Estimation de fréquences fondamentales multiples en vue de la séparation de signaux de parole mélangés dans un même canal

François Signol

N° D'ORDRE

THÈSE DE DOCTORAT

SPECIALITE : PHYSIQUE

École Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

Présentée par : **François SIGNOL**

Sujet :

Estimation de fréquences fondamentales multiples en vue de la séparation de signaux de parole mélangés dans un même canal

Soutenue le 14 décembre 2009 devant le jury composé de :

M. Alain de Cheveigné	Rapporteur
M. Philippe Martin	Rapporteur
M. Gaël Richard	Président de jury
M. Claude Barras	Examinateur
M. Jean-Sylvain Liénard	Directeur de thèse

Table des matières

Table des figures

Liste des tableaux

Remerciements

Mes premiers remerciements vont à mon directeur de thèse Jean-Sylvain Liénard. Sans lui, ce travail n'aurait tout simplement pas été ce qu'il est aujourd'hui. Il a su me conseiller, m'encadrer et m'aider. Ce travail est le fruit de nos innombrables discussions et de notre excellente entente pendant ces années. De la même manière, je remercie Claude Barras pour son aide plus que précieuse, ses idées et sa disponibilité.

Ensuite, je remercie grandement les deux rapporteurs de ce travail Alain de Cheveigné et Philippe Martin pour l'intérêt et l'attention qu'ils ont porté à mon travail. Leurs critiques et leurs remarques justifiées m'ont énormément motivé pour les suites envisageables de ce travail. Je remercie le président du Jury Gaël Richard pour ses questions piquantes et forts justes ainsi que pour ses conseils lors de la soutenance.

Je remercie à présent le Dr. Sylvain Le Beux, ami et collègue dans l'adversité de cette fin de thèse. Je remercie également tous les collègues du LIMSI et de l'IUT d'ORSAY pour les discussions caféinés et les mots-croisés qui ont sans nuls doutes participés à cette réussite.

Un merci spécial à Sibel Ruata Torres qui a su me supporter et m'encourager pendant ces années très complexes et stressantes. ¡Muchísimas gracias ! Sin ti, no hubiera podido terminar esta tesis.

Un clin d'oeil tout particulier au Dr. Olivier Joly : deux origines similaires, deux parcours bien différents, pourtant des points communs : l'humain, la guitare, ...

Je remercie également mes amis, véritables psychologues sans le savoir. Un merci tout particulier à Jean-roch Constans, Tanguy Mercier, Anthony Debray d'être venus assiter à ma soutenance. Pareillement, un grand merci à Monique et Louis sans qui je n'aurai sans doute pas connu Paris.

Je remercie évidemment mon frère et ma sœur d'avoir été là et de m'avoir soutenu le jour qui fut important pour moi.

Enfin, ce travail revient à mes parents sans qui, évidemment, tout ceci n'existerait pas.

Avant-propos

Ce travail de thèse est le fruit d'un travail d'équipe entre mon directeur de thèse Jean-Sylvain Liénard et moi-même.

Le sujet de thèse original fut la séparation automatique de parole superposée, sujet qui s'est avéré trop vaste pour une seule thèse. Il se retrouve néanmoins dans le positionnement scientifique adopté (chapitre 2) et dans la manière dont sont implémentés les algorithmes proposés. Ils sont strictement trame-à-trame et n'utilisent aucune information a priori. Ceci dans l'optique d'être utilisés comme traitement élémentaire d'un système de séparation de parole qui fera usage des propriétés de continuité du signal et de ses structures de haut niveau.

Nous avons travaillé en concertation étroite tout au long de ces quatre ans. Le principe général d'estimation de F_0 multiples que nous proposons (chapitre 4) nommé PSH est le produit de cette collaboration fructueuse. En revanche, les deux implémentations décrites dans ce manuscrit ont été écrites de manière indépendante et dans des optiques différentes (chapitre 5). La version nommée PSH_{mul} a été développée par Jean-Sylvain Liénard dans l'optique d'être rapide et robuste. La version PSH_{min} développée par mes soins a été conçue pour étudier les mécanismes internes du principe général de PSH, pour confirmer l'ensemble des mes développements théoriques (annexes A.1, A.2, A.3, A.4, A.5 et A.6) et pour produire les graphiques du manuscrit.

Mes contributions à ce travail d'ensemble se situent à tous les niveaux, théoriques comme pratiques. Elles portent plus particulièrement sur la formalisation du comportement des différents peignes et de leurs combinaisons, ainsi que sur l'élaboration et la mise en œuvre de méthodes d'évaluation prenant en compte l'influence primordiale de la décision de voisement sur les performances monopitch et multipitch.

Je vous souhaite par avance une bonne lecture.

Résumé

Mots-Clés : Estimation de F_0 multiples, parole superposée, décision voisé/non-voisé, évaluation monopitch et multipitch, suivi de F_0, situation de Cocktail Party, CASA.

Résumé : La problématique de cette thèse est inscrite dans le cadre très général de la séparation de parole. L'objectif est la détermination des fréquences fondamentales F_0 de chacun des locuteurs de mélanges de parole monauraux. Il s'agit d'estimer des hypothèses F_0 par un algorithme trame-à-trame purement fréquentiel reposant sur l'utilisation de plusieurs peignes spectraux. Nous proposons un algorithme d'estimation de F_0 multiples conçu pour traiter la parole superposée et utilisable sur des signaux mono-locuteur. Une étape d'évaluation permet ensuite de valider le bon fonctionnement de l'algorithme. Le manuscrit de thèse est organisé de la façon suivante :

Le chapitre 1 pose le problème général de séparation de parole et son lien avec l'estimation de F_0.

Le chapitre 2 est un état de l'art concernant le rôle de la F_0 dans la séparation de mélanges de parole. Il conclut sur la nécessité d'élaborer un algorithme d'estimation de F_0 multiples spécialement conçu pour traiter des mélanges de parole.

Le chapitre 3 commence par préciser les caractéristiques intrinsèques de la parole qui affectent l'estimation de F_0. Ces caractéristiques sont distinctes de celles d'un signal d'un instrument de musique si bien que l'estimation de F_0 est un problème différent dans les deux cas. Ensuite, un état de l'art des méthodes d'estimation de F_0 en situation monopitch et multipitch est détaillé. Il se conclut en proposant un algorithme d'estimation de F_0 purement fréquentiel, piste qui n'a été que peu étudiée sur de la parole superposée.

Le chapitre 4 analyse l'estimation de F_0 par peigne fréquentiel. Il décrit quatre peignes élémentaires et développe une analyse de leurs comportements, avantages et défauts. Cette analyse révèle que les pics parasites des fonctions de pitch sont indexables par un couple d'entiers non nuls (p, q) premiers entre eux. Par conséquent, ils peuvent être traités de manière sélective. Nous proposons trois nouveaux peignes : le PAL, le PDN et le PDM et analysons leurs comportements sur des signaux synthétiques monopitch et multipitch. Le chapitre se conclut en proposant d'utiliser les caractéristiques des PDN et PDM pour construire un algorithme qui élimine d'emblée les pics parasites des fonctions de pitch en conservant les pics des F_0 à estimer.

Le chapitre 5 explique dans le détail le principe du Peigne à Suppression Harmonique ou PSH, principe sur lequel repose notre algorithme d'estimation de F_0 multiples. Le principe de PSH consiste à fabriquer une fonction de pitch à partir de la combinaison des fonctions de pitch obtenues par les PDN et PDM. En effet, la nature de ces peignes permet d'atténuer de manière sélective certaines familles

de pics parasites. La fonction de pitch résultante dont les pics parasites sont atténués est utilisée pour estimer les hypothèses F_0. Le chapitre détaille deux implémentations particulières du principe de PSH nommées PSH_{min} et PSH_{mul}. Elles permettent de traiter les cas bipitch avec des F_0 à l'octave et fonctionnent bien sur des trames de parole réelle. Le chapitre se conclut par la nécessité d'évaluation des implémentations faites en traitant des signaux de parole réelle.

Le chapitre 6 commence par traiter le problème de la quantification du voisement d'un segment de parole et explique l'impact de la décision voisé/non-voisé (VnV) sur les résultats de l'évaluation. Il propose plusieurs méthodologies d'évaluation dont une nouvelle impliquant la décision VnV. Les méthodologies décrites sont utilisées en situation monopitch sur plusieurs algorithmes emblématiques de la littérature et sur l'algorithme PSH_{mul}. Il fait de même en situation bipitch sur deux algorithmes de la littérature et notre PSH_{mul}. PSH_{mul} se révèle être un bon algorithme d'estimation de F_0 multiples comparé à ceux évalués.

Le chapitre 7 conclut ce manuscrit de thèse. Il récapitule ainsi les apports de ce travail de thèse. Il esquisse aussi différentes perspectives de ce travail à court, moyen et long-terme.

Publications :

- Chapitre 3 :

François Signol. Suivi d'un flux de F_0 par programmation dynamique dans un mélange monaural de signaux de parole. In *Rencontres des Jeunes Chercheurs en Parole. RJCP07*, Paris, France, 2007.

- Chapitre 4 :

Jean-Sylvain Liénard, François Signol and Claude Barras. Speech fundamental frequency estimation using the Alternate Comb. In *INTERSPEECH 2007*, Antwerpen, Belgium, 2007.

- Chapitre 5 :

Jean-Sylvain Liénard, Claude Barras and François Signol. Using sets of combs to control pitch estimation errors. *Proceedings of Meetings on Acoustics*, 4(1), 2008.

- Chapitre 6 :

François Signol, Claude Barras, and Jean-Sylvain Liénard. Evaluation of the Pitch Estimation Algorithms in the monopitch and multipitch cases (**Abstract**). *The Journal of the Acoustical Society of America*, 123(5) : 3077, Acoustics 2008, Paris, 2008.

Chapitre 1

Introduction générale

L'environnement sonore qui nous entoure est complexe et le plus souvent le produit de plusieurs sources sonores qui se combinent pour former une scène auditive. Il est alors remarquable de voir avec quelle facilité apparente notre appareil auditif est capable d'isoler les sources sonores les unes des autres. L'un des défis actuels du CASA ou *Computational Auditory Scene Analysis* est de mettre en place des traitements automatiques qui reproduisent ces mécanismes naturels de ségrégation de sources sonores. Dans ce travail nous nous intéressons à une situation particulière de mélange sonore : la parole superposée ou problème de *Cocktail Party*. Il s'agit d'une scène auditive dans laquelle plusieurs locuteurs parlent simultanément. Cette scène auditive particulière est constituée uniquement de parole.

L'un des indices acoustiques les plus importants que notre système auditif utilise pour parvenir à séparer les sources sonores est la fréquence fondamentale F_0. La F_0 correspond à la fréquence de vibration d'une source sonore. Dans le cas de la parole, la F_0 représente la vibration des cordes vocales d'un locuteur. Dans le cas d'un instrument de musique tel que la trompette, il s'agit de la fréquence de vibration des lèvres du trompettiste.

Parmi l'ensemble des scènes auditives qui comportent une ou plusieurs F_0, les plus emblématiques sont les scènes auditives de parole superposée ou les scènes auditives musicales. De nombreuses méthodes existent actuellement pour l'estimation de F_0 de la parole et de la musique. Même si l'application souhaitée est la même pour les deux types de scènes auditives, celles-ci diffèrent suffisamment par la nature de leurs signaux pour que la transposition des méthodes d'estimation de F_0 de l'une à l'autre soit difficile. L'estimation de F_0 pour des signaux musicaux est rendue difficile par la polyphonie et le recouvrement temporel des notes. L'estimation de F_0 pour la parole superposée est rendue difficile par la non catégorisation des F_0 en notes discrètes et stables, ainsi que par ses variations spectrales. Dans cette thèse, le matériel est la parole et notre méthode d'estimation répond aux problématiques posées par la nature et les spécificités de ce signal.

Il convient de distinguer deux situations d'estimation de F_0. La situation d'estimation « monopitch » pour laquelle le signal analysé ne comporte qu'un seul locuteur et qui donc ne comporte au plus qu'une F_0 à un instant donné. Cette situation typique est la plus étudiée. Dans le cadre de la thèse nous nous intéressons à la situation de parole superposée dans laquelle les signaux émis par plusieurs locuteurs sont mélangés dans un seul canal. Nous parlons alors de situation d'estimation « multipitch » dans laquelle le signal analysé peut comporter plusieurs F_0 à un instant donné. Cette situation, par-

ticulièrement présente pour des signaux musicaux (polyphonie, orchestre), se retrouve également dans la parole superposée. Les F_0 respectives de chacun des locuteurs se retrouvent ainsi mélangées. Elles peuvent s'entrecroiser ou se confondre ce qui complexifie considérablement le traitement automatique. De plus, le mélange de signaux a des répercussions sur les amplitudes et les phases dans le domaine fréquentiel (ie. harmoniques communes) qui posent problème.

Dans cette thèse, nous distinguons clairement l'estimation de F_0 du suivi de F_0. Cette distinction n'est pas toujours très claire dans la littérature. Un signal de parole est généralement analysé par trames successives avec recouvrement. Chacune de ces trames est un court extrait du signal dans lequel le signal de parole est considéré comme stationnaire (de l'ordre de 40ms). L'estimation de F_0 a pour rôle d'estimer la ou les F_0 présentes dans chacune des trames (respectivement estimation monopitch ou multipitch). Le suivi de F_0 a pour rôle de construire les trajectoires de chacune des F_0. Ce sont deux tâches différentes. L'estimation de F_0 fait toujours des erreurs et l'état de l'art en la matière révèle que ces erreurs sont souvent corrigées par des post-traitements divers qui utilisent des contraintes implicites pour construire les trajectoires de chacune des F_0. Cette reconstitution de trajectoire relève du suivi de F_0 et non de l'estimation trame-à-trame mise en œuvre dans la thèse. Nos algorithmes d'estimation de F_0 n'utilisent aucun post-traitement ni aucune connaissance préalable sur les signaux du mélange. Les valeurs fournies à chaque trame sont destinées à être utilisées par une chaîne de traitement de plus haut niveau pouvant rassembler les F_0 d'un même locuteur en une même trajectoire. En d'autres termes, l'algorithme d'estimation multipitch (et monopitch) que nous proposons se veut comme un traitement élémentaire qui peut servir de « brique de base » dans la construction de traitements de plus haut niveau.

L'estimation de F_0 pour des signaux de parole en situation monopitch a déjà été largement étudiée. Les méthodes sont classables en trois familles d'approches : temporelles, fréquentielles, spectro-temporelles. Les approches temporelles reposent sur le principe de l'autocorrélation qui consiste à comparer la forme d'onde d'une trame avec une version décalée d'elle même. Si la forme d'onde comporte une F_0, elle est quasi-périodique et un motif temporel se répète de manière quasi-identique dans le temps. Comparer une trame avec une version décalée d'elle-même permet de mettre en évidence l'écart temporel pour lequel la ressemblance entre les formes d'onde est maximale. Les approches fréquentielles sont dites de *pattern-matching* car elles reposent sur la détection de la structure harmonique d'un signal comportant une F_0 dans la représentation fréquentielle (ie. spectre d'amplitude). Elles peuvent considérer le module spectral, le cepstre et mettre en œuvre une étape préalable de détection des partiels en relation harmonique. Les approches spectro-temporelles séparent le signal avec un banc de filtres et traitent tous les signaux de sortie (ou canaux). Cette séparation avec un banc de filtres vise à modéliser ce qui se passe dans l'oreille interne qui décompose le signal en temps et en fréquence. La modélisation peut être plus ou moins proche de la biologie. Ces approches utilisent des méthodes temporelles sur chacun des canaux et profitent de la séparation de l'information en bandes de fréquence pour raffiner la prise de décision voisé/non-voisé et l'estimation de F_0. Les trois familles d'approches précédentes sont déterministes. Certains travaux placent l'estimation de F_0 dans un cadre probabiliste et profitent ainsi d'un formalisme bien connu. Ces approches reposent généralement sur un modèle sinusoïdal de la parole et tous les paramètres du modèle sont estimés à partir de connaissances a priori apprises d'un corpus. Toutes les méthodes utilisées fabriquent ce que nous appelons une « fonction de pitch », qui

quantifie la force d'une fréquence en tant qu'hypothèse F_0. En situation monopitch, la fréquence dont l'amplitude de la fonction de pitch est maximale est considérée comme la F_0 de la trame.

Malgré la diversité et la multitude des algorithmes d'estimation de F_0 (AEP) existants, ceux-ci font toujours des erreurs dont les plus fréquentes sont les erreurs d'octave ou de sous-octave. Nous avons mené une analyse fine des erreurs faites ce qui nous a permis de les généraliser. En définitive, elles ne sont pas uniquement des erreurs double ou moitié mais elles sont fractionnaires de type $(p/q)F_0$ ((p, q) entiers non nuls et premiers entre eux). Une classification des erreurs est possible sur une échelle allant de « peu parasite » jusqu'à « très parasite ». Les erreurs les plus parasites sont les erreurs de type sous-multiple ((1/2), (1/3), etc.) et viennent ensuite les erreurs multiples de type ((2/1), (3/1), etc.). L'indexation des pics parasites par un couple (p, q) est à l'origine de notre algorithme d'estimation multipitch. Nous avons cherché à minimiser au maximum l'apparition des pics parasites dans les fonctions de pitch et le fait de connaître la position fréquentielle des pics parasites permet de les traiter spécifiquement.

Ce qui n'est déjà pas simple sur des signaux mono-locuteur l'est encore moins sur des signaux de parole superposée. L'état de l'art en situation multipitch est plus abondant en musique qu'en parole. Toutefois, il existe quelques AEP multipitch en parole qui s'appuient sur une généralisation 2-D du principe de l'autocorrélation (AMDF précisément), sur une méthode spectro-temporelle (découpage par banc de filtre) avec une analyse par autocorrélation sur chacun des canaux pour estimer les candidats F_0, ou bien encore sur la modélisation par une gaussienne de chacune des harmoniques du spectre d'amplitude de la trame et par l'utilisation d'un filtrage de Kalman pour estimer les trajectoires des F_0. L'estimation de F_0 repose toujours sur l'exploitation d'une fonction de pitch même lorsque celle-ci n'est pas explicite. En situation multipitch, ces fonctions de pitch peuvent contenir beaucoup plus de pics parasites. Il apparaît que là encore, les méthodes passent habituellement par une étape de suivi de F_0 pour corriger les erreurs faites par les AEP.

Nous proposons un algorithme d'estimation de F_0 conçu pour traiter des situations multipitch. L'algorithme repose sur le principe de Peigne à Suppression Harmonique (PSH). Pour chacune des trames analysées, l'AEP que nous proposons fournit plusieurs hypothèses F_0, leurs amplitudes et la force de voisement de la trame. Il est conçu pour traiter les situations multipitch sans connaître a priori le nombre de sources et est également utilisable en situation monopitch en ne considérant que le candidat de plus forte amplitude.

L'algorithme est d'approche purement fréquentielle et repose sur trois familles de peignes spectraux. Un peigne spectral est défini par sa fréquence, le nombre de dents et l'amplitude de chacune de ses dents. La fréquence du peigne spectral notée F_c est la fréquence à laquelle se trouve la première dent du peigne. Les autres dents du peigne spectral sont placées aux multiples de F_c. Le nombre de dents peut être limité à une certaine valeur ou bien occuper tout le spectre. L' amplitude de chaque dent peut être positive, négative, suivre une certaine décroissance ou bien être constante. Le choix des amplitudes fixe le comportement du peigne spectral. Les familles de peignes spectraux que nous utilisons sont nommées Peigne Uniforme Infini (PUI), Peigne à Dents Manquantes (PDM) et Peigne à Dents Négatives (PDN). Le PUI est un peigne fréquentiel dont l'amplitude des dents est constante et unitaire. Le nombre de dents est en principe illimité mais le spectre du signal étant borné, les dents situées au-delà de cette borne ne jouent aucun rôle. Le PUI un échantillonneur étendu sur tout le spectre. Le PDM est défini

par un paramètre supplémentaire qui est son ordre (un entier naturel strictement supérieur à un). Un PDM d'ordre deux par exemple est un PUI auquel il manque les dents dont le numéro est multiple de deux. De manière générale, un PDM d'ordre N est un PUI dont il manque les dents de numéro multiple de N. Un PDN est également défini par son ordre. Par exemple, un PDN d'ordre deux de fréquence F_c est un PUI de fréquence F_c auquel on ajoute des dents négatives aux multiples impairs de $F_c/2$. De manière générale, un PDN d'ordre N et de fréquence F_c est un PUI auquel on a ajouté N dents négatives entre deux dents positives successives.

Notre AEP utilise des trames de 40 millisecondes avec 30 millisecondes de recouvrement qui sont extraites du signal de parole superposée. Chacune de ces trames est multipliée par une fenêtre de Hann. Le spectre d'amplitude de chacune des trames est calculé. Pour chacun des peignes spectraux, une fonction de pitch est construite par produit scalaire entre un peigne spectral et le spectre d'amplitude en faisant varier la fréquence du peigne spectral sur une étendue allant de 50 à 800Hz (quatre octaves) couvrant largement l'évolution de la F_0 de la parole lue. Bien entendu, ces paramètres de trame et d'étendue de F_0 peuvent être adaptés selon les besoins et selon le matériau de parole traité.

Si une trame voisée a pour fréquence fondamentale la valeur F_0, son spectre d'amplitude est une structure harmonique dont les harmoniques sont espacées de F_0. Le résultat du produit scalaire entre un PUI et le spectre d'amplitude de la trame est une fonction de pitch qui présente de nombreux maxima locaux. Les plus importants se trouvent en F_0 et tous ses sous-multiples (moitié, tiers, quart, etc.). Des maxima locaux parasites importants se trouvent également aux fréquences multiples (double, triple). Nous montrons que les pics parasites se trouvent aux fréquences (p/q)F_0 (p,q entiers non nuls et premiers entre eux). Avec un PUI, l'amplitude d'un pic parasite est inversement proportionnel à p et ne dépend pas de q. Les pics parasites (p, q) les plus parasites sont ceux avec un p faible et les moins parasites ceux avec p grand. Il est donc possible d'établir une échelle des pics parasites allant du « peu parasite » au « très parasite ». Nous proposons un outil de quantification de l'émergence des pics de F_0 par rapport à tous les autres pics parasites. Cet outil se nomme « émergence » et correspond à l'écart relatif entre l'amplitude des pics F_0 et celle du pic parasite le plus important.

L'identification des pics parasites par le couple (p, q) nous a amené à l'utilisation des PDM et PDN dont l'objectif est de détecter et réduire les pics parasites. Les PDM permettent d'atténuer les pics parasites sous-multiples et les PDN atténuent les pics parasites multiples. Leur utilisation conjointe mène à une forte atténuation de tous les pics parasites, et permet de faire émerger le pic principal $(1, 1)$, même dans un mélange de deux sons. Une contribution de la thèse est d'avoir démontré de manière théorique les fonctionnements et les fonctions de pitch obtenue avec les PUI, les PDM et les PDN dans la situation monopitch. Ces résultats théoriques ont été obtenus à partir d'un spectre d'amplitude qui modélise une structure harmonique idéale à savoir un peigne spectral dont toutes les dents ont la même amplitude. En situation monopitch, nous donnons la valeur précise des amplitudes des pics (p, q). Dans ce cas idéal, tous les pics parasites peuvent être non seulement atténués mais supprimés. Le principe de PSH fonctionne de manière optimale lorsque les structures harmoniques mélangées sont d'amplitude proche et qu'elles possèdent approximativement le même nombre d'harmonique. Par exemple, si une voix domine complètement l'autre, il sera difficile pour PSH de détecter la périodicité de la voix dominée. Nous avons testé la robustesse de l'estimation multipitch dans un cas théorique en faisant varier le nombre de dents de la structure harmonique idéale et en faisant varier l'amplitude des

dents.

Deux implémentations du principe de PSH on été réalisées et testées. Elles se nomment PSH_{min} et PSH_{mul}. L'évaluation a été menée dans le cas mono-locuteur et bi-locuteur. Des corpus de parole réelle, lue et relativement neutre en intonation ont été utilisés (Keele, Bagshaw, Mocha). Ce sont des corpus de signaux mono-locuteur. Une des premières étapes a consisté à normaliser en intensité tous les signaux mono-locuteur en ne considérant que les zones non-silencieuses. Pour générer les signaux bi-locuteur, nous avons additionné toutes les paires possibles de signaux normalisés d'un même corpus en considérant séparément séparant les mélanges de voix de femme/voix de femme, voix de femme/voix d'homme et voix d'homme/voix d'homme. L'évaluation requiert une annotation des F_0 de référence sur laquelle les AEP peuvent être comparés. Cette annotation a été réalisée de manière automatique à partir des signaux EGG de chacun des signaux mono-locuteur des corpus. Les F_0 de référence des signaux bi-locuteur sont obtenus à partir des F_0 de référence des deux signaux mono-locuteur constituant le mélange.

Le critère d'évaluation de l'estimation de F_0 utilisé est le taux d'erreurs grossières ou *Gross Error Rate* (GER) correspondant au taux d'erreurs d'estimation à plus de 20% relatif de la valeur de référence. Le taux d'erreurs fines ou *Fine Error Rate* (FER) renseigne sur l'écart relatif moyen à la référence lorsque l'estimation est correcte. Le taux de précision ou *Precision Rate* (PRR) met en évidence la qualité d'un AEP à donner les bonnes hypothèses F_0 en premier. Dans le cas idéal de la situation monopitch, un AEP doit donner le bon candidat à la première hypothèse. S'il faut regarder jusqu'à la quatrième hypothèse pour trouver le bon candidat, le taux de précision diminue. En situation monopitch, nous avons comparé notre algorithme aux algorithmes YIN, PRAAT et SWIPE. Nous avons testé les algorithmes de quatre manières différentes. La première méthodologie dite « sans sous-voisé » est la plus équitable mais impose aux AEP évalués de fournir une estimation de F_0 pour toutes les trames, ce qui n'est pas toujours le cas. La seconde dite « standard » consiste à tester chacun des algorithmes dans leur version standard fournie par les auteurs. La troisième dite « par normalisation des trames » consiste à tester chacun des algorithmes dans leur version standard mais l'évaluation porte uniquement sur les trames qui sont détectées comme voisées par tous les algorithmes. Ces trames dont le voisement est clairement établi sont toutes placées au centre des segments voisés du signal. Cette méthode d'évaluation renseigne sur la performance d'estimation des algorithmes là où le voisement ne fait aucun doute. Or la majorité des erreurs faites par les AEP sont faites sur les extrémités des segments voisés, là où justement le voisement n'est plus aussi net. La quatrième façon d'évaluer prend en compte la décision voisé/non-voisé (VnV) faite par les algorithmes. Il s'agit de placer chaque AEP testé sur un point de fonctionnement pour lequel le taux d'erreurs de sur-voisé ou *Overvoiced Rate* (la trame de référence n'est pas voisée mais l'algorithme décide que la trame est voisée) et le taux d'erreur de sous-voisé ou *Undervoiced Rate* (la trame de référence est voisé mais l'algorithme décide que la trame n'est pas voisée) sont parfaitement contrôlés par l'expérimentateur. Ces deux taux sont notés respectivement OVR et UVR. Nous montrons que la décision VnV influe beaucoup sur le GER. Selon le point de fonctionnement choisi, le GER peut être augmenté ou diminué. Il est nécessaire qu'une évaluation fournisse l'OVR et l'UVR, tout autant que le GER. En situation monopitch les implémentations de PSH se révèlent être aussi performantes que l'état de l'art pour les quatre méthodologies d'évaluation. L'amélioration des performances de PSH_{mul} lorsque l'AEP a la possibilité de fournir plu-

sieurs hypothèse F_0 a aussi été évalué. Cette situation est nommée sur-hypothétisation. Les évaluations menées ont permis de mettre en évidence la confusion généralement présente dans la littérature concernant la quantification du voisement. Parfois, la décision VnV repose sur l'estimation de F_0 et parfois l'estimation de F_0 repose sur la VnV. Cette confusion provient certainement de la difficulté pour la communauté scientifique de définir convenablement un critère acoustique de voisement.

En situation bipitch, nous avons comparé PSH_{mul} à deux algorithmes multipitch de Wu, Wang & Brown (WWB) et à celui de Vishnubhotla & Espy-Wilson (VEW). Les aspects de décision VnV sont plus compliqués à définir en situation multipitch. Il a été décidé qu'une trame est considérée comme voisée lorsqu'au moins une de ses deux références F_0 est voisée. Les critères de qualités utilisés sont deux taux de GER : le taux d'erreur grossière sur les candidats (GER^C) et le taux d'erreur grossière sur les trames (GER^T). Le premier renseigne sur la qualité d'un AEP à correctement estimer toutes les F_0 de référence et le second renseigne sur la qualité d'un AEP à trouver toutes les F_0 de référence de chacune des trames. Deux taux de précision PRR_1 et PRR_2 sont aussi utilisés pour quantifier la qualité d'un AEP à donner les bonnes hypothèses F_0 en premier. Le taux de précision PRR_1 concerne les trames monopitch (1 valeur F_0 de référence voisée) et le taux de précision PRR_2 concerne les trames bipitch (2 valeurs F_0 de référence voisées). En situation bipitch, un AEP doit idéalement donner les deux bons candidats dans les deux premières hypothèses. Sinon, le taux de précision diminue. PSH_{mul} est évalué avec les trois méthodologies décrites en situation monopitch. La première évaluation décrite comme standard donne les résultats obtenus par PSH_{mul}, WWB et VEW dans leurs réglages par défauts. La seconde méthodologie reprend les AEP dans leur réglage standard mais les teste uniquement sur les trames considérées voisées pour tous les AEP. La troisième méthodologie donne à l'expérimentateur la possibilité de normaliser le point de fonctionnement en termes de décision VnV. Toutefois, Les deux algorithmes WWB et VEW n'étant pas paramétrables, la troisième évaluation consiste à régler PSH_{mul} sur le même point de fonctionnement que WWB puis VEW. Une quatrième évaluation reprenant la méthodologie par normalisation des trames est présentée. Elle permet de mettre en évidence l'effet positif de proposer plus d'hypothèse que de référence F_0 à estimer. Elle permet aussi de montrer les taux de PRR moyens obtenus par PSH_{mul}. Avec PSH_{mul}, les taux de GER^C moyens obtenus sont de l'ordre de 10% et les taux moyens de GER^T sont de l'ordre de 15%.

L'estimation de F_0 multiples peut être intégrée dans de nombreuses applications. L'application la plus directe est le suivi des F_0 de chacun des locuteurs du mélange. Des contraintes de continuité temporelle et fréquentielle peuvent permettre de regrouper des valeurs F_0 isolées en segments F_0 plus longs. L'estimation de F_0 multiples peut être aussi utilisée pour améliorer les systèmes de reconnaissance de parole actuels qui se heurtent à des baisses de performances sur les segments de parole superposée. Enfin, l'estimation de F_0 multiples peut aider la séparation de parole ce qui est notre objectif à long terme. En changeant de matériau pour passer de la parole à la musique, une des perspectives la plus directe est de tester notre AEP dans des tâches de transcription automatique de voix chantée ou d'instruments de musique.

Chapitre 2

Estimation de F_0 et séparation de parole

2.1 Introduction

Les environnements sonores dans lesquels nous évoluons sont extrêmement variés (campagne, rue, salle de concert, gare, etc.). Quelque soit le contexte sonore, il est captivant de constater à quel point notre perception auditive est performante dans la reconnaissance et la compréhension des sources sonores qui nous entourent. Cette compréhension est le résultat d'une faculté de séparation de sources qui intrigue et inspire la communauté scientifique. Une situation particulière de séparation de sources est la situation de *Cocktail Party* (CP) décrite par Cherry [Cherry, 1953, Cherry and Taylor, 1954]. Il introduit le terme de Cocktail Party pour désigner la situation dans laquelle plusieurs locuteurs parlent simultanément. Le terme de « parole superposée » est utilisé dans le domaine de la reconnaissance de parole pour désigner la situation de mélange de parole. Dans le reste du manuscrit, parole superposée et Cocktail Party seront utilisé indifféremment. Lorsque plusieurs locuteurs parlent simultanément nous parvenons aisément à suivre à volonté le locuteur qui nous intéresse. Ceci est rendu possible par la faculté de séparation de parole que possède notre système auditif. Cette faculté est issue de mécanismes complexes faisant intervenir d'une part des connaissances a priori sur les sources sonores et d'autre part des traitements effectués sur l'onde sonore perçue, parmi lesquels se trouve l'estimation de F_0. Cette thèse est orientée vers l'estimation de F_0 et propose un nouvel algorithme d'estimation de F_0 (AEP) conçu pour traiter des signaux de parole superposée. Cet algorithme d'estimation de F_0 a été élaboré de manière à être l'un des premiers étage d'un système automatique de séparation de parole. Il sera détaillé au chapitre 5 et évalué au chapitre 6. L'objectif du présent chapitre est de clarifier notre positionnement scientifique en posant au préalable l'ensemble des définitions nécessaires à la compréhension du manuscrit. Il convient de clarifier les termes de F_0, d'estimation de F_0 et de séparation de parole. Ces trois points sont respectivement abordés dans les paragraphes 2.1.1, 2.1.2 et 2.1.3. Une fois ces définitions posées, la section 2.2 décrit en quoi la F_0 est un indice acoustique utilisé par notre système auditif pour séparer de la parole. La section 2.3 montre en quoi l'estimation de F_0 peut être utile pour un système de séparation de parole.

2.1.1 Fréquence fondamentale, F_0 et pitch

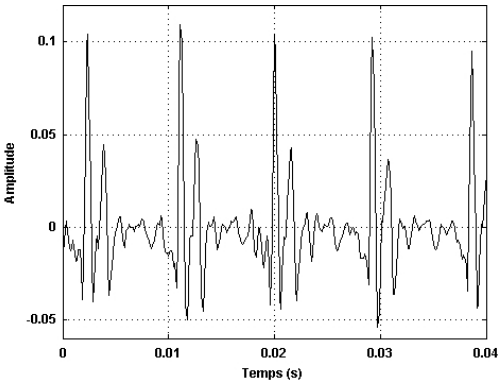

FIGURE 2.1: Illustration de la quasi-périodicité d'une voyelle réelle.

La fréquence fondamentale F_0 est une grandeur physique qui est à rapprocher de la manière dont est produit le son. Lorsqu'une source sonore émet un son produit par la vibration de certaines de ses parties (biologiques ou non), la fréquence de vibration est nommée F_0. Pour un signal de parole, elle correspond à la fréquence de vibration des cordes vocales d'un locuteur. Le signal de parole n'est pas le seul signal qui puisse comporter une F_0. Un signal de musique émis par un instrument comporte également une F_0 qui correspond à la fréquence de vibration de l'instrument (ex : vibration de la corde pour un piano ou un violon, vibration des lèvres du trompettiste pour une trompette). La F_0 est donc une caractéristique de la source sonore. Il faut ici distinguer la F_0 qui est une grandeur physique propre à la source sonore et la hauteur ou *pitch* qui est une grandeur perceptive donc propre au récepteur du signal sonore. Toutefois, cette distinction n'est pas toujours respectée strictement. Il existe des expressions consacrées qui sont apparues dans le domaine de l'estimation de F_0 comme « estimation multipitch » et pour lesquelles pitch est employé pour désigner F_0. Dans le reste du manuscrit, F_0 ou pitch pourront être utilisés indifféremment tout en gardant à l'esprit que ces termes se réfèreront toujours à la grandeur physique F_0 et donc à la vibration des cordes vocales d'un locuteur.

La périodicité d'un signal quelconque s au sens strict et mathématique du terme est donné dans l'équation 2.1 [de Cheveigne, 2006]. T représente une des périodes du signal s. T est une des périodes de s car en définitive il en existe une infinité. En effet si T est une période de s, alors tous les multiples de T le sont aussi.

$$\exists T \in \mathbb{R}_+^*, \forall t \in \mathbb{R}, s(t+T) = s(t) \tag{2.1}$$

Par définition, la période fondamentale notée T_0 correspond à la plus petite des valeurs de T, périodes de s. L'équation 2.2 formalise mathématiquement $T0$. La notation $\{T\}$ correspond à l'ensemble des périodes de s. La fréquence fondamentale F_0 est l'inverse de la période fondamentale.

$$\begin{aligned} T_0 &= min(\{T\}) \\ F_0 &= \tfrac{1}{T_0} \end{aligned} \tag{2.2}$$

La périodicité d'un signal de parole réelle ne vérifie pas strictement la périodicité mathématique. Toutefois le signal acoustique présente une série de motifs suffisamment ressemblants pour que l'on puisse parler de quasi-périodicité. Un exemple de quasi-périodicité est présentée dans la figure 2.1. Le signal présenté est un extrait de 40ms d'une voyelle /a/.

2.1.2 Estimation de F_0 monopitch et multipitch

L'estimation de F_0 est le problème central de cette thèse. Il faut distinguer deux situations d'estimation de F_0. La première situation concerne les signaux dits «monopitch». Ces signaux monopitch comportent au plus une F_0. Par exemple des signaux de parole mono-locuteur ou de musique non

polyphoniques (ex : la flute) enregistrés en solo. De manière générale, les signaux monopitch peuvent comporter plusieurs sources sonores mais exactement une seule de ces sources comporte une F_0. La situation d'estimation de F_0 pour des signaux monopitch est nommée elle-même estimation de F_0 monopitch. Dans ce cas, l'estimation revient à estimer à chaque instant la fréquence de vibration de la source sonore étudiée à partir de l'analyse de la forme d'onde du signal reçu. Cela revient à analyser le signal reçu du coté récepteur et retrouver la fréquence fondamentale de l'émetteur avec toutes les distorsions apportées par le canal de communication entre les deux. Pour la parole, l'estimation de F_0 monopitch revient à calculer en chaque instant la fréquence de vibration des cordes vocales d'un locuteur à partir de la forme d'onde d'un signal de parole. Voici quelques exemples des applications possibles de l'estimation de F_0 en situation monopitch :

- Séparation d'un signal de parole mélangé à un autre signal intrus ne comportant pas de F_0. Cette application se nomme débruitage.
- Analyse du contour intonatif d'un locuteur. Cette analyse peut participer à l'étude de l'intonation des langues ou bien encore à l'étude de l'expressivité dans la parole.
- Aide à l'élaboration d'un codage plus efficace de la parole. En effet, la connaissance de la F_0 peut permettre de coder de manière plus efficace en représentant tout une structure harmonique par une simple valeur de F_0.
- Transcription de la voix chantée ou d'instruments monophoniques.
- Requête par fredonnement dans des systèmes d'indexation multimédia.

La seconde situation concerne les signaux dits « multipitch ». Ces signaux multipitch sont des signaux qui comportent plusieurs F_0. De tels signaux se trouvent aisément dans les signaux musicaux qui sont polyphoniques par nature. Ils se trouvent également en parole dans des situations de parole superposée dans lesquelles plusieurs signaux mono-locuteur sont mélangés dans un même canal. La situation de parole superposée est souvent moins polyphonique que la musique et il est assez rare d'avoir plus de deux signaux de parole de niveau sonore comparable mélangés. Dans nos expériences, les évaluations porteront sur une situation de parole superposée à deux locuteurs que nous nommerons situation bi-locuteur ou bipitch. Toutefois, l'algorithme d'estimation que nous proposons ne requiert pas la connaissance du nombre de locuteurs et pourra donc être utilisé dans des situations où le nombre de locuteurs mélangés est supérieur à deux et n'est pas connu à priori. Actuellement, l'estimation de F_0 pour des signaux de parole mélangés est en pleine émergence sous l'impulsion de plusieurs facteurs, par exemple :

- L'importance emblématique de la situation de CP.
- L'importance de domaines de recherche comme la reconnaissance de parole qui analysent des signaux de moins en moins contraints, de plus en plus spontanés et qui se retrouvent donc dans des situations de CP. Par exemple, l'analyse de données d'interviews politiques français [Adda et al., 2007] montre que sur huit heures de débat politique, le taux de parole superposée est d'environ 5% soit environ 25 minutes. Les systèmes de reconnaissance de parole actuels fonctionnent mal sur ces segments.

L'ensemble de ces perspectives d'application justifie l'intérêt porté à l'estimation de F_0 multipitch. Des exemples de débouchés applicatifs sont donnés dans la liste suivante :

- Aide à la séparation de la parole. L'ambition initiale de la thèse est d'utiliser l'estimation de

F_0 pour de la séparation de parole. Ce point est détaillé ci-dessous et dans les sections 2.2 et 2.3.

- Aide à l'estimation du nombre de locuteurs d'une situation de parole superposée. La connaissance des évolutions du pitch de chacun des locuteurs à chaque instant apporte une information sur le nombre de locuteurs de la scène auditive.
- Aide à la séparation des instruments composants une scène musicale.
- Aide à la transcription automatique d'instruments de musique polyphoniques tels que le piano.

L'estimation de F_0 apparaît comme une étape de traitement pouvant se placer en amont de nombreuses applications. Dans cette thèse, l'estimation de F_0 est justement envisagée comme un traitement de bas niveau d'un système de séparation de parole. Cette façon de considérer le problème oblige à rendre l'étape d'estimation de F_0 la plus fiable possible. En effet, plus l'étape d'estimation de F_0 est fiable, moins les traitements de niveau supérieur seront nécessaires pour corriger les erreurs. L'élaboration d'un algorithme d'estimation de F_0 robuste et performant est une étape préliminaire obligatoire dans la construction de systèmes plus complexes.

2.1.3 Séparation de parole

La séparation de parole a pour objectif d'élaborer un système automatique capable de traiter un mélange de signaux de parole pour en extraire un, plusieurs ou tous les signaux de parole qui constituent le mélange. Plusieurs définitions de séparation de parole peuvent être envisagées :

- Retrouver tous les signaux constituants le mélange. Par exemple, séparer deux locuteurs pour être capable de restituer l'un ou l'autre au choix, comme s'il était isolé. C'est le problème le plus difficile, voisin du problème de Cocktail Party.
- Isoler une des composantes du mélange et la suivre au cours du temps. Cela ressemble à du débruitage (en anglais *enhancement*). Ce problème est bien connu et traitable quand le bruit a une structure identifiable et distincte de la cible. Pour de la parole superposée, le « bruit » est de même nature que la cible et le problème ne peut donc pas se résumer à du débruitage.

Pour chacune de ces acceptions, plusieurs niveaux de séparation peuvent être définis pour le signal (ou les signaux) résultant(s) :

- Reconstruire un signal dont la forme d'onde est la même que le signal isolé. L'évaluation peut dans ce cas utiliser des critères d'évaluation objectifs comme une distance euclidienne par exemple.
- Reconstruire un signal perceptivement équivalent au signal isolé. Il faut alors utiliser des critères d'évaluation subjectifs et passer par des expériences d'écoute.
- Identifier la présence ou non dans le mélange d'"une information spécifique à l'un des signaux. Ceci doit permet de reconnaître une source sonore d'intérêt, sans pour autant devoir reconstruire le signal correspondant de manière exacte ou approchée.

Dans ce travail, la séparation de parole est considérée comme la capacité de suivre au cours du temps un des signaux composant le mélange. Ce suivi n'implique pas forcément la reconstruction du signal mais plutôt la détection à chaque instant de tel ou tel indice acoustique caractérisant tel locuteur. En particulier, un objectif pourra être de reconstruire les trajectoires des F_0 de chacun des locuteurs au cours du temps.

Un système automatique capable de séparer la parole peut donner lieu à des applications telles que le débruitage, la reconnaissance de parole, la reconnaissance de locuteur ou l'amélioration des prothèses auditives. Depuis plusieurs décennies, des efforts de recherche conséquents ont été menés autour de la thématique du traitement automatique de la parole. Néanmoins, aucun des systèmes automatiques actuels élaborés en séparation de parole, en reconnaissance de parole ou en débruitage ne fonctionne parfaitement. Des témoignages de malentendants indiquent que l'efficacité perceptive des prothèses auditives devrait être encore largement améliorée, notamment dans la situation où plusieurs interlocuteurs parlent simultanément. Pourquoi ces systèmes restent-ils encore perfectibles ? Cela s'explique en partie par le fait que tous sont conçus pour fonctionner sous certaines conditions. Avec une gamme restreinte de signaux par exemple, ou encore avec des signaux propres et isolés. Dès lors qu'un de ces systèmes est soumis à un signal qui sort des limites de fonctionnement, il ne fonctionne plus aussi bien. Par exemple, les systèmes actuels de reconnaissance de parole fonctionnent bien lorsque le signal à analyser est un signal isolé et exempt de bruit et de réverbération (enregistrement avec un micro près de la bouche ou en salle anéchoïque) mais leurs performances se dégradent nettement quand ils sont utilisés en environnement réel. De même les systèmes actuels de séparation de parole ont généralement besoin d'avoir autant de microphones que de sources sonores dans la scène auditive. Enfin, les personnes malentendantes ont des problèmes de compréhension lorsqu'elles sont immergées dans des situations de mélange de parole. Or, en condition naturelle, les signaux que nous percevons ne sont pas isolés mais mélangés. Ces signaux ne sont pas exclusivement de la parole mais peuvent être de toute nature. Les sources sonores peuvent être mobiles et leurs intensités respectives sont fluctuantes. L'environnement sonore introduit également de la réverbération. Le système auditif des personnes normo-entendantes sait très bien traiter ces signaux divers, mélangés et bruités. Si nous voulons un jour être en mesure d'améliorer les systèmes d'aide aux malentendants il faut commencer par situer nos études dans des conditions plus proches de la communication de tous les jours, c'est-à-dire considérer que la situation normale est d'avoir à traiter plusieurs signaux simultanés.

Pour aborder ces problèmes il est judicieux d'essayer de comprendre comment le système auditif humain nous permet d'être aussi performant dans la séparation de parole. Il convient donc d'étudier les travaux menés en Analyse de Scènes Auditives ou *Auditory Scene Analysis* (ASA). Une scène auditive correspond à un contexte sonore particulier constitué de plusieurs sources sonores. L'ASA a pour objet d'étude la perception auditive humaine immergée au sein de scènes auditives diverses et notamment en situation de parole superposée. En parole superposée, une scène auditive est constituée uniquement de locuteurs. Pour le problème d'estimation de F_0, il est nécessaire d'observer les études menées par l'ASA sur l'utilité de la F_0 dans des situations de séparation de parole. La section 2.2 de ce chapitre présente les travaux clés de l'ASA dans l'analyse de l'apport de la F_0 en situation de Cocktail Party. Néanmoins, les études psycho-acoustiques de l'ASA ne sont pas suffisantes pour construire des systèmes de séparation de parole. Il est également intéressant d'étudier les systèmes de séparation de parole existants pour déterminer l'importance que revêt l'utilisation de la F_0 dans ces systèmes. La section 2.3 présente ces systèmes de séparation de parole en précisant lesquels utilisent la F_0 et lesquels non. Enfin, la section 2.4 récapitulera les points importants de ce chapitre et notre positionnement scientifique.

2.2 Importance de la F_0 dans l'ASA

Dans les travaux de Cherry [Cherry, 1953, Cherry and Taylor, 1954], les auteurs constatent que les auditeurs normo-entendants sont particulièrement doués pour écouter et comprendre à volonté le locuteur qui les intéresse. Cherry définit alors le *Cocktail Party Problem* par la question suivante :

Comment faisons-nous pour reconnaître ce que dit une personne quand d'autres parlent en même temps alors qu'un mélange de parole complexe parvient à nos deux oreilles ?

Ses expériences sont organisées en deux parties décrites ci-dessous :

1. Faire entendre le mélange de deux signaux de parole mélangés dans les deux oreilles d'un sujet. C'est l'écoute diotique. Le mélange de deux signaux de parole est le cas le plus élémentaire de situation de CP. La tâche du sujet consiste à répéter ce qui est dit par chacun des locuteurs. Le sujet ne peut pas écrire ce qu'il entend mais peut répéter l'enregistrement autant de fois qu'il le souhaite. Il est donc obligé de se concentrer tour à tour sur l'un ou l'autre des signaux de parole.
2. Faire entendre les deux signaux de parole non mélangés, un dans l'oreille gauche et un dans l'oreille droite. C'est l'écoute dichotique. La tâche des sujets consiste à répéter ce qui est dit dans une des deux oreilles. Un des signaux est le signal cible sur lequel le sujet doit se concentrer. L'autre signal est le signal interférant.

Les sons de parole utilisés sont des textes lus. Les deux expériences font intervenir une focalisation attentionnelle de la part des auditeurs. Les résultats de l'expérience 1 montrent que les sujets ressentent la tâche comme difficile et sont très concentrés durant la réalisation. Toutefois, ils réussissent convenablement à reconnaître ce qui est dit à quelques erreurs près. Cette expérience 1 souligne l'efficacité de notre faculté de séparation en situation de CP. La concentration manifeste induite par la tâche montre aussi que l'attention joue un rôle important dans sa réussite. Les sujets de l'expérience 2 indiquent ne plus ressentir de difficultés particulières et réussissent la tâche sans erreurs. Cette expérience semble montrer que notre capacité à focaliser notre attention d'une oreille à l'autre est aisée. Cette facilité de focalisation attentionnelle se traduit par de meilleures performances. Dans l'expérience 1 (écoute diotique) Cherry montre que les sujets font moins d'erreurs lorsque le mélange est constitué de phrases « clichées » c'est-à-dire de phrases simples et relativement typiques (ex. « *The man in the street* »). Cette expérience sous-entend l'existence de mécanismes de prédiction qui apporteraient une aide importante dans la séparation. L'auditeur doit en effet être capable de comprendre ce qui est dit par un des locuteurs alors même qu'à certains moments, l'autre locuteur domine complètement la voix cible. L'auditeur doit alors trouver un moyen pour combler ces « trous ». Nous pouvons émettre l'hypothèse que l'auditeur utilise ses connaissances pour prédire ce qui va être dit en fonction de ce qui a déjà été entendu. Les connaissances qu'il va utiliser ici sont de l'ordre de la connaissance de la langue (ici l'anglais) qui possède une construction bien spécifique et qui n'autorise pas certaines combinaisons de mots alors que d'autres sont fortement probables. Il semble que plus notre expertise du signal entendu est grande, plus il est facile de le séparer des signaux interférants. Dans l'expérience 2 (écoute dichotique), Cherry étudie aussi ce qui est perçu par l'auditeur dans l'oreille correspondant au message interférant. Il montre qu'aucun des auditeurs n'est capable de répéter ce qui a été prononcé

dans le message interférant. Cela sous-entend qu'une fois l'attention dirigée vers le message d'intérêt, le reste ne compte plus. Ce phénomène a été également constaté dans le domaine de la vision. Il a été montré que lorsque nous suivons un objet dans une scène visuelle il est très difficile de voir ce qui se passe ailleurs [Rensink, 2000]. Les travaux de Cherry sont antérieurs à l'ASA mais ont indéniablement initié et inspiré ce domaine de recherche. Ils montrent que des connaissances préalables sur les signaux entendus permettent de traiter plus facilement le problème de Cocktail Party. Ils montrent également le rôle essentiel des mécanismes attentionnels, sur lesquels nous ne nous attarderons pas ici.

L'ASA s'intéresse aux mécanismes perceptifs qui font émerger notre faculté de séparation de source sonore. L'approche ASA consiste à étudier les comportements et les performances d'auditeurs soumis à des simuli sonores. Par exemple, les auditeurs ont pour tâche de reconnaître des mots prononcés dans un signal de mélange de parole. Certaines des expériences psycho-acoustiques vont être présentées dans cette section. La présentation la plus complète de l'ASA se trouve dans le livre de Bregman intitulé *Auditory Scene Analysis : The Perceptual Organization of Sounds* [Bregman, 1990]. Les résultats obtenus par l'ensemble des travaux de l'ASA amènent l'auteur à poser une analogie entre l'analyse de scènes visuelles et l'analyse de scènes auditives. De la même façon qu'un humain décompose une scène visuelle en objets visuels, une scène auditive peut être décomposée en objets sonores. Ces objets sonores sont nommés « flux auditifs » ou « *streams* » en anglais. Généralement, un flux auditif correspond à une source sonore de la scène auditive mais ce n'est pas nécessairement le cas. Il se peut que plusieurs sources sonores participent à un seul flux auditif. La construction de flux auditifs s'appelle le *streaming*. C'est le streaming qui rend une scène auditive intelligible puisqu'il l'organise en objets sonores distincts. Bregman montre que le streaming fait intervenir des traitements *bottom-up* (ou ascendants) qui s'opèrent à partir du signal sonore. Ces traitements aboutissent à une structuration de bas niveau, quasi-automatique, qui se traduit par deux modes de groupement, séquentiel et simultané. Ces modes de groupement font l'objet des paragraphes 2.2.1 et 2.2.2 respectivement. Les traitements ascendants regroupent les indices acoustiques selon des critères de continuité temporelle, fréquentielle, de ressemblance, etc. Ces groupements participent au *glimpsing* présenté dans paragraphe 2.2.3. C'est dans ces traitements bottom-up qu'interviennent les F_0 de chacune des sources. Le streaming fait aussi intervenir des traitements *top-down* (ou descendants) qui s'appuient sur les connaissances d'un individu. Ces traitements aboutissent à une structuration de haut niveau dite en « schémas » propre à chaque type de signal et largement apprise au cours du développement de l'individu. Au fur et à mesure de son exposition à des stimuli sonores, l'individu apprend à reconnaître certains traits distinctifs qui marquent l'identité des sources sonores. Ces traits peuvent être de tous ordres : le timbre d'une voix familière, la structure d'une langue, etc. Les travaux de Cherry décrits précédemment illustrent ces traitements descendants. Dans cette thèse l'approche suivie est purement ascendante même si nous sommes conscients de l'importance des traitements descendants. Les deux modes de groupement de la structuration ascendante sont :

- Le groupement séquentiel qui regroupe en un même flux auditif des événements sonores ou des indices acoustiques ayant une certaine continuité dans le temps.
- Le groupement simultané qui regroupe en un même flux auditif des événements sonores ou des indices acoustiques ayant des propriétés d'harmonicité ou de continuité dans l'espace des fréquences.

Ces deux groupements s'effectuent dans les dimensions temporelle et fréquentielle ce qui est en accord avec la décomposition biologique d'un signal sonore sur la membrane basilaire de la cochlée (cf. chapitre 3 figure 3.25 pour plus de détails). Il est important de voir dans quelle mesure la F_0 participe à ces groupements. Pour cela, trois familles de paradigmes ont été étudiés. Le paradigme dans lequel des séquences de signaux sont soumis à des auditeurs, le paradigme des doubles voyelles dans lequel deux voyelles sont mélangées et sont soumises à des auditeurs. Enfin, l'analyse des différences de comportement auditif entre des auditeurs normo-entendants et malentendants pour mettre en évidence les indices acoustiques importants pour la tâche de séparation de parole. L'objectif du paragraphe 2.2.1 ci-dessous est de décrire les expériences clés utlisant la F_0 pour réaliser le streaming par groupement séquentiel. Le paragraphe 2.2.2 présente les expériences reflétant l'utilité de la F_0 pour le groupement simultané. Il est intéressant de constater que le mécanisme de streaming a été intégrée dans la théorie plus générale du glimpsing. Cette théorie est introduite dans le paragraphe 2.2.3 et explique en quoi la domination d'une source par rapport aux autres dans une région temps-fréquence donnée favorise le streaming. Enfin le paragraphe 2.2.4 montre l'utilité de la F_0 dans le streaming par l'analyse des différences de performances de séparation de parole pour des malentendants et des normo-entendants.

2.2.1 Groupement séquentiel

Un des premiers travaux mettant en évidence le groupement séquentiel [Miller and Heise, 1950] utilise le protocole présenté dans la figure 2.2. Ce protocole est un des paradigmes les plus répandus dans l'étude du groupement séquentiel. Il s'agit de faire entendre à un auditeur une séquence de motifs sonores alternés A-B-A-B dont on fait varier une caractéristique acoustique pour étudier son impact sur le streaming. La première expérience (figure 2.2 haut) est une séquence alternée A-B-A-B de sinusoïdes pures de mêmes amplitudes mais de fréquences fondamentales différentes. L'expérience montre que lorsque l'écart relatif entre les deux fréquences est inférieur à 15% environ (soit un peu plus d'un ton–12%) la séquence est perçue comme une mélodie unique ce qui met en évidence un groupement séquentiel. Au-dessus de ce seuil, la séquence a tendance à être séparée en deux mélodies distinctes. Dowling [Dowling, 1968] et Van Noorden [van Noorden, 1975] donnent un nom aux deux

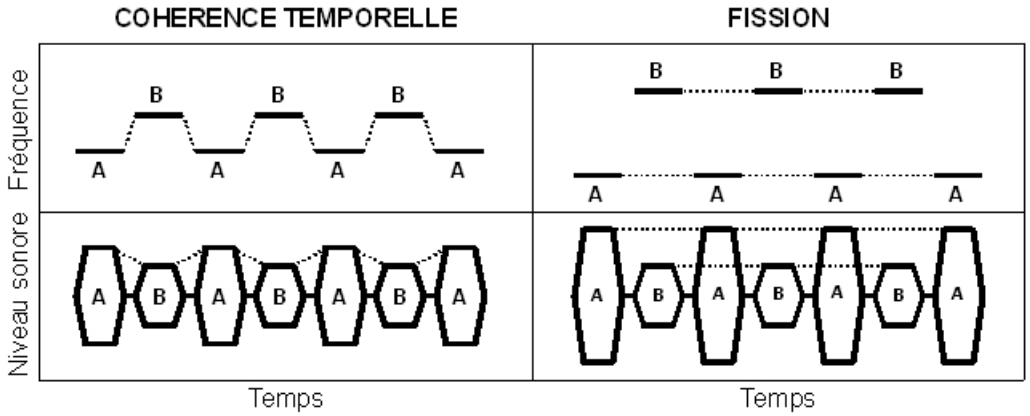

FIGURE 2.2: Protocoles expérimentaux de Van Noorden.

comportements différents pour un auditeur qui entend ces stimuli. Plus les différences entre fréquence ou bien entre amplitude sont importantes, plus le système auditif a tendance à procéder à une fission (colonne de droite). Ceci correspond à la séparation de la scène auditive en deux flux auditifs. Les auditeurs déclarent ne percevoir qu'un seul des flux auditif à la fois (soit A-A soit B-B). Dans le cas contraire, il y a cohérence temporelle (colonne de gauche) et la scène auditive est perçue comme un seul flux auditif. Les auditeurs perçoivent l'ensemble de la séquence soit A-B-A-B. La notion de flux auditif apparait dans [Bregman and Campbell, 1971] et correspond à la fission. Ces travaux illustrent bien le rôle de la F_0 dans le groupement séquentiel de sons purs. Van Noorden [van Noorden, 1977] montre également que la ségrégation en flux auditifs peut s'opérer par la différence en amplitude (cf. figure 2.2 bas). Cette seconde expérience est une séquence alternée de sinusoïdes pures de mêmes fréquences mais d'amplitudes différentes. Ce travail montre d'ores et déjà que la F_0 n'est pas le seul indice acoustique à pouvoir produire de la cohérence temporelle ou de la fission. Le phénomène de fission est-il encore présent lorsque des sons de parole sont utilisés ? Dans le signal de parole, les segments temporels qui comportent une F_0 sont les segments dits « voisés » (cf. chapitre 3 figure 3.4). Les segments voisés correspondent majoritairement à des voyelles. Il est donc intéressant d'étudier si les voyelles produisent du streaming. Le groupement séquentiel est remis en évidence dans une des expériences décrites dans l'article de Gaudrain et al. [Gaudrain et al., 2007] et dans sa thèse [Gaudrain, 2008]. Il y présente une séquence de six voyelles synthétiques à un sujet qui doit reconnaître ces six voyelles et établir l'ordre d'apparition. Trois de ces voyelles ont leur F_0 réglée à F0(1) et trois autres ont un F0(2) valant FB. La séquence est une alternance d'une voyelle dont la F_0 vaut F0(1) et d'une autre voyelle où la F_0 vaut F0(2) (cf. figure 2.3). Les auteurs proposent deux cadences de séquence à 5.7 et 7.4 voyelles par seconde. Ces correspondent à des différences de durée des voyelles 5.7 voyelles par seconde correspond à des voyelles de 175ms et 7.4 voyelles par seconde correspond à des voyelles de 135ms. La F0(1) est toujours fixée à 100Hz et les auteurs font varier F0(2) sur 10 valeurs possibles. Les résultats obtenus

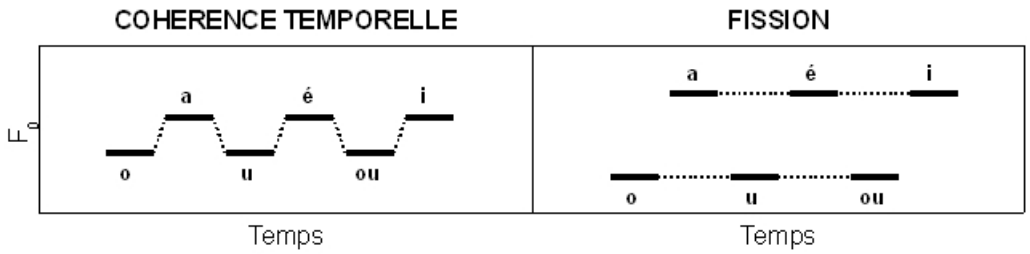

FIGURE 2.3: Séquences de voyelles synthétiques de Gaudrain et al.

sont présentés dans la figure 2.4. La partie de gauche ACROSS présente le score de reconnaissance de la séquence entière de six voyelles en fonction de la cadence des voyelles et de la différence entre F0(1) et F0(2). Pour une même différence de F_0, les scores de reconnaissance diminuent si la cadence des voyelles augmente. En effet, si la cadence des voyelles augmente, l'auditeur dispose de moins de temps pour mémoriser les voyelles et est donc moins performant. Pour une même cadence de voyelles, les scores de reconnaissance diminuent si l'écart entre les F_0 augmente. Cela s'explique par le fait qu'un écart important des F_0 défavorise le groupement séquentiel. Le streaming est donc favorisé et

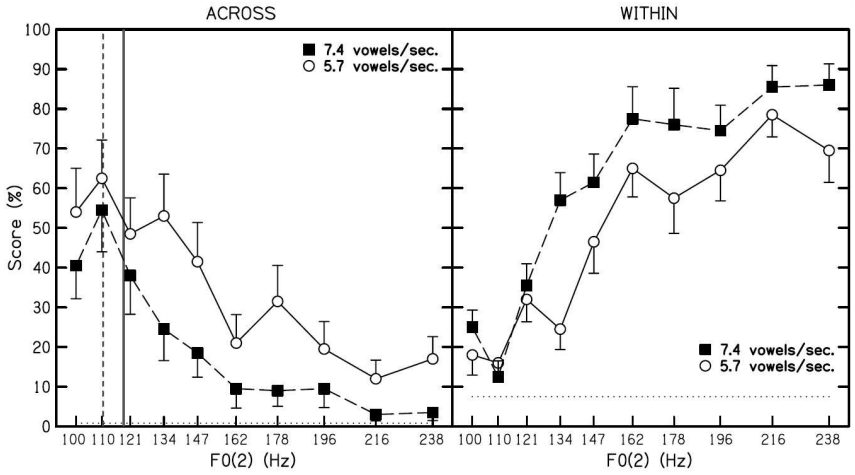

FIGURE 2.4: Groupement séquentiel sur des séquences de voyelles. Taux de reconnaissance obtenus par Gaudrain et al. Source : [Gaudrain et al., 2007].

l'auditeur ne peut plus suivre correctement la séquence de six voyelles. La tâche de reconnaissance d'une séquence de six voyelles semble être difficile car le taux maximal de reconnaissance d'une séquence de 6 voyelles se trouve autour de 50% pour de faibles écarts de F_0 et tombe rapidement autour des 20%. La partie de droite WITHIN présente le score de reconnaissance d'une séquence de trois voyelles ayant la même F_0 en fonction de la cadence des voyelles et de la différence entre F0(1) et F0(2). Pour une même cadence de voyelles et lorsque l'écart des F_0 augmente, les scores augmentent car le streaming est favorisé. L'auditeur se focalise plus facilement sur un des deux groupes de trois voyelles qui ont le même F_0 et parvient à en donner la séquence. Pour une même différence de F_0, les scores de reconnaissance de trois voyelles ont deux comportements qui dépendent de la différence entre F0(2) et F0(1). Si F0(2) est supérieur à 121Hz (soit 21% de plus que F0(1) ce qui est l'équivalent de 3 demi-tons), les scores augmentent avec l'augmentation de la cadence ce qui paraît logique dans la mesure où l'écart des F_0 étant important et le débit étant important, la ségrégation en flux est largement favorisée impliquant que trois voyelles de même F_0 vont être bien reconnues. En revanche, si la différence de F_0 est trop faible entre les deux groupes de trois voyelles, la cadence ne semble pas jouer de rôle. Les valeurs des taux de reconnaissances sont autour de 70% pour des écarts de F_0 de plus de 50%. La tâche dans ce cas est certainement plus simple puisqu'il s'agit de reconnaître l'ordre de 3 voyelles parmi 6 et non la séquence complète des 6. Pour de faibles écarts de F_0 (inférieur à 20%), le taux de reconnaissance est d'environ 20% ce qui révèle que la séquence de 6 voyelles parvient difficilement à être séparée en deux flux auditifs.

Ce paragraphe met en évidence le rôle de la F_0 dans le groupement séquentiel qui est un des mécanismes à notre disposition pour résoudre le problème de Cocktail Party. Ce groupement est utile dans des séquences temporelles successives et distinctes. Le groupement séquentiel révèle le rôle ségrégateur de la F_0 dans sa continuité temporelle. Néanmoins, dans une situation de CP, les sources ne sont plus nécessairement en séquences temporelles mais simultanées. L'objectif du paragraphe 2.2.2 ci-dessous est de décrire les expériences clés décrivant les conditions de streaming par groupement simultané.

2.2.2 Groupement simultané

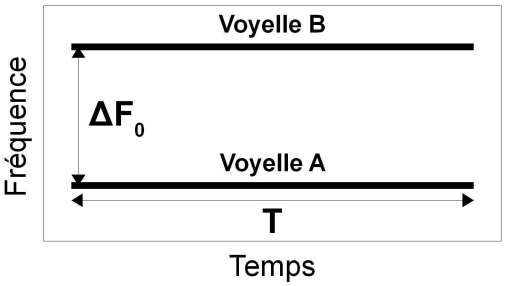

FIGURE 2.5: Paradigme des deux voyelles.

Le paradigme le plus classique qui permet d'étudier le groupement simultané est le paradigme des doubles voyelles. Il consiste à mélanger deux voyelles dans un même signal. Les auditeurs ont ensuite pour tâche de reconnaître les deux voyelles constituantes du mélange en écoute diotique. La variété des réglages de ce paradigme est importante. Le choix d'utilisation des voyelles synthétiques est justifié par le besoin de contrôle et de réglage des stimuli. La figure 2.5 illustre le principe du paradigme des doubles voyelles. Le mélange est constitué de la superposition d'une voyelle A et d'une voyelle B toutes deux stables en F_0. La durée des voyelles est réglable et vaut T dans l'exemple. Les expérimentateurs font varier T et F_0 afin de déterminer les relations entre ces paramètres et le streaming. Les travaux de Sheffers [Scheffers, 1983], Zwicker [Zwicker, 1984], Assmann et Summerfield [Assmann and Summerfield, 1990] et de Culling [Culling and Darwin, 1993] utilisent des voyelles dont la F_0 est stable tout le long de la durée du stimuli. Ces travaux montrent qu'une différence de F_0 des deux voyelles apportent une amélioration dans la reconnaissance des deux voyelles mélangées.

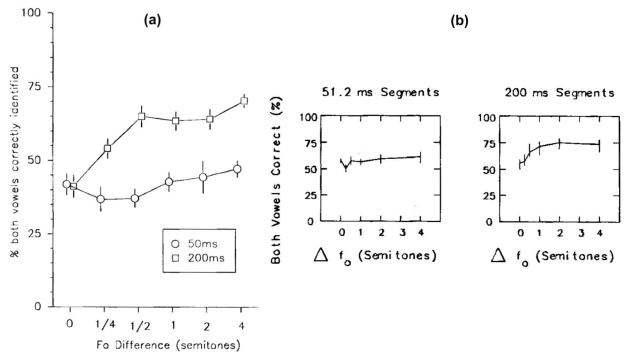

FIGURE 2.6: Taux de reconnaissance de double-voyelle en fonction de la différence des F_0. (a) Source : [Culling and Darwin, 1993]. (b) Source : [Assmann and Summerfield, 1990].

Le graphique (a) de la figure 2.6 présente les résultats obtenus par Culling et Darwin. Les deux graphiques de droite nommés (b) sont les résultats obtenus par Assmann et Summerfield. Les deux expériences testent deux durées de stimuli à 50ms environ et 200ms. Les auteurs calculent le taux de reconnaissance de doubles voyelles c'est-à-dire le nombre de couples de voyelles reconnus rapporté au nombre total de couples de voyelles entendus. Les courbes obtenues ont les mêmes comportements dans les deux expériences. Pour les voyelles stables de courte durée (≈50ms) une différence de F_0 entre les deux voyelles apportent une faible amélioration du taux de reconnaissance. Pour les voyelles stables de longue durée (200ms), il y a amélioration du taux de reconnaissance jusqu'à atteindre une différence de F_0 de un demi-ton (≈6%). Au-delà, les courbes ont un comportement asymptotique et le gain de performance est d'environ 20% en passant de 50% à 70%. La F_0 semble effectivement être un indice acoustique utilisé pour faire du groupement simultané. Il y a un autre point intéressant à étudier. Ce point concerne le taux de performance lorsque les deux voyelles mélangées ont la même fréquence fondamentale. Dans ce cas les auditeurs se montrent quand mêmes capables de reconnaître les deux voyelles à environ 50% de performance. Une différence de

F_0 aide donc à la séparation mais n'est pas la seule information utilisée pour reconnaître les deux voyelles. L'article de Darwin [Darwin, 1981] présente l'impact de la F_0 et des instants d'apparitions (*onsets*) ou de disparitions (*offsets*) sur le groupement perceptif (séquentiel ou simultané) de parole superposée. Cet article est un exemple de travaux qui présente l'effet de la coopération de plusieurs indices acoustiques sur nos facultés de séparation de parole.

Il peut être objecté que la trop grande spécificité du paradigme des double-voyelles ne permet pas de dire que les résultats trouvés sont généralisables à la parole naturelle. Néanmoins, les voyelles occupent temporellement parlant une très large majorité de la parole (en excluant les silences). Elles constituent sans doute un vecteur de transmission d'information important dans la parole naturelle [Ladefoged and Broadbent, 1957]. Cette constatation laisse à penser que la F_0 est également utilisée pour séparer de la parole superposée. Les travaux de [Brokx and Nooteboom, 1982, Assmann, 1999, Bird and Darwin, 1997] montrent effectivement l'effet de la F_0 dans le groupement simultané en utilisant non plus des mélanges de voyelles mais des phrases superposées. Là encore, les auteurs sont confrontés au besoin de contrôler au mieux leurs stimuli. Une des solutions adoptée par Brokx & Nooteboom et Bird & Darwin est d'utiliser de la parole resynthétisée par LPC (*Linear Predictive Coding* en anglais). Le principe de la LPC est décrit dans [Atal and Hanauer, 1971] et consiste à considérer que la valeur d'un échantillon du signal est une combinaison linéaire de la valeur de n échantillons précédents. Un corpus de parole est d'abord enregistré avec des phrases qui ont la particularité d'être toujours voisées. Puis les phrases superposées sont générées en additionnant deux phrases entre elles. L'une des phrases conserve son contour intonatif original et l'autre phrase du mélange est resynthétisée de manière à rendre constante la différence de F_0 des deux phrases à chaque instant. Le critère de performance utilisé est un taux de reconnaissance de mot.

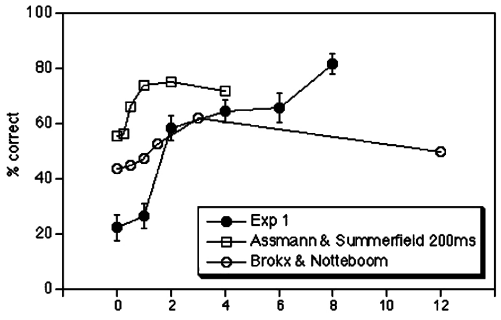

FIGURE 2.7: Taux de reconnaissance de mots dans des mélanges de deux phrases voisées. Source : [Bird and Darwin, 1997].

La figure 2.7 illustre les résultats obtenus par Bird & Darwin en comparaison avec ceux obtenus par Brokx & Nooteboom et Assmann & Summerfield. L'échelle des abscisses correspond à l'écart en demi-tons entre les deux F_0 des phrases superposées. La courbe notée Exp 1 présente les résultats de Bird & Darwin. Les taux de reconnaissance augmentent lorsque l'écart de F_0 entre les deux phrases augmente. Entre 1 et 2 demi-ton, l'amélioration du taux de reconnaissance de mot est rapide. A partir de 2 demi-tons d'écart de F_0 l'amélioration de la reconnaissance de mot se ralentit. Les trois courbes présentent une tendance asymptotique au-delà de 1 ton d'écart de F_0. Assmann approfondi ces résultats en conservant d'une part les zones silencieuses et non-voisées des signaux et d'autre part en observant l'apport du contour intonatif dans la séparation de parole. Pour cela, il synthétise une version monotone, c'est-à-dire de F_0 constante, (cf. figure 2.8 gauche) de chacune des phrases du corpus. Les résultats obtenus (cf. figure 2.8 droite) montrent qu'il est légèrement plus facile de séparer des phrases dont le contour intonatif fluctue que deux phrases de F_0 monotones. Ils montrent aussi que plus l'écart de F_0 entre les deux phrases est important, plus la séparation est facile. Dans le cas du mélange de

phrase avec intonation, l'amélioration du taux de reconnaissance de mots-clés est relativement linéaire en fonction de l'écart de F_0 en demi-ton. Le gain est de l'ordre de 10% sur 8 demi-tons. Dans le cas du mélange de parole à F_0 monotone, la relation entre le taux de reconnaissance de mots et l'écart de F_0 est moins linéaire mais la tendance est également à l'amélioration de la séparation lorsque l'écart de F_0 augmente. Le gain est de l'ordre de 20% entre le cas où les deux phrases ont même F_0 et le cas où les deux F_0 sont espacées de 8 demi-tons. Il faut noter un autre point intéressant. Comme il a été constaté pour les voyelles isolées, deux phrases synthétisées à la même F_0 restent compréhensibles à 50% du taux de reconnaissance de mots-clés ce qui implique que la F_0 est une aide à la séparation mais qu'elle n'est pas utilisée seule. L'ensemble des résultats obtenus semble indiquer que la F_0 participe de

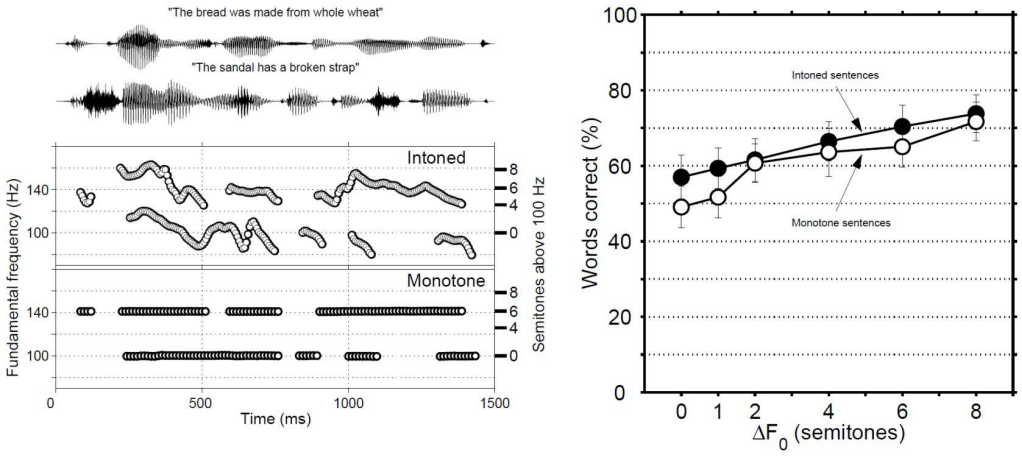

FIGURE 2.8: Résultats de séparation obtenus par Assmann. Source : [Assmann, 1999]

manière forte à la séparation de double-voyelle et même de parole. Estimer les F_0 multiples est donc utile pour pouvoir séparer de la parole.

2.2.3 Glimpsing

Les résultats obtenus par Assmann de la figure 2.8 permettent d'introduire la notion de *glimpsing* [Miller, 1947, Festen and Plomp, 1990, Cooke, 2003]. En effet, il semble plus facile de séparer des phrases dont le contour intonatif fluctue que deux phrases de F_0 monotones. Ce phénomène est à mettre en relation avec les résultats obtenus dans les travaux de Egan et al. [Egan et al., 1954]. Ils mettent en évidence l'influence de la différence de niveau sonore entre un signal cible et un signal d'interférence dans une tâche de répétition du signal cible et ce en écoute diotique ou en écoute dichotique. Plus le signal d'interférence est fort, plus la tâche est difficile. Les travaux de Mc Keown [McKeown, 1992] et de de Cheveigné et al. [de Cheveigné et al., 1997] étudient l'effet de l'intensité des voyelles concurrentes sur la capacité de les reconnaître. Il ressort de ces expériences que l'amplitude relative des voyelles induit que l'une ou l'autre est dominante et aisément reconnaissable. La réussite dans la tâche de reconnaissance des deux voyelles tient en la reconnaissance de la voyelle non-dominante. Il apparaît alors qu'une différence de F_0 entre les voyelles peut mener à faciliter la séparation de la voyelle dominée et donc à la réussite de la tâche. Une autre expérience présentée dans [McKeown and Patterson, 1995]

étudie l'impact de la durée de présentation des deux voyelles sur la performance des auditeurs dans la reconnaissance des deux voyelles mélangées. Les auditeurs disent entendre une voyelle dominante et identifient facilement cette voyelle même pour des temps de présentation courts. Pour la voyelle perçue comme non-dominante, elle nécessite plus de temps de présentation pour être reconnue avec précision. Le F_0 apporte une facilité supplémentaire dans la reconnaissance de la voyelle non-dominante pour des temps de présentation courts. Son effet s'estompe lorsque le temps de présentation augmente.

La théorie du glimsping rassemble tous ces travaux en un cadre plus général. Notre appareil auditif en condition de mélange sonore utilise les régions temps-fréquence pour lesquelles une des sources domine toutes les autres pour établir proprement les caractéristiques de la source et ainsi aider la séparation. Notre système auditif procéderait par « aperçus » successifs, comme des coups d'œil dans le paysage sonore lorsque ce paysage sonore contient une source qui domine les autres. Ceci est rendu possible car l'environnement sonore est dynamique, que les sources s'allument, s'éteignent et occupent des régions spectro-temporelles différentes et ce de manière relativement indépendante. La F_0 jouerait un rôle dans le glimsping en fusionnant en une même région temps-fréquence les régions en relation harmonique ou au contraire en séparant les régions qui ne le sont pas ou qui proviennent d'une autre structure harmonique [Coy and Barker, 2005].

2.2.4 Comparaison malentendants/normo-entendants

La littérature dans ce domaine est abondante et il n'est pas question ici de dresser un état de l'art complet mais de donner quelques pointeurs que le lecteur intéressé pourra étudier. Les malentendants équipés ou non d'implants cochléaires sont souvent nettement plus perturbés en situation de CP que des auditeurs normo-entendants. Le phénomène naturel de masquage auditif est plus important chez les malentendants ce qui s'explique en termes d'intégration temporelle et de résolution fréquentielle dégradée cf. [Hall and Fernandes, 1983, Moore, 1985]. L'intégration temporelle renvoie à la relation entre la durée d'un stimuli et le seuil à partir duquel il est perçu. La résolution fréquentielle renvoie à la différence de fréquence nécessaire pour percevoir deux sons. Les travaux de Festen et Plomp [Festen and Plomp, 1983] et de Hygge et al. [Hygge et al., 1992] illustrent et quantifient les difficultés des malentendants à traiter diverses situations bruités ou de Cocktail Party. Ces limitations s'expliquent toujours en termes de résolution temporelle ou fréquentielle.

Les travaux de Stubbs & Summerfield [Stubbs and Summerfield, 1988, 1990] testent deux méthodes automatiques de séparation de parole sur des phrases superposées. Les algorithmes testés sont l'algorithme de séparation proposé par Parsons [Parsons, 1976] détaillé au paragraphe 3.4.2 et un algorithme de séparation basé sur le cepstre dont le principe est détaillé au chapitre 3 sous-paragraphe 3.3.2.1. Ces deux algorithmes s'appuient sur la F_0 pour séparer les signaux de parole. Les auteurs présentent à des auditeurs normo-entendants et malentendants quatre types de signaux. « *Isolated sentences* » sont les phrases prises isolément. « *Unprocessed pairs* » sont les signaux de mélange de deux phrases. « *Cepstral Filtering* » sont les signaux issus de la séparation automatique par un algorithme à base de cepstre et « *Hybrid Algorithm* » se réfère aux signaux obtenus après séparation par l'algorithme de Parsons. Les auteurs évaluent la performance de séparation des auditeurs par l'intermédiaire d'un taux de reconnaissance de mots-clés. La figure 2.9 présente les résultats obtenus. Les auteurs utilisent des

phrases monotones (a) ou des phrases avec intonation (b). L'observation des colonnes « *Unprocessed*

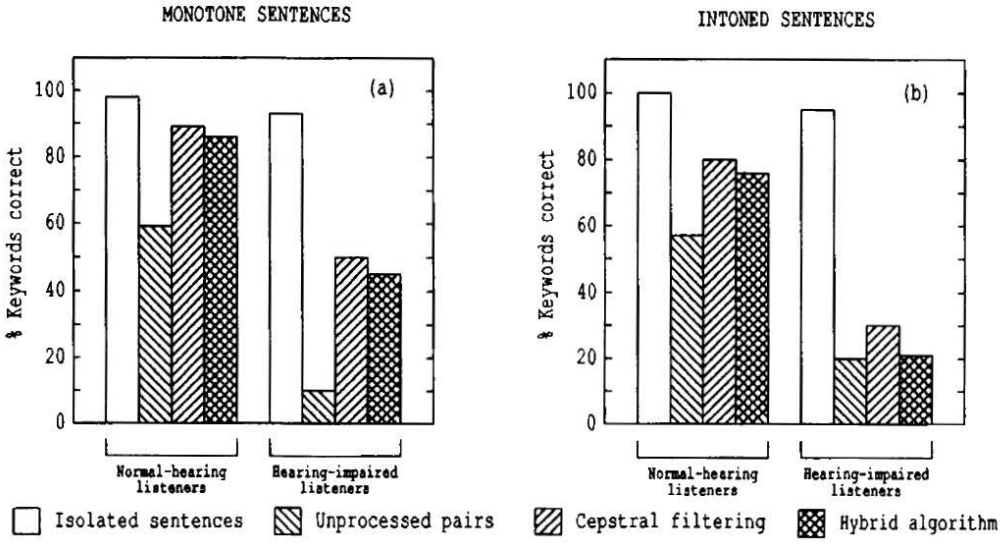

FIGURE 2.9: Performances de reconnaissance de mots après séparation automatique utilisant la F_0. Source : [Stubbs and Summerfield, 1990].

pairs » montre un résultat intéressant concernant les malentendants. Le passage de phrases mélangées de F_0 monotones à des phrases mélangées intonatives fait passer le taux de reconnaissance de mot de 10 à 20%. Ceci pourrait s'expliquer par le fait que les malentendants s'appuient sur le glimpsing pour séparer les phrases et que le glimpsing est plus important sur des mélanges de phrases intonatives. L'observation des colonnes « *Cepstral Filtering* » et « *Hybrid Algorithm* » montre que pour des mélanges de phrases à F_0 monotones (a), la séparation automatique apporte une amélioration de la performance chez l'auditeur normo-entendant comme chez le malentendant. Cela sous entend qu'il est tout à fait possible d'aider les malentendants dans des situations de mélange de parole au travers d'algorithmes automatiques. En revanche, cette amélioration n'apparait plus ou très peu chez les malentendants dans les phrases avec intonation (b). Ce résultat peut paraître contradictoire avec les résultats de Assmann en figure 2.8. Ce n'est pas le cas car Assmann fait écouter le mélange des deux phrases et montre qu'un mélange avec intonation est sensiblement plus facile à séparer. Le protocole de Stubbs & Summerfield effectue d'abord la séparation avec deux algorithmes différents avant de faire écouter le résultat de la séparation aux sujets de l'expérience. Les sujets entendent donc le résultat de la séparation et non le mélange. Les taux de reconnaissance de mots-clés sont donc dépendants de la qualité de l'algorithme de séparation. Selon les auteurs, la dégradation de performance pour les mélanges de phrase avec intonation provient non pas des malentendants mais bien des limitations des algorithmes de séparation qui fonctionnent moins bien sur des sons dans lesquels la F_0 varie. Cela justifie totalement le besoin d'algorithmes robustes d'estimation de F_0 pour la parole superposée.

Une bonne revue de l'ensemble de l'utilité de la F_0 dans le streaming est donnée dans le livre *Pitch and Auditory Grouping* [Darwin, 2005]. L'ASA a permis de mieux comprendre la faculté humaine de ségrégation de parole et notamment de mettre en évidence le rôle des groupements simultanés et séquentiels dans la construction des flux auditifs. Le groupement simultané et le groupement sé-

quentiel s'appuient sur des indices acoustiques de bas niveau parmi lesquels la F_0 revêt une grande importance. Le glimpsing, phénomène nous permettant de séparer les sources en s'appuyant sur les zones spectro-temporelles dominées par une source, utilise la F_0 comme indice de regroupement. Les expériences réalisées sur des malentendants encouragent à l'élaboration de méthodes d'estimation de F_0 de parole superposée fiables pour améliorer les systèmes automatiques de séparation de parole. L'approche ASA souffre toutefois d'une limitation importante. Pour s'assurer que les expérimentations portent bien exclusivement sur le paramètre que l'expérimentateur souhaite tester, il doit être en mesure de contrôler parfaitement les stimulus qu'il utilise. Cela implique que les signaux utilisés en ASA sont souvent synthétiques et que les protocoles sont peu réalistes (ex. répétition à cadence régulière de voyelles synthétiques). Dans quelle mesure les résultats obtenus dans ces conditions très contrôlées sont-ils transposables à des signaux réels ou des protocoles expérimentaux plus réalistes ? Ces questions amènent à regarder ce qui se fait en termes de séparation de parole sur des signaux de parole réelle. Comment ont été exploitées les découvertes de l'ASA dans la construction de systèmes automatiques de séparation de parole ?

2.3 Importance de la F_0 dans les systèmes actuels de séparation de parole

En matière de séparation de parole, il existe deux grandes familles d'approches. La première des familles est issue du domaine de recherche nommé séparation aveugle de sources ou *Blind Source Separation* (BSS) et va être décrite dans le paragraphe 2.3.1. La seconde famille d'approche est nommée Analyse de Scène Auditive Computationnelle ou *Computational Auditory Scene Analysis* (CASA). Elle va être détaillée dans le paragraphe 2.3.2. Comment les deux familles se différencient-elles ? Les approches BSS s'intéressent à la statistique du signal alors que les approches CASA s'intéressent aux caractéristiques de la source. Les approches CASA disposent de une ou deux observations de la scène sonore alors que les approches BSS ne sont pas limitées en nombre d'observations. Plus il y a d'observations, plus les approches BSS sont aisément utilisables. Les deux familles diffèrent par les hypothèses qu'elles font sur les signaux à séparer et sur la différence entre le nombre de capteurs et le nombre de signaux mélangés. Un article de van der Kouwe et al. [van der Kouwe et al., 2001] compare les performances de systèmes de séparation orientés BSS et CASA. Il conclut que les deux approches sont également intéressantes et qu'elles fonctionnent sous différentes hypothèses. Les auteurs émettent l'idée de trouver un moyen de combiner les approches pour obtenir un système de séparation encore plus performant.

2.3.1 F_0 et approches BSS

La séparation aveugle de source a pour objectif de reconstruire les signaux de chacune des sources sonores constituant une scène auditive. Le problème étant que chacune de ces sources n'est pas observable isolément mais qu'un certain nombre de micros d'enregistrement observent la scène auditive dans son ensemble. Chacun des micros capte ainsi la contribution de chacune des sources ainsi que l'ensemble des réverbérations induites par l'environnement sonore. De plus les observations faites par les micros

peuvent être bruitées par les micros eux-mêmes. La formulation la plus simple du problème de BSS qui ne considère ni la réverbération ni le bruit d'observation se pose de la manière qui suit. Soit n le nombre de sources d'une scène auditive. Soit m le nombre de micros d'observation de la scène auditive. Soit $s_j(t)$ le signal émis par la source numéro j et $x_i(t)$ le signal mesuré par le micro numéro i. Chacun des signaux observés $x_i(t)$ est la somme pondérée par a_{ij} des signaux issus de chacune des sources de la scène auditive. Les coefficients a_{ij} dépendent de l'organisation spatiale des sources par rapport à chacun des micros, des obstacles en présence, des réverbérations diverses, etc. En ne considérant ni les réverbérations ni les bruits d'observation, les signaux observés sont donnés par l'équation 2.3.

$$x_i(t) = \sum_{i=1}^{n} a_{ij} s_j(t) \tag{2.3}$$

L'hypothèse classiquement utilisée est d'avoir autant de sources à séparer que de micros d'observation de la scène auditive soit $n = m$. Cette hypothèse est forte et sera discutée un peu plus loin. L'équation 2.3 peut s'écrire sous la forme matricielle donnée en équation 2.5 en considérant les notations fournies dans les trois équations 2.4 ci-dessous.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{bmatrix}, \mathbf{s(t)} = \begin{bmatrix} s_1(t) \\ \dots \\ s_n(t) \end{bmatrix}, \mathbf{x(t)} = \begin{bmatrix} x_1(t) \\ \dots \\ x_n(t) \end{bmatrix} \tag{2.4}$$

Le problème de séparation aveugle de source résumé sous forme matricielle revient à être capable de calculer la matrice $\mathbf{A}$ qui satisfasse l'équation 2.5 à partir uniquement des signaux observés $\mathbf{x(t)}$ et éventuellement d'informations a priori sur les signaux à séparer.

$$\mathbf{x(t)} = \mathbf{A}.\mathbf{s(t)} \tag{2.5}$$

Dans cette description du problème de la BSS, la F_0 n'apparaît nulle part. Effectivement, ce type d'approche ne nécessite pas l'estimation de F_0. Les approches BSS ne sont donc pas concernées par l'estimation de F_0 et ne seront pas décrites plus précisément dans le présent travail. Voici quelques pointeurs sur des méthodes relativement récentes de séparation BSS de parole [Lambert and Bell, 1997, LeBlanc and Leon, 1998, Te-Won et al., 1998, Douglas and Sun, 2003]. Comment les auteurs résolvent-il le problème, qui paraît indéterminé ? De manière générale, les auteurs cités ci-dessus et tous ceux du domaine s'appuient sur des caractéristiques statistiques du signal de parole. L'article de Davenport [Davenport, 1952] est un des premiers travaux qui traite de la statistique du signal de parole. Les auteurs posent généralement la contrainte d'indépendance statistique des sources sonores mélangées à tout instant du temps. Jusqu'à quel niveau cette contrainte reste-t-elle valable lorsque les sons mélangés sont deux sons de parole, des sons de même nature [Cauwenberghs, 1999] ? Les méthodes BSS passent donc par un apprentissage des statistiques des signaux à traiter afin d'être les plus performantes possibles dans la séparation.

Une contrainte inhérente des approches BSS est de posséder plusieurs enregistrements de la même scène sonore. C'est une contrainte forte comme il a été dit précédemment car nous humains ne disposons

que de deux oreilles pour analyser des scènes auditives comportant bien souvent plus de deux sources sonores. Il est donc raisonnable de penser que les méthodes BSS ne sont pas adaptées pour traiter un signal de mélange unique. Il est néanmoins important de citer les travaux de Weiss & Ellis [Weiss and Ellis, 2007] qui présentent un système de séparation de parole avec un seul enregistrement. Ils y parviennent en utilisant des modèles de locuteurs. Il y a aussi les travaux de Ozerov et al. [Ozerov et al., 2005] qui présentent un système capable de séparer de la voix chantée avec un seul enregistrement. Ils calculent également des modèles de locuteur. La contrainte du nombre de micros est donc moins forte qu'auparavant et c'est pourquoi il est important de regarder ce qui se passe dans ce domaine de recherche. Les travaux de Hershey et al. [Hershey et al., 2009] parviennent également à séparer de la parole superposée dans un seul signal. Ces auteurs utilisent un corpus de parole élaboré pour le Challenge de Cooke [1]. La particularité de ce corpus est d'être construit à partir de phrases respectant une certaine structure décrite dans [Cooke et al., 2006]. Ils intègrent dans leur système de séparation en plus de modèles de locuteurs les informations connues sur la structure grammaticale et la structure des phrases. Ceci permet à leur système d'être performant. Toutefois, ce système reste-t-il aussi performant sur des phrases moins contraintes ?

Certes, les approches de séparation de parole de type BSS fonctionnent sans utiliser explicitement la F_0. Pour autant, rien ne semble montrer qu'intégrer la F_0 d'une manière ou d'une autre dans les modèles utilisés (ie. modèle de locuteur) ne pourrait pas apporter une amélioration des performances de séparation.

2.3.2 F_0 et approches CASA

La conception de séparation de parole soutenue dans cette thèse est plus proche d'une approche CASA que d'une approche BSS. Le signal de parole est produit par l'humain et par son appareil phonatoire (cf. chapitre 3 figure 3.2). Ceci confère au signal de parole des particularités qui se trouvent sous forme d'indices acoustiques dans la forme d'onde. Ces indices acoustiques, s'ils sont extractibles à partir de l'analyse du signal acoustique, renseignent sur les caractéristiques de l'émetteur donc sur l'identité du locuteur. S'il est possible d'attribuer à chaque instant un indice acoustique (comme la F_0 par exemple) à la bonne source (locuteur) alors il doit être possible de séparer de la parole. Les approches CASA s'intéressent donc aux caractéristiques physiques de l'émetteur pour séparer la parole. Le domaine CASA tente de modéliser au mieux chacun des étages de traitement de notre système auditif. Pour arriver à cela, il se donne des problèmes de séparation qui soient les plus « écologiques » possibles. Ainsi, comme nous disposons de deux oreilles, une branche du CASA est dédiée à la séparation de parole en situation binaurale c'est-à-dire en situation où le système de séparation dispose de deux signaux de la même scène auditive. Le CASA binaural est présenté dans le sous-paragraphe 2.3.2.1. Une autre branche du CASA dite monaurale étudie les moyens de séparer la parole à partir d'un signal unique de la scène auditive. C'est cette branche à laquelle le travail de thèse est rattaché puisque l'AEP proposé travaille à partir d'un seul enregistrement d'un mélange de parole. Le sous-paragraphe 2.3.2.2 détaille les travaux majeurs du CASA monaural.

1. cf. `http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm`

2.3.2.1 CASA binaural

Au sein des approches CASA, il existe des systèmes de séparation basés sur la localisation des sources sonores. La localisation est un indice acoustique important dans une scène auditive et il est certain que nous l'utilisons en situation de CP pour faciliter la séparation de parole. La situation binaurale permet effectivement d'avoir accès à l'information spatiale par l'intermédiaire des décalages temporels (ITD-*Interaural Time Difference*) et des disparités spectrales entre les sons qui parviennent à nos deux oreilles. Ces approches n'incluent pas nécessairement l'utilisation de la F_0 et sont donc en dehors du cadre de la thèse. Elles ne seront donc pas détaillées ici. Les travaux de Lyon [Lyon, 1983] et Roman et al. [Roman et al., 2003] proposent des systèmes de séparation utilisant uniquement l'information de localisation des sources. Le CASA binaural a l'avantage de pouvoir séparer d'autres mélanges que des mélanges de parole pure, des mélanges dans lesquels les sources ne comportent pas de F_0. Toutefois, la localisation n'est sans doute pas suffisante seule. Pour s'en persuader il suffit de considérer la situation dans laquelle un personne écoute une pièce musicale dans un haut-parleur unique. Cette personne parvient à séparer la scène auditive en sources alors que l'information spatiale n'est plus présente. La localisation est, comme la F_0, une aide à la séparation de parole. Tessier et Berthommier [Tessier and Berthommier, 1997] étudient l'effet cumulé de la F_0 et de la localisation pour de la séparation de doubles voyelles. Les travaux de Denbigh et Zhao [Denbigh and Zhao, 1992] proposent un système de séparation monaural basé sur la F_0 et un système binaural basé sur la F_0 et la localisation. Ces travaux concrétisent les observations faites en ASA dans lesquelles les auditeurs parviennent à séparer des voyelles mélangées qui ont la même F_0. Les auditeurs utilisent alors d'autres indices acoustiques. Un système de séparation CASA semble augmenter son efficacité lorsqu'il combine plusieurs indices acoustiques.

2.3.2.2 CASA monaural

Au sein des approches monaurales, l'ensemble des méthodes existantes peuvent être classées en deux familles. La première famille s'appuie sur des modèles auditifs ou perceptifs et tente de modéliser au mieux ce qui existe biologiquement dans l'appareil auditif humain. La seconde famille concerne les méthodes inspirées des techniques de traitement du signal. Ces méthodes n'essaient pas nécessairement de « copier » la biologie mais plutôt d'en reproduire les fonctionnalités. Ce travail de thèse est à classer dans la seconde famille.

Méthodes perceptives et glimpsing Le CASA utilise des modèles de l'oreille humaine. Le détail d'un de ces modèles est donné au chapitre 3 paragraphe 3.3.3. L'oreille humaine découpe l'onde sonore en banc de filtres et ceci est repris dans un bon nombre de systèmes de séparation de parole. Ce découpage en banc de filtres permet de représenter un signal de parole sous la forme d'une image dont les pixels sont des points temps-fréquences (T-F). La valeur du pixel pour un point (T-F) donné représente l'amplitude du signal de parole à cet instant donné et pour une fréquence donnée. Patterson et al. [Patterson et al., 1992] parlent « d'images auditives » dont deux exemples sont donnés figure 2.10. Les deux exemples présentent les images auditives obtenues par un modèle auditif sur deux voyelles isolées. Le système le plus sophistiqué de Weintraub [Weintraub, 1986] comporte trois étages de trai-

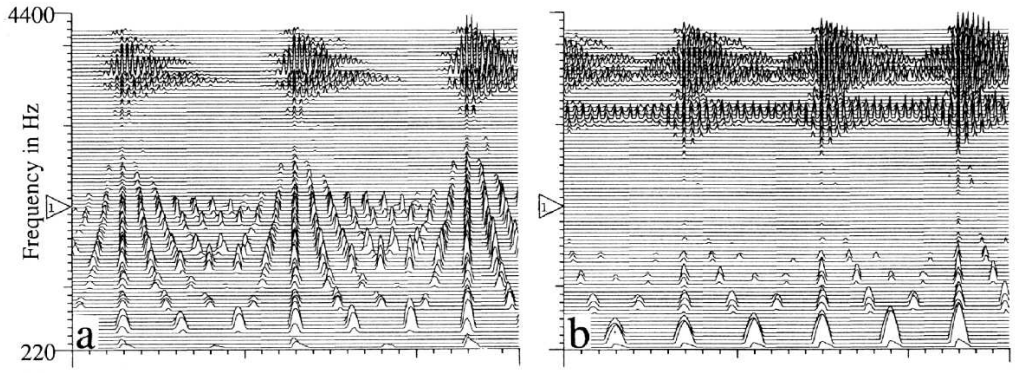

FIGURE 2.10: Images auditives de voyelles isolées /a/ et /i/ respectivement en (a) et (b). /a/ est de fréquence fondamentale à 100Hz et /i/ 125Hz. Source : [Patterson et al., 1992].

tement. Le premier étage estime les F_0 en présence dans la parole superposée en analysant les écarts temporels entre les pics des signaux obtenus en sortie d'un banc de filtre auditif. Il estime ensuite le nombre de sources par une chaîne de Markov. Son modèle de Markov peut prendre sept états différents : silence, périodique, non périodique, onset, offset, périodicité croissante et périodicité décroissante. La dernière étape consiste à estimer l'amplitude de chacune des voix sachant l'état courant de son modèle de Markov. Ce travail repose donc sur l'estimation de la F_0.

Meddis et Hewitt [Meddis and Hewitt, 1992] proposent un modèle de séparation automatique du paradigme des doubles voyelles. Leur modèle se veut proche de la biologie. Ils passent par une étape d'estimation des F_0 comme en témoigne la figure 2.11. La partie de gauche détaille le modèle de l'oreille et la partie de droite détaille comment est utilisée de l'information de F_0 pour séparer les voyelles. Ces travaux présentent l'inconvénient de ne traiter que des mélanges de deux voyelles stables. Qu'en est-il de la séparation de parole superposée, plus réaliste ? Un succès du CASA a été de concrétiser la théorie ASA du glimpsing en exploitant des images auditives à partir de masques binaires. Les images auditives sont des représentations du signal sous la forme de points temps-fréquence (T-F). Or, pour chaque point T-F d'un mélange sonore, il se peut qu'une des sources domine toutes les autres. La théorie du glimpsing précise que c'est ce genre de point T-F sur lequel notre perception s'appuie pour séparer les signaux. Un point T-F dominé par une source aura presque la même valeur que le même point T-F de la source dominante isolée[Roweis, 2003]. Pour une source donnée, il suffit donc de repérer les points T-F pour lesquels elle est dominante et construire un masque binaire correspondant à la source. La séparation de signaux dans cette optique consiste à estimer les masques binaires de chacune des sources à partir de la représentation T-F du mélange. Ce genre de séparation a l'avantage de permettre la reconstruction des signaux constituants une fois les masques binaires estimés. La question centrale de cette méthode est : comment construire les masques binaires ? C'est là qu'interviennent les indices acoustiques et notamment la F_0. En effet, les masques binaires sont construits par le regroupement de points T-F qui respectent certaines contraintes. Les contraintes utilisées sont toutes inspirées par les résultats ASA. Par exemple, lorsqu'une F_0 est détectée, l'ensemble de la structure harmonique est fusionnée en une seule zone T-F. Les onsets et offsets peuvent aussi être utilisés. Cette première étape est nommée étape de segmentation mais elle n'est pas suffisante. Il faut maintenant regrouper les zones T-F disjointes

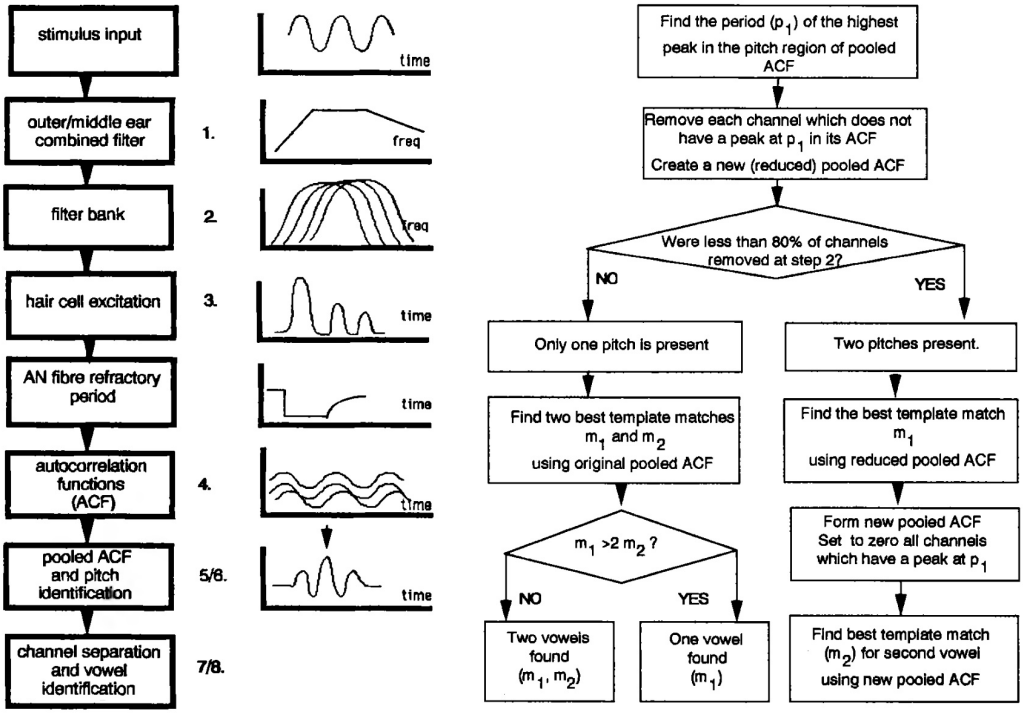

FIGURE 2.11: Modèle de séparation d'un signal de voyelles superposées. Source : [Meddis and Hewitt, 1992].

mais appartenant à une même source. Cette étape est l'étape de regroupement (ou *clustering* en anglais). Toutes ces étapes de traitements ne sont pas triviales et sont des problématiques en soi. Les travaux plus récents de Hu & Wang [Hu and Wang, 2004] proposent un schéma de séparation donné en figure 2.12. Leur système de séparation intègre à la fois un modèle de l'oreille et un fonctionnement par masque binaire. Il est important de constater que la F_0 joue un grand rôle dans leur système. Coy & Barker [Coy and Barker, 2005] proposent un système de reconnaissance de parole pour des

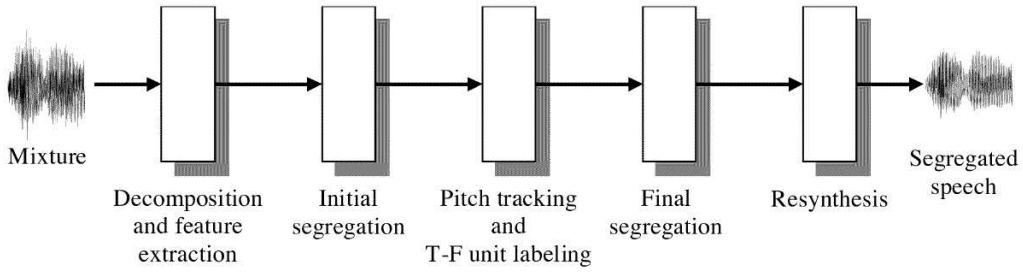

FIGURE 2.12: Modèle de séparation d'un signal de voyelles superposées. Source : [Hu and Wang, 2004].

signaux de parole superposée. Ils fabriquent ce qu'ils appellent des fragments cohérents. Ces fragments correspondent à des zones spectro-temporelle de l'image auditive qui est un spectrogramme. Ils utilisent aussi la F_0 pour construire ces fragments cohérents en regroupant les points T-F appartenant à la structure harmonique qui possède la même F_0. Pour reconstruire les trous dans le masque binaire, ils utilisent une technique dite de *Missing Data Technique* [Cooke et al., 1997] permettant de reconstruire

les régions spectro-temporelles manquantes entre les fragments. Un raffinement a été apporté à la notion de masque binaire en considérant cette fois-ci un masque flou. Le lecteur intéressé pourra consulter [Radfar et al., 2007, Reddy and Raj, 2004]. Un point T-F devient une combinaison linéaire des points T-F de chacune des sources ce qui permet d'adoucir la contrainte de binarité du masque.

