16/96

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT D'ELECTRONIQUE

MEMOIRE DE FIN D'ETUDES

THEME

Détection de la fréquence fondamentale par filtrage inverse

Projet de fin d'études en vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

Proposé et dirigé par :

M^elle M. GUERTI

Etudié par :

M^r ABDESSEMED

Mohamed Salah-eddine

PROMOTION Septembre 1996

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT D'ELECTRONIQUE

MEMOIRE DE FIN D'ETUDES

THEME

Détection de la fréquence fondamentale par filtrage inverse

Projet de fin d'études en vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

Proposé et dirigé par :

M^elle M. GUERTI

Etudié par :

M^r ABDESSEMED

Mohamed Salah-eddine

PROMOTION Septembre 1996

DEDICACES :

JE DÉDIE CE MODESTE TRAVAIL À :

♥ À MON DÉFUNT PÈRE QUI RESTERA TOUJOURS VIVANT DANS MON COEUR

♥ À MA TRÈS CHÈRE MÈRE QUI A ÉTÉ LÀ À CHAQUE INSTANT

♥ À MES FRÈRES TAREK ET YOUCEF ET PARTICULIÈREMENT À MA PETITE SOEUR ADORÉE AMINA

♥ À MES AMIS FAYÇAL, KAMAL, MOURAD ET SAÏD ET À TOUS CEUX QUI M'ONT SOUTENU ET QUI ONT CRU EN MOI

Remerciements :

Je tiens à remercier Melle M. GUERTI pour son encadrement et sa présence, les membres de l'administration de notre école pour leur aide et conseils, et tous les professeurs qui ont contribué à ma formation au sein de l'ENP.

SOMMAIRE

Introduction générale

La quasi totalité des phénomènes naturels est pseudo-périodique ; de la rotation de la terre autour du soleil à la rotation de l'électron autour de l'atome ; tout se répète d'une manière plus ou moins cyclique, et cela est bien la preuve même de l'harmonie et de la finesse de ce monde.

Albert EINSTEIN l'a bien compris en étudiant la physique relativiste en énonçant à cet effet : *" DIEU est subtil, mais pas trompeur"*

Cependant, dans la nature rien n'est parfait, en ce sens que tout signal recueilli par un capteur physique, présente un certain niveau de bruit supplémentaire. L'objectif des détecteurs de périodicité, dans ce cas, est d'extraire la fréquence fondamentale (Fo = 1/To) du signal original et d'en éliminer les bruits parasites. Le champ d'application de ces détecteurs est suffisamment vaste pour englober les domaines les plus variés tels que : l'astronomie, la géophysique, la mécanique, la thermodynamique, la physique quantique, le traitement de la parole, le traitement des signaux biomédicaux, la télécommunication en hyperfréquence, l'optique, etc.

La connaissance de cette périodicité ou de cette fréquence fondamentale apporte beaucoup d'informations sur la cause qui l'a produite (quel que soit le domaine d'application).

Certaines applications nécessitent un traitement en temps réel ; c'est le cas du traitement des signaux radar, ou du codage de la parole dans le système LPC. D'autres peuvent, aisément, être traités en temps différé, comme c'est le cas du traitement des signaux cardiaques, et des signaux de l'encéphalogramme, ou même dans le cas de l'analyse mélodique de la parole.

Cependant, ces détecteurs de fondamental se heurtent à différents obstacles, dépendant du domaine d'utilisation, et qui ont pour effet de fausser les résultats obtenus par le détecteur ; on parle alors de rapport signal sur bruit ou de qualité du signal capté.

L'utilisation correcte de notre détecteur de périodicité dans une application donnée, et dans un environnement donné, ne doit pas se faire sans une étude préalable des conditions d'expérimentation, ce qui implique la nécessité de disposer d'une base de données significative, riche, constituée de signaux naturels pseudo-périodiques, et recueillis dans de bonnes conditions d'enregistrement.

Le choix du domaine d'utilisation de notre détecteur s'est porté sur la parole, la base de données est constituée d'une trame d'échantillons de parole de 320 échantillons, prélevés à 8 kHz, prise du manuel d'utilisation du logiciel d'analyse LPC et extraction FO par SIFT [13]. Cette trame est équivalente à 40 ms prise du signal de la parole, d'où l'insuffisance d'informations pour le test de notre détecteur.

Pour remédier à ce problème, on a créé un fichier parole en se basant sur cette trame, et ceci en la reproduisant de façon continue, ce qui a aidé au test de notre détecteur.

Nous avons organisé notre projet de la manière suivante :

- Le chapitre I est un aperçu sur le mécanisme de vibrations des cordes vocales. L'accent est mis sur les raisons de la difficulté d'extraction de FO.
- Le chapitre II traite les différentes méthodes d'analyse de la parole d'une façon brève, on s'attardera sur les techniques de la prédiction linéaire.
- Le chapitre III détaille les différentes techniques de détection de la fréquence fondamentale.
- Le chapitre IV illustre notre travail de recherche.

A la fin du dernier chapitre nous donnerons des conclusions générales sur notre travail.

Et nous terminerons par des références bibliographiques pouvant mieux éclaircir certains points.

Sigles et Abréviations

AMDF	Average Magnitude Différence Function
AR	Auto Régressif
ARMA	Auto Régressif à Moyenne Ajustée
ASSP	Acoustics, Speech and Signal Processing
BBG	Bruit Blanc Gaussien
CLC	Compressed Center Clipper
DOS	Disk Operating System
EAP	Erreur Absolue sur le Pitch
EF	Erreurs Fines
EGG	Electro-Glotto-Graphique
EGV	Erreur Globale de Voisement
EMP	Erreur sur la Moyenne du Pitch
FO	Fréquence Fondamentale
GE	Grosses Erreurs
IBM	International Business Machines
LPC	Linear Prédiction Coding
PCD	Point Critique de déaillance
PDA	Pitch Detection Algorithm
RSB	Rapport Signal sur Bruit
SE	Sans Erreurs
Sgn	Codage Signe
SIFT	Simplified Inverse Filtering Technic
TPZ	Taux de Passage

CHAPITRE PREMIER

LA FREQUENCE FONDAMENTALE, SON ORIGINE ANATOMIQUE ET PHYSIOLOGIQUE. LES PROBLEMES ET OBJECTIFS DE SA DETECTION

I-1/ Introduction

La parole est la faculté de communiquer la pensée par un système de sons articulés [1]; c'est le moyen de communication le plus important entre les humains qui sont les seuls êtres vivants à profiter d'un tel système structuré.

Le support de transmission d'un message vocal c'est l'air et ceci grâce aux fluctuations de sa pression, engendrées, ensuite émises par l'appareil phonatoire et qui sont détectées par l'oreille, laquelle procède à une certaine analyse dont les résultats sont transmis au cerveau, qui les interprète.

Un message vocal est une succession d'images auditives, éléments réduits sans sens, mais dont l'association permet d'avoir les éléments qui constituent les niveaux supérieurs qui sont les syllabes, mots, phrases...,etc.
Un message vocal se caractérise par une très forte redondance, ce qui lui permet de résister aux perturbations du milieu ambiant. La redondance est aussi présente au niveau sémantique, ce qui facilite la compréhension du message par le cerveau.

Avant d'aborder les techniques de détection, nous décrirons succinctement le processus de production de la parole afin de localiser le siège du phénomène qui nous intéresse à savoir la vibration des cordes vocales. Nous préciserons également les caractéristiques de l'onde glottique et dégagerons les principales difficultés qui interviennent lors de la mesure de FO.

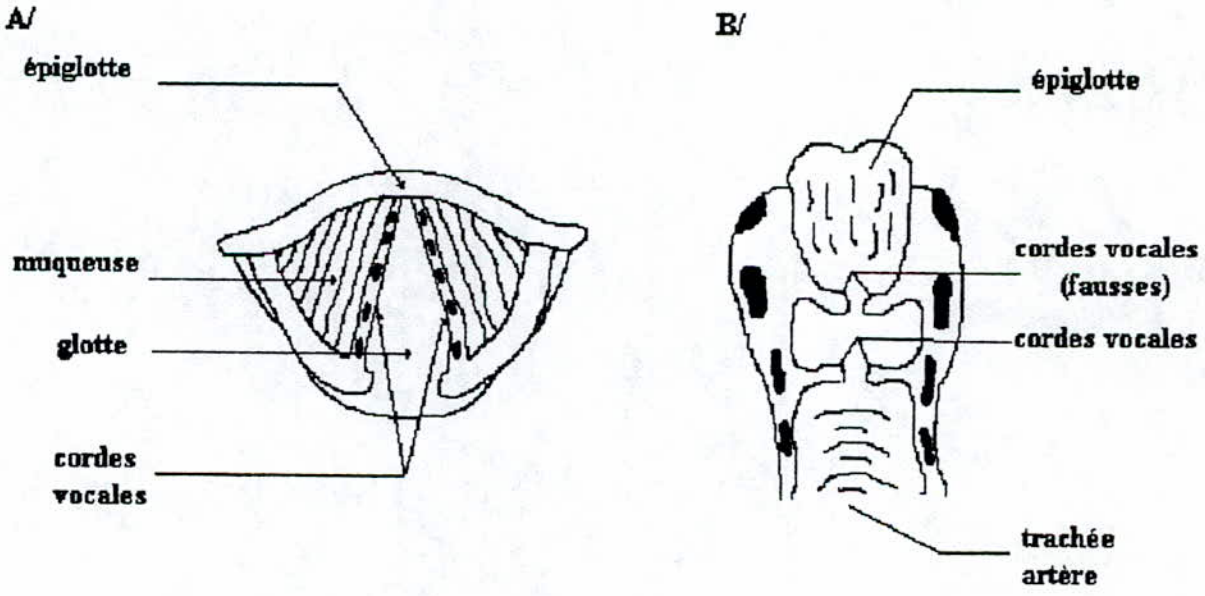
I-2/ Anatomie

Les principaux organes de la phonation sont :

1- Les poumons.

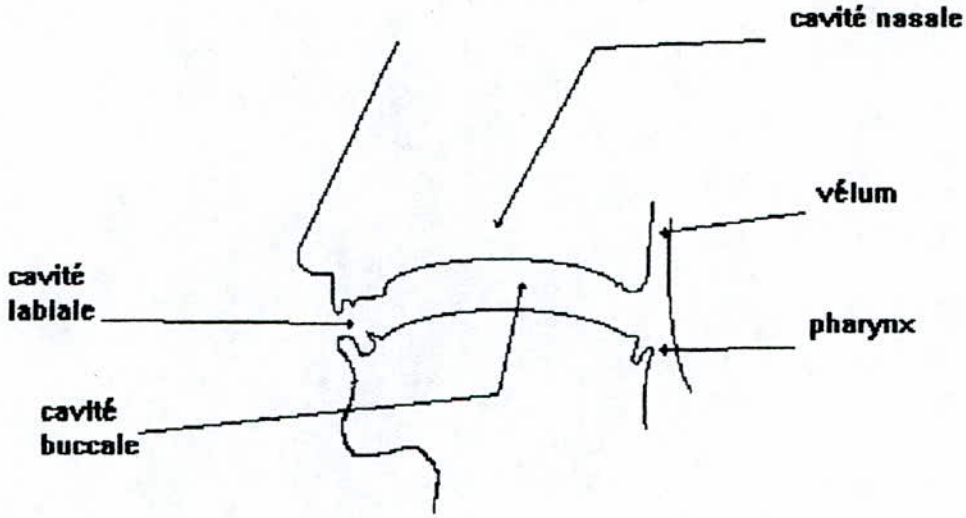
2- La trachée artère et sa partie supérieure le larynx qui supporte deux lèvres symétriques placées en travers appelées cordes vocales, ces lèvres peuvent fermer complètement le larynx et, en s'écartant, déterminer une ouverture triangulaire appelée glotte (figure 1-1)

3- Le conduit vocal est un ensemble de cavités situées entre la glotte et les lèvres qui sont : la cavité pharyngienne, buccale, nasale et labiale. Le muscle mobile qui gère le couplage entre la cavité pharyngo-buccale et la cavité nasale s'appelle le voile du palais ou vélum (figure 1-2) [6].



(A) : vue de haut **(B) : vue verticale**

figure (1-1) : Vues du larynx [6]



figure(1-2) : Le conduit vocal [6]

I-3/ Description du processus de production

Nous examinerons dans cette partie succinctement, la physiologie de l'appareil vocal humain, nous insisterons sur une unité fonctionnelle particulière, celle qui nous intéresse dans ce travail, à savoir le larynx.

Il faut noter cependant, que le phénomène de phonation ne se réalise pas au moyen d'un appareil ou organe spécifique. La production de la parole fait intervenir des organes affectés aux fonctions de respiration et même de nutrition [10].

I-3-1/ Description fonctionnelle

L'appareil phonatoire peut être décomposé en trois parties :

- Le système sub-glottique, qui se compose des poumons et du conduit trachéo-bronchique. Les poumons constituent le générateur de puissance et la trachée artère, l'élément de transmission.

- Le larynx : qui est l'endroit où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Les cordes vocales qui sont deux lèvres symétriques placées en travers du larynx, constituent l'oscillateur de relaxation.

- Le conduit vocal : c'est un système de filtrage formé par les cavités supra-glottiques. Il se compose de deux parties, une partie orale comprenant le pharynx et les diverses cavités buccales et une partie nasale comprenant les

fosses nasales. Ces dernières étant connectées au conduit oral par le voile du palais.

Une soixantaine de muscles participent à la phonation de manière coordonnée.

Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par des sensations cinesthésiques.

Cette complexité anatomique et physiologique permet la production de structures sonores variées. Cependant, seul un sous-ensemble est utilisé de façon pertinente dans la parole [10].

I-3-1-1/ La source vocale : Le LARYNX

C'est un ensemble de cartilages articulés, ligaments, muscles et muqueuses qui surmonte la trachée artère et pénètre dans le pharynx. Il a pour rôle avec l'épiglotte, d'empêcher le passage d'aliment dans la trachée lors de la déglutition. L'examen de la structure interne du larynx nous permet de distinguer les organes essentiels dans le phénomènes de production de la parole, ce sont :

- Le cartilage cricoïde : il sert de support aux cartilages articulés thyroïdes et arythenoïdes

- Le cartilage thyroïde : il peut pivoter autour d'un axe horizontal sous l'action de deux muscles antagonistes. Son échancrure antérieure est décelable au toucher : c'est la "*pomme d'Adam*". L'épiglotte et les cordes vocales sont fixées sur ce cartilage.

- Les cartilages arythenoïdes : ce sont deux petits cartilages symétriques, articulés sur deux facettes du cartilage cricoïde.

- L'os hyoïde : c'est le support flottant du larynx, relié au maxillaire inférieur, il supporte également les insertions des muscles de la langue.

- L'épiglotte : elle est reliée par un ligament à la partie antérieure du cartilage thyroïde.

La forme intérieure du larynx est conique. Au niveau du rétrécissement, se trouve les deux muscles formant les cordes vocales et une ouverture qui est la glotte. Le mouvement de cartilages est coordonné par l'action conjointe de plusieurs muscles. Il faut noter cependant, qu'un seul muscle a tendance à écarter les cordes vocales, alors que plusieurs tendent à les comprimer [10].

I-3-1-2/ Le Larynx dans la fonction de phonation

Il existe différents modes de phonation. La position ainsi que la réaction des cordes vocales permettant de différencier entre ces différents modes, sont essentiellement :
- Le voisement : Les cordes vocales vibrent, les arythenoïdes sont rapprochés.
- L'absence de voisement : Les cordes vocales sont en position écartée et ne vibrent pas.
- L'aspiration : Courte période non voisée qui se produit pendant et immédiatement après un relâchement articulatoire dans les cavités supra glottiques.
- La laryngalisation : vibration d'une partie seulement des cordes vocales, les arythenoïdes sont étroitement rapprochés.
- L'occlusion : Les cordes vocales sont en contact ou assez rapprochées à l'exception des arythénoïdes entre lesquels va naître un bruit ou friction [10].

I-3-1-3/ Fonctionnement des cordes vocales

Les cordes vocales entrent en vibration lorsqu'elles sont appliquées l'une contre l'autre et qu'un excèdent de pression existe au-dessous de la glotte.
Les cordes vocales vibrent sur toute leur hauteur, elles sont le siège de deux ondes de surface progressant de bas en haut comme le montre la figure (1-3).

Nous pouvons dire que les cordes vibrent d'une dépression au niveau intraglottique, s'ajoute à cela l'effet de Bernouilli qui tend à rapprocher les parois jusqu'à une nouvelle dépression.

Il est important de noter que la période de vibration est variable.

Contrairement à un système oscillant ayant une fréquence propre, les cordes vocales ont un mode de vibration dit de relaxation. Celui-ci dépend de plusieurs paramètres induisant ainsi la variabilité interlocuteurs et intra-locuteurs.
Parmi ces paramètres nous citerons entre autres :
- L'anatomie des cordes vocales,
- Le couplage subglottique et supraglottique,
- L'interaction entre les cordes,
- Les propriétés aérodynamiques de l'air qui excite le larynx.

Il est patent de remarquer que ces paramètres sont loin de demeurer constants au cours d'une élocution [10].

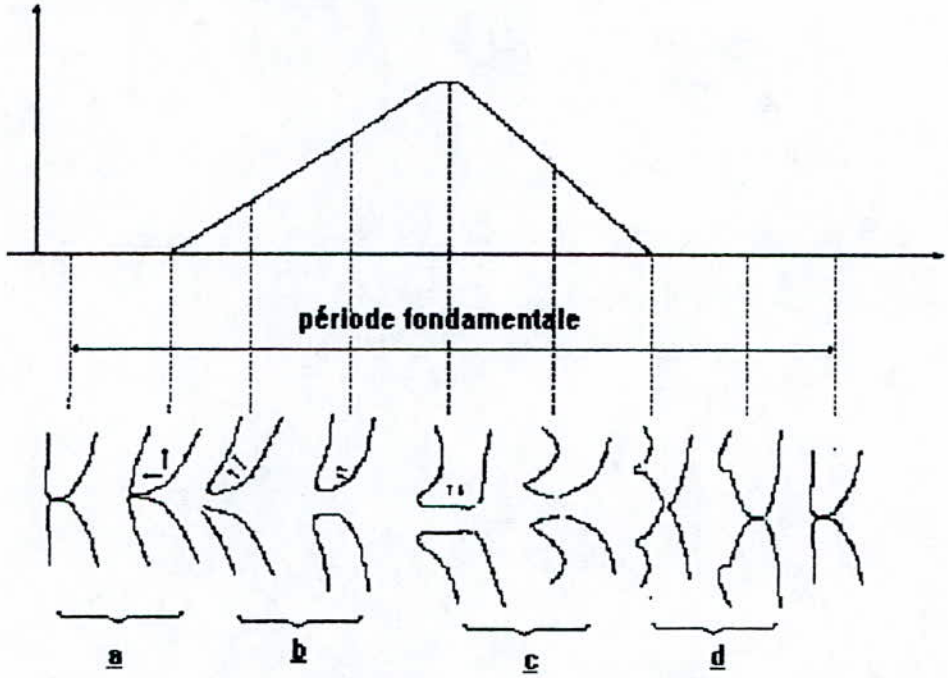


Figure (1-3) : Fonctionnement des cordes vocales [10].

a : Les cordes vocales sont appliquées l'une contre l'autre sur toute leur épaisseur.

b : la pression en dessous de la glotte tend à écarter les cordes tout en les déformant vers le haut ; la zone d'application se déplace vers le haut.

c : Les cordes vocales s'écartent brusquement, libérant au passage de l'air, s'exerce alors la force de rappel des muscles des parois.

d : Les cordes sont de nouveau en contact et le cycle recommence.

