الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة الجامعات
Ministère aux Universites

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT ELECTRONIQUE

PROJET DE FIN D'ETUDES

SUJET

RECONNAISSANCE DE MOTS ISOLES EN MODE MULTILOCUTEUR

Proposé par : Mr N.Beniddir

Etudié par : B. Khène, M. Tounsi

Dirigé par : Mr N.Beniddir

PROMOTION JUIN 92

Nous tenons à remercier vivement Mr Beniddir et Mr Derras pour leurs conseils et leur aide .

Que Mrs Kertous , Benali , Abismaïl ainsi que tous ceux qui ont contribué de prés ou de loin à la finition de ce document ,trouvent ici l'éxpréssion et l'assurance de nos remerciements .

Dedicaces

A la mémoire de mon père ;
à ma mère , mes frères et toute ma famille ;
à mon oncle ,sa femme et tous ses enfants ;
à tous(tes) mes amis(es) et mes proches ;

"" Ma teɣlim lliɣ tellam mara ɣlliɣ ""

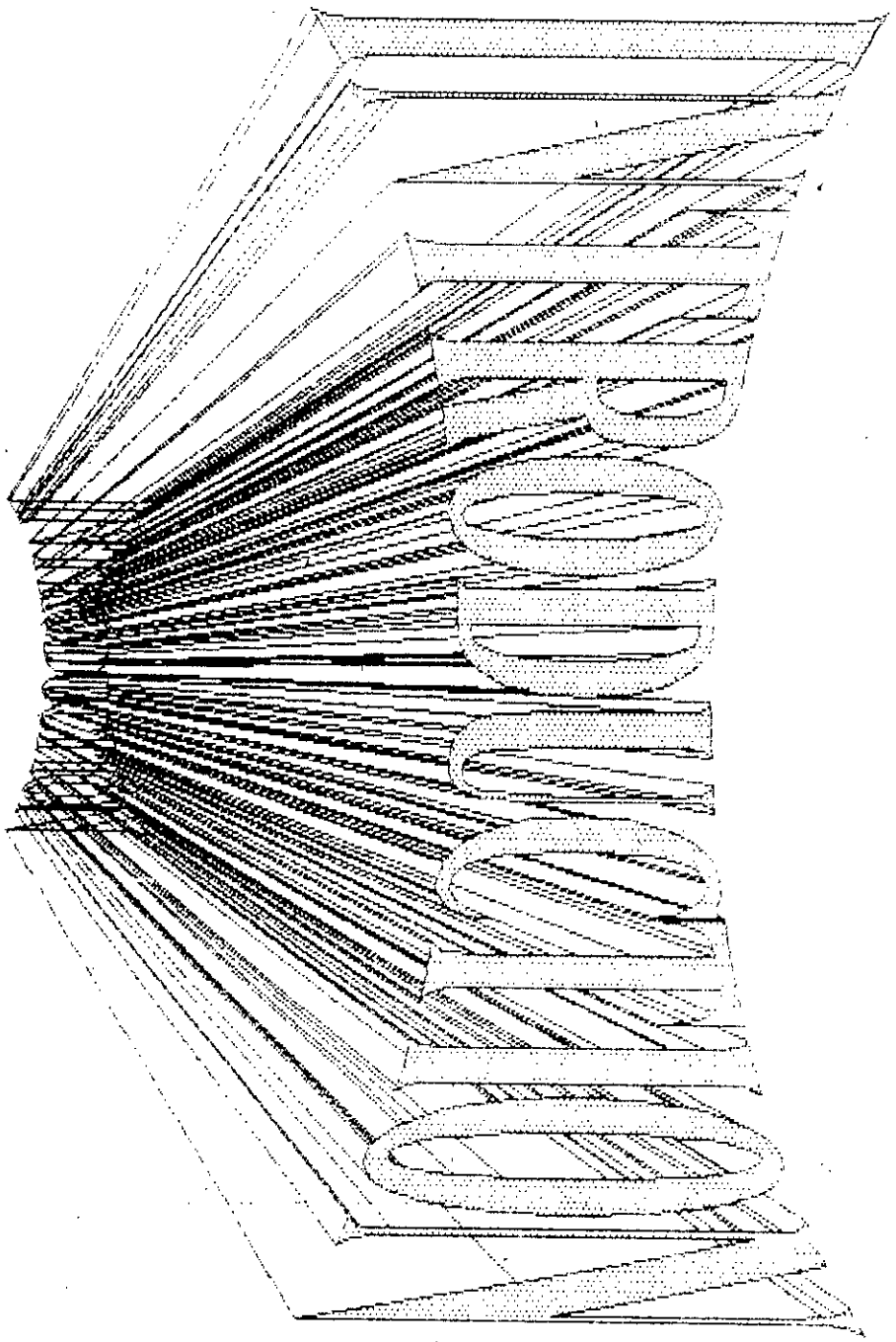
Med TOUNSI .

