UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

MÉMOIRE PRÉSENTÉ À

L'UNIVESRITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE DE

LA MAÎTRISE EN INGÉNIERIE

PAR

HASSAN EZZAIDI

DÉTECTION DE LA DOUBLE PAROLE DANS LE CONTEXTE DE RADIOTÉLÉPHONE MAIN-LIBRE EN VÉHICULE

DÉCEMBRE 97

Bibliothèque
Paul-Émile-Boulet

UQAC
Université du Québec
à Chicoutimi

Mise en garde/Advice

Afin de rendre accessible au plus grand nombre le résultat des travaux de recherche menés par ses étudiants gradués et dans l'esprit des règles qui régissent le dépôt et la diffusion des mémoires et thèses produits dans cette Institution, **l'Université du Québec à Chicoutimi (UQAC)** est fière de rendre accessible une version complète et gratuite de cette œuvre.

Motivated by a desire to make the results of its graduate students' research accessible to all, and in accordance with the rules governing the acceptation and diffusion of dissertations and theses in this Institution, the **Université du Québec à Chicoutimi (UQAC)** is proud to make a complete version of this work available at no cost to the reader.

L'auteur conserve néanmoins la propriété du droit d'auteur qui protège ce mémoire ou cette thèse. Ni le mémoire ou la thèse ni des extraits substantiels de ceux-ci ne peuvent être imprimés ou autrement reproduits sans son autorisation.

The author retains ownership of the copyright of this dissertation or thesis. Neither the dissertation or thesis, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

TABLE DES MATIÈRES

GLOSSAIRE

A.D.F : Algorithme de détection du fondamental.

A(k) : Coefficients du modèle A.R.

A.R : Analyse autorégressive.

Cepstre : Technique utilisée dans les systèmes d'analyse et de reconnaissance de parole. Les coefficients cepstraux correspondent à la transformée de Fourier inverse du logarithme du spectre d'amplitude.

C(k): Coefficients du filtre peigne pour annuler le fondamental et les harmoniques.

D.A.V : Détecteur d'activité vocale.

DbTlk : Égal à 1 si on est dans une situation de la double parole, sinon il est égal à zéro.

D.A.V.Rx : *Détection d'activité vocale du locuteur lointain.*

D.A.V.Tx : Détection d'activité vocale du locuteur local.

D.D.P : Détection de l'activité de double parole.

DSP : Processeur dédié aux traitements du signal digital.

Dyn : Opérateur lié à l'énergie proposé par Jean Rouat pour estimer les changements d'enveloppes à la sortie d'un banc de filtres cochléaires.

E.F : Estimation du fondamental.

EnEcho : Énergie de l'écho résiduel.

EnLoc : Énergie du locuteur local.

Filtrage Inverse: Il consiste à diviser la transformée de fourier du signal vocal par la réponse en fréquence du conduit vocal estimée via une analyse autorégressive.

Fréquence glottale : Fréquence de vibration de la glotte.

Fréquence glottique : Fréquence de vibration de la glotte.

Hauteur tonale : Hauteur perçue d'un son voisé.

H.P : Haut parleur.

Locuteur local : Locuteur situé au volant du véhicule.

Locuteur lointain : Le correspondant lointain dont le signal provient du haut-parleur.

Locuteur proche : Synonyme du locuteur local.

Locuteur receive : Synonyme du locuteur lointain.

LPC : Technique de prédiction linéaire utilisée dans les systèmes du codage/décodage en parole (``Lineair Predictive Coding``).

Préaccentuation : un filtrage passe-haut utilisé pour atténuer les composantes en basse fréquence et rehausser les composantes en haute fréquence.

RIF : un filtre à réponse impulsionnelle finie.

RII : un filtre à réponse impulsionnelle infinie.

RLE : Rapport Local à Écho. C'est un critère de mesure qui définit en dB le rapport entre l'énergie du locuteur local et l'énergie de l'écho résiduel pour chaque trame dans les situations où le locuteur lointain est actif.

Système Ampex : un système de détection de la hauteur tonale proposé par Van Immersel et Martens (1992).

Système Algomai94 : un système de détection de la hauteur tonale proposé par Rouat et al (1997).

Signal haut-parleur : Signal propre du locuteur lointain.

Signal lointain : Synonyme du signal haut-parleur.

Signal microphone : Signal capté par le microphone de la voiture (signal du locuteur proche + écho).

Signal proche : Signal propre du locuteur proche.

Signal pseudorésiduel : Dans le système main-libre, le pseudorésiduel est obtenu en effectuant une analyse autorégressive sur le signal haut-parleur et un filtrage inverse du signal microphone en utilisant les coefficients de l'analyse du signal haut-parleur.

Signal receive : Synonyme du signal lointain.

Signal résiduel : Dans la technique de LPC le résiduel est obtenu en effectuant une analyse autorégressive et un filtrage inverse sur un même signal.

Teager : Opérateur énergie proposé par Kaiser à partir des travaux de Teager afin d'estimer l'énergie d'un signal en bande étroite.

Tx(n) : échantillon du signal microphone à l'instant n.

Voie haut-parleur : Synonyme du signal du haut-parleur.

Voie microphone : Synonyme du signal microphone.

Voie proche : Synonyme du signal proche.

Voie ``receive`` : Synonyme du signal ``receive``.

δ_V : Seuil de voisement dans le système Ampex.

δ_T : Seuil d'évidence dans le système Ampex.

LISTE DES FIGURES

LISTE DES TABLEAUX

SOMMAIRE

L'environnement très bruité, le couplage du haut parleur (H.P) avec le microphone, ainsi que le problème de gain du H.P dans un contexte de radio mobile en véhicule font l'objet de plusieurs travaux en télécommunications. Des algorithmes pour réduire le bruit et pour annuler l'écho (A.E) ont été proposés dans la littérature scientifique. En général, tous les algorithmes d'annulation d'écho sont basés sur des filtres à coefficients adaptatifs qui fonctionnent assez bien. Cependant, la façon d'adapter les coefficients influence terriblement les performances. Nous proposons ici une technique qui permet de mieux détecter les moments de mises à jour des coefficients des filtres (paramètres). Normalement, ces filtres ne doivent pas être adaptés lorsque le locuteur local parle (locuteurs installés en véhicule). On a généralement recourt à des algorithmes à base d'énergie afin de séparer la voix du locuteur local de celle du correspondant lointain. Nous proposons une technique, qui au lieu de l'énergie, utilise un détecteur de hauteur tonale (D.H.T) et qui est basé sur un modèle auditif (Rouat et al., Speech Comm. Jour., 1997). Ce DHT est introduit en cascade avec le filtre auto-regressif (A.R.) déjà inclus dans le système. Conjointement, le DHT et le filtre A.R. nous ont permis d'annuler le fondamental, les composantes harmoniques et la contribution vocale du locuteur lointain sur le canal microphone.

REMERCIEMENT

Je tiens à remercier mon directeur de recherche, M. Jean Rouat professeur à l'U.Q.A.C, qui a contribué énormément à ma formation dans le domaine de traitement de la parole et qui m'a beaucoup aidé pour la rédaction de mon mémoire. Je le remercie également de m'avoir permis de passer un stage intéressant à Paris, chez Alcatel Phones Mobile.

Je désire remercier Ivan Bourmeyster Chef d'Études de Fonctions Vocales qui m'a initié à la mise au point des algorithmes en virgule fixe et qui a essayé de rendre mon séjour agréable à Paris. Je désire remercier également, tout le personnel du groupe d'Études de Fonctions Vocales qui m'a expliqué le fonctionnement du système mobile main-libre en véhicule et aux techniques de traitements utilisées actuellement dans l'industrie de télécommunication.

Également je remercie tous le personnel de l'U.Q.A.C qui a contribué directement ou indirectement à la réalisation de ce travail (Luc Morin, Danny Ouellet, Chantale Dumas,..).

Finalement, je remercie tous les membres de ma famille qui m'ont donné leur soutien et leur encouragement. Je remercie en particulier ma chère mère Chinig Aziza , mon frère M'hammed et ma petite soeur Madiha.

CHAPITRE 1

INTRODUCTION

1.1 Problématique

Une récente étude au Canada a démontré que 30% des accidents sont dus à l'usage d'un téléphone mobile (*voir le Point n° 1275*). Pour donner plus de sécurité aux passagers et au chauffeur de la voiture, une nouvelle loi en Europe entrera en vigueur pour remplacer le terminal mobile par un ensemble fixe. Ceci a exigé des concepteurs et des fabriquants d'optimiser leurs algorithmes pour livrer un nouveau produit hautement performant, concurrentiel et bien adapté à ce nouvel environnement. Les fabriquants de terminaux mobiles ont mis au point un système dit ``*main libre*`` pour continuer l'usage de leur produit en véhicule avec plus de sécurité. Différents systèmes ont été proposés par les fabriquants de téléphone : ceux à un prix très accessible comme le système dit ``*système mixte*`` et d'autres à un prix élevé tel que le système dit ``*système à installation complète*``. Malheureusement, les systèmes mains libres actuels ne permettent pas aux interlocuteurs de discuter

simultanément afin d'éviter les effets de l'écho et du couplage entre le microphone et le haut-parleur (la nécessité d'utiliser des algorithmes hautement robustes pour éviter ces problèmes). À l'exception d'Alcatel qui propose un système ``*full duplex*`` et qui utilise des algorithmes réducteurs de souffle (énergie) permettant à deux interlocuteurs de converser simultanément (*voir le Point n° 1280, Mars 1997*).

Dans le contexte de la radio téléphonie mobile et main libre en véhicule, l'insonorisation de l'habitacle (plusieurs sources de bruit) et le phénomène d'écho relié à la réflexion du son sur les parois de la voiture étaient et sont encore les grands problèmes techniques à résoudre. S'ajoute à ceci, le problème du gain du haut parleur (H.P) du système mobile et le problème de retard lié au canal de transmission.

Généralement, les systèmes intègrent différents algorithmes relativement efficaces pour supprimer l'écho et atténuer le bruit. Malheureusement, la performance de ces algorithmes est très dépendante de la façon dont on détecte l'activité vocale entre le locuteur lointain (le signal vocal provenant du haut-parleur) et le locuteur installé en véhicule (le signal à transmettre par le système). Dans les systèmes actuels, l'activité vocale pour chaque locuteur est évaluée par des algorithmes qui fonctionnent à base d'énergie. Ces algorithmes d'activité vocale fonctionnent bien pour la parole propre où le

rapport signal bruit est grand (bon). Or, l'environnement en voiture est très bruité (moteur, secousse, vent,..) et par conséquent la détection de l'activité vocale se trouve considérablement dégradée.

1.2 Solutions Proposées

Pour remédier à cette situation, nous proposons de remplacer le détecteur de l'activité vocale à base d'énergie par un détecteur basé plutôt sur l'information de la fréquence glottique (fréquence fondamentale). L'estimation de la fréquence glottique (E.F) est fournie par un algorithme basé sur un modèle auditif reconnu pour sa robustesse en milieu bruité et non bruité. Cet algorithme a été utilisé en conjonction avec une analyse autorégressive (A.R.).

Comme première stratégie, on a estimé la fréquence fondamentale sur le canal de réception (signal haut-parleur) et sur le canal microphone où les interférences entre les deux locuteurs peuvent apparaître. En comparant les valeurs estimées de la fréquence fondamentale, on doit normalement prévoir l'activité de chaque locuteur. Cette démarche a été analysée et interprétée selon différentes combinaisons A.R. et E.F.

Dans la deuxième stratégie, on estime uniquement la fréquence fondamentale sur la voie de réception. Cette information va servir à la synthèse d'un filtre peigne. À l'aide d'un filtrage inverse, on supprime la contribution vocale et glottale du locuteur lointain sur du signal microphone. Cette

démarche est apparue plus prometteuse. En effet, un gain discriminatoire de 3 dB en moyenne a été obtenu par rapport au système de référence. Le système de référence n'inclut pas l'algorithme d'estimation du fondamental.

La base de donnée utilisée, est fournie par Alcatel Mobile Phones . Elle est enregistrée de façon synchrone entre les deux locuteurs (lointain, local) dans divers états et conditions de déplacement de la voiture.

Enfin, on va discuter de l'implémentation en temps réel et de l'analyse de la complexité de l'algorithme. L'évaluation et la comparaison des performances seront présentées également.

1.3 Organisation du mémoire

Le chapitre 2, donnera quelques généralités du signal vocal. On donnera une bibliographie des travaux concernant la détermination du fondamental. À la fin du chapitre, on fera une introduction des méthodes de prédiction linéaire et du cepstre tout en donnant quelques références. On trouvera l'introduction au fonctionnement du système main-libre dans le chapitre 3. On fera une revue des travaux d'annulation d'écho et de suppression du bruit. La formulation du problème et les solutions proposées seront traitées au chapitre 4. On essayera de présenter clairement la méthodologie employée pour chaque stratégie énoncée. Au chapitre 5, on décrira la base de donnée utilisée pour fin d'expérimentation. On définira les critères d'évaluation et on discutera des

résultats pour chacune des méthodes proposées. Le chapitre 6, discutera de l'implémentation en virgule fixe de l'algorithme d'estimation du fondamental et évaluera les performances de cet algorithme en le comparant à d'autres systèmes reconnus par leurs robustesses dans ce domaine . Enfin, on donnera au chapitre 7 une conclusion générale du mémoire .

Figure 1 : Le système mobile main libre

CHAPITRE 2

ANALYSE DU SIGNAL VOCAL

Dans ce chapitre, on décrira le fonctionnement des principaux organes d'articulation responsables du mécanisme de la production. On donnera brièvement une description de l'appareil auditif. Ensuite, on donnera une définition de la hauteur tonale en se référant à une série d'articles scientifiques. On décrira le système de détection de la hauteur tonale en virgule flottante utilisé pour réaliser les simulations proposées au chapitre 4 et nécessitant la détection de la hauteur tonale. Finalement, on donnera un exposé concis du modèle de production le plus utilisé (modèle autorégressif) et on parlera aussi des outils de traitements numériques les plus utilisés pour estimer la hauteur tonale et la résonance du conduit vocal (Cesptre et LPC). La dernière section, sera consacrée à définir les opérateurs non linéaires Teager et Dyn qui seront utilisés ultérieurement pour trouver une solution à la problématique de la double parole dans le contexte radiotéléphone main-libre en véhicule. Pour résumer, ce chapitre nous donnera une description de la hauteur tonale et

nous familiarisera avec quelques outils de traitement numérique qu'on utilisera dans les prochains chapitres.

2.1 Introduction générale

la parole est le moyen naturel pour la communication entre les humains. Elle se caractérise des autres sons par l'intelligibilité de son contenu, sa richesse en information et sa redondance. La recherche en parole implique la coopération entre plusieurs disciplines : neurosciences, linguistiques, génies (traitement numérique du signal), etc... La citation de Gunnar Fant dans l'ouvrage [J. P. Tubach et al, 1989] relève l'aspect multidisciplinaire de la parole:

> Les techniques liées à la communication parlée connaissent actuellement un développement très actif, en liaison avec l'informatique, et présentent un potentiel important dans le domaine de l'interaction homme-machine. Nous voulons communiquer avec nos ordinateurs de la façon la plus naturelle, c'est à dire en utilisant le langage parlé, pour faciliter et accélérer l'interaction et l'échange d'informations. Nous cherchons à rendre ces machines accessibles par la voix, au téléphone, pour que l'on puisse accéder à l'information sans avoir besoin d'un clavier et d'un écran de visualisation. Ces techniques d'entrées/sorties vocales sont également très importantes pour beaucoup d'applications d'aide aux handicapés.
>
> L'histoire de cette recherche nous a apporté un enseignement : le traitement de parole n'est pas simplement affaire d'ingénieurs. Il y a de vastes domaines de connaissances fondamentales, à propos de la production et de la perception de la parole, de la linguistique et de la phonétique, que nous devons connaître et intégrer à notre travail....

Et même en ingénierie, le domaine d'activité de chaque laboratoire de recherche en technologie vocale reste vaste. Il se distingue par la nature de l'approche utilisée (acoustique, physiologique..) afin d'extraire les paramètres cibles du signal vocal. Il est donc impossible de couvrir ou d'invoquer toutes les recherches actuelles sur la parole et le lecteur pourra consulter d'autres ouvrages [Deller et Al, 1993] [J. P. Tubach et al, 1989].

2.2 Mécanisme de Production de la Parole

Dans cette section, on décrira le mécanisme de production et quelques caractéristiques du signal vocal pour établir la liaison avec les hypothèses (ou modèles) dont on discutera au chapitre 4.

La production de la parole est la mise en action d'une source d'excitation et d'un ensemble d'articulation coordonnée. Plusieurs travaux ont été réalisés pour éclaircir le mécanisme de production. Parmi eux, on trouve ceux consacrés à l'étude de l'appareil respiratoire et d'autres aux articulateurs impliqués pendant la production. Les travaux de Draper et al (1959) sur les muscles respiratoires de la parole demeurent une bonne référence. Le fonctionnement des cordes vocales a été clarifié par Van Den Berg (1970).

On mentionne Boë et Al (1980) qui ont travaillé sur la fréquence laryngienne. Hardcastle (1976), Gentil et Al (1980), Abry et Al (1980) ont étudié

les effets et implications de la langue pendant le processus phonatoire. Dans leurs travaux, ils ont étudiés le mouvement, la forme et la position de la langue. Alors que, Bognar (1980) à son tour s'attaquait à mettre en évidence l'impact des mouvements mandibulaires pendant la production de parole.

La source d'excitation est l'air généré par l'appareil respiratoire et qui est poussé jusqu'au larynx. À la sortie du larynx, l'air est parfois modulé (quasi périodique) par les vibrations des cordes vocales donnant naissance à un signal d'excitation quasi périodique. Quand l'air n'est pas modulé, le signal d'excitation prend la forme d'un bruit ou de turbulences. Ceci a permis de subdiviser les sons de parole en deux grandes catégories : les sons voisés et les sons non-voisés. On trouve dans les ouvrages de J. P. Tubach et al (1989) et Deller (1993) une classification plus détaillée des sons.

Le signal d'excitation, passe par le conduit vocal où il subit d'importantes transformations dues à ses résonances propres. On appelle les résonances (fréquences propres) du conduit vocal les formants. Les formants produisent différentes tonalités bien caractéristiques pour chaque phonème selon la forme et la restriction du conduit vocal, les lèvres et évidement la nature de la source d'excitation.

2.2.1 Les sons voisés et formants

Un son voisé est un signal modulé par les vibrations des cordes vocales.

Son spectre est caractérisé par la présence des harmoniques du fondamental. L'enveloppe du spectre présente des maximums correspondant aux fréquences propres (formants) du conduit vocal. Toutes les voyelles en français présentent cette caractéristiques des sons voisés sans oublier une bonne partie des consonnes /b/,/d/ et /g/,...

2.2.2 Les sons non voisés

Un son non-voisé ne présente pas de structure périodique tel que les consonnes /f/ et /s/... Par conséquent son spectre ne présente pas de répartition en harmoniques.

2.3 Mécanisme d'audition

Le système auditif est divisé en deux grandes parties : le système auditif périphérique et le système auditif central. Le système auditif périphérique comprend :

a) l'oreille externe (pavillon, conduit auditif);

b) l'oreille moyenne (tympan, marteau, enclume et étrier) qui transforme l'onde acoustique en impulsion mécanique. L'oreille moyenne joue le rôle d'adaptateur d'impédance du milieu acoustique au milieu liquide ;

c) L'oreille interne transforme le signal de pression dans le liquide en influx nerveux (signal électrique) ;

Le système auditif central est composé de la partie nerveuse traitant l'information auditive périphérique. En effet la membrane basilaire est très sensible aux variations de pression du liquide de la cochlée. Ces variations sont transformées en influx nerveux pour arriver jusqu'au cortex via divers noyaux et centres de traitement de l'influx nerveux. Le système auditif central est assez complexe et son fonctionnement n'est pas encore bien compris.

L'oreille humaine est très sensible à la gamme de fréquence située entre 800 Hz et 8000 Hz. Les limites extrêmes peuvent s'étendre entre 20 Hz et 20000 Hz [Boite et Kunt, 1987].

2.4 Fréquence fondamentale (ou hauteur tonale ou fréquence glottale)

Une description du système de détection de la hauteur tonale proposé par Rouat et al (1997) est l'objet de cette section. Dans une étude exploratoire, ce système à virgule flottante a été utilisé pour analyser les différentes solutions proposées à la problématique de la détection de la double parole, sans effectuer aucune modification à ce système. Ceci nous permet dans un premier temps de savoir si les solutions proposées dans ce mémoire apportent une amélioration significative pour détecter la situation de double parole. Par la suite, on a développé une autre version en virgule fixe, optimisée pour être prête à fonctionner en temps réel. Cette version en virgule fixe est décrite à

l'annexe A et a été réalisée pour Alcatel Mobile Phones. Elle reprend en grande partie les éléments de la version en virgule folttante.

2.4.1 Historique

La détermination de la fréquence glottale a connu une longue histoire. Plusieurs publications scientifiques lui ont été consacrées. Elle date des années du vocodeur [Dudley,1959] dont la qualité et les performances dépendaient énormément de l'estimation de la fréquence fondamentale. L'intérêt lui est encore accordé pour diverses raisons : l'oreille est plus sensible au changement de la fréquence fondamentale (prosodie : intonation, accents, etc.) qu'à ceux d'autres paramètres du signal vocal [Flanagan et Saslov,1958] [Klatt,1973] [Harris et Umeda, 1987] [Hess, 1983]. L'information de la prosodie est prédominée par la présence de ce paramètre. La fréquence glottale est aussi utilisée largement dans les systèmes de reconnaissance et de synthèse de parole. Elle est utilisée dans les systèmes de codage et décodage à faible débit. Denbigh et Zhao (1992) ont utilisé principalement l'information du fondamental pour séparer les signaux de deux locuteurs qui interférent dans le cas des segments voisés.

Il existe différentes approches d'analyse visant à établir ou à associer une mesure pour les variations de la source d'excitation. En général, on distingue les approches d'analyses basées sur le mécanisme de production, le mécanisme de d'audition ou sur la nature acoustique du signal vocal. Pour

chaque approche, on emploie un nom différent pour la mesure de la source d'excitation soit : la fréquence glottale, la hauteur tonale ou la fréquence fondamentale [Hess,1983], [Hess et Indefry, 1987].

Si on analyse le signal vocal à partir du mode de production alors la fréquence fondamentale est définie comme la mesure des cycles d'excitation du larynx. Autrement dit, elle mesure la fréquence de vibration des cordes vocales. Dans ce cas, le terme approprié est la fréquence laryngienne ou glottale. L'analyse peut être effectuée dans le domaine spectral ou temporel .

En mode d'audition, la fréquence du fondamental correspond à une analyse dans le domaine spectral ou temporel en se basant sur des modèles psychoacoustiques ou des modèles auditifs [Goldstein,1973], [Terhard,1979]. Dans ce cas, on utilise plutôt le terme de la hauteur tonale pour caractériser la mesure de l'effet perçu.

Et finalement, en acoustique la fréquence fondamentale est définie au sens mathématique de Fourier [Van Den Enden et Al, 1992], [Elliot, 1987] :

$$\forall\, x \in R \text{ et } \forall\, T \in N^{+} \quad si \quad f(x+T) = f(x)$$

Alors: T est la période et $F=1/T$ est le fondamental.

En présence de bruit, la détermination de la fréquence fondamentale compte parmi les problèmes les plus délicats à résoudre dans l'analyse de parole.

Conformément à l'analyse de Hess (1983) et de Rabiner et al (1976) tous les problèmes et difficultés rencontrés par les algorithmes d'extraction de la fréquence fondamentale sont causés par les raisons suivantes :

* les changements brutaux du conduit vocal ;
* Irrégularité de la source d'excitation ;
* les mouvements des lèvres et d'autres articulateurs ;
* l'interaction entre le conduit vocal et la source d'excitation ;
* l'absence du fondamental lié à la bande-passante dans le cas des commnunications téléphoniques;
* les distorsions subies par le signal original (canal de transmission).

2.4.2 Revues bibliographiques

Plusieurs auteurs ont publié des algorithmes de suivi de la hauteur tonale ou d'estimation du fondamental selon les perspectives des applications. On trouve un résumé dans les publications de Hess (1983; 1987) regroupant les algorithmes les plus intéressants jusqu'à l'année 1988. D'autres ont suivi, et se distinguent par leurs modélisations et incorporations des connaissances auditives ou de production. On peut citer entre autre les travaux de l'université de Québec à Chicoutimi [Rouat et al, 1997] ou de l'université de Gand en Belgique [Van Immerseel et Martens, 1992].

Généralement on sépare les algorithmes d'estimation du fondamental en trois catégories: ceux opérant en domaine temporel, d'autres en domaine

fréquentiel et ceux opérant dans les deux domaines à la fois.

Dans le domaine temporel la fréquence fondamentale est déterminée en analysant directement la forme du signal acoustique. Les stratégies utilisées pour évaluer les impulsions glottales sont souvent le comptage des passages par zéro du signal, la mesure des pics, des vallées ou l'extraction de l'enveloppe (les références de cette section sont citées dans le mémoire de Liu (1992)). Les performances dans cette catégorie dépendent en grande partie des marqueurs de temps qui permettent de segmenter et localiser différentes zones du signal à traiter, et par conséquent perd de sa robustesse en milieu bruité. Il y a aussi les méthodes de corrélation qui sont encore appliquées avec succès en parole téléphonique. Ces méthodes sont reconnues par leur robustesse au bruit mais introduisent parfois une confusion entre le fondamental et les harmoniques. D'après Rabiner (1977) le fait de combiner l'autocorrélation à des techniques telles que : "time domain center clipping" ou "AMDF" (average mean distance function), corrige les erreurs engendrées par les formants.

