République Algérienne Démocratique et Populaire
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

Université ABOU BEKR BELKAÏD - TLEMCEN
FACULTE DES SCIENCES DE L'INGENIEUR
DEPARTEMENT D'ELECTRONIQUE

Mémoire de Magister en Electronique

Spécialité : Signaux et Systèmes

Thème

ETUDE DE LA METHODE PLP (PERCEPTUAL LINEAR PREDICTION) EN RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Présenté par :

Lamia BENLALDJ

Soutenu en Juillet 2000 devant le jury :

Président:	*M^r T. BENDIMERAD*	*M.C. à l'université de Tlemcen.*
Examinateurs:	*M^r B. CHERKI*	*M.C. à l'université de Tlemcen.*
	M^r M. BENABDELLAH	*M.C. à l'université de Tlemcen.*
Encadreur:	*M^r F. BEREKSI REGUIG*	*Prof. à l'université de Tlemcen.*
Co-Encadreur:	*M^elle O. GAOUAR*	*C.C. à l'université de Tlemcen.*
Invité d'honneur:	*M^r S. M. GHITRI*	*M.C. au Département des Langues Arabes*

Année universitaire: 1999 / 2000

REMERCIEMENTS

Mes sincères remerciements sont adressés à mon encadreur M. BEREKSI REGUIG et mon co-encadreur Mlle GAOUAR pour leur aide, conseils, critiques et disponibilité durant l'élaboration de ce mémoire.

Je remercie M. T. BENDIMERAD de m'avoir fait l'honneur d'accepter de présider le jury. Mes remerciements vont aussi à M. CHERKI et M. BENABDELLAH de bien avoir voulu juger le travail effectué dans le cadre de ce magister.

Je remercie aussi M. GHITRI mon invité d'honneur, ainsi que tous les enseignants ayant participé à ma formation.

DEDICACES

Je dédie ce mémoire à mes très chères parents.

A ma mère que j'aime énormément, je te remercie pour ta gentillesse, ton soutien et ta patience. A mon père aussi que j'aime beaucoup, qu'il trouve ici ma sincère gratitude. Sans eux jamais je n'aurai atteint ce but.

Je le dédie ensuite à mes deux sœurs : Hayet et Souad en espérant qu'elles réussissent dans leurs vies.

Je dédie aussi ce mémoire à toute la famille Benlaldj et la famille Cherif grands et petits.

A mon amie Nawel, son mari Azzeddine et leurs familles respectives.

A toute ma promotion d'ingénieurs 1992.

A tous mes amis spécialement : Saïd, Mourad, Naïma, Abdelghani, Amaria, Hadj, Fouad et Nawel sans oublier Ihssane et Mohammed.

Lamia

ABSTRACT

Our work consists in the study, the processing and the analysis of the speech signal. This signal is very variable. This variability is due to the speaker or other factors such as the environment or the recording means. Therefore, it is not a stationary signal.

The analysis of this signal is carried out by the PLP technique (Perceptual Linear Prediction). This technique uses human hearing properties for the speech perception and the linear prediction coding (LPC) based on autoregressive all-pole modeling (AR) of the vocal tract. Such properties are the nonequal sensitivity of the human hearing at different frequencies and the non linear relation ship between the intensity of the sound and its perceived loudness. On the other hand, the application of the AR model allows us to extract some parameters (Prediction coefficients) characterising the speech signal.

A set of tests have been made to study and to analyse the PLP technique. First, we used this method in the identification of oral vowels of the French and the Arabic language. In this case, the location of the first two maxima of the PLP spectrum allows us to make the distinction between such vowels.

Secondly, and in the isolated-word identification of one speaker, the PLP technique applied on a speech signal, provide a sequence of parameters known as the template (acoustical image). This template constitute the dictionary of references.

The recognition has been done by a conventional fixed-end point Dynamic Time-Warping algorithm (DTW). This algorithm is based on a temporal alignment of the template by dynamic programming. The application of DTW algorithm on the template leads to satisfactory recognition scores.

Key words: speech signal, speech perception, linear prediction, perceptual linear prediction (PLP), DTW algorithm, automatic recognition of isolated-words.

ملخّص

إنّ عملنا يهتمّ بدراسة و تحليل الإشارة الصوتية (الصوت بصفة عامة). إنّ الصّوت في حدّ ذاته متغيّر جدّا. هذه التغيرات التي تحدث ناتجة عن المتكلم، أو خارجة عن نطاقه كالمحيط و الأدوات المستعملة لتسجيل الصوت.

قمنا بتحليل الصوت بطريقة ال PLP: (Perceptual Linear Prediction) التي تستعمل خاصيات سمع الأصوات عند الإنسان و طريقة التوقع الخطي العادي (Prédiction Linéaire Conventionnelle PLC) التي تعتمد في حدّ ذاتها على نموذج يدعى النموذج AR (modèle autoregressif) للجهاز الصوتي.
من بين الخصائص المستعملة للسمع عند البشر : إختلاف سمع الأصوات في مختلف الذّبذبات (التواترات) و العلاقة اللا خطية بين شدّة الصوت و الشدة التي نسمع بها. زيادة على هذا، إستعمال النموذج يسمح لنا بإستخلاص قيم (معاملات التوقع) لتعريف الإشارة الصوتية.

مجموعة تجارب قمنا بها لدراسة و تحليل تقنية ال PLP .
أوّلا، للتعرف على حركات اللغة الفرنسية (les voyelles) و حركات اللغة العربية. إستعملنا هذه الوسيلة للتحصّل على أول أقصى ذروتين نلاحظها على التمثيل التواتري لل : PLP و اللتي سمحتا لنا بالتفريق والتعرف على الحركات.
ثانيا، في التعرف على كلمات معزولة لناطق واحد، إستعمال تقنية ال PLP على إشارة صوتية،أعطى مجموعة من الأشعة للمعاملات مسماة صورة أكوستكية،مجموعها يكوّن معجم.
التعرف كان بإستعمال طريقة تدعى DTW (Dynamic Time Warping) مبدؤها مبني على الإستقامة الزمنية بالبرمجة الديناميكية للصور الأكوستيكية. إستعمال هذه الصور مكّننا من التحصل على نتائج جدّ مرضية في التعرف الصوتي .

أهم الكلمات : الإشارة الصوتية، سمع الصوت، التوقع الخطي، التوقع الخطي السمعي (PLP)، DTW ، التعرف الآلي على كلمات منعزلة.

SOMMAIRE

INTRODUCTION GENERALE

Dès les années 60, des systèmes de reconnaissance de la parole ont été développés suivant les demandes exigées pour différentes applications.

De nos jours avec l'introduction des méthodes numériques et l'extension des domaines : informatique et électronique, ces systèmes sont de plus en plus évolués et leur utilisation de plus en plus répandue.

Les applications de la reconnaissance automatique de la parole sont entre autres le dialogue naturel avec une machine, la commande vocale d'une machine à dicter, ou même d'un robot. Elle a aussi d'autres applications dans le domaine biomédical (rééducation des mal-entendants) et dans l'étude des langues.

Par conséquent, une étude du signal vocal ainsi que le développement d'un système de Reconnaissance Automatique de la Parole « RAP » s'avèrent importants.

Le signal vocal véhicule différents types d'informations : les sons, la syntaxe et la sémantique de la phrase, l'identité du locuteur et son état émotionnel. Toutes ces informations compliquent l'interprétation du signal et augmentent le nombre de données à traiter. De plus, le signal de la parole étant évolutif, il est nécessaire de le traiter et de l'étudier sur une période bien déterminée.

Le but de notre présente recherche est la reconnaissance automatique de mots isolés en mode monolocuteur. Pour cela, une analyse du signal vocal est faite.

Différentes méthodes d'analyse ont été largement étudiées et appliquées. Celles qui exploitent :

Les transformées à court terme temporelles, spectrales ou encore spectro-temporelle [1], [6], [7]. Des paramètres caractéristiques du signal de la parole tels les formants et le pitch peuvent être déterminés avec une résolution fortement corréler à la fenêtre de traitement.

Les méthodes homomorphiques qui tentent de séparer les contributions respectives de la source et du conduit vocal par l'application de fonctions logarithmiques au spectre du signal générant ainsi le cepstre, à partir du quel la période du fondamental et l'enveloppe du spectre sont déterminés. Cette méthode devient relativement lourde dans le cas des voix de femme où le conduit vocal est long et la fréquence élevée [1], [2].

Les méthodes qui se basent sur la modélisation du conduit vocal : la prédiction linéaire. Les coefficients de prédiction sont alors exploités pour générer une enveloppe spectrale relative au spectre de puissance [1], [2]. Cependant ce modèle approxime de manière uniforme le spectre de puissance sur toutes les fréquences de la bande d'analyse. Cette propriété est contradictoire avec l'audition humaine. En fait les détails spectraux ne sont pas toujours préservés d'après leur proéminence auditive [10], [14]. Cette limitation devient

accrue dans un système de reconnaissance en présence de diverses conditions de variabilité du signal acoustique.

Ainsi une autre technique combinant la prédiction linéaire LPC et exploitant les propriétés d'audition de l'ouïe humaine est développée. C'est la technique PLP (*Perceptual Linear Prediction*) testée et analysée dans notre travail [1], [10], [14]. En fait, cette technique en plus de la modélisation du conduit vocal, modélise les propriétés de l'ouïe humaine et plus particulièrement la propriété de la sensibilité auditive différente en différentes fréquences, la relation non linéaire entre l'intensité du son et l'intensité perçue et la manière dont elles sont utilisées pour la distinction des sons [28].

Les paramètres issus de l'étape d'analyse sont exploités par la suite dans un algorithme de reconnaissance de la parole.

Plusieurs approches peuvent être utiliser pour effectuer une reconnaissance automatique de la parole : modèles de Markov, comparaison dynamique, réseaux de neurones, ...

La comparaison dynamique effectuée à l'aide de l'algorithme DTW (Dynamic Time Warping) procède à un ajustement des échelles temporelles du mot à tester et du mot de référence. A chaque intervalle du mot test prononcé, elle tente d'associer un intervalle de l'image acoustique du mot de référence de manière à minimiser, au sens d'une certaine distance, l'écart entre le son prononcé et l'empreinte en mémoire [18], [20], [22]. C'est l'approche adoptée dans ce travail.

Les modèles de Markov cachés (HMM : Hidden Markov Model) sont des approches stochastiques où la distance est remplacée par des probabilités. Chaque référence acoustique est représentée sous la forme d'un graphe d'état (modèle de Markov). Le problème de l'alignement temporel entre un modèle de Markov et une image acoustique du mot à tester consiste à rechercher un chemin optimal (le plus probable) dans le graphe d'état [18].

Les approches connexionnistes dans lesquelles on retrouve 'les réseaux de neurones', effectuent une discrimination linéaire complexe, en considérant qu'un neurone effectue une discrimination linéaire simple. Ces approches sont basées sur une architecture classique de Perceptron multicouches (MLP : Multi Layers Perceptron) [18]. Ces derniers sont intéressants pour leur pouvoir de reconnaissance mais demandent une adaptation pour intégrer la notion de temps, d'où les réseaux de neurones à décalage temporel (TDNN : Time Delay Neural Networks) [21].

L'étape de reconnaissance est précédée d'une étape d'apprentissage dans laquelle il s'agit de constituer le dictionnaire contenant les références.

Notre travail a été réparti en quatre parties.

En chapitre I, nous décrivons le signal vocal, le système de production de la parole et le système auditif humain. Nous présentons également les propriétés de l'ouïe humaine.

Le second chapitre comporte les différentes méthodes d'analyse spectrales, temporelles et spectro-temporelles ainsi que l'analyse paramétrique (analyse par PL et analyse cepstrale). Un modèle de production de la parole est présenté. C'est le modèle tout-pôle. Les étapes de l'implémentation de ce modèle sont présentées en détails.

Nous présentons en chapitre III la méthode d'analyse PLP, qui exploite les propriétés de l'oreille humaine. Les tests que nous avons fait à ce niveau montrent l'apport de cette technique par rapport à la PLC et sa contribution dans la reconnaissance automatique de la parole.

Dans le dernier chapitre (chapitre IV), sont présentés en premier lieu les étapes d'implémentation de l'algorithme DTW basé sur le principe de l'alignement temporel par programmation dynamique. Des images acoustiques du vocabulaire à reconnaître sont obtenues par la suite en faisant une analyse par PLP de chaque mot constituant le dictionnaire de références. Ces images acoustiques sont en fait des vecteurs de paramètres caractérisant les tranches d'un signal donné.
Un ensemble de tests ont été effectués dans plusieurs langues (arabe, français et anglais) pour étudier les performances du système dans le domaine de la reconnaissance automatique de mots isolés.

CHAPITRE I

GENERALITES SUR LE SIGNAL VOCAL

Les sons de la parole sont donc le produit d'un ensemble de conditions anatomiques, physiologiques et articulatoires. Ils sont produits soit par les vibrations des cordes vocales (source de voisement), soit par l'écoulement turbulent de l'air dans le conduit vocal, soit lors du relâchement d'une occlusion de ce conduit (source de bruit).

Les sons sont décrits et classés selon l'état du larynx, la zone articulatoire, la position du voile du palais et de l'aperture articulatoire qui signifie le degré d'ouverture du conduit vocal.

L'aperture est à la base de la distinction entre voyelle et consonne. Pour les voyelles, le passage de l'air est libre et l'aperture est grande, alors que pour les consonnes, le passage de l'air est plus au moins entravé et l'aperture plus au moins grande [1].

I - 3 - 2 Mécanisme de la phonation

Les *sons voisés* résultent d'une vibration périodique des cordes vocales ; des impulsions périodiques de pression (la source) sont ainsi appliquées au conduit vocal. Un son voisé est représenté par un signal quasi-périodique (fig. I - 2). On remarque sur son spectre (fig. I - 2) des raies qui correspondent aux harmoniques du fondamental F_0 (structure du pitch); l'enveloppe de ces raies présente des maximums appelés *formants* [2] (voir § I - 4 - 4). La structure harmonique d'un son voisé peut être mise en évidence par des filtres à faible bande passante (≈ 50 Hz).

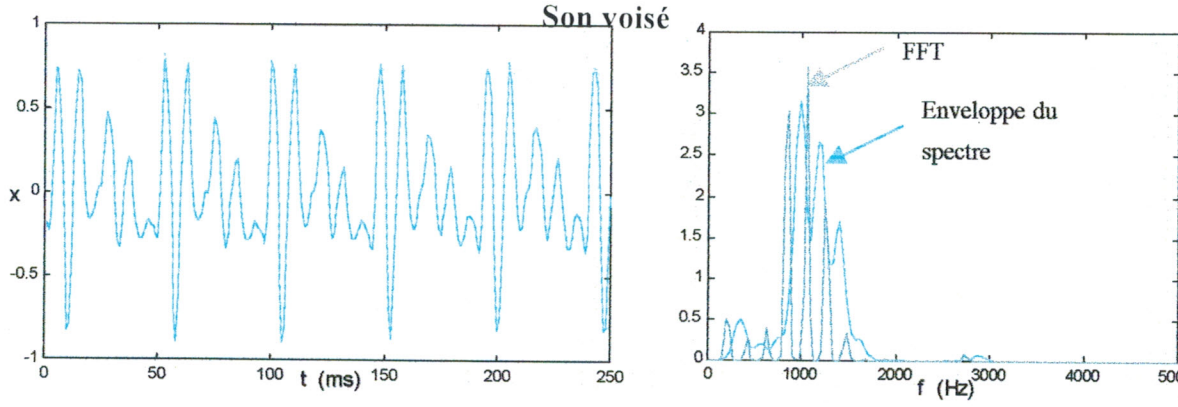


Fig. I - 2 Représentation temporelle de la voyelle /a/ et son spectre

Un *son non voisé* ne présente pas de structure périodique (fig. I - 3), il peut être considéré comme un bruit blanc [2].

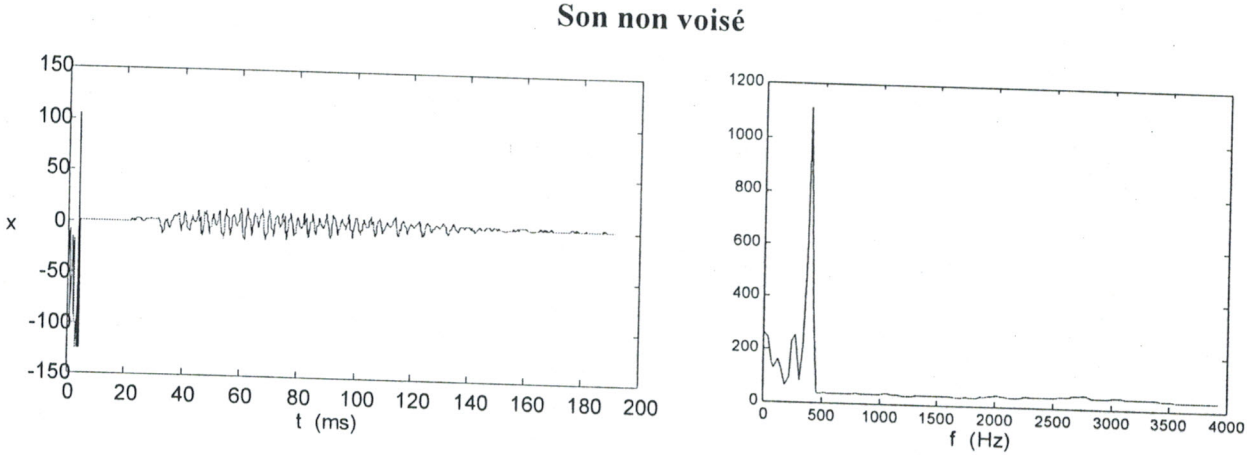


Fig. I - 3 Représentation temporelle de la consonne /t/ et son spectre

Les *sons fricatifs* résultent de l'écoulement de l'air au niveau des lèvres et des dents. Ils peuvent être non voisés (f, s, …) ou voisés (v, z, …).

Un *son occlusif* (polis) est produit par une fermeture momentanée du conduit vocal, suivie par une ouverture brusque ; il peut être voisé (b, d, …) ou non voisé (p, t, …).

I – 3 - 3 Caractéristiques de production de la parole

Le processus de production de la parole présente certaines caractéristiques (continuité, variabilité, …) :

- Continuité; lorsqu'on écoute parler une personne, on perçoit une suite de mots que l'analyse du signal vocal sépare difficilement. Le même problème de segmentation se retrouve à l'intérieur du mot, perçu comme une suite de sons élémentaires, les phonèmes.
- Variabilité; à contenu phonétique égal, le signal vocal est très variable, tant pour différents individus que pour un même locuteur, en raison des différences anatomiques.
- Le conduit vocal est un tuyau tridimensionnel qui est excité par une ou deux sources acoustiques. La source laryngienne peut être considérée comme quasi-périodique, avec une fréquence pouvant évoluer très rapidement. La seconde source génère du bruit de friction ou d'explosion (glotte, lèvres).
- Encodage; depuis l'idée jusqu'au signal sonore, interviennent plusieurs niveaux successifs de traitement : sémantique (concept), syntaxique (structure du langage), lexical (mots), morphologique et phonétique (phonèmes et leurs interactions).

I - 4 Caractéristiques phonétiques

I - 4 - 1 Définition du phonème

Un phonème est la plus petite unité présente dans la parole qui est susceptible par sa présence de changer la signification d'un mot par exemple pari/mari, pas/bas et mie/mes [1]. La notion de phonème ne tient compte que des caractéristiques acoustiques qui permettent une distinction entre des mots. On ne tient pas compte des phénomènes physiques de production du son, tant que la différence d'articulation (fonction du dialecte, de la cadence d'élocution, du contexte) ne permet pas de distinguer des mots différents. Par exemple la langue française comprend 36 phonèmes [2], et la langue arabe en comprend 31 [5].

I - 4 -2 Classification des phonèmes de la langue française

La répartition des 36 phonèmes de la langue française est faite dans la figure (fig. I - 4) [2], où on donne quelques exemples de mots dans lesquels les voyelles sont présentent. Les symboles phonétiques ont été proposés par l'API (Association Phonétique Internationale).

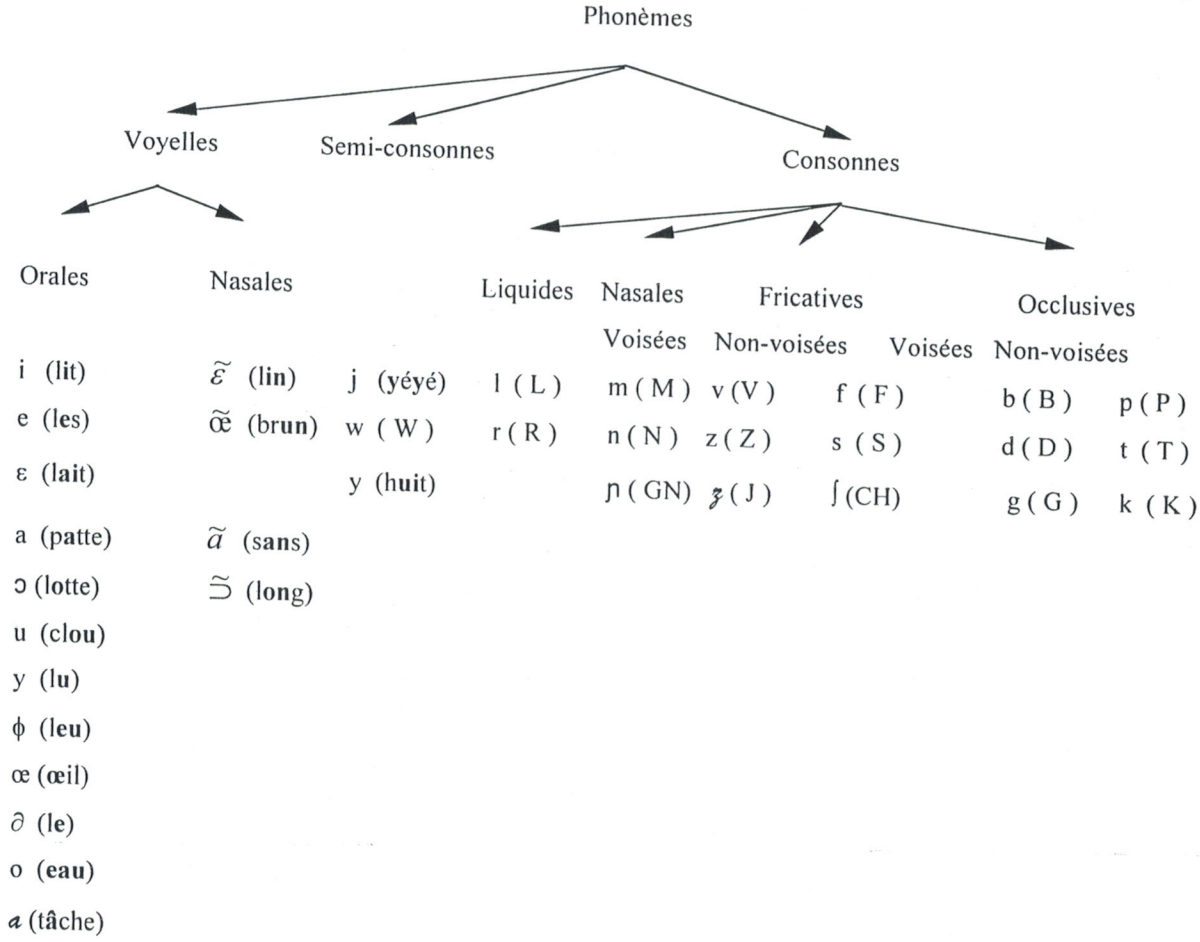


Fig. I - 4 Les phonèmes de la langue française

Les phonèmes peuvent êtres rangés en catégories selon des "traits distinctifs" qui indiquent une similitude au niveau articulatoire, acoustique ou perceptif. Les voyelles peuvent être rangées selon [6] :

- La nasalité;
- l'ouverture du conduit vocal;
- la position de la constriction du conduit vocal;
- l'arrondissement des lèvres.

Les consonnes sont classées selon :

- le voisement;
- le mode d'articulation (occlusif, nasal, fricatif);
- le lieu d'articulation (labiale, dentale, palatale).

Les voyelles orales sont des sons voisés qui impliquent le conduit vocal (cavité bucco-pharyngale) et non la cavité nasale. Tandis que pour les voyelles nasales, le conduit nasal est couplé à la cavité buccale par abaissement du voile du palais qui met en communication le conduit nasal et le conduit oral. L'émission s'effectue par la suite à la fois par les narines et par la bouche [2].

Les semi-consonnes comme les voyelles sont des sons voisés sans source de bruit [1].
Dans les liquides, l'obstruction n'existe que sur une petite largeur au centre du conduit vocal, laissant un écoulement libre sur les côtés [1].

Une source périodique liée à la vibration des cordes vocales s'ajoute à une source de bruit continue pour les fricatives voisées.
Les fricatives non voisées résultent d'une turbulence créée par le passage de l'air dans une constriction du conduit vocal.

Les consonnes occlusives se produisent quand une forte pression maintenue en un certain point du conduit vocal est relâchée brusquement. Pour les occlusives voisées, un son de basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées, la tenue est un silence [2].

I - 4 - 3 Classification des phonèmes de l'arabe standard

L'arabe standard ne contient que des voyelles orales (d'après la classification des phonèmes faite pour l'ASM, Arabe Standard Moderne, fig. I-5) qui sont des voyelles brèves : *fatha* /a/, *kasra* /i/ et *domma* /u/.

