UNIVERSITÉ DU QUÉBEC

MÉMOIRE

PRÉSENTÉ À

L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN RESSOURCES ET SYSTÈMES

PAR

YONG CHUN LIU

UN DÉTECTEUR PERCEPTIF DE LA HAUTEUR TONALE POUR LA PAROLE TÉLÉPHONIQUE

AVRIL 1992

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.

SOMMAIRE

La détection de la hauteur tonale dans un signal de parole est un problème complexe et important pour des applications en reconnaissance de parole continue. Lorsqu'on souhaite détecter la hauteur tonale sur la parole téléphonique, la difficulté est plus grande, puisque la fréquence fondamentale n'est pas claire dans le signal. Nous proposons un modèle pratique basé sur des connaissances psychoacoustiques et physiologiques de l'oreille. En effet, cette dernière est capable d'extraire la hauteur tonale du signal de parole téléphonique. Le modèle proposé comprend trois éléments: un banc de filtres auditifs qui simule les mouvements mécaniques de la membrane basilaire; un modèle fonctionnel qui calcule des pseudo-histogrammes périodiques reliés à la période de la fréquence fondamentale; l'élément final combine la sortie des histogrammes pour extraire la hauteur tonale. Ce modèle est testé sur des données de parole numérisées à travers le réseau téléphonique de la région de Montréal. Les résultats des expériences indiquent que cette approche permet d'obtenir la hauteur tonale même si l'énergie de la composante fondamentale du signal de parole est très faible.

REMERCIEMENTS

Je tiens avant tout à remercier Dr. Jean Rouat, professeur au département des sciences appliquées de l'université du Québec à Chicoutimi, qui premièrement m'a introduit au domaine du traitement de la parole. Je le remercie pour avoir dirigé ma recherche et m'avoir prodigué de précieux conseils. Je le remercie particulièrement pour avoir corrigé mon mémoire et pour m'avoir permis d'améliorer mes capacités à communiquer en français.

Que tous les membres d'ERMETIS (Equipe de Recherche en Microélectronique et Traitement Informatique des Signaux) soient remerciés pour leur soutien technique. Je suis reconnaissant au Dr. Yongke Wu du GRIPS (Groupe de Recherche en Ingénierie des Procédés et Systèmes) pour m'avoir initié à l'utilisation du logiciel "Publisher" pour composer ce mémoire. Finalement, j'aimerais remercier ma femme Keli Zhang qui a dactylographié le manuscrit.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

Ce mémoire s'organise autour d'un sujet concernant l'analyse de la parole: l'extraction des paramètres. Nous nous intéressons à l'analyse de la parole téléphonique et plus particulièrement, l'extraction de la hauteur tonale. En fait, ce dernier paramètre est très important pour la reconnaissance de la parole continue. En outre, il est très difficile de l'extraire sur la parole téléphonique, car la composante fondamentale n'est pas toujours présente dans ce type de signal, en raison de la bande étroite du téléphone (de 300 Hz à 3400 Hz).

1.1 But de ce mémoire

En reconnaissance de parole, l'analyse qui consiste à extraire les paramètres est cruciale et difficile. Les performances de l'algorithme de reconnaissance reposent en grande partie sur la qualité de l'analyse. Notamment, pour la reconnaissance de parole continue, la hauteur tonale est un paramètre important, car l'information prosodique est prédominée par ce dernier.

Le but de ce mémoire est de proposer un algorithme capable d'extraire la hauteur tonale de la parole téléphonique en utilisant un modèle auditif.

1.2 Stratégie de ce mémoire

L'originalité du travail réside dans l'alliance d'outils de traitement des signaux et de connaissance sur la psychoacoustique et la physiologie de l'oreille. L'alliance de théories permet de réaliser un extracteur de la hauteur tonale reflétant certaines propriétés perceptives humaines et manifestant une robustesse aux bruits. Notre approche consiste à réaliser un modèle auditif qui se compose d'un banc de filtres et d'un modèle fonctionnel sans avoir à modéliser de façon exacte la transduction mécanique-électrique effectuée dans l'oreille interne. L'information à la sortie de chaque canal est ensuite combinée avec les autres sorties des canaux pour en déduire la hauteur tonale.

1.3 Organisation de ce mémoire

Dans le chapitre 2, nous essayons d'introduire certains concepts nécessaires pour comprendre le mémoire, ensuite nous faisons la revue des algorithmes existants sur l'extraction de la hauteur tonale et de la fréquence fondamentale. Dans le chapitre 3, nous décrivons le modèle proposé de façon détaillée. Nous présentons les expériences effectuées et les résultats obtenus à partir du modèle proposé dans le chapitre 4. Enfin, dans le chapitre 5, nous donnons la conclusion et discutons nos résultats ainsi que nos perspectives.

CHAPITRE 2

EXTRACTION DE LA HAUTEUR TONALE

2.1 Introduction

Ce chapitre est composé de deux parties. La première partie porte sur une courte introduction au système auditif et à la définition de la hauteur tonale. La description du système auditif est d'abord présentée. Les définitions de la hauteur tonale et de la fréquence fondamentale du signal de parole sont ensuite introduites à titre comparatif. La deuxième partie, occupant la majeure partie de ce chapitre, est consacrée à la revue des algorithmes d'extraction de la fréquence fondamentale et de la hauteur tonale de la parole.

2.2 Système auditif et hauteur tonale

Dans ce paragraphe, on introduit brièvement dans un premier temps le système auditif. On décrit simplement la succession des transformations, que subit l'information acoustique, lors de son cheminement au travers du système auditif périphérique et ce jusqu'à son arrivée dans le système nerveux central. Ensuite, on définit la fréquence fondamentale et la hauteur tonale.

2.2.1 Système auditif

La figure 2.1 représente un schéma très simplifié du système auditif. On peut trouver

une description précise pour chaque partie du système auditif dans la littérature éditée par Aran et al. [3]. En général, le système auditif est un système particulièrement complexe, que l'on décompose habituellement en deux parties: le système auditif périphérique (oreille) et le système nerveux. Le système auditif périphérique est constitué par l'oreille externe (pavillon et conduit auditif), l'oreille moyenne (tympan et osselets) et l'oreille interne (cochlée). Le système nerveux est responsable du traitement de l'information acoustique venant du système auditif périphérique et achemine les signaux de retour à l'oreille interne. Contrairement à la partie périphérique, la partie nerveuse est nettement moins bien connue.

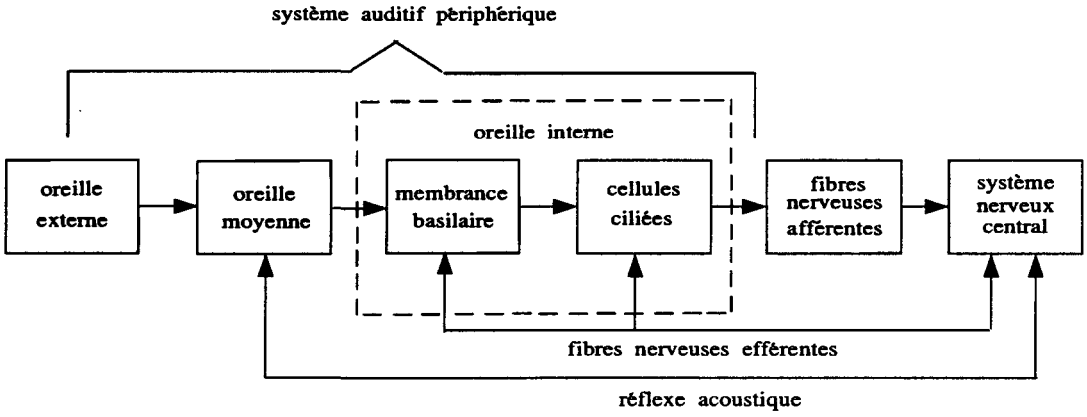

Figure 2.1 Schéma simplifié du système auditif d'après Dolmazon [3]

La vibration acoustique recueillie par l'oreille externe subit diverses transformations avant d'être acheminée sous forme d'influx nerveux vers le système nerveux central. Les vibrations d'onde acoustique mettent en mouvement le tympan. Ce dernier transforme les vibrations aériennes en vibrations osseuses. Ces vibrations solides sont transformées en ondes de pression dans les liquides de la cochlée. De plus, les changements d'onde de presssion en phase avec la pression acoustique, sont transformées en déplacements de la membrane basilaire qui accomplit une première analyse fréquentielle du son d'entrée.

Sur la membrane basilaire, on peut trouver l'organe de corti: un ensemble comprenant des cellules ciliées qui sont sensibles aux déplacements relatifs des membranes. Ces cellules ciliées (récepteurs sensoriels) convertissent les mouvements de la membrane basilaire en information électrochimique déclenchant ainsi les influx nerveux dans les fibres du nerf auditif. Ces influx sont ensuite transmis au système nerveux central par les fibres nerveuses afférentes. Notons que l'information acoustique codée est réellement interprétée lors de son arrivée dans le système nerveux central. L'information de retour ("feedback") est transmise par les fibres efférentes à l'oreille interne, présentement les fonctions des fibres efférentes sont moins connues. Il est à noter que les fibres nerveuses afférentes et efférentes constituent les fibres nerveuses auditives.

2.2.2 Fréquence fondamentale

Les sons de parole dits voisés, exemple les voyelles et les consonnes telles que */b/ /d/ /g/*, sont produits avec une vibration des cordes vocales (ou vibration laryngienne). La fréquence fondamentale (au sens de l'analyse de Fourier) du signal de parole apparaît comme la fréquence de vibration laryngienne. En d'autres termes, l'intervalle de temps entre des impulsions glottiques adjacentes correspond à la période de la fréquence fondamentale du signal de parole. La mesure de la fréquence fondamentale peut se faire à partir du signal de parole dans le domaine temporel, ou dans le domaine spectral: à partir de la fréquence fondamentale du spectre d'un son voisé. Nous allons décrire ces méthodes de mesure plus loin.

2.2.3 Hauteur tonale

La hauteur tonale (HT) d'un son n'est conceptuellement pas facile à définir d'une façon explicite et générale. Comme l'indique les travaux de Demany dans ce qui suit:

"A qui n'est pas familier de l'immense littérature consacrée à la perception de la HT et à ses mécanismes, il peut certes sembler que ce à quoi l'on se réfère en parlant de HT est une qualité sensorielle très simple. Mais la réalité est différente, comme divers auteurs l'ont déjà souligné" [5]. D'après Demany, des problèmes de définition de HT se posent dans les deux cas: les sons purs et les sons complexes.

La HT d'un son pur est l'attribut perceptif à partir duquel il est possible de lui apparier un autre son pur, en ajustant la fréquence de ce dernier. Les deux sons ne diffèrent alors que par le niveau d'intensité à condition que la différence d'intensité de ces deux sons purs ne soit pas très grande. Demany affirme qu'après appariement de leurs HT, les deux sons purs ne diffèrent que par la sonie. Ce qui implique que deux sons purs quelconques ne peuvent différer que par la HT et/ou la sonie. D'après les expériences de Girija [18], la HT d'un son pur ou d'un signal sinusoïdal de fréquence f, d'intensité normale correspond à la fréquence f.

Parmi tous les sons pour lesquels le système auditif humain est capable d'extraire une sensation de HT, les plus importants sont complexes et périodiques ou quasi-périodiques [5]. Tel est le cas de la parole voisée. Selon la théorème de Fourier, tout signal complexe périodique de période $1/f$ peut être caractérisé par une somme de signaux purs dont les fréquences sont différentes. La fréquence de chaque composante (signal pur) peut être représentée par $n \cdot f$, où n est un nombre entier et f est la fréquence fondamentale du son complexe périodique. Lorsqu'on veux apparier la HT d'un son pur de fréquence ajustable, à celle d'un son complexe, on trouve que la fréquence du son pur est ajustée à une valeur très proche de f. En d'autres termes, la HT d'un son complexe correspond généralement à sa fréquence fondamentale. Donc, jusqu'à un

certain degré, l'extraction de la fréquence fondamentale est équivalente à l'extraction de HT de la parole voisée. Notons que, plus précisement, la fréquence qui correspond à la HT et la fréquence fondamentale ne sont généralement pas identiques. Terhardt donne un bon exemple [64] pour lequel la fréquence fondamentale d'un son complexe composé de trois sinusoïdes de fréquence 520, 620 et 720 Hz respectivement, est 20 Hz, mais la fréquence correspondant à la HT perçue est autour de 104 Hz.

Il faut mentionner que l'oreille est capable de percevoir la HT d'un son complexe dont la composante fondamentale ne comporte aucune énergie. C'est souvent le cas en parole, notamment lorsqu'elle est transmise par une ligne téléphonique. C'est d'ailleurs ce type de signal qui va être traité dans ce travail.

2.3 Revue des algorithmes d'extraction de la hauteur tonale

2.3.1 Introduction

Un algorithme capable de trouver la fréquence fondamentale ou la hauteur tonale (HT) est une partie essentielle d'un système de traitement de la parole. La fréquence fondamentale (ou la HT) est en effet un paramètre important dans le signal de parole. Elle permet de fournir l'information pour comprendre la nature de la source d'excitation dans la production de parole. Elle est très utile pour la reconnaissance des locuteurs [4] [54], pour l'apprentissage de la parole chez une personne atteinte de déficience auditive [27], et pour presque tout système d'analyse-synthèse de parole (vocodeur) qui a besoin de ce paramètre [12]. L'information prosodique dans une prononciation est déterminée de façon prédominante par la HT. De plus, l'oreille est plus sensible aux changements de la fréquence fondamentale qu'à ceux d'autres paramètres du signal de parole [21]. Par ailleurs, la présence de la HT dans la parole voisée est une source majeure de

redondance qui peut être exploitée de manière efficace en codage de parole à bas-débit. Néanmoins, la mesure précise de la HT est très difficile à effectuer, particulièrement pour de la parole bruitée. En raison de l'importance de la HT, beaucoup d'algorithmes pour l'extraction de HT ont été proposés dans la littérature du traitement de la parole.

Il faut remarquer que la plupart des méthodes proposées pour l'extraction de la HT n'utilisent pas certaines propriétés perceptives, mais elles sont réalisées avec une perspective d'application. Certains chercheurs [46] [38] [60] réalisent leurs algorithmes d'extraction de la HT en y incoporant des connaissances auditives.

A la section suivante, on revoit les méthodes proposées d'extraction de la HT et de la fréquence fondamentale. Ces méthodes se regroupent de la façon suivante:

(1) méthodes basées sur la forme d'onde;

(2) méthodes basées sur l'auto-corrélation;

(3) méthodes spectrales et cepstrales;

(4) méthodes perceptives à base de modèles du système auditif périphérique;

(5) méthodes utilisant la reconnaissance des formes.

2.3.2 Méthodes basées sur la forme d'onde

Ce type de méthodes traite directement le signal de la parole (ou ses versions filtrées) pour estimer la fréquence fondamentale. Pour ces méthodes, la stratégie utilisée le plus souvent est de mesurer les pics et les vallées du signal, ou de compter les passages par zéro du signal.

Un des premiers extracteurs de la fréquence fondamentale est celui de Gold-Rabiner [20], où l'extracteur combine toutes les sorties des six estimateurs parallèles de la

fréquence fondamentale. Une seule décision finale découle en utilisant la règle de majorité.

Actuellement la performance de ce type de méthodes, basées sur la structure temporelle de la parole, dépend en grande partie du bon choix des marqueurs du temps. Dans la pratique, les marqueurs utilisés fréquemment sont les pics de l'amplitude [35] et les passages par zéro [14] ou une combinaison de ces paramètres. Mais le désavantage de l'utilisation de ces marqueurs du temps est leur sensibilité aux variations du signal de la parole. Les variations mentionnées normalement sont causées par la nature transitoire de la source d'excitation, par la structure formantique apportée par le conduit vocal, et par tous les bruits de mesure. Afin de contourner ces obstacles, Nguyen et al. [40] ont proposé un marqueur du temps, défini par la formule suivante:

$$\overline{X} = \frac{\sum_{i=1}^{N} i \cdot x(i)}{M} \qquad et \qquad M = \sum_{i=1}^{N} x(i) \tag{2.1}$$

Selon les auteurs, les excursions ou bosses du signal de parole sont traitées comme des régions géométriques repésentées par leurs centres de masse et ces derniers définis ci-dessus peuvent être les marqueurs du temps.

Récemment, Medan et al. ont proposé un bon algorithme pour la détermination de la fréquence fondamentale de parole [34]. Cet algorithme a deux avantages. Le premier avantage est que l'algorithme est capable de vaincre la propriété de non-stationnarité de la parole en introduisant un modèle de similarité pour représenter les signaux de parole de deux périodes consécutives. Pour expliquer l'idée des auteurs explicitement, on décrira brièvement leur analyse ci-dessous:

A l'instant t_0, on définit deux signaux $x_\tau(t, t_0)$ et $y_\tau(t, t_0)$ représentés par les équations suivantes

$$\begin{aligned} x_\tau(t, t_0) &= s(t) w_\tau(t - t_0) \\ y_\tau(t, t_0) &= s(t + \tau) w_\tau(t - t_0) \end{aligned} \tag{2.2}$$

où *s(t)* représente la signal de parole et $w_\tau(t)$ est une fenêtre rectangulaire de largeur τ secondes. Notons que ces deux signaux sont non nuls seulement à l'intérieur de l'intervalle $[t_0, t_0 + \tau]$.

Ensuite, on considère un segment de parole commençant à t_0 et incluant exactement deux périodes de la fréquence fondamentale, dit $\tau = T_0$, supposant que $x_{T_0}(t; t_0)$ est la première période, $y_{T_0}(t, t_0)$ est la deuxième période et T_0 dénote la période de la fréquence fondamentale à l'instant t_0. On suppose que la similarité entre deux périodes consécutives de la fréquence fondamentale est grande, donc, on considère que $y_{T_0}(t)$ est une version modulée en amplitude de $x_{T_0}(t, t_0)$. Ceci s'exprime par le modèle suivant de similarité:

$$x_{T_0}(t, t_0) = a(t_0) y_{T_0}(t, t_0) + e(t, t_0) \tag{2.3}$$

où $a(t_0)$ est un facteur de modulation en amplitude à t_0 et qui reflète le changement d'amplitude de l'impulsion glottale, $e(t, t_0)$ est une fonction d'erreur qui reflète la différence entre ces deux périodes. L'intervalle de temps $\tau = T_0$ est défini comme étant la période de la fréquence fondamentale à l'instant $t = t_0$ quand $e(t, t_0)$ est minimale sur l'intervalle $[t_0, t_0 + \tau]$ ou quand la similarité entre deux périodes du signal est maximale. Alors la détermination de la période de la fréquence fondamentale T_0 est équivalente à

l'optimisation de la fonction de coût J:

$$T_0 = \arg \min_{a(t_0)>0} \left\{ J = \frac{\int_{t_0}^{t_0+\tau} [x_\tau(t,t_0) - a(t_0)y_\tau(t,t_0)]^2 dt}{\int_{t_0}^{t_0+\tau} [x_\tau(t,t_0)]^2 dt} \right\} \tag{2.4}$$

où *arg min* représente la valeur de J qui minimise l'expression.

Le deuxième avantage de cet algorithme est qu'il extrait la fréquence fondamentale avec une résolution supérieure de telle façon que la période de la fréquence fondamentale peut se représenter par un nombre fractionaire des échantillons. D'après les résultats communiqués par les auteurs [34], cet algorithme est aussi robuste au bruit.

2.3.3 Méthodes basées sur l'auto-corrélation

La fonction d'auto-corrélation utilisée pour la mesure de la fréquence fondamentale sur une fenêtre du signal peut se calculer par la formule suivante:

$$R(l) = \sum_{i=1}^{N-l} s(i) \cdot s(i+l) \qquad 0 \le l \le N-1 \tag{2.5}$$

où N est la largeur de fenêtre (le nombre d'échantillons dans la fenêtre), *s(i)* représente le signal de parole et l représente le décalage entre le signal original et le signal décalé. Le maximum de la fonction est obtenu en principe lorsque le décalage l est égal à la période du signal.

Dans ce paragraphe on présente deux types de méthodes basées sur l'auto-corrélation: les méthodes temporelles et les méthodes spectrales.

2.3.3.1 La fonction d'auto-corrélation dans le domaine temporel (FADT)

La FADT est bien connue dans l'extraction de la périodicité du signal bruité en raison de sa robustesse, et elle est toujours largement utilisée pour la détermination de

la fréquence fondamentale de la parole [50]. L'avantage que cette approche possède en comparaison des méthodes basées sur la forme d'onde est que la localisation absolue du pic du signal peut déterminer la période de la fréquence fondamentale. Un pic dans l'auto-corrélation au délai τ correspond directement à une période τ du signal, car l'origine est explicitement définie.

Cependant, pour cette méthode, il existe un problème aux résonances du conduit vocal. Ces dernières changent parfois rapidement pendant deux périodes consécutives d'impulsion glottale. Il en résulte que le signal n'est plus bien corrélé d'une période à l'autre. Dans ce cas, le pic situé au délai correspondant à la période de la fréquence fondamentale devient le moins dominant et souvent plusieurs pics superflus apparaissent. Évidemment, ceux-ci compliquent la mesure de la fréquence fondamentale. En considérant les raisons évoquées précédement, un prétraitement convenable du signal de parole doit être effectué pour lisser les spectres ou enlever les composantes de haute fréquence en utilisant un filtre passe-bas avant de calculer l'auto-corrélation [21] [25] [34]. Parallèlement, certains algorithmes d'extraction de la fréquence fondamentale sont basés sur le signal résiduel à partir du codage de prédiction linéaire ("Linear Predictive Coding", LPC). En d'autres termes, ces algorithmes fonctionnent sur la sortie du filtre inverse LPC [31] [65].

Un des bons exemples de ce type d'approche est l'algorithme proposé par Krubsack et Niederjohn [25]. Pour améliorer les performances de la détermination de la fréquence fondamentale et la décision de voisement, les auteurs ont utilisé deux mesures de confiance: l'exactitude probable de la valeur de la fréquence fondamentale et l'exactitude probable de la décision de voisement. Comme la détermination de la fréquence fonda-

mentale et la décision de voisement, ces deux mesures de confiance sont aussi déduites séparément à partir de la même fonction d'auto-corrélation.

2.3.3.2 La fonction d'auto-corrélation spectrale (FAS)

La FAS a d'abord été présentée par Chilton et Evans [8], et puis une approche similaire a été suggérée indépendamment par Labat et al. [26]. La FAS est utilisée pour mesurer les espacements harmoniques réguliers dans le spectre du signal de parole en appliquant l'auto-corrélation à la densité spectrale de puissance (DSP). La FAS et la FADT sont étroitement apparentées, mais la FADT et la DSP sont des transformées de Fourier l'une de l'autre. D'après Chilton et Evans [9], si *p(k)* représente la densité spectrale de puissance d'un signal, sa FAS, *R(k)* peut être définie par la formule suivante:

$$R(k) = \frac{2}{N} \sum_{s=0}^{N/2} p(s) \cdot p(s+k) \tag{2.6}$$

où k représente le décalage dans le domaine fréquentiel échantillonné.

Puisque *p(k)* est une fonction réelle paire, la fonction d'auto-corrélation peut se calculer à partir de seulement $N/2$ échantillons d'une transformée de Fourier discrète d'ordre N. Lorsque le signal est périodique, on observe des pics dans la fonction d'auto-corrélation spectrale. Ces pics correspondent à des harmoniques de la fréquence fondamentale, on retrouve dans le domaine fréquentiel certaines propriétés de l'auto-corrélation dans le domaine temporel.

De même que pour la FADT, la FAS nécessite un lissage préalable du spectre. Chilton et Evans [9] ont trouvé qu'un prétraitement effectif peut se faire en utilisant une préaccentuation du signal, puis un filtrage inverse pour lequel les coefficients du filtre sont les coefficients LPC.

2.3.4 Méthodes spectrales et cepstrales

Aux paragraphe suivants, nous allons présenter deux types de méthodes pour l'extraction de la fréquence fondamentale: méthodes spectrales et méthodes cepstrales.

2.3.4.1 Méthode spectrale

La méthode spectrale est définie comme étant celle qui permet d'obtenir la fréquence fondamentale en traitant le spectre de la parole directement, en d'autres termes, en détectant une série de pics spectraux. Cette méthode déduit la fréquence fondamentale à partir de deux techniques: le calcul du plus petit commun multiple à une série d'harmoniques et la mesure de l'espacement entre les pics spectraux (harmoniques).

La première technique peut se faire en mesurant les périodes d'harmoniques individuelles et en trouvant le plus petit commun multiple à ces périodes[57] [36]. Cette approche est appelée: mesure de la fréquence fondamentale basée sur l'histogramme de période. Il existe un autre moyen pour réaliser la première technique en cherchant un bon appariement d'harmoniques de la fréquence fondamentale [64] [11] [61] [44]. Un bon exemple de ces travaux est l'algorithme proposé par Terhardt. Dans cet algorithme, les hauteurs spectrales d'harmoniques individuelles sont d'abord calculées, ensuite, la HT est déduite à partir de ces hauteurs spectrales, en utilisant le principe d'appariement de sous-harmoniques. Par exemple, pour un signal complexe consistant en trois composantes de fréquences 520 Hz, 620 Hz et 720 Hz respectivement, les hauteurs spectrales de ces trois composantes correspondent directement à leurs fréquences, et les sous-harmoniques de ces composantes sont présentées au tableau 2.1. On trouve que trois sous-harmoniques (104 Hz, 103.3 Hz et 102.9 Hz) constitutent le plus petit

Numéro de sous harmonique	Fréquence de composante (Hz)		
	520	620	720
1	520.0	620.0	720.0
2	260.0	310.0	360.0
3	173.3	206.7	240.0
4	130.0	155.0	180.0
5	***104.0***	124.0	144.0
6	86.7	***103.3***	120.0
7	74.3	88.6	***102.9***
8	65.0	77.5	90.0

Tableau 2.1 Fréquences sous-harmoniques des composantes 520, 620 et 720 Hz d'après Terhardt [64]

intervalle de l'ensemble. Selon le principle d'appariement de Terhardt, la fréquence de sous-harmonique pertinente à la plus basse composante, 104.0 Hz, correspond à la HT.

Puisque les harmoniques de la fréquence fondamentale sont bien séparées dans le spectre du signal de parole, alors, il suffit de mesurer la distance entre les pics adjacents spectraux pour déterminer la fréquence fondamentale, ce qui correspond à la deuxième technique mentionnée précédemment. Hess l'a décrite en détail [21].

2.3.4.2 Méthode cepstrale

Une des techniques populaires dans l'extraction de la fréquence fondamentale est basée sur l'utilisation du cepstre qui a été à l'origine proposée par Noll [41] [43]. Pour faciliter la compréhension de cette approche, il est nécessaire de revoir brièvement la façon dont la parole est produite. Le système de production de parole consiste en une source glottale et un conduit vocal comme indiqué à la figure 2.2. Le signal de source, $s(t)$, est produit par un débit d'air à travers les cordes vocales. Le conduit vocal est complètement specifié par sa réponse impulsionnelle $h(t)$. Donc le signal de parole $f(t)$

peut s'exprimer par l'équation suivante:

$$f(t) = s(t) * h(t) \tag{2.7}$$

où $*$ représente la convolution.

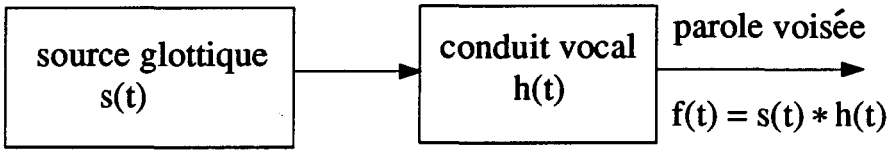

Figure 2.2 Système fondamental de production de la parole voisée, où h(t) est la réponse impulsionnelle du conduit vocal.

Alternativement, si *S(f)* est le spectre du signal de la source et *H(f)* est le spectre de la réponse impulsionnelle du conduit vocal, alors le spectre du signal de parole peut s'exprimer algébriquement par l'équation suivante

$$F(f) = S(f) \cdot H(f) \tag{2.8}$$

où

$$\begin{aligned} F(f) &= F_{TF}[f(t)] \\ S(f) &= F_{TF}[s(t)] \\ H(f) &= F_{TF}[h(t)] \end{aligned} \tag{2.9}$$

F_{TF} dénote la transformée de Fourier

Le spectre de puissance de parole s'exprime par l'équation suivante:

$$|F(f)|^2 = |S(f)|^2 \cdot |H(f)|^2 \tag{2.10}$$

En prenant le logarithme, l'équation ci-dessus devient:

$$log|F(f)|^2 = log|S(f)|^2 + log|H(f)|^2 \tag{2.11}$$

Donc, les effets de la source vocale et du conduit vocal sont séparés. En calculant la transformée de Fourier inverse au spectre de puissance logarithmique, on obtient le cepstre:

$$F_{TF}^{-1}\left[log|F(f)|^2\right] = F_{TF}^{-1}\left[log|S(f)|^2\right] + F_{TF}^{-1}\left[log|H(f)|^2\right] \qquad (2.12)$$

F_{TF}^{-1} dénote la transformée inverse de Fourier.

Dans le cesptre, la terminologie "quefrence" est utilisée et correspond à l'inverse de la période du pic dans le cesptre du signal de parole. Quand la parole est voisée, le cepstre comprend un grand pic à la quefrence correspondant à la période de source glottale. Si l'intervalle d'analyse comprend diverses périodes du signal, les pics distincts dans le cepstre vont apparaître correspondant respectivement à chaque période individuelle. De cette façon, une analyse fondée sur le cepstre est capable de détecter la fréquence fondamentale. Un important avantage de cette méthode est que: l'extraction de la fréquence fondamentale est indépendante de la présence de la composante fondamentale du signal de parole, car le pic cepstral est produit par la structure harmonique du spectre [42]. De plus, cette approche est insensible à la distorsion de phase.

Bien que les méthodes cepstrales sont largement utilisées pour l'extraction de la fréquence fondamentale, on n'a pas encore prouvé que ce type de méthode est capable de fonctionner avec succès pour la parole bruitée [43] [26]. Par ailleurs, cette méthode a un désavantage commun avec l'auto-corrélation: le choix de la longueur de la fenêtre d'analyse. De plus, dans une optique de traitement du signal, le calcul du cepstre est coûteux, car ce dernier doit être suréchantillonné pour éviter le repliement [59]. D'après Rabiner et al. [51], cette approche n'est pas une bonne méthode au regard de sa performance dans la détermination de la fréquence fondamentale.

2.3.5 Méthodes perceptives à base de modèles du système auditif périphérique

Au cours des dernières années, certains auteurs ont utilisé des modèles auditifs périphériques pour la perception ou l'extraction de HT [59] [60] [18]. Parallèlement, quelques chercheurs ont proposé des algorithmes d'extraction de HT en utilisant des connaissances perceptives [46] [49] [64]. D'après Patterson [46], le système auditif du mammifère se divise probablement en deux parties: le sous-système analogue et le sous-système numérique. Le sous-système analogue consiste en l'oreille externe, l'oreille moyenne et l'oreille interne. Sa fonction principale est d'effectuer une analyse spectrale du son d'entrée. Le sous-système numérique est compris de nerfs auditifs et de sections auditives du cerveau, et sa fonction est de convertir les sorties du sous-système analogue (l'analyse spectrale) en une série d'impulsions neurales, puis d'analyser en détail les caractéristiques de ces impulsions. Le mécanisme déterminant la HT d'un son est situé dans le sous-système numérique. Spécifiquement, Patterson maintient qu'il semble plus plausible que le sous-système numérique analyse les informations temporelles des impulsions neurales et qu'il les combine avec les informations spectrales lorsque la HT est déterminée. Cependant, certains algorithmes utilisent seulement des modèles spectraux pour lesquels la détermination de HT ne se base que sur les informations spectrales. Moore [38] croit qu'il ne suffit pas d'extraire la HT de la parole en n'utilisant que le modèle spectral, et que pour la perception ou l'extraction de HT nous avons besoin du modèle synthétique (le modèle spectro-temporel). Patterson aussi croit qu'il est difficile de comprendre comment un modèle spectral peut expliquer la perception du timbre, car dans la plupart des cas, le timbre d'un son peut être changé par la phase des harmoniques de haute fréquence, mais cette information de phase

n'existe plus dans un modèle spectral [46].

L'algorithme d'extraction de HT proposé par Patterson comprend un modèle spectro-temporel et un extracteur de HT. Le modèle spectro-temporel effectue une analyse spectrale de la parole en utilisant un banc de filtres auditifs, puis il convertit les sorties des filtres en une série d'impulsions neurales. A partir de ces impulsions neurales, on peut observer qu'il existe un patron de répétition, en d'autres termes, les impulsions neurales sont clairement modulées par ce "patron" de répétition. Il est très important que le taux de répétition corresponde à la fréquence de HT. En se basant sur cette idée, Patterson a conçu un extracteur de périodicité, nommé processeur spiral ("spiral processor", d'après l'auteur), qui est capable d'extraire la période de répétition à partir d'impulsions neurales modulées. Patterson a expliqué qu'un arrangement en spirale d'impulsions neurales peut convertir la régularité temporelle produite par le son périodique en information de position. En fait, le processeur spiral détecte la fréquence de la porteuse et la fréquence de modulation à partir d'impulsions neurales modulées de façon séparée ou combinée. Il arrange l'information en une forme qui lui permet d'être superposée à travers les canaux fréquentiels afin de produire une estimation générale de HT sans avoir besoin d'un générateur de sous-harmonique [63] ou d'une sélection d'harmonique [11].

