U.F.R. STMIA
Université de Nancy I

CRIN
INRIA

Centre de Recherche en Informatique de Nancy

ANALYSE ET RECONNAISSANCE DE LA PAROLE PAR MODELES RETRO-AUTOREGRESSIFS ET RESEAUX NEURONAUX

THÈSE

Présentation et Soutenue Publiquement le 11 Juillet 1990
à Université de Nancy I
pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE NANCY I
EN INFORMATIQUE
par

Wu Li

Composition du Jury:

Président:	Jean-Marie PIERREL
Rapporteurs:	René CARRE
	Pierre MARCHAND
Exa	Marie-Christine HATON
	Jean-Paul HATON
	François LONCHAMP

U.F.R. STMIA
Université de Nancy I

CRIN
INRIA

Centre de Recherche en Informatique de Nancy

ANALYSE ET RECONNAISSANCE DE LA PAROLE PAR MODELES RETRO-AUTOREGRESSIFS ET RESEAUX NEURONAUX

THÈSE

Présentation et Soutenue Publiquement le 11 Juillet 1990

à Université de Nancy I

pour l'obtention du grade de

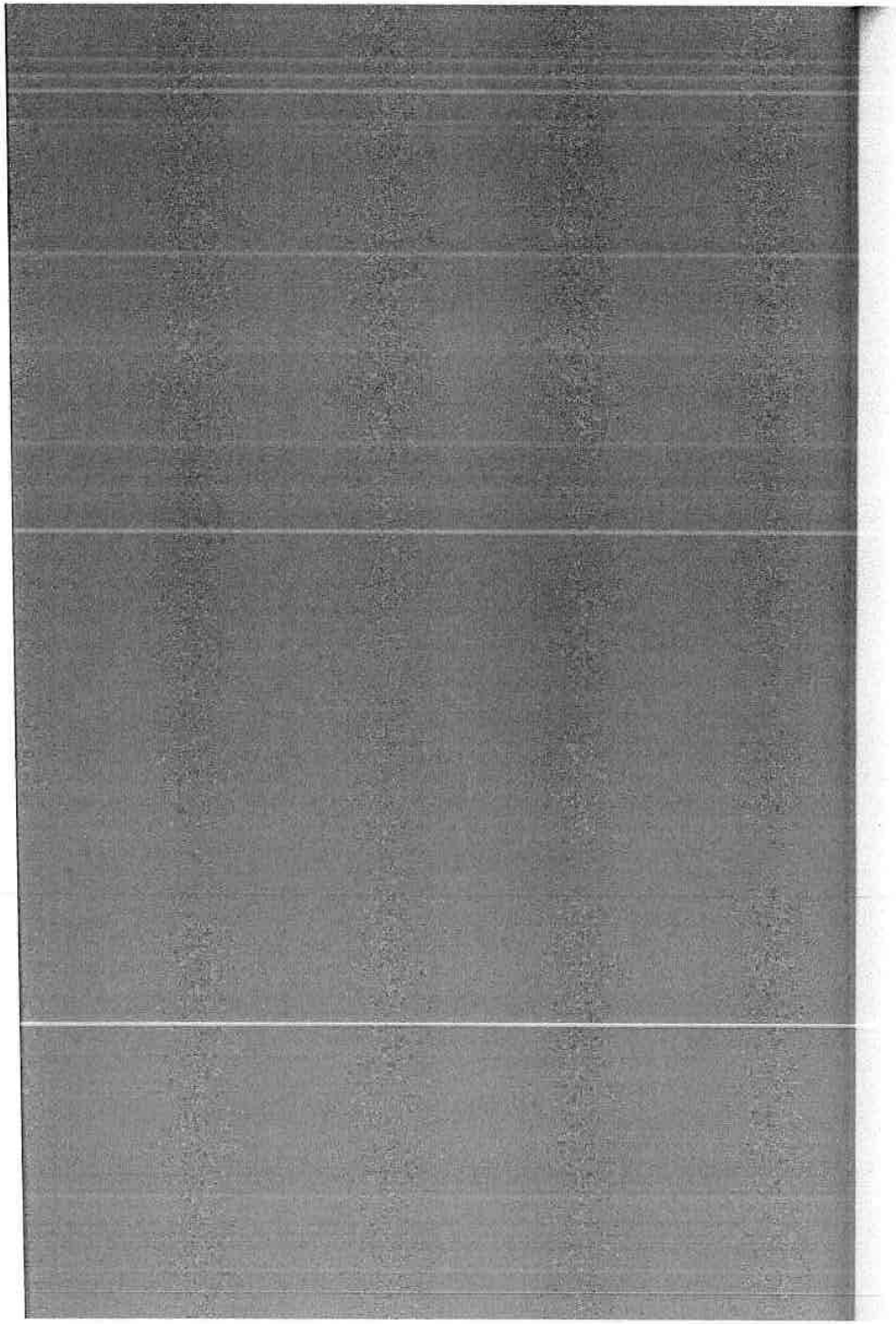
DOCTEUR DE L'UNIVERSITÉ DE NANCY I
EN INFORMATIQUE

par

Wu Li

Composition du Jury:

Président:	Jean-Marie PIERREL
Rapporteurs:	René CARRE
	Pierre MARCHAND
Examinateurs:	Marie-Christine HATON
	Jean-Paul HATON
	François LONCHAMP



U.F.R. STMIA
Université de Nancy I

CRIN
INRIA

Centre de Recherche en Informatique de Nancy

ANALYSE ET RECONNAISSANCE DE LA PAROLE PAR MODELES RETRO-AUTOREGRESSIFS ET RESEAUX NEURONAUX

THÈSE

Présentation et Soutenue Publiquement le 11 Juillet 1990
à Université de Nancy I
pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE NANCY I
EN INFORMATIQUE

par

Wu Li

Composition du Jury:

Président:	Jean-Marie PIERREL
Rapporteurs:	René CARRE
	Pierre MARCHAND
Examinateurs:	Marie-Christine HATON
	Jean-Paul HATON
	François LONCHAMP

Je tiens à remercier Monsieur Jean-Paul Haton qui a accepté de m'accueillir dans son laboratoire et de diriger mon travail de recherche. Qu'il trouve ici l'expression de ma profond gratitude pour m'avoir confié un sujet intéressant: la reconnaissance de la parole.

Je tiens à remercier Monsieur Jean-Marie Pierrel qui me fait l'honneur de bien vouloir présider ce jury.

Monsieur René Carré m'a fourni des remarques importantes durant mon rapport.

Monsieur François Lonchamp qui m'a donné des conseils nécessaires pendant mon travail. Sans lui, je n'aurais pas pu rentrer rapidement dans le sujet concernant des voyelles nasales.

Je tiens à remercier Monsieur Pierre Marchand ainsi que Madame Marie-Christine Haton qui ont accepté de juger ce travail.

Enfin, je remercie tout particulièrement Anna Boyer, Yifan Gong, Sunee Limchareun pour l'aide qu'ils m'ont apportée et pour leur soutien.

Table des matières

1
Introduction

1.1 Introduction

Le langage est une expression de la sagesse humaine. Depuis le jour où il s'est formé, il permet une communication simple et facile entre les hommes.

L'homme ne peut pas vivre sans langue. Et pourtant, si l'homme a réussi à calculer précisément l'orbite des planètes, à lancer de la terre des navettes, il doit avouer qu'il est encore un profane dans la connaissance des mécanismes de la communication.

Une langue ne se forme pas en un jour. Elle est liée étroitement au développement de l'activité sociale. Elle n'existe que par l'homme. L'apprentissage de la langue ne peut donc pas être facilement obtenue. L'informatique n'a pas réussi, jusqu'à présent, à assurer une communication orale entre l'homme et la machine. On peut cependant espérer qu'au lieu de communiquer à l'aide de clavier, l'homme pourra un jour faire comprendre sa langue à l'ordinateur, et ainsi réaliser l'ambition de dialogue entre l'homme et la machine.

Tout acte de parole suppose la présence d'au moins deux personnes: celle qui parle et celle qui écoute. L'une produit des sons, l'autre les entend et les interprète. Si nous remplaçons une de ces deux personnes par l'ordinateur, alors l'écoute de la parole correspond à la reconnaissance, et, en sens inverse, il s'agit de la synthèse.

Bien que l'on ait commencé depuis longtemps à mener des études de linguistique et de phonétique, et bien que l'on sache maintenant réaliser des systèmes de reconnaissance très sophistiqués à partir de systèmes tout simples sur les mots isolés, les résultats de reconnaissance de la plupart des

systèmes sont encore médiocres par rapport à ceux de l'homme. La reconnaissance de la langue oblige la machine à prendre toutes ses forces *brutes*, il n'empêche que le résultat est loin d'être parfait. L'avion conçu selon le mécanisme des oiseaux n'a pas réussi, ce qui vole au ciel d'azur est l'avion inconventionnel qui se base sur la théorie de l'aérodynamique. Bien que l'on n'ait pas encore trouvé la meilleur méthode pour la reconnaissance de la parole par l'ordinateur, il est sûr que l'étude de la langue humaine peut être utile pour ces travaux.

La communication orale est un processus d'*encodage* et de *décodage*. Le message codé franchit l'espace intermédiaire et est décodé par la personne qui le recoit. L'unité d'encodage est le phonème. L'unité de décodage est difficile à définir, puisque le but du décodage est de comprendre le sens véhiculé par le signal. Alors l'unité n'est pas forcément le phonème en phase de décodage. Mais elle se fonde sur le phonème.

La figure 1.1 montre plusieurs représentations du signal de parole:

(a) la parole dans le domaine temporel, ce qui nous donne une relation de l'amplitude avec le temps.

(b) la parole dans le domaine fréquentiel en fonction du temps. L'intensité des composants spectraux est donnée par le niveau de gris du tracé. Les composants spectraux sont obtenus en utilisan une analyse de type DFT (Discrete Fourier Transform), LPC (Linear Predictive Coding) ou bien Cepstre appliquée à une fenêtre de la parole à un instant donné.

(c) l'énergie sonore en moyenne totale.

(d) la fréquence d'excitation que l'on appelle fréquence fondamentale, ce qui nous fournit l'indice *voisé* ou *non-voisé* d'un phonème.

Le système phonétique minimal du français comporte environ 16 voyelles et 20 consonnes.

Les positions des formants sont déterminées par l'articulation, c'est-à-dire la forme des résonateurs. La correspondance entre articulation et acoustique se trouve dans un plan F1/F2 (F1: 1er formant, F2: 2ème formant). Comme la plupart de l'énergie se concentre autour du formant, il est raisonnable de représenter les voyelles par leurs formants et largeurs de bande.

Ce travail concerne l'analyse et la reconnaissance des voyelles nasales françaises.

Dans le classement phonétique des voyelles du français, il y a quatre nasales: /ɑ̃/ /ɔ̃/ /ɛ̃/ /œ̃/ comme dans *lent*, *long*, *lin* et l'*un*. La distinction entre /ɛ̃/ et /œ̃/ n'est plus faite par un nombre important de locuteurs du français standard [42].

Une voyelle nasale possède certaines propriétés qui la distinguent d'une voyelle orale.

Au plan articulation, les voyelles nasales se caractérisent par l'abaissement du voile du palais. Si le voile ferme le passage par le nez en s'accolant à la paroi postérieur du pharynx, on obtient une articulation *orale*. Si par contre le voile du palais laisse ce passage libre, on obtient une articulation *nasale*.

Le degré de couplage est variable. A l'extrémité de la bouche, il y a un mélange des sources orale et nasale. A cause de cette connexion, le spectre des voyelles nasales est complexe et très difficile à interpréter. La figure 1.2 nous montre un exemple. F'in est noté comme le *ième* formant nasal, F'i comme le formant oral en cas de nasalisation.

On distingue les voyelles nasales et les voyelles nasalisées avec mise en jeu du conduit nasal [57]: les voyelles *nasalisées* sont produites lors d'un léger abaissement du voile du palais. Beaucoup de voyelles anglaises, en particulier, sont nasalisées surtout au voisinage des consonnes nasales, mais elles conservent leurs caratères phonétiques de voyelles orales. Les voyelles nasales sont produites au cour d'un abaissement important du voile du palais associé à certaines configuration spécifiques du conduit vocal.

Comme vue dans la figure 1.2, la localisation des formants nasals est très difficile. Est-ce que ces formants nasals jouent un rôle important dans la reconnaissance des voyelles nasales? Comment faire pour une extraction correcte des formants oraux quand les formants nasals se présentent dans le spectre? Comme renforcer les formants nasals? Ce sont des problèmes rencontrés en reconnaissance de la parole et que nous avons tenté de résoudre dans cette thèse.

Parmi tous les groupes des sons, seules les voyelles orales ont un modèle permettant de les représenter analytiquement. Quant aux autres, une modélisation dépend de la connaissance sur l'articulation phonétique.

Du point de vue de la complexité de la parole naturelle, le problème du choix de l'ordre de prédiction du modèle utilisé est crucial. Le choix est guidé par le nombre de pôles et zéros. Leur répartition devient plus complexe au fur et à mesure de l'augmentation de la fréquence. Analyser les consonnes exige l'elargissement de la région fréquentielle. Pour pouvoir facilement analyser les voylles nasales dans le même environnement, nous proposons d'introduire la *prédiction linéaire sélective*, l'idée de Makhoul, dans le modèle ARMA. Dans la deuxième chapitre, nous présentons d'abord un rappel des techniques de traitement du signal, ensuite les modèles AR, MA et ARMA pour conclure sur notre modèle ARMA sélectif spécialement conçu pour l'analyse des voyelles nasales.

Après avoir construit un modèle, le problème se pose de savoir comment extraire des formants significatifs. Toujours à cause de la nasalisation, nous ne pouvons plus faire une extraction simple en mesurant la largeur de bande d'un formant. Nous savons qu'un expert phonéticien ne choisit pas les formants uniquement en utilisant le critère de la bande minimale, mais qu'il examine d'abord l'allure générale de tout le spectre de la portion de signal étudiée et détermine ensuite trois zones de recherche des formants. De la même manière, nous commençons par comparer le spectre inconnu à tous les spectres de la bibliothèque: ceci est équivalent à l'examen de l'allure globale du spectre. Et ensuite nous cherchons les candidats autour des formants fournis par la bibliothèque. Dans le troisième chapitre, nous présentons un algorithme d'extraction des formants de voyelle en référence à un expert phonéticien.

Dans le quatrième chapitre, nous présentons d'abord les résultats de l'analyse des voyelles nasales, extraits de la thèse d'état de François LONCHAMP[41], puis deux expériences de reconnaissance de voyelles nasales se fondant sur les données fournies par notre modèle sélectif. Les résultats montrent que, à partir des trois premiers formants les plus significatifs, la reconnaissance des voyelles nasales /ã/ et /$\tilde{\text{ɔ}}$/ peuvent atteindre un score aussi bon que pour les voyelles orales. Par contre, la reconnaissance du /$\tilde{\text{ɛ}}$/ dépend du locuteur. Le score peut varier de 0% à 100%. L'expérience montre également que l'analyse par une fenêtre à long terme devient intéressante grâce à l'utilisation du modèle ARMA. Premièment, elle demande moins de calcul. Deuxièment, l'effet de moyennage réduit énormément les pics aléatoires susceptibles d'apparaître dans le spectre lors d'une analyse à court terme, ce qui est très important dans l'extraction des formants pour une voyelle nasale.

L'analyse du signal de parole en utilisant des réseaux neuronaux est une méthode qui bénéficie depuis quelques années d'un regain d'intérêt. Dans le dernier chapitre, nous présentons nos expériences sur la reconnaissance des mots isolés en mono-locuteur et multi-locuteur fondées sur un modèle de perceptron multi-couches.

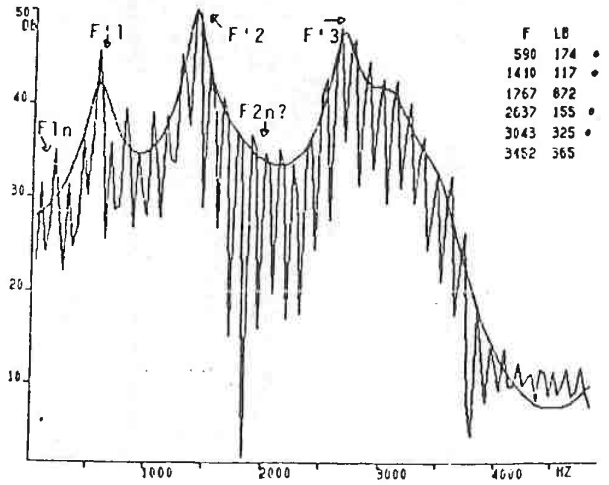


Fig 1.2 Spectre d'une voyelle nasale. .

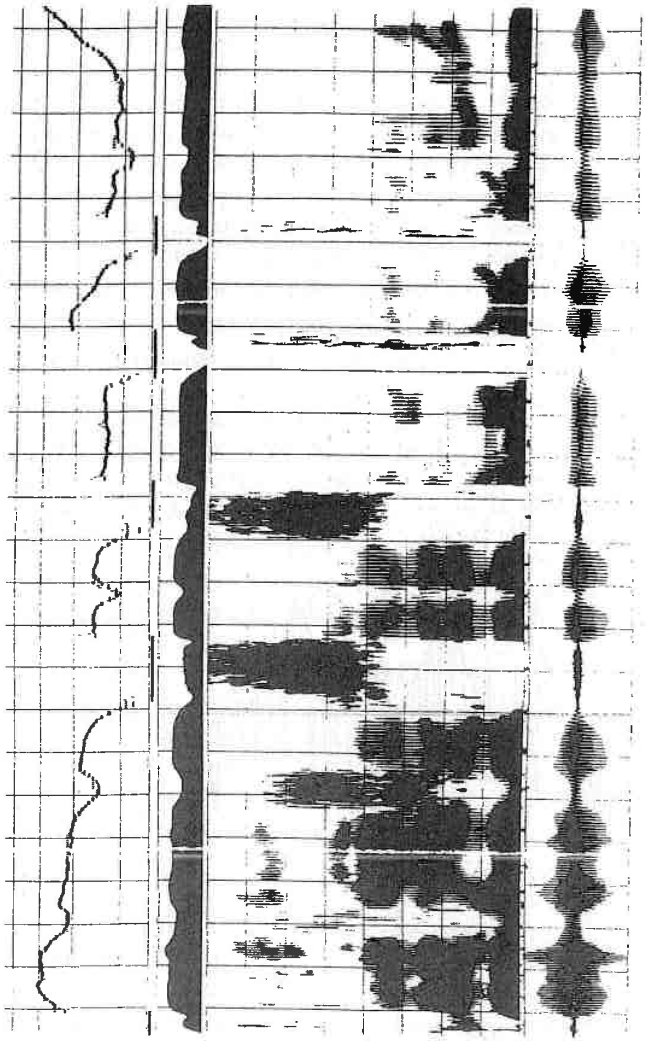


Figure 1.1 a:signal. b:spectrogramme. c:énergie. d:fréquence fondamentale.

2
Modèles pour l'analyse de la parole

2.1 Introduction

Le modèle d'un système est défini comme une relation mathématique appropriée qui relie l'entrée et la sortie du système. L'application de cette relation permet de contrôler, en fonction de la sortie, l'environnement du système afin de garantir un état optimal. L'application à l'identification des formes consiste à représenter les informations essentielles contenues dans le signal de sortie, à établir une forme selon ces informations essentielles et à la comparer avec les formes connues.

L'intérêt de l'analyse de la parole par modèle est de permettre de représenter le signal par une relation analytique et de saisir les caractères principaux. Dans ce chapitre, nous faisons d'abord quelques rappels concernant le traitement de base du signal, ensuite nous parlons des modèles AR, MA et ARMA et dans la dernière section nous présenterons notre algorithme ARMA dans les intervalles fréquentiels sélectifs, bien adapté à l'analyse des voyelles nasales.

2.2 Quelques rappels sur les méthodes classiques en analyse du signal

Le traitement de la parole par ordinateur est basé sur des théories de traitement numérique[61,68,49,69,33]. Nous indiquons ici ce qui est nécessaire la compréhension de la suite de l'exposé.

2.2.1 Discrétisation

Si nous représentons un signal continu par $x_a(t)$, et son spectre par $X_a(\Omega)$, alors ils sont reliés par la transformation de Fourier suivante:

$$x_a(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X_a(j\Omega) e^{j\Omega t} d\Omega$$

et

$$X_a(j\Omega) = \int_{-\infty}^{+\infty} x_a(t) e^{-j\Omega t} dt$$

Ω étant la fréquence en radian et t le temps.

L'ordinateur ne travaille que sur les données numérisées. Le premier pas pour traiter la parole est de la discrétiser à l'aide d'un convertisseur analogique-numérique. Soit un signal échantillonné à intervalle de temps T, notons $x(n)$ le signal numérisé, alors

$$x(n) = x_a(nT) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X_a(j\Omega) e^{-j\Omega nT} d\Omega$$

Puisque la transformée de Fourier pour un signal numérisé est

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} X(e^{j\omega}) e^{j\omega n} d\omega \qquad (2.1)$$

donc, nous avons la correspondance entre les deux spectres continu et discret:

$$X(e^{j\Omega T}) = \frac{1}{T} \sum_{r=-\infty}^{\infty} X_a(j\Omega + j\frac{2\pi r}{T}) \qquad (2.2)$$

la relation entre la fréquence continue et celle discrète vérifie

$$\omega = \Omega T \qquad (2.3)$$

$$\Omega = 2\pi f, \quad f : Hz; \quad \omega : -\pi .. + \pi; \quad T : seconde$$

On note Ω_o la **pulsation** la plus élevée des composantes spectrales du $X_a(j\Omega)$. f_s la fréquence d'échantillonnage et T la période vérifiant $T = \frac{1}{f_s}$. Pour garder l'invariance du spectre avant et après l'échantillonnage, Ω_o doit être inférieure ou égale à deux fois f_s, soit $\Omega_o \leq 2\pi f_s$ ou $\Omega_o \leq 2\pi/T$. Ceci veut dire que le signal étudié ne doit pas avoir de composante spectrale au-delà de $2\pi f_s$. Si la fréquence maximum Ω_o n'est pas inférieure à la fréquence f_s qu'on désire, alors il faut éliminer artificiellement les composantes fréquentielles à partir de $2\pi/T$ par un filtre passe-bas. Sinon il y aura des repliements du spectre, et le spectre observé pour l'analyse sera déformé. Si au contraire, la fréquence maximum Ω_o est inférieure à $2\pi/T$, on n'aura pas besoin de filtrer. Dans ce cas, la fréquence maximum du signal numérisé est supérieure à celle du signal continu. Par conséquent, la fréquence maximale d'un signal numérisé ne dépend pas du rang fréquentiel de celui-ci avant d'être numérisé. La fréquence maximum numérisée est toujours la moitié de la fréquence d'échantillonnage.

2.2.2 Transformation de Fourier discrète (DFT)

La formule (2.1) a supposé que le nombre des échantillons d'un signal est infini, par conséquent, le spectre est de type continu. Il faut donc numériser aussi le spectre parce que nous ne pouvons pas sauvegarder ce spectre continu en mémoire d'ordinateur.

La parole est un signal dont les caractéristiques varient en fonction du temps, on doit la couper en segments courts pour obtenir des paramètres stables. La transformée de Fourier discrète (2.4) permet de calculer son spectre numérisé à partir des N échantillons du signal. La longueur (nombre des échantillons) du spectre est aussi N. Les deux séries spectrale et temporelle sont liées par les relations suivantes:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)nk} \qquad (2.4)$$

et

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)nk} \qquad (2.5)$$

2.2.3 Transformée en Z

Si la transformée de Fourier ne décrit que des caractéristiques du signal au long de l'axe des fréquences réelles, la transformée en Z nous permet de représenter le signal dans un plan complexe de deux axes: l'un est la fréquence réelle, l'autre l'effet de l'amortissement.

Soit x(n) le signal échantillonné, par définition la transformée en Z de ce signal s'écrit:

$$X(z) = \sum_{n=0}^{\infty} x(n) z^{-n} \tag{2.6}$$

$$z = re^{j\omega}$$

La transformée en z est égale à la transformée de Fourier lorsque $r = 1$ ou $|z| = 1$:

$$X(\omega) = \sum_{n=0}^{\infty} x(n) r^{-n} e^{-j\omega n}|_{r=1} \tag{2.7}$$

C'est une transformée en z sur le cercle unitaire.

2.2.4 Relation entre la transformée en Z et en ω, Ω

La transformée en z joue un rôle pour analyser une série des données discrétisées, pourtant la transformation en ω est importante pour analyser les caractères du système. Les deux sont reliées par la formule (2.7). z et ω sont des fréquences normalisées pour un signal numérisé. La relation 2.3 nous permet de retrouver la fréquence réelle en Hz à partir de la fréquence normalisée.

2.2.5 Fonction de transfert du système

Supposons que le système est linéaire et invariant au cours du temps. Une réponse fréquentielle est définie comme une transformée de Fourier sur la réponse du système à une entrée d'excitation de type d'impulsion unitaire. Egalement, une fonction du système est la transformée de z sur cette réponse. Un système peut se caractériser par plusieurs représentations (Fig. 2.2). Pour la simplicité, on ne distingue plus ici la réponse fréquentielle du système ou la fonction de transfert du système.

Si le système est linéaire et invariant dans le temps, sa sortie peut être représentée par une convolution de la fonction du système avec l'entrée.

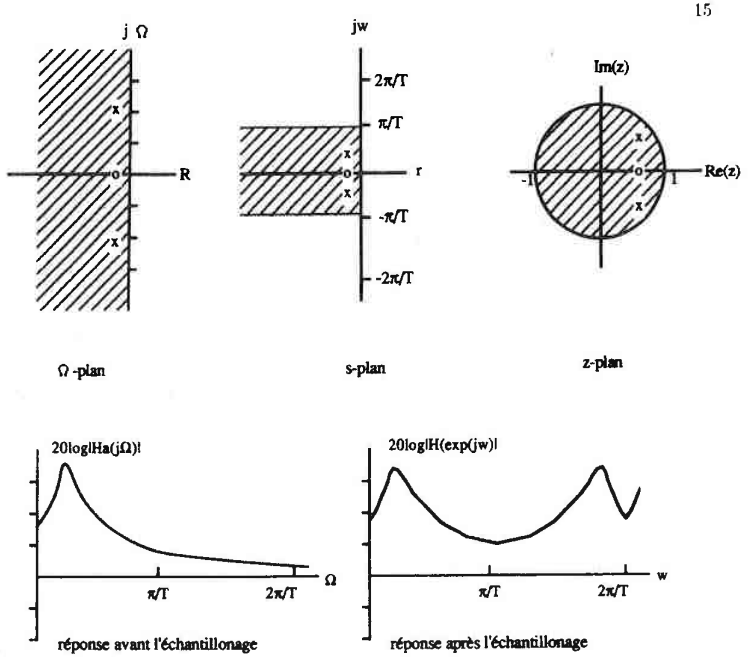


Figure 2.1. Présentation d'un sytème en z, w et Ω -plan.

Si un système est physiquement réalisable, et si sa réponse ne parvient jamais avant l'excitation qu'il subit, ce système est dit causal. C'est à dire:

si $n < n_o$, et $x_1(n) = x_2(n)$
alors $n < n_o$, $y_1(n) = y_2(n)$.

En plus, un pôle du système est défini comme une fréquence à laquelle la réponse du système devient maximale. Un zéro de système est une fréquence à laquelle la réponse devient minimale.

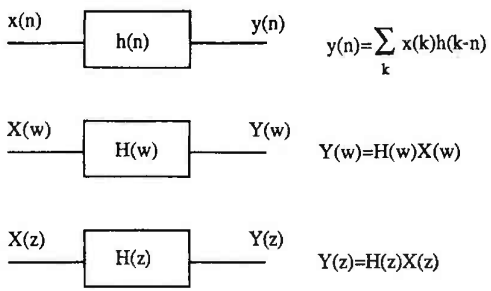


Figure 2.2. Représentations diverses d'un système. h(n):réponse d'impulsion unitaire, H(ω):réponse fréquentielle. H(z):fonction de transfert

Un système est dit stable si et seulement si la condition suivante est vérifiée:

$$\sum_{k=-\infty}^{\infty} |h(k)| \leq \infty$$

ici h(k) est la réponse impulsionnelle du système. Bien évidemment, les pôles du système doivent se trouver à la gauche du plan ω. La position des zéros n'a pas d'importance.

Pour un modèle de l'organe articulatoire, on est sûr qu'il est stable et causal, mais il ne possède pas de caractéristiques linéaires et invariantes dans le temps, ce qui se traduit par un spectrogramme évolutif. Pour pouvoir analyser ce genre de signal avec les outils exposés, nous devons restreindre la longueur de la fenêtre d'analyse et considérer qu'il n'est stable que pendant quelques centièmes de seconde.

2.2.6 Système de phase minimum

Supposons qu'un système est représenté sous une forme polaire:

$$H(z) = |H(z)|e^{jarg[H(z)]}$$

$$H(z) \longrightarrow h(n).$$

et sous une forme logarithmique complexe:

$$\hat{H}(z) = \log[H(z)] = \log|H(z)| + jarg[H(z)]$$

$$\hat{H}(z) \longrightarrow \hat{h}(n).$$

Si les deux formules suivantes sont vérifiées, on l'appelle système de phase minimum.

$$\log|H(e^{j\omega})| = \hat{h}(0) - \frac{1}{2\pi} P\int_{-\pi}^{\pi} arg[H(e^{j\theta})]cot(\frac{\theta-\omega}{2})d\theta$$

$$arg[H(e^{j\omega})] = \frac{1}{2\pi} P\int_{-\pi}^{\pi} log|H(e^{j\theta})|cot(\frac{\theta-\omega}{2})d\theta$$

Une des caractéristiques d'un système de phase minimum est que tous ses pôles et zéros sont à l'intérieur du cercle unitaire. On peut rétablir la phase à partir de l'amplitude, et également retrouver l'amplitude à partir de la phase.

Dans la pratique, dans le cas où on ne sait pas si un système est de phase minimum, on suppose alors qu'il l'est. L'avantage de cette hypothèse est de faciliter l'estimation des paramètres, car on peut ignorer l'une des deux parties.

2.3 Prétraitement du signal de la parole

2.3.1 Fenêtre

Dans la méthode d'analyse du spectre par la fonction d'autocorrélation, on a besoin de calculer en théorie une autocorrélation d'un signal de longueur infinie. Mais l'intervalle d'étude du signal étant de longueur limitée, on considère que le signal vaut zéro partout en dehors de l'intervalle d'analyse. Cette opération introduit en réalité un effet de la fenêtre $w(n) = 1(n) - 1(n-N)$, N étant le nombre du points du signal dans la fenêtre, $1(n)$ est une fonction unitaire (formule 2.8).

$$1(n) = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \qquad (2.8)$$

L'effet de cette fenêtre virtuelle est exprimé par

$$s_w^m(n) = w(m-n)\, s(n)$$

ici $s(n)$ est le signal original, $w(n)$ la fenêtre choisie, $s_w^m(n)$ le signal encadré et positionné au moment m. Alors sa transformée de Fourier sera:

$$S_w^m(e^{j\omega}) = W(e^{j\omega}) \odot S(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} W(e^{j\theta}) e^{j\theta m} S(e^{j(\omega+\theta)}) d\theta. \qquad (2.9)$$

le symbole $\odot$ signifie le calcul de convolution linéaire.

La cohérence entre S_w^m et S dépend de la forme de fenêtre $W(e^{j\theta})$.

L'intégrale de n'importe quelle fonction multipliée par la fonction de *Kronaker* $\delta(t-t_o)$ rend une valeur de fonction au moment t_o:

$$f(t_o) = \int_{-\infty}^{+\infty} \delta(t-t_o) f(t) dt. \qquad (2.10)$$

Nous voulons que S_w^m garde toutes les caractéristiques de s(n) sans encadrement. Par comparaison de (2.9) et (2.10), $W(e^{j\omega})$ doit posséder une forme étroite (soit $\delta(w)$ dans le cas idéal) pour arriver au but. Mais plus le spectre en fréquence est étroit, plus la longueur nécessaire du signal temporel est grande. Donc on ne peut choisir qu'une fenêtre relativement étroite pour un signal de longueur fixée.

Le critère de choix est de prendre une telle fenêtre dont le lobe principal est étroit, et le reste le plus faible possible par rapport au lobe principal. La figure 2.3 [22] donne le spectre de diverses fenêtres. La fenêtre rectangulaire a un lobe étroit, mais la différence de l'énergie entre ce premier lobe et le reste n'est pas très grande. Cela peut introduire une perturbation du spectre encadré. La fenêtre de *Hamming* est convenable pour nous. La différence de l'énergie entre le premier et le reste atteind 40dB, de plus, la largeur du premier lobe principal peut se réduire en élargissant la longueur de la fenêtre.

A cause de l'effet de décroisement dans les deux côtés de la fenêtre de Hamming, la longueur effective de Hamming dépend de l'énergie. Tribolet[81] a indiqué que la longueur effective dépend de la deuxième et quatrième puissance (W_2 W_4) de la fenêtre:

$$N_{eff} \cong \frac{W_2^2}{W_4} N = \beta N.$$

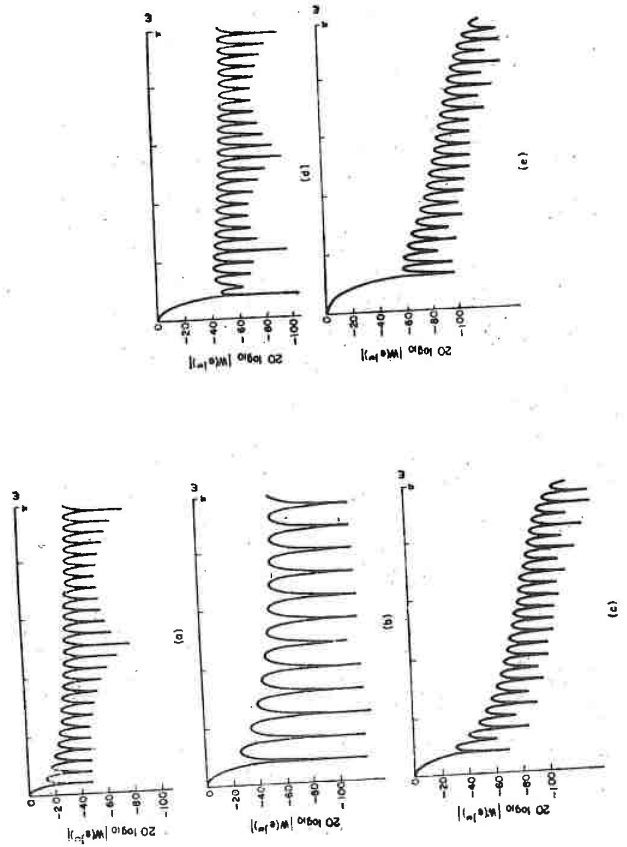


Figure 2.3. Spectres des diverses fenêtres d'après [oppe75]. (a)rectangle (b)triangle (c)Bartlett (d)Hanning (e)Hamming (f)Blackman

Pour une fenêtre rectangulaire, β prend valeur 1. Et pour *Hamming* $\beta \cong 0.55$.

Pour que la fenêtre influence moins les caractéristiques de la parole, nous devons également choisir une fenêtre de phase linéaire. Une fenêtre de longueur finie symétrique est de phase linéaire. La fenêtre *Hamming* convient.

2.3.2 Préaccentuation

Une préaccentuation est effectuée en différenciant deux échantillons successifs:

$$y(n) = y(n) - \alpha y(n-1) \qquad \alpha = 0.9 \sim 1$$

C'est comme si l'on faisait passer le signal à travers un filtre de fonction de transfert $H(z) = 1 - \alpha z^{-1}$.

La préaccentuation a deux effets sur le résultat:

1. améliorer la précision de l'analyse. Elle réduit la dynamique du spectre, donc elle réduit également l'influence des lobes secondaires de la fenêtre.
2. réduire la variation du résultat à cause de la position de la fenêtre d'analyse[67]. L'analyse devient moins sensible à la position de fenêtre.

Remarque: on fait toujours d'abord une préaccentuation et ensuite l'opération de fenêtre.

2.4 Modèle

2.4.1 Identification du modèle

L'identification est un processus grossier pour indiquer le genre du modèle représentatif en utilisant un ensemble de données ou bien des connaissances a priori sur le modèle physique.

Le modèle d'un système est déterminé simultanément par l'entrée et par la sortie. Donc avant d'avoir établi un modèle, c'est à dire avant d'établir les relations entre entrée et sortie, nous devons faire des études concernant les caractéristiques physiques de ce système. Deux catégories existent en ce qui concerne l'établissement d'un modèle physique [25].

La première classe appartient à la catégorie de problèmes appelée *boite complètement noire*. Nous n'avons aucune information sur ce système. Nous ne savons pas s'il est linéaire ou non linéaire, par exemple.

La deuxième classe est ce qu'on appelle *boite grise*. Dans ce genre de problème, certaines caractéristiques: linéarité, portée fréquentielle, etc., par exemple, sont connues par hypothèse. Mais il est possible que nous ne sachions pas l'ordre du système, ou d'autres paramètres.

Il est évident qu'un problème d'une boite grise peut être résolu plus facilement qu'un problème d'une boite complètement noire. En général, avant d'essayer de modéliser un problème, il faut toujours faire certaines hypothèses et simplifications, donc il est possible de transformer un problème de boite noire en un problème de boite grise. Un modèle de la production de parole est fait sous les hypothèses que le conduit est rigide et sans perte, que les surfaces de chaque section sont invariantes, et qu'il est linéaire et invariant dans le temps pendant un temps court.