Dans le domaine spectral, les techniques analysant directement le spectre du signal donnent des résultats souvent non précis et dépendent énormément de la taille de la fenêtre d'analyse. Les algorithmes ont plutôt orienté la recherche du fréquence fondamental en se basant sur des techniques manipulant davantage les harmoniques. On cite la technique qui extrait le

fondamental comme étant le plus petit multiple commun à une série d'harmoniques dépassant un certain seuil. D'autres auteurs ont proposé une autre technique qui consiste à utiliser une fonction de compression sur l'axe des fréquences pour avoir la même information du spectre avec différentes représentations. Ensuite, ils appliquent une transformation linéaire ou non-linéaire en combinant les différentes représentations du spectres pour extraire le fondamental. Schroedar (1968) avait utilisé le produit comme fonction de transformation (voir figure 2). Martin (1981; 1987) a utilisé une fonction peigne pour les transformations de compression sur l'axe des fréquences et Hermes (1988) de son coté a utilisé une fonction logarithmique à base 2. D'autres méthodes et techniques sont citées dans le mémoire de Liu (1992).

L'analyse spectrale est très bien connue et largement utilisée dans l'analyse du signal. Son application dans la parole est freinée par plusieurs obstacles liés à l'hypothèse de considérer les segments de parole comme étant stationnaires et quasi périodiques. Ces hypothèses sont souvent loin d'être vérifiées. De plus, l'analyse spectrale est incapable de suivre ou détecter les changements instantanés liés aux articulations pendant la phonation.

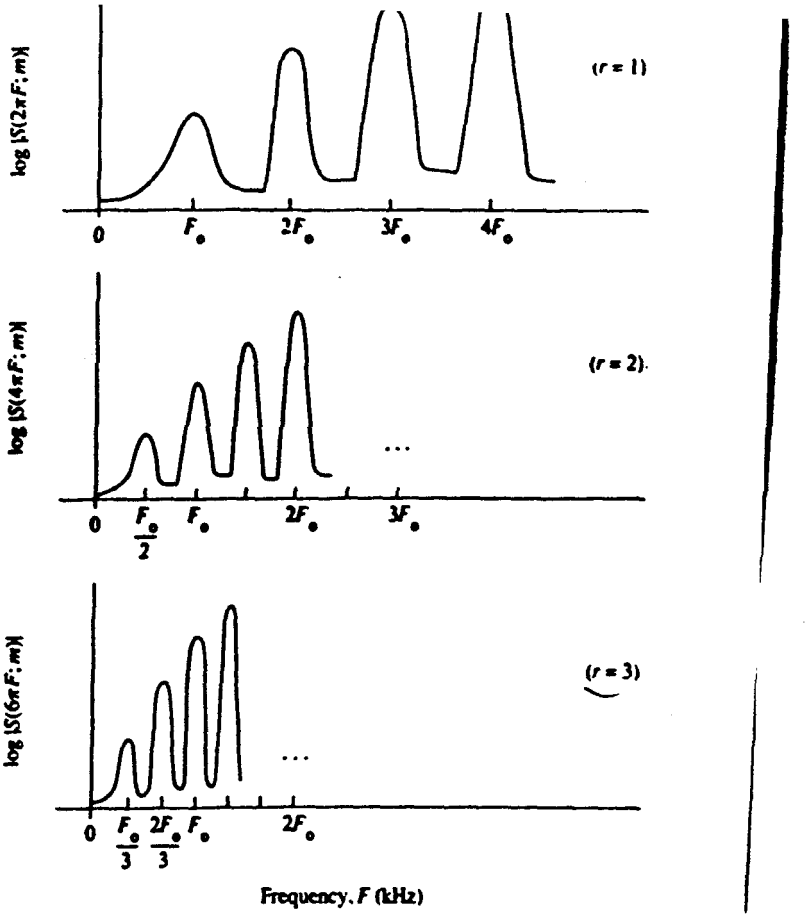


Figure 2 : Application d'une fonction linéaire ou non linéaire sur le spectre original du signal. Le graphique illustre le cas où on multiplie plusieurs fois l'axe du spectre original par différents facteurs entiers. Ensuite, en faisant une sommation sur toutes les transformations (y compris le spectre original), on détecte l'harmonique du spectre ayant l'amplitude la plus élevée qui est supposée correspondre à la fréquence fondamentale [Deller et al, 93].

L'oreille humaine procède par une analyse spectro-temporelle d'où l'apparition d'autres catégories d'algorithmes [Rouat, 1997], [Van Immerseel et Martens, 1992], [Terhard,1979]. D'autres méthodes et techniques sont citées dans le mémoire de Liu (1992).

Van Immerseel et Martens (1992) ont proposé un algorithme intéressant qui modélise approximativement bien le mécanisme d'audition. L'algorithme

fonctionne en temps réel pour de la parole propre ou bruitée. L'oreille moyenne est modélisée par un filtre passe bande à 2 pôles. La membrane basilaire est représentée par un banc de filtres. La fréquence caractéristique pour chaque filtre est répartie sur une échelle non linéaire. L'algorithme inclut aussi un modèle pour la cellule ciliée, un filtre d'extraction de l'enveloppe et une unité de traitement centrale. À la sortie du modèle, on est supposé obtenir une information équivalente à celle convoyée par le nerf auditif.

On présente à la prochaine section les grandes lignes de l'algorithme Algomai94 [Rouat et al, 1997] que nous avons utilisé.

2.5 Système de détection de fréquence glottale et de décision de voisement [Rouat et al, 1997]

Dans cette section, on décrira le fonctionnement des différents modules constituant le système de détection de la hauteur tonale à virgule flottante. Ce système sera utilisé par la suite dans tous les tests et expériences rapportées dans le chapitre 5 et l'annexe A. Également, On donnera sa version en virgule fixe dans l'annexe A qui comptera parmi l'une de nos contributions importantes de ce mémoire.

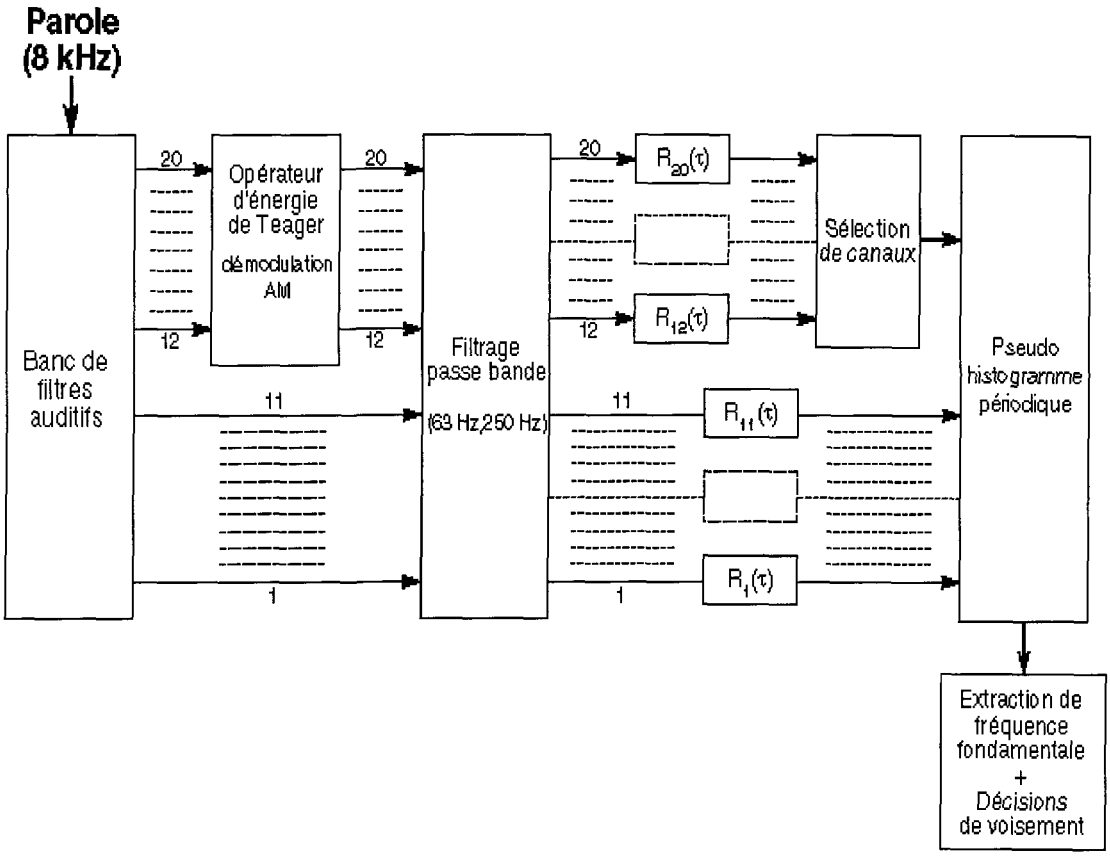


Figure 3: Schéma du système d'estimation de la fréquence glottale et de décision de voisement.

2.5.1 Principe général

Le système de suivi de fréquence glottale et de décision de voisement est basé sur le fait que les modulations d'amplitude apparaissant à la sortie d'un banc de filtres cochléaires sont caractéristiques de la hauteur tonale. On fait la distinction entre les harmoniques résolus par le système auditif (en basse

fréquence) et les harmoniques non résolus qui créent des battements lorsque la largeur de bande des filtres cochléaires est suffisamment grande vis-à-vis de la fréquence glottale. La présence de ces battements permet au système auditif de percevoir la hauteur tonale même lorsque le fondamental est absent ou trop bruité (B. Moore, 1989).

L'algorithme comprend trois modules (figure 3). Le premier module est un banc de vingt filtres cochléaires (fréquences centrales de 330 Hz à 3700 Hz), le second traite les sorties des filtres afin de rehausser la période de modulation du signal glottique et de combiner les informations des canaux sélectionnés en un pseudo-histogramme périodique. Le troisième module estime la fréquence glottale et prend la décision de voisement.

2.5.2 Description des deux premiers modules

Les onze premiers canaux (330 Hz - 1270 Hz) sont simplement filtrés passe-bande avant calcul de la corrélation normalisée entre la fenêtre centrée et cette même fenêtre décalée. Ceci est fait pour chaque canal i ($R_i(t)$, i=1, 11). Onze représentations différentes du fondamental (lorsque présent) et des premiers harmoniques sont ainsi obtenues. Les neuf derniers canaux sont prétraités à l'aide de l'opérateur énergie de Teager (J.F. Kaiser, 1990, 1993). Cette opération est équivalente à estimer le carré de l'enveloppe du signal pondéré par le carré de la pulsation instantanée (J.Rouat, 1993).

Lorsque le signal est très bruité (rapport signal à bruit de 0 dB et moins), l'algorithme peut utiliser une unité de sélection automatique des canaux. Cette unité permet de sélectionner les canaux pour lesquels le caractère harmonique du signal ressort suffisamment bien. Pour les expériences du présent mémoire, nous n'avons pas utilisé cette technique de sélection automatique car les données enregistrées en véhicule n'étaient pas assez bruitées pour justifier cette augmentation de complexité.

Le pseudo-histogramme périodique (figure 3) est noté PPH et est obtenu en réalisant la somme à travers les canaux sélectionnés des corrélations normalisées.

$$PPH(\tau) = \frac{1}{M} \sum_{i=1}^{M} R_i(\tau)$$

M est le nombre de canaux qui contribuent à la fréquence glottale, ici M=20.

2.5.3 Décision de voisement et estimation de la fréquence glottale

Les deux plus grands pics 'éligibles' dans PPH(τ) sont sélectionnés. Pour être 'éligible', un pic doit être plus grand qu'un seuil S prédéterminé. Si on ne trouve pas deux pics vérifiant ces conditions, on déclare que le segment est non voisé. On suppose que les deux pics correspondent à des valeurs du temps égales respectivement à τ_1 et τ_2. τ_1 et τ_2 sont des multiples (ou l'un d'eux peut être égal) de T qui est la période fondamentale. T est donc un des sous

multiples de τ_1 et τ_2. On cherche les sous-multiples et T est associé au plus petit sous multiple qui correspond à un pic vérifiant la relation:

$$PPH(T) \geq S_{pe}.Max\left[PPH(\tau_1);PPH(\tau_2)\right]$$

avec Spe = 0.5.

Si l'algorithme ne trouve pas T, le segment est déclaré comme étant possiblement non voisé, sinon il est déclaré voisé avec une fréquence égale à 1/T. Ces informations sont ensuite traitées afin de prendre la décision finale et définitive de voisement et de valeur de fréquence glottale. L'algorithme a été adapté afin de pouvoir fonctionner en temps réel dans le contexte de la détection de double parole tel que décrite dans ce mémoire.

2.6 Méthode du cepstre et de la prédiction linéaire (Linear Predictive Coding LPC)

La présente section apporte l'information théorique nécessaire pour la bonne compréhension des diagrammes et des stratégies qui seront présentées à la suite de ce mémoire.

2.6.1 Modèle autorégressif

Le mécanisme de production est modélisé simplement par un système de transmittance G/V(z) et d'excitation e(t). G correspond au gain du système. Les

sons voisés sont caractérisés par une excitation périodique (train d'impulsion). Dans le cas des sons non voisés, l'excitation du système est composée d'un bruit blanc, de moyenne nulle et de variance unité [Boite et Kunt, 1987][Deller et al, 1993]. (Voir figure 4)

La transformé du signal prend la forme suivante :

$$S(z) = \frac{G \times E(z)}{V(z)} \qquad (2.1)$$

où S(z), E(z) et V(z) sont respectivement les transformés en z de s(t), e(t) et v(t). Le signal s(t) est produit par un signal excitateur e(t) (source glottique) traversant un système linéaire de réponse impulsionnelle v(t) (conduit vocal). L'expression (2.1) peut être exprimée dans le domaine temporel par :

$$s(n) = e(n) + \sum_{i=1}^{p} a(i) \times s(n-i) \qquad (2.2)$$

Cette expression prédit que chaque échantillon peut être estimé à partir des p échantillons précédents.

Ceci représente le modèle autorégressif d'ordre p relatif à la définition du signal autorégressif. On lui donne souvent le nom " tout pole" puisque toutes les racines de la fonction de transfert sont des pôles. Les coefficients a(i) sont appelés les coefficients de prédiction.

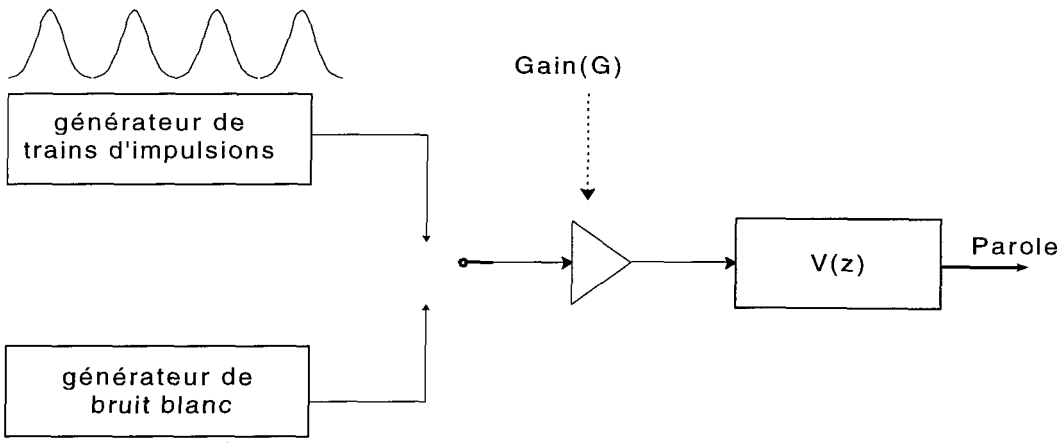


Figure 4 : Modélisation de la parole

Un autre modèle connu sous autorégressif à moyenne ajustée "ARMA" est parfois utilisé. Il se distingue par la présence à la fois des pôles et des zéros. La transmittance prend plutôt la forme suivante:

$$S(z) = \frac{E(z) \times N(z) \times G}{V(z)} \qquad (2.3)$$

En effet, dans ce modèle, on tient compte du couplage avec la cavité nasale responsable d'une autre classe de sons appelée: sons nasals tels / m/ et /n/.

2.6.2 Cepstre

Le couplage entre l'excitation glottale et le conduit vocal, rend difficile la mesure précise du fondamental. Pour remédier à ce problème, une approche originale connue par le cepstre a été proposée par Noll (1964) pour extraire le fondamental. Oppeheim (1968) l'a exploitée pour les systèmes de codage/décodage. Finalement, on trouve les systèmes d'analyse et de la

reconnaissance automatique qui ont bénéficié largement de cette technique. Le formalisme mathématique de cette méthode peut être consulté dans l'ouvrage [Oppenheim and Schafer, 1975].

Tel que mentionné, la parole peut être considérée comme étant le résultat d'une convolution entre la source d'excitation et le conduit vocal (voir section 2.6.1):

$$s(n) = e(n) * v(n) \qquad (2.4)$$

Le formalisme mathématique d'homomorphisme, permet de transposer l'opérateur convolution "*", en opérateur addition "+" dans le domaine transposé.

C'est le principe de base utilisé par le cepstre pour déconvoluer le signal vocal. La figure 5 illustre les étapes de transformation nécessaires pour le cepstre.

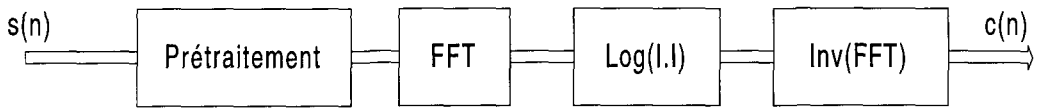


Figure 5: Déconvolution par le cepstre

Les s(n) sont les échantillons du signal temporel avant toute transformation.

Les c(n) sont les coefficients cepstraux dans un nouveau domaine réel appelé : quefrentiel.

L'avantage du cepstre, est son insensibilité au problème des formants (leur confusion avec les harmoniques dans le contexte de détection de fréquence glottale) et de la distorsion de la phase. Il permet par un simple filtrage passe bas (simple fenêtre rectangulaire) de récupérer la contribution vocale dans le domaine queférentiel. Dans le cas des voisés, le fondamental est celui ayant le pic le plus élevé.

Par contre l'inconvénient de la méthode du ceptre est sa sensibilité au bruit et à la taille de la fenêtre d'analyse. En plus la mesure du fondamental n'est pas toujours précise.

2.6.3 Prédiction linéaire (LPC)

Pour contourner les problèmes liés aux méthodes spectrales, une approche originale a été proposée par Atal (1971). Elle a été désignée pour les applications des systèmes de codage/décodage dans le laboratoire Bell. Depuis, un grand intérêt lui a été accordée dans plusieurs disciplines et ses applications sont très importantes actuellement. L'avantage du LPC est de paramétriser à chaque instant la contribution du conduit vocal par des coefficients de prédiction a(k). Les coefficients sont mis à jour à chaque fois qu'on considère une nouvelle fenêtre d'analyse du signal. En général, on

travaille avec une fenêtre de durée 10 ms jusqu'à 15 ms dépendant de l'objectif des applications. Nous allons donner quelques aspects théoriques de cette méthode.

2.6.3.1 Estimation des paramètres

Le signal vocal peut être considéré comme étant un signal issu d'un modèle autorégressif et s'exprimer dans le domaine temporel par l'équation (2.5). A partir de cette équation, on va calculer les coefficients de prédiction a(i) du modèle A.R.

$$s(n) = \sum_{i=1}^{p} a(i) \times s(n-i) + e(n) \qquad (2.5)$$

e(n) est l'excitation; a(i) les coefficients de prédiction du modèle A.R.

Ainsi, si on dispose de e(n) et des a(i) on peut prédire les échantillons suivants à partir de conditions initiales données. Or l'information de la source e(n) est quasiment inaccessible et l'approximation reste la seule alternative.

L'équation (2.5) peut être remplacée par :

$$\tilde{s}(n) = \sum_{i=1}^{P} a(i) \times s(n-i) \qquad (2.6)$$

par suite, l'erreur de prédiction peut s'exprimer par:

$$e(n) = s(n) - \tilde{s}(n)$$
$$= s(n) - \sum_{i=1}^{P} a(i) \times s(n-i) \qquad (2.7)$$

Les coefficients optimaux a(i) sont déterminés en minimisant l'erreur quadratique moyenne E:

$$E = \sum_{n=0}^{N} e^2(n) = \sum \left(s(n) - \sum_{i=1}^{P} a(i)s(n-i) \right)^2$$

En dérivant par rapport aux coefficients a(i), on obtient un système d'équations linéaires :

$$0 = \frac{\partial E}{\partial a_j}$$
$$0 = -\sum_{n=0}^{N-1} s(n)s(n-1) + 2\sum_{n=0}^{N-1}\sum_{k=1}^{P} a(k)a(n-k)s(n-j)$$
$$\Rightarrow \sum_{n=0}^{N} s(n)s(n-1) = \sum_{n}^{N}\sum_{k=1}^{P} a(k)a(n-k)s(n-j)$$

Il existe une succession d'articles visant à optimiser les méthodes de résolution de ces équations. Makhoul [1975] a écrit un bon article de synthèse sur ce sujet. Il débute par une introduction générale du modèle AR et ARMA. Il a établi ensuite, les coefficients de prédiction dans le domaine fréquentiel et temporel. Il a traité les cas où le signal est considéré déterministe (voisement) ou non déterministe (bruit, non voisé) en abordant également la question de la stabilité et du gain du système.

2.6.3.2 Détermination du fondamental à partir du LPC

Atal et Hanauer (1971) ont proposé 2 techniques pour estimer le fondamental. La première analyse l'erreur résiduelle pour déterminer le fondamental de façon synchrone. La position de l'impulsion glottale correspond aux instants où l'énergie de l'erreur résiduelle est large. La deuxième approche, effectue un filtrage passe bas du signal, suivi par une préaccentuation. Ensuite, elle applique la corrélation pour extraire le fondamental à partir du signal de l'erreur résiduelle.

2.7 Opérateurs non linéaires

L'objet de cette section est de présenter la formulation mathématique des opérateurs non linéaires Teager et Dyn car ils font partie des outils de traitements utilisés dans les solutions proposées.

2.7.1 Opérateur Teager

L'opérateur dit de Teager a été proposé par Kaiser (1990) et est inspiré des travaux de Teager. Il calcule l'énergie d'un signal par analogie avec l'énergie totale des systèmes mécaniques (masse + ressort).

Si on applique Teager (T(s(t))) à un signal analogique [Rouat, 1993]:

$$s(t)=A(t)\sin(\omega(t)t+F(t))$$

$$T(s(t))=\left(\frac{ds(t)}{dt}\right)^2-s(t)\frac{d^2s(t)}{dt^2}$$

$$=\eta(t)+\frac{1}{2}\left(\left(\left(\frac{dA(t)}{dt}\right)^2-A(t)\frac{d^2A(t)}{dt^2}\right)(1+\cos(2\upsilon(t))\right)$$

$$+\left(\frac{A^2}{2}\frac{d^2\phi(t)}{dt^2}\sin(2\upsilon(t))\right)$$

Avec:

$$\eta(t)=A^2(t)\left[\omega(t)+\frac{d\phi}{dt}\right]^2$$

$$\upsilon(t)=\omega t+\phi(t)$$

La formule numérique utilisée pour le calcul de Teager [Kaiser, 1990] est:

$$T(s_n)=s_n s_n-s_{n+1}s_{n-1}$$

Le premier terme $\eta(t)$ est l'énergie extraite par Teager. Alors que le deuxième terme négligeable est considéré comme du bruit (sous toute réserve que certaines conditions soient vérifiées). On remarque que l'énergie extraite par Teager est proportionnelle au carré de l'amplitude et au carré de la fréquence instantanée. Ainsi, l'opérateur Teager est capable de démoduler rapidement un signal modulé en AM ou FM. Il est à noter que l'utilisation de

Teager suppose que la valeur de la fréquence centrale $\left(\text{i.e } \frac{\omega}{2\pi}\right)$ est importante par rapport à la largeur de bande de l'amplitude ou de la phase.

2.7.2 Opérateur Dyn

Dyn a été proposé par J. Rouat (1993) comme étant un opérateur capable de faire ressortir les fluctuations de l'enveloppe d'un signal. Dyn calcule la dérivé de la puissance instantanée d'un signal de parole par analogie avec l'évaluation des variations d'énergie potentielle d'un système mécanique fermé (masse+ressort). Si on applique Dyn à un signal analogique [Rouat, 1993] :

$$s(t) = A(t)\sin(\omega(t)t + \Phi(t))$$

$$\mathrm{Dyn}(s(t)) = s(t)\frac{ds(t)}{dt}$$

$$= \frac{1}{4}\frac{d^2A(t)}{dt^2} + \frac{A^2(t)}{2}\sqrt{\left(\left(\frac{dA/dt}{A(t)}\right)^2 + \left(\frac{d(\upsilon(t))}{dt}\right)^2\right)}\cos(2\upsilon(t) - \zeta(t))$$

Avec:

$$\zeta(t) = a\tan\left(A^2(t)\frac{\left(\omega(t) + d\phi/dt\right)}{dA/dt}\right)$$

$$\eta(t) = A^2(t)\left[\omega(t) + \frac{d\Phi(t)}{dt}\right]^2$$

$$\upsilon(t) = \omega t + \Phi(t)$$

La formule numérique utilisée pour le calcul de Dyn est:

$$Dyn(s_n) = s_n(s_n - s_{n-1}) \quad ou \quad Dyn(s_n) = \frac{s_n}{2}(s_{n+1} - s_{n-1})$$

CHAPITRE 3

LE PHÉNOMÈNE DE BRUIT ET D'ÉCHO

Dans ce chapitre, on étudiera les problèmes du bruit et d'écho rencontrés dans le contexte radiotéléphone main-libre et on donnera les solutions et techniques utilisées pour résoudre ce problème.