Moore a proposé un modèle pour l'extraction de la HT de sons complexes [38]. Ce modèle, indiqué à la figure 2.3, comprend cinq modules. Le premier module est un banc de filtres. Le deuxième est la transduction des sorties des filtres aux impulsions neurales. Le troisième est un mécanisme qui, pour chaque canal fréquentiel, analyse les intervalles de décharge d'impulsions neurales. Le quatrième compare les intervalles de

décharge présents aux différents canaux, puis trouve les intervalles communs à travers les canaux. Généralement, l'intervalle qu'il trouve le plus souvent correspond à la période du fondamental. Enfin, le cinquième module effectue une sélection finale à partir des intervalles qui sont le plus souvent réprésentés à travers les canaux, et l'inverse de l'intervalle sélectionné correspond à la HT perçue.

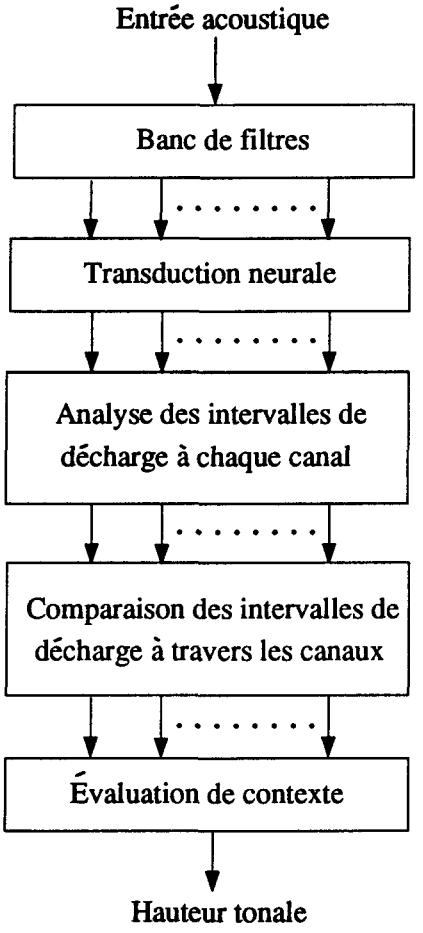

Figure 2.3 Modèle schématique pour l'extraction de HT proposé par Moore [38]

Slaney et Lyon ont proposé un détecteur perceptif de HT [60]. Ce détecteur comprend trois parties: un modèle cochléaire, un banc d'auto-corrélateurs et un mécanisme de prise de décision de HT. Nous présentons un schéma à la figure 2.4 de ce détecteur. Le modèle cochléaire convertit l'information du signal acoustique en une

représentation à canaux multiples qui peut être considérée comme étant liée aux probabilités de décharge instantanée des nerfs. Le corrélogramme, produit en calculant l'auto-corrélation indépendante de chaque canal, permet de représenter le signal de la parole dans deux dimensions: la composante fréquentielle sur la direction verticale et la structure temporelle sur la direction horizontale. Lorsqu'un son d'entrée est périodique, les fonctions d'auto-corrélation de tous les canaux montrent un pic à la même position horizontale, en d'autres termes il y a une synchronisation du pic à travers les canaux. Le délai de corrélation, relié à cette position, en général correspond à la période de HT perçue [28]. En fait, la localisation du pic synchronisé à partir du corrélogramme se réalise lors du mécanisme de prise de décision de HT.

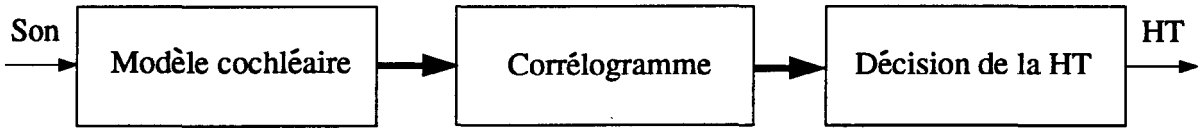

Figure 2.4 Structure du déctecteur de HT proposé par Slaney et Lyon [60]

2.3.6 Méthodes utilisant la reconnaissance des formes

Des études récentes faites autour de la détermination de fréquence fondamentale utilisant le perceptron à multi-couches ont été rapportées par certains auteurs [22] [33]. Dans ces études, Howard et Walliker ont proposé un algorithme appelé "MLP-TX" pour estimer la période du signal de parole bruitée. Dans cette étude, les auteurs utilisent une classification par perceptron multi-couches (MLP, "Multi-Layer Perceptron") pour déterminer l'instant de la fermeture des cordes vocales. L'algorithme global s'organise d'une telle façon que le signal de parole est d'abord filtré par un banc de filtres, puis une transformée non linéaire (le redressement demi-alternance) est appliquée à la sortie de chaque filtre, et finalement les informations à travers les canaux se groupent

sous la forme de vecteurs pour faire la classification par le perceptron. De la même façon, Martinez-Alfaro et Contreras-Vidal ont présenté un algorithme de détection de la fréquence fondamentale dont la classification par perceptron multi-couches a été réalisée par un réseau neural en utilisant l'algorithme d'apprentissage appelé "rétro-propagation". Cet algorithme ne nécessite pas de pré-traitement du signal de parole et la classification possède une grande capacité de discrimination, d'après les auteurs.

2.4 Conclusion

Dans le présent chapitre nous avons discuté des différentes méthodes d'extraction de la fréquence fondamentale et de la HT. Tous ces algorithmes ont des avantages et des inconvénients. Le choix doit se faire selon le but à atteindre. Pour la méthode qui utilise l'analyse par la fenêtre de courte durée (par exemple, l'auto-corrélation), le choix de la bonne longueur de la fenêtre n'est pas facile à faire car le signal de parole dans la pratique n'est pas bien corrélé d'une période à l'autre. Cependant, les méthodes basées sur le traitement direct du signal de parole (par exemple, la détection des pics sur la forme d'onde) sont influencées par les décalages de phase car ceux-ci tendent à brouiller les pic du signal [59]. Par contre, les méthodes fondées sur l'auto-corrélation ou sur le cepstre sont robustes aux distorsions de phase.

D'ailleurs, on trouve qu'il est inadéquat d'utiliser seulement la détection du pic du signal de parole lorsqu'elle est bruitée, même si la fréquence fondamentale de ce type de parole est existante. Il est probable que la fréquence fondamentale existante dans la parole bruitée peut se détecter en comparant la forme d'onde avec sa version décalée. Néanmoins, cette comparaison est difficile en considérant que les résonances du conduit

vocal sont continuellement changeantes dans le temps. En d'autres termes, le signal de parole varie d'une période à l'autre. Une solution à ce problème est d'utiliser un pré-traitement afin de lisser le spectre du signal, ou de diminuer la variabilité entre deux périodes adjacentes. Pourtant la performance globale de ce type d'algorithme n'est toujours pas satisfaisante, car le pré-traitement du signal fonctionne mal lors de la transition voisée/non voisée de la parole [59].

En résumé et conformément à l'analyse de Rabiner et al. [51] tous les problèmes recontrés par les algorithmes d'extraction de la fréquence fondamentale sont causés par les raisons suivantes:

(1) La source d'excitation glottale n'est pas parfaitement périodique;

(2) Il y a interaction entre le conduit vocal et la source d'excitation glottale;

(3) Il est difficile de définir exactement le début et la fin de chaque période du signal de parole voisée;

(4) Il est difficile de faire la distinction entre la parole non voisée et la parole moins voisée.

Pour contourner les difficultés énumérées ci-dessus, l'extraction de la fréquence fondamentale peut se faire par la réalisation d'un modèle auditif en utilisant des connaissances psychoacoustiques et physiologiques. Il est évident que l'être humain est capable de percevoir la parole beaucoup mieux que tous les systèmes informatiques. Par exemple, lorsqu'une personne écoute la parole naturelle en milieu bruité, l'intelligibilité de la parole reste relativement haute, par contre, pour certains vocodeurs, l'intelligibilité de la parole vocodée tombe rigoureusement [19]. Il est donc raisonnable d'adopter la

structure du système auditif humain afin d'obtenir de meilleurs traitements de la parole [15].

CHAPITRE 3

MODÈLE PROPOSÉ

3.1 Introduction

Lorsqu'on étudie la cadence de décharge des impulsions sur une fibre par rapport à la durée de la stimulation acoustique sinusoïdale, on observe essentiellement que les impulsions sont synchronisées sur la période de la stimulation pourvu que la cadence de décharge ne soit pas très élevée [46]. Dans le cadre de cette étude, on peut assumer que la fibre décharge aux pics des signaux de stimulation, comme si elle était stimulée par une série d'impulsions. En fait, la réponse d'une seule fibre à la parole périodique est un flot d'impulsions avec espacement régulier, en d'autres termes, un son périodique tel qu'une voyelle peut produire un flot d'impulsions régulières à la sortie des neurones auditifs primaires [46].

Notons qu'un flot d'impulsions produites par les fibres nerveuses auditives préservent des informations concernant l'intervalle sur le temps entre deux pics positifs du signal périodique d'entrée. De plus, si la caractéristique stochastique de la transduction neurale est ignorée, on peut assumer que les signaux sortant des filtres auditifs sont équivalents à ces impulsions nerveuses. Nous présentons un exemple à la figure 3.1 pour illustrer ce qui vient d'être écrit. Sur cette figure, on peut voir que les sorties des filtres auditifs sont modulées en amplitude et que la cadence de répétition est étroitement

reliée à la hauteur tonale. La tâche actuelle est de réaliser un extracteur de périodicité qui fonctionne sur les signaux modulés à la sortie des filtres auditifs.

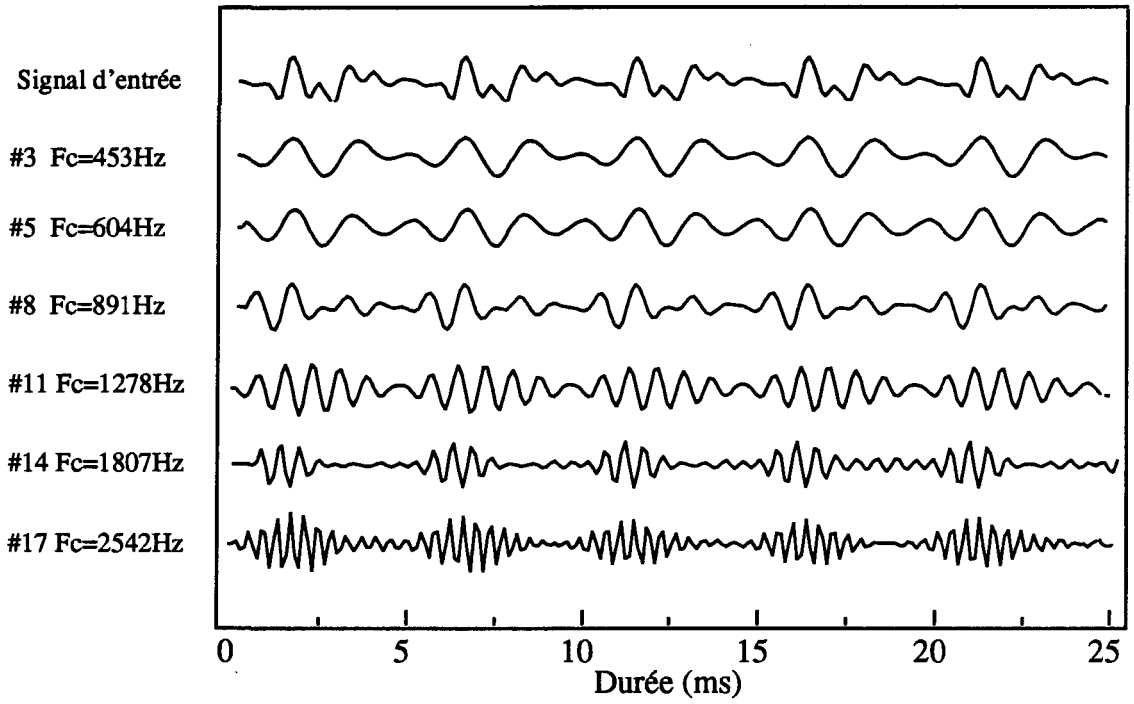

Figure 3.1 Sorties des filtres auditifs en réponse à une portion de la voyelle nasale / ã /

Il est nécessaire de souligner que le modèle proposé n'est pas un modèle auditif exact, mais plutôt un modèle fonctionnel d'extraction de HT. Ce modèle fonctionnel s'intéresse plus particulièrement au traitement des informations à la sortie des filtres auditifs en incorporant des caractéristiques perceptive auditives et des connaissances auditives. En fait, pour extraire la HT, il n'est pas nécessaire de construire un modèle auditif périphérique complet, parce qu'il nous semble que les informations à la sortie des filtres auditifs sont suffisantes pour l'extraction de la HT.

Dans ce chapitre, nous allons introduire en détail chaque module du modèle proposé. D'abord, nous allons donner une présentation générale du modèle, puis introduire la conception du banc de filtres, et du sous-modèle fonctionnel du modèle, enfin, nous

allons expliquer l'algorithme de décision et de post-traitement. Le travail porte sur des données de la parole téléphoniques dont la largeur de bande considérée est de 300 Hz à 3400 Hz, et la fréquence d'échantillonnage est de 8 kHz.

3.2 Présentation générale du modèle

Le modèle proposé comprend trois modules: un banc de filtres, un sous-modèle fonctionnel et un mécanisme de prise de décision. Le banc de filtres est composé de 19 filtres auditifs passe-bande et couvre la plage de fréquence de 300 Hz à 3400 Hz. Il simule l'analyse fréquentielle accomplie par la cochlée. Le sous-modèle fonctionnel est un modèle mathématique qui consiste en la rectification de la sortie du filtre auditif et du signal original, la mutiplication de deux signaux et l'auto-corrélation. Le mécanisme de prise de décision est basé sur un algorithme d'extraction de HT à partir de la somme des auto-corrélations et d'un post-traitement. Le schéma du modèle est présenté à la figure 3.2.

Pour extraire la HT de la parole bruitée, nous proposons un autre sous-modèle fonctionnel pour lequel il n'y a pas de multiplication des signaux, par contre nous ajoutons une sélection dans chaque canal pour décider si le canal participe à la décision finale de HT. Le schéma du modèle dans ce cas est donné à la figure 3.3.

On donne ici seulement les grandes lignes du modèle, les détails seront donnés plus loin.

3.3 Banc de filtres

Dans ce paragraphe, on présente chaque étage de façon détaillée en regard de la conception du banc de filtres: le filtre auditif, la distribution des filtres,

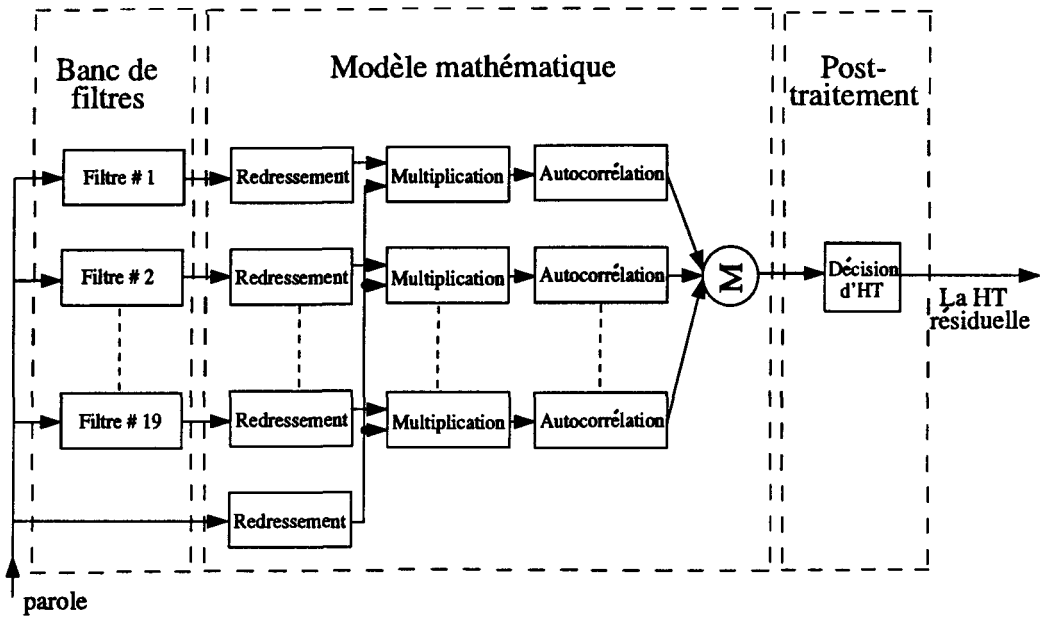

Figure 3.2 Structure générale du modèle

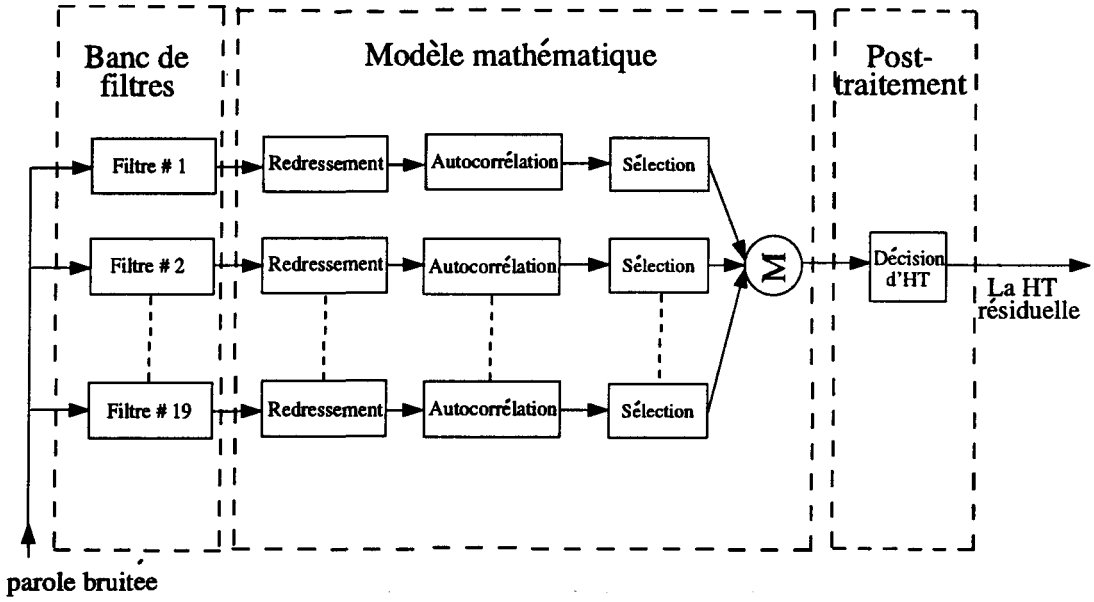

Figure 3.3 Structure générale du modèle pour la parole bruitée

les caractéristiques des filtres en amplitude et phase, le délai du filtre auditif et les performances du banc de filtres.

3.3.1 Filtre auditif

D'après les données de Patterson [45], et de Moore et Glasberg [39], les filtres

auditifs seraient des filtres à pentes exponentielles arrondis à leur sommet (représentation spectrale). Ces auteurs les caractérisent par la largeur de bande d'un filtre rectangulaire équivalent (ou ERB pour «Equivalent Rectangular Bandwidth»). En d'autres termes, il s'agit d'un filtre rectangulaire dont un «côté» aurait la même longueur que la hauteur maximale du véritable filtre (même puissance maximale transmise pour un son pur) et dont la «surface» serait égale à celle du véritable filtre (même puissance de bruit blanc transmise). La largeur de bande passante de ces filtres rectangulaires équivalents correspond théoriquement, selon les auteurs, à la largeur des bandes critiques mesurées expérimentalement. La formule suivante donne la largeur de la bande critique *ERB* en Hertz, à partir de la fréquence centrale, f_c exprimée en kHz:

$$ERB(f_c) = 6.23f_c^2 + 93.39f_c + 28.52 \tag{3.1}$$

Dans le modèle proposé, la forme du filtre auditif utilisée est une exponentielle arrondie simple dont l'expression est la suivante:

$$W(g) = (1 + pg)e^{-pg} \tag{3.2}$$

où g est la déviation de la fréquence à partir du centre du filtre, divisée par la fréquence centrale (f_c), c'est-à-dire, $g = |f - f_c|/f_c$. Le paramètre p détermine la largeur de bande du filtre, pour cette forme simplifiée de l'équation 3.2, il est égal à $4f_c/ERB(f_c)$. La figure 3.4 montre la forme schématisée d'un filtre auditif dont la fréquence centrale est de 1000 Hz.

Dans la pratique, pour réaliser le filtre, nous avons choisi la méthode par séries de Fourier. Cette méthode consiste à trouver une série de Fourier qui représente

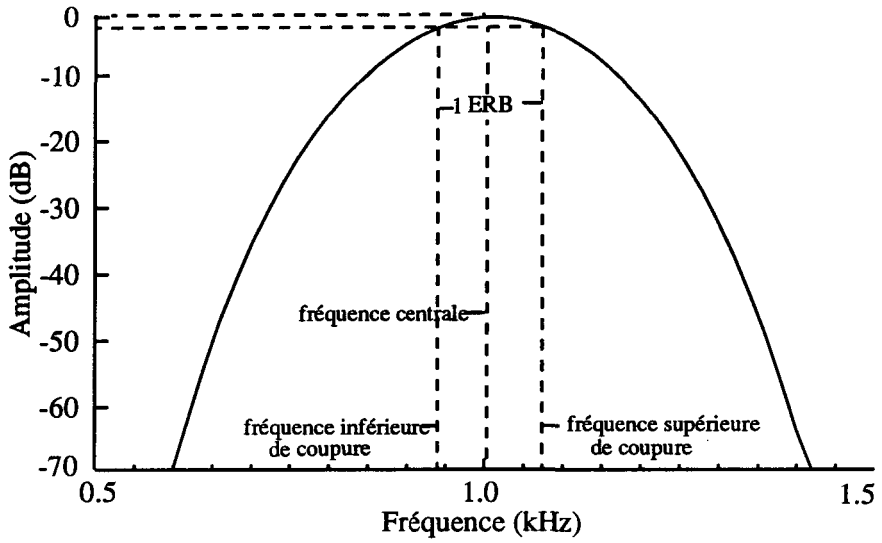

Figure 3.4 Forme schématisée d'un filtre auditif avec F_c = 1008 Hz

convenablement après troncature la caractéristique *H(f)* de la réponse fréquentielle du filtre désiré.

Etant donné la réponse impulsionnelle *h(n)*, sa transformée de Fourier est donnée par la relation:

$$H(f) = \sum_{n=-\infty}^{\infty} h(n) e^{-i2\pi nf/f_s} \tag{3.3}$$

où f_s est la fréquence d'échantillonnage.

La relation inverse nous donne:

$$h(n) = \frac{1}{f_s} \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} H(f) e^{i2n\pi f/f_s} df \tag{3.4}$$

Pour le calcul des coefficients de Fourier ou des coefficients du filtre, il faut trouver le résultat de l'intégrale d'équation 3.4 avec $-l \leq n \leq l$, où l est l'ordre du filtre. Dans notre cas, la réponse frequentielle *W(f)* est caractéristée par l'équation 3.2. Les

coefficients d'un filtre ayant la fréquence centrale f_c sont obtenus à partir de la formule suivante:

$$h(n) = \frac{1}{f_s} \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} W(f) e^{i2n\pi f/f_s} df \qquad -l \leq n \leq l \tag{3.5}$$

où

$$\begin{aligned} W(f) &= (1 + pg)e^{-pg} \\ g &= \frac{|f - f_c|}{f_c}, \quad p = \frac{4f_c}{ERB(f_c)} \\ ERB(f_c) &= 6.23f_c^2 + 93.39f_c + 28.52 \end{aligned} \tag{3.6}$$

Le résultat de l'intégrale nous donne (f_c en kHz):

$$h(n) = \frac{bc}{a}\left\{\cos(n\pi)e^{-\frac{c}{b}}\left[e^{p}k_1 - e^{-p}k_2\right] + \frac{4}{a}c^2\cos(bn\pi f_c)\right\} \quad et \quad -l \leq n \leq l \tag{3.7}$$

où

$$\begin{aligned} k_1 &= p - \frac{c}{b} - \frac{2}{a}c^2 \\ k_2 &= p + \frac{c}{b} + \frac{2}{a}c^2 \\ p &= \frac{4f_c}{6.23f_c^2 + 93.39f_c + 28.52} \\ c &= \frac{p}{f_c}, \quad b = \frac{2}{f_c} \\ a &= c^2 + (bn\pi)^2 \end{aligned} \tag{3.8}$$

l: l'ordre du filtre

f_c: la fréquence centrale d'un filtre

f_s: la fréquence de l'échantillonage

Nous donnons les détails du calcul à l'annexe A.

3.3.2 Distribution des filtres auditifs

Certains auteurs [67] [68] [69] soulignent qu'il est préférable d'utiliser une échelle spectrale basée sur des données psychoacoustiques, telle que l'échelle Bark, plutôt que d'exprimer les fréquences en Hertz. Zwicker a utilisé l'échelle de bande critique (la densité de bande critique) pour allier les fréquences en Hertz à l'échelle Bark. Cette échelle de bande critique peut être obtenue en intégrant la fonction réciproque de la fonction de la bande critique. De la même façon, une fonction comparable, pour allier les fréquences en Hertz à l'échelle de ERB (la densité de ERB), s'obtient à partir de l'équation 3.1:

$$\begin{aligned} ERB\_{ech}(f) &= \int \frac{1}{ERB(f)} df \\ ERB\_{ech}(f) &= 11.17 ln \left| \frac{f + 0.312}{f + 14.675} \right| + 43.0 \end{aligned} \tag{3.9}$$

où la fréquence est exprimée en kHz. Cette fonction est tracée à la figure 3.5.

Pour le modèle proposé, la distance entre deux filtres auditifs consécutifs est constante en longueur sur la membrane basilaire, ce qui signifie que la différence de fréquence caractéristique (centrale) entre deux filtres auditifs est constante en terme d'échelle de ERB. Notons que le banc de filtres auditifs dans le modèle proposé ne simule que la zone de 300 Hz à 3400 Hz correspondant à la bande du téléphone, mais ce qui ne signifie pas que la conception du banc de filtres se borne seulement à cette région. A l'annexe B, on verra que le banc de filtres auditifs utilisé peut fonctionner sur toute autre région spectrale en entrant les paramètres désirés lors de la génération des coefficients des filtres.

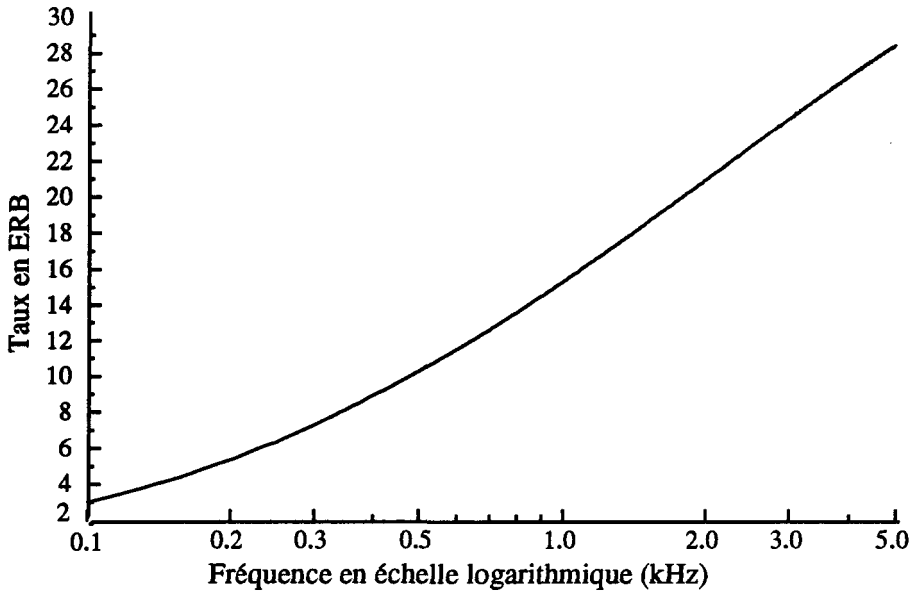

Figure 3.5 Fonction reliant la fréquence en Hz à l'échelle de ERB

La fréquence centrale d'un filtre en terme d'échelle de ERB est calculée dans le modèle proposé par la formule suivante:

$$f_{ci} = f_{min} + 0.5 + (f_{max} - f_{min} - 1.0)\frac{i - 1.0}{N - 1.0} \tag{3.10}$$

où

f_{ci}: fréquence centrale du filtre i en terme d'échelle de ERB

f_{min}:fréquence minimale en terme d'échelle de ERB dans le banc.

f_{max}:fréquence maximale en terme d'échelle de ERB dans le banc.

N: nombre total de filtres dans le banc.

En calculant l'inverse de l'équation 3.9 on obtient la fréquence centrale du filtre i exprimée en kHz:

$$f_{ci\_Hz}(f_{ci}) = \frac{14.675e^{\frac{f_{ci}-43}{11.17}} - 0.312}{1 - e^{\frac{f_{ci}-43}{11.17}}} \tag{3.11}$$

Dans la pratique, la largeur de bande de chaque filtre est égale à une unité de ERB. En d'autres termes, la distance entre les deux fréquences de coupure de chaque filtre a une valeur d'une unité de ERB. Si f_{hi} représente la fréquence supérieure de coupure du filtre i en terme d'échelle de ERB et f_{bi} représente la fréquence inférieure de coupure de ce filtre en terme de l'échelle de ERB, la fréquence centrale du filtre peut s'exprimer par:

$$f_{ci} = \frac{f_{bi} + f_{hi}}{2} \tag{3.12}$$

car la fréquence centrale f_{ci} est au milieu de la bande de fréquence en terme d'échelle de ERB.

Automatiquement, la fréquence supérieure de coupure du filtre i en terme d'échelle de ERB, f_{hi} peut s'exprimer par:

$$f_{hi} = f_{ci} + 0.5 \tag{3.13}$$

et la fréquence inférieure de coupure de ce filtre en terme d'échelle de ERB, f_{bi} peut s'exprimer par:

$$f_{bi} = f_{ci} - 0.5 \tag{3.14}$$

Les paramètres utilisés actuellement pour le banc de filtres auditifs sont reportés au tableau 3.1. Les filtres se chevauchent et les largeurs des bandes augmentent avec les fréquences centrales (mesurées en Hz). Notons que l'ordre des filtres est de 19. C'est aussi l'ordre le plus bas qui garantit que tous les filtres vérifient les spécifications souhaitées.

Numéro du filtre	Fréquence centrale (Hz)	Fréquence inférieure de coupure (Hz)	Fréquence supérieure de coupure (Hz)	Largeur de bande (Hz)	Réponse maximale après la correction de gain (dB)
1	329.27	300.00	360.00	60.00	-16.15
2	388.73	356.63	422.45	65.18	-12.71
3	453.98	418.75	491.01	72.26	-9.87
4	525.65	486.95	566.35	79.40	-8.30
5	604.44	561.88	649.20	87.32	-6.65
6	691.12	644.28	740.40	96.12	-5.08
7	786.59	734.99	840.91	105.93	-3.90
8	891.85	834.94	951.81	116.87	-2.59
9	1008.07	945.22	1074.33	129.11	-1.70
10	1136.54	1067.04	1209.88	142.84	-0.86
11	1278.79	1201.81	1360.10	158.28	-0.24
12	1436.56	1351.15	1526.85	175.70	-0.03
13	1611.85	1516.91	1712.34	195.43	1.09
14	1807.03	1701.27	1919.11	217.84	2.59
15	2024.86	1906.76	2150.18	243.42	3.82
16	2268.58	2136.36	2409.10	272.74	4.72
17	2542.06	2393.59	2700.12	306.53	5.27
18	2849.94	2682.66	3028.36	345.69	5.31
19	3197.80	3008.63	3400.00	391.37	4.87

Tableau 3.1 Les paramètres utilisés dans le banc de filtres

3.3.3 Caractéristiques des filtres auditifs en amplitude et phase

D'après des données psychoacoustiques [70] [56] [13], et la norme AFNOR NFS300–003 (1965), nous avons développé une fonction expérimentale pour caractériser le gain des filtres auditifs. En tenant compte des atténuations initiales du banc de filtres, on trouve que cette fonction peut se représenter par la formule suivante:

$$gain(f_c) = g(f_c)8.889e^{-0.1054i} \tag{3.15}$$

où f_c est la fréquence centrale en kHz du filtre i et

$$g(f_c) = \begin{cases} 0.37f_c^2 + f_c & 0.1\ kHz < f_c \leq 0.4\ kHz \\ -0.336f_c^2 + 1.12f_c + 0.076 & 0.4\ kHz < f_c \leq 1.5\ kHz \\ -0.188f_c^2 + 1.12f_c - 0.403 & 1.5\ kHz < f_c \leq 4.0\ kHz \end{cases} \tag{3.16}$$

Autrement dit, le gain augmente de façon non linéaire avec la fréquence centrale du filtre auditif, et arrive au maximum quand f_c est autour de 3 kHz, et diminue ensuite. Les réponses en amplitude (gain) du banc des filtres sont présentées à la figure 3.6. Ici le nombre total des filtres auditifs utilisés est égal à 19. C'est aussi la valeur minimale que nous avons trouvée expérimentalement pour laquelle le recouvrement entre la fréquence supérieure de coupure d'un filtre et la fréquence inférieure de coupure du filtre suivant est suffisant. La fonction de transfert de chacun de ces filtres peut s'obtenir (la figure3.6) en calculant une transformée rapide de Fourier (FFT) de leurs réponses impulsionnelles.