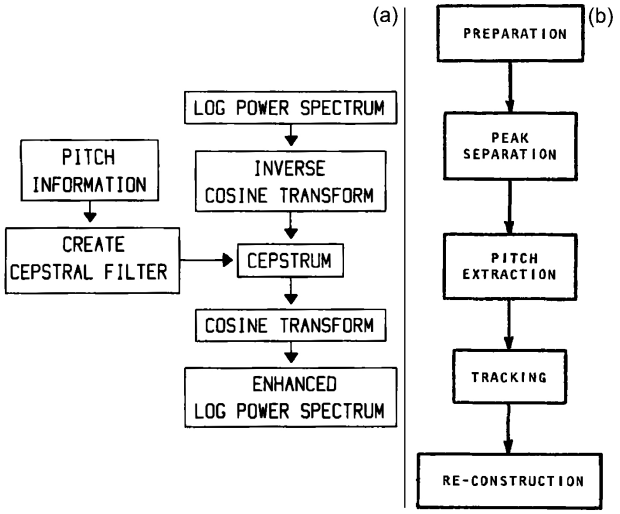

FIGURE 2.13: Schémas des systèmes de séparation évalués dans [Stubbs and Summerfield, 1990]. (a) est le synopsis de la méthode cepstrale donnée dans [Stubbs and Summerfield, 1988]. (b) est le synopsis de la méthode de Parsons [Parsons, 1976].

Méthodes par techniques classiques de traitement du signal L'outil classiquement utilisé est la transformée de Fourier discrète à court terme qui permet de représenter n'importe quel signal temporel sous la forme fréquentielle. La représentation (T-F) par transformée de Fourier glissante est appelée spectrogramme. Parsons [Parsons, 1976] ainsi que Stubbs & Summerfield [Stubbs and Summerfield, 1988] mettent au point des systèmes de séparation de parole basés sur l'extraction du pitch. L'algorithme de Parsons estime la F_0 à partir du spectre et Stubbs & Summerfield à partir du cepstre. Les synopsis des méthodes sont donnés figure 2.13. Hanson et Wong [Hanson and Wong, 1984] proposent un algorithme nommé EMS dont l'objectif est d'améliorer l'intelligibilité d'un locuteur lorsque celui-ci est mélangé à un autre. L'algorithme passe par une étape de suppression des harmoniques appartenant à une même structure ce qui fait intervenir la connaissance de la F_0 de la structure. Dans [Virtanen and Klapuri, 2001] les auteurs proposent une méthode de séparation de sons à structure harmonique s'appuyant sur un algorithme d'estimation multipitch.

Une bonne revue des méthodes CASA peut être trouvée dans [Wang and Brown, 2006] et une bonne revue de méthodes récentes de séparation de parole se trouve dans [Divenyi, 2004]. La F_0 semble jouer un rôle central dans les systèmes de séparation CASA monauraux. En revanche, elle n'est pas nécessairement utilisée en CASA binaural. La thèse est inscrite dans une approche de type CASA monaurale ce qui explique l'intérêt porté à l'étape d'estimation de F_0.

2.4 Conclusions

L'objectif de ce chapitre était de présenter en quoi l'estimation de F_0 peut être utile pour la séparation de parole. Il a montré que pour des systèmes de séparation de parole monauraux ont tous recours à cet indice acoustique. La séparation de parole est définie comme la capacité de suivre au cours du temps un des signaux qui composent le mélange. Ce suivi n'implique pas forcément d'être capable de reconstruire le signal mais plutôt d'être capable de dire que tel ou tel indice acoustique à tel instant

appartient à tel locuteur. La fréquence fondamentale de la parole est l'un de ces indices, peut-être le plus important. D'un point de vue pratique, s'appuyer sur la F_0 d'un signal à suivre suppose que cet indice puisse être isolé à partir du mélange, dès l'analyse bas niveau, sans recourir à des connaissances de plus haut niveau (lexicales, syntaxiques, sémantiques, etc.), qui font justement l'objet de la séparation. Ainsi le présent travail, tout en restant dans la perspective de la séparation de parole, s'assigne comme objectif la recherche d'une ou plusieurs périodicités dans un mélange de signaux de parole. Néanmoins, l'estimation de F_0 en situation monopitch et multipitch pose des problèmes encore mal résolus et ce même avec des signaux bien contrôlés. C'est pourquoi le travail d'investigation de cette thèse est orienté vers l'estimation de F_0.

Chapitre 3

État de l'art de l'estimation de F_0 isolée et multiple

3.1 Introduction

L'objectif de ce chapitre est d'analyser ce qui a déjà été fait dans le domaine de l'estimation de F_0 en situation monopitch et multipitch. Il est utile de rappeler les deux situations d'estimation de F_0 déjà définies dans le chapitre 2 :

- La « situation monopitch » ou de F_0 isolée correspond à des signaux comportant au maximum une seule fréquence fondamentale. Pour la parole, ces signaux sont de type mono-locuteur. Pour la musique, ces signaux correspondent à des instruments monophoniques jouant en solo (ex. trompette).
- La « situation multipitch » ou de F_0 multiple correspond à des signaux comportant plusieurs fréquences fondamentales. Ces signaux se retrouvent en parole superposée car la scène sonore est un mélange de signaux mono-locuteurs. En musique, ces signaux sont émis par des instruments polyphoniques (ex. piano) ou bien par un orchestre (mélange de plusieurs instruments).

La situation monopitch ou multipitch peut se référer à différentes échelles temporelles. Au niveau d'un signal complet, la définition est celle décrite par les deux points ci-dessus. Au niveau de la trame, c'est-à-dire d'un court segment de parole, la situation multipitch correspond à une situation dans laquelle plusieurs F_0 sont à estimer. La situation monopitch correspond à une trame ne comportant qu'une seule F_0 à estimer. Dans la suite de ce chapitre, nous nous placerons à l'échelle de la trame. La durée d'une trame est choisie de manière à ce que le signal de parole puisse être considéré comme stationnaire. L'ordre de grandeur d'une trame en parole est de 40ms. En musique les trames sont plus longues ($\approx$100ms).

L'estimation de F_0 est un des plus anciens problèmes de l'analyse de parole et de la musique. Il existe aujourd'hui une grande variété d'algorithmes d'estimation de pitch (AEP) conçus pour la parole ou pour la musique. Un état de l'art très complet à l'époque a été écrit par Hess [Hess, 1983]. Il a notamment permis de passer de techniques d'estimation de F_0 dans lesquelles l'harmonique de la F_0 est recherchée dans le spectre, à des techniques qui s'appuient sur la structure harmonique entière pour déterminer la F_0. En ce qui concerne la musique, la lecture des articles de Klapuri et

al. [Klapuri et al., 2000, Klapuri, 2003, 2005, 2006] et de la thèse de Emiya [Emiya, 2008] permet de se faire une bonne idée des méthodes existantes d'estimation de F_0 en musique. En ce qui concerne la parole des revues plus récentes sont disponibles dans les articles de Raghuram [Raghuram, 2002] et de Gerhard [Gerhard, 2003]. De Cheveigné propose un nouvel état de l'art intégrant l'estimation de F_0 en situation multipitch [de Cheveigne, 2006]. L'état de l'art proposé dans ce chapitre n'a pas vocation à être aussi exhaustif. Il est orienté vers les travaux qui sont les plus proches de ceux présentés dans la thèse.

Un signal de parole ou un signal musical est généralement découpé en trames sonores successives comme l'illustre la figure 3.1. L'exemple donné dans cette figure est un extrait de parole tiré du corpus de Bagshaw (fichier homme rl001) et la durée des trames est de 50 millisecondes. Un recouvrement entre les trames est utilisé pour améliorer la résolution temporelle d'estimation de F_0. Dans l'exemple, le recouvrement est de 25ms (la moitié de la durée d'une trame) ce qui veut dire que la moitié du signal est en commun à deux trames successives. Pour chacune des trames, il est possible d'estimer la ou les F_0 éventuellement présentes. Il s'agit d'une estimation trame-à-trame qui consiste à donner des hypothèses de F_0 pour chacune des trames. Cette définition permet de distinguer clairement l'estimation de F_0 du suivi de F_0 (*pitch tracking* en anglais). Le rôle de l'estimation de F_0 se limite à fournir des hypothèses F_0 pour chacune des trames. Le suivi de F_0 reprend ces hypothèses et les rassemble en une même trajectoire selon certains critères de continuité. Il fabrique ainsi les contours intonatifs pour la parole ou les partitions pour la musique. Le suivi de F_0 sert typiquement à la correction des erreurs générées par l'étape d'estimation de F_0. La section 3.2 va présenter les spécificités du signal de parole

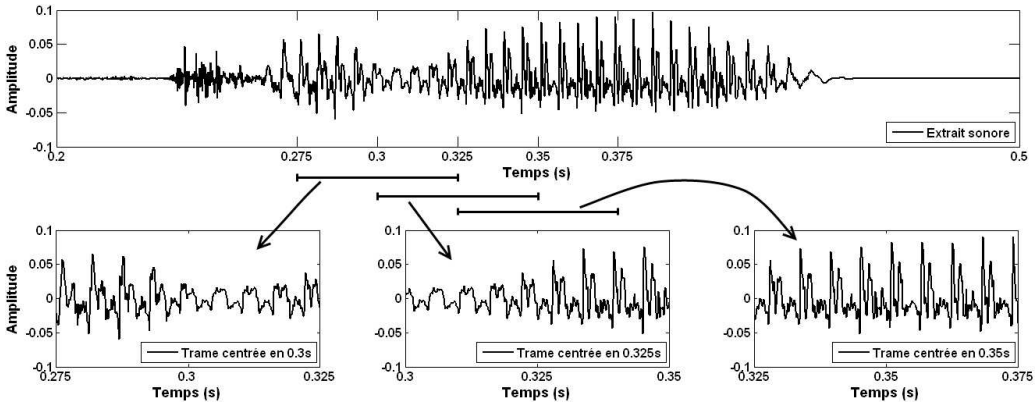

FIGURE 3.1: Illustration du découpage d'un signal en trames successives. Les segments temporels sont des trames de 50ms avec un recouvrement de 50%. Extrait du signal rl001 du corpus de Bagshaw.

et leur impact sur l'estimation de F_0. Celle-ci ayant été autant étudiée pour des signaux musicaux que pour des signaux de parole, il convient de préciser les similarités et les différences entre un signal de parole, matériau de la thèse, et un signal émis par un instrument de musique. La parole comme la musique sont émis dans un environnement particulier possédant des caractéristiques acoustiques. L'environnement affecte les signaux entre l'émission et la réception. En effet, à la réception, un signal émis est souvent entaché de bruit divers. Deux exemples de distorsions affectant l'estimation de F_0 vont être présentés. Il résulte de cette section que la conception d'un AEP est influencée par la nature et la spécificité des signaux à traiter. L'élaboration d'un AEP est également guidée par l'application à

satisfaire. L'AEP doit-il traiter des situations uniquement monopitch ? Uniquement multipitch ? Les deux ? Les signaux à traiter sont-ils exempts de bruit ou de réverbération ? Autant de questions qui conditionnent un AEP.

La section 3.3 va présenter le problème d'estimation de F_0 en situation monopitch. De nombreux AEP existent, pour la parole comme pour la musique, et toutes les approches monopitch passent par la construction d'une « fonction de pitch ». La fonction de pitch quantifie la force d'une fréquence donnée en tant qu'hypothèse F_0. En situation monopitch, le maximum (ou minimum) de cette fonction est souvent considéré comme la F_0. Les erreurs typiques faites par les AEP sont des erreurs d'octave ou de sous-octave. Les solutions adoptées pour corriger ces erreurs vont être détaillées. Elles font généralement intervenir un lissage temporel ou un suivi de F_0.

La section 3.4 va présenter l'estimation de F_0 en situation multipitch. Les signaux multipitch peuvent être des mélanges de parole ou bien des signaux polyphoniques musicaux. Dans cette situation, les AEP musicaux sont plus nombreux que les AEP de parole. C'est pourquoi il est judicieux d'observer ce qui se fait en musique pour étudier en quoi ces approches sont transposables à la parole. Il faut néanmoins garder à l'esprit que les signaux étant différents, les méthodes seront différentes. Le problème d'estimation des F_0 est plus délicat en multipitch qu'en monopitch. L'estimation de F_0 passe toujours par la fonction de F_0 mais son exploitation est plus complexe. La fonction de F_0 peut être exploitée de manière itérative ou conjointe. Dans les deux cas, le fait d'analyser un mélange sonore fait ressortir deux problèmes qui n'apparaissaient pas en situation monopitch. D'une part l'estimation de F_0 se heurte au problème des harmoniques communes. D'autre part, la connaissance du nombre de sources mélangées devient importante. Deux AEP multipitch conçus pour traiter des mélanges de parole bi-locuteurs seront finalement détaillés. Ce sont les deux AEP auxquels nous comparerons notre méthode d'estimation de F_0.

La section 3.5 conclura ce chapitre par une synthèse des points importants à retenir et en expliquant les raisons qui ont motivé notre choix d'un AEP purement fréquentiel et purement bottom-up.

3.2 Spécificité de la parole et estimation de F_0

La nature d'un signal est à rapprocher de la façon dont celui-ci est produit. La particularité du processus de production va donner au signal un certain nombre de caractéristiques. Il existe deux types de signaux qui comportent une ou plusieurs F_0 : les signaux de parole et les signaux musicaux. Le paragraphe 3.2.1 présente le signal de parole du point de vue de sa production pour expliquer quelles caractéristiques de la parole peuvent nuire à l'estimation de la F_0. Le paragraphe 3.2.2 présente les différences intrinsèques entre un signal de parole et un signal musical. Ces différences sont suffisantes pour conclure qu'estimer la F_0 sur de la parole ou sur de la musique sont deux problèmes différents. Ceci implique que les AEP conçus pour de la musique sont différents de ceux conçus pour de la parole. Le paragraphe 3.2.3 aborde l'effet de la qualité d'un signal sur l'estimation de F_0. Un signal sonore est émis dans un environnement qui va influer sur lui sous la forme de déformations diverses. Ces déformations vont influer sur la qualité du signal à traiter qui sera plus ou moins bruité. Deux exemples de déformations sont présentés : la réverbération et le filtrage sélectif (téléphone, son à travers un mur, etc.).

3.2.1 Production de la parole

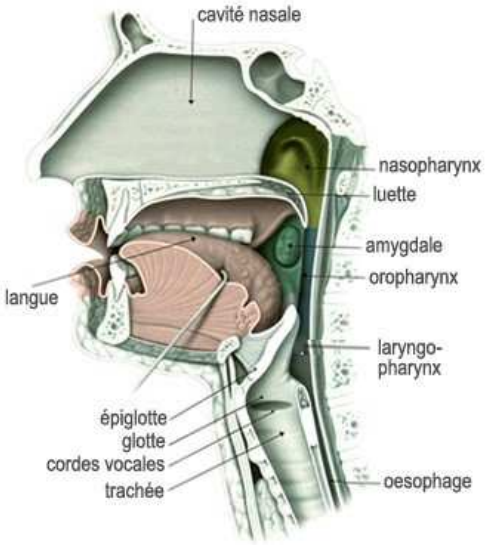

FIGURE 3.2: Anatomie de l'appareil phonatoire humain. Source : `http://lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html`

La production de la parole est assurée par un système biologique dynamique dont l'anatomie, le mouvement, l'inertie mécanique induisent un certain nombre de caractéristiques qui compliquent son traitement automatique et notamment l'estimation de la F_0. Le signal de parole est un signal issu de notre appareil phonatoire présenté figure 3.2. Pour être schématique, l'air expulsé des poumons remonte dans la trachée artère, passe au travers du larynx contenant les cordes vocales puis se disperse dans les cavités nasales et buccales avant d'être émis dans l'environnement. Lors de la production de certains segments de parole, les cordes vocales peuvent vibrer à une certaine fréquence que l'on nomme fréquence fondamentale F_0. Ces segments de parole sont nommés segments « *voisés* ». C'est cette fréquence de vibration qu'un AEP cherche à estimer. L'excursion typique de la F_0 pour une voix d'homme est comprise entre 70 et 200Hz. Pour une voix de femme, l'excursion est comprise entre 100 et 300Hz et pour une voix d'enfant, la F_0 varie entre 150 et 400Hz. Ces valeurs sont valables pour de la voix lue ou relativement neutre. Pour de la voix spontanée, les excursions peuvent être plus grandes (de l'ordre de 1kHz). Comme pour tout système biologique en mouvement, la mécanique de ces cartilages ajoute une forme de bruit à la fréquence fondamentale. Si l'on demande, même à un chanteur professionnel, de garder la note, la note ne sera pas strictement stable en F_0 mais fluctuera autour comme si un bruit de faible amplitude avait été ajouté. Ce bruit est appelé « *jitter* » et est une modulation de fréquence de faible amplitude. De la même manière, l'intensité du signal sonore est contrôlée par différents muscles permettant de moduler le débit d'air. Ces muscles ajoutent aussi à l'intensité du signal une certaine forme de bruit. Ce bruit appelé « *shimmer* » est une modulation d'amplitude de faible intensité. Le jitter et le shimmer constituent des spécificités du signal de parole et peuvent compliquer l'estimation de la F_0. La vibration produite par les cordes vocales se propage dans deux cavités : la cavité buccale et la cavité nasale. Ces cavités ont un rôle de résonateur. La structure de ces cavités est modulée par l'intermédiaire de différents muscles comme la langue ou bien le palais. L'agencement des cavités permet de contrôler les résonances. Ces résonances se retrouvent dans le signal sous forme de formants comme présenté dans la figure 3.3. Ces formants peuvent aussi parasiter l'estimation de la F_0 puisqu'ils correspondent à des zones fréquentielles qui du point de vue énergétique émergent beaucoup du voisinage. Il se peut que l'estimation de F_0 fournisse la valeur d'un formant à la place de la F_0 [McAulay and Quatieri, 1990]. Lors de la production de parole, les cordes vocales ne vibrent pas en permanence. Ces segments temporels dans lesquelles les cordes vocales ne vibrent pas sont dits « *non-voisés* », et correspondent à trois situations distinctes. D'une part, les zones silencieuses qui correspondent à des pauses dans la locution. D'autre part, des

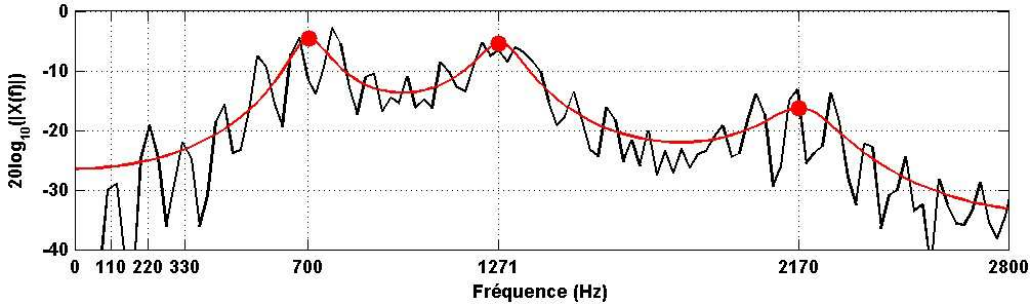

FIGURE 3.3: Spectre et formants obtenus par LPC. (trait noir) Spectre d'amplitude. (trait rouge) LPC d'ordre 10. (ronds rouge) Trois premiers formants en 700, 1271 et 2170Hz. Trame de 40ms d'une voyelle /a/ du fichier rl001 du corpus de Bagshaw. La F_0 est de 110Hz.

zones dans lesquelles certaines consonnes ne nécessitent pas la vibration des cordes vocales (ex. /f/, /s/, /ch/, /k/ en français). Enfin, les zones de respiration du locuteur. L'appareil de production de la parole nous permet d'émettre une large diversité de signaux sonores. Le signal de parole est non-stationnaire et comporte des segments voisés, non-voisés et des silences. La figure 3.4 présente un exemple de l'ensemble des non-stationnarités possibles d'un signal de parole. Le premier graphique en partant du haut est un extrait de parole (corpus de Keele, voix de femme, fichier f2nw0000). Il est séparé en trois parties nommées « Partie1 », « Partie2 » et « Partie3 » dont les zooms sont présentés dans les trois graphiques justes au-dessous. Le deuxième graphique en partant du haut correspond à la prononciation du mot anglais « *cloak* ». Le troisième graphique en partant du haut présente la zone de respiration du locuteur comprise entre les parties 1 et 3. Le quatrième graphique en partant du haut correspond à la prononciation de « *They agreed* ». Les parties 1 et 3 sont relativement similaires et

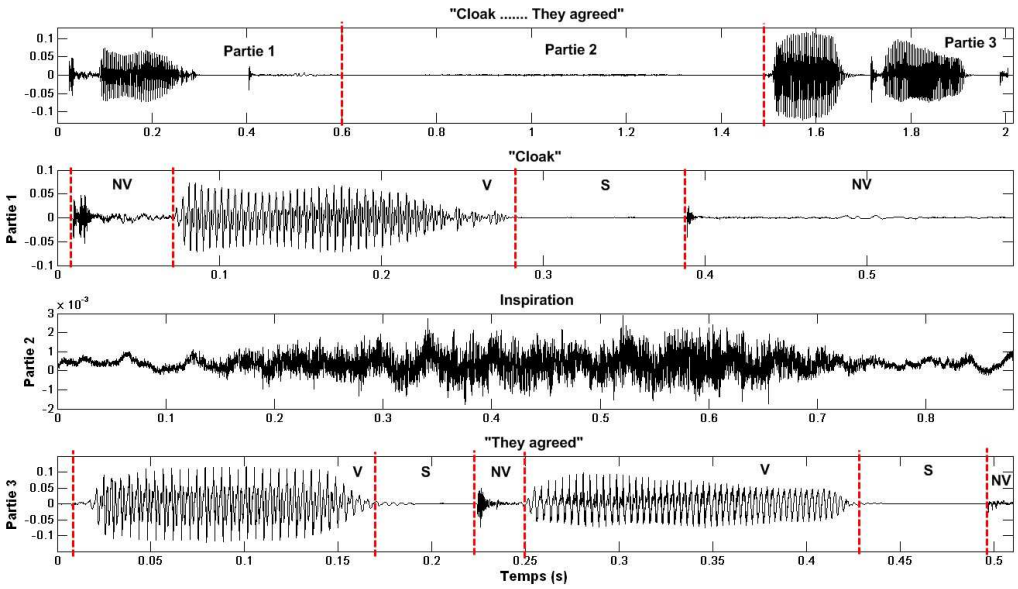

FIGURE 3.4: Exemple d'un signal de parole. Extrait de la base de Keele (fichier femme f2nw0000). Les balises (S) représentent les zones temporelles silencieuse. Les balises (V) représentent les zones temporelles voisées. Les balises (NV) représentent les zones temporelles noon-voisées.

comportent des zones silencieuses, voisées et non-voisées. Schématiquement, les zones voisées correspondent à la production de voyelles et les zones non-voisées correspondent à certaines consonnes. La partie 2 est un segment dans lequel le locuteur reprend son souffle (entre 0.2 et 0.7s dans le graphique). Cette inspiration est comprise entre la fin d'une phrase et le début d'une autre. Du point de vue des amplitudes, l'amplitude maximale du signal des parties 1 et 3 non-silencieuses se trouvent autour de 0.1 alors que l'amplitude maximale de la zone de respiration se situe autour de 0.003. Dans cet exemple la zone silencieuse est environ 30 fois moins importante. En échelle décibel (ie. $20log_{10}$ de l'amplitude) cette différence d'amplitude est d'environ 30dB. A l'écoute, cette zone est encore audible. Le fait de découper en trames de courte durée (de l'ordre de 40ms) permet de considérer le signal d'une trame comme quasi-stationnaire. Les différentes zones devront être prise en compte par les AEP car estimer la F_0 dans des segments non voisée n'a pas de sens.

Le signal de parole est riche et divers. Les AEP doivent prendre en compte l'ensemble des non-stationnarités de la parole. Ceci implique qu'ils doivent être en mesure de détecter d'une part les zones de parole ou de silence et d'autre part de détecter les zones voisées ou non au sein des zones de parole. Ceci accentue encore la difficulté de la tâche d'estimation de F_0.

3.2.2 Musique et parole : points communs et particularités

La musique est produite par un ou plusieurs instruments. La diversité des instruments de musique est très importante. Même s'il est possible de les regrouper en familles (instruments à vent, à cordes, etc.) chacun est spécifique. La parole est produite par l'appareil phonatoire humain. L'appareil phonatoire humain peut être vu comme un instrument de musique particulier. A ce titre, il y a des analogies possibles entre le signal d'un instrument de musique et le signal de parole. Il y a également des spécificités qu'il faut intégrer pour traiter de manière adéquate la parole. Les analogies sont dues à la production des deux sortes de signaux :

- Il y une source de vibration constituée par les cordes vocales ou un rétrécissement du conduit pour la parole. La source de vibration en musique dépend de l'instrument. Par exemple, pour un violon il s'agit du frottement de l'archet sur une des cordes. Pour une trompette, il s'agit de la vibration des lèvres du trompettiste.
- La vibration se propage dans des cavités. Pour la parole, il s'agit de la bouche et du nez. Pour l'instrument, il s'agit du corps de l'instrument. Par exemple, pour le violon il s'agit de la caisse de résonance et pour la trompette, il s'agit de l'ensemble colonne d'air et pavillon.

La parole comme le signal d'un instrument de musique sont des signaux qui peuvent comporter une F_0. Les différences fortes entre le signal de parole et le signal d'un instrument sont :

- La parole n'est pas constituée uniquement de segments temporels voisés mais contient aussi des segments temporels pendant lesquels les cordes vocales ne vibrent pas. Le signal émis par un instrument (hormis percussions) est toujours à peu près harmonique (avec éventuellement un certain critère d'inharmonicité ex. piano).
- Lorsqu'un instrument émet une note, la F_0 de cette note est relativement stable pendant toute la durée de la note (excepté certains cas de vibrato ou d'ornements). Cette stabilité n'existe pratiquement pas en parole courante. La F_0 de la parole évolue dans le temps et contribue à la

structure prosodique.

- La convention musicale occidentale catégorise les fréquences de vibration des instruments en notes. Cette échelle est discrète. Il y a donc une discontinuité fréquentielle des F_0 de deux notes différentes successives. Au contraire, la F_0 de la parole (hormis pour la voix chantée) n'est pas catégorisée en notes. Cette grandeur est continue.

La différence de natures des deux signaux est suffisante pour produire des systèmes automatiques capables de les discriminer. L'indexation de documents audio propose un certain nombre de critères quantitatifs permettant de différencier un signal de musique d'un signal de parole. Par exemple, les travaux de Pinquier et al. [Pinquier et al., 2002] utilisent la modulation d'énergie, la modulation de l'entropie et la durée des segments stationnaires pour établir des différences entre la musique et la parole. Les travaux de Wang et al. [Wang et al., 2003] utilisent par exemple le ratio de l'énergie contenue dans les basses fréquences par rapport à celle contenue dans les hautes fréquences. Il y a aussi des différences dans le contexte d'utilisation des signaux de parole et des signaux musicaux. A l'exception des instruments solo, un signal de musique comporte traditionnellement plusieurs instruments mélangés. En condition naturelle, la polyphonie (situation multipitch) est généralement importante. Par exemple le piano à lui seul peut émettre plusieurs notes simultanément. Le signal de parole en condition de parole superposée est nettement moins polyphonique. On considère qu'un mélange de deux à trois signaux de parole reste intelligible. Au-delà, le signal devient une sorte de brouhaha inintelligible (*babbling* en anglais). Les problèmes d'un AEP en musique sont de l'ordre du nombre de F_0 simultanées qui peut être très important, autour de 6 et plus. Les problèmes d'un AEP en parole sont de l'ordre de la dynamique des F_0 de la parole qui évoluent sans cesse. Il n'y a pas un problème plus complexe que l'autre mais deux problèmes bien distincts qui dépendent de la nature de la source sonore. Il faut retenir de ces considérations qu'un algorithme d'estimation de F_0 doit être adapté à la nature des sources auxquelles il va être confronté.

3.2.3 Qualité du signal de parole

La qualité des signaux utilisés joue un rôle important dans l'efficacité des AEP. Cette qualité peut être dégradée par l'environnement ou par la présence de bruits qui se mélangent au signal. Il convient de citer deux exemples de déformations dues à l'environnement qui nuisent à la qualité d'un AEP : la réverbération et le filtrage sélectif. Le son de parole est une onde qui subit des réflexions multiples sur les obstacles de l'environnement. La réverbération est le résultat de l'ensemble des rebonds du son de parole et par conséquent une somme du signal original avec des versions décalées et atténuées de celui-ci. Ainsi, les formes d'ondes d'un signal de parole enregistré directement à la sortie de la bouche diffère fortement de celle du même signal enregistré trois mètres plus loin. La réverbération de l'environnement modifie le signal et suscite toujours des efforts de recherche dans le domaine du speech enhancement. Ce sujet ne sera pas plus abordé ici car il n'est pas central dans la thèse. L'article récent de Wu & Wang [Wu and Wang, 2006] renseignera le lecteur intéressé par ce point. Il faut néanmoins être conscient de l'impact négatif de la réverbération quant à l'estimation de la F_0. La figure 3.5 illustre son effet perturbateur. Le signal utilisé est le signal homme rl001 du corpus de Bagshaw. La réverbération est obtenue artificiellement avec trois filtres passe-tout en série avec des retards de

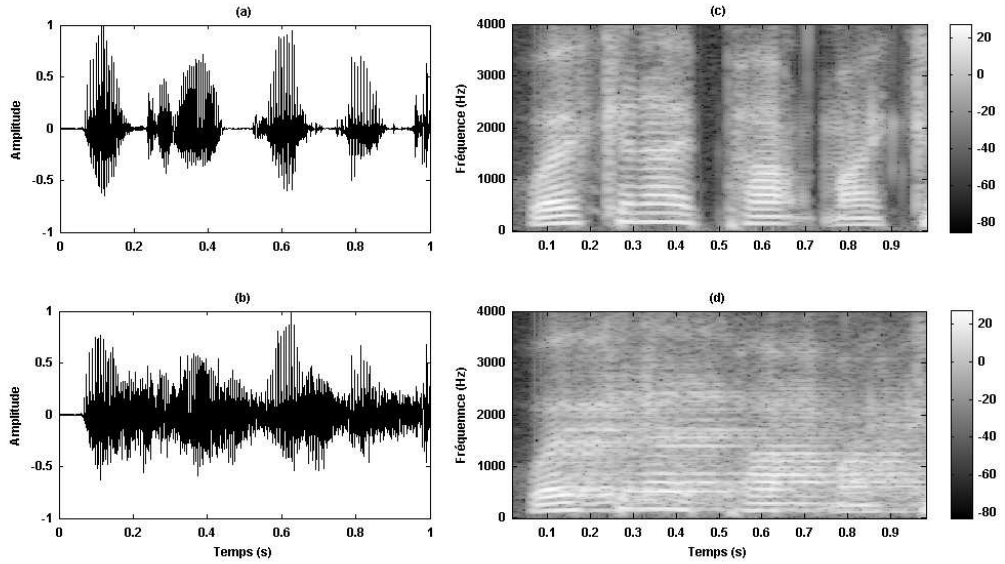

FIGURE 3.5: Effets de la réverbération. (a) Signal original. (b) Signal avec réverbération. (c) Spectrogramme du signal original. (d) Spectrogramme du signal avec réverbération.

respectivement 50ms, 25ms et 12.5ms et les gains de chacun des filtres passe-tout valant respectivement 0.5, 0.3 et 0.1. La réverbération est volontairement importante pour bien constater de son effet néfaste. La courbe (a) présente le signal de parole rl001 qui est enregistré avec un micro près de la bouche et le graphique (c) présente le spectrogramme équivalent. Les structures harmoniques des zones voisées apparaissent clairement dans le spectrogramme. Les graphiques (b) et (d) illustrent l'effet parasitaire de la réverbération. La forme temporelle (c) autant que le spectrogramme (d) sont plus bruités. Les structures harmoniques sont nettement moins claires ce qui rend plus difficile l'estimation de la F_0. Une étape de déréverbération est souhaitable si la parole est produite dans un milieu réverbérant.

Le filtrage sélectif est une autre type de déformation. Le filtrage téléphonique est le filtrage sélectif typique. Autrefois, un signal de parole téléphonique ne contenait que les fréquences comprises entre 300Hz et 3.4khz. Pour estimer la F_0 cela est gênant puisque la F_0 typique d'un homme varie entre 80 et 200Hz et que par conséquent l'harmonique correspondant à la F_0 n'est pas présente physiquement dans le signal. La qualité d'un signal de parole est aussi affectée par les bruits divers qui se mélange au signal. Ainsi, il est rare que le signal de parole soit émis tout seul dans l'environnement. En condition naturelle, un signal sonore est entouré d'une multitude d'autres sons provenant de diverses sources sonores. Ces sons complexifient l'estimation de F_0. Selon la nature du signal interférant, le traitement automatique est plus ou moins complexe. Le cas de la parole superposée est sans doute l'un des plus compliqués puisque le bruit est de même nature que le signal.

Il faut retenir de ce paragraphe que la qualité du signal est importante également dans l'estimation de F_0 et qu'il faut adapter l'AEP à la qualité du signal de parole utilisée. Utiliser un AEP sur des signaux pour lesquels il n'est pas conçu n'a pas de sens. La section 3.3 qui suit dresse un bilan des approches existantes en matière d'estimation de F_0 dans la situation monopitch.

3.3 Le problème monopitch d'estimation de F_0

La situation monopitch se retrouve majoritairement en parole et non en musique (hormis pour le cas d'un instrument seul et ne produisant qu'une note à la fois). Le cas mono-locuteur est historiquement le premier problème auquel se sont confrontés les AEP [Hess, 1983]. Il s'agit d'estimer la F_0 dans un signal comportant au plus une F_0 à un instant donné. Tous les AEP monopitch construisent une fonction de pitch c'est-à-dire une fonction qui quantifie la force d'une fréquence en tant que candidat F_0. Selon les méthodes, plus la fréquence étudiée a une force élevée, plus cette fréquence est un candidat sérieux en tant que F_0 ou bien au contraire, plus la force d'une fréquence est faible, plus la fréquence a des chances d'être la F_0. Dans le premier cas, la fonction de pitch quantifie une « force de périodicité » et dans le second cas la fonction de pitch quantifie une « force d'apériodicité ». Cette force peut être une grandeur sans unité, une probabilité, une intensité, etc. Estimer la F_0 à partir de la fonction de pitch revient à rechercher la fréquence pour laquelle la force est maximale si la fonction de pitch révèle la force de la périodicité ou minimale si la fonction de pitch révèle l'apériodicité. L'équation 3.1 formalise mathématiquement l'estimation de la F_0 dans le cas où la fonction de pitch F quantifie la force de périodicité. Dans le cas où F est indicateur d'apériodicité, le argmax de l'équation 3.1 devient un argmin.

$$F_0 = \underset{f \in [F_{0min}, F_{0max}]}{\arg\max} \Phi(f) \tag{3.1}$$

F_0 est la fréquence fondamentale estimée, F_{0min} est la fréquence minimale de recherche et F_{0max} la fréquence maximale de recherche. L'intervalle de fréquences $[F_{0min}, F_{0max}]$ est l'intervalle dans lequel l'AEP recherche la F_0. Cet intervalle doit être adapté en fonction de l'excursion possible de la F_0. Un intervalle de $[50, 800Hz]$ permet de couvrir largement l'excursion de la parole lue, quelque soit le locuteur.

Il existe trois classes principales de méthodes d'estimation de F_0 en situation monopitch qui émergent de l'ensemble des travaux : les approches temporelles, fréquentielles, par modèle cochléaire. Ces familles de méthodes peuvent ensuite être utilisées de manière déterministe ou probabiliste. Cette section présente quelques fonctions de pitch clés de chacune des approches ce qui permettra de vérifier que l'estimation faite par détection du maximum (ou minimum) de la fonction de pitch est possible. Le paragraphe 3.3.1 décrit l'approche temporelle, le paragraphe 3.3.2 décrit l'approche fréquentielle, le paragraphe 3.3.3 décrit l'approche par modèle cochléaire et le paragraphe 3.3.4 décrit le point de vue probabiliste de ces approches.

3.3.1 Approches temporelles

Estimer la F_0 dans le domaine temporel revient à trouver dans le signal un motif qui se répète plus ou moins à l'identique de façon périodique. Le motif périodique est clairement visible dans la figure 3.6. Le signal présenté est un signal artificiel crée pour l'exemple qui sera utilisé tout le long du chapitre. Ce signal est inspiré de la modélisation sinusoïdale de sons périodiques [McAulay and Quatieri, 1986, Serra, 1997], modèle très répandu dans les méthodes probabilistes d'estimation de F_0 (cf. paragraphe 3.3.4 pour plus de détails). Il est constitué de deux trames successives de 40ms et espacées de 10ms. La durée totale du signal notée d est donc de 50ms. La fréquence d'échantillonnage notée F_e est de 8kHz.

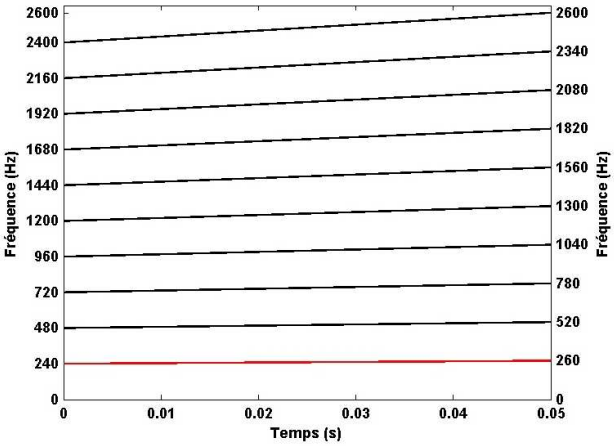

FIGURE 3.6: Schéma de l'évolution spectro-temporelle des harmoniques du signal exemple. La ligne rouge représente la F_0. Les lignes noires sont les harmoniquues. L'ensemble des harmoniques (F_0 comprise) forme la structure harmonique du signal exemple.

La F_0 du signal exemple dépend du temps et augmente linéairement tout au long des 50ms. Elle commence à 240Hz et termine à 260Hz comme le schématise la figure 3.6 (ligne rouge). L'écart entre la valeur de la F_0 en 0 et celle à 50ms est notée ΔF_0 et vaut 20Hz. Le signal exemple modélise un son voisé. Un son voisé comporte des harmoniques (multiples de la F_0) qui forment une structure harmonique. Dans l'exemple, la structure harmonique est constituée de 10 harmoniques avec F_0 comprise (cf. figure 3.6 lignes noires et rouge). Le nombre d'harmonique est noté N_h. Les harmoniques sont les multiples de la F_0 et sont modélisées par des sinusoïdes à phase aléatoire et à amplitude décroissante en fonction du numéro d'harmonique. La phase de l'harmonique numéro h notée ϕ_h est choisie aléatoirement entre $-\pi$ et π. L'amplitude de l'harmonique numéro h notée α_h décroit en racine inverse du numéro de l'harmonique $1/\sqrt{h}$. Le signal exemple $s(t)$ est la somme des 10 harmoniques et sa formule est décrite par l'équation 3.2.

$$s(t) = \sum_{h=1}^{N_h} \alpha_h \sin 2\pi h[F_0 + \frac{\Delta F_0}{2d}t] + \phi_h \tag{3.2}$$

La figure 3.7 présente la forme d'onde d'une réalisation du signal exemple. Une observation fine des amplitudes de la forme d'onde du graphique de la figure 3.7 peut laisser penser que le signal n'est pas parfaitement périodique alors qu'il devrait l'être. Ceci est dû à la résolution temporelle limitée par la fréquence d'échantillonnage de 8kHz qui implique que les instants des maxima ne sont pas nécessairement alignés avec les instants des échantillons. Ceci peut donner l'illusion d'une non-périodicité mais en réalité le signal est bien périodique.

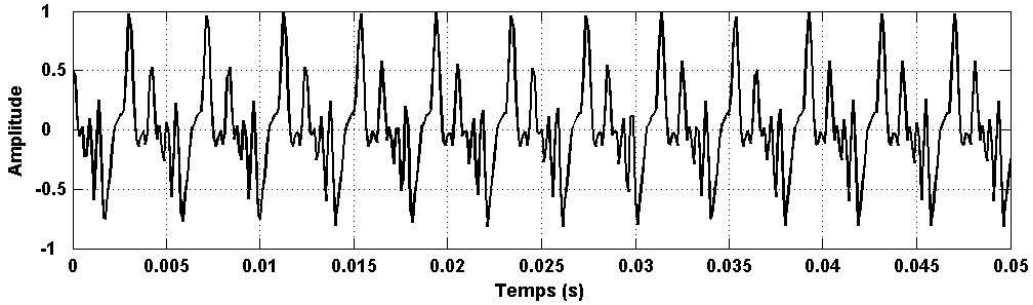

FIGURE 3.7: Forme d'onde du signal exemple avec une F_0 variant linéairement de 240Hz et 260Hz.

Pour repérer les motifs qui se répètent, il faudrait être en mesure de comparer le signal avec une

version décalée de lui-même. En mesurant la ressemblance du signal d'origine avec sa version décalée, il serait possible de quantifier de quelle manière et de quelle force il y a répétition d'un motif ou non.

3.3.1.1 Autocorrélation (ACF)

Ce principe est l'essence même de l'autocorrélation (ACF) dont l'origine se trouve dans le travail de Licklider [Licklider, 1951]. L'ACF est la méthode temporelle la plus répandue. Pour décrire le principe, soit le signal discret $x(n)$ définit pour tout n entier. L'autocorrélation de $x(n)$ est donnée par l'équation 3.3. L'ACF correspond à la corrélation d'un signal avec une version décalée de celui-ci. Le décalage qui donne la plus grande coïncidence représente la périodicité du signal.

$$ACF(\tau) = \sum_{n=0}^{N-1-\tau} x(n)x(n+\tau) \tag{3.3}$$

L'ACF donnée ci-dessus est une fonction « brute ». Elle peut être normalisée par le carré de la norme du vecteur $x(n)$ considéré dans l'intervalle $[0, N-1-\tau]$ selon ce qui est présenté équation 3.4. L'autocorrélation normalisée est notée ACF_n. Cette définition assure que l'ACF_n vaut 1 lorsque τ vaut 0.

$$ACF_n(\tau) = \frac{ACF(\tau)}{\sum_{n=0}^{N-1-\tau} x^2(n)} \tag{3.4}$$

La figure 3.8 présente les ACF_n calculées sur les deux trames du signal exemple décrit par l'équation 3.4. La première trame est centrée en 20ms et la seconde est centrée en 30ms. Les deux trames sont multipliées par une fenêtre de Hann avant de calculer l'ACF normalisée. Le fenêtrage est utilisé pour éviter les discontinuités aux extrémités des trames. Le graphique (a) représente l'autocorrélation normalisée de la première trame et le graphique (b) montre l'ACF normalisée de la seconde trame. L'échelle des abscisses représente le décalage en échantillons (ou *lag* en anglais) entre la trame et la version décalée d'elle même. Par la suite, le terme de « *lag* » sera utilisé pour parler de ce décalage. A l'intérieur de chacune des trames, la F_0 a évolué. Il faut pourtant associer la trame à une seule valeur de F_0. L'ACF permet d'obtenir la valeur moyenne de l'évolution de la F_0 au sein d'une même trame. L'estimation de F_0 en exploitant l'ACF passe par l'analyse des séries de maxima locaux régulièrement espacés sur l'échelle des lags dans les graphiques (a) et (b) de la figure 3.9. Ces maxima locaux correspondent à des coïncidences maximales entre la trame originale et une version décalée d'elle-même du lag correspondant au maximum local. L'ACF quantifie donc la force de périodicité d'une trame. Si le signal est périodique, il est normal de retrouver des maxima locaux régulièrement espacés en lag. En effet, si un signal périodique coïncide parfaitement avec lui-même lorsque le décalage est d'une période, il coïncide aussi avec lui-même lorsque le décalage est un multiple de la période fondamentale. Les méthodes d'estimation de la F_0 par ACF utilisent cette caractéristique et considèrent comme F_0 de la trame la fréquence correspondant au lag du maximum local de plus forte amplitude. L'estimation de la F_0 sera donc parasitée par les sous-multiples de la F_0 qui correspondent aux lags multiples du lag de la F_0. Par définition, l'ACF normalisée au lag 0 vaut 1. Le maximum au lag 0 n'est pas comptabilisé comme un maximum local pour estimer la F_0. Pour le graphique (a), le maximum local de plus forte amplitude (0.89) est au lag 32. Ce lag correspond à une durée de 4ms (32/8000) ce qui correspond à

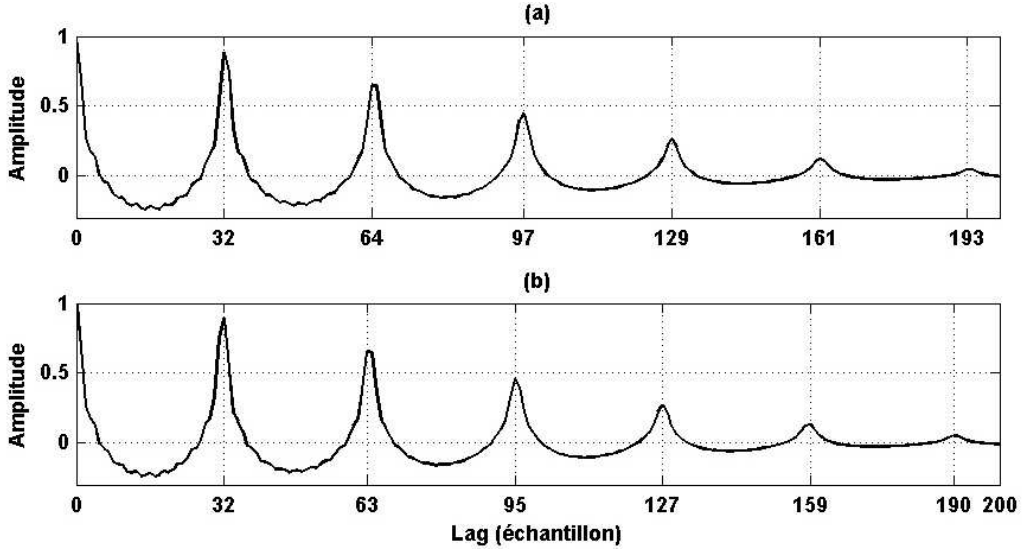

FIGURE 3.8: Autocorrélation normalisée de deux trames du signal exemple. (a) Trame centrée en 20ms. (b) Trame centrée en 30ms.

une F_0 de 250Hz (8000/32). Pour le graphique (b), le maximum local de plus forte amplitude (0.89) est au lag 32 ce qui correspond à une F_0 de 250Hz (8000/32). La précision de l'estimation est limitée par la fréquence d'échantillonnage du signal. C'est la fréquence d'échantillonnage qui impose la cadence du lag. Il se peut donc qu'un maximum local soit entre deux lag mais que la résolution temporelle ne permette pas de voir correctement ce maximum. Une interpolation peut être utilisée pour affiner l'estimation de la fréquence.

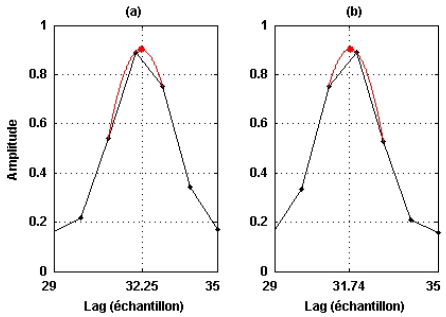

FIGURE 3.9: Interpolation quadratique. (a) Zoom autour du maximum local au lag 32 de l'ACF_n de la trame 1 (trait noir) et interpolation quadratique (trait rouge). (b) Zoom autour du maximum local au lag 32 de l'ACF_n de la trame 2 (trait noir) et interpolation quadratique (trait rouge).

En utilisant une interpolation quadratique à partir du maximum local et de ses deux échantillons voisins, la F_0 de la première trame devient 248.3Hz au lieu de 250Hz ce qui est plus proche de 248Hz la valeur théorique. De même la F_0 de la seconde trame devient 251.8Hz au lieu de 250Hz ce qui est plus proche du 252Hz théorique. Les graphiques (a) et (b) de la figure 3.9 illustrent l'amélioration de la précision obtenue avec cette simple interpolation. Les courbes en trait noir correspondent à un zoom des maxima locaux aux lags 32 des ACFn de chacune des trames du signal exemple. Le graphique (a) montre le maximum local du lag 32 de la trame 1 et le graphique (b) montre le maximum local du lag 32 de la trame 2. La F_0 de la trame 1 est de 248Hz ce qui correspond à un lag théorique d'environ 32.25 échantillons. La F_0 de la trame 2 est de 252Hz soit 31.74 échantillons de lag théorique. Les courbes en trait rouge sont les interpolations paraboliques et le point rouge est le sommet de la parabole. Dans les deux graphiques (a) et (b) le som-

met rouge est plus proche de la valeur théorique de lag que le sommet de la courbe noire. L'ACF est à l'origine de plusieurs méthodes d'estimation de F_0. Elle est utilisée dans la méthode SIFT de Markel [Markel, 1972] ainsi que dans l'article de Rabiner [Rabiner, 1977]. Un des AEP standard proposé par le logiciel PRAAT [1] est décrit dans l'article [Boersma, 1993]. Il s'agit de la fonction PRAAT « *Sound : To pitch (ac)...* ». Par la suite, cet AEP sera dénommé PRAAT. Sa particularité est de normaliser l'autocorrélation par celle de la fenêtre utilisée pour pondérer les trames (Hann), ce qui accroît considérablement la précision de l'estimation.

3.3.1.2 Intercorrélation (CCF)

L'intercorrélation normalisée est utilisée dans l'AEP fourni dans ESPS (exécutable getf0) qui est basé sur l'algorithme RAPT décrit dans [Secrest and Doddington, 1983, Talkin, 1995]. L'intercorrélation « brute » (CCF) est décrite dans l'équation 3.5 ci-dessous.

$$CCF(\tau) = \sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{3.5}$$

Elle diffère de l'ACF par les bornes supérieures. Généralement, l'intercorrélation est normalisée de la manière présentée dans l'équation 3.6. Il s'agit d'une normalisation par la norme des vecteurs $x(n)$ considérés dans l'intervalle $[0, N-1]$ et $[\tau, \tau+N-1]$. Cette forme d'intercorrélation est notée CCF_n et correspond à l'intercorrélation normalisée.

$$CCF_n(\tau) = \frac{CCF(\tau)}{\sqrt{\sum_{n=0}^{N-1} x^2(n) \sum_{n=\tau}^{N-1+\tau} x^2(n)}} \tag{3.6}$$

L'intercorrélation normalisée est reprise par Kasi [Kasi, 2002] dans un algorithme d'estimation de F_0 nommé YAAPT. L'inégalité de Cauchy-Schwarz permet de montrer que la fonction CCF_n est nécessairement comprise dans l'intervalle $[-1, 1]$. Comme pour l'ACF normalisée, la CCF présente des maxima locaux qui sont exploités pour estimer la F_0. Pour raffiner l'estimation, il est tout à fait possible d'interpoler plus finement les maxima locaux par la même interpolation quadratique décrite précédemment. L'intercorrélation est considérée comme plus stable que l'ACF_n notamment dans les lag élevés (τ élevés).

3.3.1.3 AMDF et YIN

La méthode utilisée dans YIN [de Cheveigné and Kawahara, 2002] repose sur une version modifiée de l'AMDF ou Average Magnitude Difference Function proposée par Ross [Ross et al., 1974] dont la formule est présentée dans l'équation 3.7. Le principe de comparer le signal avec une version temporellement décalée de lui-même est un point commun avec l'autocorrélation. Toutefois l'autocorrélation est basé sur un produit alors qu'ici c'est une différence. La différence correspond à la distance de Manhattan

1. cf. `http:www.fon.hum.uva.nl/praat`

des amplitudes de deux échantillons.

$$AMDF(\tau) = \sum_{n=0}^{N-1} |x(n) - x(n+\tau)| \tag{3.7}$$

La méthode YIN utilise aussi une différence entre les valeurs des échantillons comme l'AMDF. Toutefois, la distance de Manhattan de l'AMDF est mise au carré et devient le carré d'une distance euclidienne. Cette fonction est aussi connue sous le nom de SDF (*Squared Difference Function*) donnée par l'équation 3.8.

$$AMDF(\tau) = \sum_{n=0}^{N-1} [x(n) - x(n+\tau)]^2 \tag{3.8}$$

L'équation 3.9 est une réécriture de la SDF qui permet de montrer le lien entre la SDF et l'autocorrélation. Le signe moins indique que la SDF se comporte à l'inverse de l'ACF. Un pic dans l'ACF devient un creux dans la SDF. Pour exploiter la SDF il faut considérer les minima locaux au lieu des maxima. La méthode YIN quantifie donc la force d'apériodicité d'une trame.

$$SDF(\tau) = \sum_{n=0}^{N-1} [x^2(n) + x^2(n+\tau)] - 2ACF(\tau) \tag{3.9}$$

La figure 3.10 illustre la SDF sur les deux trames du signal exemple. Le graphique (a) présente la SDF de la trame 1 et le graphique (b) la SDF de la trame 2. Par définition, l'amplitude de la SDF au lag 0 est forcément nulle. Ceci se retrouve bien sur chacun des graphiques. L'estimation de F_0 passe par la localisation du minimum local de plus faible amplitude sans considérer celui au lag 0. L'estimation

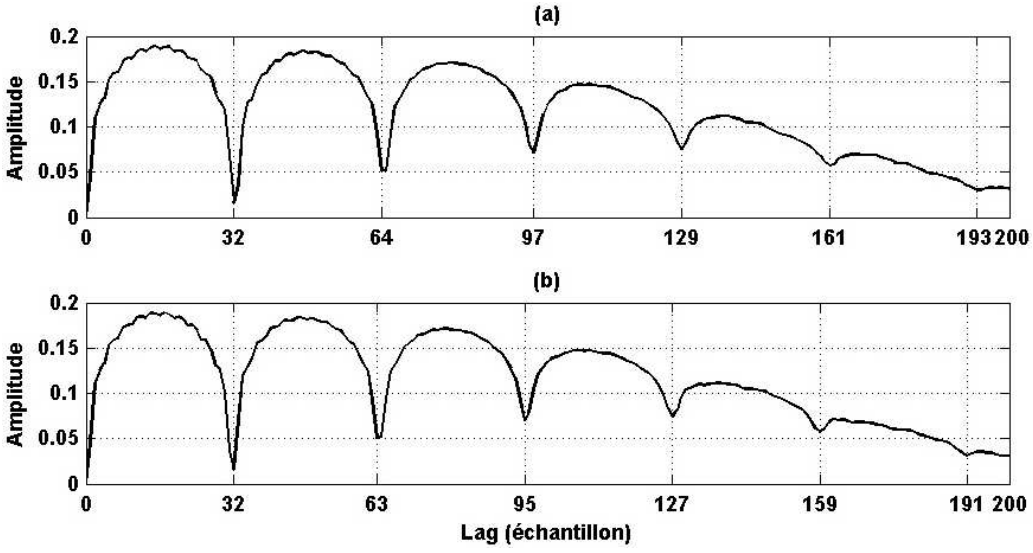

FIGURE 3.10: SDF obtenues sur les deux trames du signal exemple. (a) Trame 1. (b) Trame 2.

de la F_0 du graphique (a) fournit une F_0 de 250Hz (minimum local au lag 32). Avec interpolation quadratique, l'estimation donne 248.3Hz. L'estimation de la F_0 du graphique (b) donne une F_0 de 250Hz (251.8Hz avec interpolation quadratique). Les résultats sont identiques à ceux obtenus par

l'ACF. La F_0 mesurée par la SDF est la valeur moyenne de la F_0 sur toute la durée de la trame. Des minima locaux parasites se retrouvent très clairement dans la SDF des deux trames et sont placés aux lags multiples de 32. Cela signifie que l'estimation de la F_0 est parasitée par les sous-multiples de la F_0 comme cela a été dit pour l'ACF. Soit $m(\tau)$ la fonction définie selon l'équation 3.10.

$$m(\tau) = \sum_{n=0}^{N-1} [x^2(n) + x^2(n+\tau)] \tag{3.10}$$

Cette définition permet de dire que la SDF est une combinaison linéaire de l'ACF et de la fonction m comme l'indique l'équation 3.11.

$$SDF(\tau) = m(\tau) - 2ACF(\tau) \tag{3.11}$$

Les travaux de McLeod & Wyvill [McLeod and Wyvill, 2005] proposent une fonction de différence normalisée nommée SDF_n. Cette fonction est une manière de normaliser la SDF par le résultat de la fonction m. L'équation 3.12 présente la formule théorique de la SDF_n.

$$SDF_n(\tau) = \frac{2ACF(\tau)}{m(\tau)} \tag{3.12}$$

Il existe encore bien d'autres fonctions comme la NAF ou *Narrowed Autocorrelation Function* décrite dans [Brown and Puckette, 1989] et utilisée dans [Brown and Zhang, 1991] pour l'estimation de F_0 sur des signaux musicaux de violon, de flute et de piano. Le principe de la NAF est décrite par l'équation 3.13. Elle consiste à considérer non plus un seul décalage en τ mais plusieurs décalages multiples de τ. Ceci a pour effet de rendre les maxima locaux de la NAF plus fins et donc potentiellement plus discriminants et plus robustes au bruit. L désigne le nombre maximal de décalages de période pris en compte.

$$NAF(\tau) = \sum_{n=0}^{N-1} [x(n) + x(n+\tau) + \cdots + x(n+L\tau)]^2 \tag{3.13}$$

D'autres fonctions plus complexes sont encore utilisées dans [Xu and Principe, 2007]. Néanmoins, les principes et les comportements de ces autres fonctions sont similaires à celles présentées auparavant (ACF et SDF) et n'apportent pas de changement profond dans les résultats obtenus.

Toutes les approches temporelles présentées quantifient la similarité entre deux trames en fonction du décalage (lag) entre ces trames. La similarité est extraite soit d'un produit (ACF, CCF) soit d'une différence (SDF, AMDF). Cette façon de procéder se traduit par des courbes comportant des pics (maxima locaux) ou creux (minima locaux). La F_0 à estimer est un de ces maxima (ou minima) locaux. Dans toutes les méthodes, la F_0 est en concurrence avec les valeurs sous-multiples car les autres maxima (ou minima) peuvent être d'amplitude proche de celle du maxima (ou minima) correspondant à la F_0.

3.3.2 Approches fréquentielles

En utilisant la transformée de Fourier, il est possible de passer d'une représentation temporelle du signal à une représentation fréquentielle. Par définition de la transformée de Fourier discrète, le spectre

d'amplitude d'un signal discret quelconque $x(n)$ est donnée par l'équation 3.14.

$$|X(b)| = \left| \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{i2\pi bn}{N}\right) \right| \tag{3.14}$$

Le spectre d'amplitude de l'exemple comporte M points fréquentiels compris entre 0 et la fréquence d'échantillonnage du signal divisée par deux (théorème de *Nyquist-Shannon*). Généralement M est une puissance de 2 supérieure au nombre d'échantillons de la trame étudiée pour pouvoir d'une part utiliser l'algorithme de FFT et d'autre part améliorer la résolution fréquentielle du spectre par du zéro-padding. Le zéro-padding consiste à ajouter des échantillons à zéro à la fin d'une trame. Ceci permet d'avoir des trames artificiellement plus longues et donc une meilleure précision fréquentielle. La figure 3.11 présente trois spectres d'amplitudes. Le graphique (a) est le spectre d'amplitude de la trame 1 du signal exemple de la figure 3.11. Le graphique (b) est le spectre d'amplitude de la trame 2. Le graphique (c) est le spectre d'amplitude d'une trame construite comme le signal exemple mais dont la F_0 reste constante à 240Hz tout le long de la trame. Chacun des spectres présente des pics

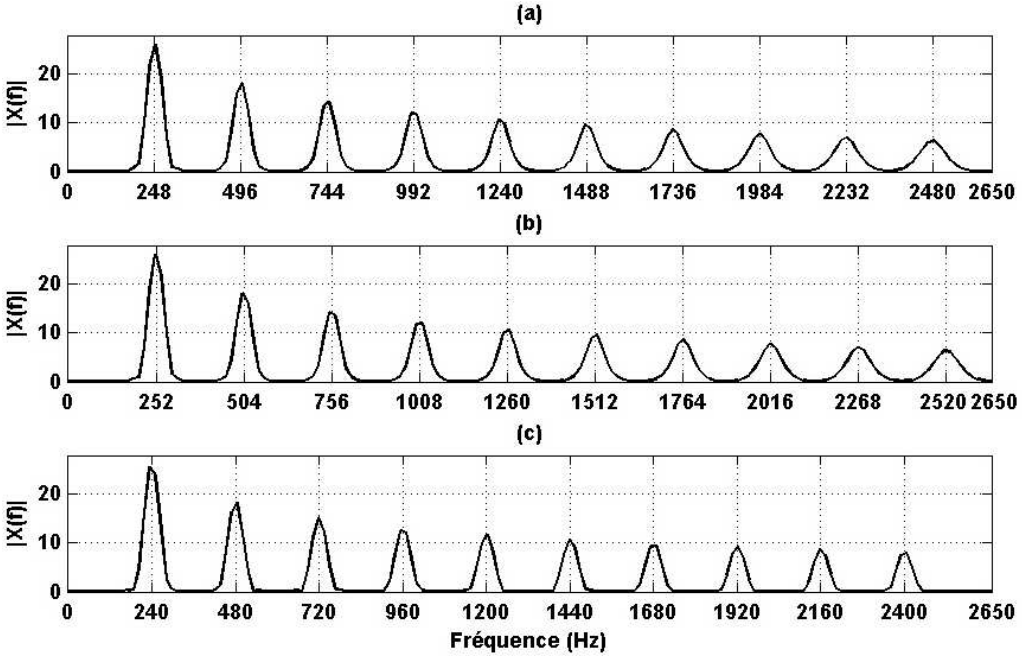

FIGURE 3.11: Exemples de spectres d'amplitude. (a) Spectre d'amplitude de la trame 1 du signal exemple. (b) Spectre d'amplitude de la trame 2 du signal exemple. (c) Spectre d'amplitude d'une trame de 50ms d'un signal avec F_0 fixe à 240Hz.

nommés « harmoniques » aux multiples de la F_0. Ces trois séries d'harmoniques forment trois structures harmoniques. L'amplitude des harmoniques décroît en inverse de racine. Dans le graphique (c) de la figure 3.11, les dix harmoniques se trouvent parfaitement alignées avec les multiples de la F_0 de 240Hz stable tout le long de la trame. Les harmoniques ont toutes la même forme et la même étendue spectrale. Dans les graphiques (a) et (b), les harmoniques sont alignés avec les multiples de la valeur moyenne de l'évolution de la F_0 pendant la trame. Ainsi, la valeur moyenne de la F_0 de la première trame est de 248Hz. Cette valeur est effectivement la fréquence de répétition des harmoniques du gra-

phique (a). Il en est de même pour le graphique (b) qui présente une structure harmonique cadencé fréquentiellement à 252Hz. Ce 252Hz correspond bien à la moyenne de l'évolution de la F_0 pendant la seconde trame. L'évolution de la F_0 a également pour effet d'étaler fréquentiellement les harmoniques qui n'ont plus une étendue spectrale constante comme dans le graphique (c). Plus le numéro de l'harmonique est grand, plus cette harmonique est large. Cet effet d'étalement est gênant car il implique qu'à partir d'une certaine fréquence, il n'est plus possible de distinguer deux harmoniques successives. Ce phénomène est important car il peut être préjudiciable à l'estimation de F_0 dans le cas où la méthode d'estimation s'appuie sur la détection des harmoniques de la structure fréquentielle de la trame. En effet, les harmoniques des hautes-fréquences seront moins facile à détecter. La problématique de détection des harmoniques dans un spectre est d'ailleurs une problématique en soi qui a été étudiée et le lecteur intéressé pourra lire [Depalle and Helie, 1997].

Comment exploiter ces spectres pour estimer la F_0 d'une trame ? La structure des spectres d'amplitude des signaux exemples est bien particulière et possède des harmoniques bien marquées qui ressortent du bruit de manière évidente. Estimer la F_0 revient à s'appuyer sur cette structure harmonique. Il suffit de reconnaître les motifs spectraux récurrents par des techniques dites de « pattern matching » et de quantifier cette récurrence pour être en mesure de quantifier la force de périodicité d'une fréquence donnée. Estimer la F_0 revient alors à trouver la fréquence maximale dont les multiples coïncident avec des harmoniques. La structure harmonique marquée du signal exemple se retrouve également sur des sons de parole voisés réels (cf. figure 3.13).

3.3.2.1 Cepstre

Le cepstre C est la transformée de Fourier inverse du logarithme du spectre d'amplitude du signal étudié. Noll [Noll, 1967] utilise le cepstre pour déterminer la F_0. Le cepstre est conséquence d'un modèle de production de la parole très répandu nommé modèle source-filtre. Ce modèle considère qu'un signal de parole $x(t)$ est le résultat de la convolution d'un signal excitateur nommé $s(t)$ avec la réponse impulsionnelle $h(t)$ du filtre correspondant au conduit vocal. Dans le cas d'un signal voisé, le signal excitateur est périodique. L'équation 3.15 montre en quoi le cepstre est un outil intéressant puisqu'il permet de linéariser le problème de séparation entre la contribution du signal excitateur et la contribution du conduit vocal. Le symbole « $*$ » est le symbole de convolution.

$$\begin{aligned} x(t) &= s(t) * h(t) \\ X(f) &= S(f).H(f) \\ |X(f)| &= |S(f)|.|H(f)| \\ \ln|X(f)| &= \ln|S(f)| + \ln|H(f)| \end{aligned} \tag{3.15}$$

La convolution de la première ligne de l'équation 3.15 devient un produit en passant dans le domaine fréquentiel par la transformée de Fourier. Ensuite, le logarithme de la dernière ligne permet de passer du produit à une somme et linéarise l'équation. La transformée de Fourier inverse étant un opérateur linéaire, le cesptre du signal $x(t)$ est donc la somme des cepstres du signal excitateur et de la réponse impulsionnelle du filtre du conduit vocal. La figure 3.12 présente le résultat du cepstre sur le signal exemple. Il apparaît un maximum local au lag 32 qui correspond à 250Hz (8000/32). Ce pic correspond

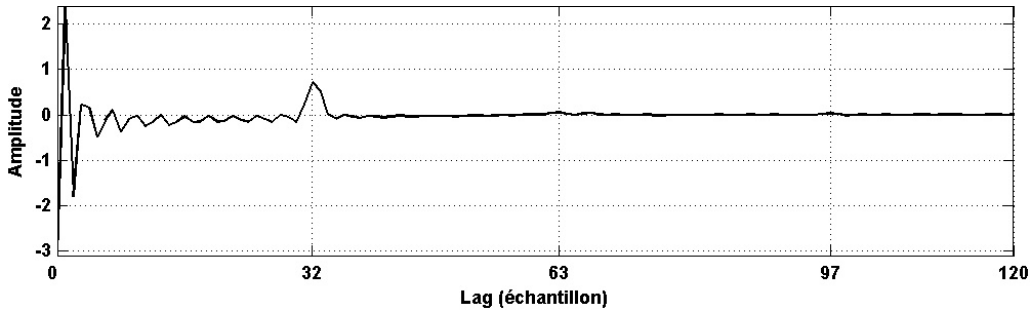

FIGURE 3.12: Cepstre de la trame 1 du signal exemple. La F_0 estimée avec interpolation quadratique vaut 248.4Hz.

à la contribution de la source périodique $s(t)$. Avec une interpolation quadratique, la maximum est placé en 248.4Hz au lieu de 248Hz. L'estimation de F_0 est donc correcte. Il y a bon nombre d'autres maxima locaux aux lags inférieurs à 32. Ceux-ci pourraient perturber l'estimation de la F_0. Il en existe également aux lags 63 et 97 qui correspondent respectivement au double et au triple de 32. Ils sont moins visibles car leur amplitude est moindre mais ils sont présents. Le cepstre est utilisé dans la méthode de séparation de voyelles de Stubbs et Summerfield [Stubbs and Summerfield, 1988, 1990] présentée dans le chapitre 2 sous-paragraphe 2.3.2.2 et est capable d'estimer la F_0 de deux voyelles mélangées. L'équation 3.16 présente le cepstre calculé sur un mélange de deux signaux $x_1(t)$ et $x_2(t)$ résultant tous deux de la convolution de leurs sources et de leurs filtres respectifs. La source de $x_1(t)$ et notée $s_1(t)$ et le filtre $h_1(t)$. La source de $x_2(t)$ est notée $s_2(t)$ et le filtre $h_2(t)$.

$$\begin{aligned} x(t) &= x_1(t) + x_2(t) \\ x(t) &= s_1(t) * h_1(t) + s_2(t) * h_2(t) \\ X(f) &= S_1(f).H_1(f) + S_2(f).H_2(f) \\ |X(f)| &= |S_1(f).H_1(f) + S_2(f).H_2(f)| \\ \ln|X(f)| &= \ln|S_1(f).H_1(f) + S_2(f).H_2(f)| \end{aligned} \tag{3.16}$$

Dès qu'il y a mélange de sources, il n'est plus possible de linéariser les contributions des sources et des filtres. D'un point de vue théorique, le cepstre ne semble pas complètement approprié pour estimer des F_0 multiples.

3.3.2.2 Histogramme de Schroeder

Le cepstre ne fait pas apparaitre explicitement la structure harmonique d'un signal voisé. Une méthode proposée par Schroeder [Schroeder, 1968] s'appuie directement sur cette structure particulière. La technique consiste à d'abord repérer les partiels, maxima locaux du spectre d'amplitude. Pour un signal de parole réel, ceci est assez simple dans les basses fréquences (<1kHz) car ces partiels sont des pics bien marqués. Néanmoins cela devient plus compliqué au-dessus de 2kHz car les pics ne ressortent plus aussi clairement du plancher de bruit pour être calculés de manière robuste. Deux raisons principales expliquent ce phénomène. D'une part, par nature la majorité de l'énergie des parties voisées de la parole se trouve en dessous de 2kHz. D'autre part, l'étalement spectral décrit précédemment rend

TABLE 3.1: Détails des valeurs fréquentielles de l'histogramme de Schroeder (F_0=248Hz) pour les quatre premiers partiels notés H_k. Les cases vides désignent des valeurs harmoniques en dehors de l'intervalle de recherche de F_0 fixé à $[50, 800Hz]$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H_1	248	124	82.6	62														
H_2	496	248	165.3	124	99.2	82.6	70.8	62	55.1									
H_3	744	372	248	186	148.8	124	106.2	93	82.6	74.4	67.6	62	57.2	53.1				
H_4		496	330.6	248	198.4	165.3	141.7	124	110.2	99.2	90.1	82.6	76.3	66.1	62	58.3	55.1	52.2

les partiels de haute-fréquence moins émergents. Dans l'exemple de la figure 3.13 la dégradation des partiels du spectre d'amplitude est clairement visible. Au-dessus de 1.5kHz, il devient difficile de placer correctement les harmoniques de la voyelle prononcée. La seconde étape consiste à diviser chacune des

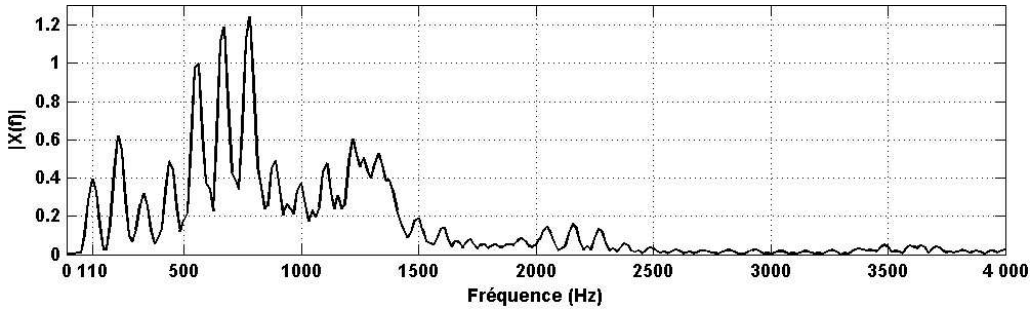

FIGURE 3.13: Spectre d'amplitude tiré de Bagshaw rl001. Extrait de 40ms de la voyelle /a/ de F_0 d'environ 110Hz. Signal identique à celui de la figure 3.3

fréquences centrales des partiels détectées par les entiers naturels allant de 1 jusqu'à une certaine limite et de dresser ensuite un histogramme des valeurs obtenues. Cette méthode fait explicitement intervenir l'exploitation de la structure harmonique au travers de la détection de partiels et de la recherche d'une relation harmonique les liant. Le tableau 3.1 illustre le processus de fabrication de l'histogramme. Chaque ligne représente les valeurs de l'histogramme obtenues par un partiel de la structure harmonique. Le tableau détaille les quatre premier partiels d'une structure harmonique dont la F_0 vaut 248Hz. Ainsi l'harmonique 4 valant 992Hz est divisé par 1 et le résultat de la division est dans la colonne 1. Seulement 992Hz est hors de la zone de recherche $[50, 800Hz]$ donc la colonne 1 est vide. La colonne 2 contient 992/2 soit 496Hz, la colonne 3 contient 992/3 soit 330.6Hz et ainsi de suite jusqu'à ce que la valeur soit inférieure à 50Hz. Une fois le tableau construit pour chacun des partiels détecté, il faut calculer le nombre d'occurrences de chacune des valeurs distinctes du tableau et l'histogramme de Schroeder est établi. Par cette technique, la fréquence fondamentale qui est une fréquence multiple commune à tous les partiels en relation harmonique doit apparaître comme le pic le plus fort de l'histogramme. Un histogramme de Schroeder est donné dans la figure 3.14. Il est construit à partir de la première trame du signal exemple (dont la F_0 évolue de 240 à 256Hz linéairement pour une moyenne de 248Hz). Les partiels sont estimés en repérant les maxima locaux du spectre d'amplitude cf. figure 3.11 (a) et en interpolant de manière quadratique la position fréquentielle de chacun pour améliorer la précision. Il faut noter qu'une légère erreur d'estimation dans leur position fréquentielle est

systématiquement faite malgré l'interpolation. Ensuite chacune des positions fréquentielles des partiels est divisée par des valeurs entières comme décrit dans le tableau 3.1. On calcule l'histogramme des valeurs obtenues entre 50 et 800Hz et pour trois résolutions fréquentielles différentes. Le graphique

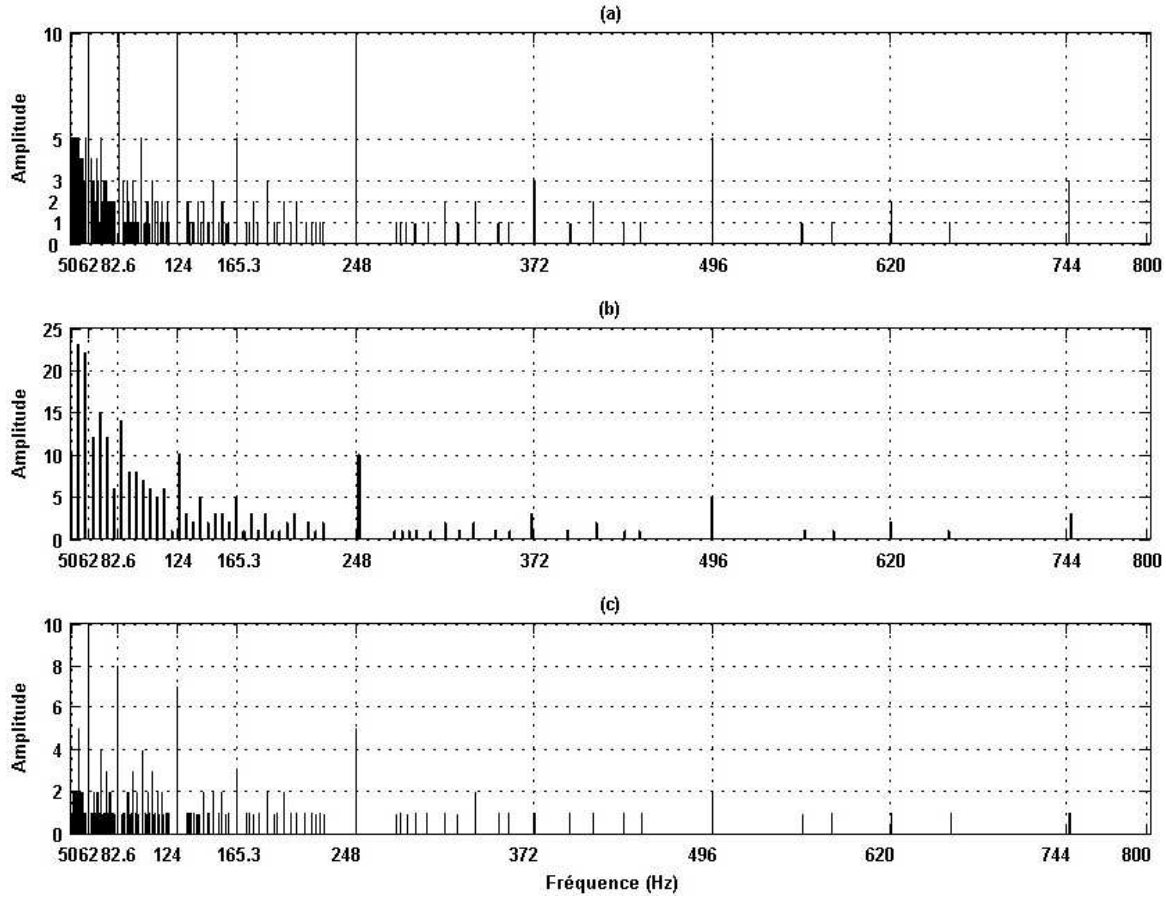

FIGURE 3.14: Histogrammes de Schroeder sur la trame 1 du signal exemple. F_0=248Hz. (a) Résolution fréquentielle de 1Hz. (b) Résolution fréquentielle de 5Hz. (c) Résolution fréquentielle de 0.1Hz.

(a) de la figure 3.14 présente l'histogramme de Schroeder pour une résolution fréquentielle de 1Hz. Le graphique (b) possède une résolution fréquentielle de 5Hz. La résolution fréquentielle du graphique (c) est de 0.1Hz. L'estimation de la F_0 est réalisée en estimant la position fréquentielle du maximum local de plus grande amplitude dans l'histogramme. Dans le graphique (a), cette manière de procéder fournit 4 maxima locaux d'amplitude 10 en 62, 82.6, 124 et 248Hz. L'amplitude maximale vaut 10 car le nombre d'harmoniques de la structure est 10. Le candidat à 248Hz est le bon candidat mais se trouve en compétition avec ses sous-multiples 2, 3 et 4. Dans le graphique (b) l'estimation de la F_0 donne 55Hz (maximum d'amplitude 23) ce qui est une erreur. Dans le graphique (c), l'estimation de la F_0 donne 62Hz (amplitude 10) ce qui est également une erreur. Ces erreurs sont des conséquences directes du choix de résolution fréquentielle. Le choix d'une résolution fréquentielle élevée comme pour le graphique (c) implique une grande précision dans l'estimation de la valeur de F_0. Néanmoins, le moindre écart, même faible, dans l'estimation de la position d'une harmonique implique qu'un pic qui apparaît comme unique à une résolution moins fine sera décomposé en un ensemble de pics de plus

faibles amplitudes.

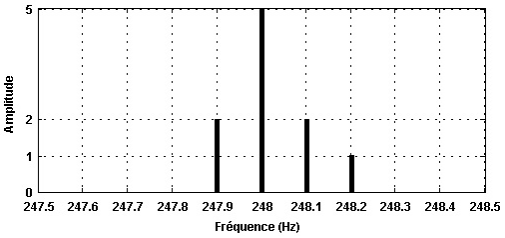

FIGURE 3.15: Zoom du graphique (c) de la figure 3.14 autour de 248Hz. Le pic théorique d'amplitude 10 est séparé en 4 pics dont la somme vaut 10.

C'est notamment le cas pour le pic en 248Hz qui ne vaut plus 10 mais 5. La figure 3.15 illustre l'effet néfaste d'une résolution trop fine. Les erreurs d'estimation des positions des harmoniques engendrent la séparation du pic principal d'amplitude 10 en 4 pics de plus faibles amplitudes autour de 248Hz. La somme des amplitudes de tous les pics vaut bien 10. Le graphique (b) présente l'histogramme avec une faible résolution fréquentielle. Cela se traduit par un manque de précision dans l'estimation de la F_0. En considérant le pic de F_0 le plus proche de 248Hz, l'estimation fournit une F_0 de 250Hz au lieu de 248Hz. Plus important encore que ce manque de précision est l'émergence importante des pics dans les basses fréquences. Ce phénomène s'explique par le fait que dans les basses fréquences, l'écart fréquentiel entre les dents se resserre. En effet, pour une harmonique donnée la série de ses sous-multiples est une série hyperbolique en $1/n$. Plus n est grand, moins l'écart entre $1/n$ et $1/(n-1)$ est grand. Étant donné que la résolution fréquentielle est constante, il y a donc de plus en plus de dents incluses dans un même pic de l'histogramme. Les erreurs des graphiques (b) et (c) induisent qu'il y a un compromis à trouver dans la résolution fréquentielle de l'histogramme pour que celui-ci ait une forme exploitable. Le graphique (a) est un bon compromis de résolution mais ce compromis n'est pas forcément aisé à fixer. D'un point de vue théorique, l'histogramme de Schroeder est un outil très intéressant mais en pratique, la résolution joue un rôle important si bien que son utilisation devient délicate. Et quand bien même le compromis est correct, il subsiste encore le problème des sous-multiples de la F_0 qui sont exactement de même amplitude que le pic en 248Hz. Ces valeurs sont des candidats parasites de l'histogramme.

3.3.2.3 Peigne de Martin

Martin dans [Martin, 1981, 1982] propose une technique d'estimation de F_0 qui consiste à calculer le produit scalaire du spectre d'amplitude d'une trame avec un peigne fréquentiel dont les dents sont régulièrement espacées en fréquence. Cette propriété permet de détecter les structures harmoniques d'un signal voisé qui présente des harmoniques régulièrement espacées en fréquence. L'écart fréquentiel entre les dents du peigne fréquentiel est réglable dans la gamme de fréquence dans laquelle la F_0 est recherchée (ex. 50-800Hz). Le peigne est limité en nombre de dents et l'amplitude des dents obéit à une certaine fonction de décroissance. La figure 3.16 illustre une implémentation possible d'un peigne inspiré de la technique de Martin. Dans cet exemple, le peigne possède 10 dents et la décroissance des dents est en avec k le numéro de la dent. Le résultat de ce peigne est donné dans la figure 3.17. La zone de recherche de la F_0 est réglée entre 50 et 800Hz. La résolution fréquentielle est de 1Hz pour cet exemple. Toutefois, cette manière de parcourir la zone de recherche n'est sans doute pas la plus astucieuse. En effet, une erreur même petite dans les basses fréquences (due à la résolution limitée à 1Hz) peut devenir importante lors du calcul de la position d'un harmonique élevé (ie. 0.5Hz d'erreur

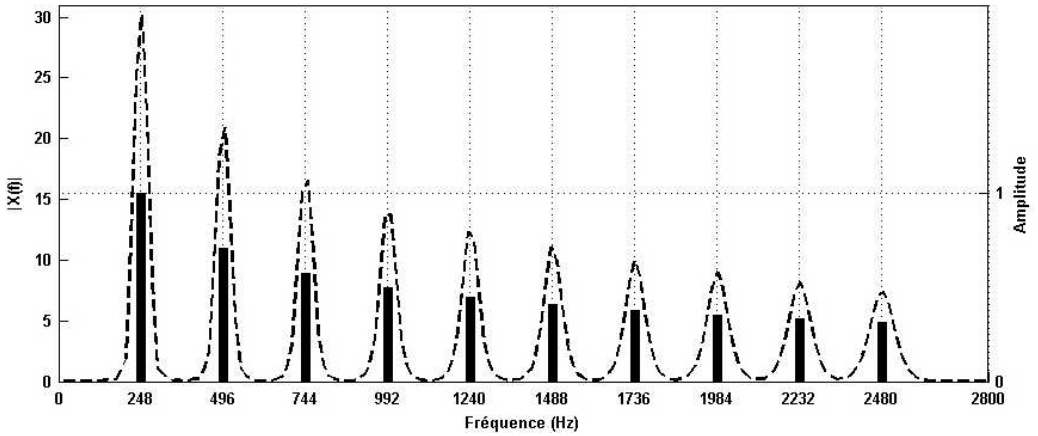

FIGURE 3.16: Spectre d'amplitude et peigne de Martin. (trait pointillé) Spectre d'amplitude de la trame 1 du signal exemple. (barres verticales) dents d'un peigne de Martin. L'axe des ordonnées de gauche correspond à l'échelle du spectre d'amplitude. L'axe des ordonnées de droite correpond à l'amplitude des dents du peigne de Martin. Dans cet exemple, les dents coïncident parfaitement avec les harmoniques (F_0=248Hz).

sur le fondamental équivaut à 5Hz d'erreur sur la 10ème harmonique). Il est donc plus approprié de calculer la fonction de pitch en commençant par les hautes fréquences en divisant la fréquence par des entiers. Une précision de 1Hz dans les hautes fréquences est suffisante et permet d'être précis dans les basses fréquences (ie. 1Hz de précision sur la 10ème harmonique est 0.1Hz sur la fondamentale). Dans la

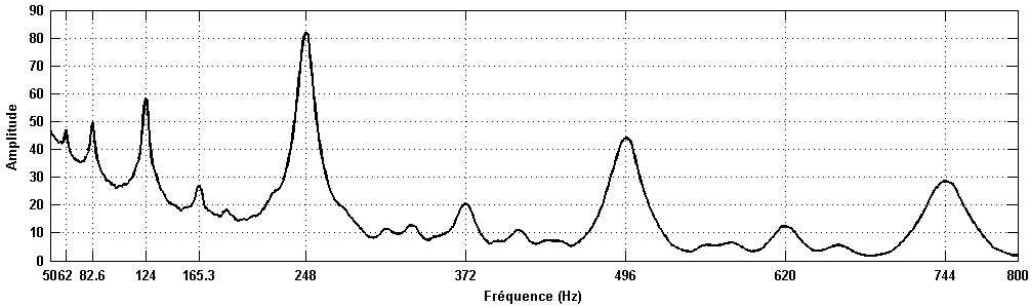

FIGURE 3.17: Fonction de pitch obtenue par notre configauration du peigne de Martin.

figure 3.17, le pic maximal correspond bien à la F_0 de la première trame du signal exemple c'est-à-dire 248Hz. D'autres maxima locaux sont apparents et ceux-ci restent d'amplitude inférieure au maximum local en F_0. Toutefois, ces pics sont parasites et peuvent induire des erreurs d'estimation. La méthode de Martin ne pose plus le même problème de résolution fréquentielle que l'histogramme de Schroeder. Certes, il faut que la résolution fréquentielle de la zone de recherche de F_0 soit suffisante pour faire apparaître les maxima locaux mais le problème d'une résolution fréquentielle trop importante n'existe plus. Plus la résolution fréquentielle de la zone de recherche est importante, plus le résultat du peigne est précis. L'exemple de son utilisé jusqu'à présent est artificiel et possède une structure harmonique bien marquée dans laquelle toutes les harmoniques sont présentes. Or, pour des signaux réels comme la parole téléphonique, la structure harmonique est altérée. Hormis pour les voix d'enfants, le fondamental sera probablement absent ou fortement atténué car ce signal est souvent filtré en dessous de 300Hz.

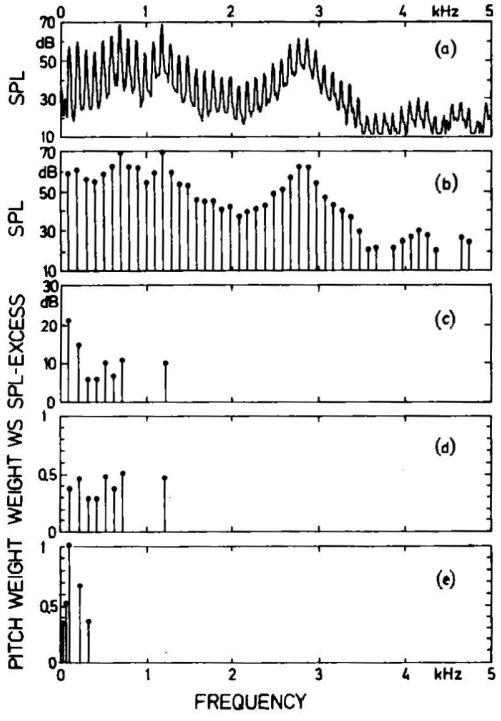

FIGURE 3.18: Principe d'obtention des huit partiels significatifs selon Terhardt et al. Source : [Terhardt et al., 1982a]

De plus, il est aisé d'imaginer des structures harmoniques à « trous » dans lesquelles il manque des harmoniques. Terhardt propose un AEP basé sur sa théorie du *Virtual Pitch* [Terhardt et al., 1982a,b]. Sa théorie distingue clairement un *spectral pitch* (ou pitch spectral) qui est une F_0 présente physiquement dans le signal, d'un *virtual pitch* (ou pitch virtuel) qui n'est pas présent physiquement dans le signal et qui pourtant est perçue. Sa méthode décrite dans la figure 3.18 consiste pour chacune des trames du signal à extraire le spectre d'amplitude cf. graphique (a), d'en calculer les harmoniques (ou partiels pour la musique) cf. graphique (b) puis d'évaluer pour chacun des partiels l'effet du masquage et de ne conserver l'harmonique que si son amplitude se trouve au-dessus d'un seuil de masquage cf. graphique (c). Seules les huit harmoniques les plus fortes sont conservées et en fonction de leur fréquence, ces harmoniques sont multipliées par une fonction de pondération qui donne plus d'importance aux fréquences entre 300 et 1000Hz cf. graphique (d). Enfin, un histogramme de Schroeder est extrait à partir de ces huit harmoniques cf. graphique (e). En s'appuyant sur la structure harmonique et pas uniquement sur une seule harmonique, sa méthode est beaucoup plus robuste aux signaux pour lesquels la fondamentale est absente où dans lesquelles la structure harmonique est incomplète. Les travaux de Terhardt et al. conceptualisent l'importance de considérer la structure harmonique et pas uniquement des harmoniques isolées. C'est ce qui est également réalisé par la méthode du peigne de Martin qui recherche une cohérence maximale entre un peigne harmonique et le spectre d'amplitude. Le peigne de Martin s'appuie donc sur la recherche d'une structure harmonique. Ceci rend ces méthodes robustes. Cette caractéristique est très importante pour nous et justifie pourquoi nous avons opté pour un AEP fréquentiel par peigne.

3.3.2.4 Méthode SHS de Hermes

Une méthode nommée SHS (ou *SubHarmonic Summation*) est proposée par Hermes [Hermes, 1988]. Les étapes de sa méthode sont illustrées figure 3.19. Le spectre d'amplitude d'une trame de signal est calculé cf. graphique (a). Les harmoniques y sont détectés et sont conservés, le reste étant considéré comme du bruit et mis à zéro cf. graphique (b). Le spectre est représenté dans une échelle logarithmique par l'intermédiaire d'interpolations cf. graphique (c). Une fonction de pondération cf. graphique (d) est appliquée au spectre pour atténuer les basses fréquences cf. graphique (e). On somme alors le spectre d'amplitude avec cinq versions décalées de lui même. La première version est décalée de $\log_2(1)$ sur l'échelle logarithmique ce qui correspond à s Hz sur l'échelle linéaire cf. graphique (f). La seconde version est décalée de $2s$ sur une échelle linéaire (g). Les graphiques (h), (i) et (j) sont des versions

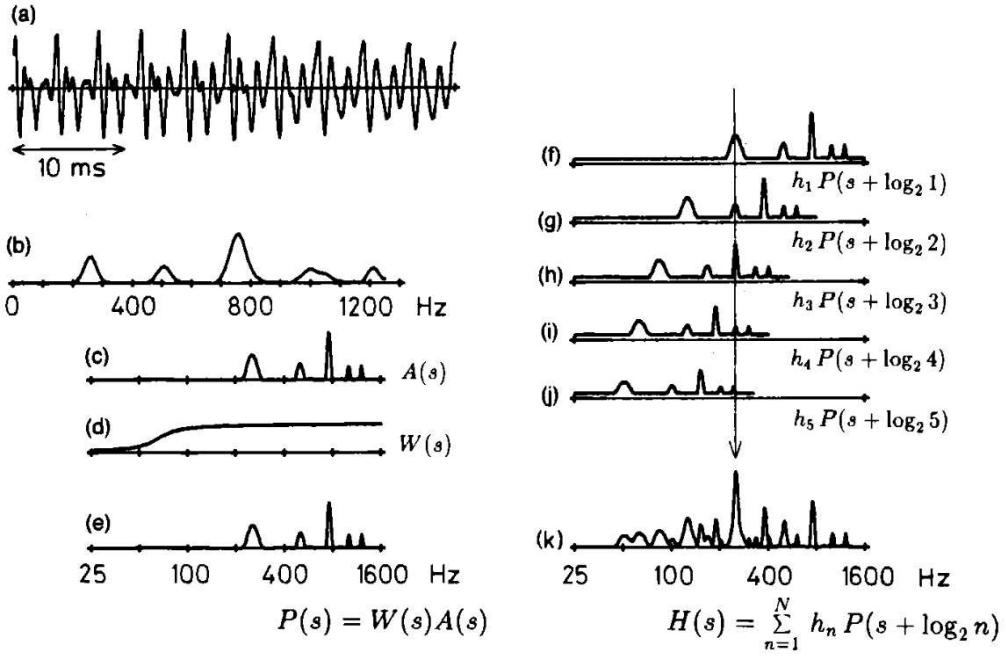

FIGURE 3.19: Chaine de traitement de l'estimateur de F_0 d'Hermes. (a) Signal de parole. (b) Spectre d'amplitude avec échelle des fréquences linéaire. (c) Spectre d'amplitude avec échelle des fréquences logarithmique. (d) Fonction de pondération. (e) Fonction P produit du spectre d'amplitude en échelle logarithmique et de la fonction de pondération. (f) Fonction P décalée de $\log_2(1)$. (g) Fonction P décalée de $\log_2(2)$. (h) Fonction P décalée de $\log_2(3)$. (i) Fonction P décalée de $\log_2(4)$. (j) Fonction P décalée de $\log_2(5)$. (k) Fonction H somme de (e), (f), (g), (h), (i) et (j). Source : [Hermes, 1988].

du graphique (e) décalées respectivement de trois, quatre et cinq s. Le principe est donc de sommer les spectres décalés de valeurs fréquentielles en relation harmoniques ($1s$, $2s$, $3s$, etc.). Ce principe est proche de celui proposé par Martin (cf. 3.3.2.3) ce qui se retrouve dans le résultat obtenu qui y ressemble cf. graphique (k). Un intérêt de la méthode d'Hermes est de représenter les fréquences sur une échelle logarithmique.

3.3.2.5 Méthode SHR de Sun

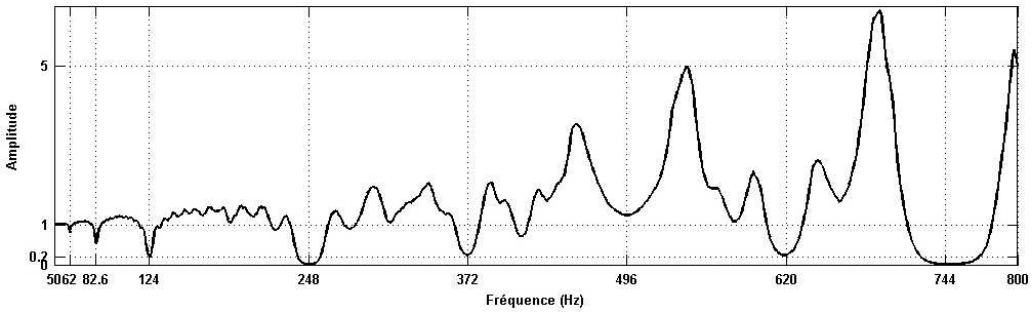

FIGURE 3.20: Fonction de pitch obtenue par la méthode de Sun. Le seuil préconisé par Sun est 0.2.

Une autre méthode d'estimation de F_0 est proposée par Sun [Sun, 2000]. Elle consiste à calculer

le *subharmonic-to-harmonic ratio* ou SHR. Le principe de la méthode est illustrée figure 3.21. Deux sommes sont extraites du spectre d'amplitude, la somme des harmoniques notée SH et la somme des sous-harmoniques notée SS. Le SHR final donné dans l'équation 3.17 est le ratio de ces deux sommes.

$$SHR = \frac{SS}{SH} \tag{3.17}$$

SS représente la somme des amplitudes des pics (ronds dans l'exemple de la figure 3.21 et SH correspond à la somme des amplitudes des creux (carrés rouges). Dans l'exemple de la figure 3.21, la fréquence à laquelle sont placées les harmoniques correspond à la F_0. Par définition de la méthode, les sous-harmoniques sont positionnées au milieu de deux harmoniques. Il serait d'ailleurs plus correct de parler d'inter-harmoniques que de sous-harmoniques. L'équation 3.18 donne les valeurs théoriques de SH et SS lorsque les harmoniques sont placées en F_0. N représente le nombre d'harmoniques et d'inter-harmoniques considérées.

$$\begin{aligned} SH &= \sum_{k=1}^{N} S(kF_0) \\ SS &= \sum_{k=1}^{N} S((k-\tfrac{1}{2})F_0) \end{aligned} \tag{3.18}$$

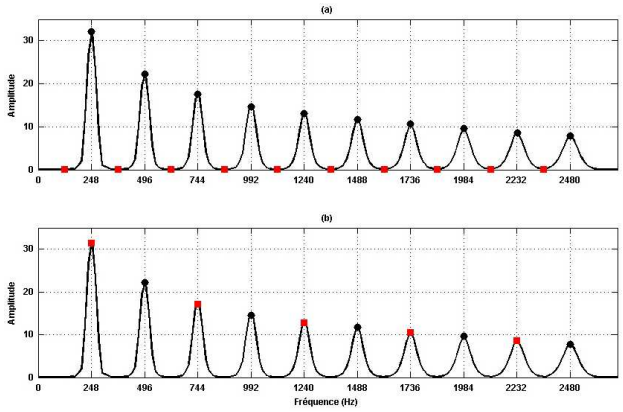

FIGURE 3.21: Illustration du principe de l'algorithme de Sun sur le spectre d'amplitude de la trame 1 du signal exemple.

SHR est un indicateur de l'émergence des inter-harmoniques par rapport aux harmoniques. En effet, en faisant varier la fréquence des harmoniques dans une zone de recherche il est possible de déterminer pour quelle fréquence le SHR est le plus faible. Dans l'article, le seuil est fixé à 0.2. Tout minimum local sous ce seuil est considéré comme une hypothèse F_0. La figure 3.20 illustre le calcul de SHR pour la trame 1 du signal exemple. Effectivement, il apparaît un minimum local proche de 0 en 248Hz. Néanmoins, d'autres minima locaux gênants se trouvent en 124Hz, 372Hz et surtout 744Hz. Il est a noter que dans le cas d'un signal non périodique comme typiquement un bruit blanc gaussien, le SHR est statistiquement proche de 1. La méthode proposée souffre donc de minima locaux parasites et également très étendus fréquentiellement. Par exemple, le minimum local autour de 744Hz est en dessous du seuil de 0.2 de 704Hz à 771Hz soit 47Hz d'étendue fréquentielle. Toutefois, cette méthode est intéressante car elle permet de rehausser le pic parasite au double de la F_0 qui est placé à 496Hz dans notre exemple. L'observation du graphique (b) de la figure 3.21 permet d'expliquer ce phénomène. Si la fréquence étudiée est au double de la F_0, alors les positions fréquentielles inter-harmonique coïncident avec des harmoniques. La valeur du SHR sera donc rehaussée. Ce point signifie que cette méthode permet d'atténuer le minimum local correspondant au double de la F_0 de manière sélective. Malgré cet avantage, la méthode de Sun ne pourra pas être utilisée telle quelle en situation multipitch car dans cette situation il n'est pas garanti que l'inter-

harmonique soit un creux. Il se peut qu'une harmonique soit placée à l'inter-dent et le SHR n'a alors plus de sens.