2.4.2 Modèle de production de la parole

Le codage direct de la parole aboutit en un codage d'au moins 20.000 bits par seconde. Cette vitesse de bits est beaucoup plus grande que la quantité d'information portée par le signal de la parole – 60 bits par seconde [69]. Le codage sur le signal n'est pas un codage des informations les plus pertinentes. En d'autres mots, il y a d'autres paramètres qui sont intrinsèques.

La parole est un signal produit dans l'organe articulatoire qui est extrêmement souple et extrêmement complexe au point de vue physique. Le parole ainsi produit varie rapidement en fonction du temps, mais le mouvement de l'organe articulatoire est relativement lent à cause des contraintes physiques. Un organe articulatoire humain peut produire environ 10 phonèmes par seconde. Donc, il est clair que la parole n'est pas un signal banal et qu'une bonne façon de l'analyser consisterait à modéliser le processus qui lui a donné naissance [19].

Il est nécessaire de passer deux phases pour obtenir le modèle:

1. établir son modèle physique

2. chercher son modèle mathématique, ceci est équivalent à faire estimation du modèle.

La modélisation de la production de parole est un problème de boite noire, car on ne connaît que la sortie, l'entrée étant enfermée à l'intérieur des organes articulatoires humains.

Le modèle de la production de la parole est basé sur les caractéristiques physiques de l'organe articulatoire. Voici la description d'un modèle physique de la production [33].

La parole est formée par une excitation appliquée au conduit vocal, qui peut être considéré comme un tuyau sonore d'une longueur approximative de 17 cm. Si les cordes vocales sont tendues, comme dans le cas des voyelles, l'air sortant de la bouche est sous forme d'impulsions quasi périodiques. Les signaux ainsi formés sont appelés *signaux voisés*. Si les cordes vocales sont écartées, soit une turbulence quasi aléatoire d'air est produite dans le conduit vocal par diminution de sa section (comme le son /s/), soit le conduit est momentanément fermé complètement pour augmenter la pression, et réouvert instantanément, produisant une impulsion décroissante (comme dans le /p/). Les signaux ainsi formés sont appelés *signaux non voisés*. On peut considérer le conduit vocal comme un système variant dans le temps, qui impose ses propriétés de transfert selon la forme de l'excitation qui lui est appliquée. Si l'on admet que l'excitation et le conduit vocal sont relativement indépendants, la production de la parole peut se résumer dans le modèle représenté sur la figure 2.4 et proposé par Schafer. Dans ce modèle, la source d'excitation est soit un générateur d'impulsions, produisant des impulsions périodiquement avec une période dont l'inverse est appelé le *fondamental* du son, soit un générateur de nombres aléatoires qui simule à la fois les turbulences quasi-aléatoires et les impulsions décroissantes. L'une de ces sources est appliquée à l'entrée d'un filtre numérique *variant dans le temps*, qui simule le conduit vocal. Les coefficients de ce filtre varient approximativement chaque 10 ms pour indiquer une nouvelle configuration du conduit vocal. Un contrôle de gain entre la source et le système ajoute une flexibilité supplémentaire dans le réglage du niveau sonore. Pour les sons produits avec des cordes vocales tendues, le filtre numérique contient un préfiltre pour tenir compte de la forme des impulsions à durée limitée, qui ne sont pas des échantillons unités.

L'estimation du modèle est basée sur le modèle physique, et permet ainsi de fournir les paramètres calculés. C'est le modèle physique qui rend compte les données et leur attribue une signification.

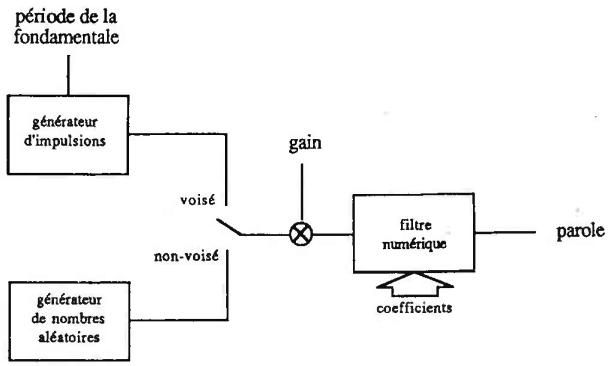


Figure 2.4. Modèle de production de la parole

La modélisation mathématique consiste à établir la fonction de transfert du système articulatoire. Le modèle ainsi créé doit respecter les deux conditions suivantes: [19]

1. Les paramètres extraits en vue de la reconnaissance devraient être attachés au sens du message émis, en même temps qu'être relativement insensibles à la personnalité du locuteur.

2. Le nombre de paramètres extraits doit cependant rester relativement faible pour ne pas engorger inutilement les étapes ultérieures du traitement.

Puisque le processus de l'établissement d'un modèle est un processus de description de certains phénomènes dans la nature, les modèles se distingueront par différentes hypothèses sur le phénomène. En tout cas, quoi que l'on fasse des hypothèses a priori, nous pouvons dire que la réalisation d'un modèle consiste à choisir un modèle spécifique parmi les modèles possibles qui sont équivalents à un système physique en mathématique.

Comme on a simplifié le signal d'excitation soit par une série d'impulsions dans le cas voisé, soit par un bruit blanc dans le cas non voisé, la question de la modélisation devient celle de la recherche des paramètres permettant de représenter le signal de sortie. C'est à dire la modélisation de la parole ici est justement une modélisation du processus qui lui a

donné naissance.

Le modèle mathématique ainsi créé selon son modèle physique est déjà largement appliqué sur un grand nombre de domaines concernant la traitement de parole.

2.4.3 Estimation du modèle

Un système de production de la parole est un système non-linéaire, dépendant du temps. C'est un système mesurable uniquement à la sortie; le signal de sortie du système reflète le caractère du système lui-même et également le caractère de la source de l'excitation. Le problème de non-linéarité peut être approché par la technique de linéarisation[45,60,32,76]. Et le problème de dépendance temporelle peut être résolu par l'analyse à court terme ($10ms \sim 30ms$)[69]. Quant à la séparation de deux caractères mélangés, nous pouvons employer la méthode d'analyse synchrone homomorphique, ou du délai de groupe de phase, ou même l'estimation directe de l'excitation[15]

Une telle identification du système n'est pas une identification complète, car le caractère d'un système se manifeste par deux caractères: l'amplitude du spectre, et la phase de chaque fréquence. Mais, si un système est de type à minimum de phase, on peut considérer uniquement l'amplitude ou la phase. Quelle que soit la partie connue, l'autre est automatiquement déterminée. Nous pouvons faire une hypothèse de phase minimale pour le système du conduit vocal[61], mais la source d'excitation n'est pas toujours un système à phase minimum[50]. Il y a beaucoup de méthodes asynchrones dans lesquelles on estime le système sans faire la séparation des différents caractères. Dans ce cas, on suppose un système à phase minimum afin de simplifier le travail.

On peut obtenir le caractère du système à partir de sa sortie, soit par la technique temporelle, ce qu'on l'appelle la *prédiction*, soit par la technique fréquentielle, ce qu'on l'appelle *mise en correspondance du spectre*.

Dans la partie suivante, nous notons $s(n)$ un signal observé, sortie d'un système H inconnu. La modélisation de H consistera à le simuler par un modèle que nous cherchons le plus adéquat possible au sens d'une norme à préciser.

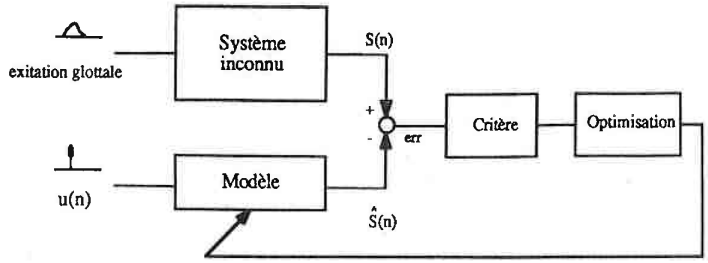


Figure 2.5. Prédiction par erreur de sortie

Prédiction

Un modèle prédictif est un modèle donnant en sortie une estimation $\hat{s}(n)$ du signal observé, la contrainte etant alors la minimisation d'un coût qui mesure l'erreur de prédiction $\epsilon_n = s(n) - \hat{s}(n)$ (Fig. 2.5). Un des critères de la prédiction est de rendre minimum ce coût au sens des moindres carrés dans toute la région où existe le signal:

$$\sum_{n=-\infty}^{\infty} |err(n)|^2 = \sum_{n=-\infty}^{\infty} |s(n) - \hat{s}(n)|^2 = min \tag{2.11}$$

Mis en correspondance du spectre

A partir du domaine fréquentiel, on peut choisir un autre critère d'estimation.

$$min_{\ p} \ \frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{S(\omega)}{\hat{S}(\omega)}|^2 d\omega \tag{2.12}$$

p étant l'ensemble des paramètres du système, $S(\omega)$ étant le spectre à référencer, $\hat{S}(\omega)$ le spectre estimé. Le résultat de l'intégrale donne une ressemblance moyenne pour toutes les fréquences.

Ces deux formules ne sont pas tout à fait équivalentes mais seulement dans certains cas ou sous certaines hypothèses simplificatives.

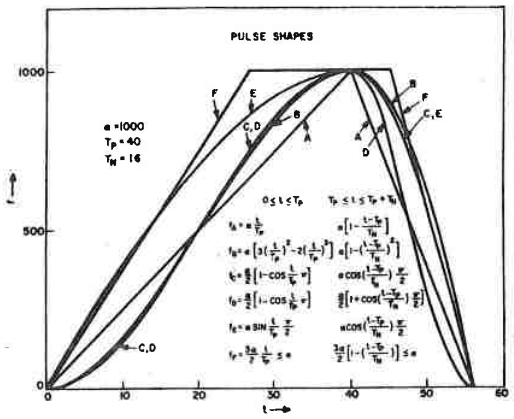


Figure 2.6. différentes formes de l'excitation

2.4.4 Excitation et Estimation

Pour un signal de parole, l'inconnue H est le filtre variable en fonction du temps dans la figure 2.4. L'estimation d'un modèle d'une parole est un travail extrêmement difficile, nous devons déterminer les paramètres internes d'un modèle à partir de sa sortie car l'entrée est complètement cachée. A partir de figure 2.5, nous voyons que la connaissance de la forme d'excitation correcte peut nous permettre d'avoir un modèle plus précis. La figure 2.6 montre quelques formes d'excitation qui ont été utilisées en synthèse. Mais le signal d'entrée est en réalité certainement loin d'être aussi simple.

La vibration des cordes est inertielle, et elle ne peut donc pas produire une impulsion idéale (impulsion de *Dirac*). Les figures 2.7 et 2.8 nous montrent une évolution des cordes vocales et des ondes produites correspondant à chaque position de l'ouverture et de la fermeture. En plus, les cordes vocales ne font non plus forcément une vibration par période à cause de différentes raisons. Dans une période, elles peuvent avoir plusieurs petites impulsions par exemple. Ces irrégularités n'ont pas une grande influence sur la perception de la parole, mais elles font changer les résultats de l'analyse automatique par ordinateur.

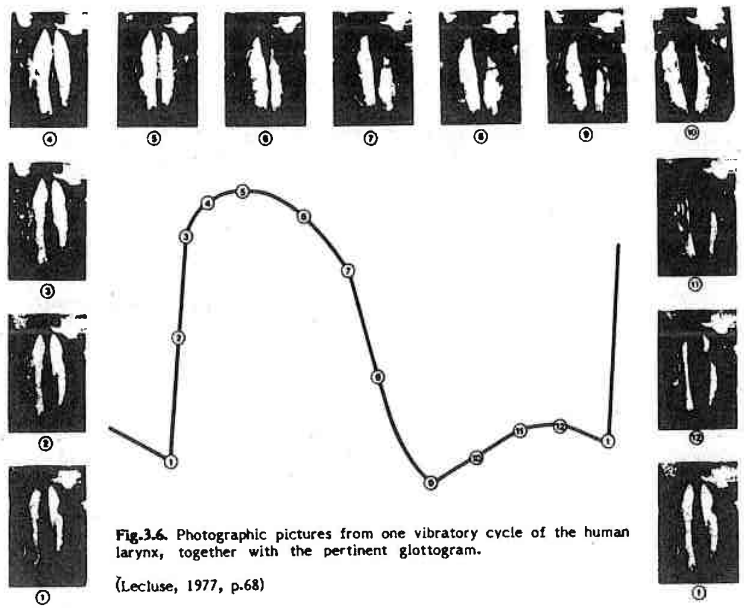


Fig.3.6. Photographic pictures from one vibratory cycle of the human larynx, together with the pertinent glottogram.

(Lecluse, 1977, p.68)

Figure 2.7. Photographie de l'évolution des cordes vocales

Si l'excitation du système reste impulsionnelle, alors les paramètres du modèle de la parole ne sont pas uniquement ceux qui reflètent les caractéristiques du conduit vocal, mais aussi ceux de la source d'excitation. Donc tous les pôles ou les zéros contenus dans le signal observé sont considérés comme les pôles ou zéros du modèle du système. Un modèle ainsi obtenu possède en réalité les caractéristiques du conduit propre et également celles de la source d'excitation. Le premier est important pour l'analyse de la parole, mais le dernier est inutile. Heureusement, les caractéristiques du conduit vocal jouent un rôle principal dans le spectre du modèle et il est possible de les distinguer par les positions et les largeurs de bande des formants.

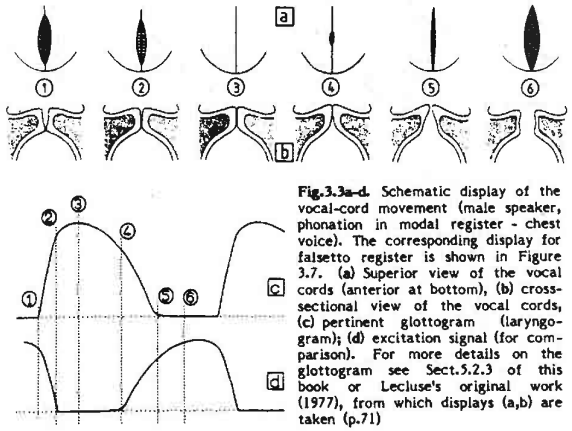


Fig.3.3a–d. Schematic display of the vocal-cord movement (male speaker, phonation in modal register - chest voice). The corresponding display for falsetto register is shown in Figure 3.7. (a) Superior view of the vocal cords (anterior at bottom), (b) cross-sectional view of the vocal cords, (c) pertinent glottogram (laryngogram); (d) excitation signal (for comparison). For more details on the glottogram see Sect.5.2.3 of this book or Lecluse's original work (1977), from which displays (a,b) are taken (p.71)

Figure 2.8. Schéma de l'évolution des cordes vocales

2.5 Trois formes du modèle

Selon les caractéristiques internes d'un système ou selon certaines hypothèses pour un système complexe, nous pouvons représenter un système avec différents modèles mathématiques.

D'après la définition d'une fonction de transfert, les caractéristiques d'un système doivent être représentées par le rapport entre la fonction de sortie et la fonction de l'entrée. C'est-à-dire $H(z) = \frac{S(z)}{U(z)}$. Son spectre est de forme $H(e^{j\omega}) = \frac{S(e^{j\omega})}{U(e^{j\omega})}$.

Koopmans [30] a indiqué que tous les spectres de puissance ayant une caractéristique continue peuvent être représentés, avec une approximation arbitraire, par un modèle rationnel ayant des ordres suffisamment grands p et q. Ici p est l'ordre de dénominateur et q l'ordre de numérateur. Donc on préfère représenter le système H(z) par le rapport de deux polynômes:

$$H(e^{j\omega}) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{q} b_i z^{-i}}{\sum_{i=0}^{p} a_i z^{-i}} \tag{2.13}$$

La relation temporelle entre l'entrée et la sortie s'écrit:

$$\sum_{i=0}^{p} a_i s(n-i) = \sum_{i=0}^{q} b_i u(n-i) \tag{2.14}$$

Nous prenons $q \leq p$, car l'énergie pour un spectre réel de parole diminue toujours lors de l'élévation de la fréquence.

Selon les valeurs de p et q, un modèle s'appelle AR (modèle Auto-Régressif) si q=0, MA (modèle à Moyenne Ajustée) si p=0, et ARMA (modèle Auto-Régressif et à Moyenne Ajustée) si $p > 0$ et $q > 0$.

2.6 le modèle AR

Le modèle AR est un modèle qui approche le système par un polynôme au dénominateur. Les zéros de ce polynôme sont les pôles du système, donc un modèle AR s'appelle également modèle tout-pôle. Puisque nous pouvons représenter approximativement un zéro du système par plusieurs pôles, il a été démontré que[64] tous les spectres AR, MA ou ARMA peuvent être représentés avec une précision arbitraire par un modèle tout pôle lorsque l'ordre du polynôme est infini. Un modèle tout-pôle est capable de caractériser la plupart des phénomènes physiques.

Le modèle AR a pour formule mathématique (2.15):

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{\sum_{i=0}^{\infty} a_i z^{-i}} \tag{2.15}$$

et sous forme temporelle:

$$s(n) + \sum_{i=1}^{\infty} a_i s(n-i) = u(n)G \tag{2.16}$$

ici G est le gain du système. On suppose que $\{u(n)\}$ est un processus centré, de variance $\sigma^2 = 1$, et ayant la caractéristique du bruit blanc. $\{s(n)\}$ est un processus stationnaire au second ordre, c'est à dire que sa moyenne est indépendante du temps et que sa covariance peut se mettre sous forme de fonction de corrélation.

Pour obtenir un modèle AR, il nous faut une méthode d'observation afin de refléter de façon optimale les caractéristiques de la parole. Rabiner [69] a énuméré quelques méthodes

parmi les plus courantes:

1. méthode de covariance
2. fonction d'autocorrélation
3. structure de treillis
4. filtrage inverse
5. estimation du spectre
6. méthode de vraisemblance

Ces méthodes diffèrent par leurs points de départ, cependant elles appartiennent à l'une des trois premières sortes.

2.6.1 méthode de covariance et autocorrélation

Bien évidemment, il n'est pas possible en réalité de représenter un spectre par un modèle tout pôle ayant un ordre infini. Nous voulons représenter un système avec moins de coefficients et assez de précision. Nous prenons une valeur suffisamment grande, notée p:

$$s(n) + \sum_{i=1}^{p} a_i s(n-i) = u(n)G \tag{2.17}$$

a_i dans la formule (2.16) n'est pas le même que celui dans (2.17). Mais nous prenons la même notation pour des raisons de simplicité typographique.

En reprenant l'hypothèse sur u(n), nous savons que son espérance est égale à zéro, soit $E\{u(n)\} = 0$. Donc, si a_i est choisi de façon à permettre que les termes $E\{s(n) + \sum_{i=1}^{p} a_i s(n-i)\}$ soient égaux à zéro et sa variance égale au minimum, les a_i seront des coefficients optimaux pour le système étudié:

$$E\{(s(n) + \sum_{i=1}^{p} a_i s(n-i) - 0)^2\} = min \tag{2.18}$$

La formule (2.18) permet de définir la notion de prédiction. On fait une prédiction de la valeur courante s(n) en regardant ses p précédentes valeurs. Supposons $\hat{s}(n)$ la vraie valeur, err(n) l'erreur de cette prédiction linéaire, nous avons:

$$err(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} a_i' s(n-i)$$

$$err(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad ici \quad a_i' = -a_i.$$

Nous devons choisir un ensemble de a_i qui permet de rendre minimale l'erreur de la prédiction au sens de l'espérance au carré. C'est bien la formule suivante qui équivaut à la forme (2.18).

$$E\{err^2(n)\} = min$$

Remplaçant l'espérance mathématique par la somme, nous avons:

$$\sum_{n=n_o}^{n_1} err^2(n) = min \tag{2.19}$$

Nous employons n_1, n_o pour indiquer la limite supérieure et inférieure de la somme.

Autocorrélation

Si $n_1 = \infty$, $n_o = -\infty$, $\sum_{n=n_o}^{n_1} err^2(n)$ devient $\sum_{n=-\infty}^{\infty} err^2(n)$. Ceci veut dire que le signal est considéré infini, et par conséquent son erreur de prédiction est également infinie.

Développant la formule ci-dessus, et la mettant en bonne forme, nous obtenons une nouvelle formule:

$$\sum_{n=-\infty}^{\infty} \{s(n) + \sum_{i=1}^{p} a_i s(n-1)\}^2 = min$$

$$\sum_{i=0}^{p}\sum_{j=0}^{p} a_i a_j \sum_{n=-\infty}^{\infty} s(n-i)s(n-i) = min \tag{2.20}$$

$$\sum_{n=-\infty}^{\infty} s(n-i)s(n-j) = \sum_{n=-\infty}^{\infty} s(n)s(n+|i-j|)$$

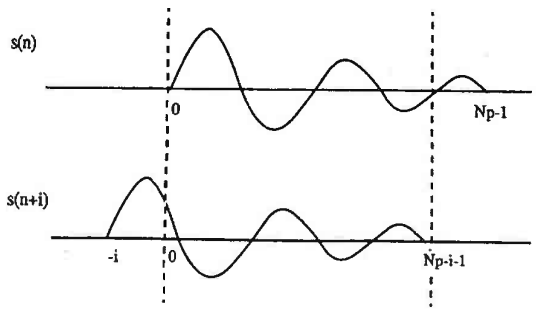


Figure 2.9. région pour le calcul d'autocorrélation

le terme $\sum_{n=-\infty}^{\infty} s(n-i)s(n-j)$ est une fonction d'autocorrélation d'arguments i et j. Puisque le signal existe en réalité entre 0 et $N_p - 1$, alors il nous faut limiter la portée du calcul pour obtenir les coefficients d'autocorrélation. La longueur de cette fenêtre a la même longueur que le signal, mais sa forme dépend de l'application. Les plus courantes sont des fenêtres rectangulaires et celle de Hamming.

Si nous notons $R(n)$ la fonction d'autocorrélation, alors:

$$R(|i-j|) = \sum_{n=-\infty}^{\infty} s(n)s(n+|i-j|)$$

introduisant les bornes existant du signal (fig.2.9), nous avons donc:

$$R(|i-j|) = \sum_{i=0}^{N_p-|i-j|-1} s(n)s(n+|i-j|)$$

N_p étant le nombre de points du signal à étudier. Avec $R(n)$ 2.20 devient:

$$\sum_{i=0}^{p}\sum_{j=0}^{p} a_i a_j R(|i-j|)$$

$$= \sum_{i=0}^{p} a_i^2 R(0) + \sum_{i=1, i\neq j}^{p} a_i a_j R(|i-j|)$$

La dérivation sur chaque a_i nous fournit l'ensemble des équations linéaires concernant a_i:

$$\frac{\partial}{\partial a_i}\{\sum_{i=0}^{p} a_i^2 R(0) + \sum_{i=1, i\neq j}^{p} a_i a_j R(|i-j|)\} = 0$$

$$\sum_{k=1}^{p} a_k R(|i-k|) = R(i), \quad 1 \leq i \leq p \tag{2.21}$$

La forme matricielle est donc:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ .. \\ .. \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ .. \\ .. \\ R(p) \end{bmatrix} \tag{2.22}$$

Cette matrice est de *Toeplitz*, nous pouvons résoudre les équations en employant l'algorithme rapide donné par Durbin[45].

Les coefficients ainsi obtenus sont appelés les coefficients de prédiction linéaire d'autocorrélation, parce que nous avons introduit la fonction d'autocorrélation dans la phase de la résolution des équations.

Covariance

Si le signal est considéré fini dans la région entre 0 et $N_p - 1$, le calcul des erreur se limite aussi à cette région. La formule (2.19) devient $\sum_{n=0}^{N_p-1} err^2(n)$. Donc nous avons:

$$\sum_{n=0}^{N_p-1} err^2(n) = \sum_{i=0}^{p}\sum_{j=0}^{p}\sum_{n=0}^{N_p-1} s(n-i)s(n-j)$$

$$= \sum_{i=0}^{p}\sum_{j=0}^{p} \Phi(i,j)$$

ici $\Phi(i,j) = \sum_{n=0}^{N_p-1} s(n-i)s(n-j)$. En tenant compte de ses bornes (Fig. 2.10), $\Phi(i,j)$ a la forme suivante:

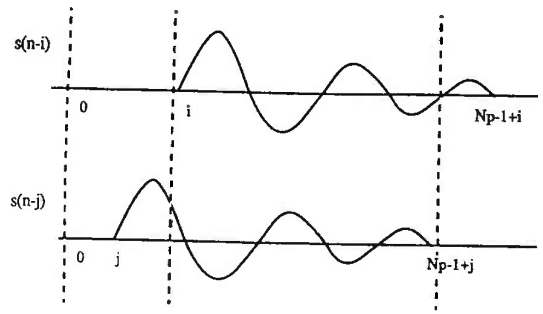


Figure 2.10. région pour le calcul de covariance

$$\Phi(i,j) = \begin{cases} \sum_{n=i}^{N_p-j+1} s(n-i)s(n-j) & j \le i \le N_p - j - 1 \\ \sum_{n=j}^{N_p-i+1} s(n-i)s(n-j) & i \le j \le N_p - i - 1 \end{cases}$$

En reprenant le même processus de traitement, nous avons:

$$\sum_{k=1}^{p} a_k \Phi(i,k) = \Phi(i,0), \quad i = 1,2,....p$$

et sous forme de matrice:

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \Phi(1,3) & ... & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \Phi(2,3) & ... & \Phi(2,p) \\ \Phi(3,1) & \Phi(3,2) & \Phi(3,3) & ... & \Phi(3,p) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ \Phi(p,1) & \Phi(p,2) & \Phi(p,3) & ... & \Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ .. \\ .. \\ a_p \end{bmatrix} = \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \\ \Phi(3,0) \\ .. \\ .. \\ \Phi(p,0) \end{bmatrix} \qquad (2.23)$$

Cette matrice n'est plus de Toeplitz, elle est seulement symétrique et demi-positive. La méthode de Cholesky fournit une solution rapide.

Gain du système

Nous avons supposé que l'entrée du système est une impulsion de forme dirac unitaire u(n). Le gain du système détermine l'amplitude de sortie (Fig.2.4). Une fois que les coefficients $\{a_i\}$ sont décidés, la différence entre les termes de la formule (2.17) dépend du gain G. Remarquons que $E\{u(n)^2\} = 1$, alors pour la méthode d'autocorrélation:

$$E\{u(n)^2\} = E\{e^2(n)G^2\} = G^2$$

$$= E\{(s(n) + \sum_{i=1}^{p} a_i s(n-i))^2\}$$

$$= R(0) - \sum_{i=1}^{p} a_i R(i)$$

Et pour la covariance:

$$G^2 = \Phi(0) - \sum_{i=1}^{p} a_i \Phi(i,0)$$

2.6.2 Structure de treillis

Les méthodes précédentes s'accomplissent en deux étapes principales: calcul de la matrice de corrélation, ensuite résolution des équations linéaires. Cependant, si on classe les erreurs de prédiction en deux sortes: erreur de prédiction progressive et erreur rétrograde, et qu'on applique à nouveau l'algorithme de Durbin, on arrivera à un autre algorithme qui est identique à la méthode de corrélation, mais différent par la structure.

$$\begin{cases} \epsilon^{(p)}(n) = s(n) + \sum_{i=1}^{p} a_i^{(p)} s(n-i) \\ b^{(p)}(n) = s(n-p) + \sum_{i=1}^{p} a_i^{(p)} s(n+i-p) \end{cases}$$

ici $\epsilon^{(p)}(n)$ étant l'erreur de la prédiction progressive de l'ordre p, $b^{(p)}(n)$ étant celle de l'erreur rétrograde de même ordre. Ces deux erreurs peuvent être représentées par les erreurs de la précédente prédiction, soit l'erreur de la prédiction de l'ordre (p-1).

$$\begin{cases} \epsilon^{(p)}(n) = \epsilon^{(p-1)}(n) + k_p b^{(p-1)}(n-1) \\ b^{(p)}(n) = b^{(p-1)}(n-1) + k_p \epsilon^{(p-1)}(n) \\ \epsilon^{(o)}(n) = b^{(o)}(n) = s(n) \end{cases}$$

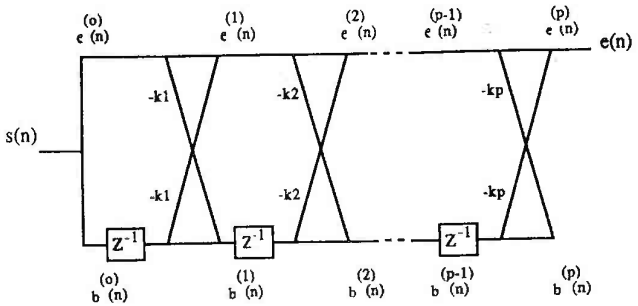


Figure 2.11. Structure de treillis

k_p est un coefficient qui relie les erreurs progressives et rétrogrades:

$$k_p = \frac{\sum_{n=0}^{N_p-1} \epsilon^{(p-1)}(n) b^{(p-1)}(n)}{\{\sum_{n=0}^{N_p-1} (\epsilon^{(p-1)}(n))^2 \sum_{n=0}^{N_p-1} (b^{(p-1)}(n))^2\}^{1/2}}$$

Ce même algorithme peut être réalisé par une structure qu'on l'appelle *structure de treillis* (Fig.2.11).

L'algorithme de Burg[7] est basé sur l'erreur de la prédiction progressive et rétrograde. L'avantage de cet algorithme est qu'il trouve toujours un filtre stable sans avoir besoin de fenêtre ni du pas intermédiaire du calcul de corrélation. Beaucoup d'études sont faites de la structure de treillis[34.74.35].

2.6.3 Estimation du spectre

Si $\hat{H}(z)$ est le modèle d'estimation du système, selon la relation entre z et ω, nous pouvons obtenir son spectre estimé:

$$\hat{H}(e^{j\omega}) = \frac{G}{\sum_{k=0}^{p} a_k z^{-k}}\Big|_{z=e^{j\omega}} = \frac{G}{\sum_{k=0}^{p} a_k e^{-j\omega k}} \qquad (2.24)$$

2.6.4 Effets de la fenêtre

Parmi ces trois méthodes principales, la fenêtre n'est introduite que pour l'autocorrélation. Comme expliqué dans la section précédente, il faut bien choisir une fenêtre pour qu'elle ne touche pas beaucoup les caractéristiques originales. Nous utilisons la fenêtre de *Hamming*.

2.6.5 Caractères de l'estimation réalisée

Par opposition à l'approche classique de l'analyse spectrale DFT, le spectre obtenu par modélisation présente les caractères suivants:

1. le spectre estimé est lissé et analytiquement décrit par l'équation du modèle. Ainsi donc, il est possible d'estimer les autres paramètres du système numériquement. Par exemple, les racines correspondent en général aux pics du spectre. A partir de cet ensemble des coefficients, il est également possible de calculer les surfaces de chaque section du conduit vocal, ...

2. tenant compte du fait que $E\{(\sum a_i s(n-i))^2\} = min$ esten équivalente, dans le domaine fréquentiel, à la formule $\int_{-\pi}^{\pi} |H(w) - \frac{G}{\sum_{i=0}^{p} a_i e^{-j\omega}}|^2 = min$, ce genre de modélisation est favorable pour bien mettre en correspondance les parties du pic dans le spectre à étudier. Pour les parties de vallée, AR donne pourtant une mauvaise correspondance. Richard[70] a décrit un algorithme utilisant le critère de *moindre absolu* (least squares value criteria). Cette estimation ne favorise pas l'ajustement des pics, et par conséquent, donne un résultat plus proche de la DFT.

2.6.6 Estimation de l'ordre du modèle AR

L'estimation de l'ordre du modèle fait partie du problème de l'identification. L'ordre d'un modèle est déterminé par les caractéristiques du système et il détermine la validité des coefficients estimés et le temps de calcul.

La méthode la plus simple pour prédire l'ordre d'un modèle AR est de calculer l'erreur en carré de la prédiction de chaque ordre, noté J_p:

$$J_p = E\{e^{(p)}(n)^2\} = \sum_{i=0}^{p} a_i^{(p)} R(i)$$

Pour chaque valeur J_p, on calcule $J_{p+1} - J_p$. Si J_p arrive à sa valeur minimale ou à une valeur constante, cela indique que l'erreur de la prédiction ne peut plus être réduite même si on continue d'augmenter l'ordre p (fig.2.12) [78]. L'ordre p dans ce cas doit être

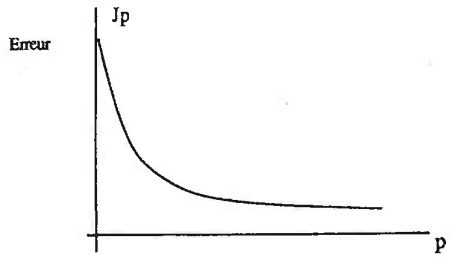


Figure 2.12. Estimation de l'ordre du AR

considéré comme un ordre de la prédiction du modèle.

En plus de cette façon mathématique, nous pouvons également choisir un ordre de la prédiction par des connaissances. Les pôles se répartissent en général tous les 1000 hertz. L'ordre de LPC est égal à deux fois le nombre de pôles possibles du conduit oral. Il faut aussi tenir compte de la présence d'un ou deux pôles réels. En pratique, l'ordre de la prédiction pour un modèle AR est égal approximativement à la fréquence d'échantillonnage divisée par 1000 plus deux à quatre pour tenir compte de la pente globale du spectre.

2.7 Le modèle MA

Le modèle MA est un modèle représenté uniquement par les zéros, c'est-à-dire $p = 0$ et $q \neq 0$ dans la formule générale (2.13), c'est pourquoi on l'appelle également modèle *tout zéro*. La valeur courante de la sortie du système est une combinaison linéaire de l'entrée sous la forme (2.7):

$$s(n) = \sum_{i=0}^{q} b_i u_{n-i}$$

Le problème est de déterminer les coefficients b_i, $0 \leq i \leq n$ de la suite de données $s(n)$ et $u(n)$. La résolution du modèle MA amenera à un problème non-linéaire à cause des coefficients inconnus $\{b_i\}$ attachés au signal de l'entrée. Ohsmann[59] a proposé un algorithme itératif et rapide en se basant sur la division de la matrice. Makhoul[45] estime sous-optimalement les paramètres du modèle MA par l'inversion du spectre du modèle AR.

A l'évidence, cette dernière méthode perd beaucoup d'informations concernant le spectre. D'abord, un modèle tout zéro doit être représenté en théorie par un modèle d'ordre infini, or nous n'avons qu'un ordre fini. L'inverse du spectre du modèle AR implique donc en réalité une tronquation.

A cause du problème de la non-linéairité, le modèle tout zéro n'a pas autant servi que celui tout pôle, surtout en traitement de la parole.

2.7.1 Prédiction de l'ordre du modèle MA

Supposons que le signal est un processus aléatoire contenant uniquement des zéros. Si le signal original contient à la fois des zéros et des pôles, nous pouvons transformer ce signal en un signal sans pôle en utilisant un prétraitement qui enlève tous les pôles. Nous supposons aussi que l'entrée (excitation) est de type bruit blanc. Nous avons les relations suivantes:

$$s(n) = b_0 u(n) + b_1 u(n-1) + \ldots + b_q u(n-q) \tag{2.25}$$

$$E\{u(n)\} = 0 \tag{2.26}$$

$$E\{u(n)u(n')\} = \delta(n-n') \tag{2.27}$$

Pour tester si la sortie s(n) du modèle correspond au processus à moyenne ajustée, nous devons observer si la corrélation des données conduit rapidement à zéro après un petit nombre de terme q. Nous savons que $E\{s(n)s(n-i-1)\} = 0$, pour $i \geq q$, et nous définissons

$$R(q+1) \triangleq \frac{1}{N_p - q - 1} \sum_{n=q+1}^{N_p-1} s(n)s(n-(q+1))$$

donc la condition nécessaire et suffisante pour que l'ordre du processus soit q est que $E\{R(q+k)\}$ soit égal à zéro dans certaine espace pour à tous $k \in 1\ldots$.

Le signal de parole est un mélange des caractéristiques de la source d'excitation et du conduit articulatoire, et contient des zéros et des pôles. Pour un tel signal, ni un modèle tout zéro, ni celui tout pôle ne peuvent être une représentation convenable. Nous devons

avoir recours à la combinaison du modèle AR et MA. Cela nous conduit au modèle ARMA.

2.8 Le modèle ARMA

Le modèle AR fournit un résultat assez précis pour les voyelles qui se caractérisent par un modèle tout pôle. Lorsque les zéros contenus dans un signal de parole ne peuvent pas en réalité être négligés, (nasales par exemple), le résultat fourni par ce modèle sera une estimation avec déviation. La conséquence se manifeste sur les aspects suivants:

- le spectre au voisinage des zéros n'est plus correct.
- la fréquence du formant près d'un zéro est déplacée, sa largeur de bande est élargie.

Les zéros du signal de parole sont produits dans les sections différentes de l'articulation ou de la propagation. Atal[2] nous donne quelques raisons d'apparition de ces zéros supplémentaires. Les zéros peuvent apparaître:

- si le conduit nasal est couplé au conduit vocal principal par l'ouverture du voile souple. C'est le cas des consonnes nasales, voyelles nasalisées et voyelles nasales.
- si la source d'excitation ne se situe pas à la glotte mais à l'intérieur du conduit vocal. C'est le cas de certaines consonnes /l/ /z/ ...
- les caractéristiques de transmission de l'enregistrement peuvent aussi facilement introduire des zéros.

