad \underbrace{+\int_{+\infty}^{-\infty} \frac{\sin(\pi fT)h\cos(\pi hT)}{a^2 + (2\pi h)^2} dh}_{I_{43}} \quad \underbrace{-\int_{+\infty}^{-\infty} \frac{\cos(\pi fT)h\sin(\pi hT)}{a^2 + (2\pi h)^2} dh}_{I_{44}} \end{aligned}$$

Nous avons $I_{42} = 0$ et $I_{43} = 0$, car les fonctions à intégrer sont impaires.

D.3 Récapitulatif

$$I = \frac{A}{T} + \frac{8\pi f}{T} A I_{31} - \frac{4\pi}{T} A I_{41} - \frac{4\pi}{T} A I_{44}$$

En remplaçant terme par terme et en remettant les bornes comme il faut, nous obtenons :

$$\begin{aligned} I &= \frac{A}{T} + \frac{8\pi f}{T} A\sin(\pi fT) \int_{-\infty}^{+\infty} \frac{\cos(\pi hT)}{a^2 + (2\pi h)^2} dh - \frac{4\pi f}{T} A\sin(\pi fT) \int_{-\infty}^{+\infty} \frac{\cos(\pi hT)}{a^2 + (2\pi h)^2} dh \\ &\quad -\frac{4\pi}{T} A\cos(\pi fT) \int_{-\infty}^{+\infty} \frac{h\sin(\pi hT)}{a^2 + (2\pi h)^2} dh \\ &= \frac{A}{T} + \frac{4\pi f}{T} A\sin(\pi fT) \underbrace{\int_{-\infty}^{+\infty} \frac{\cos(\pi hT)}{a^2 + (2\pi h)^2} dh}_{i_1} - \frac{4\pi}{T} A\cos(\pi fT) \underbrace{\int_{-\infty}^{+\infty} \frac{h\sin(\pi hT)}{a^2 + (2\pi h)^2} dh}_{i_2} \end{aligned}$$

D.4 Mise en forme de i_1

Changeons de variable :

$$\begin{cases} h' = \pi hT & h = \frac{h'}{\pi T} \\ dh' = \pi T dh & dh = \frac{dh'}{\pi T} \end{cases}$$

$$i_1 = \frac{1}{\pi T} \int_{-\infty}^{+\infty} \frac{\cos(h')}{a^2 + (2h'/T)^2} dh' = \frac{T}{4\pi} \int_{-\infty}^{+\infty} \frac{\cos(h')}{(Ta/2)^2 + (h')^2} dh'$$

Or, $a = \frac{2\alpha}{T}$, donc $\frac{Ta}{2} = \alpha$.

Finalement : $i_1 = \dfrac{T}{4\pi} \int\limits_{-\infty}^{+\infty} \dfrac{\cos(h')}{\alpha^2 + (h')^2} dh'$

D.5 Mise en forme de i_2

De la même façon :

$$i_2 = \frac{1}{(\pi T)^2} \int\limits_{-\infty}^{+\infty} \frac{h' \sin(h')}{a^2 + (2h'/T)^2} dh' = \frac{1}{4\pi^2} \int\limits_{-\infty}^{+\infty} \frac{h' \sin(h')}{\alpha^2 + (h')^2} dh'$$

D.6 Résolution numérique de i_1 et de i_2

Dans un premier temps, nous avons calculé i_1 et i_2 numériquement. Tout d'abord, MAPLE nous apprend que (nous avons testé pour un grand nombre de α) :

$$\int\limits_{-\infty}^{+\infty} \underbrace{\frac{\alpha \cos(h')}{\alpha^2 + (h')^2}}_{f(h')} dh' = \int\limits_{-\infty}^{+\infty} \frac{h' \sin(h')}{\alpha^2 + (h')^2} dh'$$

Ceci reste à prouver.

Donc il suffit de calculer l'une des deux intégrales pour avoir l'autre. Bien sûr, nous calculons i_1, qui converge plus vite que i_2 puisque la fonction à intégrer décroît en $\dfrac{1}{(h')^2}$ au lieu de décroître en $\dfrac{1}{h'}$. Nous nous limitons à utiliser la formule de SIMPSON :

$$\int\limits_{a}^{a+h} f(x)dx \simeq \frac{h}{6} \left[f(a) + 4f\left(a + \frac{h}{2}\right) + f(a+h) \right]$$

Nous constatons que la fonction passe par 0 tous les $k\pi + \dfrac{\pi}{2}$, k étant entier, et qu'à une portion large de π où la fonction est toujours positive succède une portion de la même largeur où elle est toujours négative, et inversement. De plus, l'intégrale sur une portion positive est supérieure à la valeur absolue de l'intégrale sur la partie négative qui la suit, à cause du dénominateur. Nous calculons donc i_1 sur un nombre N_c entier de fois 2π :

$$i_1 = 2 \sum_{i=1}^{N_c} \int\limits_{(i-1)2\pi}^{i2\pi} f(h')dh'$$

Le facteur 2 est dû au fait que la fonction est paire, c'est-à-dire au fait que :

$$\int\limits_{-\infty}^{+\infty} f(h')dh' = 2 \int\limits_{0}^{+\infty} f(h')dh'.$$

Nous obtenons une bonne estimation. Ainsi, avec $\alpha = 5$, $T = 0{,}3$ s, nous obtenons :

avec l'intégration numérique de MAPLE	$i_1 = 0{,}000101069204987$
avec notre intégration numérique	$i_1 = 0{,}000101069204939$

(avec, pour notre intégration, $N_c = 2000$ et $N_{pas} = 1000$, le nombre de pas par lequel nous divisons chaque portion large de 2π).

Nous avons déjà un résultat intéressant : les deux intégrales i_1 et i_2 ne dépendent pas de f, il suffit donc de les calculer une fois, où même d'avoir une table fixe contenant la solution de ces intégrales pour différents α.

D.7 Solution analytique

Dans [Bey84], page 288, nous trouvons que :

$$\int_0^{+\infty} \frac{\cos(mx)}{\alpha^2 + x^2} dx = \frac{\pi}{2\,|\alpha|} \exp\left(-\,|m\alpha|\right)$$

Soit, dans notre cas ($m = 1$, et m et α toujours positifs) :

$$\int_{-\infty}^{+\infty} \frac{\cos(h')}{\alpha^2 + (h')^2} dh' = \frac{\pi}{\alpha} \exp(-\alpha)$$

Nous avons donc :

$$I = \frac{A}{T} + \frac{4\pi f}{T} A \sin(\pi f T) \frac{T}{4\pi} \frac{\pi}{\alpha} \exp(-\alpha) - \frac{4\pi}{T} A \cos(\pi f T) \frac{1}{4\pi^2} \alpha \frac{\pi}{\alpha} \exp(-\alpha)$$

Finalement, nous obtenons :

$$\boxed{I = \frac{2a}{a^2 + (2\pi f)^2} \left\{ 1 + \left[\frac{\pi f T}{\alpha} \sin(\pi f T) - \cos(\pi f T) \right] \exp(-\alpha) \right\}}$$

qui est la transformée de Fourier de $\exp\left(-\alpha \dfrac{|n|}{T/2}\right) \mathcal{J}\left[-\dfrac{T}{2}, \dfrac{T}{2}\right]$, avec $a = \dfrac{2\alpha}{T}$.

D.8 Preuves finales

Les problèmes qu'il nous restait à résoudre étaient les suivants :

- Prouver que :

$$\int_{-\infty}^{+\infty} \frac{\alpha \cos(h)}{\alpha^2 + h^2} dh = \int_{-\infty}^{+\infty} \frac{h \sin(h)}{\alpha^2 + h^2} dh$$

- Prouver que :

$$\int_0^{+\infty} \frac{\cos(h)}{\alpha^2 + h^2} dh = \frac{\pi}{2\alpha} \exp(-\alpha) \text{ ou } \left(\int_0^{+\infty} \frac{\cos(rh)}{\alpha^2 + h^2} dh = \frac{\pi}{2\alpha} \exp(-r\alpha) \right)$$

La première équation nous a été confirmée dans [Dwi61].

Les preuves de ces deux équations nous ont été données dans [Edw22]. La seconde équation a été résolue par Laplace en 1811 (*Bulletin de la Société Philosophique*), qui utilisa une méthode légèrement compliquée et tordue, que nous pouvons trouver à la référence citée. Nous avons trouvé, pour $\alpha = 1$ et $r = 1$, une autre preuve, infiniment plus sympathique, dans [Dix84b]. Nous allons rapidement la présenter, pour α et r quelconques.

Mais, tout d'abord, nous allons donner la preuve, relativement simple, de la première équation.

D.9 Preuve de la première équation

Il suffit, pour qu'elle nous saute aux yeux, du moins plus facilement, de partir de (avec, pour simplifier, r positif et différent de 0) :

$$I = \int_0^{+\infty} \frac{\cos[rh]}{\alpha^2 + h^2} dh$$

Passons de r à $r + \delta r$. Nous avons alors :

$$I + \delta I = \int_0^{+\infty} \frac{\cos[(r + \delta r)h]}{\alpha^2 + h^2} dh$$

Puis :

$$\begin{aligned}
\frac{\delta I}{\delta r} &= \int_0^{+\infty} \frac{1}{\alpha^2 + h^2} \frac{\cos[(r + \delta r)h] - \cos[rh]}{\delta r} dh \\
&= \int_0^{+\infty} \frac{1}{\alpha^2 + h^2} \{-h \sin[rh] + \epsilon\} dh \\
&= \int_0^{+\infty} \frac{-h \sin[rh]}{\alpha^2 + h^2} dh + \int_0^{+\infty} \frac{\epsilon}{\alpha^2 + h^2} dh
\end{aligned}$$

Avec ϵ qui tend vers 0 quand δr tend vers 0. ϵ passe par son maximum ϵ_1 pour une valeur x_1 de x, compris dans l'intervalle $[0 \ldots + \infty[$. Alors le second terme est inférieur à :

$$\begin{aligned}
\epsilon_1 \int_0^{+\infty} \frac{1}{\alpha^2 + h^2} dh &= \epsilon_1 \frac{1}{\alpha^2} \int_0^{+\infty} \frac{1}{1 + \left(\frac{h}{\alpha}\right)^2} dh = \epsilon_1 \frac{1}{\alpha} \int_0^{+\infty} \frac{1}{1 + y^2} dy \\
&= \epsilon_1 \frac{1}{\alpha} [\arctan y]_0^{+\infty} = \epsilon_1 \frac{1}{\alpha} \frac{\pi}{2}
\end{aligned}$$

Et cette expression tend vers 0 avec δr.
Donc (en utilisant le résultat de la seconde équation à prouver) :

$$\frac{\delta I}{\delta r} = -\frac{\pi}{2} \exp(-r\alpha) = -I\alpha = \int_0^{+\infty} \frac{-h \sin[rh]}{\alpha^2 + h^2} \, dh$$

De plus $\cos(rh)$ et $h \sin(rh)$ sont paires et, dans notre cas, $r = 1$. Donc la démonstration est complète.