3. Le phénomène du bruit

Le bruit dégrade considérablement la performance des algorithmes d'annulation d'écho (voir section 3.4) et en principe si on n'était pas confronté au problème de bruit, la détection de la double parole serait résolue par la majorité des algorithmes d'annulation d'écho. Pour cette raison nous jugeons convenable de consacrer cette section à étudier en détail la nature et les sources du bruit dans le contexte radiotéléphone main-libre. Également, on citera de façon concise les algorithmes les plus utilisés pour annuler le bruit dans le contexte des radiocommunications.

3.1 Introduction

Il est difficile de donner une définition précise au bruit car tout signal peut l'être relativement au contexte. Si on considère à titre d'exemple

l'utilisateur du téléphone, le signal cible est représenté par la parole du locuteur. Le bruit représente les parasites introduits par les composantes du matériel du téléphone ou par toutes autres interférences extérieures (y compris la parole d'autres locuteurs). Par conséquent, le débruitage reste toujours une problématique mal définie à cause du comportement aléatoire du bruit. Les solutions proposées de nos jours, analysent le signal sous la considération de certaines hypothèses et contextes.

Dans cette partie, on va s'intéresser aux problèmes du bruit et des techniques de suppression utilisées dans le contexte radio téléphone main libre.

3.2 Source de bruit

D'après Degan et Prati (1988) les sources de bruits sont principalement dues à l'environnement extérieur, intérieur et propre au véhicule.

La majeure partie du bruit extérieur est due:

* aux contacts des pneus avec la chaussé ;
* à la forme et taille des pneus ;
* à la pénétration de l'air dans le véhicule (état des fenêtres) ;
* au croisement avec d'autres moyens de transports ;
* à différentes vitesses de la voiture.

Le bruit à l' intérieur et propre au véhicule est dû:

* au fonctionnement du moteur et de la boite de vitesse ;

* à la propriété matérielle et géométrique de l'habitacle.

3.3 Revues des techniques utilisées

L'objectif du débruitage en radiocommunication est de réduire la gêne provoquée chez l'utilisateur. On utilise plusieurs algorithmes visant principalement à réduire le Rapport Signal Bruit (RSB) sans faire subir de distorsions importantes au signal traité (Windrow, 1975 ; Boll, 1979 ; Ephraim, 1984).

3.3.1 Méthodes basées sur la soustraction des harmoniques

Berouti (1979) a présenté un algorithme intéressant pour les signaux bruités. L'algorithme fonctionne seulement dans le cas de bruits additifs (bruit non corrélé avec le signal) et stationnaires ayant une espérance mathématique nulle. On donne un résumé sur le principe de cette méthode qui est très connue dans le débruitage. Premièrement, on suppose que le signal vocal est de nature déterministe et stationnaire. Soient :

s(n) l'échantillon à l'instant t pour la parole propre.

b(n) l'échantillon à l'instant t du bruit .

y(n) l'échantillon à l'instant t pour la parole bruitée.

Dans le cas d'un bruit additif, on peut exprimer le signal par l'équation numérique:

$$y(n) = s(n) + b(n) \qquad (3.1)$$

On ramène l'équation (3.1) dans le domaine fréquentiel:

$$Y(\omega) = S(\omega) + B(\omega) \qquad (3.2)$$

avec $\omega = 2\Pi f$, ω est la pulsation et f est la fréquence.

La densité spectrale de puissance du signal est exprimée par la formule:

$$|Y^2(\omega)| = |S^2(\omega)| + |B^2(\omega)| + |S^*(\omega) \times B(\omega)| + |S(\omega) \times B^*(\omega)| \qquad (3.3)$$

où $S^*(\omega)$ et $B^*(\omega)$ dénotent respectivement les complexes conjugués de la densité spectrale de puissance pour la parole propre et le bruit.

La soustraction spectrale revient à remplacer certains membres de l'équation (3.3) par leurs espérances mathématiques:

$$E|Y^2(\omega)| = E|S^2(\omega)| + E|B^2(\omega)| + E|S^*(\omega) \times B(\omega)| + E|S(\omega) \times B^*(\omega)| \qquad (3.4)$$

Par hypothèse le bruit et la parole sont non corrélées d'où :

$$E|S^*(\omega) \times B(\omega)| = E|S(\omega) \times B^*(\omega)| = 0 \qquad (3.5)$$

Ensuite pour les segments de silence, on peut évaluer la densité spectrale du bruit seul. Cette information sur le bruit, nous permettra finalement d'estimer la parole propre comme suit:

$$\text{Soit:} \quad Q(w) = E\left[|Y(w)|^2\right] - E\left[|B(w)|^2\right]$$

Alors l'estimateur pour la parole propre sera :

$$\left|\hat{S}(w)\right|^2 = Q(w) \quad \text{si } Q(w) > 0$$
$$= a \qquad \text{autrement}$$

Cette méthode est connue sous le nom de la soustraction harmonique linéaire lorsque α est nul.

Le fait de couper net le spectre du signal en plaçant les valeurs nulles, engendre un bruit musical. Ce bruit cause un effet indésirable à l'oreille. Pour remédier à ceci, Berouti (1979) a introduit une légère modification à sa méthode proposée avant et qui sera ensuite désignée par la méthode de suppression généralisée. En effet, dans le cas où la différence de la moyenne quadratique entre le signal bruité et le bruit estimé est très faible, il remplace la valeur originale du spectre du signal par une valeur faible au lieu d'une valeur nulle. Ce bruit synthétique est

connu par le bruit de confort. D'autres auteurs ont suggéré d'autres modifications à cet algorithme. On mentionne Lockwood (1991) qui a appliqué la soustraction harmonique du bruit seulement dans les zones spectrales où le rapport signal bruit (RSB) est faible. Cette méthode a donné des résultats intéressants en radiocommunication.

3.3.2 Autres méthodes

D'autres auteurs introduisent une autre catégorie d'algorithmes pour supprimer le bruit. Ils considèrent que chaque segment du signal est stationnaire et modélisé par un système autoregréssif d'ordre p. Paliwal et Basu (1987) supposent que le bruit est blanc, centré, additif et non corrélé avec le signal original. Premièrement, Il estime les p coefficients du modèle A.R. par une des méthodes conventionnelle. Ensuite, il applique l'algorithme du filtrage de Kalman en utilisant les p coefficients estimés auparavant pour estimer de nouveau le signal de parole.

Koo(1989) et Baillargeat(1991) ont considéré une autre approche plus réaliste qui introduit une modélisation autoregréssive pour le bruit également.

Macaulay (1980) et Yang (1993) ont suggéré d'autres techniques utilisant les notions probabilistes pour estimer le spectre d'énergie.

L'ensemble des algorithmes de débruitage qu'on a cité jusqu'à présent analyse le signal parvenant d'une seule source (un seul microphone). Il existe d'autres catégories d'algorithmes qui utilisent deux sources d'informations ou plus pendant le processus d'analyse. Par exemple en véhicule, on utilise un microphone juste devant le chauffeur et un autre microphone proche de la fenêtre située à droite ou à gauche du chauffeur. Ces techniques seront citées dans ce mémoire.

3.4 Le phénomène d'écho

Le phénomène d'écho est le principal facteur engendrant la situation de la double parole dans la radiocommunication et l'objet de cette section est de présenter un exposé simplifié de ce phénomène.

Dans cette partie, on donne une introduction du phénomène d'écho et aussi les techniques classiques et récentes utilisées pour contourner ce problème. L'étude présente s'intéresse principalement au contexte de la radiocommunication.

3.4.1 Introduction

L'écho est défini comme étant la perception des ondes acoustiques sonores réfléchies par certains obstacles (les murs, le sol, habitacle de voitures..). Si la distance parcourue par l'onde réfléchie est courte alors elle est

perçue comme une distorsion spectrale. Pour les communications à grandes distances (ordre du délai est de quelque dizaine de ms), l'onde réfléchie est perçue comme un véritable écho.

Les sources de génération de l'écho peuvent être d'origine électrique liées aux impédances du canal de transmission et liées au problème du gain rencontré pendant le processus de l'amplification du signal pour établir des communications lointaines. Un tel écho est appelé écho du canal.

Une autre source de génération d'écho est le couplage entre le haut parleur et le microphone. Un tel écho est appelé écho acoustique. Il est considéré comme étant le plus important et difficile à traiter.

Si une communication téléphonique est établie entre deux terminaux mobiles alors le seul écho présent est celui du canal de transmission. Pour les appels locaux, cet écho reste insignifiant. Pour les appels à distance, un écho inévitable aura lieu en causant une grande gêne chez le locuteur lointain. L'article de M. Sondhi et W. Kellermann [Sadaoki and Sondhi, 1991] expose les problèmes reliés à l'écho du canal. Également, il cite les algorithmes de suppression d'écho utilisés.

3.4.2 Écho acoustique

Dans le contexte de la radiotéléphonie main libre, l'écho acoustique est considéré comme étant l'écho le plus important et plusieurs travaux s'y sont intéressés (Macchi, 1988 ; Ozeki, 1984 ; Haykin, 1991). L'écho est relié à la réflexion du signal émis par le haut parleur du système mobile sur la paroie de l'habitacle et est capté par le microphone du même système. En conséquence, le locuteur lointain est dans la situation où il réentend sa communication avec un certain délai. L'écho acoustique se caractérise par le couplage entre le haut-parleur et le microphone, la structure de l'habitacle et les perturbations à l'intérieur du véhicule

3.4.3 Contrôle d'écho

Pour contourner les problèmes mentionnés ci-dessus, l'intégration d'un module de contrôle est indispensable. Classiquement, on place un atténuateur sur le canal de réception et un autre sur le canal de transmission. Et, à chaque intervalle de temps, on évalue l'énergie sur les deux canaux (réception et transmission) du système pour décider où se situe l'activité vocale des deux locuteurs. Si le locuteur lointain est actif, on diminue le gain de l'atténuateur de transmission et on accroît le gain de celui en réception. Par contre, si le locuteur local est actif, on diminue le gain de l'atténuateur de réception et on accroît le gain de celui en

transmission (voir figure 6). Ainsi, nous évitons l'émission de écho acoustique.

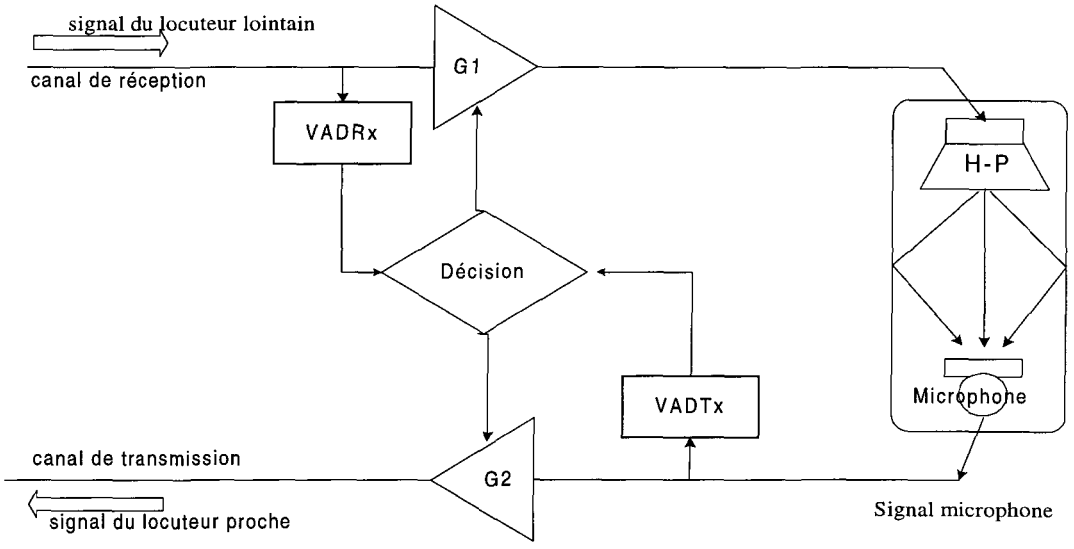


Figure 6: Suppression classique d'écho

VADRx	algorithme d'évaluation d'activité vocale du signal lointain.
VADTx	algorithme d'évaluation d'activité vocale du signal microphone.
Décision	algorithme qui décide de l'activité vocale des locuteurs lointain et local.
G1,G2	atténuateur de gain du canal.

Cette approche simple se trouve limitée par le phénomène de bruit et du gain. Le niveau de bruit élevé peut mettre la décision du contrôleur à mal. Il donne ainsi, le droit de parole uniquement au signal très bruité (communication unidirectionnelle). Sans oublier le délai, qui peut dégrader ou tronquer (ne pas considérer les segments à faible énergie) le signal transmis.

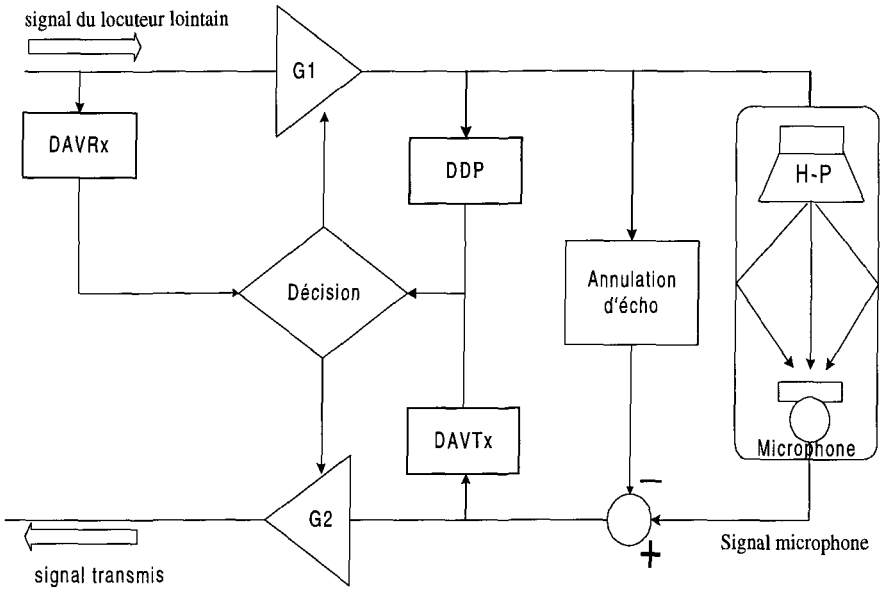


Figure 7: Système classique pour contrôle d'écho

DAVRx : Détecteur de l'activité vocale sur la voie de réception.
DAVTx : Détecteur de l'activité vocale sur la voie de transmission.
DDP : Détecteur de l'activité de double parole.
G1,G2 : Atténuateur du gain.

Par ailleurs, l'écho peut être détecté par le VadTx et les 2 canaux sont alors atténués. Il est donc nécessaire de détecter la double parole (DDP).

Le développement d'algorithmes d'annulation d'écho s'est avéré indispensable pour ce genre de problème [Gilloire, 1988]. Généralement, le contrôleur dans les systèmes mobiles est similaire à celui représenté à la figure 7.

L'ensemble des détecteurs se base sur un seuil d'énergie. Le seuil est calculé en fonction du bruit ambiant et du contexte. La décision prise par chaque détecteur est fournie au module de contrôle (Décision) qui contrôle ensuite les atténuateurs de gain en conséquence. Le contrôle peut accroître ou diminuer le gain et faire déclencher ou non l'adaptation du filtre adaptatif.

L'efficacité de chaque système dépend du fonctionnement du filtre adaptatif et surtout de l'adaptation de ses coefficients aux moments opportuns en évitant l'adaptation du filtre en situation de double parole. En cas de mauvaise décision de double parole, l'écho acoustique devient important et provoque une gêne considérable pour le locuteur lointain.

3.4.4 Techniques de détection.

En présence de double parole, l'énergie sur le canal de transmission se trouve significativement accrue. Cette observation est la base des algorithmes utilisant un seuil fixe Yang (1993).

En radiotéléphonie, on ne peut jamais adopter cette approche en raison de sa grande limite et sensibilité au bruit. Comme alternative, des méthodes basées sur la considération de deux seuils ont été envisagées [Baillargreat, 1991].

Baillargreat (1991) a montré avec cette approche, des résultats intéressant sur les données radio main libre sauf que la détermination des seuils reste une tâche rigoureuse et difficile à atteindre. Le principe de cette technique est le suivant :

Soient $P_x = \sum_{i=0}^{L} x^2(i)$ l'énergie du signal haut-parleur, $P_y = \sum_{i=0}^{L} y^2(i)$ l'énergie du signal microphone et $10 \times \mathrm{Log}_{10}\left(\frac{P_x}{P_y}\right)$ le rapport d'énergie du signal haut-parleur et microphone. Soient S1 et S2 les seuils optimisés et fixés.

si $P_x < S1$

et si $10 \times \mathrm{Log}_{10}\left(\frac{P_x}{P_y}\right) \approx 0$ alors on n'a pas d'écho et l'adaptation du filtre est non autorisée;

et si $10 \times \mathrm{Log}_{10}\left(\frac{P_x}{P_y}\right) < 0$ alors on est dans une situation de présence de l'écho seul et l'adaptation du filtre est autorisée;

si $P_x > S1$

et si $10 \times \mathrm{Log}_{10}\left(\frac{P_x}{P_y}\right) > S2$ alors on adapte le filtre (présence d'écho) ;

et si $10 \times \mathrm{Log}_{10}\left(\frac{P_x}{P_y}\right) < S2$ alors on est on présence de double parole et on ne doit pas adapter le filtre ;

Une autre technique basée sur l'analyse de l'erreur de prédiction du modèle A.R a été proposée par Ye(1991).

CHAPITRE 4

MÉTHODOLOGIE

Dans ce chapitre, on présentera le fonctionnement du système radiotéléphone main-libe et on expliquera comment la performance se dégrade dans la situation de la double parole. Ensuite, nous citons l'ensemble des solutions que nous proposons pour répondre à la problématique actuelle. Une description de la base de données utilisée et du critère d'évaluation feront également l'objet de ce chapitre.

4.1 Problématique actuelle

Les problèmes engendrés par l'écho et le bruit ont déjà été traités et discutés au chapitre 3. Nous avons aussi présenté une variété de solutions et techniques pour résoudre ces problèmes selon les perspectives de chaque application.

En radiotéléphonie, un système fréquemment mis en oeuvre dans les terminaux mobiles correspond à celui illustré par la figure 8. Il est alors

judicieux de le considérer comme un système ou une structure de référence pour comparer les performances de nos structures proposées.

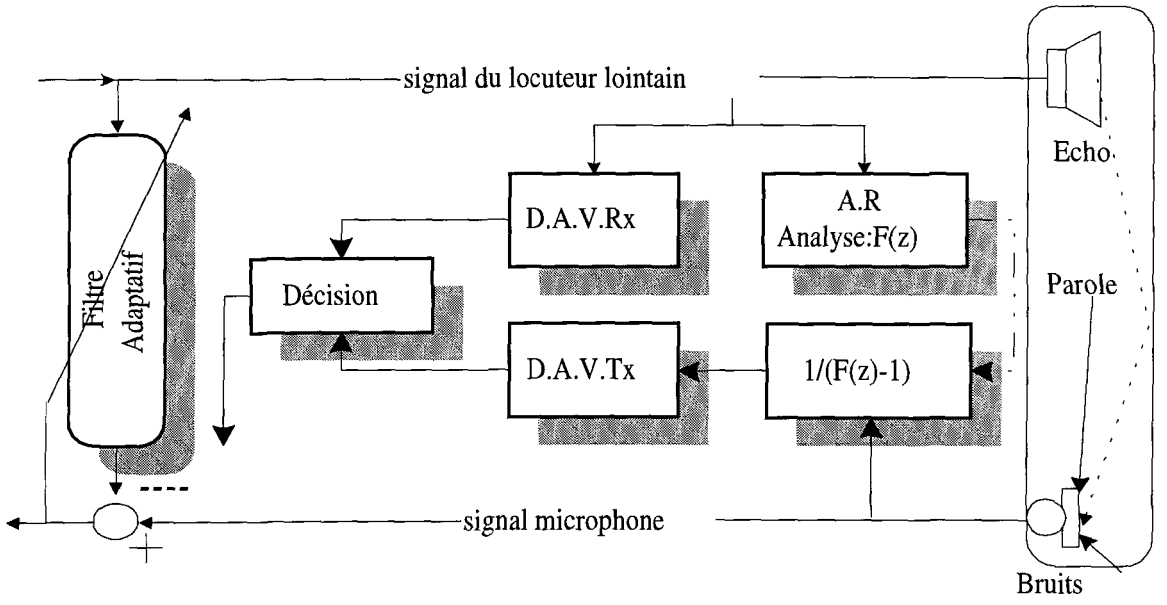


Figure 8: Système de référence

D.A.V.Rx : détection d'activité vocale du locuteur lointain,
D.A.V.Tx : détection d'activité vocale du locuteur local,
Décision : prise de décision pour adapter ou ``geler`` les coefficients du filtre,
A.R. : modélisation autorégressive.

4.2 Description du système de référence

L'écho acoustique lié au couplage entre le Haut-parleur (H.P), le microphone et l'habitacle est très important. L'écho capté par le microphone se trouve retransmis vers le locuteur lointain après un certain délai. Comme on l'avait mentionné, cela crée une gêne et rend la communication désagréable. Le rôle primordial du filtre est d'éviter ou de supprimer si possible cet effet

d'écho indésirable. Le filtre adaptatif permet de modéliser le trajet acoustique de l'écho et donne en sortie l'écho estimé. Cet écho estimé est soustrait du signal microphone pour favoriser la transmission du signal local sans interférence. Les coefficients du filtre adaptatif sont réestimés seulement dans la situation où le locuteur lointain parle et le locuteur local ne parle pas. C'est une situation pour laquelle on est en présence uniquement de l'écho réel sur la voie microphone. Dans toutes les autres situations, on se contente de ne pas modifier les coefficients du filtre.

La fonction des autres modules inclus dans le système est de permettre au système mobile de prendre les bonnes décisions pour faire adapter ou ``geler`` les coefficients du filtre. En effet, l'analyse autorégressive (A.R.) nous permet d'estimer la réponse impulsionnelle du conduit vocal du signal lointain. Cette réponse se trouve grandement atténuée sur la voie microphone par un simple filtrage (on parle de filtrage inverse). Cette technique est utilisée par les systèmes de codage/décodage de type (LPC). Une fois le filtrage inverse effectué, on obtient un nouveau signal que nous appelons ``résiduel``. Nous rappelons que le résiduel dans la technique de LPC est obtenu en effectuant une analyse autorégressive et un filtrage inverse sur un même signal alors que dans notre cas (système main libre) on considère une analyse autorégressive sur le signal lointain et un filtrage inverse sur la signal microphone ce qui génère ce que nous appelons pseudorésiduel. On parlera de pseudorésiduel

pour faire la distinction avec le résiduel du LPC (le vrai résiduel). Le pseudorésiduel se caractérise par un niveau d'écho très affaibli sans toutefois changer le niveau du locuteur local. Donc, le signal pseudorésiduel peut caractériser l'activité vocale du locuteur local seul.

Le DAVRx et DAVTx, déterminent l'activité vocale simultanée du locuteur lointain et du locuteur local. Les algorithmes utilités se basent généralement sur l'énergie du signal

Dans tout le document, la voie ``receive`` ou lointaine indiquera toujours le locuteur lointain. La voie locale ou proche ou chauffeur indiquera uniquement l'activité vocale du chauffeur . La voie micro ou signal microphone se référera à l'activité simultanée des deux locuteurs lointain et proche.

4.3 Base de données

La base de données fournie par Alcatel Mobile Phones, est enregistrée en véhicule dans des conditions et situations très différentes. A chaque simulation est associé un couple de fichiers enregistrés de façon synchrone. Le premier correspond au signal du locuteur lointain qui est capté directement à partir du canal de réception avant de passer par le haut parleur du système mobile. Le deuxième est enregistré sur le canal de transmission sans subir aucun traitement (signal microphone). Normalement, il contient la parole du locuteur

local, l'écho acoustique du locuteur lointain induit par l'habitacle et le bruit environnant.

Chaque simulation caractérise des conditions d'enregistrement bien particulières. Les paramètres variables de ces conditions sont déterminés par:

* Différentes vitesses du véhicule :

la vitesse est maintenue à 0 km/h (le moteur du véhicule est en action);

la vitesse est maintenue à 60 km/h;

la vitesse est maintenue à 90 km/h;

la vitesse est maintenue à 130 km/h;

* Différentes combinaisons d'ouverture et de fermeture des fenêtres :

les fenêtres sont toutes fermées;

seule la fenêtre du côté du chauffeur est ouverte ;

seules les deux fenêtres en avant du véhicule sont ouvertes;

* Différents locuteurs (hommes et femmes).