Pour un tel signal contenant des pôles et également des zéros, il est très adapté de prendre le modèle ARMA pour l'analyser. Un modèle ARMA n'est pas une simple addition du modèle AR et du modèle MA. AR et MA interviennent et se complètent l'un par l'autre, ce qui permet au modèle ARMA de réaliser une bonne correspondance avec le signal. Comme le problème de non-linéarité existe pour la résolution du modèle MA, il existe aussi pour le modèle ARMA. Nous savons que ce problème est causé par la non-connaissance de u(n). Par conséquent, il nous faut d'abord transformer la non-linéarité en linéarité. Les méthodes sont nombreuses: algorithme d'itération, méthode par spectre, prédiction en ordre élevé, méthode de synchronisation, etc...

2.8.1 Relation entre l'entrée et sortie d'un système causal

Un modèle ARMA concerne le signal d'entrée et de sortie. Il est important de bien comprendre les caractéristiques de corrélation dans un système causal. Cela est important

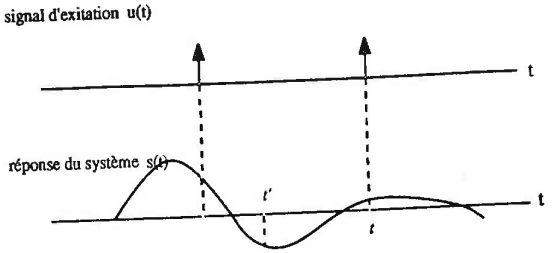


Figure 2.13. réponse d'un système causal

pour résoudre le problème de non-linéarité en phase d'identification des paramètres.

Si un système a eu une excitation à un certain moment, alors la réponse totale de ce système doit être la réponse à cette excitation en addition, la somme des réponses de toutes les excitations précédentes (fig.2.13).

Puisque nous avons fait l'hypothèse de causalité du système, la réponse du système avant le moment t ne contient pas de réponse aux excitations après le moment t. Nous avons la relation: $E\{s(t')u(t)\} = 0 \;\; si\; t' < t$. C'est-à-dire que la réponse du système avant le moment t n'est pas corrélée avec l'excitation du système après le moment t.

2.8.2 Principe du modèle ARMA

Reprenant la formule (2.14), nous avons la forme suivante:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + \sum_{i=0}^{q} b_i u(n-i) \tag{2.28}$$

La partie droite de la formule est équivalente à une estimation de la valeur courante de $s(n)$ par ses p valeurs passées et les q valeurs passées de l'entrée plus l'entrée courante.

Selon la caractéristique de causalité et l'hypothèse sur le signal d'entrée, nous avons:

$$E\{u(t)\} = 0$$

$$E\{u(t)u(t')\} = \begin{cases} 0 & t \neq t' \\ 1 & t = t' \end{cases}$$

En multipliant les deux membres de la relation (2.28) par s(n-k) et en prenant l'espérance mathématique[73], on obtient l'ensemble des équations suivantes:

$$\Phi_{ss}(k) = -\sum_{i=1}^{p} a_i \Phi_{ss}(k-i) + \sum_{i=0}^{q} b_i \Phi_{us}(k-i) \quad k \in 0,1,2...q \tag{2.29}$$

ici

$$\Phi_{xy}(k) = \sum_{n=n_o}^{n_1} x(n)y(n-k)$$

Si nous considérons que $\Phi_{us}(k) = 0$ et compte tenu du caractère causal, pour $k > 0$. nous pouvons avoir une formule différente de celle de (2.29).

$$\Phi_{ss}(k) = -\sum_{k=1}^{p} a_k \Phi_{ss}(k-i) \quad k \in q+1, q+2... \tag{2.30}$$

Cette formule est très importante. Elle nous revèle ceci: si l'ordre de zéro du système est inférieur à q (celui du modèle), nous pouvons construire des calculs sur $\{a_i\}$ à partir de la fonction de corrélation après q. Cette relation est illustrée figure2.14.

Cette équation permet théoriquement de déterminer les coefficients de la partie autorégressive du modèle, mais la stabilité n'est pas assurée car la matrice des équations est symétrique mais pas de Toeplitz. J.F.Serignat[73] a proposé une nouvelle méthode qui peut rendre la matrice de type Toeplitz, en calculant l'autocorrélation à la place de la covariance: En faisant la comparaison de la formule

$$\Phi_{ss}(k) = -\sum_{i=1}^{p} a_i \Phi_{ss}(k-i) \quad k \in q+1, q+2...$$

avec celle de

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) \quad n \in 0,1,2...$$

nous voyons que $\Phi_{ss}(k)$ est symboliquement équivalent à $s(n)$, bien que le premier soit la covariance du dernier. Considérant $\Phi_{ss}(k)$ comme un nouveau signal $s(n)$, nous pouvons

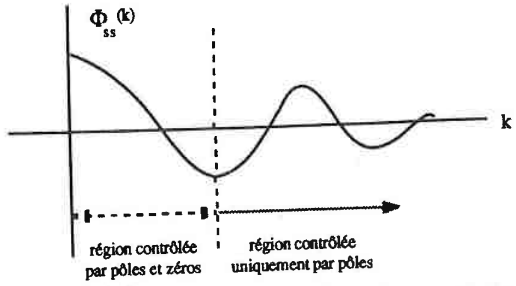


Figure 2.14. fonction de covariance d'un système de pôle-zéro

obtenir les coefficients stables $\{a_i\}$ directement par la méthode d'auto-corrélation. Une fois que les $\{a_i\}$ sont déterminés, les $\{b_i\}$ peuvent être obtenus à partir l'équation suivante:

$$\sum_{j=0}^{p}\sum_{i=0}^{p} a_j a_i \Phi_{ss}(k+j-i) = \sum_{i=0}^{q-k} b_i b_{i+k}, \quad k = 0,1,2...q$$

Dans la phase de déduction des formules, on n'a pas employé clairement la théorie des moindres carrés. Mais en réalité, elle est équivalente au processus suivant:

$$err(k) = \Phi_{ss}(k) - \hat{\Phi}_{ss}(k) \tag{2.31}$$

$$\hat{\Phi}_{ss} = -\sum_{i=1}^{p} a_i \Phi_{ss}(k-i) \quad k \in q+1, q+2...$$

Nous choisissons un ensemble de $\{a_i\}$ de façon que l'erreur $err(k)$ soit de valeur minimale dans la région où existent uniquement les pôles pour $\Phi_{ss}(k)$.

Cadzow[8] ajoute une fonction de pondération quand il emploie un critère similaire à (2.31):

$$f(a_1, a_2, ..., a_p) = \sum_{k=q+1}^{N_p-1} w(k)|err(k)|^2 \tag{2.32}$$

$w(n)$ est une fenêtre non négative. On choisit une telle fenêtre monotone et non croissante (par exemple: $w(k) \geq w(k+1)$) afin de refléter l'augmentation prévue de l'$err(k)$

lors de la croissance de k.

Selon les relations de la transformée de Fourier, si un signal possède une transformation:

$$s(n) \rightleftharpoons S(\omega)$$

alors sa fonction de covariance en possèdle aussi:

$$\Phi_{ss}(n) = \sum_k s(k)s(k-n) = |S(\omega)|^2.$$

Par rapport à la discussion sur le caractère de l'estimation réalisée d'un modèle, les méthodes proposées sont une approximation sur le spectre de puissance (spectre en carré). A cause de l'action de carré, l'accentuation de l'approximation au voisinage du pic est encore plus remarquable.

Chan[10] résume ainsi sa méthode: "l'estimation spectrale de ARMA en utilisant la transformation de Fourier sur les résidus, évite la partie MA du modèle". Ces méthodes utilisent les $N_p - q$ ($N_p \gg q$) points d'information accessibles de la fonction d'autocorrélation pour obtenir un ensemble de coefficients optimaux $\{a_i\}$. (cf: fig.2.14).

Les algorithmes précédents fournissent un processus de résolution du modèle ARMA en échappant au problème de u(n). Le calcul de la partie autorégressive et celui de la partie retrograde sont obtenus tour à tour. Nous pouvons aussi obtenir un modèle ARMA par prédiction de u(n) et calculer simultanément les zéros et pôles. Faire une prédiction sur l'entrée inconnue u(n) consiste à filtrer les pôles et les zéros appartenant au système, le signal filtré peuvent être considéré comme une entrée du système ou l'excitation du système.

2.8.3 Méthode ITIF

Konvalinka[29] propose un algorithme itératif (ITIF) pour calculer simultanément les coefficients $\{a_i\}$, $\{b_i\}$ (fig.2.15). Puisque la connaissance sur $u(n)$ est obtenue progressivement, le raffinement sur B(z) l'est également (formules (2.33), (2.34), (2.35)).

$$H(z) \approx \frac{S(z)}{U^l(z)} = \frac{B_q^l}{A_p^l} \tag{2.33}$$

$$U^l(z) = \frac{S(z)}{B_q^{l-1}(z)} C_r^l(z) \qquad avec B_q^0 \equiv 1 \tag{2.34}$$

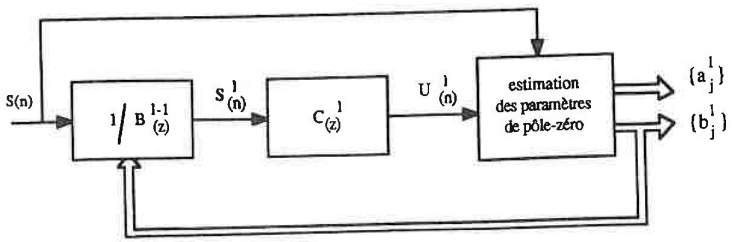


Figure 2.15. estimation itérative du ARMA

$$err^l(z) = S(z)A_p^l(z) - U^l(z)B_q^l(z) \tag{2.35}$$

Le traitement pour filtrer les zéros et les pôles se fait en deux parties: d'abord filtrer les zéros qui sont déjà prédits, ensuite filtrer les pôles du système dans le signal et les zéros résiduels. La deuxième partie est faite par un modèle tout pôle, son ordre de prédiction $r = p + o$ impliquant p pôles et q pôles qui représentent l'effet des zéros résiduels. Bien évidemment $o \gg q$. Lorsqu'on calcule $\{a_i\}$, $\{b_i\}$ par itération, c'est-à-dire $u(n)$. l'effet de l'accumulation de l'itération nous permet de prendre un petit ordre pour o. Dans l'algorithme ITIF, $o = q$. Khouri[28] fournit un algorithme modifié de *ITIF*: au lieu de calculer dans chaque boucle les $\{a_i\}$, il propose de fixer $\{a_i\}$ pour toutes les itérations car $\{a_i\}$ peut être déterminé séparément par le principe précédent (fig.2.14).

La méthode itérative nous permet d'approcher un modèle précis, mais le contrôle de l'itération est lui même un problème complexe, le choix de l'ordre de p et q au départ, la vitesse de convergence déterminent sensiblement la quantité de calcul.

En plus de la méthode itérative, il y a aussi des algorithmes non itératifs donc plus efficaces.

2.8.4 Méthode de décomposition

Song[76] a proposé un algorithme de décomposition d'un modèle tout pôle ayant un ordre élevé.

En réalité, un modèle tout pôle d'ordre fini peut représenter un spectre donné avec une précision suffisante si l'ordre est assez grand. Dans la plupart des cas, un modèle tout pôle de l'ordre 40-50 est largement suffisant pour approcher un spectre complexe[75], pôle zéro par exemple. Ce modèle tout pôle ayant un ordre fini sera référencé comme un modèle de grand ordre et sera considéré comme un modèle universel.

Une décomposition d'un modèle de grand ordre consiste à décomposer le modèle universel en une partie pôle $A(z)$ et une partie zéro $B(z)$. Nous notons le modèle tout pôle de grand ordre par $C_r(z)$:

$$C_r(z) = \sum_{k=0}^{r} C_r(k) z^{-k} \approx \frac{B(z)}{A(z)} \tag{2.36}$$

ici r étant son ordre (40-50).

La décomposition se détermine par une mimimisation de la fonction:

$$min_{A,B} \frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{A(w)}{C_r(w)B(w)}|^2 dw \tag{2.37}$$

remarquons, que si $A(w)$, $B(w)$, $C_r(w)$ ont des racines à l'intérieur du cercle unité du plan z, la procédure de minimisation (2.37) est équivalente à la minimisation de la mesure d_m donnée ci-dessous:

$$d_m = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{A(w)}{C_r(w)B(w)} - 1|^2 dw \tag{2.38}$$

la minimisation de (2.38) donnera des paramètres optimaux de pôle et de zéro, mais cette mesure est hautement non-linéaire. Il est possible d'éviter ce genre de problème en prenant une mesure similaire:

$$d_1 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |A(w) - C_r(w)B(w)|^2 dw \tag{2.39}$$

La mesure (2.39) ne contient que des opérations linéaires, pourtant les paramètres de $A(w)$, $B(w)$ obtenus ne sont plus optimaux, parce que nous avons multiplié les deux côtés de (2.39) par une modification de $C_r(w)B(w)$. d_m est pondéré fréquenciellement par $|C_r(w)B(w)|^2$. Puisque $C_r(w)$ (le spectre inverse du signal) est grand dans les vallées du spectre et petit aux formants, et que $B(w)$ est grand aux formants, petit dans les vallées mais moins significatif par rapport à l'effet de $C_r(w)$, $|C_r(w)B(w)|$ sera petit aux formants et grand aux vallées. Par conséquent, le modèle pôle-zéro obtenu selon la mesure (2.39)

fournit une meilleure correspondance aux vallées qu'aux formants.

Prenant la relation temporelle et fréquentielle: $A(z) = \sum_{i=0}^{p} a_i z^{-i}$, $B(z) = \sum_{i=0}^{p} b_i z^{-i}$, $C(z) = \sum_{i=0}^{p} C_{r_i} z^{-i}$, et notons $D(w) = A(w) - C_r(w)B(w)$. Donc:

$$D(w) = \sum_{k=0}^{p} a(k)\epsilon^{-jwk} - \sum_{k=0}^{r} C_r(k)\epsilon^{-jwk} \sum_{k=0}^{q} b(k)\epsilon^{-jwk}$$

$$d(n) = a(n)\{1(n) - 1(n-p)\} - \sum_{i=0}^{r+q} C_r(i)b(n-i)$$

selon la relation ci-dessous:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |D(w)|^2 dw = \sum_{n=-\infty}^{\infty} d^2(n)$$

Nous avons un ensemble d'équations qui permet de trouver les a_i et b_i:

$$a(i) - \sum_{k=0}^{i} C_r(k)b(i-k) = 0, \quad i = 1, 2, \ldots p$$

$$-\sum_{k=0}^{p} a(k)C_r(k-i) + \sum_{k=0}^{q} R_c(i-k)b(k) = 0, \quad i = 1, 2, \ldots q$$

$$ici \quad R_c(i) = \sum_{j=0}^{r-i} C_r(j)C_r(j+i)$$

La méthode précédente accentue sur les vallées du spectre. Ceci peut être très utile dans certains cas (détecter les fréquences de vallées quand l'effet des bruits est négligeable, par exemple). Mais dans la plupart des cas, on préfère que la correspondance soit meilleure sur les formants. Les raisons sont les suivantes:

- les formants sont à des fréquences où se concentre l'énergie, les vallées ne représentant que des zones de faible énergie. Donc si les formants sont peu sensibles aux bruits, les vallées le sont beaucoup.
- une parole est perçue par ses formants. Les oreilles humaines sont sensibles à position.

En remarquant que le filtrage d'un signal par son filtre inverse donne un spectre plat: $|C_r(w)S(w)|^2 = 1$, si on multiplie les deux côtés de (2.38) par $|B(w)C_r(w)S(w)|^2$, on obtient une mesure pondérée uniquement par $|B(w)|^2$:

$$d_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |A(w)S(w) - B(w)C_r(w)S(w)|^2 dw \qquad (2.40)$$

La mesure (2.40)' permet une meilleure correspondance du spectre aux formants.

Comme pour le traitement précédent, on a:

$$d_2(n) = \sum_{i=0}^{p} a(i)s(n-i) - \sum_{j=0}^{q} b(j)u(n-j) \qquad (2.41)$$

l'excitation $u(n)$ est le résidu du filtrage inverse du signal $s(n)$:

$$u(n) = \sum_{i=0}^{r} C_r(i)s(n-i) \qquad (2.42)$$

La mesure (2.41) est réécrite sous forme (2.43) à une signification claire:

$$err(n) = s(n) - (\sum_{i=1}^{p} a'(i)s(n-i) - \sum_{j=0}^{q} b'(j)u(n-j)) \quad ici\ a' = -a \ \ et \ \ b' = -b. \qquad (2.43)$$

C'est une forme de la prédiction sur le signal $s(n)$: $err(n) = s(n) - \hat{s}(n)$.

Puisque la minimisation de (2.40) est équivalente à celle de (2.41), en faisant:

$$\frac{\partial}{\partial a(i)} d_2 = 0 \qquad i = 1,2,...p$$

$$\frac{\partial}{\partial b(i)} d_2 = 0 \qquad i = 1,2,...q$$

nous avons:

$$\sum_{i=1}^{p} a(i)R_{ss}(k,i) - \sum_{i=1}^{q} b(i)R_{su}(k,i) = -R_{ss}(k,0) + R_{su}(k,0) \qquad k = 1,2,...p \qquad (2.44)$$

$$-\sum_{i=1}^{p} a(i)R_{us}(k,i) + \sum_{i=1}^{q} b(i)R_{uu}(k,i) = -R_{uu}(k,0) + R_{su}(k,0) \quad k = 1,2,...q \qquad (2.45)$$

ici R_{uu} est l'autocorrélation du signal $s(n)$, R_{uu} l'autocorrélation de l'excitation $u(n)$ estimée, R_{su} et R_{us} la corrélation de $s(n)$ et $u(n)$:

$$R_{ss}(i,j) = \sum_{k=-\infty}^{\infty} s(k-i)s(k-j) = R_s(|j-i|) \qquad (2.46)$$

$$R_{uu}(i,j) = \sum_{k=-\infty}^{\infty} u(k-i)u(k-j) \qquad (2.47)$$

$$R_{su}(i,j) = R_{us}(j,i) = \sum_{k=-\infty}^{\infty} s(k-i)u(k-j) \qquad (2.48)$$

(2.47) et (2.48) peuvent se représenter uniquement par les coefficients du filtre de grand ordre C_r et l'autocorrélation du signal R_s dans (2.46):

$$R_{uu}(i,j) = \sum_{k=0}^{r}\sum_{l=0}^{r} C_r(k)C_r(l)R_s(|(j-i)+(n-m)|) \qquad (2.49)$$

$$R_{su}(i,j) = R_{us}(j,i) = \sum_{k=0}^{r} C_r(k)R_s(|i-(j+m)|) \qquad (2.50)$$

les (2.44, 2.45) construisent les (p+q) équations qui permettent de trouver simultanémant les coefficients $\{a_i\}$ et $\{b_i\}$.

2.8.5 Méthode de phase dérivée

Si nous considérons qu'un système de parole est de phase minimale, l'identification du système à partir de son spectre d'amplitude est complètement équivalente à l'identification à partir de sa phase.

Yegnanarayana[87] a proposé une méthode utilisant une dérivation négative sur le spectre de phase (*NDPS*: Negative Derivative of the Phase Spectrum), c'est-à-dire par les caractéristiques du retard de groupe de la phase. Cette nouvelle méthode enlève d'abord l'effet de l'excitation par le cepstre, et sépare ensuite les pôles et les zéros par leur signe (positif ou négatif).

Supposons qu'un conduit vocal puisse être se représenté par p résonateurs et q anti-résonateurs, alors sa fonction de transfert sous une autre forme est:

$$H(w) = G\frac{\prod_{n=1}^{q}(b_n - jw)(b_n^* - jw)}{\prod_{m=1}^{p}(a_m - jw)(a_m^* - jw)} \qquad (2.51)$$

G étant le gain, a_i b_i les fréquences des résonateurs et anti-résonateurs, * dénote la conjugaison complexe.

Le retard du groupe $g(w)$ qui est défini par la dérivation sur $-arg[H(w)]$ est

$$g(w) = \sum_{m=1}^{p} \frac{-2(|a_m^2 + w^2|)Re\{a_m\}}{(|a_m|^2 - w^2)^2 + (2wRe\{a_m\})^2} + \sum_{n=1}^{q} \frac{2(|b_n^2 + w^2|)Re\{n_n\}}{(|b_n|^2 - w^2)^2 + (2wRe\{b_n\})^2}$$

$$= \sum_{m=1}^{p} g_m^p(w) + \sum_{n=1}^{q} g_n^z(w) \qquad (2.52)$$

remarquons que g_m^p a un pic fort au voisinage de la fréquence $w_m^p = \sqrt{(Im\{a_m\})^2 - (Re\{a_m\})^2}$ si $(Re\{a_m\})^2 \ll (Im\{a_m\})^2$, et g_n^z a une vallée profonde au voisinage de la fréquence $w_n^z = \sqrt{(Im\{b_n\})^2 - (Re\{b_n\})^2}$ si $(Re\{b_n\})^2 \ll (Im\{b_n\})^2$. La hauteur de g_m^p à w_m^p et la profondeur de g_n^z à w_n^z sont données ci-dessous:

$$g_m^p(w_m^p) = -\frac{1}{Re\{a_m\}}$$

$$g_n^z(w_n^z) = \frac{1}{Re\{b_m\}}$$

Les valeurs $-\frac{1}{Re\{a_m\}}$ et $\frac{1}{Re\{b_m\}}$ sont positive et négative respectivement. Selon la direction d'un pic, il est facile d'extraire les pôles et les zéros, donc les fréquences fondamentales du système.

Mikami[53] indique que, pour réduire l'effet des bruits sur le spectre, il vaut mieux utiliser l'enveloppe du spectre logarithmique au lieu de sa moyenne.

Nous avons présenté quelques méthodes représentatives. Il y en a encore beaucoup d'autres. La première classe[55,63] consiste en des méthodes qui éliminent dans la mesure possible l'influence de la fréquence fondamendale. La deuxième classe [80] consiste en des méthodes qui utilisent la technique des multipulse. Et la troisième [28,39] consiste en des méthodes qui estiment le spectre par fonctions rationnelles. Ces méthodes se fondent sur certains phénomènes physiques. Le seul point commun est qu'elles doivent fixer l'ordre du modèle a priori. Morikawa[56] a proposé une méthode adaptive qui permet d'estimer automatiquement à la fois les coefficients du modèle et son ordre.

2.9 Modèle d'analyse de régions particulières en fréquence

Dans la partie précédente, nous avons parlé de quelques techniques pour construire un modèle. La région d'analyse de ces méthodes est automatiquement fixée, l'intervalle étant entre 0Hz et $f_s/2$ (f_s est la fréquence d'échantillonnage). Pour pouvoir représenter les caractéristiques des consonnes, la fréquence d'échantillonnage doit être assez grande. On prend en général 16 kHz.

Du point de vue de la complexité de parole naturelle, le problème du choix de l'ordre de prédiction est crucial. Le choix est guidé par le nombre de pôles et zéros. Leur répartition devient plus complexe au fur et à mesure de l'augmentation de la fréquence. Analyser les consonnes exige l'élargissement de la région fréquentielle. Dans cet intervalle large, si nous fixons l'ordre de LPC à celui nécessaire pour les voyelles orales, nous ne pouvons plus faire une bonne prédiction des nasales. Et si nous choisissons un ordre optimum pour les voyelles nasales, nous obtiendrons un résultat de prédiction difficile à interpréter à cause de ce grand ordre. Il y a des algorithmes[10,11] qui permettent d'estimer l'ordre d'un signal. Nous devons savoir leur limite. Il n'y a pas raison de croire que l'estimation sur l'ordre serait égale au véritable ordre du signal, parce que la séquence des données accessibles est de longueur finie[10]. Peut-on trouver une technique avec un ordre à l'avance, qui convienent à la fois à l'analyse des voyelles orales et aux voyelles nasales?

La répartion des premiers formants est stable dans certaines régions fréquentielles, donc il est possible de déterminer l'ordre de LPC si l'intervalle de l'analyse est selectionné.

2.9.1 Région d'analyse pour les voyelles

Fujimura et Lindgvist[14] ont fait une expérience concernant le conduit articulatoire par la méthode analogique. Leur expérience nous revèle deux points importants de l'acoustique d'une voyelle nasale ou orale:

- il est possible de prédire le nombre des pôles et des zéros dans une région comprise entre de 0Hz à 3000Hz environ.
- le spectre d'une voyelle n'a pas toujours une forme régulière en haute fréquence. Il y existe des zéros dont on ne connait pas l'origine.

Par ailleurs, Lonchamp[41] a fait une synthèse des données antérieures concernant les voyelles nasales. A partir de ces données, nous remarquons qu'il existe généralement deux zéros et cinq pôles pour une voyelle nasale française dans la région de 0 à 3000Hz. Liénard[36] a remarqué que les effets du couplage se font surtout sentir entre 300Hz et 2500Hz et que le spectre est peu affecté dans la zone des 3000-4000Hz.

Plus l'intervalle d'analyse est grand, plus les caractères **supplémentaires sont nombreux,** Il existe par exemple des zéros générés par les impulsions glottales le long d'une ligne parallèle à l'axe imaginaire[50]. L'estimation des paramètres du modèle dans une région de haute fréquence est facilement faussée.

Par conséquent, faisant une restriction sur l'intervalle d'analyse, nous pouvons passer à côté de la difficulté de la détermination des ordres et bien sûr gagner du temps du calcul. Les ordres du modèle peuvent être décidés par des connaissances locales.

Makhoul[46] a donnée un algorithme de *"prediction linéaire sélective"* pour un modèle tout pôle. Nous allons, ci dessous, introduire cette idée dans le modèle pôle zéro[85].

2.9.2 Modèle pôle zéro sélectif

La méthode de LP spectre à une région sélective est décrite ci-dessous: [45]

Soit un spectre de puissance $P(w)$, $0 \leq w \leq w_b$. On désire mettre en correspondance ce spectre dans un intervalle $w_a \leq w \leq w_b$ avec un spectre $\hat{P}(w)$ tout pôle. On obtient $\hat{P}(w)$ par une transformée fréquentielle de $w_a \to 0$ et $w_b \to \pi$. Ensuite on calcule l'autocorrélation $R_s(n)$ par transformée Fourier inverse, cette $R_s(n)$ étant l'autocorrélation des composants fréquentiels du signal qui se trouvent dans la région selectionnée. Les coefficients sont obtenus par l'équation (2.21).

Nous voyons que la LP sélective permet d'estimer l'autocorrélation dans la région désirée. Pour un modèle de pôle zéro, il faut aussi se baser sur la fonction d'autocorrélation.

Nous prenons comme départ la méthode de Song qui a été décrite auparavant. Supposons que $s'(n)$ est un signal dont les composantes se limitent dans une région fréquentielle sélectionnée, que $C'_r(n)$ est le filtre inverse du système dans la même région, et que $u'(n)$ est l'estimation de l'entrée aussi dans cette région. $C'_r(n)$ peut être résolu même si les fréquences du signal original $s(n)$ ne sont pas restreintes dans cette région selectionnée, car

$$u'(n) = \sum_{k=1}^{r} C'_r(k)x'(k-n)$$

En remplaçant $u(n)$, $s(n)$ et $C_r(n)$ dans la formule (2.49) par $u'(n)$, $s'(n)$ et $C'_r(n)$ respectivement, nous avons

$$R_{u'u'}(i,j) = \sum_{k=0}^{r}\sum_{l=0}^{r} C'_r(k)C'_r(l)R_{s'}(|(j-i)+(n-m)|)$$

ici $R_{s'}(n)$ est la fonction d'autocorrélation du signal $s'(n)$. Comme le fait Makhoul, $R_{s'}(n)$ peut être obtenu de $R'_s(n)$. $R'_s(n)$ signifie l'autocorrélation *"mapped"* de $R_s(n)$. Donc nous avons:

$$R_{s'} \longrightarrow R'_s$$

Puisque les termes $R_{s's'}$, $R_{s'u'}$, $R_{u's'}$ et $R_{u'u'}$ peuvent se représenter par C'_r et R'_s, nous pouvons calculer les coefficients $\{a_i\}$ et $\{b_i\}$ par les équations(2.44, 2.45).

En résumé, l'algorithme est le suivant:

1. définir un intervalle d'analyse $w_a \leq w \leq w_b$.
2. calculer le spectre puissance $P'(w)$ dans cette région en transformant $w_a \leq w \leq w_b$ en $0 \leq w \leq \pi$.
3. calculer l'autocorrélation R'_s par transformée de Fourier inverse de $P'(w)$.
4. estimer le filtre inverse C'_r d'ordre élevé r avec R'_s.
5. calculer $R_{u's'}$, $R_{s'u'}$ et $R_{u'u'}$.
6. estimer les a_i et b_i par les équations (2.44, 2.45).

2.9.3 Complexité

La complexité de cet algorithme se basant sur [76] est déterminée par les ordres p, q et r dont les choix dépendent de l'intervalle de travail.

Cel algorithme contient cinq parties de calcul:

1. deux calculs d'une DFT pour l'estimation de R'_x
2. une LP avec d'ordre élevé pour C'_r
3. $(r+1)^2(q+2)(q+1)/2$ multiplications pour le terme $R_{u'u'}$, soit de l'ordre de $(r*q)^2$ multiplications
4. $(q+2)(q+1)/2$ multiplications pour le terme $R_{u's'}$

5. résolution de la matrice par la méthode de *cholesky*

Si 1/, 2/ et 5/ sont considérés constant en temps de calcul, alors 3/ et 4/ sont des facteurs importants. La réduction des ordres r et q signifie la réduction du temps de calcul. Comme l'ordre de q dépend principalement du nombre de zéros possible dans l'intervalle, le choix correct de cet intervalle évite des calculs inutiles. Quant à sa valeur, nous pouvons la fixer par la connaissance de cette région.

Ici, nous donnons seulement le calcul auxiliaire par rapport au modèle de Makhoul. Nous ne voulons pas comparer la complexité avec d'autre modèle.

2.9.4 Résultat

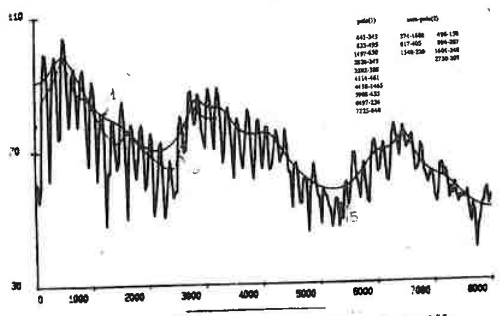
Pour pouvoir vérifier ce modèle de pôle zéro sélectif, nous traçons ici (fig.2.16, fig.2.17, fig.2.18, fig.2.19) les spectres d'une voyelle nasale naturelle /ɔ̃/ en faisant varier les ordres pour les pôles et pour les zéros.

Le signal de cette voyelle est numérisé à une fréquence d'échantillonage de 16kHz à l'aide d'un convertisseur *analogique-numérique* de 16bits. Un segment du signal pondéré par une fenêtre d type *Hamming* de longeur 25.6 msec et avec une préaccentuation de 0.95 est utilisé pour l'analyse.

Le critère du choix de l'ordre pour les pôles respecte la régle suivante: l'ordre pôle est égal à deux fois le nombre de pôles existant dans l'intervalle d'analyse plus deux ou quatre.

Pour l'ordre zéro q, il ne doit pas dépasser l'ordre p car le système de la production de la parole est de type passe-bas, donc $q < p$. Comme pour le choix des pôles, l'ordre des zéros utilise deux fois le nombre de zéros plus deux ou quatre.

Pour une nasale naturelle /ɔ̃/, nous prenons un intervalle de 0..3200Hz où existent cinq pôles (trois pôles oraux, deux pôles nasaux) et environ deux zéros causés par le couplage entre conduit oral et conduit nasal. L'ordre pour pôles est de 12, l'ordre pour les zéros est de 8. La figure2.16 montre le spectre estimé et les valeurs du formant. La figure2.17 montre le spectre obtenu par une application directe de l'algorithme Song avec q=18 p=20 et r=40 et avec q=14 p=20 r=40. La figure2.18 montre une comparaison entre les deux précédents résultats.



(1) LPC (tout pôle): ordre:p=20, région:0..8000Hz.
(2) selecARMA: ordre: r=40 q=8 p=12, région:0..3200Hz.

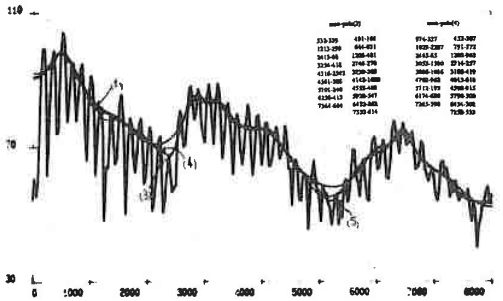
Figure 2.16. comparaison entre LP(1) selecARMA(2) et DFT(5)

Nous remarquons sur la figure2.16 que les valeurs de formant et de largeur de bande d'analyse de selective-ARMA sont plus raisonables que ceux de l'analyse LP tout pôle. La figure2.17 nous indique que si les ordres p et q ne sont pas bien choisis, le spectre obtenu par ARMA est clairement différent de celui de LP tout pôle, principalement autour des vallées. Les formants sont encore difficiles à interpréter (pour q=14 p=20). Le premier formant à 496Hz apparait jusqu'à q=18. La figure2.18 donne une différence de l'approximation. Si les ordres de selective-ARMA(2) et ceux de ARMA(3) sont choisis comme indiqué sur la figure, nous gagnons beaucoup en temps de calcul, soit $(18/8)^2 + (18/8)^2/2 = 7.5$, avec plus précision.

En utilisant le même traitement, nous donnons dans la figure2.19 le spectre au cours du temps, où l'évolution des formants est clairement tracée.

2.10 Conclusion

La modélisation du signal est une phase importante pour analyser une parole. Le modèle de AR à beaucoup servi dans le traitement de la parole grâce à la simplicité et l'efficacité du calcul.



(3) selecARMA: ordre:r=40 q=18 p=20. région:0..3200Hz.
(4) selecARMA: ordre:r=40 q=14 p=20. région:0..3200Hz.

Figure 2.17. comparaison entre LP(1) ARMA(3) et DFT(5).

La parole contient en réalité à la fois des pôles et des zéro. Même pour un son produit par un mécanisme tout pôle, les bruits en phase de transmission le transforment en un son pôle et zéro. Donc un modèle MA ou ARMA est toujours utile. Mais le calcul est très lourd en raison de la non-linéarité du problème. Il faut donc trouver un algorithme assez efficace pour réduire les exigences en temps de calcul selon le problème à résoudre.

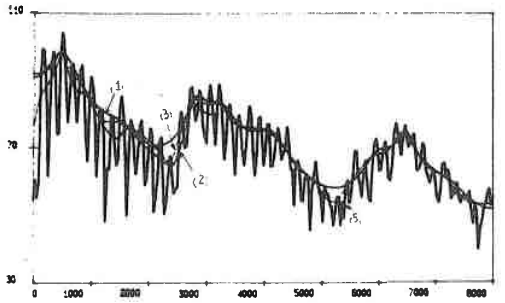


Figure 2.18. comparaison entre LP(1) selecARMA(2) ARMA(3)

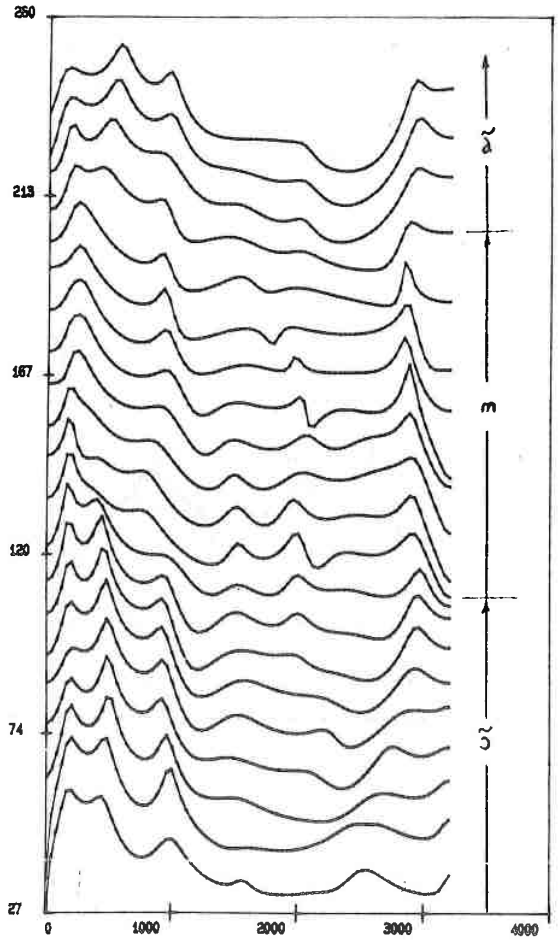


Figure 2.19. évolution des formants

3
Extraction automatique des formants

3.1 Introduction

Un formant est caractérisé par sa fréquence centrale et sa largeur de bande. La fréquence centrale indique la position d'un pic fréquentiel dans le spectre, et la largeur de bande mesure la concentration de l'énergie au voisinage de sa fréquence centrale. Une bande plus étroite correspond à une grande concentration de l'énergie autour de la fréquence centrale.