D.10 Preuve de la seconde équation

Nous utilisons dans cette section le théorème des résidus. La fonction considérée est (avec $\alpha > 0$ et $r > 0$) :

$$f(z) = \frac{\exp(irz)}{z^2 + \alpha^2} = \frac{\exp(irz)}{(z + i\alpha)(z - i\alpha)}$$

Nous définissons les sous-ensembles U, P_1 et P_2 de $\mathcal{C}$ ainsi :

$$\begin{cases} U & : \quad y > -\alpha \\ P_1 & : \quad y = 0, -R < x < R \\ P_2 & : \quad y > 0, x^2 + y^2 = R^2 \end{cases}$$

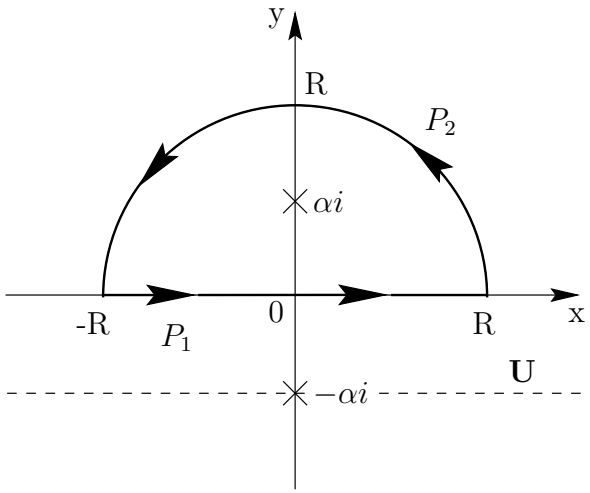


FIG. D.1 – *Espace complexe*

f est holomorphe dans $U - \{i\alpha\}$.
D'après le théorème des résidus, nous avons :

$$Res(f, i\alpha) = g(i\alpha)$$

Avec :

$$g(i\alpha) = f(i\alpha)(z - i\alpha) = \frac{\exp(iir\alpha)}{i\alpha + i\alpha} = \frac{\exp(-r\alpha)}{2i\alpha}$$

Et, $P_1 \cup P_2$ étant fermé et étant le bord orienté dans le sens direct d'un sous-ensemble de $\mathcal{C}$ contenant le point singulier de $f(z)$, nous avons :

$$\int_{P_1 \cup P_2} f(z)dz = 2i\pi Res(f, i\alpha)$$

Donc :

$$\int_{P_1} \frac{\exp(irz)}{z^2 + \alpha^2} dz + \int_{P_2} \frac{\exp(irz)}{z^2 + \alpha^2} dz = 2i\pi \frac{\exp(-r\alpha)}{2i\alpha} = \frac{\pi}{\alpha} \exp(-r\alpha)$$

Nous avons de plus :

$$\int_{P_1} \frac{\exp(irz)}{z^2 + \alpha^2} dz = \int_{-R}^{R} \frac{\exp(irx)}{x^2 + \alpha^2} dx$$

Et :

$$\int_{P_2} \frac{\exp(irz)}{z^2 + \alpha^2} dz = \int_{0}^{\pi} \frac{\exp\left[irR\exp(i\theta)\right] Ri\exp(i\theta)}{\left(R\exp(i\theta)\right)^2 + \alpha^2} d\theta$$

Or :

$$\left|\frac{\exp\left[irR\exp(i\theta)\right]Ri\exp(i\theta)}{\left(R\exp(i\theta)\right)^2+\alpha^2}\right| = \left|\frac{\exp\left[irR\cos(\theta)\right]\exp(i\theta)Ri\exp\left[iirR\sin(\theta)\right]}{\left(R\exp(i\theta)\right)^2+\alpha^2}\right|$$

$$= \frac{R\exp\left[-Rr\sin(\theta)\right]}{\left|R^2\exp(2i\theta)+\alpha^2\right|} \leq \frac{R}{R^2-\alpha^2}$$

Donc :

$$\left|\int_0^{\pi}\frac{\exp\left[irR\exp(i\theta)\right]Ri\exp(i\theta)}{\left(R\exp(i\theta)\right)^2+\alpha^2}d\theta\right| \leq \frac{\pi R}{R^2-\alpha^2}$$

De sorte que cette intégrale tend vers 0 quand R tend vers $+\infty$. Nous obtenons donc :

$$\int_{-\infty}^{+\infty}\frac{\exp(irx)}{x^2+\alpha^2}dx = \frac{\pi}{\alpha}\exp(-r\alpha)$$

De plus :

$$\int_{-\infty}^{+\infty}\frac{\exp(irx)}{x^2+\alpha^2}dx = \underbrace{\int_{-\infty}^{+\infty}\frac{\cos(rx)}{x^2+\alpha^2}dx}_{E_1} + i\underbrace{\int_{-\infty}^{+\infty}\frac{\sin(rx)}{x^2+\alpha^2}dx}_{E_2}$$

Comme la fonction à intégrer dans E_2 est impaire, l'intégrale est nulle. Et, comme la fonction à intégrer dans E_1 est paire, nous avons bien :

$$\boxed{\int_0^{+\infty}\frac{\cos(rx)}{x^2+\alpha^2}dx = \frac{\pi}{2\alpha}\exp(-r\alpha)}$$

D.11 Application à la fenêtre de pondération de Hanning-Poisson

Nous avons :

$$F(t) = \underbrace{\text{Poisson}}_{F_P(t)}\underbrace{\left[0{,}5+0{,}5\cos\left(\pi\frac{t}{T/2}\right)\right]}_{F_H(t)}$$

Et :

$$TF_F(f) = TF_P(f)\star TF_H(f)$$

Donc, $TF_P(f)$ étant I, nous avons :

$$TF_F = 0{,}5TF_P(f) + 0{,}25TF_P\left(f-\frac{1}{T}\right) + 0{,}25TF_P\left(f+\frac{1}{T}\right)$$

D.12 Les fenêtres de pondération de Hanning et de Hanning-Poisson dans le domaine fréquentiel

Sur les figures D.2 ($\alpha = 2$) et D.3 ($\alpha = 4$), sont représentées trois fenêtres de pondération dans le domaine fréquentiel. Ces trois fenêtres de pondération sont celles de Blackman (trait en tirets et points), Hanning (trait en tirets) et Hanning-Poisson (trait plein). Pour α suffisamment grand (la limite étant environ 1), la fenêtre de Hanning-Poisson est sans lobes secondaires.

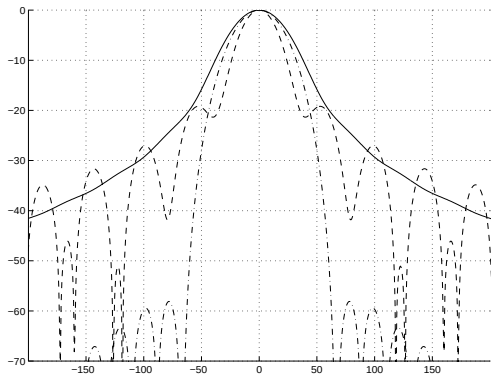


FIG. D.2 – *Fenêtres de pondération de* BLACKMAN *(trait en tirets et points),* HANNING *(trait en tirets) et* HANNING-POISSON *(trait plein) dans le domaine fréquentiel. Pour les deux dernières :* $\alpha = 2$. *La largeur des fenêtres de pondération est* 45,4 ms. *En abscisse : la fréquence en* Hz *; en ordonnée l'amplitude en* dB

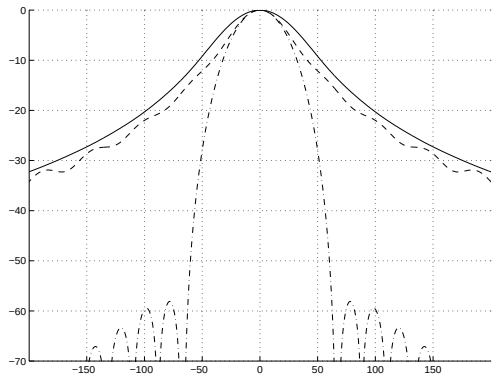


FIG. D.3 – *Fenêtres de pondération de* BLACKMAN *(trait en tirets et points),* HANNING *(trait en tirets) et* HANNING-POISSON *(trait plein) dans le domaine fréquentiel. Pour les deux dernières :* $\alpha = 4$. *La largeur des fenêtres de pondération est* 45,4 ms. *En abscisse : la fréquence en* Hz *; en ordonnée l'amplitude en* dB

Annexe E

Fonction d'atténuation du canal auditif et de l'oreille moyenne

Les fonctions d'observation telles qu'elles ont été définies dans l'exposé ne prennent pas en compte les caractéristiques de l'audition. Dans des cas extrêmes, nous pourrions imaginer que des transitions existant d'un point de vue traitement du signal ne fussent pas décelables à l'oreille. Ainsi, nous avons introduit la fonction d'atténuation du canal auditif et de l'oreille moyenne. Pour résumer rapidement et grossièrement, elle dit que nous n'entendons rien ni au-dessous de 20 Hz ni au-dessus de 20000 Hz.