I-4/ Caractéristiques de la voix

La voix se caractérise par son intensité, sa hauteur et son timbre.

- L'intensité : correspond à l'amplitude des vibrations sonores ; elle est fonction de la pression d'air expiratoire en amont du larynx et se chiffre en décibels.

- La hauteur : correspond à la fréquence des sons, elle est commandée par l'élasticité de la corde vocale, sa masse et sa largeur. Donc la hauteur est fixée par la fréquence des vibrations des cordes vocales appelée fréquence fondamentale ou pitch, cette fréquence varie de :

* 80 à 200 Hz pour une voix masculine,
* 150 à 450 Hz pour une voix féminine
* 200 à 600 Hz pour une voix d'enfant

- Le timbre : résulte des amplitudes relatives des harmoniques comprises dans le son complexe du larynx et de l'action sélective des cavités de résonance du conduit vocal [2,6].

I-5/ Caractéristiques de l'onde glottique

L'onde glottique générée se présente sous forme d'un signal de type triangulaire possédant un temps de montée, un temps de descente et un temps de relaxation qui correspond à l'accolement des cordes vocales (voir figure(1-4)) [10].

La fréquence des vibrations des cordes vocales est désignée par le terme fréquence laryngienne si on se réfère au processus articulatoire, par le terme fréquence fondamentale FO si l'on se place dans le domaine acoustique. Dans le domaine de la perception on parle de hauteur de la voix.

Le signal glottique obtenue lors d'un voisement a une forme quasi-périodique car il est bien rare que deux impulsions glottiques soient exactement identiques.

Il est illusoire de penser pouvoir disposer du signal glottique pur à cause du couplage conduit vocal-larynx et on ne peut malheureusement pas isoler acoustiquement le larynx.

La forme du signal glottique a pu être étudiée indirectement en proposant des modèles mathématiques du larynx, tel que le modèle à deux masses de Ishizaga et Flanagan, et le modèle à poutres de Perrier [10].

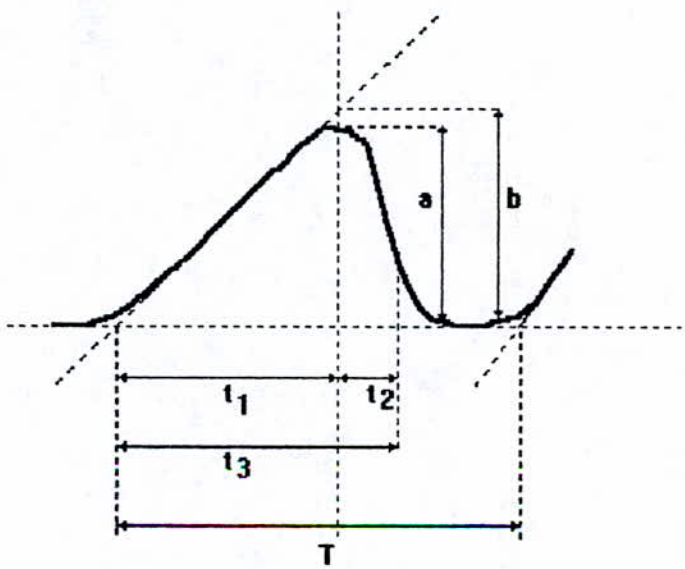


Figure (1-4) : CARACTERISTIQUE DE L'ONDE GLOTTIQUE [10].

I-6/ Relation entre les dimensions des cordes vocales et la fréquence laryngienne

- Deux aspects doivent être soulignés lorsqu'on aborde la relation entre longueur et fréquence de vibration des cordes vocales.
- Les relations directes pour un même locuteur et au cour de la phonation, entre les variations de la longueur de ses cordes vocales et l'évolution de leur fréquence de vibration.

De nombreuses expérimentations ont mis en évidence la corrélation entre la longueur des cordes vocales et la fréquence moyenne de la voix. Un régime d'oscillation plus lent s'établit lorsqu'il s'agit des cordes vocales plus longues, des études et observations, effectuées sur plusieurs locuteurs ont montré que les fréquence laryngiennes sont d'autant plus élevées que le larynx est petit [10].

I-7/ Opposition voisée - non voisée

On peut définir deux catégories de sons selon qu'ils sont dus :

- à une vibration laryngienne périodique : "son sonore ou voix"

ou

- à une génération de bruits à travers une constriction du conduit vocal : "son sourd ou non-voisé".

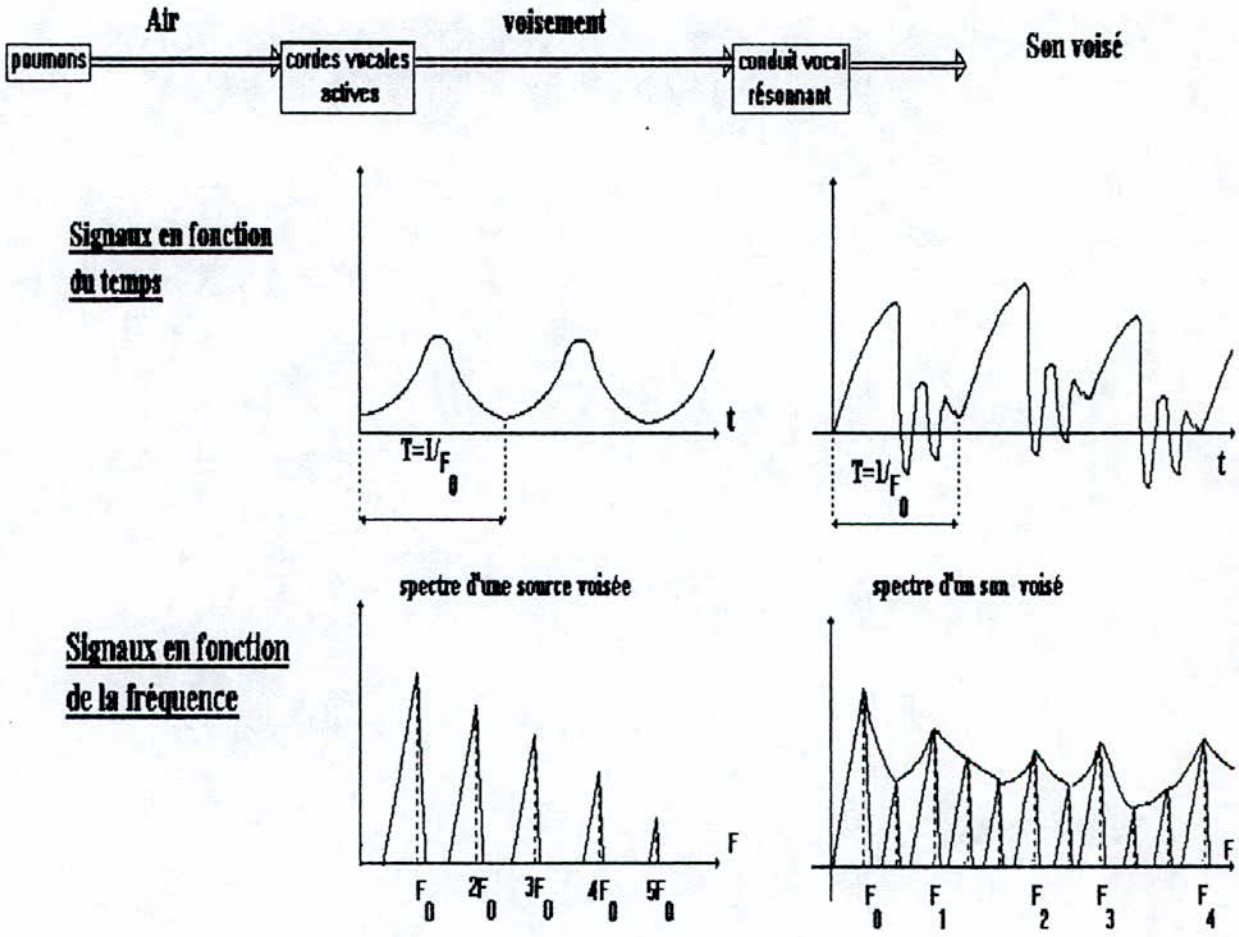
Ces deux genres de sons peuvent être combinés c'est-à-dire que la source de bruit vient s'ajouter à la source d'excitation vocale.

Dans la première catégorie de sons, l'ouverture brusque de la glotte libère la pression accumulée en amont, elle se referme ensuite plus graduellement, la forme du signal glottique est donc sensiblement triangulaire, son spectre est riche en harmoniques et présente une pente de -12 dB/octave, le son résultant dans ce cas est voisé dont la forme est quasi-périodique.

Dans la figure (1-5) on a représenté le spectre d'un son voisé, on y observe les raies qui correspondent aux harmoniques de la fondamentale F_0 (structure de pitch) ; l'enveloppe de ces raies présente des maximums appelés formants et qui correspondent aux fréquences propres F_i (i = 1,2,3,4,.....) du conduit vocal (structure formantique).

Les trois premiers formants sont essentiels pour caractériser le spectre vocal ; les formants d'ordre supérieurs ont une influence plus limitée.

Dans la seconde catégorie, le signal est apériodique, il peut être considéré comme un bruit blanc filtré par la transmittance de la partie du conduit vocal situé entre la constriction et les lèvres ; son spectre ne présente donc pas de structure de pitch et il est relativement uniforme sur une large bande de fréquence (figure 1-6) [12].



figure(1-5) production d'un son voisé[12]

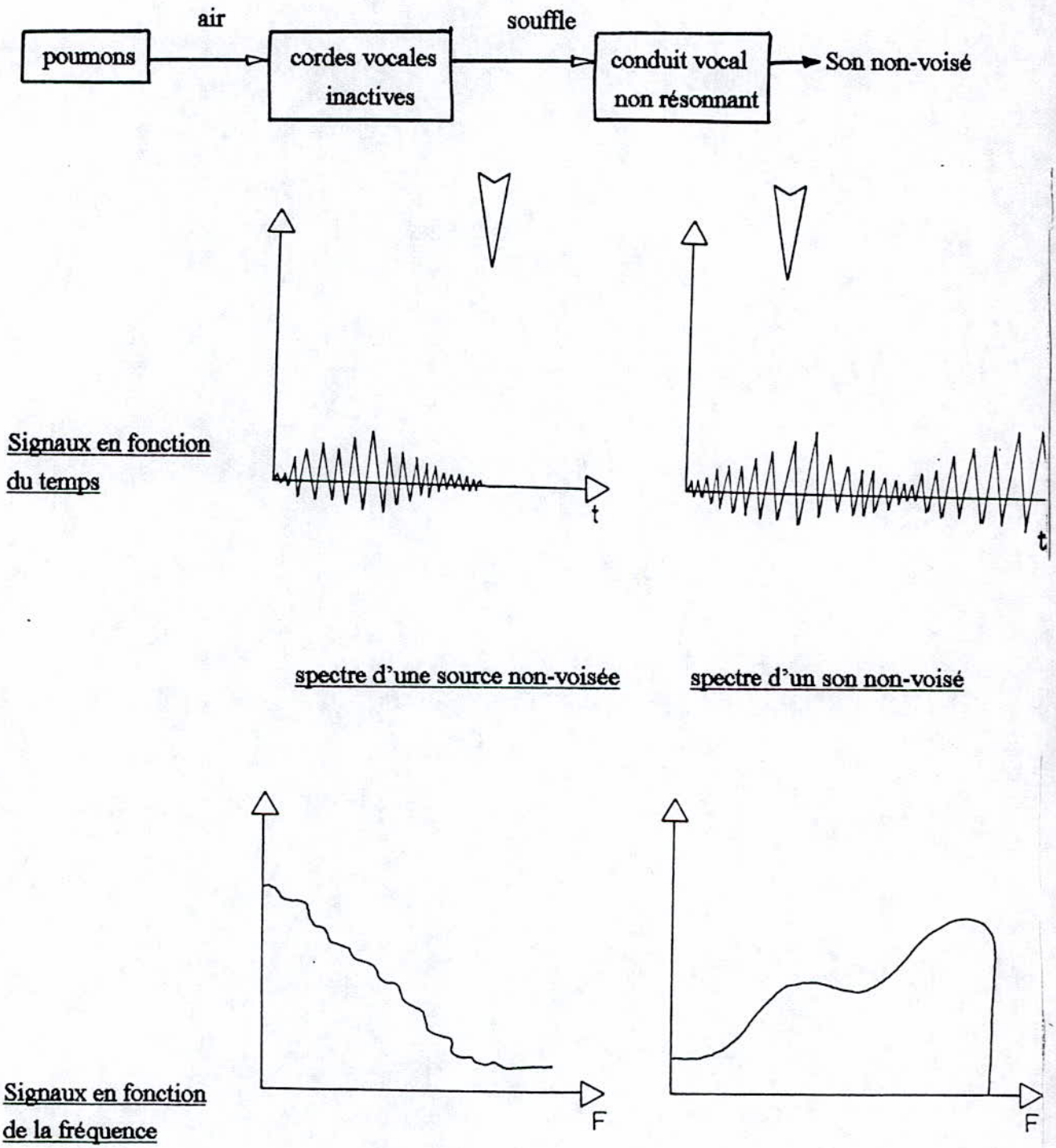


Figure 1-6 : Production d'un son non-voisé [12].

I-8/ Opposition orale-nasale

Le vélum peut se relever à l'horizontale, interdisant toute communication entre les cavités nasales et pharyngo-buccale.

On obtient dans ce cas "un son oral". Dans le cas contraire, si le vélum est abaissé, l'air passe par la cavité nasale, le son ainsi produit est dit "nasal" [12].

I-9/ Classification des sons du langage

Les sons du langage peuvent être classés selon les critères suivants :

- L'opposition sonore-sourde selon que les cordes vocales vibrent ou non.
- L'opposition nasale-orale selon que le vélum est abaissé ou non.
- Le lieu ou le point d'articulation, endroit où la constriction du conduit vocal est maximale.
- Le mode d'articulation qui permet de déterminer la manière de l'écoulement d'air à travers le canal respiratoire lors de la phonation.

La plupart des langues naturelles sont composées à partir de sons distincts, qui sont les phonèmes. Par définition, un phonème est la plus petite unité présente dans la parole susceptible par sa présence de changer la signification d'un mot.

On distingue deux classes liées au mode de production des sons.

1- Les voyelles :

Les voyelles sont caractérisées par un passage libre de l'air la seule source d'excitation du conduit vocal est la vibration laryngienne.

Lors d'une réalisation vocalique, le conduit vocal présente une configuration quasi-stable, certaines voyelles sont dites orales car leur production ne fait pas intervenir la cavité nasale (exemple : /a/,/i/,....), d'autres sont nasales et dans ce cas le conduit nasal et couplé à la cavité buccale et l'émission se produit à la fois par les narines et la bouche (exemple : /ã/,/õ/,....).

2- Les consonnes :

Les consonnes se caractérisent par une construction ou une fermeture soit momentanée complète du passage de l'air.

Ces constrictions peuvent se produire en divers points du conduit vocal.

Les sous-classes des consonnes sont :

- *Les constrictives sourdes* : une constriction importante due à une source bruitée est à l'origine de certaines consonnes telles que /f/,/s/,/
- *Les plosives (ou occlusives) sourdes* : l'interruption de l'écoulement de l'air en un point particulier du conduit vocal provoque l'augmentation de la pression en

amont de l'occlusion puis un relâchement (explosion) brusque est générateur de bruit (exemple : /t/,/k/,...)

Ces deux catégories de consonne peuvent être également voisées, dans ce cas l'excitation est due à un bruit et à une vibration des cordes vocales.
exemples :

- /z/ , /ʒ/, ... sont des constrictions voisées
- /b/, /d/, ... sont des occlusives voisées

- *Les consonnes nasales* : c'est le conduit nasal qui est principalement mis à contribution lors de la production de ce type de son.
- *Les semi-voyelles et les liquides* : Ils ont un comportement assez voisin des voyelles orales [6,12].

I-10/ Position du problème de détection de la fréquence fondamentale par rapport au problème plus général de la prosodie

Les composantes essentielles de la prosodie sont la durée des phonèmes, l'intensité et le contour de la fréquence fondamentale, c'est-à-dire l'évolution de celle-ci.

La prosodie relève de nombreuses disciplines liées à la parole, phonation, phonologie, phonétique, psycho et sociophonétique.

En traitement du signal de parole, les paramètres prosodiques ont une grande importance. En synthèse, ils contribuent à conférer une plus grande intelligibilité du signal synthétique. En reconnaissance, ils peuvent servir dans la segmentation ainsi que pour signaler le type de structure syntaxique et sémantique de la phrase.

La variation des paramètres prosodiques dépend de facteurs aussi divers que l'origine socio-géographique du locuteur, son âge, son sexe, son état émotionnel, des traits relevant du code linguistique telle que la modalité déclarative ou interrogative de l'énoncé ou tout simplement de son agencement syntaxique.

Afin de montrer le rôle que peuvent jouer les paramètres prosodiques aussi bien dans les systèmes de synthèse que dans les systèmes de reconnaissance, une approche consiste à considérer non plus la réalité acoustique de ces paramètres et de surmonter la chaîne de phonation afin de savoir par quels moyens, un locuteur code les informations linguistiques, mais de déterminer quels traits pertinents de ce message l'auditeur retient. C'est l'objectif d'une théorie de la perception auditive. Cette approche trouve sa justification dans le fait que les paramètres acoustiques mesurables objectivement et prosodiques en particulier (fréquence fondamentale, durée,

intensité) qui définissent la qualité d'un son, sont associés aux sensations physiologiques et psychologiques de l'auditeur qui sont respectivement, la hauteur, la durée perceptive et la sonie.

De nombreuses études ont montré que l'ouïe humaine est d'un ordre de magnitude plus sensible aux variations de FO qu'à ceux des autres paramètres prosodiques. Ce qui justifie par ailleurs, la pertinence du choix de ce paramètre, comme substance expressive essentielle dans la majorité des modèles de synthèse ou de reconnaissance [10].

I-11) Mesures objectives et subjectives de la fréquence fondamentale

a) Mesures objectives:

La bande audible se situe entre 20 Hz et 20 kHz. afin de situer le domaine de la parole dans cette bande, rappelons que la fréquence fondamentale moyenne est de 125 Hz pour les hommes, de 250 Hz pour les femmes et 300Hz pour les enfants, avec des variations moyennes autour de ces valeurs de l'ordre d'une octave, voir de deux octaves.

Un son pur est caractérisé par sa fréquence. Un signal aussi complexe que le signal de parole est caractérisé par sa fréquence fondamentale FO et son timbre. On définit des échelles objectives logarithmiques de FO telles que l'Octave et le Ton (1 octave = 6 tons).

b) Mesures subjectives :

Pour les sons purs, une grandeur psychologique a été définie : le MEL qui permet de respecter le rapport subjectif existant entre deux notes séparées d'une octave.

Pour la parole, il faut souligner que la sensation de hauteur dépend des autres paramètres prosodiques. De ce fait, il est très difficile d'estimer une échelle de hauteur. L'interaction entre les composantes fréquentielles rend elle aussi l'estimation de cette échelle subjective difficile.

Il a été défini par contre, un seuil différentiel de fréquence : cette limite dépend de la fréquence FO (ou F pour un son pur) et pour une plus faible part de l'intensité sonore. L'effet de celle-ci peut être négligé entre 40 et 80 dB.