A chaque voyelle brève correspond une voyelle longue de même qualité (même aperture et même lieu d'articulation) mais de durée plus longue. Il s'agit de *alif* / $\bar{a}$ /, *ya* $\bar{a}$ *-maddiyah* / $\bar{i}$ / et *waw_maddiyah* / $\bar{u}$ / [5].

Reste les consonnes qui sont au nombre de 28 pour les locuteurs d'arabe standard et classées comme suit (fig. I-5) :

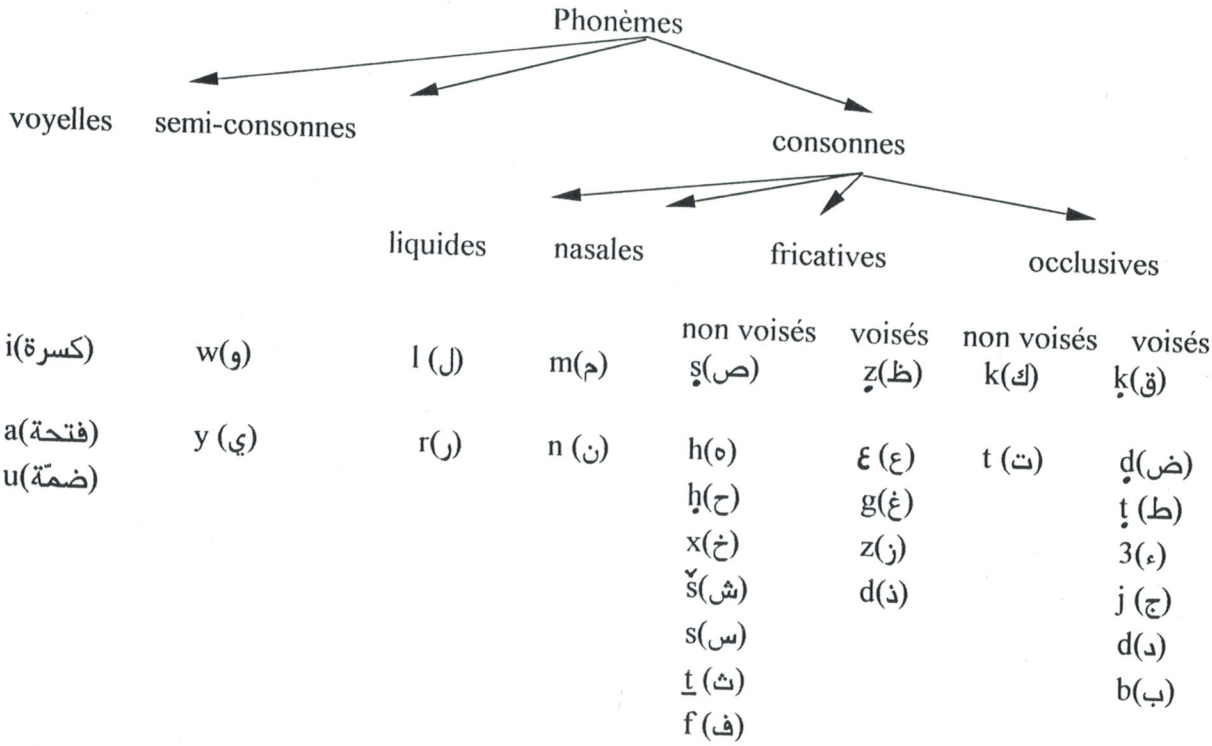


Fig. I - 5 Les phonèmes de la langue arabe

I - 4 - 4 Les résonances (formants)

Le conduit vocal constitue une cavité de résonance. Il possède une fonction de transfert. Les fréquences de résonance dépendent du volume d'air et de la forme du résonateur, elles sont dites " fréquences de formants " ou " formants ". Les formants sont numérotés dans leur ordre d'apparition depuis les fréquences basses jusqu'aux fréquences hautes : F_1, F_2, F_3, …., F_n.

Le spectre de la parole émise à la sortie des lèvres est le produit du spectre de la source et de la fonction de transfert. On parlera dans ce cas d'une représentation " *tout pôle* ".

Un des traits caractéristiques des voyelles, lorsqu'on les observe dans le domaine spectral, est la présence de formants, c'est à dire de bandes de fréquence dont l'énergie est particulièrement élevée. Les deux premiers formants suffisent pour les caractériser, comme le montre le tableau I-1 ci-après. Le premier formant renseigne sur l'aperture de la voyelle, le second sur le caractère antérieur - postérieur de la voyelle [1].

Voyelle	F_1 (en Hz)	F_2 (en Hz)	Voyelle	F_1 (en Hz)	F_2 (en Hz)
i	250	2 500	y	250	1 800
e	375	2 200	ϕ	375	1 600
ε	550	1 800	œ	550	1 400
a	750	1 350	∂	500	1 500
ɔ	550	950	o	375	750
u	250	750	ɑ	750	1 200

Tableau I-1 Les voyelles et leurs formants

Les voyelles sont souvent représentées positionnées sur un plan, dont les axes sont les formants F_1 et F_2. Elles tracent alors un triangle dont les extrémités sont occupées par les voyelles "extrêmes", c'est-à-dire [a], [u], [i]. Ce triangle représente également les positions moyennes de la langue dans la bouche selon deux axes (fig. I - 6) :

- Antérieur à postérieur;
- fermé à ouvert.

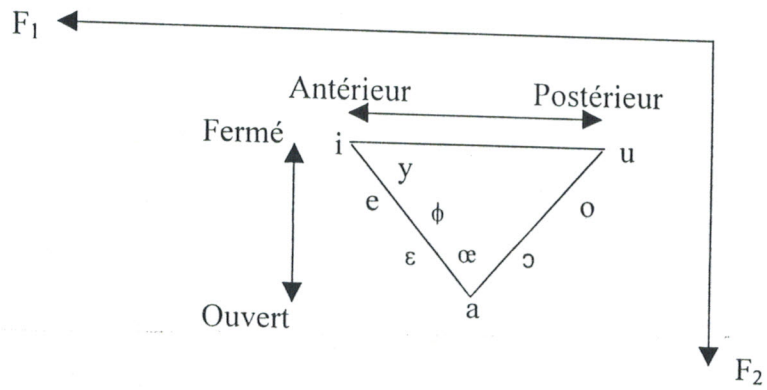


Fig. I - 6 Le triangle acoustique des voyelles orales du français

I - 5 Modélisation de la production de la parole

L'utilisation du formalisme des systèmes linéaires discrets et l'absence de couplage entre la glotte et le conduit vocal permet de modéliser séparément la source et le système de filtrage linéaire.

Ainsi la production de la parole peut être modélisée par le modèle simplifié représenté sur la figure I-7.

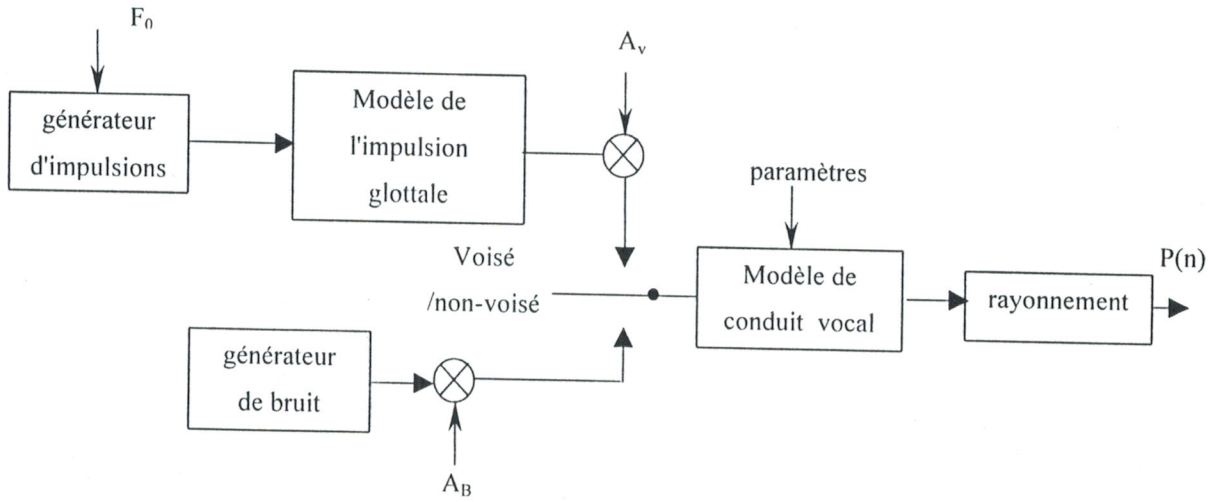


Fig. I - 7 Modèle de production de la parole

Deux types de sources sont requises suivant qu'on produit des sons voisés ou non-voisés :

- une source produisant une onde périodique (générateur de train d'impulsions), ce signal excite à son tour un filtre linéaire dont la réponse impulsionnelle approxime la forme de l'onde glottique suivie par un contrôle de gain qui permet d'ajuster l'amplitude du voisement,
- un générateur de bruit blanc à distribution gaussienne, suivi d'un paramètre de gain permettant un contrôle de l'amplitude du signal.

La modélisation source-filtre du comportement du conduit vocal est destinée surtout à représenter les caractéristiques formantiques des différents sons de la parole. En traitement du signal, on parle d'une représentation tout pôle.

Une bonne approximation du signal de la parole peut être obtenue en modélisant chaque formant par un filtre résonateur passe-bande tout pôle; caractérisé par sa fréquence de résonance F_i (formant) et sa bande passante B_i. La fonction de transfert du conduit vocal est alors obtenue en connectant les filtres résonateurs en cascade.

Cette configuration ne permet pas de modéliser correctement les sons fricatifs et plosifs. Une source périodique liée à la vibration des cordes vocales s'ajoute à la source de bruit pour les fricatives voisées, voilà une des limitations du modèle tout pôle. La deuxième étant la production des sons nasalisés qui fait intervenir deux cavités associées en parallèle, et donc la transmittance correspondante possède des zéros [1].

Le signal vocal n'est pas un signal stationnaire, les paramètres du modèle sont donc variables. Pendant des intervalles de temps de l'ordre de 20 à 30 ms, les déformations du conduit vocal sont assez lentes pour que les coefficients de la fonction de transfert puissent être maintenus constants [1].

I - 6 Perception de la parole

I - 6 - 1 L'audition

L'audition est une fonction complexe, assurée conjointement par l'oreille et le système nerveux, grâce à laquelle le monde extérieur est perçu [7].
Dans une première phase, l'oreille transforme l'information contenue dans le signal acoustique et la transmet au cerveau par l'intermédiaire du nerf auditif.
La deuxième phase correspond à la reconstitution du message linguistique[6].

I - 6 - 2 Mécanisme de l'audition

- Anatomie de l'oreille

L'oreille de l'homme est composée de trois parties anatomiques distinctes : l'oreille externe, l'oreille moyenne et l'oreille interne (fig. I - 8).

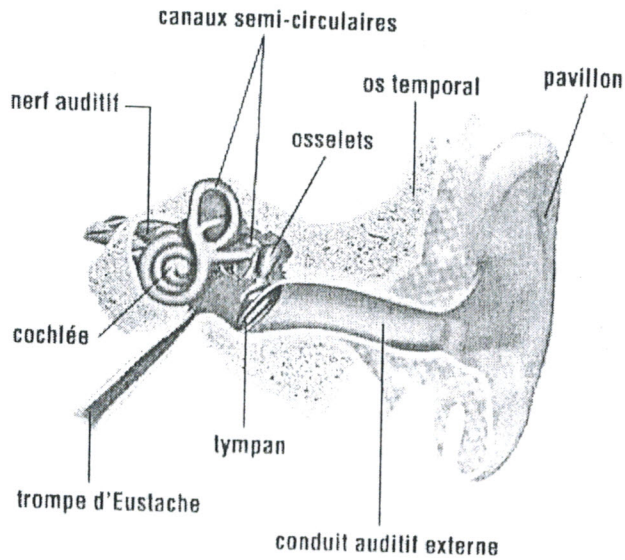


Fig. I - 8 Anatomie de l'oreille

- L'oreille externe comprend le pavillon et le conduit auditif externe, qui conduit au tympan, elle permet de recueillir les sons et de les amplifier [7]. Le conduit auditif est revêtu d'un épiderme qui contient de nombreuses glandes cérumineuses; le tympan, membrane de tissus conjonctifs fibreux, semi-rigide, située au fond du conduit, est disposé en oblique par rapport à l'axe de ce dernier. Il joue le rôle d'un microphone de pression. Sa face externe est concave et recouverte de peau; sa face interne, convexe, est recouverte d'une muqueuse [8].

- L'oreille moyenne, remplie d'air, est isolée de l'oreille externe par la membrane tympanique, et de l'oreille interne par une cloison osseuse percée de deux petites ouvertures : la fenêtre ovale (ou fenêtre vestibulaire), et la fenêtre ronde (dite aussi fenêtre cochléaire). La cavité tympanique communique avec le rhino-pharynx par la trompe d'Eustache. A l'intérieur de la caisse tympanique se trouve placée la chaîne des osselets, constituée par trois petits osselets articulés entre eux : le marteau dont le «manche» est lié à la face interne du tympan, l'enclume et l'étrier.

L'oreille moyenne assure la fonction de transmission proprement dite. Elle peut être assimilée à un transformateur d'impédance. Les vibrations aériennes mettent en mouvement le tympan, dont la vibration, transmise par la chaîne des osselets à travers la fenêtre ovale, aura pour conséquence une oscillation des liquides de l'oreille interne [8]. Un autre phénomène purement physique est lié à la position des osselets les uns par rapport aux autres, qui assure un mécanisme de levier amplificateur [7].

- L'oreille interne située dans une capsule osseuse très compacte, elle regroupe l'organe nerveux de l'audition, la cochlée, et l'organe de l'équilibration ou vestibule. La cochlée est un tube osseux, creux, enroulé en spirale et empli de liquide; appelé aussi limaçon. Il est divisé sur toute sa longueur par deux membranes, la membrane dite «de Reissner» et la membrane basilaire(fig. I - 9)[8]. Cette partie assure la fonction de perception. Comme pour d'autres organes sensoriels, le processus commence par l'élaboration d'un message nerveux à partir d'un phénomène non nerveux, en l'occurrence la vibration d'un liquide, grâce à l'intervention de cellules spécialisées [7].

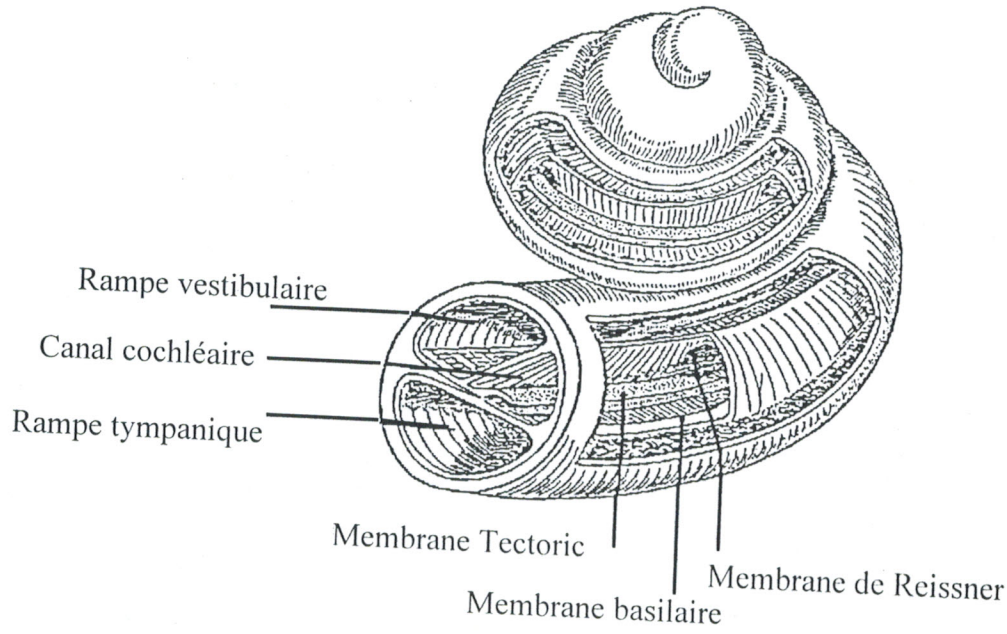


Fig. I - 9 La cochlée

L'organe sensoriel proprement dit, l'organe de Corti, repose sur la membrane basilaire. Il est composé de deux types de cellules ciliées adaptées à la réception des vibrations. Chez l'homme, on dénombre environ 3 500 cellules ciliées internes et 12 000 cellules ciliées externes. Au sommet de ces cellules se trouvent des cils rigides, les sténocils, alignés en 3 ou 4 rangées de taille croissante. Ils sont composés de filaments d'actine qui assurent leur flexibilité et leur rigidité [8].

- Sensations auditives

Les réponses attendues des auditeurs aux stimulus sont l'audibilité ou non d'un son, une modification ou non des sensations auditives, la classification des sons entendus dans une échelle, . . . etc. Le stimulus n'étant rien d'autre que la pression acoustique sur le tympan. Les principaux caractères des sensations auditives sont [1] :

- L'intensité perçue (*sonie*) : un son devient un stimulus dès qu'il atteint l'oreille. Quand le stimulus correspond d'après sa pression acoustique et sa fréquence, à un son audible, il déclenche une sensation perçue. L'intensité des sons est contrôlée par la force des vibrations, laquelle dépend du débit avec lequel l'air est expiré [7]. Les sons s'ordonnent dans une échelle de faible à fort. La sonie est appelée aussi *intensité sonore subjective*.

La sonie est caractérisée par des lignes d'isosonie, chacune de ces courbes (fig. I-10) relie les coordonnées (niveau de pression acoustique et fréquence) des sons purs qui donnent à l'oreille humaine une égale sensation d'intensité. Pour qu'un son pur de 100 Hz soit perçu avec la même intensité subjective qu'un son de 1000 Hz, sa pression sonore doit être plus élevée.

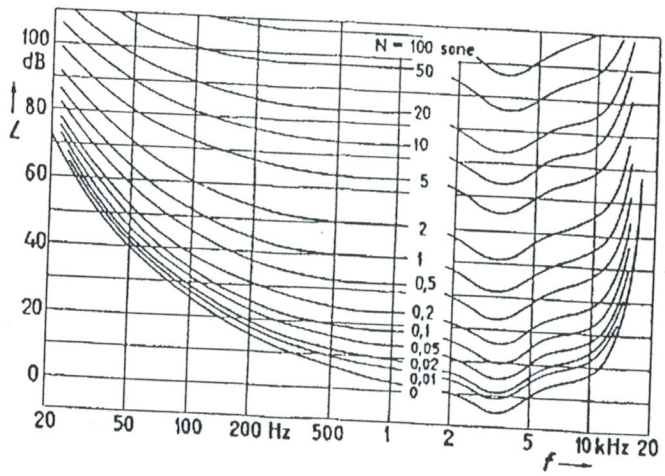


Fig. I - 10 Les courbes d'isosonie en champs libre

Sachant que pour mesurer la sonie d'un son pur, on maintient sa fréquence et sa durée. A un son de niveau acoustique L = 40 dB, de fréquence 1 KHz et de durée 1s, on attribue arbitrairement une sonie

N = 1 sone

Une échelle linéaire des intensités serait difficile à utiliser en psychoacoustique. En effet, entre le seuil d'audition et le seuil de douleur, l'étendue des intensités acoustiques susceptibles de stimuler l'oreille est immense. L'intensité juste perceptible (seuil absolu) est 10^{12} fois moins forte que l'intensité maximale (seuil de la douleur). On préfère donc utiliser une échelle logarithmique allant de 0 à 120 dB pour représenter les intensités sonores [9].

- La hauteur perçue (*tonie*) : la hauteur de la voix est déterminée par la longueur, la forme et la position des cordes vocales [7]. La tonie d'un son dépend principalement de la fréquence, de manière non linéaire, mais aussi de la pression acoustique.

La fréquence qui fixe la hauteur est dite " fréquence du fondamental " ou " pitch ". Celle-ci représente le rythme d'accolement des cordes vocales. Elle varie selon une voix masculine, féminine ou d'enfant. Les sons s'ordonnent ici dans une échelle de hauteur du grave à l'aigu [2].

La fréquence fondamentale peut varier de 80 à 200 Hz pour une voix masculine, de 150 à 450 Hz pour une voix féminine et de 200 à 600 Hz pour une voix d'enfant [1].

L'unité de tonie, correspondant à l'unité Hz de fréquence, est le ***mel***. En dessous de 500 Hz, la valeur de la tonie en mels est pratiquement identique à la valeur de la fréquence en Hz. Mais, au-delà de 1000 Hz, il faut plus que doubler la fréquence pour percevoir un doublement de la hauteur.

- Le timbre : c'est l'ensemble des caractéristiques qui permettent de différencier une voix. Il provient en particulier de la résonance dans la poitrine, la gorge, la cavité buccale et le nez [7]. Et ce sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre d'un son.
- La durée : un des caractères des sensations auditives est la durée d'un son.

La possibilité de distinguer la fréquence, les sons graves et les sons aigus, résulte de la structure même de la cochlée. En effet, la membrane sur laquelle reposent les cellules sensorielles, et qui s'enroule en coquille, n'est pas homogène, et n'a donc pas les mêmes propriétés physiques sur toute sa longueur. Elle est de plus en plus large, fine et souple de la fenêtre ovale jusqu'au sommet. Son maximum d'amplitude de vibration se situe près de la fenêtre ovale pour les sons aigus, près du sommet pour les sons graves. Or, les points de départ des différents neurones, sensibles chacun à une certaine gamme de fréquences, sont échelonnés tout au long de la membrane. Dans la zone de fréquence comprise entre 500 et 8000 Hz, l'oreille peut détecter une variation de 0.03 % dans la fréquence du son. Elle est moins sensible pour les sons de fréquences basses ou de faible intensité[7].

I - 6 - 3 Aire d'audition

L'aire d'audition de l'homme est comprise entre le seuil d'audition (qui varie de 0 à 40 dB suivant la fréquence) et le seuil de douleur (qui se situe aux alentours de 120 dB)[1]. La gamme de fréquence peut aller donc de 20 à 20 000 Hz mais elle est souvent limitée à 5 000 Hz. Soit en tous 10 octaves [2].

I - 6 - 4 Effet de masque

Lorsqu'on entend simultanément deux sons de fréquences différentes, il arrive que l'un d'entre eux devienne inaudible. Cet effet de masque, qui peut être total ou partiel, dépend des intensités et fréquences relatives des deux sons, appelés *son masqué* et *son masquant*.
Pour chaque bruit utilisé comme son masquant, on obtient une courbe d'effet de masque spécifique. Ces courbes peuvent être considérées comme des translations du seuil d'audition vers les niveaux plus élevés. Le bruit peut être un bruit blanc à bande large, uniformément masquant ou à bande étroite [1].

I - 6 - 5 Bandes critiques

L'oreille humaine peut différencier plus de 600 hauteurs différentes et a la faculté d'intégrer certaines zones de fréquence en bandes appelées *bandes critiques*.
Les courbes d'effet de masque correspondant à des bruits blancs montrent que les seuils d'audition sont plats jusqu'à environ 500 Hz puis croissent au-delà de 500 Hz de 10 dB lorsque la fréquence est multipliée par 10.
On peut expliquer l'allure des courbes si l'on admet que l'oreille prend en compte la puissance du bruit blanc non pas de 20 Hz à 20 KHz, mais sur des zones de fréquence relativement étroites. Ces zones ont approximativement la même largeur en dessous de 500 Hz et au-delà de 500 Hz leur largeur doit croître proportionnellement à la fréquence [1].
L'oreille doit donc pouvoir percevoir en dessous de 500 Hz, l'intensité acoustique mesurée dans des bandes de fréquence de largeur constante et au-delà de 500 Hz, l'intensité mesurée dans des bandes de fréquences de largeur relative constantes.
En dessous de 500 Hz, la largeur de bandes critiques est pratiquement indépendante de la fréquence et vaut à peu près 100 Hz. Au-dessus de 500 Hz, elle augmente proportionnellement à la fréquence centrale [1], [6].
La décomposition du spectre de fréquence en bandes critiques correspond à une propriété fondamentale de l'ouïe. L'ouïe peut former une bande critique en n'importe quel point de l'échelle des fréquences. En les rangeant l'une à coté de l'autre, on trouve dans la zone de fréquences de 20 Hz à 16 KHz 24 bandes critiques. Le tableau I-2 montre les fréquences centrales f_m, les largeurs de bandes Δf_G et les fréquences frontières f_g des bandes critiques juxtaposées.

N°	f_m (Hz)	Δf_G (Hz)	f_g (Hz)
1	50	80	100
2	150	100	200
3	250	100	300
4	350	100	400
5	450	110	510
6	570	120	630
7	700	140	770
8	840	150	920
9	1 000	160	1 080
10	1 170	190	1 270
11	1 370	210	1 480
12	1 600	240	1 720
13	1 850	280	2 000
14	2 150	320	2 320
15	2 500	380	2 700
16	2 900	450	3 150
17	3 400	550	3 700
18	4 000	700	4 400
19	4 800	900	5 300
20	5 800	1 100	6 400
21	7 000	1 300	7 700
22	8 500	1 800	9 500
23	10 500	2 500	12 000
24	13 500	3 500	15 500

Tableau I-2

Les bandes critiques

L'intervalle critique joue un rôle essentiel dans la sensation de force sonore. L'unité mel dans ce cas est inadaptée. Elle évoquerait une sensation de hauteur et non une sensation d'intensité. On utilise alors une unité plus grande, le ***Bark***, en mémoire de Barkhausen, tel que [1] :

1 Bark =100 mels.

D'une autre manière, la bande critique correspond à l'écartement en fréquence nécessaire pour que deux harmoniques soient discriminées dans un son complexe.