Comme nous avons utilisé la méthode des séries de Fourier pour trouver les coefficients du filtre, la phase est indépendante des coefficients du filtre et elle est linéaire dans la bande passante [62]. La relation de phase peut être exprimée par:

$$p(f) = \frac{2\pi l f}{f_s} \tag{3.17}$$

où l est l'ordre du filtre. A la figure 3.7 on présente les caractéristiques de phase pour quatre filtres. On trouve que la relation de phase du filtre avec la fréquence dans la bande passante est linéaire et elle est en concordance avec la formule théorique de l'équation ci-dessus.

3.3.4 A propos du délai de propagation sur la cochlée et du délai du filtre

Selon certains travaux de recherche sur la physiologie de la cochlée, on suppose que les cellules dont la fréquence caractéristique est en basse fréquence, ne peuvent

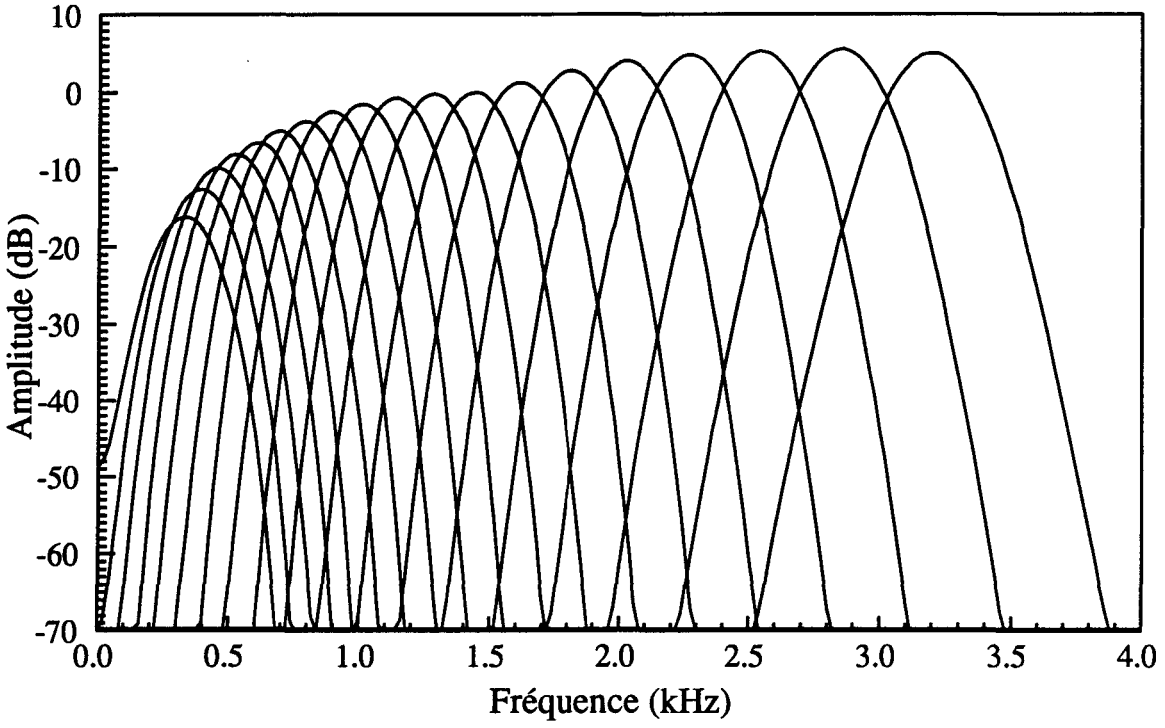

Figure 3.6 Réponses en amplitude du banc de filtres

répondre au stimulus qu'après un certain intervalle de temps, dû au délai de propagation sur la membrane basilaire dans la cochlée. Ce délai augmente avec la distance entre la position d'une cellule et l'extrémité basale de la cochlée. Certains auteurs considèrent ce phénomène de propagation dans leurs modèles auditifs périphériques [30] [29] [60] [10] [18] [66]. Particulièrement, lorsqu'ils simulent la vibration de la membrane basilaire de la cochlée en utilisant un modèle mécanique liquide, ce délai de propagation est compris automatiquement dans le modèle [37] [1] [2] [7] [6]. Cependant, d'autres auteurs ignorent ce délai de propagation dans leurs modèles auditifs [59] [16] [17] [32] [46].

Patterson [47] a effectué des expériences de perception de phase en utilisant un modèle perceptif de l'audition. Il a trouvé que le système auditif est capable de s'adapter

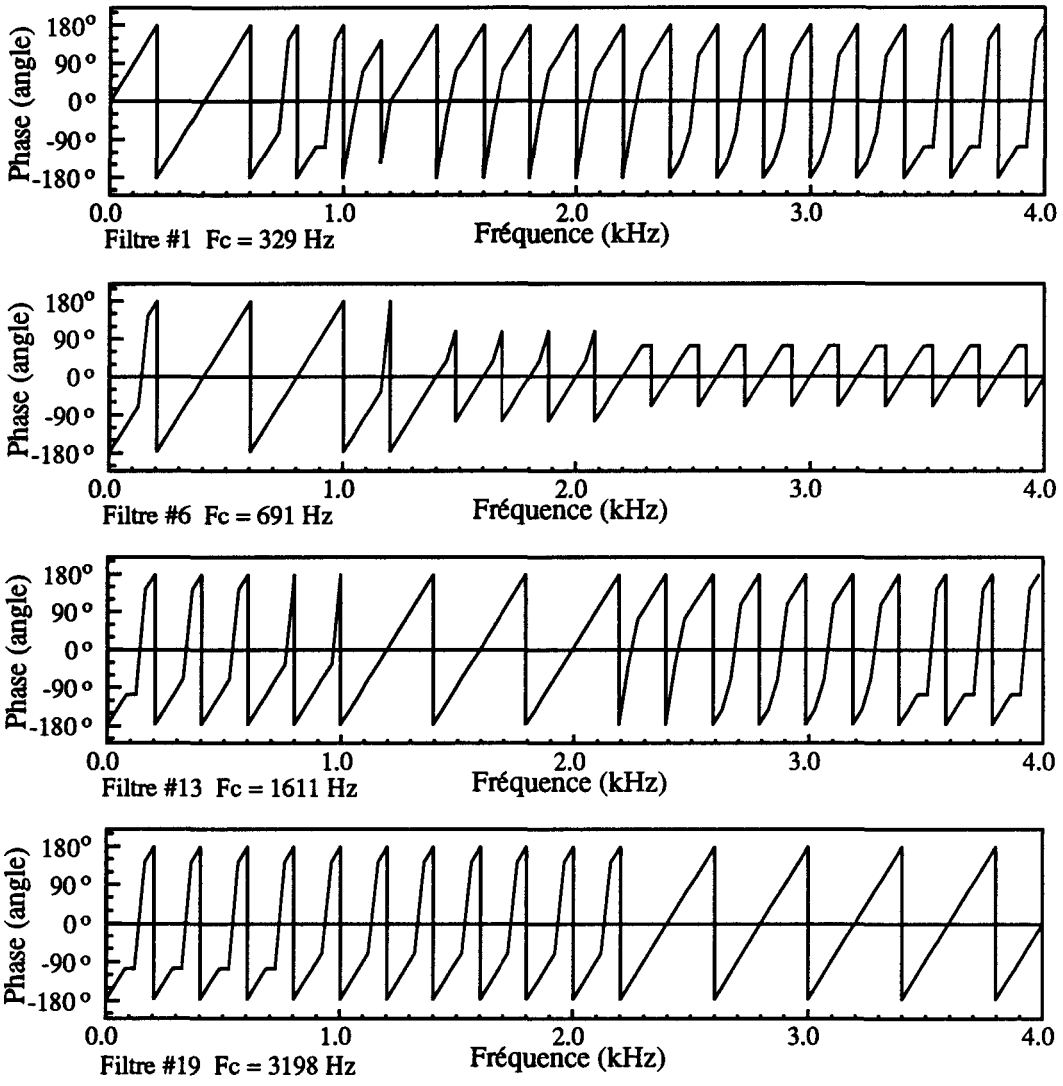

Figure 3.7 Réponses en phase de quelques filtres

au délai de propagation dans la cochlée. Et de plus, Patterson est arrivé à la conclusion que le délai de propagation peut être omis dans le modèle perceptif de l'audition. A partir des travaux de Patterson, nous avons supprimé ce type de délai dans la version finale de notre modèle. Actuellement, les performances générales du modèle proposé ne sont pas du tout influencées par le délai de propagation en raison des caractéristiques de l'algorithme utilisé. Ceci sera expliqué en détail au prochain chapitre.

Cependant, il faut considérer un autre délai: celui du filtre. A partir de la conception du filtre numérique, le délai du filtre réel est toujours égal à l'ordre du filtre. Donc, ce délai peut se corriger à l'aide de la formule suivante:

$$y(i) = y(i + l) \tag{3.18}$$

où $y(i)$ est la sortie d'un filtre au temps i et l est l'ordre du filtre.

3.3.5 Performances du banc de filtres auditifs

Dans ce paragraphe, on présente les performances du banc de filtres auditifs en observant les réponses des filtres à différents signaux.

A la figure 3.8, on présente la réponse impulsionnelle du banc de filtres auditifs. A partir de cette figure, on trouve que les réponses impulsionnelles dans les canaux inférieurs en fréquence sont plus longues que celles des canaux supérieurs en fréquence, et la fréquence d'oscillation dans les canaux inférieurs en fréquence est plus faible que celle des canaux supérieurs en fréquence. On trouve aussi que la durée des réponses impulsionnelles des canaux inférieurs en fréquence est autour de 4 ms. Notons que les saturations du pic de la réponse impulsionnelle observées dans certains canaux supérieurs en fréquence sont provoquées par le logiciel graphique et non pas par les filtres. La Réponse du banc de filtres à un sinus de 1 kHz est donnée à la figure 3.9. Ici la réponse maximale se trouve dans le canal pour lequel la fréquence centrale est la plus proche de la fréquence du signal du stimulus. A la figure 3.10, on présente les sorties du banc de filtres à un sinus pour lequel il y a deux changements d'amplitude du signal présenté. On peut voir à partir de cette figure que les canaux dont la fréquence centrale est loin de 1 kHz sont aussi très sensibles aux changements d'amplitude. Ceci signifie que le banc de filtres utilisé est capable de capturer les changements instantanés de l'enveloppe du

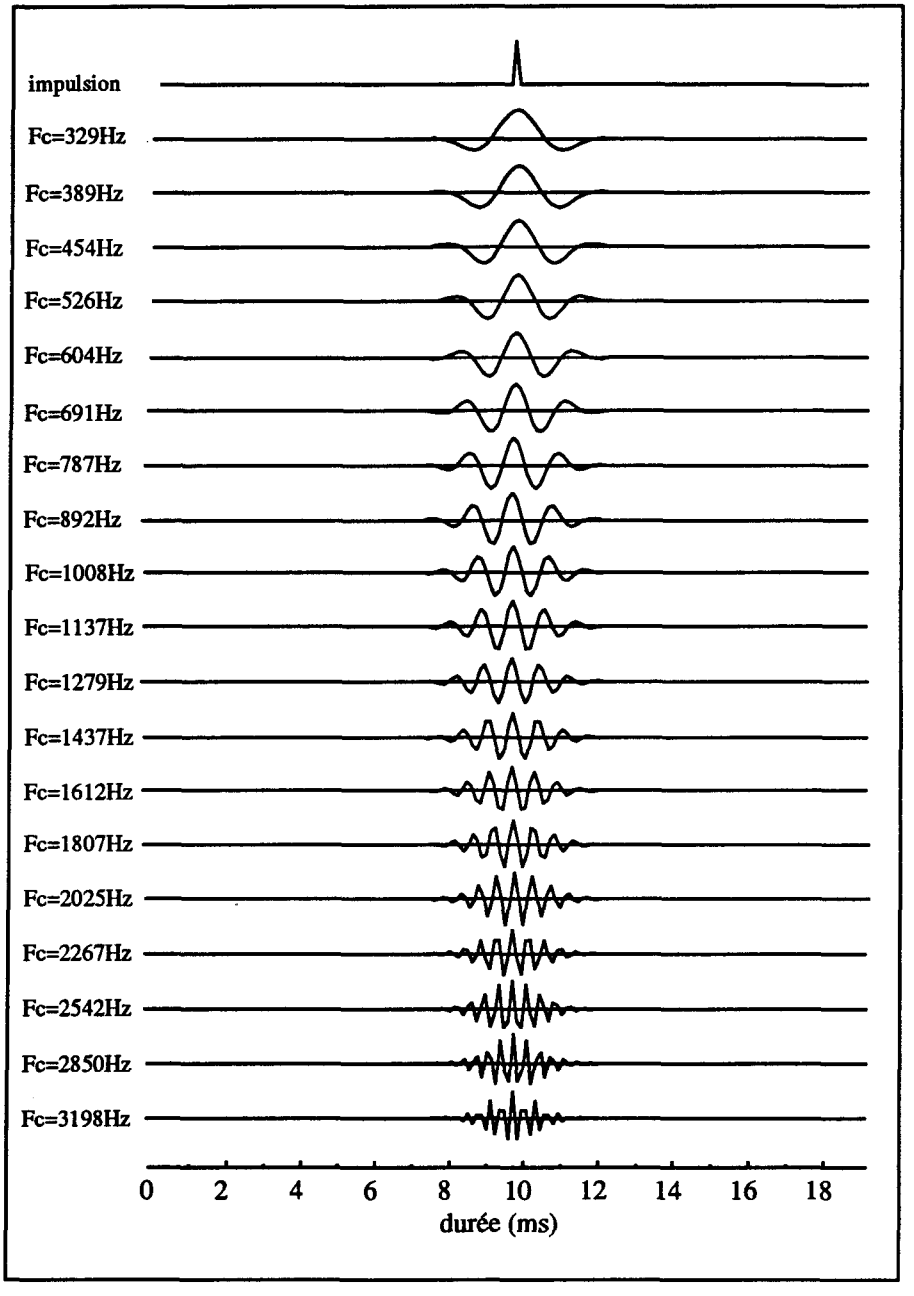

Figure 3.8 Réponse impulsionnelle du banc de filtres auditifs

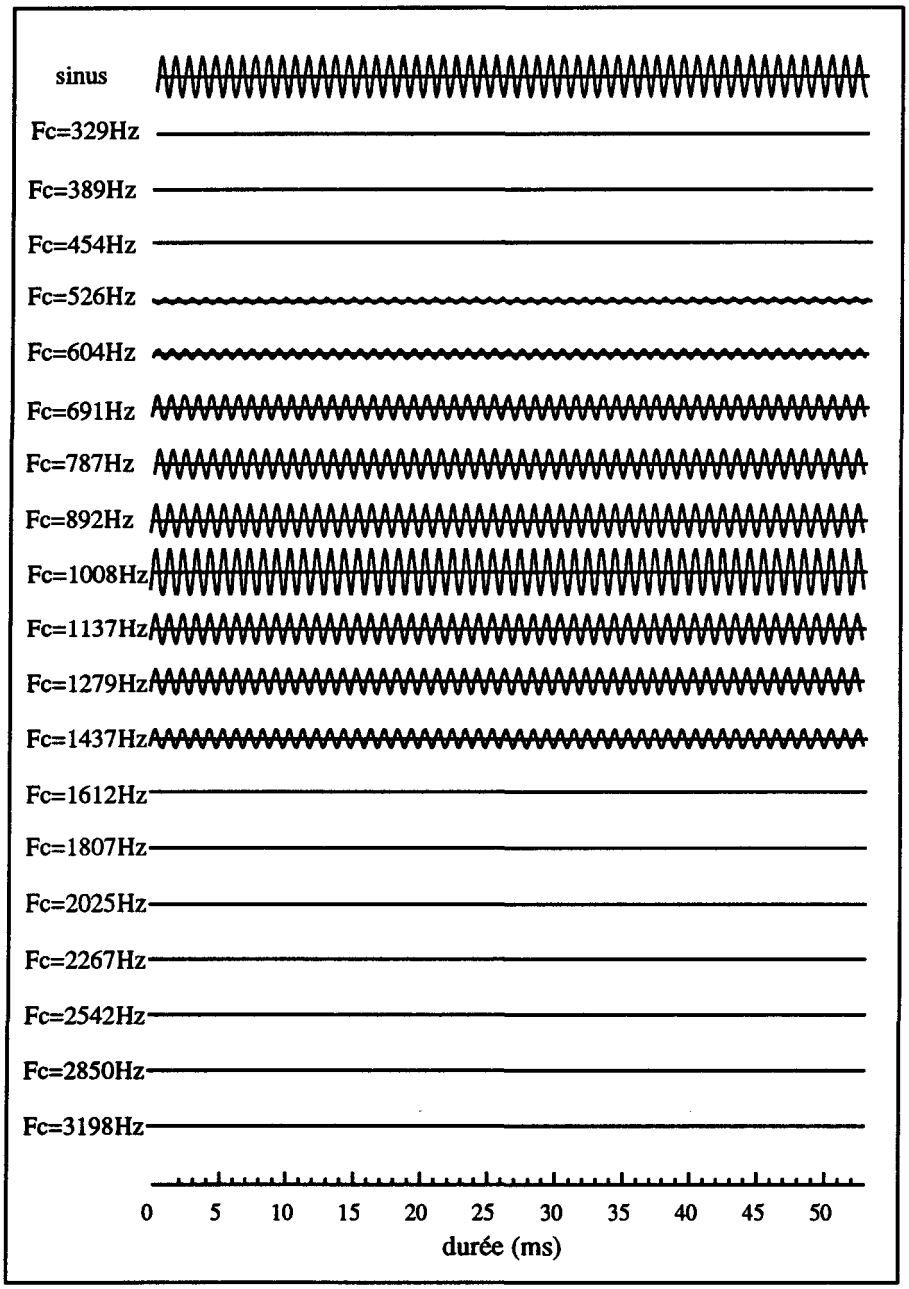

Figure 3.9 Réponse du banc de filtres auditifs à un sinus de 1kHz

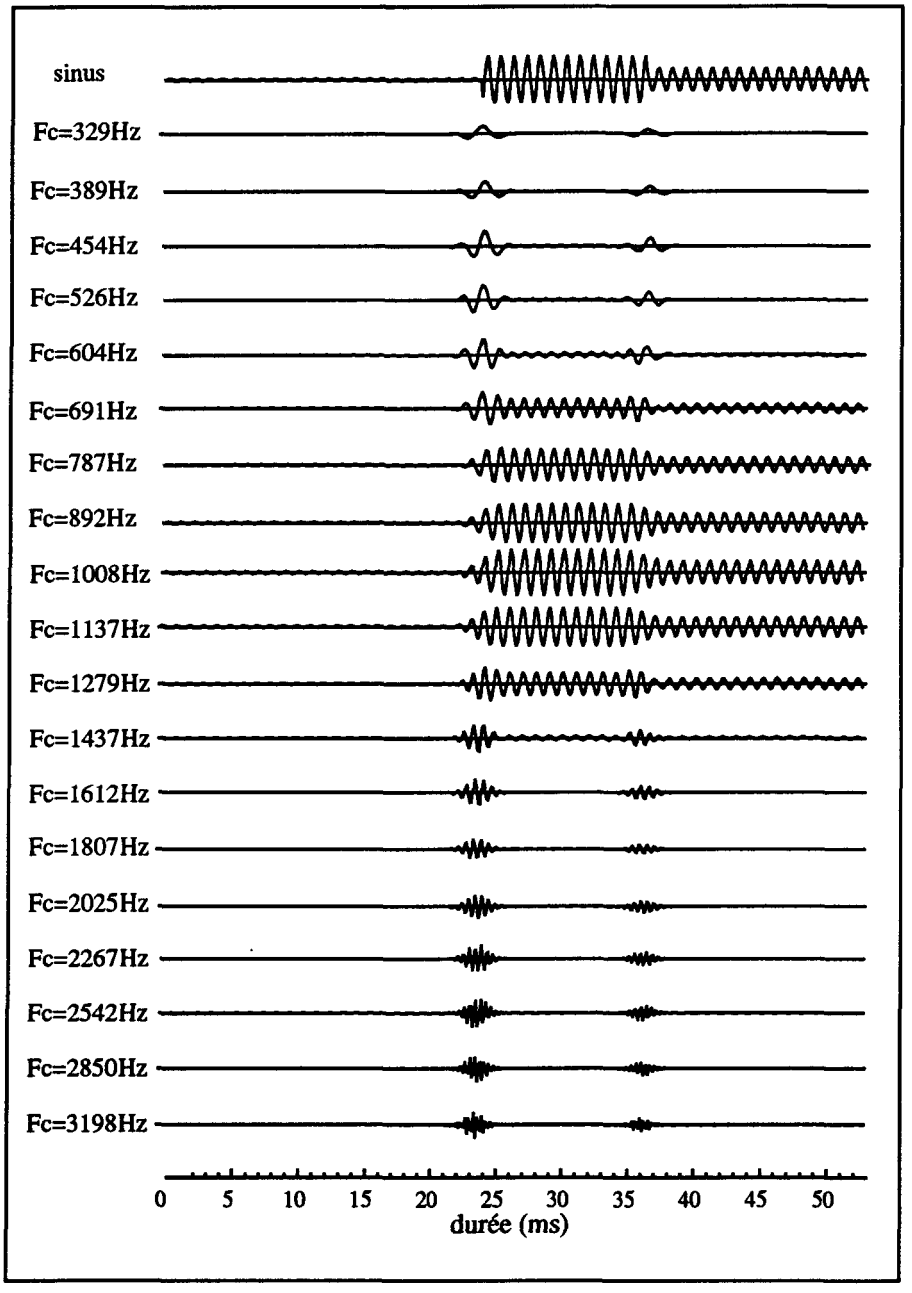

Figure 3.10 Réponse du banc de filtres auditifs à un sinus de 1kHz d'amplitude variable

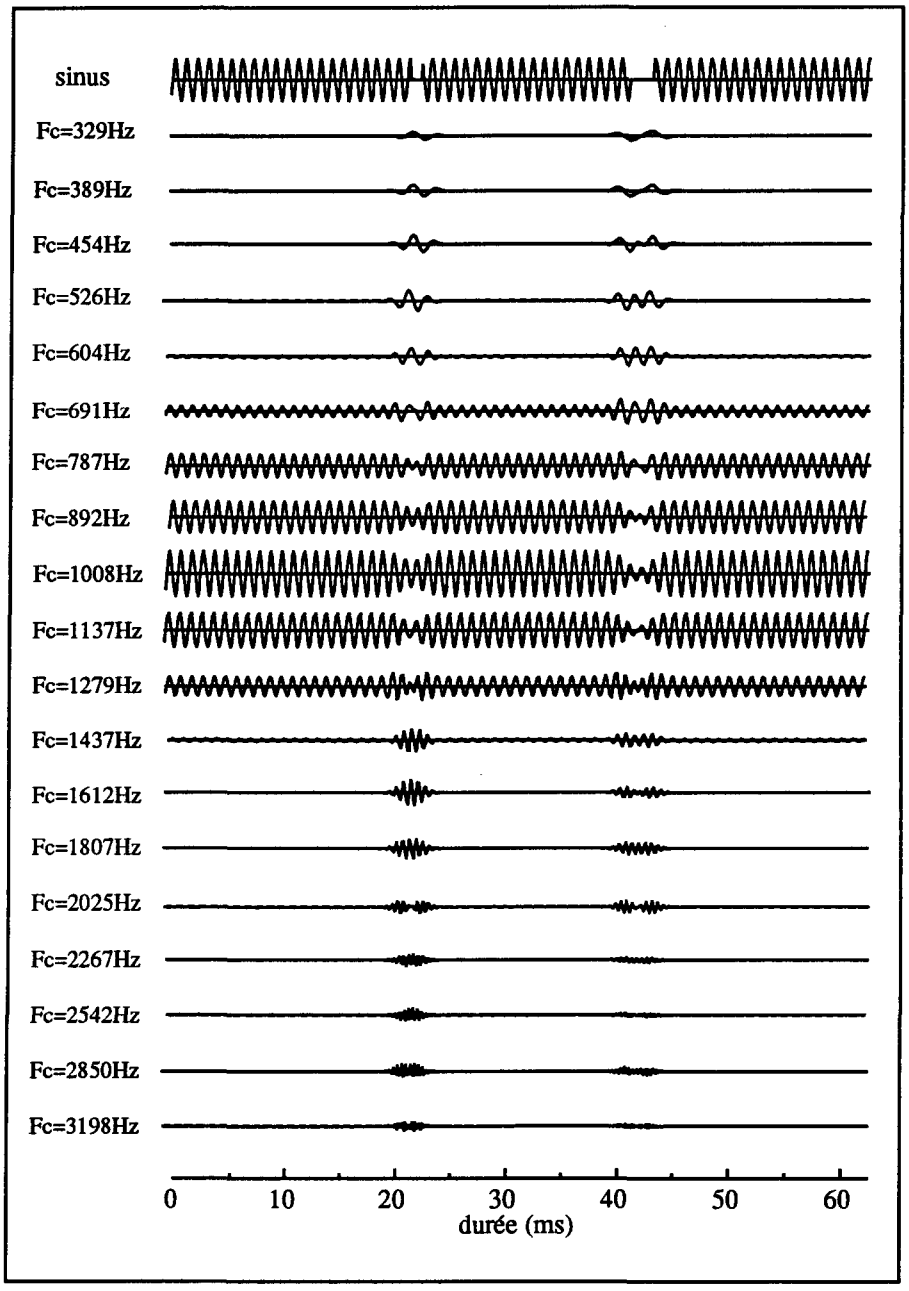

Figure 3.11 Réponse du banc de filtres auditifs à un sinus de 1kHz qui contient deux silences de durée égale à 1ms et 2ms respectivement

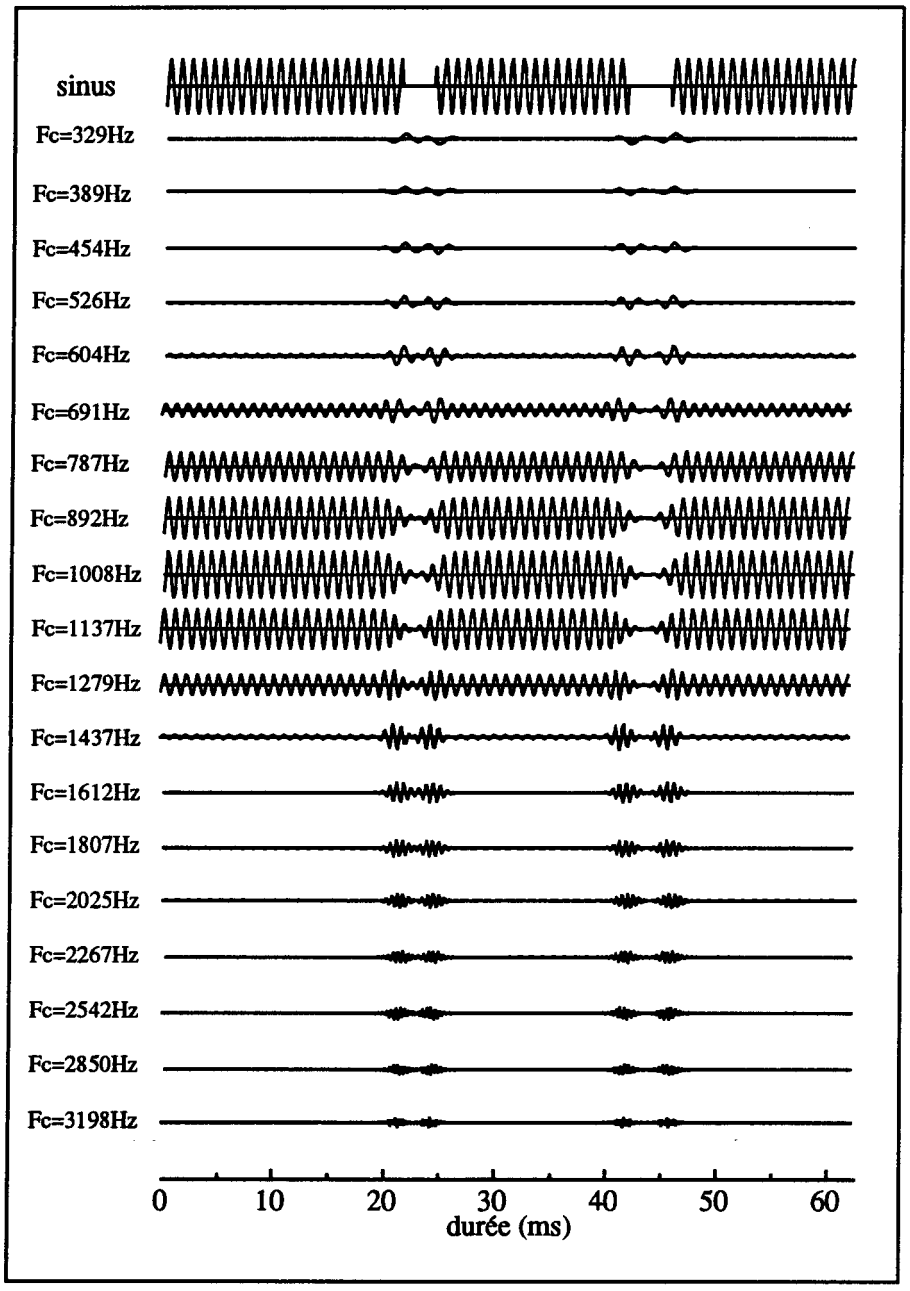

Figure 3.12 Réponse du banc de filtres auditifs à un sinus de 1kHz qui contient deux silences de durée égale à 3ms et 4ms respectivement

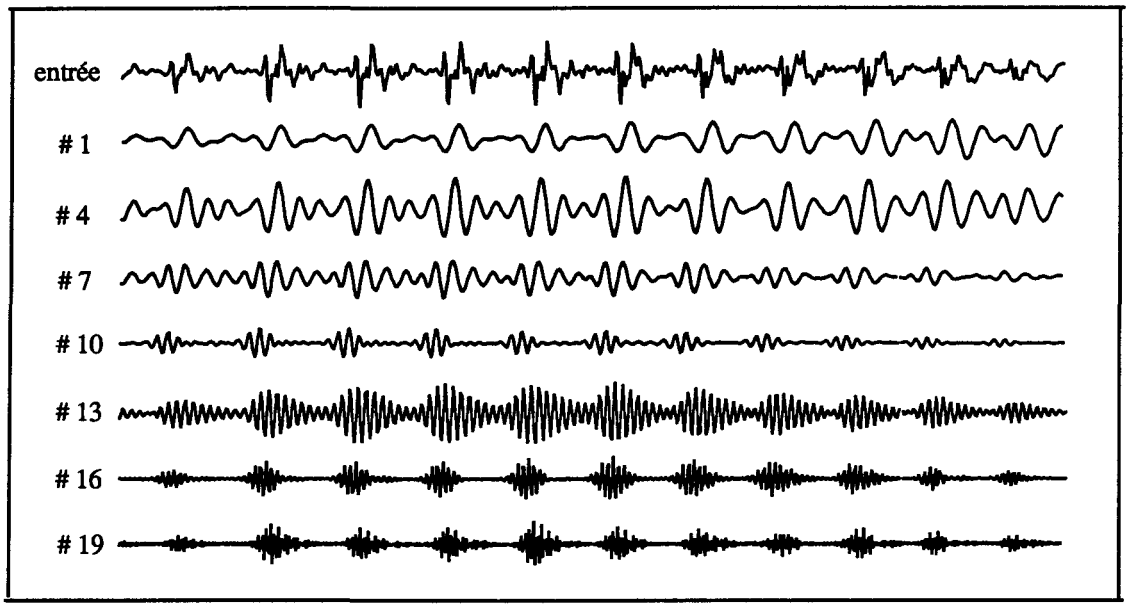

Figure 3.13 Réponse des filtres auditifs à une voyelle orale / a / de la parole téléphonique "ANNULER"

