الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة الجامعات
Ministère aux Universites

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT électronique

PROJET DE FIN D'ETUDES

SUJET

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE PAR LA METHODE STOCHASTIQUE

Proposé par :B.Bousseksou **Etudié par** :Z.Ait-kheddache **Dirigé par** :B.Bousseksou

PROMOTION 91-92

E.N.P. 10, Avenue Hacen Badi - El-Harrach - Alger

<< AU NOM D'ALLAH >>

REMERCIEMENTS

Je tiens à remercier tous ceux qui m'ont aidé et encouragé durant mes dernières années d'étude.

Je remercie vivement Mr Bousseksou, mon promoteur, qui a su me guider et me conseiller pour mener ce projet à son terme.

Je remercie, de même, le personnel de la Bibliothèque et du centre de calcul, pour leur amabilité et toute l'aide précieuse et irremplaçable qu'ils m'ont fournie.

TABLE DES MATIERES

BIBLIOGRAPHIE

[1] R.BOITE : Traitement de la Parole

[2] F.CINARE : Reconnaissance et Synthèse de Parole

[3] IEEE : A Stochastic Segment Model for Phoneme - Based Countinous Speech Recognition (Décembre 1989)

[4] IEEE : LPC Speech Coding Based On Variable-Length Segment Quantization (Septembre 1988)

[5] E. EMERIT: Cours de Phonétique Acoustique

[6] J. LIFERMAN: Les Méthodes Rapides de Transformation du Signal

[7] A. SPATARU: Fondements de La Théorie de Transmission de l'Information

INTRODUCTION GENERALE

Les machines qui parlent et qui reconnaissent la parole vont avoir beaucoup d'applications utiles dans le futur. Beaucoup de temps pourrait-être économisé, en utilisant un ensemble d'instructions parlées à la place d'un ensemble d'instructions imprimées, quand il faudra mettre en marche un équipement complexe, dans les situations où les yeux et les mains sont déjà occupés, où des actions peuvent-être représentées sous formes d'informations nouvelles qui peuvent être mieux représentées sous forme de messages parlés.

Déjà, on peut trouver sur le marché des produits aux capacités limitées certes, mais pouvant rendre de grands services aux utilisateurs potentiels de cette technologie.

En reconnaissance de la parole continue, dans les vocabulaires larges, les mots sont fréquemment modélisés comme étant des suites d'unités linguistiques telles que les phonèmes, autrement dit, un mot est modélisé acoustiquement en concaténant les modèles acoustiques des phonèmes suivant sa chaîne structurée de prononciation entraînée dans le dictionnaire contenant les "transcriptions" des différents phonèmes. L'Avantage de cette approche est qu'il n'est pas nécessaire d'entrainer tous les mots dans le vocabulaire mais uniquement les modèles des différents phonèmes.

La méthode des chaînes de Markov est une modélisation probabiliste de la réalisation acoustique d'un phonème, mais bien que cette dernière a donné de bons résultats en ce qui concerne la modèlisation des phonèmes aux longueurs variables, la reconnaissance de la parole est loin d'atteindre la performance, chose pour laquelle nous proposons à travers ce modeste travail, une nouvelle approche appelée MODELISATION DES PHONEMES EN SEGMENTS STOCHASTIQUES dans laquelle chacun des segments phonétiques, aux longueurs différentes est transformé en un segment de longueur fixée commune.

CHAPITRE I

LA PAROLE NATURELLE

I - 1/- INTRODUCTION

Le phénomène de la production de la parole suscite aujourd'hui l'intérêt des chercheurs. En effet, une meilleure connaissance de ce phénomène est devenue nécessaire pour pouvoir avancer dans les secteurs de la reconnaissance de la parole. Une telle connaissance exige l'étude de plus en plus approfondie des caractéristiques acoustiques de la parole; ainsi que celle des différentes divisions possibles de la parole.

I-2/ - PRODUCTION DE LA PAROLE NATURELLE

Le discours de la parole est une succession de différences de pression de l'air qui engendrées par le système de la phonation (fig I-1)

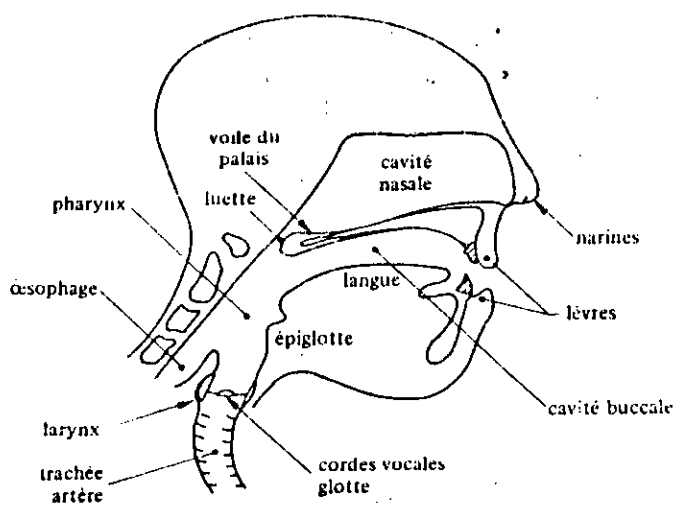


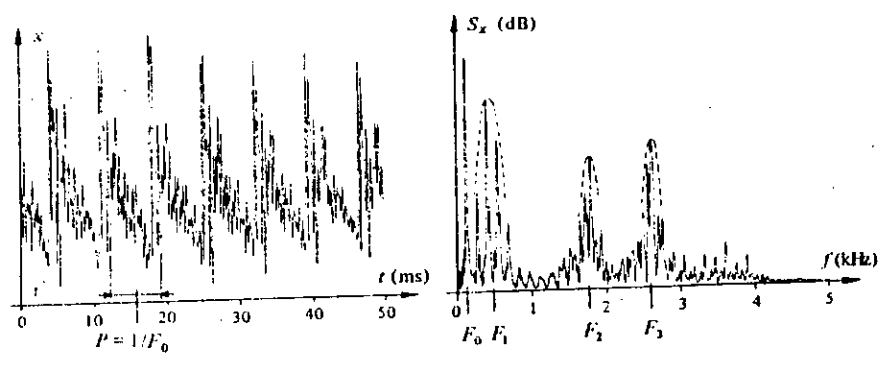
Fig I.1 Système phonatoire humain

Le conduit vocal humain peut être considéré comme une succéssion de tubes ou cavités acoustiques de réactions diverses.

Les sons voisés résultent donc de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales: l'ouverture brusque de la glotte libére la pression accumulée en amont; elle se referme ensuite plus graduèllement.

Un son voisé est un signal quasi-périodique(fig I-2 et I-3).

Sur la figure I-3, on observe les raies qui correspondent aux harmoniques du fondamental F_0. L'enveloppe de ces raies présente des maximums appelés formants et qui correspondent aux fréquences propres Fi (i=1, 2, 3...) du conduit vocal.



I-2 et I-3: Un signal vocal et son spectre

Les trois premiers formants sont essentiels pour caractériser le spectre vocal , les formants d'ordre supérieur ont une influence limitée. Un son non voisé ne présente pas de structure périodique, il peut être considéré comme un bruit blanc filtré par la transmittance du conduit vocal. Remarquons toutefois qu'un son non voisé présente la même structure formantique que le même son voisé.

I-3/- Modélisation de la production de la parole

La modélisation proposée ici utilise le formalisme des systèmes échantillonnés (Transformée en Z).

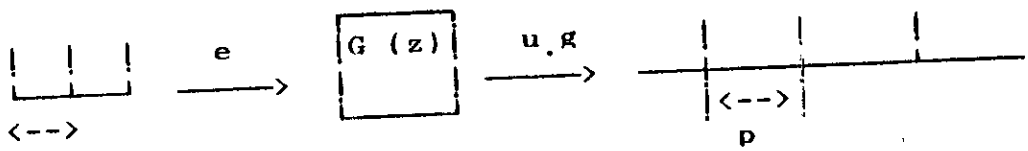


fig I-4 Modélisation de la source pour les sons voisés

Pour les sons voisés, la source est un train périodique d'ondes (fig 4). Ce train d'ondes est modélisé par un filtre passe-bas d'ordre 2 à pôles réels dont la fréquence de coupure est de l'ordre de 100 HZ:

$$G(z) = \frac{A}{(1+\alpha z^{-1})\ (1+\beta z^{-1})}$$

Pour les sons non voisés, la source est un bruit blanc.
On peut assimiler le conduit vocal à une cascade de résonateurs dont la transmittance est de la forme:

$$V(z) = \frac{B}{\prod_{k=1}^{M} (1 + b_{1k}\, z^{-1} + b_{2k}.z^{-2})}$$

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

$$F_k = (1/2\pi).f_0.\ \cos^{-1}\left(\frac{-b_{1k}}{2.\sqrt{b_{2k}}}\right)$$

f_0: fréquence d'échantillonnage.

Le son est finalement émis à travers l'ouverture des lèvres, où:

$$R(z) = C\ (1-z^{-1})$$

En résumé, la transmittance globale entre le train d'impulsions et le son emis :

$$T(z) = G(z).V(z).R(z)$$

$$= \frac{\sigma.\ (1-z^{-1})}{(1+\alpha.z^{-1}).(1+\beta.z^{-1}).\ \prod_{k=1}^{M} (1+ b_{1k}\cdot z^{-1} + b_{2k}.z^{-2})}$$

On suppose que l'un des pôles de G(z) est proche de l'unité donc:

$$T(z) = \frac{\sigma}{(1+\alpha z^{-1}).\prod_{k=1}^{M}(1+b_{1k}z^{-1}+b_{2k}.z^{-2})} = \frac{\sigma}{A(z)}$$

On a posé: $$A(z) = (1+\alpha.z^{-1}).\prod_{k=1}^{M}(1+b_{1k}z^{-1}+b_{2k}.z^{-2})$$

$$= 1+\sum_{i=1}^{2M+1}(a_i.z^{-i})$$

La transmittance de ce modèle est dite tout-pôles; son inverse le polynôme A(z) est la transmittance du filtre inverse. Ses limitations sont cependant évidentes. En premier lieu, la source est soit un train périodique d'impulsions, soit un bruit blanc; les sons fricatifs voisés (v,z...) ne peuvent pas être produits par ce modèle.