I - 7 Représentation du signal vocal

Ayant déjà vu la représentation temporelle et spectrale des signaux sons dans les paragraphes précédents il reste à montrer le sonagraph et le spectrogramme.

I - 7 - 1 Le spectrographe

Le spectrographe est le plus ancien outil utilisé par les phonéticiens pour caractériser la parole. Appareil analogique, il a été supplanté par les calculateurs mettant en œuvre des

algorithmes de TFR ou de TFD récursive. Il est ainsi possible, en utilisant des processeurs de signaux, d'obtenir des spectres en temps réel.

I - 7 - 2 Le spectrogramme

Une manière aisée de décrire le signal acoustique est d'utiliser une représentation sous forme de spectrogramme.

Le spectrogramme est une représentation tridimensionnelle, où le temps est représenté sur l'axe X, la fréquence sur l'axe Y et l'intensité de chaque composante spectrale sur l'axe Z, est symbolisé par le niveau de noir comme le montre la figure ci-dessous (fig. I - 11) [2].

Cette analyse temps-fréquence, d'abord réalisée de manière analogique à l'aide de bancs de filtres, est maintenant réalisée de manière numérique par TFR.

Les formants sont repérés par des bandes de fréquence dont l'énergie est particulièrement élevée. Dans le spectrogramme, les formants apparaissent sous forme de bandes sensiblement parallèles à l'abscisse.

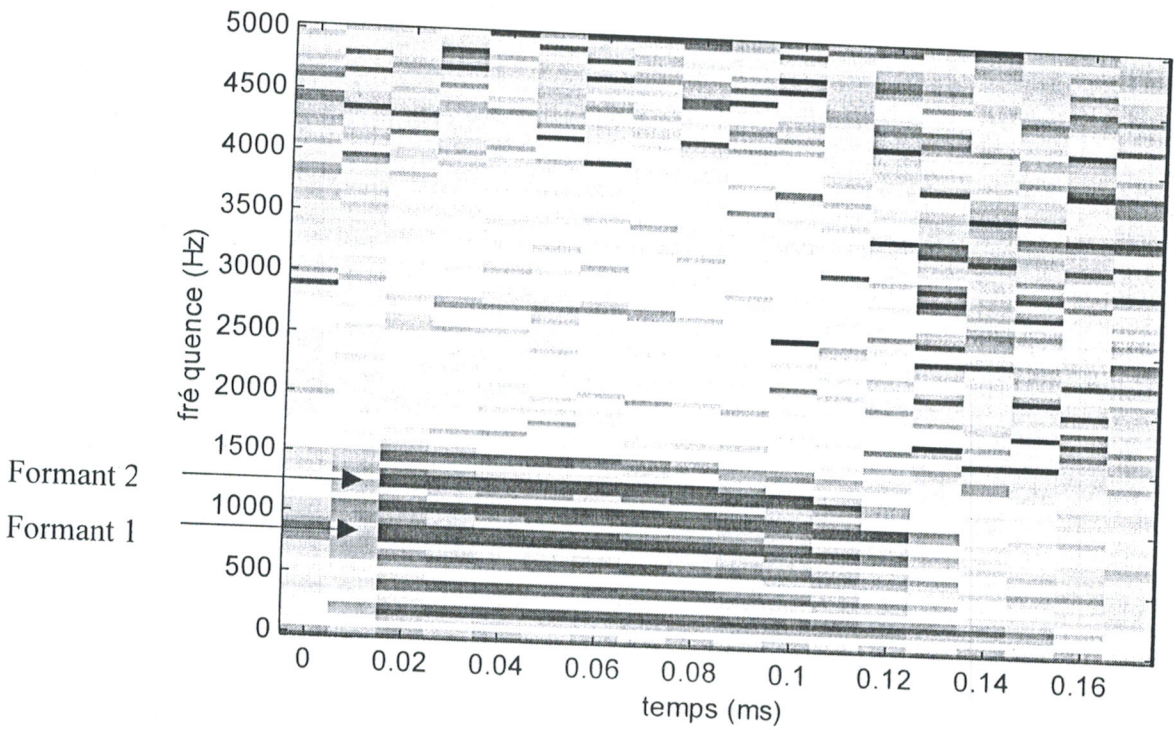


Fig. I - 11 Spectrogramme de la voyelle /a/

Chez un locuteur masculin, tous les formants sont inférieurs à 5000 Hz. Si les formants F_1 et F_2, voire F_3, sont bien marqués, les formants F_4 et F_5 sont plus difficiles à retrouver dans le spectre.

I – 8 Conclusion

Ayant passé en revue les propriétés du signal vocal (variabilité, continuité, ...) ainsi que celles de l'audition (sonie, tonie, ...), il faut faire un traitement de ce signal pour pouvoir l'exploiter dans différents domaines tels que la robotique ou la dictée (Via-voice), synthèse de la parole, rééducation des malentendants, les services vocaux interactifs, ... etc.

Un traitement est utilisé pour réduire la redondance du signal vocal. Cette réduction permet de comprimer l'onde avant le stockage ou la transmission.

L'étape de traitement est en fait une analyse qui tente de déterminer les indices (des paramètres) et à calculer leurs caractéristiques qui sont pertinents pour la reconnaissance.

Ces paramètres sont obtenus à partir du signal d'origine, après application de fonctions dans les domaines : temporel, spectral ou spectro-temporel.

Ces méthodes d'analyse sont exposées dans le chapitre suivant.

CHAPITRE II

ANALYSE DE LA PAROLE

II - 1 Introduction

La parole est un signal réel, continu, d'énergie finie et non stationnaire. Ce signal est périodique pour les sons voisés, aléatoire pour les sons fricatifs et impulsionnel dans les phases explosives des sons occlusifs.

Le traitement du signal vocal a pour but de fournir une représentation moins redondante de la parole que celle obtenue par l'onde temporelle tout en permettant une extraction précise des paramètres significatifs. L'analyse de la parole consiste à estimer les paramètres du modèle de production de la parole en se basant sur des techniques de modélisation qui varient selon l'objectif fixé.

Différentes méthodes d'analyse sont disponibles, entre autre nous citons :

- Les transformées à court terme : temporelles, spectrales et spectro-temporelles, offrent un outil mathématique rigoureux. Cependant, elles ne se réfèrent pas toujours à un modèle de production ni de perception et ne permettent pas de dissocier l'action de la source de celle du conduit vocal.
- Les méthodes fondées sur la déconvolution source/conduit : homomorphiques et celles basées sur la prédiction linéaire. Elles n'ont pas le défaut des transformées à court terme mais reposent sur un modèle de production souvent imprécis [1], [2].

II - 2 Représentation numérique du signal

La représentation numérique du signal vocal implique :

1. Un échantillonnage effectué à une fréquence F_e adéquate donc qui satisfait le théorème de Shannon,
2. La quantification de chaque échantillon suivant la précision souhaitée,
3. Un codage approprié.

Le coût d'un traitement numérique, filtrage, transmission ou enregistrement peut être réduit si l'on accepte une limitation du spectre par un filtrage préalable d'où le choix de la fréquence d'échantillonnage et d'un pas de quantification.

La fréquence d'échantillonnage doit être le double de la plus haute fréquence contenue dans le signal de la parole. Le spectre de la parole s'étend à 5 KHz ou même jusqu'à 8 KHz, il faut donc choisir une fréquence F_e entre 10 et 16 KHz. Le codage se fait sur 8, 12, 14 ou 16 bits.

Le signal de la parole contient essentiellement des fréquences basses. Il peut donc être filtré à 7 KHz par un filtre passe bas et échantillonné à 16 KHz.

Les coefficients acoustiques du signal de parole sont obtenus à partir d'un signal échantillonné et numérisé [2]. Les principales étapes de l'analyse acoustique sont représentées sur la figure II-1 ci dessous:

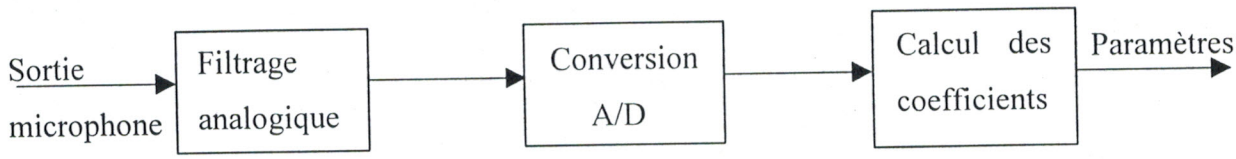


Fig. II - 1 Etapes de l'analyse acoustique

II - 2 - 1 Statistiques à long et à court terme

Un signal vocal est une réalisation d'un processus aléatoire non stationnaire; ses propriétés statistiques moyennes doivent être estimées sur des intervalles de temps de l'ordre de plusieurs dizaines de secondes et pour différents locuteurs. C'est la statistique à long terme. L'estimation des statistiques peut être faite sur les échantillons si le théorème de Shannon est respecté.

La statistique à court terme est faite sur des tranches de 10 à 30 ms durant lesquelles le signal est supposé quasi-stationnaire. Dans ce cas, les blocs successifs traités se recouvrent, on parle d'analyse par fenêtre glissante. Un tel bloc du signal constitue une *trame acoustique*[2].

II - 3 Analyse temporelle à court terme

Certaines mesures sur le signal de la parole donnent déjà suffisamment d'informations. L'évolution de l'énergie à court terme indique la succession des voyelles (très énergétiques) et des consonnes (d'énergie moindre). Le comptage et le tracé d'histogramme des passages par zéro du signal traduisent le contenu spectral [1], [2].

II - 3 - 1 Energie et puissance à court terme

L'énergie à court terme à un instant donné n d'un signal x vaut :

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m).w^2(n-m)$$

avec w(m) la fonction fenêtre (rectangulaire, Hamming, …)

La puissance à court terme est définie par :

$$P_n = \frac{1}{N}\sum_{m=-\infty}^{\infty} x^2(m).w^2(n-m)$$, où N est la longueur de la fenêtre.

Pour se rapprocher d'une résolution comparable à celle de l'oreille humaine, on est amené soit à calculer la FFT sur une échelle non linéaire de mel [2]:

$$z = \frac{1000}{\log(2)} \log(1 + f/1000), \qquad \text{où z est en mel et f en Hz,}$$

soit de prendre des fonctions de pondérations adaptées à des bandes de fréquences de plus en plus larges.

II - 4 - 4 Transformé de Fourier discrète de signaux périodiques de période N

On définit la TFD d'un signal x(n) périodique comme suit [11] :

$X(k) = \sum_{n=0}^{N-1} x(n).e^{-j(2\Pi/N)kn}$ où k=0, 1, ..., N-1 et x(n) étant le signal échantillonné.

La TFD inverse quant à elle (TFDI) est : $x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k).e^{j(2\Pi/N)kn}$ n=0, 1, ..., N-1.

La résolution spectrale [12]

La TFD comme on l'a vu, établit une relation entre deux suites périodiques, les x(n) et X(k) à N éléments différents. Pour cela, on introduit une double périodicité.
La périodicité en fréquence est introduite par l'opération d'échantillonnage du signal sachant que la fréquence d'échantillonnage doit satisfaire le théorème de Shannon.
La périodicité temporelle est introduite artificiellement en supposant que le signal se reproduit en dehors de l'intervalle de temps NT qui correspond à l'enregistrement à traiter. La TFD dans ce cas fournit un échantillon du spectre avec une période fréquentielle f égale à l'inverse de la durée de l'enregistrement et qui constitue la résolution fréquentielle de l'analyse.

D'autre part le fait que le signal ne soit pas composé uniquement de raies aux fréquences multiples de 1/NT entraîne une interférence entre les composantes spectrales obtenues. Cet effet peut être atténué par pondération des échantillons du signal avant transformation.
Cette opération revient à remplacer la fenêtre temporelle rectangulaire par une fonction dont la transformée de Fourier présente des ondulations plus faibles. Parmi ces fonctions, il existe la fenêtre de Hamming (fig. II - 2) :

$$w(t) = 0{,}54 - 0{,}46\cos(2\Pi \frac{t}{NT})$$

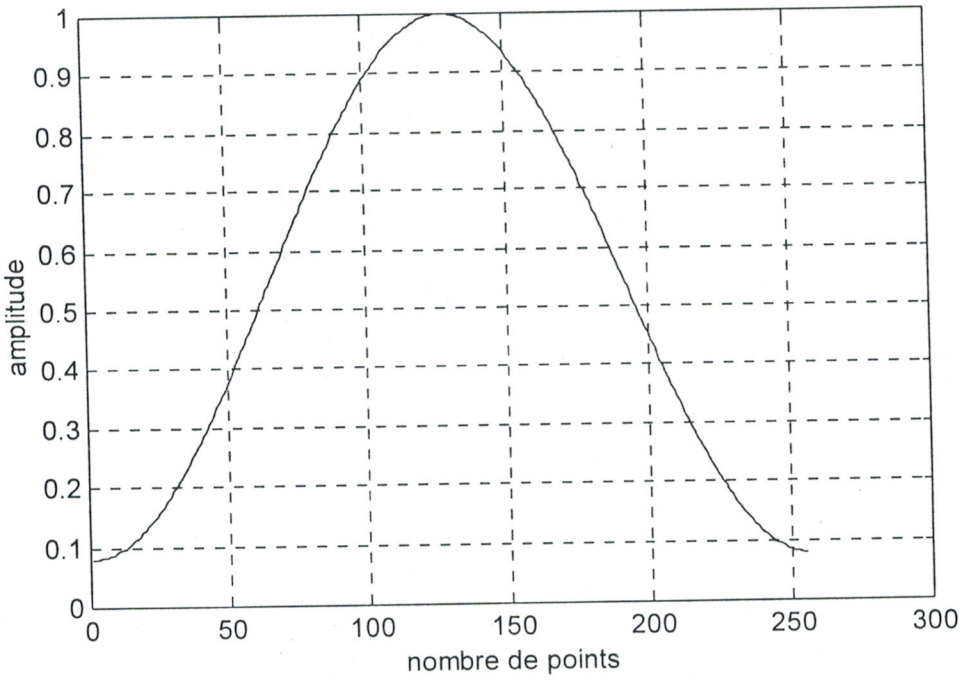


Fig. II - 2 Fenêtre de Hamming sur 256 points

Cette fonction a 99,96 % de son énergie dans le lobe principal. Le lobe secondaire le plus important se trouve à environ 40 dB au-dessous du lobe principal.

II - 4 - 5 La Transformée de Fourier Rapide (TFR)

Le nombre d'opération requis pour le calcul d'une TFD à N points définit par :

$$X(k) = \sum_{n=0}^{N-1} x(n).e^{-j(2\Pi/N)kn}$$

est de N^2 multiplications et N(N-1) additions complexes, qui reste proportionnel à N^2. Pour diminuer ce nombre d'opérations, on utilise des outils mathématiques qui nous permettent par la même occasion de réduire le temps de calcul requis. La méthode généralement suivie est de calculer des TFD de longueurs plus courtes, puis de combiner les résultats de manière appropriée. C'est ce qu'on appelle par *Transformations de Fourier Rapide TFR [11], [13].*

Les figures II-3 et II-4 ci-dessous représentent l'application d'une FFT sur les voyelles [a,i] prononcées par un locuteur féminin.

Ainsi les figures II-3 (a, b) et II-4 (a, b) illustrent les représentations temporelles, et la représentation d'une trame de 256 points relative à une durée de 20 ms durant laquelle le signal est supposé stationnaire des voyelles [a] et [i] respectivement.

Pour la voyelle /a/ on obtient le spectre FFT de la fig. II-3 (d , e) (en représentant le spectre une fois en utilisant une échelle d'amplitude et une autre en dB) en prenant une trame de 256 points du signal original, pondérée par la fenêtre de Hamming (fig. II-3 (c)).

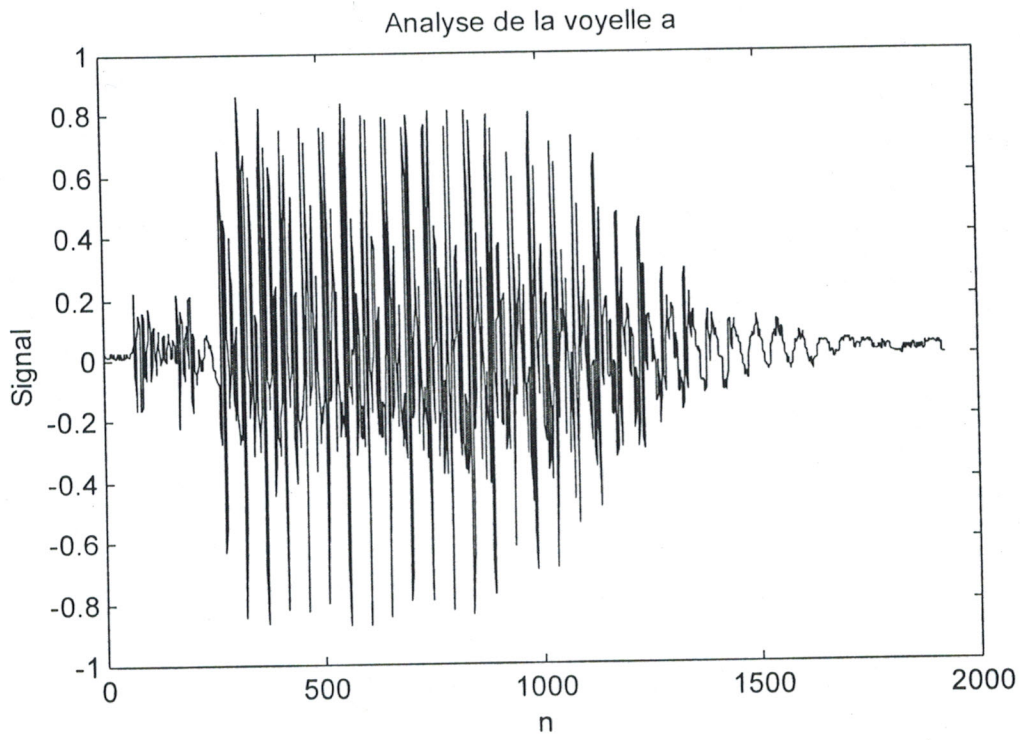


Fig. II – 3 (a) Représentation temporelle de la voyelle / a/

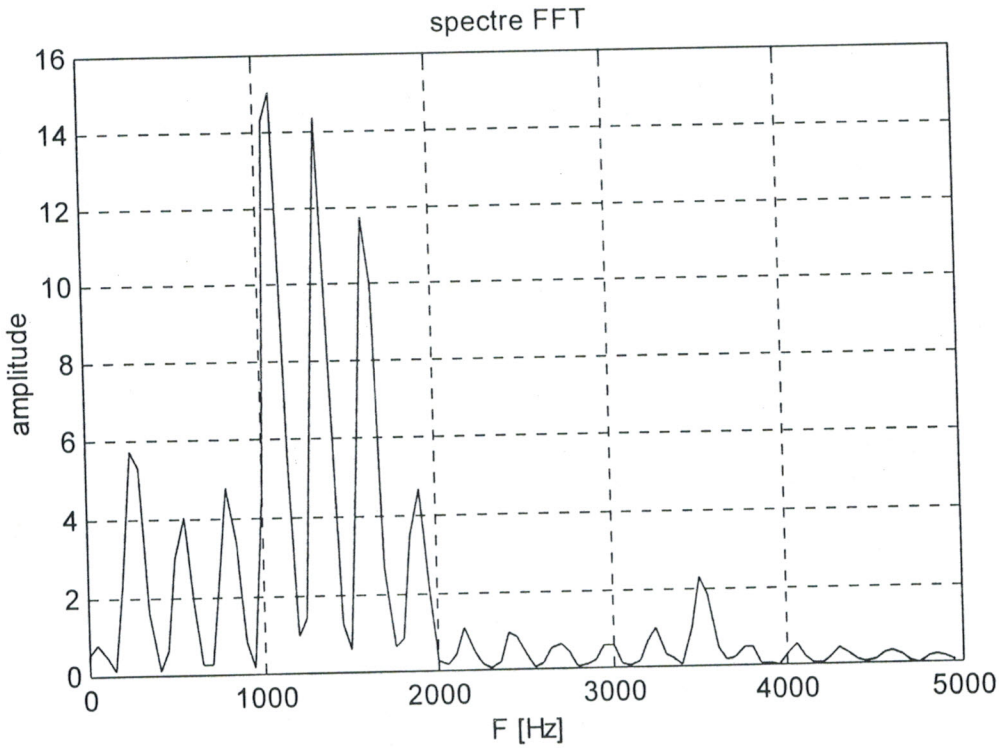


Fig. II – 3 (d) Spectre FFT de la voyelle /a/

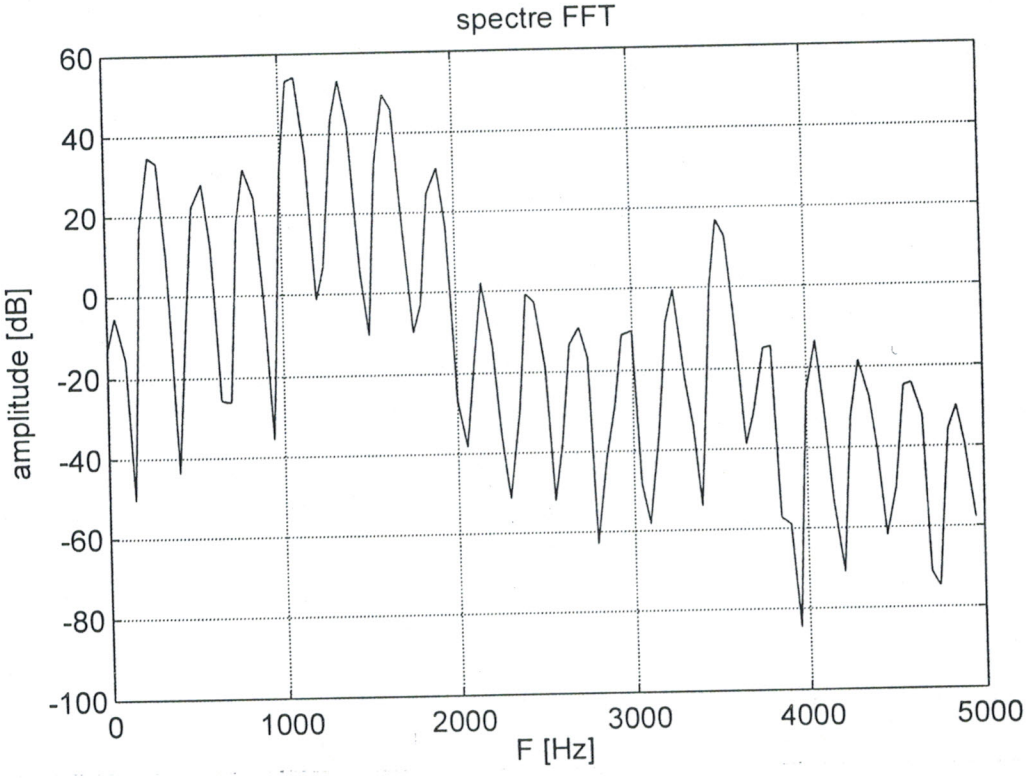


Fig. II – 3 (e) Echelle en dB pour les amplitudes

Pour la voyelle /i/ on obtient le spectre FFT de la fig. II - 8 (d , e) en prenant une trame de 256 points multipliée par la fenêtre de Hamming (fig. II-4(c)).