A ma mère qui a toujours veillé sur moi;
à mon père qui m'a toujours soutenu ;
à mes frères , mes soeurs et toute ma famille ;
à tous mes amis et mes proches ;

Brahim KHENE .

Nous dédions ce mémoire .

SOMMAIRE



A)- GENERALITES :

"LA PAROLE EST LA FACULTE DE COMMUNIQUER LA PENSEE PAR UN SYSTEME DE SONS ARTICULES " [1] . C'est le moyen de communication privilégié chez les humains qui sont les seuls êtres vivants à utiliser un tel système structuré . Cependant , nous assistons ces dernières années à une énorme profusion de machines multiples sans lesquelles , tout éxploit technique ou scientifique est désormais impossible à l'homme moderne .

Le principal problème qui reste un inconvénient pour concrétiser les larges services de ces machines est celui de la " la communication Homme-Machine " qui s'est toujours faite à travers des boites à commande ,des claviers , des écrans de visualisation ....etç .

C'est pourquoi , plusieurs chercheurs et laboratoires du monde ne cessent ,actuellement ,de développer des techniques nouvelles dans l'éspoir de rendre ces machines plus "intélligentes".

Parmi ces techniques , nous trouvons Le traitement automatique de la parole qui reste un objectif futiriste englobant :

---La Synthèse Automatique de la parole aboutissant à des machines qui "parlent" ;

---La Compréhension et la Reconnaissance de la parole aboutissant à des machines qui "entendent" et "comprennent".

Les premières tentatives en reconnaissance de la parole remontent à 1939 avec le vocodeur de Mr.Duilley aux USA . Depuis , l'intérêt économique de certaines entreprises commerciales s'est joint à l'intérêt purement scientifique surtout avec l'avènement des circuits intégrés et le développement considérable de la micro-éléctronique et de l'informatique . Cela a mené à l'un des projets les plus ambitieux du 21ème siècle : Le projet A.R.P.A lancé par le ministère de la défence Américain pour la réalisation d'un système

de compréhension de la parole de 15 millions $ avec un vocabulaire de 1000 mots .

Les difficultés à résoudre pour la concrétisation de ces projets sont considérables à cause de la compléxité du signal acoustique de la parole , de sa richesse en informations phonétiques, de sa forte variabilité ,de son aspect aléatoire et coarticulatoire .

La compréhension de la parole est actuellement abordée suivant les méthodes suivantes :

--Les méthodes globales :

Les mots y sont considérés comme des entités entières et codés commes telles et un silence suffisant doit nécéssairement séparer deux mots consécutifs . Le système compare le mot à un ensemble d'empreintes vocales préenregistrées pour reconnaitre la plus proche . Le principal désavantage de ces méthodes est que le nombre de mots à reconnaitre reste encore trés limité .

--Les méthodes analytiques :

Ces méthodes sont plus difficiles mais plus intéréssantes que les précédentes car elles sont utilisées pour reconnaitre de la parole continue . Les difficultés viènnent du fait que la parole est divisée en segments et chaque segment doit représenter l'emplacement probable de l'entité phonétique choisie .Une analyse performante est ensuite appliquée pour éxtraire les parametres pertinents de chaque segment dans le but de les comparer aux entités du dictionnaire établi . Les résultats de cette analyse ou plus éxactemment lesprobabilités correspondantes à chaque entité phonétique sont ensuites affinées en utilisant des informations syntaxiques ,sémantiques et contextuelles . Ces métodes sont trés prométteuses surtout avec l'arrivée des systèmes éxperts .