La fonction d'atténuation que nous utilisons est la suivante :

$$A(f) = 3{,}64f^{-0{,}8} - 6{,}5\exp\left[-0{,}6(f-3{,}3)^2\right] + 0{,}001f^4$$

où $A(f)$ est en dB et f en kHz.

Elle a été suggérée par Terhardt et all. (voir [Pai92] et [Col94]).

Nous l'avons appliquée dans le cas du « flux spectral calculé avec les spectres d'amplitude », sur l'extrait de flûte **flute.sf**. En fait, nous multiplions les échantillons fréquentiels des spectres d'amplitude par :

- $B(f) = 10{,}0^{-A(f)/10{,}0}$ pour $f > 0$
- $B(f) = 0$ pour $f = 0$

Nous donnons $A(f)$ et $B(f)$ respectivement sur les figures E.1 et E.2.

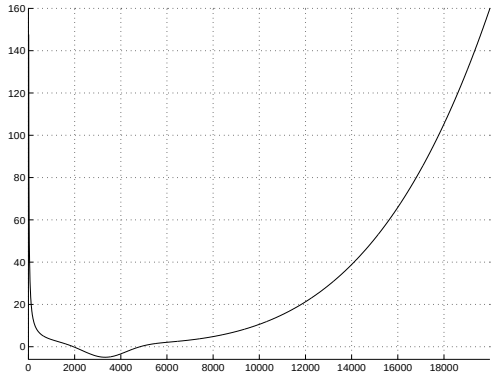


Fig. E.1 – *Atténuation $A(f)$ (dB). En abscisse : la fréquence en Hz*

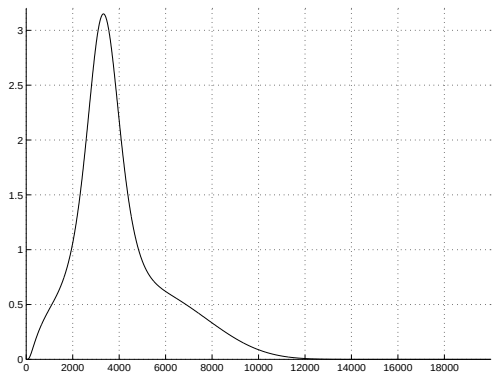


Fig. E.2 – *Atténuation $B(f)$ (linéaire). En abscisse : la fréquence en Hz*

Nous donnons sur les figures E.3 et E.4 respectivement le « flux spectral calculé en utilisant les spectres d'amplitude sur toutes les fréquences sans prendre en compte la fonction d'atténuation », et le « flux spectral calculé en prenant les spectres d'amplitude sur toutes les fréquences en prenant

en compte la fonction d'atténuation $B(f)$ ». Ces deux analyses ont été effectuées entre 0,31 et 13,3 secondes, c'est-à-dire que nous avons coupé le début et la fin du signal.

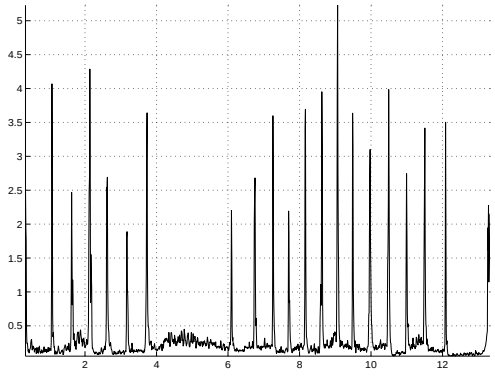


FIG. E.3 – « *Flux spectral calculé avec les spectres d'amplitude* » *sans l'atténuation pour l'extrait de flûte. En abscisse : le temps en seconde*

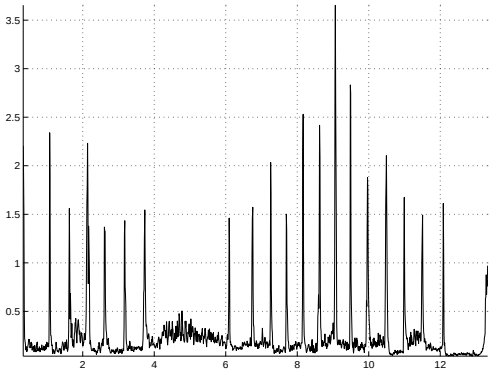


FIG. E.4 – « *Flux spectral calculé avec les spectres d'amplitude* » *avec l'atténuation pour l'extrait de flûte. En abscisse : le temps en seconde*

L'objectif n'est pas forcément d'améliorer les performances de la *segmentation en zones stables* : il est de constater les différences qui peuvent exister entre une fonction d'observation dans sa version « purement traitement du signal » et une fonction d'observation dans sa version « psychoacoustique ».

Annexe F

Moyenne et variance des estimateurs de σ^2 pour une variable aléatoire gaussienne

F.1 Moments M_k d'une variable aléatoire gaussienne

Soit un processus aléatoire $\mathcal{N}\left(m,\sigma^2\right)$. Nous nous plaçons dans le cas général où m n'est pas nulle.

Le moment d'ordre k s'écrit :

$$M_k = \int_{-\infty}^{+\infty} x^k p_x(x) dx = E\left[x^k\right]$$

où $p_x(x)$ est la densité de probabilité de la variable aléatoire x et E l'opérateur espérance. Ainsi :

$$M_k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^k \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) dx$$

Il faut faire le changement de variable $y = x - m$. Alors, nous obtenons :

$$M_k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (y+m)^k \exp\left(-\frac{y^2}{2\sigma^2}\right) dy$$

Pour chaque k, nous développons $(y+m)^k$, de telle façon que M_k soit la somme d'intégrales du type :

$$M_k = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^\alpha m^\beta \exp\left(-\frac{y^2}{2\sigma^2}\right) dy$$

avec $\alpha \in [0 \ \ldots \ k]$ et $\beta \in [0 \ \ldots \ k]$. Nous profitons alors de ce que :

$$\begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{+\infty} y^{2\alpha+1} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy &= 0 \\ \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{+\infty} y^{2\alpha} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy &= \dfrac{(2\alpha)!}{2^\alpha \alpha!}\sigma^{2\alpha} \end{cases}$$

La solution de la première intégrale est évidente puisque la fonction à intégrer est impaire ; et celle de la seconde vient de ce que :

$$\int_0^{+\infty} x^{2n} \exp\left(-ax^2\right) dx = \frac{1.3.5\ldots(2n-1)}{2^{n+1}a^n}\sqrt{\frac{\pi}{a}} \text{ (voir [Bey84] page 290)}$$

Finalement, après quelques calculs, nous obtenons :

$$\begin{array}{lcll} M_1 & = & E[x] & = & m \\ M_2 & = & E\left[x^2\right] & = & \sigma^2 + m^2 \\ M_3 & = & E\left[x^3\right] & = & 3\sigma^2 m + m^3 \\ M_4 & = & E\left[x^4\right] & = & 3\sigma^4 + 6m^2\sigma^2 + m^4 \end{array}$$

Nous n'avons pas ici besoin d'aller au-delà de M_4 [1].

F.2 Moyenne des estimateurs de la variance

Deux estimateurs de la variance sont couramment utilisés :

$$\hat{\sigma}^2_{(1)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{m})^2$$

ou :

$$\hat{\sigma}^2_{(2)} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{m})^2$$

avec :

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Le premier est biaisé, alors que le second ne l'est pas. Nous prouvons que :

$$E\left[\hat{\sigma}^2_{(1)}\right] = \sigma^2 - \frac{\sigma^2}{N}$$

et que :

$$E\left[\hat{\sigma}^2_{(2)}\right] = \sigma^2$$

Il suffit de développer $\hat{\sigma}^2_{(1)}$ et $\hat{\sigma}^2_{(2)}$ et de les exprimer en fonction des moments $M_1 = E[x]$, $M_2 = E\left[x^2\right]$... Ceci nous conduit, si nous le faisons à la main, à des calculs un peu tordus (surtout pour les variances des estimateurs de la variance : voir la section F.3), peu réjouissants, mais pas très compliqués. Nous donnons ci-dessous les détails pour $E\left[\hat{\sigma}^2_{(1)}\right]$:

$$\hat{\sigma}^2_{(1)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{m})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 + \frac{1}{N}\sum_{i=1}^{N}\hat{m}^2 - \frac{2}{N}\sum_{i=1}^{N} x_i\hat{m}$$

Donc :

$$E\left[\hat{\sigma}^2_{(1)}\right] = \underbrace{\frac{1}{N}\sum_{i=1}^{N} E\left[x_i^2\right]}_{S_1} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} E\left[\hat{m}^2\right]}_{S_2} \underbrace{- \frac{2}{N}\sum_{i=1}^{N} E\left[x_i\hat{m}\right]}_{S_3}$$

1. M_3 et M_4 sont utilisés lors du calcul de Var $\left[\hat{\sigma}^2_{(1)}\right]$ et de Var $\left[\hat{\sigma}^2_{(2)}\right]$: voir la section F.3.

Résolvons les sommes S_1, S_2 et S_3 l'une après l'autre. Nous utilisons pour cela les résultats de la section F.1.

$$S_1 = \frac{1}{N} N M_2 = \frac{1}{N} N\sigma^2 + \frac{1}{N} N m^2$$

$$S_2 = \frac{1}{N} \sum_{i=1}^{N} E\left[\frac{1}{N} \sum_{j=1}^{N} x_j \frac{1}{N} \sum_{k=1}^{N} x_k \right] = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{N^2} E\left[\sum_{j=1}^{N} \sum_{k=1}^{N} x_j x_k \right] = \frac{1}{N} \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} E\left[x_j x_k\right]$$

$$S_3 = -\frac{2}{N} \sum_{i=1}^{N} E\left[x_i \frac{1}{N} \sum_{j=1}^{N} x_j \right] = -\frac{2}{N} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} E\left[x_i x_j\right]$$

Ensuite :

$$S_2 + S_3 = -\frac{1}{N} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} E\left[x_i x_j\right] = -\frac{1}{N} \frac{1}{N} N\left(\sigma^2 + m^2\right) - \frac{1}{N} \frac{1}{N} N(N-1) m^2$$

le premier terme venant des N cas où $i = j$ (alors : $E\left[x_i x_j\right] = E\left[x_i^2\right] = M_2$) et le deuxième terme des $N(N-1)$ cas où $i \neq j$ (alors : $E\left[x_i x_j\right] = E\left[x_i\right] E\left[x_j\right] = M_1 M_1$).