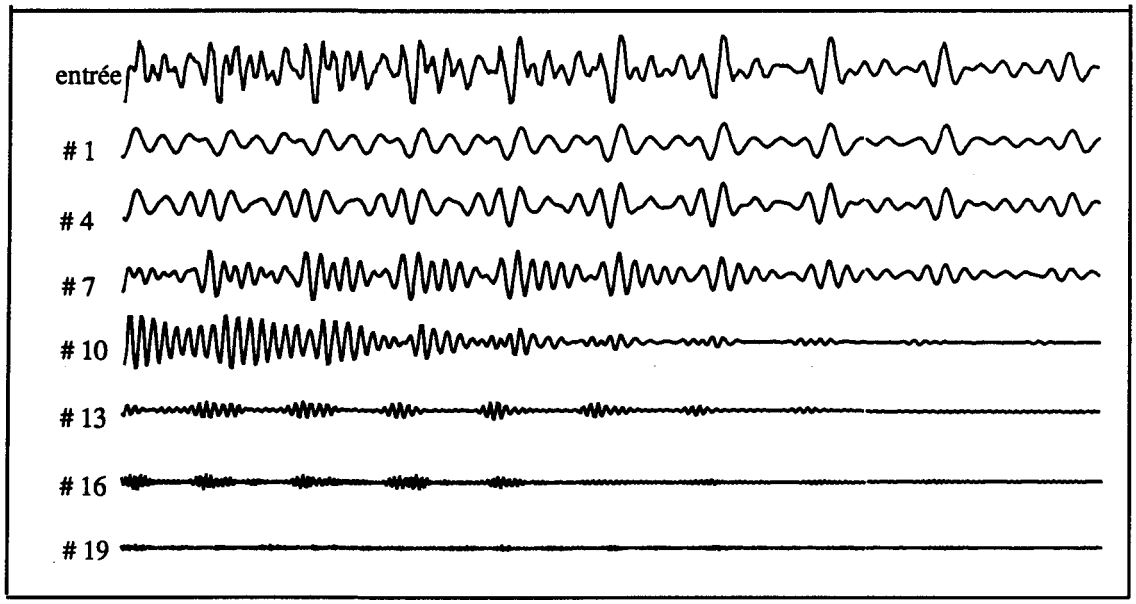

Figure 3.14 Réponse des filtres auditifs à une voyelle nasale / õ / de la parole téléphonique "NON"

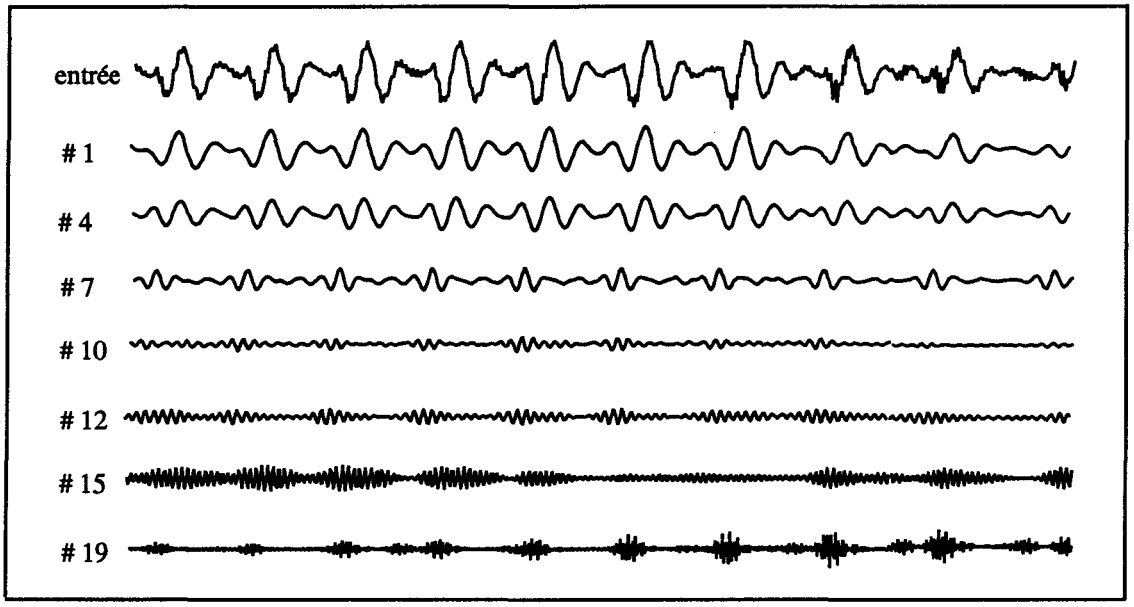

Figure 3.15 Réponse des filtres auditifs à une voyelle orale / e / de la parole téléphonique "ANNULER"

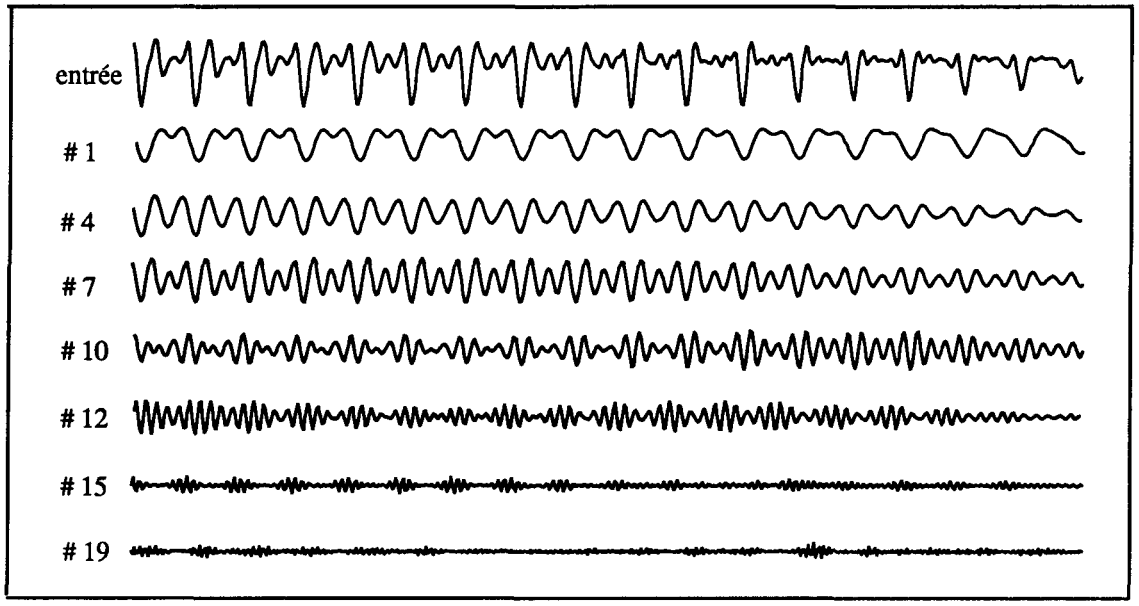

Figure 3.16 Réponse des filtres auditifs à une voyelle orale / o / de la parole téléphonique "ZÉRO"

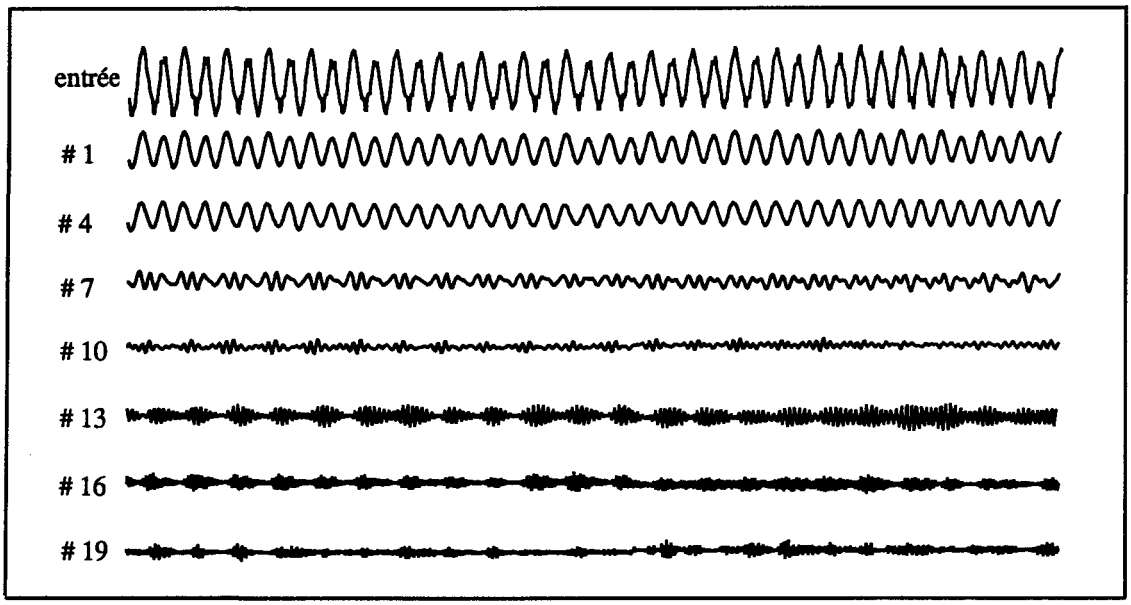

Figure 3.17 Réponse des filtres auditifs à une voyelle orale / ϕ / de la parole téléphonique "DEUX"

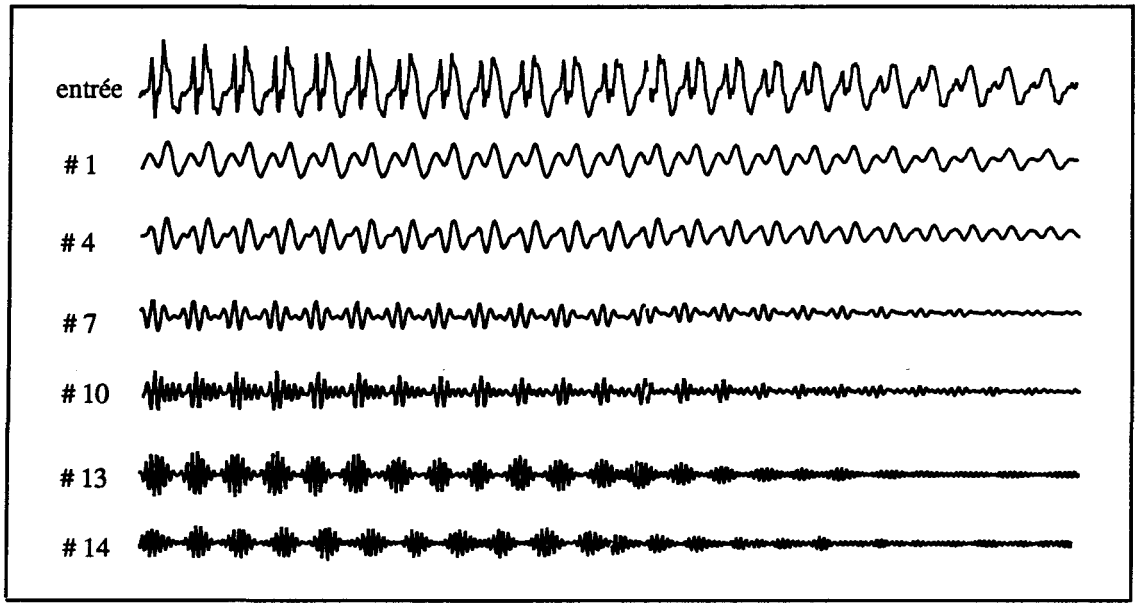

Figure 3.18 Réponse des filtres auditifs à une voyelle nasale / $\tilde{\epsilon}$ / de la parole téléphonique "UN"

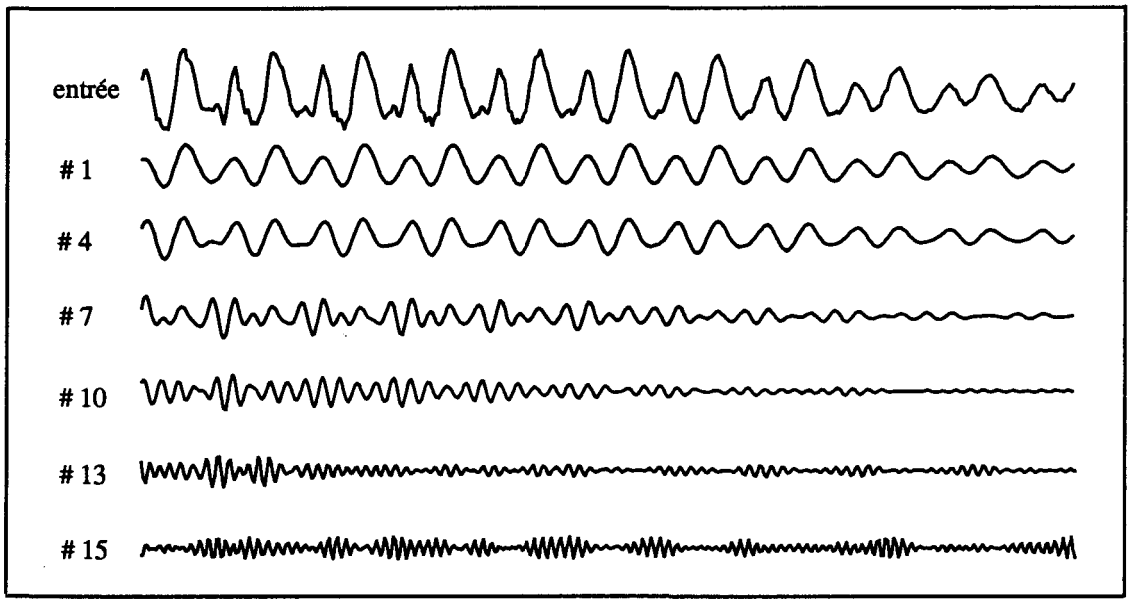

Figure 3.19 Réponse des filtres auditifs à une consonne occlusive voisée / g / de la parole téléphonique "ANGLAIS"

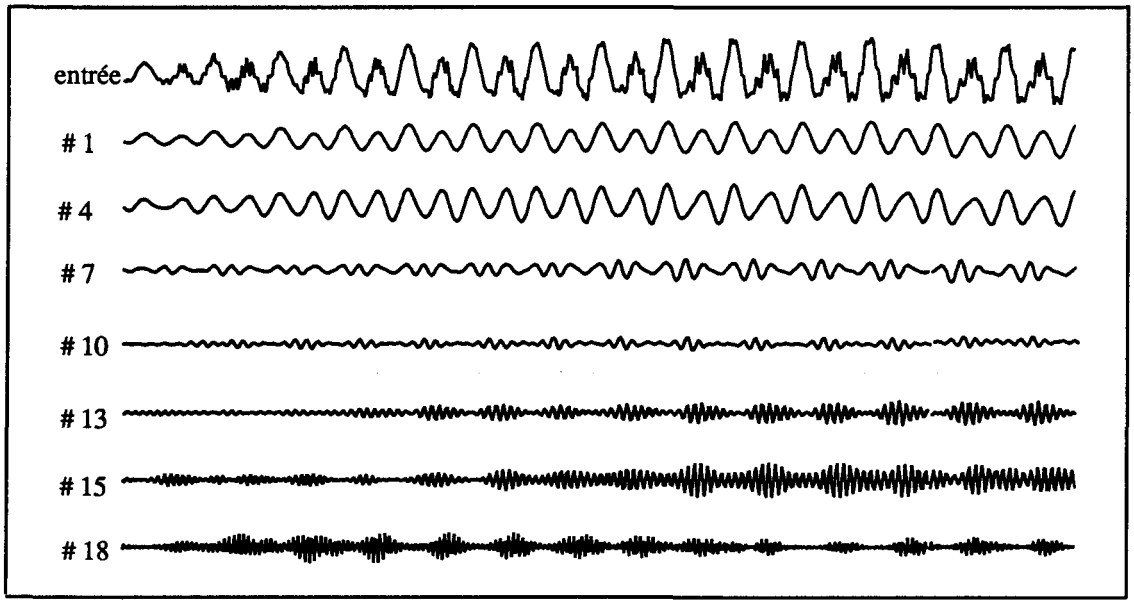

Figure 3.20 Réponse des filtres auditifs à une consonne liquide / *l* / de la parole téléphonique "ANGLAIS"

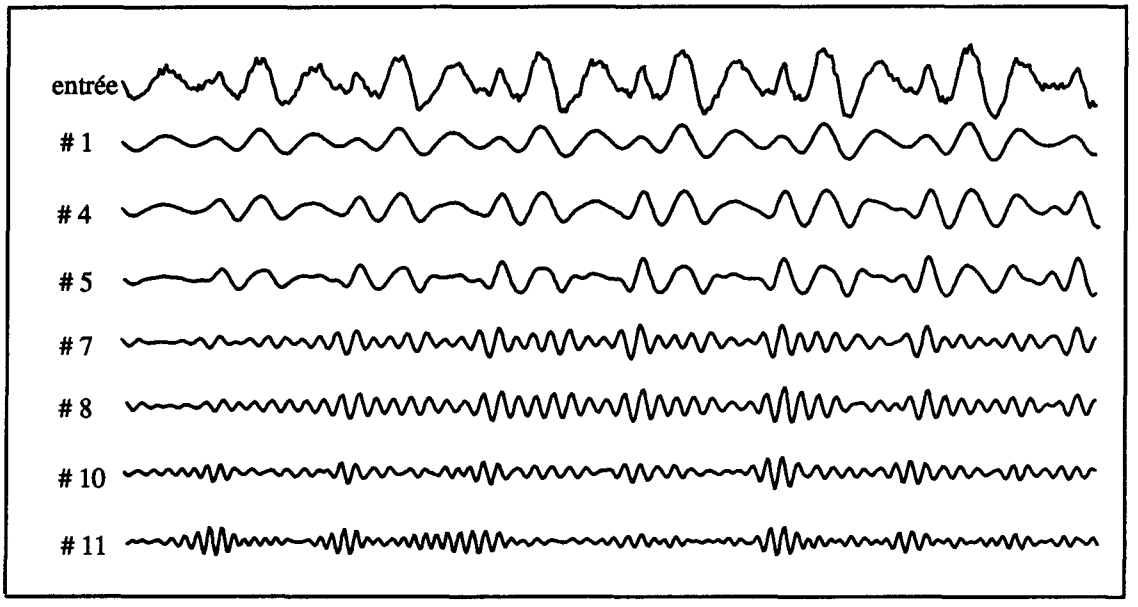

Figure 3.21 Réponse des filtres auditifs à une consonne nasale (suivie de la voyelle nasale) / n / de la parole téléphonique "NEUF"

signal du stimulus. Aux figure 3.11 et figure 3.12 nous présentons la réponse du banc de filtres à une fonction sinusoïdale interrompue par des silences. Plus spécifiquement, à la figure 3.11 le sinus contient des silences de 1 ms et 2 ms respectivement, et à la figure 3.12 le sinus contient des silences de 3 ms et 4 ms respectivement. De ces deux figures, on peut conclure que le banc de filtres auditifs est capable de détecter des silences de courte durée tout comme le système auditif est capable de le faire [38]. Ceci signifie que le banc de filtres est capable de garder la bonne information temporelle du signal d'entrée après que le signal ait été filtré.

De la figure 3.13 à la figure 3.23, on présente les sorties du banc de filtres pour les prononciations de voyelles et consonnes que nous avons pris à partir de données de parole téléphonique. Les figure 3.13–3.18 sont les réponses des filtres aux voyelles, et

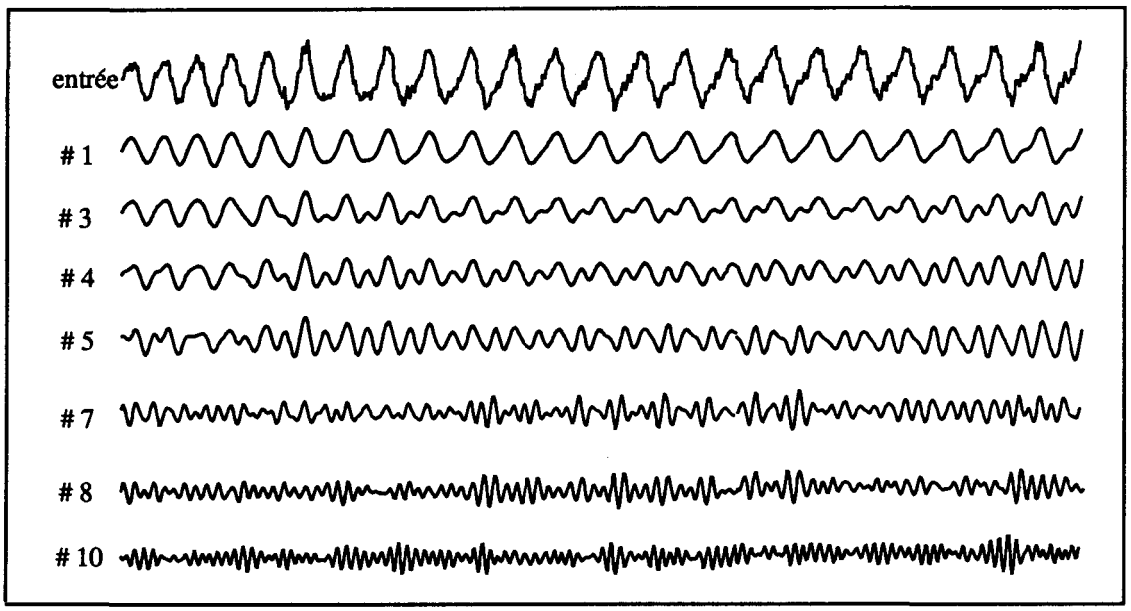

Figure 3.22 Réponse des filtres auditifs à une consonne glissante / μ / de la parole téléphonique "HUIT"

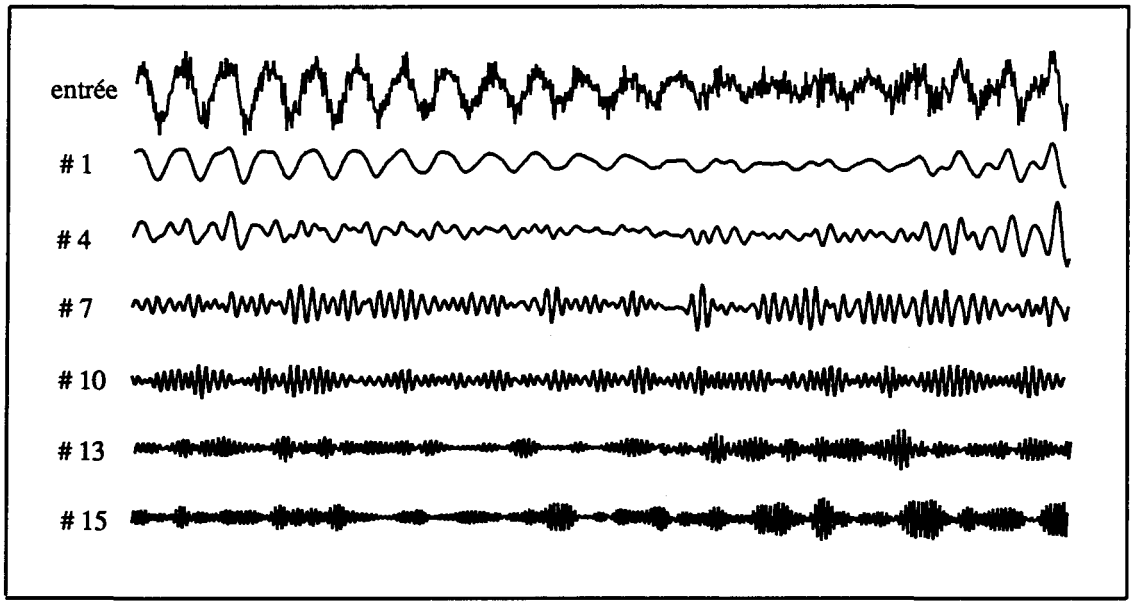

Figure 3.23 Réponse des filtres auditifs à une consonne fricative / z / de la parole téléphonique "ZÉRO"

les figure 3.19–3.23 sont les réponses des filtres aux consonnes. D'après ces figures, on peut observer que la plupart des canaux, après le filtrage, ont des patrons de répétition d'enveloppe. En d'autres termes, le signal est modulé en amplitude par la période de HT.

3.3.6 Conclusion

Le banc de filtres simule les déplacements de la membrane basilaire dans la cochlée. Chaque filtre caractérise la réponse d'un groupe de cellules. La fréquence centrale d'un filtre correspond à la fréquence caractéristique moyenne des cellules associées sur la membrane basilaire. Comme la forme du filtre est déterminée à partir d'expériences psycho-acoustiques et physiologiques, le filtre caractérise certaines propriétés auditives humaines.

Comme les filtres sont de type FIR ("finite impulse response"), ils ont une bonne résolution temporelle comme on a vu précédemment. Il est important que les filtres utilisés soient capables de capturer les changements instantanés de l'enveloppe du signal de parole, car beaucoup d'informations importantes dans la parole sont contenues aux lieux où le signal acoustique est variable plutôt qu'aux lieux où le signal est relativement stable. Il faut remarquer que beaucoup d'auteurs réalisent le banc de filtres de leurs modèles auditifs par des filtres IIR ("infinite impulse response") de phase non linéaire. Il nous semble qu'il est alors difficile de garder la bonne information temporelle du signal de parole même si le filtrage IIR est plus rapide que celui proposé.

Il faut noter ici qu'initialement nous avons considéré le délai de propagation sur la membrane basilaire de la cochlée. Ce délai a été supprimé dans la version finale de notre modèle global, car l'algorithme d'extraction de HT du modèle proposé est insensible aux changements de phase.

3.4 Sous-modèle fonctionnel

Le sous-modèle fonctionnel est en fait un modèle mathématique pour lequel la sortie d'un groupe de cellules associées sur la membrane basilaire est traitée de façon non linéaire en utilisant l'auto-corrélation. Ensuite elle est combinée avec d'autres canaux pour produire "le pseudo-histogramme périodique" comme entrée au traitement final. Aux paragraphes suivants, nous allons détailler chacune des quatre parties du sous-modèle fonctionnel: rectification, multiplication, autocorrélation et combinaison.

3.4.1 Redressement

Il est bien connu que les activités spontanées des fibres sont sensibles à la direction positive du déplacement des cellules depuis que Rose a présenté ce phénomène intéressant [52]. Pour simuler ce phénomène, tous les auteurs de modèles auditifs périphériques utilisent un redressement mono-alternance. Le redressement le plus simple est proposé par Rose [53] pour lequel l'entrée négative est simplement mise à zéro et l'entrée positive est gardée inchangée. Pour simplifier le calcul, cette forme simple de redressement est utilisée dans notre modèle, et elle est définie par l'équation suivante:

$$y(t) = \begin{cases} x[t] & x[t] \geq 0 \\ 0 & x[t] < 0 \end{cases} \tag{3.19}$$

où $x[t]$, $y[t]$ sont respectivement l'entrée et la sortie du redressement mono-alternance.

Nous présentons deux exemples à la figure 3.24 (c) et à la figure 3.25 (c) pour expliquer le processus du redressement mono-alternance.

3.4.2 Multiplication

La prochaine opération du sous-modèle fonctionnel est la multiplication entre deux signaux: le signal redressé d'entrée $x[n]$ et le signal redressé de sortie d'un filtre $y_i[n]$

où *i* est le numéro du canal. Cette opération est un processus non linéaire proposé à l'origine par Girija [18]. Le but de la multiplication ici est d'améliorer la structure périodique du signal après le redressement, car l'enveloppe du signal à la sortie du filtre est très similaire à celle du signal d'entrée lorsque cette dernière est de la parole voisée. D'après nos expériences, le modèle avec la multiplication ne fonctionne bien que sur de la parole non bruitée. C'est pourquoi pour la parole bruitée nous avons proposé un autre sous-modèle fonctionnel pour lequel la multiplication a été supprimée et remplacée par un autre traitement. Nous allons le présenter à la section 3.5. A la figure 3.24 (d) et à la figure 3.25 (d), nous avons présenté deux exemples pour illustrer les effets de la multiplication sur la parole téléphonique.

3.4.3 Auto-corrélation

Pour estimer la période du fondamental dans chaque canal, nous avons pris la fonction d'auto-corrélation qui s'est avérée être la plus robuste à tous les types de signal de parole. Dans la pratique, la fonction d'auto-corrélation est réalisée en prenant une fenêtre d'analyse (la largeur de fenêtre du signal peut être ajustée) et en la multipliant avec une version d'elle-même, mais décalées de façon temporelle. Le maximum de la fonction d'auto-corrélation est obtenu en principe lorsque le décalage entre le signal original et le signal décalé est égal à une période du signal. Deux exemples d'auto-corrélation sont présentés à la figure 3.24 (e) et à la figure 3.25 (e) respectivement.

3.4.4 Combinaison des canaux

Après le calcul de la fonction d'auto-corrélation pour chaque canal, nous combinons tous les canaux en sommant les sorties des fonctions d'auto-corrélation fenêtre à fenêtre. Ici nous avons emprunté le terme de "pseudo-histogramme périodique", introduit par

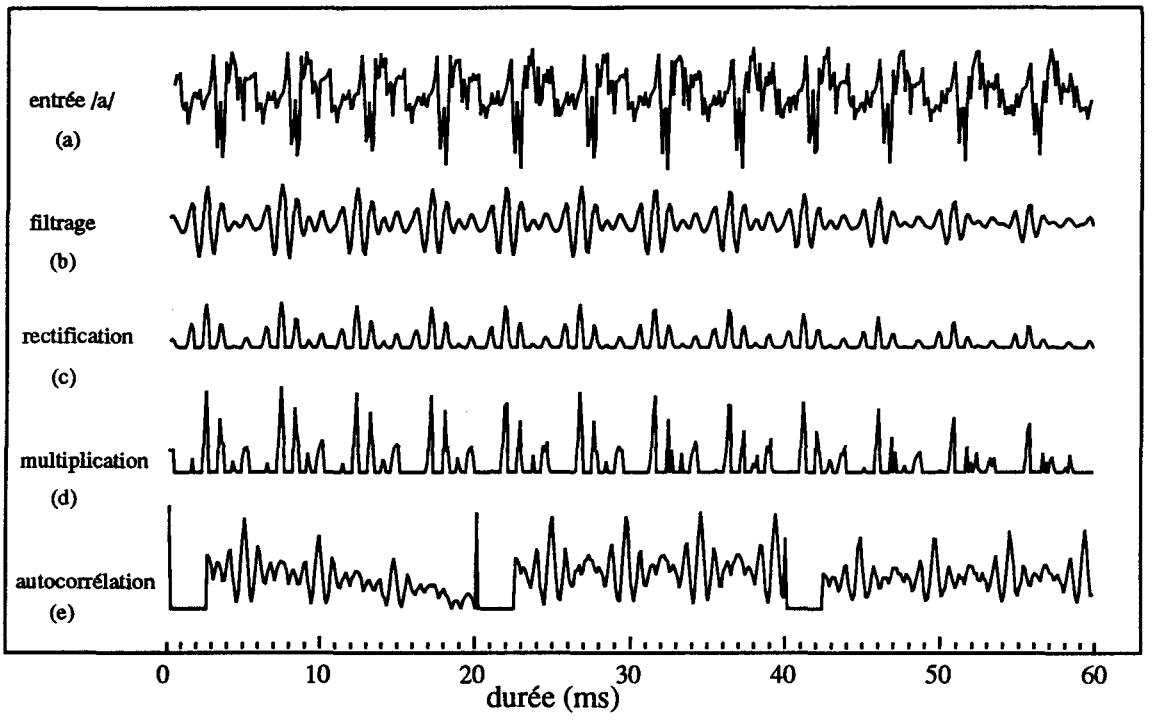

Figure 3.24 Sorties intermédiaires du canal #10 (F_c = 1136 Hz) pour la prononciation / a / de la parole téléphonique "ANNULATION"

Seneff [59], pour représenter la somme d'auto-corrélation. Au cours de nos expériences, nous avons trouvé que, lorsque l'entrée est périodique, les fonctions d'auto-corrélations de tous les canaux montrent un pic au même délai dans les fenêtres présentes, et généralement ce délai correspond à la période de HT perçue. Actuellement, cette observation est complètement conforme à la théorie proposée par Licklider [28].