Par ailleurs ,d'autres méthodes telles que les chaines de

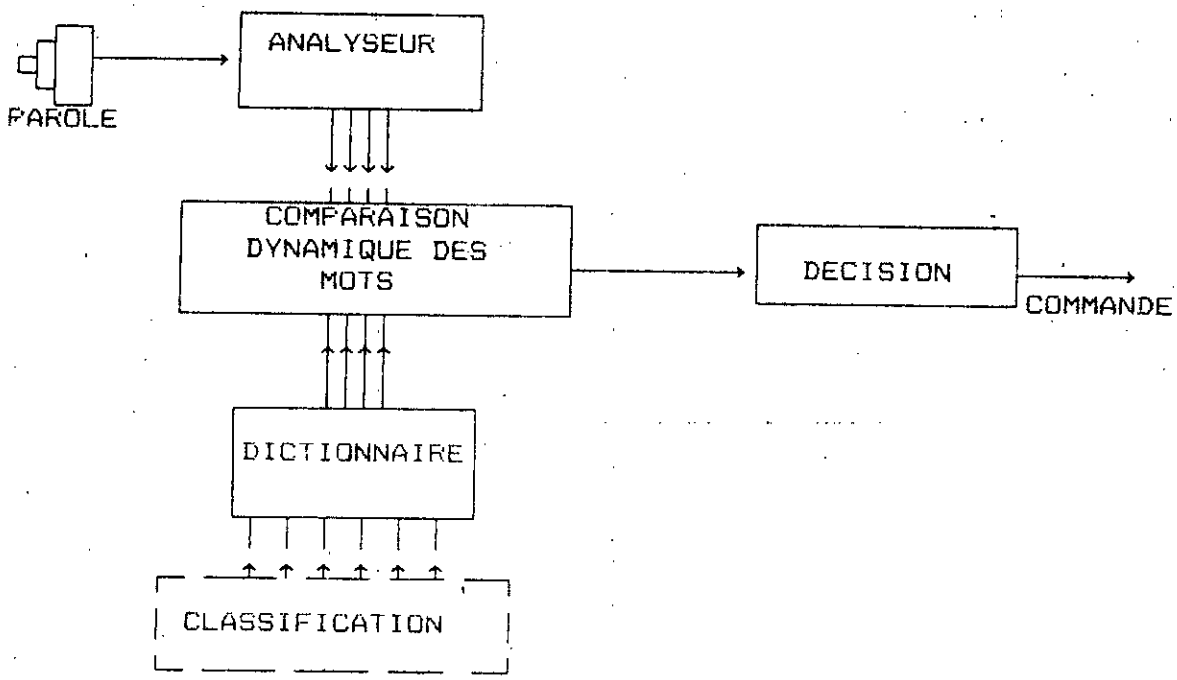
Markov et l'intélligence artificielle commencent à s'imposer de nos jours avec de grandes performances .

Notons que les systèmes disponibles en reconnaissance de la parole sont soit monolocuteurs ie que tout utilisateur étranger au système serait non reconnu ,soit multilocuteurs ie une tentative de généralisation des premiers , qui prennent en considération la variabilité des mots d'un locuteur à un autre .

B)- PRESENTATION DU TRAVAIL :

L'objectif de notre travail est l'étude d'un système de reconnaissance de la parole par les méthodes globales en mode multilocuteurs pour une éventuelle implantation sur le micro-processeur TMS 320 .

SYNOPTIQUE DU SYSTEME DE RECONNAISSANCE DE LA PAROLE :



Notre mémoire est élaboré de la façon suivante :

--Le chapitre I décrit les méthodes d'analyse que nous appliquons aux voyelles orales de la langue Française ;
--Le chapitre II éxplique la comparaison dynamique des mots avec ses différents algorithmes ;
--Dans le chapitre III , nous étalons l'apprentissage et les méthodes de reconnaissances que nous avons retenus pour le système ;
--Suit ensuite le chapitre IV où est décrite l'étape de décision sur la reconnaissance des mots ;
--Dans le chapitre V , nous faisons la simmulation des voyelles pour nous servir de fichiers pour les tests ;
--Le chapitre VI regroupe les résultats de la classification ie les les différentes configurations des échantillons de nos voyelles qui vont constituer le dictionnaire ,ainsi que plusieurs tests que nous avons traités en mode monolocuteurs et d'autres pour finaliser notre reconnaissance en mode multilocuteurs ;
--Enfin , le dernier chapitre (VII) donne nos conclusions générales .

N.B : Les organigrammes sont donnés en annexe .

CHAP I

ANALYSE DE LA PAROLE

II-A / INTRODUCTION AU SIGNAL VOCAL :

Le contenu d'un message vocal ,au sens strict , est simplement son intélligibilité ; au sens large , ce message est trés complexe car contenant une quantité importante d'informations imbriquées entre elles , donc d'éxtraction trés difficile . Ce message est aussi caractérisé par une trés grande redondance et ce , pour résister aux perturbations du milieu ambiant .

La parole est aussi un phénomène sonore produit par le système phonatoire (*fig-I-1*) . L'énergie provient de l'air envoyé par les poumons (muscles thoraciques et abdominaux) et rarement de l'air aspiré par la bouche .