Finalement :

$$S_1 + S_2 + S_3 = \frac{1}{N} N\sigma^2 + \frac{1}{N} N m^2 - \frac{1}{N} \frac{1}{N} N\left(\sigma^2 + m^2\right) - \frac{1}{N} \frac{1}{N} m^2 N(N-1) = \sigma^2 - \frac{1}{N}\sigma^2$$

F.3 Variance des estimateurs de la variance

Nous prouvons de la même façon que :

$$\mathrm{Var}\left[\hat{\sigma}^2_{(1)}\right] = 2\sigma^4 \left(\frac{N-1}{N^2} \right)$$

et que :

$$\mathrm{Var}\left[\hat{\sigma}^2_{(2)}\right] = 2\sigma^4 \left(\frac{1}{N-1} \right)$$

Ainsi, la variance de l'estimée est moins grande dans le cas biaisé que dans le cas non biaisé.

Annexe G

Densité de probabilité de $u = \sqrt{\sum_i^N x_i^2}$

G.1 Introduction

Les N variables aléatoires x_i sont indépendantes. Nous nous plaçons dans le cas où les x_i obéissent à un processus de Rayleigh. Si les x_i obéissaient à une loi normale, $z = \sum_{i=1}^{N} x_i^2$ (voir la section G.2.2) suivrait une loi du χ^2 (voir [Pap65] page 250, ou [AS70] page 940) et $u = \sqrt{\sum_{i=1}^{N} x_i^2}$ (voir la section G.2.3) une loi du χ.

G.2 Densités de probabilité

G.2.1 De $y = x^2$

Soit x_i des variables aléatoires qui obéissent au même processus de Rayleigh. La densité de probabilité de chaque x_i est :

$$p_x(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Calculons la densité de probabilité de $y = x^2$. Nous avons $x = \sqrt{u}$ et $ydy = 2xdx$. Ainsi (voir [Cha96] page 28) :

$$p_y(y) = \frac{1}{2\sqrt{y}} \frac{\sqrt{y}}{\sigma^2} \exp\left(-\frac{y}{2\sigma^2}\right) = \frac{1}{2\sigma^2} \exp\left(-\frac{y}{2\sigma^2}\right)$$

G.2.2 De $z = \sum_{i=1}^{N} x_i^2$

La densité de probabilité de $w = y_1 + y_2$ est :

$$p_w(w) = p_y(y) * p_y(y) = \frac{1}{4\sigma^4} \int_0^w \exp\left(-\frac{r}{2\sigma^2}\right) \exp\left(-\frac{w-r}{2\sigma^2}\right) dr = \frac{1}{4\sigma^4} \exp\left(-\frac{w}{2\sigma^2}\right) w$$

où $*$ représente la convolution.

La densité de probabilité de $z = w + y$ est :

$$\begin{aligned} p_z(z) &= \frac{1}{4\sigma^4}\frac{1}{2\sigma^2}\int_0^z \exp\left(-\frac{w}{2\sigma^2}\right) w \exp\left(-\frac{z-w}{2\sigma^2}\right) dw \\ &= \frac{1}{8\sigma^6}\exp\left(-\frac{z}{2\sigma^2}\right)\int_0^w w dw \\ &= \frac{1}{16\sigma^6}\exp\left(-\frac{z}{2\sigma^2}\right) z^2 \end{aligned}$$

Finalement, nous avons, pour $z = \sum_{i=1}^{N} x_i^2$:

$$p_z(z) = \alpha \frac{1}{\sigma^{2N}} \exp\left(-\frac{z}{2\sigma^2}\right) z^{N-1}$$

L'intégrale de toute densité de probabilité étant égale à 1, et le domaine de définition de z étant $[0 + \infty[$, nous avons :

$$\alpha = \frac{1}{\frac{1}{\sigma^{2N}}\int_0^{+\infty} \exp\left(-\frac{z}{2\sigma^2}\right) z^{N-1} dz} = \frac{1}{\beta}$$

Si nous faisons le changement de variable $v = \frac{z}{2\sigma^2}$, nous obtenons, en sachant que $\int_0^{+\infty} x^n \exp(-ax)\, dx = \frac{n!}{a^{n+1}}$ pour $a > 0$ et n entier positif (voir [Bey84] page 290), après quelques calculs :

$$\beta = 2^N (N-1)!$$

où ! représente la factorielle. Et donc :

$$p_z(z) = \frac{1}{2^N (N-1)!}\frac{1}{\sigma^{2N}} \exp\left(-\frac{z}{2\sigma^2}\right) z^{N-1}$$

G.2.3 De $u = \sqrt{\sum_{i=1}^{N} x_i^2}$

Finalement : $u = \sqrt{z}$, donc $z = u^2$, $zdz = 2udu$, et :

$$p_u(u) = \frac{1}{2^N (N-1)!}\frac{1}{\sigma^{2N}} \exp\left(-\frac{u^2}{2\sigma^2}\right) u^{2N-2} 2u = \frac{1}{2^{N-1}(N-1)!}\frac{1}{\sigma^{2N}} \exp\left(-\frac{u^2}{2\sigma^2}\right) u^{2N-1}$$

N	2	4	6	10	100	1000
γ	1,88	2,742	3,393	4,417	14,124	44,716
$\sqrt{2}\sqrt{N}$	2,000	2,828	3,464	4,472	14,142	44,721
$\frac{\sqrt{2}\sqrt{N}}{\gamma}$	1,064	1,031	1,021	1,013	1,0013	1,0001

TAB. G.1 – γ *en fonction de* N

G.3 Moyenne de u

Le domaine de définition de u est $[0,\ +\infty[$. La moyenne de u est alors :

$$\begin{aligned} E[u] &= \frac{1}{2^{N-1}(N-1)!}\frac{1}{\sigma^{2N}}\int_0^{+\infty}\exp\left(-\frac{u^2}{2\sigma^2}\right)u^{2N}du \\ &= \frac{1}{2^{N-1}(N-1)!}\frac{1}{\sigma^{2N}}\int_0^{+\infty}\exp(-v)\left(2\sigma^2\right)^N v^N\frac{\sigma^2}{\sqrt{2}\sqrt{\sigma^2}\sqrt{v}}dv \\ &= \underbrace{\sqrt{2}\frac{1}{(N-1)!}\int_0^{+\infty}\exp(-v)\,v^{N-\frac{1}{2}}dv}_{\gamma}\,\sigma \end{aligned}$$

Que vaut γ ? En fait, γ tend vers $\sqrt{2}\sqrt{N}$ quand N tend vers l'infini, comme le montre le tableau G.1. Nous ne donnons pas de preuve analytique.

G.4 Test avec des signaux simulés

Soit un signal de bruit gaussien $\mathcal{N}(0,\sigma^2)$ de taille $M = 32$. Nous savons que les échantillons fréquentiels du spectre réel et ceux du spectre imaginaire sont des variables aléatoires qui suivent des lois $\mathcal{N}\left(0,\frac{\sigma^2}{2}\right)$. De plus, nous savons (voir [Bri81]) qu'elles sont indépendantes. Ainsi, les échantillons fréquentiels du spectre d'amplitude (module du spectre complexe) suivent une loi de RAYLEIGH :

$$p_b(b) = \frac{b}{\sigma^2/2}\exp\left(-\frac{b^2}{2\sigma^2/2}\right)$$

dont la moyenne est : $E[b] = \sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{2}} = \frac{\sigma\sqrt{\pi}}{2}$. Pour $\sigma = 3$, nous obtenons $E[b] = 2{,}6587$: il s'agit de la droite en pointillés de la figure G.1. En trait plein est représentée l'évolution de la moyenne en fonction du nombre de points utilisés pour la calculer en pratique. Le signal de bruit blanc simulé est de longueur M. Le premier spectre, obtenu avec la première réalisation du signal nous donne $\frac{M}{2}$ points : la moyenne est calculée sur $\frac{M}{2}$ points ; le deuxième nous donne $\frac{M}{2}$ autres points : la moyenne est calculée sur M points ; etc. Nous voyons que la moyenne pratique rejoint la moyenne théorique.

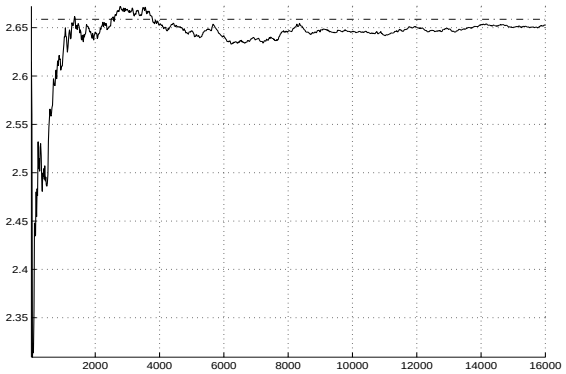


FIG. G.1 – *Évolution de la moyenne en fonction du nombre de points utilisés pour la calculer. En abscisse : le nombre de points ; en ordonnée : la moyenne. Trait interrompu : moyenne théorique ; trait plein : moyenne estimée*

Annexe H

Corrélations : quelques tests (signaux simulés)

Soient x_1 et x_2 les deux fonctions d'observation (ou les deux variables aléatoires). Il s'agit de vecteurs de taille M.

Si x_1 et x_2 obéissent à des lois normales de variances respectives 1 et σ^2 et de moyennes nulles. Et si x_1 et x_2 sont indépendantes, le coefficient de corrélation, l'information mutuelle et le test du χ^2 sont tous les trois très petits (très proches de 0).