En second lieu, la production des sons nasalisés fait intervenir deux cavités associées en parallèle; la transmittance correspondante est de la forme:

$$\frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \frac{\sigma 1.A_2(z)+\sigma_2.A_1(z)}{A_1(z).A_2(z)}$$

Toutefois, malgré ses limitations, la transmittance tous-pôles est la base de la modélisation par prediction linéaire. Pour pouvoir assimiler le numérateur à une constante, on doit surestimer le degré du dénominateur.
On remarque aussi que le signal vocal est stationnaire pendant des intervalles de temps de l'ordre de 20 ms.
Durant cet intervalle, les coefficients de T(z) sont constants.

I-4/ - LES PHONEMES

Les linguistes ont défini le phonème comme étant la plus petite unité phonétique. La langue française contient 36 phonèmes comprenant 16 voyelles, 17 consonnes, et 3 semi-consonnes. Sur la figure I.5, on peut voir une classification des phonèmes.

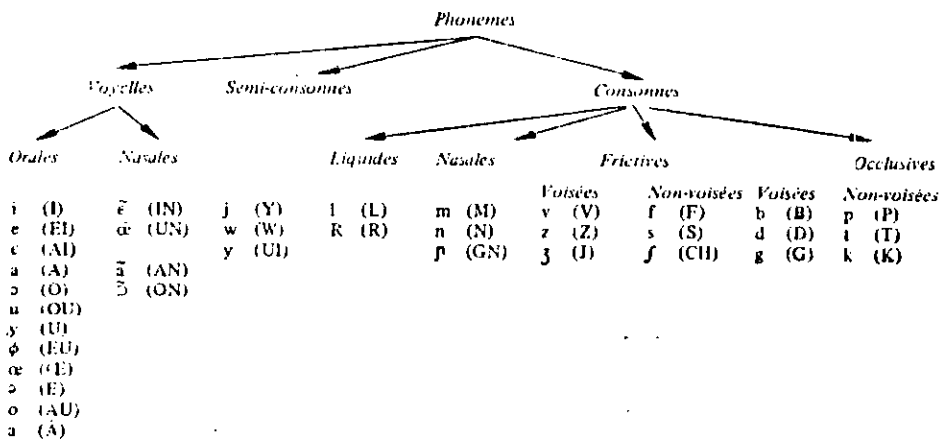


fig I-5 phonèmes de la langue française, d'après Kunt [97]

La production d'un phonème donné laisse toutefois place à une certaine variabilité sur le plan acoustique.

Les voyelles orales (i,o,u,...) sont émises sans intervention de la cavité nasale. Pour les voyelles nasales, le conduit nasal est couplé à la cavité buccale et l'émission se produit à la fois par les narines et par la bouche. Les nasales font aussi intervenir la cavité nasale.

Les consonnes fricatives résultent de l'écoulement de l'air dans une constriction étroite en un point du conduit vocal, en particulier au niveau des lèvres et des dents. Les sons fricatifs sont non voisés (f, s,..) ou voisés (v, z,...).

Les consonnes occlusives correspondent à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal puis relâchée brusquement. La période d'occlusion est appelée la phase de tenue.

Pour les occlusives voisées un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées, la tenue est un silence.

	1 er for. (Hz)	2éme for.(Hz)
a	750	1350
o	375	750
i	250	2500
e	400	2200
u	250	600

Fig I-6 les deux premiers formants des voyelles

La table I-6 représente les voyelles pour les deux premiers formants . Les deux premiers formants caractérisent déjà beaucoup mieux toutes les voyelles.

I.5/ - CONCLUSION

Le mécanisme de la production de la parole qui a été décrit dans ce chapitre conduit très naturellement à une modélisation particulière appelée MODELISATION AUTOREGRESSIVE.
Nous disposons là d'un outil très commode pour procéder à l'analyse du signal vocal. On prendra quand-même quelques précautions car l'hypothèse de stationnarité du signal n'est valable que durant de courts intervalles de temps.
Les procédures de modélisation doivent être adaptées pour en tenir compte.

CHAPITRE II

CHAPITRE II

ANALYSE DE LA PAROLE

II-1 INTRODUCTION :

Pour pouvoir réaliser un système de reconnaissance de la parole, il est très important de faire passer ce signal par une analyse qui consiste à tirer du signal les paramètres pertinents capable de représenter correctement toutes ses caractéristiques.

La première opération d'analyse consiste à tronçonner le temps en intervalles d'environ 10 à 20 ms . Ces intervalles sont appelés " fenêtres temporelles " . L'analyse du signal parole repose sur l'hypothèse fondamentale suivante :

" Le signal à analyser est stationnaire pendant toute la durée de la fenêtre d'analyse, c'est à dire que ses propriétés en particulier spectrales ne varient pas sur la longueur de la fenêtre ".

L'analyse sera faite successivement sur chacune des fenêtres . Parmi les méthodes d'analyse citons :
l'analyse cepstrale; l'analyse par prédiction linéaire et l'analyse par banc de filtre; peut être numérique (logiciel) ou matériel (vocodeur à canaux par exemple).

II-2 PRETRAITEMENT

Avant d'analyser le signal parole, on doit passer par unprétraitement qui consiste en un échantillonnage, une préaccentuation et un fenêtrage.

II-2-1 ECHANTILLONNAGE

Dans le cas des méthodes numériques, ce traitement ne s'effectue pas directement sur les signaux analogiques à temps continu fournis par un microphone.

Les signaux analogiques sont échantillonnés, codés, puis rangés sous forme numérique dans une mémoire pour être traité par ordinateur.

Le spectre du signal parole est limité à 6 KHZ, il conserve donc ses caractéristiques.

Selon le théorème de SHANNON : $Fe \geq 2 . Fm$

Avec Fe : Fréquence d'échantillonnage

Fm : Fréquence maximale du signal parole

donc en théorie 12 KHZ vont suffire comme fréquence d'échantillonnage, et si on prend 256 échantillons par fenêtre de 20 ms on aura pour fréquence d'échantillonnage :

$$Fe = \frac{256}{2 \times 10^{-2}} = 12,8 \text{ KHZ}$$

II-2-2 PREACCENTUATION :

Du fait de l'évolution du spectre du signal parole en haute fréquence dans les deux milieux : Le conduit vocal et l'air extérieur, il subit une baisse d'énergie de 6 dB par octave.

La préaccentuation est nécessaire pour rétablir le niveau énergétique . C'est le cas des sons voisés car le spectre est surtout localisé dans la partie haute fréquence.

II-2-3 FENETRAGE :

Le fenêtrage limite l'amplitude des rebonds fréquenciels dûs à la limitation temporelle du signal.

Il existe plusieurs types de fenêtres :

- Fenêtre rectangulaire :

$W_r(k) = A$ pour $k = 0,N$ Avec A : constante

pour un nombre d'échantillons N = 9, L'atténuation = -13 dB .

- Fenêtre triangulaire :

$W(k) = 1 - 2|k| / N$ pour $|k| \leq N / 2$

L'amplitude des lobes secondaires est inférieure à celle de la fenêtre rectangulaire, mais cette fenêtre ne nous parrait pas très performante du fait de l'élargissement du pic central.

- <u>fenêtre de Hanning</u> :

$W_h(k) = 1 / 2 (1 + \cos (2 \pi . K / N))$ pour $|k| \leq - N / 2$

Son premier lobe secondaire n'est pas vraiment atténué, par contre les autres lobes secondaires sont atténués à environ -43 dB du pic central.

- <u>Fenêtre de Hamming</u> :

$W_h(k) = 0,54 - 0,46 \cos (2. \pi . (k - 1) / N)$

pour k = 1, N. Elle a 99,96 % de son énergie dans le lobe principal et le lobe secondaire .

Le plus important est de 40 dB en dessous du lobe principal .

- <u>Critère de choix de la fenêtre</u> :

Le choix de la forme particulière d'une fonction fenêtre dépend principalement de la largeur du pic central et de l'amplitude des lobes secondaires relative à celle du pic central . Si le pic central est large, les transitions rapides de la transformée de Fourrier du signal original sont très mal approximées.
D'après le résultat selon lequel 99,96 % d'énergie est concentrée dans le lobe principal, le niveau du premier lobe secondaire est 43,9 au dessous du lobe principal ; toutes ces qualités nous conduit à opter pour la fenêtre de Hamming.

II-3 ANALYSE CEPSTRALE :

Le signal de parole contient des informations phonétiques essentiellement contenues dans le spectre du signal et des informations prosodiques essentiellement contenues dans le pitch (la fréquence du fondamental).
le conduit vocal module le signal pulsé qui est fourni par la vibration des cordes vocales.
Il y a combinaison par convolution des deux filtres qui sont les cordes vocales et le conduit vocal.

soit h(n) le signal issu de la source d'excitation et x(n) la fonction de transfert du conduit vocal indépendante de h(n)

$$Y(n) = \sum_{k=-\infty}^{k=+\infty} [h(n-k).x(k)] = h(n)*x(n) \qquad (II-1)$$

si D est l'opérateur de déconvolution

$$D[Y(n)] = D[h(n)*x(n)] = D[h(n)] + D[x(n)] \qquad (II-2)$$

$$Y'(n) = h'(n) + x'(n)$$

h(n) et x(n) sont régulièrement échantillonnés, * est l'opération de convolution .