Les formants sont des paramètres essentiels pour caractériser la parole. Leur extraction avec précision est très importante pour l'étude de la perception humaine de la parole, la reconnaissance et la transmission à bas débit de la parole.

3.2 Résonateur et formant

Les formants qui apparaissent dans un spectre sont causés par les différents résonateurs du conduit vocal. Imaginons le conduit articulatoire comme étant un tuyau régulier semi-ouvert, excité par un signal. Un tel conduit forme des résonateurs qui produisent des fréquences inverses de multiples de sa longueur. La figure 3.1 montre la manière dont se répartissent les ventres de déplacement pour les quatre premières résonances.

Cependant, si le tuyau n'est pas régulier, par exemple dans un tuyau semi-ouvert composé de deux parties de section et de longueur différentes, les noeuds et les ventres de vibration ne se trouvent plus répartis selon des intervalles équidistants. Donc la distribution fréquentielle des résonances n'est plus régulière. La direction de déplacement de ces fréquences respecte une règle générale dégagée par Fant(1960) (dans [36]): lorsque le

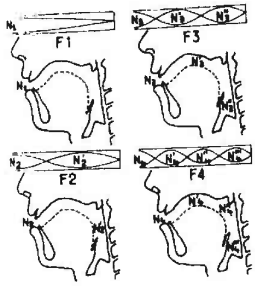


Figure 3.1. Analogie du conduit vocal avec un tuyau semi-ouvert, d'après Chiba et Kajiyama (dans [Liénard 77])

rétrécissement du conduit se trouve au voisinage d'un ventre de déplacement, la fréquence du formant correspondant baisse; lorsque le rétrécissement a lieu au voisinage d'un noeud, la fréquence augmente (Fig.3.2). Cette figure obtenue par la synthèse nous montre très bien ce que reflètent les formants d'un conduit vocal.

Chaque fréquence de résonance manifeste un maximum dans le spectre, mais les maxima dans le spectre ne coïncident pas exactement avec les formants, même pour les maxima calculés de la fonction du transfert polynomiale. Nous ne pouvons obtenir que les formants estimés.

3.3 Difficulté de la détection des formants

Lors de la détection des formants, on est souvent confronté à la présence des problèmes suivants:

1. pseudo-formants qui sont dûs en général au mauvais choix de l'ordre de la prédiction.
2. pics ayant une faible énergie et une largeur de bande étroite.
3. pics ayant une énergie importante et une largeur de bande large.

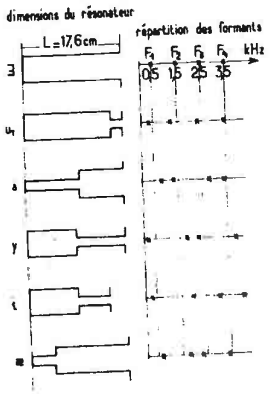


Figure 3.2. Caractéristiques géométriques d'un tuyau double fournissant les fréquences de résonance de diverses voyelles, d'après Fant 1960 (dans [Liénard 77])

Le premier problème est évident: plus l'ordre de LP est grand, plus est élevé le nombre de formants qui vont apparaitre dans le spectre. En plus des trois formants représentatifs et des formants supplémentaires, d'autres formants sont créés seulement par la mise en correspondance forcée de la courbe et du spectre.

Le deuxième est causé, dans la plupart des cas, par le couplage de pharynx et du conduit vocal. Il faut ignorer ce genre de formants.

Le troisième se produit souvent dans le spectre des voyelle nasales comme /ã/ et surtout /ɔ̃/. A cause des pôles et des zéros supplémentaires, les deux premiers formants possèdent des bandes de grande largeur. Ce type de formants est important car il est un indice de la nasalisation.

Dans le spectre d'une voyelle naturelle, l'allure du spectre est fixée par la nature de la voyelle prononcée, mais avec des variations à cause du profil complexe du conduit vocal (Fig.3.3). Un bon algorithme devrait être capable de détecter les formants qui se trouvent très proches, les formants potentiels qui n'apparaissent pas dans le spectre, et les formants

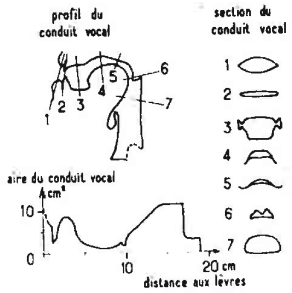


Figure 3.3. Schématisation du conduit vocal sous forme de sa fonction d'aire d'après Fant(1960) (dans [Lienard 77])

les plus représentatifs.

3.4 Détection des formants candidats

La détection des formants consiste à sélectionner les bons formants parmi plusieurs candidats. La détermination des formants candidats peut se faire par diverses méthodes.

3.4.1 Détection des maxima du spectre

Cette technique semble a priori la plus naturelle, puisqu'une résonance se traduit par un maximum d'énergie dans le spectre. Cette détection peut s'effectuer de plusieurs manières[20]:

Recherche de l'harmonique de plus grand énergie

On peut par exemple opérer à la sortie de filtres contigus, et comparer les sorties des filtres deux par deux. Les formants détectés doivent ensuite être interpolés car les fréquences du multiple de F_0 ne sont pas toujours les fréquences du maximum de spectre

surtout à un *pitch* très élevé, qui présente des harmoniques très espacés.

Utilisation des moments spectraux

Supposons le spectre échantillonné en un certain nombre de points de fréquences $f_1, f_2, ..., f_i, ...f_N$. Le moment d'ordre p de ce spectre sera défini par:

$$M_p = \sum_{i=1}^{N} f_i^p A(f_i)$$

$A(f_i)$ étant l'amplitude en chaque point.

Alors la fréquence du formant situé dans la zone définie par les indices i_1 et i_2 peut être calculée à l'aide de la formule suivante:

$$F = \frac{M_1}{M_0} = \frac{\sum_{i=i_1}^{i_2} f_i A(f_i)}{\sum_{i=i_1}^{i_2} A(f_i)}$$

Il est possible, pour chaque formant (F_1, F_2, F_3) de déterminer les limites f_{i_1} et f_{i_2} évitant au maximum les recouvrements de zones formantiques, ceci en utilisant des moments d'ordre 2.

3.4.2 Extraction des formants par analyse prédictive

La prédiction linéaire représente un système sous forme polynômiale, donc il est possible de trouver les formants et même les anti-formants à partir des racines de polynôme.

Soient $a_0, a_1, a_2, ..a_p$ et $b_0, b_1, b_2, ..b_q$ les paramètres de prédiction (voir chapitre précédent). Nous avons donc la forme en z:

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{q} b_i z^{-i}}{\sum_{i=0}^{p} a_i z^{-i}} \qquad (3.1)$$

Pour calculer les fréquences des formants et anti-formants, il faut transformer la formule (3.1) en s, ceci donnant la formule (3.6). Nous choisissons une transformation conservant l'invariance de la réponse impulsionnelle:

$$h(n) = h_a(nT)$$

— **pour les formants:**

$$H(z) = \sum_{k=1}^{N} \frac{A_k}{1 - e^{s_k T} z^{-1}} \quad \Leftrightarrow \quad H(s) = \sum_{k=1}^{N} \frac{A_k}{s - s_k} \tag{3.2}$$

Ceci vaut dire que les pôles aux $z = e^{s_k T}$ en z sont transformés en pôles s_k en s. Elle vérifie:

$$z_k = e^{s_k T} \tag{3.3}$$

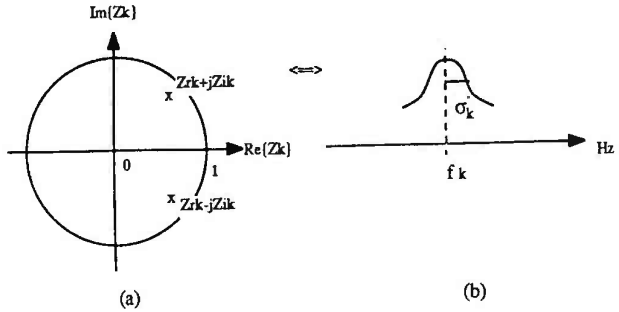
Si $z_k = z_{rk} + j z_{ik}$, alors:

$$f_k = \frac{1}{2\pi T} tg^{-1}[\frac{z_{ik}}{z_{rk}}] \tag{3.4}$$

$$\sigma_k = -\frac{1}{2\pi T} ln(z_{rk}^2 + z_{ik}^2) \tag{3.5}$$

Les formules (3.4) et (3.5) sont utiles pour calculer les fréquences des formants, étant donné les racines complexes en z (Fig.3.4).

— **pour les anti-formants:**



(a) représentation de Zk dans le plan.
(b) représentation d'un formant de fréquence f k et de largeur de bande σ_k .

Figure 3.4. relation entre (f_k, σ_k) et (z_{rk}, z_{ik})

Comme indiqué par Oppenheim[61], les zéros de la fonction du transfert sont une fonction des pôles et des coefficients A_k dans le développement de fraction algébrique. Et ils ne seront pas, en général, reflétés de la même manière que les pôles.

Alors pour un $H(z)$ donné, nous pouvons le représenter sous une forme du développement selon les pôles:

$$H(z) = \frac{B(z)}{A(z)} = \frac{B(z)}{\prod_k (1 - (z_{rk} + j z_{ik}) z^{-1})(1 - (z_{rk} - j z_{ik}) z^{-1}) \prod_l (1 - z_{rl} z^{-1})}$$

$$= \sum_k \left(\frac{a_{rk} + j a_{ik}}{1 - (z_{rk} + j z_{ik}) z^{-1}} + \frac{a_{rk} - j a_{ik}}{1 - (z_{rk} - j z_{ik}) z^{-1}}\right) + \sum_l \frac{a_{rl}}{1 - z_{rl} z^{-1}}$$

en prenant la transformation $z = e^{sT}$:

$$H(s) = \sum_k \left(\frac{a_{rk} + j a_{ik}}{s - (s_{rk} + j s_{ik})} + \frac{a_{rk} - j a_{ik}}{s - (s_{rk} - j s_{ik})}\right) + \sum_l \frac{a_{rl}}{s - s_{rl}}$$

où

$$s_{rk} = \frac{1}{T} ln[z_{rk}^2 + z_{ik}^2]$$

$$s_{ik} = \frac{1}{T} tg^{-1}[\frac{z_{ik}}{z_{rk}}]$$

$$a_{rk} + j a_{ik} = \frac{B(z)}{A(z)}(1 - (z_{rk} + j z_{ik}) z^{-1})|_{z=z_{rk}+jz_{ik}} \quad théoème\ de\ résidu$$

où s_{rk} et s_{ik} sont les parties réelle et imaginaire d'un pôle complexe. s_{rl} étant le pôle réel.

Avec tous ces coefficients, nous construisons une fonction en s:

$$H(s) = \frac{B(s)}{A(s)} = \frac{\sum_{k=0}^{M} d_k s^k}{\sum_{k=0}^{N} c_k s^k} \tag{3.6}$$

Les racines complexes de l'équation $A(s) = 0$ correspondent aux fréquences (en radian) des formants du système, les racines complexes de $B(s) = 0$ correspondant aux fréquences (en radian) des anti-formants.

3.4.3 Parcourir le spectre

Tout simplement, il suffit de parcourir le spectre et de localiser les maxima[72,48,47, 45,46]. Le spectre pourait être un cepstre[72] ou le spectre de la prédiction linéaire[51,49]. Si le spectre est obtenu par DFT sur les coefficients de la prédiction linéaire, la technique de "rehaussement" permet de mettre en évidence les pics d'*épaule* [51].

3.5 Détection des formants

Après avoir trouvé les formants candidats, il faut sélectionner les formants représentatifs. Schafer et Rabiner[72] cherchent les formants dans un spectre lissé par cepstre

et localisent les formants dans les zones prédéterminées, par exemple, premier formant entre 200..900Hz, deuxième formant dans la bande 550..2700Hz et troisième formant dans 1100..2950Hz. McCandless[51,52] cherche jusqu'à quatre formants dans une zone de 150..3400Hz. Il enlève un formant non significatif par certains critères. et à la fin lisse la trace des formants par un filtre de forme : $F_i' = \frac{1}{4}F_i(n-1) + \frac{1}{2}F_i(n) + \frac{1}{4}F_i(n+1)$ si $|F_i'(n) - F_i(n)| < 100Hz$. Papamichalis et Doddington[62] calculent les formants par les racines. et sélectionnent les formants ayant une faible largeur de bande comme formants. Le lissage est fait par minimisation d'une fonction de coût: $C(i,j) = dis(F_i, F_j') + \alpha dis(\sigma_i, \sigma_j')$. le symbole $'$ signifie la fenêtre d'analyse précedente (Fig.3.5).

R.J.Niederjohn et M.Lahat[58] ont développé un autre algorithme utilisant *les passages par zéro.* qui permet d'extraire correctement les premiers formants dans une situation des bruits de haut niveau. C'est une méthode contenant deux processus: une analyse grossière en fréquence suivie d'une analyse précise statistique temporelle. Ils ont utilisé des filtres de bande étroite afin de détecter les formants, et de réduire de plus les bruits.

Kopec[31] a cependant ouvert un nouveau chemin pour détecter les formants par un modèle markovien. Au lieu d'utiliser les paramètres heuristiques. il transforme le problème de la détection de formant en un processus markovien de décodage sur une séquence observée.

Après cette revue de tous ces algorithmes, nous voyons que toute l'importance de l'observation du signal. En s'inspirant du comportement d'un expert phonéticien quand il analyse le spectrogramme, nous proposons dans la section suivante un nouvel algorithme.

3.6 Algorithme proposé

3.6.1 Principe de l'algorithme

Nous savons qu'un expert phonéticien ne choisit pas les formants uniquement en utilisant le critère de la bande minimale. Mais il examine d'abord l'allure générale de tout le spectre du morceau de signal étudié et ensuite il détermine trois zones de recherche des formants.

Bien que le spectre varie en fonction du locuteur et du phonème étudié, il garde une certaine ressemblance avec des spectres bien définis. C'est pour cela que nous utilisons

une bibliothèque de spectre où on a déterminé manuellement la position des formants (fréquence et largeur de bande). On commence par comparer le spectre inconnu avec tous les spectres de la bibliothèque et il sera identifié au spectre de la bibliothèque qui réalise la plus petite distance. Ensuite la zone de recherche des formants est déterminée par les trois fréquences centrales du spectre de référence (Fig.3.6). Par exemple, si le spectre d'un signal est identifié comme un spectre du phonème /ɛ/, nous savons par conséquent qu'il faut chercher les formants au voisinage de 432Hz, 1693Hz et 2475Hz.

En considérant la variation du spectre selon le locuteur, et en tenant compte de la *stabilité* des formants, un spectre moyen ne se composant que des formants a le moins de facteurs dépendants de locuteur possible. Ceci nous conduit à construire le spectre *moyen* pour la bibliothèque à partir des formants moyens.

Il y a quinze voyelles dans la langue française. Puisque nous ne distinguons pas les voyelles orales /ɔ/ et /ɑ/, non plus que les voyelles nasales /ɛ̃/ et /œ̃/, il y a en réalité dans notre test treize voyelles. Pour les références, les formants d'un même signal sont calculés à l'aide du polynôme du filtre inverse pour des fenêtres d'analyse differentes. Ensuite on construit le spectre *moyen* de référence à partir des valeurs de ses trois formants.

3.6.2 Détail de l'algorithme

Préparation des références

Les références sont stockées sous forme de trois formants (fréquence de formant et largeur de bande en Hz), soit $\{(f_i^n, \sigma_i^n), i = 1,2,3;\ \ n = 1,2..13\}$, i étant l'indice de $i^{ème}$ formant et n le numéro du phonème. Les valeurs de formant d'une voyelle peuvent être les valeurs théoriques, ou les valeurs en moyenne extraites de la parole naturelle. Nous avons choisi le dernier cas.

Nous sélectionnons semi-automatiquement les trois premiers formants pour chaque voyelle, et effectuons un calcul de moyenne:

$$f_{i\ moy} = \frac{1}{N}\sum_{k=1}^{N} f_i^k, \qquad i = 1,2,3$$

$$\sigma_{i\ moy} = \frac{1}{N}\sum_{k=1}^{N} \sigma_i^k. \qquad i = 1,2,3$$

N étant le nombre éventuel de $i^{ème}$ formant.

Le spectre de référence est obtenu de ces trois formants par moyenne. Les pôles (z_{rk}, z_{ik}) en z correpondant sont calculés selon la transformation (3.3):

$$z_{rk} = cos(2\pi T f_k)e^{\pi T \sigma_k}$$

$$z_{ik} = sin(2\pi T f_k)e^{\pi T \sigma_k}$$

Nous construisons le polynôme de $A(z)$ par ces pôles (formule 3.7) et ensuite calculons le spectre *SPECREF(m)* par DFT (formule 3.8):

$$A(z) = \sum_{i=1,2,3} (1-(z_{rk}+jz_{ik})z^{-1})(1-(z_{rk}-jz_{ik})z^{-1}) \qquad (3.7)$$

$$SPECREF(m) = -20\ log|DFT\{\ A(z)|_{z=e^{jwm}}\ \}| \qquad (3.8)$$

3.6.3 Extraction des formants

Le premier pas de l'extraction est de déterminer les zones de recherche. Ces zones seront indiquées par les trois formants composant le spectre de référence qui réalise la distance minimale.

Supposons *SPECVOY(m)* le spectre d'une voyelle inconnue:

$$SPECVOY(m) = 20\ log|DFT\{\ \frac{B(z)}{A(z)}|_{z=e^{jwm}}\ \}|$$

Nous faisons une comparaison de ce spectre inconnu avec tous les spectres de référence en utilisant une distance de *Hamming*:

$$d_j = \sum_m |SPECREF^j(m) - SPECVOY(m)|$$

$$n = arg[min_{j=1,2,..13} \{d_j\}]$$

ici n est le numéro dont la référence réalise la distance minimale. Nous avons donc trois fréquences (f_1^n, f_2^n, f_3^n) qui indiquent les trois zones importantes.

Le deuxième pas est de sélectionner les formants au voisinage de chaque formant de référence.

Nous commençons par chercher le 3ème formant. La zone de recherche est définie par $(f_2^n + f_3^n)/2$ et 3200Hz. Si $((f_2^n + f_3^n)/2) < 1950Hz$, alors la borne inférieure est fixée à 1950Hz. S'il n'y a qu'un seul pic dans cet intervalle, alors c'est le 3ème formant quelque soit sa largeur de bande. S'il y a deux pics p_1, p_2 des largeurs de bande minimales, en notant 1 le formant qui a une largeur de bande minimale, et 2 celui qui a une largeur de bande plus grande, alors les distances de p_1 et p_2 à f_3^n sont calculées par $d_1 = |f_3^n - p1|$ et $d_2 = |f_3^n - p2|$. Nous ne choisissons pas uniquement selon le critère de bande minmale, mais selon les positions relatives. Si p1 est trop loin de f_3^n (350Hz par exemple), alors si p2 est deux fois plus près de f_3^n que p1, on prend le deuxième pic comme F_3, ou si p2 est plus près que p1 et si la largeur de bande du p1 (noté b1) est supérieur à celle de p2 (noté b2) de l'ordre de 100Hz, alors F_3 =p2. Voici l'algorithme qui permet d'en choisir un pic comme F_3:

```
borneInferieure = (f2 + f3) / 2;

si borneInferieure < 1950
alors
  borneInferieure = 1950;
fin_si

Choisir entre borneInferieure et 3200 deux pics
des largeurs de bande minimale, notant p1, p2;

si nombrePics = 1
alors
  F3 = p1;
sinon
  F3 = le pic le plus representatif;
fin_si
```


Si on ne trouve aucun formant entre 1950..3200Hz, on reduit la borne inférieure par pas de 200Hz jusqu'à 1400Hz:

```
borneInferieure = 1950;

tantque (pas de pic entre borneInferieure et 3200)
        et
        (borneInferieure > 1400) faire
debut
  Choisir entre borneInferieure et 3200 deux pics
  des largeurs de bande minimale, notant p1, p2;

  si nombrePics = 1
  alors
    F3 = p1;
  sinon
    F3 = le pic le plus representatif;
  fin_si

  borneInferieure = borneInferieure - 200;

fin_tantque

F3=p1;
```

Ensuite nous déterminons le 2ème formant. Puisque F_3 est déterminé, alors le 2ème formant devrait se trouver entre f_1^n et F_3. Le traitement est le même que pour F_3.

```
borneInferieure = f1 + delta;
borneSuperieure = F3 - delta;

Choisir entre borneInferieure et borneSuperieur deux pics
des largeurs de bande minimale, notant p1, p2;

si nombrePics = 1
alors
  F2 = p1;
sinon
  F2 = le pic le plus representatif;
fin_si


si on ne trouve aucun formant entre borneInferiuer et borneSuperieur
alors

  tantque pas de formant faire
    Choisir entre borneInferieure et borneSuperieur deux pics
    des largeurs de bande minimale, notant p1, p2;

    borneInferieure = borneInferieur - 50;
  fin_tantque

fin_si

F2=p1;
```

A la fin, on effectue la détermination du 1er formant. A cette étape, nous avons le 2ème formant qui permet de prédire le rang du 1er formant. Si le 2ème formant est supérieur à 1650Hz, on commence la recherche à partir de 200Hz sinon à partir de 320Hz. Ceci consiste à éviter l'effet du formant glottal.

```
si F2 > 1650
alors
```



```
  borneInferieure = 200;
sinon
  borneInferieure = 320;
fin_si

borneSuperieur = F2 - delta;

si borneSuperieure > 1100
alors
   borneSuperieure = 1100;

Choisir entre borneInferieure et borneSuperieur deux pics
des largeurs de bande minimale, notant p1, p2;

si nombrePics = 1
alors
  F2 = p1;
sinon
  F2 = le pic le plus representatif;
fin_si

F1 = p1;
```

La contrainte de la largeur de bande pour choisir un des deux formants est enlevée, car elle n'est plus une bonne indication du formant dans cet intervalle.

Si on ne trouve aucun formant, on recommence avec une borne inférieure encore plus basse, 165Hz par exemple. Si l'on trouve un pic, ce sera le 1er formant, sinon il n'y a pas de formant dans cette région, et on doit reparcourir la zone qui était résumée pour F_2, parce qu'il est possible que l'on ait enlevé le formant ayant une bande plus large que celui retenu pour F_2.

```
si pas de formant
```

```
alors

  borneInferieure = 165;

  borneSuperieur = F2 - delta;

  si borneSuperieure > 1100
  alors
     borneSuperieure = 1100;

  Choisir entre borneInferieure et borneSuperieur deux pics
  des largeurs de bande minimale, notant p1, p2;

  si pas de formant
  alors
    borneInferieure = F2 + delta;
    borneSuperieure = F3 - delta;

    Choisir entre borneInferieure et borneSuperieur deux pics
    des largeurs de bande minimale, notant p1, p2;

    tantque (pas de formant) et (borneSuperieur < F3)  faire

       Choisir entre borneInferieure et borneSuperieur deux pics
       des largeurs de bande minimale, notant p1, p2;

       borneSuperieure = borneSuperieure + 100;
    fin_tantque

    F1 = F2;
    F2 = p1;

  fin_si

fin_si
```


3.6.4 Résultat

Nous avons utilisé un corpus LABISA du GRECO-PRC communication Homme-machine (référencer au chapitre 4 pour une présentation du corpus). C'est un corpus comprenant des voyelles orales et aussi des voyelles nasales. En considérant la structure complexe d'une voyelle nasale, nous avons choisi le modèle ARMA dans une zone fréquentielle sélectionnée entre 0 et 3200Hz. L'ordre des zéros est 8, l'ordre des pôles est 12, et le signal de chaque fenêtre d'analyse est préaccentué avec un coefficient de 0.95. La longueur de la fenêtre est de 30msec. Le déplacement de la fenêtre est de 12.8msec.

Les fréquences et les largeurs de bande de treize voyelles dans notre bibliothèque sont données respctivement dans la Tab.3.1.1 et la Tab.3.1.2 ci-dessous.

	f_1/σ_1	f_2/σ_2	f_3/σ_3
i	315/120	1955/124	2758/178
ϵ	388/103	1805/112	2529/156
ε	432/119	1693/121	2475/150
a	500/125	1526/124	2473/170
$\mathfrak{o}$	459/126	1173/126	2421/126
o	389/119	1027/149	2465/142
u	453/170	1295/209	2478/179
y	364/140	1693/115	2574/203
œ	475/106	1426/112	2400/148
ə	403/143	1432/139	2422/158
$\tilde{\text{ɔ}}$	495/192	1138/203	2575/201
$\tilde{\text{ɛ}}$	567/197	1564/126	2598/166
$\tilde{a}$	516/214	1004/155	2642/199

Tab. 3.1.1 : Les fréquences et les largeurs de bande de treize voyelles dans la bibliothèque pour les locuteurs masculins.

	f_1/σ_1	f_2/σ_2	f_3/σ_3
i	365/85	1673/308	2535/151
ϵ	440/74	1762/259	2625/199
ε	471/87	1677/219	2647/218
a	539/123	1607/165	2686/247
$\mathfrak{o}$	503/104	1308/132	2718/197
o	450/113	1073/149	2734/182
u	475/143	1280/214	2647/239
y	350/78	1824/157	2604/198
œ	557/109	1615/125	2716/157
ø	444/99	1548/172	2677/213
$\tilde{\mathfrak{o}}$	492/169	1084/148	2471/271
$\tilde{\varepsilon}$	509/138	1558/150	2610/264
$\tilde{a}$	477/162	1078/138	2444/297

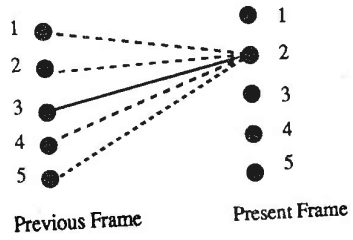
Tab. 3.1.2 : Les fréquences et les largeurs de bande de treize voyelles dans la bibliothèque pour les locuteurs féminins.

Nous avons fait le test sur sept locuteurs masculins: JB, BG, BP, BT, FC, FL, GM, et sur quatre locuteurs feminins: JF, MD, DP, CF. La figure 3.7 et 3.8 montre une partie des traces de trois premiers formants de voyelle. C'est une trace sans lissage, et indépendante des consonnes en contexte.

3.7 Conclusion

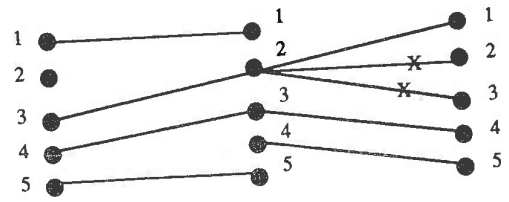
Dans ce chapitre, nous avons décrit les méthodes permettant d'extraire les formants du signal. Nous avons proposé un autre algorithme qui est capable de détecter les formants à partir des coefficients de la prédiction linéaire de grand ordre, et nous avons mis en évidence les formules utiles pour calculer les fréquences des formants et des anti-formants à partir des coefficients du modèle ARMA.

Dans notre algorithme, nous avons introduit la technique d'expert phonéticien dans le domaine fréquentiel. Il serait encore mieux que cette idée soit implantée le long de l'axe du temps. Les mouvements des l'organes articulatoires sont relativement lents, donc les fréquences de formant ne sont pas très variables le long du temps. La correction des formants selon cette information aidera à l'extraction des formants sur l'axe des fréquences.



C(2,3) = Min { C(2,i), i=1...,5 }

(a) 1-Step Dynamic Programming



C(1,2) = Min { C(1,2), C(2,2), C(3,2) }

(b) Disconnection of Multiple Branches

Figure 3.5. établissement de la continuité de trace des formants

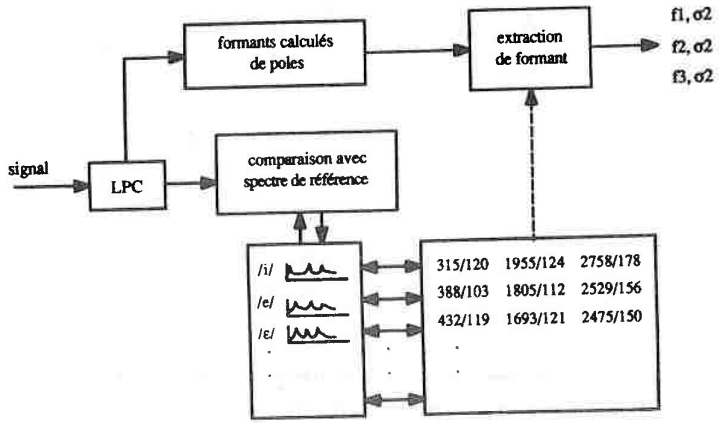


Figure 3.6. Principe de détection des formants

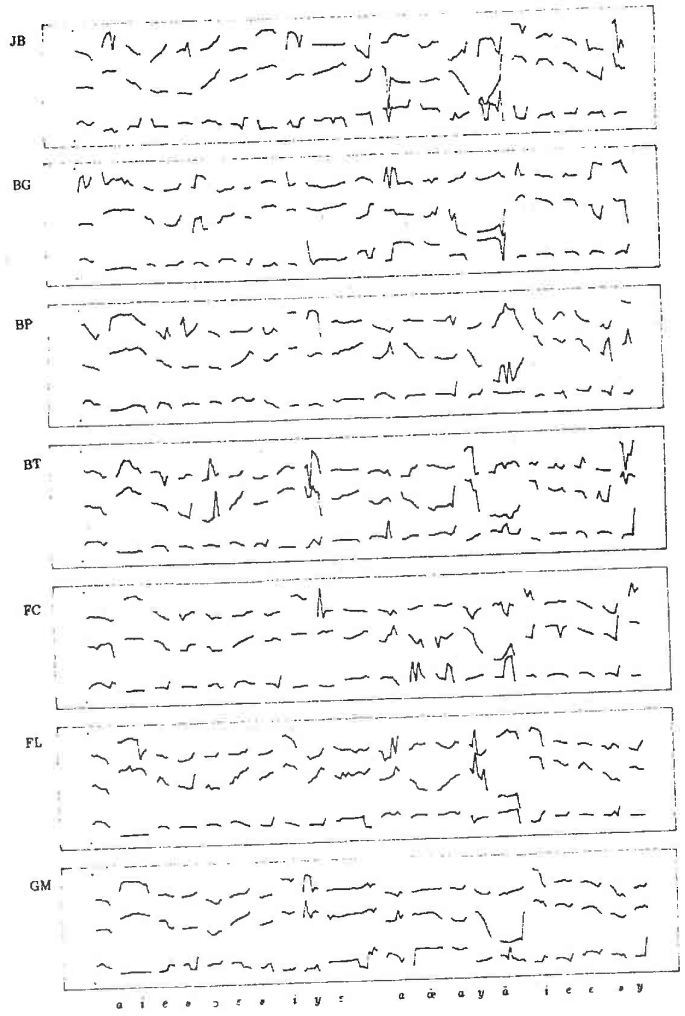


Figure 3.7. Les traces de premiers formants de sept locuteurs masculins. La phrase: la bise et le soleil se disputaient, chacun assurant qu'il était le plus ... .

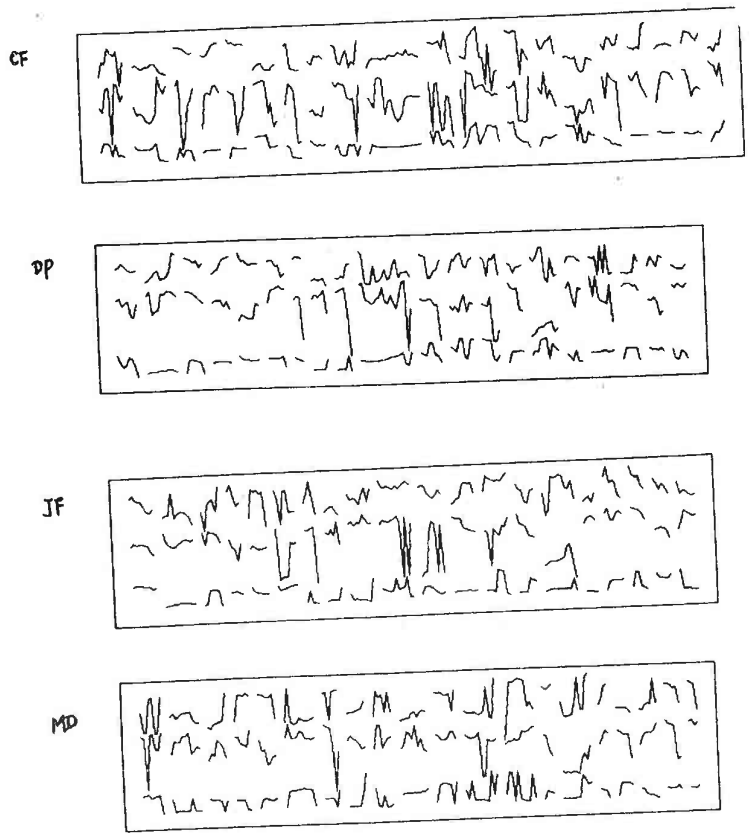


Figure 3.8. Les traces de premiers formants de quatre locuteurs féminins. La phrase: la bise et le soleil se disputaient, chacun assurant qu'il était le plus ... .

4
Les voyelles nasales et leur reconnaissance

4.1 Introduction

25% des langues au monde possèdent des voyelles nasales et toutes ont des voyelles nasalisées dues à la coarticulation des consonnes nasales [42]. La nasalité vocalique est une aspect important des phonèmes dans les langues contenant les voyelles nasales, elle constitue un trait distinctif au niveau sémantique.

Il y a 15 voyelles dans la langue française, y compris 4 nasales qui définissent une catégorie phonologique décrite dans le tableau suivant:

lèvres	position de la langue	
	postérieure	antérieure
arrondies	$\tilde{\text{ɔ}}$	$\tilde{\text{œ}}$
non arrondies	$\tilde{\text{ɑ}}$	$\tilde{\text{ɛ}}$

La nasalisation s'est développée à partir de l'influence d'une consonne nasale sur la voyelle qui la précédait (i.e. assimilation régressive)[24]. Delattre (1969) a montré une évolution progressive des voyelles nasalisées vers les voyelles nasales par le schéma suivant:

$$\begin{array}{l} \check{\text{i}} - \check{\text{e}} - \check{\text{ɛ}} - \tilde{\text{ɛ}} \\ \check{\text{y}} - \check{\text{ø}} - \check{\text{œ}} - \tilde{\text{œ}} \\ \check{\text{u}} - \check{\text{o}} - \check{\text{ɔ}} - \tilde{\text{ɔ}} - \tilde{\text{ɑ}} \end{array}$$

Le signe ˘ signifie voyelle nasalisée et ~ voyelle nasale.

Les voyelles nasales ne possèdent plus beaucoup de ressemblance avec les voyelles orales correspondantes. Elles sont des voyelles tout à fait indépendantes. A cause de l'effet de l'introduction de la cavité nasale, les voyelles nasales montrent souvent des caractéristiques un peu irrégulières et très complexes, fonction du degré de couplage.

Dans la partie suivante, nous allons présenter brièvement l'anatomie des fosses nasales, synthétiser les résultats antérieurs obtenus par des chercheurs. Ensuite nous allons analyser les voyelles nasales naturelles en utilisant un modèle de pôle-zéro, et essayer de les reconnaître automatiquement.

4.2 Anatomie de la cavité nasale

Le détail de la description des fosses nasales est donné dans la thèse d'état de F.Louchamp[41].

La cavité nasale (ou les fosses nasales) peut être divisée en trois parties de taille inégale: d'avant en arrière, on trouve la zone du nez ($\approx$ 2cm), puis les fosses nasales proprement dites (6-7 cm), divisées verticalement par une cloison paramédiane, et enfin le conduit rhinopharyngal ou nasophayngal (3-5 cm). La figure 4.1 montre la position générale de la cavité nasale dans la tête.

4.2.1 Le nez

Dans le nez, il y a une lame cartilagineuse interne verticale qui divise le nez en deux parties plus ou moins égales. La figure4.2 montre l'architecture osseuse et cartilagineuse du nez.

4.2.2 Les fosses nasales et le rhinopharynx

Les fosses nasales ont la forme d'un volume trapézoïdal allongé dont la base est plus large que le sommet. Leur longueur est d'environ 7 à 8 cm, la hauteur maximale de 5 cm et la largeur passe de 4 cm à la base à moins de 1 cm au sommet, en ne comptant pas les cavités accessoires (sinus). Elle sont divisées verticalement par une cloison cartilageuse et osseuse.

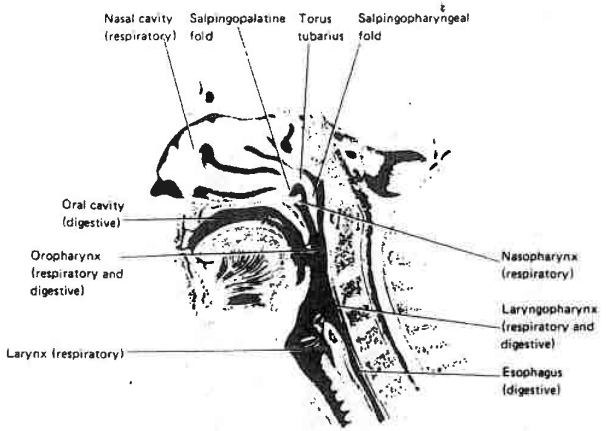


Figure 4.1. position anatomique des fosses nasales.