3.3.2.6 Méthode SWIPE de Camacho

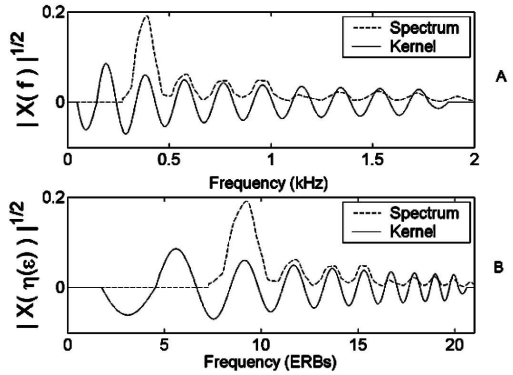

FIGURE 3.22: Ondelette utilisée par Camacho (*kernel*). (A) Fréquences en échelle linéaire. (B) logarithmique. Source : [Camacho, 2007].

Camacho [Camacho, 2007] et Camacho & Harris [Camacho and Harris, 2008] utilisent dans SWIPE la technique de produit scalaire entre un spectre et une fonction noyau (ou *kernel* en anglais) donnée figure 3.22. La fonction noyau de Camacho est d'amplitude décroissante comme le peigne de Martin. Elle en diffère car elle n'est pas discrète et qu'elle possède des valeurs négatives. L'auteur [Camacho, 2007] donne d'ailleurs un parallèle intéressant entre l'ACF, le cepstre et sa fonction kernel. Il est décrit dans la figure 3.23. La formulation fréquentielle de l'ACF revient à calculer, pour chaque lag, le produit scalaire entre le carré du spectre d'amplitude et un cosinus dont le lag étudié est la période. Le cepstre est aussi le fruit du produit scalaire entre un cosinus et le logarithme du spectre d'amplitude. Le kernel de Camacho est d'allure sinusoïdal mais d'amplitude décroissante et dont certains lobes positifs peuvent manquer (version SWIPE'). La figure 3.24 présente le résultat de l'ondelette de Camacho sur la trame

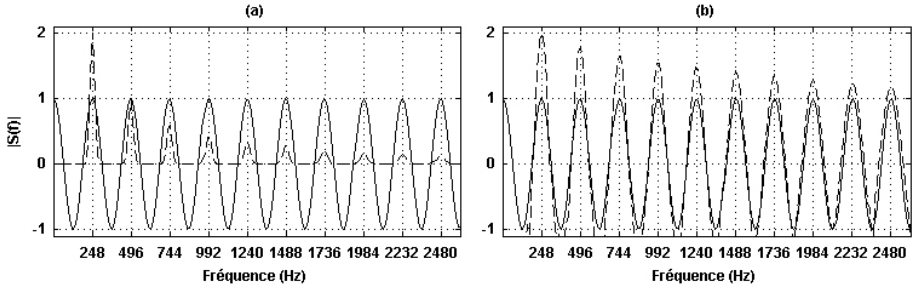

FIGURE 3.23: Interprétations fréquentielles des ACF et Cepstre. (a) ACF. (b) Cesptre.

1 du signal exemple. Le maximum local de plus forte amplitude est bien situé à 248.1Hz et l'estimation de F_0 est donc correcte. Il subsiste quelques maxima locaux parasites notamment à la sous-octave qui peuvent éventuellement produire des erreurs d'estimation. Toutefois, le nombre de pics parasites est très inférieur aux méthodes présentées précédemment ce qui implique qu'elle est sans doute une des plus aboutie de l'état de l'art pour l'estimation de F_0 en situation monopitch. En revanche, l'auteur n'a pas conçu cette méthode pour une estimation multipitch. La possibilité de donner des valeurs négatives à la fonction noyau de SWIPE est également intéressante et se retrouvera dans notre AEP.

Toutes les approches fréquentielles présentées quantifient la ressemblance d'un spectre à une structure harmonique. Ces techniques de *pattern matching* permettent de donner pour chacune des fré-

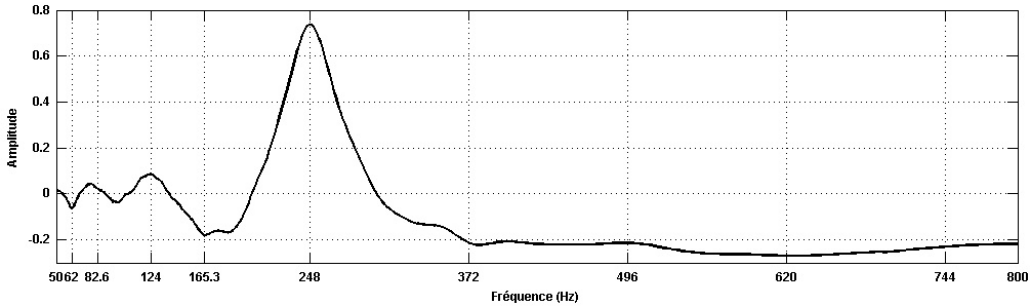

FIGURE 3.24: Fonction de pitch obtenue par SWIPE sur la trame 1 du signal exemple. La F_0 estimée est égale à 248.1Hz.

quences étudiées une valeur qui quantifie la force de la fréquence comme candidat F_0. Cette façon de procéder se traduit par une fonction de pitch comportant des maxima locaux. La F_0 à estimer est un de ces maxima. L'estimation de la F_0 consiste à trouver le maximum local de plus forte amplitude. La fonction de F_0 des approches fréquentielles comporte des pics parasites multiples et sous-multiples ce qui la différencie de la fonction de F_0 des approches temporelles.

3.3.3 Approches spectro-temporelles

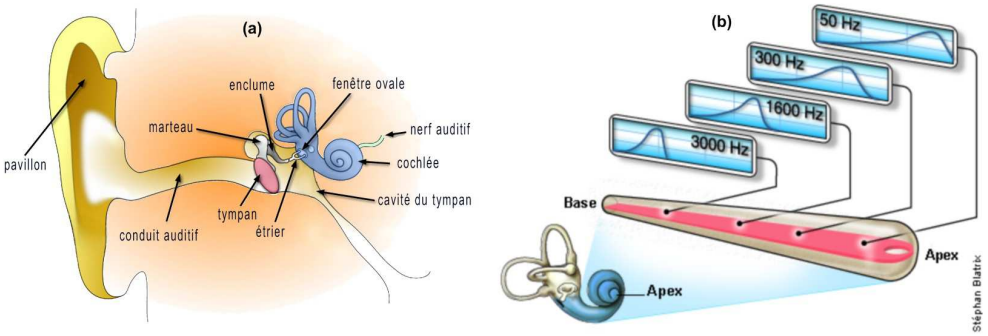

FIGURE 3.25: Fonctionnement de l'oreille. (a) Anatomie de l'oreille. (b) Schéma de la membrane basilaire et de sa sélectivité fréquentielle. Source : `http://www.iurc.montp.inserm.fr/cric51/audition/francais/cochlea/cocphys/fcocphys.htm`.

Les approches spectro-temporelles décomposent le signal dans un espace temps-fréquence. Une manière de décomposer un signal est d'utiliser un modèle cochléaire qui s'approche au mieux du fonctionnement biologique de l'oreille. Schématiquement, une onde sonore émise dans l'environnement est captée par notre pavillon dont la forme favorise la concentration de l'onde vers le conduit auditif. L'onde parvient alors jusqu'au tympan, membrane qui entre en vibration en fonction de l'onde sonore. Les osselets ont pour rôle de transmettre la vibration tympanique jusqu'à la cochlée. Les osselets transforment la vibration tympanique en mouvement mécanique excitateur de la cochlée. Ainsi, l'étrier met en mouvement la fenêtre ovale ce qui va propager la vibration dans la cochlée (cf. (a) de la figure 3.25). La cochlée contient la membrane basilaire sur laquelle sont placées des cellules ciliées

qui réagissent lorsque la membrane se déforme (cf. (b) figure 3.25). La membrane basilaire est une structure particulière conçue pour entrer en résonance, c'est-à-dire en vibration maximale pour une seule fréquence d'excitation donnée. Ainsi, l'anatomie de la membrane basilaire permet de décomposer le mouvement mécanique de l'étrier dans le domaine spectro-temporel.

Les chercheurs se sont intéressés à la modélisation de la décomposition spectro-temporelle effectuée à l'intérieur de la cochlée. Le signal de parole est ainsi séparé en canaux par un banc de filtres couvrant tout le domaine spectral. Ce banc de filtres est conçu de manière à ressembler aux filtres de la cochlée. Les filtres sont souvent modélisés par des filtres gammatones d'ordre 4 [Glasberg and Moore, 1990, Slaney, 1993] dont un exemple est présenté figure 3.26 à gauche. Il y est illustré la fonction de transfert de chacun des filtres. Leurs bandes passantes augmentent lorsque la fréquence augmente, reproduisant le comportement naturel de notre appareil auditif qui possède une bien meilleure résolution dans les basses fréquences que dans les hautes fréquences. La fréquence centrale de chacun des filtres est déterminée sur une échelle logarithmique ERB (*Equivalent Rectangular Bandwidth*) décrite dans l'équation 3.19. f_{ERB} est la fréquence centrale du filtre en échelle ERB et f_{LIN} est la fréquence centrale du filtre en échelle linéaire.

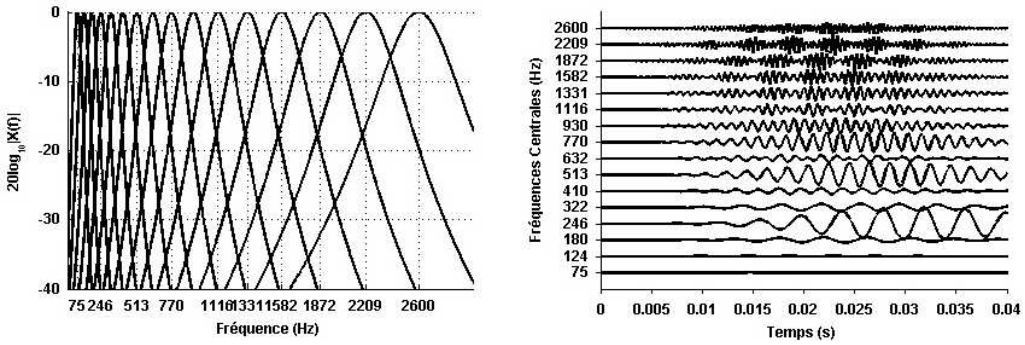

FIGURE 3.26: Filtres gammatones et canaux. (gauche) Illustration de 16 filtre gammatones. (droite) Canaux=signaux temporels issus des filtres gammatones. Le signal utilisé est la trame 1 du signal exemple.

$$\begin{aligned} f_{ERB} &= 24.7(4.37\frac{f_{LIN}}{1000}+1) \\ b_w &= 0.887 f_{ERB} \end{aligned} \tag{3.19}$$

Les fréquences centrales partent d'une fréquence minimale (ex. 50Hz) et cette valeur est incrémentée d'un ERB. Il apparaît également dans l'équation 3.19 que la bande passante notée b_w à -3dB des filtres gammatones d'ordre 4 est proportionnelle à la fréquence centrale ERB.

Les signaux de sortie de chacun des filtres sont appelés canaux. Les AEP construits sur un modèle perceptif appliquent alors des méthodes temporelles ou fréquentielles sur chacun des canaux. Les approches perceptives n'apportent pas en soi une nouvelle manière d'estimer la F_0 mais elles offrent la possibilité de séparer l'ensemble de l'information du signal en bandes de fréquence. Cette séparation possède un double avantage. Tout d'abord, les AEP temporels fonctionnent mieux sur des signaux à bande étroite. Ensuite, la séparation de l'information en canaux distincts permet de comparer les résultats individuels des canaux et d'affiner la prise de décision et l'estimation de la F_0. Si un grand nombre de canaux obtiennent la même estimation de F_0, cette valeur est probablement correcte. Ce principe général est repris non plus sur des canaux mais sur des AEP. Gold [Gold, 1962] se rend

compte que l'estimation de F_0 d'un AEP isolé donne de moins bons résultats que lorsque l'on combine les estimations de plusieurs AEP. Cette constatation est reprise dans les travaux de McGonegal et al. [McGonegal et al., 1975] qui combinent le cepstre et l'autocorrélation dans une méthode nommée SAPD (*SemiAutomatic Pitch Detector*). Multiplier les façons d'estimer la F_0, par différents AEP ou par séparation en canaux permet d'améliorer l'estimation. Friedman [Friedman, 1979] exploite une méthode simple d'estimation de F_0 nommée « *zero-crossing* ». Il s'agit de compter le nombre de passage par zéro d'un segment de parole auquel on a retiré la composante continue et auquel on a retiré les hautes fréquences (typiquement au-dessus de 1kHz). Si on divise la durée temporelle du segment par le nombre de passage par zéro on obtient alors une estimation de la période du signal. L'inconvénient de cette méthode est qu'elle est très peu robuste au bruit. L'algorithme réalise une opération de zéro-crossing sur tous les canaux et regarde si un certain nombre de canaux donnent le même candidat F_0. Cela rend l'AEP plus robuste. Il existe d'autres modèles perceptifs que ceux à base de gammatone dans la littérature [Slaney, 1988, Immerseel and Martens, 1992] mais tous amènent à une décomposition du signal selon un banc de filtre. Généralement, sur chacun des canaux construits à partir du modèle perceptif utilisé, une ACF normalisée est calculée. L'ensemble des ACF de tous les canaux forment ce qui est appelé corrélogramme. Le corrélogramme est utilisé dans [Slaney and Lyon, 1990, Meddis and Hewitt, 1991, Immerseel and Martens, 1992, Patterson et al., 1992, Patterson, 2000]. Le graphique (a) de la figure 3.27 présente le corrélogramme de la première trame du signal exemple. La somme de tous

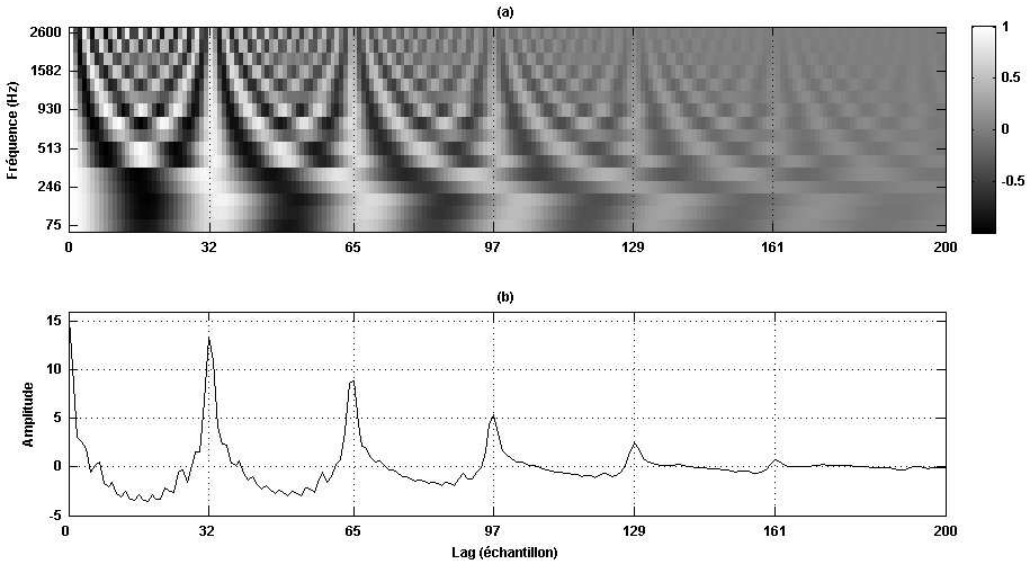

FIGURE 3.27: Corrélogramme et somme des canaux du corrélogramme. (a) Corrélogramme sur la trame 1 du signal exemple. (b) Somme de l'ensemble des canaux du corrélogramme.

les canaux permet d'obtenir une fonction qui possède les mêmes propriétés qu'une ACF. Le graphique (b) de la figure 3.27 montre la somme des canaux du corrélogramme qui sera nommé corrélogramme cumulé. L'estimation de la F_0 repose sur le maximum local de plus forte amplitude et ce maximum est toujours en concurrence avec les autres maxima locaux situés aux lags multiples du lag de la F_0 qui vaut ici 32. Rouat, Liu & Morisette [Rouat et al., 1997] proposent des améliorations dans l'utilisation du corrélogramme et du corrélogramme cumulé (qu'ils nomment *pseudo-periodic histogram*). Le schéma

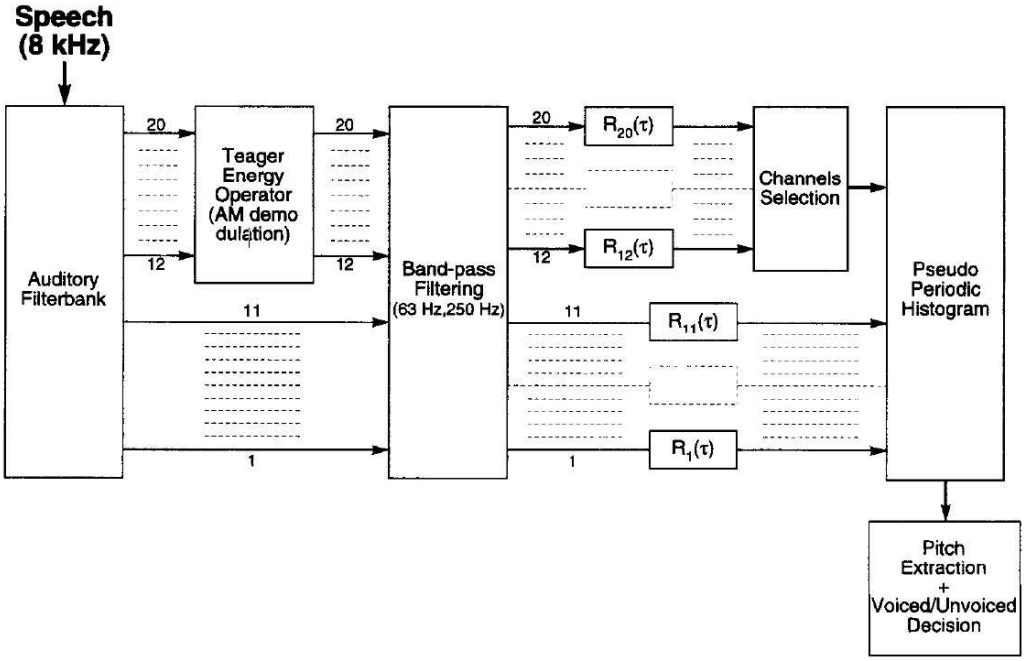

FIGURE 3.28: Schéma bloc de la méthode d'estimation de F_0 de Rouat, Liu & Morisette. Source : [Rouat et al., 1997].

bloc de leur méthode d'estimation est décrit dans la figure 3.28. Une première amélioration consiste à séparer les traitements effectués par les canaux basse fréquence des traitements effectués par les canaux haute fréquence. Les canaux dont la fréquence centrale est inférieure à 1230Hz sont considérés comme basse fréquence. Pour les canaux basse fréquence, l'autocorrélation directe est appliquée au signal. En revanche, pour les canaux haute fréquence, l'autocorrélation de l'enveloppe du signal est calculée. La seconde amélioration concerne encore les canaux haute fréquence.

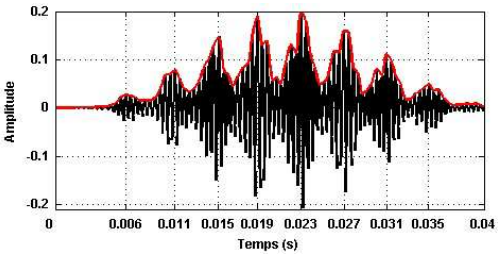

FIGURE 3.29: Enveloppe d'un canal non résolu. (noir) Signal de sortie du canal 15 de fréquence centrale à 2209Hz. (rouge) Enveloppe dont le battement est de période moyenne 4ms (environ 248.2Hz) ce qui correspond à la F_0 de la trame.

Une procédure de sélection est faite pour ne garder que ceux qui conserve une information de pitch sûre. Pour cela, une autocorrélation sur une fenêtre deux fois plus longue est calculée. Si les autocorrélations de la fenêtre courte et deux fois plus longue présentent le même pic, le canal haute fréquence est conservé. Le corrélogramme cumulé est calculé en sommant les autocorrélations des canaux basse fréquence et des canaux haute fréquence sélectionnés. Ces améliorations sont reprises dans [Hu and Wang, 2004] qui est présenté en détail dans le paragraphe 3.4.4. Les filtres de fréquence basse sont relativement fins en termes de bande passante et ne peuvent laisser passer qu'une seule harmonique. L'harmonique est alors considérée comme résolue car le signal du canal ressemble alors à une sinusoïde pure (cf. figure 3.26 droite canal 4 de fréquence centrale à 242Hz soit quasiment

la F_0). Dans l'exemple où F_0 vaut 248Hz, la limite à partir de laquelle on ne peut plus considérer qu'un canal ne contient qu'une seule harmonique correspond à la fréquence pour laquelle la bande passante du canal est supérieure à F_0. Avec la formule de l'équation 3.19, cela correspond à une fréquence de 2361Hz. Au-delà de cette fréquence, les canaux sont assurés de contenir plusieurs harmoniques dans la bande passante. Les harmoniques sont dites dans ce cas non-résolues. Si la F_0 était de 100Hz par exemple, la fréquence à partir de laquelle les harmoniques sont non-résolues serait d'environ 815Hz. Dans le cas d'un canal dans lequel les harmoniques sont non-résolues, le signal devient une somme de sinus plus complexe comme en témoigne le canal 15 de la figure 3.29. La fréquence centrale de ce canal est de 2209Hz. Toutefois, l'information de F_0 est encore présente dans l'enveloppe du signal qui est modulée à la fréquence de la F_0. Il faut donc calculer l'enveloppe des signaux issus des filtres haute fréquence avant d'estimer la F_0. La figure 3.29 présente le signal du canal de fréquence centrale à 2209Hz soit proche de la valeur théorique à partir de laquelle les harmoniques sont non-résolues. Le battement de l'enveloppe se produit à la fréquence moyenne de 248.2Hz soit la F_0 de la trame.

La séparation de l'information du signal en canaux permet de prendre une décision d'estimation par combinaison des résultats de chaque canal, donc plus fiable qu'à partir du signal entier. Malgré cet avantage, l'estimation de la F_0 repose toujours sur l'estimation de la position fréquentielle de maxima locaux dans le corrélogramme, qui en comporte plusieurs et un certain nombre sont parasites.

3.3.4 Point de vue probabiliste

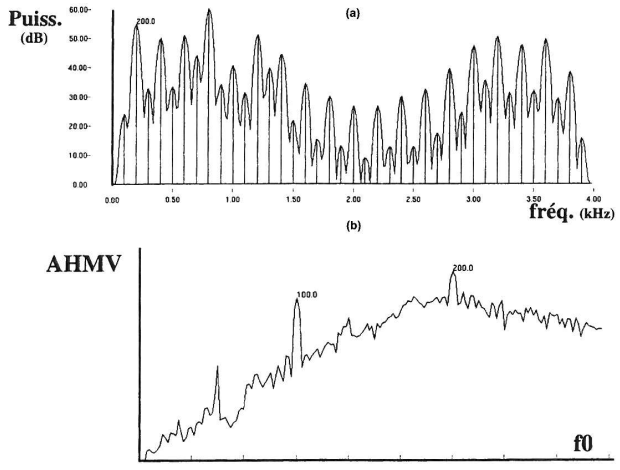

FIGURE 3.30: Méthode probabiliste AHMV. (a) Spectre d'amplitude du signal synthétique de F_0 égale à 200Hz. (b) AHMV représentant a force de pérodicité d'une fréquence donnée. Le maximum global se trouve effectivement en 200Hz. Source : [Doval, 1994].

Les approches décrites jusqu'à présent sont des approches déterministes. Elles peuvent toutes être reconsidérées avec un point de vue probabiliste et ainsi profiter d'un formalisme bien éprouvé. Les travaux de thèse de Doval [Doval, 1994] aboutissent à un AEP probabiliste. L'auteur modélise le spectre d'amplitude d'une trame en un ensemble de partiels en relation harmonique et un ensemble de partiels en relation non harmoniques. Chaque partiel est défini par sa position fréquentielle et par son amplitude qui sont considérés comme étant des variables aléatoires. Cela revient à réduire la trame à une somme de sinusoïdes harmoniques ce qui est en lien avec le modèle sinusoïdal harmonique tout en ajoutant des sinusoïdes ne faisant pas partie de la structure harmonique. L'auteur considère que le nombre d'harmoniques d'une structure harmonique est variable, que l'amplitude des harmoniques est globalement décroissante en fonction du numéro de l'harmonique, que l'enveloppe de l'amplitude des harmoniques est régulière, que les premières harmoniques sont marquées et rarement absentes, que

le nombre de partiels non harmoniques est très variable et dépend du bruit, et que l'amplitude des partiels non harmoniques est plutôt faible par rapport à celle des harmoniques. Il modélise toutes ces connaissances par l'intermédiaire de densités de probabilités comme l'écart relatif entre un partiel et la valeur théorique de l'harmonique correspondant, la probabilité de la présence d'une harmonique en fonction de son numéro ou bien la probabilité du nombre de partiels non harmoniques en fonction du nombre de partiels non harmoniques. Ces densités de probabilités sont apprises lors d'une étape d'apprentissage avec une base de données. Une fois les paramètres appris, une autre base de données peut être utilisée pour estimer la F_0.

La méthode construit une fonction nommée AHMV (Appariement d'Harmoniques selon le Maximum de Vraisemblance) qui permet d'attribuer à une fréquence donnée la probabilité qu'elle a d'être la F_0 du spectre étudié. La figure 3.30 présente un exemple d'AHMV obtenue sur un signal synthétique de F_0 valant 200Hz. Le maximum local de l'AHMV est bien de 200Hz et l'estimation est donc correcte. Il faut néanmoins noter la présence d'un maximum local non négligeable en 100Hz et en 66.6Hz qui correspondent à la sous-octave et au tiers de la F_0. Un autre phénomène intéressant concerne la décroissance générale de l'AHMV à gauche et à droite de la F_0. Les approches statistiques ne seront pas discutée plus en profondeur dans ce manuscrit. Le lecteur intéressé pourra lire l'ouvrage de Christensen & Jakobsson [Christensen and Jakobsson, 2009] et l'article de Godsill & Davy concernant des sons musicaux [Godsill and Davy, 2002]. Les AEP probabilistes s'appuient généralement sur le modèle sinusoïdal harmonique de McAulay & Quatieri [McAulay and Quatieri, 1986, 1990] utilisé pour construire le signal exemple du chapitre présenté figure 3.6. Elles ont pour grande force de passer par une étape d'apprentissage qui permet d'adapter l'estimation de F_0 aux signaux étudiés. Néanmoins, le fait de considérer le problème de l'estimation d'un point de vue probabiliste ne change pas radicalement le problème. Comme dans les méthodes déterministes, l'estimation de F_0 est basée sur une fonction de pitch qui quantifie la probabilité qu'une fréquence soit la F_0.

3.3.5 Erreurs des AEP monopitch et solutions adoptées

Quelque soit l'approche, les AEP monopitch font typiquement des erreurs d'octave et de sous-octave. L'observation des figures 3.8, 3.10 pour les approches temporelles, des figures 3.14 (a), 3.17, 3.19 (k), 3.20, 3.24 pour les approches fréquentielles, de la figure 3.27 pour les approches spectro-temporelles et de la figure 3.30 pour les approches probabilistes montre que les fonctions de pitch possèdent toutes des pics (ou creux) parasites qui peuvent selon les situations être plus forts (ou faibles) que les pics (ou creux) correspondant à la F_0. Pour remédier aux erreurs locales d'estimation de F_0, plusieurs solutions ont été envisagées. Ces solutions sont soit des post-traitements soit l'intégration du processus d'estimation dans un processus de suivi de F_0 ce qui a pour effet de rendre la distinction entre estimation et suivi moins claire. L'un des post-traitements les plus courants consiste en un filtrage médian des valeurs de F_0 estimées. Étant donné que les erreurs sont souvent des sauts d'octave ou de sous-octave, un lissage médian est parfaitement adapté pour supprimer des erreurs ponctuelles. En revanche, si les erreurs sont présentes sur un trop grand nombre de trames adjacentes, ce type de filtrage ne pourra plus les corriger. L'autre manière de corriger les erreurs d'estimation consiste à intégrer l'estimation dans un processus de suivi de F_0 ou de lissage de trajectoire de F_0 pour éviter les discontinuités et les sauts

d'octave. Par exemple, l'algorithme standard d'estimation de F_0 de PRAAT [Boersma, 1993] procède à une programmation dynamique (PD) qui permet de lisser la trajectoire de F_0. Cette PD revient à intégrer des informations temporelles dans l'estimation de la F_0. Une programmation dynamique est également présente dans la méthode RAPT [Talkin, 1995]. Le suivi d'une F_0 par dans un mélange de parole par PD a fait l'objet d'une publication [Signol, 2007]. La brosse spectrale proposée par Martin [Martin, 2000] permet également d'intégrer la continuité locale de F_0 dans l'estimation de F_0 et d'éviter les erreurs. Les travaux de Salor et al. procèdent au suivi de F_0 par filtrage de Kalman [Salor et al., 2006]. Doval [Doval, 1994] et son approche probabiliste inclut également son étape estimation dans des Modèles de Markov Cachés (*Hidden Markov Model* ou HMM en anglais) ce qui lui permet de faire du suivi de F_0 par l'intermédiaire de l'algorithme de Viterbi (recherche de meilleur chemin). Le principal inconvénient de ces méthodes se trouve dans les transitions entre trames voisées et trames non voisées. Autant le coût de passage entre deux candidats F_0 peut être modélisé par l'écart relatif entre ces deux valeurs autant il n'est pas aussi évident de donner un coût de passage entre trame voisée et trame non voisée. Quel critère peut être utilisé ? A ce bas niveau de traitement qu'est l'estimation de F_0, nous souhaiterions ne pas avoir à intégrer d'information temporelle. C'est pourquoi nous proposons d'intégrer directement un mécanisme de suppression des pics parasites dans la construction des fonctions de pitch. Si les pics parasites des fonctions de pitch sont d'amplitude très faibles par rapport aux amplitudes des F_0 à estimer, il est raisonnable de penser que l'estimation de F_0 même multiples sera grandement facilitée et moins sujette aux erreurs. Ainsi, l'AEP que nous proposons est strictement trame-à-trame, ne fait pas appel à une étape de suivi et nous retrouvons cette claire distinction entre estimation de F_0 et suivi de F_0.

La section 3.4 présente l'existant en matière d'estimation multipitch. Le fait de passer d'un problème monopitch à un problème multipitch ajoute de nouvelles problématiques et une simple transposition des méthodes monopitch n'est pas suffisant.

3.4 Problème multipitch d'estimation de F_0

Le passage à la situation multipitch fait apparaître des problèmes nouveaux et ne peut se réduire à une transposition de la situation monopitch. La section 3.4.1 présente deux de ces problèmes : l'estimation du nombre de sources mélangées dans le signal cf. 3.4.1.1 et les harmoniques communes cf. 3.4.1.2. L'estimation de F_0 multiples repose toujours sur la construction d'une fonction de pitch. Néanmoins, l'exploitation de cette fonction doit être repensée. Deux façons d'exploiter les fonctions de pitch existent dans la littérature. La première se nomme estimation itérative et est décrite dans le paragraphe 3.4.2. La seconde est nommée estimation conjointe et est décrite dans le paragraphe 3.4.3. Les paragraphes 3.4.4 et 3.4.5 ont pour objectif de détailler deux AEP multipitch auxquels nous comparerons nos propres résultats dans le chapitre 6. Les avantages et inconvénients de chacun des algorithmes seront discutés ce qui permettra de justifier nos choix quant à la conception de notre AEP.

3.4.1 Problématiques apportées par la situation multipitch

L'estimation de parole superposée ajoute deux problèmes qui n'existaient pas ou du moins n'étaient pas visible dans la situation monopitch. Le premier problème concerne la connaissance du nombre de sources mélangées dans le signal. En effet, si l'AEP connaît le nombre de F_0 à estimer, l'estimation est plus facile. Le second problème concerne l'effet de l'addition de deux sons voisés et son impact sur la représentation fréquentielle du mélange.

3.4.1.1 Estimation du nombre de sources

L'estimation du nombre de sources est un problème lié à l'estimation de F_0 en situation multipitch. Le fait de connaître le nombre de sources du mélange aide à l'estimation des F_0 puisque cette connaissance permet d'arrêter l'AEP lorsqu'il a estimé le nombre adéquat de F_0. Le contraire est aussi valable puisque le fait d'estimer un certain nombre de F_0 renseigne sur le nombre de sources présentes dans le mélange sonore étudié. Ce problème précis étant hors du cadre de la thèse, il ne sera pas plus abordé mais le lecteur intéressé pourra se référer aux articles [Jong-Shiann and Ingram, 2004, Olsson and Hansen, 2004] et à la thèse de Yeh [Yeh, 2008].

3.4.1.2 Harmoniques communes

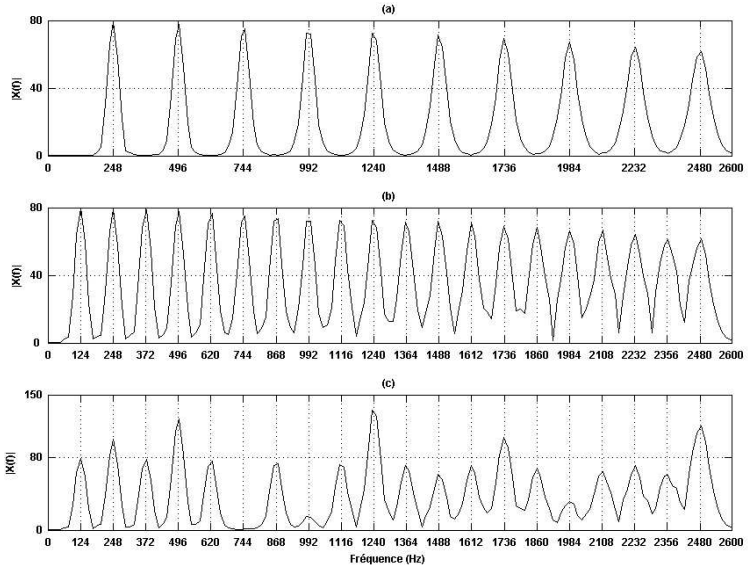

FIGURE 3.31: Effet des harmoniques communes. (a) Spectre d'amplitude d'une trame exemple de F_0 248Hz. (b) Spectre d'amplitude d'une trame exemple de F_0 124Hz. (c) Spectre d'amplitude de la somme temporelle de (a) et (b).

La figure 3.31 présente le problème des harmoniques communes sur le signal exemple que nous avons utilisé tout au long de ce chapitre. Une modification a été apporté au signal exemple et il n'y a plus de décroissance de l'amplitude des dents. Le graphique (a) présente la structure harmonique de 10 harmoniques de la première trame du signal exemple de F_0 valant 248Hz. Le graphique (b) présente le spectre d'amplitude de la première trame d'un signal exemple dont la F_0 est moitié soit 124Hz. La structure harmonique comporte 20 dents pour que l'étendue fréquentielle des deux structures soit identique. Le graphique (c) montre le spectre d'amplitude de la trame résultat de l'addition des deux trames en (a) et (b). La première chose à constater est la présence d'harmoniques communes aux fréquences multiples de 248Hz. Que se passe-t-il aux fréquences des harmoniques communes sur le spectre d'amplitude ? Le spectre d'amplitude d'un signal mélangé n'est pas la somme des spectres

d'amplitude des signaux composant le mélange. La phase intervient dans le calcul de l'amplitude des harmoniques communes. Or, il est utile de rappeler que pour chacune des harmoniques, la phase est choisie aléatoirement entre $-\pi$ et π. Dans notre exemple, pour une harmonique commune donnée, si les phases des harmoniques du signal (a) et du signal (b) sont en opposition comme pour la fréquence 744Hz alors l'harmonique du mélange disparaît. En revanche, si les phases d'une harmonique commune sont égales à 2π près, l'amplitude de l'harmonique du mélange vaudra la somme des amplitudes des deux harmoniques des signaux isolés. Cet effet des phases aux harmoniques communes est critique car il peut casser la structure harmonique du mélange en supprimant des harmoniques ce qui est néfaste à l'estimation de F_0. A l'inverse, lorsque l'effet de phase est additif sur l'amplitude de l'harmonique, la régularité naturelle de l'enveloppe spectrale des harmoniques d'un signal isolé est rompu ce qui peut perturber l'attribution de telle ou telle harmonique à telle ou telle valeur de F_0. Quelques travaux existent concernant la détection des harmoniques communes et l'estimation des phases et des amplitudes originales [Parsons, 1976, Woodruff et al., 2008, Yeh, 2008].

Les problèmes d'harmoniques communes ou de l'estimation du nombre de sources vont se retrouver dans les méthodes d'estimation de F_0 qu'elles soient itératives ou conjointes. La section présente les approches d'estimation de F_0 itératives. Une bonne explication des problèmes et des approches de l'estimation de F_0 multipitch peut être trouvée dans [de Cheveigné and Kawahara, 1999].

3.4.2 Estimation itérative

Les méthodes d'estimation de F_0 itératives ont pour principe d'estimer la F_0 de la source dominante du mélange sonore, de supprimer cette source dominante du mélange puis de réitérer l'opération jusqu'à ce que toutes les F_0 soient estimées.

L'estimation de la F_0 de la voix dominante a été étudiée sur de la parole ou de la musique. Pour la parole, les travaux de Martin [Martin, 2008a,b] apportent une solution à l'estimation de la F_0 d'une voix dominante. La méthode consiste à considérer un spectrogramme dans son intégralité et à utiliser le principe de brosse spectrale aussi décrite dans [Martin, 2000]. L'évolution temporelle de la F_0 étant soumise à des contraintes physiques, elle est continue (de l'ordre de 1% par ms pour une voix « normale »). Une structure harmonique évolue peu entre deux trames successives suffisamment proches et il est donc possible d'essayer d'aligner les structures harmoniques de deux trames. Si cet alignement respecte des contraintes de continuité, la structure harmonique est détectée et il est possible d'en estimer la F_0. L'idée de s'appuyer sur la structure complète est astucieuse car elle permet de limiter l'effet des harmoniques manquantes ou faibles. Lorsque deux structures harmoniques sont mélangées, l'algorithme proposé suit la structure de plus forte énergie, donc celle dominante. Pour la voix chantée, les travaux de Durrieu, Richard & David [Durrieu et al., 2008] proposent d'extraire la mélodie d'un chanteur mais l'estimation de F_0 n'apparait pas explicitement. Elle apparaît néanmoins implicitement dans le modèle utilisé pour décrire le chanteur. Il est possible d'imaginer un système couplant un algorithme d'estimation de F_0 utilisé pour apprendre le modèle de voix d'un chanteur et l'approche BSS proposée pour déterminer la mélodie chantée dans un mélange musical. Sur des signaux musicaux, Ils sont notamment motivés par l'application d'estimation de mélodie [Goto, 2000, 2001]. Ces travaux de suivi de F_0 dominante sont très utiles car ils peuvent justement servir à l'étape d'estimation de F_0 dans

les AEP multipitch itératifs.

Les travaux de Parsons [Parsons, 1976] proposent un algorithme de séparation de parole voisée comportant une étape d'estimation de F_0 bipitch itérative. L'AEP est purement fréquentiel. Les pics du spectre d'amplitude sont considérés comme des partiels potentiels d'une structure harmonique. L'algorithme passe par une étape de détection des harmoniques communes en utilisant la symétrie des partiels (cf. les travaux de Yeh & Roebel [Yeh and Röbel, 2004] sur le même sujet). Un histogramme de Schroeder est calculé à partir des positions fréquentielles et le maximum de l'histogramme est considéré comme la première F_0. Les partiels appartenant à la structure harmonique de cette première F_0 ne sont plus considérés et un deuxième histogramme est calculé. Le maximum sera la deuxième F_0.

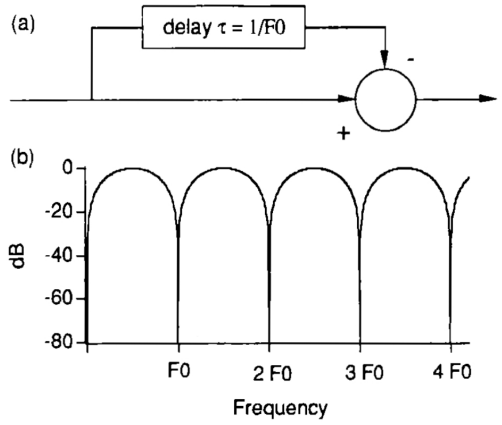

FIGURE 3.32: Exemple de filtrage en peigne par annulation. (a) Représentation temporelle. (b) Représentation fréquentielle. Source : [de Cheveigné, 1993].

Un AEP multipitch conçu pour séparer de la voix voisée mélangée est donné dans [de Cheveigné, 1993]. Cet algorithme se nomme DDF et sa version itérative consiste à calculer le pitch dominant le mélange en utilisant une AMDF. Une fois la F_0 dominante estimée, la structure harmonique de cette F_0 est annulée par un filtre en peigne illustré en figure 3.32 (b). Ces filtres en peigne semblent avoir une plausibilité biologique [de Cheveigné, 1998]. Par cette méthode, l'auteur obtient un taux d'erreur à 10% de 3.1% sur des mélanges bipitch de parole réelle (34.5 secondes environ).

Les travaux de [de Cheveigné and Kawahara, 1999] proposent une amélioration de l'algorithme DDF et utilisent la SDF au lieu de l'AMDF (cf sous-paragraphe 3.3.1.3). L'étape de suppression est effectuée par un peigne spectral de la forme de la figure 3.32. Les performances obtenues sur environ 12 minutes de mélange bipitch synthétique sont de 8.42% d'erreur. Les signaux synthétiques mélangés sont parfaitement périodiques (sinus complexes de même intensité possédant 10 harmoniques), ne sont jamais à l'octave ni égaux en termes de F_0. Les auteurs notent le fait intéressant que lorsque un des deux signaux domine l'autre en intensité, leur méthode fonctionne nettement mieux. Cela tend à montrer l'efficacité des algorithmes itératifs lorsque les signaux mélangés sont d'intensités hétérogènes.

Les travaux de Klapuri et al. [Klapuri et al., 2000, Klapuri, 2003, 2006] proposent un AEP multipitch itératif. Le spectre d'amplitude d'une trame est calculé et subit une étape de débruitage. Le spectre est alors découpé en 18 bandes de fréquences distribuées de manière logarithmique. Chaque bande occupe 2/3 d'octave environ de 50 à 6kHz. Dans chacun des canaux, l'équivalent d'une fonction de pitch par peigne de Martin est calculée. Les auteurs fusionnent ensuite les données de chacun des canaux pour calculer la F_0 dominante du signal. Le processus itératifs commence alors et les harmoniques appartenant à la structure harmonique correspondant à la F_0 sont supprimées en respectant une contrainte de régularité spectrale. Cet AEP est utilisé dans un algorithme de séparation décrit dans [Virtanen and Klapuri, 2001]. L'article de Klapuri [Klapuri, 2005] propose un AEP dont les premiers étages de traitement modélisent plus finement ce qui peut se passer dans l'oreille.

Les méthodes d'estimation de F_0 itératives itèrent le processus (estimation de F_0 dominante, suppression) jusqu'à ce que le nombre de candidats F_0 connu ou souhaité soit atteint ou que le résidu soit trop faible. Elles ont démontré qu'elles fonctionnent mais en pratique les étapes de suppression sont compliquées (ie. harmoniques communes) et la condition d'arrêt est complexe à déterminer (ie. estimation du nombre de sources).

3.4.3 Estimation conjointe

Les AEP multipitch par estimation conjointe cherchent la combinaison de F_0 qui correspond le mieux au signal observé. Le nombre de candidats N à estimer est décidé en fonction de la connaissance du nombre de source du mélange, de l'estimation du nombre de sources ou bien peut être un choix arbitraire. En considérant le problème avec l'utilisation d'une fonction de pitch, l'estimation de F_0 conjointe consiste à considérer les N premiers maxima locaux comme les candidats F_0.

Du point de vue de la parole, deux AEP multipitch par estimation conjointesont décrits précisément dans les paragraphes 3.4.4 et 3.4.5. Ils sont tous deux spectro-temporels et conçus pour les signaux bipitch uniquement. Ils cherchent soit les canaux pour lesquels l'une ou l'autre des F_0 est isolée et ressort clairement (WWB), soit ils cherchent à estimer directement les 2 F_0 sur les canaux (signal à bande limitée) VEW.

Dans l'article [de Cheveigné, 1993], l'auteur utilise DDF et le filtrage en peigne de la figure 3.32 de manière conjointe. Il recherche la combinaison de filtres en peigne mis en série qui minimise le signal de sortie. Cette méthode est principalement limitée par la connaissance du nombre de sources et donc de filtres en cascade à utiliser. De plus, une recherche exhaustive des combinaisons de filtres doit être faite ce qui peut être assez lourd en termes de calcul. L'auteur montre que le taux d'erreur fait avec cette méthode est de 1.7% sur des signaux bipitch réels (34.5 secondes).

Les travaux de de Cheveigné & Kawahara [de Cheveigné and Kawahara, 1999] déjà décrits dans le paragraphe 3.4.2 démontrent qu'avec des signaux parfaitement périodiques, en évitant soigneusement les cas de mélange à l'octave ou d'égalité de F_0, une approche d'estimation conjointe ne fait pas d'erreur. Ils s'appuient pour cela sur une évaluation portant sur environ 12 minutes de mélange bipitch et sur environ 2h de signaux tripitch.

Le point de vue probabiliste de l'estimation de F_0 conjointe est abordé dans [Bach and Jordan, 2005] en passant par la modélisation du spectre d'un signal voisé comme un peigne fréquentiel dont les partiels sont en relation harmonique. Les auteurs imposent une certaine contrainte de continuité sur l'enveloppe spectrale. Ils intègrent leur modèle dans une chaine de Markov cachée. Les paramètres du modèle sont appris au préalable. L'algorithme proposé est conçu pour traiter de la parole superposée bi-locuteur.

Les travaux de Kameoka et al. [Kameoka et al., 2004a,c] proposent un AEP multipitch par estimation conjointe basé sur l'algorithme EM (*Expectation/Maximisation*). Les auteurs modélisent chaque source voisée comme un ensemble de partiels en relation harmonique, ces partiels étant des gaussiennes. Chaque source est donc une mixture de gaussienne et les auteurs recherchent, par un algorithme d'EM et une étape d'estimation du nombre de sources, les paramètres de leur modèle qui maximisent la ressemblance avec le spectre observé. Cette méthode est générale et peut s'appliquer à la parole ou à la

musique. Le travail de Kameoka et al. décrit dans [Kameoka et al., 2004b] utilise le modélise développé précédemment (mélange de gaussiennes). Ils l'intègrent dans un processus de filtrage de Kalman ce qui permet aux auteurs de construire les trajectoires des F_0 contenues dans le mélange. Les exemples donnés sont pris sur de la parole superposée et l'algorithme semble assez efficace.

Les signaux musicaux, de nature plus polyphonique, ont été largement étudiés. Izmirli & Bilgen dans [Izmirli and Bilgen, 1996] détectent un certain nombre de partiels dans le spectre d'amplitude, chacun de ces partiels étant considérés comme une F_0 potentielle. Les auteurs cherchent alors parmi l'ensemble des partiels détectés ceux qui sont en relation harmonique. Une fonction de coût quantifie l'écart d'un partiel à la valeur théorique attendue et la force totale d'une structure harmonique est une somme pondérée des couts de chaque partiel (pondérée par le numéro du partiel). Ceci revient à une utilisation implicite d'un peigne spectral. Les auteurs considèrent les structures harmoniques les plus fortes et les F_0 de ces structures sont les F_0 estimées.

Les travaux de Tolonen & Karjalainen [Karjalainen and Tolonen, 1999, Tolonen and Karjalainen, 2000] sont des méthodes spectro-temporelles qui s'appuient sur une forme d'autocorrélation cumulée calculée sur deux canaux (un entre 0 et 1kHz et l'autre au-dessus de 1kHz). Hormis le fait que leur découpage spectro-temporel soit différent d'une séparation classique en banc de filtre de type gammatone, le point original de leur méthode,a près calcul de la fonction SACF (cumul des transformées de Fourier inverse de la racine cubique de la densité spectrale de puissance des deux canaux) est de retirer à la SACF une version dilatée temporellement par un facteur 2 de la SACF, puis par 2, etc. Cette opération permet aux auteurs de réduire les pics parasites typiques des ACF cumulés (cf. figure 3.27 situés aux lags multiples des F_0 à estimer. Les auteurs montrent le bon fonctionnement de leur méthode sur des mélanges de trois notes de clarinette et sur un mélange de voyelles.

Les travaux de Fernandez-Cid & Casjus-Quiros [Fernandez-Cid and Casajus-Quiros, 1998], de Pertusa & Iñesta [Pertusa and nesta, 2008] et de Emiya [Emiya, 2008] proposent des AEP mulitpitch conçus pour traiter des signaux musicaux polyphoniques. L'estimation de F_0 est utilisée en vue de transcription automatique. Dans les travaux de Fernandez-Cid & Casajus-Quiros, les auteurs développent une approche spectro-temporelle qui analyse la forme des pics spectraux (amplitude et phase). L'algorithme détermine un certain nombre de pics comme candidats partiels et seuls ceux répondant à certaines contraintes telles que la ressemblance avec la forme du spectre de la fenêtre ou la continuité spectro-temporelle sont conservés. Les auteurs obtiennent ainsi une liste de partiels dans laquelle ils recherchent des relations harmoniques (on retrouve l'utilisation implicite d'un peigne spectral). Les auteurs utilisent des critères perceptifs et des connaissances sur les sources pour éliminer les erreurs d'estimation faites en trame-à-trame. Ils ajoutent un post-traitement intégrant les informations trame-à-trame sur 150ms afin de supprimer les notes trop courtes, les trous ou les erreurs d'octave. L'AEP multipitch de Pertusa & Iñesta calcule le spectre d'amplitude pour chacune des trames et un certain nombre de candidats F_0 y sont extraits. Toutes les combinaisons possibles entre ces candidats sont calculées. La meilleure combinaison en terme de régularité de l'enveloppe spectrale (*spectral smoothness* en anglais) et de somme des amplitudes des harmoniques est sélectionnée par un processus trame-à-trame. Le travail de thèse de Emiya propose un AEP multipitch fondée sur un modèle sinusoïdal et sur la régularité spectrale. L'auteur estime le bruit de fond d'un spectre par un modèle MA et estime l'enveloppe spectrale d'une structure harmonique par un modèle AR. Des contraintes de régularité

spectrale lui permette de proposer des candidats F_0 et l'estimation finale est intégrée dans une étape de suivi de F_0 réalisée à partir d'une chaîne de Markov Cachée et de l'algorithme de recherche du meilleur chemin de Viterbi. Ce suivi de F_0 multiples intègre une étape de détection des onsets des notes comme dans les travaux de Duan et al. [Zhiyao et al., 2007] qui proposent un estimateur multipitch ajoutant la notion d'événement temporel (début/fin, onset/offset d'une note). La prise en compte de l'inharmonicité, notamment pour le piano, est intégré au système pour améliorer la détection des partiels appartenant à une note.

Les travaux de Yeh [Yeh, 2008] présentent un algorithme d'estimation multipitch. Le système élaboré passe par une étape d'estimation de niveau de bruit servant à l'estimation des partiels. L'amplitude des partiels est estimée en détectant les éventuels harmoniques communes. Chaque ensemble de partiels considéré est associé à une fonction de score intégrant l'harmonicité des partiels (utilisation implicite d'un peigne spectral), la continuité temporelle de l'enveloppe spectrale (spectral smoothness), la synchronie temporelle d'évolution des partiels, etc. Une étape de suivi de F_0 multiples décrite dans [Chang et al., 2008] utilise les F_0 multiples estimées au niveau trame-à-trame et les intègre dans un HMM pour réaliser le suivi des trajectoires F_0 multiples.

Les AEP multipitch musicaux récents se caractérisent très largement par l'intégration de la contrainte de continuité temporelle des enveloppes spectrales (spectral smoothing). C'est le cas notamment des travaux de Le Roux, Kameoka, Ono, de Cheveigné & Sagayama [Roux et al., 2007b] qui complètent les travaux de Kameoka et al. précédemment décrits pour la parole en intégrant dans leur modèle GMM une contrainte de continuité temporelle de l'enveloppe spectrale. L'enveloppe spectrale doit être relativement lisse et est modélisée par une spline (polynôme). Cela permet aux auteurs non seulement d'être capable d'estimer de multiples F_0 [Roux et al., 2007a] mais aussi d'intégrer cela dans un analyseur multipitch capable de donner les trajectoires des F_0 d'une pièce musicale [Kameoka et al., 2007].

Les approches par estimation conjointe souffrent généralement d'une charge calculatoire plus importante qu'une approche itérative. En revanche, les performances d'une méthode par estimation conjointe semblent être meilleures qu'en utilisant la même méthode de manière itérative [de Cheveigné, 1993]. Deux algorithmes d'estimation de F_0 auxquels nous comparerons nos résultats par la suite vont être détaillés dans les paragraphes 3.4.4 et 3.4.5. Ces deux méthodes sont des méthodes d'estimation conjointe tout comme l'AEP que nous proposons dans cette thèse.

3.4.4 Algorithme de Wu, Wang & Brown (WWB)

L'AEP multipitch proposé par Wu, Wang & Brown [Wu et al., 2003] est construit sur un modèle cochléaire et sur un corrélogramme déjà décrits dans le paragraphe 3.3.3. Cet algorithme sera dénommé WWB dans le reste du manuscrit. WWB est une méthode par estimation conjointe dont le schéma bloc est donné figure 3.33. Les auteurs séparent le traitement des canaux basse-fréquence (80-800Hz) de celui des canaux haute-fréquence (800-5kHz) en s'inspirant des travaux de Rouat et al. [Rouat et al., 1997]. L'étape nommée « *Channel/Peak Selection* » du schéma bloc est le résultat de l'analyse du comportement des ACF d'un corrélogramme. Les auteurs remarquent que lorsque qu'un canal contenant des harmoniques est bruité par une autre source, le pic de l'ACF correspondant à la F_0 est de moindre amplitude que le pic de l'ACF du même canal non bruité. Si la valeur maximale du pic de

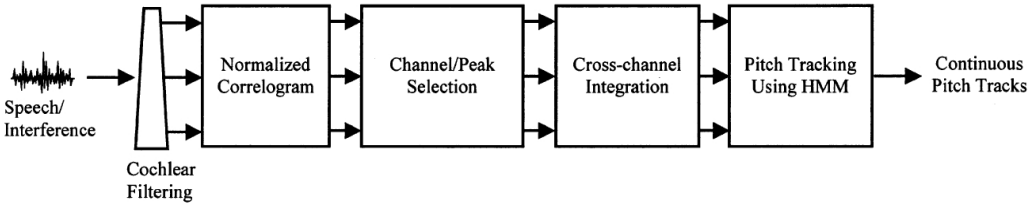

FIGURE 3.33: Schéma bloc de la méthode WWB. Source : [Wu et al., 2003].

plus grande amplitude ne dépasse pas un seuil (0.945 dans l'article), alors le canal est considéré comme bruité et sera rejeté. Ils calculent le corrélogramme sur des trames de 16ms et de 32ms et comparent les ACF obtenues pour chacun des canaux. Ils remarquent également que plus un canal est bruité, moins l'ACF du corrélogramme de 16ms ressemble à celle du corrélogramme de 32ms.

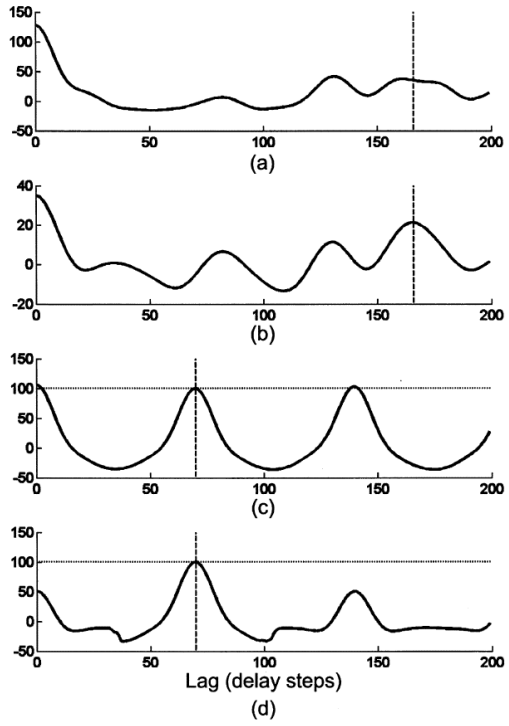

FIGURE 3.34: Comportements de la CC dans des situations de mélange. (a) CC sans sélection de canaux. (b) CC avec sélection des canaux peu bruités. (c) CC avec sélection des canaux mais sans sélection des pics. (d) CC avec sélection des canaux et sélection des pics. Source : [Wu et al., 2003].

Cette contrainte permet aux auteurs de rejeter les canaux dont la position du pic sur l'ACF de 16ms est décalé de plus d'un certain lag du pic le plus proche de l'ACF de 32ms (2 échantillons dans l'article). Les auteurs procèdent également à une sélection des pics de chacune des ACF de chacun des canaux. La sélection de pics est faite en considérant deux principes. L'ACF d'une trame voisée donne une série de pics aux multiples du lag de la F_0. Un pic au lag correspondant à F_0 sera donc accompagné d'un autre pic au lag double. Si cela n'est pas le cas, alors le pic est supprimé de l'ACF. L'écart accepté dans l'article est de 5 échantillons. La suppression consiste à mettre à 0 le pic jusqu'aux deux minima locaux adjacents. Le second principe concerne les canaux haute-fréquence qui sont construit à partir de l'enveloppe du signal de sortie. Si l'amplitude du premier pic de l'ACF de l'enveloppe est supérieur à un seuil (0.6 dans l'article), tous les pics aux multiples du lag sont supprimés. Selon les auteurs, cette deuxième manière de sélectionner les pics est critique dans la suppression des erreurs multiples et sous-multiples de la F_0. Une fois les sélections de canaux et de pics opérées, l'étape nommée « Cross-Channel Integration » peut avoir lieu. Elle consiste à additionner ensemble les ACF des canaux considérés comme non bruités, où sont conservés les pics sélectionnés. Cela permet de construire le corrélogramme cumulé ou CC. L'effet de la sélection des canaux sur le CC est décrite dans l'exemple de la figure 3.34 graphiques (a) et (b). Le signal utilisé est une trame de parole voisée à laquelle est ajouté un bruit blanc gaussien. La position théorique du lag de la F_0 est donnée par le trait vertical pointillé. Le pic avec la sélection des canaux est bien plus marqué que sans.

L'effet de la sélection des pics sur le CC est illustré dans les graphiques (c) et (d) de la figure 3.34. La sélection des pics permet de réduire de moitié le pic parasite au lag double du lag de la F_0. Dans WWB, l'estimation de F_0 est incluse dans un processus de suivi de F_0 réalisé à partir d'une modélisation par HMM et avec un algorithme de Viterbi. L'étape de suivi de F_0 est très importante dans WWB. Le système est limité à des mélanges de deux signaux. Les auteurs se comparent aux AEP proposés dans [Gu and van Bokhoven, 1991, Tolonen and Karjalainen, 2000] et leur AEP donne de meilleurs résultats. C'est pourquoi nous souhaitons comparer notre AEP à WWB. Il existe toutefois des limitations à l'algorithme WWB. C'est un système d'emblée bipitch. De quelle manière peut-il être étendu à des situations multipitch dans lesquelles plus de deux F_0 sont présents ? D'autre part, WWB passe par une étape d'apprentissage pour calculer les paramètres qui servent à réaliser son suivi de F_0. Notre AEP se démarque donc de WWB puisqu'il n'utilise pas d'étape d'apprentissage ni de suivi de F_0.

3.4.5 AEP de Vishnubhotla & Espy-Wilson (VEW)

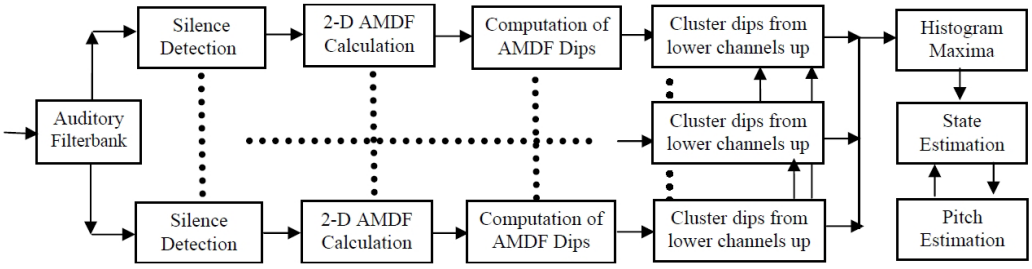

FIGURE 3.35: Schéma bloc de la méthode VEW. Source : [Vishnubhotla and Espy-Wilson, 2008].

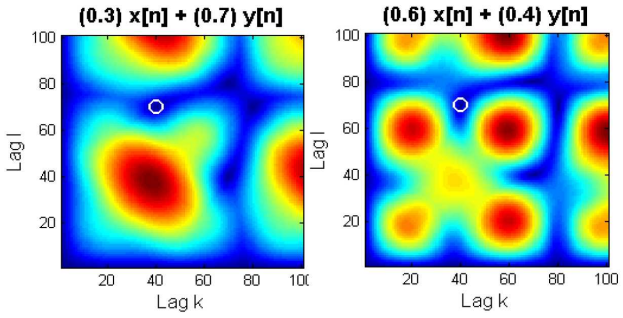

FIGURE 3.36: Fonction de pitch de la méthode VEW. Les ronds blancs sont les couples estimés de F_0. Les pixels bleus sont de faible amplitude et les pixels rouges sont de forte amplitude. Source : [Vishnubhotla and Espy-Wilson, 2008].

L'AEP multipitch proposé par Vishnubhotla & Espy-Wilson dans [Vishnubhotla and Espy-Wilson, 2008] repose sur l'AMDF décrite dans le sous-paragraphe 3.3.1.3 équation 3.7. Le principe est identique à celui de l'algorithme DDF proposé par de Cheveigné dans [de Cheveigné, 1993]. Cet algorithme sera nommé VEW dans le reste du manuscrit. Le schéma bloc de la méthode est donné dans la figure 3.35. VEW est un AEP multipitch par estimation conjointe comme WWB. VEW repose aussi sur une séparation du signal par un banc de filtres. Dans chacun des canaux, une étape de détection de silence est d'abord appliquée pour supprimer d'emblée ces zones temporelles de l'estimation de F_0. Au lieu de rechercher un seul lag pour lequel l'AMDF soit minimale, la méthode généralise la recherche à un couple de lags (k, l) qui minimise la fonction de pitch

donnée par l'équation 3.20. La fonction de pitch calculée par VEW n'est plus une courbe 1-D mais une image 2-D. Les auteurs nomment cette fonction de pitch l'AMDF 2-D. Elle sera notée AMDF2 comme dans l'équation 3.20.

$$AMDF2(k,l) = \sum_{m=0}^{N-1} |x(n+m) - x(n+m-k) - x(n+m-l) - x(n+m-k-l)| \tag{3.20}$$

La figure 3.36 présente le résultat de l'AMDF 2-D sur le mélange de deux signaux x et y qui sont respectivement des sinusoïdes pures l'une de F_0 à 40 échantillons et l'autre de F_0 à 70 échantillons.

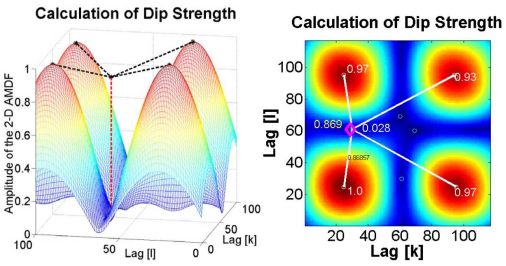

FIGURE 3.37: Méthode d'interpolation de VEW. Source : [Vishnubhotla and Espy-Wilson, 2008].

Deux mélanges sont réalisés avec des pondérations différentes sur x et y. Les cercles blancs sur chacune des images représentent les minima globaux de l'AMDF pour les deux mélanges et indiquent que les F_0 sont correctement estimés. Une AMDF 2-D est calculée pour chacun des canaux. Les auteurs recalculent ensuite les AMDF 2-D de chacun des canaux en considérant que l'AMDF 2-D final du canal numéro c est égal à la somme des AMDF 2-D des canaux inférieurs et de celui considéré c'est à dire la somme des AMDF 2-D de 1 à c. Pour améliorer la précision de la détection du minimum global de chacun des AMDF 2-D finaux, les auteurs passent par une interpolation qui s'appuie sur les quatre maxima locaux (zones en rouges sur la figure 3.37). Les couples (k,l) issus de l'estimation du maximum global de chacune des AMDF 2-D des tous les canaux sont intégrés dans un histogramme. C'est l'évolution temporelle des histogrammes calculés qui permet à VEW d'estimer la ou les F_0. Nous souhaitons nous comparer à cet algorithme car il reste d'approche majoritairement bottom-up même si l'intégration temporelle des histogrammes n'est pas clairement donnée. Le même inconvénient que WWB peut être soulevé. Comment VEW peut-il être étendu à l'estimation de plus de deux F_0. La seule solution est-elle de calculer une AMDF 3-D ? La charge de calcul risque de devenir rapidement rédhibitoire.

3.5 Conclusions

Dans ce chapitre, le matériau parole a été décrit ce qui a permis d'en montrer les caractéristiques (temporelles et fréquentielles) pouvant permettre l'estimation de F_0 et celles qui rendent cette estimation compliquée. Les différences des problématiques d'estimation de F_0 sur des signaux de parole superposée et sur des signaux musicaux sont aussi décrites. Un état de l'art de l'estimation de F_0 en situation monopitch a été développé. Les principales approches temporelles, fréquentielles et spectro-temporelles ont été détaillées. Le point de vue probabiliste a également été abordé. Il ressort de cet état de l'art que les méthodes font des erreurs et que ces erreurs sont généralement corrigées par des post-traitements (filtrage médian, programmation dynamique, etc.) qui ont pour point commun une forme de suivi de F_0. Un état de l'art de l'estimation de F_0 en situation multipitch a été développé. La

majorité des approches sont conçues pour des signaux musicaux. Quelques travaux concernent la parole superposée et deux algorithmes WWB et VEW sont détaillés. WWB et VEW sont des approches spectro-temporelles et les traitements effectués sur chacun des canaux sont purement temporels (ACF ou AMDF 2-D). Ces deux méthodes actuelles sont lourdes en coût de calcul. Les travaux concernant les AEP conçus pour traiter de la parole superposée sont encore assez peu développés. En particulier les approches strictement trame-à-trame et purement bottom-up. Il est intéressant d'aller voir du coté des approches purement fréquentielles puisqu'elle ne semble pas avoir été étudiées sur de la parole superposée. C'est pourquoi l'AEP que nous proposons est purement fréquentiel. Nous proposons d'utiliser des peignes spectraux, ce choix étant motivé par plusieurs raisons.

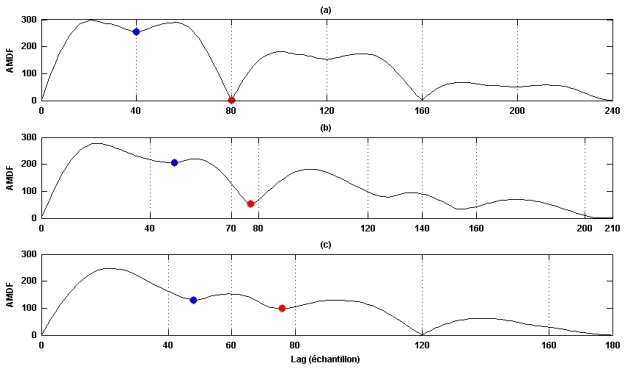

FIGURE 3.38: Exemples d'AMDF-1D obtenues sur trois mélanges de deux sinus purs de mêmes amplitudes. (a) $T_{01} = 40$, $T_{02} = 80$. (b) $T_{01} = 40$, $T_{02} = 70$. (c) $T_{01} = 40$, $T_{02} = 60$. Les points bleus sont les minima locaux correspondant à l'estimation de T_{01}. Les points rouges sont les minima locaux correspondant à l'estimation de T_{02}.

D'une part, la structure harmonique semble être une caractéristique robuste à l'estimation de F_0. Même dans une structure incomplète, il doit être possible de s'appuyer sur les harmoniques présentes pour estimer les F_0. Or, un peigne spectral permet de mettre en évidence les structures harmoniques de manière simple. D'autre part, les méthodes temporelles semblent ne pas se comporter correctement dans un mélange comme le suggère la figure 3.38. Elle présente le comportement de l'AMDF 1-D sur trois mélanges de deux sinus d'amplitudes identiques mais de périodes fondamentales différentes. La période fondamentale du premier sinus est notée T_{01} et celle du deuxième sinus est notée T_{02}. T_{01} est fixe pour les trois graphiques (a), (b) et (c) et vaut 40 échantillons. La durée du signal est fixée de manière à ce que le sinus de plus grande période fondamentale ait trois périodes. Dans le graphique (a), les points rouges et bleus coïncident avec les périodes fondamentales théoriques en 40 et 80. En revanche, la détection du minimum local en 40 n'est pas évidente car ce minimum local est d'amplitude bien supérieure à celui placé en 160. Dans les graphiques (b) et (c) les points bleus et rouges ne sont pas alignés avec les périodes fondamentales théoriques en 40 et 80. Même si l'algorithme parvient à déterminer les bons maxima locaux (les plus proches de ceux théoriques), ceux-ci sont décalés par rapport à la valeur théorique ce qui implique une erreur dans l'estimation de F_0. Les trois graphiques (a), (b) et (c) semblent indiquer qu'il est difficile d'interpréter correctement l'AMDF 1-D en tant que fonction de pitch dans une situation de mélange de parole. Soit les minima locaux sont difficilement détectables, soit ils sont décalés par rapport à la valeur théorique. Ces défauts nous amènent à privilégier une approche fréquentielle.

Le chapitre suivant présente une analyse de la pertinence et de la faisabilité de l'estimation conjointe de F_0multiple sur de la parole superposée par des peignes spectraux.

Chapitre 4

Peignes spectraux pour l'estimation conjointe de F_0 multiples

4.1 Introduction

L'objectif de ce chapitre est de démontrer la faisabilité d'un algorithme d'estimation conjointe de F_0 multiples traitant de la parole superposée et reposant sur l'utilisation de peignes spectraux. Le principe d'estimation conjointe de F_0 par un peigne fréquentiel quelconque est décrit dans le paragraphe 4.1.1. Les fonctions de pitch obtenues par plusieurs peignes spectraux dans la situation monopitch et multipitch vont être présentées. La situation multipitch sera symbolisée par la situation bipitch et par la situation dans laquelle quatre sons voisés sont mélangés. La situation de quatre sons voisés est nommée situation 4-pitch. Les trois situations monopitch, bipitch et 4-pitch sont modélisées par des spectres d'amplitude idéaux décrits dans le paragraphe 4.1.2.

La section 4.2 présente le fonctionnement du peigne spectral le plus élémentaire nommé Peigne Uniforme Infini (PUI) qui se résume à un simple échantillonneur. Les fonctions de pitch en situations monopitch et multipitch présentées sont obtenues sur les spectres d'amplitude idéaux. Les qualités et défauts du comportement de ces fonctions de pitch sont ensuite discutés. Les fonctions de pitch présentent des maxima locaux aux fréquences F_0 à estimer. Cette caractéristique est nécessaire pour l'estimation de F_0 (mais pas suffisante) si bien qu'il devrait être possible d'estimer la/les F_0 par l'utilisation de peignes fréquentiels. Néanmoins, ces maxima locaux sont entourés d'autres pics parasites, notamment dans la situation multipitch, et montrent que le PUI n'est certainement pas suffisant pour correctement estimer les F_0 multiples.

La section 4.3 présente plusieurs méthodes existantes permettant d'atténuer les pics parasites. Ces méthodes passent par l'utilisation d'autres peignes spectraux : le Peigne Uniforme Fini (PUF), le Peigne Décroissant Infini (PDI), le Peigne Décroissant Fini (PDF) et le Peigne ALterné (PAL) décrit dans [Liénard et al., 2007]. Ces peignes apportent effectivement une diminution des pics parasites mais cette diminution est soumise à un compromis inévitable. Une atténuation des pics parasites sur-harmoniques (supérieurs à la F_0) s'accompagne nécessairement d'une augmentation des pics parasites sous-harmoniques (inférieurs à la F_0) et vice et versa. Un des apports de cette thèse est de montrer que ce compromis, inévitable en utilisant un peigne spectral unique, peut disparaître en combinant

plusieurs familles de peignes spectraux.

La section 4.4 présente une nouvelle famille de peignes spectraux nommée Peigne à Dents Négatives (PDN). Ces peignes spectraux particuliers sont définis par leur ordre (entier naturel supérieur ou égal à 2) et par leur fréquence caractéristique. Ils permettent d'atténuer spécifiquement les pics parasites sur-harmoniques (au-dessus de la F_0) en fonction de l'ordre utilisé. La section 4.5 présente une autre famille de peignes spectraux nommée Peigne à Dents Manquantes (PDM) également définis par leur fréquence caractéristique et leur ordre. Ils permettent d'atténuer spécifiquement les pics parasites sous-harmoniques (au-dessous de la F_0) en fonction de l'ordre du peigne.

La section 4.6 présente le principe de PSH qui est à l'origine de l'AEP multipitch que nous proposons. La combinaison des résultats de l'ensemble des PDN et PDM permet en effet d'atténuer beaucoup de pics parasites et donc de fournir une fonction de pitch dont les maxima locaux de plus forte amplitude sont ceux des F_0 à estimer.

La section 4.7 tirera les conclusions de ce chapitre et amènera le lecteur vers un chapitre décrivant les implémentations possibles du principe de PSH.

4.1.1 Estimation conjointe de F_0 par un peigne spectral quelconque

De manière générale, la fonction de pitch d'une méthode d'estimation de F_0 à base de peigne fréquentiel est issue du produit scalaire entre un peigne fréquentiel noté $P(F_c, f)$ et le spectre d'amplitude du son étudié noté $|S(f)|$. Le peigne quelconque $P(F_c, f)$ est au moins défini par sa fréquence caractéristique notée F_c qui correspond à sa fréquence fondamentale. La fonction de pitch du PUI est la fonction traçant l'évolution du produit scalaire de $|S(f)|$ et du PUI en faisant varier sa fréquence F_c dans un intervalle de recherche de F_0 donné. L'équation 4.1 présente la valeur théorique de la fonction de pitch obtenue pour une fréquence F_c donnée.

$$\Phi(F_c) = \int_0^{F_{MAX}} |S(f)|.P(F_c, f)df \tag{4.1}$$

$\Phi(F_c)$ est la fonction de pitch. F_{MAX} est nommée fréquence de coupure et correspond à la fréquence maximale de calcul du produit scalaire entre le spectre $|S(f)|$ et un peigne fréquentiel $P(Fc, f)$ quelconque. F_{MAX} peut être réglé de manière arbitraire et peut, par exemple, valoir la moitié de fréquence d'échantillonnage F_e.

L'intervalle de recherche de la F_0 est un paramètre important et doit être choisi de manière adapté au signaux considérés. Cet intervalle est noté $[F_{0min}, F_{0max}]$ avec F_{0min} la fréquence minimale de recherche et F_{0max}.la valeur maximale. Tout le long du chapitre, F_{0min} est fixée à 50Hz et F_{0max} vaut 800Hz. Ce choix permet de couvrir quatre octaves ce qui très raisonnable pour contenir l'ensemble de la variation naturelle d'une F_0 (homme, femme, enfant) pour une parole lue et relativement neutre d'expressivité. Il existe plusieurs manières de parcourir cette zone de recherche. Celle adoptée pour les exemples présentés consiste à parcourir la zone de $[50, 800Hz]$ par pas constants de 1Hz. Autrement dit, la construction d'une fonction de pitch nécessite 800-50+1 produits scalaires soit 751 exactement. Les AEP que nous proposons chapitre 5 parcourent l'intervalle de recherche de manière logarithmique avec un nombre de pas constant par octave. L'estimation des F_0 par estimation conjointe consiste à

repérer dans la fonction de pitch $\Phi(F_c)$ les N pics de plus forte amplitude dans l'ordre décroissant de leur amplitude. Ces N pics constituent les hypothèses F_0 du signal étudié. Le cas idéal étant d'avoir exactement le même nombre de pics que de F_0 constituant le mélange et de fixer N à la valeur exacte du nombre de F_0 du mélange. Selon les besoins, rien n'empêche de fixer N à une valeur plus grande que le nombre théorique de F_0.

4.1.2 Spectres d'amplitude exemples

Les spectres d'amplitude idéaux utilisés et notés $|S(f)|$ sont des sommes de paraboles décalées en fréquence. Ces paraboles modélisent les partiels du spectre d'amplitude comme il est proposé dans [Martin, 2000]. Elles sont d'épaisseur fréquentielle b_w finie et sont centrée autour de la valeur centrale du partiel. Les paraboles sont décrites selon l'équation 4.2

$$H(f) = 1 - \frac{4}{b_w^2} f^2, f \in \left[-\frac{b_w}{2}, \frac{b_w}{2} \right] \tag{4.2}$$

Pour être certain qu'il n'y ait pas plus d'une dent du PUI par parabole, b_w doit valoir au maximum F_{0min}. Ce point est important car dès lors qu'il y a plus d'une dent dans une parabole, le PUI n'a plus le rôle de « détecteur » de structure harmonique mais se rapproche du rôle d'un intégrateur (somme des tous les échantillons du spectre d'amplitude). L'épaisseur fréquentielle b_w est donc fixée à F_{0min} soit 50Hz pour tous les exemples du chapitre. En dehors de l'intervalle $[-b_w/2, b_w/2]$, $H(f)$ est nulle. Pour simplifier l'interprétation des fonctions de pitch du chapitre, l'amplitude maximale des harmoniques est constante et unitaire.

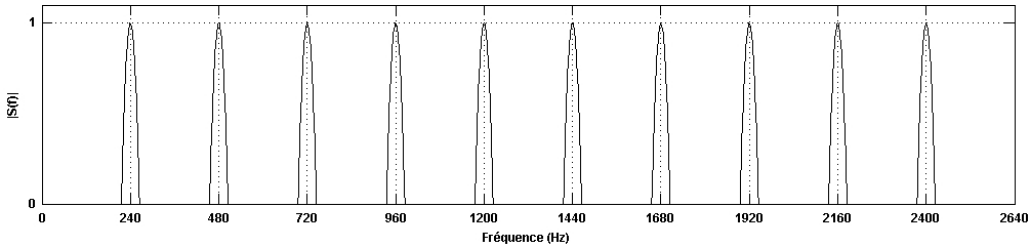

FIGURE 4.1: Spectre d'amplitude monopitch exemple.

4.1.2.1 Situation monopitch

Le spectre d'amplitude idéal d'un signal voisé est constitué de N_h partiels en relation harmonique. L'équation 4.3 présente la formule théorique du spectre d'amplitude idéal en situation monopitch.

$$|S(f)| = \sum_{k=1}^{N_h} H(f - kF_0) \tag{4.3}$$

La F_0 est constante et fixée à 240Hz. Le nombre d'harmoniques N_h est fixé à 10. La fréquence d'échantillonnage F_e est fixée à 8kHz ce qui implique que les spectres d'amplitude sont définis jusqu'à 4kHz. La figure 4.1 présente le spectre exemple de la situation monopitch.

4.1.2.2 Situation bipitch

La figure 4.2 présente la fonction de pitch d'un spectre d'amplitude en situation bipitch. Les F_0 valent 200 et 240Hz et sont respectivement notées F_{01} et F_{02}. Les deux structures comportent dix harmoniques. Les harmoniques communes sont indiquées par des points rouges. Soit $|S_1(f)|$ le spectre d'amplitude isolé de la structure harmonique F_{01} et $|S_2(f)|$ le spectre d'amplitude isolé de la structure harmonique F_{02}. Le spectre d'amplitude $|S(f)|$ comportant les deux structures harmoniques est calculé de manière à ce que pour chacune des fréquences, la valeur de $|S(f)|$ vaille le maximum entre $|S_1(f)|$ et $|S_2(f)|$ comme le témoigne l'équation 4.4.

$$|S(f)| = \max(|S_1(f)|, |S_2(f)|) \tag{4.4}$$

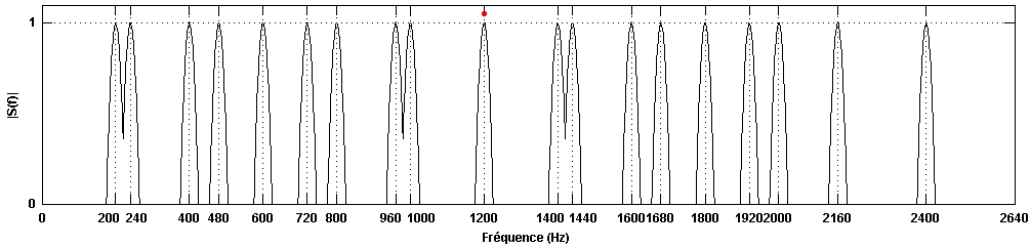

FIGURE 4.2: Spectre d'amplitude bipitch exemple.

4.1.2.3 Situation 4-pitch

En situation 4-pitch, la méthode de construction du spectre d'amplitude est identique à la situation bipitch. Le spectre d'amplitude est constitué de quatre structures harmoniques de dix harmoniques chacune et de F_0 200, 240, 280 et 320Hz. Les quatre F_0 sont respectivement notées F_{01}, F_{02}, F_{03} et F_{04}. La figure 4.3 illustre le spectre d'amplitude idéal de la situation 4-pitch exemple. Les harmoniques communes sont indiquées par des points rouges.

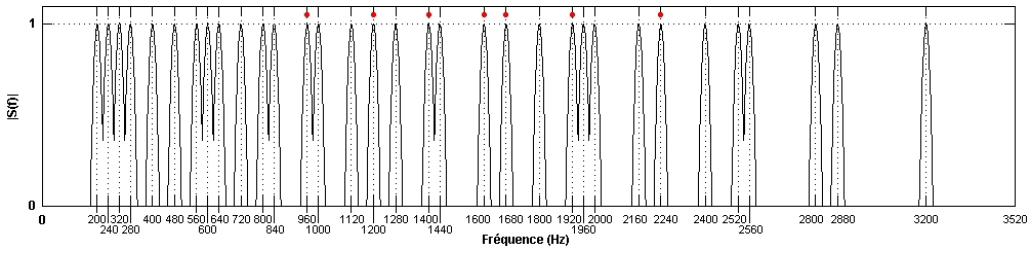

FIGURE 4.3: Spectre d'amplitude 4-pitch exemple.

Analysons à présent le fonctionnement du peigne spectral le plus élémentaire sur ces trois spectres d'amplitude idéaux représentant trois situations de mélange différentes. Ce peigne spectral fait l'objet de la section 4.2 suivante.

4.2 Peigne Uniforme Infini (PUI)

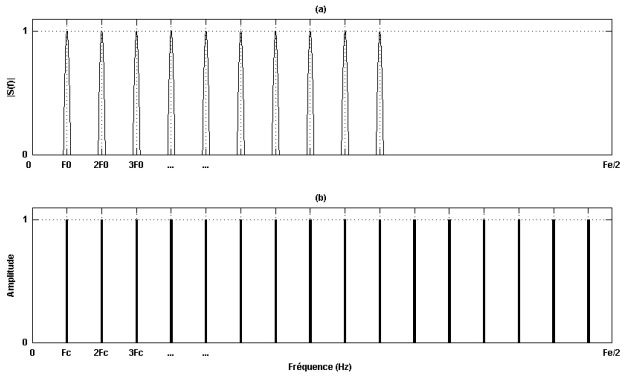

FIGURE 4.4: Spectre d'amplitude et PUI. (a) Spectre d'amplitude monopitch exemple. (b) PUI avec F_c égale à la F_0 du spectre exemple.

Le Peigne Uniforme Infini (PUI) est le peigne fréquentiel le plus élémentaire. Il s'agit d'un peigne fréquentiel comportant des dents discrètes régulièrement espacées en fréquence et d'amplitude unitaire (cf. graphique (b) figure 4.4). Le PUI peut être vu comme un simple échantillonneur de fréquence caractéristique F_c paramétrable. Ce seul paramètre F_c suffit à le définir. Le PUI s'étend sur l'ensemble du spectre d'amplitude donc de 0 à la moitié de la fréquence d'échantillonnage F_e. Dans l'exemple de la figure 4.4, la fréquence caractéristique du PUI est égale à la F_0 de $|S(f)|$ et les dents du PUI coïncident exactement avec les harmoniques du spectre. Le produit scalaire dans cette configuration est donc maximal. Les fonctions de pitch obtenues par un PUI sont aussi nommées fonctions PUI. Les paragraphes 4.2.1, 4.2.3 et 4.2.4 présentent les fonctions PUI obtenues dans les situations monopitch, bipitch et 4-pitch respectivement. Le paragraphe 4.2.2 analyse la fonction PUI monopitch pour en tirer une règle générale concernant la position des maxima locaux de la fonction de pitch.

4.2.1 Situation monopitch

La fonction PUI obtenue sur le spectre d'amplitude exemple monopitch est donné dans la figure 4.5. La fonction de pitch présente un certain nombre de maxima locaux. Il y a quatre maxima locaux d'amplitude 10 en 240, 120, 80 et 60Hz. Cette amplitude de 10 correspond à la somme des sommets de chacun des 10 harmoniques. La valeur du F_0 est effectivement le maximum local de plus forte amplitude (point rouge) mais cette valeur se retrouve en concurrence avec ses sous-multiples. De manière générale, le pic en F_0 est entouré de pics parasites. En fonction des amplitudes de ces pics, ils sont plus ou moins parasites. Les plus parasites sont ceux d'amplitude 10 qui correspondent aux sous-multiples de la F_0 soit $(1,2)$, $(1,3)$ et $(1,4)$. Ces pics parasites sont extrêmement gênants car ils sont de même amplitude que le maximum $(1,1)$ correspondant à la F_0. Il y a d'autres pics parasites d'amplitude moins élevée mais potentiellement gênants. Les pics $(2,1)$, $(2,3)$ d'amplitude 5, le pic $(3,2)$ d'amplitude 3, les pics $(4,3)$ et $(5,3)$ d'amplitude 2, etc. La zone sous-harmonique est caractérisée par des pics parasites dont l'amplitude peut atteindre celle du pic de la F_0. En effet, lorsque F_c vaut la moitié de F_0, toutes les harmoniques du spectre se retrouvent alignées avec les dents de numéro pair du PUI et le produit scalaire vaut 10. La zone sur-harmonique est caractérisée par des pics parasites de plus faible amplitude. L'amplitude des pics sur-harmoniques multiples de la F_0 (ie. $(2,1)$, $(3,1)$, $(4,1)$, etc.) suit une décroissance hyperbolique en A/k avec A l'amplitude du pic de la F_0 et k entier, le numéro du multiple. En effet, le nombre d'harmonique étant de 10, lorsque F_c vaut F_0, les 10 harmoniques sont bien alignées avec les dents du PUI. En revanche si F_c est double, seule une harmonique sur deux

est alignée avec le PUI.

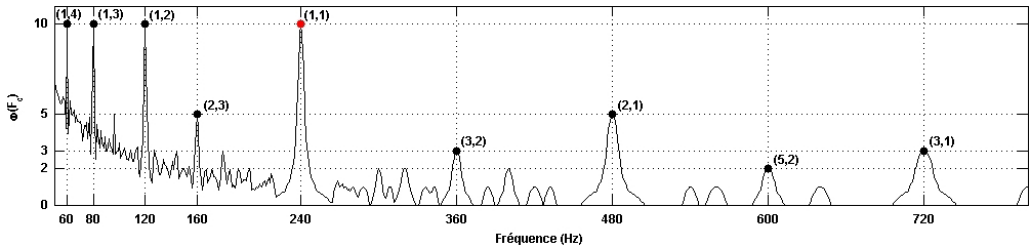

FIGURE 4.5: Fonction PUI en situation monopitch et indexation des pics. Chaque pic est défini par un couple d'entiers (p, q). Le pic $(1, 1)$ est le pic de la F_0, le pic $(2, 1)$ est le pic à l'octave et le pic $(1, 2)$ est le pic à la sous-octave.

4.2.2 Indexation des pics parasites

Les pics parasites, s'ils sont d'amplitude trop importante, induisent des erreurs d'estimation de F_0. Dans la littérature, ces erreurs sont appelées généralement erreurs d'octave ou de sous-octave. Effectivement, les pics parasites les plus forts sont généralement placés à l'octave ou à la sous-octave ce qui implique que la grande majorité des erreurs soit moitié ou double. Il est proposé dans ce paragraphe une généralisation des erreurs faites par les AEP. Une des contributions de cette thèse est de mettre en évidence que la position des maxima locaux n'est pas aléatoire. Elle est fonction de leur position fréquentielle par rapport aux F_0. La position de chacun des pics de la fonction de pitch est indexable par un couple d'entiers naturels non nuls (p, q) comme indiqué sur la figure 4.5. Ceci amène à ne plus limiter les erreurs à de simples erreurs d'octave ou de sous-octave mais à des erreurs de type fractionnaire p/q avec p, q entiers non nuls et premiers entre eux. Un couple (p, q) détermine leur position fréquentielle par rapport à la F_0 du signal par la relation $(p/q)F_0$. Ainsi $(1, 1)$ dénote le pic $(1/1)F_0$ correspondant à la F_0. $(2, 1)$ dénote le pic $(2/1)F_0$ à l'octave et $(1, 2)$ le pic à la sous-octave. La valeur théorique des amplitudes des pics parasites (p, q) avec p, q entiers non nuls et premiers entre eux est donnée par l'équation 4.5. La preuve mathématique est fournie en annexe A.1.

$$\Phi\left(\frac{p}{q}F_0\right) = E\left(\frac{N_h}{p}\right) \tag{4.5}$$

L'amplitude des pics parasites ne dépend que de p. Les pics parasites $(1, q)$ sont de même amplitude que le pic de la F_0 $(1, 1)$ ce qui rejoint parfaitement les résultats de la figure 4.5. Ces pics parasites sont donc extrêmement gênants puisqu'on ne peut déterminer que le pic en F_0 est le maximum local de plus forte amplitude. Le caractère parasite d'un pic de la fonction de pitch est quantifiable sur une échelle allant du plus parasite au moins parasite. Tout pic à la fréquence sous-multiple de F_0 $(p = 1, q)$ est le plus parasite car l'amplitude de ce pic est identique à celui de la F_0. La deuxième famille de pic la plus parasite est celle qui correspond à $p = 2$ soit $(2, 1)$, $(2, 3)$, $(2, 5)$, $(2, 7)$, etc. Autrement dit, plus p est grand, moins le pic est parasite.

L'indexation (p, q) des pics parasites a été pressentie dans l'article [Klapuri, 1998] qui propose une analyse du spectre de mélanges sonores à partir de la théorie des nombres. Elle fait aussi l'ob-

jet d'une note de bas de page dans le livre intitulé « *Multi-Pitch Estimation* » de Christensen & Jakobsson [Christensen and Jakobsson, 2009]. Toutefois, ces références ne pointent pas suffisamment l'importance que revêt une telle indexation. En effet, la dénomination (p, q) est essentielle puisque le fait de pouvoir nommer les pics parasites va nous permettre de mettre en œuvre les mécanismes pour les traiter spécifiquement.

Les pics parasites des fonctions de pitch produisent des erreurs qui sont actuellement résolues par des post-traitements (programmation dynamique, filtres médians, etc.). Il s'agit généralement de méthodes de suivi de F_0 qui appliquent des contraintes de continuité temporelle ou fréquentielle pour former des trajectoires de F_0 les plus lisses possibles. L'indexation (p, q) proposée dans cette section va permettre de ne plus s'appuyer sur de tels algorithmes pour corriger l'estimation de F_0. Le problème n'est donc plus de mettre en place des post-traitements pour corriger les erreurs (p, q). Il s'agit à présent d'utiliser l'indexation pour réduire les pics (p, q) dans le principe même de la construction de la fonction de pitch. Ainsi l'AEP que nous proposons ne fait pas appel à une étape de suivi de F_0. Au plus bas niveau de traitement, il est déjà possible d'atténuer fortement les erreurs fractionnaires (p, q) des AEP. Nous espérons ainsi repousser au plus tard l'intervention d'informations haut-niveau sur le signal dans l'estimation de F_0.

Qu'en est-il des pics parasites des fonctions PUI dans les deux situations multipitch ?

4.2.3 Situation bipitch

La fonction de pitch de la figure 4.6 présente plus de pics parasites que dans la situation bipitch. Deux maxima locaux sont bien placés aux valeurs des F_0 à savoir 200 et 240Hz (points rouges). Néanmoins, ils ne sont pas les maxima locaux de plus forte amplitude et l'estimation conjointe avec deux hypothèses F_0 sera erronée. Les pics en 80, 100 et 120Hz seront les premiers candidats F_0 données par l'AEP. Pour expliquer ce phénomène, considérons le pic à 100Hz. Les multiples de 100Hz coïncident avec les dix harmoniques de F_{01} mais également avec certaines harmoniques de la série la série harmonique F_{02} (ie. 480Hz, 720Hz, 1680Hz et 1920Hz qui sont proche d'un multiple de 100Hz à moins de $b_w/2$ soit 25Hz et 2400Hz qui tombe exactement sur un multiple de 100). Il en va de même pour tous les pics sous-harmoniques. Il est également important de noter le comportement hyperbolique de la fonction

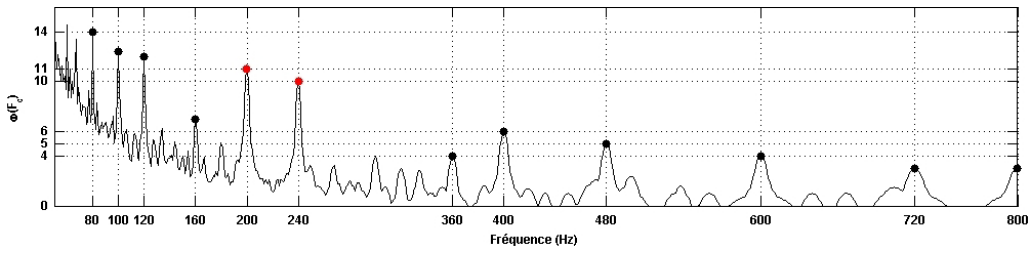

FIGURE 4.6: Fonction PUI en situation bipitch.

PUI dans la zone sous-harmonique. Ce phénomène est également présent en situation monopitch (cf. figure 4.5). Il sera explicité dans le paragraphe 4.3.1 et est manifestement un obstacle de plus pour l'estimation de F_0 puisque toutes les amplitudes de la zone sous-harmonique sont augmentées et peuvent

donc potentiellement devenir un candidat F_0.

4.2.4 Situation 4-pitch

La figure 4.7 présente la fonction de pitch du spectre d'amplitude 4-pitch. Les points rouges sont les pics des F_0 théoriques. Les pics parasites de la fonction de pitch sont très nombreux. Les fréquences inférieures à la plus petite des F_0 200Hz deviennent très parasites et sont d'amplitude supérieure aux amplitudes des pics placés aux F_0. Il est à noter que le pic en 80Hz émerge beaucoup des autres ce qui s'explique par le fait qu'il est sous-multiple commun de 240 et 320Hz. L'estimation de F_0 par exploitation de la fonction de pitch obtenue par PUI est d'autant plus complexe que le nombre de candidats pitch est important.

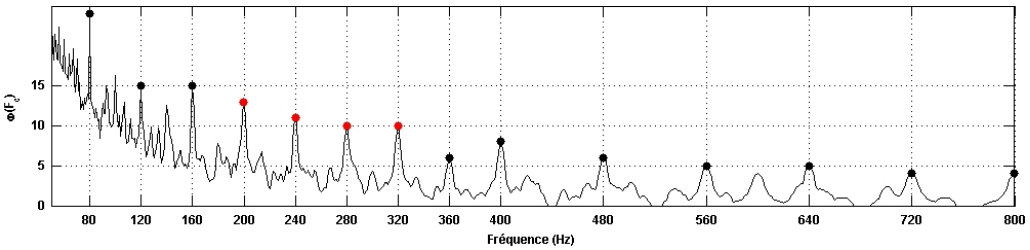

FIGURE 4.7: Fonction PUI en situation 4-pitch.

Deux effets néfastes de l'estimation de F_0 par PUI apparaissent clairement. Premièrement, les nombreux pics parasites des fonctions PUI qui engendrent des erreurs d'estimation de F_0 dès lors que leurs amplitudes sont supérieures aux amplitudes des F_0 théoriques. Ceci survient dès la situation monopitch et s'amplifie en situation multipitch. Deuxièmement, la remontée hyperbolique constatée dans les basses fréquences qui complique encore plus une approche de type recherche de maximum. Par conséquent, l'utilisation d'un PUI pour estimer des F_0 multiples ne semble pas adaptée puisqu'elle fait des erreurs alors même que les signaux exemples utilisés ont une périodicité parfaite. Comment rendre l'estimation de F_0 possible par l'utilisation de peignes spectraux ? Il est nécessaire pour cela de réduire l'amplitude des pics parasites afin de rendre les pics à estimer les plus émergents possibles. Plusieurs solutions ont déjà été envisagées en situation monopitch. Elles peuvent s'étendre telles quelles à la situation multipitch mais restent encore limitées. Les mécanismes de réduction des pics parasites font l'objet de la section 4.3.

4.3 Réduction des pics parasites

Les figures 4.5, 4.6 et 4.7 ont un comportement hyperbolique qui peut se révéler gênant pour l'estimation de F_0 lorsque celles-ci sont dans les basses fréquences. Le paragraphe 4.3.1 explique le pourquoi de ce comportement hyperbolique et donne également le moyen de le supprimer définitivement. La suppression hyperbolique se révèle d'ailleurs être une manière de réduire les pics parasites. Les paragraphes 4.3.2, 4.3.3, 4.3.4 et 4.3.5 présentent d'autres formes de peignes spectraux intégrant des mécanismes d'atténuation des pics parasites.

4.3.1 Correction du comportement hyperbolique

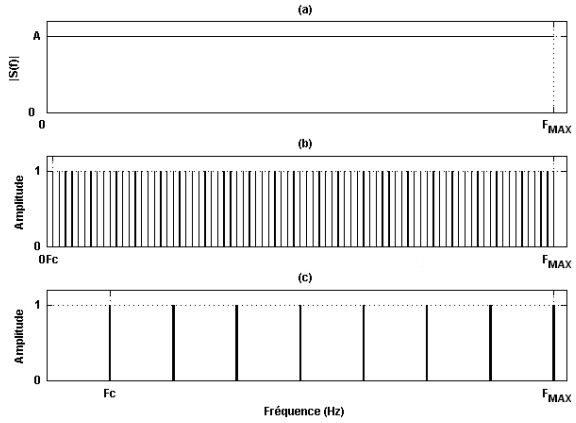

FIGURE 4.8: Pourquoi le comportement hyperbolique ?

Le comportement hyperbolique dans les fonctions de pitch obtenues par PUI s'explique à partir des schémas de la figure 4.8. Le graphique (a) présente un spectre d'amplitude $|S(f)|$ constant d'amplitude A. Ce spectre correspond à celui d'un bruit blanc gaussien. Le graphique (b) présente un PUI dont la F_c est réglée à une valeur fréquentielle faible. Le graphique (c) présente un PUI dont la F_c est réglée à une valeur fréquentielle plus importante. Le produit scalaire de (a) avec (b) donnera une valeur plus importante que le produit scalaire entre (a) et (c) puisqu'il dépend du nombre de dents du PUI comprises entre 0 et F_{MAX} par pas de F_c. Soit $M(F_c)$ la fonction qui correspond à ce nombre de dents. La valeur de $M(F_c)$ est donnée par l'équation 4.6.

$$M(F_c) = E\left(\frac{F_{MAX}}{F_c}\right) \tag{4.6}$$

Le produit scalaire de $|S(f)|$ avec un PUI est donné en équation 4.7. Étant donné que l'amplitude du spectre est constante, ce produit scalaire est proportionnel au nombre de dents du PUI. Le produit scalaire est noté $\Phi(F_c)$ et correspond à la fonction de pitch.

$$\Phi(F_c) = \sum_{i=1}^{M(F_c)} |S(iF_c)| = A.M(F_c) = A.E\left(\frac{F_{MAX}}{F_c}\right) \tag{4.7}$$

La fonction de pitch est donc une fonction de type hyperbolique en $1/F_c$ ce qui en justifie le comportement hyperbolique. En effet, un spectre $|S(f)|$ quelconque est décomposable en deux fonctions nommées $C(f)$ et $V(f)$ comme l'indique l'équation 4.8. $C(f)$ représente la moyenne du spectre d'amplitude et correspond donc à une fonction constante tout le long des fréquences. $C(f)$ est similaire à la fonction présentée dans le graphique (a) de la figure 4.8. L'amplitude constante de la fonction $C(f)$ sera notée A. $V(f)$ représente l'évolution du spectre autour de sa moyenne et correspond à une fonction quelconque mais dont la moyenne est nulle.

$$|S(f)| = C(f) + V(f) = A + V(f) \tag{4.8}$$

La fonction de pitch obtenue par un PUI est donnée par l'équation 4.9 suivante. Elle dépend de $M(F_c)$ qui correspond au nombre de dents du PUI jusqu'à F_{MAX} (cf. équation 4.6).

$$\begin{array}{ll} \Phi(F_c) & = \sum_{i=1}^{M(F_c)} |S(iF_c)| \\ \Phi(F_c) & = \sum_{i=1}^{M(F_c)} |C(iF_c)| + \sum_{i=1}^{M(F_c)} |V(iF_c)| \end{array} \tag{4.9}$$

Il convient d'analyser la valeur de la première somme avec $C(iF_c)$. $C(f)$ étant une fonction constante, la somme se simplifie. L'équation 4.10 en donne la valeur théorique.

$$\sum_{i=1}^{M(F_c)} C(iF_c) = A.M(F_c) = A.E\left(\frac{F_{MAX}}{F_c}\right) \tag{4.10}$$

Il s'agit de la même fonction hyperbolique obtenue et illustrée dans l'équation 4.7. La valeur de la fonction de pitch est donc la somme d'une fonction hyperbolique et d'une autre somme contenant $V(iF_c)$ comme l'indique l'équation 4.11. C'est ce qui implique que dans toutes les fonctions de pitch vues jusqu'à présent, le comportement hyperbolique l'emporte dans les basses fréquences, augmente considérablement les amplitudes des pics parasites et rend l'estimation de F_0 difficile.

$$\Phi(F_c) = A.E\left(\frac{F_{MAX}}{F_c}\right) + \sum_{i=1}^{M(F_c)} |V(iF_c)| \tag{4.11}$$

La valeur de la deuxième somme dépend entièrement de l'allure de $V(f)$. Lorsque Fc tend vers 0, la somme devient égale à la moyenne de $V(f)$ qui est nulle. Dans ce cas, la deuxième somme en $V(iF_c)$ disparaît. Si $V(f)$ est une structure harmonique, la deuxième somme est maximale lorsque F_c vaut la F_0 de la structure harmonique. La seconde somme en $V(iF_c)$ est donc un indicateur de la présence d'une structure harmonique dans $V(f)$. Pour une fréquence donnée, plus l'écart à l'hyperbole est important, plus la structure harmonique de cette fréquence est forte. L'écart relatif à l'hyperbole pourrait se révéler être un indicateur de voisement utile.

Avec cette analyse, la suppression du comportement hyperbolique devient aisée. L'équation 4.11 montre que la fonction de pitch est la somme d'un terme hyperbolique et d'un autre terme permettant de quantifier le voisement. Si le spectre d'amplitude est de moyenne nulle, alors A vaut 0 et l'hyperbole disparaît. Il suffit donc de retirer la moyenne au spectre d'amplitude étudié pour annuler tout effet hyperbolique. Les sous-paragraphes 4.3.1.1, 4.3.1.2 et 4.3.1.3 reprennent les fonctions de pitch en annulant le comportement hyperbolique.