Par la transformée en Z on obtient :

$$Y(Z) = H(Z).X(Z)$$

$$Ln|Y(Z)| = Ln|H(Z)| + Ln|X(Z)| \qquad (II-3)$$

Le cepstre " C " sera alors la transformée inverse de Z de" Ln|Y(n)| " telque :

$$C(n) = (1/N) \cdot \sum_{k=1}^{N} \left(\mathrm{Log} |Y(f)| e^{i2\pi nf/N} \right) \qquad \text{(II-4)}$$

avec N: nombre de points dans une fenêtre

n: rang du coefficient cepstral

Les coefficients cepstraux se calculent par l'équation :

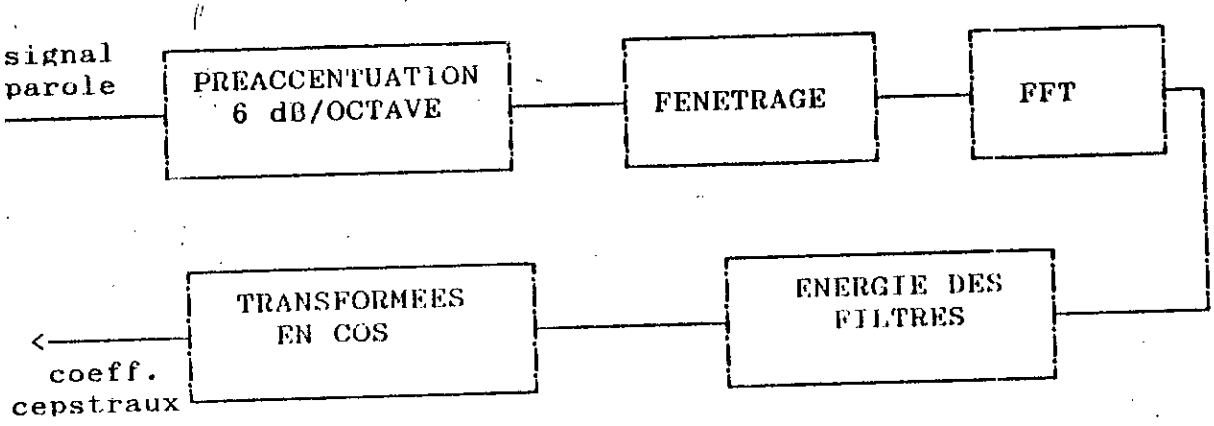
$$C(n) = (1 / N_f) \cdot \sum_{k=1}^{N_f} \mathrm{Log}\left[|E(k)| \cos[n (k-1) / N_f)] \right] \qquad \text{(II-5)}$$

avec: n: le rang du coefficient cepstral

k: le numéro du filtre d'énergie E(k)

N_f: le nombre total des (16 à 25) filtres triangulaires

ETAPES D'ANALYSE :



ETAPES DE L'ANALYSE

La préaccentuation est une dérivation numérique :

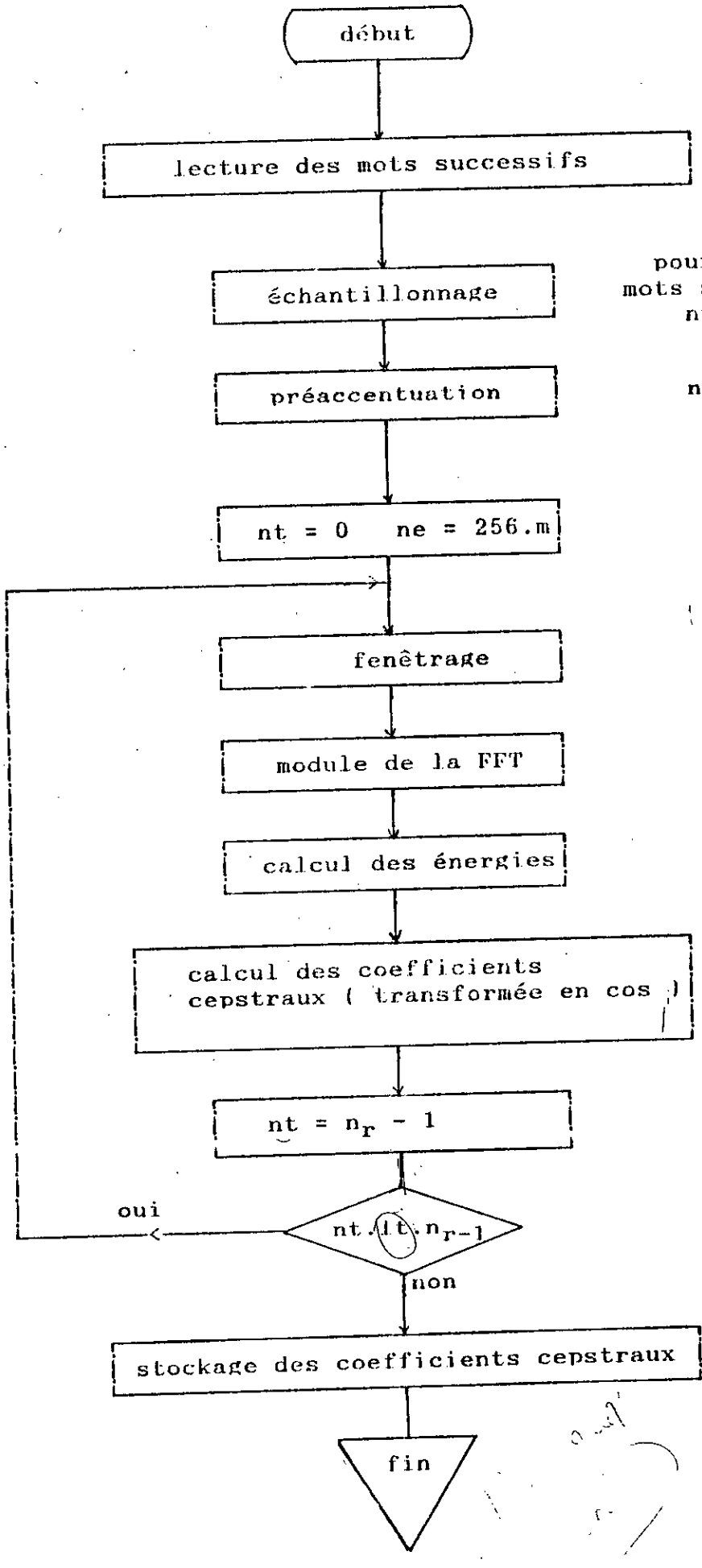
$Y(k+1) = Y(k+1) - Y(k)$

Le calcul du spectre se fait grâce à la FFT.

- Echelle Mel : l'oreille est sensible à une échelle quasi-logarithmique dite échelle Mel qui est linéaire sur le premier KHZ et logarithmique au delà . L'échelle logarithmique tente d'éviter de donner un même poids à des zones de fréquence qui n'ont pas la même densité d'information.

ORGANIGRAMME

Nous présentons ici l'organigramme de la méthode d'analyse cepstrale.
si on prend par exemple un phonème de 0,8 seconde, celui-ci sera tronçonné en 40 trames de 20 ms chacune et dans chaque trame nous avons 256 échantillons.



pour chacun des mots successifs on a
nt:nombre de trames

ne:nombre d'échantillons

n_r:nombre total d'echantillons du mot

II-4/ ANALYSE PAR PREDICTION LINEAIRE :

Cette technique permet d'estimer des paramètres comme le pitch, les formants, le spectre instantané.

Le signal est approximé par un polynôme d'ordre p.

Un échantillon de signal parole est représenté par les p échantillons qui le précèdent par une combinaison linéaire.

$$S_p(n) = \sum_{k=1}^{p} [a(k).S(n-k)] \qquad (II-6)$$

n = 1,N
p: ordre de prédiction

Le signal parole n'étant pas polynômial, les coefficients LPC a(k) sont calculés en minimisant la somme des carrés des différences entre les échantillons réels de la parole s(n) et la valeur estimée par combinaison linéaire S_p.

$$E = \sum_{n=1}^{N+p} [\, [\, S(n) - \sum_{k=1}^{p} a(k) \, . \, S(n-k) \,]^2] \qquad (II-7)$$

Il existe plusieurs méthodes de minimisation de l'erreur quadratique citons entre autres :

- La Méthode de covariance

- La Méthode d'autocorrélation

- La Méthode en treillis

La prédiction linéaire est une méthode précise d'estimation des paramètres de la parole de plus, les calculs de certaines méthodes de résolution sont relativement rapides.
Dans notre cas, la sommation se fait sur un nombre N fini d'échantillons pendant lesquels les caractéristiques du conduit vocal et de la source sont constantes, ce qui conduit à la stationnarité de l'échantillon considéré, pour cela on utilise la méthode d'autocorrélation.

- Méthode d'autocorrélation :

Détermine les coefficients a(k) pour lesquels E est minimale.

Cela se fait en annulant la dérivée de E :

$$\partial E/\partial R(i) = \sum_{n=1}^{N+p} 2[s(n) - \sum_{k=1}^{p} a(k).s(n-k)].S(n-i) = 0 \qquad (II-8)$$

$$\sum_{n=1}^{N+p} S(n).S(n-1) = \sum_{k=1}^{p} a(k).\sum_{n=1}^{N+p} S(n-k).S(n-i) \qquad (II-9)$$

on pose :

$$c(i,k) = \sum_{n=1}^{N+p} S(n-k).S(n-i) \qquad (II-10)$$

$$\sum_{n=1}^{N+p} S(n).S(n-i) = \sum_{k=1}^{p} a(k).c(i,k) \qquad \text{(II-11)}$$

Avec i = 1,p et k = 1,p

d'où :

$$c(i,k) = \sum_{n=1}^{N+i-k} S(n).S(n+i-k) \qquad \text{(II-12)}$$

c(i,k) est la matrice d'autocorrelation, c'est une matrice carrée d'ordre "p" dite aussi "matrice de Toeplitz" car les éléments situés symétriquement de part et d'autre de la diagonale sont égaux.

on note R(k) la fonction d'autocorrélation définie par :

$$R(k) = \sum_{n=1}^{N+k} S(n).S(n+k) \qquad \text{(II-13)}$$

avec k=1,p

$$R(k) = R(-k)$$

$$\text{d'où:} \quad c(i,k) = R(|\, i-k\,|) \qquad \text{(II-14)}$$

L'équation II-8 peut s'écrire sous forme :

$$\sum_{K=1}^{p} a(k).R(|i-k|) = R(i) \qquad \text{(II-15)}$$

avec i=1, p

L'équation II-14 peut s'écrire sous forme matricielle

$$\begin{bmatrix} R(0) & R(1) \ldots\ldots\ldots & R(p-1) \\ R(1) & R(0) \ldots\ldots\ldots & R(p-2) \\ \vdots & \vdots & \vdots \\ R(p-1) & R(p-2) & \ldots\ldots\ldots R(0) \end{bmatrix} \begin{bmatrix} a(1) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} R(1) \\ \vdots \\ R(p) \end{bmatrix} \qquad \text{(II-16)}$$

De nombreuses méthodes permettent de résoudre le système d'équation linéaire telles que :

- méthode de Gauss-Seidel
- Méthode de Jacobi
- Méthode de Gauss-Jordan
- Méthode de Durbin

Mais en tenant compte de la rapidité d'exécution et de l'encombrement mémoire réduit, nous avons choisi pour résoudre le système d'équation d'auto-correlation la méthode de Durbin.