L'appareil auditif est capable de percevoir une variation très faible de la hauteur d'un son. On définit une échelle, le savart, qui divise l'octave en 300 intervalles (1000 log2) . Cette unité représente approximativement, le seuil différentiel moyen à 1 kHz. A titre comparatif, ce seuil est de 0.3 % et 0.4 % entre 250 Hz et 5 kHz, pour la parole ROSSI a estimé ses limites entre 2 % et 0.75 % [10].

I-12) Problèmes rencontrés lors de la détection de la fréquence fondamentale

On a vu précédemment que le paramètre FO est d'une importance considérable dans la plupart des systèmes d'analyse. L'information prosodique est dominée par ce paramètre. Cependant son extraction est loin d'être une tâche facile.

La source vocale (excitation glottale) est variable et irrégulière; le registre sonore est étendu : murmure, chant, cri, sans parler des voix pathologiques. Au premier abord le problème apparaît simple: on ne doit détecter que la fréquence fondamentale d'un signal approximativement périodique. Cependant, il s'avère que la condition préalable que le signal soit approximativement périodique se trouve loin de la réalité lorsqu'il s'agit du signal de parole. Une vibration des cordes vocales n'induit pas automatiquement des périodicités nettes sur le signal observable. Là encore, il est nécessaire de développer des méthodes spécifiques, en y ajoutant lorsque c'est possible, une composante linguistique.

Les principales raisons de cette difficulté de détection sont les suivantes:

1/ Les caractéristiques essentiellement évolutives du signal de parole : le signal de parole est un processus non stationnaire. Il existe des cas extrêmes, celui des transitions d'un son à un autre, caractérisé par un mouvement rapide du conduit vocal et par son influence sur la source.

2/ La fréquence fondamentale peut varier dans une large gamme, d'environ 4 octaves (50 Hz à 800 Hz).

3/ Les variabilités interlocuteurs et interlocuteurs. Celles-ci sont liées aux différences physiologiques, aux effets de la coarticulation et aux latitudes variables de réalisation au plan linguistique.

4/ Il arrive, pour les voix féminines surtout, que F0 coïncide avec le premier formant (ce dernier variant entre 200 Hz et 1400 Hz).

5/ Des irrégularités d'origine non pathologiques peuvent apparaître sur le signal glottique. Le signal excitateur peut tomber temporairement dans un mode irrégulier appelé "Vocal fry". Les intervalles entre impulsions excitatrices consécutives sont irréguliers.

6/ Le problème de détermination de FO peut être vu comme un problème de déconvolution (un signal source qui excite un ensemble de cavités avec différentes fonctions de transfert). Or, on ne dispose que du signal produit (de sortie) et non du signal d'excitation peu commode à évaluer directement : le système à résoudre présente plus d'inconnues que de données.

7/ La détermination de la fonction de voisement est prise généralement indépendamment du détecteur de FO. Une défaillance dans la décision de voisement est une source d'erreur supplémentaire. Il n'existe, jusqu'à présent, aucun dispositif infaillible dans la détermination du voisement.
8/ Pour certains types de sons tels que les consonnes fricatives voisées ou les plosives voisées, l'excitation est une combinaison d'une vibration des cordes vocales et d'un bruit de friction ou d'explosion, ce qui rend la détection très délicate pour ces catégories de sons.

De nombreuses techniques de détection ont été proposées jusqu'à ce jour. Cependant, malgré l'amélioration des processus d'analyse et des moyens de calcul, il est difficile d'en dégager la meilleure technique. En général, une même méthode, n'est pas aussi performante pour les sons stables, les sons brefs, ou les sons transitoires.

La plupart de ces méthodes mettent en oeuvre des algorithmes suffisamment simples pour pouvoir être implantés sur micro-calculateurs.
Plusieurs techniques d'extraction de FO ont été comparées tant au plan de leur rapidité et précision que leur complexité de mise en oeuvre.
Certaines de ces méthodes ont été préférées à d'autres à cause de la facilité d'implantation qu'elles offraient, en respectant les contraintes temps réel. Des techniques bien que souvent plus performantes ont été délaissées en raison de leur complexité de mise en oeuvre. Cependant, un regain d'intérêt leur est accordé aujourd'hui, avec les progrès technologiques récents en matière d'intégration à grande échelle et la baisse relative des coûts de production qu'ils ont entraînés. Ainsi nous pouvons envisager la mise en oeuvre d'algorithmes de plus en plus complexes compatibles avec les objectifs temps réel. C'est ce qu'on va tenter de réaliser dans les chapitres qui suivent [10].

CHAPITRE DEUX

LES DIFFERENTES METHODES D'ANALYSE DU SIGNAL DE LA PAROLE

II-1/ Introduction

Dans ce chapitre nous allons exposer les différentes méthodes d'analyse de la parole et qui sont :

- L'analyse spectrale
- L'analyse temporelle

II-2/ L'analyse spectrale

Dans cette méthode on travaille sur le spectre instantané de la parole étant donné que, la transformation et les séries de Fourier savérent insuffisantes, car le signal voisé, où l'on suppose une certaine périodicité, n'est jamais périodique en réalité car la périodicité d'un signal s'étend sur un intervalle de temps infini.

Donc pour l'analyse spectrale de la parole on utilise:

- Les filtres numériques
- Les analyses de Fourier sur calculateur.

II-2-1/ Les filtres numériques

Le terme " filtre numérique " se rapporte à un système dans lequel un signal, échantillonné et numérisé sous forme d'une suite de nombres x(nT), ou plus simplement x_n (T est la période d'échantillonnage, et n le rang de l'échantillon), est transformé en une seconde suite de nombres y_n, qui représente alors le signal de sortie. Dans le cas d'un filtre linéaire, la relation la plus générale entre les suites x_n et y_n est de la forme :

$$y_n = \sum_{k=1}^{M} a_k y_{n-k} + \sum_{k=0}^{M} b_k x_{n-k} \qquad (2\text{-}1)$$

Le " filtre numérique " est défini par les paramètres $a_1,\ldots,a_M$ et $b_1,\ldots,b_M$.
Lorsque l'un au moins des coefficients a_k est non nul, on obtient un filtre *récursif* : l'échantillon de sortie, à l'instant t = nT, soit y_n, est fonction non seulement des échantillons, actuel (x_n), et passés ($x_{n-1},\ldots,x_{n-M}$), mais aussi des échantillons précédents ($y_{n-1},\ldots,y_{n-M}$) du signal de sortie lui-même.

Le problème à résoudre est le suivant : étant donné un filtre, dont les caractéristiques sont imposées (fréquence centrale s'il s'agit d'un filtre passe-bande, largeur de bande, forme précise de la courbe de réponse, qui peut " tomber " plus ou moins vite en dehors de la bande passante), déterminer les coefficients a_k et b_k tels que si l'on échantillonne le signal x(t) la cadence

N = 1/T, les échantillons y_n obtenus à l'aide de l'équation (2-1) soient le plus voisins possible de ceux que l'on obtiendrait en échantillonnant la sortie du filtre analogique dont on veut faire la synthèse numérique, si on appliquait à son entrée le signal continu x(t) [9].

a/ Réalisation des filtres numériques

Ayant déterminé les coefficients a_k et b_k satisfaisant aux conditions imposées à notre filtre nous pourrons réaliser l'opération de filtrage de l'équation (2-1) de deux manières différentes :

- soit en introduisant les échantillons x_n dans un calculateur universel, cette méthode est très souple, car elle permet de modifier à volonté les valeurs des coefficients, par exemple pour tester le filtrage, toutefois elle présente l'inconvénient de rendre l'ordinateur indisponible à d'autres tâches. En outre les multiplications et additions ne peuvent être effectuées que l'une aprés l'autre ; le filtrage risque d'être long.

- soit à l'aide d'un petit équipement spécialisé permettant le traitement " en temps réel " : dans le cas présent, cela signifie que la sortie des échantillons " filtrés " y_n ne présentera qu'un retard négligeable par rapport à l'entrée des x_n [9].

b/ Avantage des filtres numériques

Les composants entrants dans la réalisation des circuits numériques sont en général beaucoup plus fiables que les circuits analogiques. Les caractéristiques de ces derniers présentent souvent des dérives plus ou moins lentes, dûes par exemple aux variations de température. En outre la précision des éléments peut être nettement plus élevée dans le filtrage numérique , au prix bien entendu d'un accroissement dans la complexité du " matériel " si l'on utilise un équipement spécialisé.
Les matériels correspondants sont enfin plus aisés à simuler sur ordinateur (en vue de leurs mises au point) que les machines analogiques [9].

II-2-2/ Les analyses de Fourier sur calculateur

Cette technique des analyses de Fourier étant des plus courantes et largement diffusée dans la littérature, nous mentionnerons seulement que les transformations de Fourier (TF) peuvent être réalisées sur calculateur, dans des délais très courts permettant le traitement en temps réel, grâce à des algorithmes trés performants (" TF rapides "). Les matériels spécialisés (transformateurs de Fourier) effectuant les TF en dehors de l'ordinateur principal se répandent de plus en plus [9].

II-3/ L'analyse temporelle

Certains événements (par exemple la fermeture brusque du conduit vocal lors de la production d'une plosive) sont mieux caractérisés par l'évolution temporelle du signal que par son spectre.
Nous allons aborder trois des principales techniques qui permettent d'analyser les aspects temporels du signal de la parole en indiquant le cas échéant les rapports entre ces approches et ces techniques d'analyse spectrale [9].

II-3-1/ La fonction d'autocorrélation

Soit s(n) une fonction quelconque numérique. Sa fonction d'autocorrélation R(d) sera définie par :

$$R(d) = \sum_{n=-\infty}^{+\infty} s(n) \cdot s(n-d) \qquad (2\text{-}2)$$

La nouvelle variable d est ce que l'on appelle un " retard ".
La fonction d'autocorrélation peut être calculée sur calculateur à partir du signal numérique s(n) grâce à un ensemble de " retards " [9].

II-3-2/ Passage par zero

Le passage par zéro du signal ou de ses dérivées successives (le signal s(n) prend la valeur zero) ou, ce qui revient à dire, changement de signe du signal s(n). Ainsi, lors de la production des " fricatives ", le taux de passage par zéro par seconde est élevé .

L'information relative à l'amplitude du signal est perdue, puisque l'on ne s'intéresse qu'à son signe. Par contre la méthode est simple à mettre en oeuvre, ne nécessitant pas d'opérations arithmétiques compliquées [9].

La troisiéme technique qui traite les différentes techniques d'analyse prédictives est détaillée ci-dessous

II-4/ Techniques d'analyse prédictive

Lors de la production de la parole le conduit vocal se déforme instantanément, ceci se traduit par une variation aléatoire des paramètres de la transmittance de ce dernier.
La connaissance de ces paramètres joue un rôle important lors de l'analyse ou de la synthèse de la parole, d'où la nécessité de les prédire (voir II-4-2).

La méthode d'analyse par prédiction linéaire ou LPC repose sur l'hypothèse fondamentale selon laquelle un échantillon s(n) du signal de sortie, est une fonction linéaire des 'p' échantillons qui le précèdent, et des 'q+1' échantillons de l'excitation [6,7,9,14].

Nous aurons donc :

$$s(n) = \sum_{i=1}^{p} a_i . s(n-i) + G . \sum_{k=0}^{q} b_k . U(n-k) \qquad (2\text{-}3)$$

II-4-1/ Modélisation de la production de la parole

L'absence de couplage entre la glotte et le conduit vocal permet de modéliser séparément la source et le système de production (voir figure 2-1).

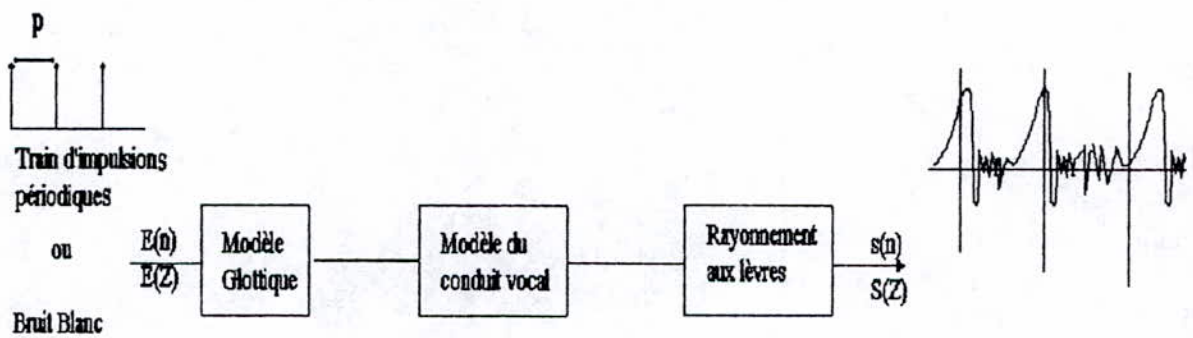


Figure (2-1) : Modéle linéaire de la production de la parole [14]

Pour les sons voisés, la source est un train périodique d'ondes de forme particulière (montée rapide en pression suivie d'une chute plus graduelle).

Ce train d'ondes est modélisé par la réponse d'un passe-bas d'ordre 2 à pôles réels et dont la fréquence de coupure est de l'ordre de 100 Hz comme le montre la figure (2-1) ; la transmittance est de la forme :

$$G(Z) = \frac{A}{(1-\alpha Z^{-1})(1-\beta Z^{-1})} \qquad (2\text{-}4)$$

α, β : coefficients de la transmittance G(Z).
A : gain de la transmittance G(Z)

Pour les sons non-voisés la source est un bruit blanc.

Le conduit vocal peut être assimilé à une succession de tubes acoustiques élémentaires, qui sont modélisés par une cascade de résonateurs dont la transmittance est de la forme :

$$V(Z) = \frac{B}{\prod_{k=1}^{K} (1 - b_{1k}Z^{-1} - b_{2k}Z^{-2})} \qquad (2\text{-}5)$$

b_{1k}, b_{2k} : coefficients de la transmittance V(Z).
B : gain de la transmittance V(Z).

Chaque résonateur correspond à un formant dont la fréquence centrale est donnée par :

$$f_k = \frac{1}{2p} f_s \cos^{-1}\left[\frac{b_{1k}/2}{\sqrt{b_{2k}}}\right] \qquad (2\text{-}6)$$

où f_s est la fréquence d'échantillonnage.

L'ouverture des lèvres représente une charge acoustique ; le rayonnement aux lèvres peut être modélisé par la transmittance :

$$R(Z) = C(1 - Z^{-1}) \qquad (2\text{-}7)$$

En résumé, la tranmittance globale entre le train d'impulsions de la figure(2-1) et le signal émis serait :

$$T(Z) = G(Z) \bullet V(Z) \bullet R(Z) \qquad (2\text{ - }8)$$

$$T(Z) = \frac{\sigma (1 - Z^{-1})}{(1 - \alpha Z^{-1})(1 - \beta Z^{-1}) \prod_{k=1}^{K} (1 - b_{1k}Z^{-1} - b_{2k}Z^{-2})} \qquad (2\text{ - }9)$$

σ : gain de la transmittance T(Z).

Si l'on considère que les deux pôles de G(Z) sont très voisins de l'unité, on obtient la forme simplifié :

$$T(Z) = \frac{\sigma}{A(Z)} \bullet \frac{1}{1 - Z^{-1}} \qquad (2\text{-}10)$$

et on pose :

$$A(Z) = \prod_{k=1}^{K} (1 - b_{1k} Z^{-1} - b_{2k} Z^{-2}) \qquad (2\text{-}11)$$

d'où :

$$A(Z) = 1 - \sum_{k=1}^{2K} a_i Z^{-i} \qquad (2\text{-}12)$$

La transmittance du modèle du conduit vocal $V(Z) = \frac{B}{A(Z)}$ est dite *tous-pôles*, son inverse, le polynôme A(Z) est la transmittance du filtre inverse.

On peut finalement écrire les deux formules suivantes :
- <u>modèle d'analyse</u> :

$$E(Z) = S(Z) \bullet A(Z) \bullet (1 - Z^{-1}) \qquad (2\text{-}13)$$

- <u>modèle de synthèse</u> :

$$S(Z) = \frac{E(Z)}{A(Z)} \bullet \frac{1}{1 - Z^{-1}} \qquad (2\text{-}14)$$

Pour le terme $1\text{-}Z^{-1}$ on a :

- pour l'analyse, on travaillera sur le signal dit "*préaccentué*" :

$$y(n) = s(n) - s(n-1) \qquad (2\text{-}15)$$

C'est à dire qu'on utilisera le modèle :

$$E(Z) = A(Z) \bullet Y(Z) \qquad (2\text{-}16)$$

En fait, l'opération de préaccentuation revient à supprimer les composantes continues, et à remonter les hautes fréquences par rapport aux basses, d'environ 6 dB/octave, ceci rend plus détectables les pics du spectre vers les hautes fréquences.

- pour la synthèse, on désaccentuera le signal obtenu à la sortie du modèle : $Y(Z) = \frac{E(Z)}{A(Z)}$ et on obtiendra ainsi :

$$s(n) = s(n-1) + y(n) \qquad (2\text{-}17)$$

A partir de maintenant, nous omettons volontairement le terme $1\text{-}Z^{-1}$ (à cause de la préaccentuation).
Les limitations de ce modèle sont cependant évidentes :
- En premier lieu, la source est soit un train périodique d'impulsions, soit un bruit blanc, les sons fricatifs voisés ne peuvent pas être produits par ce modèle.
- En second lieu, la production de sons nasalisés fait intervenir deux cavités associées en parallèle ; la transmittance correspondante présente donc des zéros en Z distincts de l'origine.

La transmittance tous-pôles est la base de la modélisation par prédiction linéaire.

Enfin, il est essentiel de rappeler que le signal vocal n'est pas un signal stationnaire : le conduit vocal se déforme d'une façon continue ; les paramètres du modèle sont donc variables dans le temps.
Toutefois, les déformations sont suffisamment lentes pour que les coefficients de T(Z) puissent être maintenus constants pendant des intervalles de temps de l'ordre de 20 ms [6,14].

Remarque : ce modèle de production est appelé Auto-Régressif (AR).

II-4-2/ Equations de prédiction linéaire

On travaillera sur un signal échantillonné et préaccentué s(n).
On veut le modéliser sur la forme :

$$\underbrace{e(n)}_{\text{erreur de prédiction}} = s(n) - \underbrace{\sum_{i=1}^{M} \alpha_i s(n-i)}_{\hat{s}(n)\ :\ \text{signal prédit}}$$

de façon à pouvoir resynthétiser le signal comme le montre la figure(2-2).

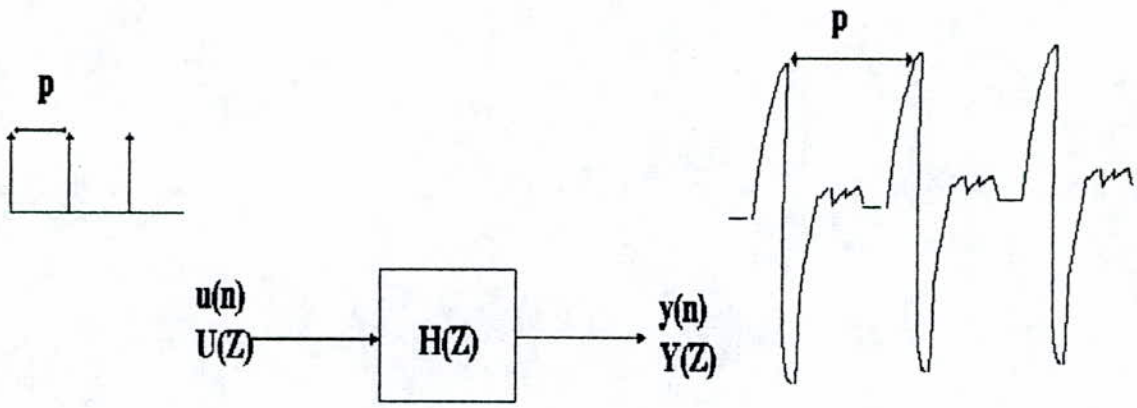


Figure (2-2) : Modèle de modèlisation de la parole [14]

$$H(Z) = \frac{G}{1 - \sum_{K}^{M} a_K Z^{-K}} = \frac{Y(Z)}{U(Z)}; \qquad (2\text{-}18)$$

G: gain du modèle.