4.4 Solutions proposées

On présente deux démarches visant à trouver une stratégie permettant la détection de la double parole dans le contexte radiotéléphone main-libre. La première démarche consiste à comparer et analyser d'une part la hauteur tonale entre le signal lointain avec le signal microphone et d'autre part entre le

signal lointain et le signal pseudorésiduel (voir section 4.2). Également, une comparaison basée sur l'extraction de l'enveloppe du signal lointain et du signal microphone a été envisagée. La deuxième démarche consiste à utiliser le principe du filtrage inverse utilisé dans les systèmes du codage/décodage pour filtrer le signal lointain à partir du signal microphone en supposant que l'information tonale est robuste aux distorsions introduites par l'habitacle du véhicule. Cette démarche s'est avérée plus intéressante et exige moins de temps de calcul et peut être facilement implémentée en temps réel.

4.4.1 Méthodologie de la Démarche 1 (comparaison de hauteur tonale)

La stratégie de cette démarche repose sur la richesse contenue dans la hauteur tonale. On a déjà parlé de ce paramètre dans le chapitre 2. Suite à cela, nous supposons que la hauteur tonale peut jusqu'à une certaine limite être résistante aux distorsions introduites par le bruit et l'écho dans le véhicule. Le calcul du fondamental à partir du canal de transmission et de réception constitue la stratégie principale de cette première démarche. Le but est de trouver un critère ou une stratégie fiable pour détecter l'activité vocale de chaque locuteur.

4.4.1.1 Stratégie 1

On a remarqué à partir du spectrogramme du pseudorésiduel (c'est à dire après avoir filtré la contribution vocale du signal lointain à partir du signal microphone) qu'il reste généralement des résidus d'énergie localisés en haute fréquence (2400 Hz et plus) et qui caractérisent surtout l'activité du locuteur lointain Tandis que l'énergie du locuteur local semble couvrir une zone surtout étendue de 700 Hz à 2400 Hz.

Ainsi un détecteur de la hauteur tonale sur la bande de fréquence entre 700 à 2400 Hz environ, devrait permettre la détection de l'activité vocale du locuteur local seul. Pour adapter correctement les coefficients du filtre adaptatif une information supplémentaire sur l'activité vocale du locuteur lointain devra être considérée. Dans un premier temps, on a considéré une détection à base d'énergie pour déterminer l'activité vocale du locuteur lointain.

Comme première stratégie, on a imposé au système d'extraction de la hauteur tonale (E.H.T) de ne considérer dans le traitements que les canaux 1 à 16 (une bande de 400 Hz à 2400 Hz). Ainsi, on serait capable d'extraire seulement la hauteur tonale du locuteur local. Ensuite, on a poursuivi avec d'autres tests en augmentant ou en diminuant le nombre des canaux pour chercher la largeur de bande qui contient le plus d'information utile afin d'extraire la hauteur tonale du locuteur local.

4.4.1.2 Stratégie 2

Dans cette étape on calcule le fondamental avec une sélection des canaux de 7 à 14 (700Hz à 2000 Hz) du signal microphone (ou pseudorésiduel). Ensuite, à l'aide d'une comparaison judicieuse entre les deux fondamentaux, on devrait pouvoir déterminer l'activité vocale de chaque locuteur.

4.4.1.3 Stratégie 3

Elle consiste à utiliser l'opérateur non linéaire Teager [Kaiser, 1990] pour extraire l'énergie du signal de parole (démodulation en AM) en moyenne et haute fréquence à partir du signal du locuteur lointain et du signal microphone. Le choix de cet opérateur Teager a été fait en raison de sa rapidité de calcul et pour les performances intéressantes (partiellement) attribuées au système de suivi de hauteur tonale Algomai94 [Rouat, 1997]. Ensuite on l'a combiné à un autre opérateur Dyn et à un calcul de la corrélation tel que utilisé par le système Algomai94 [Rouat et al, 1992], [Rouat, 1993].

4.4.2 Méthodologie de la démarche 2 (réduction de la contribution glottale du locuteur lointain)

Les résultats de la première démarche furent insatisfaisants et l'investigation d'une autre démarche a été nécessaire. L'approche et les stratégies utilisées dans la deuxième démarche constituent l'objet de cette partie.

L'approche proposée commence par estimer le fondamental du locuteur lointain via un algorithme de suivi du fondamental et de décision voisé/non-voisé basé sur un modèle auditif (Algomai94). On poursuit ensuite le traitement sur le signal lointain en effectuant cette fois-ci une analyse autorégressive pour estimer les coefficients de prédiction. Les coefficients de prédiction permettent de paramétriser la fonction de transfert du conduit vocal. Généralement on utilise entre 8 et 10 coefficients. La deuxième partie du traitement vise à filtrer l'écho acoustique réel (signal lointain) du signal microphone en utilisant les paramètres estimés. L'estimation du fondamental permet de synthétiser un filtre peigne pour supprimer les composantes harmoniques du fondamental. Ce filtrage peut s'effectuer dans le domaine temporel ou spectral. Également la suppression de la contribution vocale est réalisée de la même manière qu'avec le système de référence.

La figure 9 représente la première structure qu'on propose. Elle se distingue de la structure de référence par l'incorporation de l'algorithme d'estimation du fondamental et de suppression de ses composantes.

En fait, on a étudié cinq nouvelles structures. Chaque structure se distingue des autres par une combinaison différente des modules (algorithmes d'analyse autorégressive, algorithme d'estimation du fondamental). Les structures sont représentées dans les figures 9 à 12.

structure 1: L'élimination de la contribution vocale du locuteur lointain est réalisée avant l'annulation du fondamental (voir figure 9).

structure 2: L'annulation du fondamental est effectuée avant l'élimination de la contribution vocale (voir figure 10).

structure 3: On procède par une annulation du fondamental suivie d'une élimination de la contribution vocale. Enfin, on effectue une deuxième annulation du fondamental (voir figure 11).

Les structures 4 et 5 (voir figure 12) sont identiques aux structures 2 et 3 sauf que l'estimation du fondamental est évaluée à partir de la sortie du filtre d'annulation d'écho (filtre adaptatif).

De plus, nous avons étudié trois méthodes permettant de supprimer la contribution glottale du locuteur lointain. On donne le principe de chaque méthode de suppression comme suit:

H1: Atténuation des composantes spectrales les plus 'proches' des harmoniques du fondamental.

H2: Atténuation des composantes spectrales encadrant les harmoniques du fondamental.

H3: Atténuation des composantes spectrales en pondérant les harmoniques du fondamental avec une fonction qui est proportionnelle à leur 'éloignement' .

4.4.3 Critère d'évaluations

Nous avons utilisé le critère RLE d'Alcatel : Rapport Local à Écho. Il définit en dB le rapport entre l'énergie du locuteur local (EnLoc) et l'énergie de l'écho résiduel (EnEcho) pour chaque trame dans les situations où le locuteur lointain est actif. Le RLE est évalué à partir du signal microphone noté Tx(n), ou de signaux calculés à partir de ce signal.

On doit distinguer entre 2 situations possibles: le cas de la double parole (DbTlk(n)=1) ou de l'écho résiduel seul (DbTlk(n)=0). Le VadRx(n) indique l'activité vocale du locuteur lointain (VadRx(n)=1 s'il parle ou VadRx(n)=0 dans le cas contraire).

La mesure du RLE est déterminée par :

$$RLE = 10 * \log_{10}\left(\frac{EnLoc}{EnEcho}\right)$$

Avec:

$$EnLoc = \frac{\sum_{n}^{N} (Tx(n))^2 * VadRx(n) * (1 - DbTlk(n))}{\sum VadRx(n) * (1 - DbTlk(n))}$$

$$EnEcho = \frac{\sum_{n}^{N} ((Tx(n))^2 * VadRx(n) * DbTlk(n))}{\sum_{n}^{N} (VadRx(n) * DbTlk(n))}$$

n = 1 .. N, N est la taille de la fenêtre.

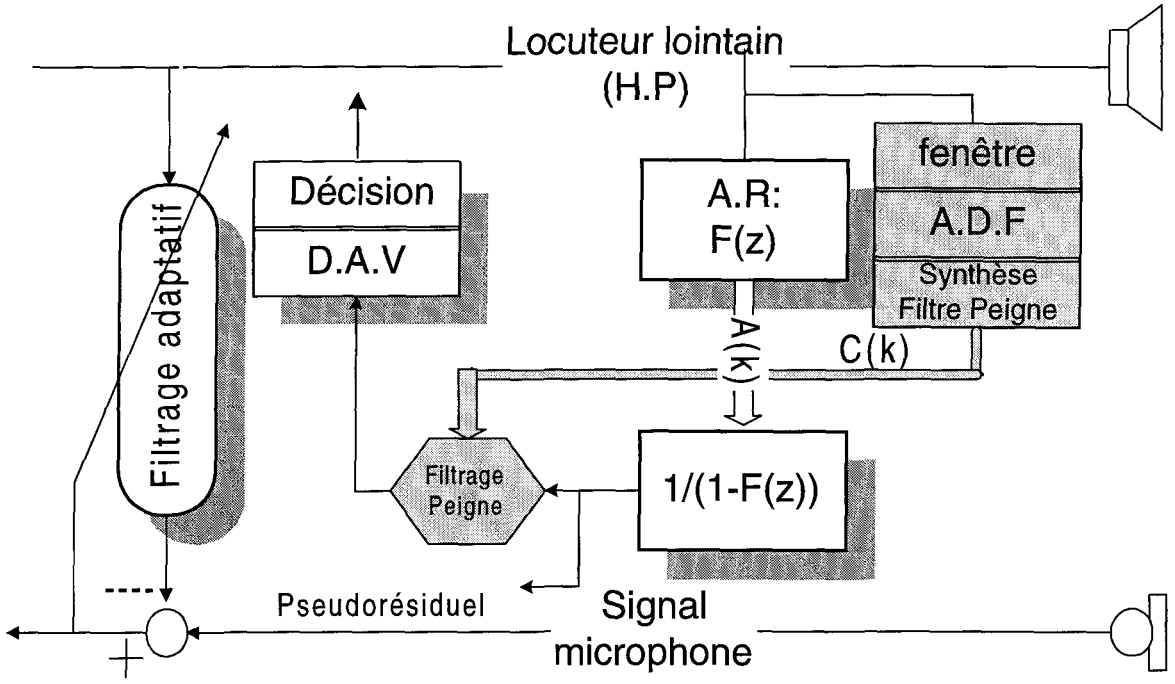


Figure 9 : Structure proposée No 1.

A.D.F : Algorithme de détection du fondamental.

D.A.V : Détecteur d'activité vocale.

A(k) : Coefficients du modèle A.R.
C(k) : Coefficients du filtre peigne.

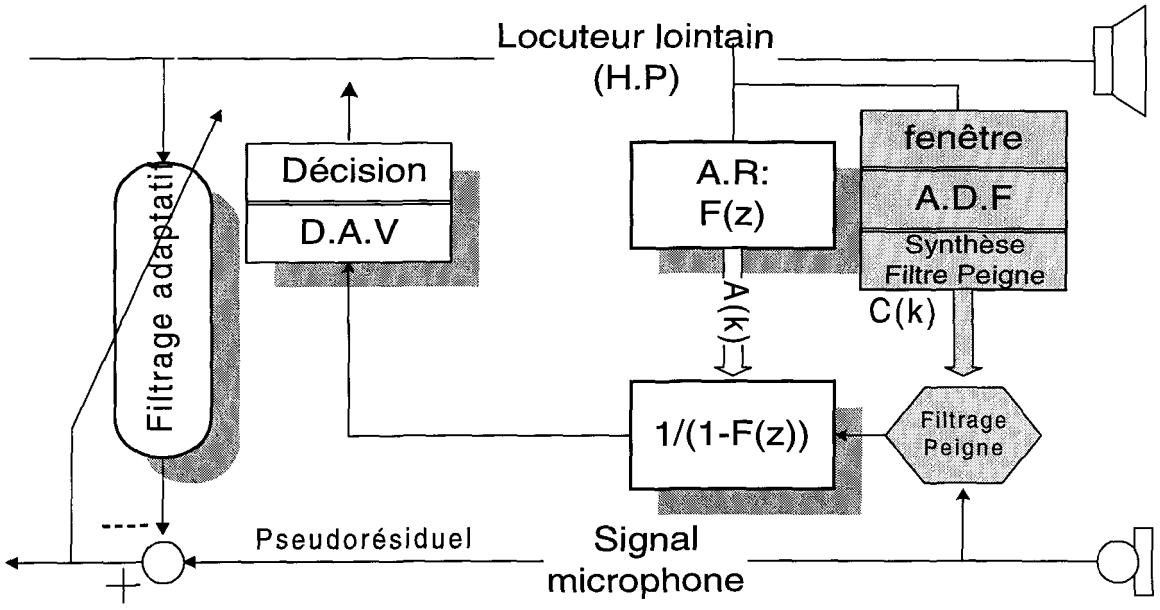


Figure 10 : Structure proposée No 2.

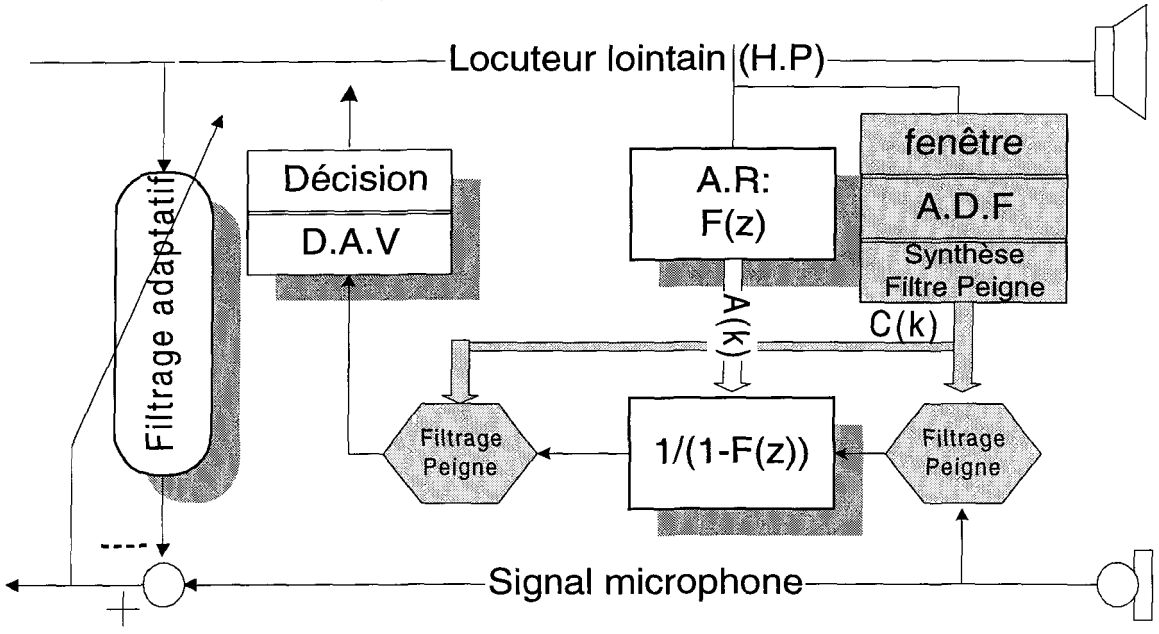


Figure 11 : Structure 3.

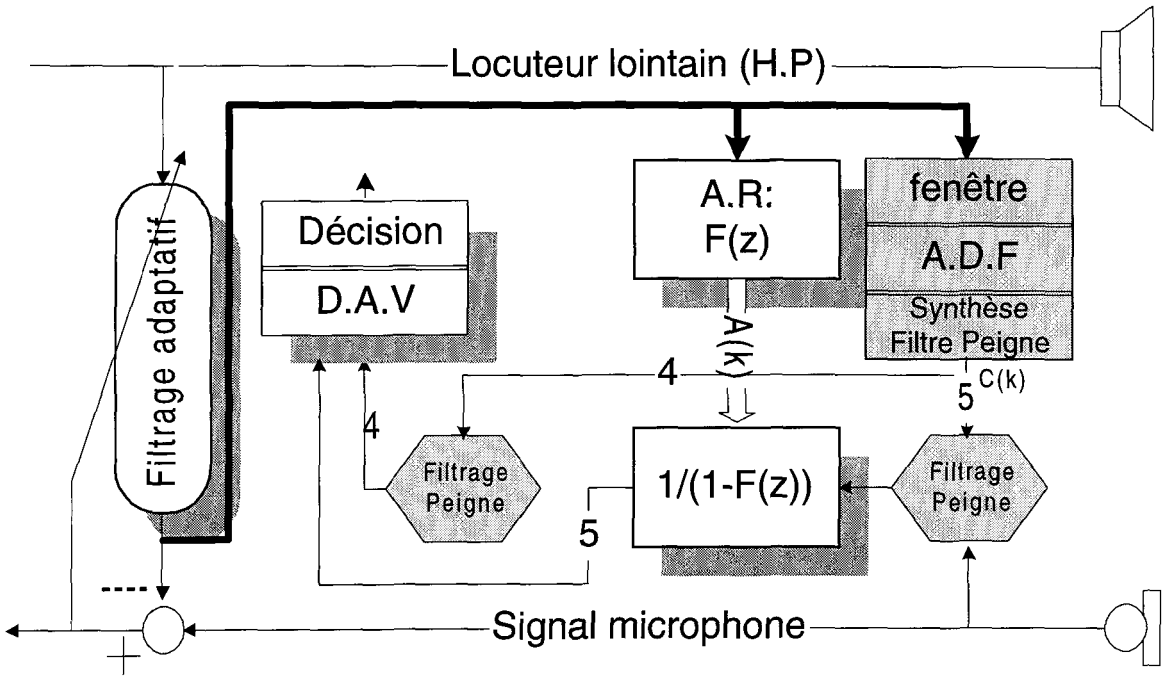


Figure 12 : Structures proposées No 4 et No 5.

Après la synthèse du filtre peigne, le chemin dans la figure portant le numéro 4 caractérise la structure 4, alors que le numéro 5 caractérise la structure 5.

CHAPITRE 5

ANALYSE DES RÉSULTATS

Dans ce chapitre, on présente les résultats obtenus pour les différentes stratégies proposées. Les résultats de la démarche 1 ont été analysés en se basant soit en comparant la valeur du fondamental estimé sur chacun des canaux de communication soit en visualisant les images à analyser dans un espace à 3 dimensions. Les images sont obtenues à partir des sorties et des enveloppes d'un banc de filtres cochléaires appliqué respectivement au locuteur lointain et à la voie microphone. Dans la démarche 2, l'ensemble des résultats des systèmes proposés sont analysés en les comparant au système de référence fourni par Alcatel Mobile Phones par le moyen du critère d'évaluation RLE également fourni par Alcatel Mobile Phones. Ce critère a été présenté au chapitre précédant.

5.1 Analyse et discussion de la démarche 1 (comparaison de la hauteur tonale).

5.1.1 Stratégie 1

On rappelle que cette stratégie 1 consiste à trouver un critère pour la détection de la double parole en comparant conjointement la hauteur tonale du

signal du locuteur lointain et celui du signal pseudorésiduel dans la bande de fréquence de 400 à 2400.

D'après les résultats, on a trouvé que dans le cas de la parole bruitée à l'intérieur de l'habitacle (contexte bruité) la stratégie 1 semble être bien fonctionnelle. Tandis que dans le contexte de la parole propre à l'intérieur de l'habitacle (contexte non bruité), la stratégie 1 semble connaître certaines limites dues au contexte d'utilisation . Plus précisément, la difficulté se manifeste au niveau du signal lointain qui reste détectable dans la bande de 400 Hz à 2400 Hz lorsque l'activité vocale du locuteur local est absente. La raison de ce résultat, est due à la présence d'une faible énergie du signal lointain sur la bande moyenne du pseudorésiduel que le détecteur de la hauteur tonale est capable de distinguer et d'extraire en absence de bruit. Alors qu'en présence de bruit, cette énergie se trouve complètement noyée ou écrasée ce qui a rendu indétectable la hauteur tonale dans ce cas.

En résumé, la détection de la hauteur tonale est fiable pour les signaux enregistrés dans les conditions bruitées. Tandis que pour les signaux enregistrés dans les conditions non bruitées, l'hypothèse proposée dans la stratégie 1 n'est pas toujours vérifiée. En conséquence, on rejette l'hypothèse de cette stratégie qui exclut la présence du signal du locuteur lointain dans la bande de fréquence de 400 Hz à 2000 Hz.

5.1.2 Stratégie 2

On rappelle que cette stratégie 2 consiste à trouver un critère pour la détection de la double parole en comparant conjointement la hauteur tonale du signal du locuteur lointain à celle du signal microphone.

La stratégie 2 a connu le même sort que la stratégie 1. Dans le cas des signaux propres la stratégie 2 semble bien fonctionner malgré les distorsions subies par le signal lointain une fois capté par le microphone. Dans le cas des signaux bruités, la mesure du fondamental du locuteur lointain subit d'importantes distorsions une fois estimée sur le canal de transmission et par conséquent on ne peut plus se fier à un simple critère de comparaison des hauteurs tonales. Cette stratégie se trouve également écartée puisque qu'elle reste dépendante du contexte d'utilisation.

5.1.3 Stratégie 3

On rappelle que cette stratégie 3 consiste à trouver un critère pour la détection de la double parole en comparant cette fois-ci conjointement l'enveloppe du signal du locuteur lointain avec le signal microphone ou le signal pseudorésiduel.

Dans cette stratégie, l'interprétation des graphiques représentée en 3 dimensions étaient utilisée comme critère pour trouver un scénario possible pour la détection de la double parole. Les graphiques nous permettaient d'analyser facilement le comportement et les transformations du signal des deux locuteurs.

Les premiers tests consistaient en un traitement des signaux de parole par un banc de filtres passe-bande du système Ampex [Van Immersel et Martens 1992], constitué par des filtres IIR. Ensuite, Teager [Kaiser, 1990] est appliqué pour extraire l'enveloppe du signal. Un autre opérateur Dyn [Rouat, 1993] a été appliqué à la place de Teager pour amplifier la modulation spécifique à la parole et éliminer les entrelacements de l'enveloppe (voir graphique 13). Les résultats visualisés par des images en 3D, étaient ambigus et difficiles à interpréter lors de la comparaison de la voie receive avec la voie microphone ou le signal du pseudorésiduel. En effet, la comparaison des résultats graphiques était supposée nous fournir une stratégie de détection pour la double parole.

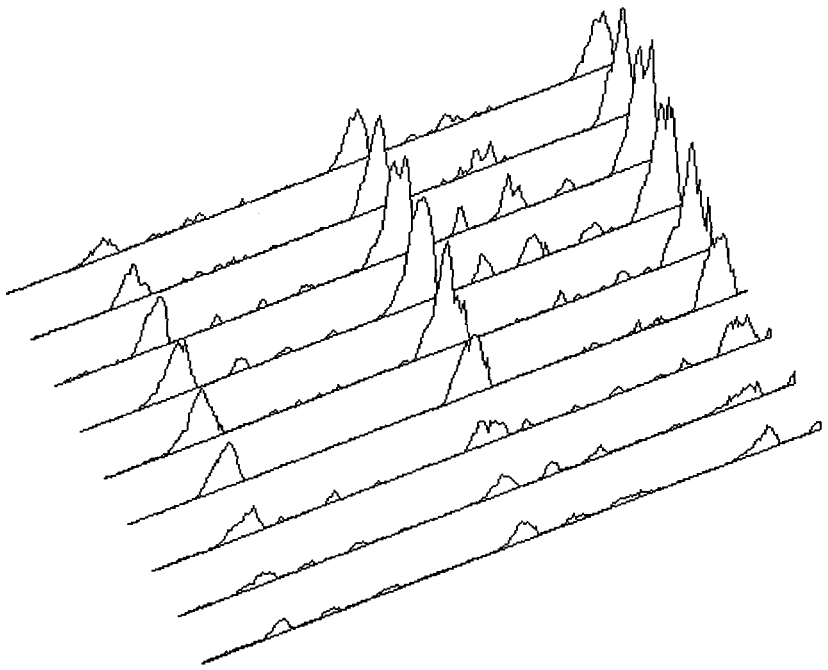


Figure 13 : Traitement de la voie micro avec l'operateur Dyn en moyenne et haute fréquence, l'alignement des pics élevés correspond aux instants d'excitation glottale.

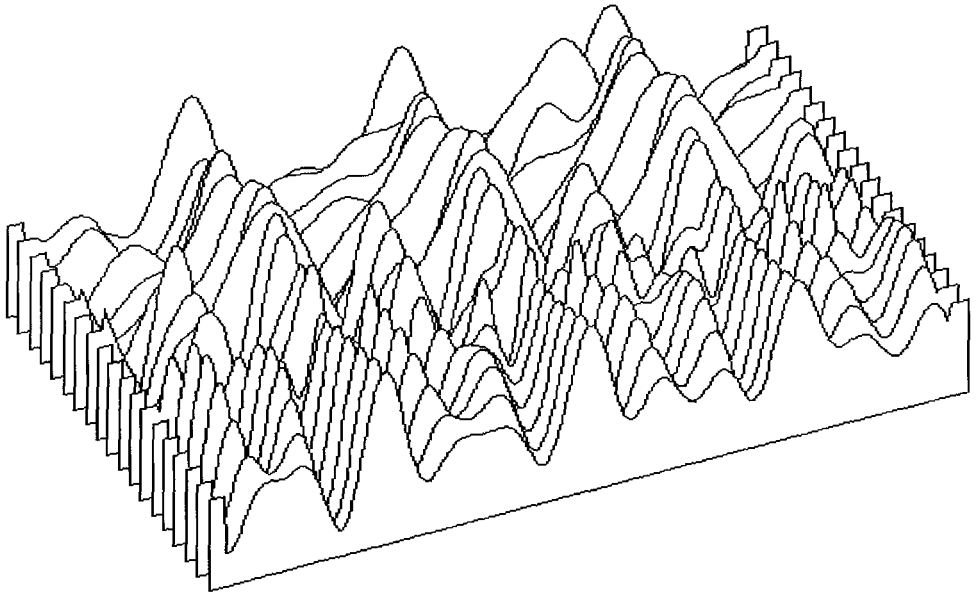


Image 3 D du signal haut parleur: consonne /d/ dans le contexte "de", (Homme).