Méthode de Durbin :

$$D(i)=R(i)-\sum_{j=1}^{i-1}\left[A(j,i-1).R(i-j)/E(j-1)\right] \qquad (II-17)$$

$$E(i)=\left[1-D^2(i)\right].E(i-1) \qquad (II-18)$$

$$A(j,i) = A(j,i-1)-D(i).A(i-j,i-1)$$

$$A_1(j) = a(j,12)$$

où

D(i) : coefficients de réflexion

E(i) : Erreur quadratique

A(j,i): coefficients de prédiction

$A_1(j)$: coefficients d'autocorrelation

avec $i=1,12$ et $j=1,i-1$

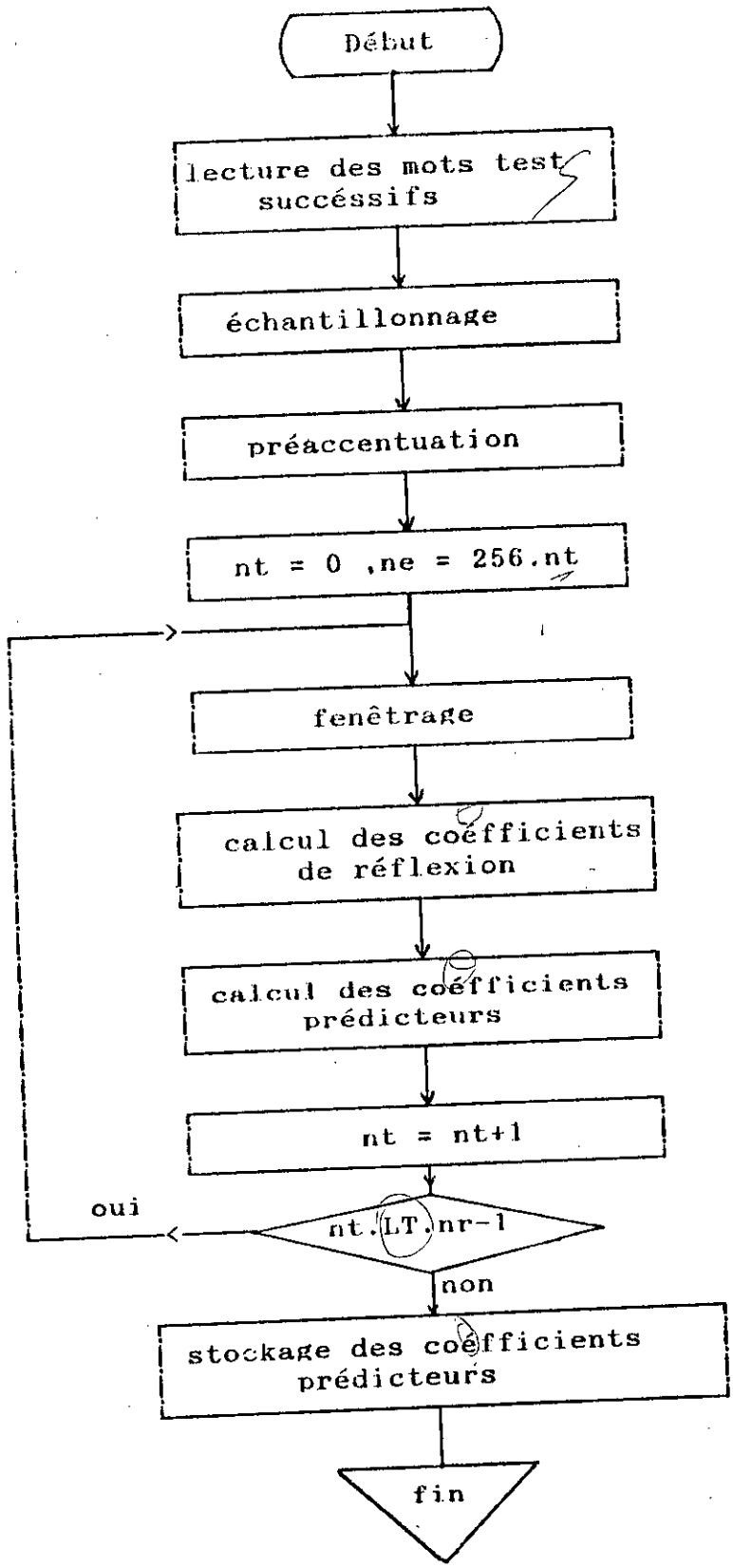
avec les conditions initiales suivantes :

$E(1) = R(1)$

$D(2) = [R(2)/R(1)]$

$A(2,2) = D(2)$

$E(2) = [1 - D^2(2)].R(1)$


Début
lecture des mots test successifs
échantillonnage
préaccentuation
nt = 0 ,ne = 256.nt
fenêtrage
calcul des coefficients de réflexion
calcul des coefficients prédicteurs
nt = nt+1
oui
nt.LT.nr-1
non
stockage des coefficients prédicteurs
fin

II-5 CONCLUSION :

Dans ce chapitre nous avons étudié deux méthodes d'analyse " la méthode cepstrale " et " la méthode de prédiction linéaire ". La méthode cepstrale inverse le cepstre du signal et favorise les hautes fréquences par rapport aux basses fréquences qui ne présentent pas le même intérêt puisqu'elles sont spécifiques au locuteur (pitch).

De plus la méthode cepstrale est économique du point de vue représentation étant donné que 8 coefficients au lieu de 12 par trame (comme c'est le cas de la LPC) suffisent à bien représenter le signal.

Il est également possible grâce à la méthode cepstrale de réaliser une économie dans le temps de calcul.

Faire un traitement pour 8 coefficients est évidement plus rapide que de le faire pour 12, ce qui nous rapproche de l'aspect temps réel qui est un des buts les plus recherchés.

CHAPITRE III

CHAPITRE III

RECONNAISSANCE DE LA PAROLE

Il a toujours été considéré que l'oreille procèdé à une certaine analyse spectacle continue dont les résultats sont transmis au cerveau. L'identification d'un mot consiste donc en une succession de comparaisons de spectres, si l'on se limite au niveau acoustique.

Un spectre peut être considéré d'une façon objective par différents ensembles de paramètres; en particulier le spectre du modèle AR, qui correspond très bien à l'enveloppe du spectre vocal, peut être caractérisé par les coefficients $a_p(i)$.

Un ensemble ordonné de paramètres caractérisent un spectre vocal est appelé vecteur acoustique ou vecteur spectral.

Un mot est bien sûr, constitué par une suite de vecteurs acoustiques.

La principale difficulté consiste à reconnaître un mot. Pour cela on dispose de trois approches.

La première exploite la notion de distance entre deux mots. L'algorithme DTW permet d'optimiser cette distance. Sur cette méthode, nous ne donnerons aucun détail. La deuxième approche apparemment très différente consiste à spéculer sur un modèle statistique de production pour chaque mot du vocabulaire. Et la troisième, qui ne diffère pas trop de la deuxième, consiste à modéliser les phonèmes en segments stochastiques.

Nous allons approfondir et détailler cette méthode.

III.1 METHODES STOCHASTIQUES

On distingue deux méthodes:

La première est celle qui a déjà été citée et qui s'appuit sur l'hypothèse qu'un automate peut emettre un mot ou un phonème. A chaque état interne de l'automate correspond l'émission d'un vecteur acoustique.

C'est donc la succession des états internes qui provoque la production d'un mot. On dit de ce modèle qu'il est doublement stochastique car d'une part les transitions d'un état à un autre se font selon certaines probabilités et d'autre part l'émission d'un vecteur acoustique obéit à une probabilité qui dépend de l'état.

Ce modèle s'identifie tout naturellement à une chaîne de Markov qui est un processus causal.

En tout état de cause les variables de sortie de ce modèle sont supposées observables, les variables d'état ne le sont pas.Le problème sera celui de déterminer à chaque instant l'état le plus probable compte tenu d'une séquence d'observations donnée et de la connaissance supposée des paramètres du modèle.

Ce problème et le modèle qui lui est associé, justifient l'utilisation courante de l'expression anglo - saxonne de "Hidden Markov Modeling" (HMM).

Rappelons encore une fois que cette méthode ne sera pas développée. La deuxième méthode stochastique est celle que nous développerons ici. Cette nouvelle approche est appelée Modélisation des phonèmes en segments stochastiques, celle-ci est introduite pour modéliser les phonèmes à durées variables. Le phonème X est observé comme étant une séquence de vecteurs acoustiques. La durée de ce phonème X est variable selon le phonème et les cordes vocales du locuteur.

La modèlisation des phonèmes en segments stochastiques consiste en 1)° une normalisation temporelle du segment X de longueur variable en un segment Y de longueur fixée, et en 2)° une densité multidimensionnelle des paramètres du segment normalisé Y; cette densité étant une gaussiène.

Le modèle segmental représente la structure spectrale/temporelle à travers tout le phonème. Ce modèle permet aussi l'incorporation dans Y des "caractéristique acoustiques - phonétiques " contenues dans le segment phonétique X, ceci en plus des habituelles caractéristiques spectrales qui sont utilisées dans la méthode des chaînes de Markov cachées et la DTW.