Puisque les grands pics dans les fonctions d'auto-corrélation sur les différents canaux arrivent au même délai, la somme des canaux permet de rehausser le pic de "synchronisation". Nul doute que ceci va faciliter la localisation du pic qui correspond à la période de HT au traitement final. Un exemple de la combinaison des canaux est

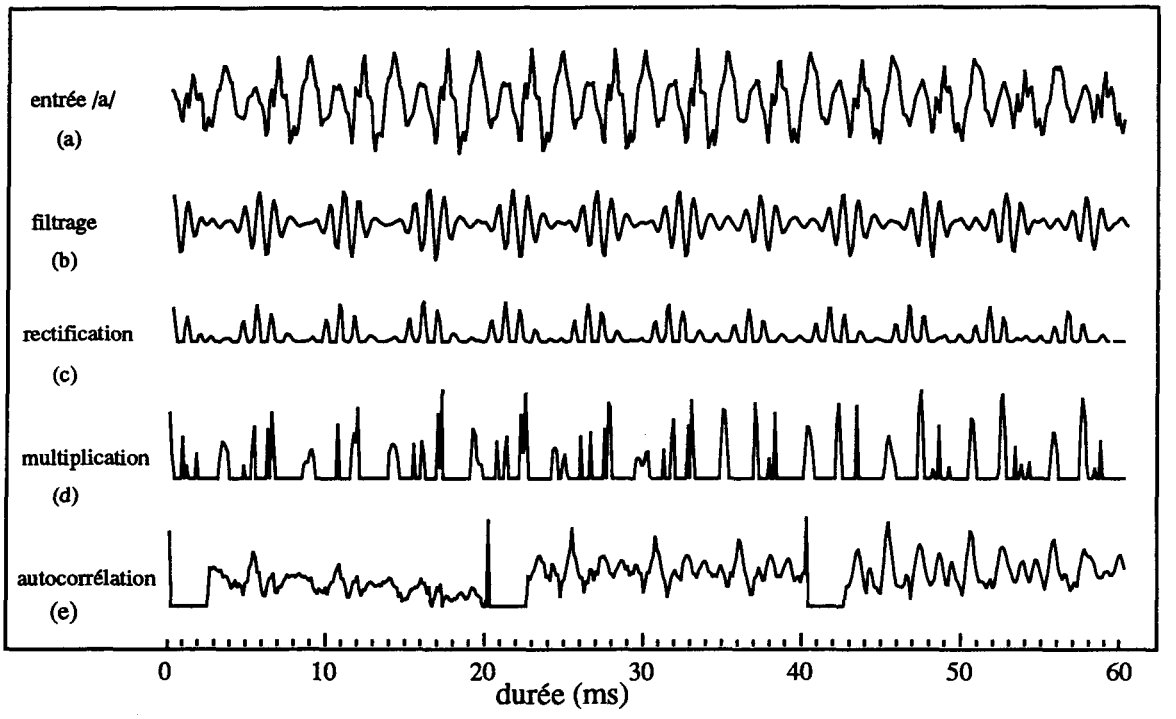

Figure 3.25 Sorties intermédiaires du canal #10 (F_c= 1136 Hz)
pour la prononciation / ϵ / de la parole téléphonique "ANGLAIS"

expliqué schématiquement par la figure 3.26.

Sur la figure 3.26, on peut voir que, d'une part, la combinaison rehausse le pic de synchronisation par rapport aux canaux, d'autre part, le procédé linéaire de somme est actuellement capable de reproduire le phénomène de la dominance des composantes du signal de parole. Par exemple, le canal 1 (F_c=329.27 Hz) contient plus d'énergie que les autres canaux et il a une dominance par rapport aux autres canaux, en conséquence, la combinaison linéaire lui permet de contribuer plus au pic final dans le pseudo-histogramme périodique que les autres canaux.

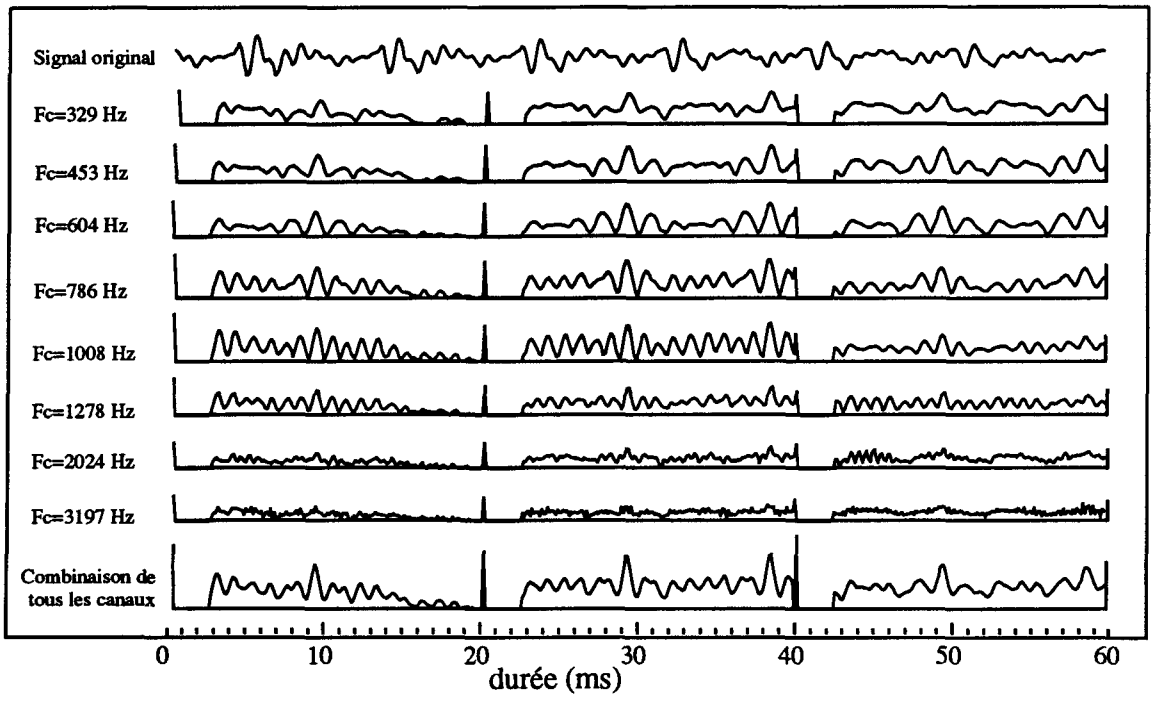

Figure 3.26 Combinaison des canaux. L'entrée est une portion (3 fenêtres) d'une prononciation / ã / de la parole téléphonique "COMMANDE"

3.5 Sous-modèle fonctionnel pour la parole bruitée

Comme nous l'avons souligné à la section 3.4.2, l'opération de multiplication entre la sortie redressée du filtrage et la version rectifiée d'entrée originale ne fonctionne plus sur de la parole bruitée. La raison est que, pour le canal qui contient le moins de bruit, la structure périodique est nettement meilleure que celle d'entrée qui est bruitée. En conséquence, la multiplication détériorera cette bonne structure périodique existante. Ainsi, pour la parole bruitée, nous avons enlevé la multiplication et ajouté un autre mécanisme appelé "sélection" qui est capable de décider si la sortie de la fonction d'auto-corrélation d'un canal dans la fenêtre considérée doit être combinée ou pas. Autrement dit, la sélection fonctionne comme un commutateur: il ferme quand le signal

dans la fenêtre présente est périodique; il est ouvert quand le signal n'est pas périodique. Évidemment, le sous-modèle fonctionnel dans ce cas ne combine que les canaux pour lesquels les signaux sont périodiques. Plus précisément, les canaux dont les signaux sont apériodiques ou corrompus par le bruit, ne sont pas périodiques, ne sont pas combinés dans la fenêtre présente. La structure du sous-modèle fonctionnel pour la parole bruitée a été présentée à la figure 3.3.

3.6 Décision de HT

Le dernier module du modèle proposé est la décision de HT. Ce module comprend l'estimation de la période de HT et le post-traitement. Nous allons les présenter respectivement dans ce paragraphe.

3.6.1 Estimation de la période de HT sur le pseudo-histogramme périodique

Habituellement, la période de HT est le délai du temps associé au plus grand pic dans le pseudo-histogramme périodique. Malheureusement, dans les situations pour lesquelles le signal est parfaitement périodique, ou lorsque la HT est élevée, il y a quelques pics dont les hauteurs sont égales à celles des multiples de la période fondamentale. La figure 3.27 et la figure 3.28 montrent un exemple de ce cas. De plus, dans le pseudo-histogramme périodique, le pic correspondant à la plus haute harmonique de HT ou correspondant à la plus basse harmonique de HT est probablement de plus grande amplitude, dues aux déviations de la périodicité. Ces déviations sont causées par le bruit ou la distorsion du signal sur la ligne téléphonique. Donc, la tâche principale ici est de trouver un pic p_t dans le pseudo-histogramme périodique à condition que l'amplitude du pic p_t soit proche de ou égale à celle du plus grand pic dans ce pseudo-

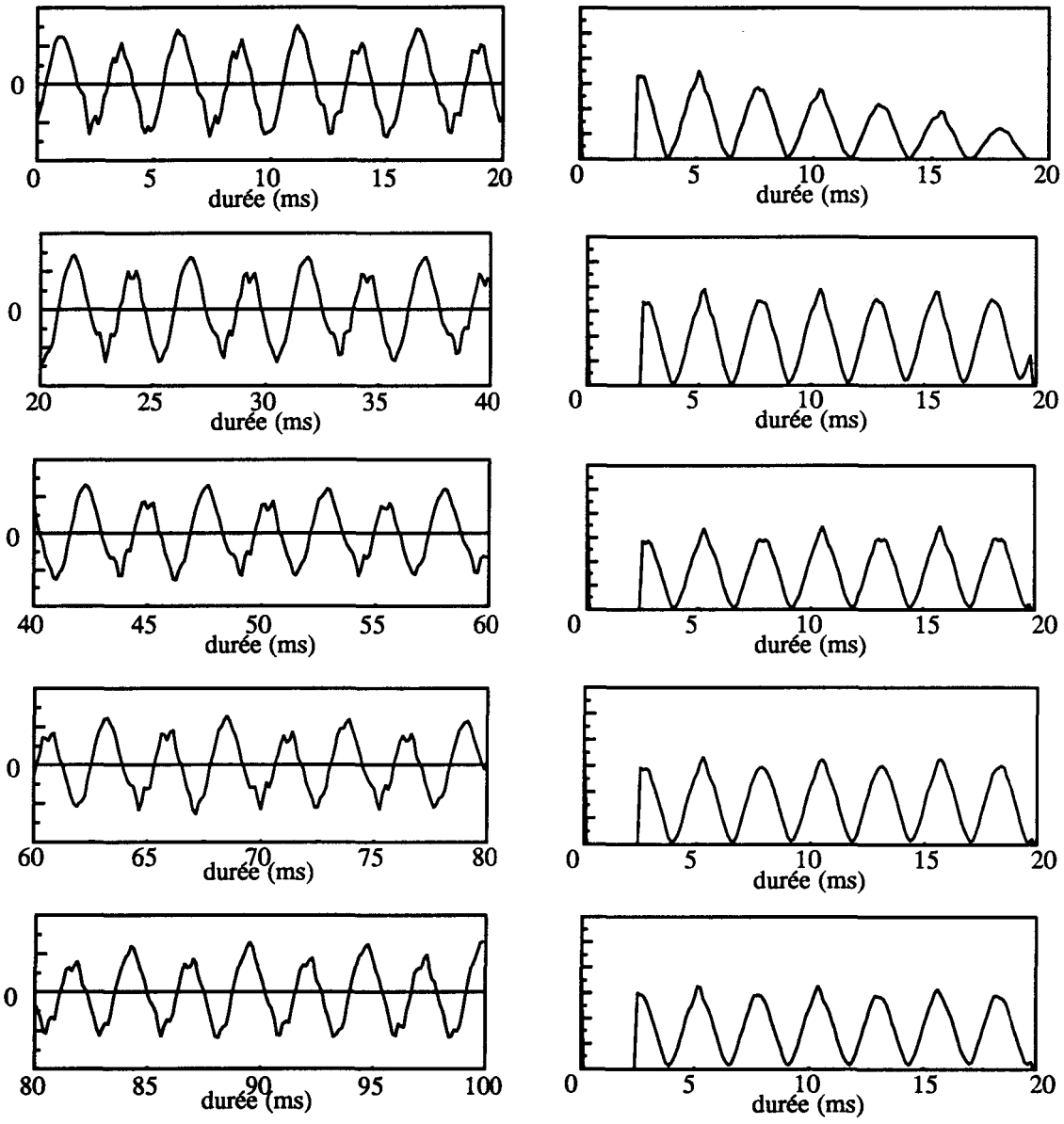

Figure 3.27 Plusieurs pseudo histogrammes périodiques (à droite) du signal d'entrée de prononciation / *dϕ* / de la parole téléphonique "DEUX" (à gauche)

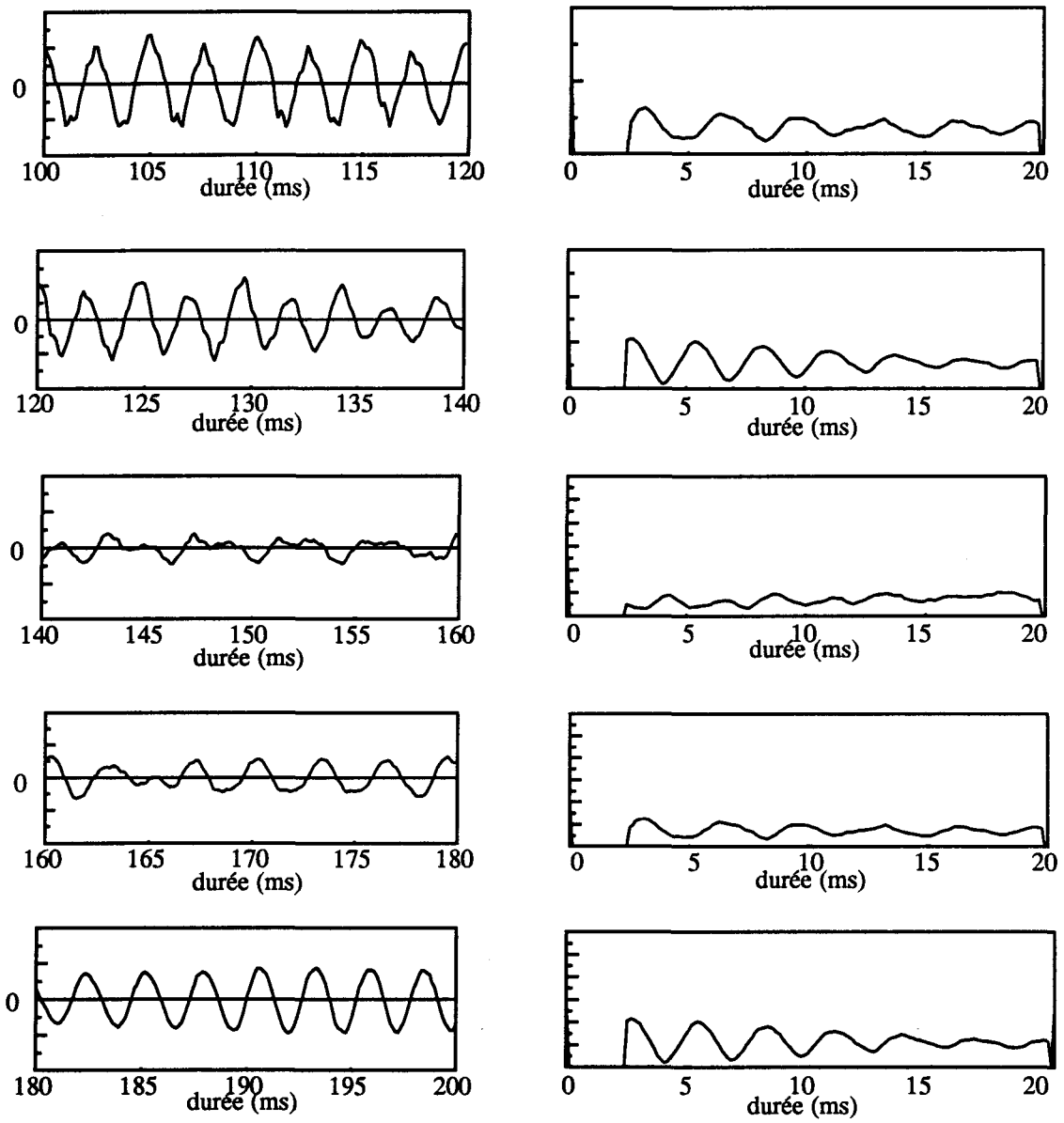

Figure 3.28 (suivie de la figure3.27) Plusieurs pseudo histogrammes périodiques (à droite) du signal d'entrée de prononciation / $d\phi$ / de la parole téléphonique "DEUX" (à gauche)

histogramme périodique, et qu'il y a au moins un autre pic existant à un instant multiple de t (avec t, l'instant associé à p_t).

Une stratégie pratique qui s'est avéré bien fonctionner dans nos expériences est la suivante (voir la figure 3.29):

(1) Trouver le premier plus grand pic entre le début et le centre dans la fenêtre considérée du pseudo-histogramme périodique, noter le décalage du pic par τ_{1P}.

(2) Trouver le plus grand pic entre τ_{1P} et la fin de la fenêtre, et noter le décalage du pic par τ_{1PL}, si τ_{1PL} est un multiple de τ_{1P}, accepter τ_{1P} comme la période de HT, sinon, continuer.

(3) Trouver le deuxième plus grand pic entre le début et le milieu de la fenêtre, noter le décalage du pic par τ_{2P}.

(4) Trouver le plus grand pic entre τ_{1P} et la fin de la fenêtre, noter le décalage du pic par τ_{2PL}, si τ_{2PL} est un multiple de τ_{2P}, accepter τ_{2P} comme la période de HT, sinon, continuer.

(5) Vérifier si il y a un pic à $2\tau_{1P}$, si oui, accepter τ_{1P} comme la période de HT; sinon, vérifier si il y a un pic à $2\tau_{2P}$, si oui, accepter τ_{2P} comme la période de HT; sinon, donner zéro à la HT de cette fenêtre.

Dans la pratique, nous avons réalisé l'estimation de la période de HT à partir du pseudo-histogramme périodique à l'aide de deux seuils pour détecter les segments de silence et pour prendre la décision reliée au voisement. Ces deux seuils S_{AMP} et S_{RAP}, sont définis comme suit:

$$S_{AMP} = PSP(\tau_P), \qquad S_{RAP} = \frac{PSP(\tau_0)}{PSP(\tau_P)} \tag{3.20}$$

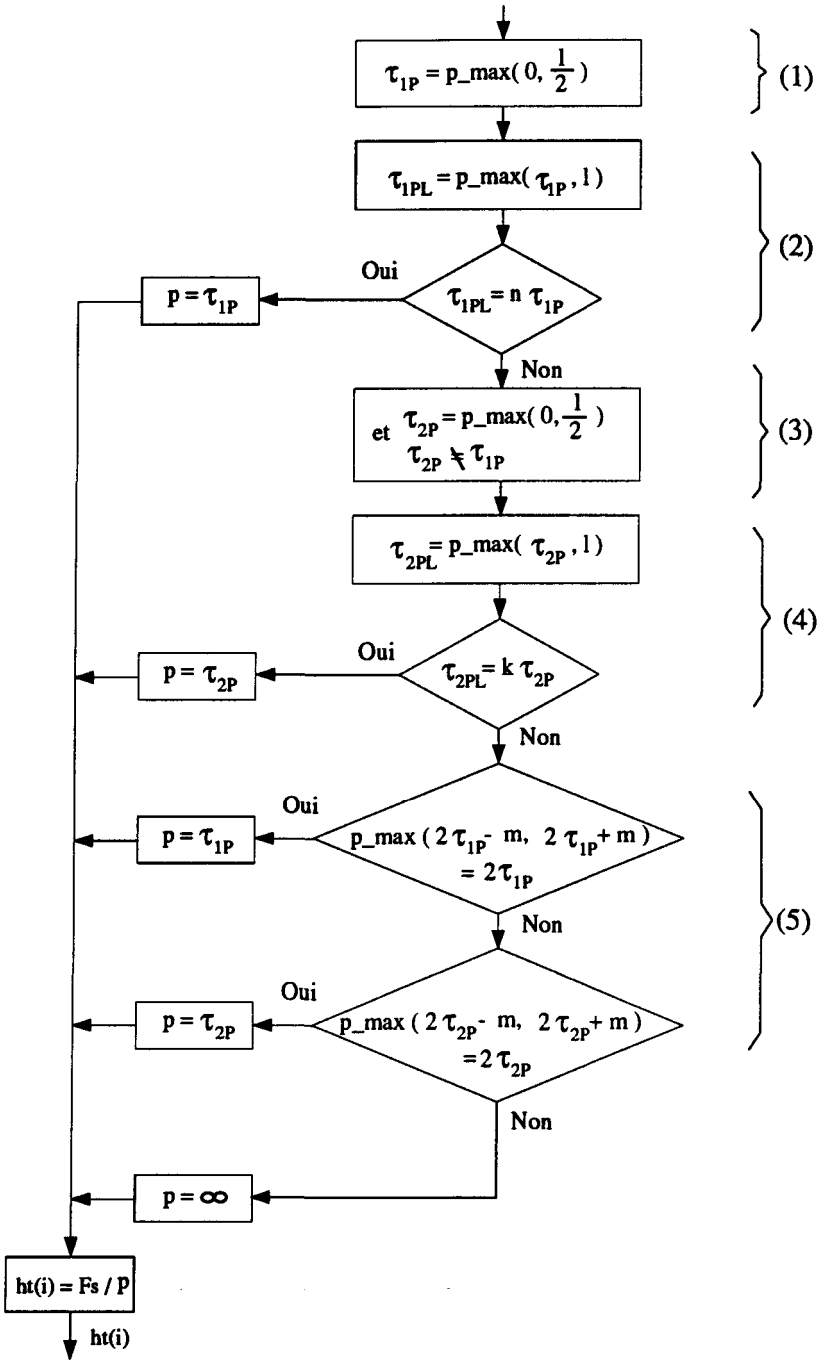

Figure 3.29 Processus de l'estimation de la période de la HT

où $PSP(\tau_0)$ est l'amplitude du pic au délai zéro dans le pseudo-histogramme périodique et $PSP(\tau_p)$ est l'amplitude du pic le plus grand au délai τ_p. Évidemment, S_{AMP} est faible lorsque le segment traité est du silence ou est non voisé. Par contre, quand le

segment traité est voisé, S_{AMP} est grand et S_{RAP} est petit. D'après l'expérience, nous avons choisi $S_{AMP} = 1000$ et $S_{RAP} = 3.00$.

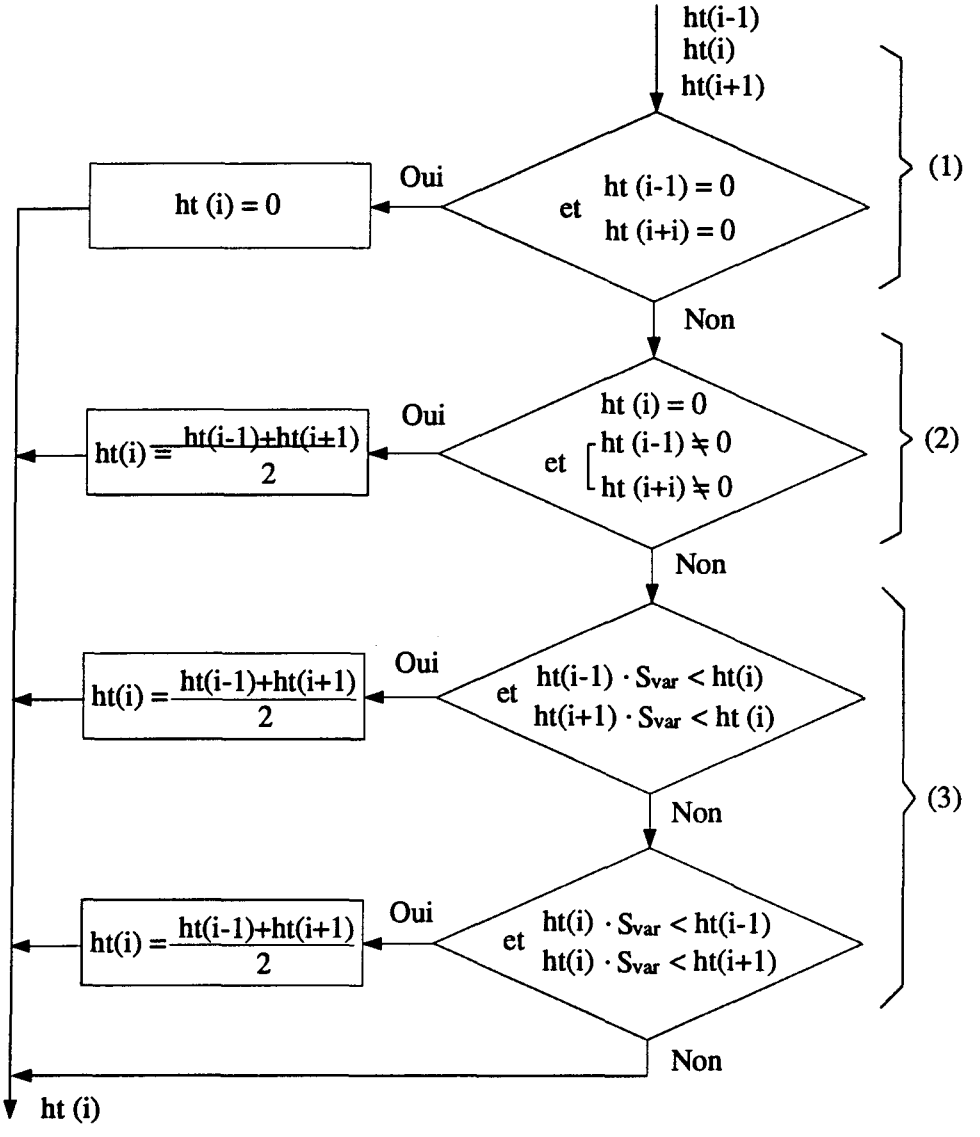

Figure 3.30 Processus du post-traitement

3.6.2 Post-traitement

Le but du post-traitement que nous avons effectué est de corriger les erreurs dans l'estimation de la valeur de HT. Celle-ci peut être trop élevée ou trop basse par rapport

à la valeur de HT réelle. L'algorithme de post-traitement se réalise en comparant la HT de la fenêtre considérée avec celle des fenêtres voisines.

Supposant que $ht(i-1)$ et $ht(i+1)$ sont les hauteurs tonales de trois fenêtres consécutives, alors, l'idée du post-traitement peut se résumer de la façon suivante (voir la figure3.30):

(1) Si $ht(i-1) = 0$, et $ht(i+1) = 0$, remplacer $ht(i)$ par zéro;

(2) Si $ht(i) = 0$, mais $ht(i-1) <> 0$ et $ht(i+1) <> 0$, remplacer ht(i) par $[ht(i-1) + ht(i+1)]/2$;

(3) Si $ht(i-1) \cdot S_{var} < h(i)$ et $ht(i+1) \cdot S_{var} < ht(i)$, ou $ht(i-1) > ht(i) \cdot S_{var}$ et $ht(i+1) > ht(i) \cdot S_{var}$, remplacer $ht(i)$ par $[ht(i-1) + ht(i+1)]/2$. S_{var} est le rapport de variation acceptable de HT entre deux fenêtres adjacentes.

3.7 Conclusion

Nous avons présenté les différents modules du modèle proposé: le banc de filtres et ses réponses aux divers signaux acoustiques; le sous-modèle fonctionnel et les performances de sorties intermédiaires du sous-modèle; le mécanisme d'extraction de la HT et le post-traitement.

Notre modèle repose en grande partie sur les résultats d'expériences. Le modèle est capable de reproduire les phénomènes exposés par Patterson [46]. Cependant, afin de ne pas alourdir notre algorithme nous n'avons pas simulé de façon exacte les activités de l'audition, mais nous avons effectué des opérations mathématiques équivalentes (sous-modèle fonctionnel). Par ailleurs, les expériences ont démontré qu'il est raisonnable d'extraire la HT en utilisant directement les informations à la sortie des filtres auditifs.

CHAPITRE 4

EXPÉRIENCES ET RÉSULTATS

4.1 Introduction

Dans ce chapitre, on présente les résultats que nous avons obtenus à partir du modèle proposé et on les compare avec ceux d'un algorithme classique qui nous sert de la référence. Cette comparaison nous permet d'estimer la capacité du modèle proposé à extraire la hauteur tonale de la parole téléphonique. Afin de vérifier que le modèle proposé possède cette capacité, on compare les performances du modèle proposé avec l'étiquetage manuel qui a été réalisé par le logiciel "Signalyze" en mesurant directement la période du signal périodique.

Les paragraphes suivants décrivent les données utilisées, l'algorithme de référence, la comparaison et l'évaluation des performances du modèle proposé.

4.2 Données utilisées dans les expériences

Le Centre Canadien de Recherche sur l'Informatisation du Travail (CCRIT) a mis à notre disposition la base de données (CRIQUB), qui comprend les prononciations de 12 chiffres isolés, 17 mots isolés et 20 séries de chiffres pour 387 locuteurs. Les données ont été numérisées à travers le réseau téléphonique de la région de Montréal et les locuteurs parlaient en français ou en anglais. Dans les expériences, nous avons utilisé

une partie de la base de données, qui comprend seulement 10 chiffres français isolés et 14 mots français isolés. Ces 24 mots, présentés au tableau 4.1, ont été aléatoirement choisis dans la base de données de façon à ce qu'ils soient prononcés par 24 locutuers différents.

Chiffre	Mot	
0	ANGLAIS	COMMANDE
1	FRANCAIS	ANNULATION
2	TERMINER	INFORMATION
3	ANNULER	RECOMMENCER
4	CHIFFRE	
5	QUITTER	
6	ARRET	
7	DÉBUT	
8	OUI	
9	NON	

Tableau 4.1 Les mots testés dans les expériences

Comme nous l'avons mentionné au début de ce mémoire, la composante fondamentale n'est pas claire dans le spectre du signal de parole téléphonique en raison de la bande étroite de la ligne téléphonique. Évidemment, ceci augmente la difficulté de l'extraction. D'ailleurs certaines prononciations dans les données testées sont plus faibles que les prononciations normales.

Pour tester les performances du modèle en milieu bruité, nous avons ajouté du bruit au signal de la parole. Ce bruit a été obtenu en amplifiant le bruit électronique de la carte d'acquisition. Le rapport de signal au bruit ("signal to noise ratio", SNR) en dB

est défini par la formule suivante:

$$SNR = 10 \cdot log \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} b^2(n)} \tag{4.1}$$

où $s(n)$ est le signal de la parole, $b(n)$ est du bruit ajouté et N est le nombre d'échantillons dans le segment considéré.

Nous donnons un exemple à la figure 4.1 pour illustrer le signal de la parole bruitée. A partir de cette figure, on trouve que le SNR est de -9.09dB, -4.70dB dans les zones de silence, il est grand dans les zones de voyelle ($+11.21$dB, $+11.91$dB et $+15.36$dB) et petit dans les zones de consonne ($+4.12$dB et $+2.92$dB). Alors le SNR moyen pour la prononciation entière de ce mot est de $+8.77$dB.

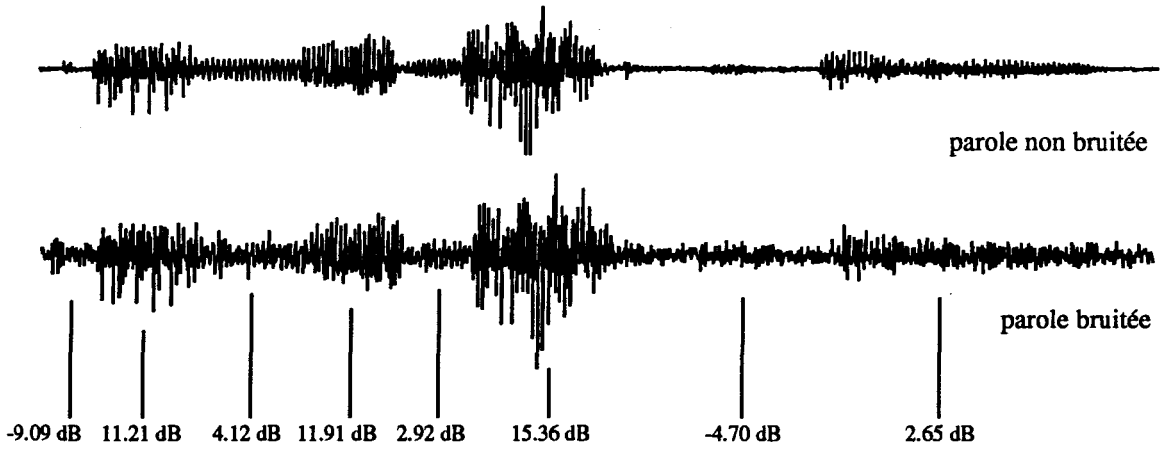

Figure 4.1 Un exemple du signal de la parole bruitée testée dans les expériences: ANNULATION

4.3 Un algorithme d'extraction de la fréquence fondamentale du signal pour l'évaluation des performances du modèle proposé

La HT d'un son complexe périodique en général correspond à sa fréquence fondamentale (cf. la section 2.2.3). Pour évaluer les performances du modèle proposé, on a développé un autre algorithme, nommé "AUTO+PT", pour lequel la fréquence fondamentale de la parole est trouvée directement en calculant l'auto-corrélation suivie d'une prise de décision sur la fréquence fondamentale. Notons que l'algorithme de décision finale de la fréquence fondamentale pour l'AUTO+PT est identique à celui utilisé dans le mécanisme de prise de décision du modèle proposé. La figure 4.2 illustre la différence de structure d'algorithme entre l'AUTO+PT et le modèle.