Pour les tests, nous prenons des observations de $M = 20000$ points. Et pour N, la taille de l'histogramme, nous prenons $N = 15$. Une méthode pour déterminer N le nombre de cases de l'histogramme à prendre en fonction du nombre de points est la règle de Sturge, qui nous donne ici : $N = \log_2(M) + 1 \simeq 15$.

H.1 Dépendance linéaire : $x_2 = x_1 + y$

H.1.1 Premier cas

x_1 est uniformément répartie entre 0 et 1 et y est normale, de variance σ^2 et de moyenne nulle. Voir les figures H.1 et H.2.

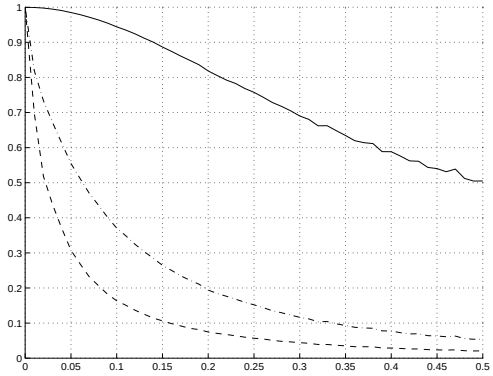


Fig. H.1 – *Variation des mesures de la corrélation en fonction de la variance du bruit. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

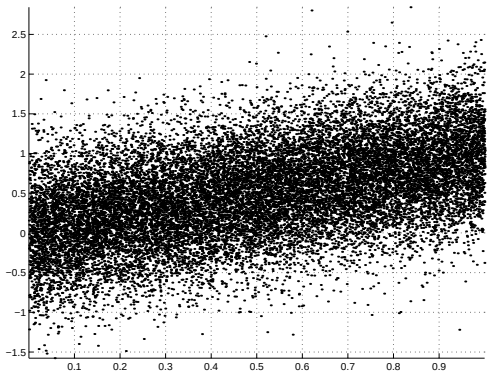


Fig. H.2 – $x_2 = x_1 + y$. *x_1 est uniformément répartie. y est normale, de variance σ^2. Une observation, avec $\sigma = 0{,}5$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Le coefficient de corrélation, l'information mutuelle normalisée $im(x_1,x_2)$ et le test du χ^2 sont égaux à 1 quand $x_2 = x_1$. $im(x_1,x_2)$ diminue plus vite que le coefficient de corrélation, et le test du χ^2 plus vite encore, au fur et à mesure que les deux variables aléatoires se décorrèlent.

H.1.2 Second cas

x_1 et y sont normales, de variances 1 et σ^2 et de moyennes nulles. Voir les figures H.3 et H.4.

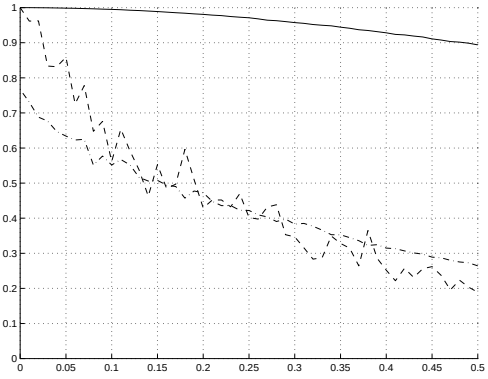


FIG. H.3 – *Variation des mesures de la corrélation en fonction de la variance du bruit. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

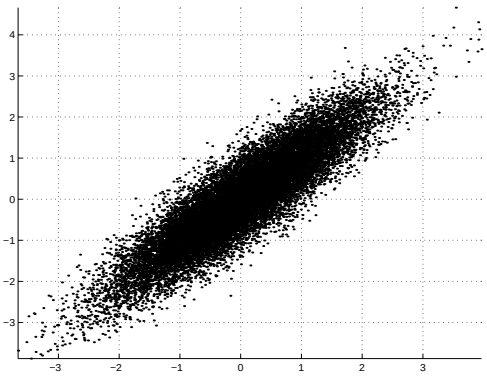


FIG. H.4 – *$x_2 = x_1 + y$. x_1 est normale, de variance 1. y est normale, de variance σ^2. Une observation, avec $\sigma = 0{,}5$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Puisque x_1 n'est pas uniformément répartie, $im(x_1,x_1)$ n'est pas égale à 1. Du fait de la normalisation, elle est toujours inférieure à 1. Dans le cas normal, et si nous n'avons à notre disposition qu'un histogramme au lieu de la vraie densité de probabilité, nous avons, avec Δx la taille d'une case de l'histogramme :

$$p_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) \Delta x$$

Dans ce cas, nous obtenons (la définition de $H(X)$ est donnée page 75) :

$$H(X) \simeq \hat{H}(X) = \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) \Delta x \left(\log_2\left(\frac{\Delta x}{\sqrt{2\pi}\sigma}\right) - \frac{x_i^2}{2\sigma^2}\frac{1}{\log_e(2)}\right)$$

Ainsi, avec $\sigma = 1$, $N = 15$ et $M = 20000$, nous obtenons $\hat{H}(X) \simeq 2{,}96$. Cette estimation de $H(X)$ a été obtenue en calculant la moyenne des $H(X)$ obtenus à partir de 1000 observations de X. En fait, Δx dépend de l'observation, donc $H(X)$ aussi : le problème est que, contrairement au cas du bruit uniforme, ici le domaine de définition de X est infini. Dans le cas du bruit uniforme, nous avons $H(X) = 3{,}9069$: alors, $im(x_1,x_1)$, dans le cas du bruit normal, est égale à $\frac{2{,}96}{3{,}9069} = 0{,}75$. Nous le constatons sur la courbe H.3, quand la variance du bruit additif est nulle, c'est-à-dire quand il n'y pas de bruit additif, c'est-à-dire encore quand $x_2 = x_1$.

Pour $N = 5$, $im(x_1,x_1) \simeq 0{,}58$; $N = 10$, $im(x_1,x_1) \simeq 0{,}71$; $N = 15$, $im(x_1,x_1) \simeq 0{,}75$; $N = 20$, $im(x_1,x_1) \simeq 0{,}78$; $N = 50$, $im(x_1,x_1) \simeq 0{,}83$; $N = 200$, $im(x_1,x_1) \simeq 0{,}87$.

Remarquons que la valeur de $im(x_1,x_1)$ est relativement peu sensible à la variance de la fonction d'observation.

Le test du χ^2 part de 1. Ensuite, son comportement ressemble à celui de l'information mutuelle.

H.2 Parabole : $x_2 = x_1^2 + y$

x_1 et y sont normales, de variances 1 et σ^2 et de moyennes nulles. Voir les figures H.5 et H.6.

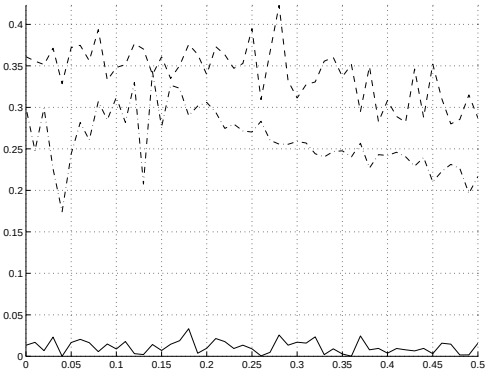


FIG. H.5 – *Variation des mesures de la corrélation en fonction de la variance du bruit. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

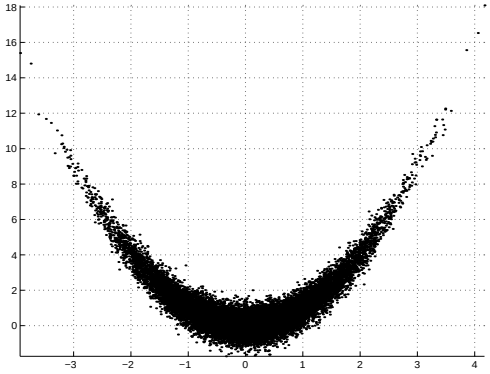


FIG. H.6 – *$x_2 = x_1^2 + y$. x_1 est normale, de variance 1. y est normale, de variance σ^2. Une observation, avec $\sigma = 0{,}5$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Nous voyons que le coefficient de corrélation n'est pas du tout efficace dans ce cas, contrairement à l'information mutuelle et au test du χ^2.

H.3 Cercle

y est uniformément répartie entre 0 et 2π. x_1 est égale à $2\cos(y)+z_1$ et x_2 est égale à $2\sin(y)+z_2$. z_1 et z_2 sont gaussiennes de variances σ^2 et de moyennes nulles. Voir les figures H.7 et H.8.

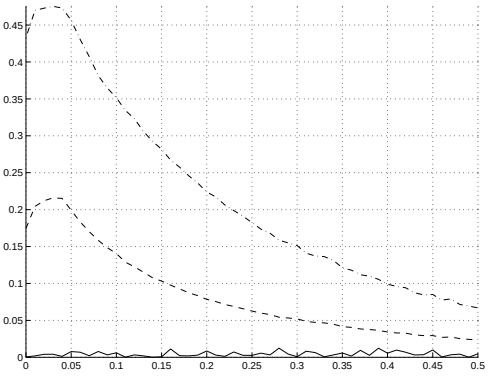


FIG. H.7 – *Variation des mesures de la corrélation en fonction de la variance du bruit. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

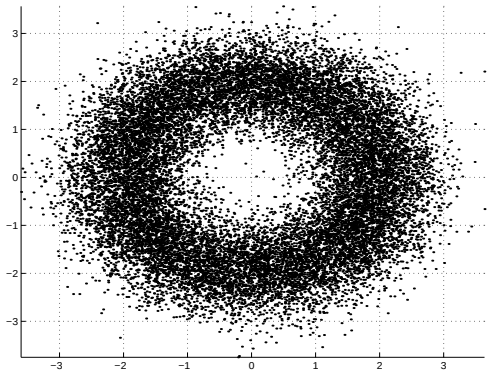


FIG. H.8 – *Voir la section H.3 pour la définition de x_1 et x_2. Une observation, avec $\sigma = 0{,}5$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Nous voyons que le coefficient de corrélation n'est pas du tout efficace dans ce cas, contrairement au test du χ^2 mais surtout à l'information mutuelle.