Dans le travail qui reste, nous exposerons quelques définitions sur la segmentation de la parole, le modèle stochastique du segment phonétique, l'algorithme de reconnaissance (qui réalise simultanément la segmentation et la reconnaissance de ces derniers) et puis c'est la phase d'apprentissage dans laquelle, à partir de la parole segmentée en phonèmes donnée, nous estimerons les paramètres à savoir les moyennes et les covariances correspondants aux différents modèles phonétiques qui vont constituer le dictionnaire, puis vient la phase de ré-estimation de ces modèles en utilisant un algorithme de segmentation automatique et finalement nous présenterons un algorithme itératif qui va converger localement vers une collection optimale de modèles phonétiques.

III.2 Segmentation de la parole

La segmentation est un des problèmes les plus difficiles à résoudre. La situation idéale serait celle où chaque segment corresponderait à un phonème. Differéntes méthodes existent baseés sur les courbes de variations d'energie, ou de variabilité du signal.

Dans ce paragraphe, nous nous sommes bornés à citer brièvement les différentes approches pour la segmentation de la parole en phonèmes car dans l'algorithme de reconnaissance qu'on vérra plus loin, la segmentation et la reconnaissance ont lieu semultanément.

Les differentes approches pour segmenter la parole en segments phonètiques sont :

a - Segmentation en segments d'états stables :

Cette approche repose sur le fait qu'il est possible d'identifier des segments d'état stable du signal de parole qui sont habituellement decrits en terme de changement de pression de l'air comme une fonction du temps.

b - Segmentation avec l'aide d'un modèle de référence :

Dans cette approche, le mot à segmenter est partagé en états spectraux. La plupart des phonèmes correspondent à un état spéctral. A titre d'exemple, le mot "nenaanene" (mot sans signification) contient 10 états spectraux:
Le silence avant le mot /#/, les phonèmes /n/, /e/, /n/, /a/, /n/,/e/,/n/, /e/, et le silence après le mot /#/.

III.3/ Modélisation des phonèmes en segments stochastiques

Cette méthode consiste en une transformation : la normalisation temporelle des segments phonétiques; et une modélisation de ces derniers par une densité gaussienne multidimensionnelle aprés que ces segments aient subit la normalisation temporelle.

- Motivations

Les principaux avantages qu'offre cette méthode sont :

1 - L'incorporation dans le segment phonétique normalisé Y, des "caractéristiques acoustiques - phonétiques" contenues dans le segment X avant normalisation (ceci en plus des habituelles caractéristiques spectrales qui sont utilisées dans la méthode des chaînes de markov cachées et la DTW).

2 - Le modèle segmental est une représentation du phonème, donc en considérant la parole à un niveau segmental, nous pouvons mieux "capturer" la structure spectrale / temporelle à travers toute la durée du phonème ceci en comparaison à ce que peut permettre une considération trame par trame.

III.3.1/ - Normalisation temporelle :

Soit X une suite de vecteurs acoustiques; cette séquence étant un segment de parole de longueur variable et qui représente un phonème, lequel phonème peut s'ecrire comme suit :

$$X = [x_1, x_2, \ldots\ldots, x_L]$$

avec x_i : Vecteur acoustique à p composantes

L : longueur du segment phonétique

Donc X peut être considéré comme une matrice pxL dont le nombre de colonnes est variables suivant le phonème et suivant la variabilité des cordes vocales des locuteurs.

X étant donné, on peut trouver sa représentation en longueur fixe qu'on notera Y.

Y est un segment phonètique normalisé et de longueur fixe c'est-à-dire la même longueur que tous les segments phonètiques ayant subi la normalisation temporelle et ceci quelle que soit la variabilité en longueur des phonèmes avant normalisation temporelle.

Y peut s'ecrire comme suit :

$$Y = XT_L = [y_1, y_2, \ldots . y_m]$$

Avec T_L : Matrice Lxm et qui permet d'effectuer la normalisation tomporelle

m : Longueur du segment phonètique normalisé et qui est la même pour tous les segments ayant subi la normalisation

y_j : Vecteur acoustique à p composantes après normalisation temporelle (refenétrage)

Donc de même, Y peut être considéré comme une matrice pxm dont le nombre de colonnes m est fixé (quel que soit le nombre de colonnes de la matrice acoustique X).

Définition

La normalisation temporelle consiste à choisir m instants équidistants, instants auxquels la trajectoire du segment X est refenetrée.

La Matrice T_L et ses propriètés :

Comme il a déjà été indiqué, la matrice T_L est de dimension Lxm, quant à la structure de celle-ci elle peut être trouvée comme suit :
Si on designe par t_{rs} un élément de la matrice T_L correspondant à la r^{eme} ligne et à la s^{eme} colonne, t_{rs} peut s'exprimer comme suit :

$$t_{rs} = 1 - \alpha \quad \text{pour} \quad r = \lceil\beta\rceil + 1 \quad ; \quad s = 1,2,\ldots,L$$
$$= \alpha \quad \text{pour} \quad r = \lceil\beta\rceil + 2$$

avec $\alpha = \beta - \lceil\beta\rceil$; $\qquad \beta = \dfrac{(s-1)(l-1)}{(m-1)}$

et $\lceil\beta\rceil$ désigne la valeur entière de β n'excedant pas β.
Quant à l'avantage de cette transformation, il reside dans le fait que celle-ci réalise une compression de données efficace, et permet d'utiliser facilement des algorithmes de reconnaissance et de segmentation automatiques.

III.3.2 / - le modèle probabiliste

Comme il a été déjà mentionné, le modèle segmental est une gaussienne multidimentionnelle basée sur le segment phonétique normalisé Y.
Si on désigne par α un modèle phonétique contenu dans le dictionnaire du système, la densité $P(Y/\alpha)$ ou bien d'une manière équivalente $Ln[P(Y/\alpha)]$, peut être interprétée comme etant le score résultant de la comparaison des caractéristiques des phonèmes à savoir le phonème normalisé Y et le modèle phonétique α (qui est lui même normalisé).

Le lagorithme de la probabilité conditionnelle correspondant au segment phonétique Y et au modèle phonètique α peut s'exprimer d'une manière sommaire comme suit :

$$\text{Ln}\,[P(Y/\alpha)] = \sum_{j=1}^{m} \text{Ln}[P_j(y_j)/\alpha] \ldots\ldots\ldots \qquad \text{(III-1)}$$

tout en sachant bien que y_j est le j^{eme} vecteur acoustique de la suite normalisée Y et m le nombre de vecteurs acoustiques constituant ladite suite.

Dans l'expression (III.1), $P_j\ (y_j/\alpha)$ désigne un modèle gaussien p-dimensionnel dans lequel est mis en jeu le j^{eme} vecteur acoustique y_j, par suite on peut poser :

$P_j\ (y_j/\alpha) \sim N(\mu_j, C_j)$ dans laquelle N désigne la loi normale de
moyenne μ et de cavariance C: μ_j et C_j sont respectivement le j^{eme} vecteur acoustique et sa matrice covariance, lequel vecteur fait partie du modèle phonétique α. Le vecteur μ_j est mis en comparaison avec le vecteur y_j appartenant à la suite normalisée Y.

III.4/ - Algorithmes de reconnaissance

Dans ce paragraphe, nous decrivons l'algorithme de reconnaissance en utilisant les modèles segmentaux gaussiens.
Le but de cet algorithme est de maximiser le taux de reconnaissance, lequel taux désigne la probabilité pour que le phonème reconnu α soit identique au modèle phonétique α.
Pour accomplir ceci, nous choisissons une suite de phonèmes qui maximise la probabilité des segments phonétiques normalisés: c'est la probabilité maximum à postériori (PMA) de la suite de segments.

Plus précisément, les segments phonètiques issus de l'analyse sont premièrement transformés en segments phonètiques normalisés, puis à l'aide du calcul des probabilités de ces segments et selon le dictionnaire des modèles phonètiques, on peut trouver la probabilité maximum à postériori de la suite de segments phonétiques à reconnaître.

Avant de détailler, nous signalons que deux cas sont envisagés ici, le premier est celui où la segmentation en phonèmes est supposée réalisée préalablement, ceci uniquement dans le but de clarifier les idées. Le deuxième cas, et c'est le plus important est celui où la segmentation est supposée non réalisée, et dans lequel on généralisera en établissant un algorithme qui permettera de réaliser simultanément la segmentation en phonèmes et la reconnaissance de ces derniers: on parle de la reconnaissance <u>automatique</u> de la parole.

1° cas : <u>la segmentation est supposée réalisée préalablement</u>

Commençons par le cas le plus simple celui où la reconnaissance ne porte que sur un seul phonème. L'algorithme de reconnaissance est tout simplement la régle de la probabilité maximale à postériori dans laquelle est mis en jeu le segment normalisé Y à reconnaître :

$$\hat{\alpha} = \arg\max_{\alpha} p(\alpha/Y) \quad \text{.....} \qquad \text{(III.2)}$$

ou bien, d'une manière equivalante, en utilisant la régle Bayesienne, on a:

$$\hat{\alpha} = \arg\max_{\alpha} \mathrm{Ln}[P(Y/\alpha).P(\alpha)] \qquad \text{(III.2)}$$

avec $\hat{\alpha}$: phonème reconnu

$p(\alpha)$: probalité à priori du modèle phontique occurent appartenant au dictionnaire .

Quant à l'expression $Ln[P(Y/\alpha).P(\alpha)]$, elle peut s'exprimer comme suit :

$$Ln[P(Y/\alpha).P(\alpha)] = Ln[P(Y,\alpha)]$$

avec :

$$Ln[P(Y,\alpha)] = -(mp/2)Ln(2\pi)-(1/2)\sum_{j=1}^{m}[(y_j-\mu_j(\alpha))^T.$$

$$.(C_j(\alpha))^{-1}.(y_j-\mu_j(\alpha))+Ln|C_j(\alpha)|] \qquad (III.4)$$

Expression dans laquelle $\mu_j(\alpha)$ et $C_j(\alpha)$ sont respectivement le j^{eme} vecteur acoustique et sa matrice covariance correspondante, lequel vecteur fait partie du modèle phonétique α occurrent, la notation $|C_j(\alpha)|$ designe le déterminant de la matrice covariance C_j.