4.3.1.1 Situation monopitch

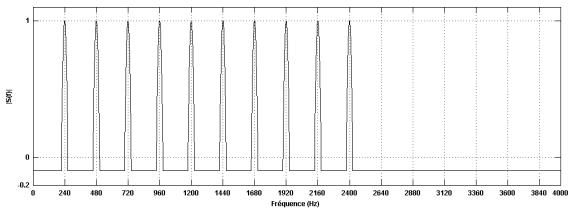

FIGURE 4.9: Spectre d'amplitude exemple dont la moyenne est nulle.

Pour illustrer la suppression du comportement hyperbolique des fonctions de pitch obtenues par PUI en situation monopitch, le spectre d'amplitude idéal dont les harmoniques sont formées à partir de paraboles est utilisé (cf. figure 4.1). Pour que la moyenne du spectre soit nulle, une constante négative est ajoutée dans les zones fréquentielles qui ne sont pas occupées par les harmoniques. La figure 4.9 illustre l'ajout de ces parties négatives sur le spectre. Cette manière de procéder permet de conserver les amplitudes maximales des harmoniques à 1 afin de faciliter l'interprétation des résul-

tats théoriques. Le spectre d'amplitude pourrait aussi être moyenné. Cette façon de procéder n'est pas tout à fait identique à celle d'ajouter des parties négatives. Dans l'exemple la constante négative vaut -0.0952. La figure 4.10 présente la fonction PUI avec correction hyperbolique. Les amplitudes négatives de la fonction PUI ne sont pas affichées car elles correspondent à des fréquences qui ne sont pas des F_0. La partie négative de la fonction PUI avec correction hyperbolique n'est pas présentée. En effet, si le produit scalaire est négatif, la fréquence n'est probablement pas une hypothèses F_0 sérieuse dans la mesure où sa structure harmonique tombe majoritairement sur les parties négatives du spectre d'amplitude, parties qui ne correspondant pas à des partiels. Pour cette raison, seules les parties positives des fonctions de pitch décrites dans ce chapitre seront présentées.

La suppression hyperbolique est effective. Cette suppression a un effet bénéfique sur l'estimation de F_0 puisque le pic le plus fort est unique et correspond à la F_0 (point rouge). L'estimation de la F_0 est donc correcte. La fréquence de coupure est fixée à 2425Hz et correspond à la fréquence du dernier partiel du spectre d'amplitude plus sa demi épaisseur fréquentielle. Pourquoi les amplitudes des sous-harmoniques de la F_0 sont-elles abaissées ? Par exemple, l'amplitude de la fonction de pitch à 120Hz prend en compte les 10 sommets de chacune des paraboles mais prend également en compte les inter-harmoniques. Or, les inter-harmoniques sont d'amplitude négatives et vont donc réduire l'amplitude du pic. Ce raisonnement s'applique à toutes les valeurs inférieures à F_0 et explique la diminution d'amplitude des pics sous-harmoniques. Malgré une estimation de F_0 correcte, certains pics restent d'amplitude élevée et donc potentiellement parasites (ie. 120Hz, 80Hz, etc.).

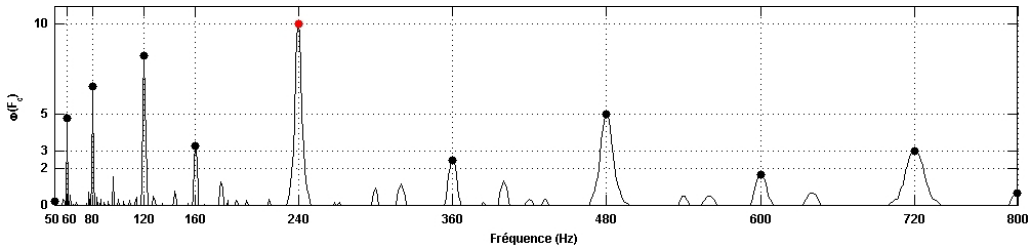

FIGURE 4.10: Fonction PUI monopitch avec correction hyperbolique.

4.3.1.2 Situation bipitch

La fonction PUI avec correction hyperbolique obtenue en situation bipitch est donnée figure 4.11. La constante négative vaut -0.2038 et prend en compte le recouvrement de certaines harmoniques (cf. figure 4.2). Les deux pics de plus forte amplitude sont effectivement les pics des F_0 en 200 et 240Hz (points rouges). L'estimation de F_0 est donc correcte. Il reste néanmoins de nombreux pics parasites (ex. 100 et 120Hz) et les plus forts d'entre eux dépassent nettement la moitié des amplitudes des F_0. L'annulation de l'hyperbole est intéressante mais n'assure pas complètement le rejet des pics parasites qui restent d'amplitude élevée. Ceci va se retrouver en situation 4-pitch mais de manière amplifiée.

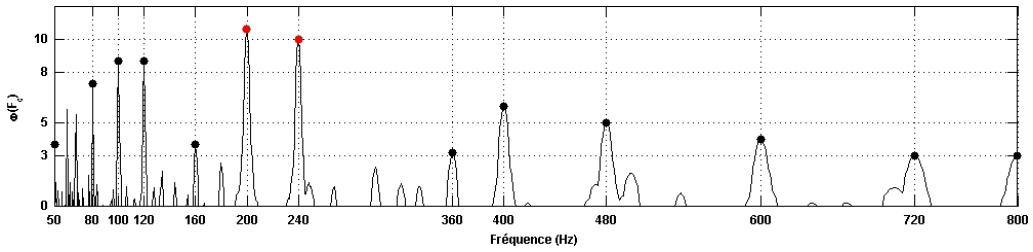

FIGURE 4.11: Fonction PUI bipitch avec correction hyperbolique.

4.3.1.3 Situation 4-pitch

Le même constat se retrouve dans la fonction de pitch présentée figure 4.12. La courbe présentée est la fonction PUI obtenue sur le spectre d'amplitude 4-pitch de 200, 240, 280 et 320Hz avec correction hyperbolique. Les maximas locaux en 200, 240, 280 et 320Hz ne sont plus les quatre pics de plus fortes amplitudes. Les quatre premiers candidats rangés par ordre décroissant d'amplitude sont 80, 160, 200 et 320Hz. L'estimation de F_0 n'est pas correcte. Le pic en 80Hz est très élevé car il est en relation de sous multiple avec 240Hz et 320Hz. En conservant plus de quatre hypothèses F_0 il est possible de retrouver les quatre F_0 correctes. L'augmentation du nombre de F_0 rend la fonction de pitch moins lisible. Le nombre de pics parasites augmente avec le nombre de candidats F_0 à trouver. Les basses fréquences situées au-dessous des valeurs des quatre F_0 sont plus chaotiques et malgré l'annulation de l'hyperbole, les maxima locaux des basses fréquences sont d'amplitude élevée. La correction du

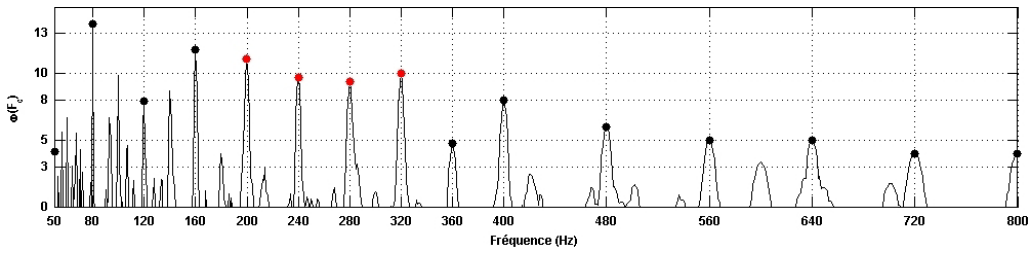

FIGURE 4.12: Fonction PUI 4-pitch avec correction hyperbolique.

comportement hyperbolique est systématiquement intégrée dans la suite du chapitre. Toutefois, la correction hyperbolique n'est pas suffisante. Pour s'en persuader, il suffit d'observer la situation 4-pitch qui échoue alors que l'exemple donné est idéal. Le même raisonnement expliquerait l'amplitude du pic en 160Hz.

Existe-t-il d'autres formes de peignes spectraux dont les caractéristiques permettent d'atténuer les pics parasites ? La réponse est oui et un certain nombre d'entre eux sont présentés dans les paragraphes suivants.

4.3.2 Peigne Uniforme Fini (PUF)

Une des solutions possibles pour réduire l'amplitude des pics parasites sous-harmonique ($q > p$) est de limiter le nombre de dents du PUI. Il ne faut plus parler de peigne uniforme infini mais de Peigne

Uniforme Fini (PUF). En limitant le nombre de dents à une valeur fixe notée N_d, pour les fréquences F_c inférieures à la F_0 le PUF n'a pas assez de dents pour prendre en compte toutes les harmoniques de la structure harmonique. Il est donc possible de prédire que ce peigne va réduire l'amplitude des pics sous-harmonique puisqu'il ne pourra pas s'étendre sur tout le spectre d'amplitude.

4.3.2.1 Situation monopitch

Dans l'exemple de la figure 4.13, le nombre de dents est limité à 8. Le pic $(1,1)$ a pour amplitude 8 car les 8 dents du PUF coïncident parfaitement avec les harmoniques et le produit scalaire vaut 8. Le pic $(1,2)$ est réduit comparé au $(1,1)$ car des 8 dents du PUF de $F_c = F_0/2$ il n'y a que 4 dents qui coïncident parfaitement avec les harmoniques. C'est pourquoi le produit scalaire vaut 4. Le comportement hyperbolique disparaît avec l'utilisation du PUF car le premier terme de l'équation 4.11 devient constant et non plus une fonction hyperbolique ($A.N_d$ au lieu de $A.M(F_c)$). Le fait de limiter

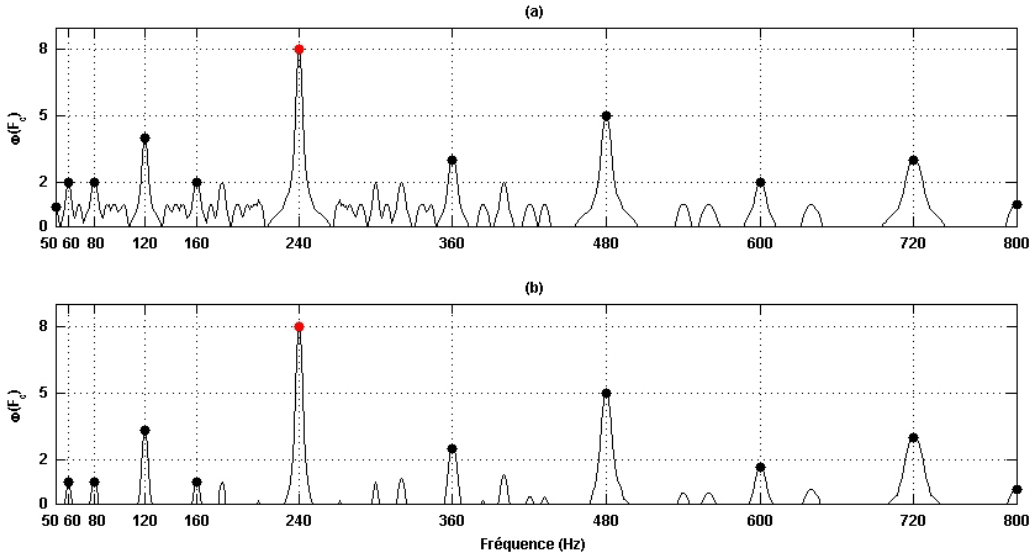

FIGURE 4.13: Fonctions PUF monopitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

le nombre de dents provoque une remontée relative des pics parasites sur-harmoniques par rapport au PUI. Pour s'en convaincre, il suffit d'observer l'amplitude des pics $(1,1)$ et $(2,1)$ de la fonction PUI figure 4.5 et des graphiques (a) et (b) de la figure 4.13. L'amplitude du pic $(2,1)$ vaut toujours 5 dans toutes les figures. En revanche, l'amplitude du pic $(1,1)$ de la fonction PUF n'est plus 10 mais 8. L'émergence relative du pic de la F_0 est donc diminuée par rapport au PUI ce qui équivaut à une remontée des pics parasites sur-harmonique. Il y a un autre inconvénient à limiter le nombre de dents. Il n'est plus possible de prendre en compte toute la structure harmonique du spectre. Ceci peut avoir pour effet d'être moins robuste à des structures harmoniques dans lesquelles manquent des harmoniques. L'idée serait de fixer le nombre de dents à une valeur suffisamment grande mais dans ce cas, on se ramène au PUI et les amplitudes des sous-harmoniques remontent. Un compromis apparaît donc. Les valeurs théoriques des amplitudes du graphique (a) de la figure 4.13 sont décrites dans l'équation 4.12. La preuve théorique est donnée dans l'annexe A.2. Celles du graphique (b) sont plus complexes à

déterminer et ne sont pas données. L'amplitude des dents ne dépend plus que de p comme pour le PUI et le q intervient désormais.

$$\Phi\left(\frac{p}{q}F_0\right) = \min\left(E\left(\frac{N_h}{p}\right), E\left(\frac{N_d}{q}\right)\right) \tag{4.12}$$

4.3.2.2 Situation bipitch

La figure 4.14 présente la fonction de pitch obtenue par un PUF de 8 dents dans la situation bipitch. Dans les deux cas, l'estimation de F_0 est correcte puisque les deux maxima locaux de plus forte amplitude sont placés en 200 et 240Hz. Les pics sous-harmoniques sont très nettement abaissés comparé à la figure 4.6 ce qui est un avantage du PUF. En revanche, les pics sur-harmonique et notamment celui en 400Hz sont remontés.

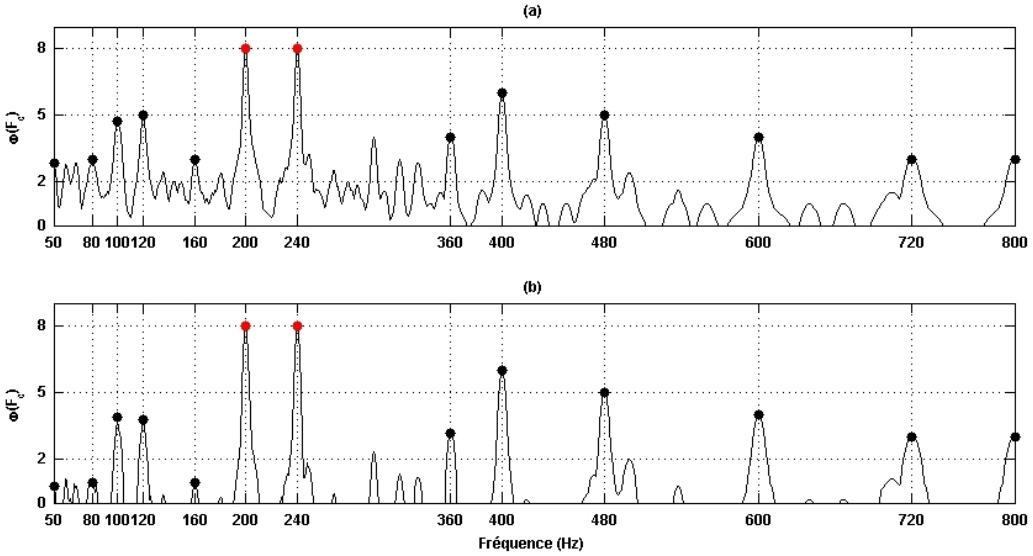

FIGURE 4.14: Fonctions PUF bipitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

4.3.2.3 Situation 4-pitch

La figure 4.15 présente les fonctions de pitch obtenues par un PUF dans la situation 4-pitch. Dans les graphiques (a) et (b), l'estimation de F_0 est presque correcte car les quatre pics de plus grande amplitude sont bien les pics en 200, 240, 280 et 320Hz. Ils sont toutefois en concurrence avec le pic à 400Hz qui est de même amplitude. L'amplitude de ces pics est de 8 ce qui correspond exactement au nombre de dents du PUF. Le pic en 400Hz s'explique dans la mesure où les 8 multiples de 400Hz correspondent effectivement avec une harmonique d'une des séries harmoniques des F_0 théoriques comme l'illustre le tableau 4.1. Le PUF est un peigne qui offre une bonne atténuation des pics sous-harmoniques même en situation 4-pitch. Toutefois, cet effet bénéfique est contrebalancé par l'augmentation relative (diminution de l'amplitude des pics F_0) des pics sur-harmoniques. Y aurait-il un compromis à trouver entre l'atténuation des pics parasites sous-harmonique et l'atténuation des pics parasites sur-harmonique ?

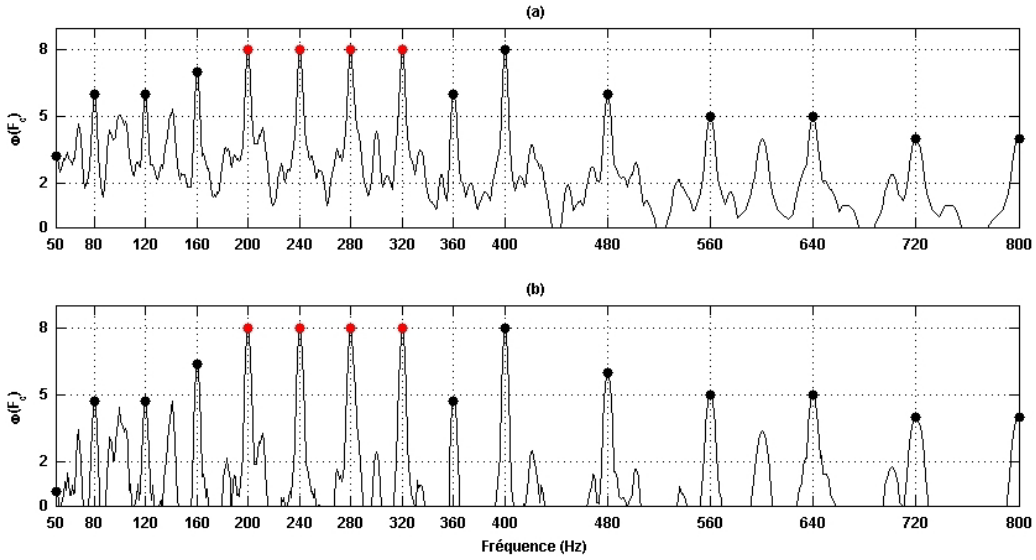

FIGURE 4.15: Fonctions PUF 4-pitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

TABLE 4.1: Lien des multiples de 400Hz avec les harmoniques des F_0 théoriques.

Fréquence (Hz)	**400**	**800**	**1200**	**1600**	**2000**	**2400**	**2800**	**3200**
Série harmonique	$2F_{01}$	$4F_{01}$	$6F_{01}$	$8F_{01}$	$10F_{01}$	$10F_{02}$	$10F_{03}$	$10F_{04}$

Il est intéressant de regarder ce qui se produit si au lieu de limiter le nombre de dents, l'amplitude de celles-ci décroît en fonction de leur numéro.

4.3.3 Peigne Décroissant Infini (PDI)

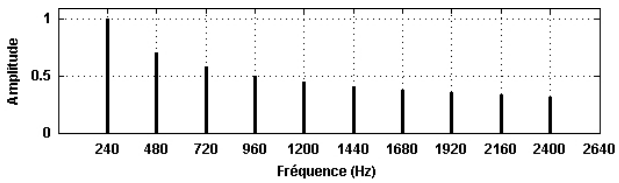

FIGURE 4.16: PDI avec décroissance en racine inverse.

Le peigne décrit dans ce paragraphe comporte un nombre illimité de dents dont les amplitudes décroissent en fonction du numéro de dents. Ce peigne est nommé PDI comme Peigne Décroissant Infini. La fonction de décroissance peut être quelconque. En donnant une décroissance aux dents du PUI, les amplitudes des pics parasites et notamment sous-harmoniques sont réduites. Le fait d'atténuer au lieu de limiter le nombre étant moins radical, on s'attend à ce que la remontée des pics sur-harmoniques soit moins marquée que pour un PUF. Si le compromis entre atténuation sous et sur-harmonique est vrai, il est attendu que les pics sous-harmoniques soient moins atténués que pour un PUF. La figure 4.16 donne un exemple de Peigne Décroissant Infini.

4.3.3.1 Situation monopitch

Dans cet exemple la décroissance des dents est en racine carrée du numéro de la dent. Ainsi l'amplitude de la dent 1 vaut $1/\sqrt{1}$, la dent 2 vaut $1/\sqrt{2}$ et ainsi de suite. La décroissance des dents donne plus d'importance aux premières dents qu'aux dernières. Cela rejoint la limitation du nombre de dents car à partir d'un certain nombre de dents, l'atténuation peut la faire quasiment disparaître. Mais l'effet doit être plus modéré. La figure 4.17 illustre les fonctions de pitch obtenue par un PDI. Le graphique (a) présente le résultat obtenu sans correction hyperbolique et le graphique (b) avec. Dans les deux graphiques, la fréquence de coupure F_{MAX} vaut 2425Hz. Dans les deux cas l'estimation de F_0 est correcte. L'équation 4.13 donne les amplitudes de la fonction de pitch du graphique (a). La valeur théorique de l'amplitude maximale en 240Hz est de 5.02 ce qui correspond à la somme des racines inverses de 1 à 10. Les amplitudes des pics (p, q) du graphique (a) sont fonction de p et de q contrairement au PUI. La preuve théorique est donnée en annexe A.3.

$$\Phi\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{E(N_h/p)} \frac{1}{\sqrt{qk}} \tag{4.13}$$

Le PDI permet de réduire les pics parasites sous-harmoniques sans relever de manière trop importante

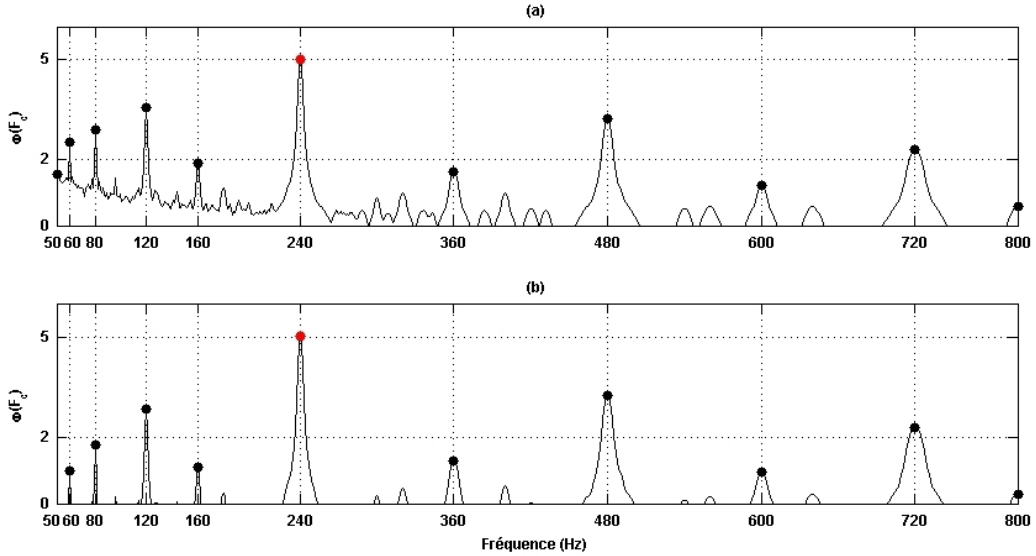

FIGURE 4.17: Fonctions PDI monopitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

les pics sur-harmoniques. Cette approche est effectivement plus modérée dans le comportement qu'un PUF. Toutefois le comportement hyperbolique réapparaît mais de manière atténuée par rapport à un PUI. Le PDI est suffisant pour déterminer correctement la F_0 dans notre exemple car la maximum local le plus élevé correspond bien à la F_0 du signal (points rouges). Une analyse poussée pourrait être menée sur la meilleure fonction de décroissance à adopter pour optimiser l'émergence du maximum local de la F_0 mais ce point ne sera pas étudié ici. La meilleure configuration serait celle qui rend l'amplitude des pics parasites sous-harmonique et sur-harmonique égales et la plus faible possible. C'est pratiquement le cas dans la figure 4.17.

4.3.3.2 Situation bipitch

La figure 4.18 présente les fonctions de pitch obtenus par un PDI dans la situation bipitch. F_{MAX} est fixée à 2425Hz. Le graphique (a) présente la fonction de pitch sans correction hyperbolique et le graphique (b) avec. Dans les deux graphiques (a) et (b), l'estimation des F_0 est correcte. Les mêmes constats qu'en situation monopitch peuvent être émis : équilibre presque parfait entre l'amplitude des pics parasites sous-harmonique/sur-harmonique et importance de la correction hyperbolique. Il est a noter qu'avec l'exemple choisi, l'estimation des deux F_0 est correcte même sans correction hyperbolique. Cela n'est plus le cas en situation 4-pitch.

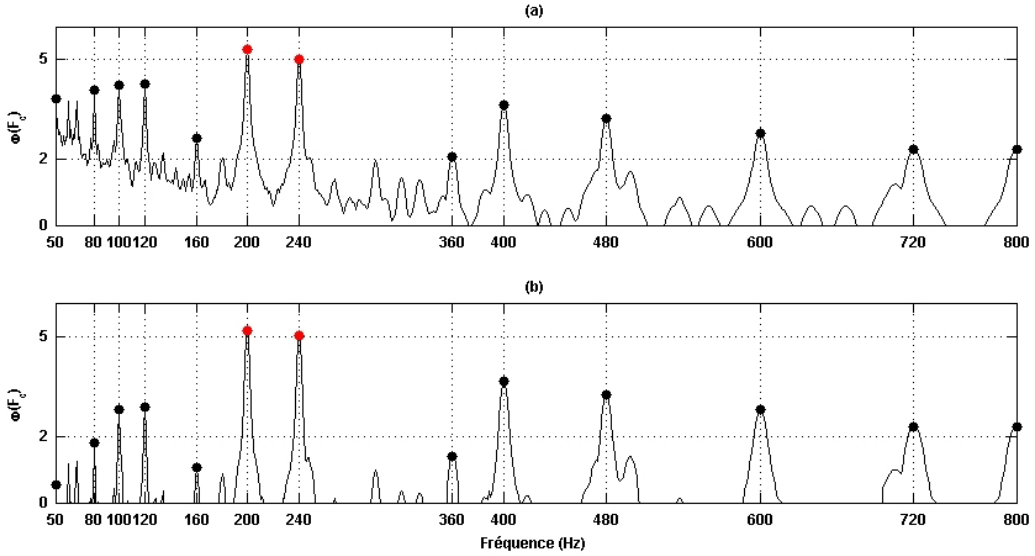

FIGURE 4.18: Fonctions PDI bipitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

4.3.3.3 Situation 4-pitch

La figure 4.19 présente les fonctions de pitch obtenues par un PDI dans la situation 4-pitch. Le graphique (a) montre la fonction de pitch sans correction hyperbolique et le graphique (b) présente le résultat obtenu avec suppression hyperbolique. La fréquence de coupure F_{MAX} est fixée à 3225Hz (10*320+25). Dans le graphique (a), l'estimation des F_0 est incorrecte car parasitée par le comportement hyperbolique. Le pic en 80Hz est de plus forte amplitude que les pics des F_0. Son amplitude s'explique par le fait que 80Hz soit le pic $(1,3)$ de la structure harmonique de 240Hz (F_{02}) et le pic $(1,4)$ de la structure harmonique du 320Hz (F_{04}). Il faudra plus de 4 hypothèses F_0 En revanche dans le graphique (b), l'estimation est correcte. Le PDI se révèle être meilleur que le PUF dans les situations bipitch et 4-pitch. Il parvient même à correctement estimer les F_0 dans la situation 4-pitch. Son comportement plus modéré que le PUF semble effectivement indiquer qu'il y a un compromis trouver entre l'émergence sous-harmonique et sur-harmonique. Un gain d'émergence d'un coté implique une perte de l'autre. Il y a un point de fonctionnement optimal. Que se passe-t-il lorsque le PUF et le PDI sont fusionnés en un même peigne ?

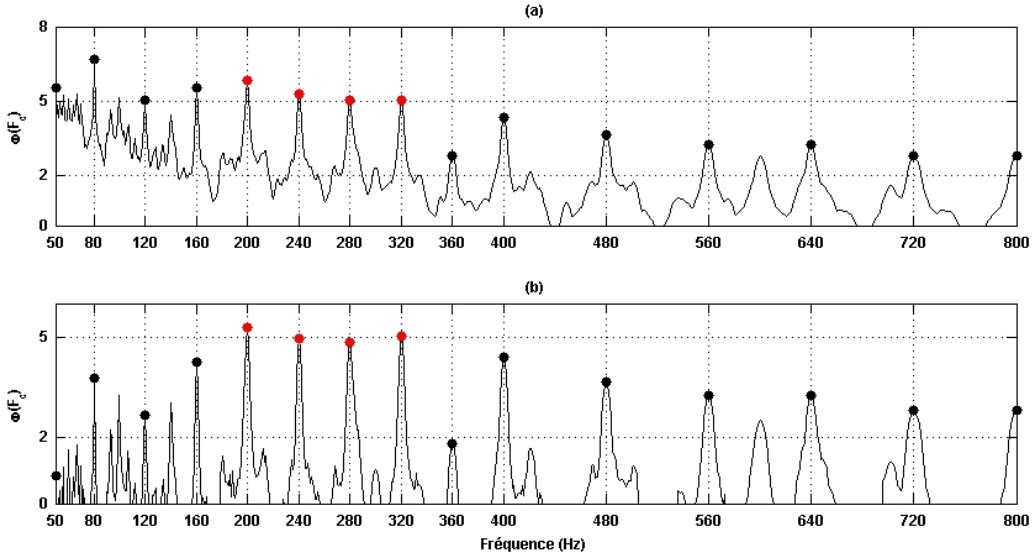

FIGURE 4.19: Fonctions PDI 4-pitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

4.3.4 Peigne Décroissant Fini (PDF)

Le Peigne Décroissant Fini (PDF) combine à la fois la limitation du nombre de dents et la décroissance des amplitudes des dents. Le PDF est un prototype de peigne qui s'apparente à une configuration du peigne de Martin [Martin, 1981, 1982] présenté dans le chapitre 3 figure 3.16. La décroissance considérée est en racine carrée inverse du numéro de la dent et le nombre de dents N_d est fixé à 8. La fréquence de coupure est fixée à la fréquence centrale de la dernière harmonique plus sa demi épaisseur fréquentielle soit 2425Hz dans l'exemple.

4.3.4.1 Situation monopitch

La figure 4.20 présente deux fonctions de pitch du PDF. Le graphique (a) représente la fonction de pitch sans correction hyperbolique et le graphique (b) avec correction. La valeur théorique des amplitudes de pics (p, q) du graphique (a) sans correction hyperbolique est donnée par l'équation 4.14. La preuve théorique est donnée en annexe A.4.

$$\begin{aligned} \Phi\left(\frac{p}{q}F_0\right) &= \sum_{k=1}^{N} \frac{1}{\sqrt{qk}} \\ N &= \min\left[E\left(\frac{N_h}{p}\right), E\left(\frac{N_d}{q}\right)\right] \end{aligned} \tag{4.14}$$

L'estimation de la F_0 est correcte. Le pic parasite $(2, 1)$ reste cependant de forte amplitude. La décroissance en racine inverse mêlée à la limitation du nombre de dents n'est pas optimale en terme d'égalisation des amplitudes des pics parasites sous et sur-harmoniques. Pour améliorer cela, il faudrait une décroissance moins forte ou une augmentation du nombre de dents. Le comportement hyperbolique n'est plus mais est remplacé par un « offset » dans les basses fréquences (cf. considérations sur le PUF dans le paragraphe 4.3.2.)

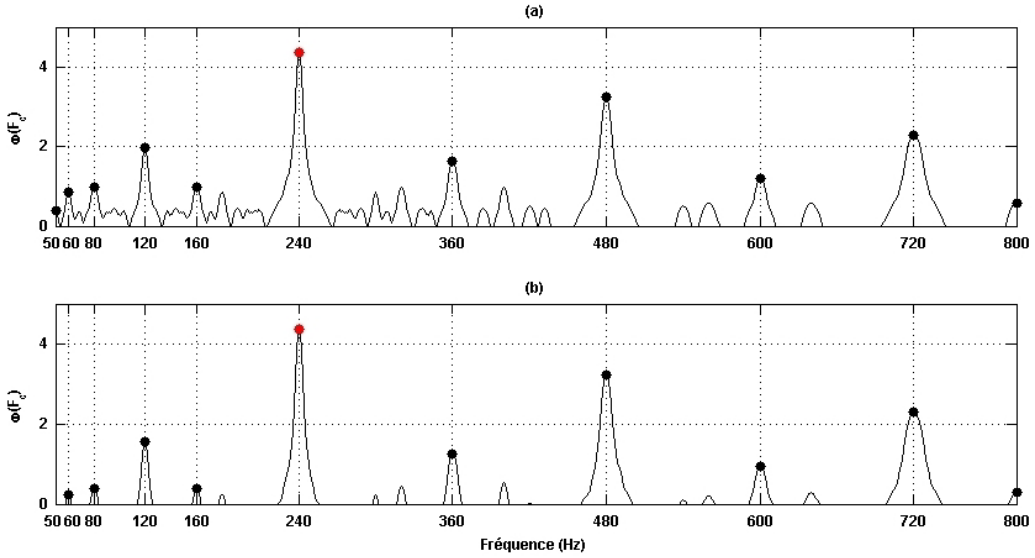

FIGURE 4.20: Fonctions PDF monopitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

4.3.4.2 Situation bipitch

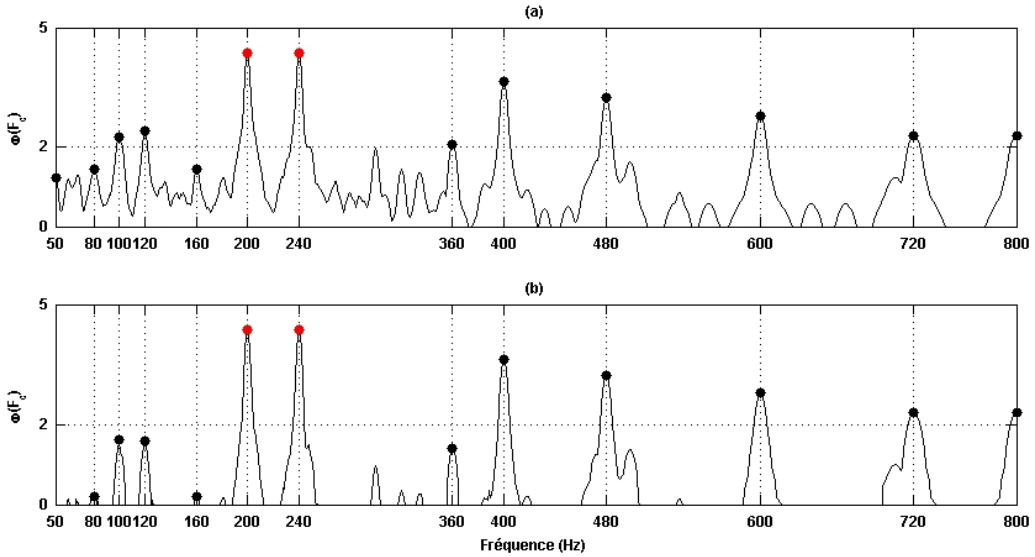

FIGURE 4.21: Fonctions PDF bipitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

La figure 4.21 présente les fonctions de pitch obtenues par le PDF dans la situation bipitch. Le graphique (a) correspond à la fonction de pitch obtenue sans correction hyperbolique et le graphique (b) est obtenu avec la correction hyperbolique. L'estimation de F_0 est correcte dans les deux cas car les deux premiers pics en terme d'amplitude sont ceux de 200 et 240Hz (points rouges). Les pics parasites sur-harmoniques restent importants ce qui est dû à la limitation du nombre de dents intrinsèque au PDF.

4.3.4.3 Situation 4-pitch

La figure 4.22 présente les fonctions de pitch obtenues par le PDF dans la situation 4-pitch. Le graphique (a) concerne le réglage sans correction hyperbolique. Le graphique (b) présente le résultat obtenu avec correction hyperbolique. L'estimation de F_0 est presque correcte. Toutefois le pic parasite en 400Hz est de même amplitude que les quatre pics de F_0 (points rouges). Le PDF s'avère être

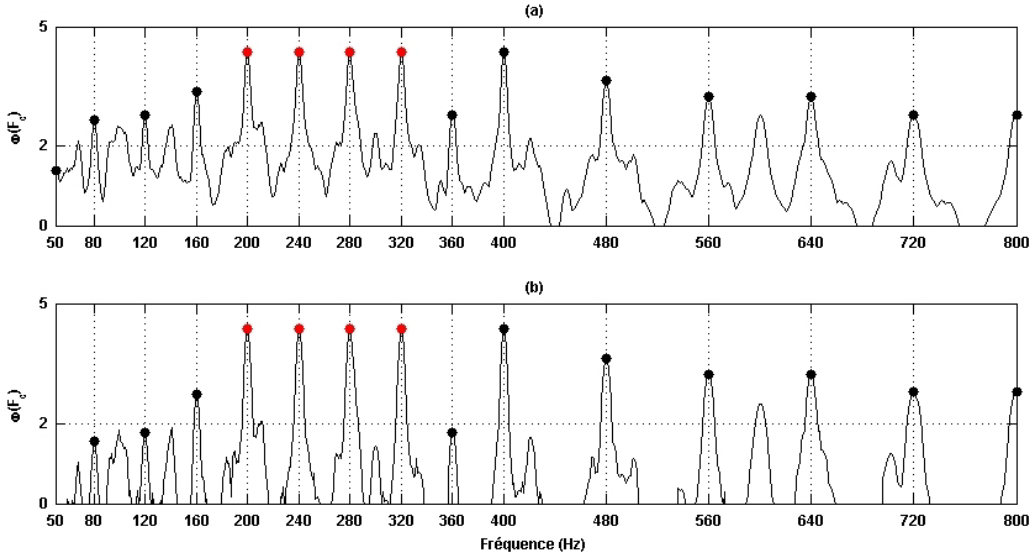

FIGURE 4.22: Fonctions PDF 4-pitch. (a) Sans correction hyperbolique. (b) Avec correction hyperbolique.

moins performant que le PDI. Ceci s'explique par le fait que le réglage utilisé (8 dents et décroissance en racine inverse) est sous-optimal et privilégie trop l'atténuation sous-harmonique. Le PDF profite d'ailleurs de la limitation du nombre de dents pour être plus performant que le PDI dans l'atténuation sous-harmonique. Ceci induit qu'il est moins bon en terme d'émergence sur-harmonique. Le compromis entre les deux émergences est confirmé. Même s'il n'est pas utilisé avec son meilleur réglage en terme d'équilibre entre atténuation sous et sur-harmonique, le PDF va servir de référence pour l'étude du Peigne Alterné (PAL) de la section 4.3.5.

4.3.5 Peigne Alterné (PAL)

Le Peigne Alterné (PAL) décrit dans l'article [Liénard et al., 2007] est un PDF auquel sont ajoutées des dents négatives. De par sa proximité avec le PDF du paragraphe 4.3.4, il va lui être comparé. Les dents négatives permettent de réduire les pics parasites sur-harmoniques (cf. section 4.4 pour explications) dont les amplitudes sont relevées par la limitation du nombre de dents. La figure 4.23 présente le PAL. Le PAL est un peigne limité à 8 dents avec une décroissance des dents en racine inverse ($1/\sqrt{k}$ avec k le numéro de la dent du peigne). L'amplitude des dents négatives d'ordre 2 vaut $a2 = 0.4$ et l'amplitude des dents négatives d'ordre 3 vaut $a3 = 0.2$. Ces valeurs ont été trouvées empiriquement et ont minimisé l'erreur d'estimation de F_0 sur le corpus de Keele (cf. tableau 4.2).

4.3.5.1 Situation monopitch

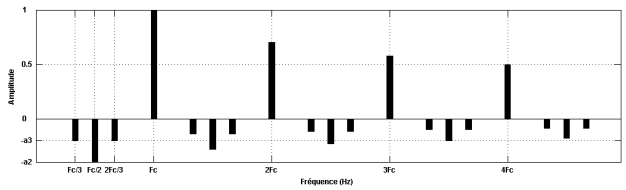

FIGURE 4.23: Exemple de Peigne Alterné. $a2$ et $a3$ représentent respectivement les amplitudes négatives des dents négatives d'ordre 2 et 3.

La figure 4.24 présente la fonction de pitch obtenue par le PAL sur le signal exemple. La fonction de pitch du PAL (rouge) est comparée avec la fonction de pitch obtenue avec un PDF (noir). Ceci permet de mettre en évidence l'apport des dents négatives sur les émergences sous et sur-harmonique par rapport à la version du PDF donnée auparavant. La valeur théorique des amplitudes est complexe à donner. Le PAL est la combinaison additive de plusieurs PDF : un PDF correspondant aux dents positives aux multiples de F_c, un PDF correspondant au dents négatives aux multiples de $F_c/2$ mais avec les dents de numéro pair manquantes et un dernier PDF aux multiples de $F_c/3$ avec les dents de numéro multiples de 3 manquantes. La formule théorique intègre cette combinaison linéaire avec les coefficients $a2$ et $a3$. Le PAL contient des dents négatives qui ont donc pour avantage de réduire les pics parasites sur-harmoniques. Toutefois, elles n'ont pas d'effet sur les pics parasites sous-harmoniques qui sont les plus gênants. Le PAL est un peigne qui donne de bons résultats lorsqu'il s'agit de ne détecter qu'une seule F_0 comme le témoigne le tableau 4.2. Les performances du PAL ont été testées pour certains couples $(a2, a3)$ sur la base de donnée Keele décrite dans le chapitre 6 sous-paragraphe 6.5.1.1. Il s'agit d'une évaluation monopitch qui porte sur environ cinq minutes trente secondes de parole dont environ la moitié est voisée. VnV correspond à l'erreur de décision Voisé/non-Voisé (combinant les erreurs de vous-voisé et les erreurs de sur-voisé) et GER_G correspond à un taux d'erreurs grossières qui représente un écart de plus de 20% en valeur absolue relative entre l'hypothèse F_0 donnée par le PAL et une valeur F_0 de référence (cf. chapitre 6 pour les détails sur les erreurs de sous/sur-voisé et le GER_G). La taille des trames est de 40ms.

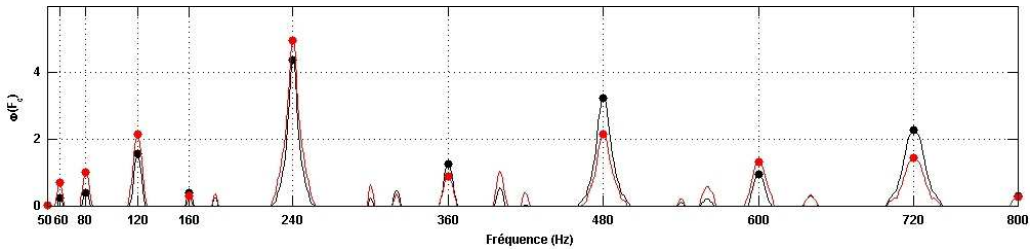

FIGURE 4.24: Comparaison entre la fonction PAL (rouge) et la fonction PDF (noir) en situation monopitch. La fonction PDF est donnée en figure 4.20.

4.3.5.2 Situation bipitch

Le PAL n'a pas été conçu pour traiter des situations multipitch. Il est néanmoins intéressant de voir les fonctions de pitch du PAL dans cette situation. La figure 4.25 présente la fonction de pitch obtenue

TABLE 4.2: Performances du PAL pour différents couples $(a2, a3)$. Testé sur la base de donnée Keele, version avec 10 dents. Source : [Liénard et al., 2007].

	VnV(%)	**GER_G (%)**
PDF $(a2 = a3 = 0)$	13.51	3.23
PAL $(a2 = 0.8, a3 = 0)$	12.04	1.99
PAL $(a2 = 0, a3 = 0.4)$	12.84	3.00
PAL $(a2 = 0.4, a3 = 0.2)$	12.35	1.85

dans la situation bipitch utilisée précédemment. L'estimation des F_0 deux est correcte. Comparé au PDF, le PAL permet de rééquilibrer les amplitudes des pics parasites sous et sur-harmoniques. Les pics parasites sous-harmoniques sont augmentés et les pics parasites sur-harmoniques sont atténués. Ceci se traduit par une émergence relative des pics F_0 à estimer de meilleure qualité (sur l'exemple considéré).

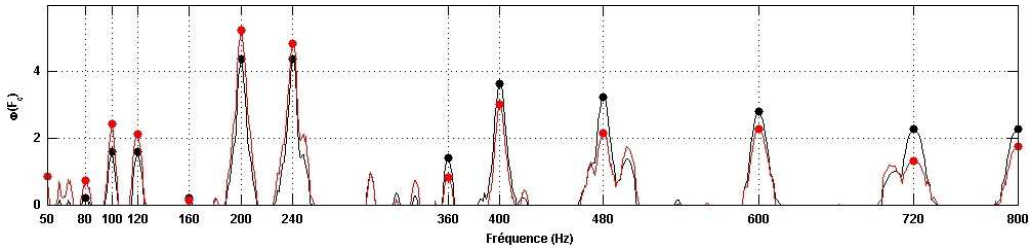

FIGURE 4.25: Comparaison entre la fonction PAL (rouge) et la fonction PDF (noir) en situation bipitch. La fonction PDF est donnée en figure 4.21.

4.3.5.3 Situation 4-pitch

La figure 4.26 présente la fonction de pitch obtenue par un PAL en situation 4-pitch. La courbe rouge est la fonction de pitch du PAL et la courbe noire est la fonction PDF avec correction hyperbolique. L'estimation de F_0 est correcte. Toutefois, le pic en 400Hz reste un concurrent très sérieux et n'est d'ailleurs presque pas diminué. Les pics en 240Hz et 320Hz sont légèrement diminués ce qui peut être gênant. Ceci est du à une interaction entre les deux structures harmoniques dont les F_0 (F_{02} et F_{04}) sont en relation $(3, 4)$ (ie. $240 = 3/4.320Hz$). Lorsque la F_c du PAL vaut 240Hz, alors les dents négatives d'ordre 3 du PAL sont placées en 80Hz, 160Hz, **320Hz**, 400Hz, 560Hz, **640Hz**, etc. Ceci implique que les dents négatives se retrouvent alignées avec les partiels du spectre d'amplitude provoquant une réduction du produit scalaire. De tous les peignes vus jusqu'à présent, le PAL est celui qui offre la meilleure atténuation des pics parasites dans toutes les situations. Néanmoins, il souffre comme les autres du compromis entre l'atténuation sous et sur-harmonique malgré l'ajout des dents négatives. De plus, lorsque les F_0 à estimer sont en relation $(2, q)$ ou $(3, q)$, les pics des F_0 sont légèrement atténués car les dents négatives du PAL se retrouvent alignées avec des harmoniques du spectre d'amplitude. Ce phénomène n'est pas désirable.

Chercher à réduire les pics parasites sous et sur-harmonique par un seul et unique peigne est

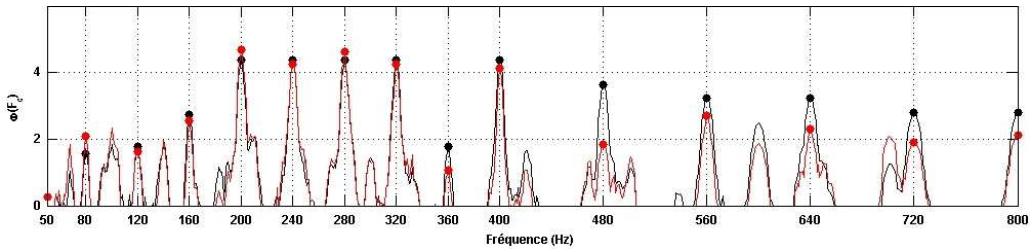

FIGURE 4.26: Comparaison entre la fonction PAL (rouge) et la fonction PDF (noir) en situation 4-pitch. La fonction PDF est donnée en figure 4.22.

probablement vain quelque soit le raffinement du peigne. Il souffrira intrinsèquement du compromis entre l'atténuation sous et sur-harmonique. Une question découle de cette observation. Est-il possible de trouver une forme de peigne capable de rendre indépendante la gestion des pics sur-harmoniques de celle des pics sous-harmoniques ? La section 4.4 présente le Peigne à Dents Négatives (PDN) reprenant les dents négatives du PAL pour atténuer les pics parasites sur-harmoniques. La section 4.5 présente le Peigne à Dents Manquantes (PDM) dont l'objectif est de traiter les pics parasites sous-harmoniques. Il s'avère que ces deux familles de peignes spectraux permettent de s'affranchir du compromis sur l'atténuation des pics parasites.

4.4 Peignes à Dents Négatives (PDN)

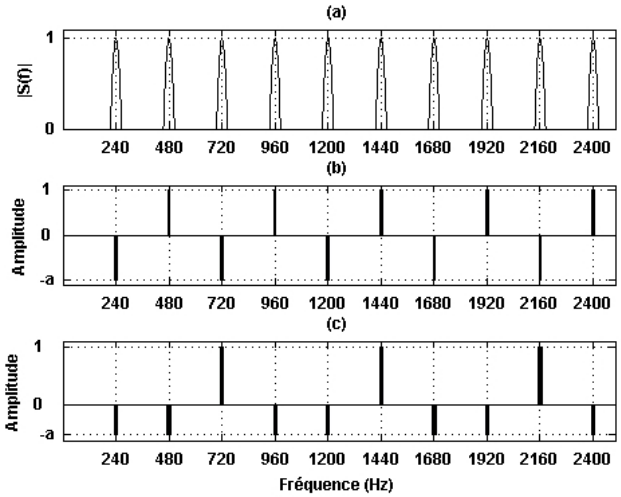

FIGURE 4.27: Illustration de PDN. (a) Spectre d'amplitude monopitch exemple. (b) PDN d'ordre 2. (c) PDN d'ordre 3.

Un PDN est un peigne au nombre de dents illimité, sans décroissance et défini par son ordre noté Δ. Un PDN d'ordre Δ est noté $PDN\Delta$. Δ est un entier naturel supérieur ou égal à 2. La figure 4.27 présente deux PDN : le PDN2 dans le graphique (b) et le PDN3 dans le graphique (c). Le graphique (a) présente un spectre d'amplitude de F_0 à 240Hz avec 10 harmoniques. Un PDN2 comporte une dent négative pour une dent positive, cette dent est placée en $\pm F_c/2$. Un PDN d'ordre 3 comporte deux dents négatives pour une dent positive. Ces dents négatives sont placées en $\pm 2F_c/3$ et $\pm F_c/3$. L'amplitude des dents négatives notée a dans les graphiques (a) et (b) est réglée de façon à rendre la somme des dents négatives et positives nulles. L'équation 4.15 indique la valeur a de l'amplitude des dents négatives en fonction de l'ordre Δ du PDN.

$$a = \frac{1}{\Delta - 1} \tag{4.15}$$

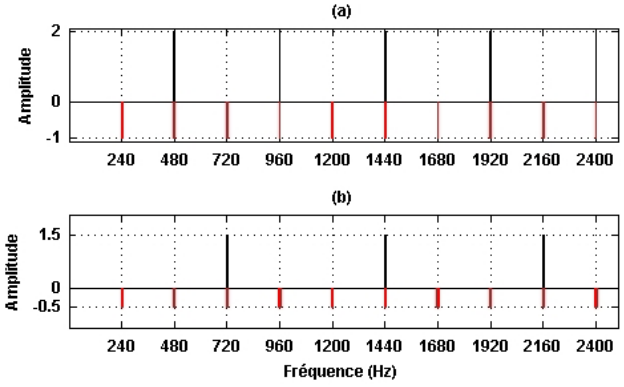

FIGURE 4.28: Illustration de PDN comme un combinaison linéaire de PUI. (a) PDN2. (b) PDN3.

Lorsque la F_c du PDN2 vaut $2F_0$ comme dans le graphique (b), le produit scalaire entre le PDN2 et le spectre vaut 0. De la même manière, lorsque la F_c du PDN3 vaut $3F_0$ comme dans le graphique (c), le produit scalaire vaut 0. Pour le PDN2 et PDN3, en fixant la fréquence caractéristique F_c à F_0, le produit scalaire est équivalent à celui obtenu par un PUI. Toutes Ces caractéristiques sont très intéressantes car elles montrent qu'un $PDN\Delta$ permet de mettre à zéro sélectivement les pics parasites sur-harmoniques. De manière générale, tous les pics parasites $(\Delta p, q)$ sont traités par un $PDN\Delta$. Les autres pics gardent la même amplitude qu'ils auraient avec un PUI. Ainsi, un PDN2 supprime les pics parasites de type $(2p, q)$ avec p, q entiers et $(2p/q)$ irréductible. De la même façon, le PDN d'ordre 4 supprime les pics parasites $(4q, p)$ mais ce PDN est inutile car les pics $(4q, p)$ sont complètement inclus dans les pics $(2q, p)$. L'ordre 4 est donc complètement inclus dans l'ordre 2. Le PDN d'ordre 5 annule les pics $(5p, q)$. L'ordre 6 est inutile car à la fois inclus dans l'ordre 2 et l'ordre 3 et ainsi de suite. En continuant le raisonnement, il apparaît que seuls les PDN dont l'ordre est un nombre premier (soit 2, 3, 5, 7, 11 etc.) sont utiles pour supprimer l'ensemble des pics sur-harmoniques et un certain nombre de pics sous-harmoniques dans lesquels $(p < q)$ soit par exemple $(2, 3)$, $(2, 5)$, etc. pour l'ordre 2 et $(3, 4)$, $(3, 5)$, etc. pour l'ordre 3. Ces considérations s'accordent avec le développement théorique donné dans l'article de Klapuri [Klapuri, 1998]. Dans les travaux de Sun [Sun, 2000, 2002] présentés dans le chapitre 3 sous-paragraphe 3.3.2.5, l'auteur parle de sous-harmoniques (terme auquel il conviendrait de préférer celui d'inter-harmonique). Sans que l'auteur ne l'indique clairement, il utilise le principe d'un PDN d'ordre 2. Un PDN d'ordre Δ est une combinaison linéaire de deux PUI. Cette combinaison linéaire est donnée dans l'équation 4.16. La figure 4.28 illustre cette combinaison linéaire pour le PDN2 et le PDN3 dans les graphiques (a) et (b) respectivement. Dans les deux graphiques, le PDN est le résultat de la somme du peigne noir et du peigne rouge. Le produit scalaire étant un opérateur linéaire, la fonction de pitch d'un PDN obéit donc à la même combinaison linéaire que celle décrite dans l'équation 4.16. Il est donc possible de donner les valeurs théoriques des amplitudes d'un PDN d'ordre Δ. L'annexe A.5 décrit dans le détail les calculs pour arriver aux deux résultats de l'équation 4.17.

$$PDN\Delta(F_c, f) = \frac{\Delta}{\Delta - 1} PUI(F_c, f) - \frac{1}{\Delta - 1} PUI\left(\frac{F_c}{\Delta}, f\right) \tag{4.16}$$

Il faut considérer que p, q sont des entiers non nuls et qu'ils sont premiers entre eux. Le raisonnement porte sur un spectre d'amplitude idéal modélisé par un peigne de Dirac et sans partie négative entre les dents du peigne.

$$\begin{array}{ll} si \frac{p}{q} = \frac{p'\Delta}{q} & -1 <= \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) <= 0 \\ si \frac{p}{q} \neq \frac{p'\Delta}{q} & \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) = \Phi_{PUI}\left(\frac{p}{q}F_0\right) \end{array} \tag{4.17}$$

Cela résultat théorique justifie les annulations des pics de type $(p\Delta, q)$ puisque dans ce cas, la valeur théorique de la fonction de pitch est bornée dans sa valeur supérieure par 0. Un $PDN\Delta$ réduit les pics parasites $(p\Delta, q)$ à une amplitude négative. Les autres pics parasites restent intacts. Il convient maintenant d'étudier le comportement des PDN en situation monopitch et multipitch.

4.4.1 Situation monopitch

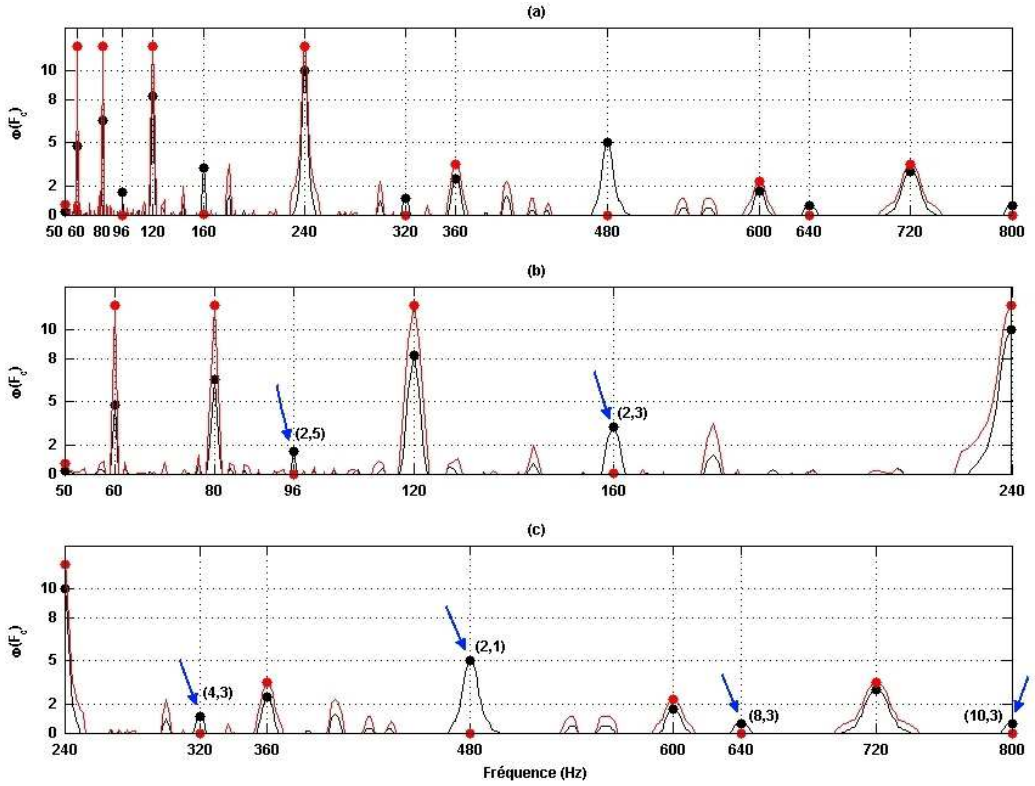

FIGURE 4.29: Comparaison des fonctions PDN2 (rouge) et fonctions PUI (noir) en situation monopitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

La fonction PDN2 est présentée dans la figure 4.29. Le graphique (a) présente la fonction de pitch obtenue par un PDN2 sur l'intégralité de l'intervalle de recherche $[50, 800Hz]$. Le graphique (b) est un zoom sur la zone sous-harmonique $[50, 240Hz]$. Le graphique (c) est un zoom sur la partie sur-harmonique $[240, 800Hz]$. La courbe noire de chacun des graphiques est la fonction PUI avec correction hyperbolique et fréquence de coupure F_{MAX} fixée à la valeur de la fréquence centrale de la dernière harmonique à laquelle on ajoute la moitié de l'épaisseur fréquentielle $b_w/2$. Un premier résultat semble en contradiction avec le résultat théorique de l'équation 4.17 car les pics sous-harmoniques en 60, 80 et 120Hz sont augmentés par le PDN2 alors que selon la théorie ils devraient être de même amplitude. Cette différence s'explique car la preuve théorique porte sur un spectre idéal sans parties négatives entre les harmoniques. Or, dans l'exemple utilisé ici le spectre d'amplitude est de moyenne nulle et possède donc des parties négatives. Lorsque des dents négatives du PDN se trouvent dans les zones négatives, le résultat du produit entre l'amplitude négative de la dent du PDN2 et de l'amplitude

négative du spectre donne une valeur positive et augmente donc le résultat final du produit scalaire. Ce phénomène se produit tout particulièrement dans les basses fréquences. Ce phénomène n'est pas important car Le comportement intéressant et utile du PDN est celui d'atténuation des dents et pas celui d'augmentation de l'amplitude. Dans l'AEP que nous proposons, seule l'atténuation est prise en compte. Cette augmentation des amplitudes se retrouvera dans tous les PDN et tous les PDM (cf. section 4.5) mais ne nuit pas à notre AEP. Le graphique (b) est un zoom sur la zone au-dessous de la F_0 $[50, 240Hz]$. Cette zone sous-harmonique correspond à des couples (p, q) dans lesquels q est supérieur à p. Le graphique (c) est un zoom sur la zone au-dessus de la F_0 $[240, 800Hz]$. Cette zone sur-harmonique correspond à des couples (p, q) pour lesquels p est supérieur à q. Dans le graphique (b), les pics en 96Hz et 160Hz sont particulièrement intéressants. Ils correspondent respectivement aux pics $(2, 5)$ et $(2, 3)$ et sont effectivement annulés par le PDN d'ordre 2. Dans le graphique (c), les pics en 320Hz, 480Hz et 640Hz. Ils correspondent respectivement aux pics $(4, 3)$, $(2, 1)$ et $(8, 3)$ et sont annulés.

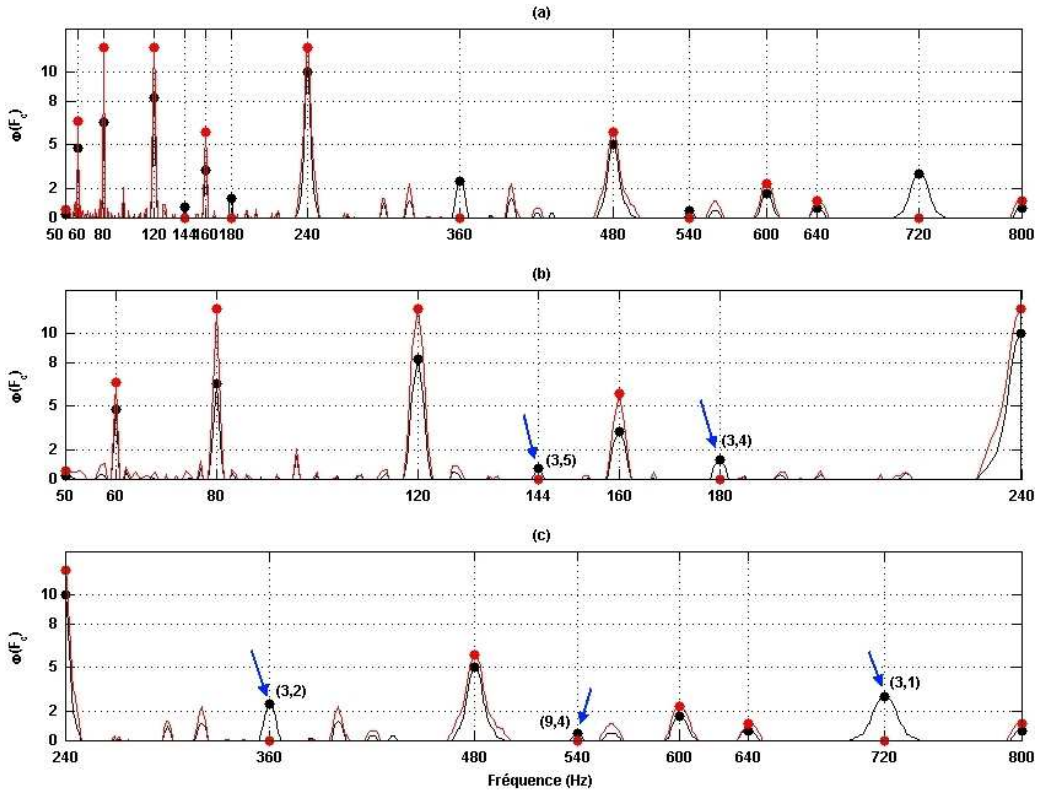

FIGURE 4.30: Comparaison des fonctions PDN3 (rouge) et fonctions PUI (noir) en situation monopitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

Les amplitudes de tous ces pics sont soit nulles soit négatives d'où le terme d'annulation. Le PDN2 permet bien de détecter spécifiquement les pics parasites répondant à l'indexation $(2p, q)$. La figure 4.30 présente la fonction PDN3. Le graphique (a) présente la fonction de pitch obtenue par un PDN3 sur l'intégralité de l'intervalle de recherche $[50, 800Hz]$. Le graphique (b) est un zoom sur la zone sous-harmonique $[50, 240\text{Hz}]$. Le graphique (c) est un zoom sur la partie sur-harmonique $[240, 800Hz]$. Dans le graphique (b), les pics en 144Hz et 180Hz sont annulés car ils correspondent respectivement aux

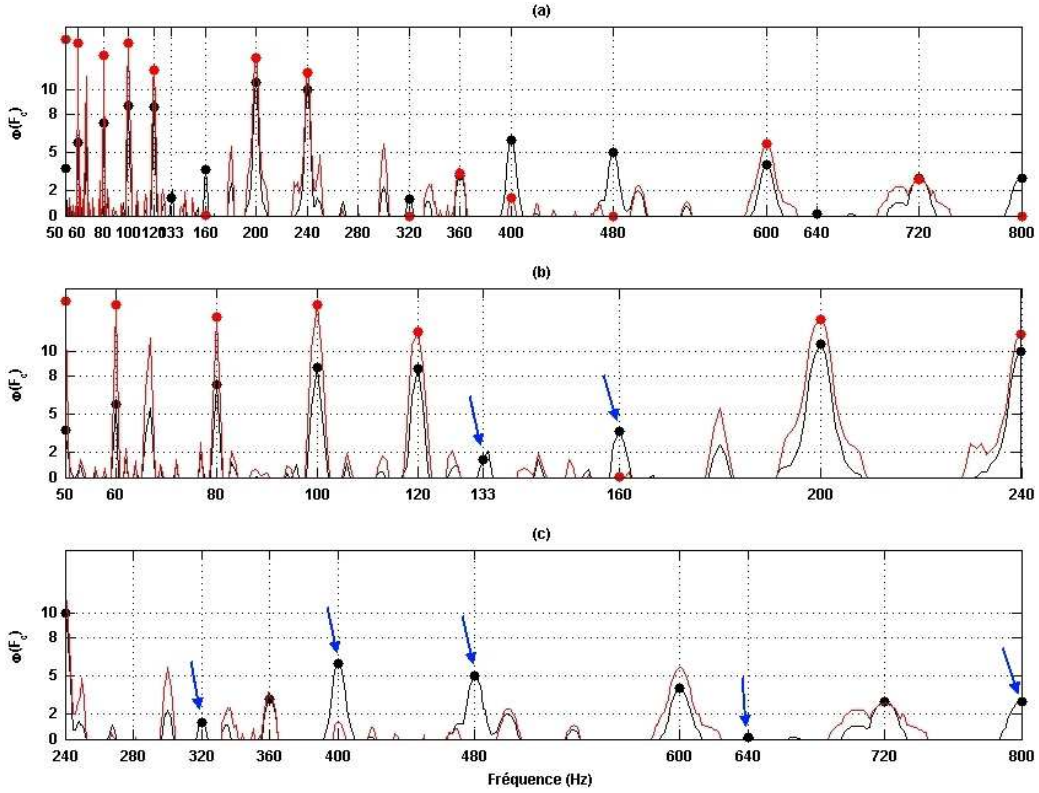

FIGURE 4.31: Comparaison des fonctions PDN2 (rouge) et fonctions PUI (noir) en situation bipitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

pics $(3,5)$ et $(3,4)$ et sont donc sensibles au PDN d'ordre 3. Dans le graphique (c), les pics en 360Hz, 540Hz et 720Hz sont annulés. Ils correspondent respectivement aux pics $(3,2)$, $(9,4)$ et $(3,1)$ et sont annulés. Le PDN3 permet effectivement la détection spécifique des pics parasites de type $(3p,q)$. Un PDN parvient-t-il encore à supprimer les pics parasites sur-harmoniques dans la situation multipitch ?

4.4.2 Situation bipitch

En situation bipitch, la figure 4.31 présente les fonctions de pitch obtenues par un PDN d'ordre 2 et 3. Le graphique (a) présente la fonction PDN2 sur tout l'intervalle de recherche. Le graphique (b) est un zoom sur la partie sous-harmonique et le graphique (c) est un zoom sur la partie sur-harmonique. La courbe noire est la fonction PUI. Les pics fléchés en bleu indiquent les pics parasites qui sont nettement atténués. Ils correspondent en effet à des pics parasites en relation $(2p,q)$ avec soit F_{01} soit F_{02}. La figure 4.32 présente la fonction PDN3. Le graphique (a) illustre son intégralité, le graphique (b) montre un zoom sur la partie sous-harmonique et le graphique (c) est la fonction PDN3 sur la zone sur-harmonique. Les pics en relation $(3p,q)$ sont fléchés en bleu dans les graphiques et démontrent le bon fonctionnement du PDN3. Les pics de la forme $(2p,q)$ pour le PDN2 et les pics $(3p,q)$ pour le PDN3. Il faut noter l'échec pour le pic $(2,5)$ de F_{01} (soit 80Hz) ce qui s'explique par le fait que 80Hz est un sous-multiple de F_{02} et cette relation empêche l'atténuation. Cet échec isolé n'est pas critique car l'estimation de F_0 ne repose pas sur un seul PDN mais sur l'ensemble des PDN et PDM de tous

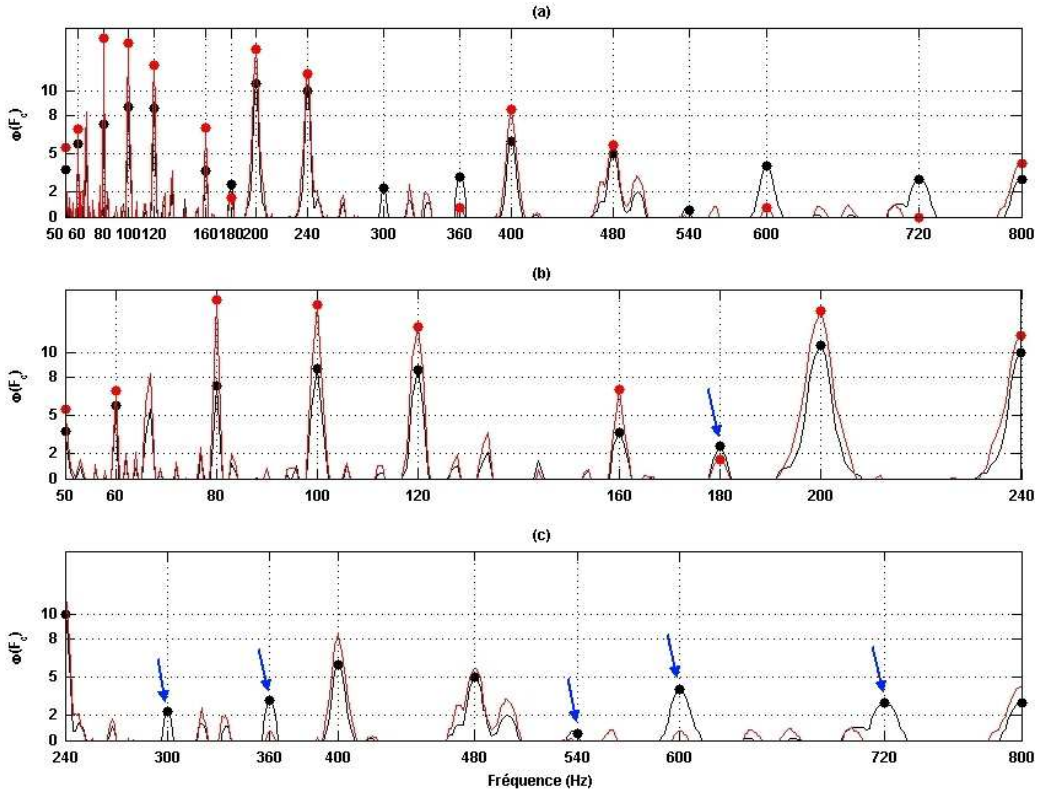

FIGURE 4.32: Comparaison des fonctions PDN3 (rouge) et fonctions PUI (noir) en situation bipitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

ordres premiers. L'ensemble des PDN et PDM devrait pourvoir le détecter comme pic parasite.

4.4.3 Situation 4-pitch

La figure 4.33 présente la fonction PDN2 dans la situation 4-pitch. La fonction PDN2 est illustrée sur tout l'espace de recherche dans le graphique (a). Les zooms sous et sur-harmoniques sont respectivement illustrées dans les graphiques (b) et (c). Les courbes rouges des PDN sont comparées avec les courbes noires d'un PUI avec correction hyperbolique. Les pics fléchés en bleu sont les pics traités et atténués par le PDN2. Le PDN2 joue toujours son rôle atténuateur même en situation 4-pitch. Toutefois, la suppression des pics parasites de type $(\Delta p, q)$ avec $q > 2\Delta$ donc sous-harmonique est moins performante. Il faut considérer qu'en situation de mélange, les PDN permettent de supprimer les pics parasites sur-harmoniques de manière fiable. L'annulation des pics sous-harmoniques de type $(\Delta p, q)$ avec $q > \Delta p$ est soumise aux relations entre les structures harmoniques mélangées et ne fonctionne pas nécessairement même si globalement dans cet exemple, la majorité de ces pics sont atténués. Le peigne à dents manquantes (PDM) qui fait l'objet de la section 4.5 suivante permet de les traiter. La figure 4.34 démontre également le bon fonctionnement du PDN3 en situation 4-pitch. Les pics fléchés en bleu sont clairement atténués et correspondent tous sans exception à des pics en relation $(3p, q)$ avec une des F_0 de la situation 4-pitch. Les pics sous-harmoniques en relation $(3p, q)$ ne parviennent plus a être atténués. Il peut être conclu que plus la situation est multipitch, moins le rôle atténuateur des

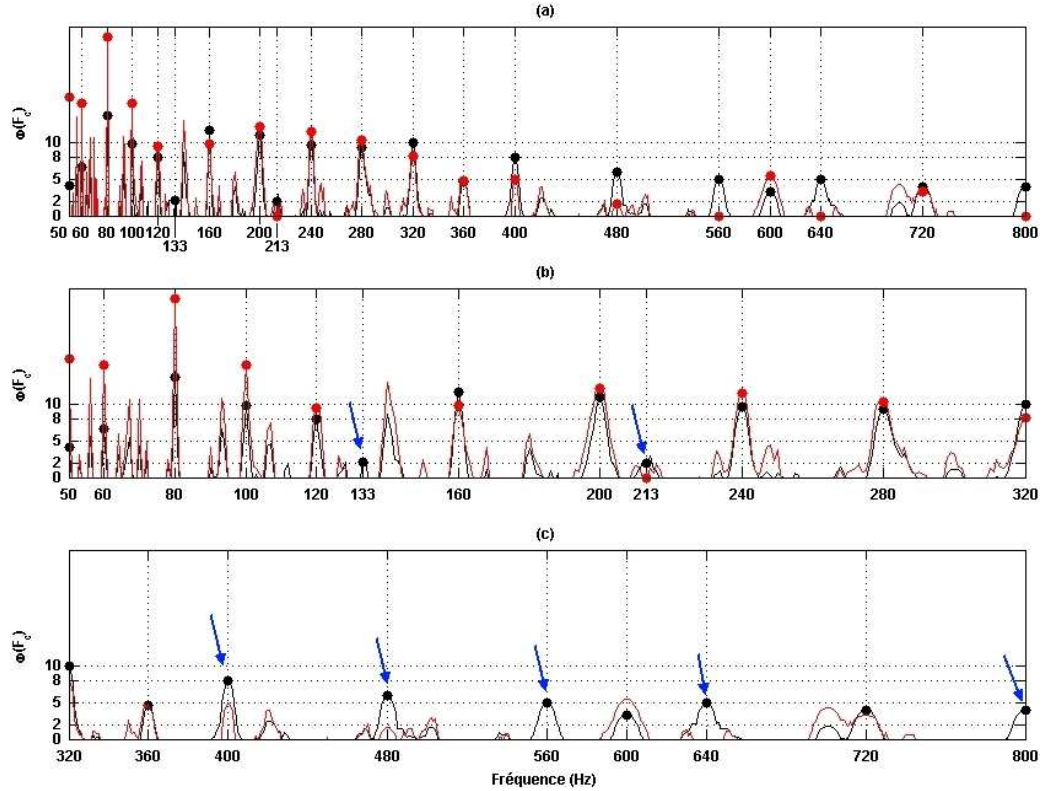

FIGURE 4.33: Comparaison des fonctions PDN2 (rouge) et fonctions PUI (noir) en situation 4-pitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

PDN est fiable dans les zones sous-harmoniques $3p < q$. Ceci n'est pas très important dans la mesure où les PDM vont traiter ces pics sous-harmoniques.

Nous sommes à présent dotés d'une famille de peignes capables d'atténuer les pics parasites de type $(\Delta p, q)$ qui correspondent majoritairement aux pics sur-harmoniques. Existe-t-il une famille de peignes spectraux capables d'atténuer les pics parasites sous-harmoniques ? La section 4.5 répond à cette interrogation par l'affirmative.

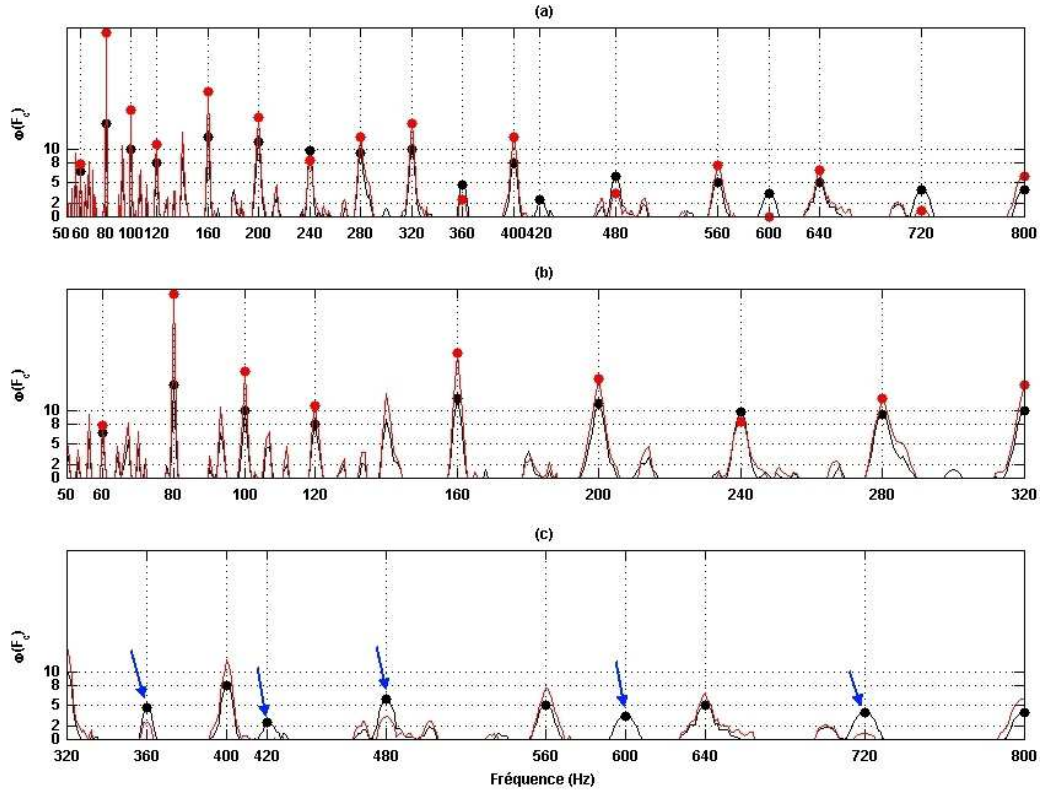

FIGURE 4.34: Comparaison des fonctions PDN3 (rouge) et fonctions PUI (noir) en situation 4-pitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

4.5 Peigne à Dents Manquantes (PDM)

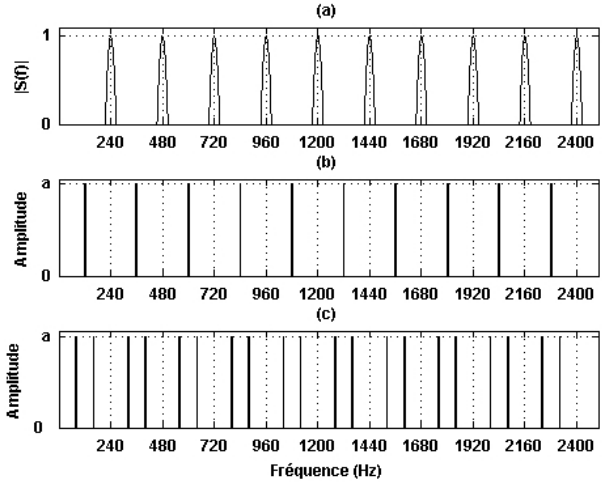

FIGURE 4.35: Illustration de PDM. (a) Spectre d'amplitude monopitch exemple. (b) PDM d'ordre 2. (c) PDM d'ordre 3.

Comme pour les PDN, les PDM sont définis par leur ordre Δ. Pour les mêmes raisons que les PDN, les seuls ordres utiles sont les ordres qui correspondent aux nombres premiers. La figure 4.35 présente un PDM d'ordre 2 dans le graphique (b). Le graphique (c) est un PDM d'ordre 3. Le graphique est un exemple idéal de spectre d'amplitude contenant 10 harmoniques. Un PDM d'ordre 2 met à zéro les dents paires et un PDM d'ordre 3 les dents dont le numéro est multiple de 3. L'amplitude a des dents non nulles est réglée de manière à ce que la somme des dents soit la même que pour le PUI de même Fc. Ainsi, s'il manque une dent sur deux, on multiplie par deux l'amplitude qui devient 2. S'il manque une dent sur trois, on multiplie par $3/2 = 1.5$ et ainsi de suite. L'équation 4.18 indique la valeur a des amplitudes des dents du PDM en fonction de l'ordre Δ.

$$a = \frac{\Delta}{\Delta - 1} \tag{4.18}$$

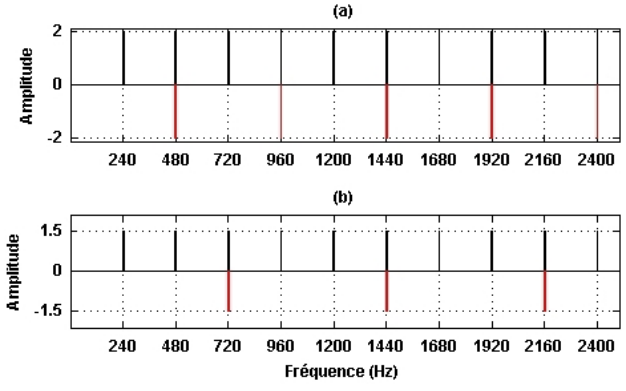

FIGURE 4.36: Illustration de PDM comme un combinaison linéaire de PUI. (a) PDM2. (b) PDM3.

La F_c du PDM2 dans le graphique (b) vaut $F_0/2$, le résultat du produit scalaire vaut 0 car les harmoniques du spectre tombent exactement dans les dents manquantes. De même, la F_c du PDM3 vaut $F_0/2$ et le produit scalaire vaut 0. Il est aisé d'imaginer la situation dans laquelle F_c est réglée à F_0. Le produit scalaire est alors identique à celui obtenu par un PUI. Ces résultats sont très intéressants car ils montrent que les PDM sont capables de supprimer spécifiquement les pics parasites sous-harmoniques. Dans la thèse de Camacho [Camacho, 2007], l'auteur propose une version de SWIPE nommée SWIPE' qui met à zéro les lobes positifs de sa fonction kernel (cf. chapitre 3 figure 3.22) dont le numéro est un nombre premier (2, 3, 5, etc.). Ce procédé est différent de celui proposé par le PDM mais le fait de retirer des dents du peigne se retrouve. Un $PDM\Delta$ est une combinaison linéaire de deux PUI comme le montre la figure 4.36. Le graphique (a) présente la combinaison linéaire pour le PDM2 et le graphique (b) illustre la combinaison linéaire pour le PDM3. Dans les deux cas, Le PDM est la somme

du PUI en noir et de celui en rouge. La fonction $PDM\Delta$ respecte la même combinaison décrite dans la figure 4.36. L'équation 4.19 en donne la formulation théorique.

$$PDM\Delta(F_c, f) = \frac{\Delta}{\Delta - 1}[PUI(F_c, f) - PUI(\Delta F_c, f)] \tag{4.19}$$

Une étude théorique démontre le comportement du $PDM\Delta$ sur le même spectre idéal (peigne de Dirac) utilisé pour prouver le comportement du PDN. La fonction $PDM\Delta$ théorique est décrite dans l'équation 4.20 et la preuve est donnée dans l'annexe A.6.

$$\begin{array}{ll} si \frac{p}{q} = \frac{p}{q'\Delta} & \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) = 0 \\ si \frac{p}{q} \neq \frac{p}{q'\Delta} & \Phi_{PUI}\left(\frac{p}{q}F_0\right) <= \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) <= \Phi_{PUI}\left(\frac{p}{q}F_0\right) + 1 \end{array} \tag{4.20}$$

Le $PDM\Delta$ met à zéro les pics de type $(p, \Delta q)$ et dans le reste des cas, il est presque équivalent à la fonction de pitch obtenue par un PUI.

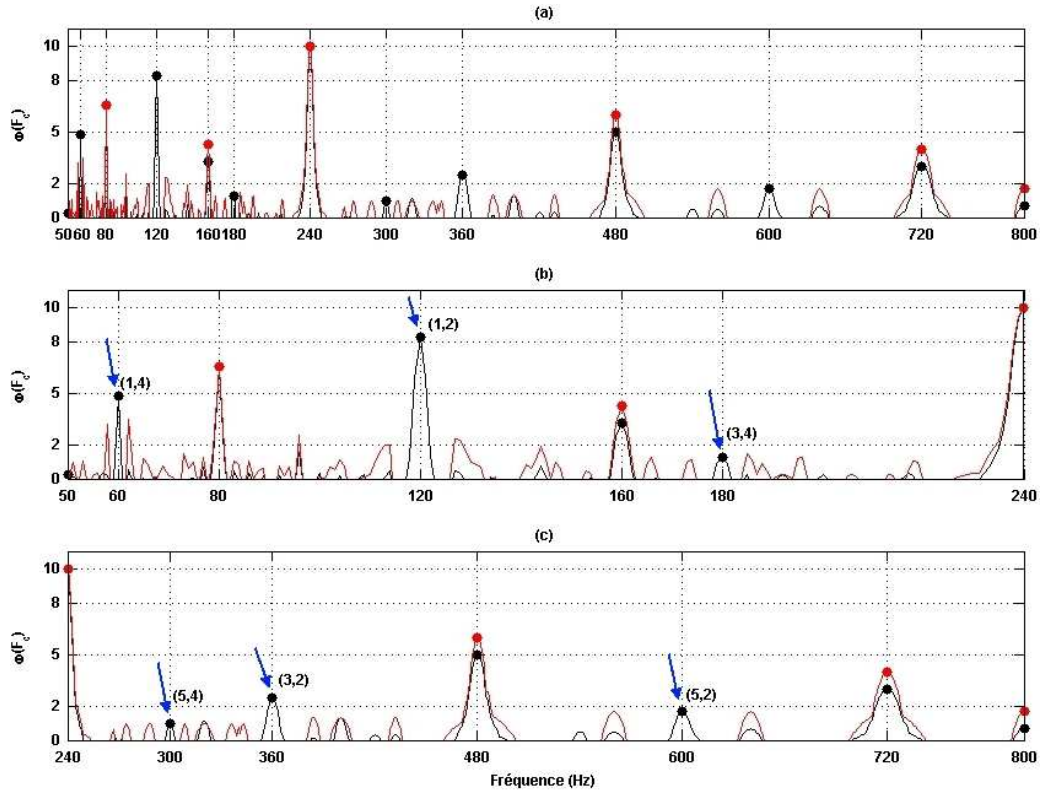

FIGURE 4.37: Comparaison des fonctions PDM2 (rouge) et fonctions PUI (noir) en situation monopitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

4.5.1 Situation monopitch

Le graphique (a) de la figure 4.37 présente la fonction de pitch obtenue par un PDM d'ordre 2. Le graphique (b) est la fonction de pitch obtenue avec le PDM d'ordre 3. La courbe noire de tous les graphiques est la fonction PUI avec correction hyperbolique. Le phénomène d'augmentation de certaines

amplitudes se retrouve comme avec les PDN. Il s'explique cette fois-ci non pas par la multiplication de deux valeurs négatives entre elles mais par l'amplitude des dents du PDM qui n'est pas unitaire. Pour compenser le manque de dents, les PDM multiplient l'amplitude des dents par un facteur supérieur à 1. Il se peut que le produit d'un nombre de dents restreint mais multiplié par un facteur supérieur à 1 soit supérieur à au produit d'un nombre de dents plus important mais non amplifié. Par exemple, le pic en 480Hz de la fonction PDM2 est supérieur à celui de la fonction PUI. Le pic de la fonction PUI vaut 5 car il y a 5 dents du PUI qui coïncident avec les sommets des harmoniques. Le pic de la fonction PDM2 vaut 6. Seule trois dents du PDM2 coïncident avec les harmoniques mais leur amplitude étant de 2, le résultat du produit scalaire vaut $2 * 3 = 6$ et est supérieur à 5. Là encore, ce phénomène n'est pas à pris en compte car la caractéristique utile est uniquement celle d'atténuation. Les pics parasites annulés sont fléchés en bleu. Le PDM2 a bien atténués tous les pics de type $(p, 2q)$. La figure 4.38 présente la fonction PDM3. Le graphique (a) la présente sur tout l'intervalle de recherche de F_0 alors que le graphique (b) présente un zoom sur la partie sous-harmonique et que le graphique présente un zoom sur la partie sur-harmonique. La courbe noire de chacun des graphiques correspond à la fonction PUI avec correction hyperbolique. Les pics parasites sous-harmoniques clairement supprimés sont les

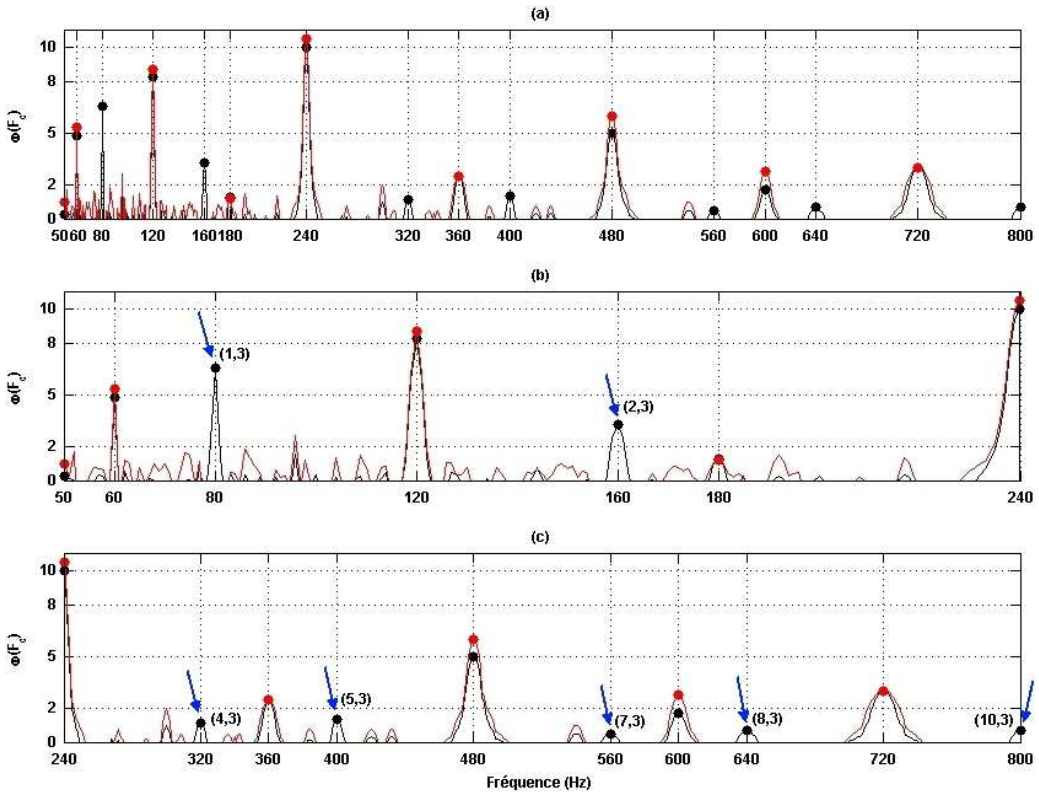

FIGURE 4.38: Comparaison des fonctions PDM3 (rouge) et fonctions PUI (noir) en situation monopitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

pics $(1, 3)$, $(2, 3)$ sont pointés par des flèches bleues. De même pour les pics parasites sur-harmoniques dans le graphique (c). Le PDM3 effectue bien l'annulation des pics de type $(p, 3q)$. Le PDM est-il encore fonctionnel en situation multipitch ?

4.5.2 Situation bipitch

Le graphique (a) de la figure 4.39 montre la fonction de pitch d'un PDM d'ordre 2 en situation bipitch sur l'intervalle [50,800Hz]. Le graphique (b) est un zoom sur la partie [50, 240Hz] et le graphique (c) est un zoom sur la partie [240,800Hz]. La courbe noire de tous les graphiques est la fonction PUI avec correction hyperbolique. Les pics parasites supprimés correspondent effectivement à des pics de type (p,2q) soit par rapport à F_{01} soit par rapport à F_{02}. L'amplitude des pics en F_{01} et F_{02} ne change quasiment pas. La figure 4.40 présente la fonction PDM3. Le graphique (a) montre l'intégralité de

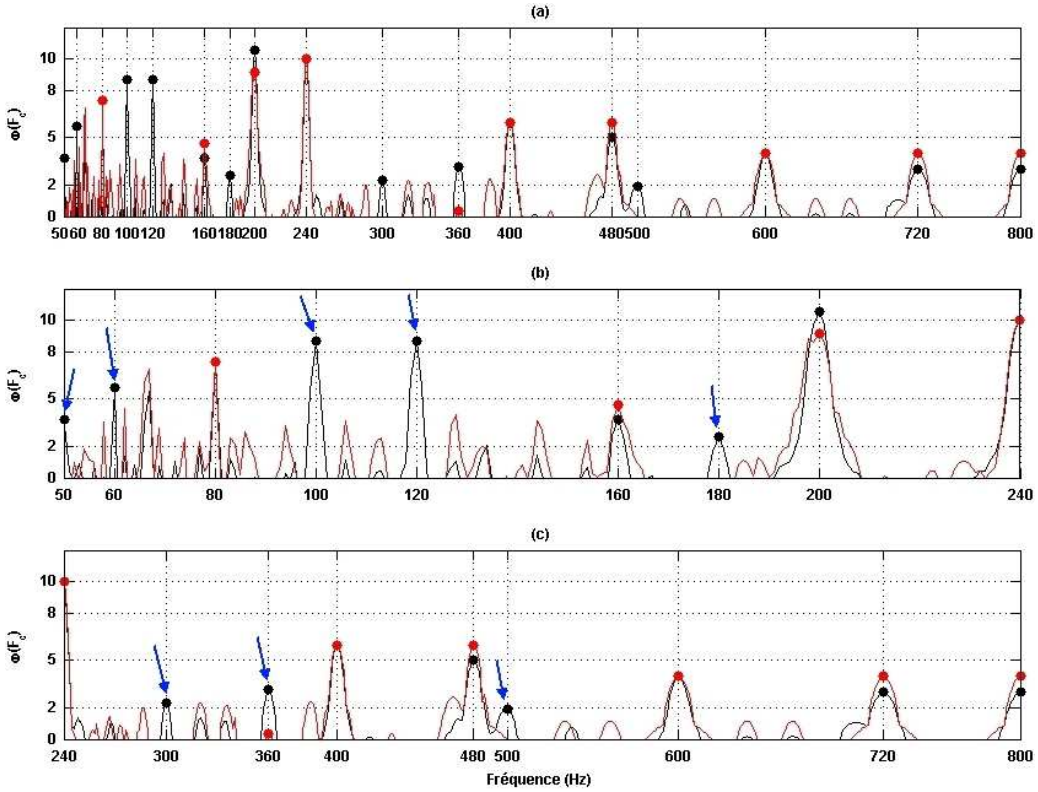

FIGURE 4.39: Comparaison des fonctions PDM2 (rouge) et fonctions PUI (noir) en situation bipitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

la fonction, le graphique (b) présente un zoom sur l'intervalle [50,240Hz] et le graphique (c) présente un zoom sur l'intervalle [240,800Hz]. La courbe noire de chaque graphique est la fonction PUI avec correction hyperbolique. Les pics parasites $(p, 3q)$ sont effectivement atténués même en situation bipitch (pics fléchés en bleu). Les amplitudes des F_0 ne changent quasiment pas ce qui les rend plus émergentes.

4.5.3 Situation 4-pitch

La figure 4.41 présente la fonction PDM2 en situation 4-pitch. Le graphique (a) la montre sur tout l'intervalle de recherche de F_0 et les graphiques (b) et (c) présentent des zooms respectivement sous et sur-harmoniques. La courbe noire est la fonction PUI avec correction hyperbolique. Il est à noter que la diminution des pics en 200Hz et 240Hz qui est gênante. Les pics parasites les plus importants sont bien

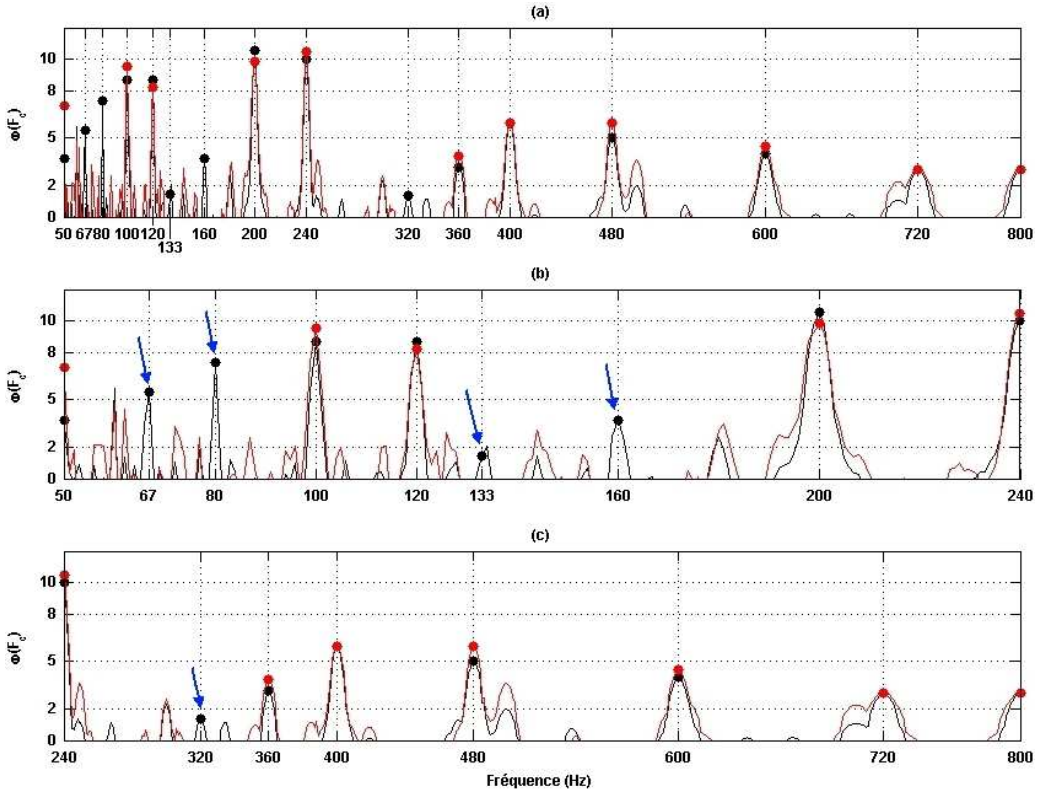

FIGURE 4.40: Comparaison des fonctions PDM3 (rouge) et fonctions PUI (noir) en situation bipitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

atténués et sont pointés par des flèches bleues. Leur position fréquentielle correspond effectivement à des relations de type $(p, 2q)$ avec au moins une des F_0 à estimer. La figure 4.42 présente la fonction PDM3 en situation 4-pitch. Son analyse démontre encore une fois l'atténuation (pics fléchés en bleu) des pics parasites correspondant aux ordres $(p, 3q)$. Les PDM parviennent encore à atténuer de nombreux pics parasites même en situation 4-pitch. Nous disposons donc de deux familles de peignes spectraux PDN et PDM. Les premiers peignes permettent d'atténuer les pics parasites de type $(\Delta p, q)$ et les seconds permettent d'atténuer les pics parasites de type de $(p, \Delta q)$. En combinant ces peignes spectraux de manière adéquate (multiplicative ou bien par un minimum), il est possible de mettre en place un AEP capable de produire des fonctions de pitch dont les pics parasites sont fortement atténués. Les pics parasites étant atténués, il devient possible d'envisager le fonctionnement d'un AEP fondé sur des peignes spectraux dans la situation multipitch. La combinaison des fonctions PDN et PDM est discutée dans la section 4.6. Cette combinaison est désignée comme étant le principe PSH ou principe de Peigne à Suppression harmonique.