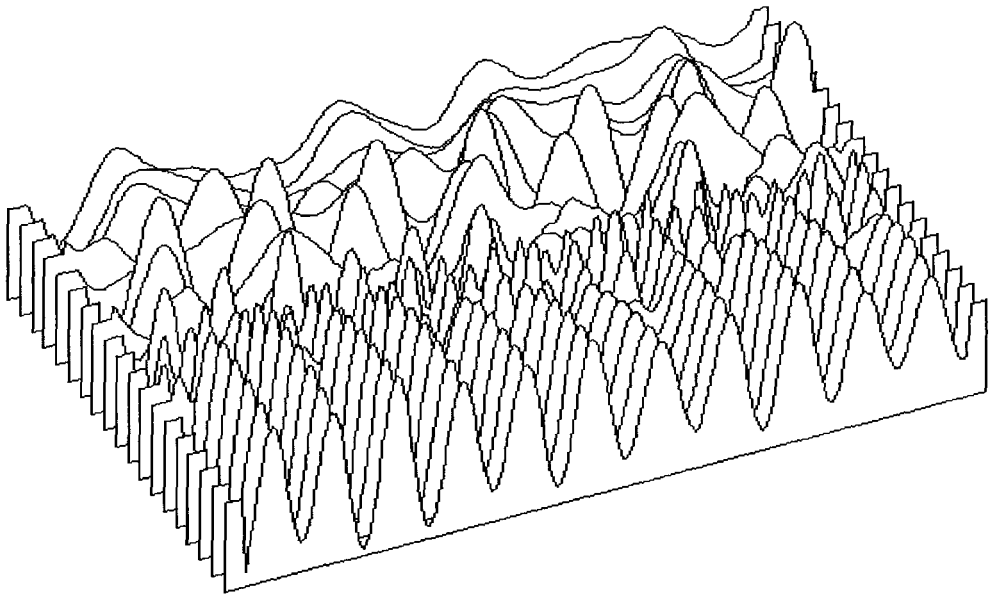


Image 3 D du signal microphone: consonne /d/ dans le contexte "de", (écho seul).

Figure 14 : Traitement du signal du locuteur lointain (16 ms) à partir de la voie haut parleur et la voie microphone dans la situation où seul l'écho est présent (aucune activité vocale du locuteur local). Les traitements effectués aux signaux haut parleur et microphone sont :application d'un banc de filtres cochléaires, extraction d'enveloppes, filtrage passe-bas, filtrage passe-haut et finalement le calcul de corrélation.

D'après les résultats, l'application seule de Teager ou Dyn n'est pas suffisante pour trouver un critère de détection de la double parole. La raison peut être due aux fluctuations introduites par le bruit sur le signal de la voie locale qui rendent la largeur de bande de l'amplitude ou de la phase non faible par rapport à celle de la porteuse et par conséquent le critère de Teager se trouve non respecté.

L'ensemble des tests ont été repris en filtrant la sortie de Teager passe-bas (fc = 250 hz, Butterworth) et passe-haut (62.5 hz, filtrage à moyenne sur une fenêtre de 16 ms). Finalement, une corrélation est calculée sur chacun des canaux telle que utilisée dans Algomai94. Les résultats étaient plus clairs et riches d'informations par rapport aux derniers tests. On arrivait facilement à distinguer la répartition des impulsions glottales (lorsqu'elles existent) et les formants d'un signal de parole à partir de la représentation graphique en 3D (voir graphiques 14, 15 et 16).

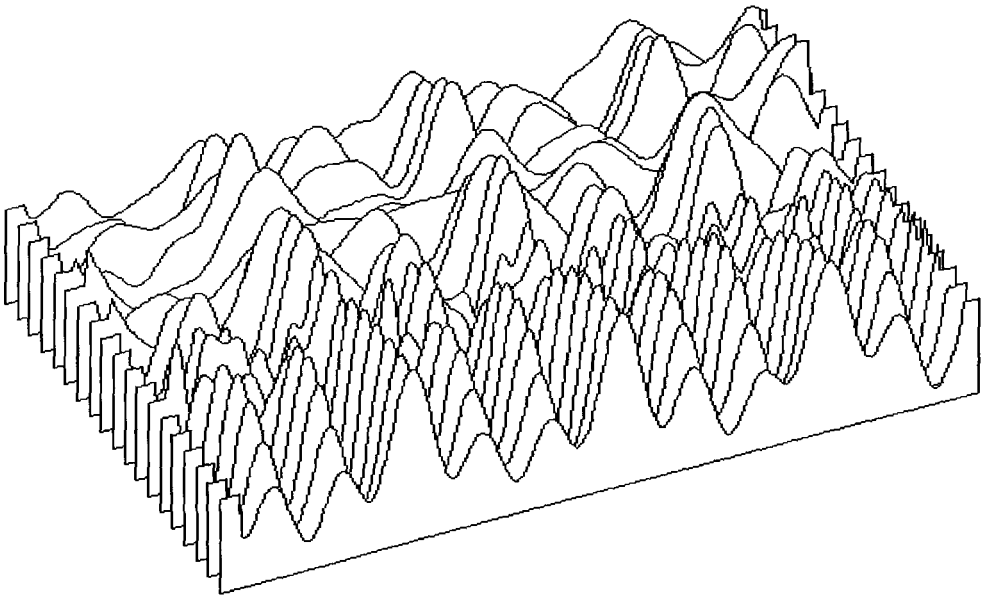


Image 3 D du signal haut parleur: voyelle /a/ dans le contexte "la", (Homme).

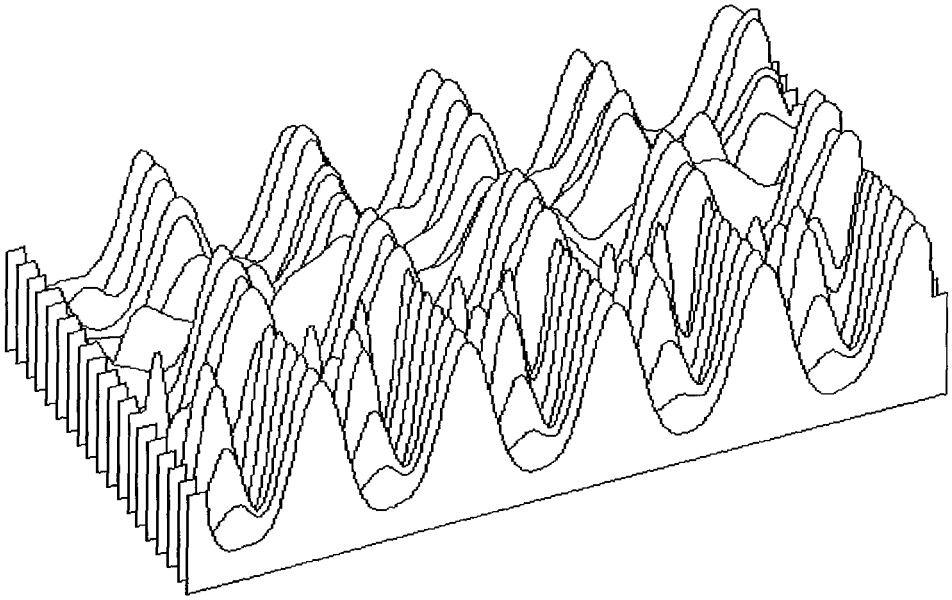


Image 3 D du signal microphone dans la situation de double parole: voyelle /a/ prononcée par le locuteur lointain et voyelle /i/ prononcée par le locuteur local.

Figure 15 : Traitement du signal du locuteur lointain (16 ms) à partir de la voie haut parleur et la voie microphone dans la situation de double parole. Les traitements effectués aux signaux haut parleur et microphone sont :application d'un banc de filtres cochléaires, extraction d'enveloppes, filtrage passe-bas, filtrage passe-haut et finalement le calcul de corrélation.

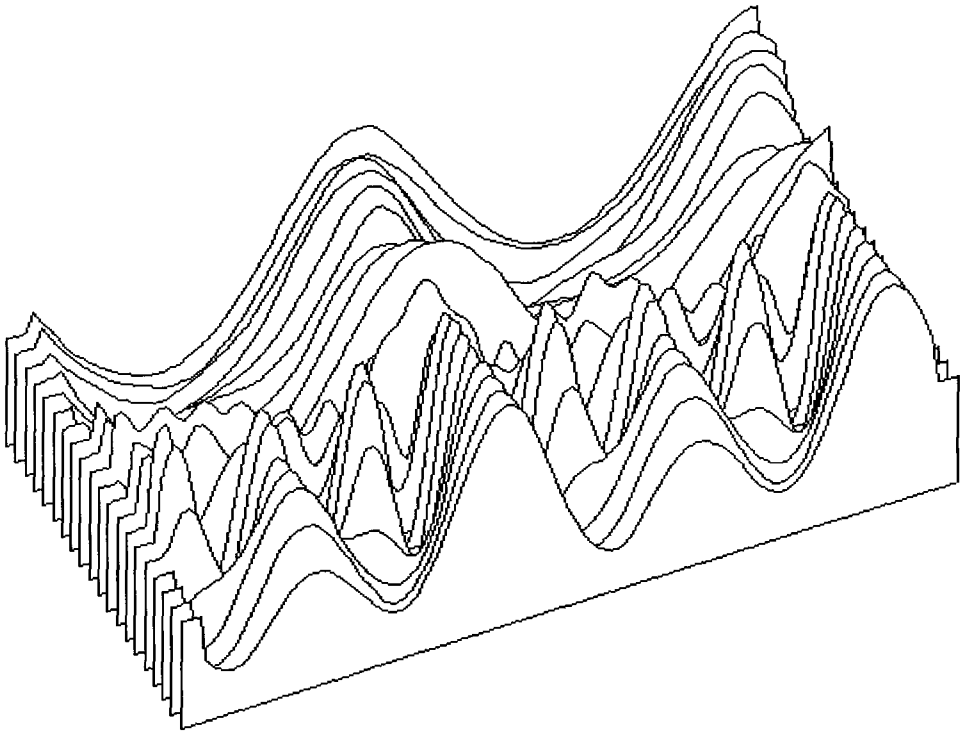


Figure 16: Résultat obtenu du signal microphone (receive silence, locale voisée) après un traitement par banc de filtres, filtrage passe-bas, fitrage passe-haut et la corrélation.

En dépit de l'information utile fournie par les résultats de cette dernière méthode proposée, l'objectif de trouver un critère fiable pour détecter l'activité vocale de chacun des locuteurs n'a pas été atteint pour différentes raisons. Parmi ces raisons, on trouve que la largeur de la bande moyenne et haute fréquence seulement ne permet pas d'extraire toute l'information spectrale caractérisant un signal. Par exemple, il existe des sons voisés caractérisés par la présence d'un premier formant en basse fréquence et dont le deuxième se situe à environ 3500 Hz. Ainsi la méthode proposée ne peut pas extraire l'information pertinente du signal localisé hors de sa bande de fréquence. De plus, on a observé dans la

situation d'écho seul que le signal receive subit des distorsions introduites par le couplage du haut-parleur et du microphone. Ces distorsions modifient considérablement les caractéristiques temporelles et spectrales du signal original et par la suite on ne peut se fier aux paramètres estimés du signal lointain à partir de la voie receive pour ensuite les retrouver sur la voie microphone. Il est à noter que la voie receive est filtrée passe-bande par le téléphone et par conséquent les derniers canaux du système proposé peuvent donner des résultats erronés. Si on veut cerner toute l'information spectrale d'un signal, il sera nécessaire de considérer les canaux en basse fréquence.

5.1.4 Discussion

On remarque que les impulsions glottales et les formants de la voie receive sont généralement visibles et répartis régulièrement sur les canaux 1 à 14. On peut dire que les distorsions spectrales subies par le signal lointain en haute fréquence sont dues au filtrage passe bande de la ligne téléphonique. A la réception de ce même signal lointain sur la voie micro dans la situation de la présence uniquement de l'écho (seul le locuteur lointain est actif), le même nombre d'impulsions glottales et de formants sont observées. Une différence importante se manifeste par un étalement des corrélations (près des formants) vers les basses fréquences ce qui implique le rétrécissement des pics de corrélation à l'impulsion glottale observée (voir graphiques 14 et 15). Cette discussion considère le cas d'un signal non bruité. Dans le cas bruité, on peut dire

que le nombre de canaux renfermant de l'information utile se réduit davantage sur la voie micro.

Pour un voisement sur la voie micro (seul le chauffeur parle), l'impulsion glottale est répartie presque de façon régulière sur tous les canaux 1 à 20 (voir graphique 16).

En ce qui concerne en général la double parole, les canaux de 14 à 20 (voie micro) gardent trace de la voie locale alors que les autres canaux sont plutôt caractérisés par une interférence des deux voies. Par exemple, on remarque que le nombre d'occurrence d'impulsion glottale sur la voie receive (voisée) est différent de la voie micro surtout pour le type voisé.

Pour les non-voisés sur la voie microphone, on remarque le changement de répartition des formants de la voie receive quand ils interféraient avec ceux de la voie locale sur la voie micro .

5.2 Analyse et discussion de la démarche 2 (réduction de la contribution glottale du locuteur lointain)

On rappelle que les structures proposées consistaient à combiner de différentes manières dans le système main libre les algorithmes d'estimation du fondamental et de la contribution vocale sur la voie receive d'une part et d'autre part les algorithmes d'annulation du fondamental et de la contribution vocale sur la voie microphone (voir section 4.2.2).

5.2.1 Résultats

Les mesures du critère RLE pour évaluer les structures proposées, sont reportées aux tableaux 1, 2 et 3. Les mesures en dB représentent une moyenne évaluée à partir d'une série de fichiers de parole. Une mesure positive peut être interprétée comme un gain apporté par la structure en question afin de discriminer davantage la situation de la double parole vis à vis de la structure de référence. Une mesure négative au contraire indiquera une perte de discrimination de la double parole par rapport à la situation de référence. La moyenne est calculée pour une série de fichiers de parole se caractérisant par la même vitesse du véhicule et pour lesquels on a utilisé une même structure (1, 2 ou 3) et une même méthode de suppression des harmoniques (H1, H2 ou H3).

	RLE (dB)		
	H1	H2	H3
0 km/h	1.31	1.90	2.05
60 km/h	0.46	0.96	0.82
90 km/h	0.23	0.44	0.95
130 km/h	1.091	0.568	0.24

Table 1: résultats de la structure 1

On remarque que la valeur moyenne en dB du critère RLE est positif (un gain en dB) dans toutes les situations d'enregistrements pour les différentes

structures et méthodes d'annulation du fondamental proposées. On en déduit que la démarche consistant à inclure un estimateur du fondamental et de procéder ensuite à son annulation par un filtre peigne excède les performances du système de référence pour les 3 premières structures. Un gain de 3 dB environ sera considéré comme une amélioration significative pour justifier l'adoption de l'approche proposée (système actuel). On mentionne qu'on a rencontré des situations où la mesure du critère RLE est négative avant d'effectuer le calcul de la moyenne RLE en dB sur l'ensemble des fichiers de parole appartenant à la même catégorie. Ces situations ont été parfois rencontrées avec des voix de femmes lorsqu'on a utilisé les structures 1 et 2 avec H1 ou H2. C'est une situation où le critère RLE donne des résultats légèrement plus élevés que la structure proposée en question.

	RLE (dB)		
	H1	H2	H3
0 km/h	1.21	1.56	2.02
60 km/h	0.22	0.53	0.77
90 km/h	0.75	0.31	0.95
130 km/h	0.51	0.22	0.24

Table 2: résultats de la structure 2

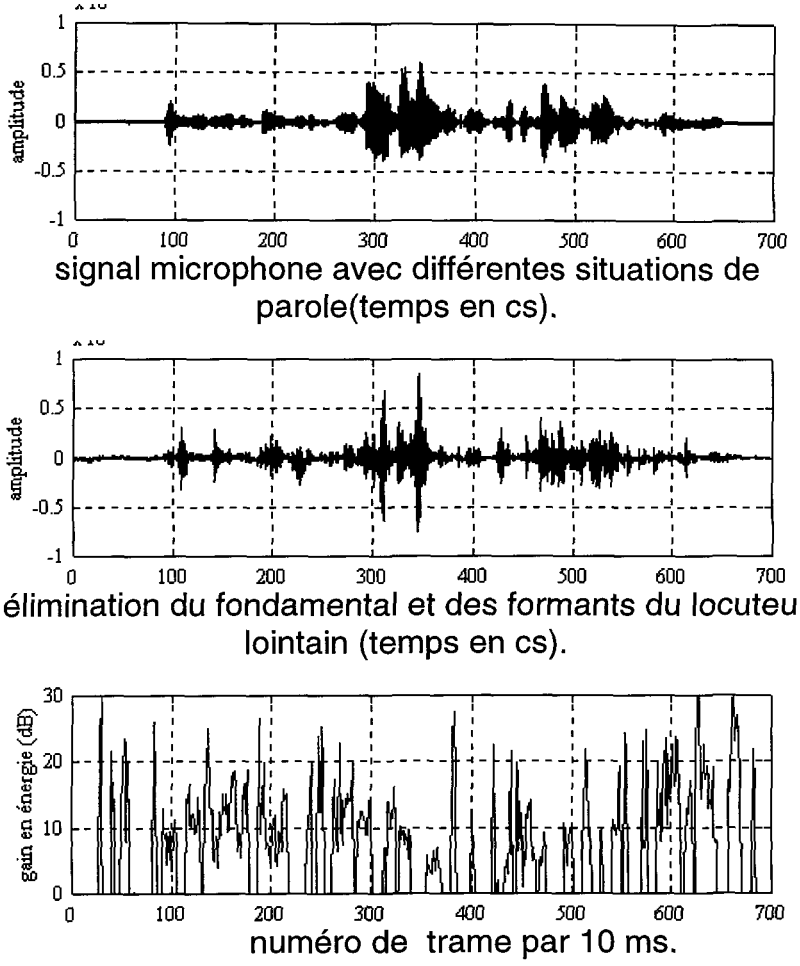


Figure 17: Traitement du signal microphone (homme, 0 km/h, fenêtres fermées) avec la structure 1 et la stratégie H1. (temps cs)

Les résultats présentés sur les tableaux, se réfèrent au cas où les fenêtres du véhicule sont entièrement fermées. Les notations H1, H2 et H3 caractérisent les stratégies de suppression des harmoniques telles que décrites précédemment.

La structure 1 est meilleure que la structure 2 pour les conditions de 0 km/h et 130 km/h (Tables 1 et 2). Ceci correspond à la situation pour laquelle la fonction

de transfert caractérisant le couplage et le bruit à l'intérieur du véhicule varie légèrement. Au contraire, la structure 2 n'est pas aussi bonne que la structure 1 à 90 Km/h et 60 Km/h car le couplage est plus fort.

	RLE (dB)		
	H1	H2	H3
0 km/h	2.90	3.53	3.5
60 km/h	3.33	4.13	4.5
90 km/h	4.00	5.31	5.0
130 km/h	4.50	4.7	4.0

Table 3 : résultats de la structure 3

La structure 3 semble donner les meilleurs résultats avec le critère RLE pour les 3 structures proposées dans les différentes conditions de conduite (voir Table 3). La méthode H3 semble être la meilleure technique de suppression du fondamental et de ses harmoniques. En fait, la structure 3 et la structure 2 sont supposées être comparables puisque la seule différence est qu'on a ajouté une annulation du fondamental dans la structure 3. La différence importante en terme du gain RLE est peut être due au processus de suppression des formants qui probablement, fait ressortir les harmoniques du locuteur lointain. D'après les graphiques, il semble que la structure 3 introduit moins de distorsions au signal

traité que les autres structures (conditions où la vitesse du véhicule est supérieure à 0 km/h).

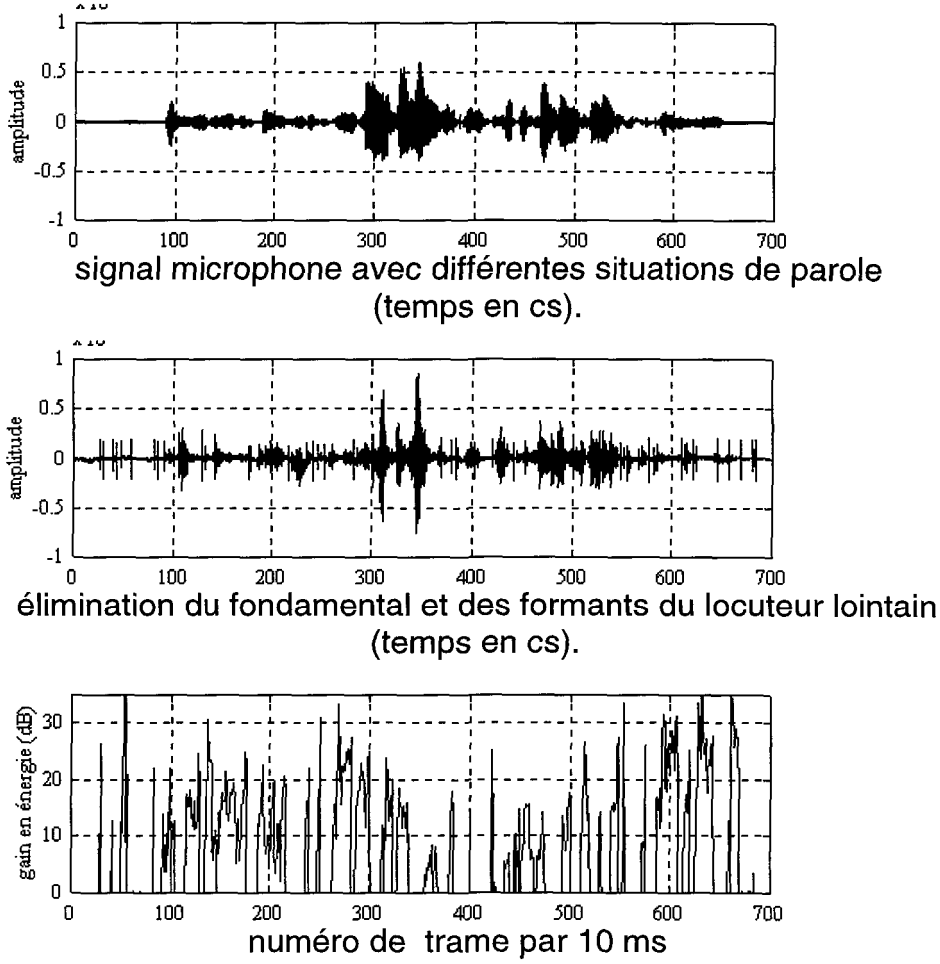


Figure 18: Traitement du signal microphone (homme, 0 km/h, fenêtres fermées) avec la structure 3 et la stratégie H1. (temps cs)

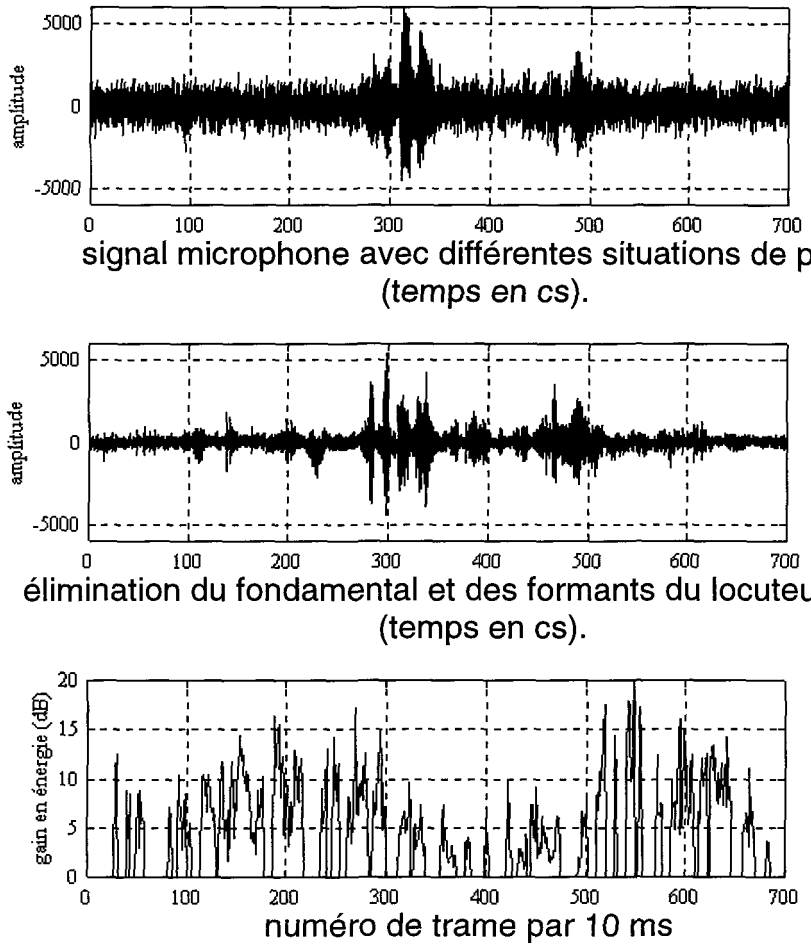
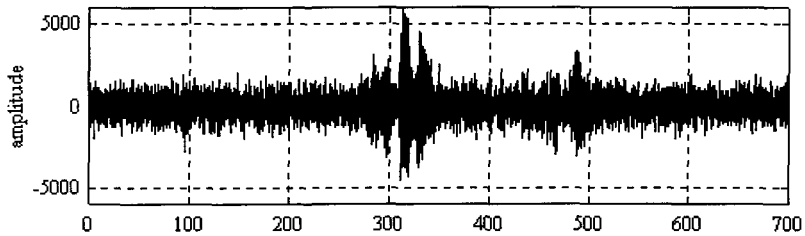
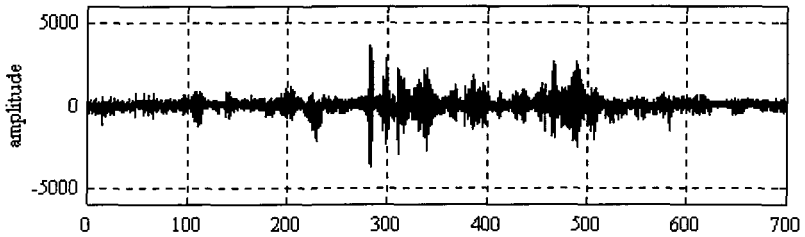


Figure 19: Traitement du signal microphone (homme, 90 km/h, fenêtres fermées) avec la structure 3 et la stratégie H1. (temps cs)

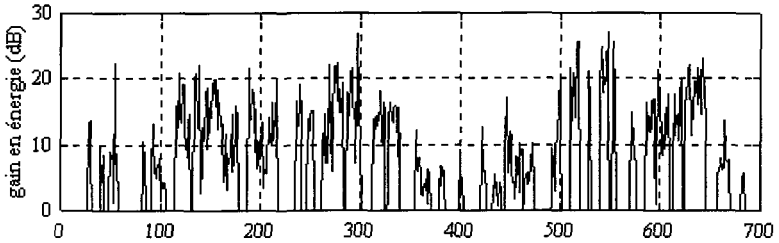
Les figures 17 à 22 illustrent le traitement réalisé pour différentes structures et différentes méthodes de suppression du fondamental. On remarque parfois dans le cas de la parole propre à 0 km/h des distorsions qui apparaissent sur le signal traité (microphone) après avoir éliminé le fondamental et les formants du locuteur lointain. Dans toutes les autres conditions de déplacement du véhicule (60 km/h,90 km/h, 130 km/h et même parfois à 0 km/h) on n'a enregistré aucune distorsion du signal après le traitement. On peut dire que les distorsions sont noyées dans le bruit.



signal microphone avec différentes situations de parole (temps en cs).