Maximiser $Ln[P(Y,\alpha)]$ revient à minimiser la fonction suivante:

$$D(Y,\alpha)=\sum_{j=1}^{m}[(y_j-\mu_j(\alpha))^T.(C_j(\alpha))^{-1}.(y_j-\mu_j(\alpha))+Ln|C_j(\alpha)|]..(III.5)$$

d'où il vient :

$$\hat{\alpha} = \arg\min_{\alpha}[D(Y,\alpha)] \qquad (III.6)$$

en définitive on a :

$$\hat{\alpha}=\arg\min_{\alpha}[\sum_{j=1}^{m}[(y_j-\mu_j(\alpha))^T(C_j(\alpha))^{-1}(y_j-\mu_j(\alpha))+Ln|C_j(\alpha)|)]](III.7)$$

Avec $\min_{\alpha}$ désignant ici la recherche du minimum en comparant Y avec tous les modèles phonètiques ocurents contenus dans le dictionnaire.

Remarque : Les modèles phonètiques sont eux-mêmes normalisés c'est-à-dire ils sont constitués du même nombre de vecteurs acoustiques que les segments normalisés Y à reconnaître.

Maitenant supposons qu'on a à reconnaitre une suite de segments phonètiques, laquelle suite sera exprimée comme suit :

$\underset{\sim}{X}=\{X_i\}_{i=1,n}$ n : Nombre de segments phonètiques

Aprés avoir subi la normalisation temporelle, on obtient une suite de segments phonètiques normalisés, qu'on peut exprimer de la manière suivante:

$\underset{\sim}{Y}=\{Y_i\}_{i=1,n}$

En supposant l'indépendance des segments X_i (donc des segments Y_i),l'algorithme de reconnaissance sera celui où chaque segment Y_i subira la comparaison avec tous les modèles phonètiques occurrents du dictionnaire pour en suite trouver la probabilité maximale à postériori correspondant au phonème reconnu αi.
Alors la suite de phonèmes reconnue est donnée par :

$$\underset{\sim}{\hat{\alpha}} = \arg \max_{\underset{\sim}{\alpha}} [Ln[P(\underset{\sim}{Y}/\underset{\sim}{\alpha})\ P(\underset{\sim}{\alpha})]] \qquad (III.8)$$

avec $\underset{\sim}{\hat{\alpha}} = \{\alpha_i\}_{i=1,n}$

$P(\underset{\sim}{\alpha})$ désigne les probalités à priori des modèles phonètiques qui subissent la comparaison avec les phonèmes normalisés qui constituent la suite $\underset{\sim}{Y}$.

$$\mathrm{Ln}[P(\underline{Y}/\underline{\alpha})] = \sum_{i=1}^{n} \mathrm{Ln}[P(Yi/\alpha i)P.(\alpha i)] \qquad \text{(III.9)}$$

L'indice i dans la notation "α_i" veut dire tout simplement que l'ensemble des modèles phonètiques sont comparés avec le i^{eme} segment à savoir Y_i, ceci d'une part, d'autre part la suite de phonèmes reconnue peut s'ecrire de la façon suivante :

$$\hat{\underline{\alpha}} = \arg \max_{\underline{\alpha}} \left[\sum_{i=1}^{n} [\mathrm{Ln}[P(Y_i/\alpha_i).P(\alpha_i)]] \right] \qquad \text{(III.10)}$$

ou bien :

$$\hat{\underline{\alpha}} = \arg \min_{\underline{\alpha}} \left[\sum_{i=1}^{n} D(Y_i,\alpha_i) \right] \qquad \text{(III.11)}$$

où $D(Y_i,\alpha_i)$ est la fonction définie par (III.5)

2° cas: la segmentation n'est pas réalisée préalablement

Soit $\underline{X}$ la suite des phonèmes à reconnaître qui peut s'ecrire comme auparavant: $\underline{X} = \{X_i\}_{i=1,n}$

aprés normalisation, on obtient une suite normalisée $\underline{Y}$:

$\underline{Y} = \{Y_i\}_{i=1,n}$

Si on désigne par $\underline{x}$ la suite de tous les vecteurs acoustiques constituant $\underline{X}$, alors on peut ecrire:

$\underline{x} = \{ x_l \}_{l=1,N}$; N: Nombre total des vecteurs acoustiques constituant $\underline{X}$

La segmentation correspondant à la séquence $\underline{x}$, peut être représentée par : $\underline{s} = \{ s(i) \}_{i=1,n}$ dans laquelle s(i) désigne l'indice du dernier vecteur acoustique contenu dans le i^{eme} segment phonètique X_i lequel, en utilisant cette notation, peut s'exprimer de la manière suivante :

$$X_i = [z_{s(i-1)+1}, \ldots\ldots\ldots, z_{s(i)}]$$

Avec la supposition que s(0)=0

X_i est un segment phonétique de longueur L(i) = s(i)-s(i-1) et bien sûr on a : $Y_i = X_i^T L(i)$.

Rappelons que ce qui vient d'être exposé n'est qu'une approche purement mathématique de la segmentation et que cela va servir dans ce qui va suivre.

En reconnaissance automatique de la parole, on détermine la segmentation du signal parole d'entrée et on reconnait la suite de phonèmes qui en correspond.

L'approche Bayesienne pour trouver la vraisemblance de la suite de phonèmes est donnée par:

$$\mathcal{L}_B(\underline{\alpha}) = \sum_{\underline{s}} [P(\underline{Y}(\underline{s})/\underline{\alpha})\ P(\underline{\alpha})] \qquad \text{(III.12)}$$

expréssion dans laquelle la notation $\underline{Y}(\underline{s})$ exprime tout simplement le fait que la suite $\underline{Y}$ depend de la segmentation. Comme cette approche est très coûteuse en temps de calcul, on lui substitue la méthode de détection par le maximum de vraisemblance lorsque certains paramètres du signal sont inconnus. Nous considérons que la segmentation est un paramètre inconnu.

Pour chaque suite hypothétique de phonèmes, on détermine la plus vraisemblable segmentation en maximisant la vraisemblance:

$$\mathcal{L}(\underline{\alpha}) = \max_{\underline{s}} \{Ln[P(\underline{Y}(\underline{s})/\underline{\alpha}).P(\underline{\alpha})]\} \qquad \text{(III-13)}$$

alors la suite de phonèmes reconnue est :

$$\hat{\underline{\alpha}} = \arg \max_{\underline{\alpha}} (\mathcal{L}(\underline{\alpha})) \qquad \text{(III-14)}$$

l'algorithme peut se résumer ainsi:

1- hypothétiser toutes les segmentations possibles:

$\underline{s}=\{s(i)\}$ $\quad i=1,n$

2- normaliser toutes les segmentations:

$X_i=[x_{s(i-1)+1},\ldots,x_{s(i)}]$

$L(i)=s(i)-s(i-1)$

$Y_i=X_i.T_{L(i)}$

3- trouver la meilleure suite de phonèmes et faire correspondre une vraisemblance pour chaque segmentation hypothétique:

$$\mathcal{L}(\alpha)=\max_{\underline{s}} \{Ln[P(\underline{Y}(\underline{s})/\underline{\alpha})].P(\underline{\alpha})\}$$

4- Parmis toutes les vraisemblances trouvées, choisir celle qui est maximale, laquelle vraisemblance correspond à la suite de phonèmes reconnue:

$$\hat{\underline{\alpha}} = \arg \max_{\underline{\alpha}} \mathcal{L}(\underline{\alpha})$$

Pour des raisons pratiques, (III-13) n'est pas utilisée alors on lui substitue une expression dans laquelle figure une somme de probabilités logarithmiques pondérées, ~~laquelle~~ cette somme exprime le score résultant de la comparaison des caractéristiques des phonèmes (les phonèmes à reconaître et les modèles phonétiques) dans ce cas le score pour la segmentation à maximiser à partir de l'ensemble des segmentations permises sera donné par :

$$J(\underline{Y}/\underline{\alpha}) = \sum_{i=1}^{n} \{Ln[P(Y_i/\alpha_i)].L(i) + Ln[P(\alpha_i) + C\} \qquad \text{(III-15)}$$

où n designe le nombre de segments.

Il est évident que $J(\underline{Y}/\underline{\alpha})$ remplace $Ln[P(\underline{Y}(\underline{s})/\underline{\alpha}).P(\underline{\alpha})]$ dans (III-13) alors celle-ci s'écrira :

$$\mathcal{L}(\alpha) = \max_{\underline{s}} \{ \sum \{Ln[P(Y_i/\alpha_i)].L(i) + Ln[P(\alpha_i)] + C\} \} \qquad \text{(III-16)}$$

par conséquent la suite de phonémes reconnue s'écrira :

$$\hat{\underline{\alpha}} = \arg \max_{\underline{\alpha}} \{ \max_{\underline{s}} \{ \sum \{ Ln[P(Y_i/\alpha_i)].L(i) + Ln[P(\alpha_i)] + C \} \} \} \qquad \text{(III-17)}$$

commentaire :
Le paramètre C désigne le coût par segment. Il est utilisé pour ajuster le taux d'insertion des phonèmes dans le système.ce

coût correspond à la probabilité logarithmique du taux de phonèmes (nombre de phonèmes/emission)
D'autre part la durée L(i)(ou longueur) du $i^{ème}$ phonème (à reconnaître) est introduite uniquement pour s'assurer qu'un segment phonétique X_i de longueur L(i) contribue au score dans la même proportion que sa longueur .