Dans le cas idéal, e(n) est nul, sauf au début de chaque période.
Il suffit donc de connaître 2M valeurs consécutives du signal pour calculer les coefficients a_i (i de 1 à M), ces 2M valeurs n'étant pas à cheval sur une impulsion de pitch, on calcule ensuite le gain en connaissant les a_i.
Evidemment, on n'est jamais dans le cas idéal et on doit avoir recours à un critère d'erreur pour calculer les coefficients du filtre.

Tout d'abord, on suppose qu'on travaillera sur une tranche de signal stationnaire, puisque le modèle qui la représente aura des coefficients constants, cette hypothèse est raisonnable la plupart du temps en raison de l'inertie des articulateurs. Le signal de parole peut le plus souvent être considéré comme stationnaire sur des tranches de longueurs de 10 à 30 ms.

Dans ce qui suit on notera par $\left(\sum_{n} \right)$ la somme sur cette tranche de signal.

Le critère d'erreur est choisi quadratique, ce qui est raisonnable et fructueux car il débouche sur des équations linéaires [6,14]:

$$E = \sum_n e^2(n) = \sum_n ((s(n) - \sum_{K=1}^{M} a_K s(n-k))^2 \qquad (2\text{-}19)$$

On calcule les a_i en annulant la dérivée partielle de E par rapport à chacun d'eux :

$$\partial E / \partial a_i = \sum_n -2s(n-i)(s(n) - \sum_{K=1}^{M} a_k s(n-k))$$

comme :

$$\partial E / \partial a_i = 0 \Leftrightarrow \sum_n s(n-i)s(n) = \sum_{K=1}^{M} a_K \sum_n s(n-i)s(n-k)$$

Si on pose :

$$\phi(i,k) = \sum_n s(n-i)s(n-k)$$

il vient :

$$\phi(i,k) = \phi(k,i); i = \overline{1,M}$$

$$\sum_{K=1}^{M} a_K \phi(i,k) = \phi(i,0)$$ Equations de YULE-WALKER

On calcule ainsi la valeur minimale du critère :

$$E_{min} = \sum_n (s(n)^2 - 2s(n)\sum_{k=1}^{M} a_k s(n-k) + \sum_{k=1}^{M}\sum_{j=1}^{M} a_k\ a_j\ s(n-j)\ s(n-k))$$

$$E_{min} = \underbrace{\sum_n s(n)^2}_{\phi(0,0)} - 2\sum_{k=1}^{M} a_K \underbrace{\sum_n s(n)s(n-k)}_{\phi(0,k)} + \sum_{k=1}^{M} a_k \underbrace{\sum_{j=1}^{M} a_j \underbrace{\sum_n s(n-k)s(n-j)}_{\phi(k,j)}}_{\substack{\phi(k,0) \\ \text{d'après YULE-WALKER}}}$$

D'où :

$$\boxed{E_{min} = \phi(0,0) - \sum_{K=1}^{M} a_K \phi(K,0)} \qquad (2\text{-}20)$$

II-4-3/ Résolution des équations de YULE-WALKER

On distingue deux méthodes principales, ces deux méthodes diffèrent par la manière de considérer la tranche de signal à traiter [14]:

-Méthode 1 (dite de covariance)

On ne fait aucune hypothèse sur le signal à traiter. Le critère d'erreur est calculé sur la fenêtre de signal entière de N points :

$$E = \sum_{n=0}^{N-1} e(n)$$

et

$$\phi(i,k) = \sum_{n} s(n-i)s(n-k) \qquad k=0,1,.....,M$$

$$i=1,.....,M$$

On voit qu'on a donc besoin de M échantillons "initiaux", pour n variant de -M à -1.

Le système à résoudre est le suivant :

$$\underbrace{\begin{bmatrix} \phi(1,1) & . & . & \phi(1,M) \\ \phi(2,1) & \phi(2,2) & . & \phi(2,M) \\ . & . & . & . \\ \phi(M,1) & \phi(M,2) & . & \phi(M,M) \end{bmatrix}}_{\phi} \begin{bmatrix} a_1 \\ . \\ . \\ a_M \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ . \\ . \\ \phi(M,0) \end{bmatrix}$$

La matrice ϕ est symétrique. De plus, le calcul montre que :

$\phi(i+1,j+1) = \phi(i,j)+$ un terme - un terme. Ce qui limite le nombre de multiplications pour le calcul de la matrice ϕ.

Mais cette matrice n'a pas les "bonnes propriétés" de la matrice utilisée pour la méthode 2 [14].

-Méthode 2 (dite d'autocorrélation)

On suppose cette fois que le signal est nul en dehors de la tranche considérée : [0,.....,N-1]. Ceci implique la multiplication du signal par une fenêtre qui tend vers 0 aux deux extrémités.

Le choix de la fenêtre se fait par un compromis entre la résolution fréquentielle et temporelle.

Le signal étant maintenant nul en dehors de [0,N-1], on peut écrire :

$$\phi(i,k) = \sum_{n=-\infty}^{\infty} s(n-i)s(n-k) = R(|i-k|) \qquad (2\text{-}21)$$

R étant l'autocorrélation du signal : $R(k) = \sum_{n=0}^{N-1-k} s(n)s(n-k)$.

Les équations de prédiction de YULE-WALKER deviennent :

$$\sum_{k=1}^{M} a_i R(|i-k|) = R(i) \qquad i=1,2,....,M$$

La matrice de ce système est non seulement symétrique, mais de Toeplitz.

$$\begin{bmatrix} R(0) & R(1) & . & . & R(M) \\ R(1) & R(0) & . & . & R(M-1) \\ . & . & . & . & . \\ . & . & . & . & . \\ R(M) & R(M-1) & . & . & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_M \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ . \\ . \\ R(M) \end{bmatrix}$$

On a besoin de (M+1) calculs de corrélations pour écrire le système [14].

* *Calcul du gain G*

Pour le modèle on a : $G.U(n) = s(n) - \sum_{K=1}^{M} a_K s(n-k)$

Pour le signal on a : $e(n) = s(n) - \sum_{K=1}^{M} a_K s(n-k)$.

Par hypothèse on a :

$U(n) = \delta(n)$ où U(n) : bruit blanc stationnaire de moyenne nulle et de variance unité.

Le modèle correspond à la fonction de transfert suivante :

$$H(Z) = \frac{G}{1 - \sum_{K=1}^{M} a_k Z^{-k}} \qquad (2\text{-}22)$$

qui correspond à la série temporelle suivante :

$$h(n) = \sum_{k=1}^{M} a_K h(n-k) + G\ U(n)$$

et h(n) doit être causal :

$h(0) = G$ si $U(n) = \delta(n)$

Le calcul de l'autocorrélation de $h(\tilde{R}(i))$ donne :

si i>0 : $\tilde{R}(i) = \sum_{i=1}^{M} a_k R(|i-k|)$

si i=0 : $\tilde{R}(0) = \sum_{k=1}^{M} a_K R(k) + G\ h(0)$

c-à-d :

$$\tilde{R}(0) = \sum_{k=0}^{M} a_k R(k) + G^2$$

On voit donc que les R(i), i=1,....,M sont solution de la même équation que les R(i). Cette équation étant linéaire, les R(i) sont donc proportionnels aux R(i).

De plus, une hypothèse raisonnable est que l'énergie de h(n) doit être égale à l'énergie du signal de parole : $\tilde{R}(0) = R(0)$.

On voit donc que les (M +1) premières corrélations du signal de parole sont égales aux (M+1) premières corrélations de la réponse impulsionnelle du modèle. De plus :

$$G = \sqrt{R(0) - \sum_{K=1}^{M} a_K R(K)} = \sqrt{E_{min}} \qquad (2\text{-}23)$$

Dans le cas de la parole non-voisée, où U(n) est un bruit blanc stationnaire, de moyenne nulle et de variance unité, on remplace les corrélations par des espérances mathématiques.

En utilisant les propriétés du bruit blanc, on obtient le même résultat que ci-dessus pour le gain [14].

II-4-4/ Solution des équations de YULE-WALKER

1/ Covariance

L'équation est du type Φ-A = Ψ, où Φ est une matrice symétrique définie positive.

-Méthode de Cholesky :

On pose : $\Phi = V.D.V^t$

où V est une matrice triangulaire inférieure, avec des "1" sur sa diagonale principale, et D est une matrice diagonale.
V et D sont déterminés par l'équation de leur définition.

On résoud ensuite en deux temps :
On a l'équation $V.D.V^t.A = \Psi$.
- On pose $Y = D.V^t.A$ et on résoud en Y.

La forme de V rend cette résolution récursive et facile : $V\text{-}Y = \Psi \Rightarrow Y$.
- On résoud ensuite : $Y = D.V^t.A$ c-à-d : $D^{-1}.Y = V^t.A$

Là encore, la forme triangulaire V rend la résolution facile. On obtient donc A [14].

Remarque

$$E_{min} = \phi(0,0) - \sum_{K=1}^{M} a_K . \phi_0(0,K) = \phi(0,0) - A^t - \Psi$$

$$= \phi(0,0) - D^{-1}.Y^t.V^{-1}.\Psi$$

Soit :

$$\boxed{E_{min} = \Phi(0,0) - Y^t.D^{-1}.Y = \Phi(0,0) - \sum_{K=1}^{M} \frac{Y_k^2}{d_K}} \qquad (2\text{-}24)$$

2/ Résolution de Durbin pour l'autocorrélation

On suppose qu'on a les coefficients $\{a_k^{(i-1)}\}$ à l'étape i-1 (k variant de 1 à i-1).

On cherche $\{a_k^{(i)}\}$ (k variant de 1 à i).

A l'étape i :

$$R(j) = \sum_{k=1}^{i} a_k^{(i)} . R(|j-i|)$$

que l'on décompose en :

(1) $$R(j) - \sum_{k=1}^{i-1} a_k^{(i)} . R(|j-k|) - a_i^{(i)} . R(|j-i|) = 0 \quad \text{pour j de 1 à i-1}$$

et

(2) $$R(i) - \sum_{k=1}^{i-1} a_k^{(i)} . R(|i-k|) - a_i^{(i)} . R(0) = 0$$

On pose :

(3) $$a_k^{(i)} = a_k^{(i-1)} - b_k \quad \text{pour k de 1 à i-1.}$$

Le problème revient donc à trouver $a_i^{(i)}$ et $\{b_k\}$ k=1,...,i-1.

On reporte (3) dans (1), on obtient :

(1)' $$\sum_{k=1}^{i-1} b_k . R(|j-k|) - a_i^{(i)} . R(|j-i|) = 0$$

Posons : l = i-j

$$\sum_{k=1}^{i-1} b_k . R(|i-l-k|) = a_i^{(i)} . R(l) \quad , l=1,.....,i-1.$$

On pose maintenant : j=i-k ; j de 1 à i-1

$$\sum_{j=1}^{i-1} b_{i-j}.R(|j-l|) = a_i^{(i)}.R(l)$$

Donc : $b_{i-j} = a_i^{(i)}.a_j^{(i-1)}$

et $b_k = a_i^{(i)}.a_{i-k}^{(i-1)}$

soit $a_K^{(i)} = a_k^{(i-1)} - a_i^{(i)}.a_{i-k}^{(i-1)}$

On reporte (4) dans (2) et on obtient :

$$R(i) - \sum_{K=1}^{i-1}\left[a_K^{(i-1)} - a_i^{(i)}.a_{i-K}^{(i-1)}\right].R(i-K) - a_i^{(i)}.R(0) = 0$$

$$R(i) - \sum_{k=1}^{i-1} a_k^{(i-1)} R(i-k) + a_i^{(i)}\left(\underbrace{\sum_{k=1}^{i-1} a_{i-k} R(i-k) - R(0)}_{E_{min}^{(i-1)}}\right) = 0$$

$$\Rightarrow k_i = a_i^{(i)} = \frac{R(i) - \sum_{K=1}^{i-1} a_k^{(i-1)} R(i-k)}{E_{min}^{(i-1)}}$$

* condition initiale

$E^0 = R(0)$

$k_1 = a_1^{(1)} = \frac{R(1)}{R(0)}$; $a_0^{(1)} = 1$

Et la récursion est initialisée.

* Remarque

$$E_{min}^{(i)} = R(0) - \sum_{K=1}^{i} a_K^{(i)} R(K) = R(0) - \sum_{K=1}^{i-1} a_K^{(i)} R(K) - a_i^{(i)} R(i)$$

$$E_{min}^{(i)} = \underbrace{R(0) - \sum_{k=1}^{i} a_k^{(i-1)} R(k)} - a_i^{(i)}\left(R(i) - \sum_{k=1}^{i-1} a_{i-k}^{(i-1)} R(k)\right)$$

$E_{min}^{(i)} =$ E_{min} $- k_i^2 E_{min}^{(i-1)}$

$$\Rightarrow E_{min}^{(i)} = (1 - k_i^2) E_{min}^{(i-1)} \qquad (2\text{-}25)$$

Comme les E_{min} sont positifs ou nuls, on en déduit que $|k_i| \leq 1$ (et même $|k_i| < 1$ puisque l'erreur résiduelle n'est jamais nulle).

De plus l'erreur résiduelle décroît à chaque pas. On a donc ici le moyen de vérifier la "convergence" et de s'arrêter suivant certains critères.

On démontre de plus que $|k_i| < 1$ est une C.N.S de stabilité pour le polynôme A(Z).

Ce qui nous amène à la rubrique suivante sur les algorithmes en treillis.
Jusqu'à présent, on a procédé en deux étapes : calcul des corrélations puis résolution des équations.

On aimerait pouvoir faire un calcul totalement récursif.

On considère l'algorithme de Durbin [14]:

On a :

$$A^{(i)}(Z) = 1 - \sum_{K=1}^{i} a_K{}^{(i)} Z^{-K} \qquad (2\text{-}26)$$

où $A^{(i)}(Z)$ la transmittance d'ordre (i).

L'entrée du système est :

$$s_n(m) = y(n+m).w(m) \qquad (2\text{-}27)$$

La sortie est :

$$e_n{}^{(i)} = e^{(i)}(n+m). \qquad (2\text{-}28)$$

On laisse tomber l'indice n à partir de maintenant :

$$e^{(i)} = s(m) - \sum_{K=1}^{i} a_K{}^{(i)} s(m-K) \qquad (2\text{-}29)$$

$$\text{et} \begin{cases} E^{(i)}(Z) = A^{(i)}(Z).S(Z) \\ A^{(i)}(Z) = A^{(i-1)}(Z) - K_i\, Z^{-i}\, A^{(i-1)}\,(Z^{-1}) \end{cases} \qquad (2\text{-}30)$$

$$E^{(i)}(Z) = \underbrace{A^{(i-1)}(Z).S(Z)}_{\substack{\text{erreur de prédiction} \\ \text{"aller" d'ordre(i-1)}}} - K_i\, Z^{-1}\, A^{(i-1)}\, S(Z) \qquad (2\text{-}31)$$

On pose :

$$B^{(i)}(Z) = Z^{-i}.A^{(i)}(Z^{-1}).S(Z)$$

$$\Updownarrow$$

$$b^i(m) = s(m-i) - \sum_{K=1}^{i} a_K{}^{(i)}.s(m+k-i) \qquad \text{erreur de pré diction "retour"}$$

$$\Rightarrow \quad E^{(i)}(Z) = E^{(i-1)}(Z) - K_i.Z^{-1}.B^{(i-1)}(Z)$$

On peut donc écrire les deux équations :

$$\begin{aligned} e^{(i)}(m) &= e^{(i-1)}(m) - k_i\ b^{(i-1)}\ (m-1) \\ b^{(i)}(m) &= b^{(i-1)}(m-1) - k_i\ e^{(i-1)}(m) \end{aligned} \qquad (2\text{-}32)$$

Avec :

$$e^{(0)}(m) = b^{(0)}(m) = s(m) \qquad (2\text{-}33)$$

On arrive à la structure d'analyse suivante [figure(2-3)] :

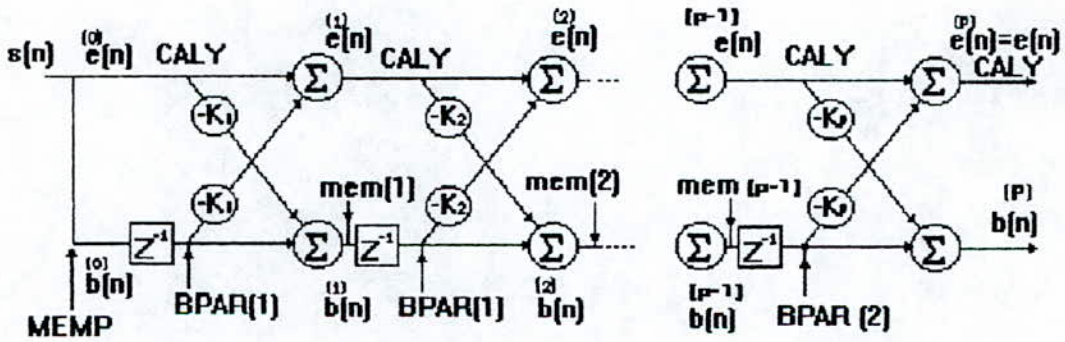


figure 2-3 : filtre en treillis inverse

Où :

CALY, BPAR, BPARS : valeur intermédiaire dans notre programme (voir Chap IV)
MEMP, MEM : mémoires du filtre en treillis inverse (voir Chapitre IV)
s(n) : entrée du filtre en treillis inverse
e(n) : sortie du filtre en treillis inverse

Comme cette structure découle de Durbin on peut calculer les k_i comme dans Durbin.

De plus comme les $a_k^{(i)}$ n'apparaissent pas dans la structure, il serait intéressant de court-circuiter leurs calculs Itakura a montré qu'on peut calculer k_i par :

$$k_i = \frac{\sum_{m=0}^{N-1} e^{(i-1)}(m).b^{(i-1)}(m)}{\sum_{m=0}^{N-1}(e^{(i-1)}(m))^2 \sum_{m=0}^{N-1}(b^{(i-1)}(m))^2} \qquad (2\text{-}34)$$

Cette méthode est dite PARCOR (corrélation partielle).
Notons encore que le problème de la stabilité du filtre est facile à régler, en surveillant les valeurs de $|k_i|$ par rapport à 1 [14].

II-5/ Conclusion

La méthode d'autocorrélation a l'avantage d'être stable

La méthode de covariance travaille sur le signal réel, mais peut être instable.

De ce fait, il est clair que le choix entre les deux méthodes porte sur la méthode d'autocorrélation à cause de sa stabilité.

CHAPITRE TROIS

LES DIFFERENTES TECHNIQUES DE DETECTION DE LA FREQUENCE FONDAMENTALE

III-1/ INTRODUCTION

Le choix d'une méthode de la fondamentale dépend de l'application envisagée et doit de ce fait, tenir compte du paramètre à privilégier.
D'une façon générale, un algorithme est jugé satisfaisant lorsqu'un compromis sur tous les paramètres avantageux est réalisée.
Si l'on se réfère à leur principe de fonctionnement, les méthodes de détection de FO sont classées dans trois catégories principales :

- Les méthodes qui utilisent les propriétés temporelles du signal de parole.