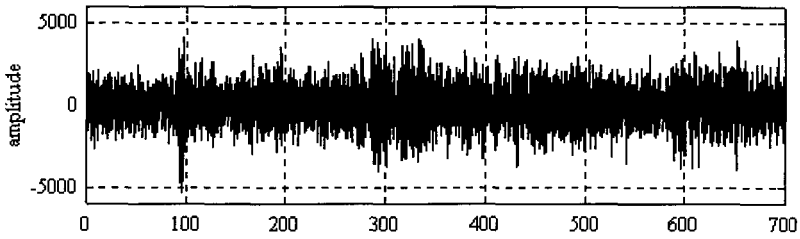


élimination du fondamental et des formants du locuteur lointain (temps en cs).

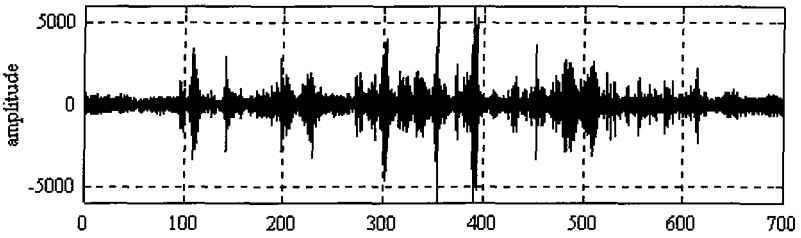


numéro par trame par 10 ms

Figure 20: Traitement du signal microphone (homme, 90 km/h, fenêtres fermées) avec la structure 1 et la stratégie H2. (temps cs)



signal microphone avec différentes situations de parole (temps en cs).



élimination du fondamental et des formants du locuteur lointain (temps en cs).

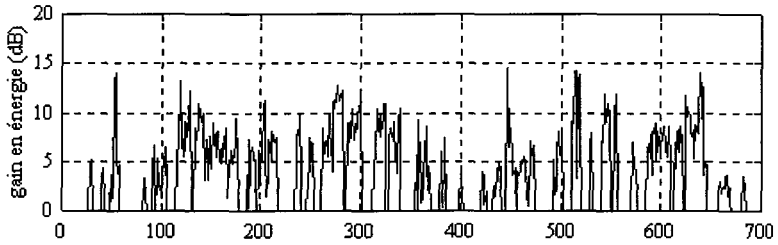
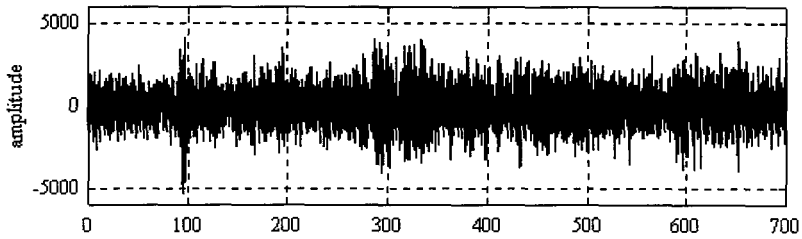


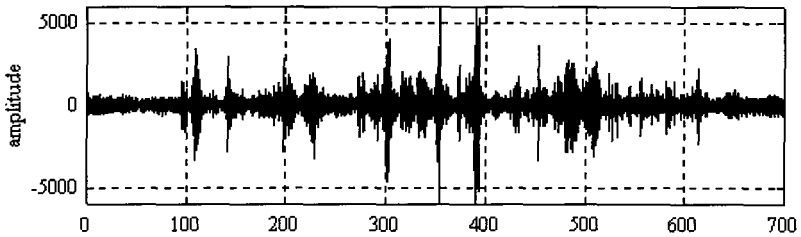
Figure 21:Traitement du signal microphone (femme, 90 km/h, fenêtres fermées) avec la structure 1 et la stratégie H3. (temps cs)

Les résultats des structures 4 et 5 ont été jugés non pertinents et non intéressants pour deux raisons principales. La première est due au fait d'estimer le fondamental à partir de l'estimé de l'écho correspondant au signal à la sortie du filtre adaptatif. Cet estimé de l'écho connaît plusieurs distorsions pour modéliser l'impulsion glottale de l'écho réel. Ces distorsions sont peut être liées au nombre limité des coefficients du filtre adaptatif utilisé pour estimer l'écho réel. Cette mauvaise estimation du fondamental peut entraîner dans certains cas un

fonctionnement désastreux du système main libre. Prenons la situation où le fondamental du locuteur lointain et du locuteur local sont respectivement de 200 et 300 hz et qu'on a estimé un fondamental de 100 hz à partir de l'estimé de l'écho. Après le filtrage inverse, la contribution glottale des deux locuteurs est annulée ainsi que la contribution vocale du signal lointain et une partie de la contribution vocale du signal local. Par conséquent, l'énergie du signal microphone se trouve considérablement affaiblie et la situation de la double parole peut être jugée à tord comme une situation d'écho seul.



signal microphone avec différentes situations de parole (temps en cs).



élimination du fondamental et des formants du locuteur lointain (temps en cs).

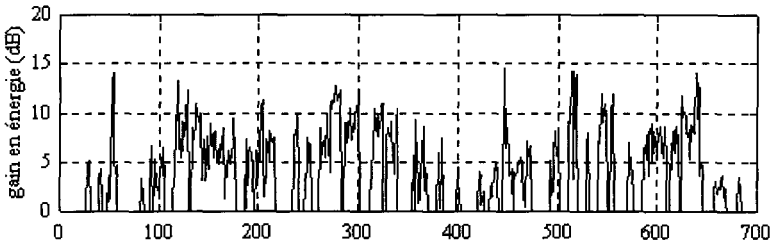


Figure 22: Traitement du signal microphone (femme, 90 km/h, fenêtres fermées) avec la structure 1 et la stratégie H3. (temps cs)

CHAPITRE 6

CONCLUSION

Différentes approches ont été présentées, utilisant la notion de fréquence glottale, pour résoudre le problème de détection de double parole dans le contexte radio mobile.

Une détection sur chacun des canaux de transmission a été étudiée. Malheureusement, l'insensibilité du fondamental aux distorsions introduites par l'habitacle n'est pas toujours vérifiée. De plus, la similarité du fondamental entre les locuteurs peut soulever d'autres problèmes.

Une détection sur un canal a été étudiée. Elle consiste à rehausser le signal du locuteur local (voie microphone) vis-à-vis de l'écho. L'introduction du détecteur de fréquence glottale dans ce cas, s'est avérée plus intéressante et un gain supérieur à 3 dB en moyenne a été atteint pour toutes les structures proposées.

L'introduction d'un réducteur de bruit, n'a pas été étudiée. Mais, il semble être un facteur important pour améliorer davantage les performances enregistrées. En effet, l'annulation du fondamental et l'élimination de la

contribution vocale du locuteur lointain sur le microphone donne naissance à un écho de nature aléatoire (bruit). Cet écho pourrait être complètement supprimé en lui appliquant un algorithme de réduction de bruit.

Nous pensons qu'il est donc possible d'intégrer cette technique à un algorithme de réduction de bruit afin d'obtenir des performances supérieures à celles décrites ici. Nous rappelons que l'évaluation a été faite après suppression de la contribution du locuteur lointain sans réduction à priori ou à posteriori de bruit. L'intégration de systèmes mixtes (intégration de connaissances perceptives avec les algorithmes standards de traitement des signaux) devrait donc permettre d'améliorer les systèmes actuels de reconnaissance de parole.

Nous avons également, mis au point un algorithme d'extraction de la hauteur tonale qui fonctionne en virgule fixe. L'algorithme proposé dans sa grande partie, est une version en virgule fixe du système Algomai94 (virgule flottante). L'évaluation des performances de cet algorithme a été effectuée avec succès comparativement à la performance des systèmes à virgule flottante Algomai94 et Ampex, connus pour leurs robustesses et efficacité. On a utilisé la même banque de données pour l'évaluation des performances. L'utilisation de l'opérateur Teager pour extraire l'enveloppe en moyenne et haute fréquence a été sensible à l'implémentation en virgule fixe. Le seuil de voisement pour distinguer entre les sons voisés et non-voisés a été implémenté temporairement en l'estimant à partir des quatre premières trames.

Puisque la version actuelle ne nécessite pas beaucoup de calculs, nous pensons que dans la prochaine version il y aura lieu de supprimer l'opérateur Teager et d'utiliser une transformée de Hilbert pour arriver à extraire l'enveloppe du signal sans être freiné par les problèmes de débordement. Ainsi, nous serions capables de chercher les harmoniques non résolus avec assez de précision et par conséquent réduire davantage les erreurs fines et les erreurs causées par les formants. Egalement, un seuil adaptatif sera envisagé pour réduire la confusion entre les segments voisés et non voisés et par conséquent augmenter la robustesse du système proposé. De plus, suivant les besoins spécifiques, il y aura lieu de modifier les stratégies de décision voisés/non-voisés et de mesure de la fréquence glottale.

BIBLIOGRAPHIE

C. Abry, L.-J Boë, R. Descout, M. Gentil, P.Graillot (1980), "Labialité et phonetique. Données fondamentales et études expérimentales sur la géometrie et la motricité labiales,". Publications de l'Université des Langues et Lettres de Grenobles, p.304.

M. Allerhand, R. Patterson (1992), "Correlograms and auditory images,". In Procedeedings of the Institute of Acoustic, Vol 14, Part 6, pp 281-288.

B.S. Atal and S. L. Hanauer (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,". J. Acoust. Soc. Amer. Vol 50, pp 637-655.

C. Baillargeat (1991). "Contribution à l'amélioration des performances d'un radiotéléphone main-libre à commandes vocales,". Thése de l'Université Paris VI.

M. Berouti R. S, R. Schwartz, J. Makhool (1979), "Enhancement of Speech corrupted by acoustic noise,". IEEE ICASSP , pp 208-211.

L.-J Boë, R. Descout, B. Guerin (1980), "Larynx et Parole,". GALF, institut de Phonétique de Grenoble.

E. Bognar (1980), "Préliminaires anatomiques à l'études phonétique des mouvements mandibulaires,". Travaux de l'institut de Phonétique de Grenoble.

R. Boite et M. Kunt (1987), "Traitement de la parole,". Presses polytechniques romandes. Lausanne.

S.F. Boll (1979), "Suppression of acoustic noise in speech using spectral substraction,". *IEEE Trans. on ASSP*, Vol. 27, No. 2, pp. 113-120, April.

N. D. Degan and C. Prati (1988), "Acoustic noise analysis and speech enhancement techniques for mobile radio applications,". *Signal Processing* 15, pp 43-56.

J.R Deller, J.G Proakis et J. H. L. Hanser (1993), "Dicrete-time Processing of Speech Signals,". Editor J. Griffin, Macmillan Publishing.

P.N Denbigh and J. Zhao (1992), ″Pitch extraction and separation of overlapping speech,″. Speech Communication 11, pp. 119-125.

M Draper, F. Ladefoged, D. Whiteridge (1959), "Respiratory Muscles in Speech,". J. Speech Hearing Res.,2,16-27.

H. Dudley (1959), "The vocoder". Bell laboratories record, pp. 122-126.

F. D. Elliot (1987), " Handbook of digital Signal Processing: Engineering Applications,", Academic Press inc.

Y. Ephraim, D. Malah (1984), "Speech enhancement using a minimum mean square error short time amplitude estimator,". *IEEE Trans. on ASSP*, Vol. 32, No. 6, pp. 1109-1121, December.

J. L. Flanagan and M. G. Saslow (1958), "Pitch discrimination for synthetic vowels". J. Acoust. Soc. Am. 30:435-442.

E.K Gary, A. J. Oppenheim and J. M. Tribolet (1977), ″Speech Analysis by Homomorphic Prediction,″. IEEE Trans. on Acoust.,Speech, and Signal Processing, Vol. ASSP-25, N0.1.

M Gentil, L.J Boë, R. Descout (1980), "Etude EMG des lèvres (OOINF, MENT, DLI et LLSA) dans la réalisation de monoyllabes du français. dans: Labialité et Phonétique,".240-262, Abry et al (1980).

A. Gilloire and M. Vitterli (1988), ″Adaptive filtering in sub-bands,″. IEEE ICASSP, pp. 1572-1575.

J. L. Goldstein (1973), "An optimum processor theory for the central formation of the pitch of complex tones,". J. Acoust. Soc. Am. 54:1496-1516.

W. -J Hardcastle (1976), "Physiology of Speech Production,". Acad. Press. London.

W. J. Harris et N. Umeda (1987), "Difference limens for fundamental frequency contours in sentences,".J. Acoust. Soc. Am. 81:1139-1145.

S. Haykin (1991), "Adaptive filter theory,". *Second Edition, Prentice-Hall*, Englewood Cliffs, New Jersey.

Hermes, Dik J.(1988), "Measurement of pitch by subharmonic summation,". J. Acoust. Soc. Am. 83, No 1:257-263.

W. J. Hess (1983), "Pitch Determination of speech Signal-Algorithms and Devices,". Springer-Verlag, Berlin.

W. J. Hess et H. Indefrey (1987), "Accurate time-domain pitch determination of speech signals by means of a laryngograph,". Speech Commun. 6:55-68.

W. J. Hess (1991), "Pich and Voicing Determination,". Advances in Speech Signal Processing, edited by Sadaoki Furui et M. Mohan Sondhi.

J. F. Kaiser (1990). "On a simple algorithm to calculate the 'energy' of a signal,". *Actes de IEEE-ICASSP'90*, Albuquerque, pp.381-384.

J. F. Kaiser (1993). "Some Useful properties of Teager's energy operators,". *Actes de IEEE-ICASSP 93* vol. 3, pp.149 -152.

D. Klatt (1973). "Discrimation of fundamental frequency contours in synthetic speech: Implication for models of speech perception". J. Acoust. Soc. Am. 53:8-16.

E. Koo, J. D. Gibson et S. D. Gray (1989), "Filtering of colord noise for speech enhancement and coding,", IEEE ICASSP 1989, pp. 349-352.

A.M. Kondoz (1994), " Digital Speech: Coding Low Bit Rate Communications Systems,". John Wiley Sons.

J. W. Kim and A. K. UN (1986), "Enhancement of noisy speech by forward/backward adaptative filtering,". IEEE ICASSP 1989, pp. 349-352.

J.S. Lim, A.V. Oppenheim (1979), "Enhancement and bandwidth compression of noisy speech,". *Proc. of the IEEE*, Vol. 37, No. 12, pp. 1586-1604, December.

Y. C. Liu (1992), "Un Détecteur perceptif de la hauteur tonale pour la parole téléphonique,". Université du Québec à Chicoutimi.

P. Lockwood and al (1991). "Noise reduction of speech enhancement in cars: non-linear spectral substraction/Kalman filtering,". EUROS-SPEECH, pp 83-86.

R. J. Macaulay and M.L Malpass (1980). "Speech enhancement using decision noise suppression filter,". IEEE trans. on ASSP, Vol. 28, No 2, pp. 137-145.

O. Macchi, M. Bellanger (1988), "*Le filtrage adaptatif transverse,*". *Traitement du Signal,* Vol. 5, No. 3, pp. 115-132.

J. Makhoul (1975), "Linear Prediction: A Tutorial Review,". *Proc. of the IEEE*, Vol. 63, No. 4, pp. 561-580.

Ph. Martin (1981), "Détection de Fo pat intercorrélation avec une fonction peigne,". Journées d'études sur la Parole 12:221-232. SFA/GALF, Lannion, France.

Ph. Martin (1987), "A logarithmic spectral comb method for fundamental frequency detection,". Proc. 11th Int. Congr. Phonetic Sciences. Tallinn, USSR, paper59.2. Estonian Academy of Sciences, Tallinn, USSR.

B.C.J. Moore (1989). "An Introduction to the Psychology of Hearing,". Academic Press, London.

A.M Noll (1964), "Short-time spectrum and cepstrum techniques for vocal-pitch detection, ". J. Acoust. Soc. Amer. Vol.36, pp. 296-302, 1964.

A. V. Oppenheim (1968), "Speech Analysis-Synthesis Based on Homomorphic Filtering,". J. Acoust. Soc. Amer. Vol 45, pp 458-465.

A. V. Oppenheim and R. W. Schafer(1975), "Digital Signal Processing,". Prentice-Hall, Inc.

K. Ozeki, T. Umeda (1984), "An adaptative algorithm using an orthogonal projection to an affine subspace and its properties,". *Electronics and Communications in Japan*, Vol. 67-A, No. 5, pp. 19-27.

K. K. Paliwal et A. Basu (1987), "A Speech enhancement method based on kalman filtering,". IEEE ICASSP 1987, pp. 177-180.

J. Prado, E. Moulines (1993), "Frequency-domain adaptive filtering with application to echo cancellation,". *Proc. of the Third International Workshop on Acoustic Echo Control*, pp. 249-258.

L. R. Rabiner et al., "A comparative performance study of several pitch algorithms,". IEEE, Trans. Acoust., Speech, Signal Processing, Vol.24, pp.399-418, 1976.

L. R. Rabiner (1977), "On the use of autocorrelation analysis for pitch detection,". IEEE, Trans. Acoust., Speech, Signal Processing, Vol.25, No.1, 1977.

R. P. Ramachandran and Petet Kabal (1989), ″Pitch Prediction Filters in Speech Coding,″. IEEE Trans. on Acoust.,Speech, and Signal Processing, Vol. ASSP-37, N0.4.

S. Rouat, S. Lemieux and A. Migneault (1992), "A Spectro-temporel Analysis of Speech Based on Non Linear Operators,". Int. Conf on Spoken Language Processing, Banff, Octobre 12 to 16, Vol. 2, pp 1629-1632.

J. Rouat (1993), "Nonlinear operators for speech analysis", in "Visual Representations of Speech Analysis", pp. 335-340, edited by Cooke, S. Beet and M. Crawford, J. Wiley and Sons.

J. Rouat, Y.C. Liu and D. Morissette (1997). "A pitch detemination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, Vol. 20, mars 1997, pp 191-207.

F. Sadaoki and M. M. Sondhi (1991), "Advances in Speech Signal Processing,". Marcel Dekker, Inc.

M.R. Schroedar, ``Period histogram and product spectrum : new methods for fundamental frequency measurement`, J. Acoust Soc Amer, Vol 43, pp829-834,1968.

E. Terhard (1979). "Calculating virtual pitch,". Hear. Res 1:155-182.

J. P. Tubach (1989), "La parole et son traitement automatique,". Collection technique et scientifique des télecommunications, Masson. Paris.

J.-W Van Den Berg (1970), "Mechanism of the Larynx and the Laryngeal Vibrations,". In Manual of Phonetics, ed. by B. Malmberg, 278-308, 2end ed. North Holland Pub. Co. Amsterdam.

A.W.M Van Den Enden et N.A.M Verhoeckx (1992), "Traitement Numérique du Signal: Une Introduction,". Masson.

L. M. Van Immersel and J-P. Martens (1992), ″Pitch and Voiced/unvoiced determination with an auditory model,″. J. Acoust. Soc. Amer. Vol 96 (6), pp 3511-3526.

M. Weintraub, (1984), ″The GRASP sound separation system,″ Proc ICASSP 84, 18A.6.1-18A.6.4.

B. Widrow et al. (1975), "Adaptive noise cancelling: principles and applications,". *Proc. of the IEEE*, Vol. 63, No. 12, pp. 1692-1716.

J. Yang (1993). "Frequency domain noise suppression approches in mobile telephone systems,". IEEE ICASSP pp. 363-366.

H. Ye and B. X. WU (1991), "A new double-talk detection algorithm based on the ortogonality theorem,". EE trans. on Communications, vol. 39, pp. 1542-1545.

Les annexes A et B ne pouvaient pas être rendues publiques. Elles sont reliées séparément de ce mémoire afin de respecter l'entente prise entre l'Université du Québec à Chicoutimi (UQAC) et Alcatel Mobile Phones.

Toute personne souhaitant disposer des annexes pourra en faire la demande à l'UQAC ou à Alcatel Mobile Phones aux adresses suivantes :

M. Jean Rouat,
Université du Québec à Chicoutimi (D.S.A)
555, boulevard de l'université
Chicoutimi, Québec
Canada.

jrouat@uqac.uquebec.ca

M. Ivan Bourmeyster,
Service DT/SCI
32 Av. Kleber
92707 Colombes Cedex
France.

bourmeys@rtbf8.art.alcatel.ca

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

MÉMOIRE PRÉSENTÉ À

L'UNIVESRITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE DE

LA MAÎTRISE EN INGÉNIERIE

PAR

HASSAN EZZAIDI

DÉTECTION DE LA DOUBLE PAROLE DANS LE CONTEXTE DE RADIOTÉLÉPHONE MAIN-LIBRE EN VÉHICULE

PARTIE CONFIDENTIELLE

ANNEXE: A et B

DÉCEMBRE 97

ANNEXE A (Confidentielle)

MISE EN OEUVRE EN VIRGULE FIXE EN VUE D'UNE MISE EN OEUVRE EN TEMPS RÉEL.

1. INTRODUCTION

Les résultats obtenus dans le chapitre 5 avec les systèmes proposés, utilisent un détecteur d'activité vocale basé sur l'information de la fréquence glottale contrairement au système de référence qui utilise plutôt un détecteur basé sur l'énergie . Dans tous les tests donnés jusqu'à présent, l'estimation de la fréquence glottale a été effectuée par le système Algomai94 [Rouat ; 1997]. Le système Algomai94 est le fruit de plusieurs travaux et recherches dirigés par le groupe E.R.M.E.T.I.S à l'université du Québec à Chicoutimi. L'objectif visé par ce système était de mettre au point un produit robuste et fiable et indépendant du contexte (bruité ou propre) sans toutefois donner d'importance ou s'intéresser à l'implémentation en temps réel. L'objet de ce chapitre est de donner toutes les étapes importantes qui nous ont permis d'optimiser le système Algomai94 et de développer une version dérivée mise en oeuvre en virgule fixe pour la machine cible d'Alcatel. Une comparaison entre la performance d'autres algorithmes et la version de l'algorithme proposé sera donnée à la fin de ce chapitre.

2. COMPLEXITÉ DE L'ALGORITHME

2.1 Introduction

L'architecture du processeur de traitement du signal digital (DSP) chez Alcatel Mobile Phones supporte seulement l'implémentation des nombres en virgule fixe. Par conséquent, il a fallu transposer tout le code écrit en virgule flottante en un nouveau code en virgule fixe sans oublier qu'il doit également fonctionner en temps réel.

La représentation des nombres avec 16 bits, se caractérise par une performance en terme de résultats inférieurs à la représentation des nombres avec 32 bits ou celle à virgule flottante. Évidemment, plus le nombre de bits dont on dispose est élevé et plus la précision du calcul est meilleure. Par conséquent, on en déduit que les performances de la version en virgule fixe seront nettement inférieures par rapport à celle en virgule flottante.