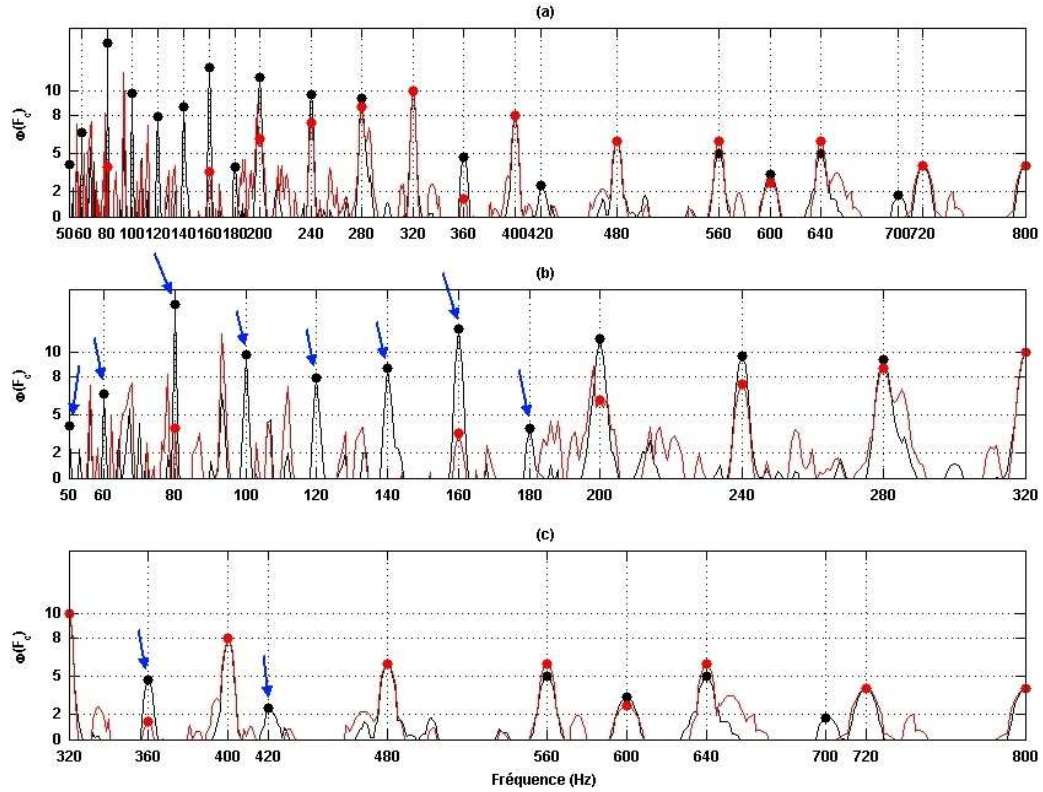

FIGURE 4.41: Comparaison des fonctions PDM2 (rouge) et fonctions PUI (noir) en situation 4-pitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

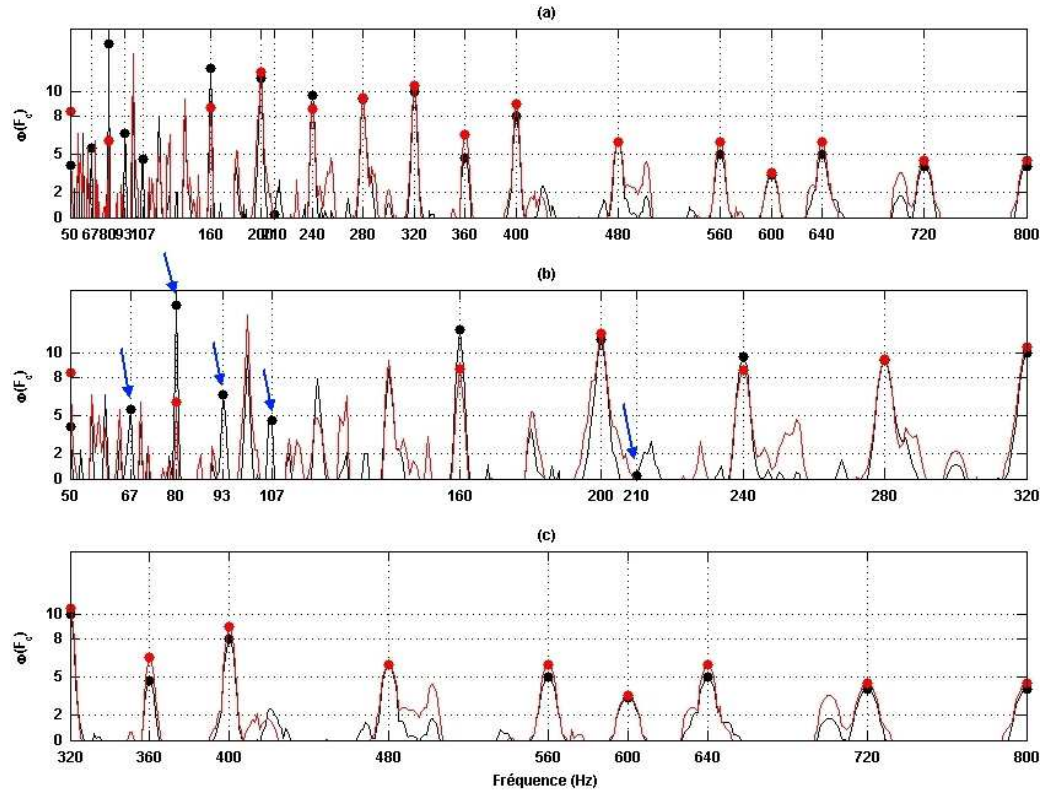

FIGURE 4.42: Comparaison des fonctions PDM3 (rouge) et fonctions PUI (noir) en situation 4-pitch. (a) Intervalle de recherche [50,800Hz]. (b) Partie sous-harmonique [50,240Hz]. (c) Partie sur-harmonique [240,800Hz].

4.6 Principe de Peigne à Suppression Harmonique

Le principe de Peigne à Suppression Harmonique (PSH) repose sur l'annulation des pics parasites des fonctions de pitch qui a été présenté dans [Liénard et al., 2008]. Il consiste à traiter isolément l'atténuation de chacun des pics parasites en utilisant les PDN et PDM à différents ordres. En combinant les résultats obtenus par chacun des PDN et PDM de tous ordres, il est possible de réduire considérablement l'amplitude des pics parasites. Les quatre points suivants présentent l'essence de PSH.

1. Calcul de la fonction PUI
2. Calcul des fonctions $PDN\Delta$ à différents ordres premiers
3. Calcul des fonction $PDM\Delta$ à différents ordres premiers
4. Calcul d'une fonction de sélection des pics F_0 par combinaisons des fonctions PUI, $PDN\Delta$ et $PDM\Delta$

Plusieurs implémentations de PSH existent et toutes respectent ces quatre étapes. Une implémentation possible nommée PSH_{min} et décrite dans le chapitre 5 paragraphe 5.2.2 donne les résultats suivants en situations monopitch et multipitch sur les exemples utilisés dans ce chapitre.

4.6.1 Situation monopitch

En situation monopitch, le graphique (a) de la figure 4.43 présente la fonction PUI avec correction hyperbolique et le graphique (b) présente la fonction PSH obtenue par l'algorithme PSH_{min} (cf. chapitre 5 paragraphe 5.2.2) en n'utilisant que les PDN et PDM d'ordre 2 et 3. La fonction PSH est calculée à partir du minimum des fonctions PDN2, PDN3, PDM2 et PDM3. Le résultat, certes obtenu

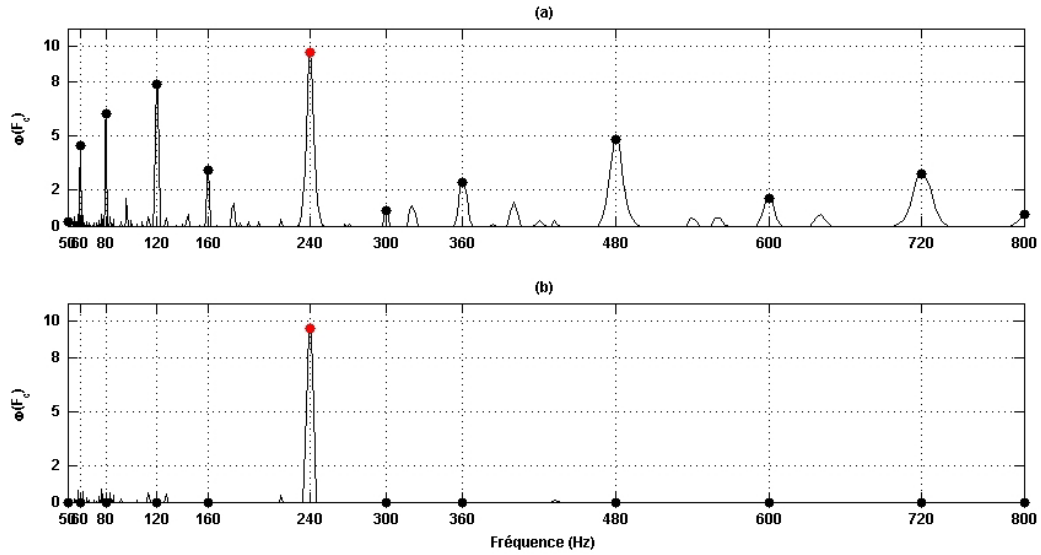

FIGURE 4.43: Comparaison des fonctions PUI et de la fonction PSH en situation monopitch. (a) Fonction PUI avec correction hyperbolique. (b) fonction PSH.

sur un spectre idéal, est clair. Le pic en F_0 émerge très largement du reste et la fonction PSH est faci-

lement utilisable. Seuls quelques pics parasites de très faibles amplitudes subsistent et principalement dans les basses fréquences. Il convient maintenant d'observer si ce principe est robuste en situation multipitch.

4.6.2 Situation bipitch

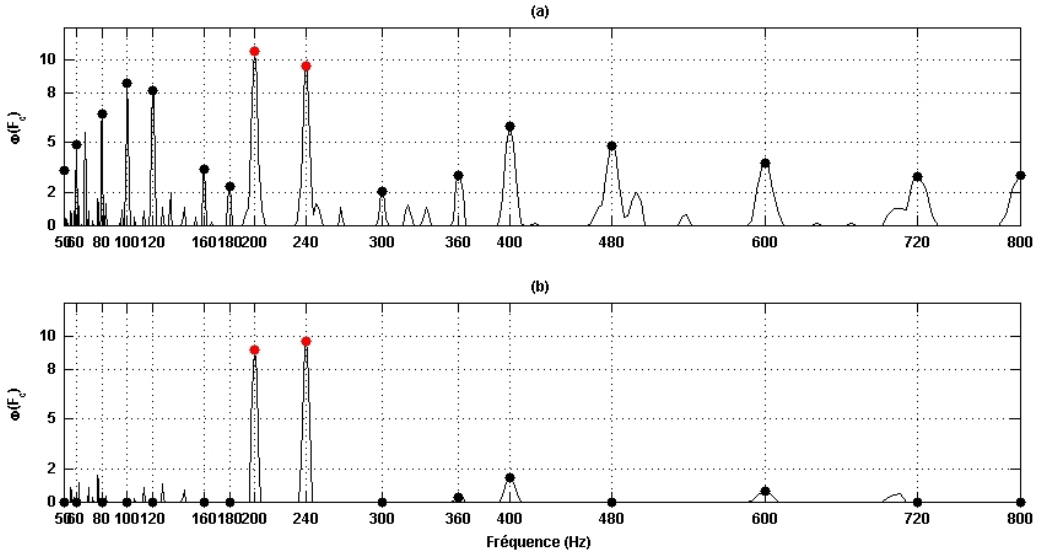

FIGURE 4.44: Comparaison des fonctions PUI et de la fonction PSH en situation bipitch. (a) Fonction PUI avec correction hyperbolique. (b) fonction PSH.

En situation bipitch, le graphique (a) de la figure 4.44 présente la fonction PUI avec correction hyperbolique et le graphique (b) présente la fonction PSH obtenue par l'algorithme PSH_{min} en n'utilisant que les PDN et PDM d'ordre 2 et 3. Là encore, le résultat est convaincant et les pics de F_{01} et F_{02} sont clairement émergent impliquant que la fonction PSH est facilement exploitable. Il faut toutefois être prudent car l'amplitude et le nombre des pics parasites ont augmentés par rapport à la situation monopitch et le spectre étudié est idéal.

4.6.3 Situation 4-pitch

En situation 4-pitch, le graphique (a) de la figure 4.45 présente la fonction PUI avec correction hyperbolique et le graphique (b) présente la fonction PSH obtenue par l'algorithme PSH_{min} en n'utilisant que les PDN et PDM d'ordre 2 et 3. Le résultat est plus nuancé que les précédents et le pic parasite en 400Hz reste important. Il reste cependant très prometteur puisque l'estimation de F_0 est correcte. Il faut observer la diminution des pics en 200Hz et 240Hz par rapport à leurs originaux du graphique (a). Ce point nuit à l'estimation de F_0 et son explication reste une piste à explorer dans l'utilisation des PDN et PDM. Les relations (p, q) entre les F_0 mélangées ainsi que le nombre d'harmoniques de chacune des structures harmoniques jouent certainement un rôle important mais ce point n'a pas pu être clairement démontré.

L'utilisation combinée des PDN et PDM aux ordres 2 et 3 permet de donner de bons résultats

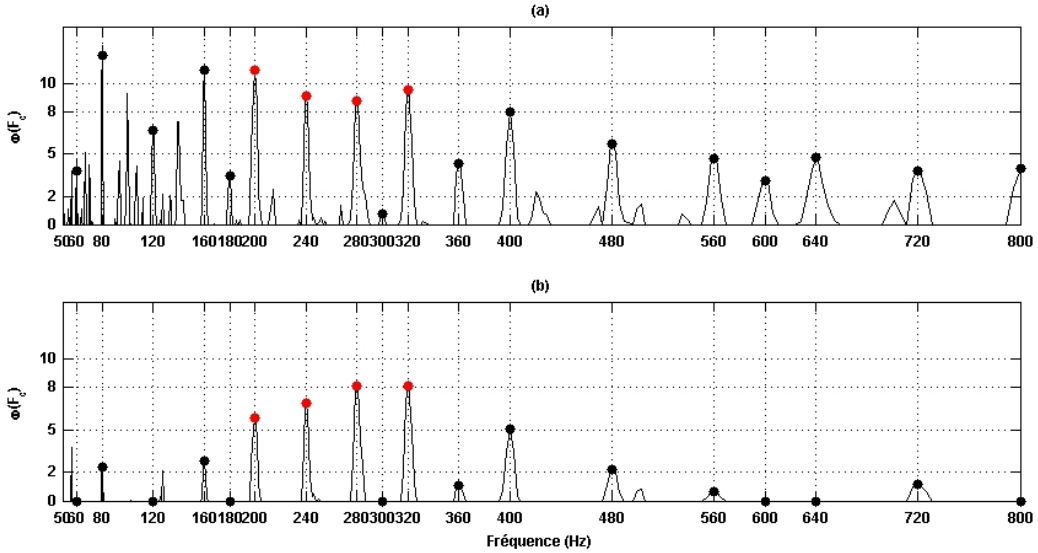

FIGURE 4.45: Comparaison des fonctions PUI et de la fonction PSH en situation 4-pitch. (a) Fonction PUI avec correction hyperbolique. (b) fonction PSH.

d'atténuation des pics parasites quelque soit la situation étudiée. Ces résultats sont certes obtenus sur des spectres idéaux mais le chapitre 6 d'évaluation montre que le principe de PSH est utilisable sur des signaux de parole superposée bipitch réels. Les performances obtenues sur des mélanges réels sont largement à la hauteur de l'état de l'art.

4.7 Conclusions

Le chapitre commence par l'analyse du comportement de l'estimation de F_0 multiples par des peignes fréquentiels élémentaires PUI, PUF, PDI et PDF. Lorsque le spectre d'amplitude n'est pas de moyenne nulle, la fonction de pitch obtenue par un de ces peignes possèdent un comportement hyperbolique très gênant pour l'estimation de F_0. La correction hyperbolique s'effectue en annulant la moyenne du spectre d'amplitude. Les pics d'une fonction de pitch ne sont pas positionnés de manière aléatoire mais suivent une règle déterminée. L'indexation (p, q) permet d'en donner la position fréquentielle exacte par rapport à la valeur de la F_0. Le pic (p, q) est de fréquence $(p/q)F_0$ avec p, q entiers non nuls et premiers entre eux. Deux familles de peignes sont finalement présentées : les peignes à dents négatives (PDN) et les peignes à dents manquantes (PDM). Les PDN permettent, en fonction de leur ordre Δ, de supprimer les pics parasites de type $(\Delta p, q)$ avec $(\Delta p/q)$ irréductible. Les PDM permettent, en fonction de leur ordre Δ, de supprimer les pics parasites de type $(p, \Delta q)$ avec $(p/\Delta q)$ irréductible. Il doit donc être possible, à partir de ces deux familles de peignes et en les utilisant à différents ordres, de supprimer sélectivement les pics parasites de la fonction de pitch d'un AEP. Ce chapitre tend à montrer la faisabilité d'un algorithme d'estimation conjointe de F_0 multiples par l'utilisation de plusieurs peignes spectraux.

Chapitre 5

Implémentations du principe de PSH

5.1 Introduction

Le principe de Peigne à Suppression Harmonique décrit dans le chapitre 4 a été présenté dans [Liénard et al., 2008]. Le principe de PSH est de combiner les fonctions PDN et PDM aux différents ordres afin de réduire l'amplitude des pics parasites tout en conservant les pics F_0 à estimer. Ce chapitre décrit dans le détail deux implémentations différentes du principe de PSH. Les deux implémentations prennent en entrée un signal de parole quelconque et un certain nombre de paramètres donnés par l'utilisateur. Les algorithmes analysent le signal par trames successives. La sortie des algorithmes est un fichier texte contenant les hypothèses F_0 estimées pour chacune des trames du signal. Le fichier contient également l'amplitude de chaque hypothèse et la force de périodicité de chacune des trames. La première implémentation est nommée PSH_{mul} et la seconde est nommée PSH_{min}. PSH_{mul} a été conçu et optimisé pour être le plus performant possible sur des signaux de parole superposée réelle tout en étant le plus rapide possible. PSH_{min} est une version « pédagogique » du principe de PSH. C'est à partir de cette version que sont obtenues tous les graphiques de ce manuscrit. PSH_{min} a été construite à partir de considérations théoriques et n'a pas été optimisée. La section 5.2 détaille le fonctionnement des deux implémentations et en donne les différences. La section 5.3 présente les résultats obtenus par PSH_{min}et PSH_{mul}dans les situations monopitch, bipitch et 4-pitch sur des trames synthétiques et sur des trames de parole réelle. La section 5.4 présente les situations dans lesquelles nos AEP sont en difficulté pour estimer correctement les F_0. Enfin, la section 5.5 présente les conclusions de ce chapitre.

5.2 Deux implémentations du principe de PSH

Cette section présente les différentes étapes de l'AEP PSH_{mul} dans le paragraphe 5.2.1. Le paragraphe 5.2.2 détaille les différentes étapes de PSH_{min}. Enfin, le paragraphe 5.2.3 discute des différences entre les algorithmes en donnant quelques prévisions quant aux avantages et inconvénients de chacun.

5.2.1 Algorithme PSH_{mul}

Cette première version de PSH a été développée sous PRAAT. L'utilisateur peut régler la zone de recherche de F_0 décrite par deux paramètres : la fréquence fondamentale minimale et la fréquence

fondamentale maximale. Le réglage standard est $[50, 800Hz]$. Le seuil de voisement est également réglable et permet d'ajuster la sensibilité de l'algorithme au voisement. Si le seuil de voisement est élevé, seules les trames dont le voisement est très marqué seront détectées. Si le seuil de voisement est faible, des trames litigieuses en termes de voisement peuvent être détectées comme voisées. L'algorithme donne aussi la possibilité de régler le nombre de candidats F_0 maximal à estimer pour chacune des trames. La figure 5.1 donne le synopsis général de l'algorithme PSH_{mul}. Les sous-paragraphes suivants détaillent l'algorithme bloc par bloc.

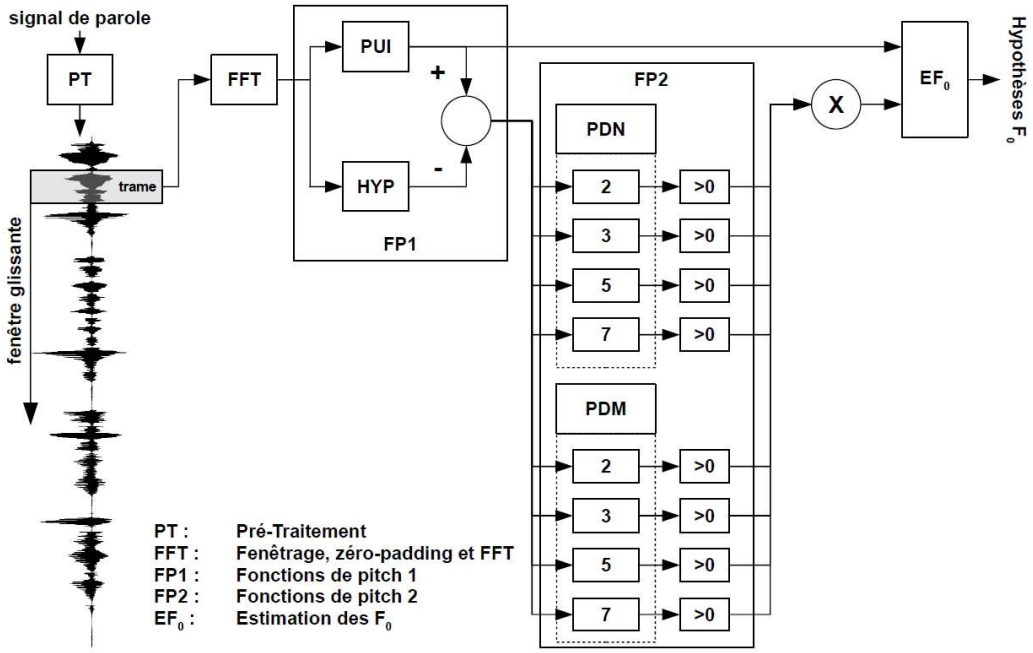

FIGURE 5.1: Schéma bloc de PSH_{mul}.

5.2.1.1 Étape de Pré-traitement (PT)

L'étape consiste à sous-échantillonner l'intégralité du signal étudié à 2kHz puisque PSH_{mul} ne considère que les informations fréquentielles inférieures à 1kHz. L'intensité du signal est normalisée au sens de la fonction « *Scale Intensity...* » de PRAAT à 70 dB. Les composantes basses fréquences du signal complet (0-40Hz) sont ensuite retirées par un filtrage passe-haut non récursif.

5.2.1.2 Étape de FFT (FFT)

Le signal est analysé par trames successives de 50 millisecondes avec un pas d'analyse entre deux trames successives de 10 millisecondes. Chaque trame est multipliée par une fenêtre de Hanning de manière à supprimer les discontinuités aux extrémités de la trame. La durée de la trame est rallongée par zéro-padding (ajout de zéro) de manière à obtenir une trame de 500 millisecondes soit 10 fois la durée de la trame initiale. Le spectre d'amplitude de la trame est alors calculé par une FFT.

5.2.1.3 Étapes de FP1

Cette étape a pour objectif de calculer la fonction PUI avec correction hyperbolique du spectre d'amplitude. Le bloc PUI consiste à calculer la fonction PUI sans correction hyperbolique à partir du spectre d'amplitude obtenu dans l'étape FFT. L'intervalle de recherche $[7,800Hz]$ est parcouru de manière logarithmique avec une précision de 50 pas par octave. Le bloc HYP consiste à calculer l'aire du spectre d'amplitude (méthode des trapèzes). L'aire ainsi estimée permet de calculer l'équation de l'hyperbole à retirer à la fonction PUI obtenue dans le bloc PUI. L'équation de l'hyperbole $H(F_c)$ est rappelée dans l'équation 5.1 avec A, l'aire du spectre d'amplitude.

$$H(F_c) = \frac{A}{F_c} \tag{5.1}$$

5.2.1.4 Étape FP2

Cette étape est le cœur de PSH_{mul} et rassemble la construction de toutes les fonctions PDN et PDM aux ordres 2, 3, 5 et 7. Ces fonctions sont calculées à partir de la fonction PUI issue de l'étape FP1. La fonction PUI est définie sur $[7,800Hz]$ de manière à ce que chacune des fonctions $PDN\Delta$ et $PDM\Delta$ soit calculables à partir d'une combinaison linéaire d'échantillons de la fonction PUI. L'équation 5.2 rappelle les valeurs des fonctions $PDN\Delta$ et $PDM\Delta$ en fonction de la fonction PUI (cf. chapitre 4). Δ est l'ordre du peigne.

$$\begin{array}{ll} \Phi_{PDN\Delta}(F_c) & = \frac{\Delta}{\Delta-1}\Phi_{PUI}(F_c) - \frac{1}{\Delta-1}\Phi_{PUI}\left(\frac{F_c}{\Delta}\right) \\ \Phi_{PDM\Delta}(F_c) & = \frac{\Delta}{\Delta-1}[\Phi_{PUI}(F_c) - \Phi_{PUI}(\Delta F_c)] \end{array} \tag{5.2}$$

Les fonctions PDN et PDM sont calculées pour l'intervalle de recherche de F_0 $[50,800Hz]$. Pour chacune des fonctions PDN et PDM, les valeurs négatives sont misent à 0 car considérées comme ne pouvant pas être des candidats F_0.

5.2.1.5 Étape d'estimation de F_0 ($\mathbf{EF_0}$)

Cette étape prend en entrée deux fonctions : la fonction PUI avec correction hyperbolique et la fonction produit des fonctions PDN et PDM issues de l'étape FP2. Les fonctions PDN et PDM atténuent certains pics parasites en fonction de l'ordre tout en laissant relativement intact les pics des F_0 à estimer. Le produit de ces fonctions permet donc de faire fortement émerger les pics des F_0 tout en réduisant notablement les pics parasites. Ce produit construit une fonction de masque binaire permettant de sélectionner les pics des F_0 et de rejeter les autres par seuillage. Les pics de la fonction PUI avec correction hyperbolique sont conservés tels quels dans la fonction PSH finale si le masque binaire les indique comme candidat F_0.

Il convient maintenant d'étudier le fonctionnement de la seconde implémentation nommée PSH_{min}.

5.2.2 Algorithme PSH_{min}

Le synopsis de l'algorithme PSH_{min}est donné dans la figure 5.2. Chaque étape va faire l'objet d'un sous-paragraphe. Le résultat de certaines étapes est analysé à partir d'un signal sinus complexe de

40ms comportant 10 harmoniques d'amplitude en décroissance racine inverse, avec une F_0 de 240Hz (cf. chapitre 3 paragraphe 3.3.1). La fréquence d'échantillonnage est de 8kHz. Les phases des harmoniques sont aléatoires selon une distribution uniforme entre $[-\pi, \pi]$. Cette version de PSH a été conçue sous

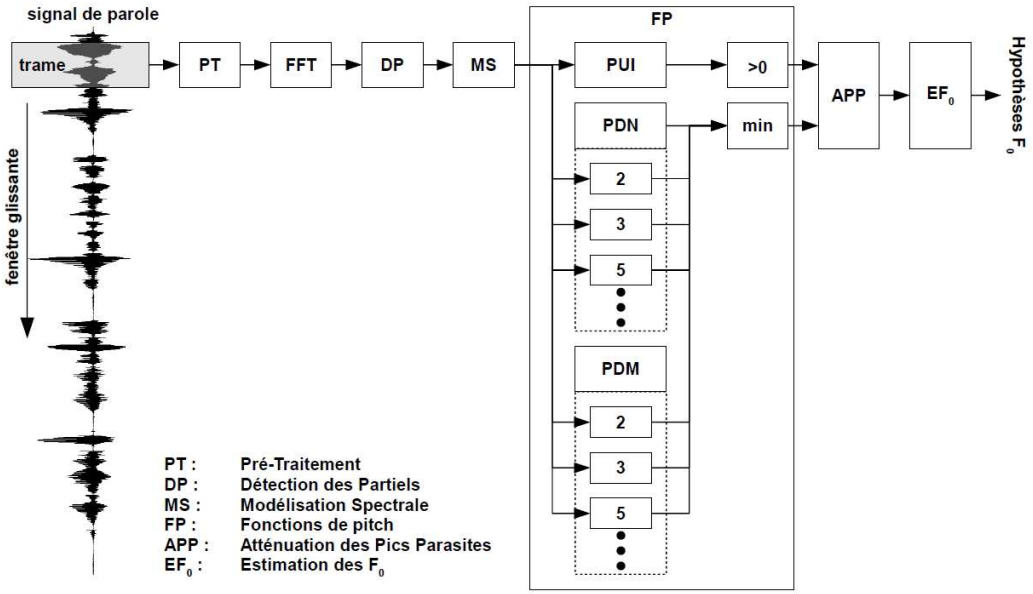

FIGURE 5.2: Schéma bloc de PSH_{min}.

MATLAB (Version 7.3.0.0267 R2006b). L'algorithme est défini par un certain nombre de paramètres permettant de régler les trames, l'intervalle de recherche de F_0, la construction du spectre d'amplitude et le nombre maximum d'hypothèses F_0 par trames. Le réglage standard de PSH_{min} considère des trames de 40ms avec un recouvrement de 30ms et un intervalle de recherche de F_0 compris entre 50 et 800Hz.

5.2.2.1 Étape de pré-traitement (PT)

La composante de chacune des trames est retirée. La trame centrée est multipliée par une fenêtre de Hanning pour supprimer les discontinuités aux extrémités. La trame fenêtrée est centrée de nouveau. Cette opération permet d'obtenir un signal de moyenne nulle et sans discontinuités aux extrémités.

5.2.2.2 Étape de calcul du spectre d'amplitude (FFT)

La durée de la trame est étendue de manière à ce que son nombre d'échantillon soit la puissance de 2 directement supérieure au nombre d'échantillon initial de la trame. La trame est complétée par des zéros (*zero-padding* en anglais). Ceci permet d'utiliser l'algorithme de Transformée de Fourier Rapide ou FFT. Le spectre d'amplitude est alors calculé.

5.2.2.3 Étape de détection des partiels (DP)

Le spectre d'amplitude obtenu est seuillé et toute valeur dont l'amplitude est inférieure de 30dB à l'amplitude maximale est mise à 0. Le choix est cohérent car un signal d'intensité de 30dB inférieure à

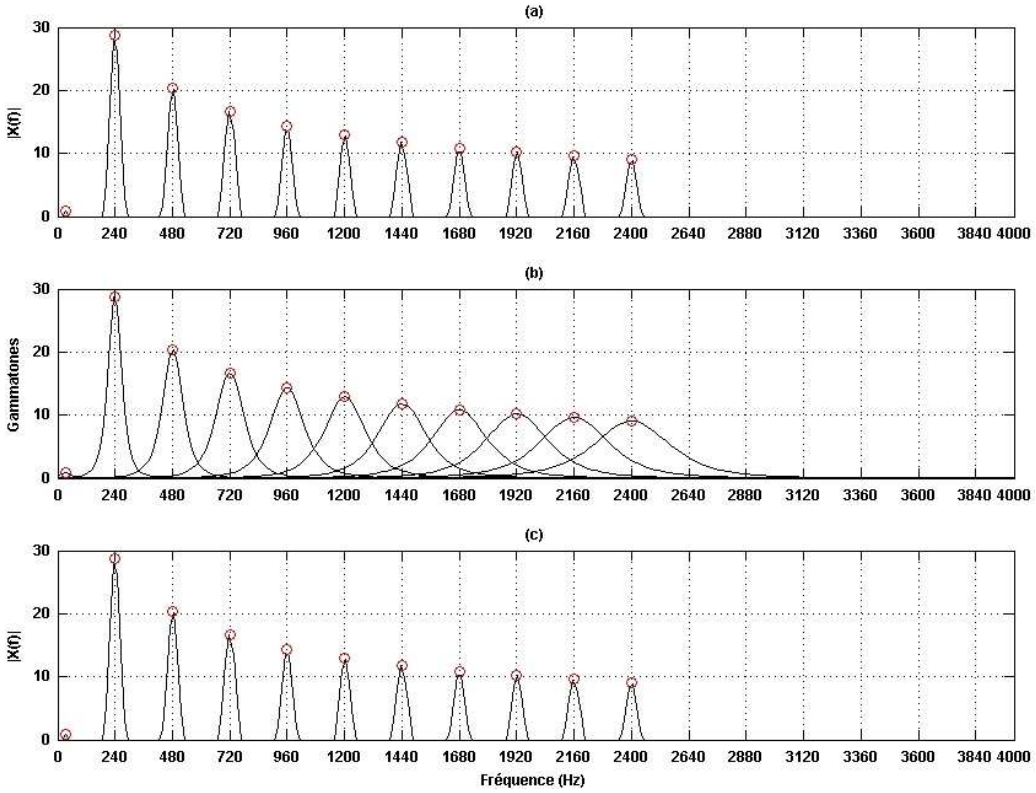

FIGURE 5.3: Détection de partiels et conservation des partiels audibles selon un modèle de masquage à base de filtres gammatones d'ordre 4. (a) Spectre d'amplitude et maxima locaux. (b) gammatones d'ordre 4 passant par chaque maximum local. (c) Partiels conservés. Les ronds rouges correpondent à ces partiels.

un autre peut être considérée comme relativement négligeable. L'équation 5.3 donne la valeur théorique du seuil noté θ. $|X(f)|$ est le spectre d'amplitude.

$$\theta = 20 \log_1 0 \max |X(f)| - 30 \tag{5.3}$$

Le spectre d'amplitude seuillé est modélisé par une somme de paraboles qui représentent les partiels contenus dans le spectre original. Une étape de détection des partiels est donc nécessaire. Cette étape est réalisée en repérant tous les maxima locaux du spectre d'amplitude. Un filtre gammatone d'ordre 4 (cf. chapitre 3 paragraphe 3.3.3) est calculé pour chaque maximum local du spectre d'amplitude. Le centre de chaque filtre passe par le maximum local auquel il correspond. Cette opération permet de construire une courbe de masquage fréquentielle au-dessous de laquelle les maxima locaux ne seront pas considérés comme des partiels perceptibles et seront exclus de la liste des partiels [Gnansia, 2005]. Les trois étapes de la détection de partiels sont illustrées dans la figure 5.3. Le graphique (a) présente l'étape de détection des maxima locaux. Le graphique (b) présente l'étape de construction des filtres gammatones. Le graphique (c) présente le résultat de la détection des partiels perceptibles.

Le tableau 5.1 donne les valeurs centrales des partiels conservés après l'étape DP. Il y a 11 partiels conservés. Dans ce cas synthétique, les valeurs fréquentielles estimées des partiels sont très proches des valeurs théoriques. Un des partiels, le premier, est incorrect. Toutefois, son impact sur l'estimation

TABLE 5.1: Liste des fréquences centrales des partiels conservés.

Numéro du partiel	1	2	3	4	5	6	7	8	9	10	11
Fréquences centrales (Hz)	31.25	239.67	480.36	719.90	959.82	1200.25	1439.79	1680.04	1920.21	2159.71	2400.25

de F_0 sera limité. La suite montrera que cette étape est relativement critique dans PSH_{min}. Tout partiel manqué affaiblit la structure harmonique et donc la qualité du pic dans la fonction de pitch. Au contraire, tout ajout de partiels amène de nouveaux pics parasites.

5.2.2.4 Étape de modélisation du spectre d'amplitude (MS)

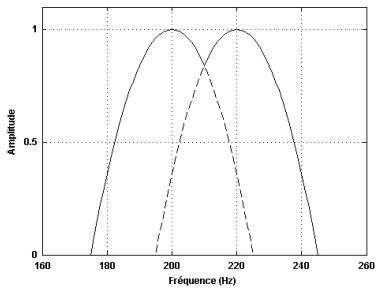

FIGURE 5.4: Recouvrement de deux paraboles. (trait continu) Amplitude résultante du recouvrement. (trait pointillé) Portions des paraboles non prise en compte.

Un partiel est défini par sa fréquence centrale et son amplitude maximale. Afin de donner la même importance à chacun des partiels, les amplitudes des paraboles doivent toutes être égales. Pourquoi vouloir donner la même importance à tous les partiels ? Ce choix est guidé par deux raisons. D'une part, la nécessité de limiter l'impact d'une éventuelle différence d'intensité entre les signaux mélangés. Si l'un domine l'autre, ses partiels sont plus marqués et de plus grande amplitude que ceux du signal dominé. Par une amplitude insuffisante, les partiels du son dominant peuvent provoquer la disparition du pic de la F_0 du signal dominé dans la fonction de pitch. Une autre raison concerne la distribution de l'énergie de la parole. La majorité de l'énergie se trouve dans les fréquences comprises entre 0 et 1kHz. Les partiels de plus haute fréquence sont d'amplitude plus faible et interviennent moins dans l'estimation de la périodicité. Le fait d'égaliser les amplitudes permet de donner autant de poids à tous les partiels, même ceux des hautes fréquences. L'inconvénient de ce choix est que tout faux partiel ou partiel du à du bruit est de même amplitude que les autres. Malgré l'égalisation des amplitudes des partiels, il faut conserver une information des amplitudes originales. Il a été décidé que l'amplitude des paraboles soit fixée à la moyenne des amplitudes des partiels. Autrement dit, chaque partiel est remplacé par une parabole centrée à la position fréquentielle du partiel et dont l'amplitude maximale vaut la moyenne de l'amplitude de tous les partiels détectés. Dans l'exemple, la moyenne de l'amplitude des partiels conservés vaut 13.25. Chaque partiel du spectre est modélisée par une parabole [Martin, 2000] d'une épaisseur fréquentielle b_w de 50Hz ($\pm$ 25Hz autour de la fréquence centrale de l'harmonique). Si pour une fréquence donnée plusieurs paraboles se recouvrent, l'amplitude de la fréquence donnée vaut l'amplitude maximale des paraboles se recouvrant (cf. figure 5.4). Les zones fréquentielles qui ne correspondent pas à des harmoniques sont remplacées par une constante négative de manière à ce que la moyenne du spectre d'amplitude modèle soit nulle. Ceci a pour effet de supprimer le comportement hyperbolique des fonctions de pitch et de maximiser les émergences comme il a été vu dans le chapitre 4. La fréquence de coupure F_{MAX} est fixée à la fréquence centrale du dernier partiel détecté à laquelle on ajoute la demi-épaisseur fréquentielle d'une parabole soit 25Hz.

Dans l'exemple F_{MAX} vaut 2425.25Hz. La figure 5.5 montre le résultat de la modélisation spectrale. La

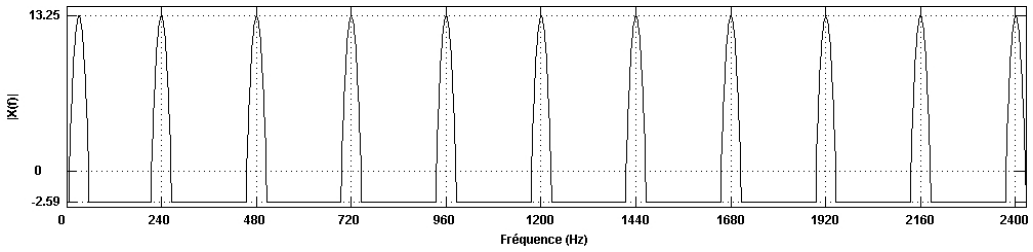

FIGURE 5.5: Modèle spectral du signal exemple.

constante négative vaut -2.59. Les paraboles sont clairement visibles et ont une amplitude maximale de 13.25. Les différentes fonctions de pitch vont être calculées à partir de ce modèle spectral. Dans PSH_{min}, l'information d'amplitude est retirée pour rendre la version uniquement sensible au nombre d'harmonique et pas à l'amplitude des harmoniques (très grossièrement le timbre ou plutôt une forme d'enveloppe spectrale). Ce choix est la conséquence de l'analyse théorique des fonctions de pitch donnée en annexe A qui porte sur des spectres idéaux de Dirac. La version PSH_{mul} prend en compte l'information d'amplitude et le nombre des harmoniques. Cette différence doit jouer un rôle important dans la différence de résultats entre les deux implémentations.

5.2.2.5 Étape de fonctions de pitch (FP)

Une fois le modèle de spectre obtenu, les différentes fonctions de pitch PUI, PDN et PDM peuvent être extraites. La fonction PUI est calculée et met en évidence les relations harmoniques entre les partiels détectés. Elle est présentée figure 5.6. Toute valeur négative dans la fonction PUI est synonyme

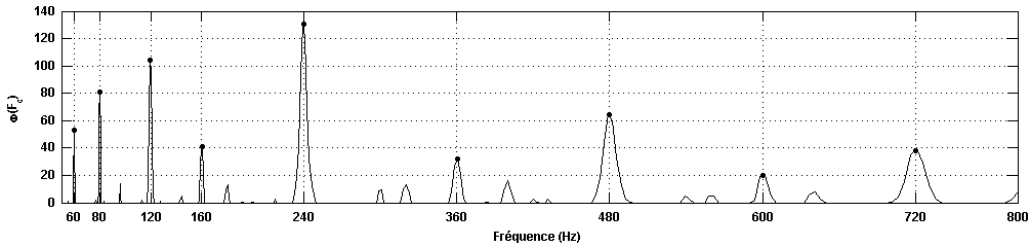

FIGURE 5.6: Partie positive de la fonction PUI.

de non périodicité (du moins en situation monopitch) puisque cela signifie que les dents du PUI étaient majoritairement dans les zones négatives du spectre d'amplitude, zones non-harmoniques. C'est pourquoi seule la partie positive est conservée intacte et les valeurs négatives sont toutes remplacées par des zéros. L'intervalle standard de recherche de F_0 est fixé entre 50 et 800Hz. Cet intervalle peut contenir au maximum 16 harmoniques (800/50). Pour couvrir l'ensemble des familles de pics parasites par les PDN et PDM, il faut donc utiliser les ordres premiers de 2 jusqu'à 16. Ainsi, dans son réglage par défaut, PSH_{min} calcule les fonctions PUI, PDN2, PDN3, PDN5, PDN7, PDN11, PDN13, PDM2, PDM3, PDM5, PDM7, PDM11 et PDM13. Comme il a été montré dans le chapitre 4, les

pics parasites sont supprimés sélectivement par les différents ordres de PDN et PDM tout en gardant relativement intact le ou les pics correspondant à la F_0. Ces caractéristiques sont reprises et permettent de construire une fonction de pitch nommée fonction de Détection des Pics Parasites. Cette fonction est notée fonction DPP. Pour chacune des fréquences de la fonction DPP, sa valeur correspond à la valeur minimale des toutes les valeurs de toutes les fonctions de pitch PDN et PDM. L'équation 5.4 formalise mathématiquement la fonction DPP. {} désigne l'ensemble et Δ est un nombre premier différent de 1.

$$DPP(F_c) = \min_{\Delta \in \{2,3,5,7,...\}} [\{\Phi_{PDN\Delta}(F_c, f)\}, \{\Phi_{PDM\Delta}(F_c, f)\}] \tag{5.4}$$

La figure 5.7 présente la fonction DPP obtenue sur le signal synthétique d'exemple. Le graphique (a) présente la fonction DPP sur tout l'intervalle de recherche $[50, 800Hz]$. Le graphique (b) présente l'intervalle sous-harmonique au-dessous de la F_0 à 240Hz et le graphique (c) illustre la partie sur-harmonique au-dessus de 240Hz. La fonction DPP possède une majorité de valeurs négatives hormis

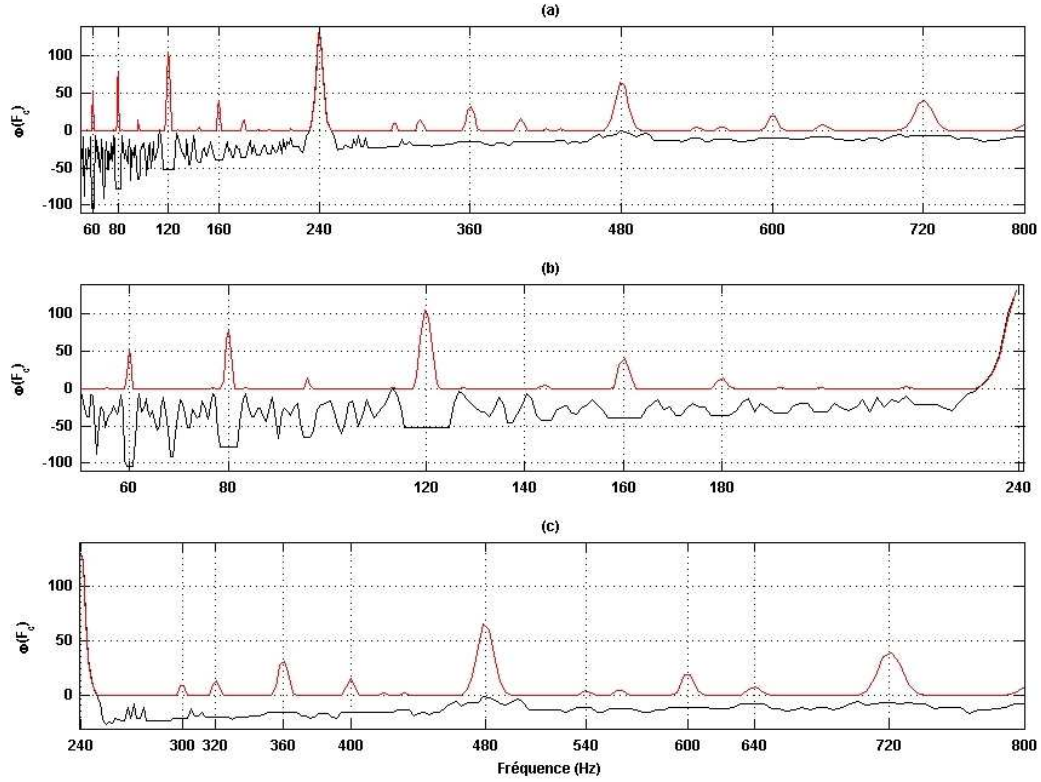

FIGURE 5.7: Fonction DPP. (a) Dans son intégralité. (b) Zoom sur l'intervalle sous-harmonique [50,240Hz]. (c) Zoom sur l'intervalle sur-harmonique [240,800Hz].

pour le pic en F_0. Le seul pic parasite restant est placé à 113.3Hz et est d'amplitude très faible. La fonction DPP est utilisée pour déterminer les pics parasites et les pics à conserver comme le précise le sous-paragraphe 5.2.2.6. Le parcours de l'intervalle de recherche $[50, 800Hz]$ est effectué de manière logarithmique. 50Hz est considérée comme l'octave 0, 100Hz est donc l'octave 1, 200Hz l'octave 2, etc. Le pas de parcours est fixé à 100 par octave. Ce choix est motivé par deux raisons. D'une part un gain de temps de calcul. Avec la version standard de PSH, il y a 13 fonctions de pitch à calculer par trame.

En conservant une échelle linéaire par pas de 1Hz avec un intervalle de recherche de $[50, 800Hz]$, il y a $13 * 751 = 9763$ itérations. Avec une échelle par octave avec 100 pas par octave, il y a $13 * 400 = 5200$ itérations soit un nombre d'itérations presque réduite de moitié. D'autre part, les maxima locaux des fonctions de pitch s'élargissent lorsque la fréquence augmente. Une échelle logarithmique permet d'adapter la résolution fréquentielle du parcours en fonction de la fréquence. La résolution est fine dans les basses fréquences et grossière dans les hautes fréquences. Avec un parcours à pas constant de 1Hz, la résolution risque d'être trop faible dans les basses fréquences (perte de précision) et risque d'être trop fine dans les hautes fréquences (perte de temps de calcul).

5.2.2.6 Atténuation des pics parasites

Un masque de suppression harmonique est calculé en comparant la fonction de pitch PUI et la fonction DPP. Chacun des maxima locaux positifs de la fonction DPP est une hypothèse F_0 potentielle. Plus un maximum local de la fonction DPP est faible par rapport à son homologue de la fonction PUI, plus le maximum local a des chances d'être un pic parasite. L'écart relatif entre la fonction PUI et la fonction DPP est le critère objectif utilisé pour quantifier la distance entre les deux fonctions. Ce critère permet de calculer le masque de suppression harmonique noté MAQ. Les valeurs de MAQ sont comprises entre 0 et 1. Plus la valeur est proche de 0, plus le pic est considéré comme parasite. La fonction MAQ est calculée seulement pour les fréquences des maxima locaux du PUI. Pour chaque valeur MAQ obtenue, elle est attribuée à tout le pic comme l'illustre le graphique (b) de la figure 5.7 pour les fréquences autour de 240Hz. L'équation 5.5 présente le calcul de la fonction masque. Si la valeur du masque est supérieure à 1, alors la valeur est saturée à 1. De même, si la valeur est inférieure à 0, elle est saturée à 0.

$$MAQ(f) = \frac{DPP(f)}{PUI(f)} \tag{5.5}$$

La fonction MAQ est utilisée comme une fonction de pondération et la fonction PSH est le résultat du produit de la fonction MAQ avec la fonction PUI. Le graphique (c) de la figure 5.8 présente la fonction PSH qui est le produit entre la fonction PUI du graphique (a) et le masque déduit de la comparaison entre la fonction PUI et la fonction DPP présenté au graphique (b). Les hypothèses F_0 sont déterminées à partir de la fonction PSH. L'amplitude des pics par ordre décroissant indique l'ordre des hypothèses F_0. Dans l'exemple, PSH donne un résultat quasi parfait puisqu'il ne reste de la fonction PUI que le pic principal d'amplitude 132.4 placé à 240Hz et un autre pic parasite d'amplitude 1.6 placé à 113Hz environ et entouré dans le graphique (c). Tout le reste est mis à zéro par la fonction MAQ.

5.2.2.7 Étape d'estimation des F_0 ($\mathbf{EF_0}$)

L'estimation est faite à partir de la fonction PSH en considérant les N premiers maxima par ordre décroissant d'amplitude. Dans l'exemple, seules deux hypothèses F_0 sont fournies. La première hypothèse vaut effectivement 240Hz (239.99Hz exactement) pour une amplitude de 132.44 et la seconde vaut 113.32Hz pour une amplitude de 1.57. L'estimation est donc correcte. Le nombre standard d'hypothèse est fixé à 5. Bien entendu, ce nombre est paramétrable à souhait.

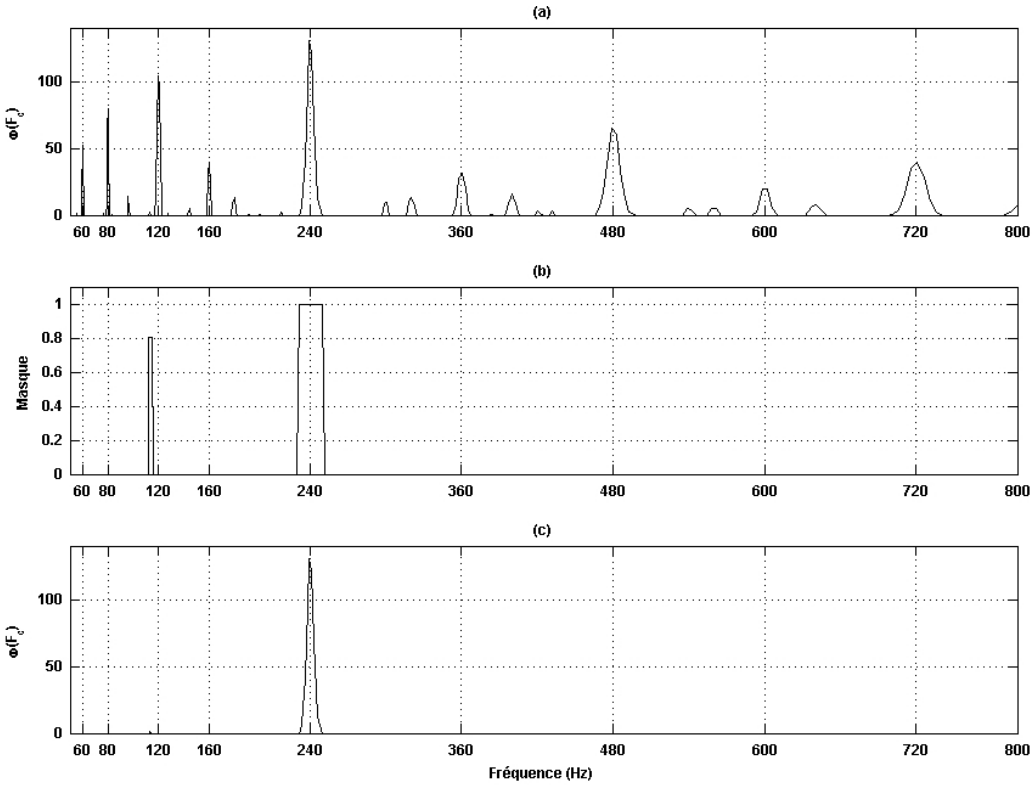

FIGURE 5.8: Fonctions PUI, MAQ et PSH en situation monopitch sur le signal synthétique exemple. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.2.3 Différences des deux implémentations

PSH_{min} et PSH_{mul} ont été écrits dans deux optiques bien différentes. PSH_{min} est un AEP construit pour vérifier des considérations théoriques, pour construire les figures du manuscrit et n'a pas été optimisé en terme de vitesse ni élaboré à partir de sons de parole réelle. C'est un outil pédagogique permettant de démontrer les principes de PSH. PSH_{mul} a été construit dans l'optique d'être rapide et relativement robuste et est le fruit de nombreux tests sur des signaux de parole superposée réelle. Les différences entre ces deux AEP sont données dans la liste suivante :

- PSH_{mul} sous-échantillonne le signal à 2kHz puisqu'il ne considère que l'information spectrale inférieure à 1kHz. Autrement fit, FMAX vaut 1kHz. PSH_{min} n'est pas limité de la sorte et son F_{MAX} dépend du nombre de partiels détectés.
- PSH_{mul} utilise l'information brute du spectre d'amplitude. PSH_{min} utilise une étape de détection de partiels visant à isoler uniquement les maxima du spectre qui s'apparente à un partiel.
- PSH_{mul} utilise l'information d'amplitudes des partiels du spectre d'amplitude et donc l'information de timbre. PSH_{min} perd cette information de timbre en uniformisant l'amplitude des partiels et ne s'appuie que sur la position fréquentielle des partiels. Il conserve néanmoins une certaine information de l'énergie de la trame en fixant l'amplitude des partiels à la valeur moyenne des partiels détectés.
- PSH_{mul} combine les fonctions PDN et PDM en utilisant un produit. PSH_{min} utilise un minimum

pour combiner ces fonctions.

- PSH_{mul} et PSH_{min} utilisent tous deux les pics de la fonction PUI avec correction hyperbolique comme base de construction de leurs fonctions PSH. PSH_{mul} utilise un masque binaire pour sélectionner les pics de la fonction PUI à conserver dans la fonction PSH. PSH_{min} calcule un masque réel issue de la différence entre la fonction DPP et la fonction PUI.

La section 5.3 présente les résultats obtenus par les deux implémentations sur quelques exemples représentant les situations monopitch, bipitch et 4-pitch. Les situations difficiles de sons à l'octave sont présentés et les deux implémentations se montrent capables, au moins sur des signaux synthétiques, de les traiter correctement.

5.3 Résultats de PSH_{mul} et PSH_{min}

Cette section présente les fonctions PSH obtenues dans différentes situations. Tout d'abord, la situation monopitch est proposée dans le paragraphe 5.3.1 pour illustrer le bon fonctionnement de l'algorithme dans ce cas de figure. Ensuite, les situations multipitch bipitch (cf. paragraphe 5.3.2) et 4-pitch (cf. paragraphe 5.3.3) sont étudiées au travers de cas généralement considérés comme problématiques : la situation de F_0 à l'octave, la situation de « *virtual multipitch* » et la situation dans laquelle la différence entre les F_0 est faible.

5.3.1 Situation monopitch

Seuls des signaux monopitch de parole réelle sont testés dans ce paragraphe. Les résultats obtenus sur une trame voisée de voix de femme sont donnés dans le sous-paragraphe 5.3.1.1. Une autre trame voisée de voix d'homme est testée dans le sous-paragraphe 5.3.1.2.

5.3.1.1 Parole réelle – voix de femme

La trame voisée est extraite du corpus de Bagshaw (cf. chapitre 6 pour des précisions sur le corpus). Le signal utilisé est sb002 et la phrase prononcée est : « *I'd like to live the signal safe* ». La trame correspond au « *i* » prononcé dans le mot « *live* » au centre de la phrase. La figure 5.9 montre la fonction PUI, la fonction MAQ et la fonction PSH obtenue par PSH_{min} dans les graphiques (a), (b) et (c) respectivement. La valeur théorique obtenue par ACF sur le signal EGG (méthode décrite dans le chapitre 6 paragraphe 6.5.2) est de 289.3Hz. La valeur estimée est de 293.1Hz soit un écart relatif de 1.3%. L'estimation est donc correcte et il reste peu de pics parasites. La figure 5.10 présente les résultats de PSH_{mul} sur cet exemple. Le graphique (a) donne la fonction PUI avec correction hyperbolique. Le graphique (b) donne la partie positive de cette même fonction PUI. Le graphique (c) donne la fonction de masque binaire et le graphique (d) donne la fonction PSH obtenue. L'estimation donne 292Hz et plus aucun pic parasite n'est présent sur la fonction PSH. L'écart relatif est de 0.9% ce qui est très correct.

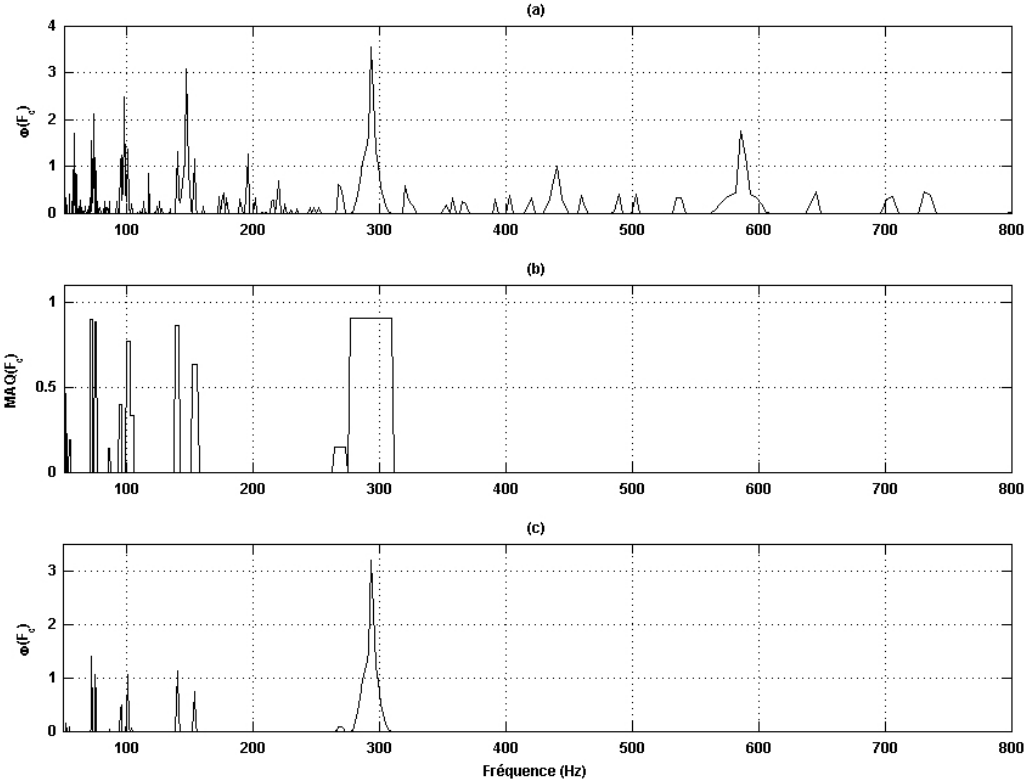

FIGURE 5.9: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation monopitch sur une voix de femme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

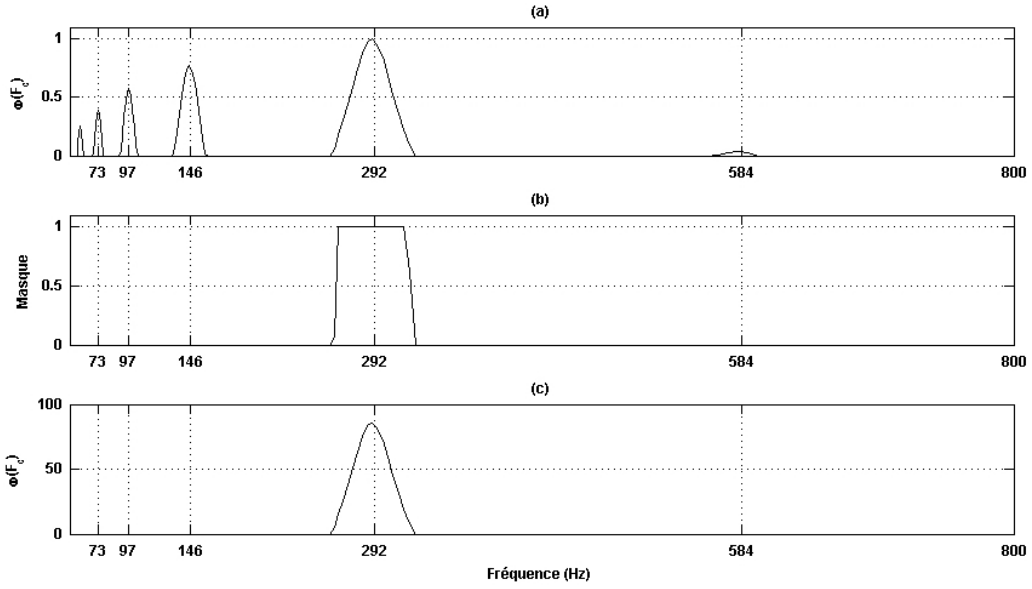

FIGURE 5.10: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation monopitch sur une voix de femme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.1.2 Parole réelle – voix d'homme

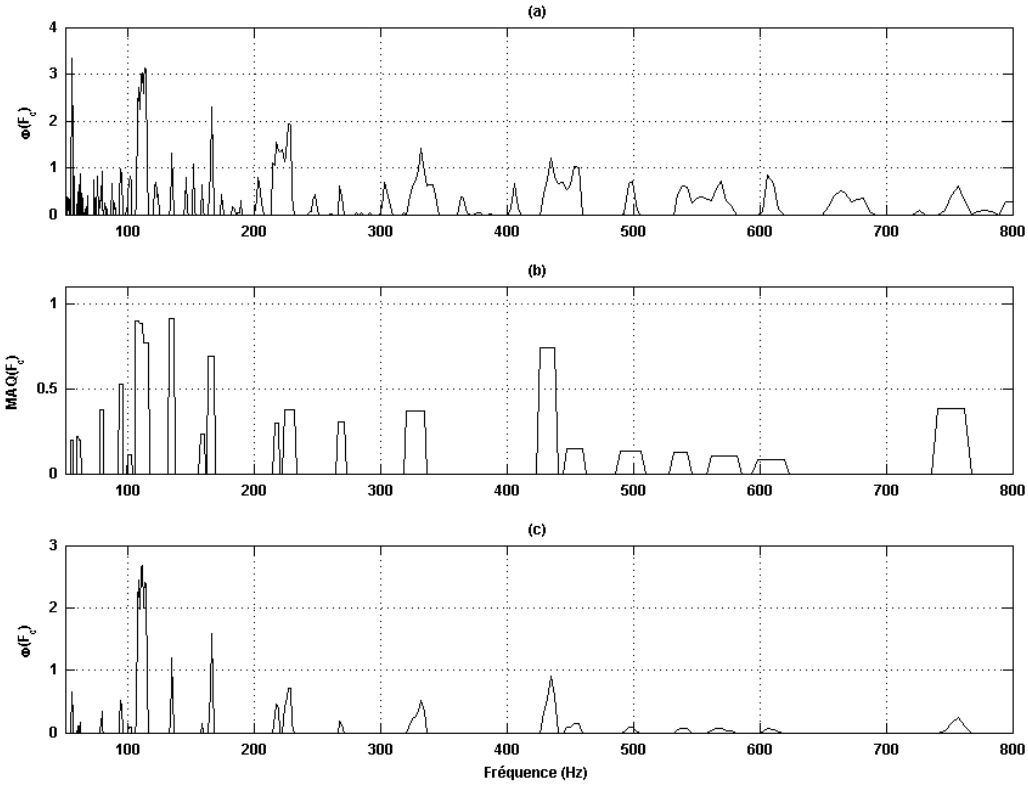

FIGURE 5.11: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation monopitch sur une voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

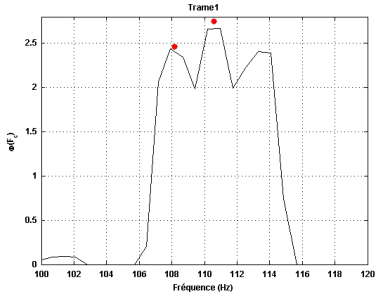

FIGURE 5.12: Illustration de pics parasites proches.

La trame utilisée est un extrait du signal rl001 appartenant au corpus de Bagshaw. La phrase prononcée est « *Where can I park my car ?* ». La trame se trouve au centre du « *a* » contenu dans le mot « *car* ». Le graphique (a) de la figure 5.11 présente la fonction PUI, le graphique (b) la fonction MAQ et le graphique (c) montre la fonction PSH. La valeur théorique est de 110.6Hz. La valeur estimée est de 110.6Hz. L'estimation est donc correcte et sans aucun décalage. En revanche, le premier pic parasite est à 108.1Hz. La figure 5.12 est un zoom autour du pic de la F_0. PSH_{min}peut produire des maxima locaux qui contiennent plusieurs autres maximum. Ces maxima ne peuvent pas être considérés à proprement parlé comme parasites car leur position fréquentielle sera proche de la position théorique. Cependant, ils vont diminuer la précision de PSH_{min}. La figure 5.13 présente les résultats obtenus par PSH_{mul}. Les graphiques (a) et (b) présentent la fonction PUI avec correction hyperbolique. Le graphique (c) présente le masque binaire et le graphique (d) présente la fonction PSH. L'estimation de F_0 donne une valeur de 113Hz ce qui est correct à 2.2% près. En situation monopitch réelle, les deux implémentations de PSH sont performantes. Qu'en est-il du passage aux situations

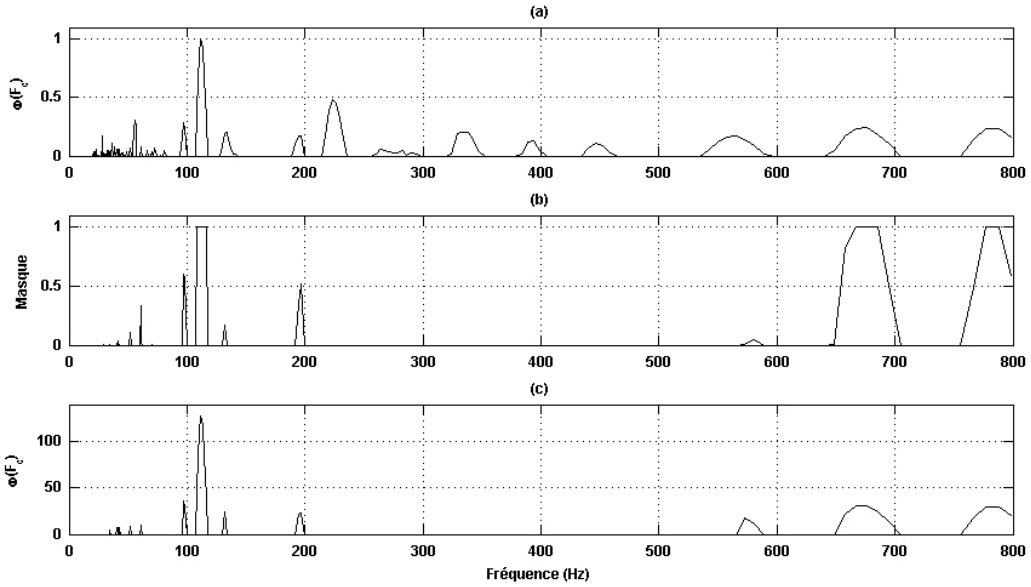

FIGURE 5.13: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation monopitch sur une voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

bipitch et 4-pitch ?

5.3.2 Situations bipitch

Dans un premier temps, des signaux bipitch synthétiques vont permettre d'étudier le comportement des deux implémentations de PSH dans les cas réputés comme difficiles à traiter. Il s'agit des cas de deux F_0 à l'octave, de F_0 en relation de « *virtual multipitch* » et de deux F_0 proches. Ces trois cas sont développés dans les sous-paragraphes 5.3.2.1, 5.3.2.2 et 5.3.2.3 respectivement. Le dernier sous-paragraphe va permettre de donner un ordre de grandeur de la résolution fréquentielle de PSH_{mul}et PSH_{min}c'est-à-dire de la différence minimale entre deux F_0 requise pour les discriminer. Dans un second temps, des signaux bipitch de parole réelle vont permettre de confronter PSH à une situation plus réaliste. Des mélanges voix de femmes/voix de femmes, voix de femme/voix d'homme et voix d'homme/voix d'homme sont présentés dans les sous-paragraphes 5.3.2.1, 5.3.2.2 et 5.3.2.3 respectivement.

5.3.2.1 Signaux synthétiques – situation à l'octave

Le signal exemple est une somme de deux signaux synthétiques. L'un est de fréquence fondamentale F_{01} à 120Hz. L'autre est de fréquence fondamentale F_{02} à 240Hz. Les deux F_0 sont à l'octave ce qui est une des situations les plus emblématiques à traiter dans la littérature. Le graphique (a) de la figure 5.14 montre la fonction PUI, le graphique (b) montre la fonction MAQ et le graphique (c) montre la fonction PSH. Il apparaît que PSH_{min} parvient très bien à estimer les F_0 à l'octave puisque les deux pics de plus fortes amplitudes sont les pics de F_{01} et F_{02}. Les valeurs estimées sont correctes : 120 et 240Hz. Le pic parasite le plus fort est situé à 56.7Hz. La figure 5.15 présente les résultats de PSH_{mul}. Les graphiques (a) et (b) montrent la fonction PUI avec correction hyperbolique. Le graphique (c) montre le masque

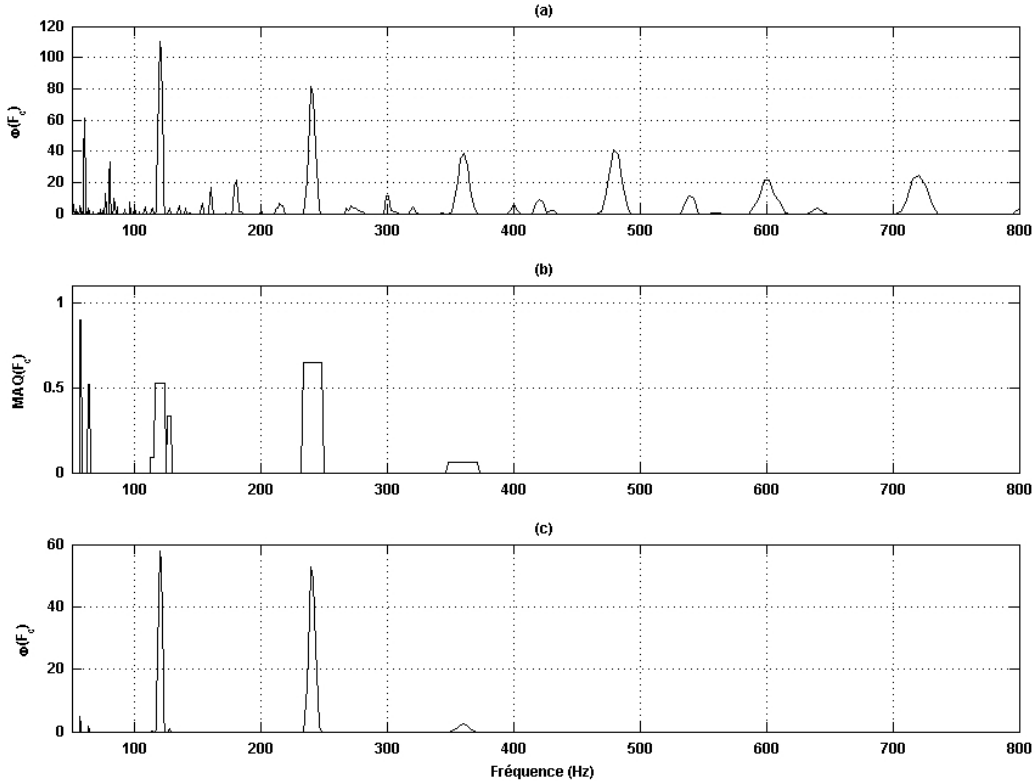

FIGURE 5.14: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch sur un signal synthétique comportant deux F_0 à l'octave. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

binaire et le graphique (d) illustre la fonction PSH. Là aussi, l'estimation est correcte et seul les deux pics des F_0 subsistent dans la fonction PSH. L'estimation donne effectivement 120Hz et 240Hz. Le cas à l'octave ne semble pas être un problème pour l'exemple synthétique étudié ni pour PSH_{min}, ni pour PSH_{mul}.

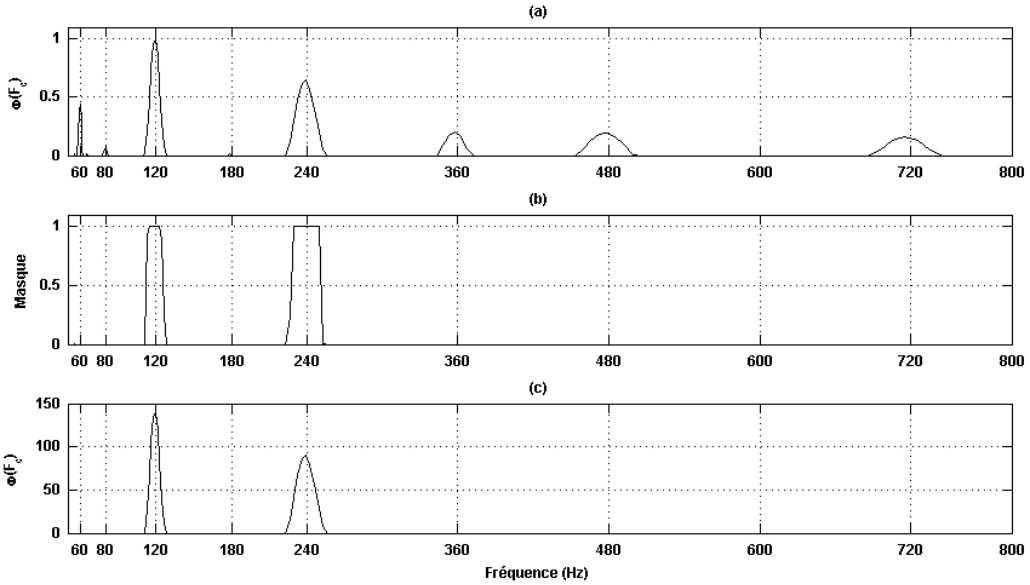

FIGURE 5.15: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch sur un signal synthétique comportant deux F_0 à l'octave. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.2.2 Signaux synthétiques – situation de *virtual multipitch*

Le signal exemple est construit de la même manière que précédemment. F_{01} vaut 160Hz et F_{02} vaut 240Hz. Cette situation représente une situation de virtual multipitch c'est-à-dire une situation dans laquelle une fréquence inférieure à F_{01} et F_{02} peut également être la F_0 des deux structures harmoniques. La situation de virtual multipitch la plus gênante est celle dans laquelle les deux F_0 sont en relation $(2, 3)$. Il s'agit du cas ici puisque $160 = (2/3) * 240$. La fréquence de 80Hz est très parasite car elle est pic $(1, 2)$ et F_{01} et le pic $(1, 4)$ de F_{02}. La figure 5.16 présente les résultats de PSH_{min}. Dans la fonction PUI du graphique (a), l'amplitude en 80Hz est effectivement plus importante que celle des F_0. Les valeurs estimées valent 160 et 240Hz. Le pic parasite le plus fort reste celui en 80Hz mais a été

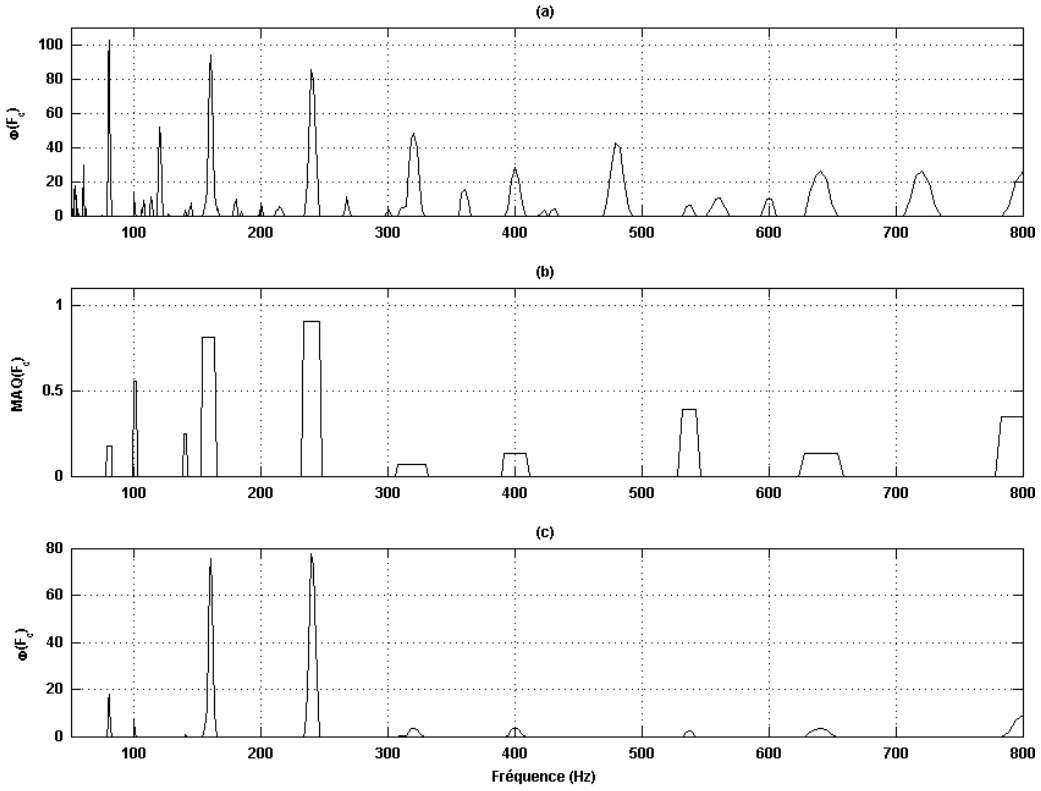

FIGURE 5.16: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch sur un signal synthétique comportant deux F_0 en situation de virtual multipitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

nettement atténué. PSH_{min} témoigne d'un comportement très satisfaisant dans cette situation puisque les deux pics des F_0 théoriques émergent très nettement du reste. Le PSH_{min} semble bien traiter tous les cas de virtual multipitch dont le cas à l'octave fait parti. La figure 5.17 présente les résultats obtenus avec PSH_{mul} dans la même situation. L'estimation est correcte et fournit 160Hz et 240Hz en hypothèse F_0. Il ne reste plus aucun pic parasite dans la fonction PSH et le comportement de PSH_{mul}est très satisfaisant.

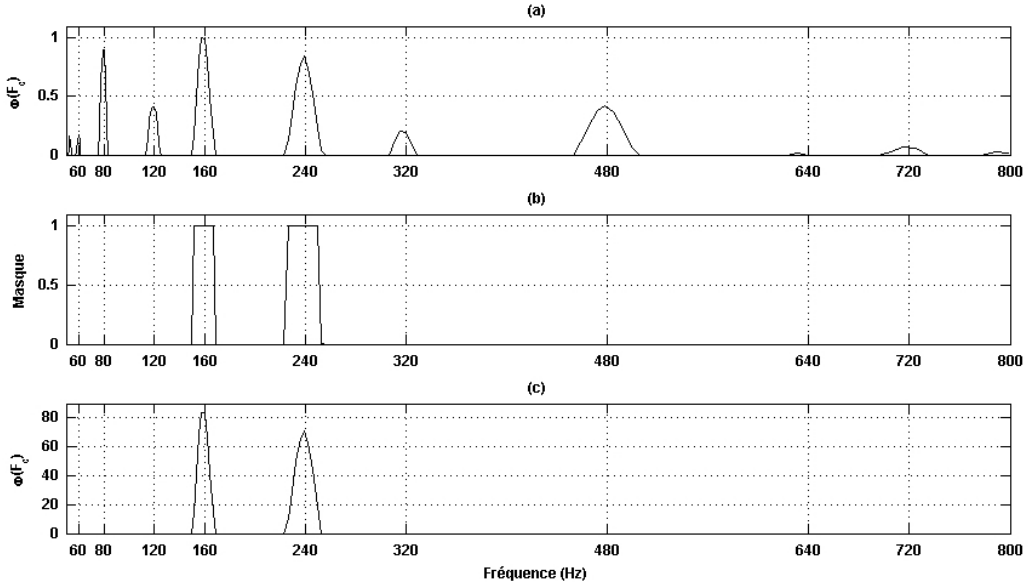

FIGURE 5.17: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch sur un signal synthétique comportant deux F_0 en situation de virtual multipitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.2.3 Signaux synthétiques – situation de F_0 proches

Le signal exemple est construit comme précédemment. Il s'agit de montrer la limite de résolution dans la distinction de deux F_0 proches. La figure 5.18 illustre la fonction PUI, la fonction MAQ et la fonction PSH obtenues dans un mélange synthétique dont une F_0 vaut 233Hz et l'autre 240Hz. L'écart entre les deux F_0 est d'environ un quart de ton ($233 * 2(0.5/12) \approx 240$) soit 3% d'écart relatif ($233 * 1.03 \approx 240$). Le graphique (a) montre la fonction PUI, le graphique (b) la fonction MAQ et le graphique (c) la fonction PSH obtenues par PSH_{min}. Les valeurs estimées sont 233.59Hz et 239.37Hz et sont donc correctes. Les pics parasites sont très fortement atténués. La limite de résolution se situe autour de ce quart de ton soit environ 3% d'écart relatif entre les deux F_0. Un test (non présenté) a été effectué avec un écart de 1.5% environ (un huitième de ton) dans le cas synthétique. PSH_{min} parvient encore à distinguer les deux F_0 mais les deux maxima locaux sont quasiment fusionnés en un seul. Si l'écart relatif est encore diminué, les deux F_0 ne sont plus distinguables. Avec un signal de parole réelle, il faut compter sur une résolution plus faible et par conséquent 3% est un bon ordre de grandeur. La figure 5.19 présente les résultats obtenus par PSH_{mul} dans une situation de F_0 proches. Les graphiques (a) et (b) donnent la fonction PUI avec correction hyperbolique. Le graphique (c) donne le masque binaire et le graphique (d) donne la fonction PSH. L'exemple n'est pas le même que celui utilisé pour PSH_{min} car PSH_{mul} ne parvient pas à discriminer une F_0 à 233 et l'autre à 240Hz. L'exemple pris est donc un mélange d'une structure harmonique de F_0 à 225Hz et l'autre à 240Hz. L'estimation de F_0 est correcte et il n'y a plus aucun pic parasite. Un test (non présenté) montre qu'à partir de 230Hz, PSH_{mul} n'arrive plus à distinguer les deux F_0 et fournit la valeur moyenne comme unique hypothèse F_0 c'est-à-dire 235Hz pour le couple de F_0 (235,240Hz). La limite de résolution de PSH_{mul} est plus faible que PSH_{min} et se situe autour de 5-6% soit autour du demi ton.

Observons à présent les capacités des implémentations de PSH soumis à de la parole réelle bipitch.

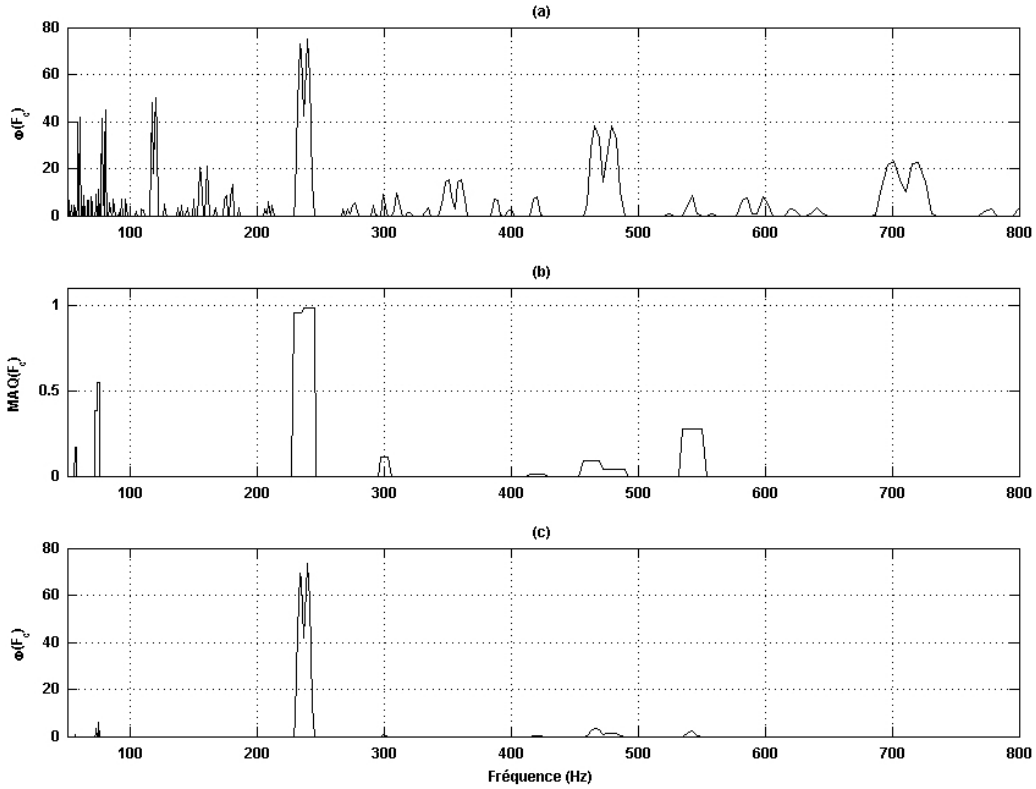

FIGURE 5.18: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch sur un signal synthétique comportant deux F_0 proches. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

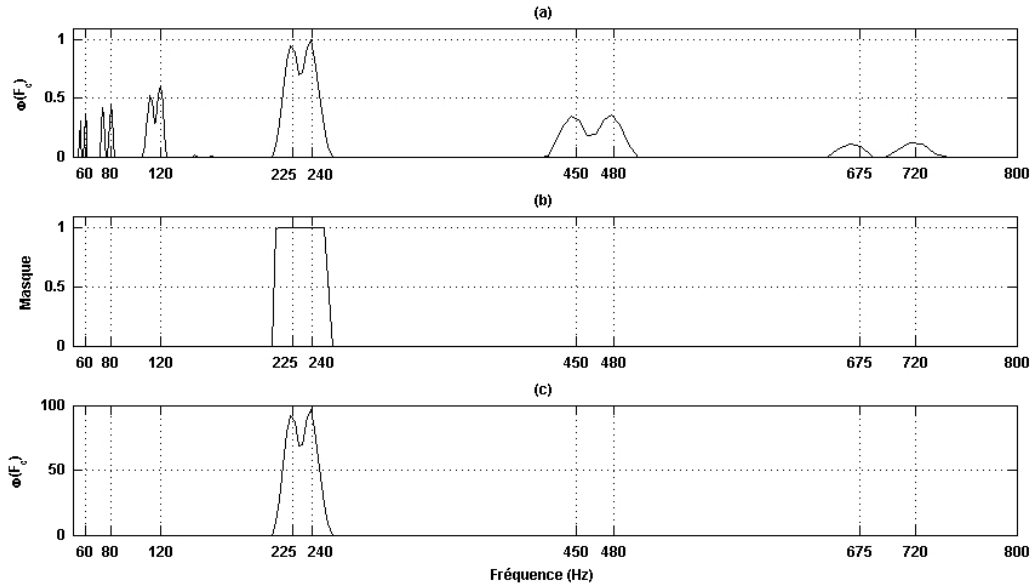

FIGURE 5.19: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch sur un signal synthétique comportant deux F_0 proches. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.2.4 Parole réelle – mélange voix de femme/voix de femme

La première trame est celle du « *i* » de voix de femme utilisé précédemment dans le sous-paragraphe 5.3.1.1. La seconde trame est un « *o* » de voix de femme. Elle est contenue dans le mot « *no* » de la phrase « *I hatch my bets and take no risks* ». Le fichier utilisé est sb034 et appartient au corpus de Bagshaw. Les signaux mélangés sont d'abord normalisés en terme d'intensité. L'intensité est entendue au sens de la fonction « *Get intensity (dB)* » fournie dans le logiciel PRAAT (cf. chapitre 6 équation 6.13 pour le détail). Il sont ensuite additionnés et mis à l'échelle pour que l'extremum du mélange vaille 0.99 afin d'éviter toute saturation (*clipping*) lors de l'enregistrement de la forme d'onde au format WAV. Cette procédure de normalisation est utilisée dans tous les mélanges utilisés dans ce chapitre. La figure 5.20 montre les résultats de PSH_{min} sur un mélange de voix de femmes. Le graphique (a) contient la fonction PUI, le graphique (b) contient la fonction MAQ et le graphique (c) présente la fonction PSH. Les valeurs théoriques à estimer sont 253.4Hz et 289.3Hz. L'estimation

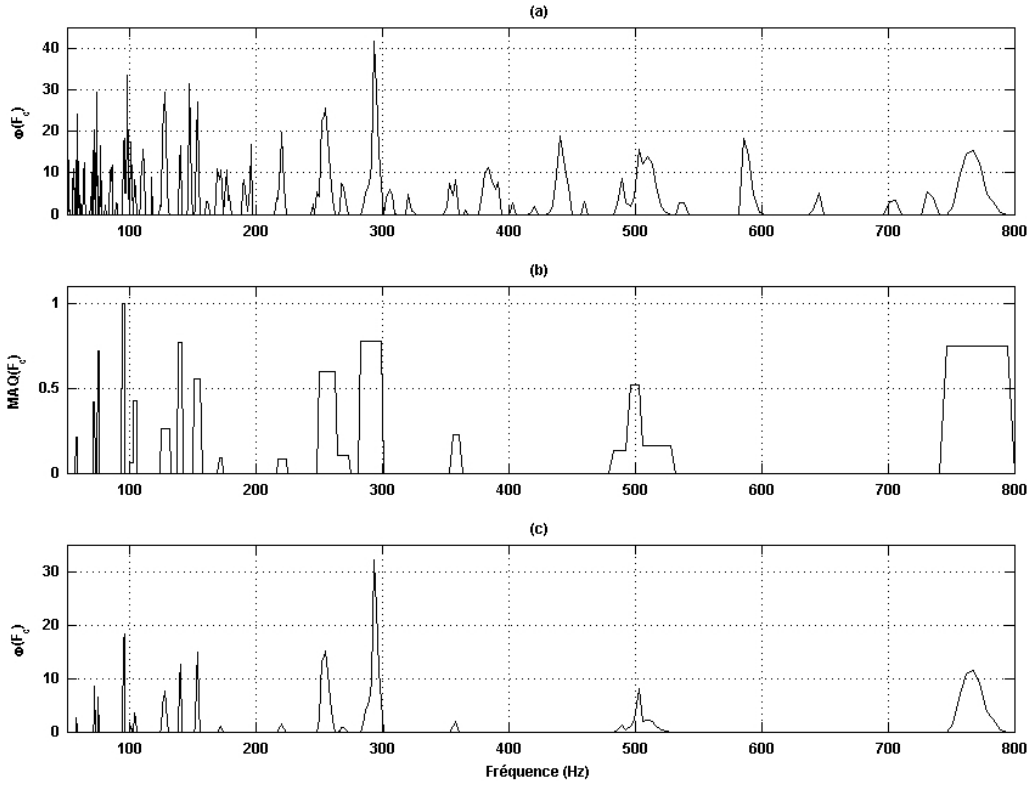

FIGURE 5.20: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch réel voix de femme/voix de femme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

donne 293.3Hz, 95Hz et 254.3Hz. Le pic en 95Hz reste plus élevé que celui en 254.3Hz. Le fait d'avoir à regarder plus d'hypothèses que de F_0 théorique diminue le critère qui sera nommé « précision » dans le reste du manuscrit. La précision d'une trame correspond au nombre d'hypothèses nécessaires pour que toutes les F_0 de références soient trouvées. Dans cet exemple, la précision est de 2/3 car il faut 3 hypothèses pour que PSH donne les deux références F_0. La précision est détaillée dans le chapitre 6 dédié à l'évaluation. La figure 5.21 présente le résultat obtenu par PSH_{mul}. Les graphiques (a) et (b)

représentent la fonction PUI avec correction hyperbolique. Le graphique (c) est la fonction de masque binaire et le graphique (d) est la fonction PSH. L'estimation de F_0 donne 256 et 290Hz en premiers

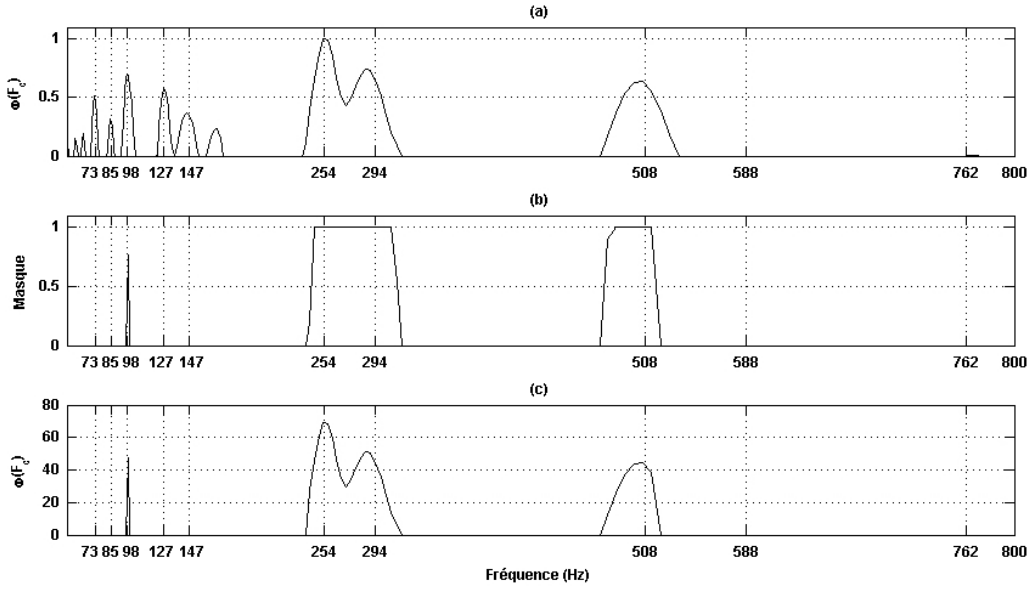

FIGURE 5.21: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch réel voix de femme/voix de femme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

candidats F_0. La précision est donc de 100%. Deux autres pics parasites restent dans la fonction PSH et sont placés en 98Hz et 508Hz.

5.3.2.5 Parole réelle – mélange voix de femme/voix d'homme

La première trame est la trame de « *i* » utilisée dans le sous-paragraphe 5.3.1.1. La seconde trame est la trame de « *a* » utilisée dans le sous-paragraphe 5.3.1.2. Après l'étape de normalisation des intensités, les signaux sont additionnés. La figure 5.22 présente les résultats de PSH_{min} sur le mélange voix de femme/voix d'homme. Le graphique (a) présente la fonction PUI, le graphique (b) présente la fonction MAQ et le graphique (c) présente la fonction PSH. Les F_0 de références valent 110.6Hz et 289.6Hz. L'estimation donne dans l'ordre des amplitudes 110.7Hz, 113.1Hz, 94.5Hz et 293.1Hz. La précision de PSH dans cet exemple est de 1/2. Le pic en 113Hz est du à un maximum local dans le même pic que celui de 110Hz, phénomène déjà décrit dans la figure 5.12. La figure 5.23 présente les résultats obtenus par PSH_{mul}. Les graphiques (a) et (b) donnent la fonction PUI avec correction hyperbolique. Le graphique (c) représente la fonction de masque binaire et le graphique (d) illustre la fonction PSH. L'estimation donne 292 et 113Hz ce qui est correct. Il reste quatre pics parasites qui sont d'amplitude inférieure à la moitié des maxima des F_0 à estimer.

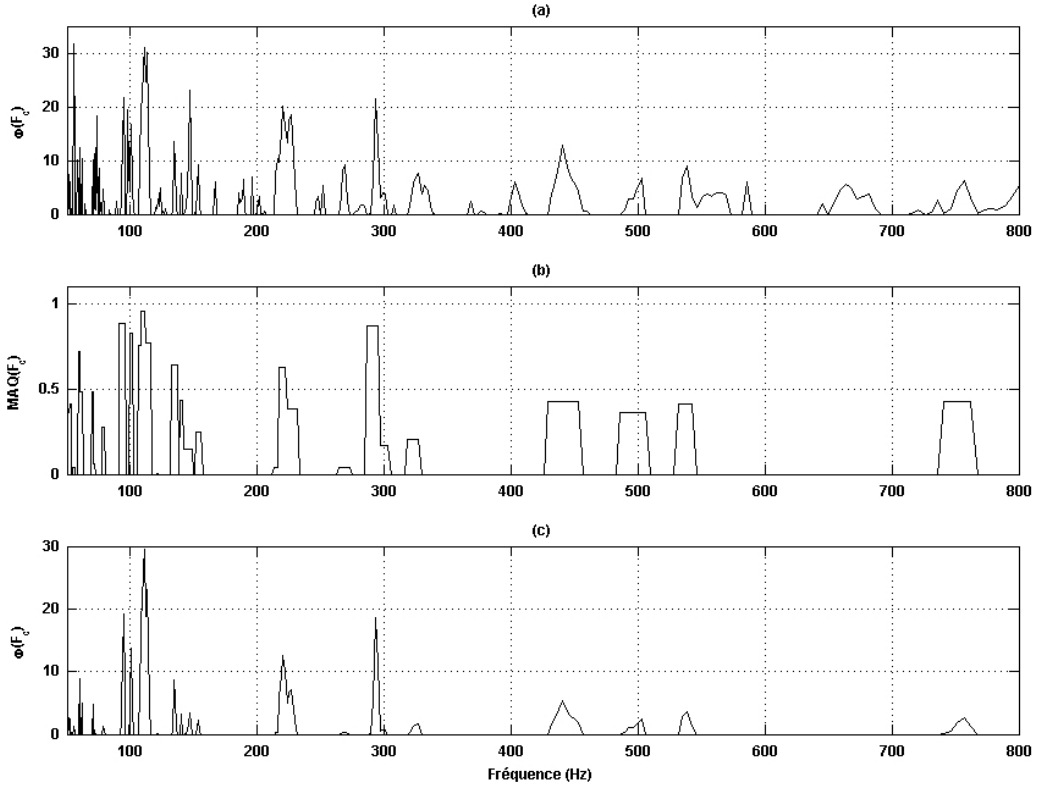

FIGURE 5.22: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch réel voix de femme/voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

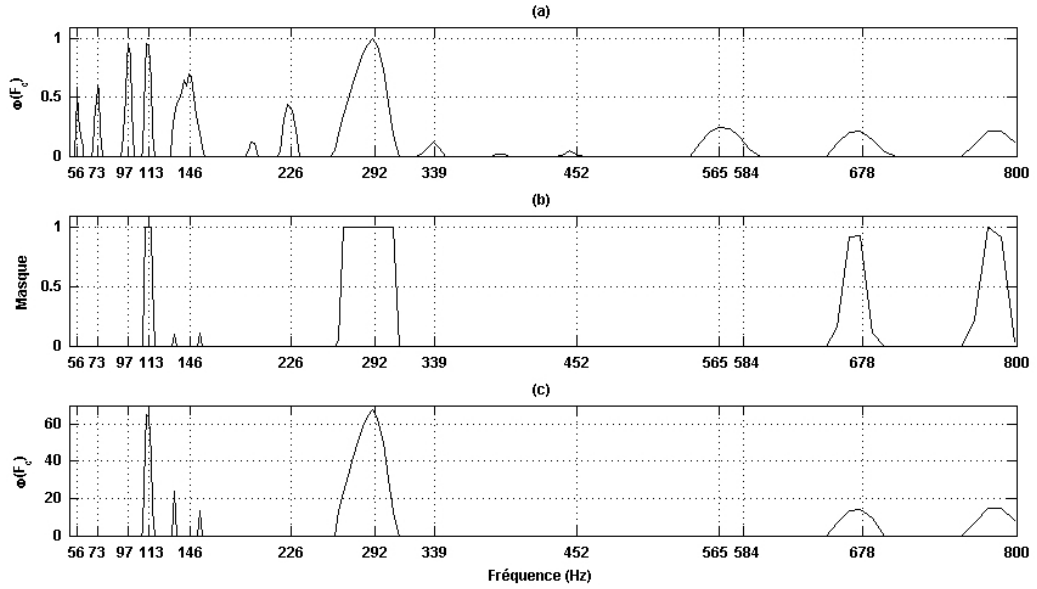

FIGURE 5.23: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch réel voix de femme/voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.2.6 Parole réelle – mélange voix d'homme/voix d'homme

La première trame contient le « *a* » utilisé dans 5.3.1.2. La seconde trame est un « *ou* » extrait du signal rl023 appartenant au corpus de Bagshaw. La phrase prononcée est : « *I can't move my legs* ». Le son « *ou* » provient du mot « *move* ». La mélange est obtenu après normalisation des intensités. La figure 5.24 présente les résultats obtenus par PSH_{min}. Les graphiques (a), (b) et (c) illustrent respectivement les fonctions PUI, MAQ et PSH. Les valeurs de F_0 références sont 101.6Hz et

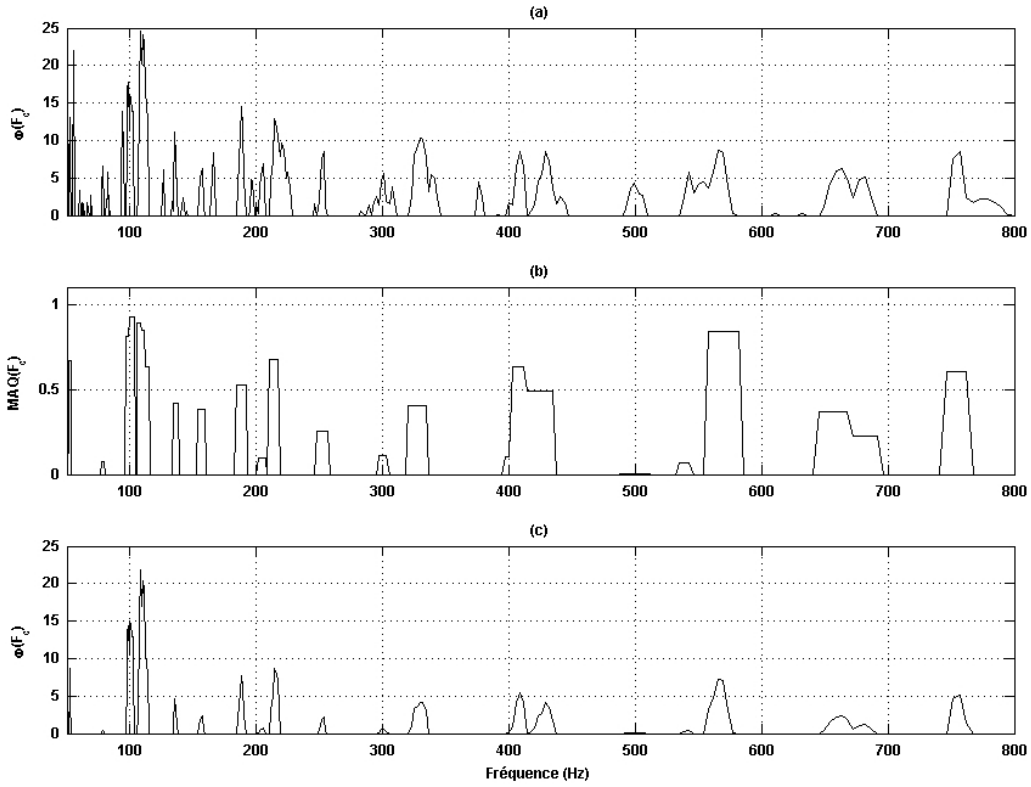

FIGURE 5.24: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation bipitch réel voix d'homme/voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

110.6Hz. L'estimation donne 108.3Hz, 110.4Hz et 100.3Hz. L'estimation de F_0 multiple est correct mais la précision est de 2/3. L'amplitude des pics parasites est rejetée à un peu plus de la moitié de celle des pics F_0 ce qui est assez satisfaisant. La figure 5.25 présente les résultats de PSH_{mul} sur le même exemple. L'estimation donne 112Hz et 99Hz. L'estimation reste correcte et la fonction PSH ne conserve cinq pics parasites. Sur les trois exemples réels, PSH_{mul} se montre plus performant que PSH_{min} notamment dans le nombre de pics parasites restants dans la fonction PSH et dans la précision requise pour retrouver les bonnes estimations de F_0. Cette observation est moins vraie sur des signaux synthétiques ce qui est logique puisque PSH_{min} est issu de considérations purement théoriques. De manière générale, le passage à une situation bipitch est bien réussi pour les deux AEP et notamment pour PSH_{mul}. La suppression des pics parasites est performante. Il apparaît logiquement une dégradation des performances entre l'utilisation de signaux synthétiques et les signaux réels. Les pics parasites restant sont plus nombreux avec des signaux de parole réelle. La précision de nos AEP diminue et il faut plus d'hypothèses pour

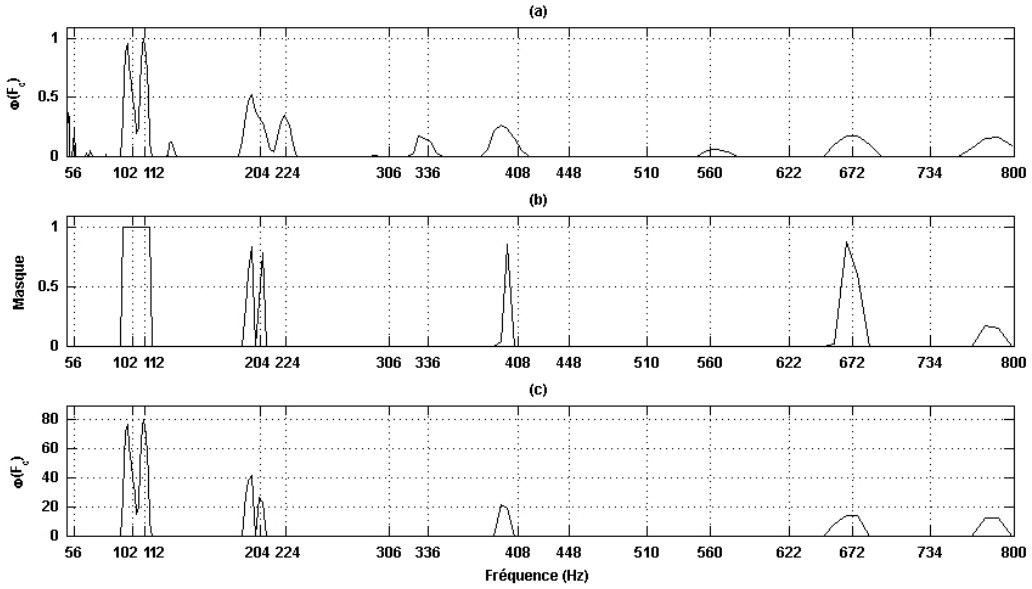

FIGURE 5.25: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation bipitch réel voix d'homme/voix d'homme. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

retrouver les références même si cela n'est pas encore trop visible pour PSH_{mul}. Qu'en est-il d'un passage à la situation plus complexe 4-pitch ?