Autre Algorithme:

En plus de l'algorithme déjà decrit ,il éxiste une autre solution qui utilise un algorithme de programmation dynamique. Plus préscisément , à chaque temps t , on calcule le score pour la meilleure suite de phonèmes reconnue qui prend fin à l'instant t :

$$(J^*)_t = \max_{\tau,\alpha}\{(J^*)_\tau + Ln[P(Y(\tau,t)/\alpha)].(t-\tau) + Ln[P(\alpha)] + C\} \quad (III-18)$$

$Y(\tau,t) = [X_{\tau+1}, \ldots\ldots, X_t]\ T_{t-\tau}$ et $(J^*)_t$ est le score de la meilleure suite de phonèmes reconnue . Le dernier phonème α^* correspondant à $(J^*)_t$, ainsi que l'instant τ^* où prend fin le phonème qui le précède sont notés pour pouvoir déterminer le ensuite la séquence de phonèmes reconnue dans sa totalité . La solution à lafin de la phrase, à l'instant tf, est donnée par $(J^*)_{tf}$. Cet algorithme est très coûteux en temps de calcul que ne l'est l'algorithme précédent. Toutefois la complexite de la recherche est proportionnelle au nombre de modèles phonétiques contenus dans le dictionnaire .

III-5/ PHASE D'APPRENTISSAGE

Dans ce paragraphe, nous abordons l'entrainement des modèles phonétiques . Nous commençons par décrire une procédure pour l'estimation statistique des modèles phonétiques à partir

d'une séquence de parole préalablement segmentée en phonèmes.Ensuite, nous décrirons un algorithme qui va réaliser la segmentation automatique laquelle segmentation va servir de ré-estimation des modèles phonétiques résultant de l'estimation tout en ayant leur transcriptions phonétiques, puis à partir de ces deux étapes on établira un algorithme itératif qui réalisera la segmentation automatique à partir de la parole continue.

III-5-1/ ESTIMATION DES PARAMETRES

A partir d'une séquence de parole segmentée en phonèmes il est possible d'estimer les modéles phonètiques à partir des statistiques des segments phonétiques normalisés. Les probabilités à priori des phonèmes sont estimées à partir de leur fréquences relatives dans la séquence de la parole d'entrainement.

Si on suppose que pour un segment normalisé tous les vecteurs acoustiques qui le constituent sont indépendants, alors un modèle phonetiques est donné par les m vecteurs acoustiques modèles indépendants.

Par conséquent, il est simplement nécessaire d'estimer le modèle gaussien pour chaque vecteur acoustique et pour chaque modèle phonétique.

Cette estimation porte sur les statistiques des vecteurs acoustiques constituant les segments phonétiques normalisés donnés.

La densité du $J^{ième}$ vecteur acoustique du phonème α est désignée par $N(\mu_j, C_j)$ [4]. Cette densité est déterminée par les méthodes d'estimation classiques pour la moyenne et la covariance (voir annexe) ; ceci à partir de l'ensemble des vecteurs acoustiques qui correspondent au $j^{ième}$ vecteur du phonème .

III-5-2/ SEGMENTATION AUTOMATIQUE :

A partir de la collection des modèles gaussiens et leurs transcriptions phonétiques on peut déterminer la segmentation d'une séquence de parole de transcription phonétique $\underline{\alpha}$ connue. Cette phase qui utilise un algorithme très simple permet de ré-estimer les modèles phonétiques obtenus à l'étape de l'éstimation, puis on établira un algorithme itératif qui permettra d'entraîner automatiquement les modèles phonétiques à partir de la parole continue .

L'algorithme est simplement une recherche parmi toutes les segmentations possibles et ensuite choisir celle qui donne la transcription α avec un maximum de probabilité .

Cette segmentation s'ecrit :

$$\underline{s} = \arg \max_{\underline{s}} \mathcal{L}(\underline{s}) \qquad \text{(III-19)}$$

où la vraisemblance de segmentation est donné par :

$$\mathcal{L}(\underline{s}) = \mathrm{Ln}[P(\underline{Y}(\underline{s})/\underline{\alpha}).P(\underline{\alpha})] \qquad \text{(III-20)}$$

$$= \sum_{i=1}^{n} \{ \mathrm{Ln}[P(Y_i/\alpha_i)].L(i) + \mathrm{Ln}[P(\alpha_i)] + C \} \qquad \text{(III-21)}$$

alors, la segmentation est donnée par :

$$\hat{\underline{s}} = \arg \max_{\underline{s}} \{ \sum_{i=1}^{n} \{\mathrm{Ln}[P(Y_i/\alpha_i)].L(i) + \mathrm{Ln}[P(\alpha_i)]+C \} \} \qquad \text{(III-22)}$$

Comme en phase de reconnaissance, le score $J(\underline{Y}/\underline{\alpha})$ (défini en III-15) remplace l'expression Ln $[P(\underline{Y}(\underline{s})/\underline{\alpha})\ P(\underline{\alpha})]$, ceci dans le but d'introduire le facteur de coût et la durée (ou longueur) de chaque phonème dans la séquence de parole .
L'introduction de ces deux facteurs va améliorer la probabilité maximum correspondant à la segmentation $\underline{\hat{s}}$.

Algorithme d'entrainement Itératif:

En utilisant les deux étapes déjà decrites , l'éstimation des paramètres et la segmentation automatique, nous pouvons définir un algorithme itératif pour entrainer automatiquement les modèles segmentaux à partir de la parole continue.
si on se donne:
* La transcription phonétique de la séquence d'entrainement
* Les modèles gaussiens initiaux pour tous les phonèmes:
$\{ P_0(Y/\alpha) \}$
* t=0

On itère:

1) determiner la segmentation à probabilité maximale $\underline{\hat{s}_t}$ correspondante à la séquence d'entrainement pour la transcription donnée et les densités de probabilité courantes $\{ P_t(Y/\alpha) \}$

2) determiner l'estimation par le maximum de vraisemblance pour les densités $\{ P_{t+1}(Y/\alpha) \}$ en utilisant $\underline{\hat{s}_t}$.

3) t <= t+1 et aller à l'étape 1)

Il apparait qu'à chaque étape de l'algorithme la vraisemblance de la segmentation $\mathcal{L}(\underline{\hat{s}_t})$ croit pour la séquence de phonèmes (d'entrainement) donnée.
Dans l'étape 1), la nouvelle segmentation est la plus vraisemblable pour les modèles phonétiques courants.

qui devrait être, au moins, aussi probable que la segmentation qui précède. Dans l'étape 2), les nouvelles densités font croitres la probabilité de la segmentation courante ceci par définition de l'éstimation par le maximum de vraisemblance.

D'autre part, le nombre de segmentations est fini car la longueur de la séquence d'entrainement est aussi finie, donc $\mathcal{L}(\hat{s}_t)$ est borné par la vraisemblance de la meilleure segmentation. Par conséquent, la vraisemblance de la segmentation $\mathcal{L}(\hat{s}_t)$ converge vers un optimum local. La séquence des segmentations converge vers un optimum local quand la vraisemblance de la segmentation de la séquence l'est du fait que le nombre de segmentations possibles est fini et la recherche de la vraisemblance maximale dans l'étape 1) est imposée.

Ce qui conduit vers une collection optimale de modèles phonétiques.

ANNEXES

A. NOTIONS DE LA THEORIE DES PROBABILITES

A - Rappel des définitions et des théoremes élémentaires:

a - Soient tous les évènements mutuellement exlusifs d'un espace d'echantillonage alors:

$$\sum_i [P(A_i)] = 1$$

b - Probabilité conditionnelle: on considére deux évènements A et B. La probabilité conditionnelle P(B/A) est la porbabilité de réalisation de B sachant que A s'est réalisé.

c - probabilité composée: C'est la probabilité de réalisation simultanée des évènements A et B:

$$P(A,B) = P(A).P(A/B) = P(B).P(B/A)$$

Pour plusieurs evenements simultanées:

$$P(A_1,B_2,....An) = P(A_1).P(A_2/A_1)...P(A_n/A_1,....,A_{n-1}$$

d - Evenements indépendants: deux evenements sont dits indépendants si:

$$P(A/B) = P(A)$$

Théorème: Soient deux evenements indépendants A et B alors:

$$P(A,B) = P(A).P(B)$$

c - Théorème des probabilités totales: on considére les évenments Hi mutuellemnt exlusis dont l'union est l'espace fondamental ... si A est un evenement quelconque alors:

$$P(A) = \sum_i P(H_i).P(A/H_i)$$

d - Théorème de Bayes
- on prend les mêmes conditions que le theorème précedent:

$$P(Hi/A) = P(Hi).P(A/Hi)/P(A)$$

A - 2 Variable Aléatoire, Loi de distribution:

a - Une variable X est dite aléatoire si elle prend ses valeurs selon une probabilité P(X).

Une variable aléatoire X peut être discréte ou continue .

Il est commode d'introduire la fonction de probabilité ou fonction de répartition P(X=x).

Dans le cas d'une variable aléatoire discréte on a bien sûr:

$$\sum_k P(X=x_k)=1$$

(x_k: chacune des valeurs que peut prendre la variable aléatoire).

Pour une variable alétoire continue on préfére définir une fonction densité de probabilité P(x) telle que:

$$P(x) >= 0$$

$$\int_{+\omega}^{+\omega} P(x)dx = 1$$

b - Loi normal ou loi de gauss: $P(x) = \exp(-\frac{x-u}{\sigma.\sqrt{2}.\pi})^2 / .(\sigma\sqrt{2}.\pi)$

u: moyenne on espérance de X; σ^2: variance loi normal multidimensionnelle pour des variables indépendants:

$$P(x_1,.....,x_n) = \prod_{i=1}^{n} . \exp(-\frac{x_i-u_i}{\sigma_i\sqrt{2}.\pi})^2 / (\sqrt{2}.\pi\sigma_i)$$

B - STATISTIQUES

a - Estimation des moyennes:

Soient le vecteur aléatoire suivant: $X = (X_1, X_2, \ldots\ldots, X_K)$

où Xi est une variable aléatoire.