H.4 Sinus

x_1 est uniformément répartie entre 0 et 4π. x_2 est égale à $\cos(x_1) + z_1$. z_1 est gaussienne de variance σ^2 et de moyenne nulle. Voir les figures H.9 et H.10.

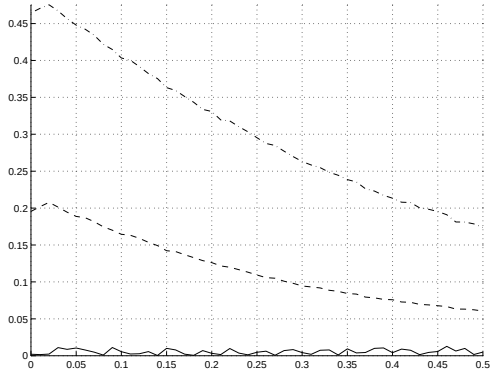


FIG. H.9 – *Variation des mesures de la corrélation en fonction de la variance du bruit. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

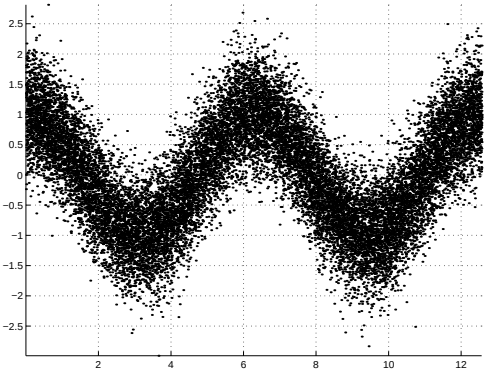


FIG. H.10 – *Voir la section H.4 pour la définition de x_1 et x_2. Une observation, avec $\sigma = 0{,}5$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Nous voyons que le coefficient de corrélation n'est pas du tout efficace dans ce cas, contrairement au test du χ^2 mais surtout à l'information mutuelle.

H.5 Arc de cercle

y est uniformément répartie entre 0 et $\alpha\pi$, α variant entre 0,01 et 2π. x_1 est égale à $2\cos(y)$ et x_2 est égale à $2\sin(y)$. Voir les figures H.11 et H.12.

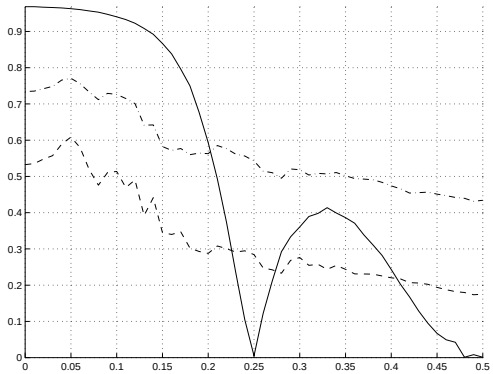


FIG. H.11 – *Variation des mesures de la corrélation en fonction de la longueur de l'arc de cercle. En abscisse : l'écart-type de y ; en ordonnée : les mesures de la corrélation. Trait plein : coefficient de corrélation ; trait en tirets et points : information mutuelle ; trait interrompu : test du χ^2*

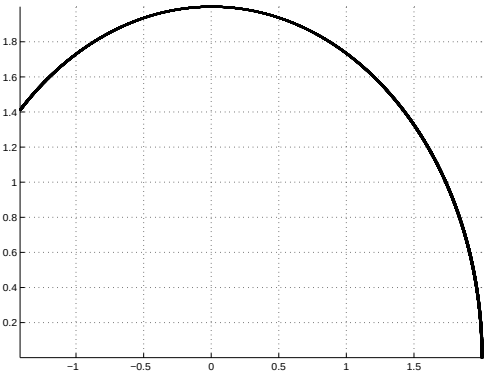


FIG. H.12 – *Voir la section H.5 pour la définition de x_1 et x_2. Une observation, avec $\alpha = \pi$, est représentée. x_1 en abscisse et x_2 en ordonnée*

Nous voyons que les performances du coefficient de corrélation se détériorent (nous passons par

un 0 de corrélation pour α=0,5), contrairement à celles de l'information mutuelle et du test du χ^2, qui restent à peu près stables.

H.6 Influence du nombre N sur l'information mutuelle

Nous avons pris $N = 5$, $N = 10$, $N = 15$, $N = 20$ et $N = 50$, et nous avons refait le premier test avec ces cinq valeurs. Nous obtenons les courbes de la figure H.13. Remarquons que dans tous les cas, avec deux variables aléatoires indépendantes, l'information mutuelle obtenue est petite ($\simeq$ 0,002).

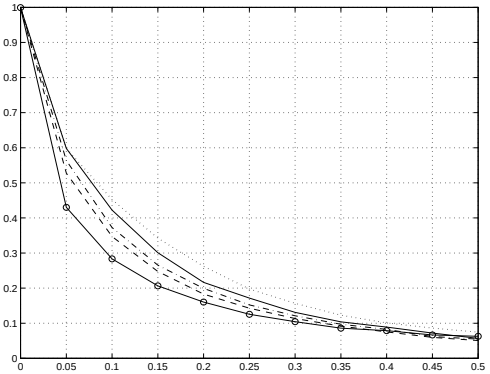


FIG. H.13 – *Résultats du premier test avec divers N. Trait en pointillés : $N = 5$; trait plein : $N = 10$; trait en tirets et points : $N = 15$; trait interrompu : $N = 20$; trait plein avec des ronds : $N = 50$. En abscisse : l'écart-type de y ; en ordonnée : l'information mutuelle*

Nous voyons que l'influence de N est très petite.

H.7 Conclusion

En fait, il faudrait utiliser une mesure MC de la corrélation entre deux variables aléatoires X et Y telle que si $Y = f(X)$ nous ayons $MC(X,Y) = 1$, avec f une « fonction » (possibilité de contraintes sur f, par exemple en ce qui concerne sa continuité, ou, à la rigueur, sa monotonie ?). Le problème est que nous pouvons toujours faire passer un polynôme par N points, du moment que l'ordre de celui-ci soit égal à $N-1$: donc, nous pouvons toujours déterminer une fonction MC telle que $MC(X,Y) = 1$, même si X et Y ne sont pas corrélées.

Faisons la remarque que dans le cas de la dépendance linéaire, sans bruit, x_2 est entièrement déterminée par x_1, et inversement. Nous voudrions donc avoir : $MC(x_1,x_2) = MC(x_2,x_1) = 1$. Ce n'est plus le cas dans le cas de la parabole (sans bruit) : x_2 est entièrement déterminée par x_1, mais il y a ambiguïté en ce qui concerne x_1 par rapport à x_2. Nous avons la même chose dans le cas du sinus. Ainsi, il faudrait que la fonction MC ne fût pas symétrique dans le cas général : $MC(x_1,x_2) \neq MC(x_2,x_1)$. Les cas du cercle et de l'arc de cercle sont encore particuliers.

Voir l'article de BASSEVILLE [Bas89]. Il existe un grand nombre de moyens pour calculer la corrélation, la distance entre deux variables aléatoires !

Bibliographie

[ABF+94] B. Angelini, D. Brugnara, Daniele Falavigna, D. Giuliani, R. Gretter, and M. Omologo, *Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus*, International Conference on Speech and Language Processing (ICSLP'94), 1994.

[AD99] D. Arfib and N. Delprat, *Alteration of the vibrato of a recorded voice*, International Computer Music Conference (ICMC'99), 1999, pp. 186 – 189.

[AS70] Milton Abramowitz and Irene A. Stegun, *Handbook of mathematical functions*, Dover publications, 1970.

[ASD97] Ali Alkulaibi, J. J. Soraghan, and T. S. Durrani, *Fast 3-level binary higher order statistics for simultaneous voiced/unvoiced and pitch detection of a speech signal*, Signal Processing, 1997, pp. 133 – 140.

[Bas89] Michèle Basseville, *Distance measures for signal processing and pattern recognition*, Signal Processing, 1989, pp. 349 – 369.

[Bat94] Roberto Battiti, *Using mutual information for selecting features in supervised neural net learning*, IEEE Transactions on Neural Networks, vol. 5, juillet 1994, pp. 537 – 550.

[BB82] Michèle Basseville and Albert Benveniste, *Détection séquentielle de changements brusques des caractéristiques spectrales d'un signal numérique*, Tech. report, INRIA, avril 1982.

[BD79] M. Baudry and B. Dupeyrat, *Speech segmentation and recognition using syntactic methods on the direct signal*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'79), 1979.

[Bey84] William H. Beyer, *CRC Standard Mathematical Tables*, 27th ed., CRC Press, 1984.

[BFO93] F. Brugnara, D. Falavigna, and M. Omologo, *Automatic segmentation and labeling of speech based on Hidden Markov Models*, Speech Communication, vol. 12, 1993, pp. 357 – 370.

[BK79] Marcia A. Bush and Gary E. Kopec, *Segmentation in Isolated Word Recognition Using Vector Quantization*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'79), 1979, pp. 17.11.1 – 17.11.4.

[BN93] Michèle Basseville and Igor V. Nikiforov, *Detection of abrupt changes: Theory and application*, Prentice-Hall Inc., 1993, voir : http://www.irisa.fr/sigma2/kniga/.

[Boa92] Boualem Boashash, *Estimating and interpreting the instantaneous frequency of a signal – part 2: Algorithms and applications*, Proceedings of the IEEE, vol. 80, avril 1992, pp. 539 – 568.

[Bre90] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, 1990.

[Bri81] D. R. Brillinger, *Time series data analysis and theory*, Holden Day, 1981.

[Bro95] Christian Brousseau, *Radar M.F./V.H.F. multiporteuse (MOSAR)*, Thèse de doctorat, 1995, Université de Rennes 1.