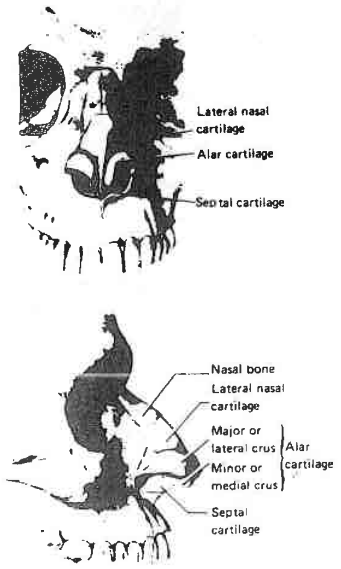


Figure 4.2. os et cartilages du nez.

Le rhinopharynx a une forme grossièrement parallèlépipédique de 3 à 4 cm de longueur lorsque le voile du palais est relevé.

4.2.3 Les sinus paranasaux

Dans la cavité nasale, il y a certaines cavités moins volumineuses qui modifient plus ou moins le spectre d'un son. Ce sont:

sinus	volume moyennes en *cm*
sinus frontaux	3 x 2.5 x 2
sinus ethmoïdaux	8.3 cm^3 au total
sinus sphénoïde	0.2 x 0.2 x 0.15
sinus maxillaires	3.3 x 2.3 x 3.4

Les sinus ethmoïdaux sont un ensemble de petites cavités aux parois très fines dans l'épaisseur de la masse latérale de l'éthmoïde. La figure 4.3 montre les formes et les positions de ces cavités.

Les sinus affectent l'acoustique par un effet semblable à un résonateur de *Helmholtz* (Fig.4.4). Un résonateur de Helmholtz est une cavité sphérique, à parois rigides, munie d'une ouverture et éventuellement d'un goulot. La fréquence de résonance dépend des volumes respectifs du goulot et de la cavité. En présence d'un son comportant de nombreuses composantes de fréquences différentes, le résonateur permet de sélectionner celles qui correspondent à sa fréquence de résonance et de les amplifier. La figure 4.4 montre la connexion d'un sinus au conduit nasal.

L'organe nasal a une structure tellement compliquée que nous ne pouvons pas savoir le rôle de chaque partie dans la prononciation des nasales. Un modèle de conduit nasal peut être obtenu, soit par etude de l'anatomie des cadavres[4], soit par analyse spectrale sur des sons produits[38.14]. Nous allons voir la correspondance du spectre à l'articulation de nasale.

4.3 Articulation et spectre des voyelles nasales

4.3.1 Articulation

L'articulation d'une voyelle nasale résulte de l'effet du passage de l'air par la bouche et par les fosses nasales. L'air venant du poumon se divise au niveau du voile, la proportion

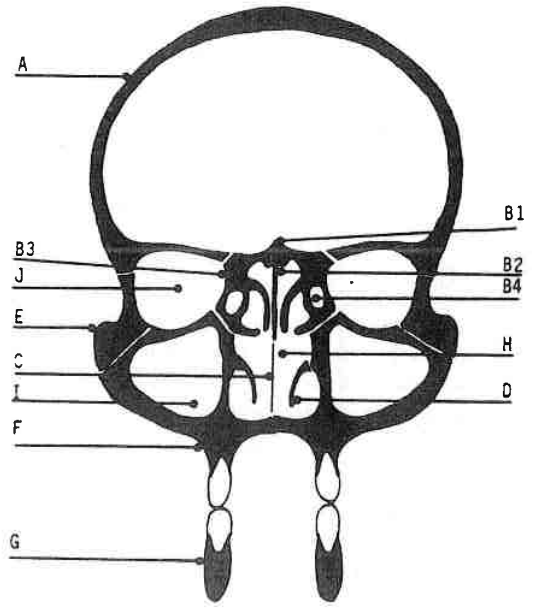


FIG 38 - TÊTE (COUPE FRONTALE)

A frontal. B ethmoïde. B1 apophyse crista-galli. B2 lame perpendiculaire. B3 os planum. B4 cellule. C cartilage de la cloison nasale. D cornet inférieur. E os malaire. F maxillaire. G mandibule. H fosses nasales. I sinus maxillaire. J orbite.

Figure 4.3. coupe frontale schématique de la tête.

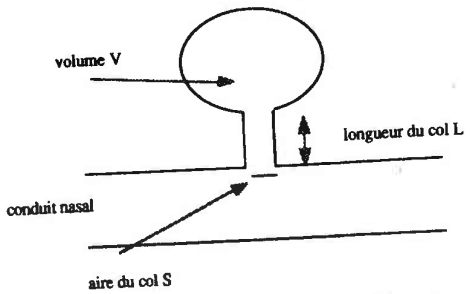


Figure 4.4. le sinus comme un résonateur de Helmholtz.

entre l'air passant vers les fosses nasales et celui passant vers la bouche est le facteur principal qui décide du degré du couplage. Pour articuler une voyelle nasale, Zerling (1984) a découvert qu'au moment de l'abaissement du voile, il se produit un recul de la langue et plus de labialisation. C'e recul de la langue implique un abaissement de F2, et un accroissement de la labialisation provoque une chute de F1 (Fig.4.5, Fig.4.6)[41].

Donc l'articulation d'une voyelle nasale est complètement différente de celle d'une voyelle orale, même d'une voyelle orale qui a la même configuration (Fig4.5). Pour mieux comprendre le mécanisme des nasales, nous analysons les voyelles nasales à l'aide d'un modèle.

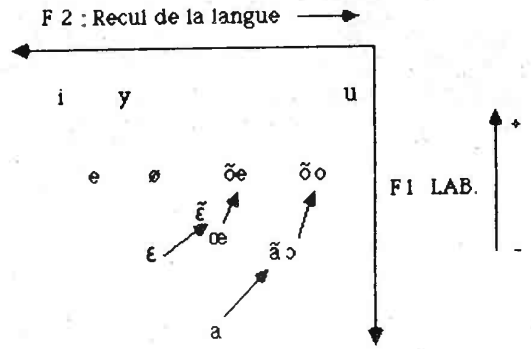


Figure 4.5. évolution schématique des deux premiers formants en fonction des différences articulatoires entre voyelles orale et nasale homologue.

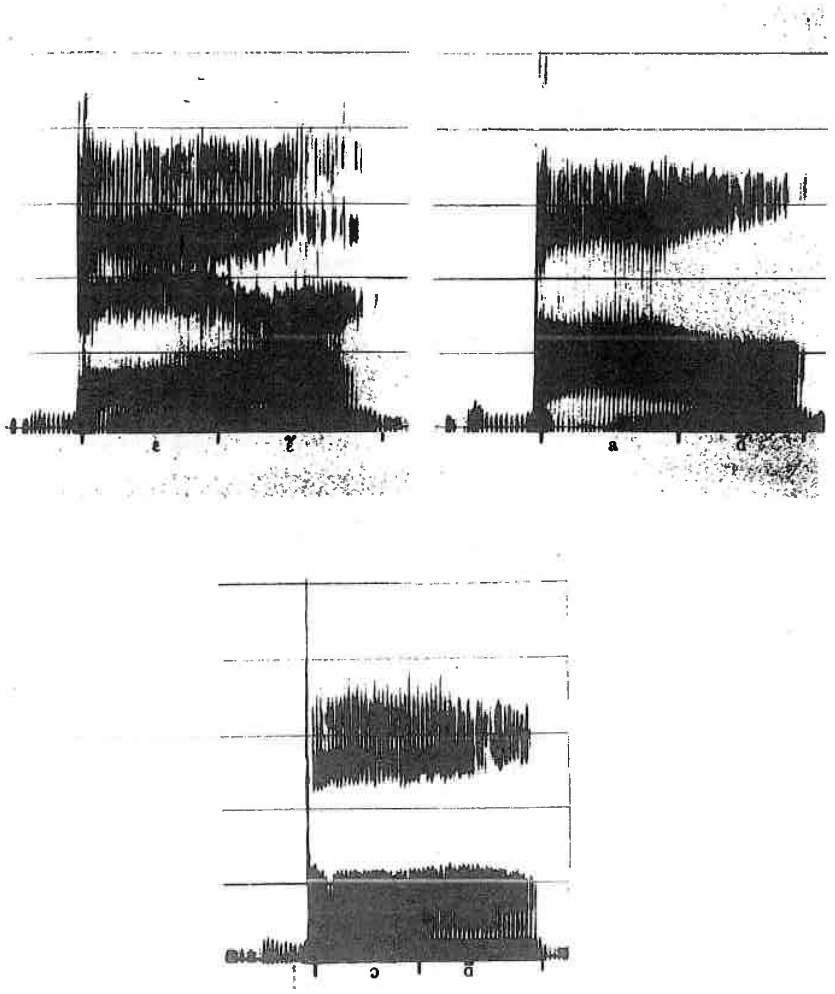


Figure 4.6. spectrogrammes des voyelles en hiatus [$\varepsilon\tilde{\varepsilon}$], [$a\tilde{a}$], [$\supset\tilde{a}$].

4.3.2 Modèle électrique

Vers 1960, G.Fant a proposé une méthode analytique pour étudier la nasalisation d'une voyelle en fonction du couplage du conduit nasal avec les cavités buccales et pharyngales.

Il est bien connu que tous les systèmes peuvent être représentés de façon équivalente par un circuit électrique. Le système de la production, après avoir fait certaines simplifications, peut être également représenté par ce genre de modèle: l'air venant des poumons est équivalent à une source de débit (volt) avec une réactance intérieure. Si on fait l'hypothèse qu'il n'y a pas de perte, alors cette réactance devient nulle. Au point du couplage, trois cavités principales, c'est à dire cavité nasale, cavité buccale et cavité pharyngale, se réunissent comme montré dans la figure 4.7.

La méthode analytique est fournie par la résolution graphique au point de couplage de l'équation $1/X_\epsilon$, où X_ϵ représente l'impédance d'entrée du conduit vocal au niveau de la glotte. De la figure 4.7 nous avons:

$$\frac{1}{X_\epsilon} = \frac{1}{X_p} + \frac{1}{X_b} + \frac{1}{X_n}$$

L'axe fréquentiel correspondant à $1/X_\epsilon$ représente les formants du système. Grâce à l'hypothèse que les pertes sont négligeables, X_b et X_n sont de la forme de l'impédance multipliée par une fonction tangente tandis que X_p est de la forme de son impédance multipliée par une fonction cotangente[57]. Les voyelles orales sont produites dans le conduit buccal et pharyngal lorsque il n'y a pas de couplage, donc l'axe de $X_{bp} = \infty$ ($\frac{1}{X_{pb}} = \frac{1}{X_b} + \frac{1}{X_p}$) correspond aux fréquences des formants oraux. De même, l'axe de $\frac{1}{X_{bp}} + \frac{1}{X_b} = 0$ correspond aux formants après couplage.

La figure 4.8 montre la relation entre le spectre de LP et la répartition de la réactance.

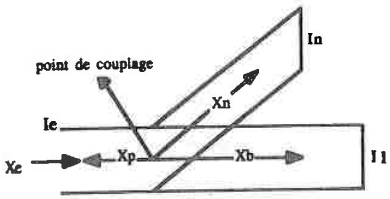


Figure 4.7. modèle de ligne de transmission équivalente à l'articulation.

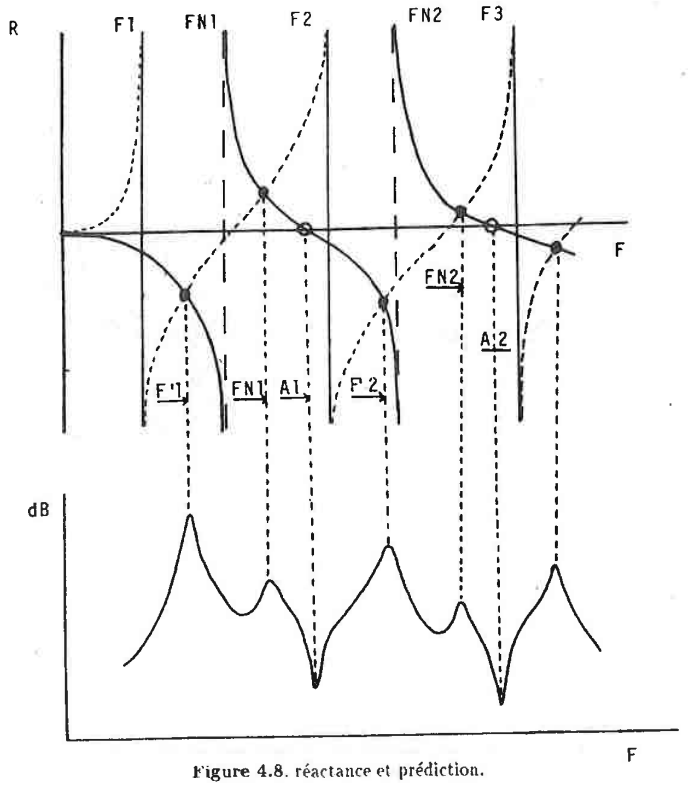


Figure 4.8. réactance et prédiction.

Dans cette figure, F_i note le $i^{ème}$ formant lié à la cavité buccale, F_{ni} note le $i^{ème}$ formant nasal, et le symbole ' est pour les formants déplacés à cause du couplage de la cavité nasale à la cavité buccale. Nous voyons qu'un formant correspond à la fréquence de croisement de X_{bp} et X_n, sa largeur de bande correspond à la distance du croisement sur l'axe fréquence, et qu'un zéro correspond à la fréquence du croisement de l'axe F et de la réactance, la largeur de bande correspond au déplacement horizontale du passage de la réactance à l'axe F.

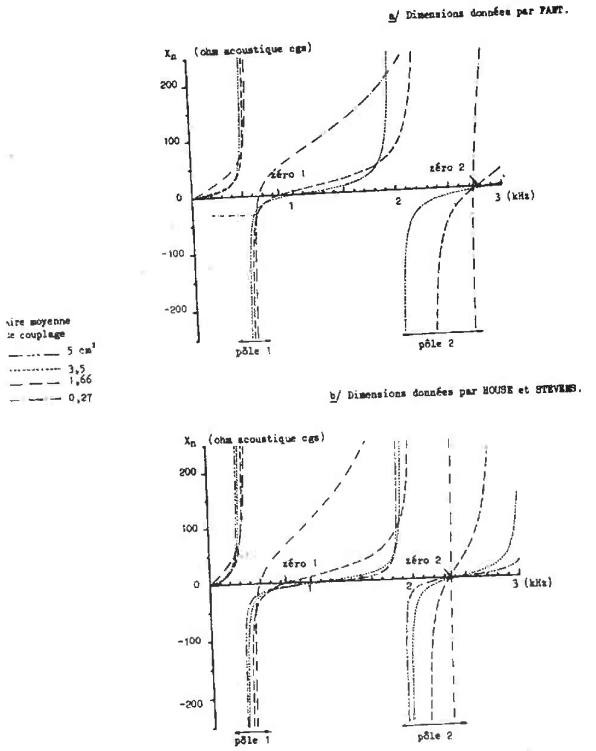


Figure 4.9. réactance d'entrée du conduit nasal au point de couplage.

4.3.3 Analyse par simulation

Mrayati[57] a calculé les réactances X_{bp} et $-X_n$ pour les quatre voyelles nasales françaises en faisant une hypothèse que les voyelles nasales /ɑ̃/ /ɔ̃/ /ɛ̃/ et /œ̃/ sont produites à partir des configurations du conduit vocal correspondant aux voyelles /ɑ/ /ɔ/ /ɛ/ et /œ/ respectivement, avec un couplage de la cavité nasale (Fig.4.9, 4.10).

A partir de ces illustrations, nous voyons très facilement les interactions du conduit buccal et du conduit nasal. Mais nous devons tenir compte des simplifications faites a priori, et la configuration vocalique qui est en général déjà déformée par le recul de la

langue (voir 4.3.1).

4.3.4 Analyse sur le spectre de voyelle nasale

L'un des aspects fondamentaux de l'analyse des voyelles nasales est l'interprétation des évènements particuliers aux nasales, se reflétant dans leur spectre. Ces évènements acompagnent souvent la nasalisation, parce que la dynamique de l'énergie spectrale est moins importante par rapport à celle des voyelles orales. Ces évènements particuliers peuvent être pris comme des indices de nasalisation, donc de reconnaissance, et également pour perfectionner nos connaissances sur la structure interne de l'articulation. Jusqu'à présent, on connait bien trois formants supplémentaires dans le spectre d'une nasale.

— *Formant glottal*

Un formant glottal est formé par la répétition de la source de parole. C'est un couplage entre source et cavité supraglottique. Ce formant est en général de faible énergie, donc il est souvent masqué par d'autres formants au voisinage. Le formant glottal paraît lorsque l'énergie autour de lui est faible. La voyelle /i/ est une voyelle qui a son énergie répartie en haute fréquence, le formant glottal devient visible dans la vallée entre F1 et F2. Fant (1980) remarque que le formant glottal est souvent plus visible pour les voyelles nasales, car la présence d'un zéro proche réalise une sorte de filtrage inverse de F1 rendant Fg plus proéminent (cf.[41]). L'effet du formant glottal est important pour F1, moins pour F2 et F3, car l'impédance de la source contient un élément inductif dont la réactance devient importante à des fréquences supérieures à F1 [57].

— *Formant causé par le couplage nasal externe*

Il peut y avoir un couplage non seulement à l'intérieur de la cavité nasale et de la cavité buccale au point du voile, mais aussi au point extérieur entre le nez et la bouche. Imaginant la cavité nasale comme un résonateur de Helmholtz, Merlier (1984) a suggéré que l'intensité de la source buccale et la proximité de l'extrêmité nasale permettaient l'existence d'un couplage nasal externe (Fig.4.11). La figure 4.12 donne un exemple.

— *Formant causé par les sinus*

Le formant causé par les sinus n'existent que pour les nasales. Lindqvist et Sundberg (1972) ont étudié l'effet des sinus maxillaires et des sinus frontaux aux résonateurs du conduit nasal (Fig.4.13).

Parmi des formants supplémentaires, certains peuvent se produire dans le cas d'une voyelle orale et également dans le cas d'une voyelle nasale. Nous devons faire très attention lorsqu'on fait l'analyse fine et automatique d' un spectre d'une voyelle.

A l'aide d'un modèle électrique, nous pouvons simuler l'articulation d'une voyelle orale ou nasale, observer les effets interactifs des différentes cavités en faisant varier certains paramètres. Mais l'analyse par modèle physique n'est pas capable de réfléter toutes les caractéristiques d'une voyelle naturelle, alors nous utilisons des analyses par DFT, LPC, modèle de pôle zéro etc. pour étudier le spectre réel d'une voyelle.

4.3.5 Analyse par spectre

Pour observer le changement du spectre d'une voyelle de orale à nasale, les spectres à différents moments sont utilisés. Hawkins et Stevens [23] ont donné des spectrogrammes de cinq couples de voyelles orale et nasale homologues reproduits à la figure4.14. La figure4.15 montre les spectres de voyelles /u/ /ɔ/ /a/ dans la condition de non-nasalisation, de glotte ouverte et de nasalité à l'aide d'un vibrateur externe[14]. En plus, la nasalisation dépend du contexte et de la perception individuelle. Dans[23], ils ont fourni aussi des critères specttraux pour distinguer entre voyelles orale et nasale synthétisées. Nous donnons, dans les figures (Fig. 4.16, 4.17, 4.18), les deux prononciations de locuteurs différents pour trois voyelles nasales /ã/ /ɔ̃/ et /ɛ̃/. Nous nous apercevons tout de suite qu'il est trop complexe pour un spectre de nasale et que l'on ne peut pas trouver des indices nasals généraux uniquement d'après le spectre. Les voyelles nasales et les voyelles orales se distinguent relativement.

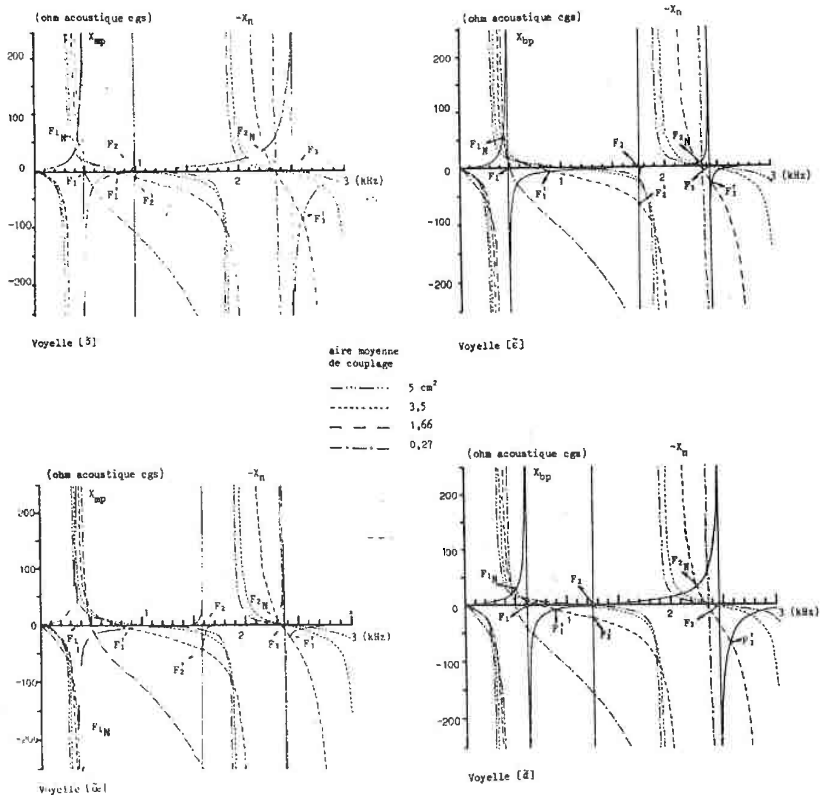
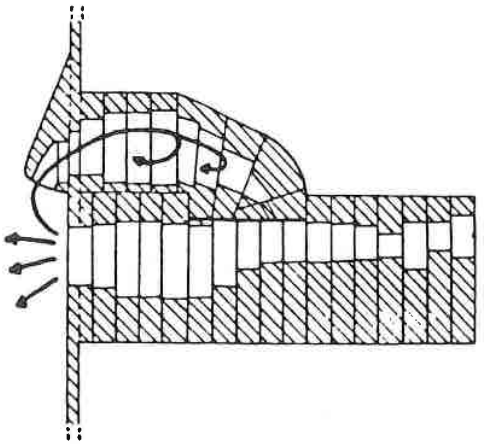


Figure 4.10. quatre degré de couplage avec la configuration vocalique équivalente à la voyelle.



Conduit vocal avec fosses nasales.

Figure 4.11. couplage nasal externe (d'après [lonc88]).

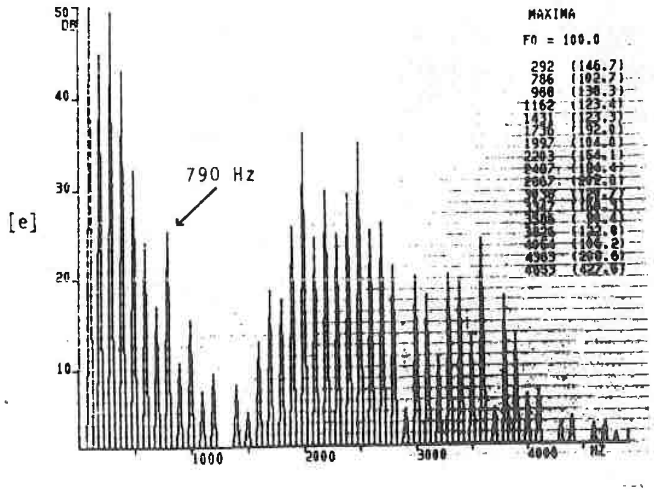


Figure 4.12. exemple d'un couplage nasal externe (d'après [lonc88]).

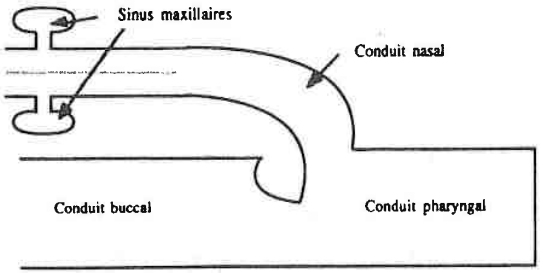


Figure 4.13. sinus attachés au conduit nasal (d'après [lonc88]).

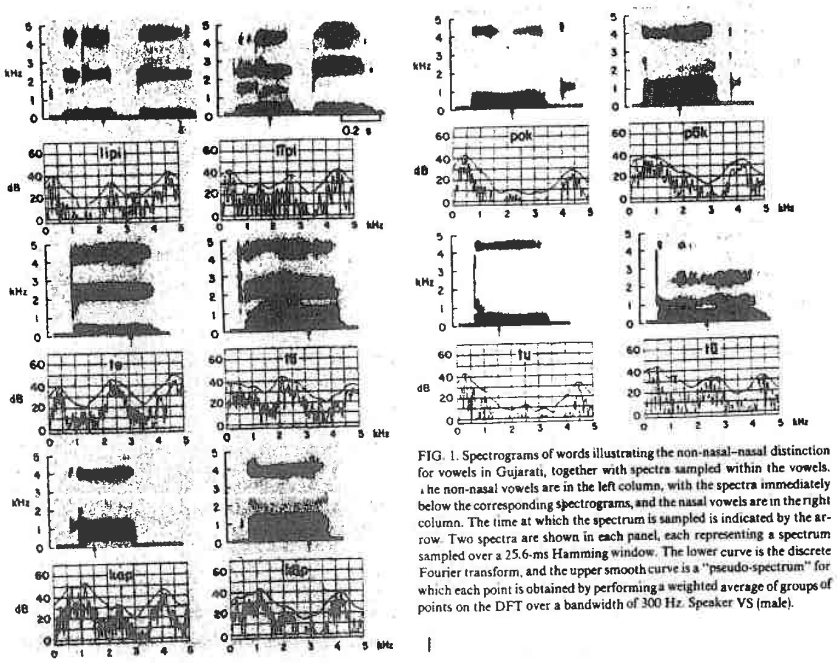


FIG. 1. Spectrograms of words illustrating the non-nasal–nasal distinction for vowels in Gujarati, together with spectra sampled within the vowels. The non-nasal vowels are in the left column, with the spectra immediately below the corresponding spectrograms, and the nasal vowels are in the right column. The time at which the spectrum is sampled is indicated by the arrow. Two spectra are shown in each panel, each representing a spectrum sampled over a 25.6-ms Hamming window. The lower curve is the discrete Fourier transform, and the upper smooth curve is a "pseudo-spectrum" for which each point is obtained by performing a weighted average of groups of points on the DFT over a bandwidth of 300 Hz. Speaker VS (male).

Figure 4.14. spectrogrammes de cinq couples de voyelles orale et nasale homologues [hawl85].

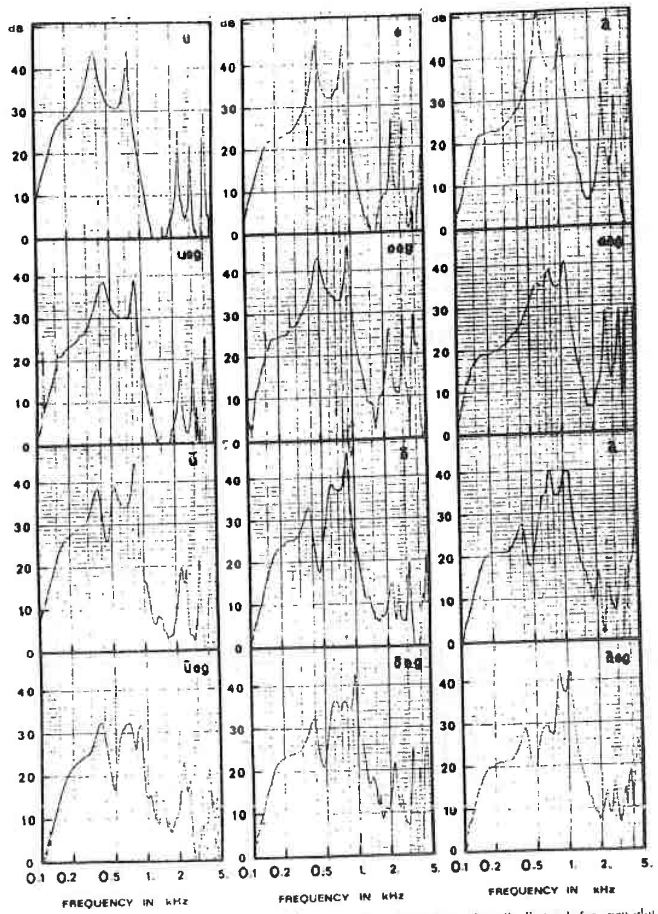


FIG. 15. Comparisons of nonnasalized, open glottis, and nasalized conditions, where "og" stands for open glottis

Figure 4.15. comparaison des spectres à la condition différente de l'articulation [fuji71].

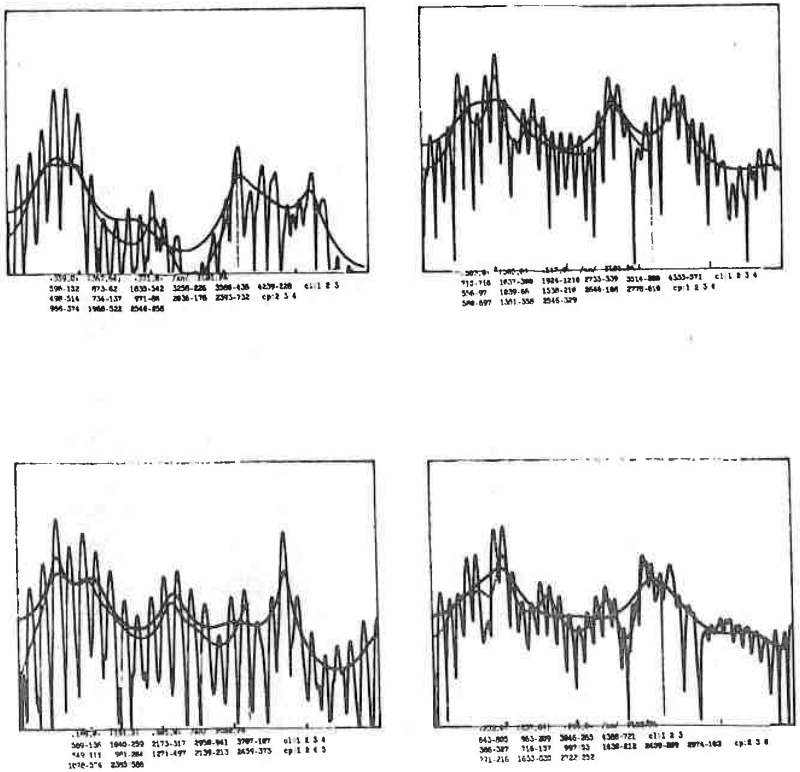


Figure 4.16. spectres de voyelle naturelle /ɑ̃/ (masculin).

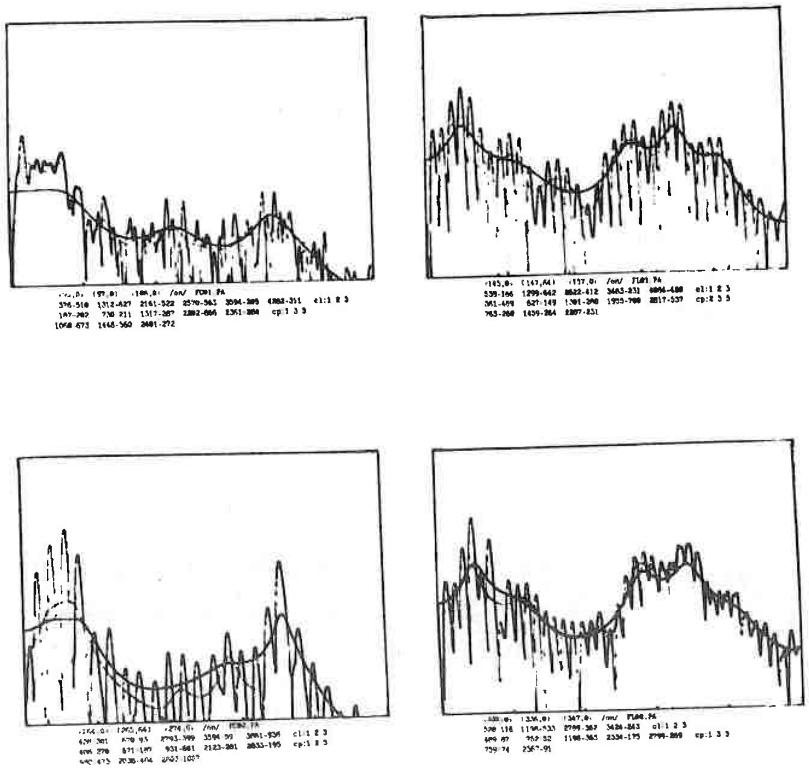


Figure 4.17. spectres de voyelle naturelle /ɔ̃/ (masculin).

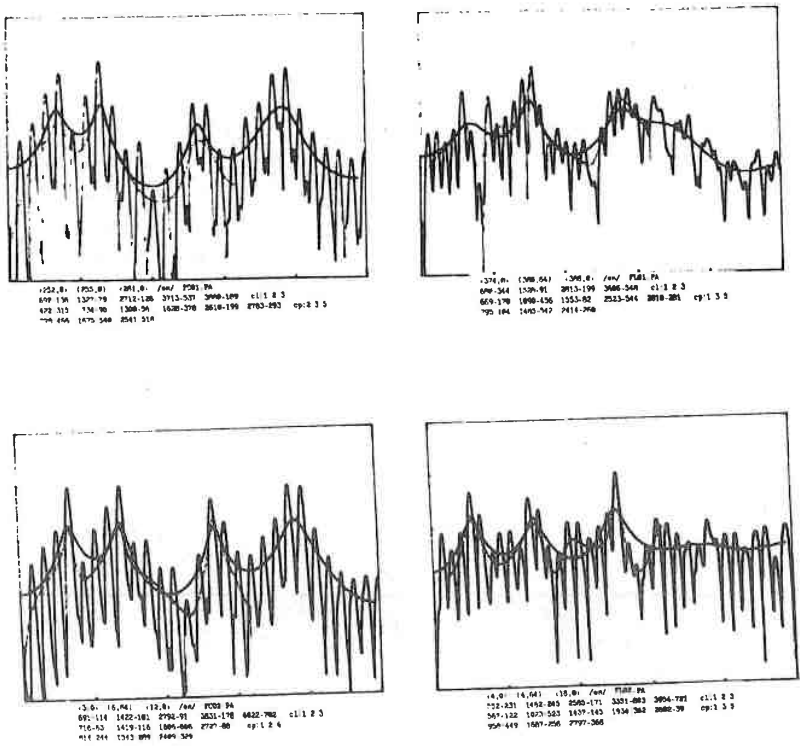


Figure 4.18. spectres de voyelle naturelle /ɛ̃/ (masculin).

4.4 Les indices de nasalisation

Les caractéristiques se manifestent principalement par les formants. Les fréquences de formants reflètent dans une certaine mesure la structure interne. Mais dans la plupart des cas, l'allure spectrale peut fournir également les informations concernant l'articulation. La raison est que l'interaction des résonateurs rend le spectre moins clair, parfois irrégulier. Donc nous distinguons deux genres d'indices: l'indice formantique et l'indice spectral.

4.4.1 Indices formantiques

Cette partie est un résumé des conclusions faites par F.Lonchamp[41,40,42]. Ce sont à notre connaissance les seuls documents qui décrivent entièrement les voyelles nasales françaises par les formants observés et théoriques.

Voyelle nasale /ɛ̃/

La voyelle /ɛ̃/ se caractérise principalement par un premier formant F_1' vers 500-700Hz, par un deuxième formant F_2' entre 1450 et 1600 Hz et par un troisième formant F_3' vers 3000Hz.

La fréquence de Fn_1 est difficile à identifier, il se confond souvent par F_1'. Fn_2 se trouve entre 2000-2200 Hz accompagné par un zéro (z_2) au dessus de 2500Hz.

Dans un spectre de /ɛ̃/, on trouve aussi un pic discret vers 1000Hz qui peut à l'occasion se confondre avec F'2, une irrégularité spectrale vers 1800-2000 Hz qui pourrait être causée par une dissymétrie des fosses nasales. Quant au pic de très basse fréquence (300Hz), il est probable qu'il représente soit le premier rebond spectral suivant le formant glottal, soit une résonance sinusale, soit les deux à la fois.

La figure4.19 montre un spectre représentatif de la voyelle nasale naturelle /ɛ̃/. La table 4.4.1 donne les formants(F) et les anti-formants(Z) obtenus à partir d'un modèle théorique par Mrayati.

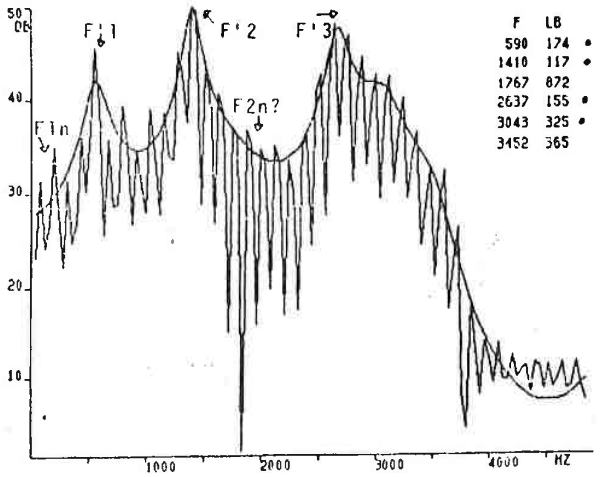


Figure 4.19. Identification des caractéristiques d'un /ɛ̃/ naturel.