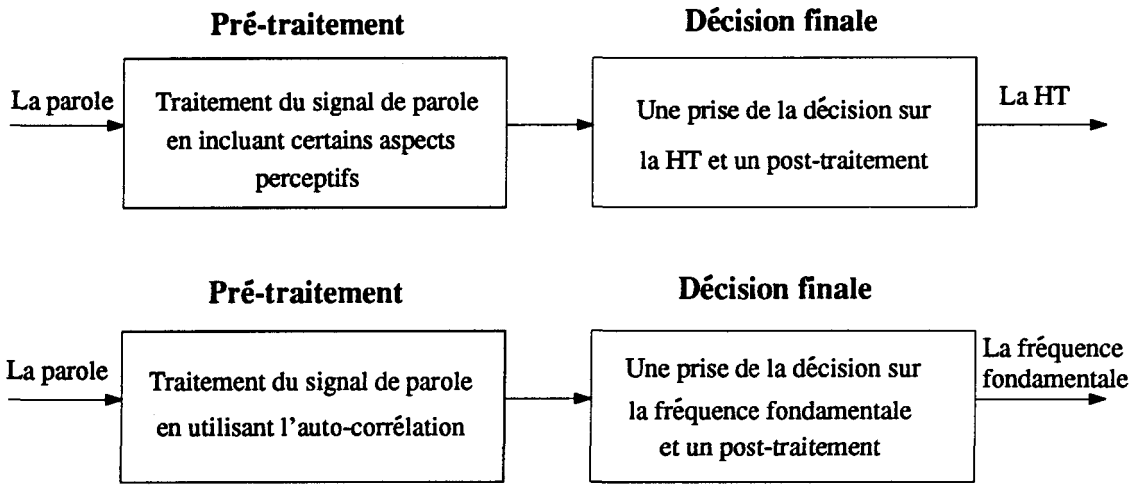

Figure 4.2 Comparaison des structures de l'algorithme entre l'AUTO+PT et le modèle proposé. En haut: la structure de l'algorithme du modèle. En bas: la structure de l'algorithme de l'AUTO+PT

A partir de la figure 4.2, on constate que la seule différence réside dans le pré-traitement du signal de parole. Le pré-traitement du signal dans le modèle proposé se réalise en exploitant certains aspects du système auditif, contrairement à l'AUTO+PT qui utilise l'auto-corrélation. L'AUTO+PT est utilisé afin de servir de référence pour

comparer les performances du modèle. Nous donnons la comparaison des performances des deux algorithmes de façon détaillée à la section suivante.

4.4 Performances du modèle proposé par rapport à celles de l'AUTO+PT

Dans ce paragraphe, on présente les performances du modèle proposé, les résultats sur les mots que nous avons sélectionnés dans les expériences, puis on donne d'autres exemples pour illustrer les performances du modèle en milieu bruité. Notons que chaque mot testé est prononcé par différent locuteur. Pour illustrer visuellement la performance du modèle proposé, nous avons placé le résultat de l'AUTO+PT et le résultat de l'étiquetage manuel sur la même figure.

4.4.1 Pour la parole téléphonique

De la figure 4.3 à la figure 4.26, nous présentons les performances du modèle proposé pour la parole téléphonique. Dans le premier exemple (figure 4.3), on observe que les valeurs de HT du modèle sont en concordance avec celles de l'évaluation manuelle, cependant, l'AUTO+PT ne peut pas trouver la fréquence fondamentale pour la consonne liquide / *l* / car le signal de cette partie est moins périodique. De même, à la figure 4.4, on constate que l'AUTO+PT n'est pas capable de détecter la fréquence fondamentale dans la consonne liquide / *r* /, de plus l'AUTO+PT estime que la consonne occlusive / *k* / et la consonne fricative / *s* / sont voisées. A la figure 4.5 et à la figure 4.6, le modèle se comporte très bien vis à vis de l'étiquetage manuel, par contre, l'AUTO+PT fonctionne mal en ce sens qu'il estime que le bruit qui précéde la voyelle nasale / *ã* / (figure 4.5) et la consonne occlusive / *t* / (figure 4.6) sont voisés. A la figure 4.7, le modèle et l'AUTO+PT se comportent bien au milieu de la consonne fricative voisée

/ *z* / où ils sont capables de détecter la fréquence fondamentale du signal, mais ils se comportent mal respectivement au début et à la fin de / *z* / pour lesquels les deux algorithmes ne peuvent pas trouver la fréquence fondamentale car l'énergie du signal est très faible. Par ailleurs, on constate que les trois courbes concordent sur cette figure. Pour le mot "UN" à la figure 4.8, les trois courbes sont semblables à la partie antérieure de la prononciation, mais au milieu de la prononciation, l'AUTO+PT chute rapidement en raison de la nasalisation. A la figure 4.9, la performance du modèle est en bonne concordance avec l'étiquetage manuel, par contre, l'AUTO+PT ne fonctionne pas pour la consonne occlusive voisée / *d* / et la partie postérieure de la voyelle fermée / ϕ /. De fait, ce mot est prononcé par une femme et la hauteur tonale monte réellement beaucoup à la fin de / ϕ /. Comme il a été mentionné à la section 3.6.1, il est difficile de détecter la fréquence fondamentale du signal en se basant sur l'auto-corrélation lorsque la hauteur tonale est très élevé, il est donc normal que l'AUTO+PT soit incapable de suivre le changement rapide de la fréquence fondamentale à la fin du / ϕ /. Les trois courbes à la figure 4.10 sont semblables sauf que l'AUTO+PT estime que la consonne occlusive non voisée / *t* / est voisée. A la figure 4.11, le modèle ne donne pas de performance satisfaite au début de la consonne occlusive non voisée / *k* / car il évalue que cette section est voisée. De la même façon, à la figure 4.12 le modèle et l'AUTO+PT se comportent mal pour la consonne occlusive / *k* /.

De la figure 4.13 à la figure 4.20, les performances du modèle sont en concordance avec l'étiquetage manuel. Plus spécifiquement, le modèle fonctionne très bien sur la consonne occlusive / *d* / de la figure 4.19 et sur la consonne fricative / *f* / de la figure 4.20, par contre, l'AUTO+PT fonctionne mal pour les mêmes consonnes.

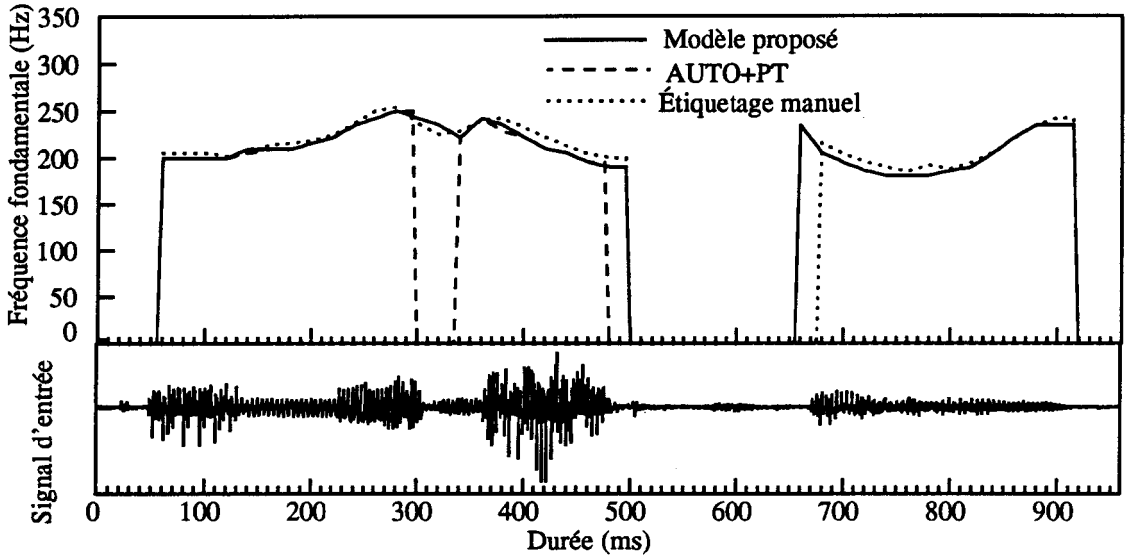

Figure 4.3 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "ANNULATION"

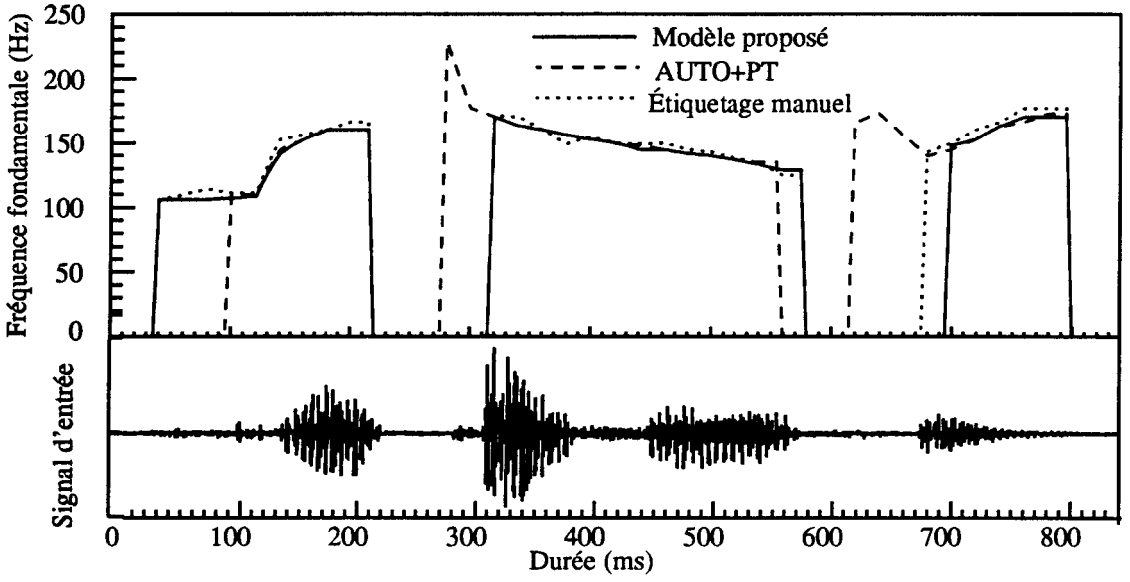

Figure 4.4 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "RECOMMENCER"

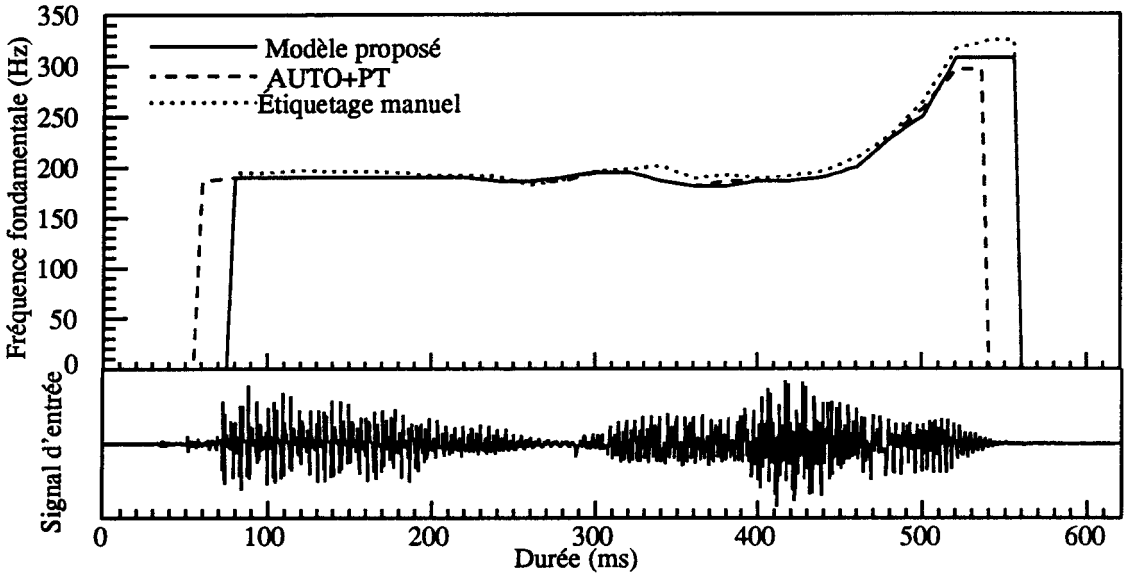

Figure 4.5 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "ANGLAIS"

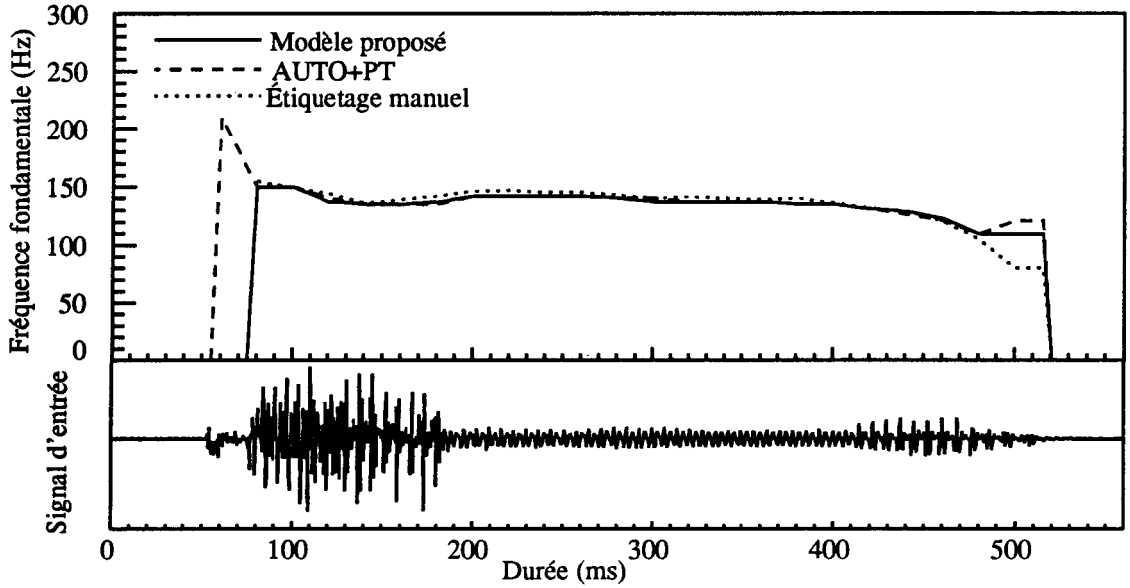

Figure 4.6 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "TERMINER"

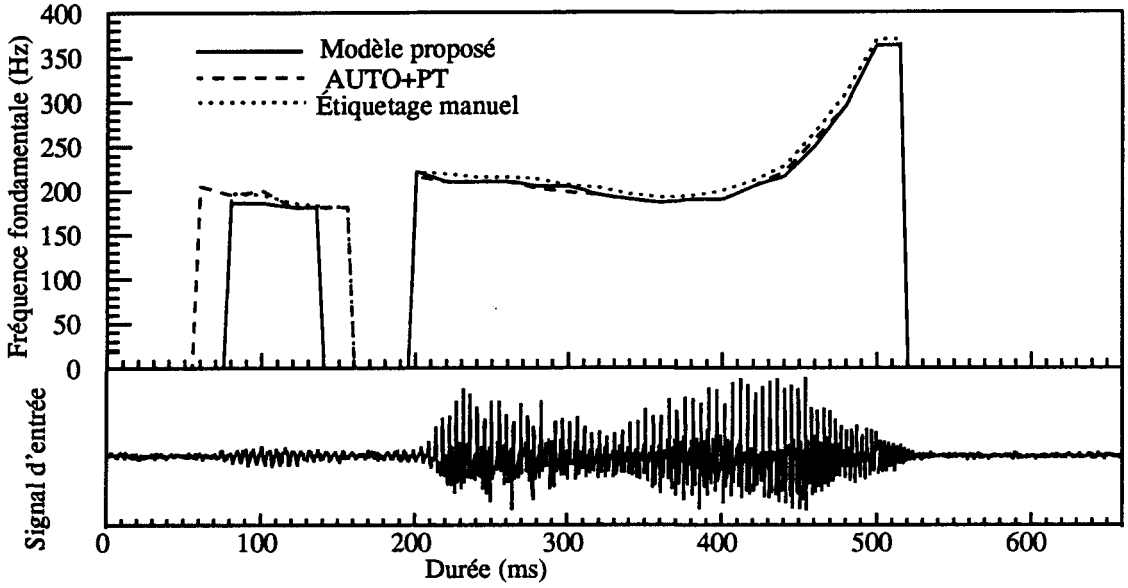

Figure 4.7 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "ZÉRO"

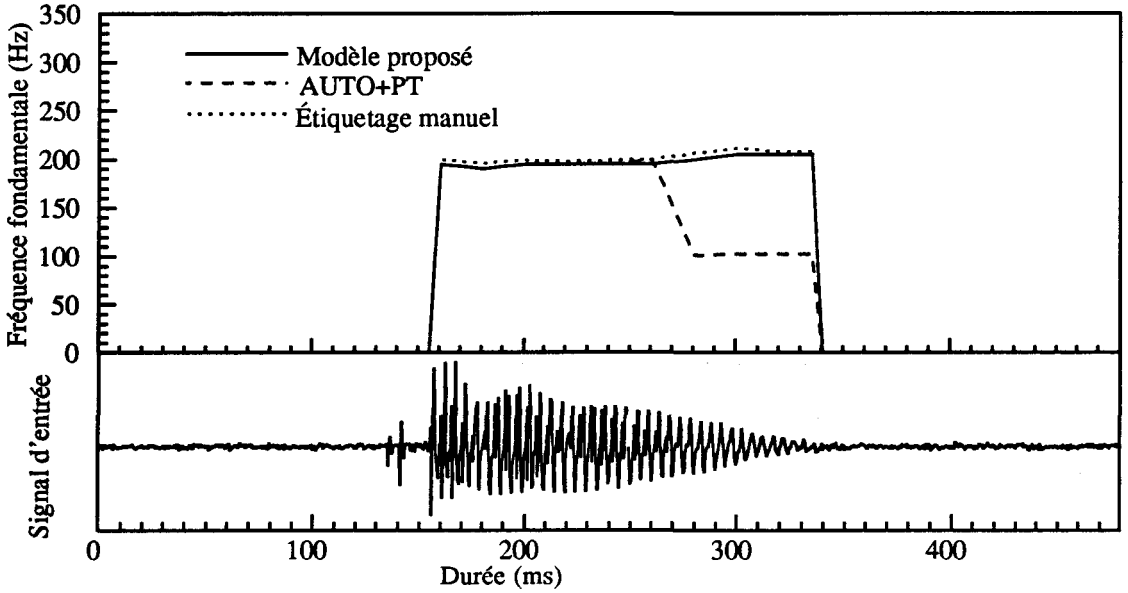

Figure 4.8 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "UN"

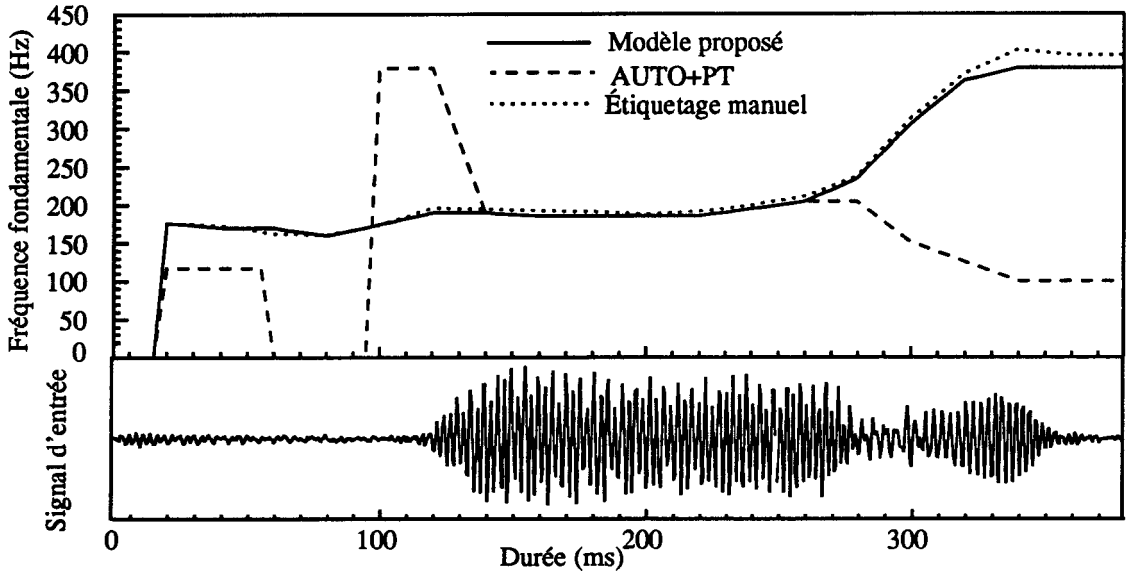

Figure 4.9 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "DEUX"

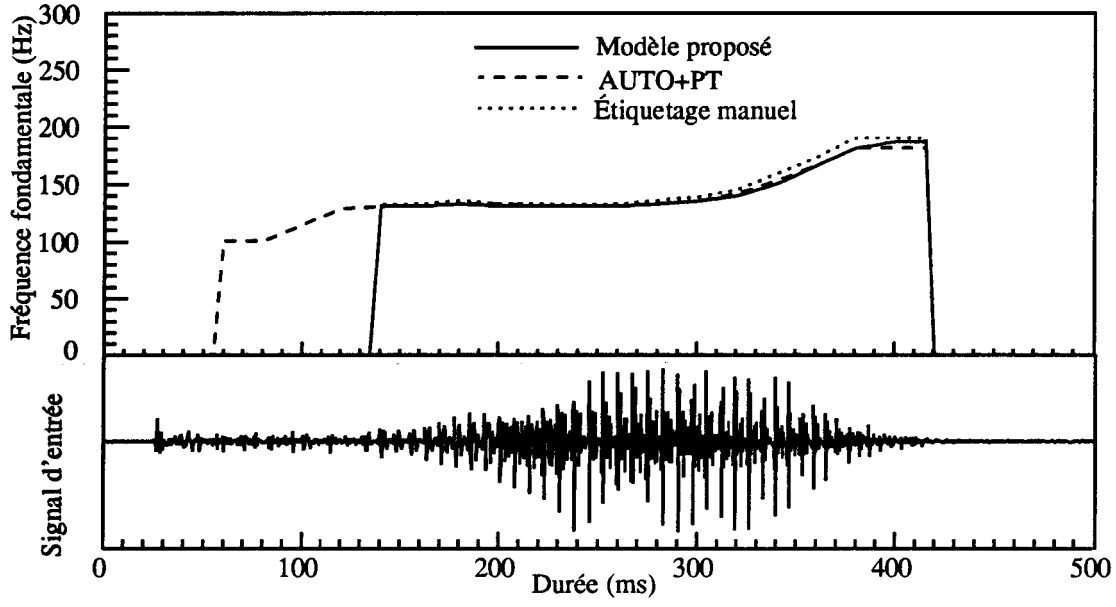

Figure 4.10 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "TROIS"

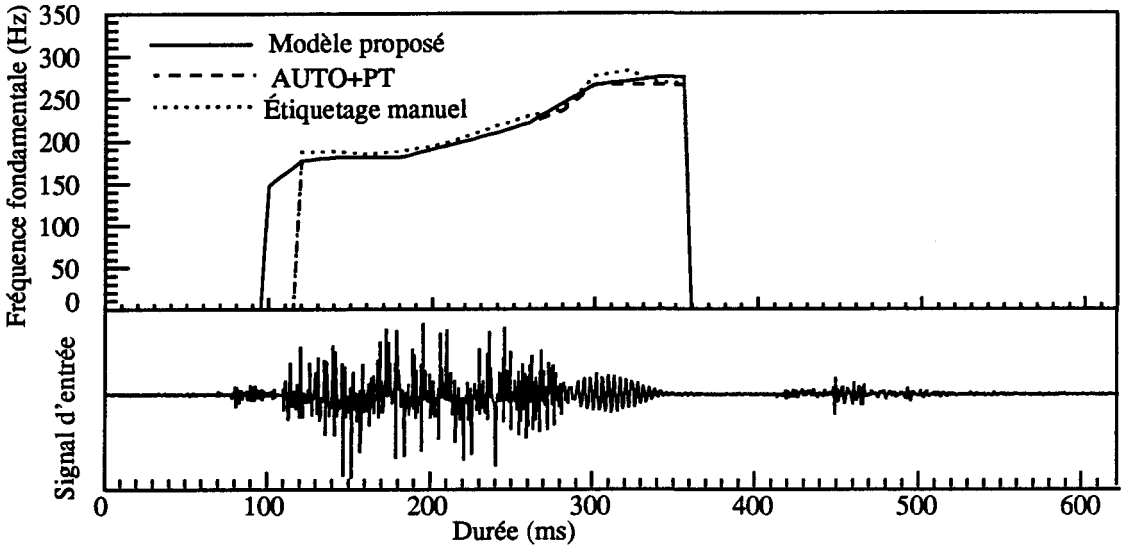

Figure 4.11 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "QUATRE"

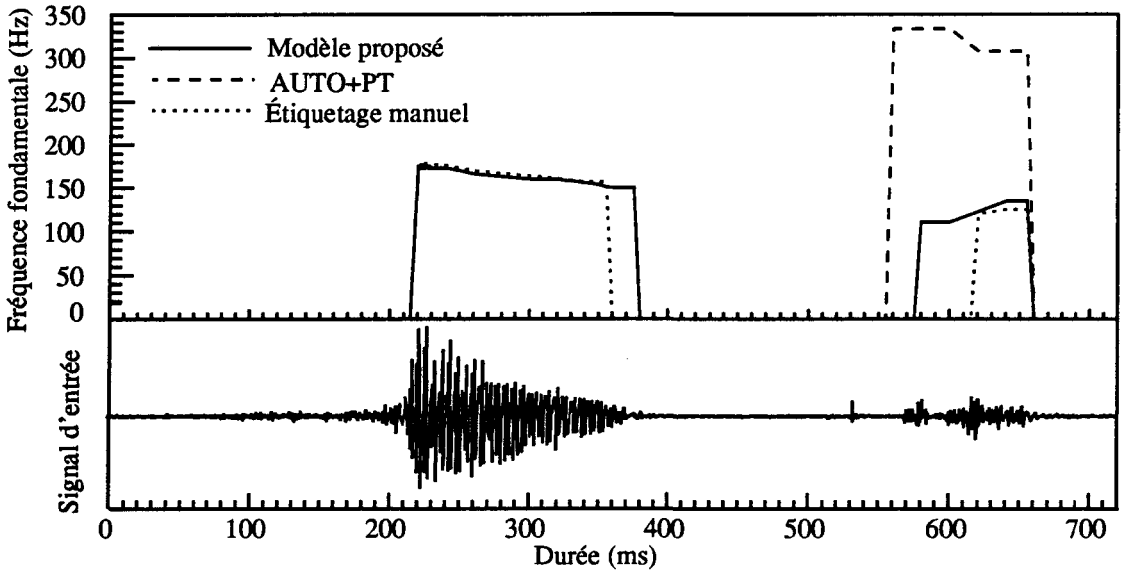

Figure 4.12 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "CINQ"

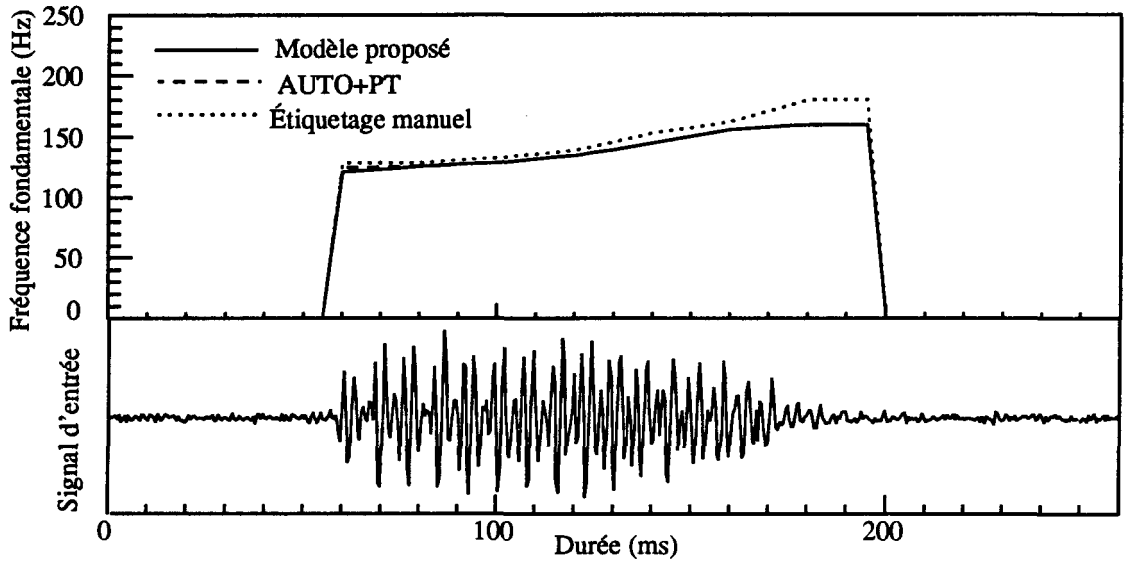

Figure 4.13 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "SIX"

Par ailleurs, on constate que l'AUTO+PT n'est pas capable de trouver la fréquence fondamentale à la partie postérieure de la voyelle nasale / *ã* / à la figure 4.19. A la figure 4.21 le problème réside surtout à la partie de / *sjõ* / où la période du signal que l'AUTO+PT donne est le double de la période réelle du signal. Ce type d'erreur, appelé "erreur double", est une erreur fréquente dans la méthode d'auto-corrélation. Notons qu'à la figure 4.23, il y a un phénomène de saturation de courbe du modèle à la fin de la prononciation car nous avons mis une limite supérieure à la hauteur tonale qui est de 400 Hz pour nos expériences (c'est un paramètre modifiable par l'utilisateur). La performance du modèle n'est pas satisfaisante à la fin de la voyelle nasale / *õ* / à la figure 4.24 où le modèle évalue que le bruit est voisé. Pour le mot "CHIFFRE" à la figure 4.25, on observe que la courbe du modèle est très proche de celle de l'étiquetage manuel, cependant, l'AUTO+PT se comporte mal à / *fr* / où l'AUTO+PT estime que

ces deux consonnes non voisées sont voisées. Dans le dernier exemple (figure 4.26), le modèle et l'AUTO+PT estiment que la partie antérieure de la consonne occlusive non voisée / *t* / est voisée, par ailleurs, il y a une erreur double à / *ki* / pour l'AUTO+PT.

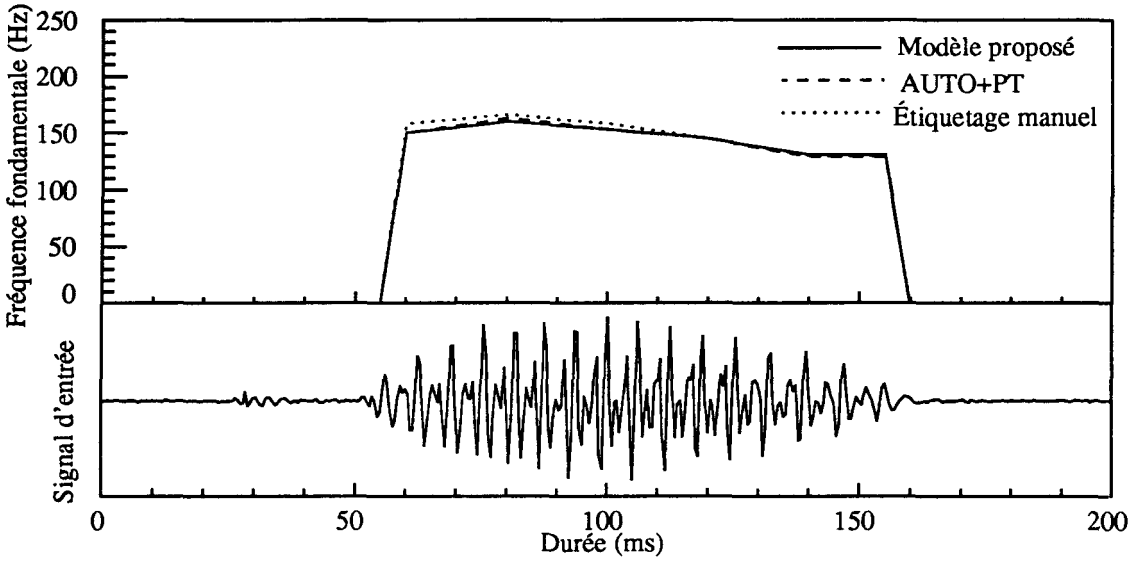

Figure 4.14 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "SEPT"

4.4.2 Pour la parole téléphonique bruitée

De la figure 4.27 à la figure 4.30, nous présentons quatre exemples pour illustrer les performances du modèle proposé pour la parole téléphonique bruitée (la structure du modèle dans ce cas a déjà été montrée à la figure 3.3). Globalement, la performance du modèle est acceptable pour ces quatre exemples. Nous n'avons pas fait beaucoup d'expériences pour la parole bruitée en considérant que la principale tâche est le traitement de la parole téléphonique dans le cadre de ce mémoire.