[BS94] K. Brandenburg and G. Stoll, *ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio*, Journal of the Audio Engineering System (AES'94), vol. 42, numéro 10, octobre 1994, pp. 780 – 791.

[BS97] Daniel S. Benincasa and Michael I. Savic, *Co-channel speaker separation using non-linear optimization*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 1195 – 1198.

[Buh95] Joachim M. Buhmann, *Learning and data clustering*, Handbook of Brain Theory and Neural Networks, 1995.

[Car95] Jean-François Cardoso, *Séparation adaptative de sources dans l'espace signal*, 15ème Colloque sur le Traitement du Signal et des Images (GRETSI'95), 1995.

[CBCG94] Malcolm Crawford, Guy J. Brown, Martin Cooke, and Phil Green, *Design, collection and analysis of a multi-simultaneous-speaker corpus*, Proceedings of The Institute of Acoustics, vol. 16, 1994, pp. 183 – 190.

[Cer94] Laurent Cerveau, *Segmentation de phrases musicales à partir de la fréquence fondamentale*, Stage de DEA, IRCAM, juillet 1994.

[CGG+92] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt, *First Leaves: a tutorial introduction to Maple V*, Springer-Verlag, 1992.

[Cha96] Maurice Charbit, *éléments de théorie du signal : les signaux aléatoires*, Ellipses, 1996.

[CJ86] C. Chafe and D. Jaffe, *Source separation and note identification in polyphonic music*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'86), 1986, pp. 1289 – 1292.

[Col94] Catherine Colomes, *Étude d'un modèle d'audition et d'une mesure objective de la qualité sonore dans le contexte du codage à réduction de débit*, Thèse de doctorat, septembre 1994, Université de Rennes 1.

[Cor99] Francis Corson, *étude de la décomposition des sons en sinusoïdes et bruit*, Stage de DEA, IRCAM, juillet 1999.

[Dan95] Jérôme Daniel, *Structuration de signaux de contrôle*, Stage de DEA, IRCAM, septembre 1995.

[DBN92] Abdulkadir Dinc and Yeheskel Bar-Ness, *Bootstrap: a fast blind adaptative signal separator*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'92), 1992, pp. 325 – 328.

[dC94] Alain de Cheveigné, *Strategies for voice separation based on harmonicity*, International Conference on Speech and Language Processing (ICSLP'94), 1994.

[DGR93] Philippe Depalle, Guillermo Garcìa, and Xavier Rodet, *Tracking of partials for additive sound synthesis using hidden markov models*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'93), vol. I, avril 1993, pp. 225 – 228.

[DGR94] ———, *A virtual castrato (!?)*, International Computer Music Conference (ICMC'94), 1994.

[Dix84a] Jacques Dixmier, *Cours de mathématiques du premier cycle – première année*, 2ème ed., Gauthier-Villars, 1984.

[Dix84b] ———, *Cours de mathématiques du premier cycle – seconde année*, 2ème ed., Gauthier-Villars, 1984.

[DJ94] D. L. Donoho and I. M. Johnstone, *Adaptating to unknown smoothness via wavelet shrinkage*, Tech. Report 426, Stanford University, 1994.

[Don94] D. L. Donoho, *On minimum entropy segmentation*, Tech. Report 450, Stanford University, avril 1994.

[Dov94] Boris Doval, *Estimation de la fréquence fondamentale des signaux sonores*, Thèse de doctorat, mars 1994, Université de Paris VI.

[Dwi61] Herbert Bristol Dwight, *Tables of integrals and other mathematical data*, 4th ed., pp. 224 – 225, New York – The Macmillan Company, 1961.

[Edw22] J. M. A. Edwards, *A Treatise on the Integral Calculus*, 1th ed., vol. 2, pp. 214 – 227, New York – The Macmillan Company, 1922.

[Ell96] David P. W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Massachusetts Institute of Technology, juin 1996.

[Fal95] Daniele Falavigna, *Comparison of different HMM based methods for speaker verification*, Eurospeech, vol. 1, 1995, pp. 371 – 374.

[FC84] Takeshi Fukabayashi and Chiu-Kuang Chuang, *Speech Segmentation and Recognition Using Adaptative Linear Prediction Algorithm*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'84), 1984, pp. 17.12.1 – 17.12.4.

[FCM+96] Pascal Faudemay, Liming Chen, Claude Montacié, Marie-José Caraty, Christine Maloigne, Xiao Wei Tu, Mohsen Ardebilian, and Jean-Luc Le Floch, *Multi-channel video segmentation*, Conference on Multimedia Storage and Archiving Systems, SPIE Symposium, novembre 1996.

[Fri90] Benjamin Friedlander, *A Sensitivity Analysis of the MUSIC Algorithm*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'90), vol. 38, octobre 1990, pp. 1740 – 1751.

[FW88] Benjamin Friedlander and A. J. Weiss, *Eigenstructure methods for direction finding with sensor gain and phase uncertainties*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'88), 1988.

[GM96] Masataka Goto and Yoichi Muraoka, *Beat tracking based on multiple-agent architecture - a real-time beat tracking system for audio signals*, International Conference on Multi-Agent Systems, 1996, pp. 103 – 110.

[HÝ7] Thomas Hélie, *Amélioration de l'extraction de partiels dans les signaux sonores*, Stage ingénieur de troisième année, IRCAM/ENST, janvier 1997.

[Haj96] John Hajda, *A New Model for Segmenting the Envelope of Musical Signals: The relative Salience of Steady State versus Attack, Revisited*, Journal of the Audio Engineering System (AES'96), novembre 1996.

[Hal85] Pierre Hallé, *Segmentation syllabique et reconnaissance des tons du chinois en parole continue*, Thèse de doctorat, mai 1985, Université de Paris-Sud.

[Har78] Fredric J. Harris, *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*, Proceedings of the IEEE, vol. 66, Janvier 1978, pp. 51 – 83.

[HB98] Perfecto Herrera and Jordi Bonada, *Vibrato Extraction and Parameterization in the Spectral Modelling Synthesis Framework*, Workshop on Digital Audio Effects (DAFx'98), 1998, pp. 107 – 110.

[Hes83] Wolfgang Hess, *Pitch determination of speech signals*, pp. 154 – 181, Springer-Verlag, 1983.

[HJBK78] Seppo Haltsonen, Matti Jalanko, Kalle-J. Bry, and Teuvo Kohonen, *Application of novelty filter to segmentation of speech*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'78), 1978, pp. 565 – 568.

[HL98] Jürgen Herre and Adam Lindsay, *ISO/IEC JTC1/SC29/WG11 Coding of moving pictures and audio*, International organisation for standardisation - Organisation internationale de normalisation, 1998.

[Hon94] Henkjan Honing, *The Vibrato Problem: Comparing Two Solutions*, Tech. report, Institute for Logic, Language and Computation, 1994.

[JEA99] Stephen Jacobs, Alexandros Eleftheriadis, and Dimitris Anastassiou, *Silence detection for multimedia communication systems*, ACM/Springer Verlag Multimedia Systems Journal, vol. 7, numéro 2, mars 1999.

[Jeh97] Tristan Jehan, *Musical signal parameter estimation*, Tech. report, CNMAT/IFSIC, 1997.

[JNN95] Huang Jie, Ohnishi Noboru, and Sugie Noboru, *Sound Separation Based on Perceptual Grouping of Sound Segments*, International Symposium on Nonlinear Theory and its Applications, décembre 1995.

[JW88] Jean-Claude Junqua and Hisashi Wakita, *SAIPH: a segmentation system for automatic labeling of a large speech database – Application to speech recognition*, Journal of the Acoustical Society of America (JASA'88), 1988, pp. 1 – 34.

[Kay88] S. Kay, *Statistically/computationally efficient frequency estimation*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'88), 1988, pp. 2292 – 2295.

[KHM96] Reinier W. L. Kortekaas, Dik J. Hermes, and Georg F. Meyer, *Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multilayer perceptrons*,

Journal of the Acoustical Society of America (JASA'96), vol. 99, numéro 2, février 1996, pp. 1185 – 1199.

[Kla97] Anssi Klapuri, *Automatic transcription of music*, Ph.D. thesis, Tampere university of technology, 1997.

[KM81] S. M. Kay and S. L. Marple, *Spectrum Analysis – A Modern Perspective*, Proceedings of the IEEE, novembre 1981, p. 1380.

[KR93] R. Kumaresan and C. S. Ramalingam, *On separating voiced-speech into its components*, Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers, 1993, pp. 1041 – 1046.

[Lar89] Jean Laroche, *Étude d'un système d'analyse et de synthèse utilisant la méthode de Prony – Application aux instruments de musique du type percussif*, Thèse de doctorat, octobre 1989, École Nationale Supérieure des Télécommunications.

[LO95] Philippe Lepain and Régine André Obrecht, *Micro-segmentation d'enregistrements musicaux*, Deuxièmes Journées d'Informatique Musicale, LAFORIA, 1995, pp. 81 – 90.

[LR91] A. Ljolje and M. D. Riley, *Automatic segmentation and labeling of speech*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'91), 1991, pp. 473 – 476.

[Mah90] Robert C. Maher, *Evaluation of a Method for Separating Digitized Duet Signals*, Journal of the Audio Engineering System (AES'90), vol. 38(12), décembre 1990, pp. 956 – 979.

[Mar80] S. L. Marple, *A new Autoregressive Spectrum Analysis Algorithm*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'80), août 1980, p. 441.

[Mas96] Paul Masri, *Computer modelling of sound for transformation and synthesis of musical signals*, Ph.D. thesis, University of Bristol, décembre 1996.

[MB96] Paul Masri and Andrew Bateman, *Improved modelling of attack transients in music analysis/resynthesis*, International Computer Music Conference (ICMC'96), 1996.

[MC98] Claude Montacié and Marie-José Caraty, *A silence/noise/music/speech splitting algorithm*, International Conference on Speech and Language Processing (ICSLP'98), 1998.

[Mee94] N. Meeus, *De la forme musicale et de la segmentation*, Musurgia, 1994, pp. 7 – 23.