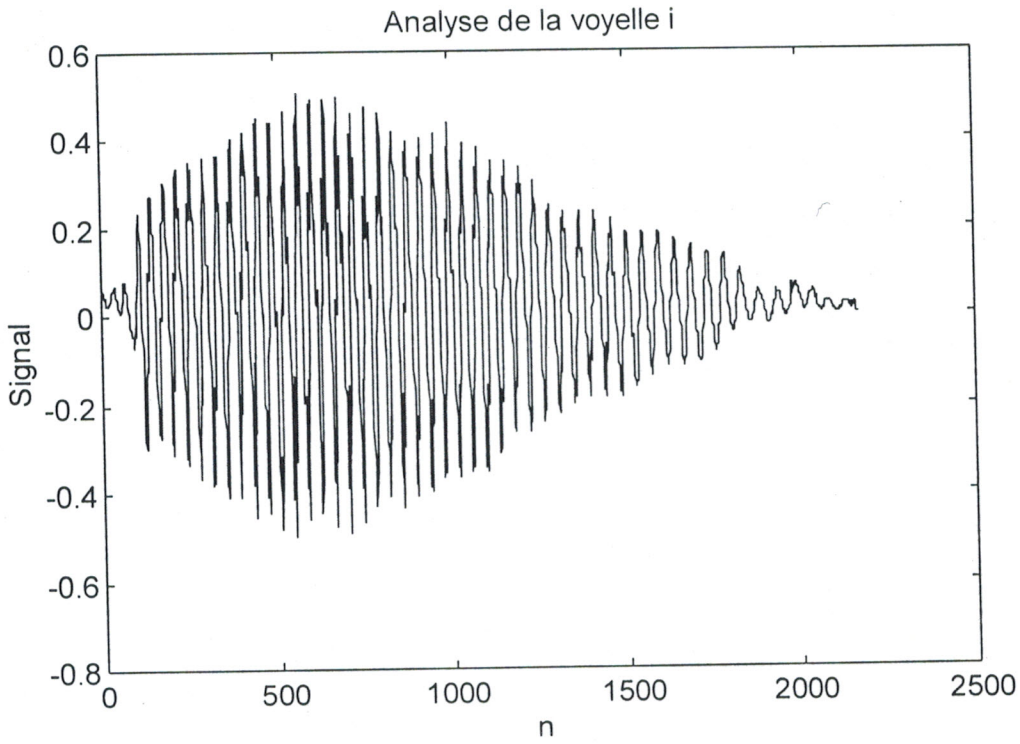


Fig. II – 4 (a) Représentation temporelle de la voyelle /i/

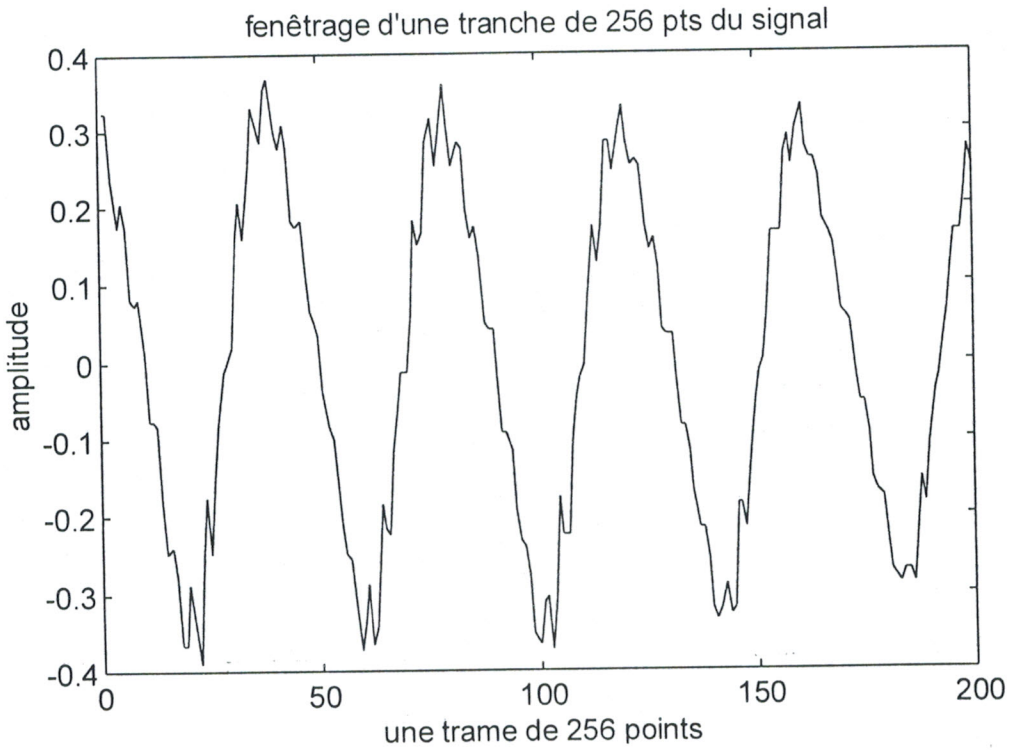


Fig. II – 4 (b) une trame de 256 points de la voyelle /i/

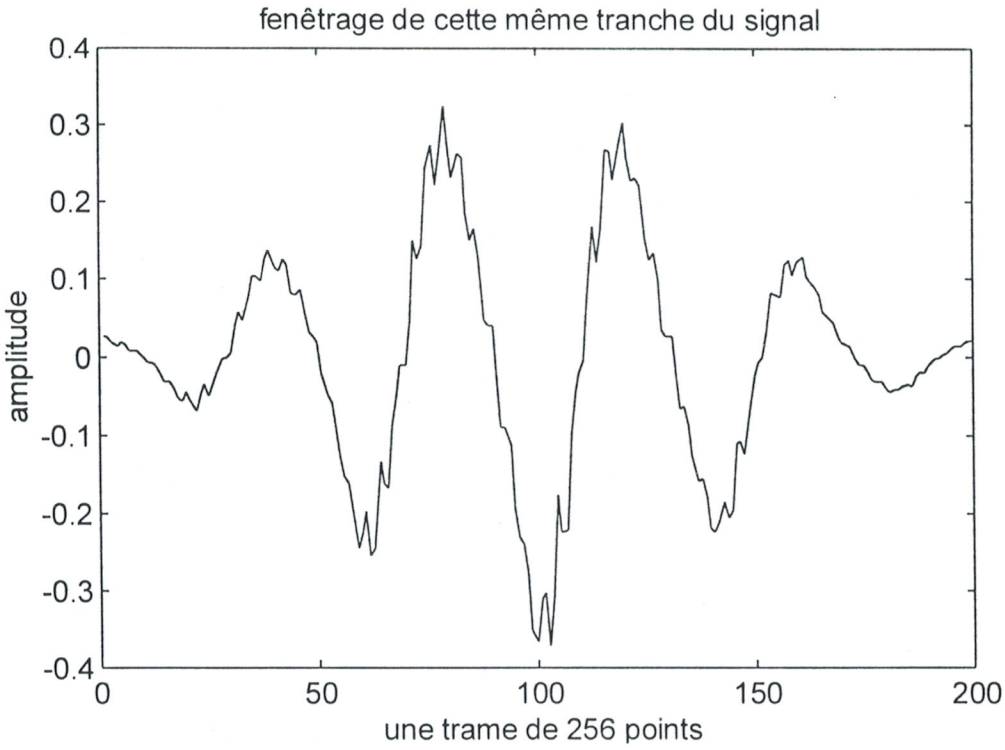


Fig. II – 4 (c) Après multiplication par la fenêtre de Hamming

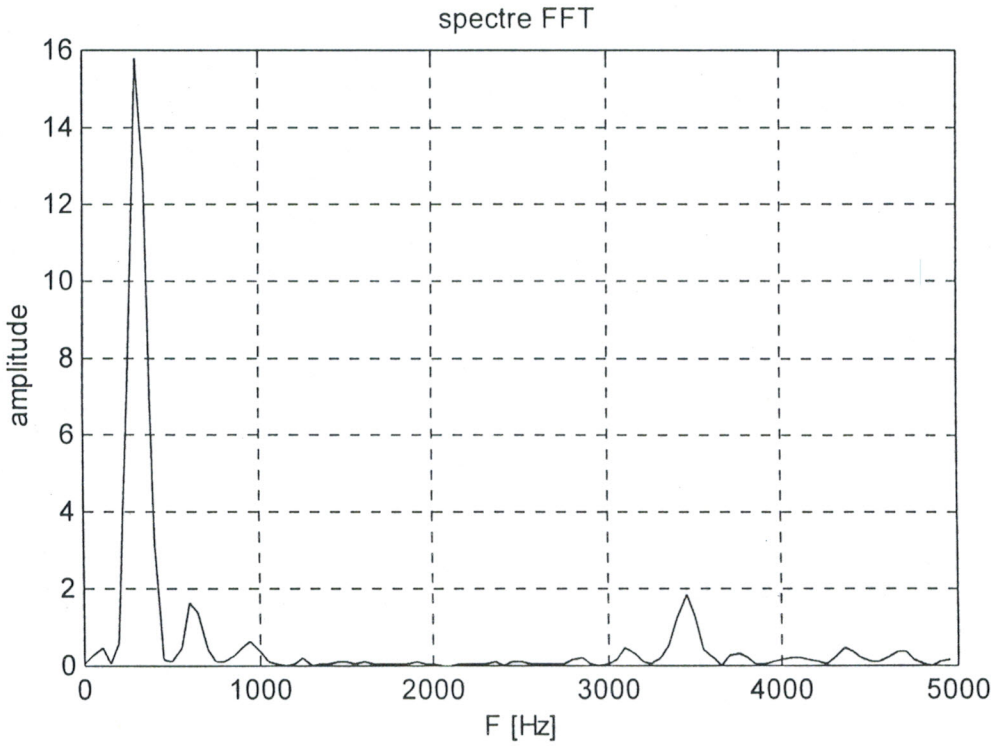


Fig. II – 4 (d) Spectre FFT de la voyelle /i/

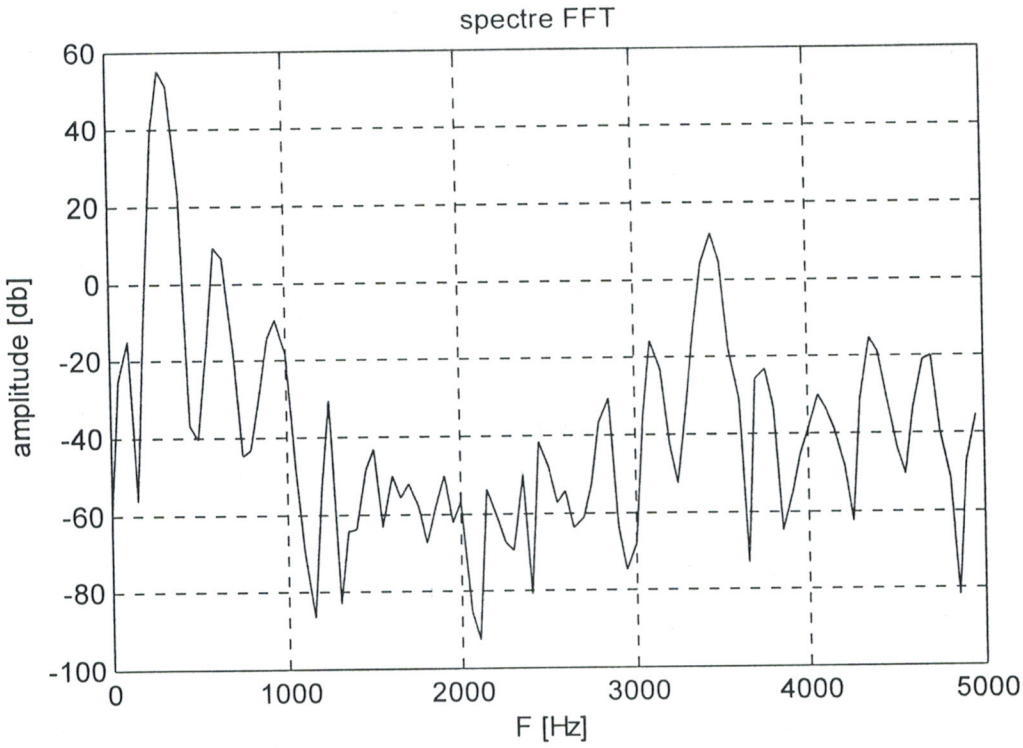


Fig. II – 4 (e) Echelle en dB pour les amplitudes

II - 5 Analyse spectro-temporelle

II - 5 - 1 Bancs de filtres

Les transformations spectro-temporelles sont une généralisation du problème des décompositions en sous-bandes des signaux. Des transformations telles que : les bancs de filtres, les ondelettes, … permettent une meilleure maîtrise de la résolution d'analyse en temps et en fréquence, la STFT en est un cas particulier [1].

II - 6 Analyse homomorphique

II - 6 -1 Principe général

Dans le cas général de la parole, un signal observé x résulte de la convolution d'une excitation u (source glottique) et d'une réponse impulsionnelle h (conduit vocal) : $x(nT)=u(nT) \otimes h(nT)$, où T représente la période d'échantillonnage.

L'analyse homomorphique tente de séparer les contributions respectives de la source et du conduit vocal par "déconvolution" afin d'obtenir le cepstre.

Un ensemble d'étapes est appliqué au signal temporel x(nT) pour générer le cepstre.

- En premier, une transformée de Fourier est appliquée au signal à traiter x(nT). Ainsi une conversion du produit de convolution en multiplication est effectuée : $X(e^{jw_kT}) = H(e^{jw_kT}).U(e^{jw_kT})$, avec X, H, U les transformées de Fourier de x, h et u respectivement, $w_k = \frac{2\Pi}{NT}k$ et N le nombre de points de l'algorithme de la FFT.
- L'étape suivante est l'application d'une fonction logarithmique à X(w) :

$$\log\left|X\left(e^{jw_kT}\right)\right| = \log\left|H\left(e^{jw_kT}\right)\right| + \log\left|U\left(e^{jw_kT}\right)\right|.$$

Une séquence $\log\left|X\left(e^{jw_kT}\right)\right|$ ainsi obtenue qui lie de façon additive les contributions du conduit et de l'excitation.

- Finalement, des traitements linéaires sont mis en œuvre : une TFI qui repasse le signal dans le domaine temporel. Le signal ainsi obtenu est appelé *cepstre* (anagramme du mot spectre) [18].

D'où :

$$u(n) \otimes h(n) \xrightarrow{TF} U(\omega).H(\omega) \xrightarrow{Ln} \hat{U}(\omega) + \hat{H}(\omega) \xrightarrow{TF^{-1}} \hat{u}(n) + \hat{h}(n).$$

L'analyse homomorphique étant utilisée pour estimer la période du fondamental et l'enveloppe du spectre, on utilise le plus souvent le "*cepstre réel*" [1], [2].

Pour un signal x(n) extrait par une fenêtre rectangulaire de longueur N et w = (2Π / N).k, le cepstre réel est défini par :

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \ln\left|X\left(\frac{2\Pi}{N}k\right)\right| . e^{j\frac{2\Pi}{N}kn}$$

Ce calcul est effectué par un algorithme FFT.

L'espace de représentation du cepstre (espace quéfrentiel : *quefrency*) est homogène au temps. Il est possible par un filtrage temporel (liftrage : *lifter*), de séparer dans le signal, la contribution de la source de celle du conduit. Les premiers coefficients cepstraux contiennent l'information relative au conduit. Cette contribution devient négligeable à partir d'un échantillon n0. Les pics périodiques visibles au-delà de n0, reflètent les impulsions de la source [1], [2].

II - 7 Analyse basée sur la prédiction linéaire

II - 7 - 1 Modèle de production du signal vocal

Le signal est produit par un système dont la transmittance est en première approximation de la forme σ/A(z), soumis à une excitation idéalisée u(n). Pour les sons dits voisés ou sonores (V), cette excitation est un train périodique d'impulsions d'amplitude unité : $u(n) = \sum_k \delta(n - kp)$ où p désigne la période du fondamental (ou pitch) exprimée en nombre de périodes d'échantillonnage.

Pour les sons non voisés (NV), l'excitation est un bruit blanc de moyenne nulle et de variance unité [2], [10], [13].

La transmittance H(z)=σ/A(z) est celle d'un filtre tout pôle. La transformée du signal peut s'écrire : $X(z) = U(z).\sigma / A(z)$ et le polynôme A(z) sera noté par :

$$A(z) = \sum_{i=0}^{p} a(i) z^{-i} \quad , a(0)=1.$$

Ce modèle de production d'un signal est appelé "autoregressif" (AR), car l'expression $X(z) = U(z).\sigma / A(z)$ correspond dans le domaine temporel à la récurrence linéaire:

$$x(n) + \sum_{i=0}^{p} a(i) x(n-i) = \sigma \; u(n).$$

L'interprétation est qu'un échantillon x(n) quelconque est une combinaison linéaire des P échantillons qui le précèdent, à laquelle il faut ajouter le terme d'excitation.

- Les coefficients a(i) sont appelés " coefficients de prédiction ",
- Le coefficient σ est appelé " gain du système AR".

Ce modèle de production est représenté par la figure II-5 ci-dessous :

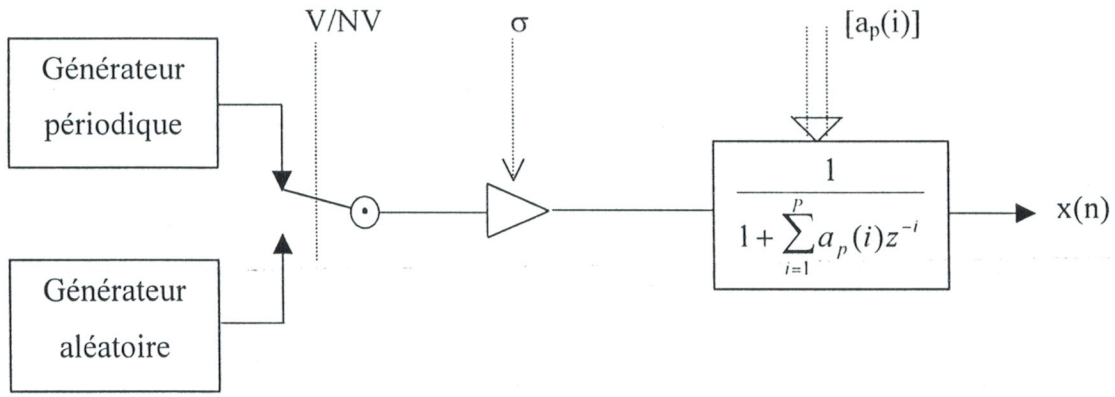


Fig. II - 5 Modèle autoregressif de production de la parole

II - 7 - 2 Prédiction linéaire

Soit un signal x engendré par un certain système autoregressif (fig.II-6). L'excitation de ce système est inaccessible : l'estimation des paramètres du modèle sera donc basée exclusivement sur l'observation du signal. Sachant que le modèle autoregressif obéit à une récurrence, on peut définir une prédiction (estimation) de chaque échantillon x(n) à partir des p échantillons qui le précèdent : $\hat{x}(n) = -\sum_{i=1}^{p} \hat{a}(i)x(n-i)$.

Les coefficients â(i), (i =1, …, p) sont les estimés des coefficients a(i) de la récurrence. L'erreur commise par la prédiction vaut : e(n)=x(n)−$\hat{x}(n)$.

Si â(i) = a(i), (i =1, 2, …, p) l'erreur de prédiction coïncide avec l'excitation à un facteur près : e(n) = σ u(n) [10].

II - 7 - 3 Définition. Filtre inverse

La transmittance du filtre inverse vaut : $A(z) = \sum_{i=0}^{p} a(i)z^{-i}$, a(0) = 1.

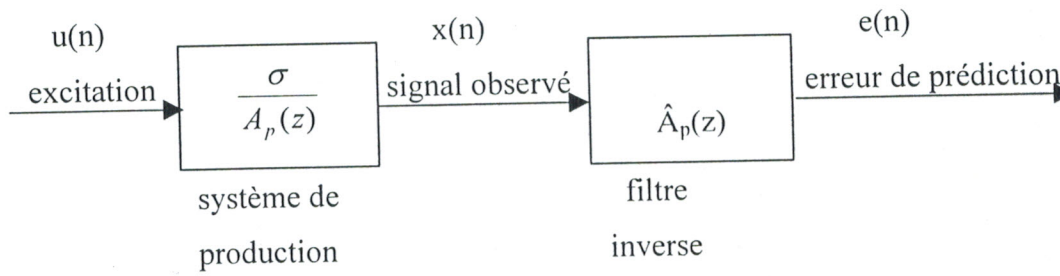


Fig. II - 6 Système AR et filtre inverse

Il ne faut pas non plus oublier la stabilité du filtre étudié, un filtre tout pôle est stable si tous ces pôles sont à l'intérieur du cercle unité, donc un système à phase minimale[10], [14].

II - 7 - 4 Estimation des coefficients de prédiction

Les coefficients de prédiction linéaire doivent être estimés sur une courte durée du signal pendant laquelle il est considéré comme stationnaire [1], [14]. Cette portion du signal est dite 'trame d'analyse'. Il faut choisir les coefficients â(i) qui minimisent l 'énergie résiduelle de prédiction E (erreur quadratique) :

$$E = \sum_{n1}^{n2} e^2(n) = \sum_{i=0}^{p} \hat{a}(i).x(n-i).\sum_{j=0}^{p} \hat{a}(j).x(n-j).$$

Deux méthodes d'analyse par prédiction linéaire découlent du choix des limites de sommation n_1 et n_2 : la ***méthode de corrélation*** et la ***méthode de covariance*** [1], [2], [15].

a) Méthode de corrélation

Le signal est supposé connu de 0 à N-1 et nul ailleurs, ce qui nous ramène à écrire :

$$x(n) = \hat{x}(n) + e(n) = -\sum_{i=1}^{p} \hat{a}(i).x(n-i) + e(n) \quad n = 0, ..., N + p - 1,$$

sous forme matricielle : X . â = e.

Dans ce cas : $E = \sum_{n=0}^{N+p-1} e^2(n) = \sum_{i=0}^{p} \hat{a}(i) \sum_{j=0}^{p} \hat{a}(j) x(n-i) x(n-j)$

$$= \sum_{i=0}^{p} \hat{a}(i) \sum_{j=0}^{p} \hat{a}(j) r(|i-j|).$$

La matrice R de corrélation est de dimension (p+1) × (p+1). Elle est de type Toeplitz, matrice symétrique dont tous les éléments sur une diagonale sont égaux.

$$R = X^T.X = \begin{bmatrix} r(0) & r(1) & \cdots & \cdots & r(p) \\ r(1) & r(0) & \cdots & \cdots & r(p-1) \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ r(p) & r(p-1) & \cdots & \cdots & r(0) \end{bmatrix}$$

tel que : $r(i) = \sum_{n=0}^{N-1-|i|} x(n).x(n+i)$; i = 0, 1, ..., p.

Les coefficients de prédiction peuvent être déterminés par une minimisation au sens des moindres carrés de l'erreur quadratique E.

Cette variance est une forme quadratique définie positive. Le minimum est donc unique, et est obtenu par :

$$\frac{\delta E}{\delta\, a(i)} = \sum_{j=0}^{p} \hat{a}(j) r(|i-j|) = 0 \quad i = 1, 2, ... p \text{ ,qui peut s'écrire :}$$

$$\sum_{j=1}^{p} r(|i-j|) \hat{a}(j) = -r(i) \quad i = 1, 2, ..., p \text{ et } \hat{a}(0)=1.$$

L'écriture matricielle est :

$$\begin{bmatrix} r(0) & \cdots & r(p-1) \\ \vdots & \ddots & \vdots \\ r(p-1) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \vdots \\ \hat{a}(p) \end{bmatrix} = -\begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix}$$

On obtient donc un système d'équations linéaires d'ordre p dit de *YULE – WALKER.*

L'expression du minimum est donnée par :

$$E_{\min} = \sum_{j=0}^{p} \hat{a}(j) r(j)$$

L'écriture matricielle sera :

$$\begin{bmatrix} r(0) & \cdots & r(p) \\ \vdots & \ddots & \vdots \\ r(p) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ \hat{a}(1) \\ \vdots \\ \hat{a}(p) \end{bmatrix} = \begin{bmatrix} E_{\min} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Cette erreur est due à l'hypothèse de base de la méthode de corrélation. On dit que l'opérateur «sort des données ».

Afin de minimiser cet effet, on pondère les échantillons du signal vocal par une fenêtre w(n) symétrique.

Le signal vocal après pondération sera = $\begin{cases} x(n).w(n) & 0 \le n \le N-1 \\ 0 & ailleurs \end{cases}$

On peut prendre comme fenêtre de pondération la fenêtre de Hamming.

b) Méthode de covariance

L'hypothèse faite à ce niveau est que le signal x(n) est connu de 0 à N-1 et il est non défini ailleurs.

$$x(n) = \hat{x}(n) + e(n) = -\sum_{i=1}^{p} \hat{a}(i).x(n-i) + e(n) \quad , \text{n = p, ..., N - 1}$$

La matrice R de covariance est de dimension (p+1) × (p+1). Elle est symétrique mais pas de type Toeplitz.

$$R = X^T.X = \begin{bmatrix} r(0,0) & r(0,1) & \cdots & \cdots & r(0,p) \\ r(1,0) & r(1,1) & \cdots & \cdots & r(1,p) \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ r(p,0) & r(p,1) & \cdots & \cdots & r(p,p) \end{bmatrix}$$

tel que : $r(i,j) = \sum_{n=p}^{N-1} x(n-i).x(n-j)$; i, j = 0, 1, ..., p .

II - 7 - 5 Algorithme de Levinson-Durbin

Cet algorithme est utilisé pour résoudre le système linéaire de p équations à p inconnues, réclamant $O(p^2)$ opérations. Il est récursif sur l'ordre : connaissant le prédicteur optimal d'ordre p-1, on obtient le prédicteur d'ordre p.

Pour cela on suppose connus les coefficients de corrélation r(i), il s'agit maintenant de résoudre le système d'ordre p :

$$\begin{bmatrix} r(0) & \cdots & r(p-1) \\ \vdots & \ddots & \vdots \\ r(p-1) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \vdots \\ \hat{a}(p) \end{bmatrix} = -\begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix}$$

c'est à dire inverser la matrice R d'ordre p, puis calculer : $E = E_{\min} = \sum_{i=0}^{p} \hat{a}(i).r(i)$.

Afin de calculer cette dernière nous allons introduire deux types d'erreurs :

- Erreur de prédiction progressive (avant) d'ordre p : elle prédit une valeur future x(n) à partir des p échantillons précédents x(n-i), i=1, 2, …, p ;

$$f(n) = e(n) = x(n) + \sum_{i=1}^{p} \hat{a}(i)x(n-i) \text{ avec } \hat{a}(0) = 1$$

- Erreur de prédiction rétrograde (arrière) d'ordre p : vérifie une valeur passée x(n-p) à partir des mêmes échantillons ;

$$g(n) = x(n-p) + \sum_{i=1}^{p} \hat{b}(p-i)x(n-p+i) \text{ avec } \hat{b}(p) = 1$$

on note : $E_{\min}^{p} = \alpha_p = \sum_{n=0}^{N+p-1} f^2(n)$

et $E_{\min}^{r} = \beta_p = \sum_{n=0}^{N+p-1} g^2(n)$

on obtient alors les mêmes résultats lorsqu'on travaille avec l'erreur progressive ou rétrograde : $E_{\min}^{p} = E_{\min}^{r} = E_{\min}$.