- Les méthodes qui utilisent les propriétés spectrales.

- Les méthodes qui utilisent à la fois les propriétés spectrales et temporelles ou méthodes hybrides.

La majorité des méthodes de détection de FO se déroule en trois phases essentielles : le prétraitement, le traitement et le post-traitement.

♦ Le prétraitement

Ce prétraitement est utilisé pour la mise en forme du signal brut avant la détection. Il comprend généralement :

- un filtrage Passe-Bas de Tchebycheff (d'ordre 3), de 500 Hz. La fréquence très basse de 500 Hz a été choisie afin d'éliminer, le plus possible, l'effet des harmoniques et du bruit additif ;
- un calcul de la puissance moyenne du signal (sur une durée de 20 ms), dans le but de localiser les zones de silence (non énergétiques). Ainsi, si cette puissance moyenne est inférieure à 25 dB (seuil expérimental), on admettra qu'il s'agit d'un silence ;
- le calcul du TPZ, ou "taux de passage par zéro", défini par le nombre de changements de signe du signal vocal sur 20 ms (il a été prouvé, expérimentalement, que les sons voisés changent moins fréquemment de signe que les sons non-voisés), dans le but de localiser les régions non voisées.
  Par conséquent, si ce TPZ est supérieur à 70 (seuil expérimental), on admettra qu'il y a absence de voisement;
- un fenêtrage médian pondéré par la fonction de Hamming [10].

♦ **Le traitement**

Celui-ci représente l'algorithme principal, utilisé par le PDA (Pitch Detection Algorithm), et qui est propre à ce dernier.

♦ **Les post-traitements**

Ces derniers sont constitués du suivi dynamique, des filtrages logiques, et de certains lissages. Le suivi dynamique aide à sélectionner les candidats optimaux pour FO : les filtrages logiques sont utilisés pour corriger les erreurs de détection, et les lissages permettent de raffiner les résultats du pitch.

Il est évident que toute détection de FO ne pourrait se faire que dans une séquence qui contient du signal utile , la détection n'est pas effectuée dans les trames de silence.

Il est à noter aussi que la classification des signaux à analyser en sons voisés et non voisée est primordiale. La détermination de FO n'est faite que dans le cas des sons voisés [10,6].

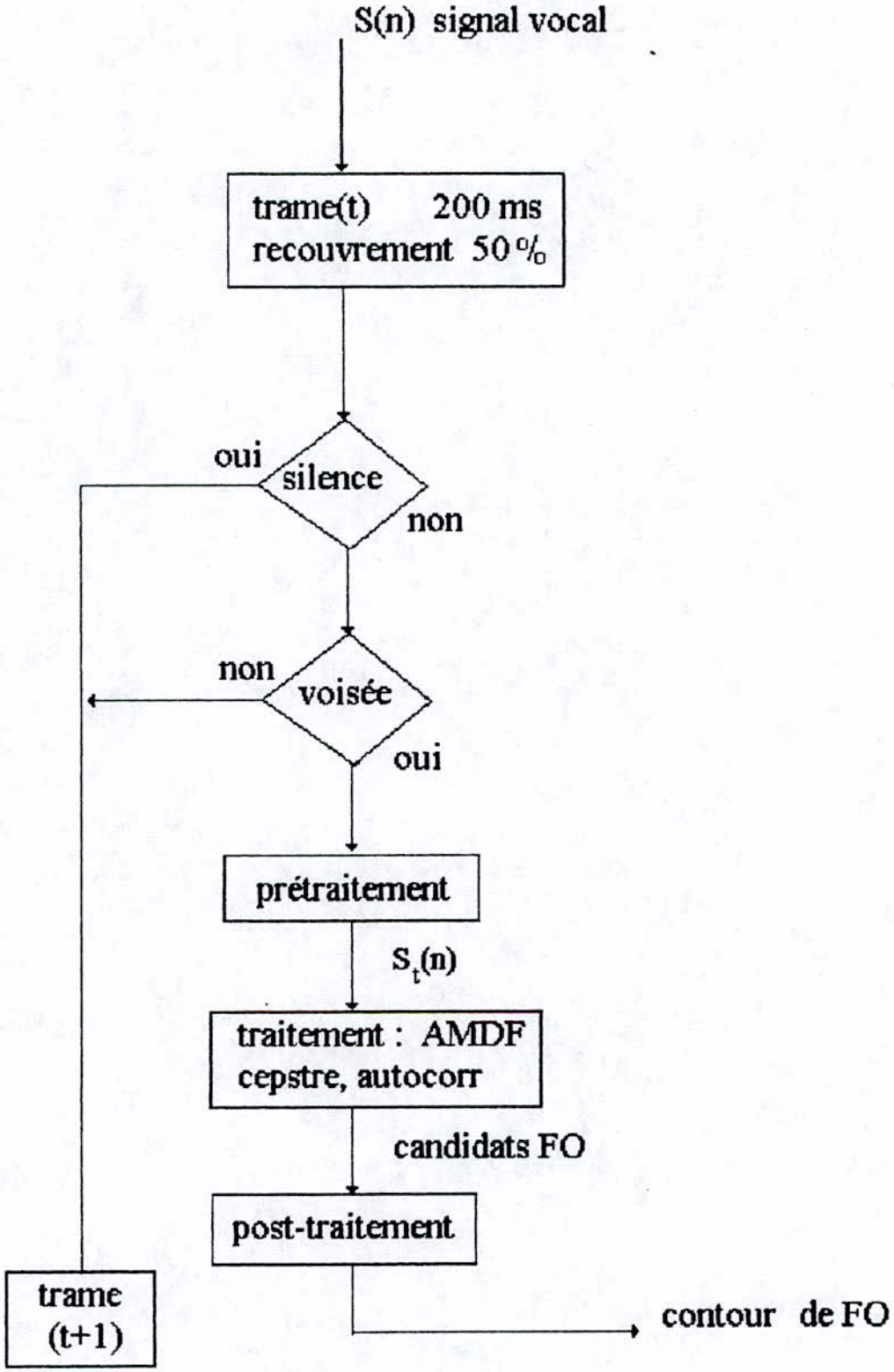


figure (3-1) : METHODES DE DETECTION EGG

III-2/ Méthodes de détection de FO

III-2-1/ Méthode EGG

Appelée EGG par abus de langage, cette méthode est utilisée pour traiter le signal électroglottographique. Elle se veut méthode référentielle, vue la précision atteinte qui avoisine les 100 % .

Les problèmes rencontrés avec cette méthode sont d'ordre purement physiologiques, tels que les mouvements des poumons ou du cou du locuteur ayant pour conséquence une modulation, voire même une perturbation du signal.

De rares erreurs de détection ont été observées au niveau des zones de transition entre deux régions de natures différentes (voisement / non voisement)

Les traitements utilisés par la méthode EGG sont :

1 - filtrage Passe-Haut de 50 Hz du type Tchebycheff, pour éliminer le rythme cardiaque, et respiratoire ;
2 - filtrage Passe-Bas de 1.000 Hz du type Tchebycheff, pour éliminer les harmoniques :
3 - accentuation, pour une synchronisation avec la vitesse des cordes vocales;
4 - filtrage Passe-Bas de 60 Hz du type Tchebycheff, pour limiter la plage de FO ;
5 - choix des pics positifs ou négatifs, selon leur importance ;
6 - calcul du seuil dynamique optimal (Seuil = 0,3 * PicMax) ;
7 - mémorisation de la position des pics d'amplitude supérieure à ce seuil ;
8 - considération des régions sans pics consistants comme non voisées ;
9 - compter le nombre de pics consistants par trame de 20 ms ;
10 - calculer la distance moyenne K entre 2 pics successifs ;
11 - si la distance mesurée K est dans l'intervalle [20, 200], incrémenter alors le compteur de voisement ;
12 - finalement si le compteur de voisement est supérieur ou égal à 3 , alors on considère que c'est une zone voisée et

FO = Fe / K , avec Fe : Fréquence d'échantillonnage ;

13 - répéter ces étapes pour la trame suivante avec un recouvrement de 50 % , sur une fenêtre de 20 ms ;
14 - post-traitements .

Le signal E.G.G est insensible à l'effet des formants et des anti-formants dus à la physiologie du conduit vocal [10,11].

III-2-2/ Méthode cepstrale

Le problème fondamental, lors de la détection de la fréquence fondamentale est la déconvolution de l'excitation du conduit vocal. En effet, le signal microphonique peut être assimilé au résultat convolutif de l'excitation glottale e(t) et du conduit vocal v(t) :

$$s(t)=e(t)*v(t)$$

En d'autres termes, s(t) est bruité convolutivement par v(t).
L'équation précédente donne, après transformée de Fourier, l'équation suivante :

$$S(f)=E(f).V(f) \qquad (3\text{-}1)$$

Le conflit convolutif est devenu multiplicatif. Une méthode ingénieuse, pour isoler E(f), est alors l'emploi d'un filtrage non-linéaire, appelé encore traitement Homomorphique. Dans notre cas on utilisera l'opérateur " CEPSTRE " qui est défini comme suit :

$$\text{Cepstre}[x(t)] = TF^{-1}\{\text{Log}|X(f)|\} = \tilde{X}(q)$$

$$\text{avec} \quad X(f) = TF\{x(t)\}$$

L'unité de la variable cepstrale 'q' est appelé " quefrence ", elle a la même unité de mesure que le temps, et elle est mesurée en secondes.

On peut vérifier aisément, la transformation du produit dans l'équation (3-1) en une somme, par l'adjonction du logarithme :

$$\text{Log}|S(f)| = \text{Log}|E(f)| + \text{Log}|V(f)| \qquad (3\text{-}2)$$

Par transformée de Fourier inverse, on obtient :

$$\mathbf{TF^{-1}}\{\text{Log}|S(f)|\} = \mathbf{TF^{-1}}\{\text{Log}|E(f)|\} + \mathbf{TF^{-1}}\{\text{Log}|V(f)|\}$$

Soit enfin :

Cepstre(s(t))=Cepstre(e(t))+Cepstre(v(t))

La résolution de cette équation ne nécessite qu'un filtrage, dès que l'on admet que les cepstres de e(t) et de v(t) sont isolés.

Heureusement, l'expérience est venue confirmer cette hypothèse, en démontrant que la structure est localisée aux basses quefrences, alors que la structure harmonique est localisée aux hautes quefrences.
Ainsi un fort pic représentant le fondamental est situé au niveau de la zone des hautes quefrences, comme le montre la figure (3-2) [9,6,10,11].

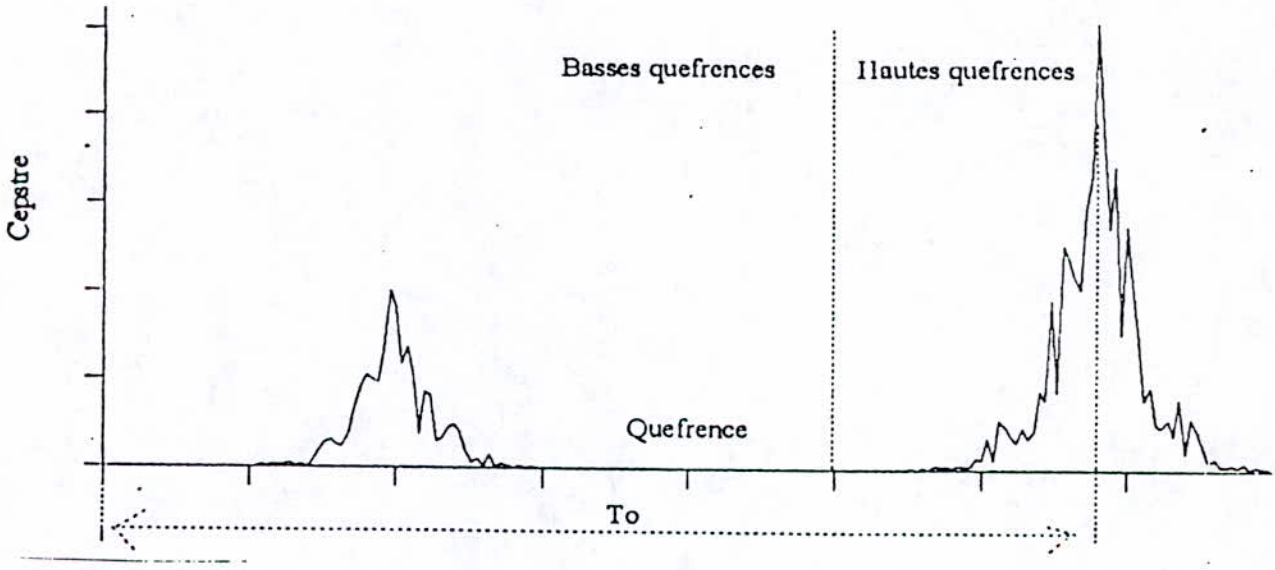


Figure (3-2) : Représentation du cepstre d'un signal de parole

Les étapes de la méthode cepstrale sont :

1- Pré-traitement ;
2- Calcul du cepstre du signal obtenu (Cepstre = Transformée inverse de Fourier du Logarithme du spectre) :

$$\text{Cepstre}[x(t)] = TF^{-1}\{\text{Log}|X(f)|\} = \tilde{X}(q)$$

3- Filtrer les basses quefrences (Filtre P-Haut cepstral);
4- Localiser le plus grand pic;
5- Mémoriser la position du pic (q0);
6- FO sera estimé, alors, par

FO = Fe/q0 où Fe est la fréquence d'échantillonnage;

7- recouvrement à 50 % et passage à la trame suivante (20 ms);
8- Post-traitements [11].

III-2-3/ Méthode de DUBNOWSKY

Cette méthode est une version modifiée de l'autocorrélation classique. Dubnowsky a inventé cette méthode, dans le but d'éliminer l'effet du bruit additif et des formants de basse amplitude qui s'ajoutent au signal vocal. De plus, le codage choisi par Dubnowsky lui confère un temps de calcul très réduit. Il s'agit, bien entendu, du codage SIGNE (sgn).
Rappelons que dans la méthode d'autocorrélation classique, on calcule l'autocorrélation dans une fenêtre de 20 ms, environ, dans laquelle on cherche le plus grand pic correspondant à un décalage kmax, se trouvant dans la bande de fréquence parlée (entre 80 Hz et 400 Hz).
La fréquence correspondante est donnée par FO = Fe/kmax, où kmax est la valeur de k pour laquelle l'autocorrélation passe par un maximum.

La modification apportée par Dubnowsky consiste à coder d'abord le signal de parole par un codage non linéaire du type signe :

$$y(n) = \mathrm{sgn}[x(n)] = \begin{cases} 1, & \text{si..}\, x(n) > \text{Seuil} \\ 0, & \text{si..}\, |x(n)| < \text{Seuil} \\ (-1), & \text{si..}\, x(n) < -\text{Seuil} \end{cases}$$

où seuil vaut : 0,64 x Min(Max1,Max3)
dans laquelle Max1 et Max3 représentent respectivement les maximums d'amplitude, dans le 1er et le 3ème tiers de la fenêtre.

Par la suite, on calcule l'autocorrélation, de la manière suivante :

$$R(k) = \sum_{i=1}^{N-k} y(i).y(i+k) \qquad (3\text{-}3)$$

Ce qui conduit, puisque y(i) prend les valeurs 1,0, ou (-1), à la formule itérative suivante :

Compteur = Compteur + 1 si y(i) = y(i+k)
Compteur = Compteur si y(i) = 0 ou y(i+k) = 0
Compteur = Compteur - 1 si y(i) = -y(i+k)

pour i = 1..N-k.

Et à la fin on pose :

R(k) = Compteur

La suite étant la même que dans la méthode d'autocorrélation classique.
Les étapes de Dubnowsky sont :

1. Prétraitement;
2. Calcul des maximas Max1 et Max3 du 1er et 3ème tiers de la trame;
3. Seuil = 0,4 . Min(Max1,Max3);
4. Ecrêtage et codage SIGNE;
5. Calcul de l'autocorrélation R(k) pour un décalage K entre 25 et 200;
6. Recherche de la position du plus grand pic Kmax et la valeur maximale de l'autocorrélation Rmax;
7. Calcul du coefficient de confiance Cf = Rmax/R(0);
8. Si Cf > 0,5 alors FO = Fe/Kmax, (Fe : fréquence d'échantillonnage), sinon
   Si Cf > 0,15 alors considérer 3 candidats pour FO (les 3 maximas de R(k),
   Sinon
   Si Cf < 0,15 alors considérer la fenêtre comme étant non-voisée;
9. Recouvrement à 50 % et passage à la trame suivante (20ms);
10. Post-traitements

La méthode de Dubnowsky est conseillée dans les applications en temps réel, vue sa rapidité de traitement.
Cependant le double écrêtage et le codage signe apporte une imprécision supplémentaire aux calculs, ce qui se traduit par des glissements de période (dédoublement de la période) [11].

III-2-4/ Méthode d'AMDF

Le terme d'AMDF provient de la littérature anglo-saxonne " Average Magnitude Différence Function " qui signifie Fonction de la valeur moyenne de la valeur absolue de la différence.

Dans cette perspective on admet que la différence d'une fonction avec elle même décalée, caractérise le degré de similitude de celle-ci avec sa courbe décalée.

De plus, si une fonction est périodique, alors ce degré de similitude varie de manière périodique, de période égale à celle du signal lui-même.

Par conséquent, une fonction complexe comprenant n fréquences différentes aura une AMDF comprenant n minimums distincts, voir figure (3-3) [14,10,11].

La formule de l'AMDF est :

$$AMDF(k) = 1/N. \sum_{n=1}^{N-k} |S(n) - S(n+k)|$$

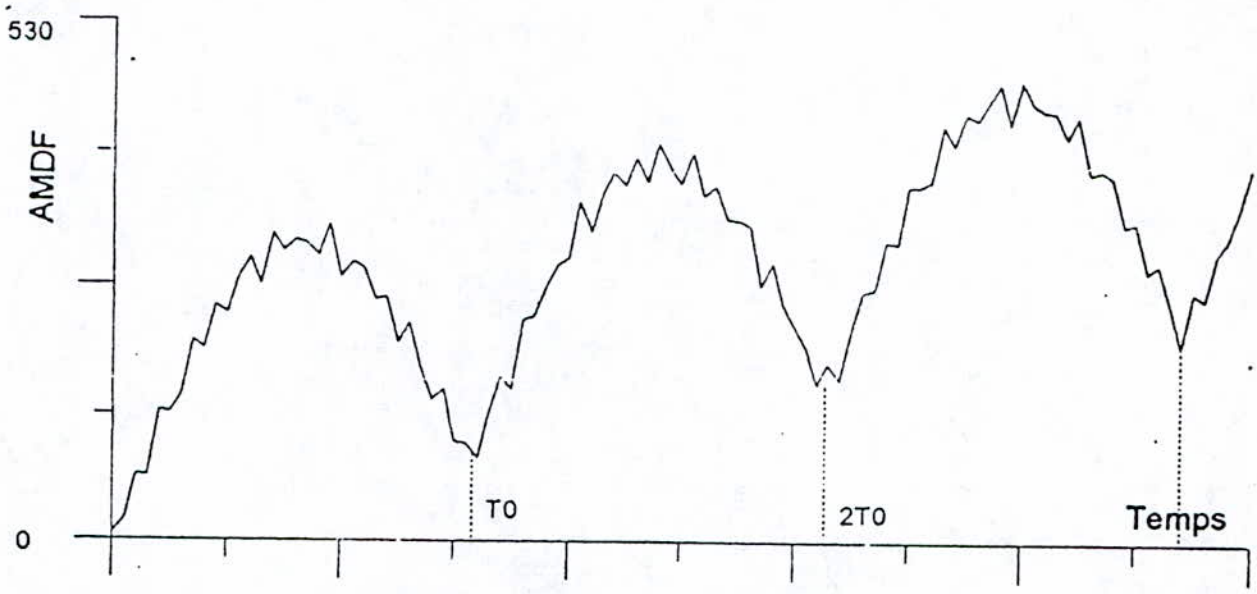


Figure (3-3) : AMDF d'un signal de parole

Dans le cas de la parole, le fondamental est localisé par le plus petit minimum après le zéro (To = kmin/Fe), et FO est donné par :

FO = Fe / kmin où kmin est la position du minimum sur l'axe des décalages.