5.3.3 Situations 4-pitch

Dans un premier temps, deux trames 4-pitch synthétiques sont utilisés pour analyser le comportement de PSH dans un cas relativement général puis dans le cas de 4 F_0 en série octave. Les sous-paragraphe 5.3.3.1 et 5.3.3.2 traitent ces deux cas. Dans un second temps, une trame 4-pitch de parole réelle est utilisé pour confronter PSH à une situation réaliste et compliquée à résoudre. Cette situation est traitée dans le sous-paragraphe 5.3.3.3.

5.3.3.1 Signaux synthétiques – situation générale

Le signal synthétique est un mélange 4-pitch dans lequel F_{01} vaut 200Hz, F_{02} vaut 240Hz, F_{03} vaut 280Hz et F_{04} vaut 320Hz. Les structures harmoniques ont toutes 10 harmoniques qui décroissent en racine inverse. La figure 5.26 présente les résultats obtenus par PSH_{min}. Les graphiques (a), (b) et (c) illustrent respectivement les fonctions PUI, MAQ et PSH. L'estimation donne 279.9Hz, 320.1Hz, 240Hz, 399.9Hz, 78Hz et 199.7Hz. La résolution de PSH_{min} dans cet exemple est de 2/3 car il faut 6 hypothèses pour retrouver les 4 F_0 de référence. Le pic en 400Hz est un pic parasite important dont la suppression est difficile car il est le double de 200Hz (F_{01}) et d'autre part le pic $(5, 4)$ de F_{04}. La figure 5.27 présente les résultats obtenus par PSH_{mul}. Le graphique (a) présente la fonction PUI et le graphique (b) en présente la partie positive. Le graphique (c) présente la fonction de masque binaire et le graphique (d) illustre la fonction PSH. L'estimation donne les hypothèses F_0 suivantes : 80, 66, 240, 161, 320, 93, 277 et 199Hz. Les pics des F_0 à estimer sont bien présents mais ne sont plus les

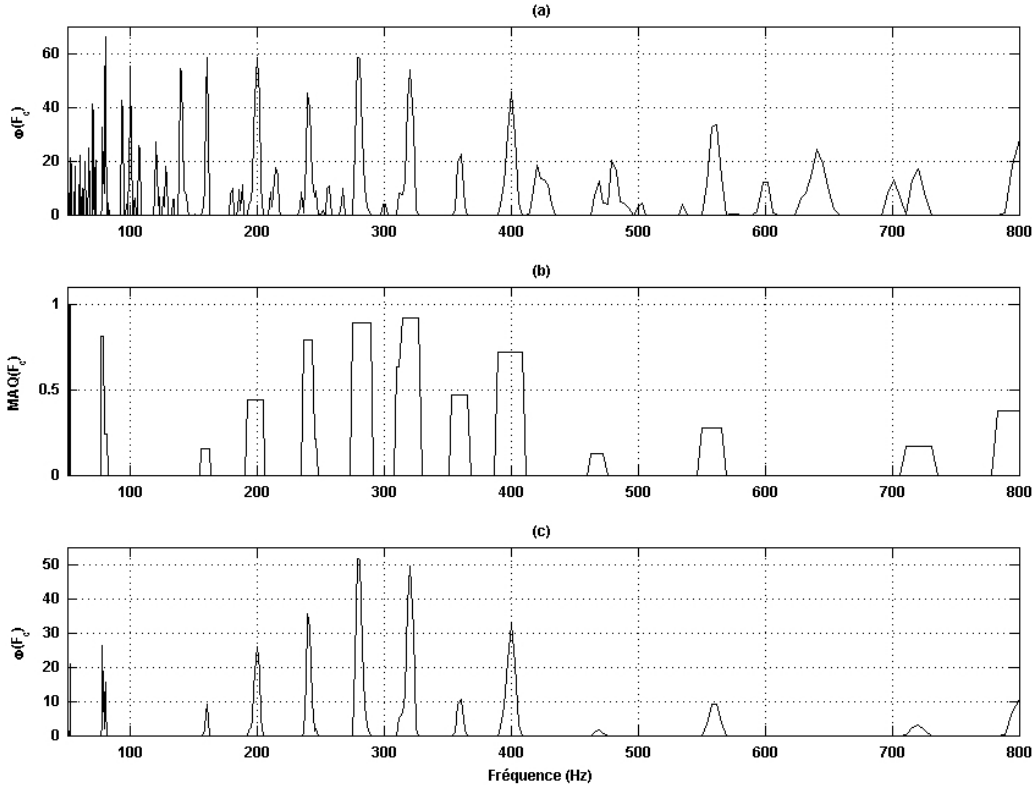

FIGURE 5.26: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation synthétique 4-pitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

pics de plus forte amplitude. Ceci se traduit par une chute de la précision qui vaut 1/2 pour l'exemple considéré.

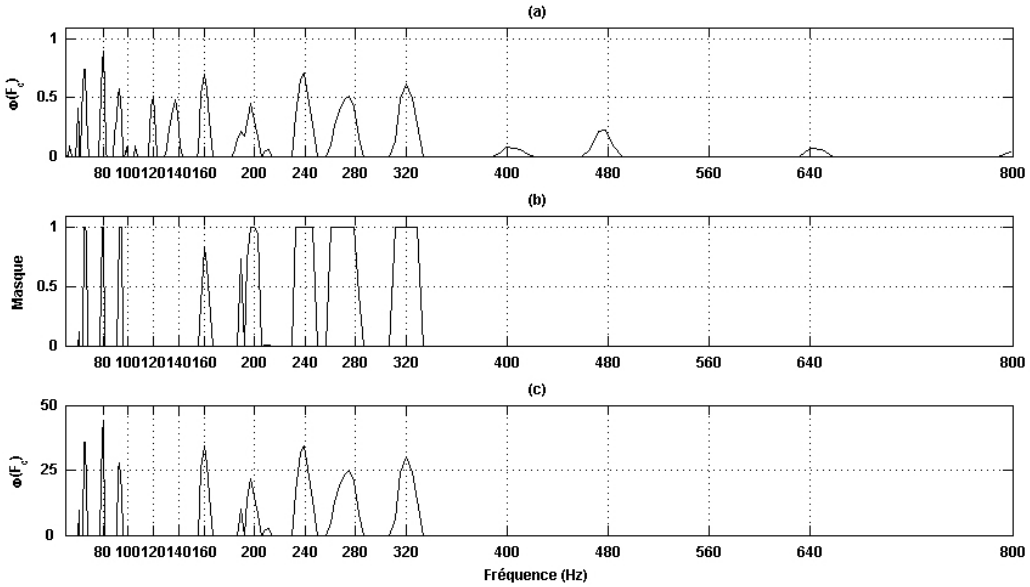

FIGURE 5.27: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation synthétique 4-pitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.3.3.2 Signaux synthétiques – situation en série d'octave

Le signal synthétique est un mélange 4-pitch dont les quatre F_0 sont en série octave. F_{01} vaut 60Hz, F_{02} vaut 120Hz, F_{03} vaut 240Hz et F_{04} vaut 480Hz. Le nombre d'harmonique est limité à 8 pour que l'harmonique la plus aiguë (3840Hz) soit située au-dessous de 4kHz. La figure 5.28 présente les résultats obtenus par PSH_{min}. Les graphiques (a), (b) et (c) illustrent respectivement les fonctions PUI, MAQ et PSH. L'estimation donne 240Hz, 479.9Hz, 120Hz, 52.5Hz et 61.9Hz. La précision du PSH est de

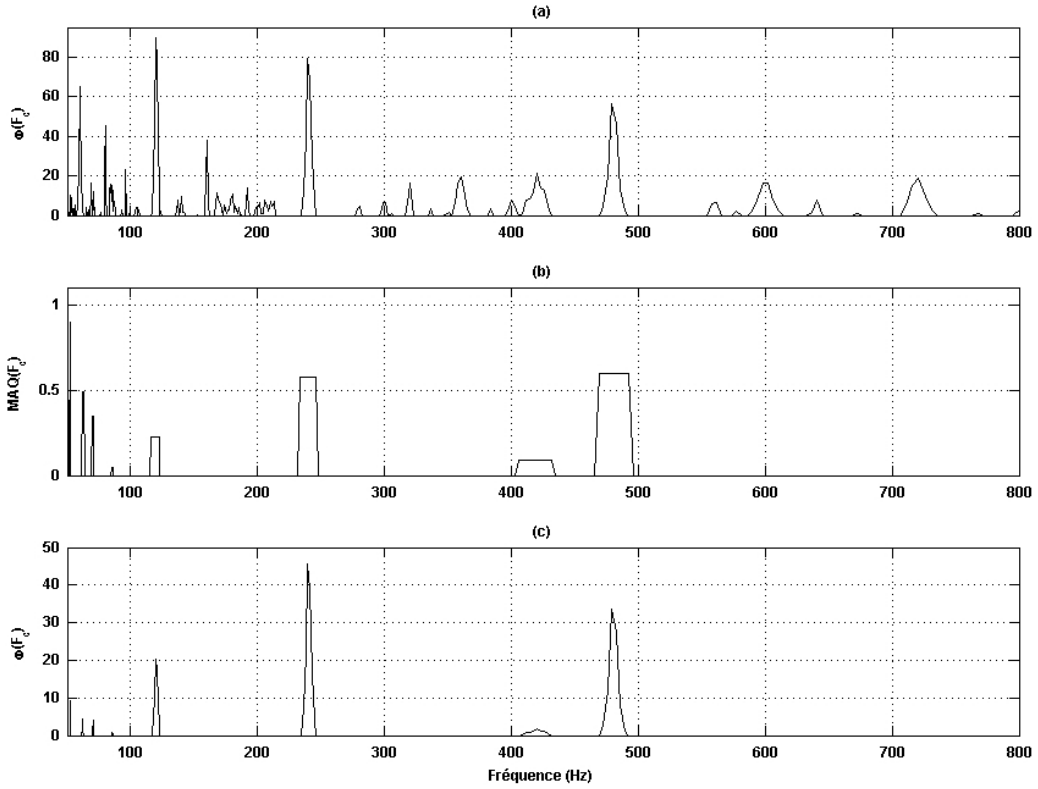

FIGURE 5.28: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation synthétique 4-pitch avec les F_0 en série octave. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

4/5 dans cet exemple. Toutefois, le pic en 61.9Hz est un pic parasite qui sera pris pour une « bonne » hypothèse (à 3.2% près) car le véritable pic en 60Hz a été supprimé dans la fonction DPP. La figure 5.29 présente le résultat obtenu par PSH_{mul}. Le graphique (a) présente la fonction PUI et le graphique (b) en présente la partie positive. Le graphique (c) présente la fonction de masque binaire et le graphique (d) illustre la fonction PSH. Les trois premiers candidats de l'estimation de F_0 sont 120Hz, 241Hz et 57Hz. L'hypothèse F_0 en 57Hz peut être considérée comme correcte (à 5% près de 60Hz) mais est en réalité une erreur. Il apparaît en effet dans le masque binaire du graphique (c) que le pic en 60Hz est rejeté, comme pour PSH_{min}. L'hypothèse en 480Hz est perdue mais était de toute façon fortement atténuée dès la fonction PUI du graphique (b). Ceci s'explique sans doute par le fait que FMAX soit fixé à 1kHz ce qui autorise la fonction PUI à ne prendre en compte que deux harmoniques (480Hz et 960Hz). Sur ces signaux synthétiques, PSH_{min} semble donner sensiblement de meilleurs résultats. Ce phénomène s'explique justement par le caractère synthétique des signaux utilisés sur lesquels PSH_{min} a

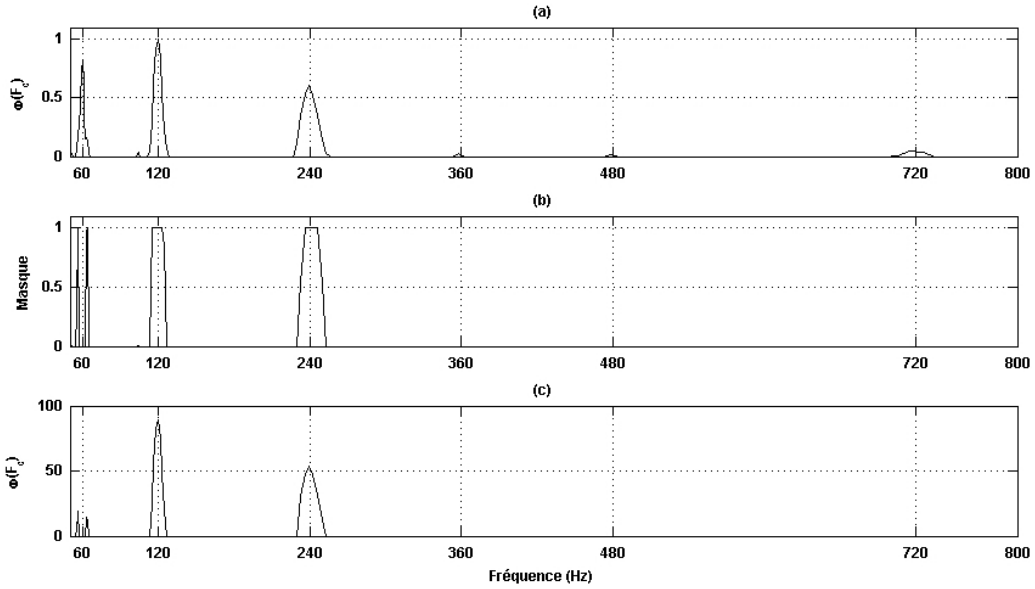

FIGURE 5.29: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation synthétique 4-pitch avec les F_0 en série octave. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

été développé. Il est probable que PSH_{mul} soit meilleur en situation réelle et c'est ce qui se vérifie dans le sous-paragraphe suivant présentant une trame 4-pitch de parole réelle ?

5.3.3.3 Parole réelle – mélange de 4 locuteurs

Le signal utilisé est le mélange des quatre trames utilisées dans les différentes situations bipitch (2 trames de voix de femme « *i* » et « *o* » et 2 trames de voix d'homme « *a* » et « *ou* »). L'étape de normalisation des intensités est réalisée avant de sommer les trames. La figure 5.30 présente les fonctions PUI, MAQ et PSH obtenues par l'algorithme PSH_{min}. Ces fonctions sont présentées respectivement dans les graphiques (a), (b) et (c). Les F_0 de référence sont 101.6Hz, 110.6Hz, 253.4Hz et 289.6Hz. L'estimation donne 293Hz, 100.6Hz, 134.1Hz, 438.1 Hz, 101.9Hz, 219.3Hz, 402.5Hz, 52.6Hz, 113.2Hz et 248Hz. La précision est de 2/5. Le nombre et l'amplitude des pics parasites est plus importante et le problème de multiples maxima locaux dans un seul et même pic se retrouve largement ici (cf. figure 5.12). Le pic en 255Hz est perdu dans la fonction DPP mais l'hypothèse en 248Hz sera associée à cette valeur de référence à 2.75% près. La figure 5.31 présente (CONTINER) L'estimation donne par ordre de décroissance des amplitudes : 99, 259, 57, 133, 292 et 113Hz. L'estimation de F_0 est correcte à condition de considérer 6 hypothèses et non 4. La précision est donc de 2/3 ce qui est bien meilleur que pour PSH_{min}. La fonction PSH est bien plus claire que celle de PSH_{min} et nous retrouvons le résultat attendu montrant que PSH_{mul} est plus performant dans les situations réelles de mélange de parole. Bien sûr ceci n'est qu'un exemple mais le chapitre 6 d'évaluation démontre le bon fonctionnement de PSH_{mul}.

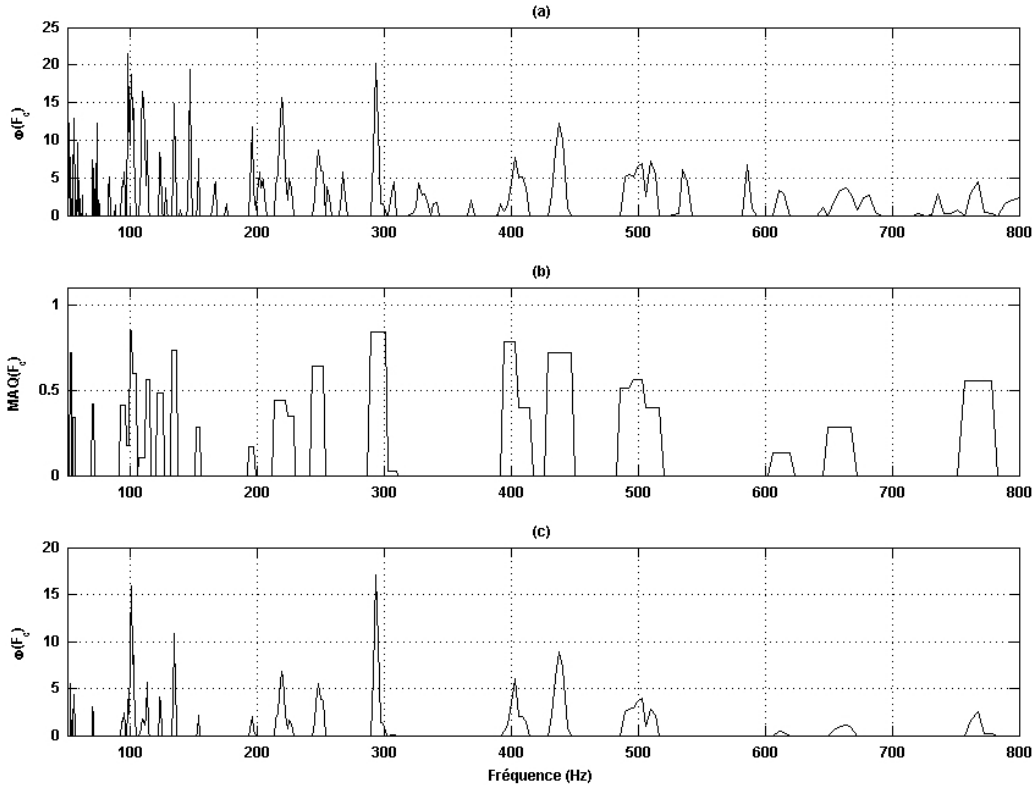

FIGURE 5.30: Fonctions PUI, MAQ et PSH obtenues par PSH_{min} en situation réelle 4-pitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

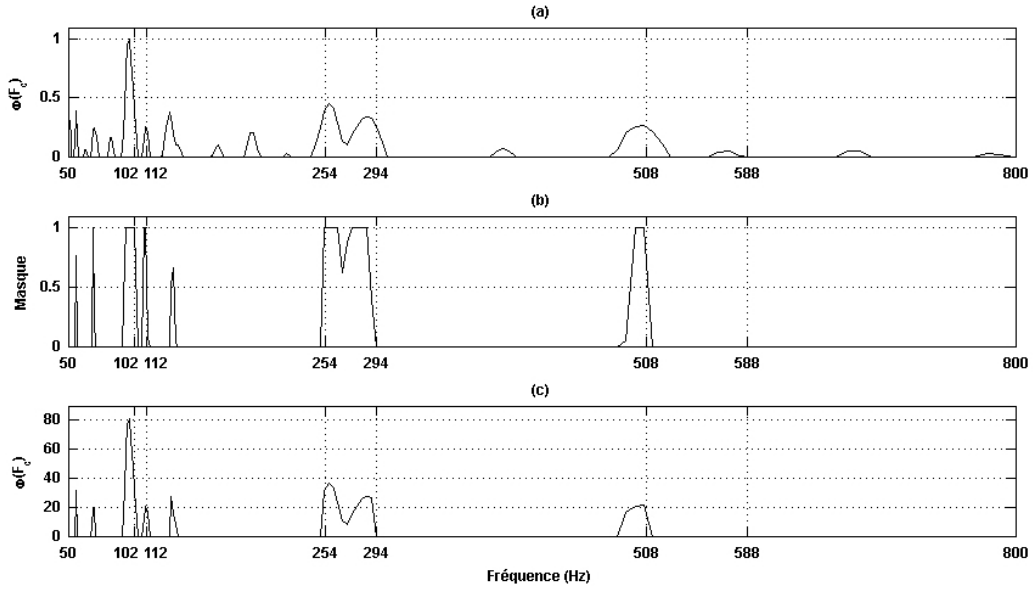

FIGURE 5.31: Fonctions PUI, MAQ et PSH obtenues par PSH_{mul} en situation réelle 4-pitch. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

Il a été vu dans cette section que les deux implémentations de PSH fonctionnent bien dans la situation monopitch. En situation bipitch, le cas à l'octave ne semble pas poser trop de problèmes. Il en va de même pour la situation de virtual multipitch au moins en situation de mélanges synthétiques. La résolution fréquentielle de PSH_{mul} est de l'ordre du demi-ton soit 6% d'écart relatif. Celle de PSH_{min}est de l'ordre du quart de ton. Une étude approfondie sur ce point serait utile et appartient aux perspectives directes du travail de thèse. Les résultats de la situation 4-pitch sont tout à fait corrects mais plus nuancés. PSH_{min} semble en mesure de traiter des cas synthétiques mais semble limité lorsqu'il est confronté à des cas de mélanges réels. PSH_{mul} parvient à traiter le cas réel 4-pitch ce qui est encourageant pour le chapitre 6 d'évaluation. Il y parvient au détriment de sa précision. Cette baisse de performance est imputable au caractère non stationnaire de la parole (F_0 non stable, jitter, détection des partiels haute-fréquence plus difficile, présence de bruit, etc.) et au fait que dans un mélange, les informations spectro-temporelles d'un signal dominant occultent les informations spectro-temporelles des signaux dominés. Ceci se traduit par la difficulté de retrouver les composantes harmoniques d'un signal dominé et donc par un pic de faible amplitude dans la fonction PSH. La section 5.4 va discuter des limitations au bon fonctionnement d'algorithmes fondés sur le principe de PSH.

5.4 Situations litigieuses

Il existe plusieurs sources de limitations d'algorithme basés sur PSH. Ces limitations sont étudiées au travers de l'algorithme PSH_{min}. Trois d'entre elles sont précisément décrites dans cette section. Le paragraphe 5.4.1 présente les limites dans l'estimation de F_0 multiples imposées par les différences de niveaux sonores entre signaux en relation dominant/dominé. Le paragraphe 5.4.2 discute des limites imposées le nombre d'harmonique et l'agencement des structures harmoniques mélangées. Ils peuvent être tels qu'il devient compliqué d'estimer correctement toutes les F_0 puisque les partiels sont difficilement détectables. Le paragraphe 5.4.3 présente un exemple de robustesse au bruit en utilisant un bruit rose.

5.4.1 Intensité des signaux concurrents

Les intensités respectives des signaux mélangés influent sur le spectre d'amplitude et par conséquent sur l'étape de détection des partiels. Il est bien évidement plus difficile de détecter les partiels d'un son dominé en intensité. Dans l'exemple utilisé ci-dessous, la structure harmonique de F_{01} (200Hz) est d'amplitude moitié à celle de la structure harmonique de F_{01} (240Hz). Les deux structures possèdent le même nombre d'harmoniques qui est 10. La différence d'intensité des deux signaux est de 6dB au sens de l'équation 6.13 du chapitre 6. La figure 5.32 présente le spectre d'amplitude du mélange des deux structures harmoniques. Les ronds rouges sont les partiels conservés par l'étape de détection des partiels de PSH_{min}. La structure harmonique de F_{01} à 200Hz est effectivement moins détectée que celle de F_{01}. Ceci se traduit dans la fonction PSH comme l'illustre la figure 5.33. L'amplitude du pic en 200Hz est réduite. L'estimation donne 239.9Hz, 400Hz et 199.9Hz. La précision de PSH_{min} vaut 2/3. Une différence d'amplitude des deux signaux diminue la précision de PSH jusqu'au moment où la différence de niveau est telle que le pic F_0 du son dominé disparaît. Ce phénomène explique pourquoi

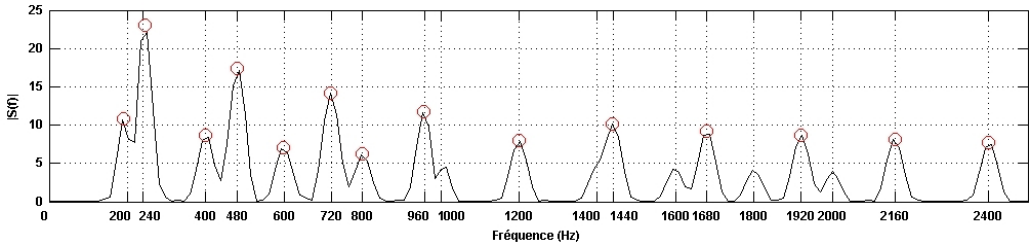

FIGURE 5.32: Spectre d'amplitude du mélange (noir) et résultat de la détection des partiels (ronds rouges).

nous nous sommes efforcés de mélanger uniquement des signaux de mêmes intensités.

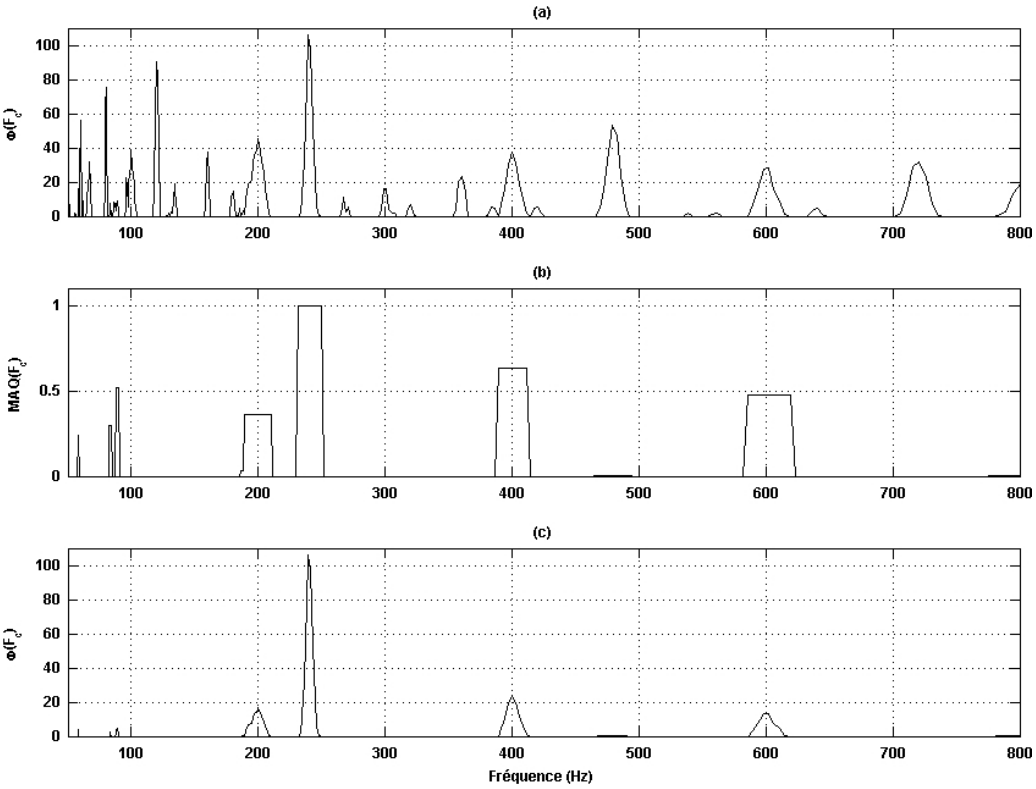

FIGURE 5.33: Fonctions PUI, MAQ et PSH de PSH_{min}. Les signaux synthétiques mélangés ont le même nombre d'harmoniques mais une différence d'intensité de 6dB. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

5.4.2 Qualité des structures harmoniques

La figure 5.34 présente le spectre d'amplitude du mélange (trait noir) ainsi que les partiels détectés (ronds rouges). Il est à noter que lorsque deux harmoniques sont proches en amplitude et en fréquence, ils ont tendance à fusionner en une seule harmonique. Dans ce cas, l'étape de détection de partiels ne détecte pas deux partiels distincts mais un seul dont la fréquence centrale correspond environ à la moyenne des deux fréquences. Ce phénomène de fusion harmonique est préjudiciable à l'estimation de F_0. La fusion harmonique se produit à deux reprises dans la figure 5.34 aux fréquences 220Hz et 980Hz

c'est-à-dire aux fréquences harmoniques communes (ou presque). La figure 5.35 présente le résultat de

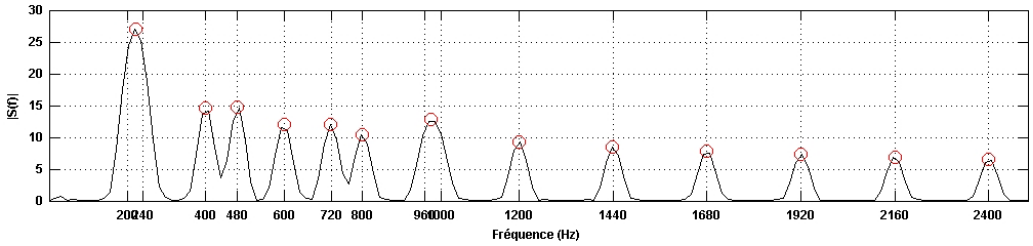

FIGURE 5.34: Spectre d'amplitude du mélange (noir) et résultat de la détection des partiels (ronds rouges).

PSH_{min} sur un tel signal. L'estimation de F_0 donne 240.1Hz, 400Hz, 600.Hz et 199.6Hz. La précision est de 1/2. Une différence du nombre d'harmonique a relativement le même effet qu'une différence d'amplitude. Étant donné que l'une des deux structures harmoniques est plus faible que l'autre, elle est moins présente dans la fonction PSH. Les structures harmoniques interagissent entre-elles parti-

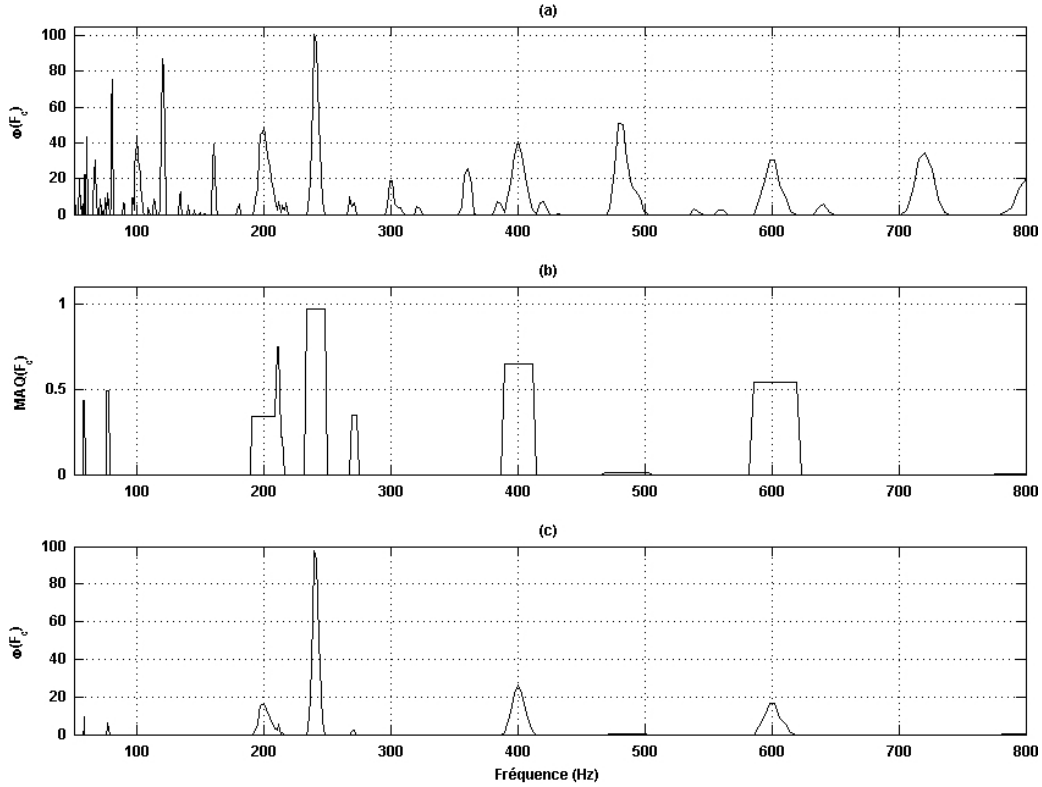

FIGURE 5.35: Fonctions PUI, MAQ et PSH de PSH_{min}. Les signaux synthétiques mélangés ont la même intensité mais un nombre d'harmoniques différent (10 contre 5). (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

culièrement aux harmoniques communes. Ceci peut avoir pour effet de détruire les harmoniques et donc d'affaiblir la structure harmonique. PSH détectant les structures harmoniques, il lui sera difficile d'estimer correctement les F_0 à l'échelle de la trame. C'est dans ce genre de situation que le suivi de F_0 peut intervenir. Ce cas limite met en lumière l'inconvénient d'une méthode basée uniquement sur le spectre d'amplitude. Il y a par exemple des cas ambigus par nature. Soit une structure harmonique

de 10 harmoniques à 240Hz et une autre de 5 harmoniques à 480Hz, les deux structures étant de relativement mêmes amplitudes. Dans ce cas très particulier, il est impossible pour le PSH et même pour n'importe quel AEP exploitant uniquement le spectre d'amplitude de voir la structure en 240Hz. La seule possibilité de voir la structure en 480Hz serait dans le cas où celle-ci serait d'amplitude bien supérieure à la structure en 240Hz. Une piste à explorer serait de voir en quelle mesure la phase peut apporter une aide dans l'estimation de F_0.

5.4.3 Robustesse au bruit

La robustesse à un bruit rose sera testée ici. L'utilisation d'un bruit rose est justifiée car la majorité de l'énergie d'un signal de parole se situe dans les basses fréquences. Un bruit blanc gaussien étend son énergie sur tout le spectre. Le bruit rose concentre son énergie dans les basses fréquences (fréquence de début de décroissance en -3dB par octave : 50Hz). Un bruit important doit nuire au niveau de la détection des partiels. Le rapport signal sur bruit ou SNR (*Signal-to-Noise Ratio* en anglais) est de 0dB ce qui signifie que le bruit est de même intensité que le signal. Quand le bruit est rose, un tel SNR correspond à un signal assez bruité. La figure 5.36 présente la fonction PSH obtenue par PSH_{min} dans le graphique (c). Le graphique (a) est la fonction PUI qui est beaucoup plus bruitée que son équivalente sans bruit de la figure 5.22. La fonction MAQ est donnée dans le graphique (b). Les valeurs théoriques

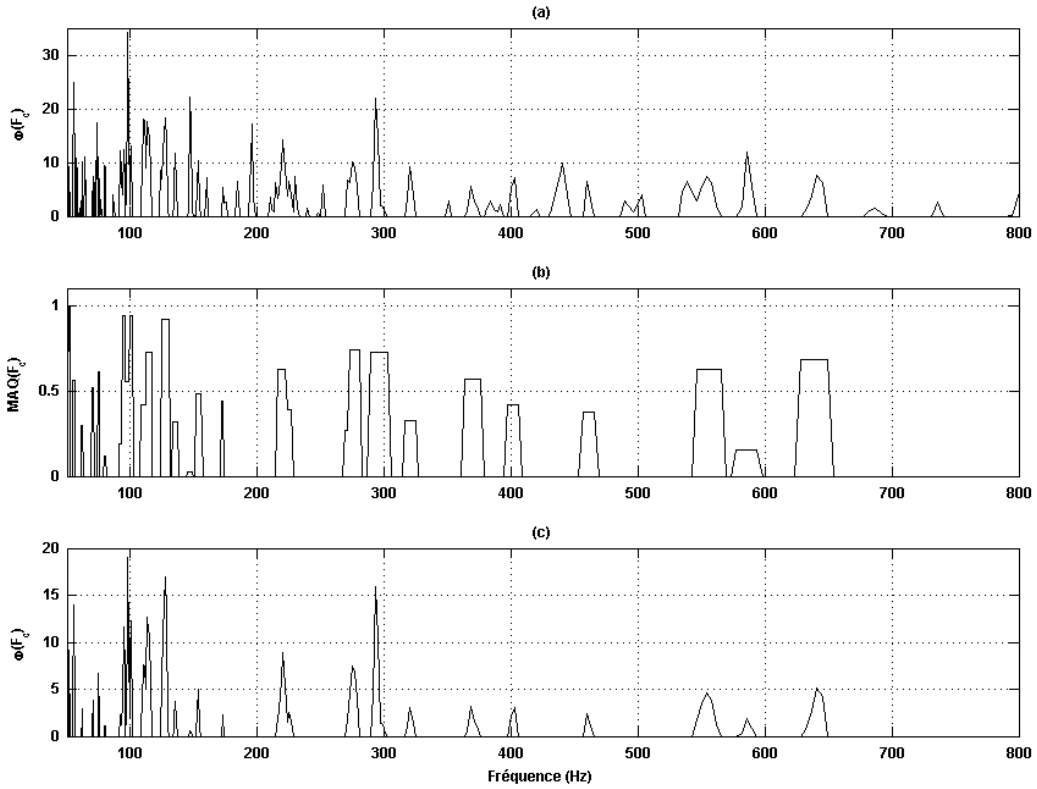

FIGURE 5.36: Fonctions PUI, MAQ et PSH de PSH_{min}. Signal de parole réelle bipitch avec bruit rose de même intensité. (a) Fonction PUI. (b) Fonction MAQ. (c) Fonction PSH.

sont 110.6Hz et 289.6Hz. L'estimation par PSH donne 97.9Hz, 127.4Hz, 293.2Hz, 55.2Hz, 113.5 Hz.

La précision est réduite à 2/5 et il faut attendre la troisième hypothèse pour trouver l'une des deux références F_0. Si les signaux traités sont bruités, il faut augmenter le nombre d'hypothèses fournies par PSH garder une bonne estimation de F_0.

5.5 Conclusions

Ce chapitre a présenté le principe d'un Peigne à Suppression Harmonique et deux implémentations particulières nommées PSH_{min} et PSH_{mul}. Ces algorithmes trame-à-trame extraient le spectre d'amplitude de chaque trame, en construisent une fonction PUI quantifiant la force de la structure harmonique de chaque fréquence dans un intervalle de recherche de F_0. Les deux implémentations utilisent alors le principe de PSH par la combinaison des fonctions PDN et PDM aux différents ordres. Les propriétés des PUI, PDN et PDM permettent d'atténuer les pics parasites de la fonction PUI. Les résultats obtenus sur des sons synthétiques et réels sont convaincants même dans des situations de mélanges à l'octave. Les limites des AEP fondés sur le principe de PSH sont de l'ordre des intensités des signaux mélangés et de la qualité des structures harmoniques du spectre d'amplitude. Dans le chapitre suivant, l'implémentation PSH_{mul} est évaluée.

Chapitre 6

Évaluation de la qualité des AEP

6.1 Introduction

Après l'implémentation du principe de PSH, le souci a été d'évaluer nos AEP et de les comparer à l'existant. Le processus d'évaluation d'AEP a fait l'objet d'une publication [Signol et al., 2008] et l'article est donné en annexe E. Une procédure d'évaluation dépend très largement de l'application que l'on désire en faire. L'évaluation peut être simplement conçue comme un outil de mesure de performance d'un algorithme. Dans ce cas un seul algorithme est évalué à un instant donné. L'évaluation peut aussi être conçue comme un outil de mesure des améliorations apportées à un algorithme. Dans ce cas, l'évaluation devient comparative puisqu'elle compare les performances de différentes versions d'un même algorithme au cours du temps. Enfin, l'évaluation peut être conçue comme un outil de comparaison de différents algorithmes. C'est le point de vue adopté et présenté dans ce chapitre. Il oblige à être équitable vis à vis de tous les algorithmes comparés. Cette équité impose le contrôle strict de l'ensemble des variabilités qui sont inhérentes au processus d'évaluation, des stimuli jusqu'aux spécificités des AEP. Tous les paramètres qui peuvent influencer les résultats doivent être identifiés et « normalisés ». Ce chapitre vise davantage à proposer des méthodologies d'évaluation comparative rigoureuses qu'à entrer dans la compétition simpliste décidant d'un meilleur et d'un plus mauvais AEP.

La première étape d'une évaluation concerne les signaux de parole mis en entrée des AEP qui doivent être évidemment les mêmes pour tous les AEP. Le choix des corpus doit être fait en fonction des caractéristiques des AEP. Il est évident que si un AEP n'est pas conçu pour traiter un certain type de signaux et qu'un des corpus en contient, il sera désavantagé. Chacun des corpus peut contenir des natures de parole différentes. Il est donc nécessaire d'utiliser des corpus homogènes en termes de nature de signaux. Par exemple, fusionner les résultats d'évaluation sur des signaux de parole lue et de parole spontanée paraît déraisonnable. De plus, chaque corpus est enregistré de manière différente et il est fort probable qu'il y ait des disparités d'intensité sonore. Il est donc nécessaire de procéder à une étape de normalisation d'intensité pour tous les signaux des corpus. Cette étape est d'autant plus importante que les signaux sont mélangés. La procédure de normalisation est détaillée dans le sous-paragraphe 6.5.1.2. La seconde étape concerne l'annotation des références F_0 utilisée. Ces références doivent évidemment être les mêmes pour tous les AEP. Elles peuvent être construites manuellement ou bien automatiquement (cf. paragraphe 6.5.2). Ces deux premières étapes normalisent des paramètres

qui sont indépendants du fonctionnement interne des AEP.

La troisième étape concerne les paramètres internes des AEP qu'il convient d'uniformiser au maximum. Les paramètres qui doivent nécessairement être identiques pour tous les AEP sont les instants d'acquisition des hypothèses F_0 et l'intervalle de recherche de la F_0. Dans la mesure du possible, la durée des trames et les post-traitements utilisés dans certains AEP doivent être maitrisés mais cela n'est pas toujours possible.

La section 6.2 présente les métriques d'évaluation utilisées dans ce manuscrit. Ces métriques ont pour objectif de tester les deux processus inhérents à tout AEP : la décision voisé/non-voisé (ou décision VnV) prise sur chacune des trames et l'estimation de F_0. Il est aussi important d'évaluer le processus de décision VnV que celui d'estimation de F_0 car le premier affecte les résultats du second. La section 6.3 présente le lien étroit entre la quantification de la force de périodicité (ou force de voisement), la décision VnV et les résultats d'évaluation des AEP, mettant en évidence l'influence de la décision VnV sur les résultats d'évaluation. Ceci nous amène à proposer plusieurs méthodologies évaluatives dans la section 6.4 valables dans toute situation monopitch et multipitch. La section 6.5 présente en détail la construction de l'annotation des références F_0 qui ont servi à l'évaluation. Elle est réalisée de manière automatique. La section 6.6 présente les évaluations menées en situation monopitch. La section 6.7 présente les évaluations menées en situation bipitch.

6.2 Métriques d'évaluation

L'évaluation des algorithmes d'estimation de F_0 est à l'origine de nombreux travaux. Il n'est pas question ici de détailler l'ensemble des travaux mais de donner quelques pointeurs que le lecteur intéressé pourra consulter. De nombreuses évaluations d'AEP en situation monopitch ont été produites dans la littérature [Rabiner et al., 1976, Immerseel and Martens, 1992, Bagshaw et al., 1993, Mousset et al., 1996, Rouat et al., 1997, de Cheveigné and Kawahara, 2001, Veprek and Scordilis, 2002, Luengo et al., 2007, Camacho, 2007]. Il ressort de ces travaux l'utilisation systématique d'un taux d'erreur grossière (*GER*) qui quantifie la qualité d'estimation de la F_0. Les évaluations de parole superposée multipitch (bipitch exactement) sont nettement moins nombreuses [Wu et al., 2003, Vishnubhotla and Espy-Wilson, 2008] si bien qu'il n'existe pas, à ce jour, de concensus clair dans la manière d'évaluer des AEP multipitch conçus pour la parole superposée. Pour trouver plus de travaux concernant l'estimation de F_0 multiples en général, il est intéressant de se tourner vers la musique. Une campagne d'évaluation a été mise en place sous l'impulsion du domaine de transcription automatique de signaux polyphoniques musicaux. Cette campagne se nomme MIREX [1] (*Music Information Retrieval Evaluation eXchange*). Un lecteur intéressé pourra lire [Poliner et al., 2007, Daniel et al., 2008, Fonseca and Ferreira, 2009, Bay et al., 2009]. Les auteurs utilisent majoritairement des taux de rappel et de précision (*recall rate*, *precision rate*) combinés linéairement en une *F-mesure* qui peut être vue comme une forme de *GER*. Il ressort notamment une analogie intéressante entre le problème de la détection des débuts et fins de note pour les AEP multipitch musicaux et le problème de la décision VnV pour les AEP conçus pour la parole (discuté dans la section 6.3). MIREX existe depuis 2005 et, outre le fait qu'elle ait motivée l'élaboration de nombreux systèmes d'estimation de F_0 multiples pour la musique, elle a aussi permis

1. `http://www.music-ir.org/mirex/2009/index.php/Main_Page`

de mettre en place des méthodologies d'évaluation comparatives « normalisées » et équitables. Il serait fort utile de mettre en œuvre le même genre de campagne évaluation pour la parole superposée.

L'ensemble des métriques utilisées dans ce manuscrit sont inspirées de la lecture des articles cités précédemment. Les métriques en situation monopitch et multipitch sont légèrement différentes et font respectivement l'objet des paragraphes 6.2.1 et 6.2.2. La décision VnV et l'estimation de F_0 doivent être évalués pour que l'évaluation soit complète. Il convient d'introduire des notations ensemblistes qui seront utiles par la suite. La trame de référence numéro i notée R^i est constituée de N_r valeurs F_0 nommées individuellement « valeur F_0 de référence » et notées r_x^i. L'équation 6.1 décrit mathématiquement cette trame. $\{\ldots\}$ désigne l'opérateur « ensemble de ».

$$R^i = \{r_1^i \ldots r_{N_r}^i\} \tag{6.1}$$

La trame d'hypothèse numéro j notée H^j est constituée de N_h valeurs F_0 nommées individuellement « valeur F_0 d'hypothèse » et notées h_x^j comme l'indique l'équation 6.2.

$$H^j = \{h_1^j \ldots h_{N_h}^j\} \tag{6.2}$$

Par convention, une valeur F_0 de référence ou une valeur F_0 d'hypothèse est voisée lorsque sa valeur est différente de 0. Si sa valeur est nulle, alors la valeur F_0 est non-voisé. Une trame est dite voisée lorsqu'au moins une de ses valeurs F_0 est voisé. En situation monopitch, cette définition est naturelle. En revanche, elle l'est moins en situation multipitch car une trame ne comporte plus une seule valeur F_0 mais plusieurs. Ceci se justifie car les trames mélangées étant approximativement de même intensité, si une d'entre-elles possède une structure harmonique, cette structure sera visible dans le mélange. Si plusieurs structures harmoniques sont mélangées, les harmoniques communes vont certes interagir mais les structures harmoniques devraient rester partiellement intactes conservant ainsi le caractère voisé de la trame. C'est pourquoi il peut être raisonnablement considéré que dès qu'une des valeurs F_0 d'une trame est voisé, la trame est considérée comme voisée.

6.2.1 Situation monopitch

Dans la situation monopitch, une trame de référence peut être confondue avec la valeur F_0 de référence qu'elle contient. En effet, N_r vaut 1 ce qui implique que la trame R^i est assimilable à la valeur F_0 de référence r_1^i.

6.2.1.1 Décision VnV

La décision VnV étant à même de produire des erreurs, des critères les quantifiant sont requis. Deux erreurs peuvent être commises : les erreurs de sur-voisé et les erreurs de sous-voisé. Ces erreurs ont été explicitées dans les équations 6.3 et 6.4 respectivement. Une évaluation doit toujours présenter ces deux taux sous peine d'être incomplète. Les trames de référence et d'hypothèse peuvent prendre chacune deux états : voisée ou non-voisée. Si une trame d'hypothèse est voisée et que la trame de référence ne l'est pas, il y a une erreur nommée erreur de sur-voisé. A l'inverse, si une trame d'hypothèse n'est pas voisée alors que la trame de référence l'est, il y a une erreur nommée erreur de sous-voisé. Les erreurs

de sur-voisé sont des fausses alarmes et les erreurs de sous-voisé sont des faux rejets. Le nombre de trames en erreur de sur-voisé rapporté au nombre de trames de référence non-voisées est dénommé taux de sur-voisé (noté OVR comme *Over-Voiced Rate*). L'OVR est une valeur sans unité comprise entre 0 et 100%. Il vaut 100% lorsque toutes les trames de référence non-voisées sont détectées comme voisées par l'AEP. L'équation 6.3 présente mathématiquement le taux de sur-voisé.

$$OVR = 100\frac{\Omega[\{K_{RU}\} \cap \{K_{HV}\}]}{\Omega[\{K_{RU}\}]} \tag{6.3}$$

Ω désigne le cardinal ensembliste c'est-à-dire le nombre d'éléments d'un ensemble. $\cap$ désigne l'opérateur ensembliste d'intersection. $\{K_{RU}\}$ désigne l'ensemble des numéros des trames de références qui sont non-voisées. $\{K_{HV}\}$ désigne l'ensemble des numéros des trames d'hypothèse voisées. Le nombre de trames en erreur de sous-voisé rapporté au nombre de trames de référence voisées est le taux de sous-voisé (noté UVR comme *Under-Voiced Rate*). Il vaut 100% lorsque toutes les trames de référence voisées sont détectées comme non-voisées par l'AEP. L'équation 6.4 présente mathématiquement le taux de sous-voisé.

$$UVR = 100\frac{\Omega[\{K_{RV}\} \cap \{K_{HU}\}]}{\Omega[\{K_{RV}\}]} \tag{6.4}$$

$\{K_{RV}\}$ désigne l'ensemble des numéros des trames de référence voisées. $\{K_{HU}\}$ désigne l'ensemble des numéros des trames d'hypothèse non-voisées.

6.2.1.2 Estimation de F_0

L'estimation de F_0 est à même de faire des erreurs, des critères les quantifiant sont donc requis. Une erreur grossière (ou *Gross Error* en anglais) est commise lorsque la distance relative au sens de l'équation 6.5 entre une valeur F_0 d'hypothèse F_0 h_y^i et une valeur F_0 de référence voisée r_x^i est supérieure à un certain seuil noté GET (seuil d'erreur grossière ou *Gross Error Threshold*). Un consensus sur la valeur de ce seuil a été trouvé dans la littérature. Le seuil est généralement de 20% ce qui peut sembler élevé mais elle est parfaitement adapté aux erreurs typiques des AEP qui sont majoritairement des erreurs de multiple ou de sous-multiple (de type (p/q) pour être exact). L'équation 6.5 présente la condition pour laquelle une erreur grossière est commise.

$$\frac{|r_x^i - h_y^i|}{r_x^i} > GET, r_x^i > 0 \tag{6.5}$$

Lorsqu'une référence ne parvient pas à être correctement estimée, c'est-à-dire qu'il n'existe pas d'hypothèses F_0 à moins de 20% relatif, alors cette référence est dite « *non trouvée* ». A l'inverse, une référence est « *trouvée* » lorsqu'une hypothèse est à moins de 20% relatif.

Deux taux d'erreurs grossières sont utilisés dans les évaluations : le GER dit « global » noté GER_G et le GER dit « local » noté GER_L. Le GER_G est le rapport entre le nombre de valeurs F_0 de référence non trouvées et le nombre de valeurs F_0 de références voisées. Il a l'avantage d'être une valeur comprise entre 0 et 100%, le 0% signifiant que toutes les références sont bien estimées et le 100% indiquant qu'aucune référence n'est bien estimée (par erreur de sous-voisé ou bien par erreur d'estimation). Il a pour désavantage d'inclure les erreurs de sous-voisé. L'équation 6.6 formalise mathématiquement le

GER_G.

$$GER_G = 100\frac{\Omega[\{r_{KO}^G\}]}{\Omega[\{r_V^G\}]} \tag{6.6}$$

$\{r_{KO}^G\}$ désigne l'ensemble des valeurs F_0 de référence non trouvées considéré sur l'ensemble des trames de référence voisées et $\{r_V^G\}$ désigne l'ensemble des valeurs F_0 de référence voisés considéré sur l'ensemble des trames de référence voisées. L'exposant G peut être compris comme l'ensemble des numéros de trame de référence voisée noté $\{K_{RV}\}$ comme le présente l'équation 6.7.

$$G = \{K_{RV}\} \tag{6.7}$$

Pour mesurer exclusivement les erreurs d'estimation de F_0, il faut introduire le GER_L qui correspond au rapport entre le nombre de références F_0 mal estimées pour les trames qui sont voisées et pour la référence et pour l'hypothèse. L'avantage de ce GER est de ne mesurer que la performance d'estimation en excluant toutes les erreurs dues à la décision VnV. Le désavantage est de limiter le calcul du GER_L aux trames qui sont voisées pour la référence et pour l'hypothèse. A l'extrême si une seule trame est voisée pour l'hypothèse et la référence F_0 et que la F_0 de cette trame est mal estimée alors le GER est de 100% alors que ce résultat n'est absolument pas représentatif de la qualité réelle de l'AEP. Pour être complet dans la description du GER, une évaluation doit fournir le GER_G et le GER_L. L'équation 6.8 formalise mathématiquement le GER local.

$$GER_L = 100\frac{\Omega[\{r_{KO}^L\}]}{\Omega[\{r_V^L\}]} \tag{6.8}$$

$\{r_{KO}^L\}$ désigne l'ensemble des valeurs F_0 de référence non trouvées considéré sur l'ensemble des trames voisées en référence et en hypothèse. $\{r_V^L\}$ désigne l'ensemble des valeurs F_0 de référence voisées considéré sur l'ensemble des trames voisées en référence et en hypothèse. L'exposant L peut être compris comme l'ensemble des numéros de trames de référence voisées noté $\{K_{RV}\}$ en commun avec les numéros de trames d'hypothèse voisées noté $\{K_{HV}\}$ comme le présente l'équation 6.9.

$$L = \{K_{RV}\} \cap \{K_{HV}\} \tag{6.9}$$

Un critère de taux d'erreur fine ou *Fine Error Rate* (FER) est aussi utilisé dans nos évaluations. Il renseigne sur l'écart moyen entre les valeurs F_0 d'hypothèse et les valeurs F_0 de référence lorsque l'estimation est correcte. Autrement dit, le FER est calculé à partir des trames pour lesquelles la valeur F_0 de référence est trouvée. Les erreurs fines sont typiquement des erreurs de 1 ou 2% autour de la valeur de pitch de référence. Elles sont nettement moins gênantes que les erreurs grossières à moins que l'application ne requière une grande précision dans l'estimation de F_0.

Un taux de précision noté PRR_1 est également utilisé. Ce taux de précision quantifie l'efficacité d'un AEP à trouver les bonnes estimations de F_0 dans les premières valeurs F_0 d'hypothèse. Concrètement, si un AEP doit estimer une F_0 et qu'il lui faut fournir 3 hypothèses avant de correctement estimer la valeur, le taux de précision est de 1/3. Ce taux de précision n'est calculé que sur les trames dont la F_0 à été trouvée. Ainsi, il peut être calculé lorsque N_h vaut 1 mais il n'est pas informatif car il

vaudra nécessairement 100%. Il devient informatif lorsque l'AEP fournit plusieurs hypothèses en sortie ($N_h > 1$).

6.2.2 Situation multipitch

Le passage à un problème multipitch ajoute de nouveaux critères d'évaluation. Les deux processus de décision VnV et d'estimation de F_0 sont toujours évalués mais la simplification (situation monopitch) consistant à confondre une trame de référence avec son candidat F_0 n'est plus valable.

6.2.2.1 Décision VnV

La définition adoptée est de considérer qu'une trame est voisée si au moins un de ses candidats F_0 l'est. A partir de cette définition, les calculs des taux sous-voisé et sur-voisé peuvent être repris de la situation monopitch cf. équations 6.3 et 6.4.

6.2.2.2 Estimation de F_0

Le passage à la situation multipitch amène à distinguer deux nouveaux types de GER. Un GER sur les candidats noté GER^C et un GER sur les trames noté GER^T. Les deux nouveaux types de GER peuvent être tous deux calculés de manière locale ou globale (cf. situation monopitch). La combinatoire aboutit à considérer quatre GER différents. Le GER^C peut être calculé de manière globale ou locale si bien que deux taux d'erreur sont utilisés pour le définir complètement : le GER^C_G et GER^C_L. De la même façon, le GER^T est décomposable en deux GER qui sont le GER^T_G et GER^T_L. Évaluer complètement l'estimation de F_0 d'une situation multipitch par GER requiert absolument ces quatre GER. Le taux d'erreur grossière sur les candidats est l'équivalent du taux de GER dans la situation monopitch. Dans le cas du GER^C_G, il s'agit de compter la totalité des valeurs F_0 de références voisées non trouvées et de rapporter ce nombre au nombre de valeurs F_0 de références voisées comme l'indique l'équation 6.6. Dans le cas du GER^C_L, il s'agit de ne considérer que les trames voisées en référence et en hypothèse et de compter le nombre de valeurs F_0 de référence non trouvées en rapportant ce nombre au nombre de valeurs F_0 de référence voisées. L'équation 6.8 traduit mathématiquement ces dires. La nouveauté apportée par la situation multipitch concerne le GER sur les trames. Ce taux GER_T prend son sens dans la mesure où l'objectif poursuivi est d'être capable de reconstruire les différentes trajectoires F_0 dans un mélange de parole. Le GER_T quantifie la qualité d'un AEP à correctement estimer toutes les valeurs F_0 de référence des trames de référence. Plus ce taux est important, plus la reconstruction des trajectoires F_0 est facile. Le ou taux d'erreurs grossières global sur les trames est donné dans l'équation 6.10. Il correspond au nombre de trames de référence non trouvées rapporté au nombre de trames de référence voisées. Une trame de référence est considérée comme « trouvée » lorsque toutes ses valeurs F_0 voisées sont trouvées. Si une des valeurs ne l'est pas, la trame est donc « non trouvée ».

$$GER^T_G = 100\frac{\Omega[\{R^G_{KO}\}]}{\Omega[\{R^G_V\}]} \tag{6.10}$$

$\{R^G_{KO}\}$ désigne l'ensemble des trames de référence non trouvées et $\{R^G_V\}$ désigne l'ensemble des trames de référence voisées. Le GER^T_L est donné dans l'équation 6.11. Les trames considérées sont uniquement

les trames qui sont voisées en référence et en hypothèse.

$$GER_L^T = 100\frac{\Omega[\{R_{KO}^L\}]}{\Omega[\{R_V^L\}]} \tag{6.11}$$

Parmi l'ensemble des trames voisées en référence et en hypothèse, $\{R_{KO}^L\}$ désigne l'ensemble des trames de référence non trouvées et $\{R_V^L\}$ désigne l'ensemble des trames de référence voisées. Lorsqu'une trame de référence possède deux valeurs F_0 de référence voisées identiques (ie. ce qui correspond à un instant où les deux locuteurs ont la même F_0), les AEP ne donnent qu'une seule hypothèse F_0. Pour éviter que ce phénomène ne soit considéré comme une erreur, il est considéré qu'une même hypothèse peut servir pour toutes les références voisées. Ce point est certes discutable, mais le cas de deux sons à l'unisson reste un véritable défi pour les AEP et il est plus raisonnable de ne pas pénaliser les AEP de ne pas savoir traiter ce cas particulièrement complexe.

Le même taux de FER explicité dans la situation monopitch est calculable en situation multipitch en considérant toutes les valeurs F_0 de référence trouvées et en calculant l'écart relatif moyen des valeurs F_0 d'hypothèse correspondant.

Plusieurs taux de précision peuvent être considérés en situation multipitch. Un taux de précision pour chaque situation c'est-à-dire un taux pour des trames monopitch, un autre taux pour des trames bipitch, encore un autre taux pour des trames tripitch, etc. Il est bienvenu d'indicer ces PRR par la situation multipitch qu'il est censé évaluer. Autrement dit, le taux PRR_1 concerne les trames monopitch, le taux PRR_2 concerne les trames bipitch et ainsi de suite. Il est utile de rappeler que le taux de précision moyen permet de quantifier l'efficacité d'un algorithme à obtenir les bonnes hypothèses F_0 en premier sur les trames qui ont été trouvées. Plus le taux de précision est faible, plus rapidement sont trouvées les bonnes hypothèses.

Ce tour d'horizon des métriques utilisées dans ce chapitre a permis de spécifier comment quantifier la qualité des AEP sur leur processus de décision VnV et sur leur processus d'estimation de F_0 à proprement parlé. Ces deux processus sont interdépendants et l'objectif de la section suivante est de montrer l'influence de la décision VnV sur les résultats d'évaluation de l'estimation de F_0.

6.3 Lien entre décision VnV et estimation de F_0

Tous les AEP doivent, à un moment ou un autre de la chaine de traitement, prendre une décision VnV sur chaque trame du signal. L'objectif de cette section est de montrer que cette décision affecte les résultats de l'évaluation de l'estimation de F_0 et que par conséquent, les deux processus sont indissociables dans une évaluation qui se veut équitable. La décision VnV est généralement prise à partir d'une courbe représentant la force de périodicité de chaque trame du signal. Le paragraphe 6.3.1 présente les méthodes existantes traitant de quantification de la force de périodicité et explique la manière dont cette quantification est réalisée dans PSH_{mul}. La décision VnV est prise à partir de la force de voisement extraite. Le paragraphe 6.3.2 présente quelques approches existantes de prise de décision VnV et celle de PSH_{mul}. Le paragraphe 6.3.3 présente l'impact de la décision VnV sur les résultats de l'évaluation d'un AEP. Cette influence est à l'origine d'une nouvelle méthodologie d'évaluation intégrant la décision VnV dans le processus d'évaluation proposée dans le paragraphe 6.4.4. La méthodologie sera

utilisée dans l'évaluation monopitch (cf. paragraphe 6.6.4) et bipitch (cf. paragraphe 6.7.3).

6.3.1 Quantification de la force de périodicité

Dans la littérature, la quantification de la force de périodicité (ou quantification de voisement) d'un segment de signal est un problème étudié mais qui reste néanmoins une tâche difficile. Du point de vue psycho-acoustique, les travaux de Fastl et Stoll [Fastl and Stoll, 1979] s'attachent à quantifier la force de périodicité (ou quantification de voisement) pour des signaux divers : des sons purs, des sons purs en structure harmonique, des structures harmoniques filtrées passe-bande, des sons purs AM et d'autres bruits divers, etc. Les auteurs établissent une échelle relative des forces de périodicité en partant du principe qu'un son pur a une force de périodicité de 100%. Les auteurs montrent que les structures harmoniques possèdent une force de périodicité comprise entre 75 et 100% alors que la force de voisement de différents bruits non-périodiques tourne autour de 25%. Les travaux de Houtsma et al. [Beerends and Houtsma, 1989, Houtsma and Smurzynski, 1990] montrent que l'identification de la F_0 s'améliore lorsque l'on renforce la structure harmonique d'un son. Ceci suggère que la force de périodicité est liée à la structure harmonique, point qui est repris dans le principe même du PSH puisque l'amplitude des fonctions de pitch est une force de la détection d'une éventuelle structure harmonique.

L'extraction automatique de la force de voisement dans la littérature a été étudiée. PRAAT [Boersma, 1993] utilise l'amplitude de l'ACF (cf. chapitre 3 sous-paragraphe 3.3.1.1) divisée par l'autocorrélation d'une fenêtre de Hanning pour déterminer la force de périodicité d'une fréquence donnée. SWIPE [Camacho, 2007, Camacho and Harris, 2008] utilise le produit scalaire entre une ondelette et la racine carrée du spectre d'amplitude (cf. chapitre 3 sous-paragraphe 3.3.2.6). L'amplitude du produit scalaire lui permet de quantifier le voisement. YIN [de Cheveigné and Kawahara, 2002] utilise une fonction nommée *Squared Difference Function* ou SDF (cf. chapitre 3 sous-paragraphe 3.3.1.3). L'amplitude de cette fonction lui fournit un critère d'apériodicité et donc une manière de quantifier le voisement. La force de périodicité de PSH_{mul} est issue du produit scalaire entre le PUI et un spectre d'amplitude et reflète donc la force de la structure harmonique à la fréquence étudiée. Il s'agit d'attribuer à la trame la force de périodicité correspondant à l'amplitude du candidat F_0 d'hypothèse le plus fort. D'autres travaux existent sur ce point comme par exemple ceux de McAulay & Quatieri [McAulay and Quatieri, 1990] qui proposent de quantifier le voisement par le rapport signal sur bruit (SNR) présenté dans l'équation 6.12. Les auteurs passent par la modélisation paramétrique du signal dans laquelle il est considéré comme étant une somme de sinus variables en amplitude et en phase. Après estimation des paramètres, ils calculent le SNR comme ci-dessous. Le SNR est le rapport de l'énergie du signal original par rapport à l'énergie du signal correspondant à la différence entre le signal original et celui estimé.

$$SNR = \frac{\sum_n |s(n)|^2}{\sum_n |s(n) - \hat{s}(n, w_0)|^2} \tag{6.12}$$

$s(n)$ est le signal discret, $\hat{s}(n, w_0)$ le modèle du signal et w_0 la fréquence fondamentale estimée du signal (exactement $2\pi F_0$).

Les approches de quantification de la force de voisement sont diverses et il ne semble pas exister de consensus clair sur la manière de la calculer. Ceci est certainement le reflet d'une difficulté pour

la communauté scientifique d'attribuer un critère acoustique au voisement. Ce problème est encore un problème ouvert et n'est pas facilité par le fait que le caractère voisé dépend de la communauté scientifique qui en parle. Le voisement d'un acousticien et celui d'un phonéticien sont différents. La qualité d'une structure harmonique semble tout de même être un bon indice acoustique pour quantifier la force de périodicité et $\mathrm{PSH}_{\mathrm{mul}}$ l'utilise.

6.3.2 Décision VnV

La décision VnV est prise à partir de la quantification de voisement. La décision VnV indique si une trame donnée comporte une F_0 ou non. La décision VnV est une décision binaire. PRAAT inclut la décision VnV dans une programmation dynamique visant à lisser la trajectoire de F_0. Camacho dans SWIPE utilise un algorithme de classification à deux états présenté dans [Camacho, 2008]. YIN propose un simple seuil paramétrable. Si la force d'apériodicité est inférieure au seuil alors la trame est voisée. Ceci offre l'avantage de pouvoir régler le seuil en fonction des besoins de l'application. La décision VnV de $\mathrm{PSH}_{\mathrm{mul}}$ est prise par un seuil calculé à partir de l'amplitude du candidat F_0 d'hypothèse le plus fort. Ce processus de décision volontairement simple est repris de YIN et SWIPE. Si la force de voisement dépasse ce seuil, alors la trame est voisée. Le réglage du seuil est critique et dépend des besoins de l'expérimentateur. Une application qui requiert des valeurs sûres doit fixer le seuil à une valeur importante. Au contraire, lorsqu'une application requiert toutes les hypothèses F_0, même les plus incertaines, le seuil doit être fixé à une valeur faible. D'autres travaux sur cette décision VnV existent. Dans l'article [Thomson and Chengalvarayan, 1998] les auteurs utilisent l'autocorrélation et le jitter pour décider du caractère voisé d'une trame. L'article d'Atal et Rabiner [Atal and Rabiner, 1976] propose une méthode de classification des trames en trois catégories : voisé/non-voisé/silence. Pour ce faire, les auteurs utilisent cinq paramètres : le zéro-crossing, l'énergie, l'autocorrélation, le premier coefficient du filtre LPC et l'erreur de prédiction de la LPC. Une étape d'entrainement préalable permet de construire les densités de probabilités des cinq paramètres pour les trois catégories. Ils élaborent ensuite un processus de décision qui combine ces paramètres. Rabiner & Sambur [Rabiner and Sambur, 1977] proposent de classifier les trames dans les trois mêmes catégories en calculant une distance entre la LPC de la trame et le LPC de gabarits de trame voisée, non-voisée ou silencieuse appris au préalable. Les erreurs de classification faites sont de l'ordre de 5%. La classification voisé/non-voisé/silence ne satisfait pas Siegel & Bessey [Siegel and Bessey, 1982] qui ajoutent la catégorie « mélange ». Ce raffinement prend en compte les zones où le caractère voisé d'une trame est litigieux. C'est le cas pour certains types de phonèmes comme les fricatives par exemple. Le processus de décision se base sur quatorze paramètres acoustiques. Là aussi, une étape d'entrainement est requise. Toutes ces approches illustrent bien la difficulté d'obtenir une décision VnV fiable et dans les systèmes les plus élaborés, une étape d'apprentissage est utilisée ce qui est synonyme d'adaptation aux données pour minimiser les erreurs. De manière générale, la décision VnV dépend souvent de l'estimation de la F_0 et inversement. Le processus de décision VnV est circulaire et la séparation des deux processus n'est pas clairement établie. C'est pourquoi il est difficile de mettre en œuvre une décision VnV.

TABLE 6.1: Ensemble des couples (Référence, Hypothèse) de décision VnV.

Référence	Hypothèse	Erreur
nV	nV	-
nV	V	sur-V
V	nV	sous-V
V	V	-

6.3.3 Influence de la décision VnV sur l'évaluation

Le tableau 6.1 présente la combinatoire complète des états de décision VnV des trames de référence et d'hypothèse. Le tableau 6.1 est un exemple typique de problème de décision. « nV » signifie que la trame n'est pas voisée et « V » signifie que la trame est voisée. « - » signifie qu'il n'y a pas d'erreur. « sur-v » désigne une erreur de sur-voisé et « sous-v » une erreur de sous-voisé. L'évaluation de l'estimation de la F_0 est effectuée sur les trames considérées comme voisées pour la référence ou bien sur les trames qui sont voisées en référence et en hypothèse. Ceci implique que l'évaluation est dépendante de la décision VnV et que si cette décision n'est pas fiable, l'évaluation sera affectée. Les erreurs de décision VnV se situent principalement aux extrémités des segments de parole voisés là où la force de voisement devient plus faible car l'estimation de F_0 y est moins fiable. Les AEP sont donc susceptibles de faire plus d'erreurs (et d'être en désaccord) sur les extrémités de segments voisés. Un AEP qui prend en compte ces zones au voisement litigieux s'expose à faire plus d'erreur d'estimation de F_0 que le même AEP qui ne les prendrait pas en compte. Cette constatation montre que le seuil à partir duquel on considère qu'une trame est voisée influe sur l'évaluation de l'estimation de F_0. Lorsque l'évaluation ne porte que sur des segments temporels fortement voisés, les AEP font peu d'erreurs. Plus l'évaluation intègre des trames litigieuses, plus l'estimation est mauvaise. La décision VnV est donc un levier à partir duquel il est possible d'améliorer artificiellement les résultats d'évaluation d'un AEP. La figure 6.1 atteste de ce phénomène en présentant le taux d'erreur grossière GER_G (cf. section 6.2) en fonction du seuil de décision VnV pour quatre AEP différents sur le corpus de Bagshaw monopitch au complet. Le graphique (a) de la figure 6.1 présente le comportement du GER en fonction du seuil de décision VnV de YIN dans la situation monopitch. Le graphique (b) présente le comportement de SWIPE, le graphique (c) montre le comportement de PSH_{min} et le graphique (d) présente le comportement de PSH_{mul}. Les allures des courbes de la figure 6.1 sont identiques : le GER_G diminue et atteint même 0 si le seuil de voisement est suffisamment élevé. Ce phénomène implique qu'une évaluation d'AEP ne peut être réduite à l'observation du GER_G. Elle doit absolument indiquer les taux d'OVR et d'UVR. Pour être équitable, il faut placer les AEP sur un même point de fonctionnement en termes de décision VnV. Comment introduire ce point de fonctionnement ? Une possibilité passe par l'introduction du rapport $\mu = (OVR/UVR)$ et d'utiliser ce rapport comme critère à uniformiser pour tous les AEP comparés. La figure ci-dessous permet de justifier cette possibilité. La figure 6.2 présente les comportements des OVR et UVR en fonction du seuil de décision VnV. Le graphique (a) de la figure 6.2 présente l'OVR (bleu) et l'UVR (rouge) obtenue par de YIN dans la situation monopitch. Le graphique (b) présente ceux de SWIPE, le graphique (c) ceux de PSH_{min} et le graphique (d) ceux de PSH_{mul}. Les courbes sont des moyennes obtenues sur l'intégralité du corpus de Bagshaw monopitch. Il apparaît que les évolutions

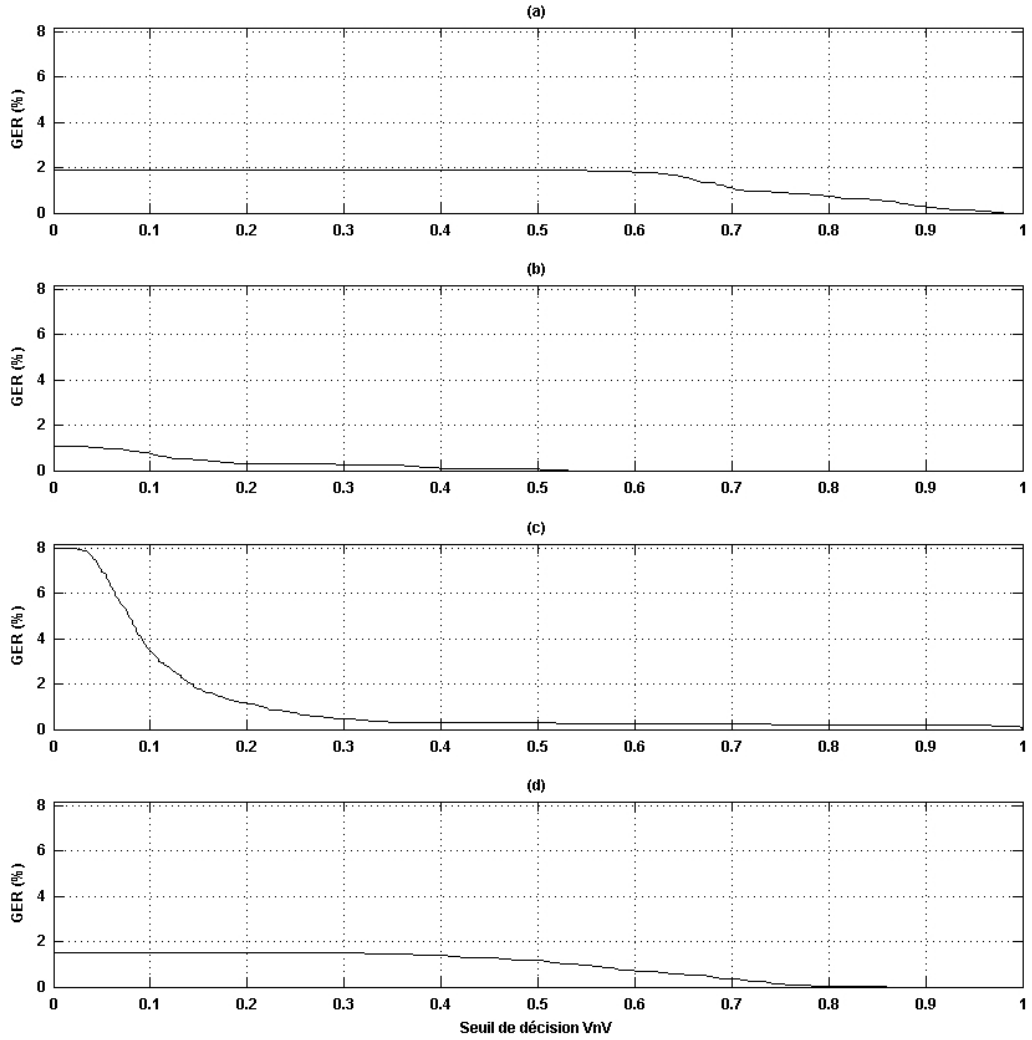

FIGURE 6.1: Évolution du GER_G en fonction de la décision VnV. (a) YIN. (b) SWIPE. (c) PSH_{min}. (d) PSH_{mul}.

des OVR et UVR sont très différentes selon l'AEP observé. De même, pour un seuil de décision VnV donné, ces taux sont très différents selon les AEP. L'évolution du GER_G (cf. figure 6.1) suit la même évolution que le taux de sur-voisé OVR. De plus, l'évolution du GER_G suit l'évolution inverse du taux de sous-voisé UVR. Avec ces observations, le point de fonctionnement de la décision VnV peut être fixé par le rapport $\mu = (OVR/UVR)$. La normalisation de la décision VnV consiste à fixer μ à une valeur constante. La valeur de μ dépend entièrement de ce que l'on souhaite mesurer. Si on souhaite tester les AEP dans des conditions où seules les trames très voisées sont considérées alors on peut fixer μ à une valeur supérieure à 1. Si on souhaite observer le comportement des AEP dans des conditions plus litigieuses, alors on peut envisager de fixer μ à une valeur inférieure à 1. Il est aussi possible de se placer à l'Equal Error Rate ou EER c'est-à-dire à l'égalité entre OVR et UVR qui équivaut à un équilibre entre le taux de sur-voisé et le taux de sous-voisé. La figure6.3 présente le rapport μ en fonction du seuil de décision VnV. Notre choix de μ est cohérent car il suit effectivement la même évolution que le GERG. On pourrait tout autant utiliser comme point de fonctionnement un taux constant d'UVR ou d'OVR. Ce choix appartient à l'expérimentateur et dépend de ce qu'il souhaite montrer. L'important

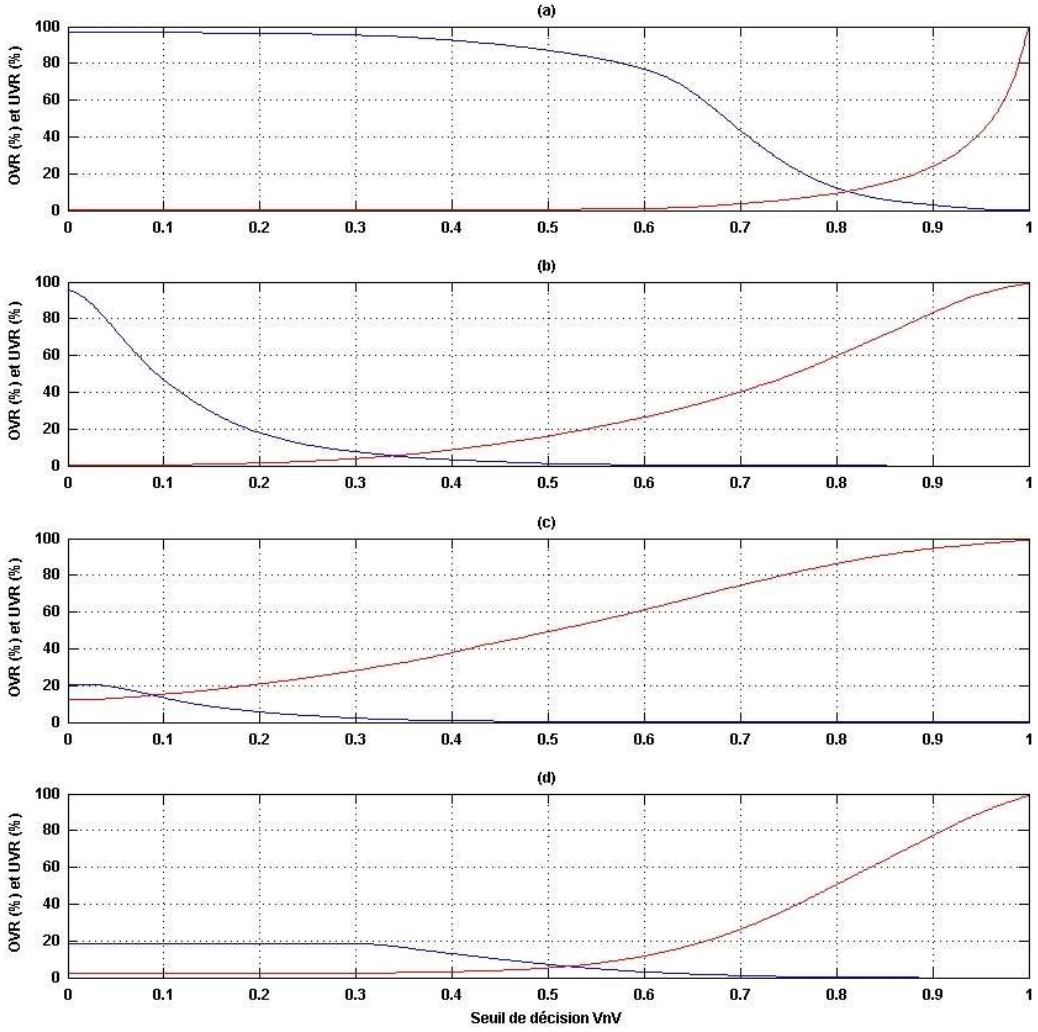

FIGURE 6.2: Évolution des OVR et UVR en fonction de la décision VnV. Les courbes bleues sont les courbes d'OVR. Les courbes rouges sont les courbes d'UVR. (a) YIN. (b) SWIPE. (c) PSH_{min}. (d) PSH_{mul}.

est d'intégrer la décision VnV dans l'évaluation et de la faire clairement apparaître dans les résultats. Un point de fonctionnement à l'EER donnerait un seuil de décision VnV à 0.81 pour YIN, 0.34 pour SWIPE, 0.09 pour PSH_{min} et 0.52 pour PSH_{mul}. Bien sur, ces seuils sont des moyennes obtenues sur l'ensemble de Bagshaw monopitch. Il est tout à fait possible que pour un fichier donné, le seuil soit différent. C'est pourquoi le seuil de décision doit être réglée pour chaque signal analysé.

Cette section a permis de rendre compte de l'importance de la décision VnV sur les résultats d'une estimation de F_0. L'évaluation ne peut se résumer à de simples GER mais doit aussi clairement indiquer le point de fonctionnement de l'algorithme et les taux d'OVR et d'UVR. La décision VnV peut être introduite dans l'évaluation sous la forme d'un point de fonctionnement VnV dépendant des OVR et UVR (rapport μ par exemple). Normaliser ce point de fonctionnement permet de rendre l'évaluation plus équitable. La section 6.4 suivante présente quatre méthodologies d'évaluation.

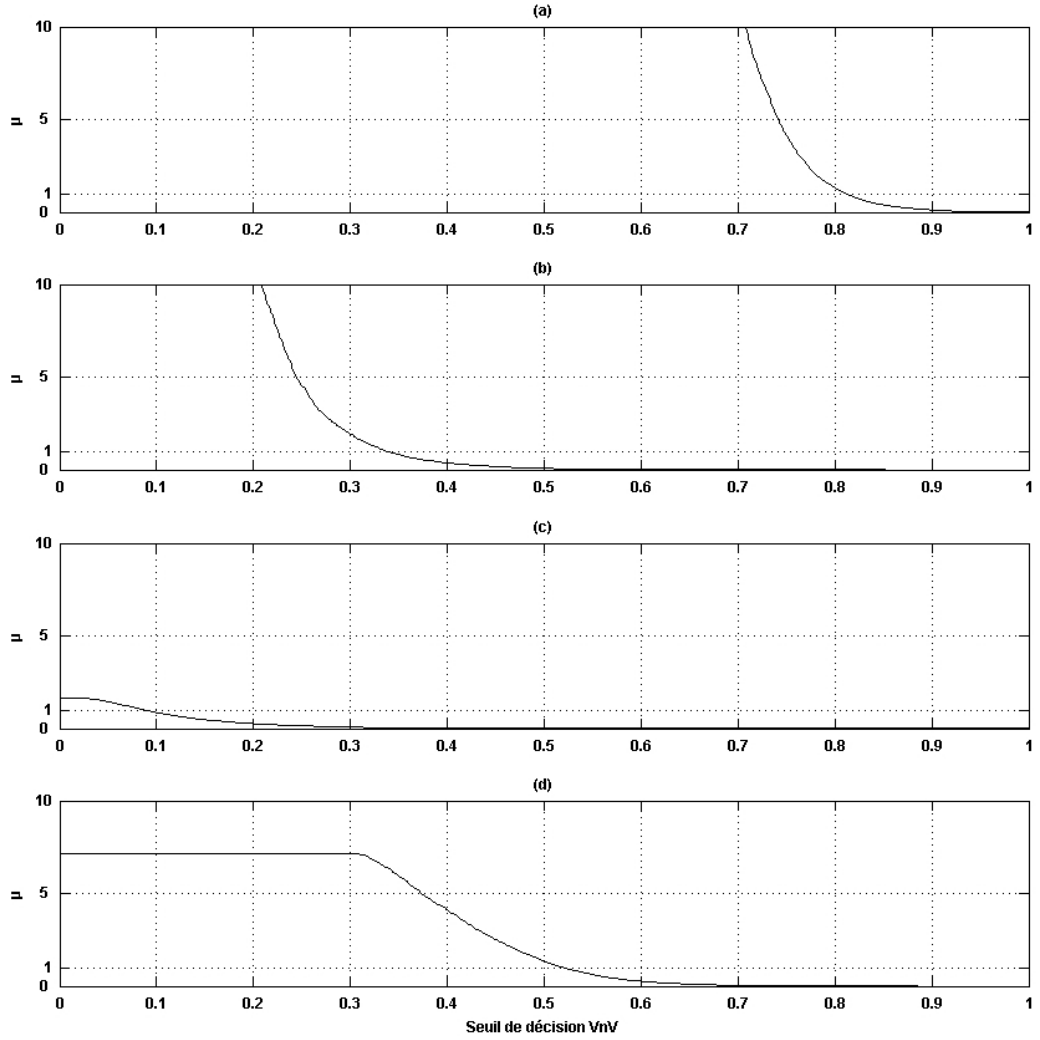

FIGURE 6.3: Évolution de μ en fonction de la décision VnV. (a) YIN. (b) SWIPE. (c) PSH_{min}. (d) PSH_{mul}.

6.4 Méthodologies d'évaluation

Quatre méthodologies sont décrites ici. La première méthodologie proposée est la plus équitable de toutes mais ne peut pas toujours être utilisée selon l'hétérogénéité de fonctionnement des AEP comparés. Les trois autres méthodologies posent moins de contraintes concernant le fonctionnement des AEP mais ont des inconvénients qu'il est nécessaire de prendre en compte dans l'interprétation des résultats d'évaluation.

6.4.1 Méthodologie par 0-UVR (0-UVR)

La méthodologie la plus équitable vis-à-vis de tous les AEP comparés consiste à ajuster les paramètres de chaque algorithme pour fournir une estimation de F_0 à toutes les trames. Ceci à pour effet de rendre le taux d'UVR nul. Il suffit alors de comparer l'estimation de F_0 obtenue à la référence. L'évaluation portera sur les trames voisées de la référence. Cette approche est la plus juste, car elle teste tous les algorithmes sur les trames pour lesquelles ils sont sensés fournir une estimation. Cette

méthodologie est nommée « méthodologie par 0-UVR » et est notée « méthodologie 0-UVR ». Elle comporte les quatre étapes suivantes :

1. Normalisation des signaux monopitch décrit dans le sous-paragraphe 6.5.1.2.
2. Génération des références F_0 décrit dans le paragraphe 6.5.2.
3. Les AEP sont réglés de manière à avoir un UVR de 0%, soit car intrinsèquement l'AEP fournit des estimations pour toutes les trames, soit par réglage du seuil de décision VnV.
4. Extraction des critères de qualité en fonction de la situation monopitch ou multipitch.

L'inconvénient de cette méthodologie est qu'elle n'est applicable que dans le cas où les AEP offrent la possibilité de régler le seuil de décision VnV ou bien lorsqu'ils donnent une estimation de F_0 pour toutes les trames quelles qu'elles soient. Or, un certain nombre d'AEP ne donnent pas ces possibilités comme PRAAT pour les AEP monopitch et WWB et VEW pour les AEP bipitch. Étant testés dans nos évaluations, il est nécessaire de mettre en œuvre des méthodologies d'évaluation alternatives. Trois nouvelles méthodologies sont proposées : une dites « standard » notée méthodologie ST, une dites par « normalisation des trames » notée méthodologie NT ou encore une dites par « normalisation de la décision VnV » notée méthodologie VnV. Ces trois méthodologies sont moins idéales que celle de « 0-UVR » mais permettent de faire une évaluation comparative entre des AEP qui proposent des réglages hétérogènes.

6.4.2 Méthodologie standard (ST)

La méthodologie ST consiste à tester chacun des AEP dans leur version standard. Cette façon d'évaluer compare simplement les hypothèses F_0 aux références F_0. Les étapes de la méthodologie ST sont données dans la liste suivante :

1. Normalisation des signaux monopitch décrit dans le sous-paragraphe 6.5.1.2.
2. Génération des références F_0 décrit dans le paragraphe 6.5.2.
3. Génération des hypothèses F_0 avec les versions standards des AEP.
4. Extraction des critères de qualité en fonction de la situation monopitch ou multipitch.

L'avantage de cette méthode est sa simplicité que n'importe quel AEP, quelque soit son fonctionnement, peut être testé et comparé aux autres. L'inconvénient de la méthodologie ST est qu'elle ne garantie pas que les AEP soient testés sur les mêmes trames. Les résultats obtenus sont donc relativement indépendants les uns des autres car ils sont obtenus sur des points de fonctionnement en termes de décision VnV différents. Dans un cas extrême, il se peut qu'un AEP obtiennent les mêmes performances qu'un autre AEP alors que les erreurs des AEP ne portent pas sur les mêmes trames. L'évaluation ST n'a de sens que si les taux d'OVR et d'UVR sont renseignés.

6.4.3 Méthodologie par normalisation des trames (NT)

La méthodologie NT consiste à évaluer l'estimation de F_0 sur les trames qui sont considérées comme voisées par tous les AEP testés. La méthodologie ressemble à celle de Camacho [Camacho, 2007]. Les étapes sont données dans la liste suivante :

1. Normalisation des signaux monopitch décrit dans le sous-paragraphe 6.5.1.2.
2. Génération des références F_0 décrit dans le paragraphe 6.5.2.
3. Génération des hypothèses F_0 avec les versions standards des AEP.
4. Seules les trames voisées pour **tous** les AEP sont conservées pour l'évaluation. Ainsi tous les AEP estiment la F_0 sur les mêmes trames.
5. Extraction des critères de qualité en fonction de la situation monopitch ou multipitch.

Les avantages de cette méthodologie est de fixer le point de fonctionnement en termes de décision VnV identiquement pour tous les AEP (μ constant). De plus, il assure que l'évaluation porte sur les mêmes trames. Toutefois, l'évaluation NT ne permet pas de donner une idée équitable du comportement des algorithmes dans les zones où le voisement est litigieux. De plus, l'expérimentateur n'a aucun contrôle sur le point de fonctionnement final des AEP. Un des inconvénients de cette méthodologie est que le résultat dépend des AEP testés, il suffit d'ajouter un AEP à l'évaluation pour que le résultat ne porte plus sur les mêmes trames. L'autre inconvénient de cette méthodologie provient du fait que la majorité des désaccords entre AEP sont produits aux extrémités des segments voisés, là où le caractère voisé de la trame est litigieux. Ainsi, cette méthodologie ne compare les AEP que pour les trames dont le voisement est indéniable et permet donc de mettre en évidence les qualités intrinsèques des AEP uniquement dans des conditions de voisement clairement établies.

6.4.4 Méthodologie par normalisation de la décision VnV (VnV)

Dans la méthodologie NT précédente, l'expérimentateur n'a aucun contrôle sur la décision VnV des AEP puiqsue cette décision est collective et dépend des AEP comparés. Ce besoin de contrôle de la décision VnV amène cette méthodologie originale proposée dans ce paragraphe. Elle intègre la décision VnV dans le processus d'évaluation en fixant μ (cf. paragraphe 6.3.3), le point de fonctionnement en termes de décision VnV, à une valeur constante pour tous les AEP. La méthodologie comporte les six points suivants :

1. Normalisation des signaux monopitch décrit dans le sous-paragraphe 6.5.1.2.
2. Génération des références F_0 décrit dans le paragraphe 6.5.2.
3. Calcul des courbes d'OVR et UVR en fonction du seuil de voisement pour chacun des signaux.
4. Pour chaque signal, le seuil de décision VnV de chaque AEP est fixé de manière à avoir un μ constant. μ est choisi par l'expérimentateur selon ce qu'il souhaite montrer.
5. Estimation des hypothèses F_0 avec le seuil de décision VnV calculé.
6. Extraction des critères de qualité en fonction de la situation monopitch ou multipitch.

L'avantage de cette méthode est d'offrir à l'expérimentateur le soin de régler le point de fonctionnement de la décision VnV à la valeur désirée. Cette manière de procéder est plus juste qu'une simple évaluation ST. En revanche, cette méthodologie ne garantie plus de tester les mêmes trames comme une méthodologie NT.

Les métriques d'évaluation, le lien entre décision VnV et estimation de F_0 et les méthodologies utilisées dans nos évaluations étant à présent énoncés, il convient de présenter dans la section 6.5

suivante les corpus utilisés et la manière dont ils ont été traités et combinés pour former nos signaux bipitch. Il décrit également la manière dont ont été élaboré les candidats F_0 de référence.

6.5 Choix des corpus et annotations de référence

Le choix des corpus de parole et la construction des références F_0 qui vont servir de vérité terrain (*groundtruth* en anglais) est une étape importante d'une méthodologie d'évaluation. Le paragraphe 6.5.1 présente les corpus utilisés dans nos évaluations. Le paragraphe 6.5.2 détaillera le processus de calcul des références F_0.

6.5.1 Choix des corpus

Les corpus utilisés sont des corpus de voix lue et relativement neutre. Ils sont décrits dans le sous-paragraphe 6.5.1.1. Les signaux n'ont pas nécessairement la même intensité sonore. Un travail préalable consiste à normaliser l'intensité de tous les signaux des corpus. Cette étape est d'autant plus importante que les signaux vont être mélangés et que nous souhaitons les mélanger à intensités égales. La méthode de normalisation d'intensité utilisée est présentée dans le sous-paragraphe 6.5.1.3.

6.5.1.1 Corpus monopitch

Trois corpus de signaux mono-locuteur composent notre base de données. Il sont tous enregistrés dans des conditions de laboratoire. Le corpus de Keele [2] comprend 10 locuteurs anglais (5 hommes, 5 femmes) qui lisent tous le même texte (« *The north wind and the sun...* » sur un ton assez neutre. Keele comporte 337 secondes de parole. Pour plus de détail sur ce corpus, il faut se référer à [Plante et al., 1995]. Le corpus de Bagshaw [3] comprend 2 locuteurs anglais (1 homme, 1 femme). Les 2 locuteurs prononcent les mêmes 50 phrases courtes pour un total de 332 secondes de parole. Le corpus de Mocha [4] comprend 460 phrases courtes prononcées par deux locuteurs, une femme et un homme. Le corpus contient 3665 secondes de parole. Au total environ 1h10minutes de parole est collectée dans la base de données monopitch. Le choix de ces corpus a été guidé par le fait que tous contiennent les signaux EGG associés. Les auteurs des corpus Keele et Bagshaw fournissent également une annotation de référence en F_0 corrigée manuellement.

6.5.1.2 Normalisation d'intensité

Le procédé de normalisation des signaux mono-locuteur consiste à donner la même intensité globale à tous les fichiers. L'intensité globale d'un signal $s(n)$ est donnée par l'équation 6.13. Ce calcul est identique à la méthode « *Get Intensity (dB)...* » du logiciel PRAAT.

$$I = 10 \log_{10} \left(\frac{1}{N P_0^2} \sum_{n=0}^{N-1} s^2(n) \right) \tag{6.13}$$

2. `http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html`
3. `http://www.cstr.ed.ac.uk/research/projects/fda`
4. `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html`

Or, si deux signaux sont très différents en termes de zones silencieuses, normaliser au sens de l'équation 6.13 revient à amplifier beaucoup les signaux « creux » (beaucoup de silences). Ce phénomène est d'autant plus important que des mélanges de parole vont être fabriqués. Les amplitudes des signaux mélangés doivent être comparables pour éviter qu'un des signaux domine l'autre en permanence. L'intensité d'un signal est donc calculée en ne considérant que les segments non-silencieux. Un étape de détection de silence est donc nécessaire. Pour ce faire, l'intensité de chacun des échantillons du signal est extraite en respectant l'équation 6.14. s est le signal étudié, $I(k)$ est l'intensité du signal à l'échantillon k. La valeur de N est fixée de manière à avoir une fenêtre glissante de 80ms. P_0 vaut 2e-5 et est un seuil d'audibilité à 1kHz.

$$I(k) = 10 \log_{10} \left(\frac{1}{(2N+1)P_0^2} \sum_{n=k-N}^{k+N} s^2(n) \right) \tag{6.14}$$

Cette façon de procéder ressemble au calcul de l'intensité de la méthode « *Sound_ To_ Intensity...* » du logiciel PRAAT. Tous les échantillons dont l'intensité est inférieure à 50dB sont considérés comme du silence. L'intensité globale du signal (cf. équation 6.13) est de nouveau calculée en excluant tous les échantillons silencieux. Pour normaliser l'intensité d'un signal de parole, il est multiplié par un coefficient nommé a. La valeur de a est fonction de l'intensité de normalisation I_N et de l'intensité originale I du signal est donnée par l'équation 6.15.

$$\alpha = NP_0^2 10^{\frac{I_N - I}{10}} \tag{6.15}$$

Chacun des signaux a donc été normalisé en intensité à 65dB en ne considérant que les zones non-silencieuses. Dans un mélange de plusieurs signaux, cette façon de normaliser ne protège pas de différences d'amplitudes localisées mais, sur toute la durée du signal, ces segments devraient se compenser.

6.5.1.3 Corpus bipitch

Le corpus bipitch a été élaboré à partir de paires de sons normalisés comme décrit dans le sous-paragraphe 6.5.1.2. Deux sons normalisés sont simplement additionnés et le signal de mélange est tronqué au plus court de ses constituants. Deux sons de deux corpus différents ne sont pas mélangés. Pour chaque corpus Keele, Bagshaw et Mocha, trois familles de mélanges sont effectués : les mélanges voix de femme/voix de femme, voix de femme/voix d'homme et voix d'homme/voix d'homme. Au total environ 30 minutes de mélange par corpus ont été tirés aléatoirement. Il y donc environ 1h30 de signaux bipitch à évaluer.

6.5.2 Annotation des références F_0

Un soin particulier doit être apporté à cette étape puisque l'ensemble des mesures de qualité vont dépendre de la vérité terrain extraite. L'annotation des références F_0 peut être faite manuellement ou automatiquement. L'annotation manuelle est une tâche fastidieuse dès que la taille des corpus dépasse quelques minutes. Elle est précise et a priori aucune erreur n'y est faite. L'annotation automatique permet de traiter facilement et rapidement une grande quantité de données. Elle utilise un AEP pour

calculer les références F_0 à partir du signal d'électroglottogramme (EGG). De ce fait, l'annotation des références contiendra probablement des erreurs (octave, sous octave, etc.) qui sont quasiment inévitables avec des algorithmes automatiques simples. Il faut aussi veiller à utiliser un algorithme de calcul des références F_0 qui soit différent de tous les algorithmes que l'on souhaite évaluer. Sinon l'évaluation peut être biaisée. Nous avons opté pour une annotation automatique car la masse de données à traiter est supérieure à une heure de parole. Nous sommes donc confrontés aux erreurs d'estimation de F_0. Toutefois, ces erreurs seront les mêmes pour chacun des AEP testé. On peut donc penser raisonnablement que ces erreurs n'affectent pas l'équité de l'évaluation.

Les références F_0 sont donc construites à partir du signal EGG. Le graphique (b) de la figure 6.4 présente le signal EGG correspondant au signal de parole du graphique (a). L'algorithme d'estimation des F_0 de références choisi est un algorithme typique de la littérature à savoir le calcul du pic maximum dans l'ACF. Cet algorithme a l'avantage d'être extrêmement simple et aisément reproductible. L'avantage d'utiliser le signal EGG est qu'il est beaucoup plus aisé à traiter automatiquement. Les

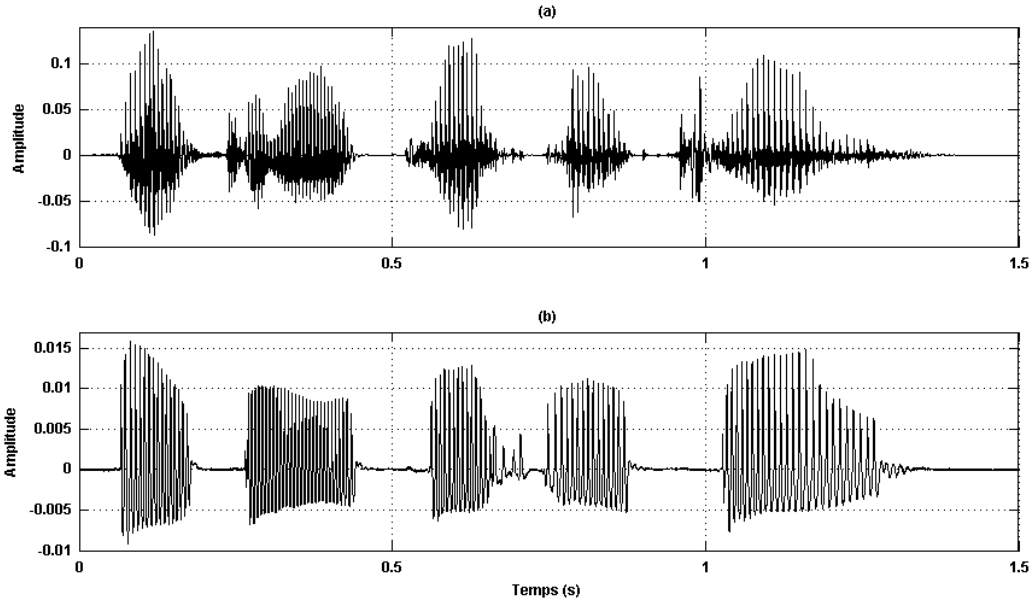

FIGURE 6.4: Exemple d'un signal EGG. (a) Signal de parole original rl001 tiré du corpus de Bagshaw. (b) Signal EGG équivalent.

périodicités sont mieux marquées que sur le signal de parole. Les références F_0 seront donc de meilleure qualité. L'inconvénient est que les corpus dotés de signaux EGG sont assez rares et de courte durée car ils sont plus compliqués à produire.