Par la suite, si les résultats obtenus lors des simulations en virgule fixe sont jugés satisfaisants alors le code en virgule fixe pourra être traduit en langage assembleur propre à chaque DSP utilisé. Cette traduction de code permettra de relever d'autres défis liés aux contraintes imposées par le matériel du DSP.

Parmi ces contraintes, on ne doit pas dépasser l'espace mémoire disponible pour stocker le code. Il faut contrôler le débordement des registres

souvent rencontrés pendant le calcul et faire respecter le nombre d'instructions par seconde (nips) accordé par le processeur pour l'exécution du code. Finalement un bon code à virgule fixe doit éviter au maximum l'utilisation de la division car cette opération nécessite plus de cycles machine et peut souvent engendrer des erreurs dommageables (division par 0).

2.2 Analyse

Dans cette partie une comparaison de la complexité de l'algorithme d'estimation du fondamental en virgule flottante (Algomai94) et celle de l'algorithme qu'on propose pour fonctionner en virgule fixe sera donnée. La démarche consiste à analyser la complexité de chacun des modules du système Algomai94 et à présenter par la suite la version optimisée en virgule fixe.

Algomai94 se compose de plusieurs modules dont chacun réalise une tâche précise. On va attribuer aux différents modules les noms suivants:

. Module à banc de filtre cochléaires.

. Module de Teager.

. Module de filtrage passe bande.

. Module de corrélation.

. Module d'extraction du fondamental et de décision de voisement.

La fonctionnalité et la liaison entre chacun des modules sont illustrées par la figure 2 (chapitre 2).

2.2.1 Optimisation du module à banc de filtre cochléaires.

Version actuelle en virgule flottante:

La version à virgule flottante consiste à appliquer un banc de filtres composé de onze filtres passe bande à réponse impulsionnelle finie (RIF) et neuf à réponse impulsionnelle infinie (RII). L'ordre des filtres RIF est de 38 et celui des filtres RII est de 5 au numérateur (zéros) et 5 au dénominateur (pôles).

Version proposée en virgule fixe:

Pour la version optimisée, on a proposé l'utilisation de seulement 10 filtres RII d'ordre(1,4) ayant 1 zéro et 4 pôles repartis dans le domaine spectral de façon à avoir le minimum de chevauchement avec les bandes critiques où apparaît plus de bruit dû au contexte propre de la radio mobile. Par exemple, on a évité d'avoir des filtres concentrés au voisinage de 1000 Hz car le bruit du moteur du véhicule présente une forte énergie spectrale dans cette zone. Pour arriver à sélectionner les 10 meilleurs filtres on a mis au point 16 filtres en virgule fixe dérivés des filtres en virgule flottante. Les 11 premiers filtres optimisés correspondent exactement aux 11 premiers filtres FIR du système Algomai94. Les 5 derniers filtres correspondent aux filtres 14 à 18 en moyenne et haute fréquence de la version

flottante.

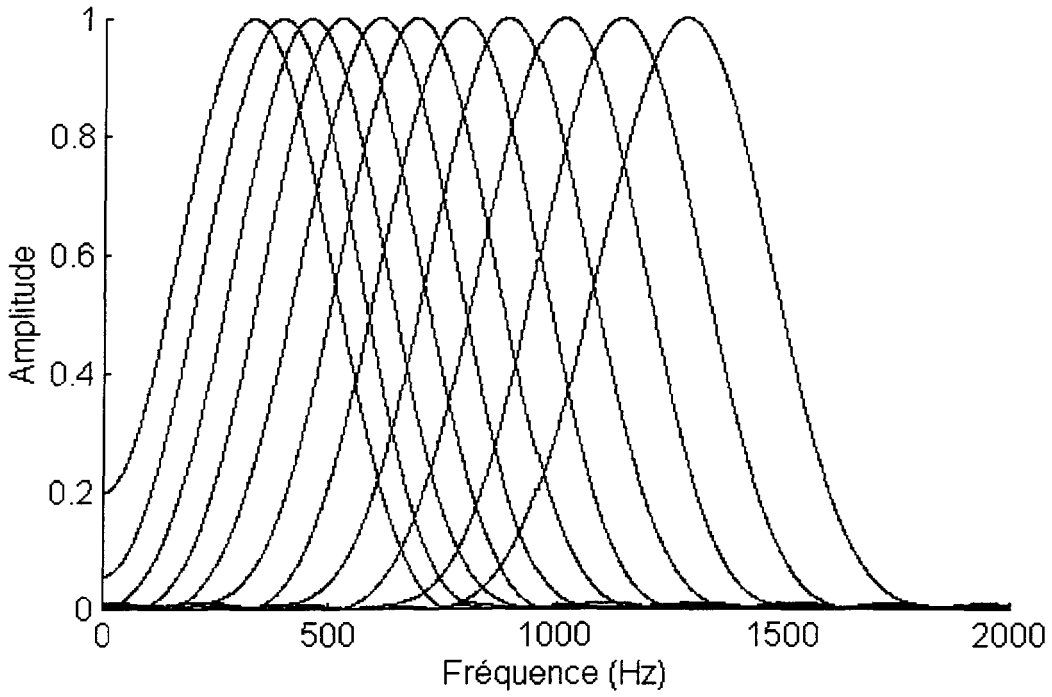


Figure 23:Réponse en amplitude du banc de filtres (Algomai94).

On livre dans l'annexe B de ce mémoire les coefficients des 10 filtres à virgule fixe qui ont été optimisés. Les 10 filtres sélectionnés sont composés d'une part de 7 filtres parmi les 10 filtres optimisés en virgule fixe et qui normalement correspondent au filtres FIR à virgule flottante. D'autre part les 3 derniers filtres sont choisis parmi les 5 filtres optimisés en moyenne et haute fréquence.

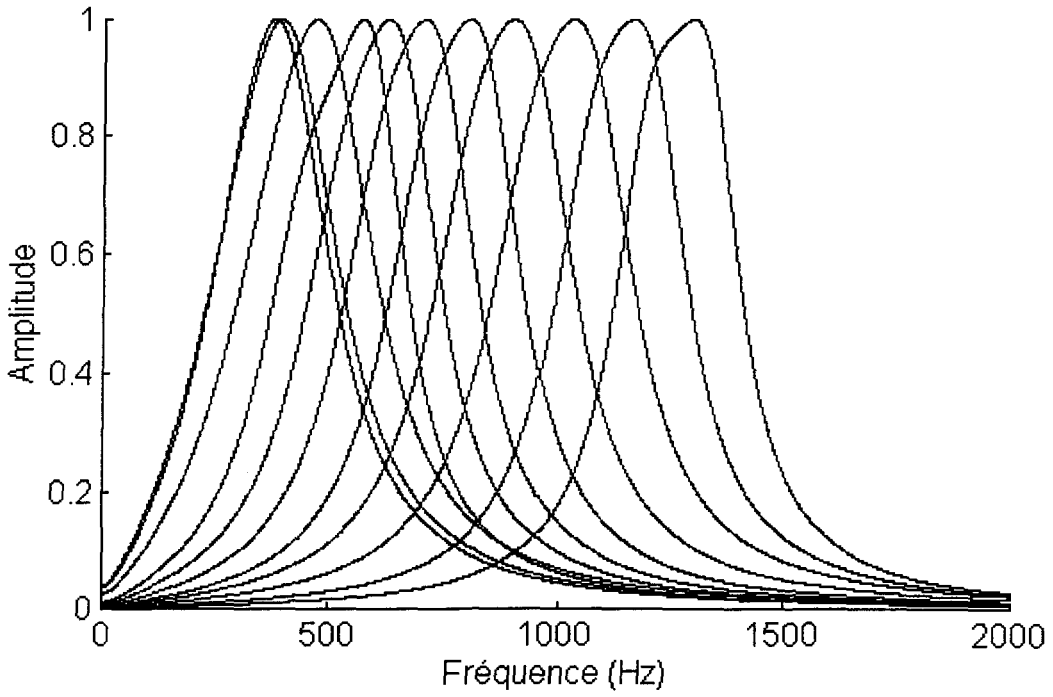


Figure 24:Réponse en amplitude du banc de filtres (version proposée).

Dans la figure 24, le deuxième filtre qui semble rendre la répartition non régulière n'est pas inclus dans la version en virgule fixe.

En général les propriétés (pente, fréquence de coupure..etc) des filtres optimisés correspondent en grande partie à ceux utilisés par AlgoMai94. Le graphique illustré à la figure 24 montre que la réponse en fréquence des filtres optimisés s'accorde bien avec les filtres RIF en version flottante à la figure 23. La figure 25 montre également la concordance de la réponse en fréquence de nos filtres (virgule fixe) en moyenne et haute fréquence avec ceux du système Ampex.

On note que la largeur de bande des filtres optimisés est légèrement inférieure à celle des filtres en version flottante et par conséquent la version proposée a tendance à être légèrement sélective.

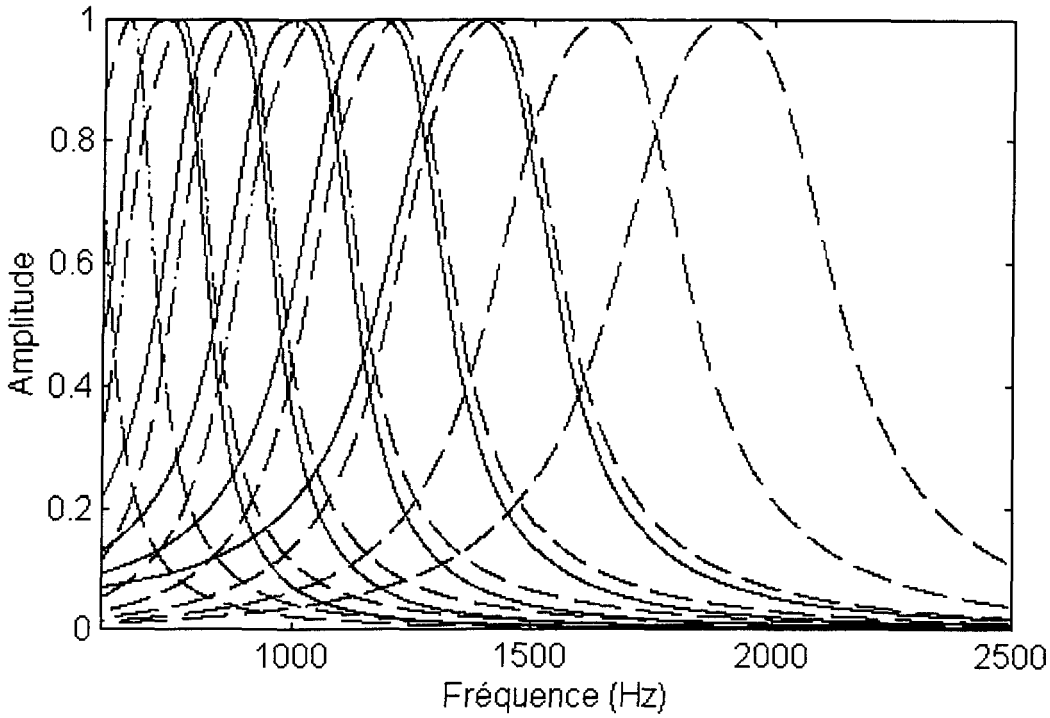


Figure 25:Réponse en amplitude du banc filtres en haute et moyenne fréquence.
version AMPEX en tiret: -----
version à virgule fixe en ligne pleine: ___

2.2.2 Optimisation du module de Teager

Version actuelle en virgule flottante:

Dans cette version, on applique l'opérateur de Teager pour extraire l'énergie du signal en moyenne et haute fréquence Cette opération est appliquée seulement pour les 3 filtres en moyenne et haute fréquence. L'opération de calcul nécessite 2 opérations de multiplication et une addition pour chaque échantillon. L'utilisation de Teager nécessite un suréchantillonage préalable de 8 KHz à 16 KHz. Une fois que Teager a été calculé on doit ensuite sous-échantillonner à 8 Khz.

Version proposée en virgule fixe:

La version optimisée utilise seulement 3 filtres en haute et moyenne fréquence auxquels l'opérateur Teager devrait être appliqué. On a limité à trois le nombre des filtres car l'implémentation de Teager s'est avérée être très sensible lors de l'implémentation en virgule fixe. Le problème provient du débordement rencontré pendant le calcul de Teager pour extraire la pseudo-enveloppe (voir la formule numérique dans la section 2.7.1). Or une bonne extraction de l'enveloppe du signal a l'avantage d'apporter à l'algorithme d'estimation de la hauteur tonale la capacité d'être robuste au bruit et de faire ressortir mieux le fondamental par rapport aux harmoniques. Le scénario inverse se produit si une mauvaise extraction de la pseudo-enveloppe est effectuée.

2.2.3 Optimisation du module de filtrage passe bande

Version proposée en virgule fixe:

On a utilisé un filtrage à 250 Hz de fréquence de coupure avec un ordre plus bas (Butterworth, ordre 3). Au lieu d'un filtrage passe haut, seule une préaccentuation a été utilisée pour atténuer les composantes en basse fréquence et rehausser les composantes en haute fréquence. Cette préaccentuation est effectuée directement sur le signal après avoir filtré le signal par le banc de filtres et le filtre passe-bas.

2.2.4 Optimisation du module de corrélation

La corrélation est la plus exigeante en nombre d'opérations et d'espace mémoire puisqu'on effectue un calcul en parallèle sur tous les canaux cochléaires. Une implémentation en temps réel sera quasiment impossible sans songer à optimiser le nombre d'opérations de la corrélation. Dans l'ouvrage de Blahut (1985), on trouve une liste des méthodes et d'algorithmes rapides dédiés aux traitements du signal digital en général. Une autre approche intéressante qui se base plutôt sur une méthode récursive a été proposée par [Allerhand et Patterson ; 1992] et qui d'après l'auteur prend moins de temps de calcul que les méthodes conventionnelles (FFT pour le calcul de la corrélation).

Dans notre cas, on a choisi pour la version proposée une technique utilisée par le système Ampex. Ce choix se trouve être justifié, d'une part par la simplicité et l'efficacité de la méthode et d'autres part en raison du fait que cette technique a été utilisée et testée avec Ampex afin d'extraire la hauteur tonale.

On va donner grossièrement le nombre d'opérations nécessaires pour calculer la corrélation pour la version actuelle en virgule flottante et celle de la version proposée en virgule fixe.

Version actuelle en virgule flottante:

AlgoMai94 utilise une fenêtre d'analyse de 16 ms (Fe=8KHz), décalée de façon à calculer l'autocorrélation normalisée pour un délai allant jusqu'à 16 ms. On a déjà donné les grandes lignes du système Algomai94 dans le chapitre 2.

Le calcul de la corrélation $R(\tau)$ pour chaque canal est établi par les formules suivantes où Ex est l'énergie du segment actuel et Ey est l'énergie du segment décalé de τ.

Pour $n = 1,..,N = 128$ et $\tau = 13,..,2N$

$$R(\tau) = \frac{1}{\sqrt{Ex}} * \frac{1}{\sqrt{Ey}} * \Phi(\tau)$$

avec:

$$Ex = \sum_{n=-N/2}^{N/2} x^2(n)$$

$$Ey(\tau) = Ex + \sum_{128}^{141} x^2(n) - \sum_{n=1}^{12} x^2(n) \quad \text{si } \tau = 13$$

$$= Ey(\tau - 1) + x^2(128) - x^2_{t-1}(1) \quad \text{si } \tau > 13$$

$$\phi(\tau) = \sum_{n=-N/2}^{N/2} x(n) * x(n+\tau)$$

$Ey(\tau-1)$ est l'énergie de la trame décalée à l'instant τ-1.

$x^2(n)$ est l'énergie de la trame courante à l'instant t.

$x^2_{t-1}(1)$ est l'énergie du premier échantillon de la fenêtre d'analyse. L'indice t-1 indique le temps de la fenêtre d'analyse et non pas de l'échantillon.

Ex est l'énergie de la fenêtre d'analyse actuelle (128 échantillons). Son calcul nécessite 128 multiplications et 127 additions . Ey est l'énergie de la fenêtre d'analyse de 16 ms décalée de τ par rapport à la fenêtre d'analyse actuelle. À chaque fois qu'on décale la fenêtre d'analyse chevauchante de τ à τ+1, on décale tous les échantillons à l'instant τ à gauche de cette fenêtre et on leurs substitue un nouvel échantillon. En effet, on exploite cette caractéristique pour simplifier le calcul de l'énergie Ey à l'instant τ +1 qui consiste à supprimer l'énergie du premier échantillon de la trame à l'instant τ et à la remplacer par l'énergie du nouvel

échantillon substitué à l'instant τ+1. Le calcul nécessaire de Ey revient à : 2 multiplications et 2 additions à chaque instant τ=14 ..256. A τ= 13 l'énergie Ey est déduite directement de Ex et nécessite 26 multiplications et 26 additions

Donc pour chaque canal et à chaque trame de 16 ms, la corrélation nécessitera approximativement les d'opérations suivantes :

$\frac{1}{\sqrt{Ex}}$ $\Rightarrow$ *128 multiplications + 127 additions + 1division + 1 racine carrée,*

$\frac{1}{\sqrt{Ey}}$ $\Rightarrow$ *(26+484) (multiplications+ additions) + 243 (divisions+ racine carrée),*

$\phi(\tau)$ $\Rightarrow$ (256-13) 128 = 15552 *(multiplications+ additions en considérant la symetrie).*

Le ``+`` signifie l'addition arithmétique.

Ce calcul est effectué en parallèle pour les 20 canaux ce qui rend le coût d'opération énorme. Ce même calcul est repris à chaque 10 ms.

Version proposée en virgule fixe:

Au lieu de traiter tous les échantillons du signal vocal après le filtrage passe-bande, Ampex procède à un sous échantillonnage de 20 kHz à 2,5 kHz. D'après l'auteur une telle résolution est suffisante pour calculer la corrélation avec assez de précision. De plus, un opérateur comparateur qui retourne le minimum des deux valeurs en argument (min(a,b)) a été utilisé à la place de l'opérateur

produit usuel ``*`` [Weintraub; 1984]. Ceci est très intéressant puisque la multiplication nécessite plus de cycles machine du processeur que ne l'exige l'opérateur de comparaison. Le fait de procéder par une décimation a permis de réduire le nombre d'opérations total par un quart (1/4) dans le cas d'Ampex.

Dans le cas de la version proposée à virgule fixe, on a modélisé le signal par un train d'impulsions positives séparées par une résolution temporelle minimale de 1 ms. Cette résolution temporelle minimale de 1 ms (Fe=8 KHz) nous permet d'aller chercher le fondamental jusqu'à 500 Hz.

La modélisation du signal par des impulsions consiste à sélectionner l'intervalle de temps où tous les échantillons ont des valeurs positifs, et à le représenter par une seule impulsion ayant une amplitude A à l'instant t. On va préciser, comment on sélectionne un intervalle dans notre contexte d'utilisation. Le début d'un intervalle est considéré à chaque fois que la valeur de l'échantillon du signal passe d'une valeur négative ou nulle vers une valeur positive. On peut représenter ceci par la formule suivante :

s(n-1)<= 0 et s(n) s(n-1) <0 alors on ouvre un nouvel intervalle.

s(n) est l'échantillon du signal s à l'instant n.

La fin d'un intervalle est considérée à chaque fois que la valeur de l'échantillon du signal passe des valeurs positives vers une valeur nulle ou négative. On peut représenter ceci également par la formule suivante :

s(n-1) > 0 et s(n) s(n-1) <=0 alors on ferme l'intervalle .

Pour qu'un intervalle soit sélectionné, il faut que la différence entre le premier et le dernier échantillon du même intervalle soit supérieure à 1 ms.

Ensuite, on cherche la position de l'échantillon ayant l'amplitude la plus élevée dans l'intervalle sélectionné pour l'attribuer à la position de l'impulsion qui va représenter cet intervalle. L'amplitude de l'impulsion correspond à l'aire de l'intervalle (la somme de tous les échantillons dans l'intervalle). Donc, chaque intervalle est représenté par une position dans le temps et une amplitude. Finalement, on a ajouté une contrainte qui consiste à ce que la distance entre les impulsions soit supérieure à 1 ms. Ceci résume la procédure qui modélise le signal par un train d'impulsions.

On avait mentionné au début de cette section qu' Ampex avait remplacé la multiplication dans le calcul de la corrélation par l'opérateur min(). Dans notre cas, on a suivi la même direction en comparant la performance des deux opérateurs min() et max(). D'après nos résultats expérimentaux, l'utilisation de l'opérateur min() a donné de meilleurs résultats que l'opérateur max(). Il nous semble que l'opérateur min() provoque moins de débordement pendant les calculs. D'un autre coté, la stratégie de représentation de l'amplitude des impulsions par l'aire a donné de meilleurs résultats que si on le représente par l'amplitude la plus élevée dans l'intervalle.

Dans la section suivante, on va analyser la complexité de la version proposée à virgule fixe. Puisque la durée temporelle minimale entre 2 impulsions est de 1ms, ceci nous permet de modéliser la durée fenêtre de 16 ms nécessaire pour évaluer la corrélation par 16 impulsions au lieu de 128 échantillons.

Pour éviter toute confusion, on pose Ap(i) l'amplitude de l'impulsion numéro i et Tp(i) le temps associé. L'indice du canal est représenté par p.

La corrélation optimisée a été calculée par la formule suivante :

$$Rp(Tp(\tau) - Tp(i)) += \sum_{\tau=1}^{N_\tau} \min(Ap(Tp(i)), Ap(Tp(\tau)))$$

p = {1 .. 10} indique le numéro du canal en cours de traitement ;

i = {1 .. Ni} indique le numéro d'impulsions de la fenêtre d'analyse actuelle de 16 ms, le maximum d'impulsion est : max(Ni)=16 ;

τ = {i .. Nτ} indique le numéro d'impulsion à l'instant τ de la fenêtre d'analyse décalée pour calculer la corrélation, le maximum d'impulsion est de 16 (max(Nτ)=16) ;

Min() est l'opérateur utilisé pour remplacer l'opérateur ``x`` produit ;

$Rp(Tp(\tau)-Tp(i))$ correspond à la contribution à la corrélation pour chaque période candidate $Tp(\tau)-Tp(i)$ à l'instant $Tp(\tau)$.

La contribution à la corrélation pour une impulsion nécessite : 16 additions + 16 comparaisons . Une fenêtre d'analyse de 16 ms (Fe=8KHz), décalée de façon à calculer l'autocorrélation pour un délai allant jusqu'à 16 ms, nécessitera 32 fois le calcul exigé pour une seul impulsion. Ce calcul est repris pour les 10 canaux à chaque 10 ms. La corrélation permet le calcul de la contribution glottique entre chacun des 2 pics pour ensuite l'accumuler à la distance séparant les 2 pics. A chaque distance $Tp(\tau)-Tp(i)$ correspond une période candidate. La période du signal doit correspondre à l'instant où la contribution de la corrélation est maximale. En opérant ainsi, nous évitons d'utiliser la technique d'Algomai94 qui recherche la période par la technique des plus petits sous-multiples communs .

Le pseudo-histogramme périodique noté PPH (différent de celui utilisé par Algomai94) est obtenu en réalisant la somme à travers les canaux des corrélations non normalisées qui sont représentées par des impulsions.

Ensuite, un lissage est appliqué à toutes les impulsions contribuant à la formation du PPH(n) ayant une fenêtre d'analyse de taille 32 ms. Le nombre d'opérations dans ce cas n'est pas critique puisque le traitement est ramené à un

vecteur à une seule dimension contrairement à 10 dimensions (qui correspondent aux nombre des canaux).

$$PPH(n)= 0.25\ PPH(n-1) + 0.5\ PPH(n) + 0.25\ PPH(n+1).$$

2.2.5 Optimisation du module d'extraction du fondamental et de décision de voisement.

Ce module a été implémenté de manière différente à celui d'Algomai94. En effet, on a rencontré souvent des problèmes de débordements pendant le calcul en utilisant des registres de 16 bits (virgule fixe), particulièrement lors de l'extraction de l'enveloppe par l'opérateur de Teager. Par conséquent, il était difficile de trouver facilement des seuils dans la version proposée pour distinguer entre le cas des segments voisés et les segments non-voisés. Par contre, on a utilisé temporairement un seuil estimé à partir des 6 premières trames qui consiste à chercher le pic le plus élevé à partir du pseudo-histogramme périodique. Ensuite, on a normalisé la valeur du pic par l'énergie du canal le plus élevé parmi les 10 canaux utilisés. Cette valeur normalisée est considérée pour toute la suite de traitement comme un seuil de voisement fixe et qu'on note : δv, pour distinguer entre les segments voisés ou d'autres segments (silences et non-voisés). L'énergie du canal qu'on utilise, correspond simplement à la somme des amplitudes des impulsions modélisant la fenêtre d'analyse initiale. Il est évident

que si on procède par détermination de seuils de façon adaptative, les performances de l'algorithme seront meilleures.

Après avoir défini le seuil de voisement, on a testé une première technique qui consiste à extraire les 2 valeurs les plus élevées du pseudo-histogramme périodique PPH(t) à l'instant t1 et t2. Ensuite, le plus grand pic est normalisé par l'énergie du canal le plus élevé parmi les10 canaux. Et, si cette valeur normalisée est inférieure ou égale au seuil de voisement δv (estimé à partir des 6 premières trames), alors le segment est jugé non périodique. Sinon, le segment est déclaré voisé de période t1 dans la mesure où la fréquence 1/t1 est comprise dans l'intervalle [90 Hz .. 500 Hz]. Si, t1 n'a pas été sélectionné comme période alors on reprend les même tests avec t2. Dans tous les autres cas le segment est déclaré non périodique.