On démontre que l'AMDF n'est rien d'autre qu'une autre version de l'autocorrélation, ce la était prévisible, vue la liaison avec le degré de similitude.

$$AMDF(n) \approx \beta_n \sqrt{2(R(0) - R(n))} \qquad (3\text{-}4)$$

où R(n) est l'autocorrélation pour un décalage n, et β_n est une constante.

Les étapes de l'AMDF sont :

1- Prétraitement
2- calcul de la fonction d'AMDF;
3- recherche de la position du plus petit minimum kmin;

4- recherche de la valeur du plus grand maximum Max, et celle du plus petit Min;
5- calcul du coefficient de confiance Cf = Max / Min;
6- Si Cf < 250 alors la région est considérée non voisée, Sinon
   Si 250 < Cf < 1200 alors considérer 3 candidats pour FO (position des 3 plus petits pics de l'AMDF), Sinon
      Si Cf > 1200 alors considérer un seul candidat Kmin;
7- FO = Fe / kmin (Fe : fréquence d'échantillonnage);
8- recouvrement à 50% et passage à la trame suivante (20ms);

9- Post-traitements [10,11] .

III-2-5/ LA METHODE DE SONDHI

Cette méthode est issue de l'autocorrélation classique, à laquelle on ajoute un codage non linéaire du type CLC " Compressed Center Clipper ". D'aprés Sondhi; le signal de parole est mieux libéré de sa structure formantique (1er formant), si les déformations générées par ces formants sont partiellement éliminées, et sachant que la plus grande partie affectée par cette déformation est la partie de basses amplitudes.
Il est plus interessant de ne garder que la partie des hautes amplitudes; la raison qui a porté Sondhi à choisir le codage CLC qui n'est rien d'autre qu'un filtre passe haute d'amplitude.

Codage CLC :

Soit une fonction x(n) quelconque,

$$y(n) = CLC[x(n)] \text{ si et seulement si}$$

$$y(n) = \begin{cases} x(n) - Seuil & \text{si } x(n) > Seuil \\ 0 & \text{si } |x(n)| < Seuil \\ x(n) + Seuil & \text{si } x(n) < -Seuil \end{cases}$$

Seuil est pris comme étant une fraction fixe du maximum de x(n) sur son domaine de définition. Il représente le niveau moyen du signal au-dessus duquel les amplitudes de ce dernier ne sont pas trés déformés par le conduit vocal.

La forme générale de la fonction d'Ambiguïté est donnée par la formule suivante :

$$A(r,f) = \int_{-\infty}^{+\infty} x(t).x^*(t-r)\ e^{-2\pi jft}.dt \qquad (3\text{-}5)$$

où r représente le paramètre temps en secondes, f représente le paramètre fréquence en Hertz, et x(t) le signal temporel à étudier.

Propriétés de la fonction d'ambiguité

a/ Si f = 0 alors :

$$A(r,0) = \int x(t)\ .\ x^*(t-r)\ .\ dt = R(r)$$

R(r) est, alors, l'autocorrélation de x(t).

Ce qui implique que si le signal x(t) est T0 périodique alors sa fonction d'ambiguité en f=0, possédera un maximum de similitude et sera maximale pour tout r multiple de T0.

b/ Si r = 0 alors :

$$A(0,f) = \int |x(t)|^2 . e^{-j2\pi ft} . dt$$

soit

$$A(0,f) = T.F\left\{|x(t)|^2\right\}$$

Mais nous savons que si x(t) est de période T0, sa T.F. est maximale pour f = F0 = 1/T0 ; or si x(t) est périodique (complexe), alors $|x(t)|^2$ le sera aussi. Donc T.F. sera maximale pour f = F0, de la même manière. La fonction d'ambiguité presentera, alors, un maximum en f = 1/T0.

Cas d'un signal périodique (de fréquence F0)

On démontre, que pour une fréquence f égale à la fréquence de périodicité F0 (f = F0), la fonction d'ambiguité prend la forme suivante :

$$A0 = A(r,f) = e^{-j\pi r.F0}.\delta(fi - F0).\sum C_n.C_{n-1}^{*}.\cos[(2n-1).\pi.rF0 \qquad (3\text{-}6)$$

C_n représente le 'n' iéme coefficent de Fourier, complexe, de x(t), considéré de période T0.

Ceci nous aménе à considérer différents cas :

1er cas :

Si r = 1/F0 alors :

$$A0 = A(1/F0,F0) = \delta(f - F0)\sum_{n=2}^{N} C_n.C_{n-1}^{*}$$

de même si r = 0

$$A0 = A(0,F0) = \delta(f - F0)\sum_{n=2}^{N} C_n.C_{n-1}^{*}$$

Donc pour r = 1/F0 ou r = 0, et f = F0, on a A(r,f) est alors maximale.

2ème cas :

Si r = 1/(2F0) (cas du 1er harmonique), alors :

$$A(1/(2F0), F0) = 0$$

donc pour r = 1/(2F0) et f = F0 le module de A(r,f) est minimal.

3ème cas :

Pour tout f multiple de F0 (f = p.F0), on démontre que l'expression de A(r,f) s'annule pour r = 1/(pF0).

Application à la parole

L'équation (3-6) offre, contrairement à la fonction d'autocorrélation, une triple condition, en f = F0.

D'une part $\Rightarrow |A(1/F0,F0)|$ est maximal

$\Rightarrow |A(1/(2F0),F0)|$ est minimal,

et d'autre part $\Rightarrow |A(0,F0)|$ est maximal

Si plusieurs maxima de la fonction d'autocorrélation sont candidats à F0 (par la méthode d'autocorrélation classique de détection), il est certain qu'un seul de ces maxima vérifiera la triple condition ci-dessus.
Ces conditions conférent un caractére de vérification trés sûr,pour améliorer la précision du détecteur mélodique surtout dans le cas des sauts d'octaves, causés par des harmoniques.
La formule discréte utilisée en programmation est :

$$A(k.Te,l.Fe/N) = A(k,l) = \sum_{n=0}^{N-1} X(n).X(n-k).\exp(-j.2.\pi.n.l/N) \qquad (3\text{-}7)$$

où N est le nombre d'échantillons et Te le pas d'échantillonnage [6].

Le chercheur S.A SELOUANI a mis au point une nouvelle méthode qui est une approche de la méthode d'ambigûité classique et baptisée :

" METHODE D'AMBIGUITE MODIFIEE "

Dans cette nouvelle approche, et dans le but d'accélérer les calculs et de réduire sa compléxité, l'auteur a procédé à un codage spécifique de type SIGNE (sgn), défini par :

$$Sgn[x(n)] = \begin{cases} 1, si... x(n) > seuil \\ 0, si... |x(n)| < seuil \\ -1, si.. x(n) < -seuil \end{cases}$$

ainsi la formule de l'ambiguîté devient :

$$A(k,l) = \sum_{n=0}^{N-1} Sgn[x(n)].Sgn[x(n-k)].\cos(2\pi n.l/N) - j.\sum_{n=0}^{N-1} Sgn[x(n)].Sgn[x(n-k)].\sin(2\pi n.l/N)$$

Par conséquent, le nombre de multiplications (au nombre de 4(k+1)(N-k/2)); qui était exigé par la méthode classique, se trouve transformé (dans cette nouvelle approche) en des additions et des décalages (au nombre de 4(k+1)(N-k/2)); ceci représente un gain en temps considérable pour des applications fonctionnant en temps réel.

Les étapes de traitement dans la méthode d'ambiguîté modifiée, sont :

1- Mémorisation de la table des cosinus et des sinus.
2- pré-traitement
3- codage Signe
4- calcul de l'autocorrélation R(k) par la méthode des cepstres

5- détection des trois pics les plus importants k1, k2 et k3 de R(k)
6- sélection du pic qui vérifie les trois conditions de l'ambiguîté suivantes

a-	AMBIG(ki,N/ki)	maximale
b-	AMBIG(0,N/ki)	maximale
c-	AMBIG(ki/2,N/ki)	minimale

7- F0 = Fe/ki (Fe fréquence d'échantillonnage)
8- recouvrement à 50% et passage à la trame suivante (20 ms)
9- post-traitements [10].

III-2-7/ Méthode du SIFT

Cette méthode est à la base de notre application, elle est plus détaillée au chapitre IV, mais nous allons quand même citer les étapes de cette méthode :

1/ Pré-traitement
2/ Sous-échantillonnage à 2 KHZ (décimation) et filtrage anti-décimation
3/ Fenêtrage
4/ Calcul des paramétres AR, a1, a2, a3 et a4 par la méthode de BURG
5/ Filtrage Passe-Bas à 600 Hz
6/ Filtrage du signal décimé par le filtre inverse, en utilisant la formule :

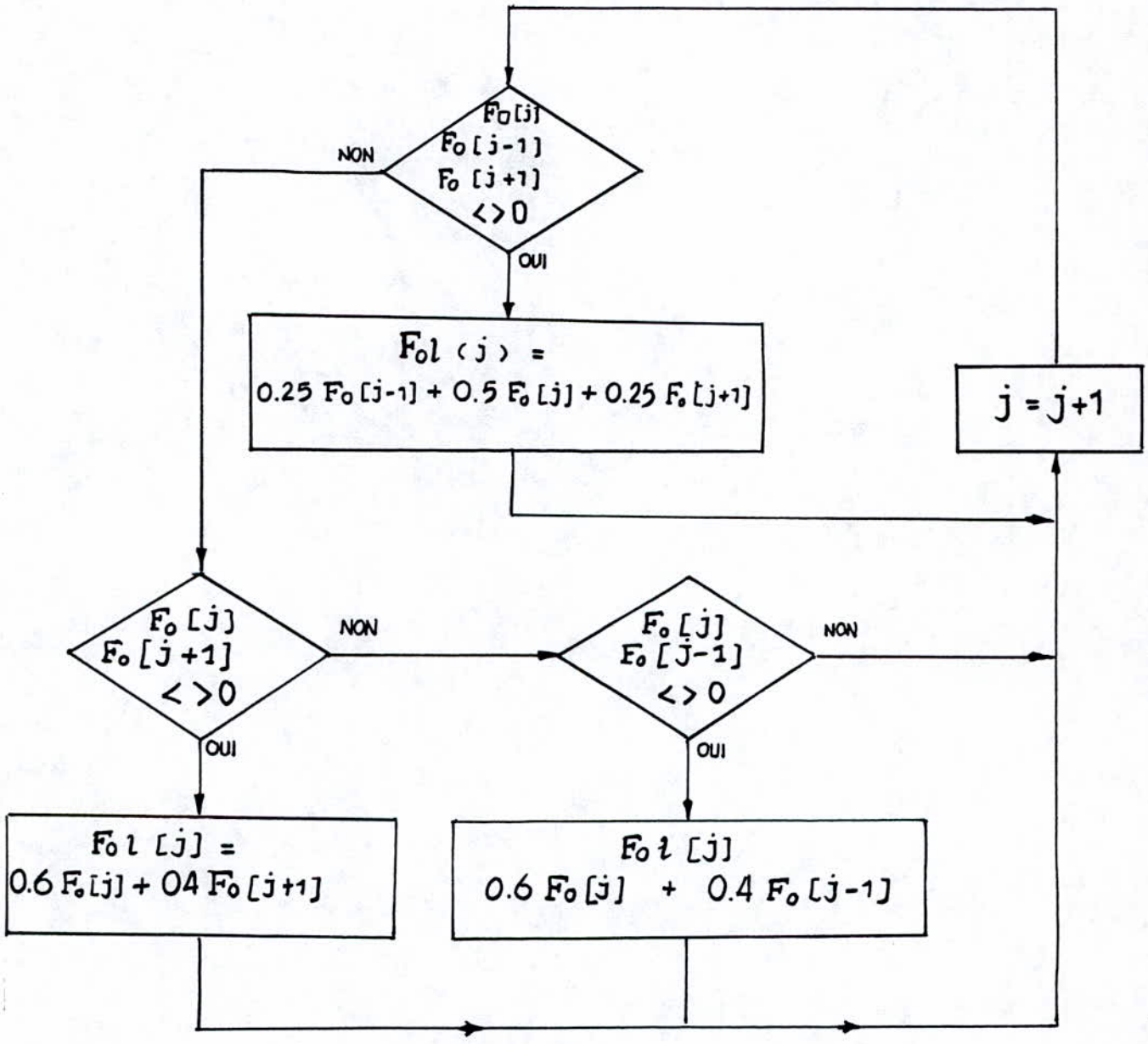
$$y(k) = \sum_{i=0}^{4} a(i).x(k-i) \qquad (a(0) = 1)$$

7/ Calcul de l'autocorrélation de y(k) soit R(k)
8/ Faire une interpolation parabolique d'ordre 3 pour sur-échantillonner R(k), et une expansion à 1/5
9/ Localisation de la position du plus long pic soit kmax
10/ F0 = Fe/kmax (Fe : fréquence d'échantillonnage)
11/ Recouvrement à 50% et passage à la trame suivante (20 ms)
12/ Post-traitements [6,11].

III-3/ Types de post-traitments

Le post-traitement est un filtrage logique dont le rôle consiste à corriger d'éventuelles erreurs de détection en tenant compte des résultats antérieurs.
Les techniques actuelles s'orientent de plus en plus vers cette démarche car elle s'est avérée efficace pour la résolution du probléme de sauts d'octaves.
Il existe trois types de post-traitements qui sont :
- Les techniques de suivi (dynamique, moyenneur, de continuité).
- Les filtrages logiques (anti-harmoniques, ilôts, etc.)
- Les lissages.

Nous allons pas nous étaler longuement sur ses différents types de post-traitements mais nous mettons l'accent sur le lissage médian car il est la base de la méthode de correction de notre application. Nous donnons ci-dessous son algorithme (figure 3-4) [10,11].



F0[j] : valeur de la fréquence fondamentale sur la fenêtre j
FOL : valeur lissée de la fréquence fondamentale sur la fenêtre j
nb : nombre de fenêtres considérées

Figure (3-4) : Organigramme du lissage médian

III-4/ Etude comparative

Tout d'abord commençons par définir les taux des trames sans erreur, ce taux représente le nombre de trames dont la valeur du fondamental détecté, est, soit correcte, ou soit dans un intervalle d'erreur n'éxcédant pas 5 %, c'est à dire :

$$\frac{\left|F0[i] - F0_{réf}[i]\right|}{F0_{réf}[i]} < 5\% \qquad (3\text{-}8)$$

F0[i] : valeur mesurée du fondamental.
$F0_{réf}[i]$: valeur référence du fondamental

Le classement des différents détecteurs vus dans ce chapitre, par ordre de précision décroissante, est le suivant [11]:

METHODES	Taux de trames SANS ERREUR %
1. Sondhi	96,7
2. AMDF	96,2
3. Ambiguïté	95,7
4. Dubnowsky	95,0
5. SIFT	89,8
6. Cepstrale	88,8

Tableau 3.1 : Classement des différents PDA (en chambre sourde)

III-5/ Conclusion

Nous avons entrepris l'étude de différents types de détecteurs de périodicité que l'on peut classer ainsi :

-Deux méthodes spectrales (SIFT, et la méthode cepstrale)

-Trois méthodes temporelles (Dubnowsky, AMDF, et la méthode de Sondhi), qui possédent, chacune d'elles son propre codage.

-Une méthode mixte " ambiguîté modifiée " dont la double analyse temps-fréquence devrait lui donner une meilleure finesse de détection.

CHAPITRE QUATRE

SIMULATION D'UN DETECTEUR DE FREQUENCE FONDAMENTALE PAR LA METHODE DE FILTRAGE INVERSE

IV-1/ Introduction

Le but de ce chapitre est d'élaborer un programme de détection de la fréquence fondamentale en se basant sur la méthode SIFT.
Les différents blocs de notre programme sont les suivants :

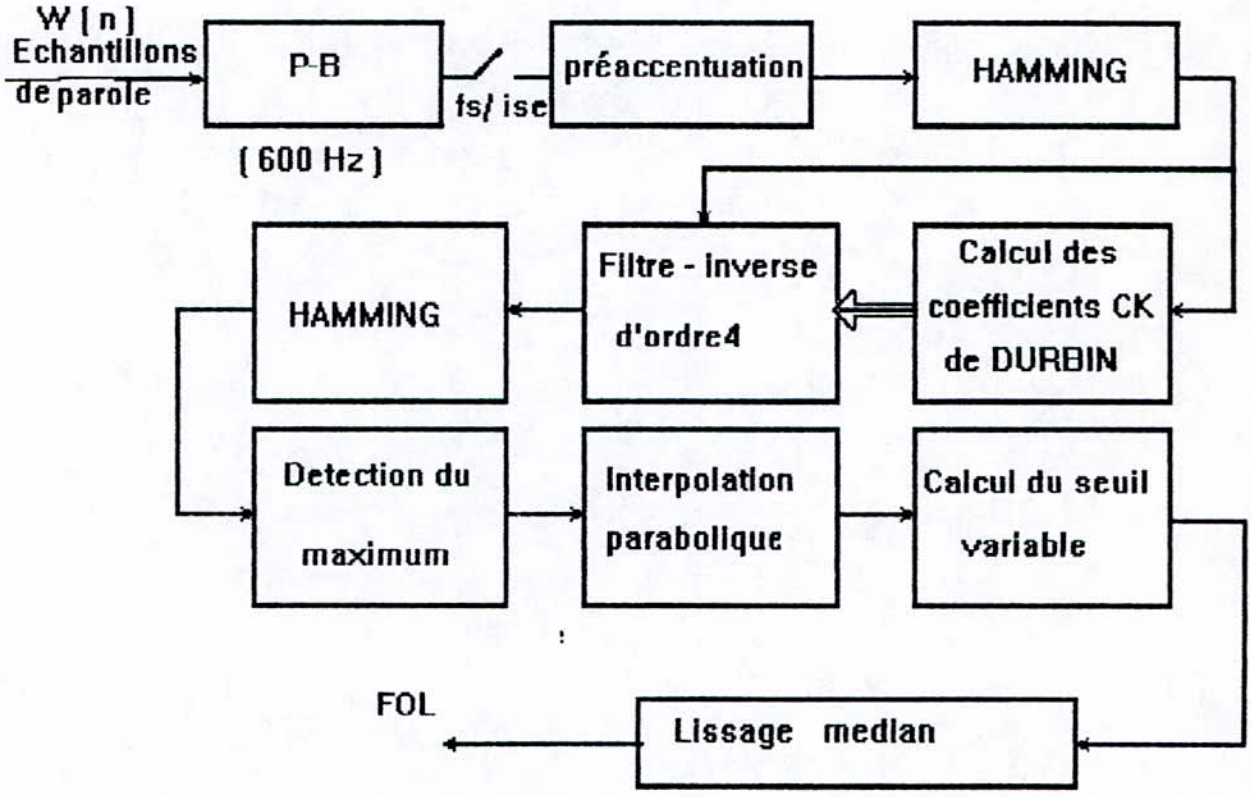


Figure (4-1) : Les différents blocs du programme proposé

Le signal de parole est filtré par un passe-bas (Butterworth d'ordre 2) de fréquence de coupure proche de 600Hz, puis le taux d'échantillonnage est réduit à 2Khz.