Soit N le nombre de valeurs que prend ce vecteur aleatoire:

$$x_1 = [\ (x_1)_1,\ (x_2)_1,\ (xs)_1,\ \ (x_i)_1 \ldots,\ (x_k)_1\ ]$$

$$x_2 = [\ (x_1)_2,\ (x_2)_2,\ (x_3)_2,\ \ldots(x_i)_2 \ldots,\ (x_k)_2\ ]$$

$$\vdots$$

$$x_N = [\ (x_1)_N,\ (x_2)_N,\ (x_3)_N,\ \ldots(x_i)_N \ldots,\ (x_k)_N\ ]$$

alors la moyenne correspondante à la variable aleatoire Xi:

$$\mu x_i = \frac{\Sigma[(x_i)_1}{N}$$

avec $(x_i)_1$ = valeur de la variable x_i correspondant à la 1^{eme} réalisation

on aura alors le vecteur moyenne:

$$\mu = (\mu_{xi},\ \mu_{x2},\ \mu_{X3}, \ldots\ldots,\ \mu_{xi}, \ldots\ldots \mu_{xk})$$

b - Estimation de la matrice covariance:

Soit X le vecteur aléatoire définie précedement, alors sa matrice covariance s'ecrira:

$$C_X = \begin{bmatrix} \sigma^2_{X1} & c_{X1X2} \cdots\cdots c_{X1Xk} \\ c_{X2X1} & \\ \vdots & \sigma^2_{X2} \\ \vdots & \ddots \\ c_{Xkx1} & \sigma^2_{xk} \end{bmatrix}$$

avec $C_{xix.j} = \sum_{l=1}^{N} [(x_i)_l - \mu_{xi}][(x_j)_l - \mu_{X.j}]/N$

et $C_{xixi} = \sigma^2_{Xi} = \sum_{l=1}^{N} [(x_i) - \mu_{Xi}]^2/N$

Où μ_{Xi}, $\mu_{X.j}$ désignent les moyennes corespondantes aux variables adjacentes X_i, X_j.

C - RESULTATS DE LA SIMULATION

a - Analyse:

La metthode d'analyse utilisée est la methode par prédiction linéaire. Le signal parole simulé renferme les cinq voyelles: a, o, i, e, u, chacune de ces voyelles est generée par une some de six sinisoides où sont exprimés l'amplitude et les deux formants de la voyelle.

La durée de la voyelle intervient dans les phases de l'echantillonnage et du fenêtrage.

Toutes ces opéraions sont effectuées dans les conditions suivantes:

- Fréquence d'echantillonnage: fe = 12.8 KHZ
- Nombre d'echantillons de la fenêtre d'analyse :

N= 256

- Ordre de prédiction linéaire : P = 12

Caractéristiques des voyelles générées:

pour le phonème a:

amplitude:80 mv

premier formant:750 Hz

deuxième formant: 1350 Hz

durée : 0.8 ms

Pour le phonème o:

amplitude :80 mv

premier formant: 375 Hz

deuxième formant: 750 Hz

durée: 0.8 ms

Pour le phonème i:

amplitude: 80 mv

premier formant: 250 Hz

deuxième formant: 2500Hz

durée: 0.8 ms

Pour le phonème e:

amplitude: 80 mv

premier formant: 400 Hz

deuxième formant: 2200 HZ

durée: 0.8 ms

Pour le phonème u:

amplitude: 80 mv

premier formant: 250 Hz

deuxième formant; 600 Hz

durée: 0.8 ms

b - Reconnaissance:

Dans cette phase, on a d'abord crée un dictionnaire qui contient la collection des cinq modèles phonétiques (voyelles) où chaque modèle est crée en estimant les moyennes et les covariances des différents vecteurs acoustiques correspondants aux différentes versions du phonème. Ces versions sont obtenues en perturbant légérement les caractéristiques des phonèmes; puis vient l'étape de ré-estimation dans laquelle on cherche à reconnaitre chaque phonème individuellement tout en perturbant légérement ses caractéristiques. nous avons repeté plusieurs

fois cette opération de reconnaissance jusqu'à ce que la reconnaissance reste parfaite malgré les légéres perturbations qu'on subi les caractérestiques à chaque repétition. Toutefois les perturbations ne devront pas dépasser une certaine limite pour rester toujours dans le cadre d'un système phonatoire humain. Une fois l'opération de ré-estimation s'est soldée par de bons résultats, on a pris comme optimaux les modèls estimés statistiquement(par les moyennes et les covariances).

Nous précisons que chaque modèle optimal est affectée d'une étiquette qui va servir comme moyen de décodage dans la reconnaissance simulée.

Concernant la reconnaissance simulée proporement dite, les mots à reconnaitre sont des suites de phonèmes (ici des voyelles).

Nous avons utilisé les étiquettes suivantes:

phonème	étiquette
a	"a"
o	"o"
i	"i"
e	" e"
u	"u"

nous avons effectué les tests suivants où chaque mot est représenté par une suite de cinq phonèmes(taux d'insertion)

TEST1

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'aoieu'

Donner les amplitudes des phonèmes successifs:

80 80 80 80 80

Donner les premiers formants des phonèmes successifs:

750 375 250 400 250

Donner les deuxièmes formants des phonèmes successifs:

1350 750 2500 2200 600

Donner les durées des phonèmes successifs:

0.8 0.8 0.8 0.8 0.8

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est: 'aoieu'

TEST2

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'ueioa'

Donner les amplitudes des phonèmes successifs:

80.5 80.9 80.7 80.6 81

Donner les premiers formants des phonèmes successifs:

255 405 255 380 755

Donner les deuxièmes formants des phonèmes successifs:

605 2205 2505 755 1355

Donner les dureés des phonèmes successifs:

0.81 0.82 0.83 0.84 0.85 ms

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'ueioa'

TEST3

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'ioaeu'

Donner les amplitudes des phonèmes successifs:

79.5 78.5 82 81 83

Donner les premiers formants des phonèmes successifs:

360 385 760 410 260

Donner les deuxièmes formants des phonèmes successifs:

2510 760 1360 2210 610

Donner les durées des phonèmes successifs:

0.75 0.78 0.74 0.82 0.82

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'ioaeu'

TEST4

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'euaoi'

Donner les amplitudes des phonèmes successifs:

75 76 77 78 80

Donner les premiers formants des phonèmes successifs:

390 240 760 365 240

Donner les deuxièmes formants des phonèmes successifs:

2190 590 1340 760 2490

Donner les dureés des phonèmes successifs:

0.75 0.76 0.77 0.78 0.79

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'euaoi'

TEST5

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'oieua'

Donner les amplitudes des phonèmes successifs:

75 76 78 77 81

Donner les premiers formants des phonèmes successifs:

370 245 395 245 745

Donner les deuxièmes formants des phonèmes successifs:

745 2495 2195 595 1345

Donner les dureés des phonèmes successifs:

0.75 0.78 0.76 0.74 0.73

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'oieua'

TEST6

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'ieuao'

Donner les amplitudes des phonèmes successifs:

75 76 80 82 83

Donner les premiers formants des phonèmes successifs:

253 405 255 755 380

Donner les deuxièmes formants des phonèmes successifs:

2505 2206 608 1356 755

Donner les durees des phonèmes successifs:

0.81 0.82 0.79 0.75 0.78

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'euiao'

TEST7

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'aiuoe'

Donner les amplitudes des phonèmes successifs:

80 80.4 79.6 78.8 77.8

Donner les premiers formants des phonèmes successifs:

755 254 257 380 408

Donner les deuxièmes formants des phonèmes successifs:

1357 2508 610 759 2208

Donner les dureés des phonèmes successifs:

0.78 0.79 0.75 0.76 0.81

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'aiuoe'

TEST8

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'eueue'

Donner les amplitudes des phonèmes successifs:

80.2 80.3 80.4 78.25 75.66

Donner les premiers formants des phonèmes successifs:

400 250 405 255 412

Donner les deuxièmes formants des phonèmes successifs:

2205 600 2206 610 2209

Donner les durées des phonèmes successifs:

0.78 0.79 0.75 0.81 0.80

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'eueue'

TEST9

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'auaua'

Donner les amplitudes des phonèmes successifs:

80.4 80.56 78.5 79.55 82.1

Donner les premiers formants des phonèmes successifs:

750 250 755 255 746

Donner les deuxièmes formants des phonèmes successifs:

1350 605 1348 609 1346

Donner les dureés des phonèmes successifs:

0.8 0.75 0.79 0.76 0.78

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'auaua'

TEST10

Donner le nombre de phonèmes du mot à reconnaitre:

5

Donner la suite des phonèmes du mot:

'iiiii'

Donner les amplitudes des phonèmes successifs:

70 71 72 73 74

Donner les premiers formants des phonèmes successifs:

230 220 200 205 206

Donner les deuxièmes formants des phonèmes successifs:

2500 2400 2350 2300 2250

Donner les durees des phonèmes successifs:

0.70 0.71 0.72 0.73 0.74

Donner le nombre de vecteurs acoustiques avant la normalisation:

40

Donner le nombre de vecteurs acoustiques aprés la normalisation:

10

Donner le nombre de coefficients prédicteurs:

12

Le mot reconnu est:'iiiee'

Interprétation des résultats des tests:

A travers les résultats de ces tests, on voit bien l'efficacité du caractère statistique de la méthode car en variant les caractéristiques des phonèmes dans une plage de ± 1,5% des valeurs de référence, on continue quand-même à reconnaître le mot. Ce fait confère à la méthode d'être utilisée dans un système de reconnaissance multilocuteur.

D'autre part, nous avons constaté que le temps de reconnaissance est très important et il serait d'autant plus important que le nombre de modèles phonétiques est grand ce qui nous permet bien de faire une suggestion en préconisant d'implanter la méthode dans des microprocésseurs puissants dans le but d'augmenter la vitesse de reconnaissance qui est l'un des plus grands objectifs en reconnaissance de la parole.

Pour finir, on peut ajouter que dans cette méthode la reconnaissance porte sur l'unité fondamentale indivisible de la parole qu'est le phonème, donc le taux d'erreurs de reconnaissance est trés faible ce qui n'est pas le cas avec les méthodes globales.