Cette première technique simple, a connu quelques problèmes pour extraire avec précision le fondamental et pour détecter correctement les segments voisés. Les sources majeurs de ces problèmes sont dues à la mauvaise estimation des seuils fixes, aux problèmes de confusion causés par les formants et les harmoniques et à la façon dont on calcule l'énergie pendant la normalisation.

La confusion de la valeur du fondamental estimé avec celle des harmoniques, se présente dans la situation où le signal acoustique se caractérise

par une répartition périodique des pics secondaires (liés au contexte de parole) plus importante que celles des pics principaux du fondamental (liés à l'excitation de la source). Et par conséquent, la modélisation par train d'impulsions sera représentée davantage par les pics secondaires qui vont masquer la contribution du fondamental.

La deuxième technique découle de la première afin de corriger les erreurs de la première technique et restreindre les problèmes de confusions engendrés par les formants et les harmoniques et le problème de seuil de voisement. Elle consiste à ajouter un traitement supplémentaire qui analyse la continuité du fondamental sur 4 trames Cette analyse permet de rehausser l'évidence des segments voisés par rapport aux non-voisés et également de mieux ressortir le fondamental par rapport à ses harmoniques. On va décrire maintenant en quoi consiste l'analyse de continuité du fondamental. Premièrement, on recherche les pics dans le PPH calculé de la trame actuelle qui dépassent un certain pourcentage du pic le plus élevé. On a utilisé un seuil de 40% pour sélectionner ou rejeter les pics du PPH. Ensuite, pour chaque pic, on évalue son évidence en fonction des pics déjà sélectionnés des 3 dernières trames (sauvegardées en mémoire) par le pseudo-code suivant :

```
Pour i = 1:n
   Pour k = 1:m
     Pour j = 1:q
       Si (T(i) - T_k(j)) < δ_T(T(i) + T_k(j))
          alors
             evid(T(i)) = PPH(T_k(j) + evid(T(i));
```

i ={1 .. n} indique le numéro d'impulsion de la trame actuelle; Max(n)=16 ;
k = {1 .. m} indique le numéro de la trame précédente; m=4 ;
j = {1 .. q} indique le numéro d'impulsion de la trame k ; Max(q)=16 ;
T(n) donne le temps de l'impulsion numéro n de la trame actuelle ;
$T_k(n)$ donne le temps de l'impulsion numéro n de la trame précédente k ;
δ_T est un seuil fixé à 0.045.

Pour extraire le fondamental, on continue avec le même traitement décrit dans la première technique. La seule différence réside dans le fait qu'on sélectionne cette fois-ci, les 2 pics ayant les évidences les plus élevées et non ceux ayant les valeurs du PPH les plus élevées.

Il nous semble que l'utilisation de l'évidence a permis d'ajouter à l'algorithme plus d'efficacité lui permettant de distinguer davantage entre les segments non-voisés (ou silences) et les segments voisés par comparaison à la première technique. Ceci permet de rendre l'algorithme moins sensible aux erreurs introduites lors du calcul du seuil fixe de voisement. La figure 27 montre à titre d'exemple la courbe d'évidence et celle du PPH qu'on a obtenu comme résultats pour le même signal vocal. D'après la figure, la différence moyenne de

l'évidence entre les segments voisés et non-voisés est plus importante que celle du PPH. Il suffit de comparer les maximums et les minimums des deux courbes.

Un filtre médiane à trois points à été utilisé pour corriger les erreurs fines rencontrées lors de l'extraction du fondamental.

3 Critères d'évaluations et résultats de l'algorithme proposé

3.1 Base de données

On a utilisé la même banque de mots et de phrases employées par Rouat (1997) pour comparer les performances d'Algomai94 à celles d'Ampex (Van Immerseel, 1992). On a considéré seulement de la parole téléphonique enregistrée en véhicule à la vitesse de 0 km/h (parole propre) et 60 km/h (parole bruitée).

3.2 Critère d'évaluation

Les formules utilisées pour évaluer les résultats sont les mêmes que celles décrites dans l'article de Rouat (1997).

La déviation pour chaque fondamental estimé Fe est calculée par :

$$\Delta f = \frac{|Fe - Fo| * 100\%}{Fo}$$

Fo est le fondamental de référence estimé manuellement.

La moyenne de l'erreur fine est définie par :

$$\text{Aver}(\Delta f) = \frac{\text{Somme des } \Delta f \text{ sur le nbre des segments déclarés voisés avec } \Delta f \leq 20\%}{\text{nbre des segments déclarés voisés avec } \Delta f \leq 20\%}$$

Soit N le nombre total des trames de tous les fichiers de parole constituant la base de données utilisée :

$$N = Nv/uv + Nuv/v + Ncor + Ngros$$

avec :

Nv/uv = nombre des voisés classés comme non-voisés ;

Nuv/v = nombre des non-voisés classés comme voisés ;

Ncor = nombre de trames correctement classés avec $\Delta f \leq 20\%$;

Ngros= nombre de trames correctement classés avec $\Delta f > 20\%$;

Les erreurs de mesures sont définies par les formules suivantes :

$$\varepsilon_{v/uv} = \frac{Nv/uv}{N}$$

$$\varepsilon_{uv/v} = \frac{Nuv/v}{N}$$

$$\varepsilon_{gros} = \frac{Ngros}{N}$$

3.3 Résultats

	$\varepsilon_{v/uv}$	$\varepsilon_{uv/v}$	ε_{gros}	Aver(ΔF)
Ampex δ_v=1.6 ;δ_T=0.05	3.6	3.64	0.75	1.91
AlgoMai94 **S**=0.45	4.35	4.60	0.87	1.61
Version Proposée δ_v=8	2.42	3.2	6.6	4.88

Table 4 : Comparaison des performances de l'algorithme proposé à 0 km/h en se basant sur les valeurs du PPH.

	$\varepsilon_{v/uv}$	$\varepsilon_{uv/v}$	ε_{gros}	Aver(ΔF)
Ampex δ_v=0.4 ;δ_T=0.35	5.55	5.58	7.89	3.08
AlgoMai94 **S**=0.3	4.87	4.1	1.89	1.04
Version Proposée δ_T=0.045	5.3	17	11	3.97

Table 5 : Comparaison des performances de l'algorithme proposé à 60 km/h en se basant sur les valeurs de l'évidence.

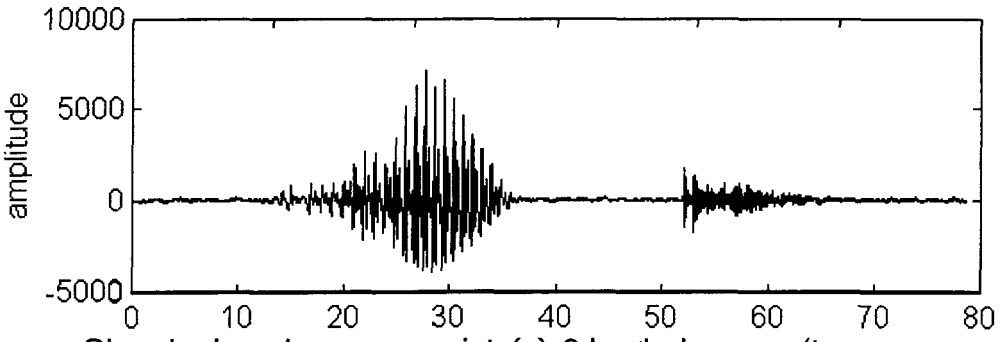
Dans cette expérience, on a utilisé uniquement les 10 filtres optimisés en basse fréquence qui correspondent aux filtres FIR du système Algomai94. On n'a

pas employé les 3 filtres auxquels on devait appliquer l'opérateur Teager comme on l'avait proposé pour la version finale au début de ce chapitre. La raison est que l'opérateur Teager introduit souvent un débordement pour extraire la modulation en amplitude. Provisoirement, on a proposé une solution pour remédier à ce problème en décalant à gauche d'un certain nombre de bits responsables de ce débordement. Ensuite, on s'est rendu compte que l'extraction de l'enveloppe par Teager introduit parfois de brusques transitions qui réduisent la performance de notre algorithme.

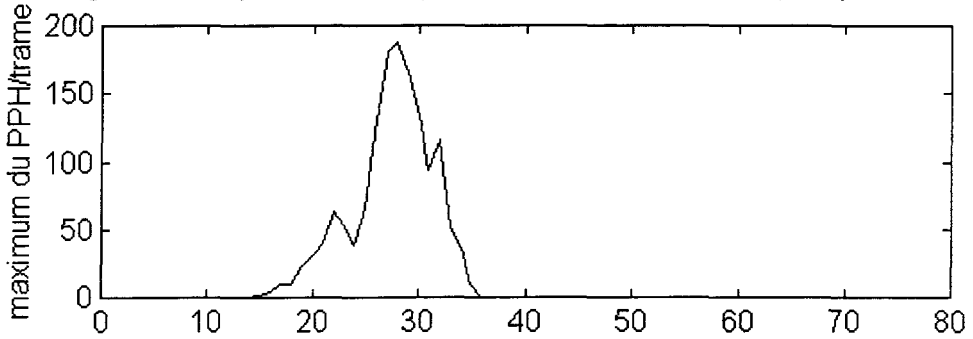
Pour les expériences à 0 km/h (parole propre) (voir figure 26), on a utilisé la première technique d'extraction du fondamental basée sur le pic le plus élevé du PPH et dépassant un certain seuil évalué manuellement. Dans ce cas, le seuil de voisement δ_v a été fixé à 8 (voir Table 4). Pour les expériences à 60 km/h (parole bruitée, voir figure 27), on a utilisé la deuxième technique d'extraction du fondamental basée sur le fondamental ayant l'évidence la plus élevée et dépassant le seuil δ_T=0.0045 . Le seuil de voisement δ_va été déduit à partir des 4 premières trames.

On peut dire d'après les résultats des Tables 4 et 5 que la performance de l'algorithme proposé est satisfaisante en la comparant au système Algomai94 et le système Ampex. On trouve aussi que notre système permet d'extraire le fondamental avec assez de précision dans le cas de la parole bruitée ou propre

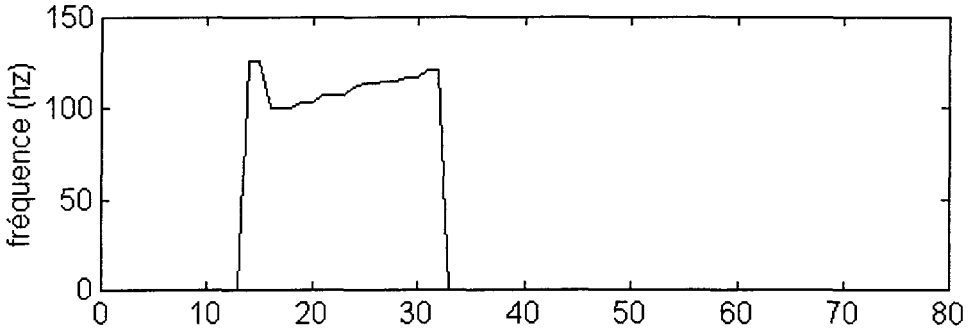
tout en fonctionnant en temps réel. Par contre, on enregistre un taux d'erreurs de 17% de classer les segments non-voisés comme étant des segments voisés. L'origine de cette erreur grossière est due à la mauvaise estimation du seuil de voisement et peut être du fait qu'on a pas utilisé les filtres en moyenne et haute fréquence.



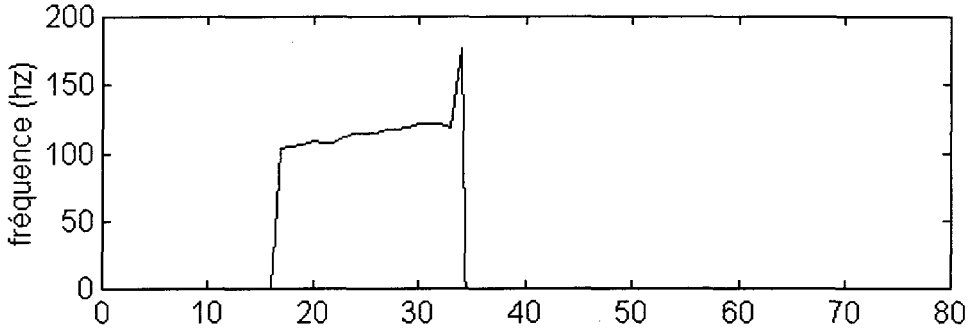
Signal microphone enregistré à 0 km/h, homme (temps en ms).



La valeur de la corrélation maximale à chaque 10 ms après avoir effectuée une sommation de la corrélation sur les 10 canaux (temps en cs).



Le fondamental estimé par la version à virgule fixe(temps en cs).



Le fondamental estimé par la version à virgule flottante (algomai94) (temps en cs).

Figure 26 : Extraction du fondamental en utilisant un seuil de voisement fixe

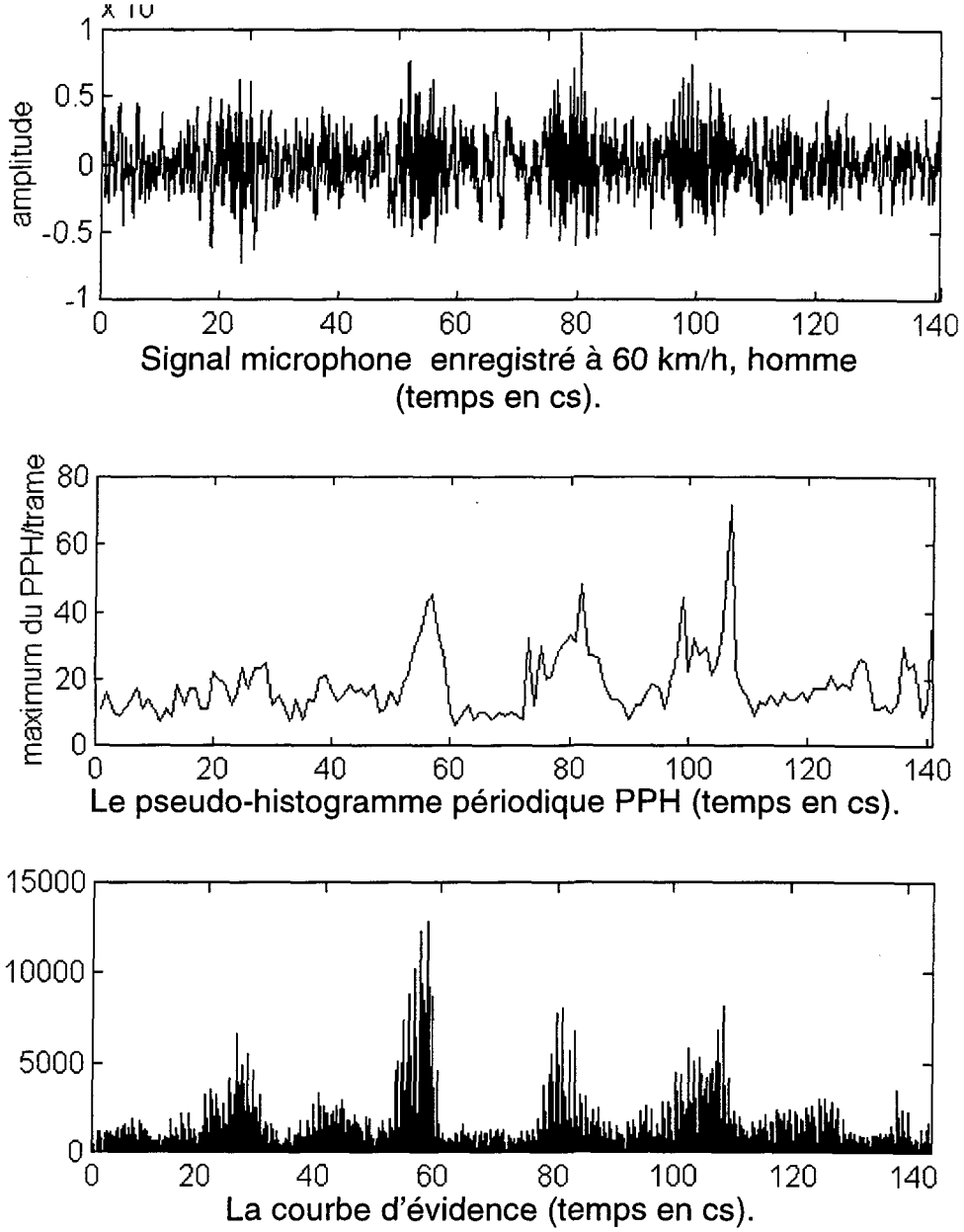
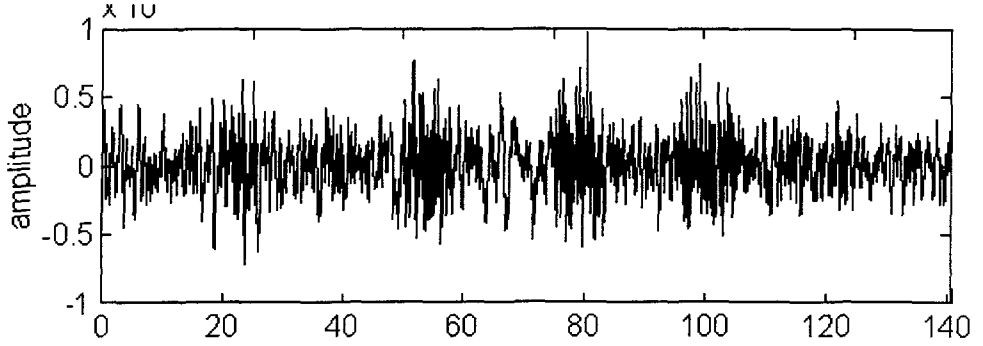
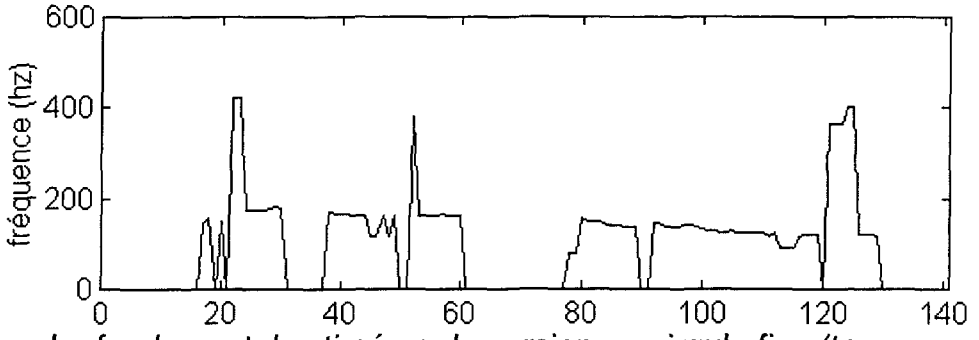


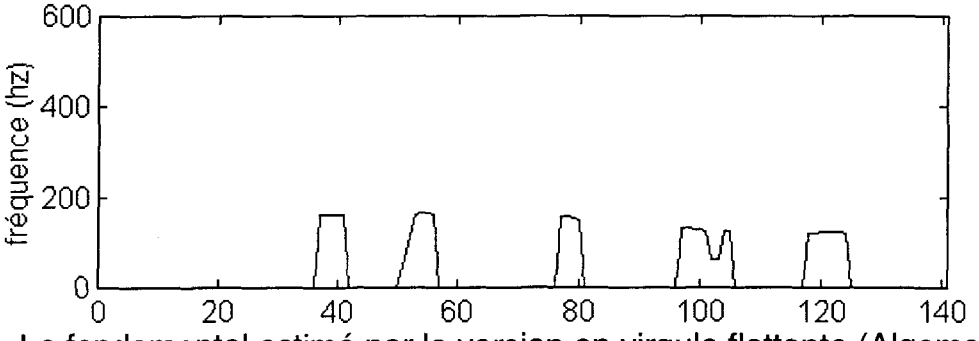
Figure 27 : Extraction du fondamental avec la technique estimant le seuil de voisement en se basant sur le calcul d'évidences (voir Annexe A.2.2.5) et la continuité du fondamental à partir de 4 trames.



Signal microphone enregistré à 60 km/h, homme (temps en cs).



Le fondamental estimé par la version en virgule fixe (temps en cs).



Le fondamental estimé par la version en virgule flottante (Algomai94)

Suite de la Figure 27

ANNEXE B

Les coefficients Ai et Bi des filtres cochléaires optimisés sont utilisés sous la forme suivante :

a(1)*y(n) = b(1)*x(n) + b(2)*x(n-1) + ... + b(nb+1)*x(n-nb)
-a(2)*y(n-1) - ... - a(na+1)*y(n-na)

avec Ai= a(1) a(2) a(3) ...
Bi= b(1) b(2) b(3) ...
Où ``*`` correspond à l'opérateur multiplication.

VERSION 01:

Les filtres en basse fréquence sont représentés par les coefficients Ai et Bi ayant l'indice de 0 à 10. Ils correspondent en principe aux filtres (FIR) en basse fréquence du système Algomai94. Les filtres en moyenne et haute fréquence sont représentés par les coefficients supérieur à 10 et qui correspondent aux filtres du système Ampex.

```
B0=[ 0.01268185909378 -0.01242822191191];
A0=[ 1.0 -3.42697862899177 4.53213267949739 -2.73457352826607 0.63633566014225];

B1=[ 0.01430646172030 -0.01402033248589];
A1=[1.0 -3.40705472211283 4.48408989200251 -2.69456297700905 0.62475429139600];

B2=[ 0.01681231212068 -0.01647606587827];
A2=[ 1.0 -3.35446458107554 4.40683312371937 -2.67305833023146 0.63401406250000];

B3=[ 0.01295960007593 -0.01270040807441];
A3=[ 1.0 -3.40642847212854 4.62097288870212 -2.93223779416141 0.739772010];

B4=[ 0.01574405411664 -0.01542917303431];
A4=[ 1.0 -3.28458028246369 4.38678997034275 -2.77793179773396 0.71425739390625];

B5=[ 0.01725279920970 -0.01690774322551];
A5=[ 1.0 -3.17907744658713 4.22234015180290 -2.69971804469348 0.72016439062500];

B6=[ 0.01904244764343 -0.01866159869057];
A6=[ 1.0 -3.02848255096123 3.99218384679503 -2.57890997665081 0.7241159025000];

B7=[ 0.02353540759744 -0.02306469944549];
A7=[ 1.0 -2.82292575408601 3.66334846923940 -2.36577101892035 0.70124964624225];
```

B8=[0.02628734790642 -0.02576160094829];
A8=[1.0 -2.56998079793249 3.31800904689192 -2.15030940584304 0.69897960250000];

B9=[0.02361341813881 -0.02314114977603];
A9=[1.0 -2.32253478546435 3.05708829116106 -1.99381792128815 0.73603530562500];

B10=[0.02079951328244 -0.02038352301679];
A10=[1.0 -2.05424811801233 2.80365765981277 -1.80640117804457 0.77246521];

B13=[0.0053 -0.0100];
A13=[1.0 -3.3912 4.7185 -3.1274 0.8503];

B14=[0.0063 -0.0088];
A14=[1.0 -3.2051 4.3760 -2.9003 0.8185];

B15=[0.0071 -0.0058];
A15=[1.0 -2.9621 3.9639 -2.6274 0.7861];

B16=[0.0080 -0.0024];
A16=[1.0 -2.6460 3.4818 -2.2990 0.7540];

B17=[0.0097 -0.0001];
A17=[1.0 -2.2384 2.9381 -1.8984 0.7179];

VERSION 02:

Dans cette section, on donne une autre version améliorée des coefficients des filtres cochléaires optimisés. Les filtres en basse fréquence sont représentés par les coefficients Ai et Bi ayant l'indice de 1 à 7. Les filtres en moyenne et haute fréquence sont représentés par les coefficients supérieur à 7.

B1 =[0.0125 -0.0115];
A1 =[1.0 -3.4388 4.5402 -2.7238 0.6274];

B2 =[0.0179 -0.0162];
A2 =[1.0 -3.3380 4.3689 -2.6441 0.6274];

B3 =[0.0221 -0.0204];
A3 =[1.0 -3.2568 4.2341 -2.5797 0.6274];

B4 =[0.0269 -0.0246];
A4 =[1.0 -3.1636 4.0832 -2.5059 0.6274];

B5 =[0.0289 -0.0265];
A5 =[1.0 -3.0532 3.9112 -2.4184 0.6274];

B6 =[0.0197 -0.0195];
A6 =[1.0 -2.8710 3.7846 -2.4831 0.7481];

```
B7 =[ 0.0228   -0.0226];
A7 =[1.0 -2.6354    3.4578   -2.2793    0.7481];

B8 =[  0.0252   -0.0003];
A8 =[1.0 -2.2384    2.9381   -1.8984    0.7179];

B9=[0.04638260869565  0.00422492753623  -0.04131864734300];
A9=[1.0          -0.2148          1.46825632     -0.16635168     0.5449648];

B10=[ 0.04744509480627        0.00187646331410     -0.03676187139324];
A10=[ 1.0        1.0791          1.70980708           0.7683192            0.506944];
```

Les coefficients optimisés des filtres passe-bas et passe-haut :

```
A_bas=[  0.0084 0.0169 0.0084];
B_bas=[  -1.7238 0.7575];

A_haut=[  0.9565  -1.9131  0.9565];
A_haut=[ -1.9112   0.915];
```