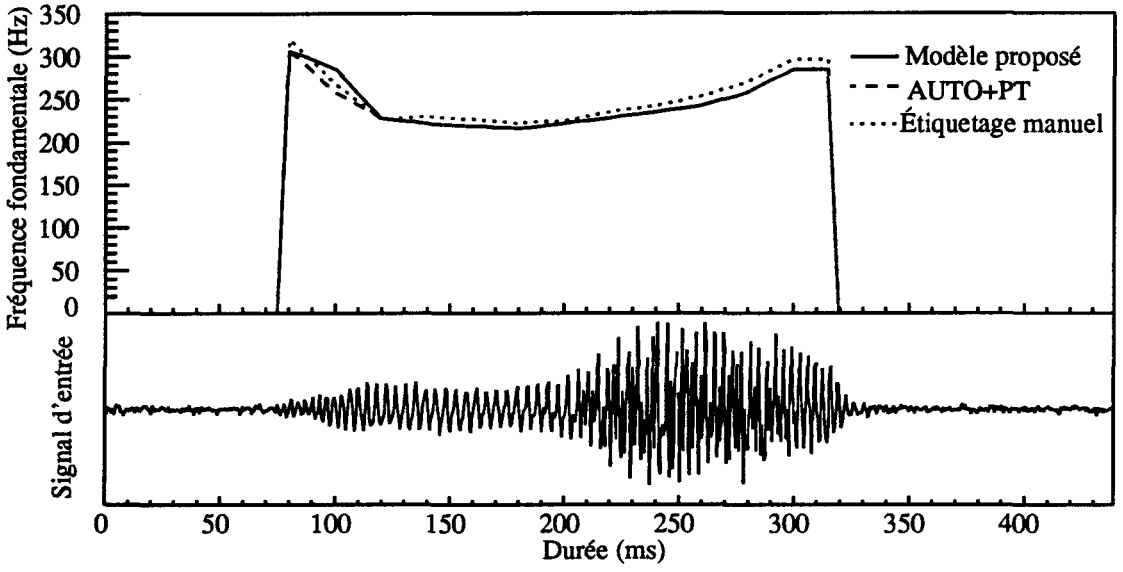

Figure 4.15 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "HUIT"

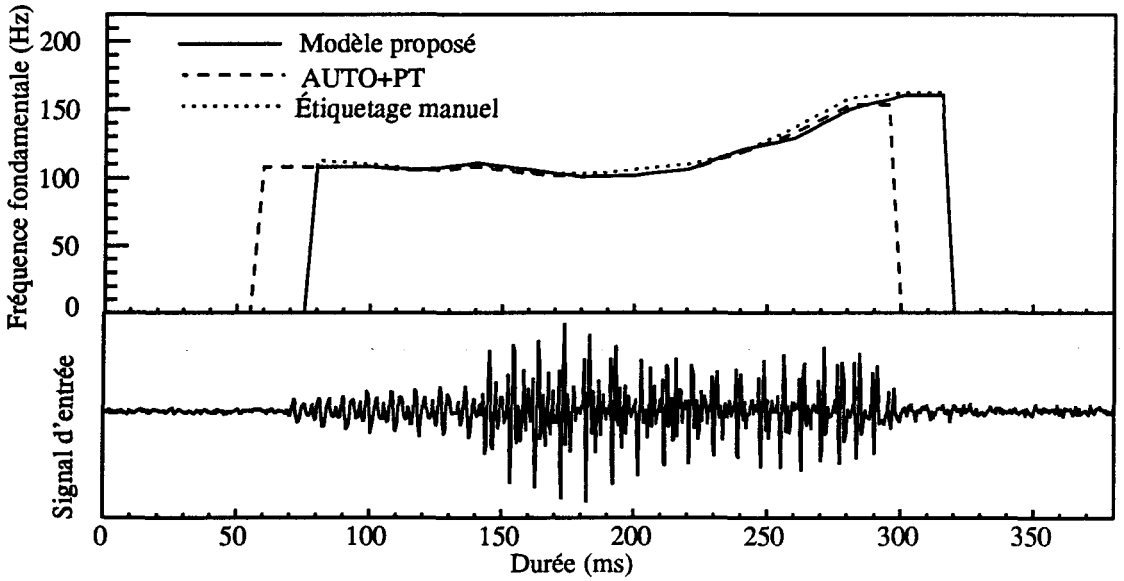

Figure 4.16 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "NEUF"

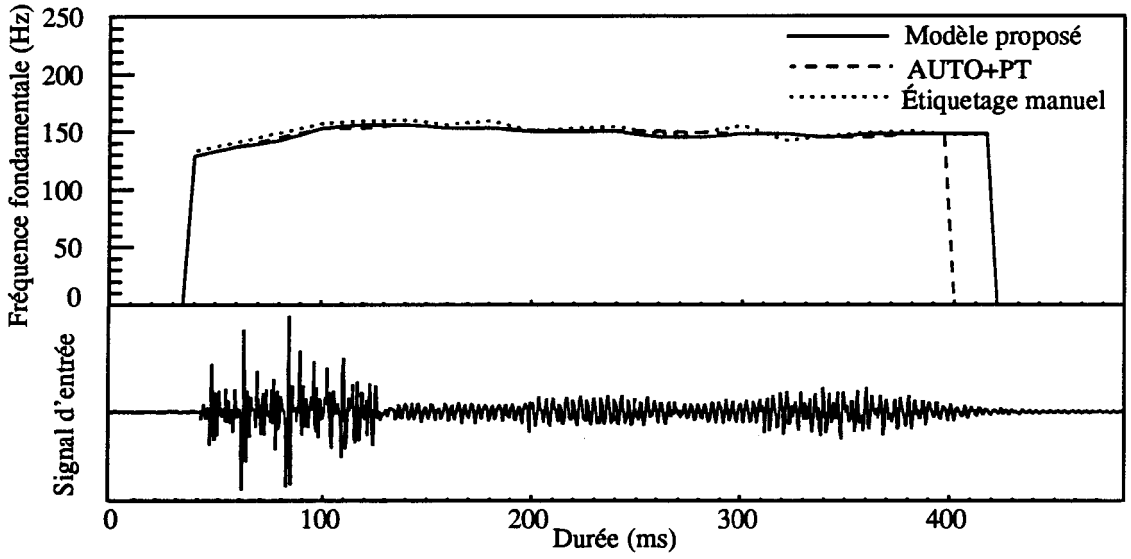

Figure 4.17 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "ANNULER"

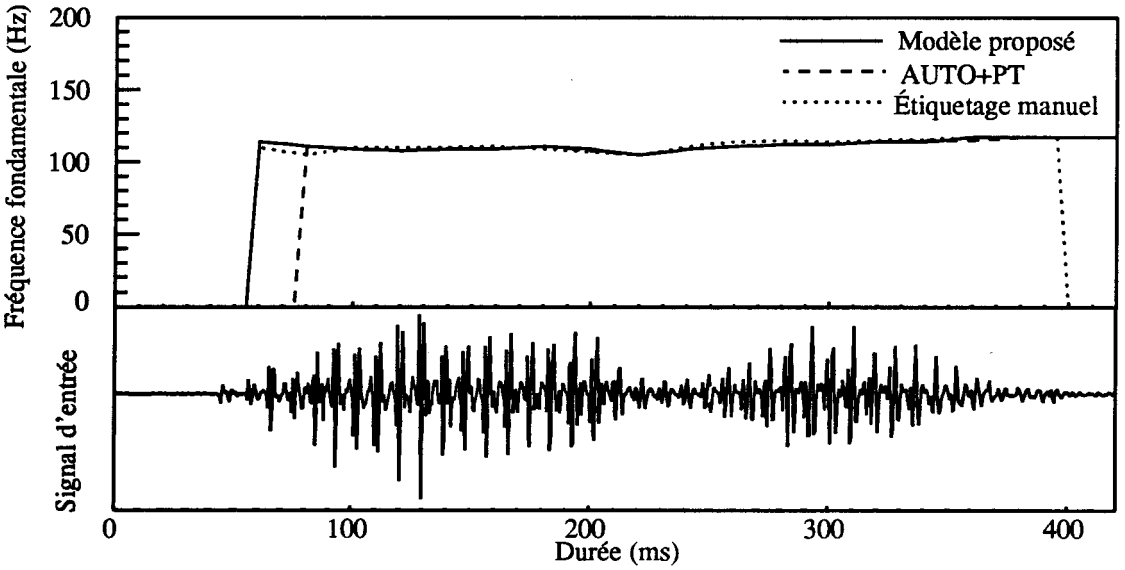

Figure 4.18 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "ARRET"

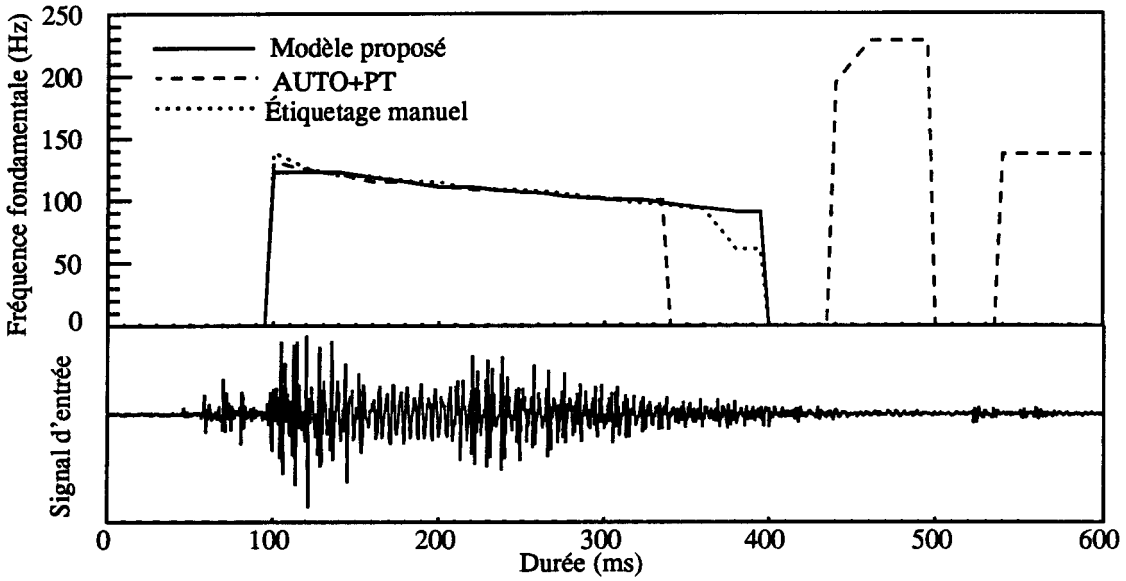

Figure 4.19 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "COMMANDE"

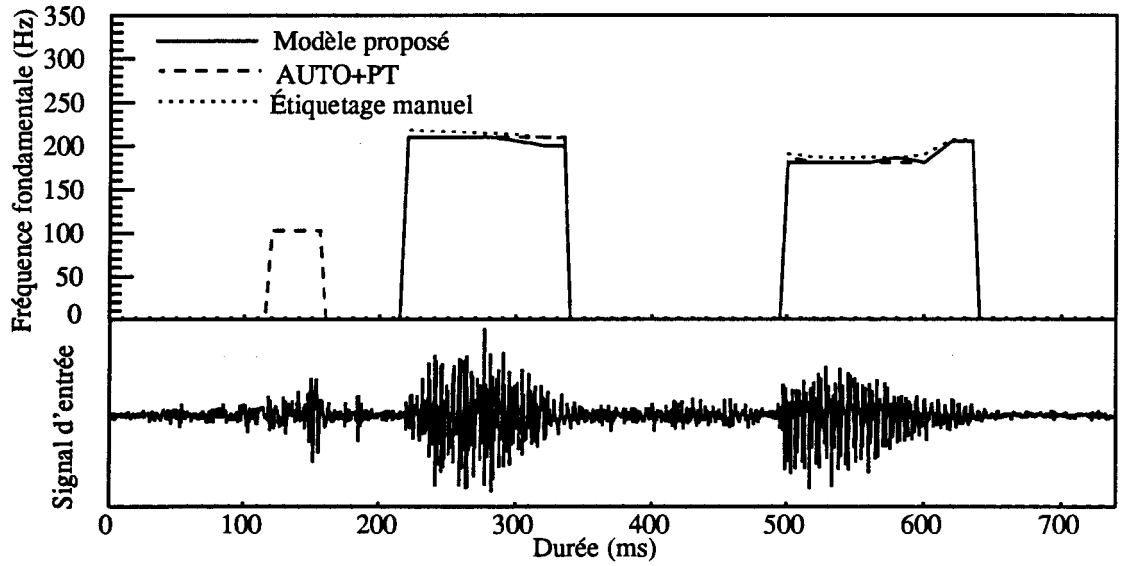

Figure 4.20 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "FRANÇAIS"

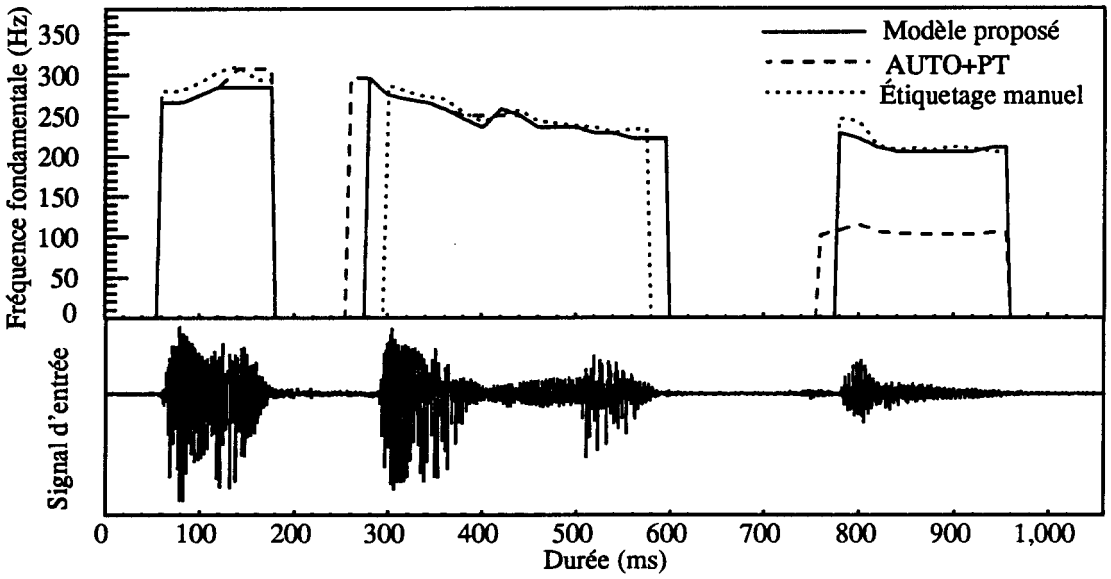

Figure 4.21 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "INFORMATION"

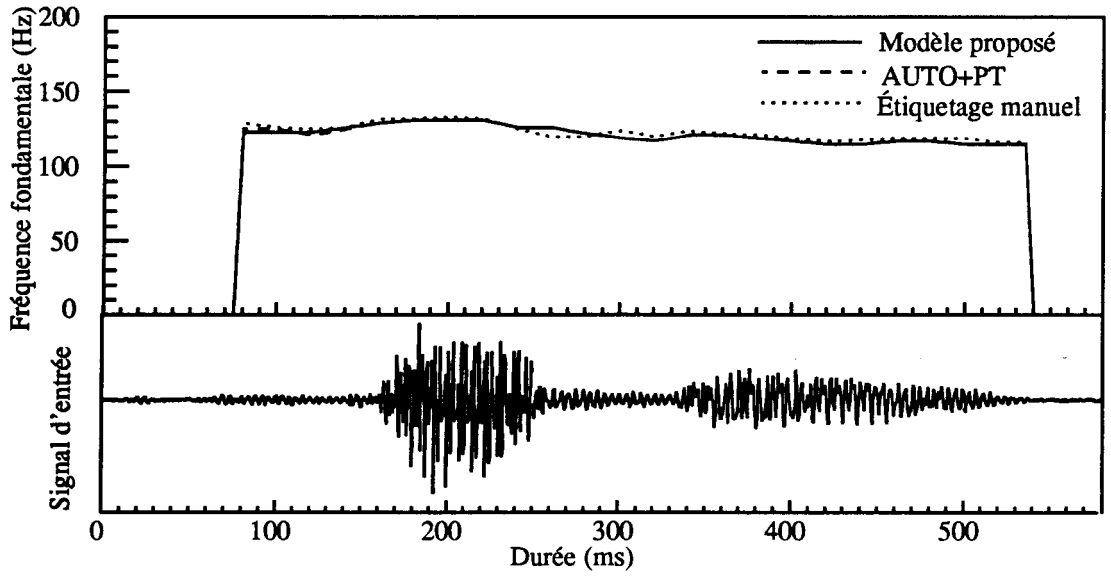

Figure 4.22 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "DÉBUT"

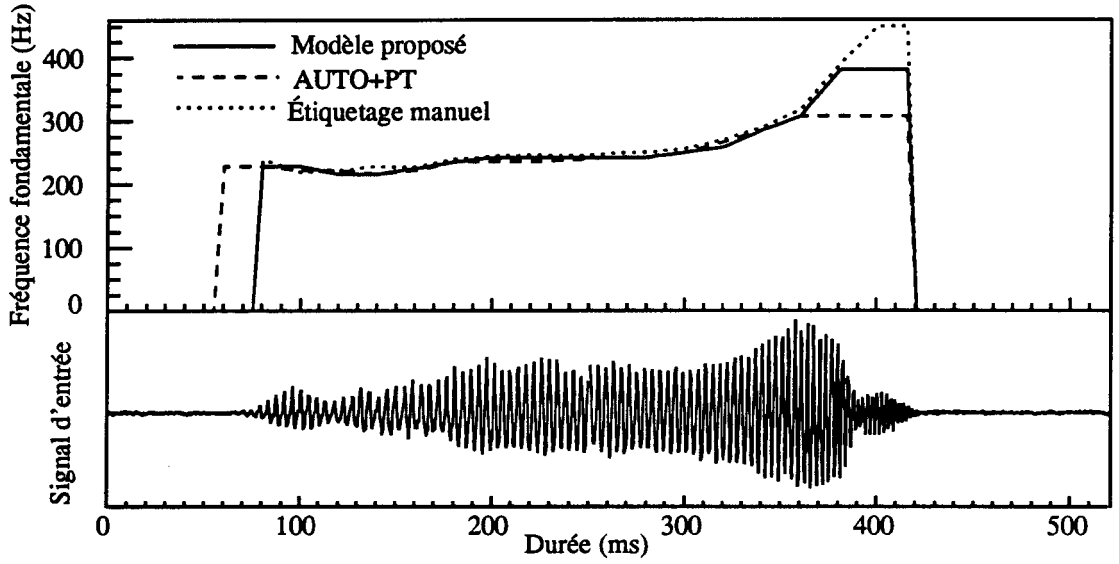

Figure 4.23 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "OUI"

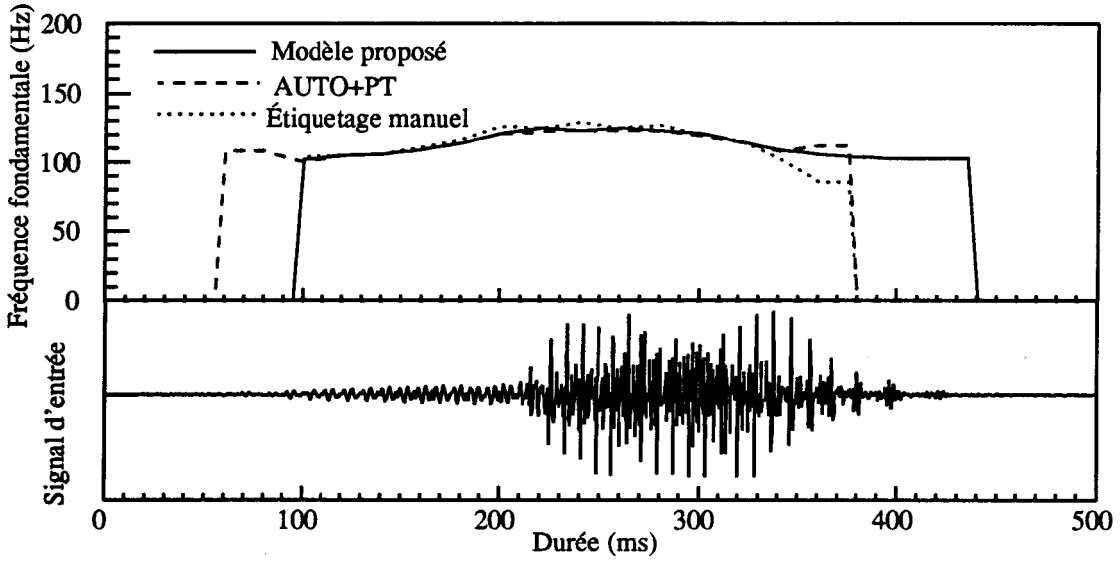

Figure 4.24 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "NON"

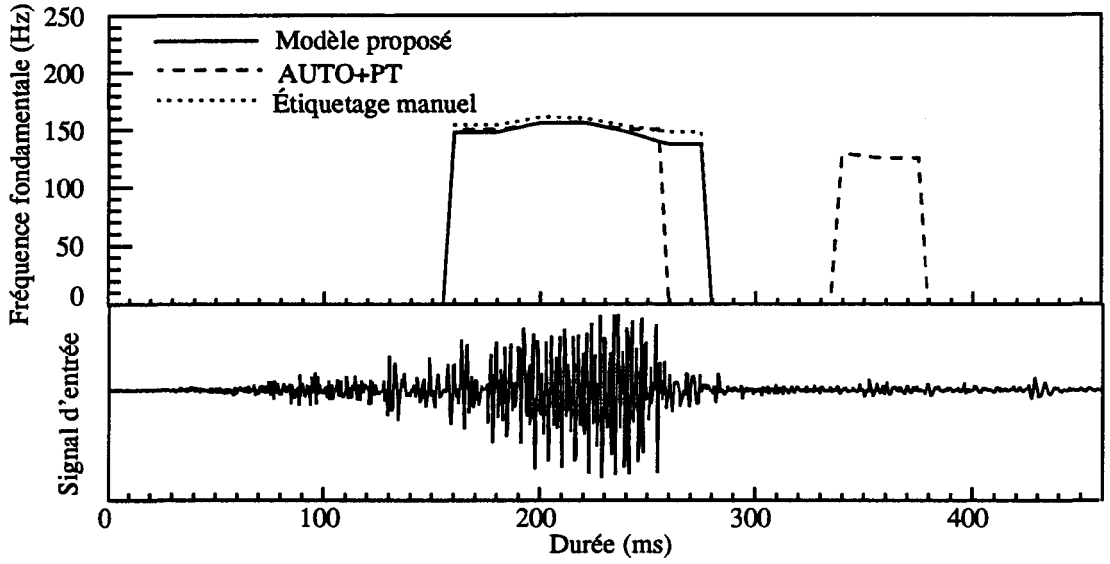

Figure 4.25 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "CHIFFRE"

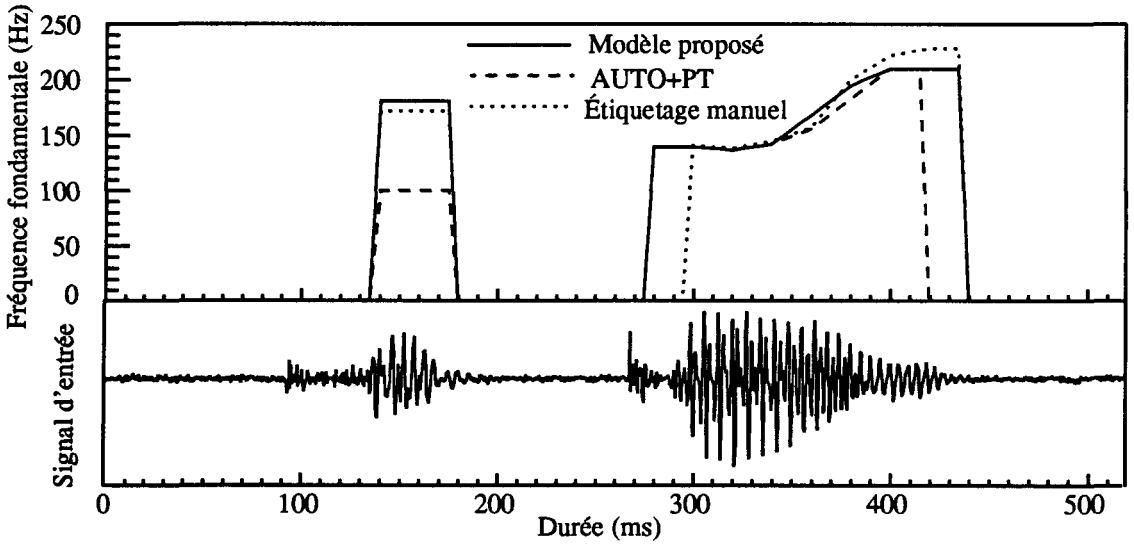

Figure 4.26 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel pour la parole téléphonique "QUITTER"

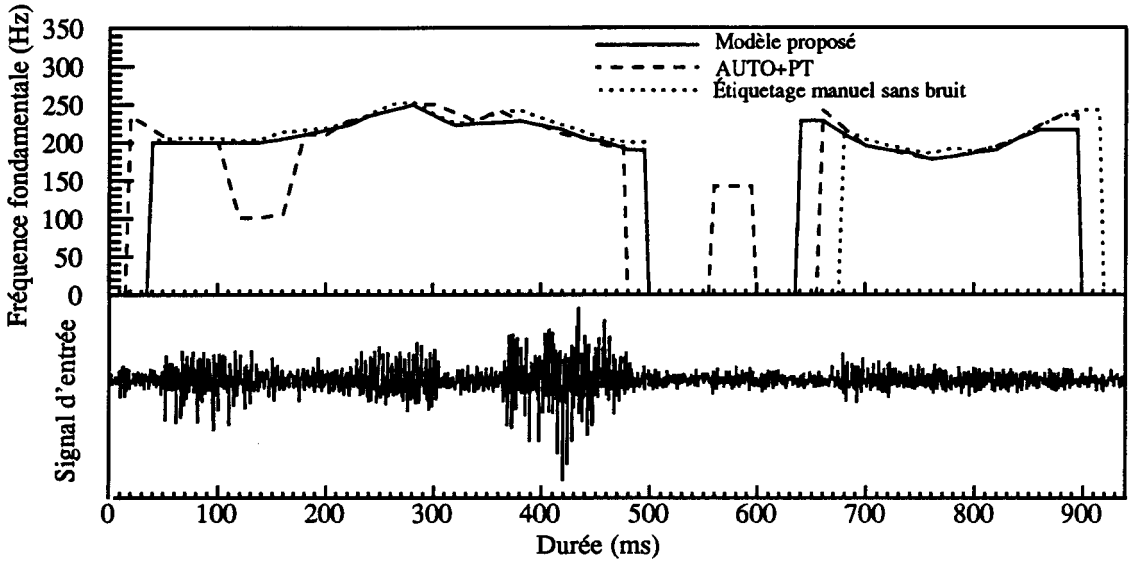

Figure 4.27 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel sans bruit pour la parole téléphonique bruitée "ANNULATION"

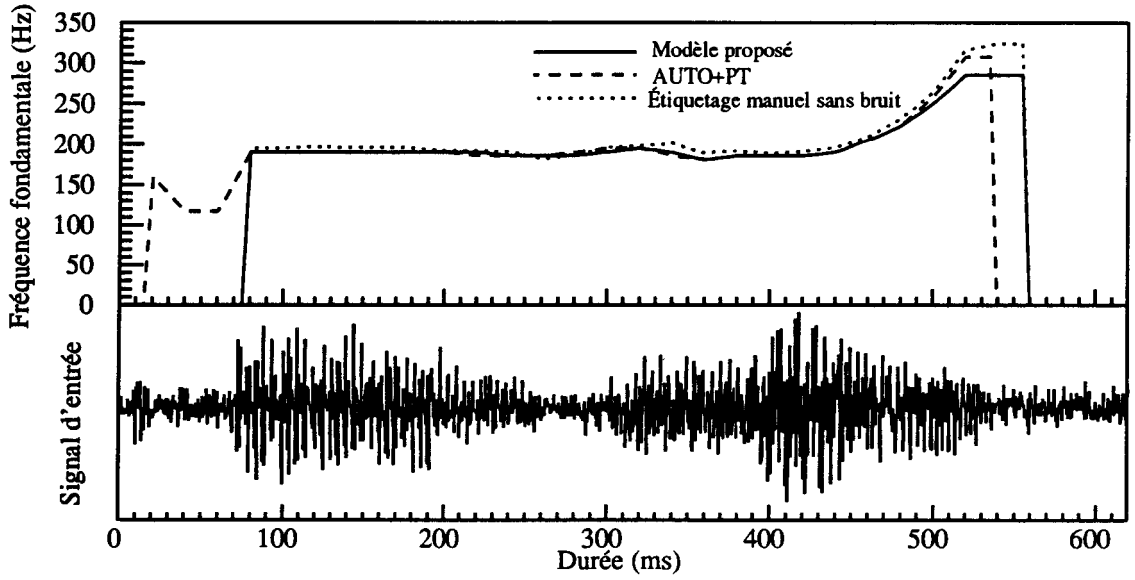

Figure 4.28 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel sans bruit pour la parole téléphonique bruitée "ANGLAIS"

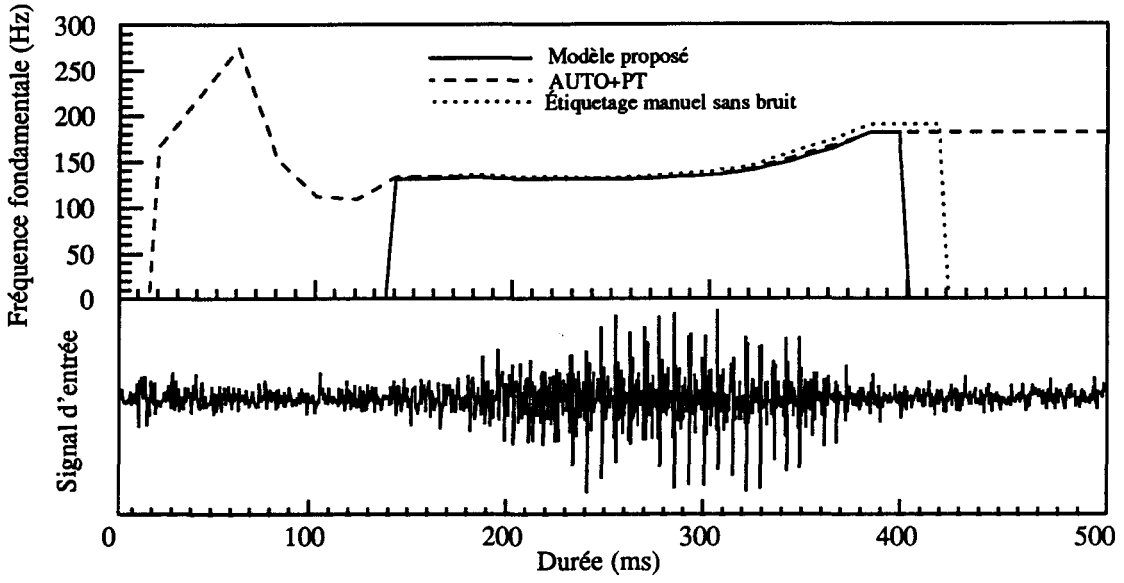

Figure 4.29 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel sans bruit pour la parole téléphonique bruitée "TROIS"

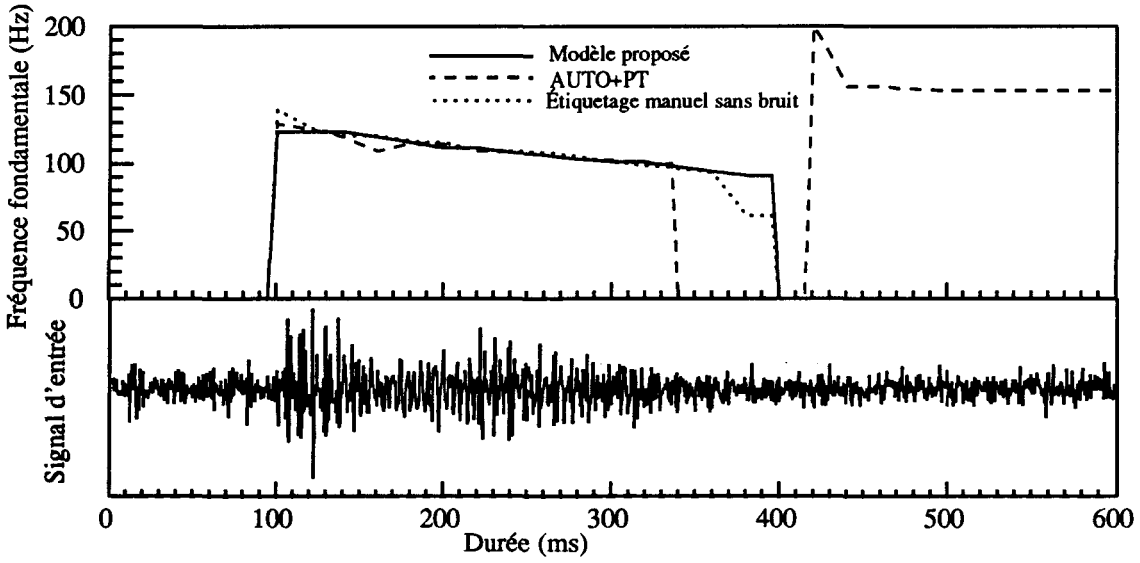

Figure 4.30 Comparaison de la performance du modèle proposé avec l'AUTO+PT et avec l'étiquetage manuel sans bruit pour la parole téléphonique bruitée "COMMANDE"

4.5 Évaluation de l'algorithme proposé

Il est très difficile d'évaluer la performance du modèle proposé parce que la fréquence fondamentale véritable du signal de parole n'est directement pas disponible. Afin de réaliser l'évaluation, on doit trouver un moyen qui permet de calculer le fondamental. Dans notre cas, la fréquence fondamentale "véritable" a été obtenue en utilisant le logiciel "Signalyze" pour mesurer de façon manuelle la période du signal. Évidemment, cette mesure manuelle à plus grande échelle est très pénible. C'est pourquoi nous avons sélectionné 24 mots (décrits à la section 4.2) pour l'évaluation des performances du modèle. Comme nous avons remarqué à la section 4.2, ces 24 mots de test ont été choisis de façon aléatoire à partir de la base de données, et ils sont prononcés par 24 locuteurs différents.