Pour un ordre m=1, 2, …, p on connaît α_{m-1} et $\begin{bmatrix} \hat{a}_{m-1}(1) \\ \vdots \\ \hat{a}_{m-1}(m-1) \end{bmatrix}$ solution de :

$$R_{m-1}\begin{bmatrix} 1 \\ \hat{a}_{m-1}(1) \\ \vdots \\ \hat{a}_{m-1}(m-1) \end{bmatrix} = \begin{bmatrix} \alpha_{m-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$, on cherche maintenant α_m et $\begin{bmatrix} \hat{a}_m(1) \\ \vdots \\ \hat{a}_m(m-1) \\ \hat{a}_m(m) \end{bmatrix}$.

Après transformation apportée sur la matrice de corrélation et multiplication de celle ci par une constante notée k_m (coefficient de corrélation partielle d'ordre m) tel que :

$\gamma_m + k_m \alpha_{m-1} = 0$ avec $\gamma_m = r(m) + \hat{a}_{m-1}(1) r(m-1) + \ldots + \hat{a}_{m-1}(m-1) r(1)$

on obtient :

$$R_m \begin{bmatrix} 1 \\ \hat{a}_m(1) \\ \vdots \\ \hat{a}_m(m) \end{bmatrix} = \begin{bmatrix} \alpha_m \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

L'algorithme suit alors les étapes ci-dessous :

Etape (0) : $\hat{a}(0)=1$ et $\alpha_0 = r(0)$

Etape (1) :

le 1er coefficient de corrélation partielle k_1 et l'erreur quadratique α_1 sont donnés par :

$$k_1 = \hat{a}_1(1) = -\frac{r(1)}{r(0)} \quad \text{et} \quad \alpha_1 = (1 - k_1^2).r(0).$$

Pour les autres étapes, les autres coefficients de corrélation partielle k_2, …, k_p proviennent de :

$$k_m = -\frac{\gamma_m}{\alpha_{m-1}} \quad m = 2,..,p \text{ sachant que } \gamma_m + k_m \alpha_{m-1} = 0.$$

Les coefficients de prédictions sont donnés par $\hat{a}_m(i)$:

$$\begin{cases} \hat{a}_m(0) = 1 \\ \hat{a}_m(i) = \hat{a}_{m-1}(i) + k_m \hat{a}_{m-1}(m-i) \\ \hat{a}_m(m) = k_m \end{cases}$$

L'énergie résiduelle de prédiction :

$$\alpha_m = \alpha_{m-1} + k_m \gamma_m = \alpha_{m-1}(1 - k_m^2)$$

Cette énergie diminue lorsque m augmente et elle reste constante lorsque $m \geq p$.

L'algorithme de Levinson-Durbin est représenté plus en détails sur le diagramme dans la page suivante.

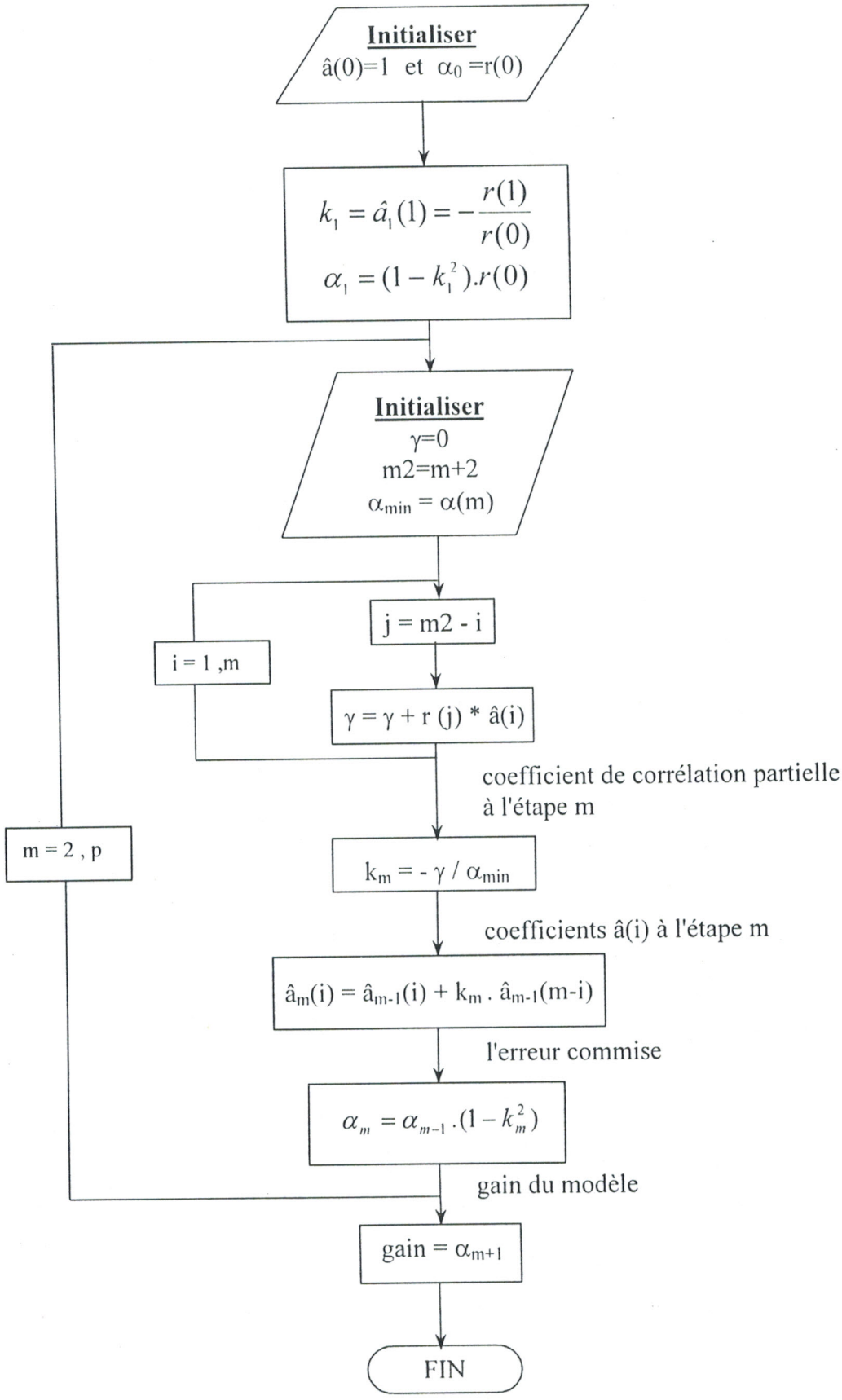


Diagramme de LEVINSON DURBIN

Il est toujours important de vérifier la stabilité du système de transmittance H(z)(§ II-7-3), pour cela il faut que les erreurs successives calculées soient positives, $E_m > 0$ avec $1 \leq m \leq p$, une condition équivalente est que les $|k_m| < 1$, $1 \leq m \leq p$.

Les coefficients de corrélation partielle k_m représentent les coefficients de réflexion des ondes stationnaires dans un conduit (tube). Les aires A de deux sections consécutives du conduit vérifient le rapport : $$\frac{A(m-1)}{A(m)} = \frac{1+k_m}{1-k_m}.$$

Généralement A(p) =1 est l'aire à la glotte, A(p-1) correspond à la section la plus proche de la glotte et A(0) à la section aux lèvres[1], [10].

- Spectre du modèle

Le spectre du modèle AR est le carré du module de sa transmittance, $P_p(\omega) = \frac{\sigma_p^2}{|A_p(\omega)|}$ où

$$\sigma_p = E_{\min} = \sum_{i=0}^{p} \hat{a}(i).x(n-i).$$

Le spectre peut être obtenu par application d'une FFT, avec une longueur qui est une puissance de deux (sinon rajouter des zéros), sur la suite de coefficients â(i).

Des exemples de modélisation de spectre par la méthode LPC (Linear Prediction Coding) appliqué sur les mêmes voyelles [a,i] sont illustrés à la figure II - 7 (a) et (b). Le traitement est fait sur des trames pondérées par une fenêtre de Hamming de longueur 20 ms (200 points), sachant que la fréquence à la quelle ils ont été enregistrés est de 10 KHz.

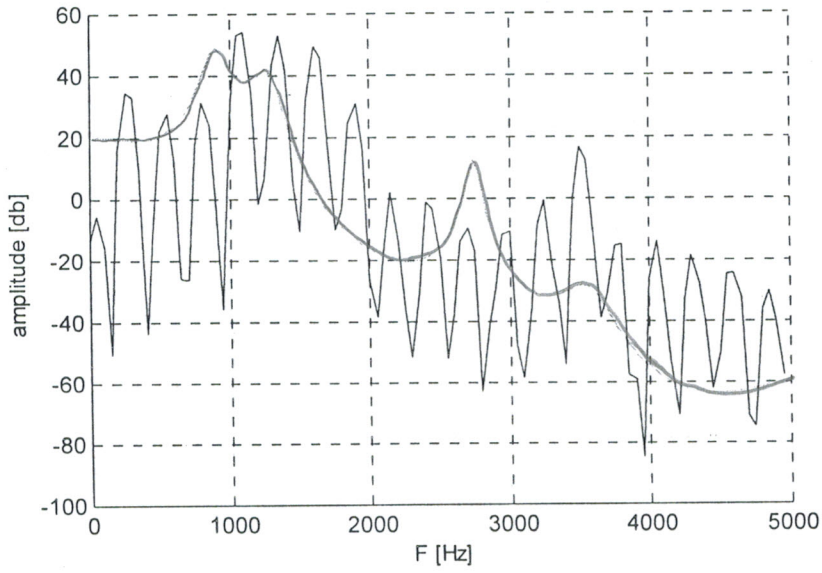


Fig. II – 7 (a) Spectre par FFT et PLC de la voyelle /a/

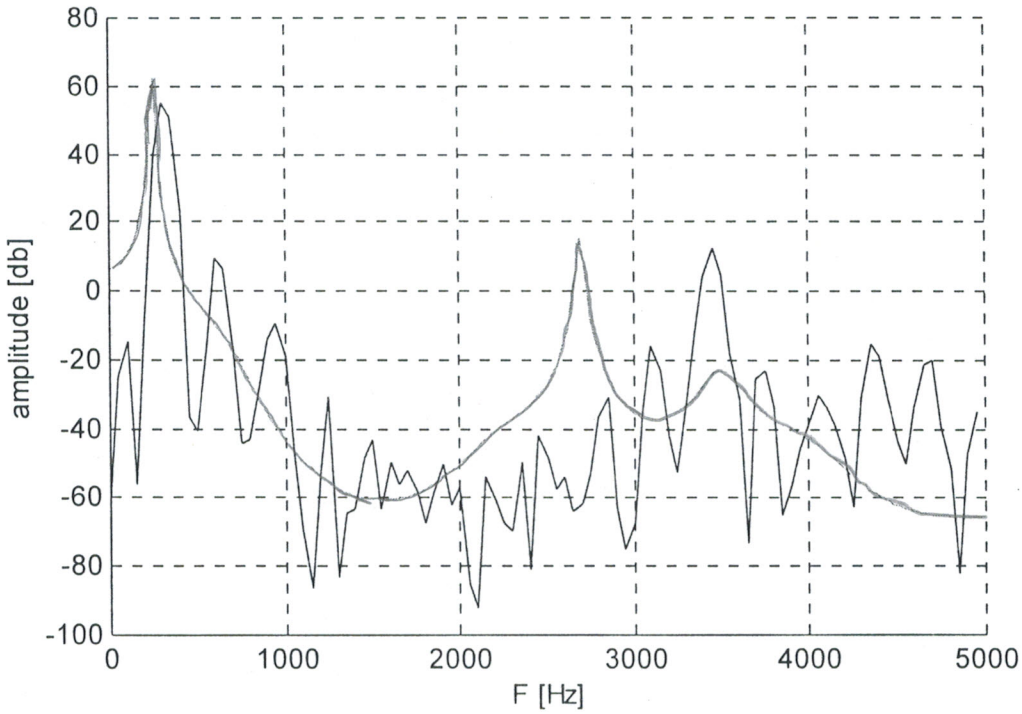


Fig. II – 7 (b) Spectre par FFT et PLC de la voyelle /ı/

II - 8 Ordre du modèle

L'ordre du modèle AR ne dépend pas de la méthode d'analyse (méthode d'autocorrélation ou de covariance) mais il est fonction de la fréquence d'échantillonnage. On estime d'une façon générale que la fonction de transfert du modèle H(z) comporte une paire de pôle par KHz de bande passante. Chaque paire de pôles caractérise un formant. L'excitation glottique d'une part et la radiation des lèvres d'autre part exigent de 3 à 4 pôles. Pour une fréquence d'échantillonnage de 10 KHz par exemple, l'ordre p serait de 13 ou 14 [14].

II – 9 Relation entre coefficients de prédiction et cepstres

Le cepstre réel peut être estimé à partir des coefficients de prédiction $\hat{a}_p(i)$ [2], [10].

On peut écrire : $Ln\left[1/A_p(z)\right] = \sum_{n=1}^{\infty} c(n) z^{-n}$.

Si l'on dérive chaque membre par rapport à z^{-1} :

$$-A'_p(z)/A_p(z) = \sum_{n=1}^{\infty} nc(n) z^{-n+1} \qquad avec \qquad A_p(z) = \sum_{i=0}^{p} a_p(i) z^{-i} ,$$

$$\text{donc : } -\sum_{i=1}^{p} i a_p(i) z^{-i+1} = \left[\sum_{j=0}^{p} a_p(j) z^{-j}\right]\left[\sum_{n=1}^{p} nc(n) z^{-n+1}\right].$$

Soit : $-ia_p(i) = \sum_{n=1}^{i-1} nc(n)a_p(i-n) + ic(i)$.

On obtient donc la récurrence :

$$c(i) = -a_p(i) - \sum_{n=1}^{i-1}(1-n/i)a_p(n)c(i-n) \qquad ;i > 0$$

avec : $c(0) = Ln\sigma^2$ où σ est le gain du modèle AR.

II – 10 préaccentuation

L'étape de prédiction linéaire est généralement précédée de celle de la préaccentuation. L'idée de base de cette phase est d'extraire la dérivée du signal enregistré en utilisant un filtre de préaccentuation dont la fonction de transfert est donnée par : $L(z) = 1 - \mu\ z^{-1}$ avec $0 \le \mu \le 1$.

Le rôle de la préaccentuation est principalement de rehausser (d'égaliser) les amplitudes faibles (les aigus) par rapport aux hautes amplitudes (les graves) afin de tenir compte de l'ensemble du signal [16].

II - 11 Modèle ARMA

La modélisation AR du mécanisme de la phonation présente des limitations et ne caractérise que d'une manière approchée la production de la parole, en particulier pour les sons nasalisés. Le modèle du conduit nasal est en réalité un filtre pôles-zéros (ARMA : autoregressif à moyenne ajustée ou Auto-Regressive Moving Average) et celui du rayonnement aux lèvres est du type tous-zéros (MA : moyenne ajustée ou encore FIR: Finite Impulse Response).

La transmittance devient alors celle d'un modèle ARMA :

$$X(z) = \frac{B(z)}{A(z)}$$

où A(z) est la partie AR et B(z) représente la partie MA.

Cela donne dans le domaine temporel la récurrence suivante :

$$x(n) + \sum_{i=1}^{p} a(i)x(n-i) = \sum_{i=0}^{q} b(i)u(n-i)$$

Chaque échantillon x(n) est la combinaison linéaire de p échantillons passés, et de q+1 échantillons présents et passés de l'excitation [2], [4].

II - 12 Limitation du modèle ARMA

Si le modèle ARMA est souvent retenu pour modéliser la parole suivant le principe déjà exposé, il n'est pas sans limitations.

Le modèle ARMA est plus délicat à estimer qu'un modèle AR. Cela amène parfois à préférer, pour une qualité donnée de la modélisation, un modèle AR avec un ordre relativement surestimé.

Mais la principale limitation réside dans l'hypothèse de stationnarité du signal acoustique qui est faite. Il faut réaliser un compromis entre la longueur de la fenêtre d'analyse et la durée pendant laquelle l'hypothèse de stationnarité est raisonnable. Ce compromis est réalisable pendant les zones stables (voyelles), mais il n'est pas satisfaisant durant les phases transitoires et injustifié sur les plosives [2].

II – 13 Conclusion

Les méthodes d'analyse de type temporel permettent d'extraire des informations du signal issu directement du microphone. Le nombre de passage par zéro du signal dans une fenêtre temporelle, le calcul de l'énergie à partir des échantillons de ce signal ou le calcul de la fréquence de variation des cordes vocales permettent de déterminer si un son est voisé ou non.

Les méthodes de type fréquentiel considère le signal comme étant composé d'une somme de sinusoïdes ou d'exponentielles. Ces méthodes permettent de mettre en évidence des propriétés du signal tel que les formants, qui sont difficiles à observer dans le domaine temporel mais qui sont essentiel pour la distinction des voyelles.

L'analyse cepstrale permet de séparer les paramètres d'excitation de ceux du conduit vocal. L'inconvénient de ce traitement est l'étape de séparation dans le cas des voix de femme (fréquence élevée et conduit vocal long) de plus elle nécessite deux transformées de Fourier et un calcul logarithmique, elle est donc relativement lourde.

L'analyse par PL est un moyen pour obtenir une enveloppe spectrale relative au spectre de puissance P (ω). Cependant elle présente des inconvénients.

En effet, le modèle tout pôle de la PL, approxime de manière uniforme le spectre de puissance P(ω) sur toutes les fréquences de la bande d'analyse.

Cette propriété est contradictoire avec l'audition humaine. Au-delà de 800 Hz, la résolution spectrale de l'audition diminue avec la fréquence. Par ailleurs, pour les niveaux d'amplitudes rencontrés uniquement dans un discours, l'audition est plus sensible aux fréquences

appartenant au milieu du spectre auditif. Par conséquent les détails spectraux ne sont pas toujours préservés par l'analyse de prédiction linéaire d'après leur proéminence auditive.

De nouvelles approches peuvent être exploitées afin de lever cette limitation. L'une d'elle consiste à introduire les propriétés du modèle auditif humain. C'est la PLP (Perceptual Linear Prediction) technique analysée et étudiée dans notre travail.

CHAPITRE III

ETUDE, IMPLEMENTATION ET ANALYSE DE LA TECHNIQUE PREDICTIVE LINEAIRE PERCEPTUELLE PLP

III – 1 Introduction

L'étude présentée au chapitre précédent a montré que le modèle tout pôle A(ω) obtenu par l'analyse de prédiction linéaire classique reproduit l'enveloppe du spectre vocal.
Les résultats présentés dans ce chapitre montrent que le spectre de puissance du signal vocal est approximé par le spectre de ce modèle. Cette approximation est faite d'une manière uniforme sur toute la bande d'analyse en ne tenant pas compte des propriétés de l'audition humaine pour la perception des sons, principalement celle de sélectivité de l'oreille humaine pour les différentes fréquences [17].

Il serait donc intéressant de modéliser ces propriétés d'audition afin de les prendre en considération lors de l'analyse du signal vocal. En fait, cette modélisation nous permettra d'extraire les paramètres significatifs de l'analyse.
L'approche PLP répond à ces spécifications, et c'est la technique que nous avons étudiée, implémentée et testée pour l'analyse de la parole.

III – 2 Analyse spectrale par PLC

L'application de l'analyse par prédiction linéaire sur des voyelles, telle qu'elle a été implémentée en chapitre II en utilisant l'approche d'autocorrélation, a permis d'obtenir des spectres (fig. III-1 et fig. III-2).

A titre d'exemple, on a pris la voyelle (phonème) /a/ antérieure et la voyelle /e/ postérieure, acquises à une fréquence d'échantillonnage de 10 KHz. L'analyse par PL a été faite sur des trames de 20 ms (durant laquelle le signal est supposé stationnaire) pondérées par la fenêtre de Hamming et préaccentuées par un filtre de préaccentuation de fonction de transfert (§ II – 10) :

$$L(z) = 1 - 0.9z^{-1}.$$

En figure III-1 (a) et (b) sont illustrés respectivement, les spectres des voyelles /a/ et /e/ prononcées par un locuteur féminin adulte. Ces spectres sont obtenus en appliquant une FFT de 256 points sur la tranche du signal temporel et une analyse par prédiction linéaire d'ordre 14.

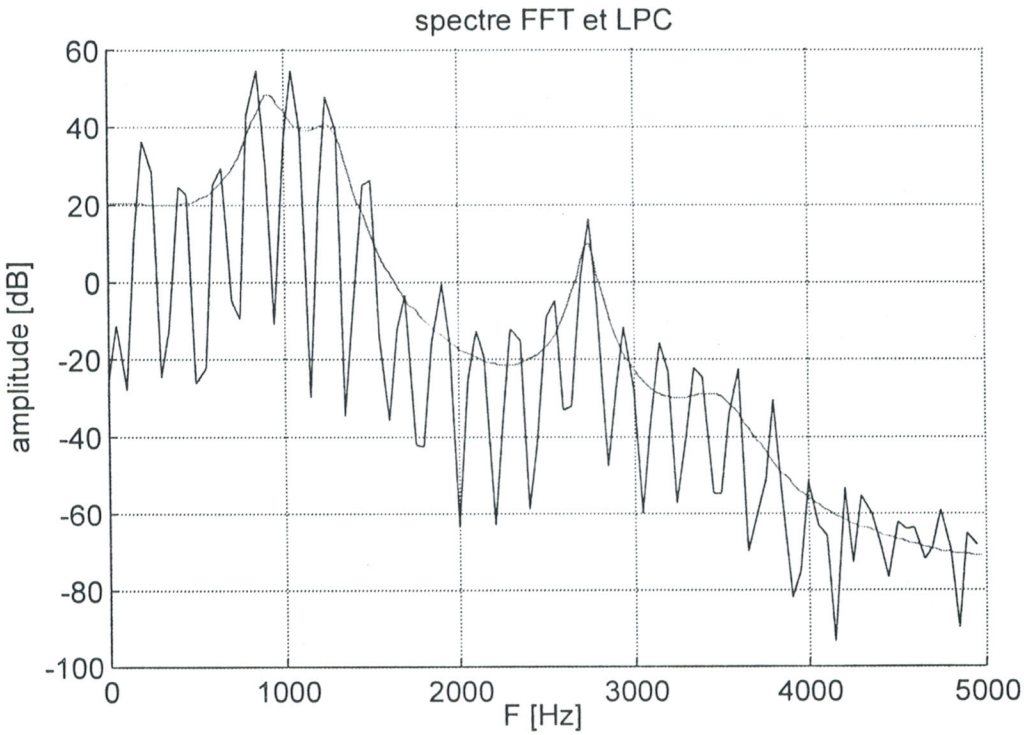


Fig. III - 1 (a) Spectres de la voyelle /a/
(FFT et PLC d'ordre 14).

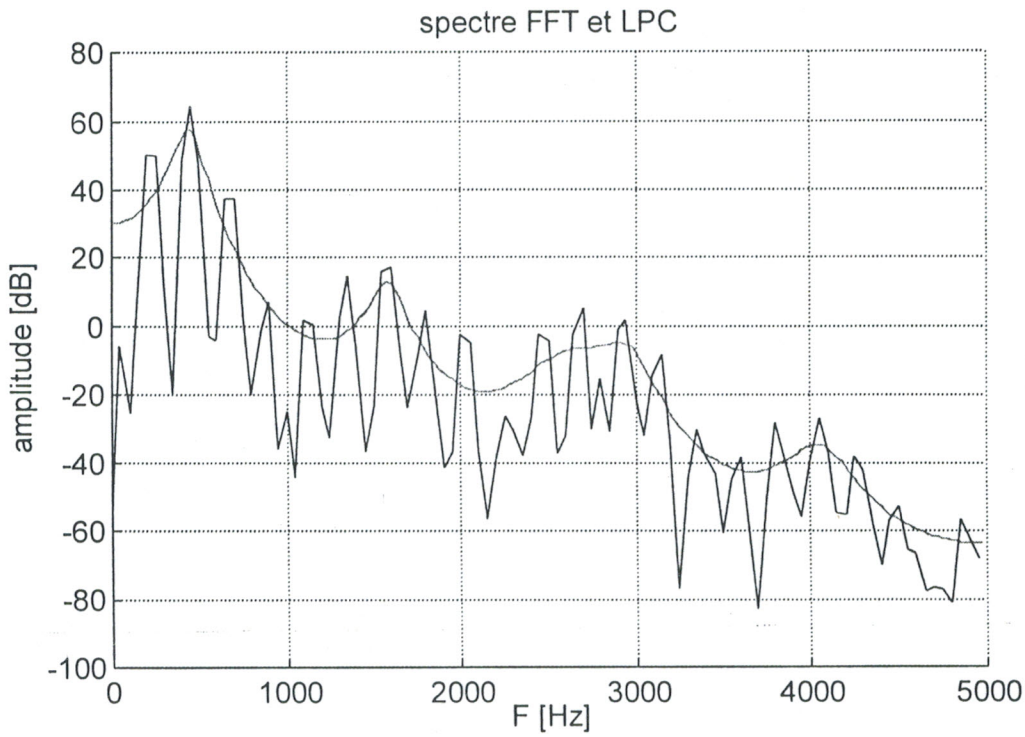


Fig. III - 1 (b) Spectres de la voyelle /e/
(FFT et PLC d'ordre 14).