Le signal W(n) est alors préaccentué et pondéré par une fenêtre de Hamming.

On applique ensuite le signal préaccentué et pondéré à l'algorithme de Durbin pour le calcul des coefficients CK du filtre inverse qui a la structure en treillis vue au chapitre 2.

La sortie du filtre (signal erreur sous-échantillonné) est alors pondéré par une fenêtre de Hamming, on réalise ensuite l'autocorrélation du signal obtenu et la position maximum de cette fonction donne la valeur du pitch.

Pour une meilleure résolution la fonction d'autocorrélation est interpolée paraboliquement dans les régions de la valeur maximale.
Une comparaison a un seuil variable permet la distinction son voisé/son non-voisé.

On réalise ensuite un lissage de la période calculée sur trois trames de façon à supprimer les erreurs de fréquences de pitch multiple de FO.

Le déroulement de notre programme se fait en trois phases qui sont :

- le pré-traitement
- le traitement
- le post-traitement.

IV-2/ Pré-traitement

La phase de pré-traitement comporte trois étapes qui sont :
- Filtrage Passe-Bas (Butterworth d'ordre 2) à 900Hz
- Décimation
- Préaccentuation.

Voir [6,10,11].

IV-2-1/ Filtrage Passe-Bas

On fait un filtrage passe-bas à 600Hz pour limiter la bande fréquence et de travailler dans la bande Basse (0-600)Hz (plage de l'existence de la fondamentale), le filtre utilisé est celui de Butterworth d'ordre deux (2) de fréquence de coupure 600Hz.
Pour une fréquence d'échantillonnage égale à 8KHz la transmittance du filtre est :

$$T(Z) = \frac{0{,}7157.(1+Z^{-1})^2}{1+1{,}1059.Z^{-1}+0{,}7569.Z^{-2}} \qquad (4\text{-}1)$$

Voir [7,8].

IV-2-2/ Décimation

La décimation ou le sous-échantillonnage permet de changer la vitesse d'échantillonnage du signal de la parole et de réduire le nombre d'échantillons à traiter [7].

IV-2-3/ Préaccentuation

Le bloc de préaccentuation a une transmittance égale à : $1 - bZ^{-1}$ (où b opt = 0,95) Voir chapitre2 [6,10,11,14].

IV-3/ Traitement

La phase de traitement comporte les étapes suivantes :

- Pondération par une fenêtre de Hamming
- Calcul des coefficients ck(i) par DURBIN du filtre en treillis inverse
- Calcul du signal d'erreur par le filtre en treillis
- Calcul de la fonction d'autocorrélation du signal d'erreur
- Détection du Max de la fonction d'autocorrélation du signal d'erreur
- Interpolation parabolique et correction du Max de la fonction d'autocorrélation
- Comparaison du Max corrigé à un seuil variable
- Décision (signal visé ou non visé)

La pondération par une fenêtre de Hamming revient à considérer le signal de la parole nul aux deux extrémités de la trame d'échantillons en cours de traitement.

L'hypothèse de la stationnarité des paramètres du circuit vocal sur une plage d'environ 10 à 35 ms nous conduit à utiliser une fréquence d'échantillonnage d'environ 10 Khz et le nombre d'échantillons de la trame avant la décimation est de 300 échantillons à 320 .

Après la pondération par la fenêtre de Hamming, on calcule les coefficients du filtre inverse par la méthode de Durbin (chapitre 2). Ceci revient à calculer d'abord l'autocorrélation du signal pondéré et préaccentué.

Vu que l'ordre du filtre a été optimisé à 4 à cause de la décimation, nous calculerons que les cinq premières valeurs de la fonction d'autocorrélation. Ces valeurs sont utilisées pour calculer les quatre coefficients du filtre inverse. On applique ensuite le signal préaccentué et pondéré à l'entrée du filtre, on aura à la sortie le signal d'erreur.

L'expérience montre que les quatre premières valeurs de l'erreur de prédiction présentent des valeurs exagérées. Ce qui nous conduit à ne pas les prendre en compte. Donc , on fait un décalage de quatre échantillons. Le même phénomène se présente vers la fin du signal d'erreur.

Ensuite, on multiplie le signal d'erreur par une fenêtre de Hamming, et on calcule l'autocorrélation du signal d'erreur pondéré. Après, on cherche le maximum de cette fonction sur l'intervalle 6-32 (pour une fréquence d'échantillonnage égale à 10 Khz). Le max s'il existe présentera les valeurs du "Pitch".

On fait ensuite une interpolation parabolique du max pour le corriger et corriger le "Pitch".

Pour être accepté, un Maximum (rapporté à la première valeur de la fonction d'autocorrélation) doit dépasser un seuil variable, et ceci pour qu'il n'ait pas un Maximum non lié à la période de Pitch; la loi représentée à la figure (4-2) donne les seuils variables.

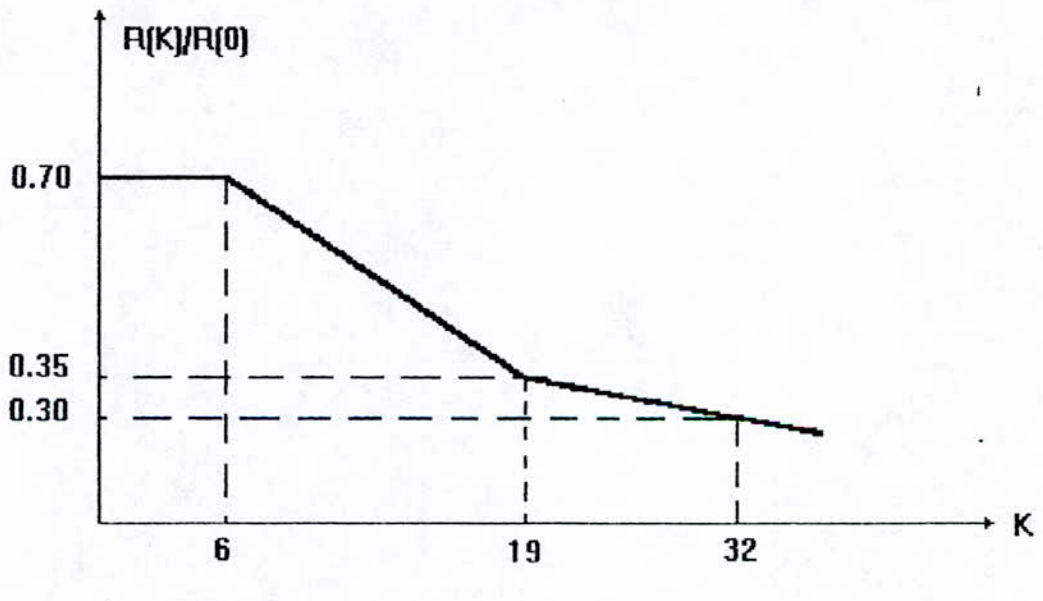


Figure (4-2) : Loi du seuil variable [6]

où R(k) la fonction d'autocorrélation du signal d'erreur.
Si R(0) est inférieur à R(5) et/ou le maximum est égal à zéro, le signal est non-voisé.

IV-4/ Post-traitement

La phase de post-traitement est réalisée par le lissage médian (voir chapitre 2).

IV-5/ Algorihmes des différents blocs

Le déroulement de notre programme se fait selon l'organigramme global de la figure (4-3).

La lecture des trames de la parole se fait sur un fichier parole.
Ce programme est rédigé en TURBO-PASCAL version 3. Nous donnerons tout d'abord les sous-programmes de notre programme principal qui sont :

* sous-programme de préaccentuation (Long1 : entier, X1, Y1) :

```
Debut
  Y1(0) = X1(O)
  Pour I = 1 à Long1-1
  Faire
    Y1(I) = X1(I) - 0,95 x 1.(i-1)
  Fin
Fin
```

ce sous programme fait la préaccentuation

* Sous-programme HAMMING (Long, Window)

```
Debut
      PI = 4 Arctg(1)
      Pour I = 0 à Long -1
     Faire
        Window (I) = 0,54 - 0,46 . cos(2.PI. (I-1)/Long)
      Fin
Fin
```

Ce sous-programme fait le calcul de la fenêtre de Hamming.

* Sous-programme WINMULTR (Long, X2, Window, Y2)

```
Long : Entier
Debut
      Pour I = 0 à Long-1
            Faire
                  Y2(i) = Window(i) . X2(i)
            Fin
```

```
Fin
```
Ce sous-programme fait la pondération d'un signal par une fenêtre de Hamming.

```
*Sous-programme correl( M, long, X3, Y3)
M, Long : Entier
Debut
        Pour I = 0 à M-1
                Faire
                        Y3(i) = 0
                        Pour j = 0 à Long-I
                                Faire
                                        IDEC = i-1
                                        Y3(i) = X3(j).X3(j+IDEC)+Y3(i)
                                Fin
                Fin
Fin
```
Ce sous-programme fait l'autocorrélation d'un signal X3.

```
* Sous-programme Durbin (P, CK, Y)
P : Entier              Y : Signal d'autocorrélation
CK : coéfficient du filtre inverse
Début
        Emin(0) = Y(0)
        CK(1) = Y(1)/Y(0)
        EMIN(1) = (1-CK(1))2
        A(1) = CK(1)
        Pour I = 2  à P
                Faire
                        S = 0
                        Pour J = 1 à I-1
                                Faire
                                        S = A(J).R(I-J)
                                Fin
                        CK(I) = (Y(I)-S)/EMIN(I-1)
                        A(Z) = CK(I)
                        Pour J = 1 à i-1
                                Faire
                                        A(j) = A(j) - CK(I). A(I-J)
                                Fin
                Fin
Fin
```

Ce sous-programme calcule les P paramétres du filtre inverse.

```
* Sous-programme Direct (Long, X, Y)
X : entrée du filtre P.B
Y : sa sortie
Long : Entier
Début
        af = 0,7151
        bf = 1,1059
        cf = 0,7569
        Y(0) = af.x(0)
        Y(1) = af.x(1) + 2.af.x(0)
        Pour I = 2 à Long
                Faire
                        Y(i) = af . (x(i) + 2.x(i-1) + x(i-2)) - bf.y(i-1) - cf.y(i-2)
                Fin
Fin

Sous-programme TREINV (P3, N3, CK, SNO, RES)
P3 : ordre du filtre
N3 : nombre d'échantillons traités
SNO : signal préaccentué et pondéré
RES : signal d'erreur
Dans le programme les BPAR sont les mémoires du filtre
Début
        Pour I = 0 à N3-1
                Faire
                 BPARS(i) = BPAR(i)
                 MEMP = SNO (i)
                 CALY = SNO (i)
                      Pour j = 1 à P3
                              Faire
                                      MEM = BRARS(j) - CK(j) - CALY
                                      CALY = CALY - CK(j).BPARS(j)
                                      BPARS(j) = MEMP
                                      MEMP = MEM
                              Fin
                      RES(i) = CALY
                      Fin
Fin
ce sous-programme calcule le signal d'erreur.
```

*** Algorithme du programme principal du traitement**

Algorithme du post-traitement est donné par l'organigramme du lissage médian dans le Chapitre 2.

Jusqu'ici on a cité les principaux Algorithmes de notre programme.

Pour plus de détail voir le listing dans l'Annexe.

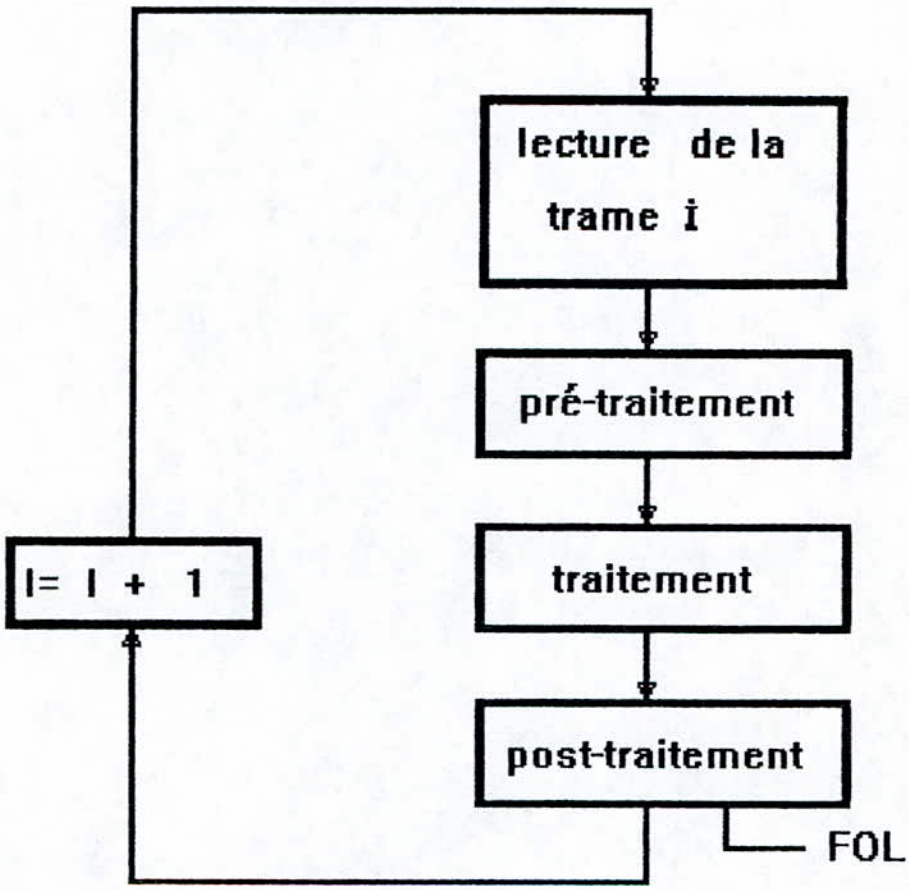


Figure (4-3) organigramme global du programme

IV-6/ Commentaires sur le programme

Vu qu'on a pas d'échantillons de parole le TEST du programme c'est fait sur une TRAME d'échantillons de parole de taille 320 prise du manuel d'utilisation des logiciels d'analyse LPC et extraction des F0 par SIFT [13] et nous avons gardé la même trame à chaque incrémentation des trames comme si le son échantillonné était une reproduction d'une même trame, ces échantillons de parole ont été échantillonnés à une frequence égale à 8 KHZ.

Les résultats obtenus de la fréquence fondamentale lissée est :
FOL = 113,485 HZ et ceci pour toutes les trames .
Pour la même personne dont on a prélevé la trame de test, le manuel d'utilisation des logiciels d'analyse LPC et d'extraction F0 par SIFT [13] donne une fréquence lissée égale à 126,98 HZ en moyenne.
D'où le bon fonctionnement de notre programme est garanti.

IV-7/ Conclusion :

La différence des résultats trouvés par notre programme comparés à ceux du manuel d'utilisation des logiciels d'analyse LPC et l'extraction par SIFT [13] se traduit par l'utilisation du filtre de Butterworth dans notre programme au lieu du filtre elliptique ou de Tchébycheff qui présentent une nette supériorité par rapport à notre filtre grâce à leur bonne sélectivité en fréquence et leur bande de transition qui est petite et leur atténuation en bande atténuée.

Conclusions Générales

Notre travail s'est porté sur la simulation d'un détecteur de fréquence fondamentale. Pour cela on a adopté la méthode dite de filtrage inverse, plus connue sous l'appellation SIFT.

Il est certain que pour mener à bien notre mission, on a dû nous familiariser avec d'autres méthodes (AMDF, EGG) et cela pour une meilleure compréhension de notre propos.

On a réussi à mettre au point un programme de simulation qui nous a donné entière satisfaction, dans la mesure où les résultats obtenus sont très proches de ceux obtenus par le manuel d'utilisation des logiciels d'analyse LPC et d'extraction F0 par SIFT [13].
Cette différence de résultats s'explique par l'utilisation du filtre de Butterworth d'ordre 2, l'approche sera meilleure si l'on remplace par un filtre elliptique ou celui de Tchebycheff d'ordre trois.

On espère voir suite à notre travail de recherche, et ceci par la combine de la méthode de filtrage inverse avec celle de l'ambiguïté modifiée, et en prenant en considération si le locuteur est de sexe masculin ou féminin, ainsi que son âge et ceci a pour effet de localiser la plage de détection en fréquence.