[Mel91] David K. Mellinger, *Event formation and separation in musical sound*, Ph.D. thesis, Stanford University, décembre 1991.

[MH80] D. Marr and E. Hildreth, *Theory of edge detection*, Proceedings of the Royal Society of London, vol. 207, 1980, pp. 187 – 217.

[MHJ95] Carl D. Mitchell, Mary P. Harper, and Leah H. Jamieson, *Using explicit segmentation to improve HMM phone recognition*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'95), 1995, pp. 229 – 232.

[MJM95] T. Moudenc, D. Jouvet, and J. Monné, *On using a priori segmentation of the speech signal in an n-best solutions post-processing*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'95), 1995, pp. 580 – 583.

[MKQ93] Petros Maragos, James F. Kaiser, and Thomas F. Quatieri, *Energy separation in signal modulations with application to speech analysis*, IEEE Transactions on Signal Processing, vol. 41, octobre 1993, pp. 3024 – 3051.

[MMR91] D. K. Mellinger and B. M. Mont-Reynaud, *Soundexplorer: A workbench for investigating source separation*, International Computer Music Conference (ICMC'91), octobre 1991, pp. 90 – 93.

[Moo75] James A. Moorer, *On the segmentation and analysis of continuous musical sound*, Ph.D. thesis, Stanford University, 1975.

[MPB97] G. F. Meyer, F. Plante, and F. Berthommier, *Segregation of concurrent speech with the reassigned spectrum*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 1203 – 1206.

[MSE93] Andrew Morris, Jean-Luc Schwartz, and Pierre Escudier, *An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram*, Computer Speech and Language, vol. 2, 1993, pp. 121 – 136.

[Nor] Michael Norris, *Design decisions in an oscillatory model of primitive auditory segregation*, (ébauche).

[NPM97] Douglas Nunn, Alan Purvis, and Peter Manning, *Source separation and transcription of polyphonic music*, Tech. report, Durham Music Technology, 1997.

[ONK96] Hiroshi G. Okuno, Tomohiro Nakatani, and Takeshi Kawabata, *A new speech enhancement: speech stream segregation*, International Conference on Speech and Language Processing (ICSLP'96), 1996.

[OS75] Alan V. Oppenheim and Ronald W. Schafer, *Digital signal processing*, pp. 239 – 250, Prentice Hall, 1975.

[Pai92] Bruno Paillard, *Codage perceptuel des signaux audio de haute qualité*, Thèse de doctorat, février 1992, Université de Sherbrooke.

[Pap65] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes*, International student edition – McGraw-Hill, 1965.

[PD94] B.I.(Raj) Pawate and Eric Dowling, *A new method for segmenting continuous speech*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'94), vol. 1, 1994, pp. 53 – 56.

[PGV97] Paolo Prandoni, Michael Goodwin, and Martin Vetterli, *Optimal time segmentation for signal modelling and compression*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 2029 – 2032.

[Phi95] Pierrick Philippe, *Codage audionumérique et transformation en ondelettes*, Thèse de doctorat, novembre 1995, Université de Paris-Sud.

[PMMS92] B. Paillard, P. Mabilleau, S. Morissette, and Joël Soumagne, *PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals*, Journal of the Audio Engineering System (AES'92), vol. 40, numéro 1/2, janvier/février 1992, pp. 21 – 31.

[Rap95] *SNNS - Stuttgart Neural Network Simulator – User Manual*, Tech. report, University of Stuttgart, 1995, voir : http//www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html.

[Red93] Darren Redfern, *The Maple Handbook*, Springer-Verlag, 1993.

[RK94] C. S. Ramalingam and R. Kumaresan, *Voiced-speech analysis based on the residual interfering signal canceler (RISC) algorithm*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'94), vol. 1, 1994, pp. 473 – 476.

[Ros94] Stéphane Rossignol, *Calibrage et autocalibrage de réseaux d'antennes – Application au radar MOSAR*, Stage de DEA, Université de Rennes 1, juillet 1994.

[Ros97a] _____, *Segmentation – Extraction du vibrato*, rapport technique, janvier 1997, IRCAM.

[Ros97b] _____, *Supplément au premier rapport d'activité*, rapport technique, avril 1997, IRCAM.

[Ros97c] _____, *Supplément au premier rapport d'activité*, rapport technique, août 1997, IRCAM.

[Ros98] _____, *Segmentation en notes ou en phones – Prises de décisions – Les deux autres niveaux de segmentation*, rapport technique, septembre 1998, Supélec – Campus de Metz.

[Ros99a] _____, *Segmentation suite*, rapport technique, février 1999, Supélec – Campus de Metz.

[Ros99b] _____, *Segmentations, séparation : suite*, rapport technique, juin 1999, IRCAM.

[Roz96] Jean-Philippe Rozé, *Document de synthèse*, rapport technique, mai 1996, IRCAM.

[RRD+99] Stéphane Rossignol, Xavier Rodet, Philippe Depalle, Joël Soumagne, and Jean-Luc Collette, *Vibrato: detection, estimation, extraction, modification*, Workshop on Digital Audio Effects (DAFx'99), 1999.

[RRS+] Stéphane Rossignol, Xavier Rodet, Joël Soumagne, Jean-Luc Collette, and Philippe Depalle, *Automatic characterisation of musical signals: feature extraction and temporal segmentation*, Journal of New Music Research, december.

[RRS+98] _____, *Feature extraction and temporal segmentation of acoustic signals*, International Computer Music Conference (ICMC'98), 1998.

[RRS+99] _____, *Segmentation, indexation et manipulation des signaux sonores*, Réunion des Théoriciens des Circuits de Langue Française (RTCLF'99), 1999.

[Sau96] John Saunders, *Real-time discrimination of broadcast speech/music*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'96), 1996, pp. 993 – 996.

[SBZD99] Mouhamadou Seck, Frédéric Bimbot, Didier Zugaj, and Bernard Delyon, *Two-class signal segmentation for speech/music detection in audio tracks*, Eurospeech'99, septembre 1999.

[Sch96] Eric D. Scheirer, *Bregman's chimerae: Music perception as auditory scene analysis*, 4th International Conference on Music Perception and Cognition, août 1996.

[Sch98] _____, *Tempo and beat analysis of acoustic musical signals*, Journal of the Acoustical Society of America (JASA'98), 1998, pp. 588 – 601.

[SCL96] Malcolm Slaney, Michele Covell, and Bud Lassiter, *Automatic audio morphing*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'96), mai 1996.

[Ser89] Xavier Serra, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*, Ph.D. thesis, Stanford University, octobre 1989.

[Smi94] L. S. Smith, *Sound segmentation using onsets and offsets*, Journal of New Musical Research, vol. 23, numéro 1, mars 1994, pp. 11 – 23.

[SNL94] Malcolm Slaney, Daniel Naar, and Richard F. Lyon, *Auditory model inversion for sound separation*, vol. 2, 1994, pp. 77 – 80.

[Spa98] Dragos Spataru, *Classification voix parlée/musique*, rapport technique, septembre 1998, Supélec – Campus de Metz.

[SS97] Eric D. Scheirer and Malcolm Slaney, *Construction and evaluation of a robust multifeature speech/music discriminator*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), 1997, pp. 1331 – 1334.

[SSW88] P. K. Sahoo, S. Soltani, and K. C. Wong, *A survey of thresholding techniques*, Computer Vision, Graphics, and Image Processing, vol. 41, 1988, pp. 233 – 260.

[SW96] Ludger Solbach and Rolf Wöhrmann, *Sound onset localization and partial tracking in gaussian white noise*, International Computer Music Conference (ICMC'96), 1996.

[TC97] Jean-Yves Tourneret and Marie Chabert, *Off-line detection and estimation of abrupt changes corrupted by multiplicative colored gaussian noise*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'97), avril 1997, pp. 3693 – 3696.

[Tem96] Stan Tempelaars, *Signal Processing, Speech and Music*, Swets & Zeitlinger, 1996.

[Vak96] David Vakman, *On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency*, IEEE Transactions on Signal Processing, vol. 44, avril 1996, pp. 791 – 797.

[Var96] Pramod K. Varshney, *Distributed detection and data fusion*, Springer, 1996.

[vB83] A. von Brandt, *Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'83), 1983, pp. 1017 – 1020.

[VM99] Jérôme Vannier-Moreau, *Speech/music/noise detection*, Tech. report, IRISA/Supélec – Campus de Metz, juin 1999.

[Wan94] Avery Li-Chun Wang, *Instantaneous and frequency-warped signal processing techniques for auditory source separation*, Ph.D. thesis, Stanford University, août 1994.

[Wan96] DeLiang Wang, *Primitive auditory segregation based on oscillatory correlation*, Cognitive Science, vol. 20, 1996, pp. 409 – 456.

[WE99] Gethin Williams and Daniel P. W. Ellis, *Speech/music discrimination based on posterior probability features*, EUROSPEECH'99, 1999.

[Wei84] Mitchel Weintraub, *The GRASP sound separation system*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'84), 1984, pp. 18.A.6.1 – 18.A.6.4.

[WRD92] Peter Wyngaard, Chris Rogers, and Philippe Depalle, *UDI 2.0 – A Unified DSP Interface*, Tech. report, juin 1992.

[YvVH99] Howard Yang, Sarel van Vuuren, and Hynek Hermansky, *Relevancy of time-frequency features for phonetic classification measured by mutual information*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'99), mars 1999.

[ZC81] Rainer Zelinski and Fritz Claus, *A segmentation procedure for connected word recognition based on estimation principles*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP'81), 1981, pp. 960 – 963.

[ZF81] E. Zwicker and R. Feldtkeller, *Psychoacoustique – l'oreille, récepteur d'information*, Masson – CNET/ENST, 1981.

[ZO96] T. C. Zhao and Mark Overmars, *Forms library – A graphical user interface toolkit for X*, Tech. report, 1996.

[ZWG99] Puming Zhan, Steven Wegmann, and Larry Gillick, *Dragon systems' 1998 broadcast news transcription system for mandarin*, DARPA Broadcast News Workshop, 1999.