Degré de couplage	Fn_1	Z_1	F_1'	F_2'	Fn_2	Z_2	F_3'
0.27 cm^2	450	500	575	1750	2250	2350	2450
1.66 cm^2	435	750	900	1750	2250	2350	2450
3.5 cm^2	415	1000	1250	1750	2250	2350	2525
5.0 cm^2	410	1100	1300	1750	2250	2350	2600

Tab 4.4.1: formants et anti-formants de /ɛ̃/

Voyelle nasale /ɑ̃/

La voyelle nasale /ɑ̃/ se caractérise par un premier formant 600-700 Hz et puis un ou deux formants. vers 800-1000 Hz, mélangés de F_2'. et Fn_1 ou F_1'. Ils sont souvent beaucoup plus intenses que le premier formant. F_3 est plus intense et à une fréquence élévée.

Les formants nasals sont moins visibles pour /ɑ̃/. Fn_2 est marqué par un formant discret vers 2200Hz.

La voyelle /ɑ̃/ a un indice de nasalité remarquable connu sous le nom d'*affaiblissement*

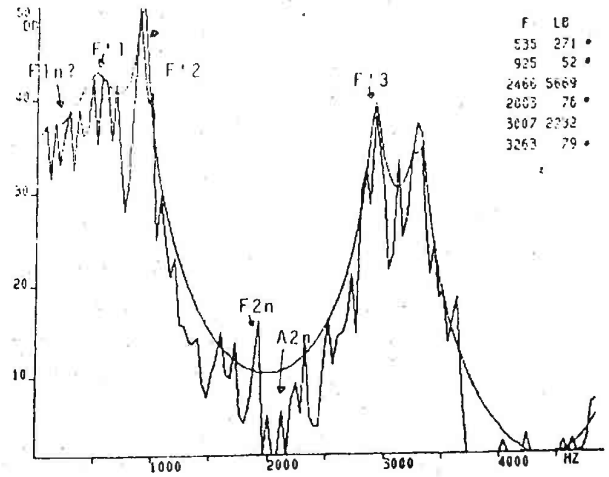


Figure 4.20. Identification des caractéristiques d'un /ɑ̃/ naturel.

de F_1.

La figure4.20 montre un spectre représentatif de la voyelle nasale naturelle /ɑ̃/. La table 4.4.2 donne les formants(F) et les anti-formants(Z) obtenus à partir d'un modèle théorique par Mrayati.

Degré de couplage	Fn_1	Z_1	F_1'	F_2'	Fn_2	Z_2	F_3'
0.27 cm^2	475	500	550	1250	2350	2350	2450
1.66 cm^2	460	650	800	1250	2250	2350	2550
3.5 cm^2	450	1000	1100	1250	2150	2350	2750
5.0 cm^2	450	1100	1200	1250	2050	2350	2850

Tab 4.4.2: formants et anti-formants de /ɑ̃/

Voyelle nasale /ɔ̃/

La voyelle nasale /ɔ̃/ est une voyelle ayant le spectre le plus difficile à interpréter. Les formants sont concentrés dans une zone fréquencielle inférieure à 1000 Hz. Elle ne présente

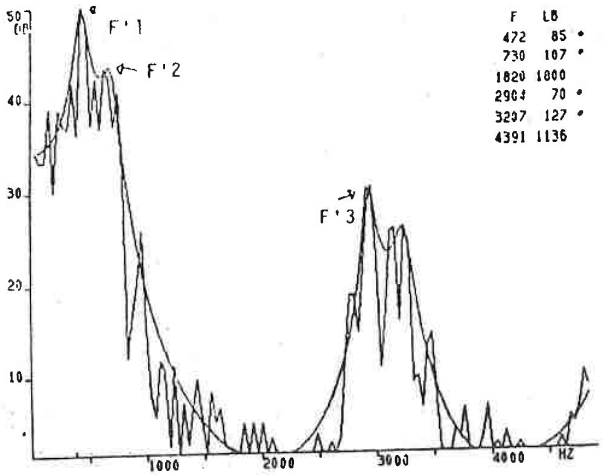


Figure 4.21. Identification des caractéristiques d'un /ɔ̃/ naturel.

souvent, sous 1000Hz, qu'un seul formant et une *épaule* sur le flanc droit de ce dernier.

On pourrait dire qu'il y a un ou deux formants à l'exclusion d'un formant vers 300Hz qui peut être, comme /ɛ̃/, d'une résonance sinusale et/ou d'un rebond du spectre glottal.

La figure4.21 montre un spectre représentatif de la voyelle nasale naturelle /ɔ̃/. La table 4.4.3 donne les formants(F) et les anti-formants(Z) obtenus à partir d'un modèle théorique par Mrayati.

Degré de couplage	Fn_1	Z_1	F_1'	F_2'	Fn_2	Z_2	F_3'
0.27 cm^2	450	500	600	1000	2350	2350	2500
1.66 cm^2	430	730	850	1000	2250	2350	2600
3.5 cm^2	415	1000	1000	1000	2100	2350	2750
5.0 cm^2	400	1100	1000	1000	2000	2350	2875

Tab 4.4.3: formants et anti-formants de /ɔ̃/

Strictement, la détermination de l'origine d'un formant est difficile. Par exemple, le

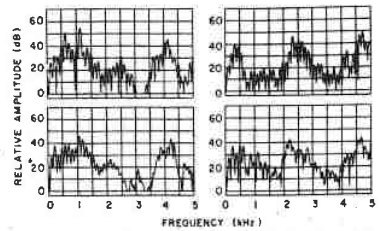


Fig. 1. Spectra of nonnasal and nasal vowels produced by a male speaker of Gujarati. The spectra at the left are sampled in the vowels in the words [kap] (top) and [kãp] (bottom). Those at the right are sampled in the vowels in the words [tɛ] (top) and [tɛ̃] (bottom). Spectra are calculated with a Hanning window of duration 26 msec. (Adapted from Hawkins and Stevens [14].)

Figure 4.22. spectres des voyelles non-nasales et nasales d'un locuteur masculin Gujarati (d'après [stev85]).

formant le plus bas pourrait être le formant nasal ou le formant oral déplacé en dépendant d'un certain critère fréquenciel[14]. Dans notre expérience plus loin, nous négligeons la nature des formants pour ne considérer que leur fréquence.

4.4.2 Indices spectraux

Les indices formantiques décrivent les voyelles nasales en distinguant et mesurant les fréquences de formant nasal et de formant oral déplacé. Pourtant les indices spectraux décrivent les évènements de nasalité plus grossiers, même statistiques. Ce sont les deux façons de décrire le même phénomène de *nasalisation*.

Quand une voyelle est nasalisée, il apparaît certaines conséquences acoustiques. La conséquence principale et la plus cohérente est une modification du spectre des sons à basse fréquence au-dessous de 1000Hz[79]. Cette modification est due à l'interaction des zéros et des pôles nasals ou additionnels qui *lissent* le spectre entre 0.5 et 1.5 KHz [42]. K.N.Stevens a utilisé le terme *proéminent* pour décrire les caractéristiques d'une voyelle nasale dans la zone de basse fréquence. Il a découvert que le degré de proéminence spectral diminue quand il y a la nasalité: le spectre devient plat et les vallées des deux côté de F_1 deviennent moins profondes (Fig.4.22).

L'un des effets de nasalité se manifeste sur la largeur de bande du formant. Les bandes passantes du premier formant des voyelles nasales sont nettement plus larges que celles correspondant aux voyelles orales (Fig.4.23)[57].

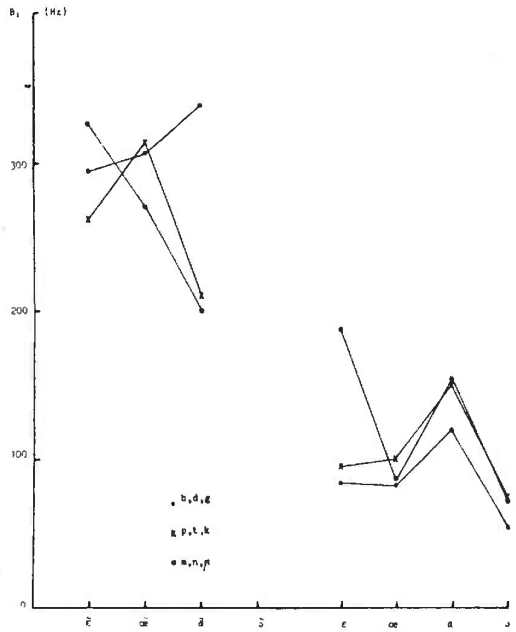


Figure 4.23. Bandes passantes du premier formant pour les voyelles nasales et orales correspondantes (d'après [Mray76]).

Un autre indice clair de la nasalité. est la fréquence élevée de F_3'[42].

J.R.Glass et V.W.Zue[17] ont utilisé la variabilité du spectre pour identifier si une voyelle est precédée ou suivie par une consonne nasale. Chaque spectre est découpé en trois segments: initial. milieu et final. Ils détectent si une voyelle est nasalisée par une comparaison des mesures dans les segments initial et final. Voici les mesures indicatives qu'ils ont choisies. L'entrainement du système est *par rotation*. Dans chaque étape, les prononciations de 5 locuteurs ont servi de références et les prononciations d'un locuteur restent a servi pour le test (6 locuteurs au total).

- *Centre de gravité*. Centre de gravité en moyenne dans le segment milieu.

- *déviation standard*. La valeur maximale de la déviation standard en moyenne.
- *Pourcentage maximum et minimum de résonance*. Le pourcentage maximum et minimum des cas où il y a une résonance additionnelle en basse fréquence.
- *Baisse maximum de résonance*. La valeur maximum de la baisse en moyenne entre le premier formant et le formant additionnel.
- *Différence minimum de résonance*. La valeur minimum de la différence en moyenne entre le premier formant et le formant additionel.

Ici la variabilité spectrale est considérée comme un indice de nasalisation. Mais la variabilité spectrale ne constitue pas un indice potentiel de nasalité pour l'ensemble des voyelles nasales du français[41].

Bien sûr, il y a d'autre indices importants.

- Les durées intrinsèques distinguent clairement les voyelles orales et nasales[41]. La durée des voyelles nasales est une à deux fois plus longue que celle des voyelles orales[57].
- Les voyelles nasales ont une fréquence intrinsèque très proche d'une voyelle à l'autre pour un locuteur. Or les voyelles orales ont des fréquences intrinsèques différentes[57].

4.5 Analyse-synthèse et perception de voyelles nasales

L'effet de la position relative entre pôle et zéro pour la nasalité peut être analysé par la perception du spectre généré de façon synthétique. Beaucoup de chercheurs [23,41,5,44,43] ont fait des expériences de ce genre. Nous citons ici seulement l'un des résultats de S.Hawkins et K.N.Stevens (Fig.4.24) pour voir comment le paramètre des pôle-zéros affectent la perception.

Prenons l'exemple d'un continuum de 9 voyelles entre [a – $\tilde{a}$]. Dans le cas de $a1$ (oral à la perception), le spectre se constitue par les pôles. Dans le cas de $a9$ (nasal à la perception), le premier formant se déplace de la fréquence 700Hz à 810Hz, le pôle nasal et le zéro nasal s'écartent. Cet éloignement de pôle et zéro nasals constitue dans la region de base fréquence un nouveau formant à la proximité d'une vallée, ce qui donne une perception nasale de cette voyelle.

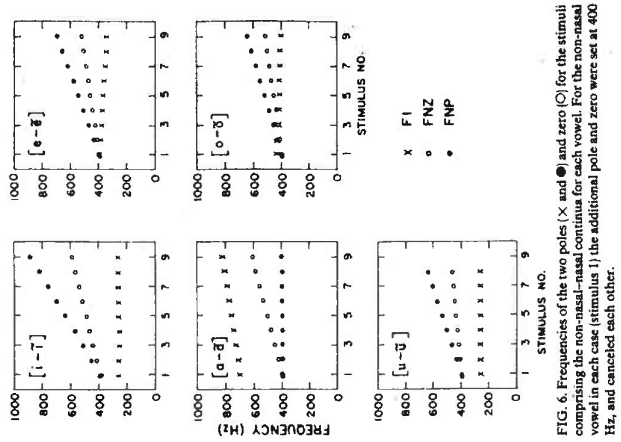


FIG. 6. Frequencies of the two poles (× and ●) and zero (○) for the stimuli comprising the non-nasal–nasal continua for each vowel. For the non-nasal vowel in each case (stimulus 1) the additional pole and zero were set at 400 Hz, and canceled each other.

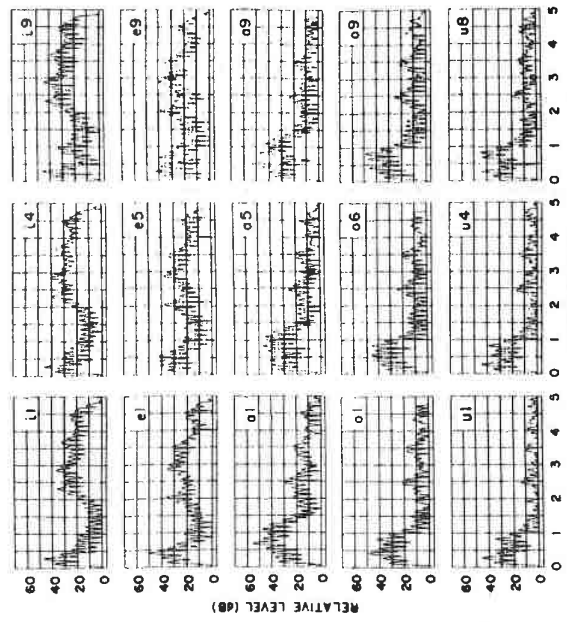


FIG. 7. Spectra of a number of the stimuli from the continua used in identification tests. Each row gives examples of spectra from one of the five vowels as shown. The left panel in a row represents the stimulus at the non-nasal end of the continuum, and the right panel represents the most nasal stimulus. The middle panel is the spectrum of the intermediate stimulus that is closest to the average 50% crossover point in the identification function for that vowel. The stimulus number is indicated in each panel. Spectra are discrete Fourier transforms calculated from the waveform weighted with a Hamming window of duration 26 ms.

Figure 4.24. Les positions des pôle-zéros et les spectres.

4.6 Reconnaissance des voyelles nasales

Nous avons vu que les caractéristiques d'une voyelle nasale présentent souvent une irrégularité due au du degré de couplage et aux cavités supplémentaires (les sinus). Les formants supplémentaires troublent l'observation normale du spectre. Donc il faut prendre les indices les plus stables et représentatifs comme référence.

Notre premier travail est de voir l'influence des zéros avec un modèle tout pôle sur un petit corpus des mots isolés. Le deuxième est de faire la reconnaissance par formants en utilisant un corpus de parole continue.

4.7 Première expérience

Le modèle tout pôle ne peut pas prendre en compte toutes les caractéristiques d'une voyelle nasale si son ordre est choisi de la même manière que pour une voyelle orale. Donc si nous faisons la comparaison du résultat fourni par un modèle tout pôle avec celui fourni par un modèle pôle-zéro, la différence entre eux doit être plus grande pour une nasale que pour une orale, surtout dans la région inférieure à la fréquence de deuxième formant.

4.7.1 Description du corpus

Nous travaillons sur un petit corpus de mots isolés. Les mots sont choisis de telle façon que la consonne touche le moins possible le caractère de la voyelle suffixe. Il y a treize locuteurs masculins et deux listes de mots dans un ordre différent. Chaque liste est prononcée une seule fois. Les enregistrements du corpus ont été effectués à 10 kHz sur 12 bits. Il est constitué d'un corpus de 13*(13+12) mots (il y a un enregistrement abimé). Voici les deux listes de mots:

Liste No 1	Liste No 2
PIRE	PAIN
PEAU	PEPE
PEUR	PEUR
PONT	PUS
PERE	PAR
PEPE	PORT
PORT	PAN
PEU	PERE
PAIN	PEAU
PAR	PONT
PUS	PIRE
POUX	PEU
PAN	POUX

4.7.2 Position de la fenêtre d'analyse

Pour mieux observer la différence manifestée par les résultats du modèle AR et modèle ARMA, nous positionnons pour chaque mot une fenêtre d'analyse dans la région vocalique. La position est déterminée selon la courbe d'énergie.

On peut se servir de la courbe d'énergie pour fixer un point d'analyse, c'est le moment quand l'énergie atteint sa valeur maximum (fig4.25a). Mais ce maximum fournit souvent une fausse indication. Par exemple, le plosive /*b*/ dans le mot *beau* faisant souvent avancer le point maximum d'énergie vocalique, si ce point est utilisé, il existera encore un résidu de /*b*/ qui perturbe le caractère de la voyelle (fig.4.25b).

Pour le cas (b) dans la figure 4.25, le meilleur moment d'analyse est de $100 \sim 160ms$ après le maximum, si la durée est suffisante ??. Nous vérifions les durées de mot dans notre corpus, et nous remarquons que les mots ne sont pas tous assez longs. Le meilleur moment d'analyse sera en dehors de cette plage pour certains mots. Par conséquent, nous fixons la position de la fenêtre entre le point de l'énergie maximum et le point du spectre le plus stable:

- calculer le moment t_{max} quand la courbe atteint sa valeur maximale
- calculer le moment t_{stable} quand la dérivée du spectre soit minimale
- alors $t_{meilleur} = (t_{max} + t_{stable})/2$

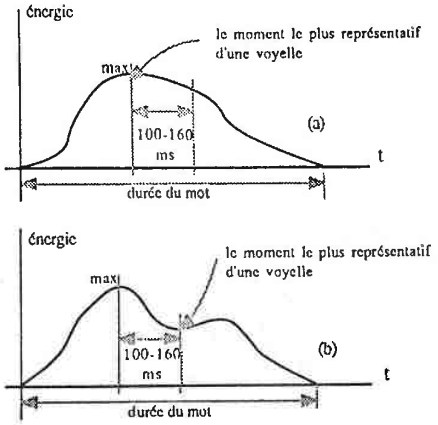


Figure 4.25. maximum et position d'analyse.

4.7.3 Choix de modèle et traitement du signal

La fenêtre d'analyse est de type Hamming avec une longeur de 25.6ms. Le signal dans la fenêtre est accentué avec un coefficient de 0.9. Nous calculons les formants par le modèle de LP tout pôle, qui a un ordre de 14, et également les formants par ARMA dans l'intervalle 0 – 3200Hz avec r=40, q=8, p=12. Comme le montrent les figures, les résultats (deux courbes dans la Fig. 4.26) donnés par ces deux modèles sont cohérents car les formants se manifestent de la même forme, alors que ceux de figure 4.27 sont incohérents.

D'abord nous sélectionnons comme formants les pics fournis par modèle AR, qui satisfont la condition en fréquence d'être des pics maximaux, et celle temporelle: de continuité. Nous prenons ces formants comme les références pour chaque fenêtre d'analyse de ARMA.

Pour chaque formant de référence i, nous le mettons en correspondance avec un des pics k fournis par le modèle ARMA. La distance entre deux fréquences formantiques est absolue, et celle entre deux largeurs des bandes est la valeur absolue pondérée:

$$dis[(f,b)^i_{ar}, (f,b)^k_{arma}] = |f^i_{ar} - f^k_{arma}| + \lambda(b^i_{ar}, b^k_{arma})|b^i_{ar} - b^k_{arma}| \qquad (4.1)$$

ici λ est une pondération favorisant les largeurs étoites:

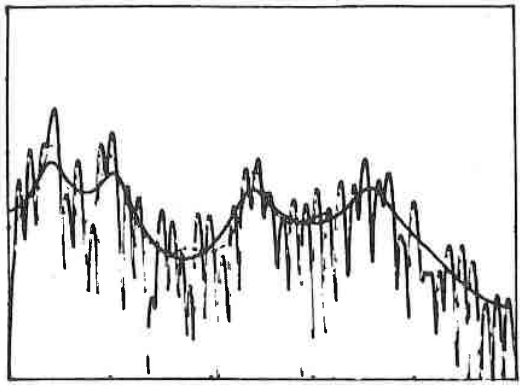


Figure 4.26. cas cohérent.

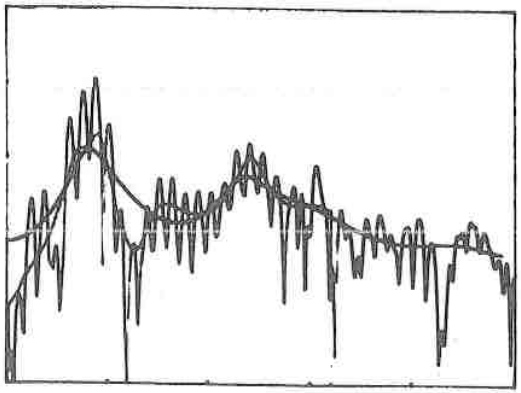


Figure 4.27. cas incohérent.

$$\lambda(x,y) = \begin{cases} -1 & \text{si } x-y < 0 \\ 0.5 & \text{si } 0 \le x-y < 300 \\ 0.25 & \text{si } 300 \le x-y < 500 \\ 0.1 & \text{si } x-y \ge 500 \end{cases} \quad (4.2)$$

Notons: *pli* l'indice de i_{eme} formant de AR. *ppi* l'indice de i_{eme} formant de ARMA, f^i_{ar}, f^i_{arma} la fréquence du i_{eme} formant; b^i_{ar}, b^i_{arma} la largeur du i_{eme} formant. Les deux spectres de AR et de ARMA ne sont pas cohérents si:

(1) $pp2 - pp1 > 1$ et $b^{pp2}_{arma} \ge b^{pp1+1}_{arma}$ et $f^{pp1}_{arma} > 390$

c'est à dire qu'il y a au moins un pic de ARMA qui n'a pas été détecté par le modèle AR. Si sa largeur de bande est moins étoite par rapport au deuxième formant marqué de ARMA, et si le 1er formant marqué de ARMA est supérieur à 390Hz. alors on dira que le résultat AR n'est pas cohérent avec celui ARMA.

(2) si l'un des deux premiers formants ARMA correspond à un formant AR ayant une largeur de bande importante, et si $b^{pp1}_{arma} > b^{pp2}_{arma}$ alors il y a incohérence, si on trouve un pic ARMA inférieur à f^{pp1}_{arma} ayant une largeur de bande étoite, alors également il y a incohérence.

La première condition est pour le cas où le 1er pic est plat, la deuxième est pour le cas où le 2ème pic est plat.

Ces deux conditions ne décrivent qu'une déformation grossière du spectre AR par rapport à celui ARMA.

4.8 Résultat de la première expérience

Avec ces deux critères de déformation, nous avons testé tous les mots dans le corpus. Le résultat est résumé dans la table 4.8.1. La première ligne est les symboles des phonèmes. La deuxième est le nombre de mots où l'incohérence est notée. La troisième le nombre de chaque phonème dans le corpus. Dans le tableau, il y a treize /ɔ̃/ qui sont marqués parmi 25, il n'y a que quatre /ɑ̃/ marqués, quant à /ɛ̃/, le résultat est zéro. Pourtant nous avons un /ɔ/, deux /o/ et deux /ø/ qui sont marqués incohérents.

voyelle	ɔ̃	ɑ̃	ɛ̃	i	e	ɛ	a	ɔ	o	u	y	ø	œ
incohérence	13	4	0	0	0	0	0	1	2	0	0	2	0
Nomb. de voyelles	25	25	25	25	25	25	25	25	25	25	25	25	25

Tab 4.8.1: incohérence entre AR et ARMA

4.9 Conclusion

Vu le tableau de résultats, nous remarquons que:

- l'incohérence entre spectres AR et celui ARMA se manifeste surtout pour le phonème /ɔ̃/.
- l'indication de l'incohérence n'est pas suffisante pour distinguer nasale et orale. C'est peut-être à cause de la grossièreté de la description.
- pour les voyelles orales, les spectres AR et ARMA sont assez proches bien que l'ordre éventuel de ARMA est beaucoup plus grand que celui de AR.

4.10 Deuxième expérience

C'est une expérience en reconnaissance des voyelles sur la parole continue en utilisant le modèle ARMA.

La variation multilocuteur due à l'accent affecte la source du spectre, mais elle affecte peu les fréquences de formant. En plus, la dégradation du signal introduite dans la transmission préserve les fréquences de formant [27]. Caraty[9] et Hunt[27] ont proposé une distance interspectrale intégrant la notion de formants. En tenant compte de l'avantage évident des formants et du fait que les fréquences de formant sont une base de la perception humaine, nous utilisons une distance interspectrale. Pour réduire l'effet de la personnalité et pour refléter certaines caractéristiques de la perception humaine, notre mesure de distance ne se base que sur les fréquences de formant, l'amplitude de formant, étant complètement ignorée.

Nous avons choisi les trois premiers formants inférieurs à 3200Hz, quelque soit le formant oral ou le formant déplacé à cause de la nasalisation, même si le formant est en réalité un mélange de formants. Nous utilisons la méthode décrite dans le chapitre précédent pour l'extraction des formants.

4.10.1 Stratégie de reconnaissance

Le processus de reconnaissance devient simple grâce au choix des formants comme paramètres.

— *Références de reconnaissance*

Notre reconnaissance est multilocuteur, par conséquent, les références doivent être universelles. Comme la description d'extraction des formants dans le chapitre précédent, nous prenons les moyennes des trois premiers formants et leur largeur de bande en parcourant les fichiers de parole. Dans notre bibliothèque de références, chaque voyelle (orale ou nasale) est associée aux trois formants et trois largeurs de bande.

— *Formes inconnues*

Etant donné un segment de la parole, (la parole est segmentée a priori, chaque segment représente une voyelle, y compris les transitions), les formes inconnues sont les vecteurs contenant 3 formants et 3 largeurs de bande, qui sont obtenus par le même processus de traitement numérique à court terme et de l'extraction de formant.

— *Mesure de distance et Décision*

La mesure de distance entre référence et inconnue est de type interspectrale. En utilisant les six paramètres (3 formants et 3 largeurs de bande), nous construisons un spectre logarithmique soit en échelle *linéaire*, soit en échelle de *barks*. La décision est faite par une comparaison des spectres inconnus avec les références, on garde pour chaque segment les trois candidats, qui donnent les formes reconnues en comptant la fois la plus fréquentées.

4.10.2 Description du corpus

Les enregistrements du corpus (GRECO *Communication Parlée*) ont été effectués à 16 kHz sur 16 bits dans d'excellentes conditions (chambre sourde et dynamique importante). il est constitué d'un texte de 120 mots.

Voici le texte du corpus:

La bise et le soleil se disputaient,
chacun assurant qu'il était le plus fort,
quand ils ont vu un voyageur qui s'avancait,
enveloppé dans son manteau.
Ils sont tombés d'accord,

que celui qui arriverait le premier à faire oter son manteau au voyageur,
serait regardé comme le plus fort.
Alors la bise s'est mise à souffler de toute sa force;
mais plus elle soufflait,
plus le voyageur serrait son manteau autour de lui,
et à la fin la bise a renoncé à le lui faire oter.
Alors le soleil a commencé à briller,
et au bout d'un moment le voyageur, réchauffé, a oté son manteau.
Ainsi la bise a dû reconnaitre que le soleil était le plus fort des deux.

et les transcriptions phonétiques (la prononciation du locuteur FL):

l a b i z e l ø s ɔ l ɛ j s ø d i s p y t ɛ,
ʃ a k œ̃ a s y R ã k i l e t ɛ l ø p l y F ɔ R,
k ã t i l z ɔ̃ v y œ̃ v w a j a ʒ œ R k i s a v ã s ɛ,
ã v l o p e d ã s ɔ̃ m ã t o.
i l s ɔ̃ t ɔ̃ b e d a k ɔ R,
k ø s ø l t i k i a R i v R ɛ l ø p R ø m j e a F ɛ R o t e s ɔ̃ m ã t o o v w a j a ʒ œ R,
s ø R ɛ R ø g a R d e k ɔ m l ø p l y F ɔ R.
a l ɔ R l a b i z(= s) s e m i z a s u F l e d ø t u (t) s e F ɔ R s,
m ɛ p l y z ɛ l s u F l ɛ,
p l y l ø v w a j a ʒ œ R s ɛ R ɛ s ɔ̃ m ã t o o t u R d ø l y i,
e a l a F ɛ̃ l a b i z a R ø n ɔ̃ s e a l ø l y i F ɛ R o t e.
a l ɔ R l ø s ɔ l ɛ j a k ɔ m ã s e a b R i j e,
e o b u d œ̃ m o m ã l ø v w a j a ʒ œ R R e ʃ o F e a o t e s ɔ̃ m ã t o,
ɛ̃ s i l a b i z a d y R ø k ɔ n e t R k o l ø s ɔ l ɛ j e t ɛ l ø p l y F ɔ R d e d ø.

Le corpus a été étiqueté manuellement par un expert phonéticien (F.Lonchamp) à l'aide de la visualisation spectrographique sur *Masscomp*. Le corpus comprend 24 voyelles nasales et 152 voyelles orales, donc 176 voyelles au total. Dans notre étude, aucune distinction n'est faite entre /ɛ̃/ et /œ̃/. non plus entre /ø/ et /ə/. donc il y a 13 voyelles différentes. Chaque phonème est repéré par son début et sa fin dans le signal temporel comme montré par la figure4.28. La sélection des segments correpondant aux voyelles se fait de manière automatique à partir des fichiers d'étiquettes.


```
‹   1,  21 ›  /#/        ‹ 343, 348 ›  /*/        ‹ 632, 635 ›  /*/
‹  21,  29 ›  /l/        ‹ 348, 358 ›  /un/       ‹ 635, 644 ›  /an/
‹  29,  39 ›  /a/        ‹ 358, 370 ›  /a/        ‹ 644, 647 ›  /t/
‹  39,  52 ›  /b/        ‹ 370, 386 ›  /s/        ‹ 647, 654 ›  /*/
‹  52,  67 ›  /i/        ‹ 386, 395 ›  /y/        ‹ 654, 662 ›  /i/
‹  67,  81 ›  /z/        ‹ 395, 403 ›  /R/        ‹ 662, 669 ›  /l/
‹  81,  87 ›  /ø/        ‹ 403, 416 ›  /an/       ‹ 669, 677 ›  /z/
‹  87,  91 ›  /l/        ‹ 416, 426 ›  /k/        ‹ 677, 694 ›  /on/
‹  91, 100 ›  /&/        ‹ 426, 429 ›  /*/        ‹ 694, 704 ›  /v/
‹ 100, 116 ›  /s/        ‹ 429, 436 ›  /i/        ‹ 704, 718 ›  /y/
‹ 116, 124 ›  /l/        ‹ 436, 440 ›  /l/        ‹ 718, 731 ›  /un/
‹ 124, 127 ›  /l/        ‹ 440, 448 ›  /ai/       ‹ 731, 738 ›  /r/
‹ 127, 138 ›  /ai/       ‹ 448, 454 ›  /t/        ‹ 738, 742 ›  /u/
‹ 138, 149 ›  /j/        ‹ 454, 461 ›  /*/        ‹ 742, 749 ›  /a/
‹ 149, 165 ›  /s/        ‹ 461, 469 ›  /ai/       ‹ 749, 758 ›  /j/
‹ 165, 173 ›  /&/        ‹ 469, 474 ›  /l/        ‹ 758, 768 ›  /a/
‹ 173, 179 ›  /d/        ‹ 474, 483 ›  /&/        ‹ 768, 778 ›  /gh/
‹ 179, 182 ›  /*/        ‹ 483, 492 ›  /p/        ‹ 778, 793 ›  /oe/
‹ 182, 189 ›  /l/        ‹ 492, 495 ›  /*/        ‹ 793, 801 ›  /R/
‹ 189, 200 ›  /s/        ‹ 495, 499 ›  /l/        ‹ 801, 808 ›  /k/
‹ 200, 207 ›  /p/        ‹ 499, 507 ›  /y/        ‹ 808, 811 ›  /*/
‹ 207, 209 ›  /*/        ‹ 507, 522 ›  /f/        ‹ 811, 817 ›  /l/
‹ 209, 218 ›  /y/        ‹ 522, 544 ›  /)/        ‹ 817, 830 ›  /s/
‹ 218, 229 ›  /l/        ‹ 544, 557 ›  /R/        ‹ 830, 841 ›  /a/
‹ 229, 233 ›  /*/        ‹ 557, 570 ›  /#/        ‹ 841, 847 ›  /v/
‹ 233, 254 ›  /ai/       ‹ 570, 574 ›  /!/        ‹ 847, 860 ›  /an/
‹ 254, 312 ›  /#/        ‹ 574, 601 ›  /#/        ‹ 860, 878 ›  /s/
‹ 312, 326 ›  /ch/       ‹ 601, 619 ›  /hh/       ‹ 878, 902 ›  /ai/
‹ 326, 337 ›  /a/        ‹ 619, 625 ›  /#/        ‹ 902, 952 ›  /#/
‹ 337, 343 ›  /k/        ‹ 625, 632 ›  /k/

                                  ‹LABISA.FL›

‹   1,  28 ›  /#/        ‹ 356, 364 ›  /k/        ‹ 690, 694 ›  /*/
‹  28,  28 ›  /l/        ‹ 364, 368 ›  /*/        ‹ 694, 700 ›  /l/
‹  28,  36 ›  /a/        ‹ 368, 387 ›  /un/       ‹ 700, 704 ›  /l/
‹  36,  49 ›  /b/        ‹ 387, 397 ›  /a/        ‹ 704, 715 ›  /z/
‹  49,  50 ›  /*/        ‹ 397, 414 ›  /s/        ‹ 715, 731 ›  /on/
‹  50,  69 ›  /l/        ‹ 414, 424 ›  /y/        ‹ 731, 742 ›  /v/
‹  69,  82 ›  /z/        ‹ 424, 432 ›  /R/        ‹ 742, 759 ›  /y/
‹  82,  92 ›  /ø/        ‹ 432, 447 ›  /an/       ‹ 759, 765 ›  /?/
‹  92,  97 ›  /l/        ‹ 447, 455 ›  /k/        ‹ 765, 779 ›  /un/
‹  97, 106 ›  /&/        ‹ 455, 461 ›  /*/        ‹ 779, 786 ›  /v/
‹ 106, 121 ›  /s/        ‹ 461, 468 ›  /i/        ‹ 786, 790 ›  /w/
‹ 121, 131 ›  /)/        ‹ 468, 472 ›  /l/        ‹ 790, 794 ›  /a/
‹ 131, 137 ›  /l/        ‹ 472, 483 ›  /ai/       ‹ 794, 806 ›  /j/
‹ 137, 149 ›  /ai/       ‹ 483, 492 ›  /t/        ‹ 806, 818 ›  /a/
‹ 149, 156 ›  /j/        ‹ 492, 496 ›  /*/        ‹ 818, 829 ›  /gh/
‹ 156, 175 ›  /s/        ‹ 496, 509 ›  /ai/       ‹ 829, 847 ›  /oe/
‹ 175, 183 ›  /&/        ‹ 509, 512 ›  /l/        ‹ 847, 859 ›  /R/
‹ 183, 191 ›  /d/        ‹ 512, 521 ›  /&/        ‹ 859, 865 ›  /k/
‹ 191, 192 ›  /*/        ‹ 521, 529 ›  /p/        ‹ 865, 868 ›  /*/
‹ 192, 200 ›  /i/        ‹ 529, 532 ›  /*/        ‹ 868, 874 ›  /l/
‹ 200, 210 ›  /s/        ‹ 532, 535 ›  /l/        ‹ 874, 889 ›  /s/
‹ 210, 216 ›  /p/        ‹ 535, 543 ›  /y/        ‹ 889, 899 ›  /a/
‹ 216, 219 ›  /*/        ‹ 543, 560 ›  /f/        ‹ 899, 905 ›  /v/
‹ 219, 228 ›  /y/        ‹ 560, 581 ›  /)/        ‹ 905, 918 ›  /an/
‹ 228, 239 ›  /l/        ‹ 581, 594 ›  /R/        ‹ 918, 935 ›  /s/
‹ 239, 243 ›  /*/        ‹ 594, 660 ›  /#/        ‹ 935, 960 ›  /ai/
‹ 243, 274 ›  /ai/       ‹ 660, 666 ›  /k/        ‹ 960, 965 ›  /hh/
‹ 274, 330 ›  /#/        ‹ 666, 673 ›  /*/        ‹ 965, 986 ›  /#/
‹ 330, 345 ›  /ch/       ‹ 673, 684 ›  /an/
‹ 345, 356 ›  /a/        ‹ 684, 690 ›  /t/

                                  ‹LABISA.GW›

/&/ : e muet      /*/ : burst      /hh/ : souffle      /#/ : pause
                  /??/ : inconnu   /!/ : bruit

Etiquetage manuel de la phrase : "La bise et le soleil se disputaient, chacun
assurant qu'il était le plus fort, quand ils ont vu un voyageur qui s'avancait".
```

Figure 4.28. Etiqetage manuel du corpus servant de test.

4.11 Résultat de la deuxième expérience

Nous avons testé sept fichiers de parole: BG, BP, BT, FL, FC, GM et JB. La mesure de distance est soit linéaire, soit en Barks. De plus, nous avons également montré le résultat de la reconnaissance par l'analyse à long terme.