On remarque d'après ces figures que le lissage du spectre vocal est meilleur pour les sommets de l'enveloppe que pour les creux. Ces sommets correspondent souvent aux fréquences de résonances (formants).

Dans ce cas, un modèle d'ordre 14, a permis de faire apparaître 4 sommets sur l'enveloppe du spectre qui s'étend de 0 à 5 KHz.

Pour la voyelle /a/ (fig. III-1(a)), le premier pic est au voisinage de 1000 Hz, le $2^{ème}$ se trouve entre 1000 et 1500 Hz, le $3^{ème}$ est entre 2500 et 3000 Hz, par contre le $4^{ème}$ et dernier pic est au-delà de 3000 Hz.

De même pour la voyelle /e/ (fig. III-1(b)), le premier pic est bien au-dessous de 1000 Hz ($\cong$500 Hz), le $2^{ème}$ est localisé entre 1000 et 2000 Hz, le $3^{ème}$ pic est aux environs de 3000 Hz et le dernier sommet dans la zone de 4000 Hz.

D'après ces valeurs on remarque que les relations suivantes sont vérifiées :

$$F_1(/a/) > F_2(/e/)$$

$$\text{et } F_2(/a/) < F_2(/e/).$$

Ces positions relatives des formants que nous avons obtenues pour les voyelles /a/ et /e/ sont en accord avec les résultats rapportés par Itahashi [18] et Bladon [19] et données en tableau I-1 (§ I- 4 -4).

Nous rappelons que ce tableau rassemble des valeurs moyennes des formants. Les valeurs exactes dépendant en fait du locuteur.

III – 3 Etude de la technique PLP

La technique PLP (Perceptual Linear Prediction) introduite à l'origine par Hermansky [21], fait appel à une modélisation tout pôle et à des échelles de fréquence non linéaires [18], [20] pour tenir compte des propriétés de l'oreille.

Ainsi cette modélisation est appliquée sur un spectre auditif obtenu par :

a) Une convolution de $P(\omega)$ (spectre de puissance du signal vocal) avec une courbe de masquage en bande critique,
b) échantillonnage du spectre obtenu à des intervalles de 1 Bark,
c) préaccentuation par une courbe isosonique,
d) compression du spectre obtenu pour simuler la relation non linéaire qui existe entre l'intensité d'un son et son intensité perçue.

Nous décrivons ci-dessous les différentes étapes de l'analyse PLP que nous avons implémentée et étudiée.

III – 3 – 1 Spectre de puissance du signal vocal

Soit un signal vocal x(n) représentant la voyelle /a/ (fig. III - 2 (a)).

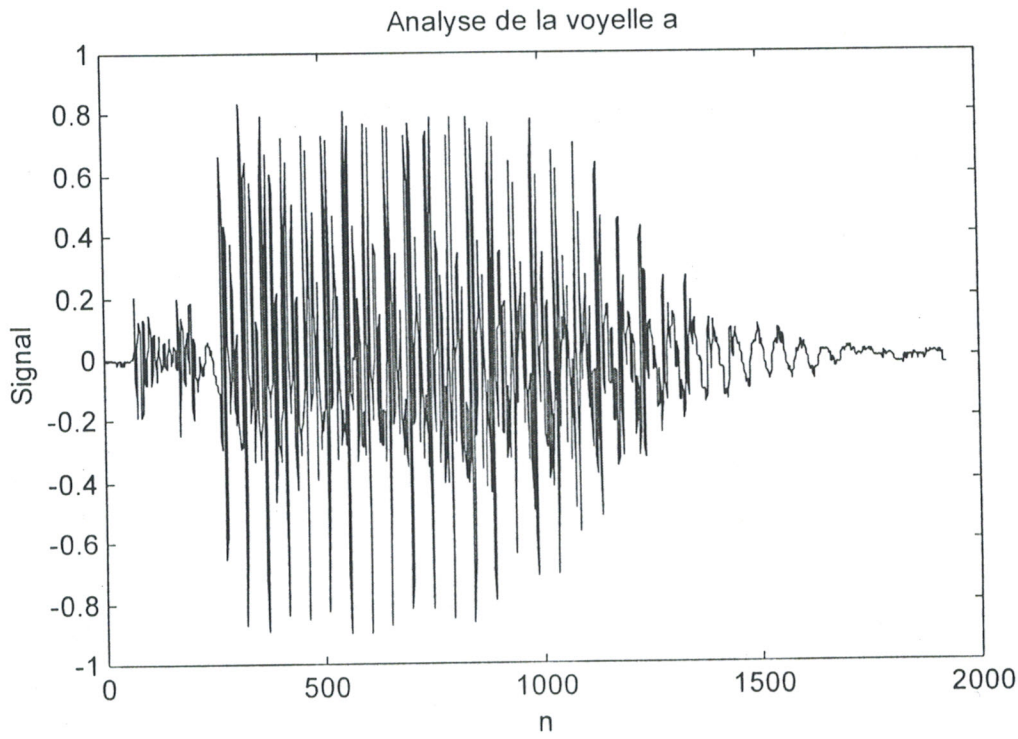


Fig. III - 2 (a) Représentation temporelle d'une voyelle /a/

Un segment de ce signal est multiplié par la fenêtre de Hamming, s(n) = x(n). w(n) (fig. III - 2 (b)) avec :

$$w(n) = 0.54 - 0.46\cos[2\Pi n/(N-1)] \qquad (1)$$

où N est la longueur de la fenêtre. Généralement, la longueur choisie correspond à une durée de 20 ms. Durée pour laquelle le signal est supposé stationnaire.

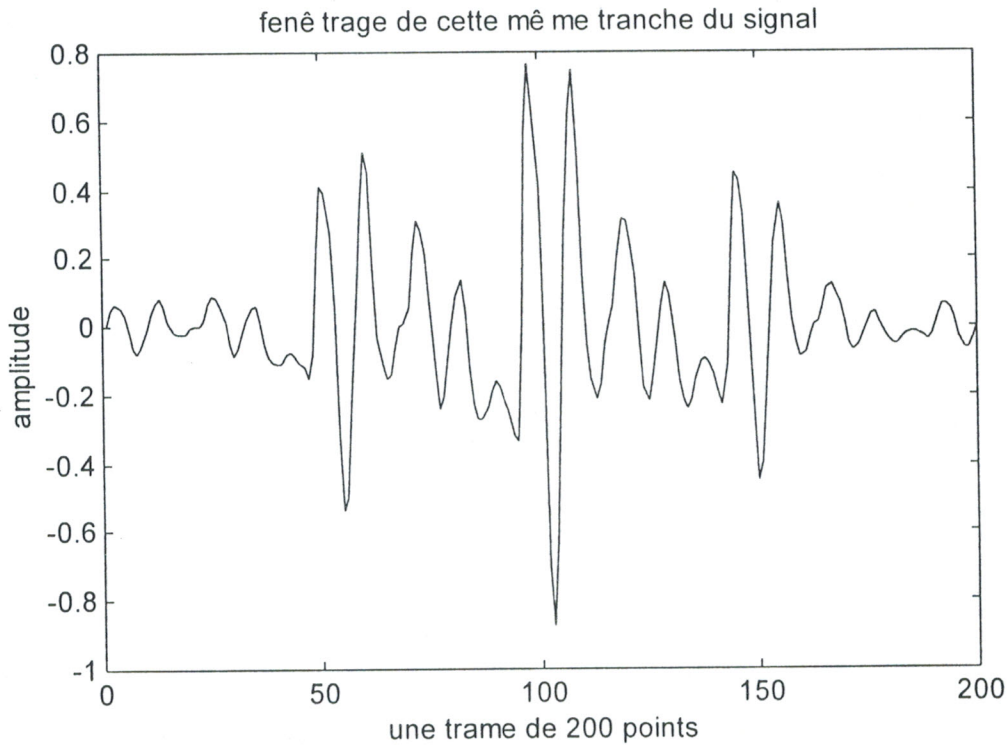


Fig. III - 2 (b) Après multiplication par la fenêtre de Hamming

Pour passer au domaine fréquentiel, une transformée de Fourier discrète est utilisée (TFD). Compte tenu que la fréquence d'échantillonnage f_e=10 KHz et que la durée pour laquelle le signal est supposé stationnaire est de 20 ms, une FFT de 256 points est alors utilisée (il faut noter que les 56 points restants sont fixés à zéro).

Les composantes réelle et imaginaire du spectre de parole court terme S(ω) sont élevées au carré et additionnées pour obtenir le spectre de puissance court terme :

$$P(\omega) = \Re[S(\omega)]^2 + \Im[S(\omega)]^2 \qquad (2)$$

Le spectre de puissance court terme d'une trame de la voyelle a est représenté en figure III – 3 ci-après.

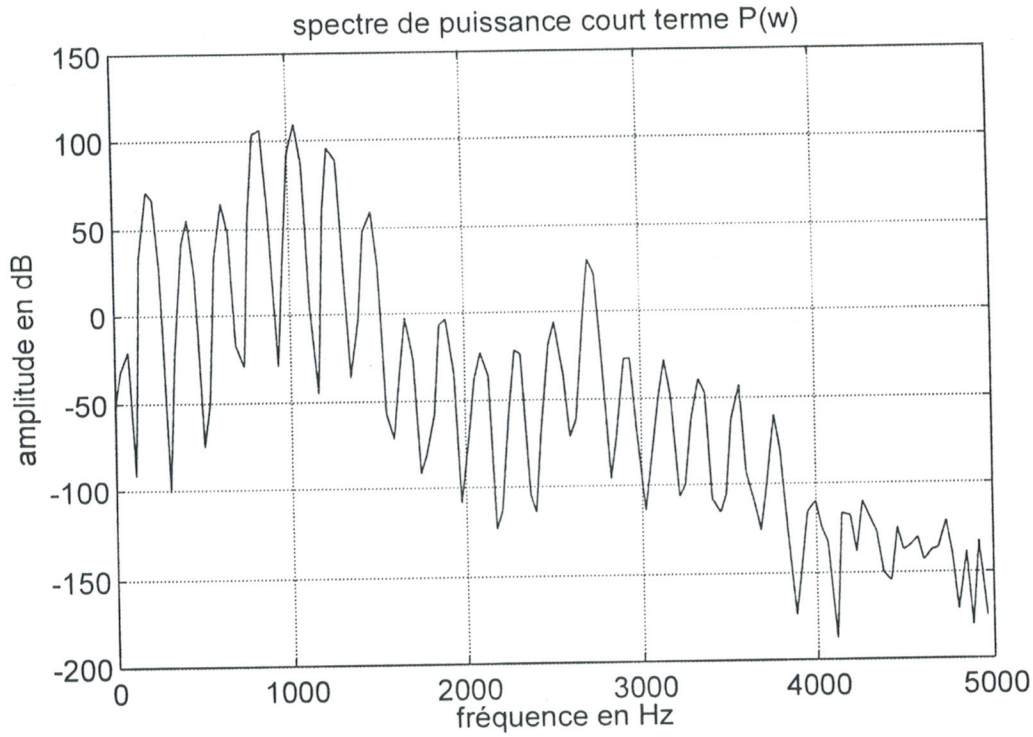
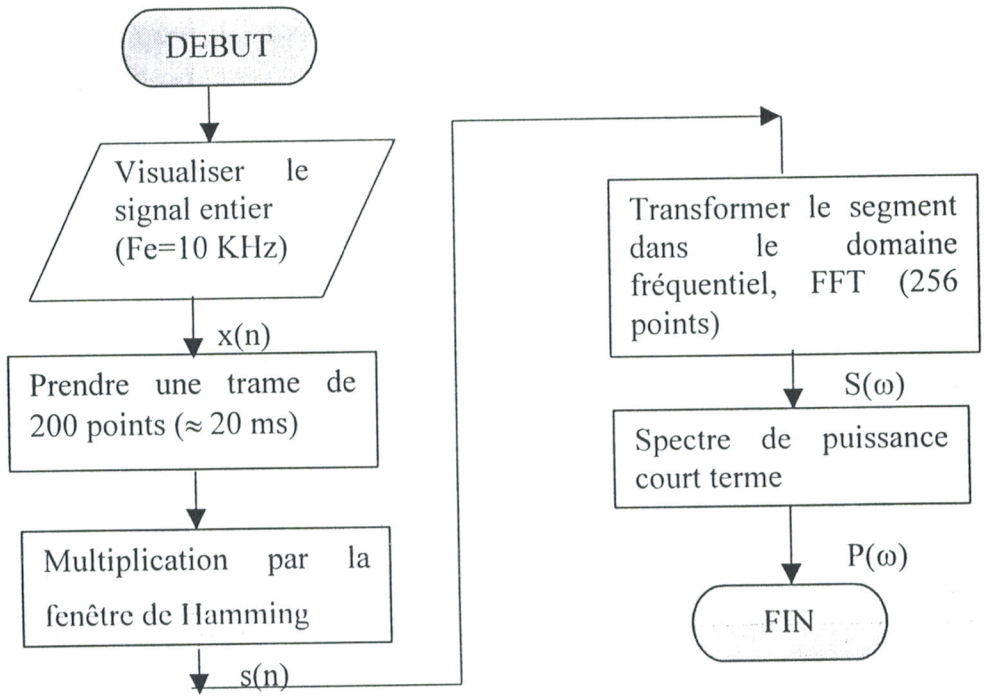


Fig. III - 3 Spectre de puissance court terme

Ce spectre (fig. III – 3) montre que la puissance de la voyelle /a/ est concentré autour de 1000 Hz, ce qui correspond au premier formant sur le spectre LPC (fig. III – 1(a)).

Ainsi les étapes d'obtention de P(ω) sont résumées sur l'organigramme III-1 ci-dessous :



Organigramme III-1

Sachant que l'oreille humaine a la faculté d'intégrer des fréquences en bandes critiques (§ I-7-5), il s'agit maintenant de fixer la résolution spectrale.

III – 3 – 2 Résolution spectrale en bande critique

L'étape suivante du traitement est une transformation non linéaire de l'échelle des fréquences, due à la sensibilité de l'ouïe humaine. L'homme entend différemment des sons selon leurs composantes fréquentielles (§ I-7-5).

Cette transformation est donnée par :

$$\Omega(\omega) = 6Ln\left\{\omega/1200\Pi + \left[(\omega/1200\Pi)^2 + 1\right]^{0.5}\right\} \qquad (3)$$

où ω est la fréquence angulaire en rad/s et Ω est exprimée en Barks.

Le résultat obtenu illustré en figure III - 4 montre bien la relation non linéaire entre l'échelle des fréquences Hertz et celle des Barks :

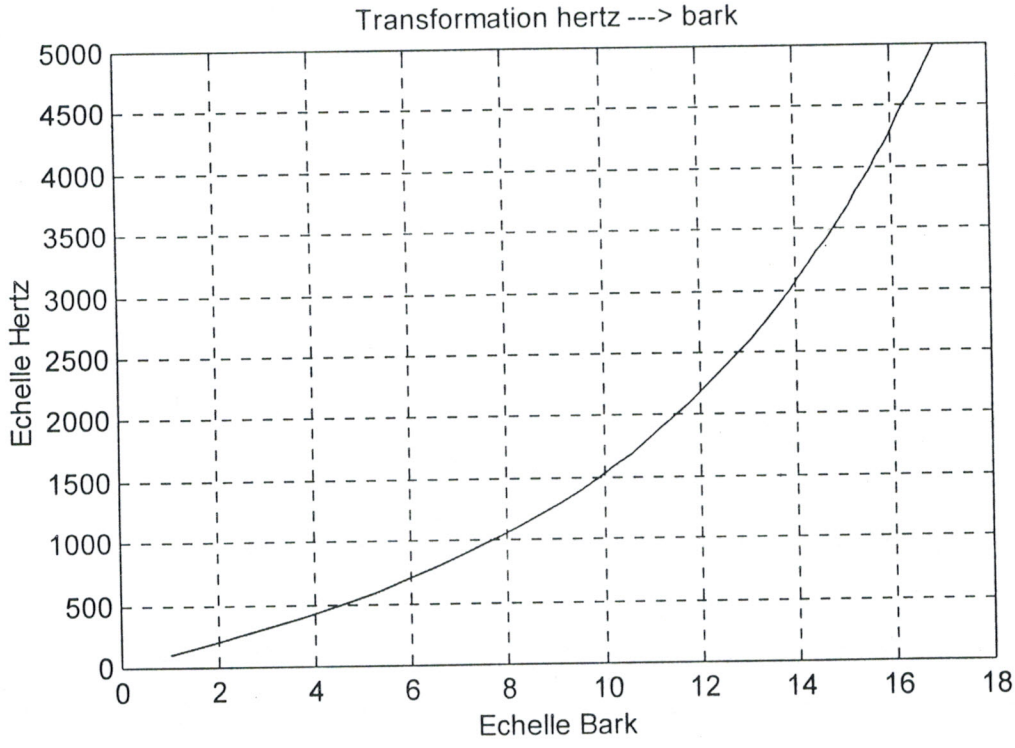


Fig. III - 4 Relation non linéaire entre Hertz et Barks

Cette convolution réduit la résolution spectrale de $\theta(\Omega)$ en comparaison avec $P(\omega)$. Ceci nous permet de sous-échantillonner $\theta(\Omega)$. Ici, $\theta(\Omega)$ est échantillonné à des intervalles de 1 Bark environ. La valeur exacte de l'intervalle d'échantillonnage est choisie de telle manière à ce qu'un nombre réduit d'échantillons du spectre couvrent la totalité de la bande d'analyse.

Sachant que la bande d'analyse s'étend de 0 à 5 KHz ce qui correspond à une bande de 0 à 16.9 Bark, il faut 18 échantillons du spectre $\theta[\Omega(\omega)]$ pour couvrir la bande d'analyse par pas de 0.994 Barks.

III – 3 – 3 Préaccentuation isosonique

Après le passage à l'échelle Bark, le spectre $\theta[\Omega(\omega)]$ échantillonné est préaccentué par une courbe isosonique simulée par :

$$\Xi[\Omega(\omega)] = E(\omega)\,\theta[\Omega(\omega)] \qquad (6)$$

La fonction $E(\omega)$ est une approximation de la sensibilité de l'ouïe humaine qui diffère en fonction de la fréquence. Elle simule la sensibilité de l'ouïe à un niveau de 40 dB à peu prés. Cette approximation est donnée par [24]:

$$E(\omega) = [(\omega^2 + 56.8.10^6)\omega^4]/[(\omega^2 + 6.3.10^6)^2(\omega^2 + 0.38.10^9)] \qquad (7)$$

Le tracé de Bode, réel et non asymptotique de cette courbe isosonique est illustré sur la figure III – 6 ci-dessous :

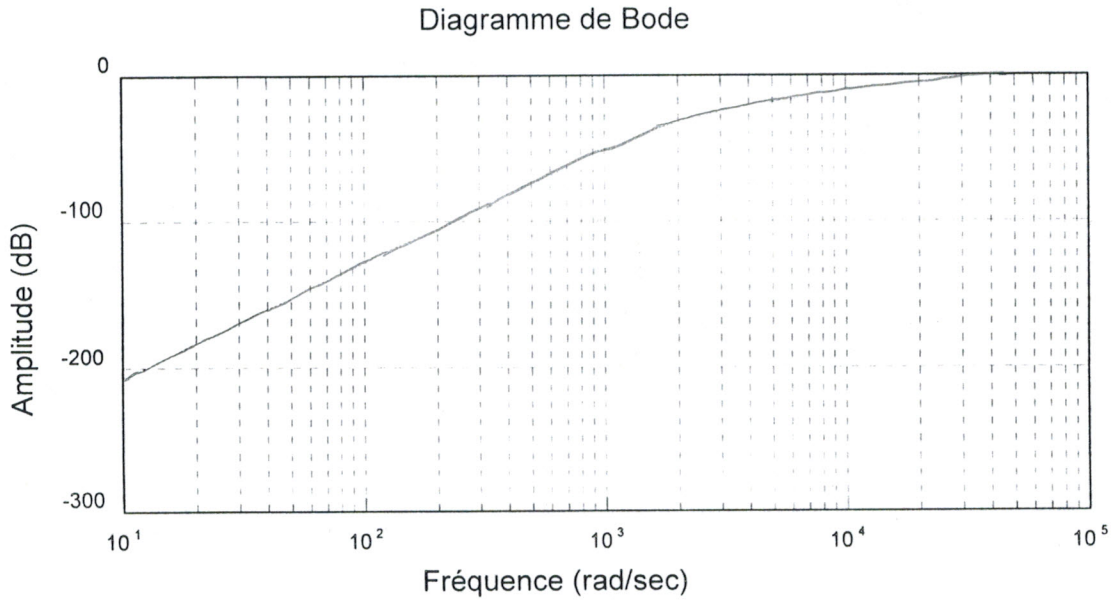


Fig. III - 6 Diagramme de Bode de la courbe isosonique

$E(\omega)$ est une fonction de transfert d'un filtre avec des asymptotes de 12 dB/oct entre 0 et 400Hz, 0 dB/oct entre 400 et 1200 Hz, 6 dB/oct entre 1200 et 3100 Hz et 0 dB/oct entre 3100 et la fréquence de Nyquist (la fréquence de Nyquist représente la fréquence maximale du spectre qui est ici de 5000 Hz).
Pour des niveaux de sons modérés, cette formule (éq.7) est applicable jusqu'à 5000 Hz. Au-delà de cette fréquence, un facteur supplémentaire est introduit dans l'expression de $E(\omega)$ pour représenter la décroissance de la sensibilité de l'audition pour des fréquences supérieures à 5000 Hz [21].

Les valeurs du premier échantillon de $\Xi[\Omega(\omega)]$ ainsi que du dernier sont prises égales à celles de leurs plus proches voisins.

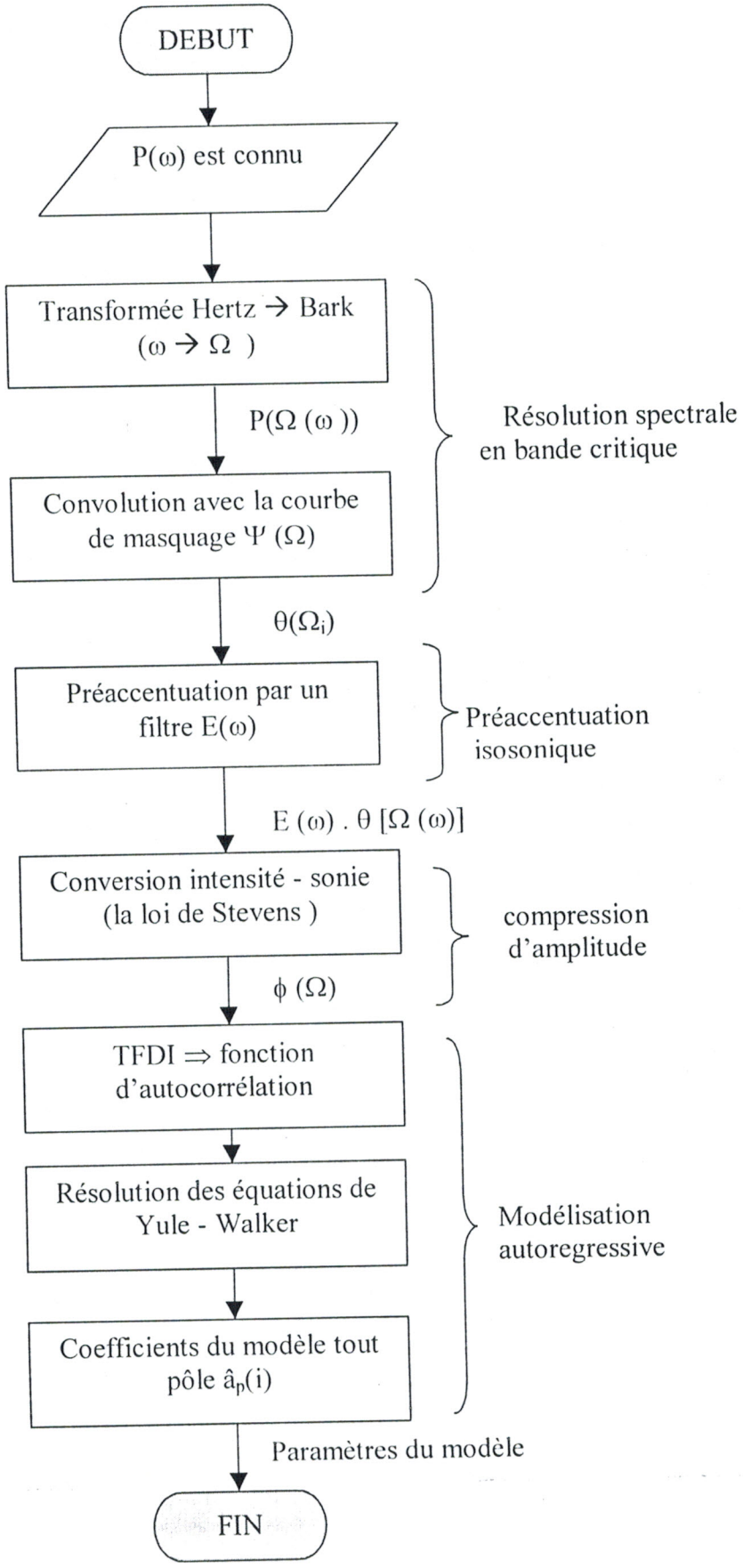
III – 3 – 4 Loi de compression

La dernière étape avant la modélisation autoregressive (tout pôle) est la compression par racine cubique de l'amplitude :

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \qquad (8)$$

Cette opération est une approximation de la loi de puissance de Stevens de l'ouïe[17]. Elle simule la relation non linéaire entre l'intensité du son et sa force perçue (voir § Sensations auditives en chapitre I).