6.5.2.1 Pré-traitement

Le signal EGG entier est filtré par un filtre passe-bande non-récursif entre 50 et 1600Hz afin de supprimer la composante continue, les bruits basses fréquences et hautes fréquences qui peuvent perturber l'estimation de F_0. Le prétraitement permet d'obtenir un signal de la forme illustrée dans la figure 6.4 graphique (b).

6.5.2.2 Détection des trames silencieuses et ACF

Une trame glissante de 40ms parcourt le signal par pas de 2ms. L'intensité globale de la trame est calculée selon l'équation 6.13. Toute trame dont l'intensité (cf. équation 6.14) est inférieure à l'intensité globale du signal moins 20dB est considérée comme silencieuse. Toutes les trames non silencieuses sont centrées, pondérées par une fenêtre de Hanning et centrées de nouveau. L'ACF normalisée est alors calculée entre 50 et 400Hz. Dans les corpus considérés, il n'y a pas de F_0 au-dessus de 400Hz.

6.5.2.3 Force de voisement et décision VnV

L'amplitude du pic le plus important de l'ACF normalisée dans l'intervalle [50,400Hz] est considéré comme la force de périodicité de la trame. La fréquence correspondante est considérée comme la F_0 de la trame. Une interpolation quadratique est utilisée pour raffiner la valeur de F_0. Si la force de périodicité de la trame est supérieure à 0.5, la trame est considérée comme voisée.

6.5.2.4 Post-traitements et correction des erreurs d'octave/sous-octave

Tout segment voisé dont la durée est inférieure à 25ms est supprimé. Trois trames sont retirées de chacune des extrémités des zones voisées restantes ce qui représente 6ms à chaque extrémité. Les principales erreurs étant faites aux extrémités des segments voisés, ces deux post-traitements permettent d'éviter un bon nombre d'erreurs simplement. La correction des erreurs d'octave ou de sous-octave se fait par un lissage temporel sur 500ms. La fonction « *pchip* » de MATLAB est utilisée pour construire une trajectoire F_0 sur l'ensemble su signal en faisant passer une spline par les valeurs F_0 déjà extraites et dans les segments non-voisés. La trajectoire F_0 obtenue est lissée sur 500ms par une fenêtre de Hanning glissante. Pour chaque trame, la sous-octave et l'octave de la F_0 déjà estimée sont calculées. Chaque trame possède donc trois possibles F_0 et la valeur finale de la trame est la valeur plus proche de la trajectoire F_0 lissée. Cette opération est efficace pour supprimer les erreurs d'octave et de sous-octave courante par l'utilisation d'une ACF.

6.5.2.5 Validité de l'annotation automatique

Les corpus Keele et Bagshaw fournissent une annotation manuelle des références F_0. Dans ce paragraphe, l'annotation automatique est comparée à celles existantes afin de montrer sa validité. Les résultats d'évaluation sont obtenus en considérant que l'annotation des auteurs est la référence et que notre annotation automatique est l'hypothèse. Il convient de donner quelques détails concernant les annotations des auteurs. L'annotation en F_0 des auteurs de Keele est calculée à partir de l'EGG en utilisant la valeur maximale d'une ACF calculée sur une trame de 25.6ms. Les auteurs ont corrigés manuellement leurs annotations F_0 en observant les signaux EGG et s'autorisent à remplacer les valeurs issues de l'ACF par des -1 ou des valeurs négatives. Les candidats F_0 de référence données par les auteurs de Bagshaw sont données sous le format temporel c'est-à-dire qu'un candidat F_0 est associé à un instant temporel. Les zones voisées sont ainsi décrites par leur instant de début et de fin.

Le tableau 6.2 présente la cohérence de l'annotation donnée par Keele et celle obtenue par notre algorithme automatique. Seul le GER local GER_L est fourni car ce qui intéresse ici est de mesurer

TABLE 6.2: Cohérence entre l'annotation de Keele et la notre.

	OVR (%)	**UVR (%)**	μ	**GER_L (%)**	**FER (%)**
Femme	10.3	8.12	1.27	0.01	2.27
Homme	7.09	17.79	0.40	0.00	1.93
Moyenne	8.70	12.95	0.67	0.01	2.10

TABLE 6.3: Cohérence entre l'annotation de Bagshaw et la notre.

	OVR (%)	**UVR (%)**	μ	**GER_L (%)**	**FER (%)**
Femme	7.83	1.87	4.19	0.01	1.72
Homme	4.49	6.59	0.68	0.00	0.83
Moyenne	6.16	4.23	1.46	0.01	1.28

si des erreurs grossières sont commises dans les segments qui sont voisés en référence (annotation des auteurs) et en hypothèse (notre annotation automatique). Les différences entre les deux références ne proviennent pas d'un écart important entres les valeurs F_0 puisque le GER_L est quasiment nul et que l'écart moyen relatif entre les valeurs estimées est de 2%. Les différences sont majoritairement faites sur la décision VnV avec des erreurs de l'ordre de 10%. Ce décalage s'explique pour trois raisons :

- L'effet des 3 trames (6ms) systématiquement retirées de chaque coté qui est un choix discutable et arbitraire. Il se justifie pas le fait qu'estimer la F_0 sur des trames trop litigieuses en terme de voisement n'a pas vraiment de sens et qu'on peut s'intéroger sur l'utilité d'évaluer les AEP dans ces zones.
- Les auteurs de Keele précisent que leur estimation de la F_0 est effectuée par pas de 10ms. En revanche ils ne précisent pas l'instant de début de leur estimation. Une observation fine des différences entre notre annotation et la leur a révélé un léger décalage (de maximum 10ms soit une trame). Ceci se traduit par une augmentation du FER et des OVR et UVR. En corrigeant ce décalage, il devrait être possible d'avoir une meilleure cohérence entre les deux.
- Dans la comparaison, les valeurs -1 ou négatives de l'annotation des auteurs ont été remplacées par des zéros ce qui est sans doute un peu « brutal » et qui doit ajouter des erreurs de décision VnV.

Le tableau 6.3 présente la cohérence de l'annotation donnée par Bagshaw et celle obtenue par notre algorithme automatique. Pour ce corpus, il n'y a aucun décalage constaté. La cohérence est meilleure qu'avec Keele. Le GER_L est quasiment nul et l'écart moyen relatif entre les valeurs estimées est de 1.3%. Là encore, c'est dans la décision VnV que les différences apparaissent. Ici la différence est autour de 5% (6.16% et 4.23% respectivement pour l'OVR et l'UVR). Les désaccords de décision VnV s'expliquent pour les raisons suivantes :

- Effet des 3 trames explicité auparavant.
- L'annotation des auteurs est continue (instants de début et de fin de segments voisés) alors que notre annotation F_0 automatique est discrète et est effectué toutes les 2ms. Il est très probable que cette discrétisation ne coïncide pas toujours parfaitement avec le début et la fin des marqueurs temporels des auteurs. Ceci a pour effet de produire des erreurs de décision VnV.

Sur les deux corpus, la moyenne d'OVR est de 7.43%, la moyenne d'UVR est de 8.59% et le rapport

moyen μ (OVR/UVR) est de 1.94% ce qui signifie que notre manière de construire les références favorise le sur-voisement et donne plus de trames voisées que les annotations faites par les auteurs. Le GER_L moyen est presque nul et l'écart relatif moyen FER est de 1.69% ce qui révèle que lorsque la décision VnV est partagée par les deux références, les mêmes estimations F_0 sont données à 1.69% prêt.

Il est nécessaire de s'interroger sur la cohérence des résultats obtenus. Les taux d'OVR et d'UVR sont-ils trop élevés, trop faible ou relativement corrects ? La valeur moyenne des écarts dans la décision VnV entre les annotations données par les auteurs et la méthode automatique est de 7.5% pour l'OVR et de 8.6% pour l'UVR. Quand on analyse les résultats de PRAAT qui utilise un post-traitement puissant censé affiner la prise de décision VnV, les taux d'OVR et d'UVR sont respectivement de 6.82% et 9.47% sur l'ensemble des 3 corpus en monopitch. Il n'est donc pas spécialement choquant d'avoir des désaccords de décision VnV de l'ordre de 8%. Ces résultats permettent de valider le bon fonctionnement de notre algorithme automatique d'annotation des références F_0. Un autre avantage d'utiliser une procédure automatique est de s'appliquer à toute base de données, quelque soit sa taille. Elle fait certes des erreurs mais l'annotation manuelle en fait aussi. Les erreurs faites par une procédure automatique ont l'avantage d'être reproductibles. La section 6.6 s'attaque à présent à l'évaluation d'AEP en situation monopitch.

6.6 Évaluations monopitch

L'objectif des évaluations monopitch n'est pas de tester un grand nombre d'AEP différents mais de valider les méthodologies proposées avec quelques AEP bien connus en intégrant les performance de PSH_{mul} en situation monopitch.

Les AEP testés ici sont YIN, PRAAT, SWIPE et PSH_{mul}. Une hypothèse F_0 est estimée toute les 10ms pour tous les AEP. L'instant de début des estimations est fixé à 20ms. La zone de recherche est fixée à [50,800Hz]. Dans la mesure du possible, la taille des trames est fixée à 40ms. Le réglage standard de PRAAT est le suivant : « *Sound : To Pitch (ac)... 0.01 50 15 off 0.03 0.45 0.01 0.35 0.14 800* ». Avec ce réglage, PRAAT utilise des trames de 60ms mais ce paramètre dépendant de la fréquence minimale de recherche, il ne peut être réglé à 40ms. Le réglage standard de SWIPE est le suivant : « *swipep(s,fs,[fmin fmax],ds,[],1/20,0.5,0.2* ». SWIPE a pour particularité d'utiliser plusieurs tailles de fenêtres différentes pour chaque instant d'estimation considéré. Pour YIN, seules les options *minf0*, *maxf0*, *hop* et *wsize* sont ajustées à respectivement 50Hz, 800Hz, 10ms et 40ms. PSH_{mul} utilise des trames de 40ms par pas de 10ms avec la même zone de recherche de la F_0.

Le seuil de décision VnV de chacun des algorithmes est fixé entre 0 et 1 par normalisation. PRAAT n'offre pas la possibilité de fixer facilement ce seuil car la décision est intégrée dans une programmation dynamique. YIN fournit une force d'apériodicité *ap0* dont la valeur minimale est 0. *ap0* est normalisé de façon à ce qu'il soit compris entre 0 et 1. Ensuite, la force de périodicité utilisée pour YIN est égale à 1-*ap0*. SWIPE fournit une force de périodicité dont les valeurs négatives sont mises à 0. Ensuite la force de périodicité est normalisée pour que son maximum vaille 1. La même normalisation est faite sur PSH_{mul}. Le seuil de décision VnV de tous les AEP est donc une valeur comprise entre 0 et 1 comme le présentent les figures 6.1, 6.2 et 6.3 précédentes. Le seuil VnV de SWIPE est fixé à 0.2 comme

TABLE 6.4: Résultats récapitulatifs de l'évaluation 0-UVR monopitch sur YIN et SWIPE.

AEP	OVR (%)	UVR (%)	$\mathbf{GER_G}$ (%)	$\mathbf{GER_L}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)
YIN	100.00	0.00	3.96	3.96	1.59	100.00
SWIPE	100.00	0.00	2.64	2.64	1.53	100.00

TABLE 6.5: Résultats récapitulatifs de l'évaluation ST monopitch.

AEP	OVR (%)	UVR (%)	μ	$\mathbf{GER_G}$ (%)	$\mathbf{GER_L}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)
YIN	16.39	7.42	2.21	8.60	1.39	1.45	100.00
PRAAT	6.82	9.47	0.63	10.70	1.44	1.21	100.00
SWIPE	20.16	2.80	7.21	3.47	0.81	1.49	100.00
PSH_{mul}	9.81	6.48	1.52	7.37	1.03	1.08	100.00

préconisé par l'auteur, le seuil VnV de YIN est fixé à 0.8 car les auteurs préconisent une apériodicité de 0.2 ($1 - 0.2 = 0.8$), le seuil de PSH_{mul} est fixé à 0.5 et le seuil de PSH_{min} est fixé à 0.2.

Les quatre méthodologies présentées auparavant sont illustrées dans les paragraphes 6.6.1, 6.6.2, 6.6.3 et 6.6.4.

6.6.1 Résultats d'évaluation 0-UVR

Ce paragraphe n'est pas critique pour le manuscrit mais il est important de montrer la validité et la faisabilité de la méthodologie que nous considérons comme « la plus juste de toutes ». Étant donné que les seuls AEP monopitch offrant intrinsèquement une estimation de F_0 pour toutes les trames sans décision VnV sont YIN et SWIPE, les résultats d'évaluation ne portent que sur ces deux algorithmes, à titre d'exemple. Cette méthodologie a été utilisée sur beaucoup plus d'AEP dans d'autres travaux comme par exemple [de Cheveigné and Kawahara, 2001]. Le tableau 6.4 récapitule les moyennes de chaque AEP sur toute la base de données monopitch. Dans le cas d'une évaluation 0-UVR, il n'y a plus de distinction entre taux de GER global et taux de GER local car l'UVR est tombé à 0%. L'UVR à 0% implique également qu'il n'est pas possible de donner le point de fonctionnement μ. Cette comparaison montre que sur les corpus testés (1h10 de parole monopitch) et en comparaison avec notre annotation automatique, SWIPE réalise 1% de GER de moins que YIN. Il faudrait procéder à une analyse statistique pour mettre en évidence si cette différence est significative ou non (un test de McNemar par exemple). Du point de vue de la précision, les algorithmes sont identiques et le taux de précision est de 100% car les algorithmes ne fournissent qu'un seul candidat F_0 d'hypothèse par trame.

6.6.2 Résultats d'évaluation ST

Dans leur réglage standard, les AEP diffèrent principalement par les OVR et UVR obtenus. Ceci rejoint les observations de la figure 6.2. Le tableau 6.5 récapitule les moyennes de chaque AEP sur toute la base de données monopitch. Le taux de précision PRR_1 est systématiquement de 100% car tous les AEP ne fournissent systématiquement qu'un seul candidat F_0 d'hypothèse. Pour tous les AEP, les GERL sont faibles et de l'ordre de 1%, résultat classique dans la littérature. Les FER sont comprises entre 1 et 2%. Les GER_G incluant les erreurs de sous-voisé sont naturellement corrélés au taux d'UVR

TABLE 6.6: Synthèses des OVR et UVR en évaluation NT monopitch.

OVR (%)	UVR (%)	μ
3.78	12.94	0.29

TABLE 6.7: Résultats récapitulatifs de l'évaluation NT monopitch.

AEP	$\mathbf{GER_G}$ (%)	$\mathbf{GER_L}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)
YIN	13.80	1.07	1.34	100.00
PRAAT	14.02	1.33	1.15	100.00
SWIPE	13.21	0.34	1.29	100.00
PSH_{mul}	13.45	0.65	1.01	100.00

et plus ce taux est élevé, plus le GER_G est élevé. Dans leur version standard, SWIPE se détache des autres et est le plus performant (hormis pour le taux de FER) avec un taux de GER_G de 3.47%. Les différences notables concernent les taux d'OVR de SWIPE et YIN qui sont relativement élevés. Cela s'explique par un seuil standard trop faible mais n'affecte pas les GER car ces trames sur-voisées n'entrent pas dans les calculs de GER. Il apparaît également une très forte disparité du point de fonctionnement μ des AEP dans leur version standard.

6.6.3 Résultats d'évaluation NT

Les résultats sont ordonnés par corpus. Le tableau 6.6 présente la synthèse des OVR, UVR et μ obtenus sur les trois corpus. Ces OVR et UVR sont les mêmes pour tous les AEP par définition de la méthodologie d'évaluation NT. Comme il était attendu, la méthodologie NT favorise le sous-voisement puisque seule les trames au voisement sûr sont considérées. Le tableau 6.7 récapitule les moyennes de chaque AEP sur toute la base de données monopitch. Il est attendu que les GER_L soit légèrement inférieurs à ceux obtenus dans l'évaluation ST puisque seules les trames sûres sont prises en compte. Par contre, le GER_G doit augmenter car les sous-voisement est nettement augmenté et que ce taux d'erreur intègre les erreurs de sous-voisé. Les comportements attendus sont bien présents. Pour chacun des AEP, le GERL est systématiquement inférieur à celui obtenu dans l'évaluation ST. SWIPE se montre toujours le plus performant sur ce point avec un GER_L de 0.34%. En effet, les trames sélectionnées étant de voisement sûr, les erreurs d'estimation sont moins probable. Les GER_G sont quasiment tous identiques (entre 13% et 14%) et supérieurs au GER_G du plus mauvais AEP de l'évaluation ST (10.70% pour PRAAT). En effet, l'évaluation NT favorise beaucoup le sous-voisement par définition car c'est l'AEP le plus restrictif (ie. de plus fort UVR) qui a tendance à imprimer sa décision VnV aux autres AEP (et donc par voie de conséquence son GER_G).

6.6.4 Résultats d'évaluation VnV

Le nombre de points des courbes d'OVR et d'UVR a été fixé à 1001 (ie. intervalle entre 0 et 1 par pas de 0.001). Le point de fonctionnement μ choisie est fixé à 0.5 ce qui signifie que l'UVR est double de l'OVR. On favorise ainsi légèrement le sous-voisement. PRAAT n'est pas utilisé dans cette l'évaluation VnV car son seuil de décision VnV n'est pas aisément accessible. Les AEP évalués sont

TABLE 6.8: Résultats récapitulatifs de l'évaluation VnV monopitch.

AEP	OVR (%)	UVR (%)	μ	$\mathbf{GER_G}$ (%)	$\mathbf{GER_L}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)
YIN	6.72	13.48	0.50	14.35	1.12	1.34	100.00
SWIPE	4.47	8.93	0.50	9.27	0.44	1.37	100.00
PSH_{mul}	5.07	10.79	0.47	11.61	0.97	1.02	100.00

TABLE 6.9: Performances de PSH_{mul} avec une sur-hypothétisation de 5.

$\mathbf{N_h}$	OVR (%)	UVR (%)	μ	$\mathbf{GER_G}$ (%)	$\mathbf{GER_L}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)
1	9.81	6.48	1.52	7.37	1.03	1.08	100.00
5	**9.81**	**6.48**	**1.52**	**6.72**	**0.29**	**1.06**	**98.24**

YIN, SWIPE et PSH_{mul}. Cette méthodologie ne garantie plus que les AEP traitent les mêmes trames mais elle garantie que la décision VnV est la même pour tous ce qui rend la comparaison des GER_L et GER_G équitable. Cette manière de procéder permet de mettre en évidence la qualité de la décision VnV des AEP. Le tableau 6.8 récapitule les moyennes de chaque AEP sur toute la base de données monopitch. SWIPE reste le plus performant des AEP testés. Des tests statistiques seraient nécessaires pour vérifier que SWIPE est significativement « meilleur » que les autres en situation monopitch.

6.6.5 Performances de PSH_{mul} en « sur-hypothétisation »

La situation maladroitement nommée de « sur-hypothétisation » désigne la situation dans laquelle un AEP propose plus de candidats F_0 d'hypothèse que de candidats F_0 de référence à estimer ($N_h > N_r$). Dans la situation monopitch, seul PSH_{mul} est prévu pour fournir plusieurs hypothèses F_0. Le tableau 6.9 démontre le gain d'une sur-hypothétisation de 5 candidats F_0 d'hypothèse (ligne en gras) sur les résultats de PSH_{min} en version standard comparé aux résultats obtenus sans sur-hypothétisation ($Nh = 1$). Les taux d'OVR et d'UVR ne changent évidemment pas car dans les deux cas l'algorithme est utilisé dans sa version standard et que la décision VnV est identique. En revanche, les GER sont bien diminués, ce qui est logique puisque si le premier candidat F_0 d'hypothèse n'est pas la bonne estimation, il se peut que le second ou le troisième le soit. Dans ce cas monopitch, la sur-hypothétisation est bienvenue car le taux de précision moyen ne diminue que d'à peine 2% pour un gain de 0.6% absolu sur le GER_G et de 0.7% absolu sur le GER_L et pour un temps de calcul quasiment identique.

En situation monopitch, tous les AEP testés offrent des GER locaux relativement proches et compris entre 0.29 et 1.44%. Avec ces ordres de grandeur, il n'est pas forcément utile de chercher à apporter une amélioration à tout prix du GER_L. Il est plus raisonnable de considérer comme état de fait que n'importe quel AEP, dans les zones voisées sûres, fera entre 0.5 et 1.5% d'erreurs grossières locales. Les GER globaux sont plus disparates et sont compris entre 3.47% et 14.02%. Pour ce taux d'erreur qui révèle véritablement la capacité de l'AEP à bien estimer la référence, des progrès semblent encore possibles. Toutefois ce GER_G étant intimement lié au taux d'UVR, tout progrès sur ce point est synonyme d'un progrès dans la décision VnV donc dans la quantification de la force de périodicité. Les FER sont très similaires pour tous les AEP et sont de l'ordre du pour-cent. SWIPE semble être « meilleur » que les autres même si cela reste à confirmer par des tests statistiques plus poussés. La

TABLE 6.10: Résultats des méthodes WWB et VEW dans la littérature. Sources : [Wu et al., 2003, Vishnubhotla and Espy-Wilson, 2008].

Catégorie d'erreur		**VEW (%)**	**WWB (%)**
I	$0 \rightarrow 1,2$	0.5	2.7
	$1 \rightarrow 2$	1.3	0.9
D	$1 \rightarrow 0$	2.8	5.7
	$2 \rightarrow 0,1$	11.6	28.4
S	Monopitch	1.2	5.2
	Bipitch	2.1	6.1
E	Energy domination	52.1	11.3
	Pitch matching	3.1	1.7

sur-hypothétisation de PSH_{mul} dans sa version standard est relativement efficace et permet à notre AEP d'atteindre 6.72% de GER_G et 0.29% de GER_L sur 1h10 de signaux monopitch.

6.7 Évaluation bipitch

Les algorithmes testés sont VEW, WWB et PSH_{mul}. Ces trois algorithmes étant trop hétérogènes dans le fonctionnement et VEW et WWB n'étant pas suffisamment paramétrables, il est impossible d'utiliser la méthodologie d'évaluation par 0-UVR. Nous proposons donc des résultats au travers des trois autres méthodologies ST, NT et VnV dans les paragraphes 6.7.1, 6.7.2 et 6.7.3 respectivement. L'objectif n'est toujours pas d'obtenir un classement du meilleur ou du moins bon algorithme. Il s'agit de valider les méthodologies décrites et les critères de qualité en situation multipitch. A ce jour, les méthodologies et les critères de qualité multipitch pour la parole superposée ne sont pas encore clairement établis dans la littérature.

6.7.1 Résultats d'évaluation ST

L'évaluation standard consiste à tester PSH_{mul}, WWB et VEW dans les versions standards fournies par leurs auteurs. Les AEP testés ici sont PSH_{mul}, WWB et VEW. Le pas d'estimation de F_0 est de 10ms, l'intervalle de recherche de F_0 est de [50,800Hz]. La normalisation de la durée des trames est difficile. Les auteurs de VEW préconisent de ne pas utiliser de trames inférieures à 45ms et les auteurs de WWB utilisent des trames de 16ms puis 32ms. Nos deux AEP utilisent des trames de 40ms. PSH_{mul} est utilisé avec un seuil de décision VnV fixé à 0.5. Les seuils de décision VnV de VEW et WWB ne sont pas accessibles et sont donc tels que les auteurs les ont déterminés. Tous les AEP fournissent deux hypothèses F_0 par trame. Des résultats d'évaluation menées par les auteurs sont accessibles dans [Wu et al., 2003] pour WWB et dans [Vishnubhotla and Espy-Wilson, 2008] pour VEW. Ils sont répertoriés dans le tableau 6.10 suivant : « I » correspond à une erreur d'insertion ce qui revient à notre critère d'OVR, « D » correspond à une erreur de suppression ce qui revient à notre critère d'UVR, « S » correspond à une erreur de substitution ce qui revient à notre GER. Les auteurs séparent l'analyse des GER pour les situation monopitch et bipitch. « E » est une analyse fine des erreurs de suppression $2 \rightarrow 0,1$ et les auteurs montrent que la majorité de ces erreurs sont dues au fait qu'une des sources de la

TABLE 6.11: Récapitulatif des l'évaluation ST bipitch.

AEP	OVR (%)	UVR (%)	μ	GER_G^C (%)	GER_L^C (%)	GER_G^T (%)	GER_L^T (%)	FER (%)	PRR_1 (%)	PRR_2 (%)
WWB	2.28	25.92	0.10	32.35	10.74	36.94	14.89	1.85	99.58	100.00
VEW	18.78	21.82	1.22	39.13	24.34	45.32	30.24	3.34	67.87	100.00
PSH_{mul}	20.43	4.74	4.94	12.60	9.26	16.62	12.59	1.95	96.48	100.00

trame domine l'autre. Une étude plus poussée est nécessaire pour bien comprendre la transformation possible de leurs résultats dans nos critères d'évaluation. Les méthodologies d'évaluation des deux articles sont différentes de celle adopté dans la thèse. Ceci justifie notre besoin de refaire tourner ces deux algorithmes avec nos propres méthodologies et nos propres données. De plus les corpus utilisés sont de plus petite taille que dans ce manuscrit. WWB utilise 10 phrases courtes soit environ 1 minute de mélange et VEW utilise 600 phrases courtes de Mocha soit environ 30 minutes. Cette disparité des corpus testés est d'ailleurs un problème puisque les résultats obtenus dans le tableau 6.10 ne sont pas issus des mêmes signaux et sont donc à comparer avec mille précautions.

L'utilisation des algorithmes WWB et VEW a posé bon nombre de problèmes pratiques. L'algorithme WWB ne supporte pas des fichiers trop longs (supérieurs à 10 secondes) sinon il est trop gourmand en mémoire (plusieurs Go de RAM). Pour tester Keele, il a donc fallu couper chaque signal de Keele manuellement. Ceci a été réalisé en coupant manuellement chaque fichier mono-locuteur entre deux inspirations et en ne conservant que les segments dont la durée est supérieure à 1 seconde et inférieure à 6 seconde. Le corpus ainsi créé est nommé KeeleChunk. L'algorithme VEW a un sérieux problème de temps de calcul puisque pour calculer ne serait-ce que 10 minutes de mélange, il faut bien plus d'une journée de calcul. Additionné à cela, les deux algorithmes bloquent sur certains fichiers qui doivent alors être retiré de l'évaluation. WWB est écrit en C et peu d'information sont données sur son éventuelle optimisation. VEW est écrit en MATLAB et n'est certainement pas du tout optimisé pour être rapide.

Le tableau 6.11 présente la synthèse des résultats obtenus sur les trois corpus. Après élimination des fichiers bloquant, environ 1h de parole superposée bipitch est évaluée. D'après ces résultats, les points de fonctionnement μ des AEP sont tous bien différents. VEW fonctionne en moyenne à l'EER (1.22) alors que WWB pratique clairement le sous-voisement (0.10) et que PSH_{mul} sur-voise de manière importante (4.94). Les taux de représentent le taux de valeurs F_0 de référence correctement estimées et pour ce GER, PSH_{mul} offre les meilleures performances avec 12.6% d'erreur. Il réalise environ trois fois moins d'erreurs que WWB et VEW. Toutefois, ces résultats doivent être mis en perspective avec le taux de sous-voisé de PSH_{mul} qui est environ 5 fois moindre que les deux autres AEP testés. Du point de vue des , WWB et PSH_{mul} se situent dans les mêmes ordres de grandeur ce qui implique que, lorsqu'ils sont en accord de voisement avec la référence, les deux algorithmes sont aussi performants (14.89% et 12.59% respectivement). En revanche, VEW se révèle nettement moins performants que les deux autres sur tous les points. Les taux de GER sur les trames suivent les mêmes observations que les GER sur les candidats. PSH_{mul}, avec ses 16.62% de (trames de références dont toutes les valeurs F_0 voisées sont bien estimées), fait environ deux fois moins d'erreurs que les deux autres AEP. Les taux de FER sont comparables entre WWB et PSH_{mul}. Le taux PRR_1 (trames monopitch) est très proche de 100% pour WWB comme pour PSH_{mul} ce qui signifie que la première valeur F_0 d'hypothèse

TABLE 6.12: Synthèses des OVR et UVR en évaluation NT bipitch.

OVR (%)	UVR (%)	μ
0.95	35.41	0.03

TABLE 6.13: Résultats récapitulatifs de l'évaluation NT bipitch.

AEP	$\mathbf{GER_G^C}$ (%)	$\mathbf{GER_L^C}$ (%)	$\mathbf{GER_G^T}$ (%)	$\mathbf{GER_L^T}$ (%)	FER (%)	$\mathbf{PRR_1}$ (%)	$\mathbf{PRR_2}$ (%)
WWB	40.32	10.80	45.13	15.21	1.84	99.56	100.00
VEW	48.84	23.19	54.31	29.02	3.00	68.24	100.00
PSH_{mul}	38.93	8.69	43.30	12.33	1.77	98.73	100.00

donnée par ces AEP est très souvent la bonne. En revanche le taux de PRR1 de VEW est d'environ 2/3 ce qui signifie qu'en moyenne, VEW doit fournir 1.5 valeurs F_0 d'hypothèse pour correctement estimer la valeur F_0 de référence. Le taux de PRR_2 n'est pas informatif ici car tous les AEP sont paramétrés pour donner deux valeurs F_0 d'hypothèses. Un point supplémentaire vient expliquer les performances de WWB. Cet algorithme utilise un post-traitement de suivi de F_0 à base de HMM qui doit être puissant. Or, les statistiques utilisées pour ce suivi de F_0 ont été optimisées sur une toute autre base de données (cf. [Wu et al., 2003] pour les détails). Les auteurs ne fournissant pas le code de la procédure permettant d'adapter les paramètres de leur système aux données, les performances de WWB sont certainement sous-optimales.

6.7.2 Résultats d'évaluation NT

Les résultats d'évaluation des AEP WWB, VEW et PSH_{mul} sont donnés. Il est attendu que les résultats de l'évaluation aient le même comportement que dans l'évaluation NT monopitch du paragraphe6.6.3, c'est-à-dire que les GER_G soient presque identiques et supérieurs au GER_G du plus « mauvais » AEP en terme d'UVR soit WWB. Le tableau 6.12 présente la synthèse des OVR, UVR et μ obtenus sur les trois corpus KeeleChunk, Bagshaw et Mocha (1h environ). Ces OVR et UVR sont les mêmes pour tous les AEP par définition de la méthodologie d'évaluation NT. L'évaluation NT favorise toujours très largement le sous-voisement comme il a déjà été vu en situation monopitch. Le tableau 6.13 récapitule les moyennes de chaque AEP sur toute la base de données bipitch . Il est attendu que les GER_L^C et GER_L^T soient légèrement inférieurs à ceux obtenus dans l'évaluation ST puisque seules les trames sûres sont prises en compte. Par contre, il est attendu que les GER_G^C et GER_G^T soient nettement dégradés étant donné qu'ils intègrent les erreurs de sous-voisé. Les comportements attendus sont confirmés. VEW apparaît comme étant le moins performant des trois alors que WWB et PSH_{mul} sont relativement équivalents. Lorsque PSH_{mul} et l'annotation de référence sont d'accords sur la décision VnV et que les trames sont voisées, PSH_{mul} ne fait que 8.69% d'erreurs grossières sur les candidats et 12.33% sur les trames. Bien entendu, ces performances dépendent du seuil d'erreur grossière qui est de 20%. En diminuant ce seuil, les performances chuteront. Une analyse statistique serait requise pour quantifier la significativité de ces résultats.

TABLE 6.14: Récapitulatif de l'évaluation VnV bipitch du couple (WWB/PSH_{mul}).

AEP	OVR (%)	UVR (%)	μ	GER_G^C (%)	GER_L^C (%)	GER_G^T (%)	GER_L^T (%)	FER (%)	PRR_1 (%)	PRR_2 (%)
WWB	2.28	25.92	0.10	32.35	10.74	36.94	14.89	1.85	99.58	100.00
PSH_{mul}	2.84	24.80	0.14	28.74	8.97	34.25	12.66	1.92	98.13	100.00

TABLE 6.15: Récapitulatif de l'évaluation VnV bipitch du couple (VEW/PSH_{mul}).

AEP	OVR (%)	UVR (%)	μ	GER_G^C (%)	GER_L^C (%)	GER_G^T (%)	GER_L^T (%)	FER (%)	PRR_1 (%)	PRR_2 (%)
VEW	18.78	21.82	1.22	39.13	24.34	45.32	30.24	3.34	67.87	100.00
PSH_{mul}	8.53	11.37	1.17	17.47	8.93	22.16	12.34	1.96	97.31	100.00

6.7.3 Résultats d'évaluation VnV

Les auteurs de VEW et WWB n'offrent pas la possibilité de régler la décision VnV. Néanmoins, PSH_{mul} est paramétrable. Il est donc possible de régler le point de fonctionnement μ de PSH_{mul} sur celui de WWB puis sur celui de VEW. Cette manière de procéder sera l'évaluation la plus équitable des trois algorithmes en fixant le point de fonctionnement μ à l'identique pour chaque couple d'algorithme à l'exception du couple non paramétrable (WWB/VEW). Le tableau 6.14 présente la synthèse des résultats obtenus sur les trois corpus pour le couple (WWB/PSH_{mul}). Le réglage sur un même point de fonctionnement μ montre qu'il est difficile de départager PSH_{mul} et WWB même si notre algorithme donne systématiquement des taux d'erreurs un peu plus faibles (hormis pour PRR_1). Il faudrait mener une étude statistique pour démontrer la significativité des résultats. Le léger décalage de point de fonctionnement μ (0.14 contre 0.10) provient de certains fichiers pour lesquels il n'est pas possible de trouver exactement le même point de fonctionnement. Dans ces cas très peu nombreux, PSH_{mul} est réglé de manière à être au plus proche du μ de WWB. Ces cas étant très isolés, les résultats sont valables et équitables. Le tableau 6.15 présente la synthèse des résultats obtenus sur les trois corpus pour le couple (VEW/PSH_{mul}). En comparaison avec VEW sur le même point de fonctionnement μ, notre algorithme fait deux fois moins d'erreurs grossières globales sur les candidats et sur les trames. Il y a encore un léger décalage de point de fonctionnement μ (1.22 contre 1.17) mais ce décalage n'altère en rien l'équité des résultats. Les tendances dans cette comparaison sont fortes et il est fort probable que PSH_{mul} soit plus performant que VEW dans sa version actuelle. VEW doit être défendu dans la mesure où cet algorithme est encore en phase de développement et qu'il n'est en aucun cas un « produit » fini.

6.7.4 Performances de PSH_{mul} en sur-hypothétisation

La sur-hypothétisation est importante dans les situations multipitch puisque les fonctions de pitch étant plus parasités, il est probablement nécessaire de considérer plus de maxima locaux que de candidats F_0 de référence. La sur-hypothétisation doit permettre, comme en situation monopitch, d'améliorer les performances d'estimation de F_0 de l'AEP mais au détriment de la précision de l'algorithme. La dégradation de la précision doit être d'autant plus importante que la situation est multipitch. Seul PSH_{mul} autorise plus de deux hypothèses par trames. Ses performances sont donc testées dans ce para-

TABLE 6.16: Performances de PSH_{mul} avec une sur-hypothétisation de 5.

N_h	OVR (%)	UVR (%)	μ	GER_G^C (%)	GER_L^C (%)	GER_G^T (%)	GER_L^T (%)	FER (%)	PRR_1 (%)	PRR_2 (%)
2	20.43	4.74	4.94	12.60	9.26	16.62	12.59	1.95	96.48	100.00
10	**20.43**	**4.74**	**4.94**	**8.67**	**5.02**	**11.40**	**6.85**	**1.78**	**92.98**	**82.93**

graphe. La décision VnV de PSH_{mul} est fixée à 0.5, sa valeur standard. Le nombre d'hypothèses F_0 par trame est fixé à 10. Il est attendu que les GER^C et GER^T diminuent comparés à ceux obtenus dans l'évaluation NT. Le tableau 6.16 présente un récapitulatif des résultats obtenus sur l'ensemble de la base de données de mélange bipitch. Les résultats sont cohérents. D'une part, les taux d'OVR et d'UVR sont bien les mêmes ce qui est logique puisque le même réglage que celui standard de l'évaluation ST est utilisé (cf. tableau 6.9). Une baisse est observée sur tous les GER et PSH_{mul} ne fait plus que 8.67% d'erreurs grossières globale sur les candidats ce qui est très prometteur. La baisse de précision du taux de PRR_2 est bien visible puisqu'une chute de presque 20 % est observée (82.93%). Ceci signifie qu'en moyenne, il faut environ 2.5 valeurs F_0 d'hypothèse par trame bipitch pour correctement estimer les valeurs F_0 de référence.

6.8 Conclusions

Ce chapitre a dressé un rapide état de l'art en matière de détection et de quantification de voisement et montre qu'il est encore mal défini. Ceci implique des erreurs de décision VnV sur les trames qui implique elle-même des problèmes dans l'évaluation. Un apport de ce chapitre est de proposer une méthodologie intégrant une normalisation du comportement des AEP au niveau de la décision VnV. Cette normalisation passe par μ, un point de fonctionnement qui correspond au rapport entre OVR et UVR. Un autre apport a été d'évaluer comparativement trois AEP multipitch dans la situation bipitch sur environ 1h de mélange, durée qui à notre connaissance n'avait pas été encore testés. Quatre méthodologies ont été proposées et validées :

- La méthodologie 0-UVR qui est la plus « équitable » mais aussi la plus difficile à mettre en pratique. En effet, elle requiert que les AEP testés puissent être paramétrés pour donner une estimation des F_0 quelque soit l'état de voisement de la trame. Or, bien souvent les AEP prennent une décision VnV et mettent à 0 les trames considérées comme non-voisées et n'autorisent pas le paramétrage de cette décision.
- La méthodologie standard permet de tester et de comparer facilement des AEP différents. Toutefois, il doit **nécessairement** être accompagné des taux d'OVR et d'UVR pour avoir du sens. Sinon les AEP peuvent être placés à n'importe quel point de fonctionnement et l'équité de l'évaluation est compromise.
- La méthodologie NT permet de normaliser le point de fonctionnement des AEP et d'évaluer les AEP sur les mêmes trames. Cette évaluation est donc équitable mais présente deux inconvénients. D'une part, il est possible que les résultats de l'évaluation soient « alignés » sur l'AEP de plus « mauvaise » qualité en termes de décision VnV. D'autre part, l'expérimentateur ne peut pas contrôler le point de fonctionnement sur lequel il se trouve puisque ce point dépend des AEP

testés. Il suffit d'ajouter ou de retirer un AEP à l'évaluation et ce point de fonctionnement est changé.

- La méthodologie VnV permet à l'expérimentateur de fixer μ à la valeur qu'il souhaite. Le fait d'ajouter ou de retirer un AEP devient transparent. Le désavantage est de ne plus strictement évaluer les AEP sur les mêmes trames. Toutefois, si ce point de fonctionnement favorise suffisamment le sous-voisement, on se rapproche de l'évaluation NT et les trames évaluées sont presque identiques sur chacun des AEP.

Les performances de PSH_{mul} ont été mises en évidence lorsque celui-ci fournit plus de candidat F_0 d'hypothèse que de candidats F_0 de référence à estimer. Au stade actuel du développement de PSH_{mul}, nous disposons d'un AEP capable d'estimer correctement les F_0 de trames bipitch avec 83.48% $(100-16.62)$ de réussite dans le réglage standard de sa version d'étude actuelle. Ce résultat est très encourageant pour réaliser du suivi de F_0 à partir de cet AEP d'autant qu'il peut être améliorer en pratiquant la sur-hypothétisation. L'ensemble de ces évaluations justifie la faisabilité d'un algorithme d'estimation de F_0 multiples à partir de peignes spectraux.

Chapitre 7

Conclusions et Perspectives

Ce travail apporte deux contributions principales. La première concerne le concept de PSH qui a permis de mettre en œuvre un nouveau système automatique d'estimation de F_0 conçu pour traiter des mélanges de parole. La seconde concerne la description et la validation de méthodologies d'évaluation des AEP valables en situation monopitch et multipitch permettant d'évaluer équitablement les AEP. Ces deux contributions principales résultent d'autres contributions plus ciblées. Nous pouvons citer l'analyse détaillée de différents peignes fréquentiels et la description de deux nouveaux peignes : les PDN et PDM. L'indexation (p, q) des pics parasites d'une fonction de pitch est une nouveauté dans la mesure où ce point n'a jamais été montré aussi central que dans cette thèse. Le comportement hyperbolique des fonctions de pitch est clairement expliqué et une solution est donnée pour l'éviter. La normalisation de la décision VnV dans l'évaluation des AEP sous la forme d'un rapport $\mu = (OVR/UVR)$ doit être soulignée. Enfin, un effot a été porté sur le positionnement scientifique de ce travail et sur les différents états de l'art du manuscrit s'étendant de l'utilisation de la F_0 dans les systèmes actuels de séparation de parole jusqu'à la quantification de voisement et la décision VnV en passant par la description de quelques AEP principaux en situation monopitch et multipitch. Après ce résumé des apports de la thèse, il convient de conclure chapitre par chapitre l'organisation du manuscrit, le cheminement des raisonnements menés et les résultats qui ont été obtenus.

Le chapitre 2 a dressé un bilan de l'utilité de la fréquence fondamentale dans un processus de séparation automatique de parole. La séparation automatique de parole dans sa forme la plus complète consiste en un système recevant un mélange sonore en entrée et reconstruisant en sortie les signaux constituants. L'Analyse de Scène Auditive (ASA) a mis en œuvre un grand nombre d'expérience psycho-acoustiques visant à comprendre la faculté humaine de séparation de sources. L'ensemble des travaux converge vers l'existence d'un processus de construction de flux auditifs (streaming). Dans le streaming, la F_0 est un indice acoustique important. Ses effets se retrouvent dans le groupement séquentiel qui regroupe par continuité temporelle. Ses effets se retrouvent également dans le groupement simultané qui regroupe par continuité fréquentielle. La F_0 favorise le glimpsing, phénomène d'émergence localisée d'une source sonore dans un espace temps-fréquence. Des études montrent également l'importance de la hauteur de voix qui, si elle est mal retransmise aux malentendants, diminue l'intelligibilité de la scène auditive. De manière générale, la F_0 se trouve être utilisée dans nos mécanismes de perception et particulièrement dans le streaming, un des mécanismes de la séparation de sources. Malgré l'évidente

utilisation de la F_0 chez l'homme, les systèmes automatiques actuels n'utilisent pas systématiquement la F_0. C'est le cas des approches de séparation aveugle de source (BSS) qui reposent sur des connaissances a priori concernant les sources mélangées et sur leur indépendance statistique. Les approches CASA sont divisées en approches binaurales et monaurales. Les premières approches (binaurales) n'utilisent pas toujours la F_0 puisqu'elles disposent de l'information d'ITD. Néanmoins des travaux montrent qu'ajouter la F_0 dans ces systèmes binauraux augmente leurs performances. La F_0 est utilisée par toutes les approches CASA monaurales, approche adoptée dans ce travail. Pour qui souhaite séparer de la parole par des méthodes CASA monaurale, la F_0 est un critère acoustique quasiment incontournable. Le problème d'estimation de F_0 s'est alors naturellement posé. Avec la contrainte supplémentaire de traiter des mélanges multipitch, le problème d'estimation de F_0 multiples est devenu incontournable.

Le chapitre 3 commence par analyser les caractéristiques intrinsèques du signal de parole, puis les compare à celles d'un signal d'instrument de musique. Les deux types de signaux sont suffisamment différents pour que les AEP soit spécialisés pour l'un ou l'autre. Il décrit également les perturbations induites par le canal de communication entre l'émetteur et le récepteur. Ce point est important car l'AEP que nous proposons se situe à la place du récepteur et capte donc le signal de l'émetteur auquel se sont ajoutés bon nombre de bruits dus à l'environnement sonore. Il procède ensuite à un état de l'art des méthodes d'estimation de F_0. Tout d'abord, il présente les méthodes monopitch qui sont classables en trois grandes familles d'approches : les approches temporelles, fréquentielles et spectro-temporelles. Ces approches peuvent également être observées d'un point de vue déterministe ou probabiliste. Le point de vue que nous avons suivi est déterministe. Quelque que soit la famille d'approche et le point de vue adopté, tous les AEP monopitch construisent ce que nous nommons « fonction de pitch » qui quantifie la force d'une fréquence donnée en tant qu'hypothèse F_0. Lorsque le segment sonore analysé est voisé, des maxima locaux apparaissent dans cette fonction. L'estimation de la F_0 consiste à repérer la fréquence du maximum local de plus forte amplitude et de le considérer comme la F_0 du segment.

Les fonctions de pitch sont composées d'une multitude de pics parasites dont les plus importants sont situés aux fréquences des F_0 mais également à leurs multiples et sous-multiples. Ces pics parasites sont parfois d'amplitude plus grande que celle des pics des F_0 et produisent alors des erreurs d'estimation typiquement nommées erreurs d'octave et de sous-octave. Une des contributions de la thèse est d'établir clairement que la position de ces pics parasites et donc des erreurs n'est pas aléatoire. Les méthodes classiques de correction de ces erreurs font intervenir des post-traitements par filtrage médian, programmation dynamique. Le suivi de F_0 intervient alors dans l'estimation de F_0. Notre AEP propose de régler le problème des pics parasites directement dans la construction de la fonction de pitch et pas par un post-traitement quelconque.

Dans une situation multipitch, l'estimation de F_0 est compliquée par deux phénomènes principaux : les harmoniques communes et l'estimation du nombre de sources mélangées. L'estimation de F_0 est faite de manière itérative ou bien conjointe. L'estimation itérative consiste à estimer la F_0 dominante d'un mélange en exploitant une fonction de pitch, puis de supprimer les harmoniques de la F_0 estimée et de réitérer l'opération jusqu'à trouver toutes les F_0. Ainsi, elle requiert la connaissance du nombre de sources mélangées ou bien un critère d'arrêt. Dans la pratique, l'étape de suppression de structure harmonique est difficile à mettre en œuvre. L'estimation conjointe consiste à se donner un certain nombre d'hypothèses F_0 par trame en exploitant une fonction de pitch. Il s'agit ensuite selon certaines

heuristiques (contraintes de continuités temporelles ou fréquentielles) de trouver la combinaison de ces hypothèses qui offre le moindre coût. La recherche d'un meilleur chemin parmi un ensemble de chemins possibles est un problème connu pour sa complexité algorithmique ce qui implique que généralement les AEP par estimation conjointe sont plus lourds que les AEP itératifs.

Ce chapitre détaille ensuite deux AEP multipitch récents qui ont été évalués. Il s'agit de WWB et VEW. Les deux sont des algorithmes spectro-temporels et sont conçus pour traiter uniquement des mélanges monopitch ou bipitch. Ils ne peuvent pas être aisément étendus à plus de deux F_0. Le premier repose sur le calcul d'une ACF sur chacun des canaux. Il utilise ensuite un algorithme de suivi de F_0 à base de Modèles de Markov Cachés pour calculer les meilleures trajectoires de F_0. Le second algorithme repose sur une AMDF 2-D qui permet de construire un plan dont les abscisses et ordonnées sont les lags (décalages) en échantillons. L'algorithme sépare le signal en canaux. Pour chaque canal il calcule le couple de lags qui donne le plus faible minimum local dans l'AMDF 2-D et ce couple correspond aux hypothèses F_0. Le chapitre se conclut en justifiant notre choix d'AEP purement fréquentiel. En effet, les approches spectro-temporelles sont déjà étudiées dans VEW et WWB qui, de plus, utilisent des approches temporelles (ACF ou AMDF 2-D) sur chaque canal. L'utilisation d'un AEP purement fréquentiel s'est donc logiquement posé.

Le chapitre 4 commence par justifier le choix d'un AEP par peigne fréquentiel et par estimation conjointe. L'estimation conjointe se justifie par la problématique étape de suppression des estimations itératives et par la facilité de calcul des hypothèses F_0. En effet, elles se calculent en détectant les N premiers maxima locaux d'une fonction de pitch, N étant paramétrable à souhait. Un AEP par peigne fréquentiel se justifie par le fait que les peignes fréquentiels permettent de considérer une structure harmonique complète et pas seulement quelques harmoniques. Cette caractéristique doit les rendre robustes même aux signaux dont la qualité des harmoniques est dégradée. Les fonctions de pitch des approches d'estimation de F_0 par peignes fréquentiels sont construites à partir de produits scalaires entre un peigne et un spectre d'amplitude. La fréquence F_c des peignes varie dans un intervalle de recherche de F_0 [50,800Hz]. La meilleure façon de faire varier F_c est de manière logarithmique en pas constant par octave.

Le chapitre présente une analyse détaillée du comportement de peignes simples tels que le PUI, PUF, PDI, PDF. Le PUI est un Peigne Uniforme Infini, c'est-à-dire un échantillonneur dont la fréquence F_c est réglable. Le PUF est un Peigne Uniforme Fini c'est-à-dire un échantillonneur au nombre de dents limité. Le PDI est un Peigne Décroissant Infini c'est-à-dire un échantillonneur dont l'amplitude des dents décroit en fonction du numéro de la dent. Le PDF est un Peigne Décroissant Fini c'est-à-dire un échantillonneur dont le nombre de dents est limité et dont l'amplitude des dents décroît avec le numéro de la dent. Lorsque la F_c d'un de ces peignes est égale à F_0, un maximum est produit dans la fonction de pitch. Seulement, ce pic n'est pas unique et de nombreux pics parasites apparaissent également. Avec un PUI, les pics parasites sont de même amplitude que le pic en F_0 ce qui produit des erreurs d'estimation. De plus, lorsque le spectre d'amplitude n'est pas de moyenne nulle, un comportement hyperbolique en $(1/F_c)$ est introduit dans les fonctions de pitch ce qui rend les basses fréquences très parasites. Un apport de la thèse est de proposer une solution pour supprimer définitivement ce comportement hyperbolique. Une des contributions de la thèse est d'établir clairement que la position de ces pics parasites est indexable par un couple d'entiers non nuls (p, q) premiers entre eux. Ce couple (p, q) défini

la position fréquentielle d'un pic parasite par rapport à une F_0 suivant la relation $(p/q)F_0$. L'analyse des comportements des PUI, PUF, PDI et PDF montre un compromis entre l'atténuation des pics parasites sous-harmonique et l'atténuation des pics parasites sur-harmonique. Un PAL a été proposé et ajoute des dents négatives au peigne fréquentiel. Ces dents négatives ont pour effet d'atténuer les pics parasites sur-harmoniques en laissant relativement intacts les pics sous-harmoniques. Toutefois ce PAL souffre toujours du compromis entre atténuation sous et sur-harmonique même si ses performances par rapport à des peignes plus simples sont meilleures. Ainsi, les PDN et PDM sont introduits. Ils sont définis par F_c et par un ordre Δ (nombre premier). Le PDN est une combinaison de deux PUI dont un est d'amplitude négative. Le PDM est un peigne uniforme infini dont il manque les dents de numéro multiple de Δ. Il est démontré que ces peignes permettent d'annuler sélectivement les pics parasites. Les PDN atténuent les pics de type $(\Delta p, q)$ avec Δp et q entiers et fraction irréductible. Les PDM atténuent les pics de type $(p, \Delta q)$ avec p et Δq entiers et fraction irréductible. Chacun des peignes a été étudié dans une situation monopitch et multipitch (bipitch et 4-pitch) à partir d'un signal synthétique. Les comportements annulateurs des PDN et PDM sont repris dans notre AEP. Le principe est d'annuler de manière sélective chaque famille de pics parasites pour construire une fonction de pitch dans laquelle les pics parasites sont atténués et qui est donc facilement exploitable et peu enclin aux erreurs. Ce principe est appelé principe de Peigne à Suppression Harmonique ou PSH.

Le chapitre 5 détaille les différentes étapes de deux implémentations particulières du principe de PSH. La première implémentation se nomme PSH_{mul} et a été évaluée. Elle a été construite pour être la plus rapide et la plus robuste possible. Après extraction du spectre d'amplitude de chaque trame, la fonction de pitch par PUI avec correction hyperbolique est calculée. Cette fonction de pitch est utilisée pour obtenir les différentes fonctions de pitch des PDN et PDM aux ordres 2, 3, 5 et 7. Les résultats individuels des fonctions PDN2, PDN3, PDN5, PDN7, PDM2, PDM3, PDM5 et PDNM 7 sont multipliés (fonction DPP) ce qui laisse émerger fortement les pics des F_0 à estimer. La seconde implémentation est se nomme PSH_{min} et son objectif a été avant tout pédagogique. Il fabrique un modèle de spectre d'amplitude de moyenne nulle (correction hyperbolique) issu d'une détection de partiels effectuée sur le spectre d'amplitude d'une trame. Chaque partiel conservé est remplacé par une parabole. A partir de ce modèle de spectre d'amplitude, PSH_{min} calcule les fonctions de pitch PDN et PDM aux ordres premiers 2, 3, 5, 7, etc. Chacun des PDN et PDM annule ou du moins atténue une famille de pics parasites tout en laissant relativement intact les pics des F_0. Une fonction DPP (Détection des Pics Parasites) est extraite de tous ces peignes en prenant pour chaque F_c la valeur minimale. Ainsi, les pics parasites sont atténués et les pics F_0 restent intacts. Dans les deux implémentations fondées sur le principe de PSH, la fonction de pitch est appelée fonction PSH. Elle est le résultat de la comparaison entre la fonction DPP et la fonction obtenue par un PUI. Les comportement de PSH_{mul} et PSH_{min} sont analysés de manière pédagogique sur des exemples synthétiques en situation monopitch, bipitch et 4 pitch et montre sa capacité d'estimation de F_0 multiples. En particulier, les deux implémentations se montrent capables de bien traiter le cas d'école de deux F_0 à l'octave ou en situation de virtual multipitch. La résolution fréquentielle actuelle de PSH est de l'ordre de 3 à 6%. Ils sont aussi testés sur des trames de parole réelle monopitch, bipitch et 4-pitch. Il parvient à trouver les bons candidats F_0 mais les fonctions de pitch sont de moins bonne qualité. Les limitations des approches fondées sur le principe de PSH proviennent de la différence des intensités dans les signaux mélangés, de la qualité

des structures harmoniques du signal, du résultat de l'interaction des harmoniques communes, etc. Un exemple avec bruit rose est aussi décrit pour montrer que PSH_{min} y est relativement robuste. Les études sur des cas isolés doivent être mises à l'échelle et illustrées sur des corpus de mélanges de parole réelle. Le chapitre 6 d'évaluation s'est imposé naturellement.

Le chapitre 6 a proposé plusieurs méthodologies d'évaluation monopitch et multipitch. Il a davatage visé à valider ces méthodologies qu'à établir un classement des AEP évalués. D'ailleurs un tel classement n'a pas de sens dans la mesure où tous les AEP évalués n'ont nullement la prétention d'être des outils commercialisables. Il commence par soulever le problème de quantification du voisement. La décision voisé/non-voisé qui découle de l'étape de quantification de voisement est ensuite abordée. Ces considérations nous amènent à proposer une nouvelle méthodologie d'évaluation valable en situation monopitch et multipitch qui intègre la décision VnV. En effet, l'évaluation de la qualité d'estimation porte uniquement sur les trames qui sont à la fois voisées pour la référence et voisées pour l'AEP. Si l'AEP ne considère voisé que les trames dont la force de voisement est sûre, alors l'évaluation portera sur peu de trames sûres et le taux d'erreur sera artificiellement diminué. Nous proposons de normaliser la décision VnV de tous les AEP en fixant le point de fonctionnement de chacun à une valeur constante noté μ correspondant à un rapport (OVR/UVR) constant. OVR est le taux d'erreur de sur-voisé (référence non-voisée et hypothèse voisée – fausse alarme) et UVR est le taux d'erreur de sur-voisé (référence voisée et hypothèse non-voisée – faux rejet). L'évaluation de l'estimation est réalisée par deux critères : le GER ou taux d'erreurs grossières et le FER ou taux d'erreurs fines. Le GER est un taux d'erreur relative à 20% entre une valeur de référence et une hypothèse. L'évaluation monopitch est réalisée sur quatre algorithmes qui sont PRAAT, YIN, SWIPE et PSH_{mul}. Trois méthodologies différentes sont utilisées. La première est dite standard et consiste à calculer l'OVR, l'UVR, le GER et le FER obtenus par les AEP dans leur version standard. La seconde est dite « par normlisation des trames » et consiste à calculer l'OVR, l'UVR, le GER et le FER seulement sur trames qui sont considérées voisées pour tous les AEP. On mesure ainsi les performances des AEP sur les zones dont le voisement est indéniable. La troisième méthodologie permet à l'expérimentateur de fixer le point de fonctionnement μ de la décision VnV à la valeur qu'il souhaite. L'évaluation multipitch est réalisée sur trois algorithmes qui sont VEW, WWB et PSH. Les critères de qualité utilisés sont les OVR et UVR concernant la décision VnV et les GER^C, GER^T, FER et PRR concernant l'estimation des F_0. Le GER^C renseigne sur la qualité d'un AEP à correctement estimer toutes les références F_0 et le GER^T renseigne sur la qualité d'un AEP à correctement estimer toutes les références F_0 de chacune des trames. Les trois méthodologies sont appliquées. La première est une méthodologie standard. La seconde est la méthodologie par normalisation des trames et la troisième consiste à régler les implémentations PSH_{mul} sur le point de fonctionnement de décision VnV de VEW puis de WWB pour pouvoir les comparer équitablement. La dernière évaluation menée en multipitch a permis de calculer le taux de précision moyen PRR de PSH_{mul}, WWB et VEW c'est-à-dire le nombre moyen d'hypothèses par trame nécessaire aux AEP pour qu'ils détectent les F_0 de manière optimale. PSH_{mul}se révèle être un AEP performant tant en situation monopitch qu'en situation multipitch. Il est capable d'estimer correctement les F_0 d'une situation bipitch pour 83% des trames. Cela est encourageant pour son utilisation comme brique de base à un système automatique de séparation de parole.

Les perspectives directes de ce travail de thèse concernent dans un premier temps les améliorations

à apporter au système actuel. Ces améliorations passent par la diffusion de PSH la plus large possible. Nous avons pour ambition de rendre une implémentation accessible à tous et espérons qu'une utilisation intensive puisse permettre de pointer plus rapidement les étapes sensibles de l'algorithme. D'ores et déjà, il montre des signes de faiblesse dans les situations de déséquilibre des intensités des signaux mélangés. Une étude intéressante à mener concernerait les limites de fonctionnement de PSH en testant divers mélanges à plusieurs différences d'intensité pour établir des bornes optimales de fonctionnement de l'AEP. Dans la même optique, il serait très utile d'analyser finement sa résolution fréquentielle. Jusqu'à quelle différence de F_0 PSH est-il capable de discriminer deux F_0 ? Une étude pourrait être menée sur la capacité de PSH à traiter des situations très multipitch. Jusqu'à combien de locuteurs peut-on considérer que la situation de parole superposée est résoluble ? La limite doit se situer autour de quatre mais cette intuition mériterait des investigations plus poussées. Il faudrait également une analyse de sa méthode de calcul de la force de voisement qui est originale. Il serait utile également d'analyser l'étape de détection de partiels qui est relativement rudimentaire et de l'améliorer si nécessaire.

Un AEP fondé sur le principe de PSH peut servir à construire d'autres applications. En quoi l'estimation de F_0 peut-elle aider à détecter les zones de parole superposée. Leur détection apporterait une aide aux systèmes actuels de reconnaissance de parole qui sont mal adaptés pour les traiter. Une autre application concerne le problème de suivi et de reconstruction de flux de F_0 dans un mélange de parole. Si une implémentation de PSH est capable de donner les bonnes hypothèses en trame-à-trame, alors il doit être possible de reconstruire les flux de F_0 d'un mélange facilement sans avoir à traiter trop d'hypothèses F_0 parasites. La publication qui traite de cette application propose un algorithme simple qui permet de regrouper ensemble des segments de trajectoires de F_0 en fonction d'une contrainte simple sur l'évolution du F_0. L'algorithme s'appuie sur une programmation dynamique mais ne paraît pas complètement adapté. Une piste envisageable serait de reconstruire les flux de F_0 petit à petit en regroupant d'abord en segments homogènes les candidats qui respectent une évolution maximale du F_0 trame à trame. A partir de ces segments F_0 cohérents, l'aide d'autres critères tels que de des distances spectrales, des modèles de locuteur, etc. pourrait permettre de désambiguïser certaines situations. Des études psycho-acoustiques ont montré que pour une voix non pathologique et dans des conditions de lecture, l'évolution maximale d'un F_0 est de l'ordre de 1% par milliseconde. S'il y a la moindre ambiguïté c'est-à-dire plusieurs possibilités de continuer un segment F_0, on préférera à ce niveau ne pas prendre de décision et générer de nouveaux segments. On reporte ainsi la décision pour des niveaux supérieurs qui auront intégré plus d'informations pour mieux décider.

Dans un horizon un peu plus lointain, une implémentation de PSH pourrait être testé sur de la voix chantée solo, duo et voir plus. Cela rejoindrait d'ailleurs notre ambition de vouloir diffuser l'algorithme le plus largement possible. Les F_0 de la voix chantée ont l'avantage de pouvoir être catégorisées en notes discrètes ce qui peut favoriser PSH. En effet, dans la musique occidentale, les notes étant en relation (p, q) ex. quinte environ $(3, 2)$, octave $(2, 1)$, tierce environ $(5, 4)$, etc, il est possible que le comportement du PSH soit parfaitement adapté à ce genre de signaux pour lesquels les structures harmoniques sont en relation (p, q). En revanche, dans quelle mesure un algorithme PSH est-il capable de suivre le vibrato ou autre effets de la voix chantée ? Est-il robuste à la dynamique importante de l'intensité sonore ? Dans le même temps, il serait intéressant d'observer le comportement d'un algorithme PSH sur des sons d'instruments de musique. Il faudra tout de même résoudre les problèmes d'inharmonicité qui sont

propres à quelques instruments tels que le piano. PSH est-il en mesure d'améliorer les applications de transcription automatique de partitions ?

D'autres perspectives à moyen-terme de cette thèse concernent l'extraction d'autres indices acoustiques tels que les instants d'apparition/disparition de sources, le timbre, la localisation, etc. Comment la coopération de plusieurs indices acoustiques peut-elle amener à la séparation de parole ? Certains travaux existent déjà en la matière et notamment ceux qui s'appuient sur le glimpsing et le masque binaire dans une représentation spectro-temporelle sur signal.

Les perspectives à long-terme de ce travail s'inscrivent dans ce qui a été présenté dans le chapitre 2 c'est-à-dire l'élaboration d'un système automatique de séparation de parole capable de traiter avec les signaux de mélange les plus spontanés et naturels possible. La perception auditive est un processus à double sens et l'un des sens de traitement est bottom-up. Un système cognitif perçoit une scène auditive et traite les informations sonores qui y sont contenues. L'oreille capte l'onde physique, la transforme en vibration puis en message nerveux. Des mécanismes de plus en plus complexes lui permettent d'établir des généralisations à partir des stimulations sonores reçues du monde extérieur. Le second sens de traitement est nommé top-down. Un système cognitif immergé dans un environnement possède ses propres motivations, ses buts, ses attentes. Il projette sur son environnement une forme d'anticipation, de prédiction du monde qu'il a appris au fur et à mesure des situations qu'il a rencontrées. Il observe alors la pertinence de ses anticipations et raffine ses apprentissages en fonction. Un système cognitif est donc actif dans la construction de sa perception sonore. Être capable de construire un tel système de séparation automatique est un vrai défi. L'estimation de la F_0 est un traitement bas-niveau très utile à notre perception dans la séparation de sources et le principe de PSH peut tout à fait être considéré comme une brique de base d'un tel système. Cette thématique de recherche est en lien avec l'apprentissage de connaissances puisque les traitements top-down les utilisent pour anticiper. Ces perspectives soulèvent beaucoup de questions. Comment modéliser des connaissances sonores ? Comment fusionner des données aussi hétérogènes que la F_0, l'intensité, les instants d'apparitions/disparitions, le timbre, pour que cette fusion produise un mécanisme de streaming ? Autrement dit, sous quelle forme entreprendre la fusion des informations bottom-up et top-down pour faire émerger une faculté de séparation de sources ?

Une application pourrait être intégrée sur un agent robotique quelconque afin de le doter d'un système auditif artificiel. L'application la plus noble à nos yeux serait de parvenir à la construction d'algorithmes de séparation de parole efficaces et implantables sur des prothèses auditives et capable de donner aux malentendants un confort auditif qu'il n'ont pas encore aujourd'hui. Le principe de PSH reste certes une étape élémentaire d'un tel système mais c'est avec chaque goutte d'eau que se forment les océans.

Annexe A

Preuves mathématiques

Les raisonnements portent sur une modélisation idéale du spectre d'amplitude d'un signal voisé. Le spectre d'amplitude est considéré comme un peigne uniforme fini (PUF) de fréquence fondamentale F_0, les amplitudes des harmoniques sont toutes d'amplitude unitaire et le nombre d'harmoniques est noté Nh. Cette modélisation peut s'écrire sous la forme d'un peigne de Dirac comme l'indique l'équation A.1. La fonction δ désigne la distribution de Dirac.

$$|S(f)| = \sum_{n=1}^{N_h} \delta(f - nF_0) \tag{A.1}$$

Il est nécessaire de réintroduire la notation donnée dans le chapitre 4 concernant le nombre de dents d'un PUI dans un intervalle donné. F_{MAX} désigne la fréquence de coupure du spectre d'amplitude et par conséquent la fréquence maximale de calcul du produit scalaire. F_{MAX} impose un nombre limité de dents aux peignes dits « infinis » tel que le PUI ou le PDI. Le nombre de dents contenues dans l'intervalle de 0 à la fréquence de coupure F_{MAX} est donné par l'équation A.2. La fonction du nombre de dents est notée $M(F_c)$. E désigne la partie entière.

$$M(F_c) = E\left(\frac{F_{MAX}}{F_c}\right) \tag{A.2}$$

A.1 PUI

Un PUI est un échantillonneur de fréquence caractéristique F_c. L'équation A.3 formalise mathématiquement le PUI.

$$PUI(F_c, f) = \sum_{n=1}^{N_h} \delta(f - nF_c) \tag{A.3}$$

Le produit scalaire d'une fonction quelconque avec un échantillonneur est la somme des échantillons de la fonction quelconque pris aux instants d'échantillonnage de l'échantillonneur. Ceci se traduit dans l'équation A.4 qui donne la formule générale d'une fonction de pitch par PUI pour un spectre

d'amplitude quelconque $|S(f)|$.

$$\Phi_{PUI}(F_c) = \sum_{k=1}^{M(F_c)} |S(kF_c)| \tag{A.4}$$

Le spectre d'amplitude $|S(f)|$ étant un PUF de N_h dents et de fréquence fondamentale F_0, il vaut 1 pour tous les multiples de F_0 jusqu'à la dent de numéro N_h et zéro ailleurs. Pour que la fonction de pitch $\Phi_{PUI}(F_c)$ de l'équation A.4 ne soit pas nulle, il faut qu'au moins une dent du spectre soit située à la même fréquence qu'une dent du PUI. Cette proposition est formalisée dans l'équation A.5.

$$\exists (p,q) \in [1, N_h] \times [1, M(F_c)], pF_0 = qF_c \tag{A.5}$$

L'équation A.5 est justement la preuve théorique de l'indexation (p, q) discutée dans le manuscrit dans le chapitre 4 section (index). Pour que la fonction PUI soit non nulle, il faut nécessairement que l'équation A.6 se vérifie.

$$F_c = \frac{p}{q} F_0 \tag{A.6}$$

Si cette équation n'est pas respectée alors la fonction PUI est nulle. N'importe quelle fraction (p/q) doit être réduite sous sa forme irréductible pour éviter les doublons (ex. $(2/2) = (1/1)$). Soit F_c égale à une fraction irréductible de F_0 tel que $Fc = (p/q)F_0$ alors l'équation A.7 donne le point de départ du calcul de la fonction de PUI.

$$\Phi_{PUI}\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{M\left(\frac{p}{q}F_0\right)} \left|S\left(k\frac{p}{q}F_0\right)\right| = \sum_{k=1}^{M\left(\frac{p}{q}F_0\right)} \sum_{n=1}^{N_h} \delta\left(k\frac{p}{q}F_0 - nF_0\right) \tag{A.7}$$

La fonction PUI est non nulle lorsque le Dirac vaut $\delta(0)$. Ceci se traduit par l'équation A.8.

$$k\frac{p}{q}F_0 - nF_0 = 0 \Leftrightarrow kp = nq \tag{A.8}$$

Ceci se produit lorsque k est multiple de q et n est multiple de p et que les multiples sont identiques. k variant entre 1 et $M(F_c)$, il y a donc $E(M(F_c)/q)$ multiples de q. De même, n variant entre 1 et N_h, il y a donc $E(N_h/p)$ multiples de p dans cet intervalle. Dans l'équation A.9 ci-dessous, k est remplacé par $k^{'}q$ et n est remplacé par $n^{'}p$ avec $(k^{'}, n^{'}) \in \mathbb{N}^{*2}$.

$$\Phi_{PUI}\left(\frac{p}{q}F_0\right) = \sum_{k'=1}^{E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right)} \sum_{n'=1}^{E\left(\frac{N_h}{p}\right)} \delta\left(k^{'}q\frac{p}{q}F_0 - n^{'}pF_0\right) = \sum_{k'=1}^{E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right)} \sum_{n'=1}^{E\left(\frac{N_h}{p}\right)} \delta(k^{'}pF_0 - n^{'}pF_0) \tag{A.9}$$

On retrouve effectivement le fait que les multiples doivent être égaux pour que le Dirac vaille $\delta(0)$. La fonction PUI se simplifie alors en une seule somme donnée dans l'équation A.10.

$$\begin{aligned} \Phi_{PUI}\left(\tfrac{p}{q}F_0\right) &= \textstyle\sum_{m=1}^{N_{min}} 1 = N_{min} \\ N_{min} &= \min\left[E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right), E\left(\frac{N_h}{p}\right)\right] \end{aligned} \tag{A.10}$$

La valeur de $M(F_c)$ peut aussi être simplifiée sous certaines conditions. Supposons que la fréquence de coupure F_{MAX} soit fixée à la fréquence de la dernière harmonique du spectre d'amplitude. Cette valeur vaut donc dans l'exemple $F_{MAX} = N_h F_0$ et représente la fréquence minimale sans perte d'information dans le spectre d'amplitude. La valeur de $M(F_c)$ est donnée dans l'équation A.11.

$$M\left(\frac{p}{q}F_0\right) = E\left(\frac{F_{MAX}}{\frac{p}{q}F_0}\right) = E\left(\frac{N_h F_0}{\frac{p}{q}F_0}\right) = E\left(\frac{qN_h}{p}\right) \tag{A.11}$$

N_{min} peut alors être simplifié dans l'équation A.12.

$$N_{min} = \min\left[E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right), E\left(\frac{N_h}{p}\right)\right] \tag{A.12}$$

Or, une propriété intéressante de la partie entière est décrite dans l'équation A.13.

$$\forall (n,x) \in \mathbb{N}^* \times \mathbb{R}, E\left(\frac{1}{n}E(nx)\right) = E(x) \tag{A.13}$$

En posant $n = q$ et $x = N_h/p$ et en utilisant l'équation A.13, N_{min} se simplifie encore selon l'équation A.14.

$$N_{min} = \min\left[E\left(\frac{N_h}{p}\right), E\left(\frac{N_h}{p}\right)\right] = E\left(\frac{N_h}{p}\right) \tag{A.14}$$

L'équation A.15 donne la valeur théorique finale de la fonction PUI dans notre cas idéal. L'amplitude des pics (p,q) ne dépend que de p. Dans les autres cas, la fonction de pitch est nulle. p et q sont premiers entre eux et non nuls.

$$\Phi_{PUI}\left(\frac{p}{q}F_0\right) = E\left(\frac{N_h}{p}\right) \tag{A.15}$$

A.2 PUF

Après le développement détaillé du PUI, la preuve théorique de la fonction PUF est plus simple. Soit N_d le nombre de dents du PUF. Comme pour le PUI, la fonction PUF est nulle si F_c ne vaut pas $(p/q)F_0$ avec p, q entiers et premiers entre eux. Le principe de calcul de la fonction PUF ne change pas de celle du PUI hormis pour $M(F_c)$ qui disparaît au profit de N_d puisque le nombre de dents n'est plus infini mais fixe. Il suffit de reprendre l'équation A.10 et de remplacer $M(Fc)$ par N_d. L'équation A.16 donne le résultat final de la fonction PUF.

$$\Phi_{PUF}\left(\frac{p}{q}F_0\right) = \min\left[E\left(\frac{N_h}{p}\right), E\left(\frac{N_d}{q}\right)\right] \tag{A.16}$$

A.3 PDI

Les dents d'un PDI suivent une fonction de décroissance notée $d(k)$ qui est fonction du numéro k de la dent. L'équation A.17 donne la formule théorique d'un PDI.

$$PDI(F_c, f) = \sum_{k=1}^{M(F_c)} d(k)\delta(f - kF_c) \tag{A.17}$$

La fonction PDI est donnée dans l'équation A.18.

$$\Phi_{PDI}(F_c) = \sum_{k=1}^{M(F_c)} d(k)|S(kF_c)| \tag{A.18}$$

De la même façon que pour un PUI, la fonction PDI est non nulle si avec p, q entiers non nuls et premiers entre eux. L'équation A.19 détaille le calcul de la fonction PDI si $F_c = (p/q)F_0$.

$$\Phi_{PDI}\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{M\left(\frac{p}{q}F_0\right)} d(k) \sum_{n=1}^{N_h} \delta\left(k\frac{p}{q}F_0 - nF_0\right) \tag{A.19}$$

En déroulant le calcul de la même manière que pour le PUI (cf. équations A.8 et A.9), la fonction PDI devient l'équation A.20.

$$\Phi_{PDI}\left(\frac{p}{q}F_0\right) = \sum_{k'=1}^{E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right)} d(k'q) \sum_{n=1}^{E\left(\frac{N_h}{p}\right)} \delta\left(k'q\frac{p}{q}F_0 - n'F_0\right) \tag{A.20}$$

Avec les mêmes conditions utilisée lors de la démonstration du PUI, la fonction PDI se simplifie sous la forme donnée dans l'équation A.21.

$$\begin{aligned} \Phi_{PDI}\left(\tfrac{p}{q}F_0\right) &= \textstyle\sum_{k=1}^{N_{min}} d(kq) \\ N_{min} &= \min\left[E\left(\frac{M\left(\frac{p}{q}F_0\right)}{q}\right), E\left(\frac{N_h}{p}\right)\right] = E\left(\frac{N_h}{p}\right) \end{aligned} \tag{A.21}$$

Dans l'exemple du chapitre 4, la décroissance est en racine inverse. L'équation A.22 donne la formulation de la fonction PDI. p et q sont premiers entre eux et non nuls.

$$\Phi_{PDI}\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{E\left(\frac{N_h}{p}\right)} \frac{1}{\sqrt{kq}} \tag{A.22}$$

A.4 PDF

Le nombre de dents d'un PDF est noté N_d. La démonstration de la fonction PDF suit la même logique que celle suivie entre le PUI et le PDF. Seul N_{min} change de valeur dans l'équation A.21.

L'équation A.23 donne la nouvelle valeur de N_{min} qui dépend à présent de N_d.

$$N_{min} = \min\left[E\left(\frac{N_h}{p}\right), E\left(\frac{N_d}{q}\right)\right] \tag{A.23}$$

La fonction PDF générale dépendant de la fonction décroissante $d(k)$ est décrite par l'équation A.24. N_{min} est donné dans l'équation A.23.

$$\Phi_{PDF}\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{N_{min}} d(kq) \tag{A.24}$$

Avec une fonction de décroissance en racine inverse, la formulation théorique du PDF est décrite dans l'équation A.25. p et q sont premiers entre eux et non nuls.

$$\Phi_{PDF}\left(\frac{p}{q}F_0\right) = \sum_{k=1}^{N_{min}} \frac{1}{\sqrt{kq}} \tag{A.25}$$

A.5 PDN

Un PDN d'ordre Δ est défini par la combinaison linéaire de l'équation A.26.

$$PDN(\Delta, F_c, f) = \frac{\Delta}{\Delta - 1}PUI(F_c, f) - \frac{1}{\Delta - 1}PUI\left(\frac{F_c}{\Delta}, f\right) \tag{A.26}$$

L'opérateur de produit scalaire étant linéaire, la fonction de pitch d'un $PDN\Delta$ respecte donc la même combinaison linéaire. La formule théorique est donnée dans l'équation A.27. La fonction PUI en F_c est déjà connue de l'équation A.15, il ne reste qu'à calculer la fonction PUI en Fc/Δ.

$$\Phi_{PDN\Delta}(F_c) = \frac{\Delta}{\Delta - 1}\Phi_{PUI}(F_c) - \frac{1}{\Delta - 1}\Phi_{PUI}\left(\frac{F_c}{\Delta}\right) \tag{A.27}$$

Lorsque $F_c = (p/q)F_0$, l'équation précédente peut se réécrire sous la forme de l'équation A.28.

$$\Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) = \frac{\Delta}{\Delta - 1}\Phi_{PUI}\left(\frac{p}{q}F_0\right) - \frac{1}{\Delta - 1}\Phi_{PUI}\left(\frac{p}{q\Delta}F_0\right) \tag{A.28}$$

On considère que p est différent de $p^{'}\Delta$ ($p^{'}$ entier non nul). Ceci équivaut au fait que p n'est pas multiple de Δ. Le rapport $(p/q\Delta)$ du second terme de la somme est donc irréductible car (p/q) l'est par définition et (p/Δ) l'est aussi par l'hypothèse qui vient d'être posée. La fonction $PDN\Delta$ est donnée par l'équation A.29 avec $P = p$ et $Q = q\Delta$ et (P/Q) irréductible.

$$\begin{array}{ll} \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) & = \frac{\Delta}{\Delta-1}\Phi_{PUI}\left(\frac{p}{q}F_0\right) - \frac{1}{\Delta-1}\Phi_{PUI}\left(\frac{P}{Q}F_0\right) \\ \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) & = \frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p}\right) - \frac{1}{\Delta-1}E\left(\frac{N_h}{P}\right) = E\left(\frac{N_h}{p}\right) = \Phi_{PUI}\left(\frac{p}{q}F_0\right) \end{array} \tag{A.29}$$

On considère maintenant que p est égal à $p^{'}\Delta$ ($p^{'}$ entier non nul). Le rapport $(p/q\Delta)$ du second terme de la somme est réductible et vaut $(p'\Delta/q\Delta) = (p^{'}/q)$. La question est de savoir si $(p^{'}/q)$ est irréductible.

Le théorème de Bachet-Bezout (équation A.30) dit que si deux nombres entiers a et b sont premiers entre eux (équivalent à (a/b) irréductible) alors :

$$\exists(x,y) \in \mathbb{Z}^2, xa + yb = 1 \tag{A.30}$$

En transposant l'équation A.30 aux entiers p et q, on obtient l'équation A.31.

$$\exists(x,y) \in \mathbb{Z}^2, xp + yq = 1 \tag{A.31}$$

Il suffit de remplacer la valeur de p par $p^{'}\Delta$ pour obtenir l'équation A.32 suivante :

$$\exists(x,y) \in \mathbb{Z}^2, xp^{'}\Delta + yq = 1 \tag{A.32}$$

A présent, posons $X = x\Delta$ et $Y = y$. X comme Y sont des entiers relatifs. L'équation A.33 découle donc logiquement et prouve que la fraction $(p^{'}/q)$ est irréductible.

$$\exists(x,y) \in \mathbb{Z}^2, Xp^{'} + Yq = 1 \tag{A.33}$$

L'équation A.34 donne la forme de la fonction de pitch dans l'hypothèse où p est multiple de Δ et vaut $p^{'}\Delta$.

$$\begin{array}{ll} \Phi_{PDN\Delta}\left(\frac{p^{'}\Delta}{q}F_0\right) & = \frac{\Delta}{\Delta-1}\Phi_{PUI}\left(\frac{p^{'}\Delta}{q}F_0\right) - \frac{1}{\Delta-1}\Phi_{PUI}\left(\frac{p^{'}\Delta}{q\Delta}F_0\right) \\ \Phi_{PDN\Delta}\left(\frac{p^{'}\Delta}{q}F_0\right) & = \frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p^{'}\Delta}\right) - \frac{1}{\Delta-1}E\left(\frac{N_h}{p^{'}}\right) \end{array} \tag{A.34}$$

Une des propriétés de la partie entière est donnée dans l'équation A.35.

$$\forall(n,x) \in \mathbb{N}^* \times \mathbb{R}, 0 \leq E(nx) - nE(x) \leq n-1 \tag{A.35}$$

Cette inégalité se retrouve dans l'équation A.34 dans laquelle il est posé que $n = \Delta$ et $x = (N_h/p^{'}\Delta)$. L'équation A.36 donne un encadrement de la valeur de $\Phi_{PDN\Delta}$.

$$\begin{array}{l} 0 \leq E\left(\Delta\frac{N_h}{p^{'}\Delta}\right) - \Delta E\left(\frac{N_h}{p^{'}\Delta}\right) \leq \Delta - 1 \\ 0 \leq \frac{1}{\Delta-1}E\left(\frac{N_h}{p^{'}}\right) - \frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p^{'}\Delta}\right) \leq 1 \\ -1 \leq \frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p^{'}\Delta}\right) - \frac{1}{\Delta-1}E\left(\frac{N_h}{p^{'}}\right) \leq 0 \end{array} \tag{A.36}$$

Le résultat final est donné équation A.37. C'est bien le même donné dans le chapitre 4. $p^{'}$ est un entier naturel non nul.

$$\begin{array}{ll} si \frac{p}{q} = \frac{p^{'}\Delta}{q} & -1 \leq \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) \leq 0 \\ si \frac{p}{q} \neq \frac{p^{'}\Delta}{q} & \Phi_{PDN\Delta}\left(\frac{p}{q}F_0\right) = \Phi_{PUI}\left(\frac{p}{q}F_0\right) \end{array} \tag{A.37}$$

A.6 PDM

Un PDM d'ordre Δ est défini par la combinaison linéaire de l'équation A.38.

$$PDM\Delta(F_c, f) = \frac{\Delta}{\Delta - 1}[PUI(F_c, f) - PUI(\Delta F_c, f)] \tag{A.38}$$

L'opérateur de produit scalaire étant linéaire, la fonction de pitch d'un $PDM\Delta$ suit donc la même combinaison linéaire. La formule théorique est donnée dans l'équation A.39. La fonction PUI en F_c est déjà connue de l'équation A.15, il ne reste qu'à calculer la fonction PUI en ΔF_c.

$$\Phi_{PDM\Delta}(F_c) = \frac{\Delta}{\Delta - 1}[\Phi_{PUI}(F_c) - \Phi_{PUI}(\Delta F_c)] \tag{A.39}$$

L'équation A.40 montre la valeur de la fonction $PDM\Delta$ lorsque F_c vaut $(p/q)F_0$.

$$\Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) = \frac{\Delta}{\Delta - 1}\left[\Phi_{PUI}\left(\frac{p}{q}F_0\right) - \Phi_{PUI}\left(\Delta\frac{p}{q}F_0\right)\right] \tag{A.40}$$

On considère que q est différent de $q^{'}\Delta$ ($q^{'}$ entier non nul). Ceci équivaut au fait que q n'est pas multiple de Δ. Le rapport $(\Delta p/q)$ du second terme de la somme est donc irréductible car (p/q) l'est par définition et (Δ/q) l'est aussi par l'hypothèse qui vient d'être posée. La fonction $PDM\Delta$ est donnée par l'équation A.41 avec $P = \Delta p$ et $Q = q$ et (P/Q) irréductible.

$$\Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) = \frac{\Delta}{\Delta - 1}\left[E\left(\frac{N_h}{p}\right) - E\left(\frac{N_h}{\Delta p}\right)\right] \tag{A.41}$$

Cette inégalité se retrouve dans l'équation A.42 dans laquelle n vaut Δ et x vaut $(N_h/p\Delta)$.

$$\begin{array}{c}
0 \leq E\left(\Delta\frac{N_h}{p\Delta}\right) - \Delta E\left(\frac{N_h}{p\Delta}\right) \leq \Delta - 1 \\
-E\left(\frac{N_h}{p}\right) \leq -\Delta E\left(\frac{N_h}{p\Delta}\right) \leq -E\left(\frac{N_h}{p}\right) \\
-\frac{1}{\Delta-1}E\left(\frac{N_h}{p}\right) \leq -\frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p\Delta}\right) \leq 1 - \frac{1}{\Delta-1}E\left(\frac{N_h}{p}\right) \\
\frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p}\right) - \frac{1}{\Delta-1}E\left(\frac{N_h}{p}\right) \leq \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) \leq \frac{\Delta}{\Delta-1}E\left(\frac{N_h}{p}\right) + 1 - \frac{1}{\Delta-1}E\left(\frac{N_h}{p}\right) \\
E\left(\frac{N_h}{p}\right) \leq \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) \leq E\left(\frac{N_h}{p}\right) + 1 \\
\Phi_{PUI}\left(\frac{p}{q}F_0\right) \leq \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) \leq \Phi_{PUI}\left(\frac{p}{q}F_0\right) + 1
\end{array} \tag{A.42}$$

L'équation A.42 encadre donc la fonction $PDM\Delta$ par la fonction PUI. On considère maintenant que q est égal à $q^{'}\Delta$ ($q^{'}$ entier non nul) ce qui équivaut à ce que q soit multiple de Δ. Par le même théorème de Bachet-Bezout utilisé pour la démonstration du PDN (cf. équation A.34), on montre que dans ce cas, le rapport $(p/q^{'})$ est irréductible. L'équation A.43 montre alors la valeur de la fonction $PDM\Delta$ dans ce cas particulier.

$$\begin{array}{ll}
\Phi_{PDM\Delta}\left(\frac{p}{q^{'}\Delta}F_0\right) & = \frac{\Delta}{\Delta-1}\left[\Phi_{PUI}\left(\frac{p}{q^{'}\Delta}F_0\right) - \Phi_{PUI}\left(\frac{p\Delta}{q^{'}\Delta}F_0\right)\right] \\
\Phi_{PDM\Delta}\left(\frac{p}{q^{'}\Delta}F_0\right) & = \frac{\Delta}{\Delta-1}\left[E\left(\frac{N_h}{p}\right) - E\left(\frac{N_h}{p}\right)\right] = 0
\end{array} \tag{A.43}$$

Le résultat final est donné équation A.44. C'est bien celui proposé dans le chapitre 4.

$$\begin{array}{ll} si \frac{p}{q} = \frac{p}{q'\Delta} & \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) = 0 \\ si \frac{p}{q} \neq \frac{p}{q'\Delta} & \Phi_{PUI}\left(\frac{p}{q}F_0\right) <= \Phi_{PDM\Delta}\left(\frac{p}{q}F_0\right) <= \Phi_{PUI}\left(\frac{p}{q}F_0\right) + 1 \end{array} \tag{A.44}$$

Annexe B

Article RJCP 07 – Le suivi de F_0

Cet article de référence bibliographique [Signol, 2007] a été présenté à la Rencontre des Jeunes Chercheurs en Parole RJCP 2007 à Paris en juillet 2007. Il traite du suivi de F_0 multiples avec une programmation dynamique.

Suivi d'un flux de F_0 par programmation dynamique dans un mélange monaural de signaux de parole

François Signol

LIMSI – CNRS
Bâtiment 508, Université Paris Sud XI – 91400 Orsay, FRANCE
Tél. : +33 (0)1 69 85 81 25 – Fax : +33 (0)1 69 85 80 88
Courriel : francois.signol@limsi.fr

ABSTRACT

In this paper we propose a F_0 trajectories tracking method designed to work with monaural speech mixtures. A two speaker's speech mixture is firstly pre-processed in order to extract some F_0 candidates in each signal's frame. A dynamic programming framework is then presented to generate each speaker's F_0 streams. The dynamic programming aims at selecting and clustering the F_0 candidates given by the pre-processing

1. INTRODUCTION

L'effet de Cocktail Party (ECP) [Che53] constitue un défi majeur du domaine CASA (Computational Auditory Scene Analysis). L'ECP est une faculté naturelle qui se manifeste en écoute monaurale et qui nous permet d'isoler et de suivre à volonté un flux sonore (un bruit de klaxon, une partition de piano, un locuteur particulier, etc.) lorsque celui-ci se trouve dans un environnement sonore complexe (rue, orchestre, restaurant, etc.) c'est-à-dire mélangé à d'autres flux sonores. Cette faculté ségrégative repose sur des mécanismes perceptifs de focalisation attentionnelle qui intègrent des traitements ascendants (bottom-up) et descendants (top-down), ces derniers étant à considérer dans le cadre d'une situation et d'un comportement donnés. Les traitements ascendants s'opèrent à partir des caractéristiques acoustiques extraites d'un signal sonore. Les traitements descendants font intervenir des indices cognitifs issus de la mémoire sonore d'un individu ou des intérêts de l'individu placé dans un contexte sonore particulier. Bien qu'il soit aisé pour un être humain normo-entendant de réaliser cette tâche, il s'avère très complexe de la modéliser.

Les études psycho-acoustiques menées dans le domaine ASA (Auditory Scene Analysis) ont mis en évidence un certain nombre d'indices acoustiques impliqués dans la ségrégation de flux sonore. Le voisement et la fréquence fondamentale (notée F_0 ou pitch) figurent comme les plus importants. Viennent ensuite d'autres indices tels que les instants de départ (onset) ou d'extinction (offset) d'une source sonore [Hu07], la localisation, la modulation d'amplitude ou encore l'enveloppe spectrale d'un segment de parole voisé [Roa07]. Les systèmes existants de séparation de parole en enregistrement monauraux utilisent différentes méthodes temporelles, fréquentielles ou spectro-temporelles afin d'extraire ces indices pour ensuite les regrouper en flux cohérents [Bar06].

Nous nous sommes intéressés à une forme restreinte du problème Cocktail Party visant à séparer deux signaux de parole mélangés artificiellement en contexte monaural. Nos travaux se sont d'abord concentrés sur l'élaboration d'algorithmes robustes de calcul de F_0. L'estimation de F_0, même pour des signaux mono-locuteur, est une tâche difficile qui fait des erreurs. Dans un cadre d'application multi-locuteurs, les algorithmes standards d'extraction de F_0 sont soit inopérants soit beaucoup moins performants. Nous avons travaillé sur une approche « multiF_0 » présentée en partie 2 conçue pour donner à chaque segment temporel (ou trame) un certain nombre de candidats F_0 potentiels. La partie 3 présente un algorithme de programmation dynamique (PD), algorithme aussi utilisé dans [Boe93] et [Eve06]. Son rôle est de reconstruire les flux de F_0 propres à chaque locuteur en s'appuyant sur des caractéristiques intrinsèques de la parole. Il est présenté dans un cadre mono-locuteur par souci de clarté. La partie 4 étend ce modèle dans le cadre bi-locuteurs puis multi-locuteurs.

2. EXTRACTION DES CANDIDATS F_0

Tous les algorithmes d'extractions de pitchs peuvent se regrouper en trois familles de méthodes : temporelle, spectrale et spectro-temporelle [Che06]. Quelle que soit l'approche utilisée, toutes se heurtent à un même type d'erreur : les erreurs sous-harmoniques et sur-harmoniques pour lesquelles les algorithmes ne donnent pas le F_0 correct mais une fraction rationnelle de celui-ci. Or, l'extraction de F_0 n'est qu'une partie de la chaîne de traitement totale. Il convient de rendre cette étape la meilleure possible pour éviter d'avoir à pallier ses défauts par des traitements en aval qui pourraient être évités ou allégés. La chaîne de traitement présentée figure 1 considère ce problème de sous et sur-harmoniques et propose de réduire son impact dès l'extraction des candidats F_0 d'un mélange de parole.

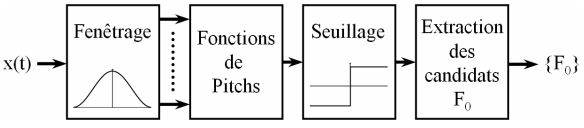

Figure 1 : Chaîne de traitement pour l'extraction de l'ensemble des candidats pitchs $\{F_0\}$

Le signal de parole x(t) est d'abord fenêtré (hanning) puis la fonction de pitch PP_t (cf. 2.1) est calculée pour chacune des trames t du signal. Le continuum temporel de l'ensemble des PP_t forme une fonction de trajectoire notée PT dont une illustration est donnée figure 3. PT est ensuite utilisée pour seuiller chaque fonction de pitch. Ce seuillage permet de garder uniquement les zones fréquentielles de forte amplitude. Puis, trame à trame, les principaux pics de la fonction de pitch sont sélectionnés en tant que candidats F0 potentiels.

2.1. Fonction de pitch et calcul multi-F_0

Notre méthode de calcul de pitch est une méthode de peigne spectral, inspirée des travaux de Schroeder [Sch68], qui consiste à calculer l'intercorrélation entre le spectre d'amplitude du signal et un peigne spectral composé d'une série de dents régulièrement espacées en fréquence. La fonction de pitch résulte de cette intercorrélation calculée pour toutes les valeurs que peut prendre le pitch dans un intervalle prédéfini (ex. 70 à 700 Hz). En mono-pitch la valeur recherchée se trouve à la fréquence du pic maximum de la fonction de pitch. Cependant cette technique ne suffit pas à éliminer complètement les erreurs harmoniques. Pour en réduire le nombre nous avons adjoint aux dents positives du peigne simple des dents négatives qui neutralisent une partie notable des erreurs (cf. figure 2). Cette technique améliore les taux d'erreur en mono-pitch, et réduit le nombre de candidats en multi-pitchs. Pour plus de détails et pour une évaluation sur ce type de peigne cf. [Lie07].

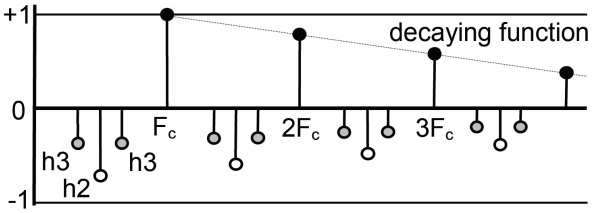

Figure 2 : Peigne fréquentiel alterné. Le peigne simple ne comporte que les dents positives.

2.2. Seuillage

Le seuillage de chaque PP_t détermine le voisement ou non de la trame t. Il détermine aussi si la trame possède des candidats F_0 potentiels ou non. Pour éviter des décisions locales trames à trames erronées, le seuillage intègre une dimension temporelle en utilisant la fonction PT. Pour la fonction PP_t de la trame t, on recherche le maximum absolu de PT sur un certain environnement temporel autour de la trame t. Cet environnement temporel T_{env} est donné en nombre de trames. Typiquement, on fixera le nombre de trames nécessaires pour un environnement de 500ms soit T_{env}=50 pour un pas temporel de fenêtrage de 10ms. Le seuil s_t (cf. équation 1) de la trame est alors le résultat du produit entre ce maximum local et une constante α (typiquement 0.1 soit -20dB).

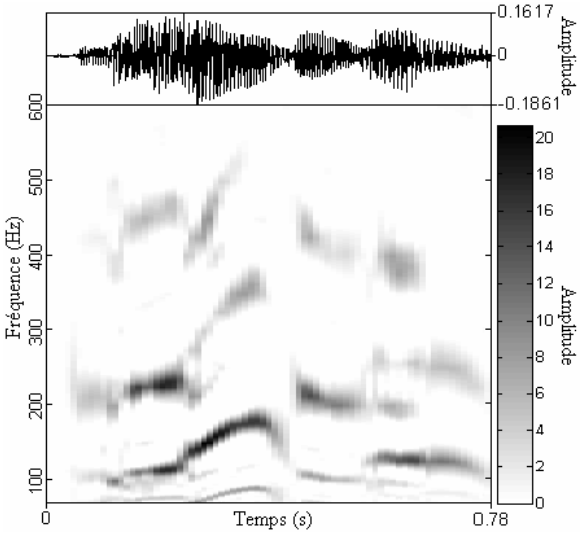

Figure 3 : PT (au-dessous) obtenu sur un extrait sonore bi-locuteurs masculin/féminin du corpus de Keele « The north wind » (au-dessus). Les deux trajectoires des F_0 de chaque locuteur apparaissent en noir. On remarque aussi la présence des sous ou sur-harmoniques de ces trajectoires.

$$s_t = \alpha . \underset{t\in[t-T_{env}/2,t+T_{env}/2],f\in[F_{0MIN},F_{0MAX}]}{\arg\max} \left[PP_t(f)\right] \quad \textbf{Eq. 1}$$

2.3. Extraction des candidats F_0

L'extraction des candidats F_0 consiste à chercher dans chaque PP_t seuillé les N_C (ex. 12) plus grands pics. On calcule le maximum de la courbe :

$$C_t = \underset{f\in[F_{0MIN},F_{0MAX}]}{\arg\max} \left[PP_t(f)\right] \quad \textbf{Eq. 2}$$

On supprime alors de PP_t le pic d'amplitude correspondant à la fréquence candidate C_t en annulant à droite et à gauche toutes les valeurs de PP_t jusqu'à la fin de la décroissance du pic. On contraint toutefois l'annulation à valoir une certaine fenêtre fréquentielle. Cet environnement fréquentiel est proportionnel à C_j et a été fixé à plus ou moins 3% ce qui correspond à un demi ton. Ceci a pour conséquence de limiter la résolution de l'extraction à un demi ton. Si deux candidats pitchs sont situés à moins d'un demi ton d'écart, un seul des deux sera détecté. On itère ce processus (équation 2 + annulation) jusqu'à ce qu'on ait obtenu les N_C candidats ou bien que la courbe soit nulle, tout ayant été supprimé.

3. SUIVI DES TRAJECTOIRES DE F_0

Chaque trame possède maintenant une liste de candidats F_0 rangés par ordre d'amplitude. Mais cet ordre ne correspond pas à la trajectoire correcte du F_0 d'un locuteur. En mono-locuteur par exemple, le premier candidat F_0 n'est pas forcément correct mais peut être un candidat harmonique. Ceci est d'autant plus vrai lorsqu'on se place dans un contexte multi-locuteurs où les candidats harmoniques d'un flux et les candidats F_0

de l'autre flux peuvent se juxtaposer (cf. figure 3). Ce traitement directement en aval de la chaîne de traitement de la partie 2 doit reconstruire les trajectoires de F_0 en sélectionnant les bons candidats parmi l'ensemble des périodicités calculées. Une PD est proposée pour réaliser ce regroupement des F_0 en flux composants le mélange de locuteurs. Cette PD s'appuie sur deux connaissances dont on dispose à propos du signal de parole. Il s'agit du nombre de locuteurs que l'on souhaite rechercher et une double contrainte de continuité qui permet de générer les coûts de passage entre états de la PD.

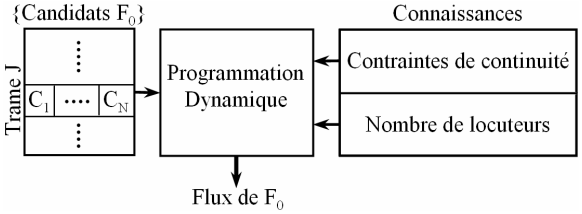

Figure 4 : Schéma fonctionnel de la programmation dynamique.

Le signal de parole est continu temporellement et fréquentiellement. Il est donc très improbable que des discontinuités temporelles apparaissent dans l'évolution des F_0. De même, il est peu probable que les transitions entre segments voisés et segments non-voisés soient trop rapides dans le temps. Ces caractéristiques intrinsèques de la parole permettent de quantifier la continuité d'un flux de F_0. La PD utilise ces informations pour reconstruire les flux de F_0 ayant les continuités temporelles et fréquentielles les plus marquées (cf. figure 4).

Plaçons nous dans un cadre de signaux mono-locuteur pour expliquer le principe général de la PD. Nous verrons dans la partie 4 que le modèle posé ici peut s'étendre facilement aux cas bi-locuteurs puis multi-locuteurs. Notons N_M^t le nombre de candidats voisés de la trame t et N_{MAX} le nombre maximal de candidats voisés attribuables pour une trame. Ainsi $N_M^t \leq N_{MAX}, \forall t \in [1,T]$. Représentons le problème de suivi d'un flux de F_0 parmi un ensemble de chemins possibles tel qu'il est communément posé dans le cadre d'une PD. Un segment temporel ou trame devient un ensemble d'états. Cette décomposition en trames et états est illustrée dans la figure 5 où les cercles sont les différents états (ordonnées) d'une trame (abscisses). Chaque état contient un candidat pitch. Le pitch 0 représente l'état non voisé d'une trame. Chaque état d'une trame t est relié à un autre état de la trame t+1 par une transition représentée par des flèches sur la figure 5. Un coût de passage est associé à chacune des transitions. Ce coût symbolise la facilité (coût faible) ou au contraire la difficulté (coût élevé) d'une transition.

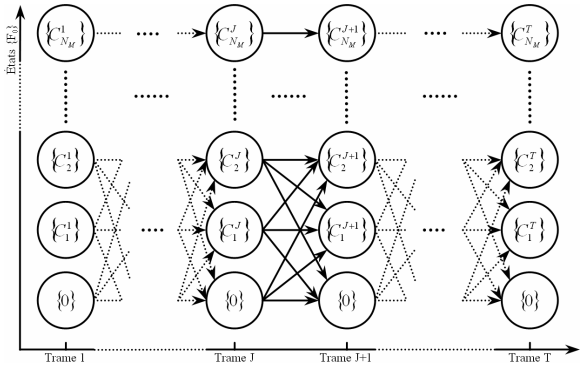

Figure 5 : Modèle de programmation dynamique dans le cadre d'un suivi de F_0 mono-locuteur.

Cette modélisation permet de générer un nombre de chemins distincts Ω_P qui vaut :

$$\Omega_p = \prod_{t=1}^{T} \left(N_M^t + 1\right) \leq \left(N_{MAX} + 1\right)^T$$

Si on associe les coûts de transition à des critères de continuité, on peut rechercher parmi les Ω_p chemins celui dont le cumul des coûts de transition est minimal. Le problème ainsi posé se restreint à un algorithme standard de calcul de meilleur chemin. L'évolution de F_0 est soumise à la continuité. Le coût associé à cette continuité peut être modélisé par la valeur absolue de l'écart relatif entre la valeur de F_0 pour la trame t et sa valeur dans la trame t+1 (cf. table 1). La continuité s'applique aussi aux transitions trames voisées/trames non voisées. Un changement d'état de voisement a un coût. Nommons C_{vnv} cette constante fixée empiriquement à 0.2. La table 1 récapitule ces coûts en fonction des transitions entre les différents états possibles.

Trame t / Trame t+1	0	C_x^{t+1}
0	0	C_{vnv}
C_x^t	C_{vnv}	$\left\|C_x^{t+1} - C_x^t\right\| / C_x^t$

Table 1 : Les coûts de passage entre états de deux trames successives dans le cadre mono-locuteur.

4. DANS LE CADRE MULTI-LOCUTEURS

La modélisation du suivi de F_0 mono-locuteur est extensible au cas bi-locuteurs. Chaque état ne contiendra plus une valeur unique C_x^t mais un couple de valeur $\{C_x^t, C_y^t\}$ dans lequel l'un ou les deux candidats peuvent être nuls. Pour conserver l'ordre des candidats F_0 indispensable pour générer les deux trajectoires de pitchs, il faut considérer que l'état $\{0, C_x^t\}$ est différent de l'état $\{C_x^t, 0\}$. Ainsi, le coût de passage D entre l'état $\{C_x^t, C_y^t\}$ et $\{C_x^{t+1}, C_y^{t+1}\}$ peut s'écrire sous la forme :

$$D = F\left(C_x^t, C_x^{t+1}\right) + F\left(C_y^t, C_y^{t+1}\right)$$

Où F n'est autre que la fonction de coût définie dans la table 1 du contexte mono-locuteur. La table 2 ci-dessous illustre tous les types de changements d'états et leurs coûts de passage associés.