Pour effectuer l'évaluation, trois types d'erreurs sont étudiées [44]: l'erreur de décision voisée/non voisée, l'erreur grossière et l'erreur fine. Le premier type d'erreur comprend tous les cas où les segments voisés de la parole sont détectés par le modèle comme des segments non voisés et vice versa. L'erreur grossière est définie comme étant la suivante: supposons que $F_{oa}(i)$ est la fréquence fondamentale mesurée manuellement pour le segment i (ou la fenêtre i) et $F_{om}(i)$ est la fréquence fondamentale que le modèle estime pour le même segment, alors

$$F_e(i) = \frac{|F_{oa}(i) - F_{om}(i)|}{F_{oa}(i)} \tag{4.2}$$

si $F_e(i) \geq 10\%$, alors $F_e(i)$ est appelée erreur grossière, sinon $F_e(i)$ est appelée petite erreur. L'erreur fine se trouve en calculant la racine carrée de la moyenne des petites erreurs au carré ("root mean square", RMS), elle peut s'exprimer par la formule

Algorithme	Erreur décicion Voisée/non voisée	Erreur grossière	Erreur fine (RMS)
Modèle proposé	3.18%	0.94%	6.6888 Hz
AUTO+PT	10.11%	5.43%	5.8993 Hz

Tableau 4.2 Évaluation des performances pour le modèle proposé et l'AUTO+PT (le nombre de segments analysés: 534)

suivante:

$$F_{ef} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} F_e^2(i)} \tag{4.3}$$

où $F_e(i)$ est la petite erreur pour le segment i et N est le nombre de petites erreurs dans tous les segments étudiés.

En utilisant les définitions précédentes d'erreur, nous avons effectué l'évaluation pour le modèle et l'AUTO+PT en analysant ses performances sur 24 mots que nous avons montrés à la section 4.4.1. Notons que chaque mot comprend divers segments (fenêtres) et que la largeur des segments est de 20 ms. Le tableau 4.2 donne les résultats de l'analyse pour 24 mots (au total: 534 segments).

En examinant le tableau 4.2, on constate que le modèle proposé fonctionne beaucoup mieux que l'AUTO+PT en regard de l'erreur de décision voisée/non voisée et de l'erreur grossière, cependant, il semble qu'il fonctionne moins bien que l'AUTO+PT en regard de l'erreur fine. Ceci pourrait s'expliquer par le fait que la hauteur tonale n'est pas identique à la fréquence fondamentale du signal de parole d'un point de vue perceptif et certaines opérations du pré-traitement dans le modèle proposé (le filtrage et la multiplication) peuvent introduire des erreurs supplémentaires par rapport à l'AUTO+PT qui n'utilise pas ces opérations. En considérant que l'erreur de décision voisée/non voisée et l'erreur grossière sont toujours plus importantes que l'erreur fine,

on peut conclure que la performance globale du modèle proposé est supérieure à celle de l'AUTO+PT.

Par ailleurs, en examinant le tableau 4.2, on voit que la précision du calcul des deux algorithmes (l'erreur fine) n'est pas élevée. Le problème précédent est évoqué par l'utilisation de l'auto-corrélation car l'algorithme d'extraction de la fréquence fondamentale qui se base sur la fonction d'auto-corrélation estime la fréquence fondamentale moyenne du signal des deux fenêtres ajacentes et non pas la fréquence fondamentale "instantané".

Comme nous avons remarqué à la section 4.4.1, les données (24 mots) utilisées lors des expériences et des évaluations ont été choisies de façon aléatoire à partir de la base de données; chaque mot est prononcé par un locuteur différent. En conséquence, l'analyse qui a été faite ci-dessus est une analyse échantillonnée à partir de toute la base de données.

4.6 Conclusion

Nous avons présenté les performances du modèle proposé pour la parole téléphonique, effectué les comparaisons avec un autre algorithme "AUTO+PT", puis réalisé une évaluation en utilisant trois types d'erreur (l'erreur de décision voisée/non voisée, l'erreur grossière et l'erreur fine). Nous avons vérifié que le modèle proposé est capable d'extraire la hauteur tonale de la parole que nous avons testée. Par rapport à l'AUTO+PT, la performance du modèle est satisfaisante même si parfois il ne fonctionne pas très bien sur certains segments de consonne.

CHAPITRE 5

CONCLUSION

Dans ce chapitre nous allons discuter les résultats dans un contexte plus général, ainsi que nos perspectives.

5.1 Discussion

Dans ce mémoire, nous nous sommes intéressés à l'analyse de la parole et plus particulièrement à l'extraction de la hauteur tonale sur de la parole téléphonique. Nous avons proposé un modèle fonctionnel pratique qui est capable de remplir cette tâche.

Une revue des algorithmes d'extraction de la fréquence fondamentale et de la hauteur tonale nous a conduit à penser qu'un modèle fonctionnel qui utilise des connaissances psychoacoustiques et physiologiques de l'oreille fonctionnerait mieux que les modèles qui traitent directement le signal de la parole. En effet, nous avons vérifié l'idée précédente en comparant la performance du modèle proposé avec celle d'un autre algorithme "AUTO+PT" pour lequel le pré-traitement du signal se réalise par l'auto-corrélation et le processus de l'estimation de la fréquence fondamentale du signal est identique à celui du modèle proposé.

L'algorithme proposé est basé sur l'hypothèse que l'information à la sortie des filtres auditifs est suffisante pour l'extraction de la hauteur tonale. Plus précisément, nous

avons conçu un banc de filtres auditifs pour simuler les mouvements mécaniques de la membrane basilaire; afin de ne pas alourdir notre algorithme, nous n'avons pas simulé de façon exacte la transduction mécanique-électrique, mais nous avons effectué des opérations mathématiques (multiplication, auto-corrélation); enfin nous avons développé un algorithme pour l'estimation de la période de HT de la parole et un post-traitement.

Le modèle proposé est testé sur des données de parole téléphonique. Les résultats indiquent que l'approche utilisée permet d'obtenir la hauteur tonale de la parole téléphonique même si l'énergie spectrale liée à la composante fondamentale de ce type de signal n'est pas claire. Nous avons comparé la performance du modèle proposé avec celle de l'AUTO+PT et l'étiquetage manuel. Les résultats comparatifs démontrent que le modèle fonctionne mieux que l'AUTO+PT et l'estimation de la fréquence fondamentale est similaire à l'étiquetage manuel. Nous avons aussi testé le modèle avec quelques données de parole bruitée, dont le SNR est de +8dB, le résultat obtenu est acceptable.

Nous sommes satisfaits de la performance du modèle proposé parce qu'il est capable d'extraire la hauteur tonale sur de la parole téléphonique. Néanmoins, nous reconnaissons que la robustesse du modèle au bruit est loin de ce qui est espéré dans la pratique. Par ailleurs, il est nécessaire de réduire le temps de traitement.

5.2 Extension et travaux ultérieurs

Une amélioration possible du modèle peut se faire au niveau des filtres. En effet, le filtrage du modèle n'est pas rapide car l'ordre minimal des filtres est élevé (19). Par ailleurs, l'analyse à l'annexe A indique que la réponse des filtres est très dépendante du changement de la fréquence d'échantillonnage. Il faut donc, si possible, développer

un autre type de filtre pour lequel l'ordre est moins élevé et la réponse moins sensible à la fréquence d'échantillonnage.

Une autre amélioration possible du modèle est de considérer la propriété de la dominance des composantes du signal lors de la combinaison des canaux. Il a déjà été trouvé dans nos expériences que la contribution de chaque canal à l'extraction de la hauteur tonale est différente. Ce phénomène est très clair lorsque l'entrée est une consonne ou de la parole bruitée. Nous croyons que cette amélioration sera très importante vis à vis d'un fonctionnement en milieu bruité. Par contre, nous savons que cette amélioration n'est pas facile à réaliser.

Nous allons essayer de remplacer la fonction d'auto-corrélation dans le modèle avec l'algorithme proposé par Medan et al.[34] car, comme nous avons mentionné au chapitre 2, cet algorithme, qui est similaire à l'inter-corrélation, est robuste au bruit et est capable d'extraire la fréquence fondamentale avec une meilleure précision. Par ailleurs, nous allons essayer d'utiliser l'algorithme de Teager [23] [24] [55] à la sortie des filtres car il nous semble que cet algorithme est capable d'extraire l'enveloppe du signal qui est modulée par la hauteur tonale de la parole.

ANNEXE A

CALCUL DES COEFFICIENTS DU FILTRE

Étant donné la réponse impulsionnelle $h(n)$ sa transformée de Fourier est exprimée par la relation suivante:

$$H(f) = \sum_{n=-\infty}^{n=+\infty} h(n) e^{-i2n\pi f/f_s} \qquad \text{A-1.1}$$

où f_s est la fréquence d'échantillonnage. La relation inverse nous donne les coefficients du filtre comme suit:

$$h(n) = \frac{1}{f_s} \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} H(f) e^{i2n\pi f/f_s} df \qquad \text{A-1.2}$$

Dans notre cas, la réponse fréquentielle est caractérisée par $W(f)$:

$$W(f) = (1 + pg)e^{-pg} \qquad \text{A-1.3}$$

où

$$g = \frac{|f - f_c|}{f_c}$$
$$p = \frac{4f_c}{6.23f_c^2 + 93.39f_c + 28.52} \qquad \text{A-1.4}$$

Ainsi,

$$h(n) = \frac{1}{f_s} \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} W(f) e^{i2n\pi f/f_s} df \qquad \text{A-1.5}$$

$$h(n) = \frac{1}{f_s}\{A + B + C + D\}$$

où

$$A = \int_{f_c}^{\frac{f_s}{2}} \left[\left(1 + p\frac{f - f_c}{f_c} \right) e^{-p\frac{f-f_c}{f_c}} e^{i2n\pi f/f_s} \right] df$$

$$B = \int_{0}^{f_c} \left[\left(1 + p\frac{f_c - f}{f_c} \right) e^{-p\frac{f_c-f}{f_c}} e^{i2n\pi f/f_s} \right] df$$

$$C = \int_{-f_c}^{0} \left[\left(1 + p\frac{f + f_c}{f_c} \right) e^{-p\frac{f+f_c}{f_c}} e^{i2n\pi f/f_s} \right] df$$

$$D = \int_{-\frac{f_s}{2}}^{-f_c} \left[\left(1 + p\frac{-f - f_c}{f_c} \right) e^{p\frac{f+f_c}{f_c}} e^{i2n\pi f/f_s} \right] df$$

$$A = e^{p} \int_{f_c}^{\frac{f_s}{2}} \left\{ \left(1 - p + \frac{p}{f_c} f \right) e^{-\frac{p}{f_c} f} \{ \cos(2n\pi f/f_s) + i \cdot sin(2n\pi f/f_s) \} \right\} df$$

$$B = e^{-p} \int_{0}^{f_c} \left\{ \left(1 + p - \frac{p}{f_c} f \right) e^{\frac{p}{f_c} f} \{ \cos(2n\pi f/f_s) + i \cdot sin(2n\pi f/f_s) \} \right\} df$$

$$C = e^{-p} \int_{0}^{f_c} \left\{ \left(1 + p - \frac{p}{f_c} f \right) e^{\frac{p}{f_c} f} \{ \cos(2n\pi f/f_s) + i \cdot sin(-2n\pi f/f_s) \} \right\} df$$

$$D = e^{p} \int_{f_c}^{\frac{f_s}{2}} \left\{ \left(1 - p + \frac{p}{f_c} f \right) e^{-\frac{p}{f_c} f} \{ \cos(2n\pi f/f_s) + i \cdot sin(-2n\pi f/f_s) \} \right\} df$$

$$A + D = 2e^{p} \int_{f_c}^{\frac{f_s}{2}} \left\{ \left(1 - p + \frac{p}{f_c} f \right) e^{-\frac{p}{f_c} f} \cos\left(2n\pi f / f_s\right) \right\} df$$

$$B + C = 2e^{-p} \int_{0}^{f_c} \left\{ \left(1 + p - \frac{p}{f_c} f \right) e^{\frac{p}{f_c} f} \cos\left(2n\pi f / f_s\right) \right\} df$$

Alors, le résultat de l'intégration est:

$$h(n) = \frac{bc}{a} \left\{ \cos(n\pi) e^{-\frac{c}{b}} \left[k_1 e^{p} - k_2 e^{-p} \right] + \frac{4}{a} c^2 \cos(bn\pi f_c) \right\} \qquad \text{A-1.6}$$

où

$$c = \frac{p}{f_c}, \quad b = \frac{2}{f_s}$$

$$a = c^2 + (bn\pi)^2$$

$$k_1 = p - \frac{c}{b} - \frac{2}{a} c^2$$

$$k_2 = p + \frac{c}{b} + \frac{2}{a} c^2$$

Théoriquement, la réponse impulsionnelle du filtre se basant sur des coefficients $h(n)$ est exprimée par:

$$H(f)_t = \sum_{n=-\infty}^{n=+\infty} h(n) e^{-i2n\pi f / f_s} \qquad \text{A-1.7}$$

Cependant, dans la pratique, la réponse dépend toujours de l'ordre du filtre. En d'autres termes, la réponse impulsionnelle actuelle se représente par:

$$H(f)_p = \sum_{n=-l}^{n=+l} h(n) e^{-i2n\pi f / f_s} \qquad \text{A-1.8}$$

où l est l'ordre du filtre. Donc, l'erreur peut s'évaluer par la formule suivante:

$$E = \left| H(f)_t - H(f)_p \right|$$

A-1.9

$$E = \left| \sum_{n=l+1}^{n=+\infty} h(n) e^{-i2n\pi f/f_s} + \sum_{n=-\infty}^{n=-l-1} h(n) e^{-i2n\pi f/f_s} \right|$$

Comme $h(n)$ est symétrique, l'équation précédente peut s'exprimer comme suit:

$$E = 2 \left| \sum_{n=l+1}^{n=+\infty} h(n) e^{-i2n\pi f/f_s} \right| \qquad \text{A-1.10}$$

Évidemment, plus l'ordre du filtre est élevé, plus l'erreur est faible. Par ailleurs, Nous avons constaté que dans la pratique si l'ordre demeure inchangé, quand la fréquence d'échantillonnage augmente, la réponse devient mauvaise. Ceci signifie que la réponse du filtre est sensible aux changements de la fréquence d'échantillonnage, conformément à l'équation A-1.10

ANNEXE B

CONCEPTION DU BANC DE FILTRES

La conception du banc de filtres comprend deux étapes:

. la distribution des filtres;

. le calcul des coefficients de chaque filtre.

La distribution des filtres (dans le modèle) est linéaire en regard de la répartition des cellules sur la membrane basilaire. En d'autres termes, la différence de la fréquence caractéristique (centrale) de deux filtres consécutifs est constante en échelle de ERB ("Equivalent Rectangular Bandwidth") pour tous les filtres. Si f_{min} et f_{max} représentent la fréquence minimale et la fréquence maximale respectivement en échelle de ERB du banc, la fréquence centrale du premier filtre f_{c1} et la fréquence centrale du dernier filtre f_{cN} peuvent se représenter en échelle de ERB par les formules suivantes:

$$f_{c1} = f_{min} + 0.5$$
$$f_{cN} = f_{max} + 0.5 \qquad \text{A-2.1}$$

car la largeur de bande du filtre est égale à une unité de ERB et que la fréquence centrale du filtre exprimée en échelle de ERB est toujours au milieu de la bande. Alors la représentation universelle de la fréquence centrale du f_{ci} (en échelle de ERB) peut s'exprimer comme suit:

$$f_{ci} = f_{min} + 0.5 + (f_{max} - f_{min} - 1.0)\frac{i - 1.0}{f_n - 1.0} \qquad \text{A-2.2}$$

où f_n est le nombre total de filtres dans le banc et i est le numéro du filtre.

Le calcul des coefficients des filtres peut se réaliser en utilisant les résultats de l'intégration de l'équation A-1.6 à l'annexe A. Nous avons développé un programme qui génère automatiquement les filtres. La figure ci-dessous donne l'architecture du programme.

(1) entrer les paramètres désirés;

(2) vérifier si le nombre de filtres est suffisant pour la plage de fréquence demandée;

(3) produire les coefficients des filtres.

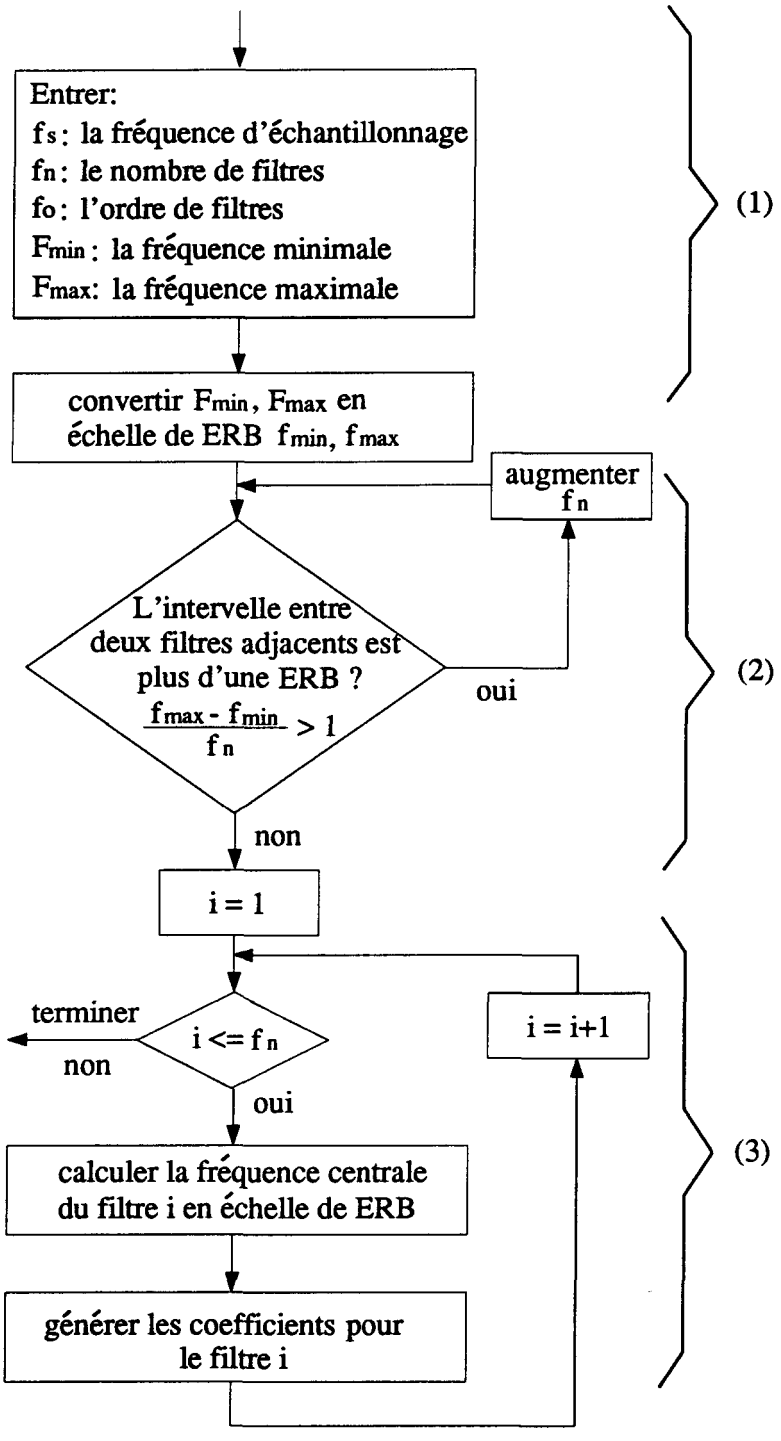

Figure B.1 Processus de génération automatique des filtres

BIBLIOGRAPHIE

[1] Allen, J.B. , "Cochlear modeling", *IEEE Trans. Acoust., speech and Signal Processing Magazine*, Vol.2, No.1, pp.3–29, 1985.

[2] Ambikairajah, E., Black, N.D., et Linggard, R., "Digital filter simulation of the basilar membrane", *Computer Speech and Language*, Vol.3, pp.118–195, 1989.

[3] Aran, J.M., Dancer, A., Dolmazon, J.M., Pujol, R. et Tran Ba Huy, P., "Les modèles d'oreille" dans la *Physiologie de la cochlée*, INSERM/SFA Paris, France, pp.149–168, 1989.

[4] Atal, B.S., "Automatic speaker recognitor based on pitch contours", *Journal of the Acoustical Society of America*, Vol.54, pp.1687–1697, 1972.

[5] Botte, M.C., Canévet, G., Demany, L. et Sorin, C., "Perception de l'intensité sonore" et "Perception de la hauteur tonale", dans la *Psychoacoustique et Perception Auditive*, INSERM/SFA Paris, Frence, pp.25–34, 43–81, 1989.

[6] Caelen, J., "Space/time Data-information in the A.R.I.A.L. project ear model", *Speech Communication*, Vol.4, pp.163–179, 1985.

[7] Cancelli, C. et al., "Experimental results in a physical model of the cochlea", *J. Fluid Mech.*, Vol.153, pp.361–388, 1985.

[8] Chilton, E. et Evans, B.G., "Performance comparison of five pitch determination algorithms on the linear prediction residual of speech", *European Conf. on Speech*

Technology, Edinburgh, 1987.

[9] Chilton, E. et Evans, B.G., "The spectral autocorrelation applied to the linear prediction residual of speech for robust pitch detection", *Proc. ICASSP*, 1988.

[10] Cooke, M.P., "A computer model of peripheral auditory processing incorprating phase-locking, suppression and adaptation effects", *Speech Communication*, Vol.5, pp.261–281, 1986.

[11] Duifhuis, H. et Willems, L.F., "Measurement of pitch in speech: an implementation of Gildstein's theory of pitch perception", *Journal of the Acoustical Society of America*, Vol.71, pp.1568–1580, 1982.

[12] Flanagan, J.L., *Speech analysis, synthesis, and perception*, New York, Springer-Verlaz, 1973.

[13] Fletecher et Munson, "Loudness, its definition, measurement and calculation", *Journal of the Acoustical Society of America*, Vol.5, pp.82–108, 1933.

[14] Geckinli, N.C. et Yavuz, D., "Algorithm for pitch extraction using zero-crossing interval sequence", *IEEE Trans. Acoust., speech and Signal Processing*, Vol.25, pp.559–564, 1977.

[15] Ghitza, O., "A measure of in-synchrony regions in the anditory nerve firing patterns as a basis for speech vocoding", *Proc. ICASSP*, 1985.

[16] Ghitza, O., "Auditory nerve representation criteria for speech analysis/synthesis", *IEEE Trans. Acoust., speech and Signal Processing*, Vol.35, No.6, pp.736–740, 1987.

[17] Ghitza, O., "Auditory neural feedback as a basis for speech processing", *Proc. ICASSP*, 1988.

[18] Girija, Y., "A new model of hearing and it's performance in pitch perception", une thèse doctorale de "University of Delaware", 1985.

[19] Gold, B. et Tierney, J., "Vocoder analysis based on properties of the humain auditory system", un rapport technique de "M.I.T. Lincoln Laboratory", No.670, 1983.

[20] Gold, B. et Rabiner, L.R., "Parallel processing techniques for estimating pitch periods of speech in the Time-Domain", *Journal of the Acoustical Society of America*, Vol.46, pp.442–448, 1962.

[21] Hess, W., *Pitch determination of speech signals (Algorithms & Devices)*, Munich, Germany, 1983.

[22] Howard, I.S. et Walliker, J.R., "The implementation of a protable real-time mutli-lager perceptron speech fundamental period estimator", —, pp.206–209, 1990.

[23] Kaiser, J. F., "On a simple algorithm to calculate the 'energy' of a signal", *Proc. ICASSP*, Vol.1, pp.381–384, 1990.

[24] Kaiser, J. F., "On Teager's Energy Algorithm and its Generalization to Continuous Signals", IEEE Signal Processing Society, 1990 Digital Signal Processing Workshop, Mohonk Mountain House, New Paltz, New York, September, 1990.

[25] Krubsack, D.A. et Neiderjohn, R.J., "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech", *IEEE, Trans. Acoust., Speech, Signal Processing*, Vol.39, No.2, pp.319–329, 1991.

[26] Labat, M. et al., "A spectral autocorrelation methode for the mesaurement of the-fundamental frequency of noise-corrupted speech", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.35, No.6, 1987.

[27] Levitt, H., "Speech processing aids for the deaf: An overview", *IEEE Trans. Audio Electroacoust. (Special Issue on 1972 Conference onSpeech Communication and Processing)*, Vol.21, pp.269–273, 1973.

[28] Licklider, J.C.R., "A dublex theory of pitch perception", *Experientia*, No.7, pp.128–134, 1951.

[29] Lyon, R.F. et Lounette, D., "Experiments with computational model of the cochlea", *Proc. ICASSP*, pp.1975–1978, 1986.

[30] Lyon, R.F., "A computational model of filtering, detection, and compression in the cochlea", *Proc. ICASSP*, pp.1282–1285, 1982.

[31] Markel, J.D., "The SIFT algorithm for fundamental frequences estimation", *IEEE Trans. Audio Electroacoustics*, Vol.20, pp.367–377, 1972.

[32] Martens, J.P. et Immerseel, L.V., "An auditory model based on the analysis of the envelope patterns", *Proc. ICASSP*, pp.401–404, 1990.

[33] Martinez-Alfaro, H. et Contreray-Vidal, J.L., "A robust real-time pitch detector based on neural networks", *Proc. ICASSP*, 1991.

[34] Medan Y. et al., "Super resolution pitch determination of speech signals", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.39, No.1, pp.40–48, 1991.

[35] Miller, N.J., "Pitch detection by data reduction", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.23, pp.72–79, 1975.

[36] Miller, R.L., "Performance characteristics of an experimental harmonic identification pitch extraction (HIPEX) system", *Journal of the Acoustical Society of America*, Vol.47, pp.1593–1601, 1970.

[37] Monderer, B. et Lazar, A., "Detection of speech signals at the output of a cochlear

model", *Proceeding: Twenty-fifth Annual Allerton Conference on Communication, Control and Computing*, pp.182–191, 1987.

[38] Moore, B.C.J., *An introduction to the psychology of hearing*, Academic Press, pp.158–193, 1989.

[39] Moore, B.C.J. et Glasberg, B.R., "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *Journal of the Acoustical Society of America*, Vol.73, No.3, pp.750–753, 1983.

[40] Nguyen, T.D. et al., "A geometric approach to real time pitch detection", *Proc. ICASSP*, 1988.

[41] Noll, A.M., "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection", *Journal of the Acoustical Society of America*, Vol.36, pp.296–302, 1964.

[42] Noll, A.M., "Cepstrum pitch determination", *Journal of the Acoustical Society of America*, Vol.41, pp.293–309, 1967.

[43] Noll, A.M., "Pitch determination of human speech by the harmonic product spectrum,the harmonic sum spectrum and a maximum likelihood estimate", *Proc. Symp. Comput. Processing Commun.*, pp.779–798, 1969.

[44] Paliwal, K.K. et Rao, P.V.S., A synthesis-based method for pitch extraction", *Speech Communication*, Vol.2, pp.37–45, 1983.

[45] Patterson, R.D., "Auditory filter shapes derived with noise stimuli", *Journal of the Acoustical Society of America*, Vol.59, No.3, pp.640–654, 1976.

[46] Patterson, R.D., "Spiral detection of periodicity and the spiral form of musical scales", *Psychology of Music*, Vol.14, pp.44–61, 1986.

[47] Patterson, R.D., "A pulse ribbon model of monaural phase perception", *Journal of the Acoustical Society of America*, Vol.82, No.5, pp.1560–1586, 1987.

[48] Patterson, R.D. et al., "The deterioration of hearing with age: frequency selectivity, thecritical radio, the audiogram and speech threshold", *Journal of the Acoustical Society of America*, Vol.72, No.6, pp.1788–1803, 1982.

[49] Patterson, R.D. et Nimmo-Smith, I., "Thinning periodicticity detectors for modulated pulse streams", *Auditory Frequency Selectivity*, Éditeurs: Moore, B.C.J. et Patterson, R.D., pp.299–307, 1986.

[50] Rabiner, L.R., "On the use of autocorrelation analysis for pitch delection", *IEEE, Trans. Acoust., Speech, Signal Processing*, Vol.25, No.1, 1977.

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

[52] Rose, J.E., "Discharges of single fibers in the mammalian auditory nerve", *Frequency analysis and periodicity detection in hearing*, Éditeurs: Plom, K. et Smoorenburg, G.F., Netherland, pp.172–176, 1970.

[53] Rose, J.E. et al., "Observation on phase-sensitive neurons of anterovertral cochlear nucleus of the cat: nonlinearity of cochlear output", *Journal of Europhysiology*, Vol.37, pp.218–253, 1974.

[54] Rosenberg, A.E. et Sambur, M.R., "New techniques for automatic speaker verification", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.23, pp.169–176, 1975.

[55] Rouat, J., Liu, Y. C. et Lemieux, S., "A non-linear analysis for clean and noisy

speech", *Canadian Acoustics*, Vol.19, No.4, pp.117–118, 1991.

[56] Scharf et Houtsma, "Audition II: Loudress, pitch, localizaiton, aural distortion, pathology", *Handbook of Perception and Human Performance*, Vol.1, Éditeurs: Boff, K.R., Kaufman, L. et Thomas, J.P., Wiley, Now York, 1986.

[57] Schroedar, M.R., "Period histogran and product spectrum: new methods for fundamental frequency measurement", *Journal of the Acoustical Society of America*, Vol.43, pp.829–834, 1968.

[58] Seneff, S., "Real-time harmonic pitch detector", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.26, pp.358–365, 1978.

[59] Seneff, S., *Pitch and spectral analysis of speech based on an auditory synchrony model*, Une thèse doctorale de le M.I.T., 1985.

[60] Slaney, M. et Lyon, R.F., "A perceptual pitch detector", *Proc. ICASSP*, 1990.

[61] Sreenivas, T.V. et Rao, P.V.S., "Pitch extraction from corrupted harmonics of the power spectrum", *Journal of the Acoustical Society of America*, Vol.65, pp.223–228, 1979.

[62] Strum, R.D. et Kirk, D.E., "Nonrecursive filter design", *First principles of discrete systems and digital signal processing*, Éditeurs: Strum, R.D. et Kirk, D.E., Addison-Wesley publishing company, pp.529–562, 1988.

[63] Terhardt, E., "Pitch, consonance and harmony", *Journal of the Acoustical Society of America*, Vol.55, pp.1061–1069, 1974.

[64] Terhardt, E., "Calculating virtual pitch", *Hearing Research*, Vol.1, pp.155–182, 1979.

[65] Un, C.K. et Yang, S., "A pitch extraction algorithm based on LPC inverse filtering

and AMDF", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.25, No.6, pp.565–572, 1977.

[66] Wu, Z.L., *Peut-on 'entendre' des événements articulatoires ? Traitement temporel de la parole dans un modèle du système auditif*, Une thèse doctorale de l'Institute National Polytechnique de Grenoble, 1990.

[67] Zwicker, E. et al., "Critical bandwidth in loudness summation", *Journal of the Acoustical Society of America*, Vol.29, pp.548–557, 1957.

[68] Zwicker, E. et Scharf, "A model of loudness summation", *Psychological Review*, Vol.72, pp.3–26, 1965.

[69] Zwicker, E. et Terhardt, "Analytical expressions for critical-band rate and critrical bandwidth as a function of frequency", *Journal of the Acoustical Society of America*, Vol.68, pp.1523–1525, 1980.

[70] Zwicker, E. et Feldlkeller, R., "L'oreille recepteur d'information", *Psychoacoustique*, Masson, 1981.