L'étude faite sur l'effet de la compression et de la préaccentuation montre que cette opération réduit encore plus la variation d'amplitude spectrale du spectre en bandes critiques. Cela est clairement illustré en fig. III–7 ci-après, où le spectre obtenu pour la voyelle /a/, en appliquant l'analyse PLP sans l'étape de compression a une amplitude plus importante que celui obtenu par PLP avec l'étape de compression.



Organigramme III-2

III – 4 Implémentation

Lors de l'implémentation de la technique PLP les étapes de l'organigramme III-2 de convolution et de préaccentuation sont combinées.

Ainsi les opérations de convolution et de préaccentuation sont effectuées pour chaque échantillon de $\Xi[\Omega]$ dans le domaine $P(\omega)$.

Un échantillon $\Xi[\Omega(\omega_i)]$ est donné par :

$$\Xi[\Omega(\omega_i)] = \sum_{\omega=\omega_{il}}^{\omega_{ih}} W_i(\omega) \, . \, P(\omega) \qquad (9)$$

Les limites ω_{il} et ω_{ih} (l :low, h :high) de la sommation, et les fonctions de pondération W_i sont exprimées à partir des équations : de la courbe de masquage (éq. (4)), la courbe isosonique (éq. (6)) et l'inverse de la transformation Hertz-Bark (éq. (3)) qui est donnée par [21] :

$$\omega = 1200\Pi \, . \sinh(\Omega / 6) \qquad (10)$$

Les fonctions de pondérations $W_i(\omega)$ sont calculées pour une fréquence d'échantillonnage et un certain nombre de points de la FFT, donnés.

Ces fonctions de pondération sont obtenues dans le cadre de cette étude pour une fréquence d'échantillonnage de 10 KHz et des trames d'analyse de 20 ms (FFT de 256 points) leur forme est illustrée en figure III – 8 ci-dessous.

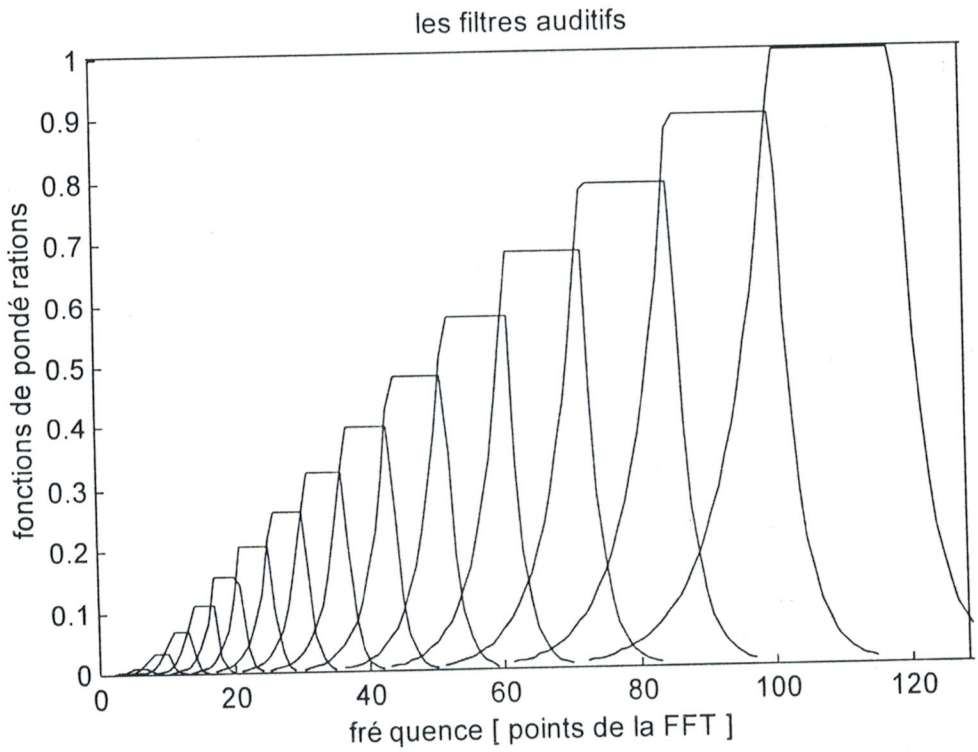


Fig. III – 8 Les filtres auditifs

On remarque sur cette figure que la largeur de $W_i(\omega)$ (c'est à dire l'intervalle de résolution) augmente avec la fréquence et ceci selon l'équation (3). Les $W_i(\omega)$ se chevauchent, ils ont un sommet plat et une pente en basse fréquence moins raide que celle en haute fréquence. C'est ce qui résulte de l'équation (4) après inversion de la fréquence par convolution et transformation du domaine 'Ω'(Bark) à 'ω' (Hertz).

Sur une bande d'analyse de 0 à 5000 Hz, qui correspond à 0 - 16.9 Bark, on obtient 18 filtres auditifs (fonctions de pondérations) pour un intervalle de résolution en bande critique de 0.9942 Bark (§ III-3-2). Ces filtres ont une fréquence basse (f_l), une fréquence centrale (f_0) et une fréquence haute (f_h) dont les valeurs sont indiquées ci-dessous. La largeur de bande B_i de ces filtres est aussi mentionnée.

f_l (Hz)	f_0 (Hz)	f_h (Hz)	B_i (Hz)
1.0e+003 *			
-0.1522	0.0999	0.2351	0.0829
-0.0512	0.2025	0.3456	0.3968
0.0483	0.3107	0.4656	0.4173
0.1492	0.4275	0.5984	0.4492
0.2542	0.5560	0.7477	0.4935
0.3661	0.6998	0.9176	0.5515
0.4882	0.8629	1.1127	0.6245
0.6237	1.0497	1.3384	0.7147
0.7763	1.2654	1.6010	0.8247
0.9503	1.5159	1.9077	0.9574
1.1505	1.8082	2.2668	1.1163
1.3823	2.1502	2.6884	1.3061
1.6522	2.5514	3.1839	1.5317
1.9675	3.0228	3.7670	1.7995
2.3370	3.5774	4.4538	2.1168
2.7708	4.2305	5.2632	2.4924

Le premier et dernier filtre (filtre 18) ont les mêmes valeurs que celles de leurs voisins. La forme de ces filtres auditifs est due à l'expression de la courbe de masquage (éq.(4)). De plus leur amplitude croît avec la fréquence selon l'équation de la courbe isosonique (éq.(7)).

L'opération la plus coûteuse (en temps de calcul) dans l'implémentation de la technique PLP est le calcul spectral de la FFT, suivi par la résolution spectrale en bande critique et la compression par racine cubique. Le coût de la modélisation autoregressive est négligeable vu le nombre d'échantillons spectraux (18 échantillons) du spectre auditif à approximer.

Les exigences d'implémentation pour la PLP sont comparables avec ceux de la PL conventionnelle, 3000 et 3400 multiplications respectivement. Voici en tableau III – 1, le nombre de multiplications nécessaires par trame de 200 échantillons [21].

PLC d'ordre 14	**Nombre de multiplications**	**PLP d'ordre 5**	**Nombre de multiplications**
Préaccentuation	200	Fenêtre	200
Fenêtre	200	FFT	2100
Autocorrélation	2800	Bande critique	450
Modèle autoregressif	200	Racine cubique	150
		IDFT	30
		Modèle autoregressif	30
Total	3400	Total	3000

Tableau III – 1 Coût de la PLP et PLC

Ainsi la méthode d'analyse par PLP est mise au point pour être appliquée sur une trame du signal vocal et fournit par la suite le vecteur de paramètres la caractérisant. Pour faire le traitement du signal en entier il suffit de prendre plusieurs trames successives, on aura ainsi un ensemble de vecteurs de paramètres qui définissent le son étudié. Le nombre de ces vecteurs correspond au nombre de trames sur lequel l'analyse PLP a été appliquée.

III – 5 Choix de l'ordre du modèle AR

Le choix de l'ordre du modèle spécifie les détails dans le spectre auditif qui doivent être préservés dans le spectre du modèle PLP. Et en traitement de la parole, on est intéressé de représenter l'information linguistique dans le signal.

Pour déterminer l'ordre optimal du modèle, des expériences ont été faites pour reconnaître la parole d'un locuteur en utilisant celle d'un locuteur différent [21]. Ainsi toute information extra-linguistique; facteurs spectraux dépendant du locuteur; ne sera pas utilisée lors de l'identification. Ces facteurs, en principe, réduisent la précision de l'identification.

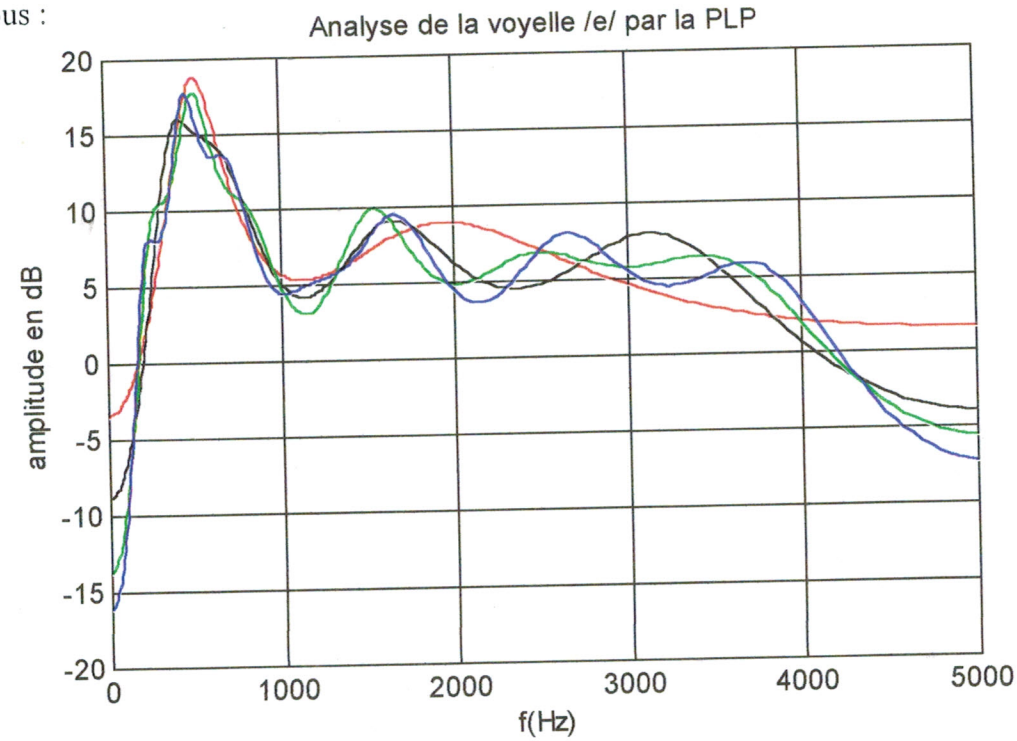
Des expériences faites par Hermansky [21] ont montré que la précision de la PLP croît jusqu'à approximativement le $5^{ème}$ ordre du modèle autoregressif, puis décroît avec une augmentation supplémentaire dans l'ordre du modèle : cas de l'identification indépendante du locuteur de phonèmes et des mots alphanumériques. Tandis que pour l'identification dépendante du locuteur, cet ordre ($5^{ème}$) reste le meilleur, la précision reste pratiquement la même en augmentant d'avantage l'ordre [21].

III – 6 Résultats

Après présentation et implémentation de la technique PLP sous l'environnement Matlab, un ensemble de tests de cette technique ont été effectués en vue d'étudier ces performances. Les tests préliminaires consistent en l'identification de phonèmes.

III – 6 – 1 Influence de l'ordre du modèle AR

Nous avons appliqué l'analyse PLP sur une tranche de la voyelle /e/, en choisissant différents ordres pour le modèle AR. Les spectres obtenus sont illustrés sur la figure III – 9 ci-dessous :



Légende :

PLP d'ordre 5

PLP d'ordre 8

PLP d'ordre 12

PLP d'ordre 14

Fig. III - 9 La PLP à différents ordre

On observe sur ces courbes que l'allure globale du spectre PLP est préservée même avec un ordre faible qui est de 5.

Or on sait que l'information essentielle pour l'identification de voyelles réside dans la forme globale du spectre [6].

En outre comme nous l'avons signalé au paragraphe précédent (§ III – 5) l'identification indépendante ou dépendante du locuteur donne les meilleurs résultats lorsque l'ordre du modèle AR dans l'analyse PLP est égale à 5 [14].

Par conséquent, nous avons utilisé cet ordre 5 dans tous les tests qui suivent.

Par ailleurs, ce faible ordre nous permet de réduire le nombre de paramètres $\hat{a}_p(i)$ à calculer.

III – 6 – 2 Identification d'un phonème sur une seule trame

Nous avons appliqué l'analyse PLP d'ordre 5 à quelques voyelles orales françaises à savoir : /a/, /e/, /i/, /o/, /u/ et /y/. L'analyse par PLP d'une seule trame de 20 ms de chaque phonème, prise du signal entier, permet d'obtenir les paramètres du modèle à partir desquels une représentation fréquentielle est possible. Il apparaît sur une telle représentation des pics pour les quels l'énergie est relativement élevée.

Ces tests ont été effectués sur plusieurs répétitions d'une même voyelle, acquises à une fréquence d'échantillonnage de 10 KHz. Les valeurs moyennes des fréquences où les pics spectraux sont localisés sont notées F_1 et F_2. Leurs valeurs, renvoyés par notre programme sont données dans le tableau III - 1 qui suit :

	F_1 (Hz)	F_2 (Hz)	F_1 (Bark)	F_2 (Bark)
/a/	1378.8		7.57	
/o/	558.2		4.97	
/i/	401.7	3043.9	4.52	13.95
/e/	499.5	2001.5	4.54	11.50
/u/	511.7		4.63	
/y/	419.5	2483.8	3.90	12.76

Tableau III – 1 Formants de voyelles par PLP

Si on reprend ces valeurs sur une figure où la première fréquence (F_1 (Bark)) est sur l'axe horizontal et la 2ème (F_2 (Bark)) sur l'axe vertical, on obtient la fig. III - 10 (a) dans le plan F_1F_2 en Bark.

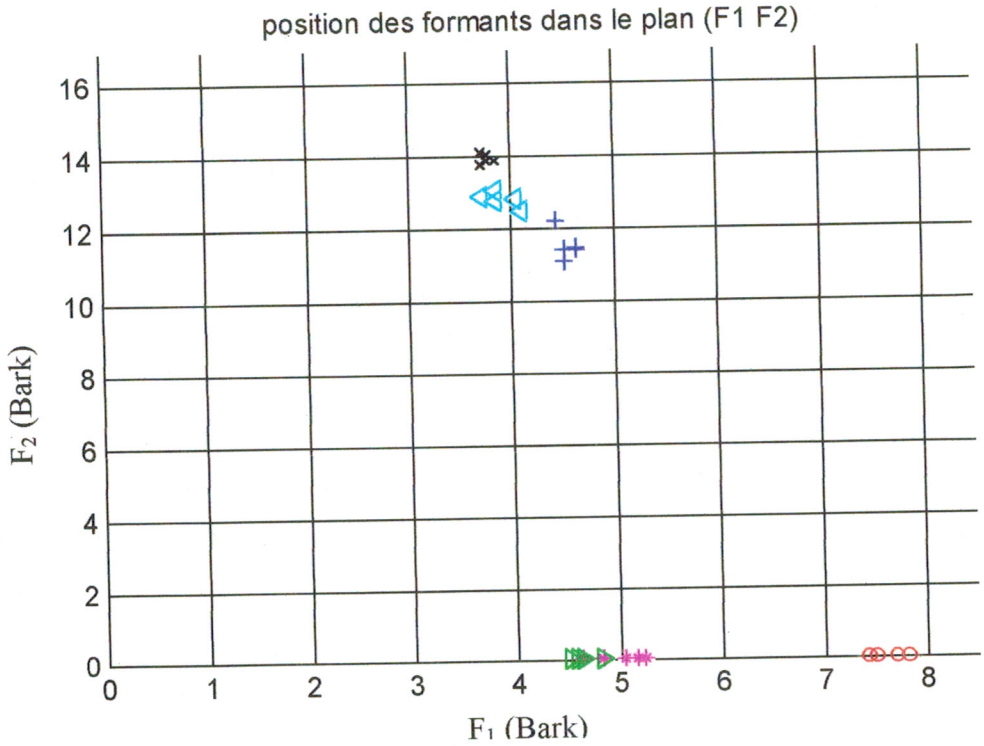


Fig. III – 10 (a) Position des voyelles dans le plan F_1F_2 (Bark)

Légende : o la voyelle /a/ + la voyelle /e/ x la voyelle /i/
* la voyelle /o/ < la voyelle /y/ > la voyelle /u/

Cette figure nous permet de faire la distinction entre voyelles à partir des deux pics que nous renvoie l'analyse PLP d'ordre 5. Sauf peut être pour la voyelle /o/ et /u/ (le son 'ou'), proches phonétiquement, où les zones se chevauchent. Chaque voyelle occupe une zone (un nuage) sur la grille propre à elle.

On voit bien aussi que certaines voyelles ont un seul pic spectral et d'autres on ont deux. La plus grande valeur obtenue pour la fréquence du premier pic spectral ne dépasse pas 8 Barks, par contre le deuxième pic s'il existe se trouve plus loin, entre 10 et 14 Barks.

La même représentation est utilisée, cette fois-ci en utilisant une échelle Hertz des fréquences (fig. III - 10 (b)).

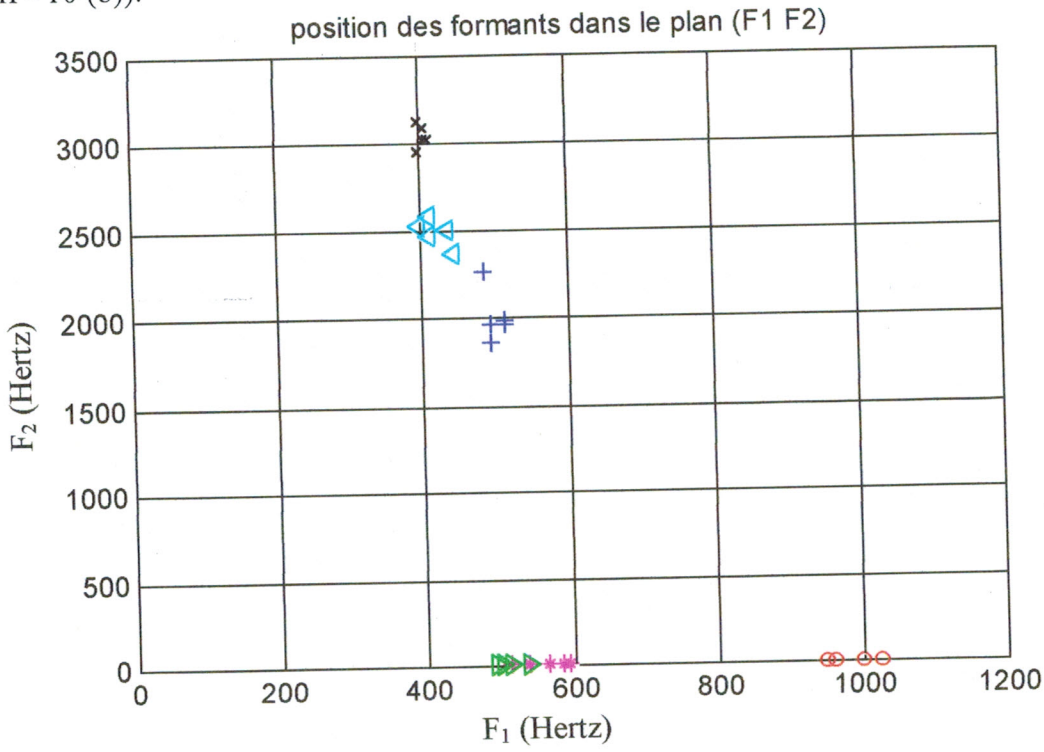


Fig. III – 10 (b) Position des voyelles dans le plan F_1F_2 (Hertz)

Légende : o la voyelle /a/ + la voyelle /e/ x la voyelle /i/
* la voyelle /o/ < la voyelle /y/ > la voyelle /u/

La voyelle /a/ (postérieure) a un seul pic aux environs de 7 Bark, 1300 Hz à peu prés, de même que le phonème /o/ pour lequel le pic est presque à 5 Bark (≅550 Hz) et pour le phonème /u/ il est supérieur à 4.5 Bark (≅500 Hz).

Les premiers formants F1 de ces voyelles vérifient la relation d'ordre suivante :

$$F_1(/a/) > F_1(/o/) > F_1(/u/)$$

qui correspond à l'ordre indiqué sur le triangle acoustique des voyelles orales du français (fig. I-6 § I – 4 – 4).

Par contre le reste des voyelles (/e/, /i/, /u/) pour lesquelles nous avons appliqué l'analyse PLP, ont deux pics spectraux. Par exemple pour la voyelle /e/ le premier pic est supérieur à 4.5 Bark (≅500 Hz) et le 2ème au-dessous de 12 Bark (<2000 Hz). Le 1er pic de la

voyelle /i/ est inférieur à 4.5 Bark (≅400 Hz), le 2ème se trouve à 14 Bark (≅ 3000 Hz). Pour la dernière voyelle étudiée /y/, on a le premier pic qui vérifie :

$$F_1(/i/) < F_1(/y/) < F_1(/e/)$$

et le 2ème pic aux environs de 13 Bark (2500 Hz).
De plus ces trois dernières voyelles sont antérieures (d'après leur position sur le triangle acoustique § I - 4 - 4).

Les valeurs relatives des formants que nous avons obtenues pour les différentes voyelles sont en accord avec celle données en tableau I – 1(§ I –4 -4) et raportées par Itahashi [18] et Bladon [19].

III – 7 PLP et perception de voyelle

La PLP du 5ème ordre est compatible avec quelques concepts de perception humaine des voyelles. Ces concepts sont, le second formant effectif F_2' et la théorie d'intégration du pic spectral à 3,5 Bark qui expliquent les résultats obtenus précédemment et donnés en § III-6-1.

III – 7 – 1 Le second formant effectif

De nombreuses expériences ont permis de montrer que l'on peut "approximer perceptivement" les voyelles naturelles par des stimuli synthétiques n'ayant que 1 ou 2 formants, caractérisant les voyelles postérieures et antérieures respectivement.

L'information essentielle pour l'identification des voyelles réside dans la forme globale du spectre. L'identification avec des stimuli synthétiques à 1 formant semble possible uniquement pour des voyelles naturelles présentant 2 premiers formants séparés de moins de 3.5 à 4 Barks en fréquence : le formant unique du stimulus doit correspondre alors au centre de gravité du spectre naturel restreint à la zone des deux premiers formants.
Pour des voyelles ayant leurs premiers formants plus éloignés, l'identification ne peut se faire qu'avec des stimuli synthétiques à 2 formants, F_1 et F_2' : F_1 ayant la même fréquence que le formant « naturel », F_2' dépendant de la répartition des fréquences et des amplitudes des

formants auditifs supérieurs. Ce pic $F_2^{'}$ est appelé *«second formant effectif»* (*Effective Second Formant*) [25][6].

Les résultats que nous avons obtenus par l'analyse PLP des voyelles sont en accord avec la réduction spectrale que peut faire l'ouïe humaine sur le spectre des voyelles.
En effet, nous avons observé que les spectres PLP des voyelles antérieures (Front) [/e/, /i/, /y/] présentent deux pics (fig. III–13 (c)(d)), alors que les spectres PLP des voyelles postérieures (Back) [/a/, /o/, /u/] en présentent un seul (fig. III–13 (a)(b)) (tableau III – 1).

Sur la figure III – 11 où sont représentés les spectres FFT, LPC (d'ordre 14) et PLP (d'ordre 5) de la voyelle /e/ on remarque que le premier pic sur le spectre PLP correspond au premier pic du spectre LPC. Le second pic qui apparaît dans le spectre PLP (voir aussi fig. III-13 (c)) ne correspond ni au 2ème ni au 3ème pic du modèle LPC.