4.11.1 Distance en échelle linéaire

La mesure de distance en échelle linéaire est une distance absolue de spectre. Si la fonction du filtre inverse est représentée par $H(z)$ qui est synthétisé à partir des formants, et le spectre représenté par $SPEC(k)$ obtenu par $DFT[H(z)]$ (N points), la distance entre un inconnu et une référence est de la forme suivante:

$$d = \sum_{k=0}^{N-1} |SPEC_{inconnu}(k) - SPEC_{refe}(k)|$$

Pour pouvoir regarder le taux de reconnaisance de chaque locuteur, les sept tables 4.11.1-4.11.7 donnent les résultats individuels, et la table 4.11.8 donne le résultat total. La table 4.11.9 montre les formants moyens de /ã/, /ɔ̃/ et /ɛ̃/.

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	17	17	20	30	13	13	5	9	4	25
1cand	4	1	5	14	11	7	6	0	3	0	6	4	4
2cand	7	7	5	17	14	13	15	7	9	1	7	4	6
3cand	9	8	5	17	15	16	15	11	12	1	8	4	11
taux	82%	100%	100%	100%	88%	80%	50%	85%	92%	20%	89%	100%	44%

Tab 4.11.1: reconnaissance des voyelles (locuteur: BG)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	17	19	16	30	13	13	5	9	4	25
1cand	7	5	0	12	9	2	8	4	6	1	5	3	10
2cand	10	6	1	16	16	9	15	8	10	1	6	4	15
3cand	10	6	1	17	17	13	17	8	11	1	8	4	16
taux	91%	100%	20%	100%	89%	87%	57%	62%	85%	20%	89%	100%	64%

Tab 4.11.2: reconnaissance des voyelles (locuteur: BP)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	16	16	20	30	14	12	5	9	4	30
1cand	8	0	0	10	4	6	5	3	10	0	2	1	6
2cand	9	5	0	16	6	9	16	6	12	0	3	3	12
3cand	9	7	1	16	11	11	20	9	12	1	9	3	11
taux	82%	88%	20%	100%	61%	55%	67%	64%	100%	20%	100%	75%	47%

Tab 4.11.3: reconnaissance des voyelles (locuteur: BT)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	16	16	19	31	13	14	5	9	5	21
1cand	8	1	2	12	14	8	3	3	10	0	4	4	11
2cand	11	6	2	15	16	16	13	6	12	0	6	4	17
3cand	11	8	2	16	16	16	19	7	13	2	6	4	16
taux	100%	100%	40%	100%	100%	84%	61%	56%	93 %	40%	67%	100%	76%

Tab 4.11.4: reconnaissance des voyelles (locuteur: FL)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	9	4	16	17	20	31	14	12	5	9	4	22
1cand	10	4	2	8	12	8	13	1	11	0	5	2	12
2cand	10	6	4	10	13	13	19	5	12	0	6	4	13
3cand	11	8	4	14	16	16	22	8	12	0	9	4	16
taux	91%	89%	100%	88%	94%	80%	71%	57%	100%	0%	100%	100%	82%

Tab 4.11.5: reconnaissance des voyelles (locuteur: FC)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	17	16	19	31	13	13	5	9	4	22
1cand	8	3	4	14	13	6	8	3	12	1	3	3	14
2cand	10	2	4	17	17	13	16	6	12	1	6	4	17
3cand	11	2	4	17	16	14	19	9	12	2	6	4	19
taux	100%	25%	80%	100%	100%	74%	61%	69%	92%	40%	89%	100%	86%

Tab 4.11.6: reconnaissance des voyelles (locuteur: GM)

voyelle	ã	ɔ̃	ɛ̃/œ̃	i	ɛ	e	a	ɔ	o	u	y	ɔ/œ	ə
nombre	11	8	5	17	19	16	31	12	13	5	9	4	22
1cand	5	1	4	14	13	4	3	2	10	0	7	0	14
2cand	8	2	4	16	16	8	6	7	11	1	8	2	15
3cand	10	3	4	16	16	10	13	9	13	3	8	4	15
taux	91%	38%	80%	94%	95%	56%	42%	75%	100%	60%	89%	100%	68%

Tab 4.11.7: reconnaissance des voyelles (locuteur: JB)

voyelle	$\tilde{a}$	$\tilde{\jmath}$	$\tilde{\varepsilon}/\tilde{\text{œ}}$	$\tilde{\imath}$	ϵ	ε	a	$\jmath$	o	u	y	$o/\text{œ}$	∂
nombre	77	57	34	116	126	134	214	91	90	35	63	26	166
1cand	48	17	17	84	76	81	48	19	64	27	52	17	71
2cand	66	34	20	107	102	91	100	47	76	4	44	25	91
3cand	71	44	21	115	115	95	125	61	85	10	56	27	109
taux	92%	77%	62%	97%	90%	71%	56%	67%	94%	29%	89%	96%	65%

Tab 4.11.8: reconnaissance des voyelles de sept locuteurs

	$\tilde{a}$			$\tilde{\jmath}$			$\tilde{\varepsilon}$		
	f1-b1	f2-b2	f3-b3	f1-b1	f2-b2	f3-b3	f1-b1	f2-b2	f3-b3
BG	697-260	1032-115	2592-109	535-312	1021-119	2733-176	664-259	1550-105	2530-221
BP	436-94	967-223	2761-192	419-124	1111-197	2522-214	149-91	1474-123	2547-137
BT	463-161	957-131	2533-154	483-126	940-170	2485-177	436-149	1392-106	2370-136
FL	557-180	1059-127	2731-160	459-184	1062-160	2666-226	533-136	1407-102	2526-160
FC	491-254	986-146	2630-232	609-225	1067-156	2700-205	658-165	1484-177	2463-162
GM	527-267	961-150	2493-242	622-210	1149-163	2463-232	604-176	1436-118	2508-127
JB	578-200	1060-176	2707-154	604-177	1300-231	2650-160	554-196	1498-126	2617-138

Tab 4.11.9: les trois premiers formants de voyelle nasale

4.11.2 Distance en échelle de Barks

Nous avons testé les fichiers de parole en utilisant une distance en échelle de Bark. Le spectre synthétisé des formants est découpé en 32 bandes critiques (fig.4.29). L'intérêt de l'approximation de cette échelle est qu'on élimine une bonne part des variantes inutiles du spectre pour ne conserver que celles qui ont un intérêt perceptif [16]. Malgré l'avantage non-linéaire en échelle de Barks, nos résultats ne sont pas meilleurs que ceux en échelle linéaire. La raison est que le spectre obtenu par formants est tellement lissé que la variation est réduite au minimum. Nous montrons ici dans la table 4.11.12 seulement les résultats moyens.

4.11.3 Reconnaissance par l'analyse à long terme

La prononciation d'une voyelle nasale n'est pas une combinaison simple de la prononciation nasale et de celle orale. Nous pouvons examiner une voyelle nasale sous deux aspects: les zéros et la transition.

- La connexion du conduit nasal affecte directement le spectre d'une voyelle. Les experts phonéticiens estiment que cette connexion introduit des zéros, et ils espèrent qu'il est possible de les retrouver par une approximation mathématique dans certaines régions prévues. Mais la réalité est qu'il est difficile d'observer à l'aide du spectre leur distribution clairement et sûrement à cause de la complexité de parole et la limite de l'analyse.

- Une voyelle nasale résulte de l'effet commun du conduit oral et du conduit nasal. Il existe probablement une transition de cette voyelle nasale soit vers la nasalité pure, ce qui est le cas de la consonne nasale, soit vers le son oral. Si cette transition peut servir comme une indication de la nasalisation, comment mesure-t-on son évolution? Une façon est de l'évaluer fenêtre par fenêtre (à court terme). Mais le résultat obtenu est difficile à interpréter dans la plupart des cas. On ne sait pas très bien si un pic appartient à un formant oral, ou à un formant nasal, ou est seulement une déformation du spectre. Par conséquent, il n'est pas pratique de détecter la transition selon l'évolution des formants au cours du temps.

Nous devons utiliser toutes les indications possibles dans la *durée* d'une voyelle nasale. On sait très bien que pour qu'un son puisse être perçu et être compris, l'information caractéristique doit garder une certaine durée et rester stable. C'est à dire, si nous faisons une analyse à long terme, le spectre contiendra toutes les autres déformations utiles, et les caratéristiques principales grâce à l'accumulation de l'information stable.

N°	f_1	f_2
1	50.06	151.57
2	100.46	203.72
3	151.57	257.97
4	203.72	313.66
5	257.97	370.19
6	313.66	430.20
7	370.19	493.30
8	430.20	559.91
9	493.30	630.33
10	559.91	705.12
11	630.33	784.61
12	705.12	869.96
13	784.61	961.15
14	869.96	1059.02
15	961.15	1164.25
16	1059.02	1277.57
17	1164.25	1399.76
18	1277.57	1531.68
19	1399.76	1674.25
20	1531.68	1828.45
21	1674.25	1995.35
22	1828.45	2176.12
23	1995.35	2372.00
24	2176.12	2584.37
25	2372.00	2814.70
26	2584.37	3064.59
27	2814.70	3335.77
28	3064.59	3630.12
29	3335.77	3949.70
30	3630.12	4296.73
31	3949.70	4673.61
32	4296.73	5082.97

Figure 4.29. Les 32 bandes critiques.

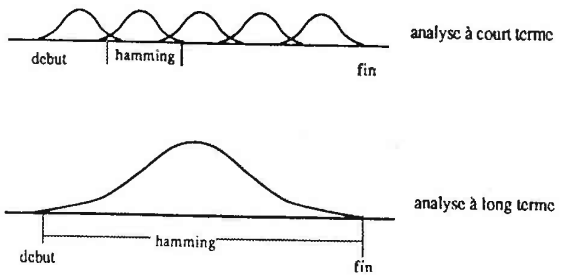


Figure 4.30. Fenêtres d'analyse à long terme et à court terme.

L'analyse à long terme ici consiste à prendre un segment complet d'une voyelle (fig. 4.30). la largeur de la fenêtre d'analyse est la durée de la voyelle qui est étiquetée manuellement comme décrit précédemment. Le traitement ensuite reste le même que pour l'analyse à court terme.

Voici les formants en moyenne de voyelles nasales de chaque locuteur:

	$\tilde{a}$			$\tilde{\jmath}$			$\tilde{\varepsilon}$		
	f1-b1	f2-b2	f3-b3	f1-b1	f2-b2	f3-b3	f1-b1	f2-b2	f3-b3
BG	691-205	949-85	2598-200	599-350	943-88	2688-172	702-249	1576-105	2376-114
BP	439-70	832-143	2600-174	407-94	976-227	2574-194	435-71	1406-169	2573-127
BT	456-111	947-109	2558-154	432-80	901-128	2456-150	427-80	1369-97	2334-141
FL	562-148	974-106	2793-172	497-120	969-154	2640-308	536-140	1443-135	2547-223
FC	379-228	881-135	2744-263	431-218	980-163	2705-176	672-402	1391-147	2434-147
GM	591-318	895-123	2360-300	460-157	935-149	2412-271	555-162	1551-84	2529-124
JB	667-273	921-204	2775-120	435-236	897-156	2726-117	646-164	1457-136	2637-131

Tab 4.11.10: les formants de voyelle nasale à long terme

Voici le taux de reconnaissance par locuteur:

voyelle	$\tilde{a}$	$\tilde{\jmath}$	$\tilde{\varepsilon}/\tilde{\text{œ}}$	i	ϵ	ε	a	$\jmath$	o	u	y	o/œ	∂
BG	64%	88%	60%	94%	94%	56%	57%	54%	62%	40%	100%	100%	36%
BP	91%	88%	0%	100%	95%	67%	63%	89%	62%	100%	100%	100%	72%
BT	91%	100%	0%	94%	67%	45%	77%	43%	75%	100%	60%	100%	57%
FL	100%	100%	40%	94%	100%	79%	55%	67%	79%	60%	60%	100%	64%
FC	82%	76%	50%	94%	100%	75%	67%	44%	60%	60%	100%	75%	66%
GM	82%	88%	40%	100%	100%	89%	61%	77%	85%	100%	60%	100%	66%
JB	91%	75%	100%	94%	95%	75%	66%	67%	85%	80%	100%	75%	77%

Tab 4.11.11: score en moyenne de la reconnaissance à long terme

4.11.4 Analyse des résultats

La table 4.11.12 montre le taux de reconnaissance de trois façons. la ligne *linéaire* est le résultat par la mesure de distance linéaire. la ligne *Bark* est celui de la mesure de Bark. et la ligne *lg terme* est celui par l'analyse à long terme. Le résultat est la moyenne de tous les locuteurs en prenant les trois candidats.

voyelle	$\tilde{a}$	$\tilde{\jmath}$	$\tilde{\varepsilon}/\tilde{\text{œ}}$	i	ϵ	ε	a	$\jmath$	o	u	y	o/œ	∂
linéaire	92%	77%	82%	97%	90%	71%	56%	67%	94%	25%	69%	96%	85%
Bark	92%	66%	59%	99%	91%	72%	70%	92%	96%	17%	69%	93%	62%
lg terme	86%	84%	50%	96%	93%	69%	69%	69%	76%	60%	94%	93%	67%

Tab 4.11.12: score en moyenne obtenu par la distance linéaire et de Bark. et à long terme

Dans les trois approches précédentes. celle par la distance spectrale linéaire montre le meilleur score pour les voyelles nasales. Mais si nous regardons la table 4.11.6 et 4.11.7 (GM et JB). le taux de reconnaissance de /ɔ̃/ est très faible. De plus. le taux de reconnaissance de /ɛ̃/ pour les locuteurs BP. BT et FL est aussi faible. Ceci est dû à une confusion entre /ɔ̃/ et /ɛ̃/. Remarquons que le deuxième formant de /ɔ̃/ est beaucoup plus bas que celui de /ɛ̃/ ($\geq$ 1300Hz), la valeur absolue de deuxième formant est suffisante pour enlever cette confusion. Un remplacement de /ɛ̃/ par /ɔ̃/ à la condition que $f_2 < 1300$ ou un remplacement de /ɔ̃/ par /ɛ̃/ à la condition contraire peut augmenter le taux de reconnaissance pour /ɔ̃/ et pour /ɛ̃/. Prenant les mêmes données. le remplacement sous ces conditions rend le taux de /ɔ̃/ pour locuteurs GM égal à 100% et pour locuteur JB à 63%, le remplacement de /ɔ̃/ par /ɛ̃/ donne un taux de 60% pour BP et FL.

L'expérience par l'approche de l'analyse à long terme montre une possibilité: même pour une voyelle de structure complexe. /ɔ̃/ par exemple. le score de reconnaissance est assez bon. Grâce à l'effet d'accumulation des informations. les formants sont plus clairs que ceux par l'analyse à court terme. Ceci peut se voir par la comparaison de la table 4.11.9 et de la table 4.11.10. Le deuxième formant de /ɔ̃/ pour locuteur JB est 1300Hz (dans Tab. 4.11.9). C'est une erreur lors de l'extraction du formant. Pourtant f_2 obtenu par l'analyse à long terme est 897Hz (dans Tab. 4.11.10).

L'analyse à long terme dans la parole continue donne le résultat plus clair parce qu'elle

- réduit le couplage avec les phonèmes précédent et suivant.
- évite l'influence de la position de fenêtre.
- moyenne la source d'excitation et rend important dans le spectre les informations stables.

Le spectre à long terme contient l'information stable et l'indication représentative, et surtout il ne demande pas beaucoup de calcul. Donc le résultat obtenu du spectre à long terme sera un indice très utile pour les traitements de la suite.

La table 4.11.11 résume le taux de reconnaissance à long terme. Nous voyons que le taux de BP et BT pour voyelle /ɛ̃/ est 0%. Le premier formant /ɛ̃/ de ces deux locuteurs est de 435-71 et de 427-80, la largeur de bande est très faible. Cette "mauvaise" reconnaissance révèle que la largeur importante de bande n'est seulement qu'une des indications de nasalisation.

Dans la deuxième expérience, nous avons observé la distribution correcte des zéros détectés par le modèle ARMA. Le résultat est désespéré, l'apparition des zéros est irrégulière, leur largeur de bande est présque non-significative. En théorie, la connexion du conduit nasal au conduit oral forme certainement des zéros. Mais leur forme de manifestation sur un spectre est peu connue. La première hypothèse est qu'ils se manifestent par une vallée. La deuxième est qu'ils se manifestent par l'affaiblissement du formant oral afin d'insister relativement sur les formants nasals. La deuxième pourait être plus raisonable au vu de nos résultats.

4.12 Conclusion

Grâce à l'utilisation du modèle ARMA, nous pouvons toujours trouver les trois formants représentatifs que nous nous utilisons pour la reconnaissance. Dans la mesure de distance, ignorer l'amplitude des formants réduit l'effet des caractéristiques personelles, l'introduction de la largeur de bande reflète naturellement la caractéristique du spectre causé par de la nasalisation. Une analyse à long terme est efficace pour enlever les effets supplémentaires.

4.13 Discussion

La perception est parfois une compréhension déformée de la vérité. Un bon exemple est la perception d'une image. Etant donné deux segments de ligne, l'une ayant aux extrémités deux flèches vers le milieu, l'autre ayant aux extrémités deux flèches vers l'extérieur, alors un ordinateur de haute précision donne toujours comme réponse l'égalité des longueurs. Pourtant la perception humaine donne une réponse de non-égalité. La même explication peut sans doute s'appliquer pour la perception d'une voyelle. D'une même voyelle, certains la perçoivent comme orale, et certains la perçoivent comme nasale. Un /ã/ peut être perçu comme /ã/ dans certains contextes et comme /ɔ̃/ dans des contextes différents. Les formants seuls ne sont pas suffisants pour la détection de la nasalisation et la distinction entre les voyelles nasales et orales.

L'acoustique d'une voyelle nasale ne peut pas être découpée en l'effet du conduit oral et en celui du conduit nasal. Bien qu'il est faisable de mesurer respectivement le signal de sortie par la bouche et celui par le nez, le signal reste encore un signal mélangé parce que le couplage a eu lieu à l'intérieur de l'organe articulatoire. Une clé importante est de mesurer la sortie au moment où le couplage est le plus faible. Si t_{ferme} représente la période de la fermeture de glotte et t_{ouvert} représente la période de l'ouverture de glotte, alors la période t_{ferme} est évidemment le moment où le couplage est fort. Le moment du couplage faible est, quand il existe, le moment d'excitation glotale parce que le conduit oral et le conduit nasal sont excités par la source glotale.

5
Réseau de neurones

5.1 Introduction

Après avoir été abandonnés il y a vingt ans, les réseaux de neurones artificiels réapparaissent dans le monde de la recherche en Intelligence Artificielle suite à leurs récents succès. Un réseau de neurones est dit aussi *une mémoire associative*. C'est une abstraction mathématique à partir de la structure associative familière de l'apprentissage humain. Comme le comportement intelligent humain, un réseau de neurones fonctionne par interaction des neurones simulés en nombre. Chaque neurone, ou unité, joue le rôle d'un simple processeur comme montré par la figure 5.1.

Ce neurone simulé a quatre composantes importantes:

- *connexions d'entrée* (synapse), desquelles le neurone reçoit des autres neurones l'activation.
- *une fonction d'intégration*, qui combine les diverses activations d'entrée dans une seule activation.

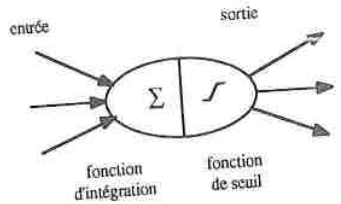


Figure 5.1. un neurone simulé

- *une fonction de seuil* qui convertit la somme des activations d'entrée en une activation de sortie.

- *des connexions de sortie* (chemin axonal), par lesquelles l'activation de sortie d'un neurone est propagagée, comme l'activation d'entrée, aux autres neurones du système.

La connexion entre deux neurones se représente par un coefficient de connexité. La valeur zéro signifie l'absence de connexion entre les deux neurones. Les réseaux de neurones artificiels sont formés de simples processeurs interconnectés qui communiquent entre eux en se transmettant que des signaux d'activation ou d'inhibition. Dans un réseau de neurones, les neurones que nous utiliserons ont pour rôle de recevoir les données du réseau. Ils constituent la *couche d'entrée*. On définit de même une *couche de sortie* qui collecte les réponses du réseau. Entre ces deux couches peuvent exister d'autres neurones qui ne sont pas reliés à l'extérieur et forment les couches *cachées*. Les formes sont stockées dans ces connexions. La détermination des coefficients de connexité est effectuée au cours d'un *apprentissage*.

Les réseaux de neurones artificiels sont fondés sur des modèle théoriques qui tentent d'expliquer de façon très imparfaite comment les cellules du cerveau et leur interconnexion parviennent à exécuter des calculs complexes non linéaires. D'abord, McCulloch et Pitts introduisirent en 1943 une fonction booléenne à un neurone. Ensuite dans l'année 1960, un simple réseau de neurones de deux couches sans couche cachée a été construit, appellé *perceptron*[54]. Mais c'était un réseau de performance très faible. Il n'etait pas capable d'effectuer des taches complexes. Il faut attendre la découverte de l'algorithme d'apprentissage plus général pour qu'on puisse édifier un réseau avec des couches cachées. L'article[71] de 1986 de Rumelhart et McClelland, présenté comment des réseaux de neurones peuvent devenir une architecture d'application nouvelle.

Un réseau de neurones se comporte complètement différemment d'un système informatique classique. Les différences remarquables sont:

- Pour un ordinateur basé sur l'architecture de Von Neumann, le calcul est physiquement séparée de ses mémoires.
  Dans une architecture de réseau de neurones, la capacité de calcul est liée aux mémoires.

- Dans les ordinateurs, les informations sont organisées de manière séquentielle, sous la forme de séries de bits.

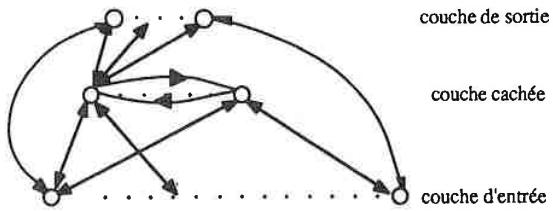


Figure 5.2. réseau multi-couche pour la Machine de Boltzmann

Dans un réseau de neurones, les informations ne peuvent pas être localisées à des emplacements déterminés, mais chaque donnée mémorisée est distribuée sur l'ensemble de la structure qui constitue la mémoire.

- L'apprentissage d'un ordinateur est un apprentissage de type "*par coeur*".
  Le réseau de neurones fait un apprentissage par l'exemple (comme un modèle stochastique markovien).

5.2 Présentation générale du réseau de neurones

Normalement, les connexions entre les neurones sont arbitraires (Fig.5.2) dans un réseau composé de plusieurs couches. La détermination de poids peut être faite par l'algorithme de la Machine de *Boltzmann*[1]. Pour simplifier, nous supposerons que seules les connexions reliant une couche aux couches d'indice supérieur sont possibles et qu'il n'y a pas de connexion intra-couche (Fig.5.3). Dans ce réseau, les données circulent de la couche d'entrée vers la couche de sortie.

Après avoir déterminé l'architecture, on détermine les paramètres du réseau. C'est la phase d'apprentissage. L'apprentissage consiste à présenter au réseau une série d'entrées, et à modifier les connexions du réseau pour qu'à chacune de ces entrée corresponde la sortie souhaitée. Pour un réseau possédant la structure de la figure 5.3, les poids peuvent être facilement calculés par l'algorithme de *rétro-propagation du gradient*[77].

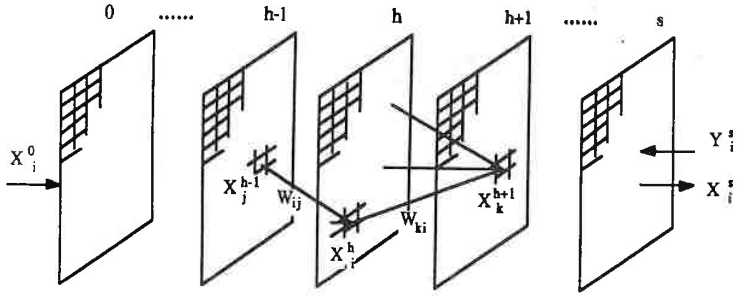


Figure 5.3. réseau multi-couches

5.2.1 Notation

Voici les notations utilisées dans la partie suivante pour présenter l'algorithme de rétro-propagation de gradient.

n : $n^{ème}$ itération pour calculer successivement le coût global.

h : Numéro de couche. $h = s$ si pour la dernière couche (couche de sortie) et $h = 0$ pour la première (couche d'entrée).

i, j, k : indice de neurones.

x_i^h : un neurone i à la couche h.

W_{ij} : la valuation de la connexion qui va de la cellule j à la cellule i. Comme les connexions du réseau sont symétriques, alors $W_{ij} = W_{ji}$.

f : une fonction sigmoïde.

A_i^h : la somme des activations des neurones à la couche *h-1* comme l'entrée de neurone i à la couche h. $A_i^h = \sum_j W_{ij} x_j^{h-1}$.

Un neurone est à la fois un noeud d'entrée et un noeud de sortie. Lorsque les données circulent d'une couche à la suivante, chaque neurone reçoit une entrée et émet son activation:

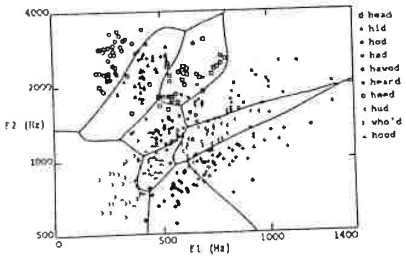


Figure 5.4. Région de décision formée par un classificateur d'un perceptron de deux couches: 2 entrée, 100 noeuds cachés et 10 sorties pour deux données de formants

$$x_i^h = f(A_i^h) \quad \text{avec} \quad A_i^h = \sum_j W_{ij} x_j^{h-1} \tag{5.1}$$

La fonction sigmoïde $f(x) = \frac{1}{e^x+1}$ a des caractéristiques désirées pour les bornes de sortie, qui correspondent à une saturation unitaire, des caractéristiques non-linéaires de la réponse impulsionnelle, et un seuil de réponse. Elle a aussi une dérivée continue, ce dont on aura besoin pour l'algorithme de rétro-propagation.

5.2.2 Applications dans la domaine de traitement de parole

Le réseau de neurones, comme un nouvel outil, a trouvé tout de suite des applications diverses en traitement de la parole.

Les deux premiers formants sont suffisants pour la perception des voyelles orales. Donc la machine doit aussi reconnaitre les voyelles orales par ces formants. Les deux formants construisent un plan de décision, la frontière entre différentes voyelles est non linaire 5.4. W.Huang, R.Lipmann et B.Gold[26] ont montré que les algorithmes traditionnels comme les *classificateur Gaussien, la k-means quantification vectorielle* etc. peuvent être implantés en utilisant un réseau de neurones. Et ils suggèrent que la coopération d'un réseau de neurones avec une HMM peut augmenter la performance du système. R.W.Prager, T.D.Harrison et F.Fallside[66] ont fait une reconnaissance non seulement sur

les voyelles en état stable, mais aussi sur les formes variant dans le temps (cf: figure5.5). Pour ce problème des caractéristiques spectrales/temporelles de la parole, A.Waibel etc.[82] a construit un réseau de neurones possédant le caractère du delai temporel (Fig.5.6). En utilisant ce réseau, ils ont réussi à reconnaître les consonnes /b/ /d/ /g/ avec un score meilleur que celui obtenu par une HMM. D'autres chercheurs[65,6] ont fait la reconnaissances des voyelles, mots isolés, mots enchainés avec un réseau de neurones.

Normaliser la parole est un problème très important. Il y a certains paramètres qui ne sont pas sensibles aux locuteurs: le lieu d'articulation. L'extraction correcte des formants est difficile par rapport au lieu d'articulation. Mais la classification est une décision floue. Bengio et De Mori[3] ont fait une identification de la place d'articulation avec réseau de neurones. Leur conclusion est qu'un réseau de neurones est capable de maitriser les connaissances et les régles de HMM.

Sejnowski et Rosenberg (présentés dans[13]) ont construit un synthétiseur de parole de texte en anglais avec un réseau de neurones. Ce réseau "prononçait" correctement 95% pour des mots du texte modèle et 90% pour un texte nouveau.

5.3 Reconnaissance de mots isolés par un réseau de neurones

5.3.1 Interface entre les mots isolés et le réseau

Dans l'utilisation d'un réseau de neurones pour la reconnaissance de mots isolés, le premier problème est de relier le signal de parole au modèle. Au niveau des mots isolés, la façon la plus directe est de représenter un mot par un tableau de dimention de M*N [65], M étant le nombre de valeurs de chaque vecteur acoustique, N étant le nombre maximum de vecteurs composant un mot. Une autre méthode plus proche du processus naturel du décodage de la parole est d'utiliser un réseau de neurones prenant en compte le facteur de déformation temporelle. Le TDNN (Time Delay Neural Network) est un tel réseau qui est capable d'apprendre les décision non-linéaires automatiquement en calculant l'erreur de rétro-propagation, et égalemet de retrouver les caractéristiques acoustiques et leurs relations temporelles indépendantes de la position du temps [82,84]. En considérant qu'il n'est pas essentiel de conserver la contrainte temporelle dans la production et la perception de la parole, H.Bourlard et C.J Wellekens [6] ont employé une méthode plus simple et plus

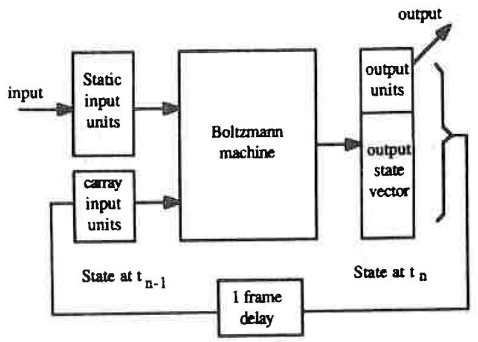


Figure 5.5. Boltzmann machine with carry units

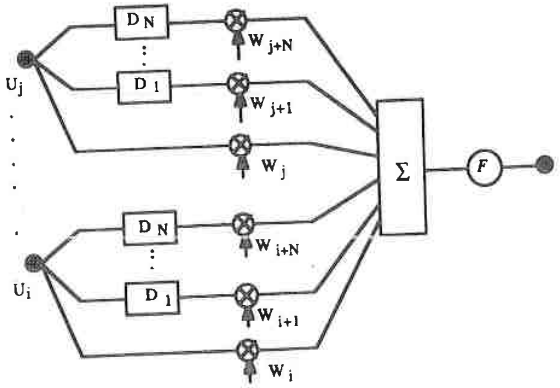


Figure 5.6. A Time Delay Neural Network

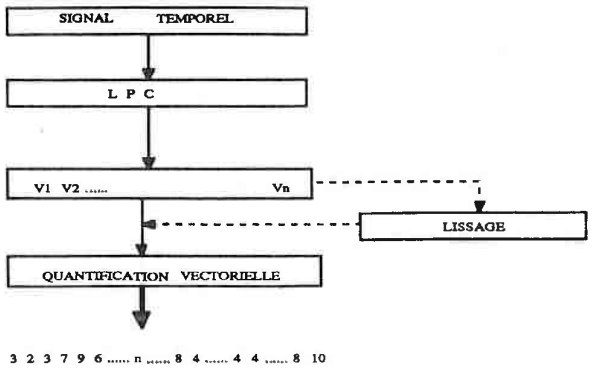


Figure 5.7. le signal temporel est d'abord transformé en une série de codes de QV qui seront ensuite utilisés par le réseau de neurones

efficace. Ici, nous avons choisi cette dernière méthode pour représenter (Fig.5.7) un mot par une suite de codes obtenus par analyse LPC du signal et quantification vectorielle(QV) en utilisant la distorsion d'Itakura-Saito à gain normalisé [37.18].

La suite de codes de quantification vectorielle est rentrée parallèlement dans le réseau des neurones Ne-Nh-Ns. *Ne* étant le nombre de cellules d'entrée qui est égal au nombre des prototypes QV, *Nh* étant le nombre de cellules de la couche cachée. *Ns* étant le nombre de cellules de sortie qui indiquent la reconnaissance d'un des dix chiffres (Fig.5.8).

Une suite de codes de QV d'un mot isolé est toujours un sous-ensemble des codes de la couche d'entrée. La machine examine donc les codes de QV, et si elle rencontre un code N, elle active la cellule du $N^{ème}$ neurone.

Une cellules d'entrée déjà activée le reste si elle apparaît à nouveau dans la forme à reconnaître.

L'ordre d'apparition des codes d'entrée n'est pas pris en compte; de même, on néglige

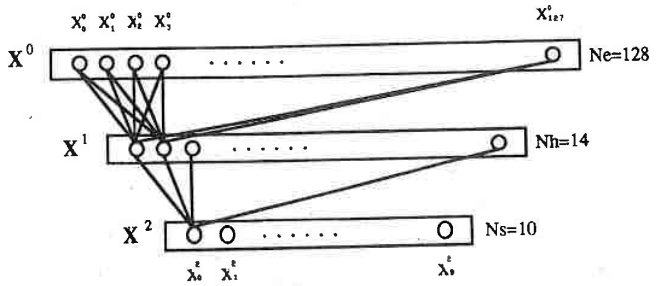


Figure 5.8. Le réseau de neurones à trois couches ayant une couche de sortie de 10 cellules, une couche cachée de 14 cellules et une couche d'entrée de 128 cellules

totalement la fréquence d'apparition d'un code d'entrée. L'ignorance de ces deux facteurs simplifie donc l'entrée d'un signal dans le réseau, mais nous devons tenir compte des conséquences impliquées par ces deux simplifications.

Dans la production de la parole, les phonèmes fondamentaux sont intégrés à une séquence. Chaque phonème porte ses propres caractéristiques mais aussi les effets de coarticulation que lui imposent ses voisins.

Un phonème apparaissant à plusieurs reprises dans différents contextes correspond donc à autant de réalisations acoustiques différentes, surtout les phonèmes ayant une durée très courte. Le contenu acoustique est suffisant pour la description de chaque mot[6].

La suppression de l'importance de la fréquence d'apparition d'un mot posera pourtant un réel problème. Dans ce cas, un code qui apparaît même une seule fois joue un rôle aussi important que celui qui apparaît plusieurs fois. Ceci est équivalent à affaiblir la résistance du système aux parasites. Si nous ne considérons pas les bruits extérieurs au système, il reste deux sortes de bruits qui peuvent influencer les résultats: Premièrement, le bruit dû à la position de la fenêtre d'analyse de LPC qui peut faire varier le vecteur des coefficients LPC, et deuxièmement, le bruit de la Quantification Vectorielle. Nous proposons

donc deux méthodes pour réduire ces deux bruits avant d'entrer les codes dans le réseau (Fig.5.7).

— *Pondération*

Pour restaurer partiellement la fréquence d'apparition d'un code dans la série, nous faison une pondération sur les valeurs affectées aux cellules correspondantes. Voici la fonction de pondération selon le nombre d'apparition:

$$X_c^0 = \begin{cases} 0.5 & \text{si } nc = 1 \\ 0.6 & \text{si } nc = 2 \\ 0.8 & \text{si } nc = 3 \\ 0.9 & \text{si } nc = 4 \\ 1.0 & \text{si } nc > 4 \end{cases}$$

nc étant le nombre d'apparition du code "C" dans la série. X_c^0 étant la cellule de la couche d'entrée correspondante au code "C". Les valeurs choisies sont *empiriques*.

Normalement, les valeurs d'entrée aux cellules sont soit 1 si elles sont activées, soit 0 si elles ne sont pas activées. Après avoir employé cette pondération, les valeurs d'entrée ne sont plus 0 ou 1, mais elles deviennent des valeurs réelles positives et inférieures à 1. Grâce à la fonction de seuil *sigmoïde*, les valeurs des entrées additionées sont transformées aussi en une valeur réelle entre 0 et 1. Donc la cellule correspondante à un code "C" active les cellules de la couche suivante selon sa valeur pondérée. Nous donnerons les résultats expérimentaux dans la partie suivante.

— *Lissage*

Une autre méthode pour réduire l'effet des bruits est de faire un lissage (Fig.5.7). Il n'y a pas de raison de faire un lissage directement sur les codes de QV de la série, parce que les codes ne sont pas une description acoustique, mais seulement des symboles. Alors nous le faisons sur la suite des vecteurs de la prédiction. Les paramètres déduits de la fonction d'autocorrélation (les coefficients de la prédiction linéaire, les coefficients de la réflection du filtre inverse, les surfaces

du conduit vocal, par exemple) sont équivalents mathématiquement. Mais seul un lissage sur les surfaces du conduit vocal leur donne un sens physique: la forme de la cavité vocale ne peut pas avoir un mouvement brutal. En plus, les surfaces lissés représentent toujours un filtre stable. En raison de la quantité du calcul, nous n'avons pas choisi les surfaces, mais nous faisons un lissage directement sur les coefficient d'autocorrélation. Ce lissage est un simple calcul de moyenne sur les coefficients, et il rend un filtre stable d'après la condition de stabilité d'un système.

Comme nous allons faire un lissage sur les vecteurs de LP avant la quantification vectorielle, pour avoir un code QV raisonnable, les prototypes de QV doivent aussi être lissés. Nous avons sauvegardé deux sortes de prototype QV, l'un est sans lissage et l'autre avec lissage. Dans la partie suivante, le rôle du lissage est évident.