ANNEXE

CODE SOURCE DU PROGRAMME REALISE :

```
{$N+;$E+}
PROGRAM SIMULATION(INPUT,OUTPUT);
USES CRT,DOS;
CONST
      IDECAL=320;                    /*NOMBRE D'ECHANTILLONS PAR TRAME*/
      NF=10;                         /*NOMBRE DE TRAMES */

TYPE
      TAB1=ARRAY[1..10] OF REAL;
      TAB2=ARRAY[1..400] OF REAL;
      TAB3=ARRAY[1..3714] OF REAL;
      TAB4=ARRAY[1..400] OF DOUBLE;
      TAB5=ARRAY[1..10] OF DOUBLE;
VAR
      I,IFEN :INTEGER;
      P0:DOUBLE;
      W:TAB2;
      F0,FOL:TAB4;

PROCEDURE SIFT(LONG,FS:INTEGER;VAR P0:DOUBLE;X:TAB2);
/*CE SOUS-PROGRAMME CALCULE LE PITCH PAR DETECTION DU MAXIMUM DE LA
FONCTION D'AUTOCORRELATION */
VAR
      ISE,L1,L1M4,I8,J8,L,LONG0,K:INTEGER;
      AMAX,AA,BB,AP,AL:DOUBLE;
      V,D,CC,VV:REAL;
      XR,SOUT,BUF,RES,UU,WINDOW,WINDOW0,U,FOUT,Y:TAB2;
      CK1,BPAR,BPARS:TAB1;
      ABUF:TAB4;

PROCEDURE DIRECT (LONG1:INTEGER;VAR X1,Y1:TAB2);
/*CETTE PROCEDURE REALISE LE FILTRAGE PASSE-BAS DE BUTTERWORTH D'ORDRE
 DEUX DE FREQUENCE DE COUPURE 600HZ*/

CONST
      AF=0.7157;
      BF=1.1059;
      CF=0.7569;
```

```
VAR
        I1:INTEGER;

BEGIN
Y1[1]:=AF*X1[1];
Y1[2]:=AF*X1[2];
Y1[2]:=Y1[2]+2*AF*X1[1];
Y1[2]:=Y1[2]-BF*Y1[1];
FOR I1:=3 TO LONG1 DO

BEGIN
  Y1[I1]:=AF*X1[I1];
  Y1[I1]:=AF*2*X1[I1-1]+Y1[I1];
  Y1[I1]:=AF*X1[I1-2]+Y1[I1];
  Y1[I1]:=Y1[I1]-BF*Y1[I1-1];
  Y1[I1]:=Y1[I1]-CF*Y1[I1-2];
  Y1[I1]:=Y1[I1];
  END;
END;

PROCEDURE PREACS(LONG2:INTEGER;VAR X2,Y2:TAB2);
/*CETTE PROCEDURE REALISE LA PRE-ACCENTUATION DU SIGNAL DE LA PAROLE
FILTRE*/
VAR
        I2:INTEGER;

BEGIN
Y2[1]:=X2[1];
FOR I2:=2 TO LONG2 DO
 BEGIN
 Y2[I2]:=X2[I2]-0.95*X2[I2-1];
END;
END;

PROCEDURE HAMMING(LONG3:INTEGER;VAR WINDOW1:TAB2);
/*CETTE PROCEDURE CALCULE LA FENETRE DE HAMMING*/
VAR
        I3:INTEGER;
        PI:REAL;

BEGIN
PI:=4*ARCTAN(1);
FOR I3:=1 TO LONG3 DO
 BEGIN
 WINDOW1[I3]:=(0.54-0.46*COS(2.*PI*(I3-1)/LONG3));
 END;
```

```
END;



PROCEDURE WINMULTR(LONG4:INTEGER;VAR X3,WINDOW2,Y3:TAB2);
/*CETTE PROCEDURE FAIT LA MULTIPLICATION DU SIGNAL DE LA PAROLE
PRE-ACCENTUE PAR UNE FENETRE */

VAR
      I4:INTEGER;

BEGIN
 FOR I4:=1 TO LONG4 DO
  BEGIN
  Y3[I4]:=X3[I4]*WINDOW2[I4];
  END;
 END;

PROCEDURE CORREL(M,LONG5:INTEGER;VAR X4,Y4:TAB2);
/*CETTE PROCEDURE CALCULE LES CORRELATIONS D'UN SIGNAL X4*/
VAR
      I5,IDEC,J5:INTEGER;
      D1:REAL;

BEGIN

 FOR I5:=1 TO M DO
      BEGIN
      Y4[I5]:=0;
      FOR J5:=1 TO LONG5-I5+1 DO
            BEGIN
            IDEC:=I5-1;
            D1:=X4[J5]*X4[J5+IDEC];
            Y4[I5]:=(Y4[I5]+D1);
            END;
      END;
END;

PROCEDURE DURBIN(P2:INTEGER;VAR CK2:TAB1;Y5:TAB2);
/*CETTE PROCEDURE CALCULE LES COEFIICIENTS DE DURBIN DU FILTRE EN
 TREILLIS INVERSE (CK(I))*/
TYPE
      TAB=ARRAY[1..10,1..10]OF REAL;

VAR
```

```
I6,J6:INTEGER;
A,EMIN:TAB1;
S:REAL;

BEGIN
EMIN[1]:=Y5[1];
CK2[1]:=Y5[2]/Y5[1];
EMIN[2]:=(1-(CK2[1]*CK2[1]))*EMIN[1];
A[1]:=CK2[1];
FOR I6:=2 TO P2 DO
 BEGIN
 S:=0;
 FOR J6:=1 TO I6-1 DO S:=A[I6-1]*Y5[I6-J6+1]+S;
 CK2[I6]:=(Y5[I6+1]-S)/EMIN[I6];
 IF CK2[I6]<-0.000001 THEN CK2[I6]:=0;
 EMIN[I6+1]:=SQR(CK2[I6]);
 EMIN[I6+1]:=(1-EMIN[I6+1]);
 EMIN[I6+1]:=EMIN[I6+1]*EMIN[I6];
 A[I6]:=CK2[I6];
 FOR J6:=1 TO I6-1 DO
   BEGIN
   A[J6]:=A[J6]-CK2[I6]*A[I6-J6];
   END;
 END;
END;

PROCEDURE TREINV(P3,N3:INTEGER;VAR
BPAR2,BPARS2,CK3:TAB1;SN0:TAB2);
/* CETTE  PROCEDURE REALISE LE FILTRE EN TREILLIS INVERSE POUR LE
CALCUL DU SIGNAL D'ERREUR A  PARTIR DU SIGNAL  DE LA PAROLE
FILTRE , PREACCENTUE ET PONDERE */

VAR
I7,J7:INTEGER;
MEM1,MEMP1,CALY1:REAL;

BEGIN
BPARS2:=BPAR2;
FOR I7:=1 TO N3 DO
  BEGIN
  MEMP1:=SN0[I7];
  CALY1:=SN0[I7];
  FOR J7:=1 TO P3 DO
    BEGIN
    MEM1:=BPARS2[J7]-CK3[J7]*CALY1;
    CALY1:=CALY1-CK3[J7]*BPARS2[J7];
```

```
    BPARS2[J7]:=MEMP1;
    MEMP1:=MEM1;
    END;
   RES[I7]:=CALY1;
   END;
  END;

  PROCEDURE CORREL1(M1,LONG9:INTEGER;VAR X9:TAB2);
/*CETTE PROCEDURE REALISE L'AUTO CORRELATION DU SIGNAL
D'ERREUR  ABUF */
  VAR
  I9,IDEC1,J9:INTEGER;
  D2:DOUBLE;
  BEGIN
   FOR I9:=1 TO M1 DO
    BEGIN
    ABUF[I9]:=0;
    D2:=0;
    FOR J9:=1 TO LONG9-I9+1 DO
     BEGIN
      IDEC1:=I9-1;
      D2:=X9[J9]*X9[J9+IDEC1];
      ABUF[I9]:=(ABUF[I9]+D2);
     END;
    END;
  END;

  BEGIN
  LONG0:=LONG;
  ISE:=Fs DIV 2000; /*LE TAUX DE SOUS-ECHANTILLONNAGE */
  L1:=LONG0 DIV ISE ; /* LONGUEUR DU SIGNAL SOUS-ECH */
  L1M4:=L1-4;
  XR:=X;
  DIRECT(LONG0,XR,SOUT);
  FOR I8:=1 TO LONG0 DO
      BEGIN
     IF I8 MOD ISE=0 THEN
        BEGIN
        J8:=I8 DIV ISE;
        BUF[J8]:=SOUT[J8]; /* BUF SIGNAL  SOUS-ECH */
        END;
      END;
   PREACS(L1,BUF,UU);
   HAMMING(L1,WINDOW);
   WINMULTR(L1,UU,WINDOW,U);
   CORREL(5,L1,U,Y);
```

```
   DURBIN(4,CK1,Y);
   FOR I8:=1 TO 4 DO BPAR[I8]:=0;/* INITIALISATION DESMEMOIRES DU
FILTRE EN TREILLIS INVERSE  */
   TREINV(4,L1,BPAR,BPARS,CK1,U);
   FOR I8:=5 TO L1 DO
     BEGIN
     RES[I8-4]:=RES[I8];/* LES QUATRE PREMIERES VALEURS NE SONT PAS
CONSERVEES */
     END;
   HAMMING(L1M4,WINDOW0);
   WINMULTR(L1M4,RES,WINDOW0,FOUT);
   CORREL1(32,L1M4,FOUT);
    /*  RECHERCHE DU MAXIMUM DE L'AUTO CORRELATION DU SIGNAL
D'ERREUR  SUR L'INTERVALLE [6,32] */
   AMAX:=ABUF[6];
   L:=0;
   K:=6;
   FOR K:=6 TO 32 DO
   BEGIN
     IF ABUF[K] >AMAX THEN
      BEGIN
       AMAX:=ABUF[K];
       L:=K;
      END;
   END;
    /* ..................................................................... */
    /* TEST DE LA PRESENCE D'UN SON VOISE */
     IF (AMAX<>0) AND (ABUF[L]>=ABUF[L-1]) THEN
      BEGIN
    /*  IMISE EN OEUVRE DE L'INTERPOLATION PARABOLIQUE  */
      AA:=ABUF[L-1]-ABUF[L];
      AA:=(AA+ABUF[L+1]-ABUF[L])/2;
      BB:=(ABUF[L+1]-ABUF[L-1])/4;
      AP:=ABUF[L]-BB*(BB/AA);
      AL:=L-(BB/AA);
      V:=AP/ABUF[1];
    /*  ..................................................................... */
    /* CALCUL DU SEUIL VARIABLE */
      IF (L>19) OR (L=0) THEN
       BEGIN
       D:=0.05/13;
       CC:=0.3+32*D;
       VV:=CC-D*L;
       END;
      IF L<19 THEN
       BEGIN
```

```
D:=0.35/13;
CC:=0.35+19*D;
VV:=CC-D*L;
END;
/* ..... */
/* DECISION SUR L'ACCEPTATION DE LA PRESENCE DE LA
   FONDAMENTALE */
IF V>=VV THEN P0:=(AL-1)*ISE ELSE P0:=0;
END;
/* ..... */
/* DECISION SUR LE NON VOISEMENT DE LA TRAME */
IF (AMAX=0) OR (ABUF[L]<ABUF[L-1]) THEN
P0:=0;
END;
/* ..... */

        /*  PROGRAMME PRINCIPAL.  */

BEGIN
 FOR IFEN:=1 TO NF DO  /* BOUCLE SUR LES NF TRAME DU SIGNAL */
  BEGIN
   /* LES W[I]  REPRESENTENT LES 320 ECHANTILLONS DE NOTRE TRAME
      DE TEST DU DETECTEUR */
      W[1]:=195;
      W[2]:=182;
      W[3]:=182;
      W[4]:=182;
      W[5]:=164;
      W[6]:=150;
      W[7]:=127;
      W[8]:=96;
      W[9]:=42;
      W[10]:=2;
      W[11]:=-35;
      W[12]:=-77;
      W[13]:=-98;
      W[14]:=-105;
      W[15]:=-111;
      W[16]:=-115;
      W[17]:=-118;
      W[18]:=-125;
      W[19]:=-118;
      W[20]:=104;
      W[21]:=-87;
      W[22]:=-62;
      W[23]:=-32;
```

```
W[24]:=-5;
W[25]:=20;
W[26]:=43;
W[27]:=57;
W[28]:=64;
W[29]:=73;
W[30]:=84;
W[31]:=87;
W[32]:=83;
W[33]:=75;
W[34]:=56;
W[35]:=30;
W[36]:=3;
W[37]:=-21;
W[38]:=-43;
W[39]:=-62;
W[40]:=-73;
W[41]:=-82;
W[42]:=-81;
W[43]:=-79;
W[44]:=-76;
W[45]:=-62;
W[46]:=-43;
W[47]:=-21;
W[48]:=-2;
W[49]:=14;
W[50]:=23;
W[51]:=22;
W[52]:=14;
W[53]:=3;
W[54]:=-12;
W[55]:=-24;
W[56]:=-35;
W[57]:=-44;
W[58]:=-55;
W[59]:=-57;
W[60]:=-16;
W[61]:=23;
W[62]:=55;
W[63]:=123;
W[64]:=151;
W[65]:=170;
W[66]:=199;
W[67]:=195;
W[68]:=205;
W[69]:=210;
```

```
W[70]:=203;
W[71]:=190;
W[72]:=165;
W[73]:=125;
W[74]:=69;
W[75]:=19;
W[76]:=-21;
W[77]:=-62;
W[78]:=-86;
W[79]:=-102;
W[80]:=-119;
W[81]:=-137;
W[82]:=-143;
W[83]:=-138;
W[84]:=-133;
W[85]:=-109;
W[86]:=-80;
W[87]:=-53;
W[88]:=-23;
W[89]:=3;
W[90]:=16;
W[91]:=34;
W[92]:=52;
W[93]:=65;
W[94]:=76;
W[95]:=79;
W[96]:=82;
W[97]:=69;
W[98]:=49;
W[99]:=29;
W[100]:=7;
W[101]:=-12;
W[102]:=-27;
W[103]:=-48;
W[104]:=-60;
W[105]:=-77;
W[106]:=-86;
W[107]:=-87;
W[108]:=-80;
W[109]:=-67;
W[110]:=-47;
W[111]:=-29;
W[112]:=-8;
W[113]:=3;
W[114]:=10;
W[115]:=9;
```

```
W[116]:=5;
W[117]:=2;
W[118]:=-1;
W[119]:=-6;
W[120]:=-16;
W[121]:=-26;
W[122]:=-39;
W[123]:=-48;
W[124]:=-25;
W[125]:=21;
W[126]:=45;
W[127]:=112;
W[128]:=143;
W[129]:=142;
W[130]:=167;
W[131]:=160;
W[132]:=179;
W[133]:=195;
W[134]:=199;
W[135]:=192;
W[136]:=165;
W[137]:=131;
W[138]:=76;
W[139]:=27;
W[140]:=-3;
W[141]:=-35;
W[142]:=-64;
W[143]:=-86;
W[144]:=-111;
W[145]:=-131;
W[146]:=-147;
W[147]:=-146;
W[148]:=-132;
W[149]:=-112;
W[150]:=-83;
W[151]:=-61;
W[152]:=-36;
W[153]:=-16;
W[154]:=-6;
W[155]:=14;
W[156]:=38;
W[157]:=59;
W[158]:=74;
W[159]:=82;
W[160]:=77;
W[161]:=63;
```

```
W[162]:=49;
W[163]:=34;
W[164]:=18;
W[165]:=4;
W[166]:=-11;
W[167]:=-33;
W[168]:=-56;
W[169]:=-75;
W[170]:=-89;
W[171]:=92;
W[172]:=-87;
W[173]:=-72;
W[174]:=-54;
W[175]:=-37;
W[176]:=-25;
W[177]:=-11;
W[178]:=-2;
W[179]:=5;
W[180]:=7;
W[181]:=11;
W[182]:=7;
W[183]:=4;
W[184]:=-6;
W[185]:=-14;
W[186]:=-28;
W[187]:=-29;
W[188]:=25;
W[189]:=57;
W[190]:=82;
W[191]:=129;
W[192]:=134;
W[193]:=140;
W[194]:=158;
W[195]:=148;
W[196]:=170;
W[197]:=183;
W[198]:=175;
W[199]:=165;
W[200]:=138;
W[201]:=96;
W[202]:=39;
W[203]:=13;
W[204]:=-14;
W[205]:=-45;
W[206]:=-63;
W[207]:=-92;
```

```
W[208]:=-118;
W[209]:=-143;
W[210]:=-146;
W[211]:=-136;
W[212]:=-117;
W[213]:=-87;
W[214]:=-57;
W[215]:=-40;
W[216]:=-22;
W[217]:=-6;
W[218]:=10;
W[219]:=35;
W[220]:=53;
W[221]:=69;
W[222]:=81;
W[223]:=77;
W[224]:=65;
W[225]:=50;
W[226]:=39;
W[227]:=26;
W[228]:=13;
W[229]:=-1;
W[230]:=-20;
W[231]:=-41;
W[232]:=-63;
W[233]:=-80;
W[234]:=-91;
W[235]:=-88;
W[236]:=-84;
W[237]:=-73;
W[238]:=-59;
W[239]:=-48;
W[240]:=-36;
W[241]:=-24;
W[242]:=-16;
W[243]:=-4;
W[244]:=1;
W[245]:=9;
W[246]:=9;
W[247]:=6;
W[248]:=-1;
W[249]:=-11;
W[250]:=-6;
W[251]:=47;
W[252]:=80;
W[253]:=106;
```

```
W[254]:=142;
W[255]:=148;
W[256]:=145;
W[257]:=154;
W[258]:=144;
W[259]:=156;
W[260]:=160;
W[261]:=153;
W[262]:=142;
W[263]:=114;
W[264]:=73;
W[265]:=29;
W[266]:=1;
W[267]:=-21;
W[268]:=-40;
W[269]:=-54;
W[270]:=-74;
W[271]:=-100;
W[272]:=-123;
W[273]:=-129;
W[174]:=-129;
W[275]:=-116;
W[276]:=-92;
W[277]:=-68;
W[278]:=-46;
W[279]:=-22;
W[280]:=-8;
W[281]:=7;
W[282]:=31;
W[283]:=43;
W[284]:=62;
W[285]:=74;
W[286]:=71;
W[287]:=60;
W[288]:=47;
W[289]:=31;
W[290]:=18;
W[291]:=6;
W[292]:=16;
W[293]:=-10;
W[294]:=-38;
W[295]:=-48;
W[296]:=-68;
W[297]:=-69;
W[298]:=-80;
W[299]:=-77;
```

```
    W[300]:=-64;
    W[301]:=-51;
    W[302]:=-48;
    W[303]:=-41;
    W[304]:=-32;
    W[305]:=-33;
    W[306]:=-28;
    W[307]:=-16;
    W[308]:=-17;
    W[309]:=-9;
    W[310]:=-10;
    W[311]:=-7;
    W[312]:=-3;
    W[313]:=16;
    W[314]:=60;
    W[315]:=112;
    W[316]:=136;
    W[317]:=166;
    W[318]:=180;
    W[319]:=175;
    W[320]:=174;
    SIFT(IDECAL,8000,P0,W);
     /* CALCUL DE LA FREQUENCE FONDAMENTALE */
    IF (P0<>0) THEN F0[IFEN]:=8000/P0 ELSE F0[IFEN]:=0;
  END;
    /* ..... */
 /* LISSAGE DE LA FREQUENCE FONDAMENTALE SUR TROIS TRAMES */
 FOR I:=2 TO (NF-1) DO
    BEGIN
     IF((F0[I]<>0)AND(F0[I-1]<>0)) AND (F0[I+1]<>0)
     THEN BEGIN
     FOL[I]:=0.25*(F0[I+1]+F0[I-1])+0.50*F0[I];
     WRITELN('FOL=',FOL[I],' ','HZ');
     END;
     IF(F0[I]<>0) AND( F0[I+1]<>0) AND (F0[I-1]=0) THEN
     BEGIN
     FOL[I]:=0.6*F0[I]+0.4*F0[I+1];
     WRITELN('FOL=',FOL[I],' ','HZ');
     END;
     IF (F0[I]<>0)AND(F0[I-1]<>0) AND (F0[I+1]=0)  THEN
     BEGIN
     FOL[I]:=0.6*F0[I]+0.4*F0[I-1];
     WRITELN('FOL=',FOL[I],' ','HZ');
     END;
    END;
END.
```

REFERENCES BIBLIOGRAPHIQUES

[1] *Le petit Robert* Edition 1972.

[2] *Nouveau Larousse Médical* Edition 1990.

[3] *Turbo-Pascal : Manuel de référence Version 5.0* Borland 1988.

[4] *Turbo-Pascal : Guide d'utilisation* . Borland 1990.

[5] M. Kunt, *Traitement numérique des signaux,* Presses polytechniques romandes, Lausanne 1984.

[6] R. Boite et M. Kunt, *Traitement de la parole,* Presses polytechniques romandes, Lausanne 1987.

[7] J. Max, *Méthodes et techniques de traitement du signal* Tome 1 et 2 , Masson, Paris 1987.

[8] M. Bellanger, *Traitement numérique du signal*, Masson, Paris 1987.

[9] J. Guibert, *La parole : compréhension et synthèse par les ordinateurs* Presses Universitaires de France, Paris 1979.

[10] S.A Selouani, *Contribution à l'extraction des paramètres prosodiques du signal de parole : cas de la fréquence fondamentale,* Thèse de magister, Université des Sciences et de la Technologie Houari Boumédiene, Alger 1991.

[11] H. Sayoud, *Robustesse des systèmes de détection de périodicité de signaux complexes en milieu bruité,* Thése de magister, Université des Sciences et de la Technologie Houari Boumédiene, Alger 1994.

[12] M. Guerti, *Contribution à la synthèse de la parole en Arabe standard. Synthèse par diphones et technique de prédiction linéaire,* ILP, Alger 1984.

[13] *Logiciels d'analyse LPC et extraction F0 par SIFT : Manuel d'utilisation,* Française d'Electronique Recherches et Mathématiques, Paris 1985.

[14] *Codage et analyse-synthèse de la parole,* CNET-ENST, Brest 1983.