Tr t / Tr t+1	$\{0,0\}$	$\{0, C_x^{t+1}\}$	$\{C_x^{t+1}, 0\}$	$\{C_x^{t+1}, C_y^{t+1}\}$
$\{0,0\}$	0	C_{vnv}	C_{vnv}	$2C_{vnv}$
$\{0, C_x^t\}$	C_{vnv}	d_1	$2C_{vnv}$	d_2
$\{C_x^t, 0\}$	C_{vnv}	$2C_{vnv}$	d_1	d_3
$\{C_x^t, C_y^t\}$	$2C_{vnv}$	d_4	d_5	d_6

Avec

$$d_1 = \frac{\left|C_x^{t+1} - C_x^t\right|}{C_x^t} \qquad d_2 = c_{vnv} + \frac{\left|C_y^{t+1} - C_x^t\right|}{C_x^t}$$

$$d_3 = d_1 + c_{vnv} \qquad d_4 = c_{vnv} + \frac{\left|C_x^{t+1} - C_y^t\right|}{C_y^t}$$

$$d_5 = \frac{\left|C_x^{t+1} - C_x^t\right|}{C_x^t} + c_{vnv}$$

$$d_6 = \frac{\left|C_x^{J+1} - C_x^J\right|}{C_x^J} + \frac{\left|C_y^{J+1} - C_y^J\right|}{C_y^J}$$

Table 2 : Les coûts de passage entre états de deux trames successives dans le cadre bi-locuteurs.

La généralisation pour un suivi de F_0 multi-locuteurs découle naturellement de ce modèle puisque les coûts de transitions pour le contexte L-locuteurs dépendent directement des coûts définis pour le contexte (L-1)-locuteurs. Les états seront l'ensemble des n-uplets possibles pour chaque trame temporelle. La complexité de calcul de notre modèle est exponentielle. Par exemple un signal L-locuteurs impliquera un nombre d'états Ω_e pour une trame t qui vaut : $\Omega_e(J) = \left(N_c^t + 1\right)^L$

Le nombre de coûts Ω_c à calculer par transition entre la trame t et la trame t+1 vaut :

$$\Omega_c(t) = \Omega_e(t).\Omega_e(t+1)$$

5. CONCLUSIONS ET PERSPECTIVES

La première partie de ce papier a présenté une méthode originale d'extraction de F_0 dans des signaux de parole multi-locuteurs. Dans la deuxième partie nous avons développé un modèle théorique de suivi de trajectoire de pitchs basé sur une programmation dynamique. Nous avons montré à partir d'un exemple mono-locuteur que ce modèle est générique et peut s'étendre facilement au cadre d'application multi-locuteurs. Deux pistes d'explorations découlent de ce travail. La première résulte d'un défaut du modèle qui dans l'état actuel ne permet pas de prédire si le suivi est robuste à des croisements de trajectoires F_0. L'autre piste de travail déjà en cours de réalisation consiste à évaluer les performances de notre modèle sur des bases de données de parole F_0-annotées telles que Keele et Bagshaw.

En réponse à la première interrogation, nous travaillons aussi à l'intégration de la continuité fréquentielle dans les coûts de transitions de la PD. En associant un candidat F_0 à son enveloppe spectrale (ensemble des harmoniques de ce F_0) nous pourrons quantifier la continuité spectrale entre deux trames et donc intégrer ce coût aux transitions de la PD. Dans les zones de croisement de trajectoires, ces enveloppes spectrales doivent jouer un rôle non négligeable.

REFERENCES

[Che53] Cherry E.C., (1953), *Some experiments on the recognition of speech, with one and with two ears*. Journal of Acoustic Society of America, 25, pp. 975-979.

[Hu07] Hu G. and Wang D.L., (2007), *Auditory segmentation based on onset and offset analysis*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 396-405.

[Che06] De Cheveigné A., (2006), *Multiple F0 Estimation*, Computational Auditory Scene Analysis, IEEE PRESS, pp. 45-79.

[Eve06] Every M.R. and Jackson P.J.B., (2006), *Enhancement of Harmonic Content of Speech Based on a Dynamic Programming Pitch Tracking Algorithm*, Interspeech Proceedings, 4, pp. 81-84.

[Roa07] Roa S., Bennewitz M. and Behnke S., (2007), *Fundamental Frequency Estimation Based on Pitch-Scaled Harmonic Filtering*, IEEE ICASSP, Vol. 4, pp. 397-400.

[Boe93] Boersma P. (1993), *Accurate Short-Term Analysis of the Fundamental Frequency and The Harmonics-to-Noise Ratio of a Sampled Sound*, IFA Proceedings, 17, pp. 97-110.

[Lie07] Liénard J-S, Signol F. and Barras C., (2007), *Speech fundamental frequency estimation using the Alternate Comb*, Proceedings of Interspeech, in press.

[Bar06] Barker J., Coy A., Ma N. and Cooke M., (2006), *Recent advances in speech fragment decoding techniques.* In Proc. Interspeech 2006, pp. 85-88.

[Sch68] Schroeder M.R., (1968), *Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement*, J. Acoust. Soc. Amer., Vol. 43, pp. 829-834.

Annexe C

Article Interspeech 07 – Le PAL

Cet article de référence bibliographique [Liénard et al., 2007] a été présenté à la conférence internationale INTERSPEECH 2007 à Anvers en août 2007. Il traite de l'estimation de F_0 monopitch par un nouveau peigne nommé Peigne Alterné (PAL) et décrit dans le chapitre 4 paragraphe 4.3.5.

Speech Fundamental Frequency estimation using the Alternate Comb

Jean-Sylvain Liénard, François Signol and Claude Barras

LIMSI-CNRS 91403 Orsay Cedex, France

{jean-sylvain.lienard, francois.signol, claude.barras}@limsi.fr

Abstract

Reliable estimation of speech fundamental frequency is crucial in the perspective of speech separation. We show that the gross errors on F0 measurement occur for particular configurations of the periodic structure to estimate and the other periodic structure used to achieve the estimation. The error families are characterized by a set of two positive integers. The Alternate Comb method uses this knowledge to cancel most of the erroneous solutions. Its efficiency is assessed by an evaluation on a classical pitch database.

Index Terms: F0 estimation, spectral comb, speech separation

1. Introduction

Separating two speech signals mixed in a single channel, although easy for a human listener, proves to be difficult for an automatic processing. Fundamental Frequency F0 is considered as the main usable indices for this task. Thus it is necessary to work out robust Pitch Estimation Algorithms (PEA) able to give satisfactory results even when several voiced signals are mixed (multipitch estimation). However F0 estimation is a difficult, error-prone operation, even when one is certain that there is only one single voice in the signal. A recent review of this problem can be found in [1].

Our objective is to analyze the nature of the errors produced by a PEA and to design a mechanism able to reduce them. The errors can be classified into 3 categories: voicing decision, gross errors and fine errors.

Voicing decision is ambiguous. The phonological point of view demands a binary decision, namely Voiced or UnVoiced, either in the production or in the perception perspective. From the Signal Processing point of view one also uses to consider that a given frame should be voiced or not, although physical reality shows that there is always some progressivity in the signal transition between the voiced and unvoiced states. Thus it is necessary to fix some threshold, above which the frame is declared voiced. It is well known that such a threshold cannot be valid for any kind of speech signal and any situation.

In a voiced frame F0 estimation is performed by a particular function (periodicity indicator) which computes a non-dimensional value for any value Fc comprised between the arbitrary limits F0min and F0max. Periodicity is indicated by the position of a given extremum of this function. The estimation can be biased in two ways. First, the extremum decision may choose a wrong one, which is easy because periodicity indicators often happen to be periodic themselves. This produces what is usually called gross errors. Second, when the system effectively chooses the right extremum, it may produce fine errors, which may have multiple causes: small voice fluctuations, presence of noise, window too narrow or too wide, computational precision. Usually the limit between the two types of errors is fixed at +- 20% of the reference F0, corresponding approximately to +- 3 semitones. Gross errors can happen with any type of periodicity indicator, spectral, temporal or spectro-temporal. In the present study we use a purely spectral method, in the line of [2], [3], [4], among others.

First we explain the principles by which gross errors are formed by the spectral structure that we call Simple Comb. We propose a modification of its structure which reduces some of those errors. The functioning of the new device called Alternate Comb is illustrated with real signals. Then we propose a monopitch evaluation in comparison with a popular autocorrelation-based PEA freely available with the Praat software [5].

2. Origin and structure of the gross errors

Let us consider a spectral function |S| composed of N harmonic peaks, of fundamental frequency F0 and amplitude unity, and a spectral comb C unlimited in frequency, i.e. an infinite series of pulses of height unity and fundamental frequency Fc. There is no spectral component between the peaks. Let us vary Fc.

When Fc = F0 all of the spectral peaks are matched by the N first teeth of the comb (Figure 1), the scalar product of both functions is maximum and equals N. When Fc = 2*F0 there is still another product maximum, equaling the integer part of N/2. Choosing this peak to represent the fundamental frequency of |S| yields an octave error. By proceeding upwards one can see that peaks of decreasing amplitude appear each time that Fc becomes a multiple of F0. These peaks correspond to the harmonic errors of order p = 2, 3 ...

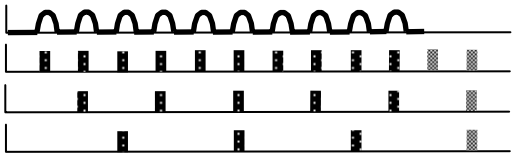

Figure 1: *series of 10 spectral peaks of fundamental frequency F0 (top) and uniform infinite combs of fundamental frequencies Fc = F0, 2*F0 and 3*F0. The matching teeth are painted in dark.*

If we move backwards from the starting position we see easily that we encounter a new peak at Fc=F0/2, although the first tooth does not match any peak (Figure 2). This is the order 2 subharmonic error, actually the sub-octave error. The problem is that, because we use an infinite comb, the scalar product amounts to the same value as for the main peak at Fc=F0.

There is a similar peak at Fc=F0/3, which produces an order 3 subharmonic error. Again, its scalar product equals N. In addition, we see that there is another related peak at Fc=2*F0/3. It produces a second order 3 subharmonic error. Thus we have to use two orders, the harmonic order p and the subharmonic order q, to specify a peak (p, q). The previous peaks are labeled (1, 3) and (2, 3). It is easy to identify other subharmonic peaks such as (1, 4), (2, 4) and (3, 4), (1, 5), (2,

5) etc. As N is limited, the amplitudes of the peaks (p, q) for which p is greater than 1 do not reach the value of the main peak (1, 1). We have to notice that peaks (1, 2) and peaks (2, 4) are two different labels for the same entity and should preferably be designated by the simplest form (1, 2).

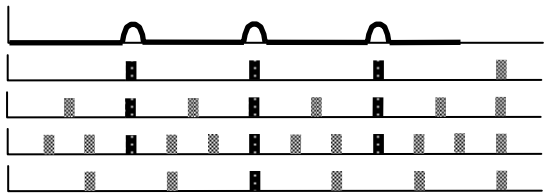

Figure 2: *series of 3 spectral peaks of fundamental frequency F0 (top) and uniform infinite combs of fundamental frequencies Fc=F0, F0/2, F0/3 and 2*F0/3. The matching teeth are painted in dark.*

Finally we observe that the subharmonic peaks observed for Fc<F0 have replicas in all of the intervals between successive multiples of F0. They are characterized by p>q. Their amplitudes are globally decreasing, due to two causes: i) the scalar product tends to take 1/p peaks in the summation when N tends to infinity and ii) N is limited.

The above considerations come very close to the basic notions developed by Schroeder in [2]: period histogram, frequency histogram, Harmonic Product Spectrum. Let us call PitchPeaks (PP) the generalization of the above scalar product as a function of Fc, which differs from HPS mainly by the fact that the products are not expressed in log units. Figure 3 shows the PP function of a physical signal (series of pulses at F0=250 Hz), analyzed by a uniform comb (all teeth equals, infinite).

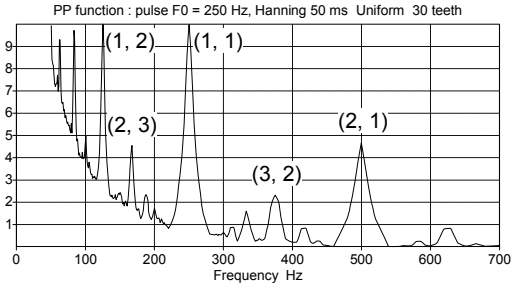

Figure 3: *Uniform Comb applied to a 250 Hz Hanning windowed pulse series. Some of the peaks are labeled with their (p, q) orders.*

3. The Simple Comb

The PP function presented above is prone to gross errors, as it exhibits many peaks having the same maximum value, especially in the region <F0. In order to make the main peak (1, 1) dominate the others there are two solutions. One is to limit the number of teeth, so that when decreasing Fc the set of tooth encompasses a smaller part of the spectrum. The other is to apply a decaying shape to the teeth. Both may be implemented together. Common values are 10 for the number of teeth and 1/m or 1/sqrt(m) for the decaying function (m is the tooth index). Figure 4 shows the same sound as in Figure 3, analysed with a 10-teeth Simple Comb decaying in 1/m. Generally, the subharmonic peaks are somewhat attenuated and become less confusing than the harmonic ones.

There is a problem concerning the unit in which the spectrum module is best expressed in the PP calculation: linear (related to amplitudes), quadratic (related to energy and autocorrelation) or logarithmic (related to the decibel scale). As noticed in [6], as the voiced speech spectrum is globally less intense in the high frequencies, the quadratic units exaggerate the importance of the lowest part of the spectrum, and the logarithmic units gives too much weight to the highest part or to the weakest spectral components. According to our experience the linear units are better adapted to the problem.

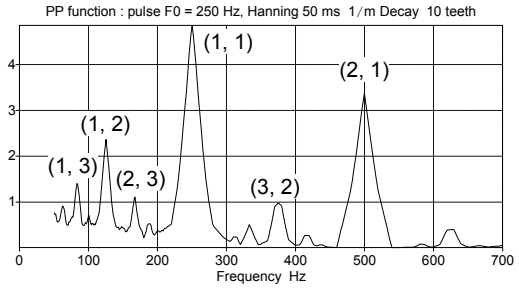

Figure 4: *Simple Comb applied to a 250 Hz Hanning windowed pulse series. Peaks of subharmonic order q>1 are attenuated compared to harmonic peaks q=1*

The Simple Comb, as well as the equivalent methods based on the accumulation of spectral shifts (for instance [4], gives good results, even for telephone voice or in the presence of noise. The implementations differ in several respects: units of spectral magnitude, F0min and F0max limits, number of teeth, decaying function, spectrum pre-processing, selection and accumulation process. These variants aim at reducing the magnitude of the secondary peaks compared to the main one. But no one eliminates them completely. This is not a real drawback in the perspective of single pitch estimation, because by definition there is only one periodicity of interest in the signal. Ensuring the existence of a maximum corresponding to the right periodicity is sufficient.

However, in the perspective of speech separation, reliable multiple pitch estimation is necessary. Mixing two periodic signals of fundamental frequencies F01 and F02 produces in PP two peak families interfering in complex ways. Although one can presume that the main peak represents one of the two periodicities, identifying the other or assessing its absence is a difficult task, for which the pitch estimator has to produce the smallest possible number of reliable candidates.

4. The Alternate Comb

In order to reduce the amplitude of the harmonic peaks we propose the Alternate Comb. To the positive teeth of the simple comb we adjunct some intermediary negative teeth, positioned at the exact frequencies that may produce the harmonic errors (Figure 5).

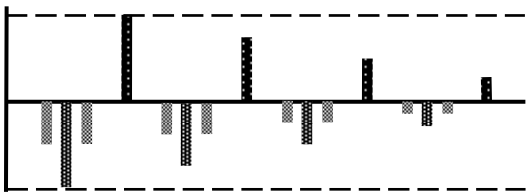

Figure 5: *Alternate Comb. The positive teeth are the same as in the Simple Comb. The negative teeth contribute to reducing the harmonic errors of orders (2, 1) and (3, 1).*

Subtracting from the PP summation the spectral components placed halfway from two successive positive teeth produces a large reduction of the octave error Fc=2*F0. The negative teeth placed at 1/3 and 2/3 of the positive teeth intervals reduce the error at Fc=3*F0. As the optimal height of the negative teeth cannot be computed a priori, weighting coefficients h2, h3 ... hp are attached to each harmonic order. These coefficients are the main parameters of the Alternate Comb. Fixing them to 0 transforms it back into a simple comb. By changing them gradually one can evaluate the impact of the proposed strategy. Figure 6 shows the PP function obtained with the Alternate Comb on the same signal as above.

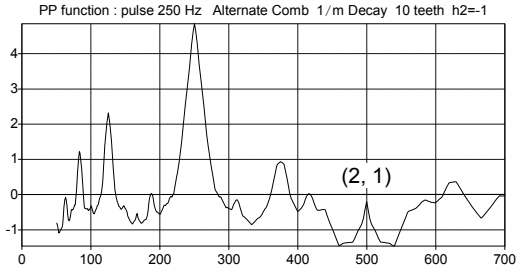

Figure 6: *Alternate Comb applied to a 250 Hz Hanning windowed pulse series. Coefficient h2 (octave error) has been set to 1. As a consequence peak (2, 1) gets cancelled out. Compare to Figure 5.*

The function PP can now take some negative values. In order to ensure the existence of positive peaks the mean value is subtracted. The amplitude of the peak retained as possibly representing F0 is compared to a threshold depending on the maximum surrounding level, within a +- 1 second interval.

Figures 8 shows the function PP obtained on a frame selected in the sum of two speech signals of equal level: /a/ (male voice 120 Hz) and /i/ (female voice 266 Hz). The Alternate Comb was tuned with h2=-0.4 and h3=-0.4. The peak at 600 Hz corresponds to the 5 th harmonic of the first vowel (p=5, q=1). It is not cancelled because the coefficient h5 was not used in this tuning (h5 set to zero).

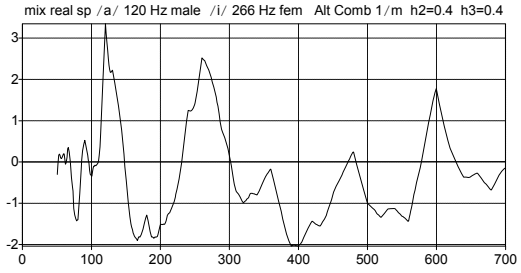

Figure 9: *Alternate Comb applied to a mix of two synthetic vowels, /i/ and /a/ (50 ms Hanning windowed), with their F0 exactly one octave apart (80 and 160 Hz)*

Figure 9 demonstrates the capacity of the Alternate Comb to simultaneously process two synthetic speech signals of equal level, that have F0s exactly at an octave interval. One can observe that octave cancellation does not wipe out the 160 Hz peak. Most of the undesired peaks are strongly attenuated, with the exception of the one located at 640 Hz.

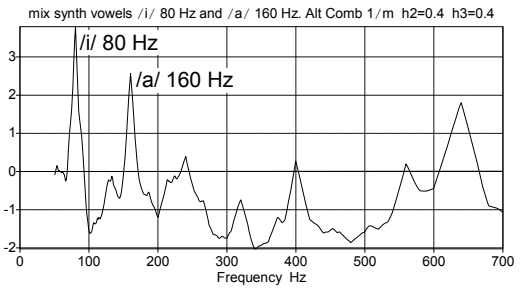

Figure 9: *Alternate Comb applied to a mix of two synthetic vowels, /i/ and /a/ (50 ms Hanning windowed), with their F0 exactly one octave apart (80 and 160 Hz)*

The Alternate Comb method bears some similarities with other published work, particularly [7], where the author implements a processing devoted to the elimination of the octave error. Our method differs in three respects: i) it is based on the analysis of the different types of gross errors and not on considerations related to voice quality; ii) we use linear units in the spectral magnitude computation, and iii) we place our study in the perspective of multiple pitch estimation.

5. Evaluation

For preliminary studies we used speech data extracted from the Speech Separation Challenge [8], in particular 10 sentences (5 males, 5 females) totalling 17 seconds. The tests reported here have been conducted with the Keele database [9], totalling 337.1 seconds of speech uttered by 10 speakers (5 males, 5 females), ie 33710 frames, of which 14936 were considered voiced by the reference algorithm.

We chose to compare several tunings of the Alternate Comb to an algorithm widely used in the speech community. The Praat AC PEA is based on autocorrelation and uses an efficient post processing. Prior to any other measurements, we compared the results given by the same algorithm on the audio signal (reference) and on the egg signal (test). As a result we observed a rather large rate of voicing errors and a rather small rate of gross errors (table 1, first line). This indicates that, as long as the gross error rate remains larger, taking as reference the standard Praat AC algorithm on the audio signal is legitimate.

As indicated above, the results obtained by a given PEA on a given database may differ according to the voicing criterion used. We minimized the corresponding bias by adjusting the voicing threshold so that the undervoicing rate (the PEA tested declares less voiced frames than the reference) is of the same order of magnitude than the overvoicing rate (the PEA tested declares more voiced frames than the reference). We checked that the gross error rates do not vary much if the undervoicing and overvoicing rates are kept within the interval of 2 to 8%.

Our evaluation was not directed towards any rigorous performance comparison with other PEAs, the results of which have been published in several papers such as [6], [7] or [10]. Instead, it aims at investigating the parameters of the Alternate Comb when gradually introducing negative teeth of orders hp (p=2 and p=3) in the Simple Comb. As some parameters are interdependent, the general idea was to seek the best result for each setting of the hp and voicing threshold parameters from a trial set (a part of the whole database

comprising 6147 frames out of 33710). The values given in table 1 were computed from the whole database with those values. All other settings were kept constant across measurements and algorithms. In particular the window width was fixed at 40 ms and the F0 interval was fixed at 75-600 Hz, which are the default values of the Praat standard algorithm.

Table 1: summary of the evaluation results

	VUV %	GER%
Praat egg signal vs audio	12.18	1.13
Simple Comb h2=0 h3=0	8.87	14.37
Alt Comb h2=-1.0	7.99	1.90
Alt Comb h3=-0.4	7.74	1.85
Alt Comb h2=-0.4 h3=-0.4	7.29	1.43

VUV is the ratio between the number of frames that have been misclassified regarding the voicing state, and the total number of frames of the database. GER represents the ratio between the number of gross errors and the number of frames declared voiced by both reference and tested PEAs.

We did not report here the mean deviation of the F0 values found in the fine error category. In all the situations examined, the average difference was less than 0.07 semitone, with a standard deviation of less than 0.30 semitone. In other words, when there is no gross error, the value found for F0 is practically exact.

The first line shows the result of the reference algorithm applied to the egg signal band-pass filtered between 50 and 1000 Hz. The result shows large discrepancies concerning the voicing decision. The audio signal is declared less voiced than the egg signal, which casts a doubt on the value of the egg signal as a voicing ground truth: in most cases the vocal folds vibrate but the sound produced is inaudible or too low in frequency to correspond to the perceptive voicing. On the other hand, when both signals are declared voiced, the rate of gross errors is quite low.

The second line corresponds to the Simple Comb. The surprise comes from the rather high rate of gross errors. This could probably be improved by adjusting more precisely the number of teeth and their decaying function. However, there is a very large gap to fill to compete with the next case.

The 3rd line shows the drastic effect of a perfect cancellation of the octave error, with the coefficient h2 equal to 1. This confirms the observations reported in [7] and [11].

The 4th line shows the effect of partially cancelling the p=3 harmonic error. This effect is as strong as the previous one. It may be explained by the fact that for low-pitched voices and short frame durations the spectral peaks tend to merge. Their processing with the order 3 interteeth produces more or less the same effect than the single negative tooth of order 2.

Finally, using both orders yields the best result (line 5). It must be noted that the final gross error rate is still superior to the one obtained on the egg signal, which confirms the statistical validity of our results.

Although our evaluation was not done in order to compete with other PEAs, it should be noted that other authors using a very similar setup and the same database obtain results in the same range. For instance, on the Keele database, with width=40 ms, F0min=50 and F0max=550 Sun [11] gets a gross error rate of 2.08% for male speakers and 1.74% for female speakers, i.e. around 1.9% for the whole database, for which we found a best rate of 1.43%. However this difference is to be appreciated with caution, due to the difference in the choice of the reference data, as well as in the many small differences that occur from one experimental setup to another.

6. Conclusions

We have presented an approach to the problem posed by the gross errors in the F0 estimation of speech signals. This approach was motivated by the multipitch perspective. Even in the monopitch case, the problem is error-prone, and we tried to understand why.

We enumerated and counted the coincidences occurring when a periodic structure of fundamental frequency F0 is confronted to a periodic set of pulses of variable fundamental frequency Fc (simple comb). We found that the confusions were maximally plausible at certain locations, indexed with two positive integers p anq q, named respectively the harmonic and subharmonic orders. Thus, as we knew where the gross errors could happen, we could reduce from the start the nocivity of these locations. This was the basis of the Alternate Comb method, in which some negative teeth indicate where the spectral amplitude should be reduced to minimize the danger of confusion.

Evaluation on a popular database proved the method to give satisfactory results, thus validating our approach in the monopitch framework.

7. References

[1] De Cheveigné, A., "Multiple F0 estimation", in *Computational Auditory Scene Analysis*, Wang and Brown eds, IEEE Press, Wiley-Interscience, 2006.

[2] Schroeder, M. R., "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement", J. Acoust. Soc. Amer., **43**, 829-834, 1968.

[3] Martin, P., "Comparison of pitch detection by cepstrum and spectral comb analysis", IEEE ICASSP, 180-183, 1982.

[4] Hermes, D. J., "Measurement of pitch by subharmonic summation", J. Acoust. Soc. Amer., **83**, 257-263, 1988

[5] Boersma P. and Weenink,D. "Praat: doing phonetics by computer", http://www.praat.org/

[6] Camacho, A. and Harris, J. G., "A spectral-based pitch estimation algorithm and pitch perception model using an integral transform with a truncated decaying cosine kernel", 4th joint meeting of ASA and ASJ, Honolulu, 2006.

[7] Sun X., "A pitch determination algorithm based on subharmonic-to-harmonic ratio", 6th ICSLP, Beijing, 2000.

[8] Cooke, M., Barker, J., Cunningham,S. and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", J. Acoust. Soc. Amer., **120**, 2421-2424, 2006.

[9] Plante, F., Ainsworth, W.A. and Meyer, G.,"A Pitch Extraction Reference Database", Eurospeech Madrid, 837-840, 1995.

[10] De Cheveigné, A., "YIN, a fundamental frequency estimator for speech and music", J. Acoust. Soc. Amer., **111**, 1917-1930, 2002.

[11] Sun X., "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio", IEEE ICASSP, 333-336, Orlando, 2002.

Annexe D

Article Acoustics 08 – Le PSH

Cet article de référence bibliographique [Liénard et al., 2008] a été présenté au congrès ACOUSTICS 2008 à Paris en juillet 2008. Il présente un algorithme d'estimation multipitch nommé PSH.

Proceedings of Meetings on Acoustics

Volume 4, 2008 http://asa.aip.org

155th Meeting
Acoustical Society of America
Paris, France
29 June - 4 July 2008
Session 1pSCc: Speech Communication

1pSCc34. Using sets of combs to control pitch estimation errors

Jean-Sylvain Lienard*, Claude Barras and Francois Signol

***Corresponding author's address: LIMSI-CNRS, bp 133, ORSAY, 91403, Essone, France, jean-sylvain.lienard@limsi.fr**

We analyze the errors of a Pitch Estimation Algorithm using the Pitch Function (response of the periodicity estimator as a function of the frequency parameter Fc). The estimator's maximum response for a single signal of fundamental frequency F0 is expected to occur for Fc=F0. Actually the pitch function exhibits many secondary peaks which occasionally cause the errors. When several signals are mixed the main peaks do not reliably represent the F0s of the component signals. By taking as periodicity estimator the correlation of the spectrum module with a uniform infinite spectral comb of fundamental frequency Fc we show that each peak corresponds to a particular value of the ratio Fc/F0=p/q (p and q positive integers). It follows that some secondary peaks can be cancelled either by augmenting the comb with intermediary negative teeth, or by setting to zero some of its teeth. These modified combs can be viewed as combinations of uniform combs of different Fcs. The present study aims at precisely defining and combining the modified combs so that the main peaks of the new Pitch Function reliably indicate the F0s of the components. Examples are given on mixtures of voiced segments extracted from natural speech.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.2998757]
Received 14 Aug 2008; published 23 Sep 2008

Using sets of combs to control pitch estimation errors

Jean-Sylvain Liénard, Claude Barras and François Signol

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

[jean-sylvain.lienard,claude.barras,francois.signol]@limsi.fr

1. Introduction

Despite some recent improvements [1,2,3,4], pitch estimation of speech signals remains an error-prone process. The main sources of errors are i) the voiced-unvoiced decision and ii) the selection of harmonics or sub-harmonics instead of the fundamental frequency, yielding what is named the gross errors (i.e. estimation error > 20%). The problem is still more complex if one considers that several voices may be mixed in the signal [5].

In the present study we focus on the gross errors problem in voiced frames of short duration, in the multipitch perspective. Pitch is estimated in the spectral dimension by use of a spectral comb [1]. We define the Pitch Function PF as the response of a given pitch estimator in a given pitch interval. This notion embodies the notions of Period Histogram and Product Spectrum [6], as well as what is called "pitch strength" or "pitch saliency" by some authors. In the monopitch case, each peak of this function can be labeled by a couple of positive integers p and q, named respectively the harmonic and sub-harmonic indices. For the pitch estimation to be correct one has to select the peak (1,1). However this peak is not always the highest one, even in the monopitch case, and this is the main cause of the gross errors. Our efforts aim at increasing the prominence of the main peak by reducing the magnitude of the secondary peaks. To do so we define some particular combs, with missing or negative teeth, which produce particular pitch functions called Suppression Functions, each one addressing a given peaks family. They are combined into the final PF, in which the main peak gets reinforced, thus reducing the risk of gross errors. We present two ways of combining the Suppression Functions, the Alternate Comb and the Suppression Comb.

2. Infinite Uniform Comb

The Infinite Uniform Comb IUC(F_c) is made of an infinite number of unitary teeth regularly spaced in frequency. We call comb frequency F_c the frequency of its first tooth. The Pitch Function value for F_c is obtained by the scalar product of the comb IUC(F_c) with the module |S(F)| of the sound spectrum.

Although this comb is theoretically infinite, the scalar product is finite because the spectrum is bounded in frequency, as for any physical sound of limited energy. If {Bmin, Bmax} is the frequency band of the sound, and {$F_c$min, $F_c$max} the frequency

range in which the estimator computes the PF, a finite comb gives the same PF as an infinite comb if it has at least nteeth:

$$nteeth >= Bmax/F_c min$$

2.1 Functioning

The functioning of the IUC has been described in [7]. It is briefly illustrated in Fig. 1.

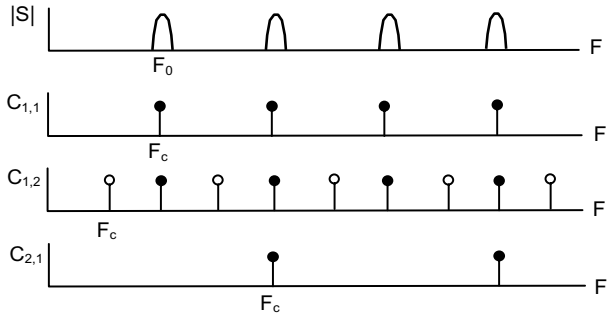

Fig 1. Functioning of the Infinite Uniform Comb on a schematic spectrum F_0. The IUC is represented for 3 values of its frequency F_c: $F_c=F_0$, $F_c=F_0/2$ and $F_c=2F_0$

When the comb frequency F_c is equal to the fundamental frequency F_0 of the sound (here a schematic sound made of a series of spectral peaks) each tooth of the comb matches a spectral comb. Thus the scalar product is maximum, yielding a peak of the PF. When F_c equals a multiple p of F_0 it also produces a peak of the PF, but only one peak out of p matches a tooth; thus the scalar product gets p times smaller. When F_c equals a sub-multiple q of F_0 all of the spectral peaks are matched by some teeth. Because the comb extends up to Bmax, the scalar product takes the same maximum value as in the $F_c=F_0$ case. More generally, the PF peaks appear when the following relation is observed:

$$F_c/F_0 = p/q \qquad p, q, integers >0$$

2.2 The peaks of the Pitch Function

Let us consider a physical sound made of a series of pulses at F_0=250 Hz, Hanning windowed, of duration 50 ms. Fig. 2 shows the PF obtained by use of an IUC, normalized to the height of the main peak. Some of the numerous peaks obtained are identified in terms of the harmonic and sub-harmonic indices p and q. In the following we shall consider that p and q are the integer terms of the irreducible fraction F_c/F_0. The peak (4,2) does not exist; it is just a redundant notation for the peak (2,1).

As mentioned above the sub-harmonic peaks (1,q) have the same maximum magnitude $A_{1,1}$ as the main peak (1,1). However, the magnitude of the harmonic peaks

(p,1) decreases theoretically according to the inverse of their index p, because only one tooth out of p matches a spectral peak:

$$A_{p,1} = A_{1,1} / p$$

Besides the harmonic (h) and sub-harmonic peaks (sh) one observes secondary (or fractional) peaks, the frequencies of which are given by:

$$F_{p,q} = (p/q)\, F_{1,1}$$

Their magnitude may vary with the spectral composition of the sound, but theoretically they tend toward the values given by:

$$A_{p,q} = A_{1,q} / p$$

In the case of the IUC the magnitude $A_{1,q}$ of the sub-harmonic peaks (1,q) equals that of the main peak, thus:

$$A_{p,q} = A_{1,1} / p$$

In other words, the relative magnitude of a fractional peak (p,q) is an inverse function of the harmonic index p and does not depend on the sub-harmonic index q.

In Fig. 2, the magnitude of the peak (3,2) at 375 Hz should be close to $A_{1,1}/3$ but the observed value is lower, due to the decreasing spectral envelope of the series of pulses used in our example sound.

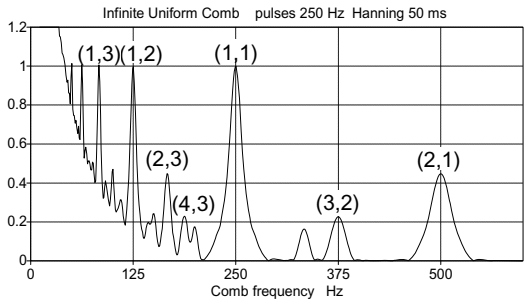

Fig. 2 Pitch Function, Infinite Uniform Comb applied to a 50 ms Hanning-windowed 250 Hz series of pulses

The fact that the sub-harmonic peaks have the same value as the main peak makes the use of the IUC problematic for the estimation of F_0, even in the monopitch case. Before examining some possible solutions we have to mention the role of the signal windowing in the PF shape.

2.3 Signal duration and windowing

When the signal duration gets shorter, the spectral peaks get thicker. For a 100 ms Hanning window, for instance, the width of a spectral peak at half height is about 20 Hz. It becomes about 40 Hz when the window gets shortened to 50 ms. The peaks of the PF, as they result from the accumulation of several spectral peaks, evolve in the same way.

This widening alters the resolution of the pitch estimation, especially in the low F_c range.

Another effect of this widening appears when F_c gets very low, so that several teeth of the comb can take place in the width of a single spectral peak. This produces the decreasing hyperbolic component observed in the low range of the PF (Fig. 2). The resulting masking effect may be tolerated to some extent if one is more interested in the peaks of the PF than in its valleys. In the example illustrated above the PF peaks remain unmasked for F_c>40 Hz, which constitutes the real lower bound of the estimator's frequency range. Several solutions may be implemented to reduce this effect, such as using a better window or modeling the spectral peaks in order to diminish their apparent width.

3. Control of the sub-harmonic decrease

In order for the comb to function as a pitch estimator it is mandatory to reduce the magnitude of the sub-harmonic peaks (1,q). Two techniques, among others, have been used in the past to achieve this task.

3.1 Limiting the number of teeth

Limiting the number of teeth to a fixed small number N_t<nteeth practically ensures the magnitude of the main peak $A_{1,1}$ to be the real maximum of the PF, in the monopitch case. For $F_c=F_0$ each tooth matches a spectral peak, so that $A_{1,1}$ is proportional to N_t. For $F_c=F_0/2$ only the $N_t/2$ even teeth match spectral peaks. The spectral peaks located above $N_t/2$ do not contribute any more to $A_{1,2}$, which yields $A_{1,2}<=A_{1,1}$. The "equal" case may occur if the part of the spectrum matched by the teeth located between $N_t/2$ and N_t has zero magnitude. The same line of reasoning holds for the other sub-harmonic peaks. Setting the number of teeth to a value comprised between 5 and 10 yields a progressive decrease of the sub-harmonic peaks, sufficient in practice to promote (1,1) as the main peak of the PF.

This technique implies a loss of information because the spectral peaks lying above the last tooth are not taken into account in the periodicity estimation. This loss is not critically important for speech because i) those high-frequency peaks convey less energy than the low ones and ii) they are to some extent packed down by the jitter.

3.2 Decreasing the teeth magnitudes

Another way to ensure the decrease of the sub-harmonic peaks is to reduce the magnitude of the teeth as a function of their rank r. An exponential decrease in r^{-k} is often chosen. This technique is just a variation of the previous one in that it favors the effect of the lower teeth. However its action is smoother, as the teeth contribution to the PF decreases with their rank. The counterpart is that it may be more sensitive than the previous one to any alteration of the energy carried by the fundamental (telephone speech, or partial masking by a low frequency noise).

The exponent k is often empirically given the value 0.5. This produces a decrease

of √2 of the peak (1,2), and a similar increase of the peak (2,1), compared to the values they had by use of the IUC. Thus the two peaks presenting the maximum risk of gross error (sub-octave and octave errors) get the same height. Both techniques can be used

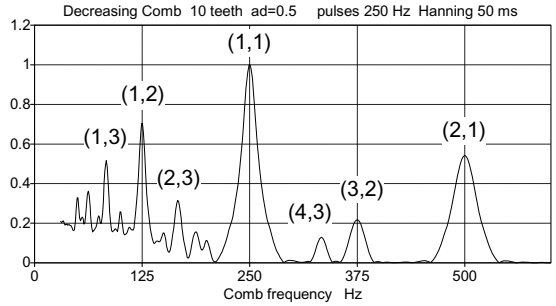

Fig 3. Pitch Function of a 10-teeth comb decreasing in √r. Same sound as in Fig. 2

together, as illustrated in fig 3. The comb used has a limited number of teeth (10 teeth), decreasing in √r. The goal of reducing the sub-harmonic peaks is attained, but the rejection parasitic peaks remains poor and the risk of octave or sub-octave errors is still high, especially in the multipitch case.

4. Control of the gross errors by use of irregular combs

In this section we present a novel approach of the control of the gross errors, based on the use of irregular combs. The particular Pitch Functions obtained with those combs are called Suppression Functions.

4.1 Missing Teeth Comb and sub-harmonic suppression

Let us consider an Infinite Uniform Comb, from which the even-numbered teeth of rank

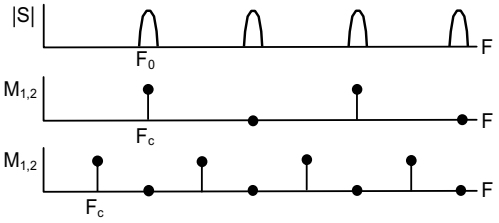

Fig. 4 Missing Teeth Comb of order 2 (odd comb $M_{1,2}$), positioned at $F_c=F_0$ and $F_c=F_0/2$. This comb cancels the sub-harmonic peaks (p,2k), p,k integers >0 p<2k

2k have been removed (odd comb, Fig. 4). The PF generated in the presence of a harmonic spectrum of fundamental F_0 exhibits a minimal response (close to zero) for the peaks (p,2k):

$$F_c/F_0 = p/2k \qquad p,\ k \text{ integers} > 0 \qquad p < 2k$$

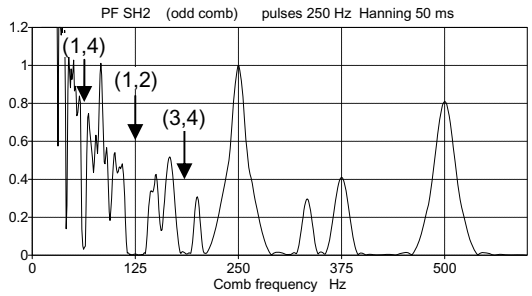

Fig 5. Suppression function SH2 obtained by processing the 250 Hz sound with the odd comb. Peaks (1,2), (1,4), (3,4)... at 125, 62.5, 187.5 Hz have been cancelled

Figure 5 shows that the peaks (1,2), (1,4), (3,4), (1,6), (5,6)... have been practically cancelled, to the extent that they were distinct in the initial PF (compare with Fig. 2). Let us denote (sh2) this order 2 sub-harmonic peaks family. The corresponding PF, after processing of the spectrum |S|, is called order 2 sub-harmonic Suppression Function and denoted SH2.

Similarly, a Missing Teeth Comb of order 3 is defined by removing the teeth of rank 3k. This comb cancels the peaks:

$$(p,3k) \qquad p,\ k \text{ integers} > 0 \qquad p < 3k$$

Those peaks (1,3), (1,6), (5,6), (1,9)... constitute the order 3 sub-harmonic family denoted (sh3). The PF obtained with the spectrum |S| is the Suppression Function SH3.

Continuing with the orders greater than 3, we have to observe that the (sh4) and (sh8) families are totally included in the (sh2) family. Similarly, the (sh6) family is included in both (sh2) and (sh3) families. Thus the corresponding suppression functions are redundant. Only the families of prime order are necessary to cancel all the error-prone sub-harmonic peaks.

If $F_c$max has been set at 600 Hz and $F_c$min at 75 Hz, then it is useless to cancel any peak beyond order 8. As we only need the prime orders, computing SH2, SH3, SH5 and SH7 is sufficient. We note that all those functions let the other peaks, including (1,1), at their proper location, while reducing their absolute magnitude in the proportion of the missing teeth.

4.2 Alternate Teeth Comb and harmonic suppression

In order to control the harmonic peaks (p,1) an approach similar to the previous one consists in placing some negative teeth between the regular teeth of the Infinite Uniform Comb.

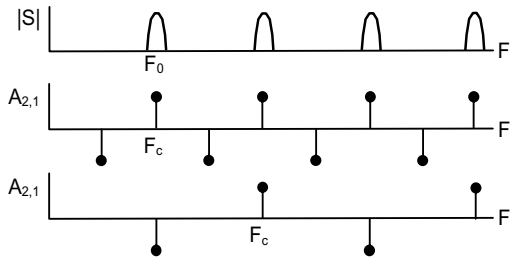

Fig. 6 Alternate Teeth Comb of order 2 $A_{2,1}$, positioned at $F_c=F_0$ and $F_c=2F_0$. This comb cancels the harmonic and fractionary peaks (2k,q) with $q<2k$

Let us consider the order 2 (Fig. 6). The positive teeth are located at frequencies kF_c (k integer >0). When the IUC is positioned at the octave, $F_c=2F_0$, it produces the parasitic peak (2,1) with a magnitude lower than the magnitude $A_{1,1}$ of the main peak. To eliminate it one has to subtract from the PF a part of $A_{1,1}$. This may produce a negative value in place of the peak (2,1). A thresholding strategy will be necessary to use this new suppression function H2 in a multiplicative combination to get the final PF.

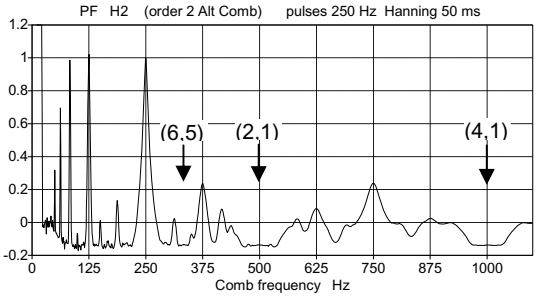

Fig 7. Suppression function H2 obtained by processing a periodic spectrum with the order 2 Alternate Teeth Comb. The function has been normalized to 1

Actually the family (h2) is not limited to the peaks (2k,1), but includes also some fractional peaks such as (6,5), which is actually the 6th multiple of the 5th sub-multiple of F_0 (Fig.7). In symbolic terms the (h2) family may be denoted:

(h2) = (2k,q) q<2k k, q integers >0

At this point we have to mention that subtracting the middle part of the interpeak parts of the spectrum to improve the pitch estimation has been investigated with success in [3] and [4], by following approaches different from ours.

At order 3 the Alternate Teeth Comb comprises 2 intermediary negative teeth, equally spaced between the regular (positive) teeth. This comb cancels the (h3) peaks family defined by:

(h3) = (3k,q) k, q integers >0 q<3k

Among the members of this family, one finds peaks (3,1), (6,1)... (3,2), (9,2)... (6,1), (6,5)... (9,1), (9,2), (9,4)...

As before, we note that some harmonic suppression families are redundant: (h4), (h6) and (h8) are included in (h2), (h6) is included in (h2) and (h3) etc. In other terms the prime order families (h2), (h3), (h5) and (h7) are sufficient in order to cancel all the harmonic and fractional peaks appearing in a $\{F_c\text{min}, F_c\text{max}\}$ interval less than 11 times $F_c$min.

4.3 Alternate Comb

The Pitch Estimation Algorithm presented in [7] is an implementation of two of the approaches described above. The reduction of the sub-harmonic peaks is done by the limitation of the number of teeth and their exponential decrease, while the reduction of the harmonic peaks is done by using an additive combination of two Alternate Teeth Combs, of orders 2 and 3. Two coefficients a_2 and a_3 were used to control their part on the harmonic reduction process, and another coefficient a_d (the exponent of the teeth decrease) was used to control the decrease of the teeth magnitude. An experimental work yielded the optimal values of a_2, a_3 and a_d. After evaluation on two classical pitch databases and comparison with other PEAs, the performance of the Alternate Comb appeared to be at the level of the best published results.

4.4 Suppresssion Comb

The Suppression Comb aims at consistently implementing the harmonic and sub-harmonic suppression functions described in the preceding sections. Each one of those functions preserves the main peak and controls some families of parasitic peaks. Thus a multiplicative combination - completed by the use of positive thresholds to avoid the negative parts - looks more appropriate than the additive one implemented in the Alternate Comb.

A first implementation consists in using as many irregular combs as the number of suppression functions needed. If we need to cover a 10 times pitch interval we need to achieve the suppression up to the order 10, and a set of 8 combs is convenient: 4 Missing

Teeth Combs to get the functions SH2, SH3, SH5, SH7, and 4 Alternate Teeth Combs to get H2, H3, H5 and H7. This method requests some amount of computing but does not bother - or only marginally - with the problem of the low frequency hyperbolic increase of the PFs.

Another implementation is based on the fact that the suppression functions can be computed from the PFs of the Infinite Uniform Comb, according to the following formulae. PF designates the Pitch Function of the IUC, i designates the order of the reconstructed irregular comb. The coefficients of the H functions ensure that the Alternate Teeth combs are centered (null mean value):

$$SH_i(f_c) = PF(f_c) - PF(if_c)$$

$$H_i(f_c) = \frac{i}{i-1} PF(f_c) - \frac{1}{i-1} PF(\frac{f_c}{i})$$

This method is faster than the first one, but it implies the computation of the PF at very low frequencies ($F_c$min/7 for the function H7), and thus encounters the problem mentioned above.

In the following examples the final PF, obtained by multiplying the 8 suppression functions, was weighted in frequency by √fc , in order to get a supplementary gain on the SH peaks such as the one evoked in section 3.2.

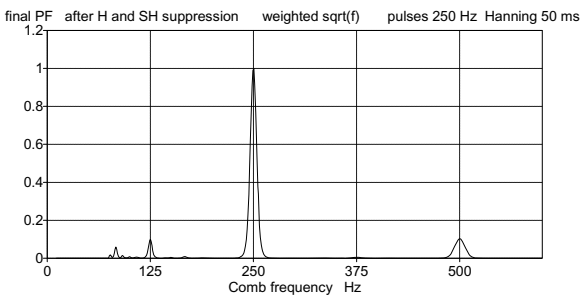

Fig 8. Final PF after H and SH suppression and √f weighting

Figure 8 represents the final PF of the same sound used throughout the paper. Comparing it to figures 2 and 5 gives an idea of the efficiency of the whole suppression process. The parasitic peaks (1,2) and (2,1) are now rejected at some -20 dB below the main peak instead of -3dB (cf Figure 3). The consequence is that the risk of gross F0 errors is highly reduced.

Figure 9 shows the final PF obtained with the mixture of same sound and an equally intense white noise. The consequence is an increase of the sh peaks (1,2) and (1,3) and a decrease of the h peak (2,1). Those parasitic peaks remain small and do not prevent the main peak to emerge.

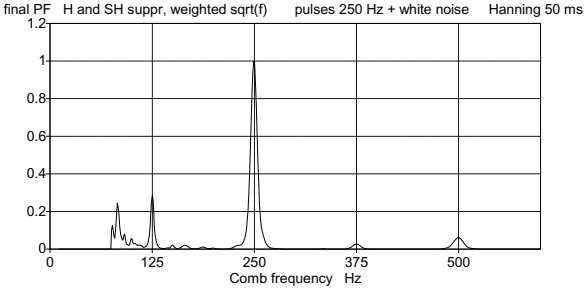

Fig 9. Same as in fig 8, but the sound has been mixed with white noise of same intensity

Figure 10 represents the final PF obtained from the mixture of two vocalic sounds uttered by a single male voice. Both sounds were extracted from continuously pitch-varying and spectrum-varying sequences. The main peak of the most acute sound is less prominent, but it appears nevertheless in second position.

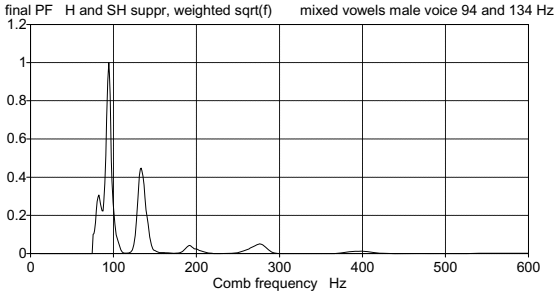

Fig 10. Mixture of two vowels at 94 and 134 Hz from the same male speaker

Finally, figure 11 shows a mixture of two artificial vowels of same intensity, whose F_0s are exactly one octave apart. The correct peaks emerge clearly, which can be seen as a positive feature for a harmonic suppression algorithm.

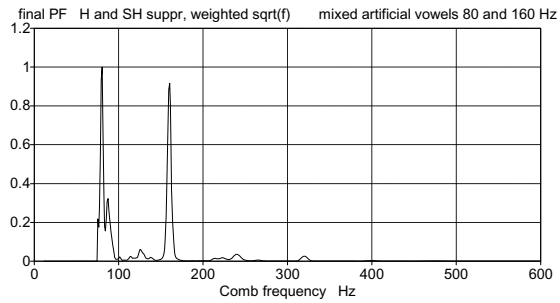

Fig 11. Mixture of two artificial vowels of same intensity, with F_0s exactly one octave apart (80 and 160 Hz)

5. Conclusion

The main goal of this work was to improve the ability of the spectral comb methods to provide an error-free pitch estimator, in order to be usable in both mono and multipitch estimation. We proposed to use irregular combs, either with some missing teeth or with alternate negative teeth, which have the property of canceling some families of erroneous solutions. An evaluation of this approach in the mono and multipitch cases is in progress [8].

References

[1] Martin, P., "Comparison of pitch detection by cepstrum and spectral comb analysis", IEEE ICASSP, 180-183, 1982.

[2] De Cheveigné, A., "YIN, a fundamental frequency estimator for speech and music", J. Acoust. Soc. Amer., 111, 1917-1930, 2002.

[3] Sun X., "A pitch determination algorithm based on sub-harmonic-to-harmonic ratio", 6th ICSLP, Beijing, 2000.

[4] Camacho A., "SWIPE: a sawtooth waveform inspired pitch estimator for speech and music", PhD dissertation, Univ. of Florida, 2007.

[5] De Cheveigné, A., "Multiple F_0 estimation", in Computational Auditory Scene Analysis, Wang and Brown eds, IEEE Press, Wiley-Interscience, 2006.

[6] Schroeder, M. R., "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement", J. Acoust. Soc. Amer., 43, 829-834, 1968.

[7] Liénard J.S., Signol F., Barras C., "Speech Fundamental Frequency estimation using the Alternate Comb", InterSpeech 2007, Antwerpen.

[8] Signol F., Barras C., Liénard J-S., "Evaluation of the Pitch Estimation Algorithms in the Monopitch and Multipitch cases", Acoustics 2008, Paris

Annexe E

Article Acoustics 08 – Évaluation

Cet article de référence bibliographique [Signol et al., 2008] a été présenté au congrès ACOUSTICS 2008 à Paris en juillet 2008. Il traite de l'évaluation du PSH et d'une méthodologie d'évaluation intégrant la décision VnV.

Evaluation of the Pitch Estimation Algorithms in the monopitch and multipitch cases

F. Signol, C. Barras and J.-S. Lienard

LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
jean-sylvain.lienard@limsi.fr

Reliably tracking the fundamental frequency F_0 of the components is an important step in the separation of superimposed speech signals. Several Pitch Estimation Algorithms (PEAs) are potentially usable and a rigorous evaluation method is needed. However, even in the monopitch case, many variations between them render such a comparison difficult. The F_0min-F_0max interval extent, the use of a priori information on the whole sequence or database and above all the arbitrary voicing threshold setting lead to large differences in the results. These biases can be removed by setting the F_0 bounds to fixed values acceptable for many voices, by proceeding with the evaluation on a strictly frame-to-frame basis, and by fixing the voicing threshold in order to get an equal error rate for overvoiced and undervoiced frames. In the multipitch case any frame may exhibit 0, 1 or 2 valid voicing according to the coincidence between the voiced and unvoiced parts of both signals. This problem is treated by defining a metric linking the PEAS's hypotheses to the pitch values of the isolated signals. The proposed methodology is applied to several PEAs on several databases in the monopitch case.

1 Introduction

In the last five decades, numerous PEAs have been developed so that a proper evaluation is mandatory. Most PEAs process the input speech signal as successive short-time frames. A F_0 estimate is produced for each voiced frame. For unvoiced frames the F_0 estimate is replaced by a zero value. This involves two distinct processes: Voiced/Unvoiced (VuV) frame decision and F_0 estimation. In some cases those processes are performed sequentially. In [1,3] the first step is the computation of a voicing criterion (typically between 0 and 1). Then this voicing criterion is compared to a threshold and the Voiced/Unvoiced (VuV) decision is taken. This is a typical sequential frame-to-frame approach. In some other cases the two processes are not distinct and sometimes they are mutually dependent. For instance in [2,6] several F_0 candidates are selected for each frame and a dynamic programming algorithm is used to select the globally optimal sequence of candidates. In this case, the F_0 estimation influences the VuV decision and vice-versa. Eventually evaluating a PEA means simultaneously evaluating the two above processes: VuV decision and F_0 estimation.

PEAs evaluation requires a database, VuV references, F_0 references and quality measures. Thus several variability sources should be addressed, which means numerous features to set and control. First of all, one has to control the database specificities such as the speech recording conditions (noise level, anechoic room, telephone, car...) or the type of speech (read, broadcast news, conference, spontaneous, expressive...). The reference voicing and F_0 labeling has to be perfectly reliable. It may differ according to the protocol used (manual, fully automatic or manually corrected); it may come from the analysis of the speech signal itself, or from an electroglottographic waveform (EGG) recorded simultaneously. The VuV decision may have been taken on the basis of the experience of trained phoneticians, or from the examination of a waveform on physical criteria, which may be quite different. Finally each PEA's specificity should be taken into account. It may or may not offer the operator the capability of disabling some pre- or post-processing, or of modifying the main parameters (frame duration, frame hop, voicing threshold, lower and upper bounds of F0 estimation, number of F0 candidates per frame). For all the above reasons, it is extremely difficult to perform a fair comparison between algorithms.

In this paper we propose an evaluation methodology that could help containing some of the main sources of variability. Section 2 presents what has been done in PEA's evaluation, highlights how the VuV decision is problematic to evaluate and proposes a way to deal with it. The usual PEA use is the monopitch case where it has to deal with a one-speaker speech signal. Section 3 proposes a new evaluation methodology and illustrates it on a database in the monopitch case. Section 4 extends our monopitch evaluation approach to the multipitch case, in which several voices may be mixed in a single signal.

2 The role of the VuV decision in the evaluation of F_0 estimation

The first quality measure for which a consensus in literature exists is the quality of the pitch estimation. What is generally reported is a gross error rate (GER). A gross error is raised when the pitch estimate lies more than a certain distance away from the reference (typically 20%). This rather high threshold has to be put in perspective with the typical pitch errors which are mainly octave or sub-octave errors and most seldom third, triple or two thirds errors [5]. The gross error rate does indeed capture the principal harmonic or sub-harmonic errors.

F_0 estimation is performed on the frames which are considered as voiced. In that sense, the GER calculation is subjected to the VuV decision. However, the voicing state may be ill-defined for many frames, often located at the bounds of the voiced speech segments. In those regions F_0 estimation may remain unsteady, whereas for frames where the voicing is strong, the F_0 estimation is quite reliable. For any PEA, avoiding the litigious frames may drastically reduce the F_0 error rate. This phenomenon is illustrated on Fig. 1, which shows that the GER decreases and falls to 0% when the VuV threshold is increased so that only the reliably voiced frames are taken into account. This example was performed on a short sample of speech with the YIN algorithm, but the same trend can be observed with any algorithm, on any speech excerpt.

Therefore, for a given PEA the GER calculation should not be presented alone. It should be linked with an evaluation of the VuV decision.

The results published in [3] are established on the frames declared as voiced by all PEAs involved in the evaluation. Thus they discard from the statistics the frames whose voicing is not reliably established. This protocol may reflect the intrinsic quality of the algorithm, in dealing with perfectly voiced sounds. But it does not give a fair idea of the PEA's behavior in the difficult situation encountered at both ends of the voiced segments. Thus, we have to normalize the algorithms behaviors in terms of VuV decision.

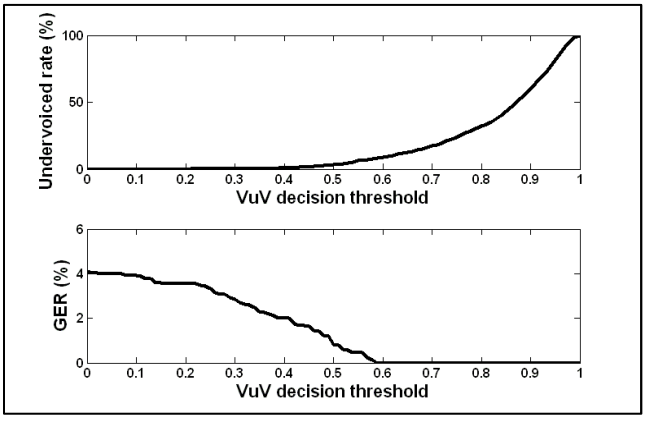

Fig. 1. Rate of frames declared as voiced and Gross Error Rate as functions of the VuV decision threshold

To formalize our voicing behavior normalization let us look at table 1 which illustrates the kinds of errors that could appear between the VuV state of a reference frame and the VuV state of an hypothesis frame. A false alarm (FA) called overvoiced error occurs when the hypothesis frame is voiced and not the reference frame. A false rejection (FR) called undervoiced error occurs in the opposite situation.

Reference	Hypothesis	Agreement
UV	UV	OK
UV	V	FA
V	UV	FR
V	V	OK

Table 1. Full set of reference/hypothesis couples and their agreement result.

Table 1 refers to a typical problem of decision theory. This problem may be addressed within the Equal Error Rate (EER) statistical framework. A way to compare the different PEAs without the bias attached to their specific VuV behavior is to tune it so that the FA/FR ratio gets a constant value. Let EER be the ratio of the overvoiced error rate to the undervoiced error rate. The value chosen for the EER depends on the voicing behavior requested by the application where we want to use a PEA. For instance, if the application is only interested in strongly voiced segments, we should set EER at a value between 0 and 1 in order to get more undervoiced than overvoiced frames. For the present evaluation purpose we shall fix it at the value 1.

3 Monopitch case

In this chapter we present our evaluation methodology beginning with the database presentation. Then we explain precisely how we build our reference. Then we provide a general formalization of how could be represented a set of reference annotations and F_0 hypotheses. At last, we propose some evaluating features and formalize them in the monopitch case.

3.1 Evaluation method

3.1.1 Database

Our mono-speaker database is composed of three speech corpora, all recorded in clean acoustic conditions. The Keele database (noted K) comprises 10 English speakers (5 males, 5 females) reading the same text one time and quite neutrally, for a total duration of 337 seconds. The Bagshaw database (noted B) comprises 2 English speakers (1 male, 1 female). Both are saying the same 50 shorts utterances, for a total of 331 seconds. The Daless database (noted D) comprises 4 French speakers (2 males, 2 females) reading various texts quite neutrally. This database contains 1359 seconds of speech. For all the databases, the authors give the EGG waveform for each speech signal.

To build a 2-speaker mixture, two normalized signals are simply added. Then, the whole 2-speaker mixtures databases are built by forming all possible signals combinations.

3.1.2 Reference annotations

This step needs a particular care because it determines the quality of the evaluation results. It raises the “groundtruth” problem.

Manual or manually corrected reference annotation is a very tedious task. It must be performed by a group of specialized operators to be reliable. However, the protocols and groups of operators differ from one database to the next, so that it may happen that a given algorithm gets different results when evaluated on different databases. This is why we chose to automatize the annotation process, even if we know that it may produce more errors than the manual annotation of a given database. Also, as there are few manually annotated databases, this approach permits the use of a non-annotated database. For instance our largest database is not manually annotated; an automatic annotation procedure was the only way to use it with the same annotation quality than the other two. In any case, a careful examination of each database is needed in order to adapt some parameters such as the F_0 range.

As EGG signals are available for all databases, we generate the VuV decision with them. Although it is not perfect, the EGG waveform is a direct trace of the vocal cords vibration so it should give a more reliable and steady voicing measure than the acoustic speech signal.

The EGG waveform is first shaped by a bandpass filtering between 50 and 1600 Hz to cancel undesired noises and low-frequency components. Then, a positive saturation threshold is computed which corresponds to the amplitude exceeded by 5% of the samples. The same negative saturation threshold is computed with the negative part of the waveform. If the negative saturation threshold's absolute value is bigger than the positive saturation threshold then only the absolute value of the waveform's negative part is kept (negative half-wave rectification). Else a positive half-wave rectification is done. The obtained waveform is normalized at 0.9 and then the temporal envelope is computed by linear interpolation between local maxima. A 18-point histogram is computed from the envelope. The histogram exhibits two peaks, corresponding respectively to the unvoiced and to the voiced segments. The local minimum between these two peaks is chosen as the voicing threshold. Below this threshold the sample is considered unvoiced and above the sample is voiced. A refinement is added to avoid too short voiced or unvoiced segments. All unvoiced segments shorter than a first

threshold and surrounded by voiced segments become voiced segments (voiced fusion). Then the isolated voiced segments shorter than a second threshold become unvoiced.

This algorithm returns a series of voiced segments defined by a beginning time and a finishing time. A frame is voiced if at least half of its width is included in a voiced segment.

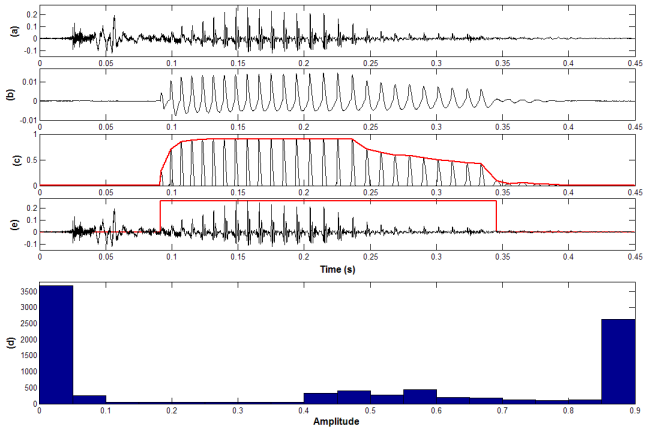

Fig 1. (a) speech signal (b) egg signal (c) filtered, saturated, half-wave rectified, normalized egg (black line) and temporal envelope (red line) (d) speech signal (black line) and voiced segment (red line) (e) envelope 18-points histogram. The voicing threshold is 0.1.

To compute the F0 reference value for a voiced frame, a basic and easily reproducible algorithm is used. The EGG signal is windowed. On each window, normalized autocorrelation is computed, its first main lobe is removed and the abscissa of the highest peak is considered as the period of the fundamental for this frame.

This algorithm returns a series of frames described by two values: the frame's time which is the middle instant of the EGG window studied, and the estimated F_0 which is strictly positive if the frame is voiced and 0 else.

3.1.3 Quality measures

Let us define some notations.

We note N_S the number of speakers mixed in the speech signal. N_R corresponds to the number of F_0 values in each reference frames. N_R is equal to N_S because there are as much references values as speakers. N_H is the number of F_0 hypothesis candidates given by a PEA per frame. N_H is superior or equal to N_R. We also note N_F the total frame's number.

Let us note $\cap$ the intersection operator, $\cup$ the union operator, Ω the cardinal operator and | the writing shortcut which signifies "such as".

We define R as the references frames set and R^t as the t^{th} reference frame. R^t contains the set of F_0 references values denoted R_x^t. R_x^t is the x^{th} F_0 reference value in the t^{th} reference frame. The same kind of notation is applied to the hypotheses. H is the set of hypotheses frames. H^t is the t^{th} hypotheses frame and H_y^t is the y^{th} F_0 hypothesis value in the t^{th} hypotheses frame. The range of t, x and y are given in Eq.(1).

$$\begin{matrix} R = \{R^1 \dots R^{N_F}\} \\ H = \{H^1 \dots H^{N_H}\} \end{matrix} \text{ and } \begin{matrix} R^t = \{R_1^t \dots R_{N_R}^t\} \\ H^t = \{H_1^t \dots H_{N_H}^t\} \end{matrix} \quad (1)$$

An unvoiced F_0 candidate is set to zero. Else, it is strictly positive. We note VC_R the set of voiced F_0 references values and uVC_R the set of unvoiced F_0 references values. VC_H is the set of voiced F_0 hypotheses values and uVC_H the set of unvoiced F_0 hypotheses values. A mathematical formulation is given in Eq.(2).

$$\begin{matrix} VC_R = \{(x,t) | R_x^t > 0\} \\ VC_H = \{(y,t) | H_y^t > 0\} \end{matrix} \text{ and } \begin{matrix} uVC_R = \{(x,t) | R_x^t = 0\} \\ uVC_H = \{(y,t) | H_y^t = 0\} \end{matrix} \quad (2)$$

Concerning the VuV decision, we consider a frame as voiced if at least one of its F_0 values is strictly positive. Thus let us note (R^t>0) as a voiced references frame and (H^t>0) as a voiced hypotheses frame. We also note (R^t=0) an unvoiced references frame and (H^t=0) and unvoiced hypotheses frame. These definitions are illustrated in the Eq.(3).

$$\begin{matrix} R^t > 0 \Leftrightarrow \exists x | R_x^t > 0 \\ H^t > 0 \Leftrightarrow \exists y | H_y^t > 0 \end{matrix} \text{ and } \begin{matrix} R^t = 0 \Leftrightarrow \forall x | R_x^t = 0 \\ H^t = 0 \Leftrightarrow \forall y | H_y^t = 0 \end{matrix} \quad (3)$$

VF_R is defined as the set of voiced references frames and VF_H is the set of voiced hypotheses frames. uVF_R is the set of unvoiced references frames and uVF_H the set of unvoiced hypotheses frames.

$$\begin{matrix} VF_R = \{t | R^t > 0\} \\ VF_H = \{t | H^t > 0\} \end{matrix} \text{ and } \begin{matrix} uVF_R = \{t | R^t = 0\} \\ uVF_H = \{t | H^t = 0\} \end{matrix} \quad (4)$$

We define two series of quality measures: the first one deals with the VuV decision, and the second one deals with F_0estimation.

In the monopitch case, N_R and N_S are equaled to 1 and the number of references frames is the same than the number of F_0 references values.

The overvoiced frame rate (OVR) corresponds to the number of FA frames over the number of unvoiced references frames as explained in Eq.(5). The undervoiced frame rate (UVR) is the ratio between the FR frames number and the number of voiced references frames. It is given in Eq.(6).

$$OVR = 100 \frac{\Omega[uVF_R \cap VF_H]}{\Omega[uVF_R]} \quad (5)$$

$$UVR = 100 \frac{\Omega[VF_R \cap uVF_H]}{\Omega[VF_R]} \quad (6)$$

The voiced decision agreement rate (V_{ok}) is given in Eq.(7). It corresponds to the number of common voiced frames between the reference and the hypothesis over the number of voiced frames in the reference or the hypothesis. The unvoiced decision agreement rate (uV_{ok}) is described in Eq.(8).

$$V_{ok} = 100 \frac{\Omega[VF_R \cap VF_H]}{\Omega[VF_R \cup VF_H]} \quad (7)$$

$$uV_{ok} = 100 \frac{\Omega[VF_R \cap VF_H]}{\Omega[VF_R \cup VF_H]} \quad (8)$$

The F_0 estimation quality measure involves a Gross Error Threshold (GET) fixed at 20%. A F_0 reference value is said "in accordance" with a F_0 hypothesis value whether their absolute relative distances is inferior to GET. Eq.(9) presents the mathematical formalization of this definition. We note the "in accordance" operator by the ≈ symbol.

$$R_x^t \approx H_y^t \Leftrightarrow \left(R_x^t > 0\right) \wedge \left(\exists y \left| \frac{\left|R_x^t - H_y^t\right|}{R_x^t} \leq GET \right. \right) \quad (9)$$

The "in accordance" operator is extendable to frames. A reference frame is in accordance with a hypothesis frame if all the F_0 references values are in accordance as showed in Eq.(10).

$$R^t \approx H^t \Leftrightarrow \forall x, \exists y \left| R_x^t \approx H_y^t \right. \quad (10)$$

The set of references frames in accordance is noted R_{ok} and is described in Eq.(11).

$$R_{ok} = \left\{ t \left| R^t \approx H^t \right. \right\} \quad (11)$$

The "in accordance" rate corresponding to a recall rate is noted RER and is given by Eq.(12). It is the ratio between the number of "in accordance" reference frames and the number of common voiced frames between reference and hypothesis.

$$RER = \frac{\Omega\left[R_{ok}\right]}{\Omega\left[VF_R \cap VF_H\right]} \quad (12)$$

The precision rate (PRR) corresponds to the number of "in accordance" reference frames over the number of voiced F_0 hypotheses values needed to put in accordance the reference frame. It is an indication of the PEA efficiency to find the rights F_0 values in the first hypotheses candidates. PRR is explained in the Eq.(13) below.

$$PRR = \frac{\Omega\left[R_{ok}\right]}{\sum_{y \in [1, N_H], t \in H_{ok}} \Omega\left[\left\{H_y^t > 0\right\}\right]} \quad (13)$$

In the case of PEAs which provide a single F_0 hypothesis value per frame, PRR always equals to 100%.

The Gross Error Rate (GER) is theF_0 estimation accuracy indicator. It corresponds to the number of frames not "in accordance" over the total number of voiced references frames. We note R_{ko} the set of reference frames not "in accordance". The GER is given in Eq.(14).

$$GER = \frac{\Omega\left[R_{ko}\right]}{\Omega\left[VF_R\right]} \quad (12)$$

An hypothesis F_0 can be associated with several reference F_0 values. Thus if some F_0 reference values are equal in a same frame (crossing streams of F_0), PEAs probably return a single hypothesis for the ensemble of equal F_0 reference values. This multi-association allows the evaluation to deal with this problem.

Once a hypothesis F_0 value is associated to one or more reference F_0 values, the hypothesis F_0 value is locked and can not be associated with other reference F_0 values.

In the monopitch case, a F_0 candidate, a frame and a stream are the same. Thus, it remains only four voicing quality measures.

3.2 Monopitch evaluation results

Four PEAs are evaluated and compared: YIN, SWIPE, PRAAT and PSH. YIN [1] provides an aperiodicity criterion which can be easily converted into a voicing feature between 0 and 1. It gives one hypothesis per frame. SWIPE [3] provides a voicing strength criterion between 0 and 1 and also gives a single hypothesis per frame. PSH [4] provides a voicing strength which needs to be converted in a voicing feature between 0 and 1. It gives a tunable number of hypotheses and is adapted to the multipitch case. In the monopitch case, we fixed it at 1 like the others tested PEAs. PRAAT [2] provides an autocorrelation coefficient between 0 and 1 as voicing feature. In our series of algorithms, it is the only PEA with a post-processing. We tuned its parameters to remove this post-processing effect. All PEAs are tuned to provide a single F_0 hypothesis value so in Table 1 the PRR column is removed because PRR is always equal to 100%.

The particularities of each PEAs may lead to some small time misalignments between the references and the hypotheses so that a time alignment step is needed. It consists in a linear interpolation between the first before hypothesis value and the first after value. There is an undefined interpolation in the case of an unvoiced hypothesis frame and a voiced one. These frames are discarded from evaluation. As the reference F_0 range differs from the hypothesis F_0 range, all the F_0 references out of the hypothesis F_0 range are discarded from the evaluation.

	EER	V_{ok}	uV_{ok}	OVR	UVR	RER	GER
YIN	1.00	76.5	74.7	12.3	12.4	93.9	5.1
PSH	1.01	85.6	85.7	6.3	6.3	95.5	4.2
SWIPE	0.98	80.1	77.8	10.3	10.4	96.4	3.0
PRAAT	1.00	62.2	60.6	17.1	16.7	94.1	3.7

Table 2. Evaluation results on all the corpora.

For a given PEA, all statistics are means obtained on the whole of the three corpora. The results given in table 2 do not aim at a classification of the algorithms tested. They show that the performance differences are not as large as they look in the original publications, and they may encourage a detailed analysis of their respective strengths and weaknesses. The PRAAT settings that we adopted to remove the post-processing are clearly not adapted to this algorithm.

4 Multipitch case

In the multipitch case, several F0 streams need to be tracked simultaneously. This complicated task is usually performed by top-down approaches. These approaches use continuity hypotheses from one frame to the next or other general knowledge concerning the whole sequence. As there is also a lack of consensus in literature concerning these top-down processes, it is still difficult to evaluate their contributions to voicing and pitch estimation. That is why we are interested in a frame-to-frame evaluation without knowing more than the information extracted from the studied frame. We are aware of the fact that the better are the frame-level feature estimators, the simplest will the higher-level processes be.

4.1 Evaluation method

4.1.1 Database an reference annotations

To build a 2-speakers mixture, two 60 dB energy normalized signals are simply added. The energy normalization method is the "Scale Intensity..." PRAAT

function. Then, the 2-speakers mixtures databases are built by adding several possible signals combinations. The mixtures are cut to the shortest component. Three types of mixtures are available: the female/female mixtures, the male/male mixtures and the male/female mixtures. K contains 1411 seconds of speech mixtures, B contains 13693 seconds of speech mixtures, D contains 79869 seconds of speech mixtures for a total of 26 hours 22 minutes of speech mixtures artificially collected.

The reference annotations in the multipitch case are directly based on the monopitch reference annotations. The annotations consist in a simple concatenation of the monopitch reference annotations.

4.1.2 Quality measures

All the notations and definitions posed in 3.1.3 are still available in this section. In the monopitch case, a speech signal contains only one speaker so that it was not necessary to introduce the notion of F_0 stream. A F_0 stream corresponds to the set of F_0 values belonging to one speaker in a speech signal. In the multipitch case we have to introduce it due to the mixing of F_0 streams. One defines S_R the set of F_0 streams contained in the signal and S_R^S the particular F_0 stream of the s^{th} speaker. There are as many F_0 streams as speakers so there are N_S F_0 streams. An ideal PEAs should be able to reproduce the exact references annotations on his hypotheses output.

$$S_R = \left\{ S_R^1 ... S_R^{N_S} \right\} \text{ and } S_R^S = \left\{ R_S^1 ... R_S^{N_F} \right\} \qquad (13)$$

An F_0 hypothesis value should be associated with several reference F_0 values. Thus if some F0 reference values are equal in a same frame (crossing streams of F_0), PEAs probably return a single hypothesis for the ensemble of equals F_0 reference values. This multi-association allows the evaluation to deal with this problem. Symmetrically once a F_0 hypothesis value is associated to one or more F_0 reference values, the hypothesis F_0 value is locked and can not be associated anymore.

With the VuV definition adopted in 3.1.3, the VuV decision evaluation criteria remain the same in the multipitch case than in the monopitch case. One frame is voiced if one of her value is strictly positive. Mathematical definitions of V_{ok}, uV_{ok}, OVR an UVR remain unchanged. Nevertheless as the reference change, a new adequate VuV decision threshold has to be computed.

There is an evolution in the F_0 estimation evaluation criteria. Contrary to the monopitch case where a F_0 reference value is the same than a frame reference, in the multipitch case we can clearly distinguish between them and any simplification is no longer available.

The "in accordance" F_0 references values rate RECR given in Eq.(15) is the number of F_0 references values in accordance over the number of voiced F_0 reference values (VC_R defined in Eq.(2)). Eq.(14) provides the definition of RC_{ok} which is the set of F_0 reference values in accordance.

$$RC_{ok} = \left\{ (x,t) \middle| (\exists y \middle| R_x^t \approx H_y^t) \right\} \qquad (14)$$

$$RECR = \frac{\Omega[RC_{ok}]}{\Omega[VC_R \cap VC_H]} \qquad (15)$$

One can define the same rate for frames. We call REFR the "in accordance" reference frames rate. It corresponds to the number of reference frames in accordance over the number of voiced references frames. It is a first feature to quantify a PEA quality in reconstructing F_0 streams. Indeed higher is REFR higher is the number of frames where all F0 references values are associated with an F0 hypothesis value and better may be the F_0 streams tracking. Eq.(16) and (17) formalize this feature.

$$RF_{ok} = \left\{ t \middle| R_x^t \approx H_y^t) \right\} \qquad (16)$$

$$REFR = \frac{\Omega[RF_{ok}]}{\Omega[VF_R \cap VF_H]} \qquad (17)$$

The precision rate on F_0 values explained by Eq.(18) is the same than in Eq.(13) except that here it may exists more than one F_0 reference values.

$$PRR = \frac{\Omega[RC_{ok}]}{\sum_{t \in VF_R \cap VF_H} \Omega\left[\left\{ H_y^t > 0 \right\}\right]} \qquad (18)$$

5 Conclusions

In this paper we provided an evaluation methodology to compare PEAs not only on their F_0 estimation efficiency but also on the VuV decision which is a PEA step at least as important as the F_0 estimation. We illustrated this view in the monopitch case, by using several PEAs on a set of 3 different databases including the EGG signals, for which an automatic annotation protocol has been worked out. The methodology has been elaborated to extend easily to the multipitch case.

This work reinforces the idea that F_0 estimation and VuV decision are deeply related, in complex ways. Voicing was considered here as a two-state, binary problem. In the future it may become necessary to treat separately two notions of voicing, one at the signal level, essentially continuous, and the other, binary, incorporating some upper-level perceptive and linguistic considerations.

References

[1] de Cheveigné A., Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", *J. Acoust. Soc. Am.* 111, 1917-1930 (2002).

[2] Boersma P., "Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound", *IFA Proceedings* 17, 97-110 (1993).

[3] Camacho A., "SWIPE: a Sawtooth Waveform Inspired Pitch Estimator for Speech and Music", *Ph. Dissertation*, University Of Florida (2007).

[4] Liénard J-S., Signol F., Barras C., "Speech Fundamental Frequency Estimation Using the Alternate Comb", *INTERSPEECH 2007*, 2273-2276 (2007).

[5] Liénard J-S., Barras C., Signol F., "Using sets of combs to control pitch estimation errors", *ACOUSTICS 2008*, (2008).

[6] Secrest, B., Doddington, G., "An integrated pitch traking algorithm for speech systems", *Proc. ICASSP 1983*, 1352-1355 (1983).

Bibliographie

Gilles Adda, Martine Adda-Decker, Claude Barras, Philippe Boula de Mareüil, BenoÃ®t Habert, and Patrick Paroubek. Speech overlap and interplay with disfluencies in political interviews. In *International workshop on Paralinguistic Speech - between models and data. ParaLing 2007*, pages 41–46, Saarbrücken, Germany, 2007.

Peter F. Assmann. Fundamental frequency and the intelligibility of competing voices. In *14th International Congress of Phonetic Sciences*, pages 179–182, San Francisco, 1999.

Peter F. Assmann and Quentin Summerfield. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 88(2): 680–697, 1990.

B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

B. S. Atal and Lawrence R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):201–212, 1976.

F.R. Bach and M.I. Jordan. Discriminative training of hidden markov models for multiple pitch tracking [speech processing examples]. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2005*, volume 5, pages 489–492, Philadelphia, USA, 2005.

Bagshaw, Hiller, and Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *European Conference on Speech Communication and Technology. EUROSPEECH93*, pages 1003–1006, Berlin, Germany, 1993.

Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of multiple-F0 estimation and tracking systems. In *10th International Society for Music Information Retrieval Conference. ISMIR 2009*, pages 315–320, Kobe, Japan, 2009.

John G. Beerends and Adrianus J. M. Houtsma. Pitch identification of simultaneous diotic and dichotic two-tone complexes. *The Journal of the Acoustical Society of America*, 85(2):813–819, 1989.

J. Bird and C. J. Darwin. Effects of a difference in fundamental frequency in separating two sentences. In *Psychophysical and Physiological Advances in Hearing*, pages 263–269. A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis, London, whurr edition, 1997.

Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA proceedings*, volume 17, pages 97–110, 1993.

A. S. Bregman and J Campbell. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89(2):244–249, 1971.

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990.

J. P. L. Brokx and S. G. Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10:23–26, 1982.

Judith C. Brown and Miller S. Puckette. Calculation of a "narrowed" autocorrelation function. *The Journal of the Acoustical Society of America*, 85(4):1595–1601, 1989.

Judith C. Brown and Bin Zhang. Musical frequency tracking using the methods of conventional and "narrowed" autocorrelation. *The Journal of the Acoustical Society of America*, 89(5):2346–2354, 1991.

Arturo Camacho. *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD thesis, University of Florida, 2007.

Arturo Camacho. Detection of Pitched/Unpitched sound using pitch strength clustering. In *International Society of Music Information Retrieval. ISMIR 2008*, pages 533–537, 2008.

Arturo Camacho and John G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652, 2008.

G. Cauwenberghs. Monaural separation of independent acoustical components. In *IEEE International Symposium on Circuits and Systems. ISCAS 1999*, volume 5, pages 62–65, 1999.

Wei-Chen Chang, Alvin W. Y. Su, Chunghsin Yeh, Axel Röebel, and Xavier Rodet. Multiple-F0 tracking based on a high-order HMM,. In *Proceedings of the 11th International Conference on Digital Audio Effects. DAFx-08*, Finland, 2008.

E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

E. Colin Cherry and W. K. Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 26(4):554–559, 1954.

Mads Christensen and Andreas Jakobsson. *Multi-Pitch Estimation*. Morgan and Claypool Publishers, 2009.

M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP 97*, volume 2, pages 863–866, 1997.

Martin Cooke. Glimpsing speech. *Journal of Phonetics*, 31(3-4):579–584, 2003.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120 (5):2421–2424, November 2006.

A. Coy and J. Barker. Recognising speech in the presence of a competing speaker using a "Speech fragment decoder". In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2005*, volume 1, pages 425–428, 2005.

John F. Culling and C. J. Darwin. Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0. *The Journal of the Acoustical Society of America*, 93(6):3454–3467, 1993.

Adrien Daniel, Valentin Emiya, and Bertrand David. Perceptually-based evaluation of the errors usually made when automatically transcribing music. In *International Conference on Music Information Retrieval. ISMIR 2008*, pages 550–555, 2008.

C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology*, 33(2):185–207, 1981.

Christopher Darwin. Pitch and auditory grouping. In *Pitch Neural Coding and Perception*, pages 278–305. Christopher J. Plack, Andrew J. Oxenham, Richard R. Fay, Arthur N. Popper, springer handbook of auditory research edition, 2005.

Jr. Davenport. An experimental study of Speech-Wave probability distributions. *The Journal of the Acoustical Society of America*, 24(4):390–399, 1952.

Alain de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *The Journal of the Acoustical Society of America*, 93(6):3271–3290, 1993.

Alain de Cheveigné. Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, 103(3):1261–1271, 1998.

Alain de Cheveigne. Multiple F0 estimation. In *Computational Auditory Scene Analysis*, pages 45–79. Deliang Wang and Guy J. Brown, John Wiley and sons edition, 2006.

Alain de Cheveigné and Hideki Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27(3-4):175–185, 1999.

Alain de Cheveigné and Hideki Kawahara. Comparative evaluation of f0 estimation algorithms. In *Proceedings of Eurospeech 2001*, volume 4, pages 2451–2454, 2001.

Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

Alain de Cheveigné, Hideki Kawahara, Minoru Tsuzaki, and Kiyoaki Aikawa. Concurrent vowel identification. i. effects of relative amplitude and f0 difference. *The Journal of the Acoustical Society of America*, 101(5):2839–2847, 1997.

P. N. Denbigh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Communication*, 11(2-3):119–125, 1992.

P. Depalle and T. Helie. Extraction of spectral peak parameters using a short-time fourier transform modeling and no sidelobe windows. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA 1997*, 1997.

Pierre Divenyi. *Speech Separation by Humans and Machines*. New York, kluwer academic publishers edition, 2004.

Scott C. Douglas and Xiaoan Sun. Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, 39(1-2):65–78, 2003.

Boris Doval. *Estimation de la fréquence fondamentale des signaux sonores*. PhD thesis, Université Paris VI, March 1994.

W. J. Dowling. Rhythmic fission and perceptual organization. *The Journal of the Acoustical Society of America*, 44(1):369 (A), 1968.

J.-L. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 169–172, 2008.

James P. Egan, Edward C. Carterette, and Edward J. Thwing. Some factors affecting Multi-Channel listening. *The Journal of the Acoustical Society of America*, 26(5):774–782, 1954.

Valentin Emiya. *Transcription automatique de la musique de piano*. PhD thesis, Université Paris VI - Pierre et Marie Curie, 2008.

Hugo Fastl and Gerhard Stoll. Scaling of pitch strength. *Hearing Research*, 1(4):293–301, 1979.

P. Fernandez-Cid and F.J. Casajus-Quiros. Multi-pitch estimation for polyphonic musical signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1998*, volume 6, pages 3565–3568, 1998.

J. M. Festen and R. Plomp. Relations between auditory functions in impaired hearing. *The Journal of the Acoustical Society of America*, 73(2):652–662, 1983.

Joost M. Festen and Reinier Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736, 1990.

Nuno Fonseca and Anibal Ferreira. Measuring music transcription results based on a hybrid decay/sustain evaluation. In *Proceedings of the 7th Conference of European Society for the Cognitive Science of Music. ESCOM 2009*, pages 119–124, Finland, 2009.

D. Friedman. Multichannel zero-crossing-interval pitch estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1979*, volume 4, pages 764–767, 1979.

Etienne Gaudrain. *Rôle de la ségrégation séquentielle pour la séparation de voix concurrentes*. PhD thesis, Université Lyon 1 - Claude Bernard, April 2008.

Etienne Gaudrain, Nicolas Grimault, Eric W. Healy, and Jean-Christophe Béra. Effect of spectral smearing on the perceptual segregation of vowel sequences. *Hearing Research*, 231(1-2):32–41, 2007.

David Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical report, Department of Computer Science, University of Regina, Canada, November 2003.

B R Glasberg and B C Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–38, 1990.

Dan Gnansia. Modèle auditif en temps réel. Mémoire de master sciences et technologie, Département d'étude cognitive, Equipe Audition, Paris, 2005.

S. Godsill and M. Davy. Bayesian harmonic models for musical pitch estimation and analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2002*, volume 2, pages 1769–1772, Orlando, Florida, USA, 2002.

Bernard Gold. Computer program for pitch extraction. *The Journal of the Acoustical Society of America*, 34(7):916–921, 1962.

M. Goto. A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000*, volume 2, pages 757–760, 2000.

Masataka Goto. A Predominant-F0 estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2001*, 5:3365–3368, 2001.

Y.H. Gu and W.M.G. van Bokhoven. Co-channel speech separation using frequency bin non-linear adaptive filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1991*, volume 2, pages 949–952, Toronto, Ontario, Canada, 1991.

Joseph W. Hall and Mariano A. Fernandes. Temporal integration, frequency resolution, and off-frequency listening in normal-hearing and cochlear-impaired listeners. *The Journal of the Acoustical Society of America*, 74(4):1172–1177, 1983.

B. Hanson and D. Wong. The harmonic magnitude suppression (EMS) technique for intelligibility enhancement in the presence of interfering speech. In *IEEE International Conference on Acoustics,*

Speech, and Signal Processing. ICASSP 1984, volume 9, pages 65–68, San Diego, California, USA, 1984.

Dik J. Hermes. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech & Language*, 2009.

Wolfgang Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer-Verlag, Germany, heidelberg edition, 1983.

A. J. M. Houtsma and J. Smurzynski. Pitch identification and discrimination for complex tones with many harmonics. *The Journal of the Acoustical Society of America*, 87(1):304–310, 1990.

Guoning Hu and DeLiang Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15(5):1135–1150, 2004.

Staffan Hygge, Jerker Ronnberg, Birgitta Larsby, and Stig Arlinger. Normal-Hearing and Hearing-Impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research*, 35(1):208–215, 1992.

Luc M. Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with an auditory model. *The Journal of the Acoustical Society of America*, 91(6):3511–3526, 1992.

Özgür Izmirli and Semih Bilgen. A multiple fundamental frequency tracking algorithm. In *3èmes Journées d'Informatique Musicale. JIM96*, France, 1996.

Jiang Jong-Shiann and M.A. Ingram. Robust detection of number of sources using the transformed rotational matrix. In *IEEE Wireless Communications and Networking Conference. WCNC 2004*, volume 1, pages 501–506, 2004.

Kameoka, Nishimoto, and Sagayama. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):982–994, 2007.

Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Accurate f0 detection algorithm for concurrent sounds based on EM algorithm and information criterion. In *Proceedings of Special Worshop in MAUI .SWIM*, 2004a.

Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Multi-Pitch trajectory estimation of concurrent speech based on harmonic GMM and nonlinear kalman filtering. In *International Conference on Spoken Language Processing. ICSLP 2004*, Jeju Island, Korea, 2004b.

Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Multi-pitch detection algorithm using constrained gaussian mixture model and information criterion for simultaneous speech. In *Speech Prosody 2004*, pages 533–536, 2004c.

M. Karjalainen and T. Tolonen. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1999*, volume 2, pages 929–932, 1999.

Kavita Kasi. *Yet Another Algorithm for Pitch Tracking (YAAPT)*. PhD thesis, Faculty of Old Dominion University, India, 2002.

Anssi Klapuri. Number theoretical means of resolving a mixture of several harmonics sounds. In *European Signal Processing Conference. EUSIPCO*, Rhodes, Greece, 1998.

Anssi Klapuri. A perceptually motivated multiple f0 estimation method. In *Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA 2005*, pages 291–294, New Paltz, New York, USA, 2005.

Anssi Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *International Conference on Music Information Retrieval. ISMIR 2006*, pages 216–221, Victoria, Canada, 2006.

Anssi Klapuri, Tuomas Virtanen, and Jan-Markus Holm. Robust multipich estimation for the analysis and manipulation of polyphonic musical signals. In *Digital Audio Effects. DAFx 2000*, pages 141–146, Verona, Italy, 2000.

A.P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing.*, 11(6):804–816, 2003.

Peter Ladefoged and D. E. Broadbent. Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29(1):98–104, 1957.

R.H. Lambert and A.J. Bell. Blind separation of multiple speakers in a multipath environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1997*, volume 1, pages 423–426, Munich, Germany, 1997.

J.P. LeBlanc and P.L. De Leon. Speech separation by kurtosis maximization. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1998*, volume 2, pages 1029–1032, 1998.

J. Licklider. A duplex theory of pitch perception. *Experientia*, 7(4):128–134, 1951.

Jean-Sylvain Liénard, François Signol, and Claude Barras. Speech fundamental frequency estimation using the alternate comb. In *Interspeech 2007*, Antwerpen, Belgium, 2007.

Jean-Sylvain Liénard, Claude Barras, and François Signol. Using sets of combs to control pitch estimation errors. *Proceedings of Meetings on Acoustics*, 4(1), 2008.

I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, and I. Sainz. Evaluation of pitch detection algorithms under real conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2007.*, volume 4, pages 1057–1060, 2007.

R. Lyon. A computational model of binaural localization and separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1983*, volume 8, pages 1148–1151, 1983.

J. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics.*, 20(5):367–377, 1972.

P. Martin. Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne. In *12ème Journées d'Etudes sur la Parole. SFA*, Montréal, 1981.

P. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1982.*, volume 7, pages 180–183, 1982.

P. Martin. Peigne et brosse pour f0 : Mesure de la fréquence fondamentale pour alignement de spectres séquentiels. In *23èmes Journées d'Études sur la parole. JEP 2000*, pages 245–248, Aussois, 2000.

P. Martin. Crosscorrelation of adjacent spectra enhances fundamental frequency estimation. In *Interspeech 2008*, Brisbane, Australia, 2008a.

P. Martin. A fundamental frequency estimator by crosscorrelation of adjacent spectra. In *Speech Prosody 2008*, Campinas, Brazil, 2008b.

R. McAulay and T. Quatieri. Speech analysis/Synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

R.J. McAulay and T.F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1990*, volume 1, pages 249–252, 1990.

C. McGonegal, L. Rabiner, and A. Rosenberg. A semiautomatic pitch detector (SAPD). *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 23(6):570–574, 1975.

J. Denis McKeown. Perception of concurrent vowels: The effect of varying thei relative level. *Speech Communication*, 11(1):1–13, 1992.

J. Denis McKeown and Roy D. Patterson. The time course of auditory segregation: Concurrent vowels that vary in duration. *The Journal of the Acoustical Society of America*, 98(4):1866–1877, 1995.

Philip McLeod and Geoff Wyvill. A smater way to find pitch. In *International Computer Music Conference. ICMC 2005*, Barcelona, Spain, 2005.

Ray Meddis and Michael J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. *The Journal of the Acoustical Society of America*, 89(6): 2866–2882, 1991.

Ray Meddis and Michael J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 91(1):233–245, 1992.

G. A. Miller. The masking of speech. *Psychological Bulletin*, 44(2):105–129, 1947.

George A. Miller and George A. Heise. The trill threshold. *The Journal of the Acoustical Society of America*, 22(5):637–638, 1950.

B. C. Moore. Frequency selectivity and temporal resolution in normal and hearing-impaired listeners. *British Journal of Audiology*, 19(3):189–201, 1985.

E. Mousset, W.A. Ainsworth, and J.A.R. Fonollosa. A comparison of several recent methods of fundamental frequency and voicing decision estimation. In *Fourth International Conference on Spoken Language. ICSLP 1996*, volume 2, pages 1273–1276, 1996.

A. Michael Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41 (2):293–309, February 1967.

Rasmus Kongsgaard Olsson and Lars Kai Hansen. Estimating the number of sources in a noisy convolutive mixture using BIC. In *Independent Component Analysis and Blind Signal Separation*, pages 618–625. 2004.

A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pages 90–93, 2005.

Thomas W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60(4):911–918, 1976.

R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand. Complex sounds and auditory images. In *Auditory Physiology and Perception*, pages 429–446. Oxford, y cazals, l. demany, k. horner, pergamon edition, 1992.

Roy D. Patterson. Auditory images:How complex sounds are represented in the auditory system. *Acoustical Science and Technology, The Acoustical Society of Japan*, 21(4):183–190, 2000.

A. Pertusa and J.M. I nesta. Multiple fundamental frequency estimation using gaussian smoothness. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2008*, pages 105–108, 2008.

Julien Pinquier, Jean-Luc Rouas, and Régine André-Obrecht. Robust speech/music classification in audio documents. In *International Conference on Spoken Language Processing. ICSLP 2002. Interspeech 2002*, Denver, Colorado, USA, 2002.

F. Plante, G. F. Meyer, and W. A. Ainsworth. A pitch extraction reference database. In *ESCA EUROSPEECH 95*, pages 837–840, Madrid, 1995.

G.E. Poliner, D.P.W. Ellis, A.F. Ehmann, E. Gomez, S. Streich, and Beesuan Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256, 2007.

L. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 25(1):24–33, 1977.

L. Rabiner and M. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 25(4):338–343, 1977.

L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5): 399–418, 1976.

Mohammad H. Radfar, Richard M. Dansereau, and Abolghasem Sayadiyan. Monaural speech segregation based on fusion of source-driven with model-driven techniques. *Speech Communication*, 49 (6):464–476, 2007.

R. Raghuram. Pitch and voicing determination of speech signals. Technical report, Deparment of Electrical Engineering, Madras, India, May 2002.

Aarthi M. Reddy and Bhiksha Raj. Soft mask estimation for single channel speaker separation. In *SAPA 2004*, Korea, 2004.

Ronald A. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40(10-12):1469–1487, 2000.

Nicoleta Roman, DeLiang Wang, and Guy J. Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley. Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 22(5):353–362, 1974.

Jean Rouat, Yong Chun Liu, and Daniel Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21(3):191–207, 1997.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Harmonic-Temporal clustering of speech for single and multiple f0 contour estimation in noisy environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, volume 4, pages 1053–1056, 2007a.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Single and multiple f0 contour estimation through parametric spectrogram modeling of speech in noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing.*, 15(4):1135–1145, 2007b.

Sam T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Eurospeech 2003*, pages 1009–1012, Geneva, Switzerland, 2003.

O. Salor, M. Demirekler, and U. Orguner. An efficient algorithm for pitch determination of speech signals - kalman filter approach. In *IEEE 14th Signal Conference on Processing and Communications Applications.*, pages 1–4, 2006.

M. T. M. Scheffers. *Sifting Vowels: Auditory Pitch Analysis and Sound Segregation.* PhD thesis, Rijksuniversiteit te Groningen, 1983.

M. R. Schroeder. Period histogram and product spectrum: New methods for Fundamental-Frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

B. Secrest and G. Doddington. An integrated pitch tracking algorithm for speech systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1983*, volume 8, pages 1352–1355, 1983.

X Serra. Musical sound modeling with sinusoids plus noise. *Musical Signal Processing*, pages 497–510, 1997.

L. Siegel and A. Bessey. Voiced/Unvoiced/Mixed excitation classification of speech. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 30(3):451–460, 1982.

François Signol. Suivi d'un flux de f0 par programmation dynamique dans un mélange monaural de signaux de parole. In *Rencontres des Jeunes Chercheurs en Parole. RJCP07*, Paris, France, 2007.

François Signol, Claude Barras, and Jean-Sylvain Liénard. Evaluation of the pitch estimation algorithms in the monopitch and multipitch cases (Abstract). *The Journal of the Acoustical Society of America*, 123(5):3077, 2008.

M. Slaney and R.F. Lyon. A perceptual pitch detector. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 1990*, volume 1, pages 357–360, 1990.

Malcom Slaney. Lyon's cochlear model. Technical Report 13, Apple, Cupertino, California, USA, 1988.

Malcom Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Apple Computer Technical Report 35, Perception Group - Advanced Technology Group, 1993.

Richard J. Stubbs and Quentin Summerfield. Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 84(4):1236–1249, 1988.

Richard J. Stubbs and Quentin Summerfield. Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 87(1):359–372, 1990.

Xuejing Sun. A pitch determination algorithm based on Subharmonic-to-Harmonic ratio. In *International Conference on Spoken Language Processing. ICSLP 2000*, pages 676–679, Beijing, CHina, 2000.

Xuejing Sun. *The Determination, Analysis, and Synthesis of Fundamuntal Frequency.* PhD thesis, Northwestern University, Evanston, Illiois, December 2002.

D. Talkin. A robust algorithm for ptich tracking (RAPT). In *Speech Coding and Synthesis*. Elsevier, New York, w. b. kleijn & k. k. paliwal (eds.) edition, 1995.

Lee Te-Won, A. Ziehe, R. Orglmeister, and T. Sejnowski. Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1998*, volume 2, pages 1249–1252, 1998.

Ernst Terhardt, Gerhard Stoll, and Manfred Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, 71(3):679–688, 1982a.

Ernst Terhardt, Gerhard Stoll, and Manfred Seewann. Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. *The Journal of the Acoustical Society of America*, 71 (3):671–678, 1982b.

Emmanuel Tessier and Frédéric Berthommier. A model of the cumulative effect of pitch and interaural delay differences for double vowel segregation. In *ICSP 1997*, Séoul, 1997.

D.L. Thomson and R. Chengalvarayan. Use of periodicity and jitter as speech recognition features. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1998*, volume 1, pages 21–24, 1998.

T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing.*, 8(6):708–716, 2000.

A.J.W. van der Kouwe, D. Wang, and G.J. Brown. A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech and Audio Processing.*, 9(3):189–195, 2001.

Leon P. A. S. van Noorden. *Temporal Coherence in the Perception of Tone Sequences.* PhD thesis, Eindhoven University of Technology, The Netherlands, 1975.

Leon P. A. S. van Noorden. Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *The Journal of the Acoustical Society of America*, 61(4):1041–1045, 1977.

Peter Veprek and Michael S. Scordilis. Analysis, enhancement and evaluation of five pitch determination techniques. *Speech Communication*, 37(3-4):249–270, 2002.

Tuomas Virtanen and Anssi Klapuri. Separation of harmonics sounds using multipitch analysis and iterative estimation. In *Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA 2001*, New Paltz, New York, 2001.

Srikanth Vishnubhotla and Carol Espy-Wilson. An algorithm for Multi-Pitch tracking in Co-Channel speech. In *ICSLP 2008. Interspeech 2008*, Brisbane, Australia, 2008.

DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley-IEEE Press, 2006.

W.Q. Wang, W. Gao, and D.W. Ying. A fast and robust speech/music discrimination approach. In *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information,*

Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, volume 3, pages 1325–1329, 2003.

M. Weintraub. A computational model for separating two simultaneous talkers. In *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 1986*, volume 11, pages 81–84, 1986.

Ron J. Weiss and Daniel P. W. Ellis. Monaural speech separation using Source-Adapted models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pages 114–117, 2007.

John Woodruff, Yipeng Li, and Wang DeLiang. Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation. In *International Conference on Music Information Retreival. ISMIR 2008*, Philadelphia, USA, 2008.

Mingyang Wu and DeLiang Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing.*, 14(3):774–784, 2006.

Mingyang Wu, DeLiang Wang, and G.J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11(3):229–241, 2003.

Jian-Wu Xu and J.C. Principe. A novel pitch determination algorithm based on generalized correlation function. In *IEEE Workshop on Machine Learning for Signal Processing.*, pages 270–275, 2007.

Chunghsin Yeh. *Extraction de fréquences fondamentales multiples dans des enregistrements polyphoniques.* PhD thesis, Université Paris VI - Pierre et Marie Curie, 2008.

Chunghsin Yeh and Alex Röbel. A new score function for joint evaluation of multiple f0 hypotheses. In *7th Conference on Digital Audio Effects (Dafx'04)*, Naples, Italy., 2004.

Duan Zhiyao, Zhang Dan, Zhang Changshui, and Shi Zhenwei. Multi-Pitch estimation based on partial event and support transfer. In *IEEE International Conference on Multimedia and Expo.*, pages 216–219, 2007.

U. Tilmann Zwicker. Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication*, 3(4):265–277, 1984.