5.3.2 Expérimentation

Nos expériences[87,86] sont scindées en trois parties, celle pour les chiffres chinois dans un contexte monolocuteur, celle pour les chiffres français en multilocuteurs, et celle sur la comparaison de la performance entre la méthode DTW (Dynamic Time Warping) et la méthode RN (Réseau de Neurones).

Notre réseau de neurone de base se compose des couches suivantes:

- la couche d'entrée possède 128 neurones
- la couche cachée possède 14 neurones
- la couche de sortie possède 10 neurones

Notre travail est réalisé sur *MASSCOMP*, la fréquence d'echantillonnage est 16KHz avec une précision de 12 bits. L'analyse acoustique LP détermine, pour chaque fenêtre de Hamming de 12.8$msec$ avec un facteur de préaccentuation de 0.95, un vecteur LPC. Le nombre de points de lissage sur les vecteurs de LP est fixé à 5.

5.3.3 Reconnaissance de mots isolés mono-locuteur

Nous avons fait des essais pour la reconnaissance monolocuteur des dix chiffre isolés (chinois 0-9). Chaque chiffre a été prononcé vingt fois, soit au total deux cents chiffres, la

moitié servant à l'apprentissage, le reste servant de formes inconnues à reconnaître.

Nous avons d'abord fait une expérience concernant la performance de la pondération sur les vecteurs acoustiques en prenant le réseau de base. Le taux de reconnaissance sans pondération est de 93%, cependant celui avec pondération est de 98%. Ceci nous indique que cette pondération est utile pour augmenter les performances du système.

Par la suite, nos expériences ont consisté à examiner les performances du réseau en faisant varier les différents paramètres, tels le nombre de neurones de la couche d'entrée et de la couche cachée, et aussi l'ordre de LPC.

La première expérimentation est de mesurer la performance du système en fonction du nombre de points de lissage. La courbe (Fig.5.9) montre l'évolution du taux de reconnaissance. Prenant la courbe de Nh16 par exemple, nous voyons que le taux de reconnaissance sans lissage est égal à 96% et qu'il arrive à son meilleur score lorsque NBL=14 ou NBL=15. En fixant NBL, NBL=4 par exemple, nous obtenons le meilleur résultat quand Nh=14, le pire résultat est obtenu quand Nh=5.

Puisque NBL=5 correspond au nombre de points de lissage des prototypes QV, alors Nh=14 est fixé comme un des paramètres de notre réseau de base. Et à partir de cette courbe, nous pouvons constater qu'après lissage, les résultats de reconnaissance peuvent être améliorés.

Ces courbes montrent une variation irrégulière. Ceci veut dire que les paramètres consernant la structure du réseau sont sensibles aux mêmes entrées que la parole.

Nous avons aussi fait varier le nombre d'unités de la couche d'entrée(Ne) en fixant le nombre d'unités de la couche cachée à Nh=14 (Fig.5.10). Nous constatons une reconnaissance optimale quand Ne=128. Si le nombre d'unités de la couche d'entrée est fixé à Ne=64, le codebook de quantification vectorielle n'est pas suffisant. En revanche, le résultat avec le codebook de Ne=256 est aussi moins bon, cela etant probalement dû à l'insuffisance du nombre de références d'apprentissage.

La figure 5.11 montre l'évolution du taux de reconnaissance en fonction de l'ordre de LPC. Les résultats de l'ordre de 20 et de 24 sont similaires. Nous prenons donc 20.

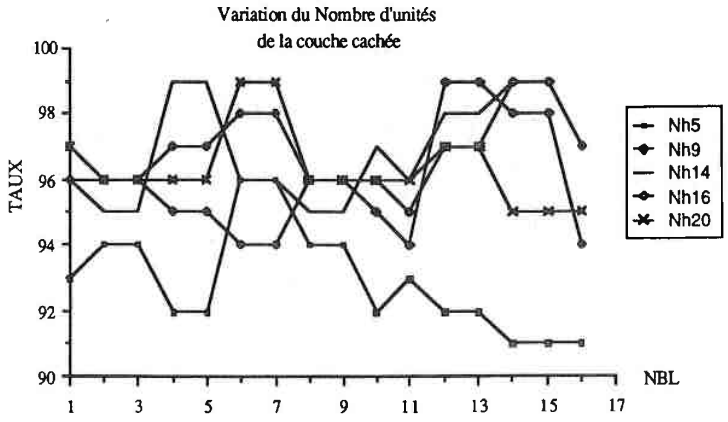


Figure 5.9. les courbes de l'évolution du taux de reconnaissance en fonction du nombre de points de lissage (NBL) selon le nombre d'unités de la couche cachée (Nh)

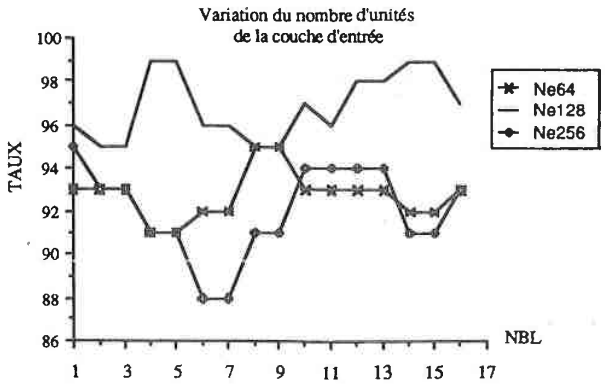


Figure 5.10. les courbes du taux de reconnaissance en fonction du nombre points de lissage (NBL) selon le nombre d'unités de la couche d'entrée (Ne)

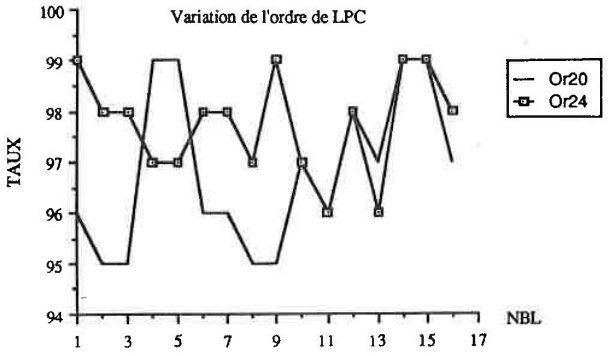


Figure 5.11. les courbes du taux de reconnaissance en fonction du nombre points de lissage (NBL) selon l'ordre de LPC (Ne)

5.3.4 Reconnaissance de mots isolés multi-locuteur

Avec les mêmes traitements sur le signal, nous faisons ici une reconnaissance multi-locuteur des chiffres français (0-9) en reprenant le réseau de base.

Pour six locuteurs féminins et six locuteurs masculins, chaque chiffre a été prononcé quatre fois par personne. Cela fait au total 12*10*4=480 mots.

Le locuteur fait partie de l'apprentissage

(1) Les deux première prononciations de trois sujets féminins et de trois sujets masculins ont servi à construire les références d'apprentissage, soit 120 chiffres. Le taux de reconnaissance pour les deux autres prononciations est de 97.3%.

(2) Comme en (1), on prend deux prononciations pour tous les locuteurs (six féminins et six masculins) comme références, soit 12*10*2=240 chiffres, et les deux autres comme mot à reconnaitre. Nous obtenons 98.8% de bonne reconnaissance.

Le locuteur ne fait pas partie de l'apprentissage

Les deux premières prononciations de trois sujets féminins et de trois sujets masculins ont servi à construire les références d'apprentissage, soit 120 chiffres. Ici, nous prenons comme chiffres inconnus deux prononciations de trois sujets féminins et de trois sujets masculins qui n'ont pas fait partie de l'apprentissage. Le résultat est de 84.2% en moyenne. Voici les résultats pour chaque locuteur:

$$féminin \begin{cases} om : & 70\% \\ md : & 100\% \\ pv : & 95\% \end{cases}$$

$$masculin \begin{cases} jph : & 75\% \\ pd : & 80\% \\ pn : & 85\% \end{cases}$$

5.4 Reconnaissance de voyelles découpées dans des mots isolés

Nous avons aussi fait une expérience de reconnaissance de voyelles, qui sont tirées du petit corpus de mots isolés (cf partie 4.10.2). Le traitement de reconnaissance de voyelles est pratiquement identique à celui de mots isolés, sauf que le nombre d'unités de la couche cachée devient 20 et le nombre d'unités de la couche de sortie devient 13 (treize voyelles différentes).

Le locuteur fait partie de l'apprentissage

La première prononciation de voyelles a servi à construire les références d'apprentissage, soit 12(locuteurs)*13(voyelles) voyelles pour les sujets masculins, et 5(locuteurs)*13(voyelles) pour les sujets féminins. Donc deux ensembles de référence sont préparés.

Pour les sujets masculins, il y a, parmi les références, trois voyelles qui ne sont pas reconnues après 100.000 passes d'apprentissage. Pour les sujets féminins, la reconnaissance des formes de référence est de 100%. Les résultats de la reconnaissanc des voyelles inconnues sont donnés dans les tables Tab.5.1 et Tab.5.2.

En regardant ces deux tableaux, on remaque qu'un réseau de neurones est capable de distinguer facilement dans la plupart des cas les voyelles orales des nasales pour un sujet masculin ou féminin. On remarque également que les voyelles orales les plus souvent confondues avec une voyelle nasale sont /ɔ/ /o/ et /u/, qui ont une forme d'articulation assez proche d'une voyelle nasale.

voyelle	*i*	*e*	*ɛ*	*a*	*ɔ*	*o*	*u*	*y*	*œ*	*ø*	·	*ɔ̃*	*ɛ̃*
i	11	0	1	0	0	0	0	0	0	0		0	0
e	0	12	0	0	0	0	0	0	0	0	·	0	0
ɛ	0	0	11	0	0	0	0	0	0	1		0	0
a	0	0	0	9	1	0	2	0	0	0		0	0
ɔ	0	0	0	0	7	0	0	0	1	1		2	0
o	0	0	0	0	0	7	3	0	0	2	0	0	0
u	0	0	0	0	0	4	8	0	0	0	0	2	0
y	0	0	0	0	0	0	0	12	0	0	0	0	0
œ	0	0	0	0	0	1	1	0	11	0	0	0	0
ø	0	0	0	0	0	0	0	0	1	11	0	0	0
ã	0	0	0	0	4	1	1	0	0	0	6	2	0
ɔ̃	0	0	1	0	0	1	1	0	0	1	1	7	0
ɛ̃	0	0	1	1	1	0	0	0	0	1	1	0	7

Tab. 5.1 rés

voyelle	*i*	*e*	*ɛ*	*a*	*ɔ*	*o*	*u*	*y*	*œ*	*ø*	*ã*	*ɔ̃*	*ɛ̃*
i	4	0	0	0	0	1	0	0	0	0	0	0	0
e	0	5	0	0	0	0	0	0	0	0	0	0	0
ɛ	0	0	4	0	0	0	0	0	1	0	0		0
a	0	0	0	5	0	0	0	0	0	0	0		0
ɔ	0	0	0	0	5	0	0	0	0	0	0	0	0
o	0	0	0	0	0	3	0	0	0	1	0		0
u	0	0	0	0	0	0	5	0	0	0	0	0	0
y	0	0	0	0	0	0	0	5	0	0	0	0	0
œ	0	0	0	0	0	1	0	0	5	0	0	0	0
ø	0	0	0	0	0	0	0	0	0	5	0	0	0
ã	0	0	0	0	0	1	0	0	0	0	3		0
ɔ̃	0	0	0	0	0	1	0	0	0	0	0	4	0
ɛ̃	0	0	0	0	0	0	0	0	1	0	0	0	4

Tab l s

Le locuteur ne fait pas partie de l'apprentissage

Une partie des locuteurs est prise pour la préparation des références, l'autre partie pour les rests. Il s'agit donc de reconnaissance de voyelles indépendante du locuteur. Les Tab.5.3 et Tab.5.4 donnent une vue globale des performances de reconnaissance.

Nous remarquons toute suite que le résultat se dégrade un peu. Dans la Tab.5.3. /ɔ̃/ est confondue avec /ã/. /o/ est confondue avec /ɔ/ /u/.

voyelle	*i*	*e*	*ɛ*	*a*	*ɔ*	*o*	*u*	*y*	*œ*	*ø*	*ã*	*ɔ̃*	*ɛ̃*
i	9	2	0	0	1	0	0	1	0	0	0	0	0
e	0	9	0	0	1	0	0	1	0	0	0	0	1
ɛ	0	0	8	0	0	1	1	0	1	0	0	0	1
a	0	0	0	10	0	0	0	0	1	0	0	0	1
ɔ	0	0	0	0	9	0	0	0	0	0	0	0	0
o	0	0	0	0	2	5	2	0	1	0	1	0	0
u	0	0	0	0	0	1	8	0	0	0	0	0	0
y	0	0	0	0	0	0	1	7	0	0	0	0	1
œ	0	0	0	0	0	0	2	0	7	1	0	0	2
ø	0	0	0	0	0	0	2	0	5	4	0	0	1
ã	0	0	0	0	2	0	0	0	0	0	9	0	0
ɔ̃	0	0	1	0	2	0	1	0	0	0	0	9	0
ɛ̃	0	0	0	2	0	0	0	0	1	0	0	0	9

Tab. 5.3 résultats pour les sujets masculins.

voyelle	*i*	*e*	*ɛ*	*a*	*ɔ*	*o*	*u*	*y*	*œ*	*ø*	*ã*	*ɔ̃*	*ɛ̃*
i	5	0	0	0	0	0	0	0	1	0	0	0	0
e	0	4	0	0	0	0	0	0	0	0	0	0	0
ɛ	0	0	4	0	0	0	0	0	0	0	0	0	0
a	0	0	0	3	0	0	0	0	1	0	0	0	0
ɔ	0	0	0	0	2	0	0	0	0	0	0	0	2
o	0	0	0	0	0	0	0	0	2	1	0	0	1
u	0	0	0	0	0	1	2	0	1	0	0	0	0
y	0	0	0	0	0	0	1	3	0	0	0	0	0
œ	0	0	2	0	0	0	0	0	2	0	0	0	0
ø	0	0	0	0	0	0	0	0	2	0	1	0	1
ã	0	0	0	0	0	0	0	0	0	0	4	0	0
ɔ̃	0	0	0	0	0	0	0	0	0	0	1	0	2
ɛ̃	0	0	0	0	0	0	0	0	0	0	1	0	3

Tab. 5.4 résultats pour les sujets féminins.

5.5 Comparaison entre DTW et RN

Nous avons choisi une programmation dynamique (formule 5.2) avec une contrainte simple et symétrique (Fig.5.12).

$$D_{pp}(i,j) = min \begin{cases} D_{pp}(i-1,j) + d(i,j) \\ D_{pp}(i-1,j-1) + 2d(i,j) \\ D_{pp}(i,j-1) + d(i,j) \end{cases} \quad (5.2)$$

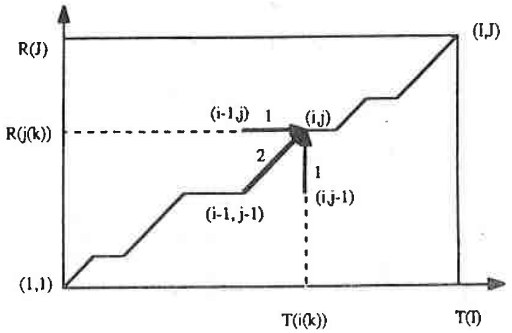


Figure 5.12. programmation dynamique et la contrainte locale

Les paramètres du réseau de neurones RN sont: Ne=128, Nh=14, Ns=10, NBL=4, ordre=20.

La comparaison entre les deux systèmes est faite uniquement pour une reconnaissance monolocuteur.

Les deux systèmes ont été testés en prenant trois ensembles de formes de référence:

- 10 formes par chacun des 10 chiffres, soit 100 formes
- 5 formes par chacun des 10 chiffres, soit 50 formes
- 1 forme par chacun des 10 chiffres, soit 10 formes.

Les résultats sont donnés figure 5.13. Dans les trois cas, l'ensemble de formes inconnues est constitué des mêmes 100 chiffres n'ayant pas servi à l'apprentissage. Les temps indiqués sont des temps CPU sur MASSCOMP, y compris l'analyse LPC et la quantification vectorielle.

On constate que dans la méthode de RN, le temps de reconnaissance et l'espace nécessaire sont indépendants du nombre de formes de référence et que le taux de reconnaissance dépend beaucoup du nombre de références.

Et dans la méthode de DTW, le temps de reconnaissance est important, tandis que le taux de reconnaissance est moins sensible au nombre de références.

La comparaison ici est une comparaison des performances externes de deux systèmes. Bien que le nombre de références soit identique pour les deux, la quantité d'information véhiculée n'est pas la même. Le DTW garde tous les éléments acoustiques en mémoire, tandis que le RN sauvegarde uniquement les structures dans ses connexions. En plus, la performance d'un système neuronal dépend de ses connexions internes entre les neurones, et une bonne projection de l'acoustique à l'entrée du système.

5.6 Conclusion

Cette expérience limitée montre l'intérêt des modèles connexionnistes pour la reconnaissance de la parole. En regardant les résultats obtenus, nous constatons que:

- Pour le DTW, le temps de décodage (reconnaissance ou classification) dépend du nombre de références stockées. Plus les références sont nombreuses, plus le temps du décodage est long.
  Ce n'est pas le cas pour un réseau de neurones. A la fois mode de stockage et organe de traitement de l'information, un réseau de neurones, étant composé parallèlement par des simples processeurs, fournit un décodage en temps réel quasi constant.
- Dans notre expérience, la reconnaissance des références est toujours 100%. Ca veut dire que le nombre de références n'a pas encore dépassé la capacité du système.

- La reconnaissance de mots isolés en multi-locuteur est moins bonne que celle pour mono-locuteur. Ceci est dû à l'insuffisance du nombre de références.

Si nous pouvons sauvegarder toutes les connaissances sur les synapses, nous pouvons également y sauvegarder une connaissance. Un des caractères du réseau de neurones est qu'on peut faire *l'apprentissage par préférence*. C'est à dire qu'une forme est tenue par importante si on augmente le nombre de références pour elle. On construit un réseau de deux sorties seulement. Si l'entrée appartient à une forme désirée, l'un des deux neurones sera activé et l'entrée acceptée, sinon elle est refusée. Un tel réseau répondra par oui ou non pour un mot donné. Donc on utilise autant d'ensembles de poids pour un réseau que de mots inconnus. Cette méthode consiste en réalité à spécialiser la fonction d'un réseau pour une seule distinction.

Il y a beaucoup de capacités potentielles d'un réseau de neurones que nous n'avons pas encore employées. De nombreux problèmes restent à résoudre comme, par exemple, la structure interne du réseau, ou le contrôle automatique du nombre d'apprentissage. Quant à un système de reconnaissance, il est important de transformer un signal temporel en une forme bien adaptée aux réseaux.

Nous pensons que le réseau de neurones est utile pour la reconnaissance des voyelles nasales, parce qu'une nasale ne peut pas être décrite clairement par des règles régulières.

Le réseau que nous utilisons est un réseau entièrement connecté d'une couche à la couche suivante. Ce genre du réseau pose certains problèmes:

- la quantité du calcul est énorme
- puisque les neurones sont entièrement connectés, une perturbation par n'importe quelle cause sera propagée à tous les nerrones associés.

Des chercheurs ont commencé à étudier un reseau de neuronnes partiellement connectés[83] pour améliorer le performance.

Une nouvelle approche a été également proposée par notre équipe à Nancy [21]. Au lieu de se baser sur des unités uniformes, de type neurones, on construit un reseau avec des

unités beaucoup plus sophistiqué, dites Colonnes Corticales. Une colonne corticale correspond, d'apres les données neurobiologiques, à une association d'une centaine de neurones, possède une activité spécifique et fonctionnelle. Un réseau de neurones de telles unités rédnit la combinaison exponentielle des connexions et est plus adaptée pour la résolution des problèmes symboliques.

En fait, le réseau de neurones qu'on utilise est seulement une simulation simple de la structure du cerveau humain. "Chacun des 10^{10} à 10^{11} neurones du cerveau est en contact avec 10^2 à 10^3 autres neurones en moyenne; les organigrammes de câblage correspondants sont rigoureusement programmés, mais la multiplicité des contacts rend théoriquement possible l'interaction de chaque neurone avec n'importe quel autre à travers trois ou quatre jonctions synaptiques"[12]. La forme de pensée dépend probablement des formes de connexions. Sans doute, l'amélioration de reconnaissance par les réseaux de neurones passe par la recherche sur la structure des réseaux.

Ensemble de références: 10 * 10 = 100 chiffres		
	DTW	RNA
TAUX	99%	99%
TEMPS (minutes)	41:23.6	5:38.3
ESPACE (octet)	477396	31308

Ensemble de références: 5 * 10 = 50 chiffres		
	DTW	RNA
TAUX	97%	91%
TEMPS (minutes)	15:22.4	5:38.3
ESPACE (octet)	236844	31308

Ensemble de références: 1 * 10 = 10 chiffres		
	DTW	RNA
TAUX	59%	41%
TEMPS (minutes)	7:21.9	5:38.3
ESPACE (octet)	47276	31308

Figure 5.13. comparaison de la reconnaissance de deux systèmes dans trois cas de figure

6
Conclusion

L'objectif de cette thèse était l'analyse des voyelles nasales à partir d'un gros corpus de parole isolée et continue. Pour réaliser ce but, nous avons d'une part construit un modèle ARMA spécialement pour l'analyse des nasales dans une région fréquencielle sélectionné; et d'autre part proposé une méthode d'extraction des formants à partir de ce modèle en imitant un expert phonéticien. La combinaison du modèle ARMA sélectionné avec notre méthode d'extraction des formants nous permet d'ignorer la plupart des formants irréguliers qu'on rencontre surtout lorsque l'ordre du modèle devient élevé.

L'articulation des nasales introduit un conduit sonore supplémentaire qui rend difficile l'application des méthodes d'analyse classiques. C'est pour cela que nous utilisons un réseau de neurones dans la reconnaissance des voyelles. Dans notre expérience, nous avons d'une part comparé la performance des réseaux neurones avec celle de DTW, et d'autre part fourni les resultats du teste. Nous pouvons dire que l'analyse par les réseaux de neurones est une nouvelle méthode puissante, qui possède une capacité à apprendre et à traiter des informations bruitées ou imcomplètes. Elle fournit ainsi des résultat plus proches de ceux donnés par un cerveau humain au sens du comportement.

Avec le modèle ARMA sélectionné, nous avons testé le corpus "la bise et le soleil". Les résultats nous montre que le modèle ARMA est capable de fournir des formants précis avec des largeurs de bande toujours raisonnable. Ce qui est intéressant, c'est qu'un traitement à long terme sur un phonème donne une analyse de la même qualité que celle à court terme. On peut conclure à partir de ce résultat que la variation des paramètres d'un conduit vocal est faible dans une durée de phonème, tandis que la cumulation des informations de prononciation joue un rôle important pour la reconnaissance: la nasalisation d'une voyelle peut se réaliser dans toute la phase de sa formation. Nous pensons que l'analyse à long terme est un traitement efficace et économique pour séparer globalement la classe nasale et orale.

Les résultats de la reconnaissance nous montre également que porter l'attention uniquement sur les formants, ou l'amélioration de la performance du modèle pôle-zéro ne sont pas suffisants pour révéler les indices des nasales. Ceci est dû au fait que l'indice de nasalité est distribué en temps et en fréquence. Un tel indice ne dépend pas seulement du facteur de paramétrage, mais aussi de facteurs psychologiques. Nous pensons qu'il faut utiliser un réseau de neurones pour effectuer une analyse globale, fournir les points les plus significatifs à certains instants, et utiliser ensuite un modèle pôle-zéro pour prendre les décisions finales.

Bibliographie

[1] D.H. Ackley, G.E. Hinton, et T.J. Sejnowski. A learning algorithm for Boltzmann machines. 1985. Cognitive Science.

[2] B.S. Atal. Linear prediction analysis of speech based on a pole-zero representation. *J.Acoust.Soc.Am.* 65(5):1310–1318, Nov. 1978.

[3] Y. Bengio et R. De Mori. Use of neural networks for the recognition of place of articulation. *IEEE-ICASSP*, pages 103–106, 1988.

[4] G. Bjuggren et G. Fant. The nasal cavity structures. STL-QPSR 4/1964.

[5] E. Bognar et H. Fujisaki. Analysis, synthesis and perception of the French nasal vowels. *Proc. IEEE-ICASSP*, pages 1601–1604, 1986.

[6] H. Bourlard et C.J. Wellekens. *Speech Pattern Discrimination and Multilayer Perceptrons.* Rapport no. Manuscript M.211, Philips Research Laboratory, Sept. 1987.

[7] J. Burg. A new analysis technique for time series data. *NATO A.S.I on Signal Proc.*, pages 39–743, 1968.

[8] J.A. Cadzow. High performance spectral estimation – a new Arma method. *IEEE Trans. ASSP*, 28(5):524–529, Oct. 1980.

[9] M.J. Caraty et X. Rodet. Distance interspectrale à critères perceptifs. *14e JEP. GALF*, 1985.

[10] Y.T. Chan et J.C. Wood. A new order determination technique for Arma processes. *IEEE Trans. ASSP*, 32(3):517–521, June 1984.

[11] J.C.OB Chow. A preliminary algorithm for determining the order of a linear stochastic dynamic system. *Proc.IEEE Decision Contr. Conf.*, 1971.

[12] CPE. La maitrise du vivant. Sciences et Techniques numéro spécial pour la révolution de l'intelligence, 1988.

[13] C. Durand. Recherche: vers le neuro-ordinateur. *Micro-Système*. Oct. 1987.

[14] O. Fujimura et J. Lindqvist. Sweep-tone measurements of vocal-tract characteristics. *J.Acoust.Soc.Am.* 49:541–558. 1971.

[15] H. Fujisaki et M. Lingqvist. Estimation of voice source and vocal tract parameters based on arma analysis and a model for the glottal source waveform. *Proc. IEEE-ICASSP*. 1987.

[16] C. Gagnoulet. *Reconnaissance de la parole - Cours destiné aux étudiants de l'E.N.S.T - Brest.* Rapport. C.N.E.T — Lannion. Jan. 1986.

[17] J.R. Glass et V.W. Zue. Detection of nasalized vowels in american english. *Proc. IEEE-ICASSP*. pages 1569–1572. 1985.

[18] A. Gourinda. Codage et reconnaissance de la parole par quantification vectorielle. Thèse de l'Université de Nancy I. 1988.

[19] C. Gueguen. Introduction à l'analyse de la parole. *7e Journées du GALF*. pages 59–83. 1979.

[20] J. Guibert. *La Parole — Compréhension et synthèse par les ordinateurs.* Presses Universitaires de France. 1979.

[21] F. Guyot. F. Alexandre. et J.P. Haton. Toward a continuous model of the cortical column: application to speech recognition. *Proc. ICASSP*. 1989.

[22] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*. pages 51–83. 1978.

[23] S. Hawkins et K.N. Stevens. Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels. *J.Acoust.Soc.Am.* 77(4):1560–1575. April 1985.

[24] J.M. Hombert. A propos des 'universaux' de la nasalisation. *16ème Journée d'Etude sur la Parole*. pages 84–87. Hammamet. Oct. 1987.

[25] T.C. Hsia. *System Identification — Least-Square s Methods.* D.C Heath and Company. Lexington, Massachusetts Toronto. 1977.

[26] W. Huang. R. Lipmann. et B. Gold. A neural net approach to speech recognition. *IEEE-ICASSP*. pages 99–102. 1988.

[27] M.J. Hunt. A robust formant-based speech spectrum comparison measure. *Proc. IEEE-ICASSP*. 1985.

[28] S.A.K. Ibrahim. R.G. Martin. et B.G. Olmedo. New contributions to the spectral estimation by means of parametric modeling using rational transfer functions. *Signal Processing*. 231–241. Dec. 1987.

[29] I.S. Konvalinka et M.R. Matausek. Simultaneous estimation of poles and zeros in speech analysis and iterative inverse filtering algorithm. *IEEE Trans. ASSP*. 27(5):485–492. Oct. 1979.

[30] L.H. Koopmans. *The Spectral Analysis of Time Series.* Academic. 1974.

[31] G.E. Kopec. Formant tracking using hidden Markov models and vector quantization. *IEEE. Trans. ASSP*. 34:709–729. August 1986.

[32] G.E. Kopec. A.V. Oppenheim. et J.M. Tribolet. Speech analysis by homomorphic prediction. *IEEE. Trans. ASSP*. 25:40–49. Feb. 1977.

[33] M. Kunt. *Traitement numérique des signaux.* Dunod. 1984.

[34] D.T.L. Lee, M. Morf. et B. Friedlander. Recursive least squares ladder estimation algorithms. *IEEE. Trans. ASSP*. 29:627–641. June 1981.

[35] S. Li et B W. Dickinson. Application of the lattice filter to robust estimation of AR and ARMA models. *IEEE. Trans. ASSP*. 36:502–512. April 1988.

[36] J.S. Liénard. *Les processus de la communication parlée: introduction à l'analyse et la synthèse de la parole.* MASSON. 1977.

[37] Y. Linde. A. Buzo. et R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Commun.*, 28(1):84–95. Jan. 1980.

[38] J. Lindqvist et J. Sundberg. Acoustic properties of the nasal tract. STL-QPSR 1/1972.

[39] R. Linggard et F.J. Owens. Pole-zero modelling of speech spectra. *Speech Communication*. 125–133. Jan. 1982.

[40] F. Lonchamp. Analyse acoustique des voyelles nasales françaises. *Verbum*. II(1):9–54. 1979. Revue de linguistique publiée par l'Université de Nancy II.

[41] F. Lonchamp. Etudes sur la production et perception de la parole — les indices acoustiques de la nasalité vocalique; la modification de timbre par la fréquence fondamentale. Thèse d'Etat. Université de Nancy II. 1988.

[42] F. Lonchamp. *Les Sons Du Français — Analyse Acoustique Cours de formation sur le Traitement de la Parole.* Rapport. GALF. 1985.

[43] S. Maeda. *Acoustic cues of vowel nasalization: a simulation study.* Rapport, CNET. 1983.

[44] S. Maeda. The role of the sinus cavities in the production of nasal vowels. *Proc. IEEE-ICASSP*, pages 911–914. 1982.

[45] J. Makhoul. Linear prediction: a tutorial review. *Proc. IEEE*, 63:561–580. 1975.

[46] J. Makhoul. Spectral linear prediction: properties and applications. *IEEE. Trans. ASSP*, 23:283–296. June 1975.

[47] J.D. Markel. Application of a digital inverse filter for automatic formant and analysis. *IEEE. Trans. Audio Electroacoust.*, AU-21:149–159. 1973.

[48] J.D. Markel. Digital inverse filtering — a new tool for formant trajectory estimation. *IEEE. Trans. Audio Electroacoust.*, AU-20:129–137. 1972.

[49] J.D. Markel et A.H. Gray. *Linear Prediction of Speech.* Springer-Verlay. 1976.

[50] M.V. Mathews. J.E. Miller. et E.E. David Jr. Pitch synchronous analysis of voiced sounds. *J.Acoust.Soc.Am.*, 33:179–186. Feb. 1961.

[51] S.S. McCandless. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE. Trans. ASSP*. 22:135–141. April 1974.

[52] S.S. McCandless. Modifications of formant tracking algorithm of avril 1974. *IEEE. Trans. ASSP*, 192–193. April 1976.

[53] N. Mikami et R. Ohba. Pole-zero analysis of voiced speech using group delay characteristics. *IEEE Trans.ASSP*. 32(5):1095–1097. Oct. 1984.

[54] M.L. Minsky et S. Papert. *Perceptrons.* Rapport. MIT. 1969. Press.

[55] Y. Miyanaga. N. Miki. N. Nagai. et K. Hatori. A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system. *IEEE Trans.ASSP*, 30(1):88–96. Feb. 1982.

[56] H. Morikawa et H. Fujisaki. Adaptive analysis of speech based on a pole-zero representation. *IEEE Trans.ASSP*. 77–88. Feb. 1982.

[57] M. Mrayati. Contribution aux études sur la production de la parole: modèles électriques du conduit vocal avec pertes. du conduit nasal et de la source vocale. Thèse de l'Université de Grenoble. 1976.

[58] R.J. Niederjohn et M. Lahat. A zero-crossing consistency method for formant tracking of voiced speech in high noise levels. *IEEE. Trans. ASSP*. 33:349–355. April 1985.

[59] M. Ohsmann. An iterative method for the identification of MA systems. *IEEE. Trans. ASSP*. 36:106–109. Jan. 1988.

[60] A.V. Oppenheim. G.E. Kopec. et J. Tribolet. Signal analysis by homomorphic prediction. *IEEE trans. ASSP*. 24:327–332. Aug. 1976.

[61] A.V. Oppenheim et R.W. Schafer. *Digital Signal Processing.* Prentice-Hall. Inc.. 1975.

[62] P.E. Papamichalis et G.R. Doddington. Time encoding of LPCroots. *IEEE-ICASSP*. pages 589–592. 1982.

[63] S. Parthasarathy et D.W. Tufts. Exitation synchronous modeling of voiced speech. *IEEE Trans.ASSP*. 35(9):1241–1249. Sep. 1987.

[64] E. Parzen. Some recent advances in time series modeling. *IEEE. Trans.Automat.. Control*. 19:723–730. Dec. 1974.

[65] S.M. Peeling. R.M. More. et M.J. Tomlinson. The multi-layer perceptron as a tool for speech pattern processing. *Proc.IoA Autumn Conf on Speech and Hearing*. 1986.

[66] R.W. Prager. T.D. Harrison. et F. Fallside. Boltzmann machine for speech recognition. *Computer Speech and Language*. 3–27. Jan. 1986.

[67] L.R. Rabiner. Lpc prediction error — analysis of its variation with the position of the analysis frame. *IEEE trans. ASSP*. 25(5):434–442. Oct. 1977.

[68] L.R. Rabiner et B. Gold. *Theory and Application of Digital Signal Processing.* Prentice-Hall. 1975.

[69] L.R. Rabiner et R.W. Schafer. *Digital Processing of Speech Signal.* Prentice-Hall. 1978.

[70] J.M Richard. K. Wang. et S. Gay. Lpc speech analysis using the l1 norm. *Proc. IEEE-ICASSP*. pages 485–488. 1985.

[71] D.E. Rumelhart et J.L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Rapport. Cambridge. MA. 1986. MIT Press.

[72] R.W. Schafer et L.R. Rabiner. System for automatic formant analysis of voiced speech. *J.Acoust.Soc.Am.*. 47:634–648. Feb. 1970.

[73] J.F. Seriguat. Modélisation du système vocal par pôles et zéros. *11e ICA. Paris*. pages 111–114. 1983.

[74] M.J. Shensa. Recursive least squares lattice algorithms — a geometrical approach. *IEEE. Trans. Automat. Contr.*, 26:695–702. June 1981.

[75] K.H. Song et C.K. Un. On pole-zero modeling of speech. *Proc. IEEE-ICASSP.* pages 162–165. 1980.

[76] K.H. Song et C.K. Un. Pole-zero modeling of speech based on high-order pole model fitting and decomposition method. *IEEE trans.ASSP.* 31(6):1556–1565. Dec. 1983.

[77] F.F. Soulie. *Le connexionisme.* Rapport. Ecole des Hautes Etude en Informatique Université de Paris 5.. 1987. Support de cours MARI 87-COGNITIVA.

[78] M.D. Srinath et P.K. Rajasekaran. *An Introduction to Statistical Signal Processing with Applications.* John Wiley and Sons, 1979.

[79] K.N. Stevens. Spectral prominances and phonetic distinction in language. *Speech Communication.* 137–144. 1985.

[80] I.M. Trancoso. L.B. Almeida, et J.M. Tribolet. Pole-zero multipulse speech representation using harmonic modeling in the frequency domain. *IEEE-ICASSP*, pages 260–263. 1985.

[81] J.M. Tribolet, L.R. Rabiner. et M.M. Sondhi. Statistical properties of an LPC distance measure. *ICASSP.* pages 739–743. 1979.

[82] A. Waibel. T. Hanazawa. G. Hinton. K. Shikano. et K. Lang. Phoneme recognition: neural networks vs. hidden Markov models. *Proc. IEEE-ICASSP.* pages 107–110. 1988.

[83] S. Wang. H. Ye. et F. Robert. A PCMN neural network for word recognition. *Neuro.88.* Juin 1988.

[84] R.L. Watrous et L. Shastri. Learning phonetic features using connectionist networks: an experiment in speech recognition. *MS-CIS-86-87 LINC LAB44.* Oct. 1986.

[85] L. Wu et J.P. Haton. An arma model with selective frequency range. *Proc. SPEECH'88. Edinburg.* Aug. 1988.

[86] F. Yang, L. Wu. et J.P. Haton. Reconnaissance des mots isolés en utilisant un réseau de neurones. *17e journées d'étude sur la parole. Nancy*, Sep. 1988.

[87] F. Yang. L. Wu. et J.P. Haton. Utilisation d'un réseau de neurones pour la reconnaissance des mots isolés. *Proc. SPEECH'88. Edinburg.* Aug. 1988.

[88] B. Yegnanarayana. Speech analysis by pole-zero decomposition of short-time spectra. *Signal Processing*, 3:5–17. Jan. 1981.

NOM DE L'ETUDIANT : WU Li

NATURE DE LA THESE : Doctorat de l'Université de NANCY I en Informatique

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le - 2 JUIL. 1990 n°1430

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

UNIVERSITE NANCY I
LE PRESIDENT
M. BOULANGE