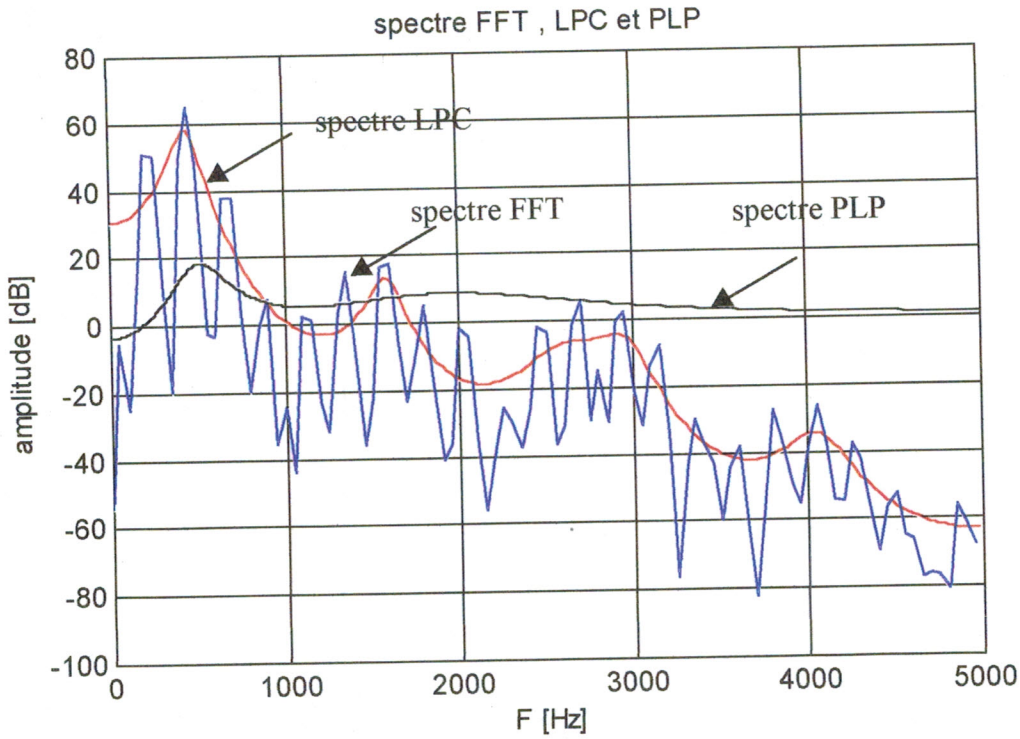


Fig. III - 11 Spectres FFT, LPC et PLP du phonème /e/

Si on essaye de représenter les même spectres sur une même figure mais cette fois-ci pour la voyelle /a/ (fig. III-12), et bien on ne voit l'apparition que d'un seul pic sur le spectre obtenu par la PLP.

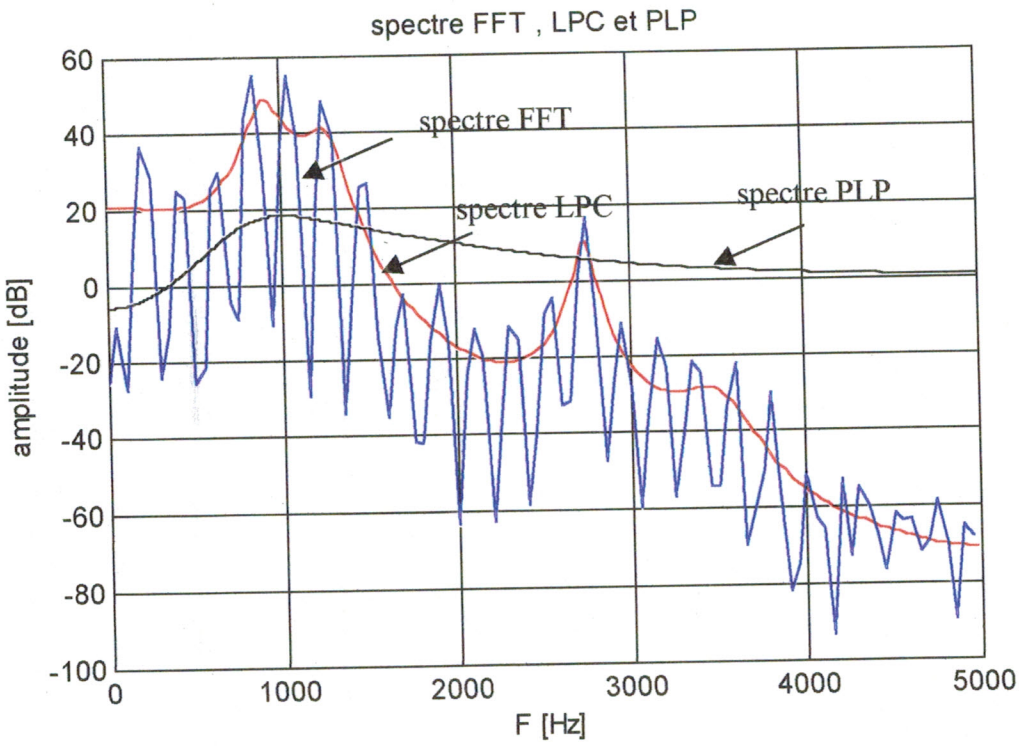


Fig. III - 12 Spectres FFT, LPC et PLP du phonème /a/

Dans l'identification de voyelles postérieures comme /a/, le seul pic présent sur le spectre PLP (fig. III – 13 (a)) est en relation avec le premier pic du spectre LPC (fig. III – 12). Tandis que pour les voyelles antérieures tels que /i/ ou /e/, où il y a 2 pics sur le spectre PLP, le second est en relation avec le $3^{ème}$ ou même le $4^{ème}$ formant du spectre LPC.

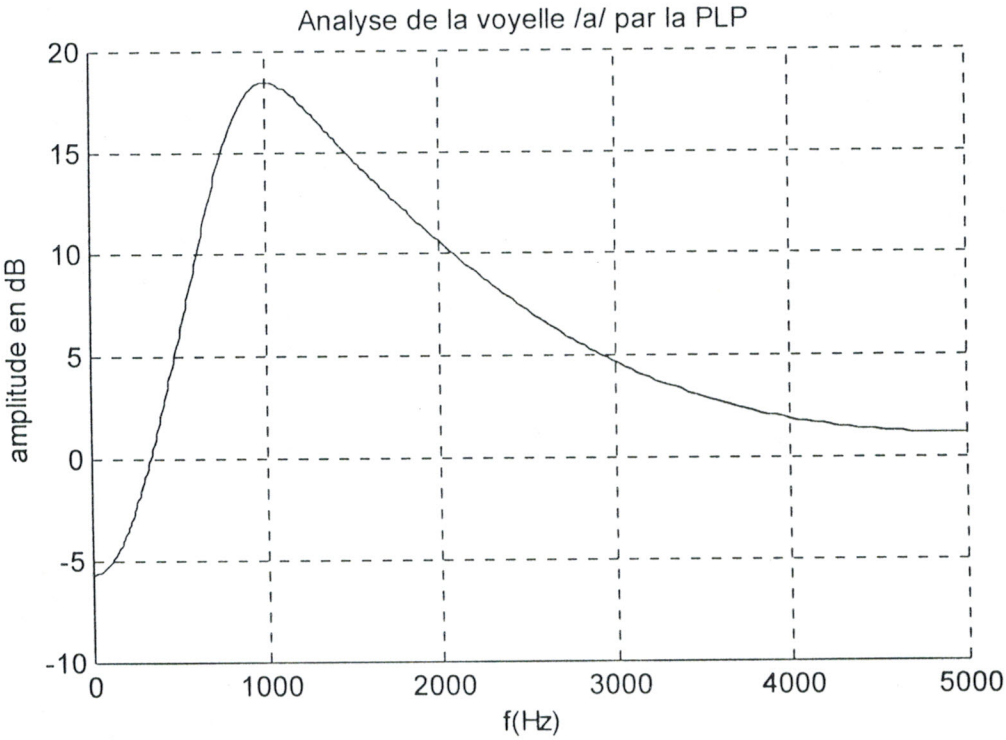


Fig. III - 13 (a) Les formants du phonème postérieur /a/

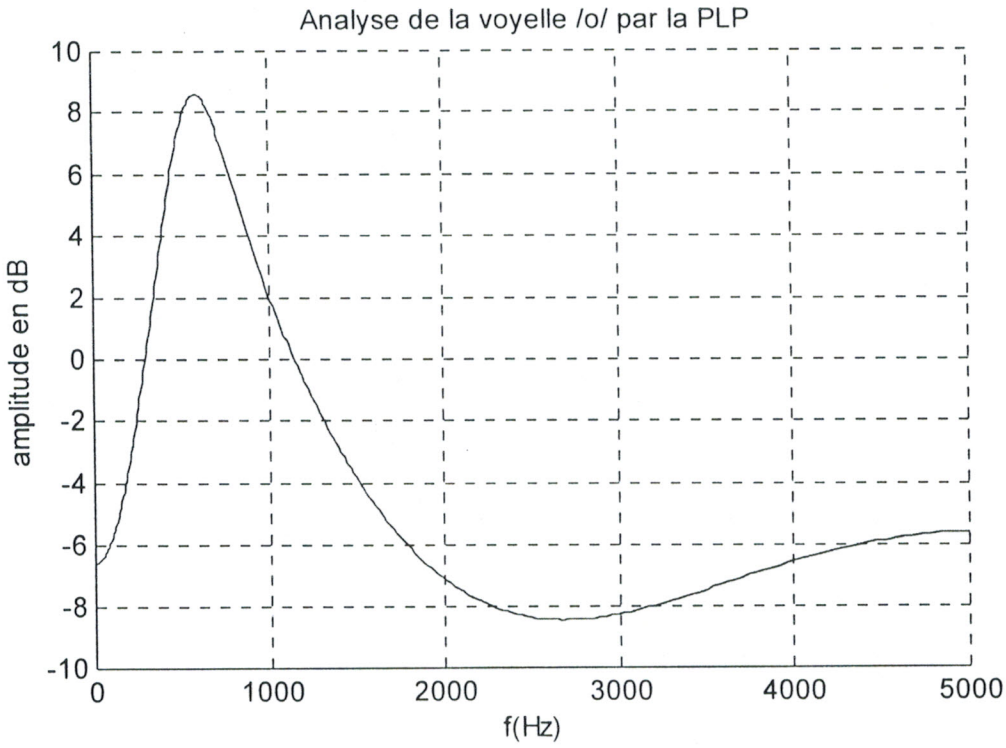


Fig. III - 13 (b) Les formants du phonème postérieur /o/

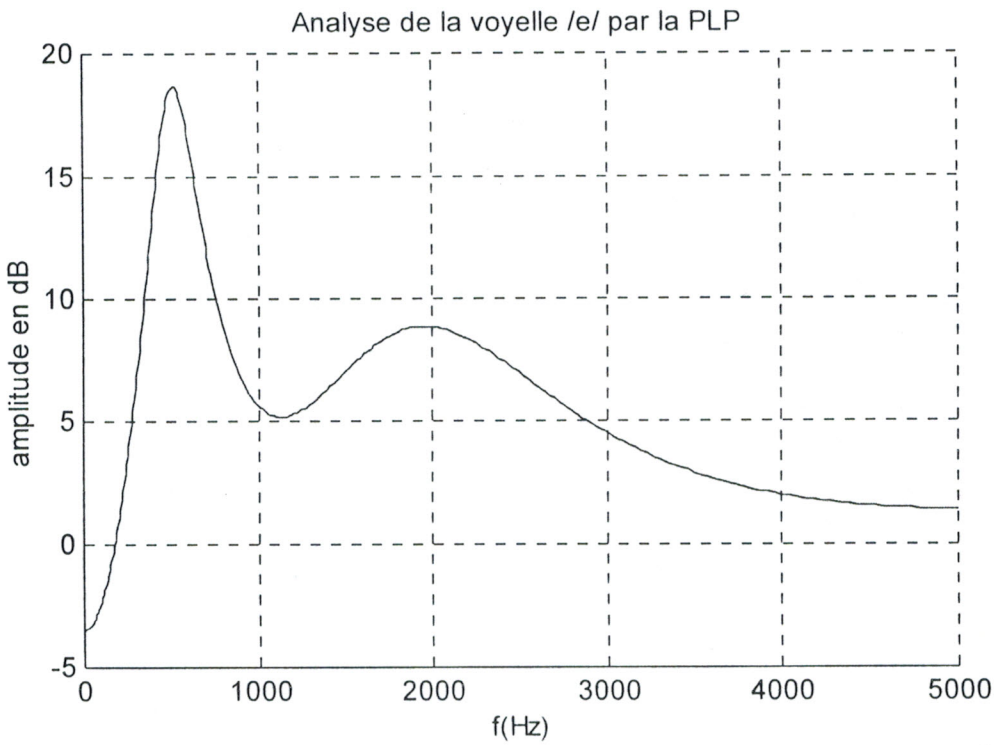


Fig. III - 13 (c) Les formants du phonème antérieur /e/

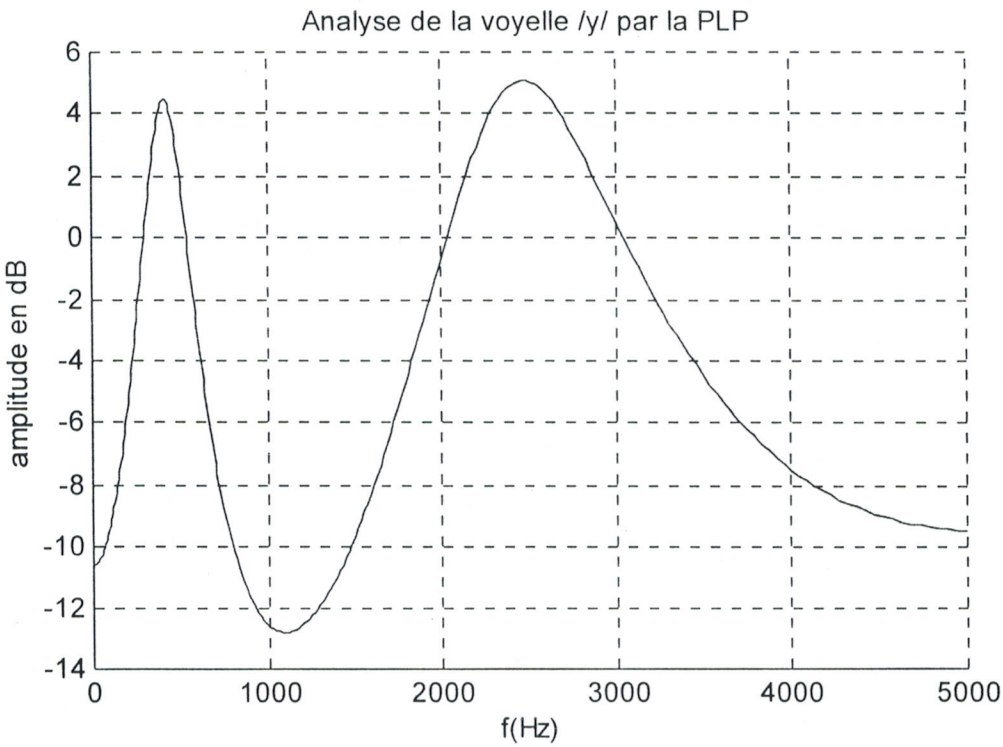


Fig. III - 13 (d) Les formants du phonème antérieur /y/

III – 7 – 2 Théorie d'intégration du pic spectral

Deux fréquences de formants suffisamment proches l'un de l'autre, sont représentés par un seul pic sur le spectre PLP. La condition pour que deux formants donnent lieu à deux pics sur le spectre PLP est que la distance $F_2 - F_1$ soit plus grande que 3 à 4 Bark.

Cette théorie d'intégration du pic a à peu prés 3,5 Bark a été reporté par Chistovich [26] après une série d'expériences puis par Hermansky [21].

Si on reprend les valeurs des formants de quelques voyelles données en tableau I-1 (§ I-4-4) pour essayer de voir l'écart en Bark entre le premier formant et le second, on obtient le tableau III –2 donné ci-dessous :

Voyelles	F_1(Hz)	F_2(Hz)	F_1(Bark)	F_2(Bark)	$\Delta F(Bark) = F_2 - F_1$
/i/	250	2500	2.43	12.80	10.37
/e/	375	2200	3.54	12.06	8.52
/a/	750	1350	6.28	9.30	3.02
/u/	250	750	2.43	6.28	3.85
/y/	250	1800	2.43	10.91	8.48
/o/	375	750	3.54	6.28	2.74

Tableau III – 2 Ecart $\Delta F = F_2 - F_1(Bark)$

Les fréquences de formants en Bark sont obtenues en utilisant l'équation (3) de la transformation Hertz_ Bark (§ III-3-2).

Dans les tests que nous avons effectués, nous obtenons pour la voyelle /a/ par exemple, un seul pic sur le spectre PLP (fig. III –13 (a)). Par contre sur le spectre LPC on observe 2 premiers pics très rapprochés (fig. III- 12). D'après le tableau III-2 on a pour cette voyelle un écart $\Delta F = 3.02$ Barks inférieur à 3.5 Barks.

Nos résultats sont donc en accord avec la théorie d'intégration du pic spectral, selon laquelle 2 formants séparés de moins de 3.5 Barks environ ne sont plus représentés que par un seul pic par la PLP.

Nous avons obtenu des résultats similaires pour les voyelles /u/ et /o/.

Par contre sur le spectre PLP de la voyelle /e/ (fig. III-13 (e)) on remarque la présence de 2 pics. Si on compare avec le spectre LPC (fig. III-11), on voit bien que les deux premiers pics qui apparaissent sont assez éloignés l'un de l'autre, cette différence entre les deux premiers formants est supérieure à 4 Barks. D'après le tableau III-2, on a pour cette voyelle un $\Delta F = 8.52$ Barks largement supérieur à 4 Barks.

Ce qui nous permet de préserver sur le spectre PLP un premier pic qui correspond plus au moins au premier formant du spectre LPC, tandis que le $2^{ème}$ pic du spectre PLP est en relation avec les formants supérieurs ($2^{ème}$ et $3^{ème}$) du spectre LPC.

Le même principe est retrouvé pour les voyelles /i/ et /y/.

Donc nos résultats restent en accord avec la théorie d'intégration du pic spectral dans laquelle 2 pics assez éloignés dans le spectre LPC ne sont pas fusionnés.

Le concept F_2' et la théorie d'intégration du pic spectral à 3,5 Bark sont tous les deux difficiles à expliquer sur des terrains purement psychophysiques et sont reliés avec ce qu'on appelle « speech mode » (mode de parole) de l'ouïe. De même, alors que les opérations pour obtenir le spectre auditif avant la modélisation autoregressive dans la PLP peuvent être justifiées par les propriétés psychophysiques de l'ouïe la modélisation autoregressive ne peut pas l'être.

Cette modélisation est utilisée dans le but de débarrasser le spectre auditif des détails qui doivent être éliminés pour supprimer l'information dépendante du locuteur [21].

III – 8 Comparaison de la PLP avec une PLC d'ordre faible

Une analyse LPC d'ordre 6 effectuée sur des voyelles postérieures [/a/, /o/, /u/] et des voyelles antérieures [/e/, /i/, /y/] montrent toujours la présence de deux pics sur leurs spectres. A titre d'exemple, nous illustrons ci-dessous les spectres obtenus par une analyse LPC d'ordre 6, appliquée sur la voyelle /a/ (fig. III-14 (a)) et sur la voyelle /e/ (fig. III-14 (b)) :

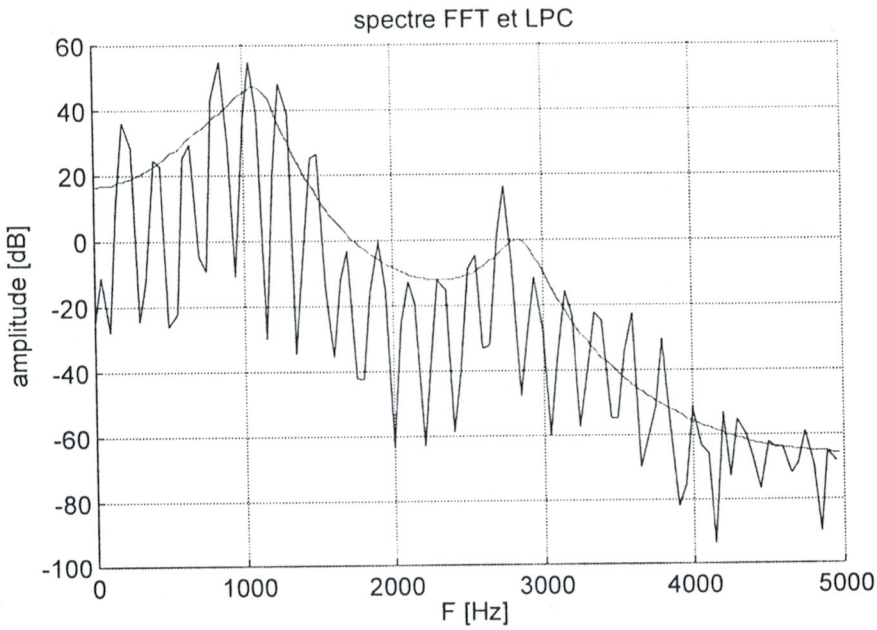


Fig. III - 14 (a) Spectres de la voyelle /a/
(FFT et PLC d'ordre 6).

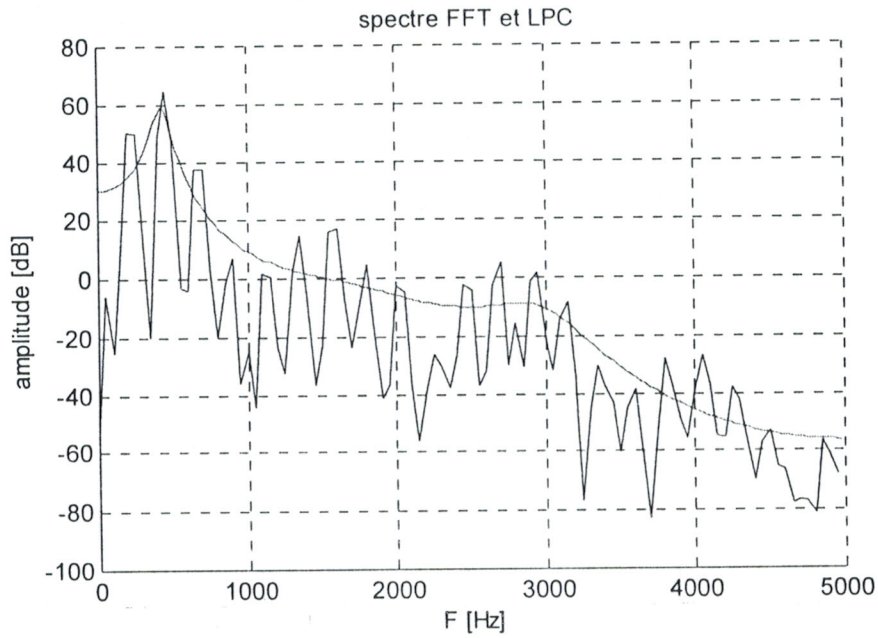


Fig. III - 14 (b) Spectres de la voyelle /e/
(FFT et PLC d'ordre 6).

Or une analyse PLP d'ordre 5 de ces même voyelles montre qu'un seul pic apparaît dans le spectre des voyelles postérieures (fig. III-13 (a) (b)), alors qu'il y en a deux dans celui des voyelles antérieures (fig. III-13 (c) (d)).

L'analyse PLP est donc en accord avec les expériences de perception des voyelles que nous avons citées au § III – 7 – 1 et selon lesquelles un auditeur peut identifier les voyelles postérieures à l'aide de stimuli synthétiques à un formant alors que pour identifier les voyelles antérieures des stimuli à deux formants sont nécessaires.
Comme le montre les fig. III-14 (a) et (b), l'analyse PLC n'est pas en accord avec ces expériences de perception des voyelles.

III – 9 Conclusion

Le but poursuivi dans ces études, et attendu de l'analyse, est de déterminer les mécanismes qui permettent à l'auditeur à la fois de différencier des voyelles acoustiquement proches et reconnaître une voyelle prononcée dans différentes conditions et différents contextes. C'est ce qui est possible pour l'ouïe humaine.

L'étude présentée dans ce chapitre montre que le principe de l'analyse PLP est d'approximer le spectre auditif de la parole par un modèle tout pôle en introduisant des principes connus dans la perception humaine de la parole (concepts de base de la psychophysique) tels que : la courbe à effet de masque, la courbe isosonique et la compression d'amplitude par la racine cubique (la loi de puissance). Ce que ne permet pas de faire une analyse par prédiction linéaire classique.

De plus l'ordre du modèle relativement réduit pour la PLP par rapport à la PLC, fixé dans cette étude à 5 nous a permis de retrouver et d'identifier les phonèmes comme une première application de la PLP avec un minimum d'espace mémoire (nombre de paramètres est proportionnel à l'ordre du modèle). En fait les résultats obtenus montrent que l'information essentielle pour l'identification de voyelles réside dans la forme globale du spectre et plus précisément dans la localisation des maximas spectraux correspondant aux 2 ou 3 premiers formants.

Ceci dit, l'analyse PLP n'est pas parfaite. Un de ses points faibles est la dépendance du résultat de toute la balance spectrale de $P(\omega)$ (l'amplitude des formants). La balance spectrale (*spectral balance*) est facilement affectée par des facteurs tels que l'équipement d'enregistrement, le canal de communication ou les bruits additifs. Mais on peut y remédier par une bonne mesure de distorsion à posteriori. Quand un changement de fréquence survient

dans la bande critique toute la balance spectrale est influencée et par conséquent la forme du spectre du modèle tout pôle aussi est influencée.

L'analyse PLP admet aussi le concept du pic F_2' *the effectif second formant* et la théorie d'intégration du pic spectral de 3,5 Bark des voyelles. Les résultats de nos tests sont accord avec ces théories. En effet, nous avons trouvé pour les voyelles antérieures (/e/, /i/, /y/) des spectres à deux maximas alors que pour les voyelles postérieures (/a/, /o/, /u/) il n'y a qu'un seul pic sur le spectre.

Après étude, implémentation et tests par l'analyse PLP dans le domaine de l'identification de phonèmes, une investigation des performances de cette technique dans le domaine de la reconnaissance de la parole est présentée dans le chapitre suivant.

CHAPITRE IV

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE RAP