II-A-1°/ MECANISME DE PHONATION :

A-1-a / Le conduit vocal :

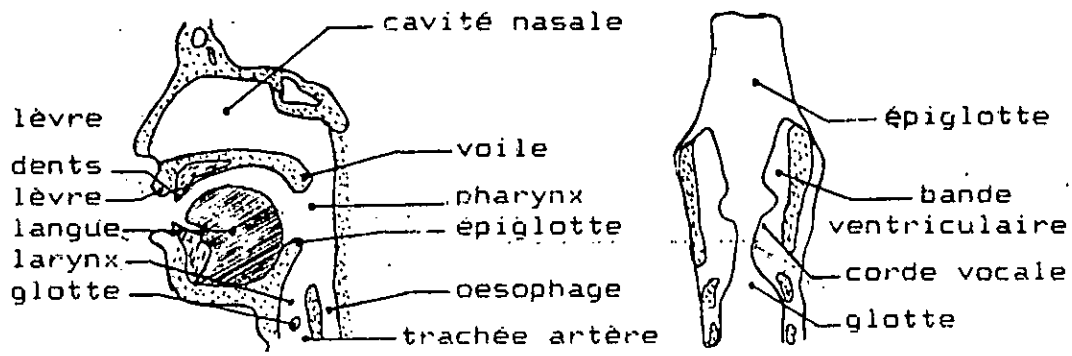


fig-I-1) a) appareil phonatoire b)section du larynx

Le signal vocal est le résultat de l'action volontaire et coordonnée des appareils respiratoires et masticatoires . Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par des sensations cénésthésiques .

Le conduit vocal est une suite de cavités : pharyngienne ,nasale

et buccale qui servent de résonnateurs car selon leurs formes , nous obtenons des résonnances à des fréquences variabes appelées FORMANTS . La cavité nasale est soit inutilisée , soit mise en parallèle sur la cavité buccale . La forme de ces cavités est affectée par des articulateurs tes que la langue ,les lèvres ,les machoires ...etç . La combinaison de ces divers élements permet plusieurs modes de production sonore et d'émission de sons variés .

A-1-b) l'éxcitation du système phonatoire :

Elle peut être causée par :

-une brusque variation de préssion ;

-la mise en vibration des cordes vocales ;

-ou par un flot d'air crée en certains points de ressèrement du coduit vocal .

II-A-2 / SONS VOISES et SONS NON-VOISES :

Quand le conduit vocal est excité par une suite d'impulsions périodiques liées aux vibrations des cordes vocales ,l'ouvèrture brusque de la glotte libère la préssion accumulée en aval et nous obtenons des *sons voisés* . Si par contre les cordes vocales s'écartent du chemin de l'air envoyé par les poumons ,nous obtenons des *sons non-voisés* .

II-A-3 / PITCH et FORMANTS :

L'intensité du son émis est liée à la préssion de l'air en amont du larynx , sa hauteur est fixée par la fréquence de vibration des cordes vocales , appelées fréquence du fondamental ou *Pitch* (Fo) Cette fréquence peut varier :

-- de 80 à 200 Hz pour une voix d'homme ;

-- de 150 à 450 Hz pour une voix de femme ;

-- de 200 à 600 Hz pour une voix d'enfant .

Sur e spectre d'un son voisé , nous pouvons observer des raies qui correspondent aux harmoniques du pitch . L'enveloppe de ces raies , présente des maximums aux fréquences propres Fi du conduit vocal (i = 1 , 2 .....) ; ces fréquences sont *Formants* .

II-B / ANALYSE ACOUSTIQUE :

II-B-1°) INTRODUCTION :

L'analyse acoustique est la partie essentièlle du traitement du signal vocal en vue de réaliser un système de reconnaissanse de parole.En éffet,cette analyse nous permet d'extraire les parametres et les coéfficients pertinents qui modélisent correctement les caracterisques du signal vocal .

Initialement,les chercheurs considéraient le signal de parole, pour le traiter,comme tout autre signal.Puis,ils ont développé des techniques trés fines spécialement conçues pour le signal vocal,parmi lesquelles nous citons :

- le prétraitement de la parole;
- l'analyse cepstrale ;
- l'analyse prédictive (LPC);
- analyse par banc de filtres .

II-B-2°) ANALYSES CLASSIQUES :

B-2°-A) PRETRAITEMENT DU SIGNAL VOCAL : [2]

Avant d'entamer l'analyse du signal vocal par les techniques LPC et cepstrales,nous lui éffectuons un prétraitement pour rendre plus éxploitable son contenu .

2°-A-1) Echantillonnage :

Puisque nous adoptons dans notre étude des méthodes numériques ,

le traitement,que nous allons faire,ne s'éffectuera pas directement sur les signaux analogiques à temps continu ; ces signaux doivent etre échantillonnés ,codés puis rangés dans une mémoire pour etre traités par la suite . Notons bien,que par traitement du signal de parole,nous entendons dans notre thèse,le traitement de l'information qui y est contenue et l'éxtraction des paramètres éssentiels . Notre objectif étant sa reconnaissance et beaucoup moins sa synthese .

D'aprés le théorème de **SHANON** [2],un signal analogique ayant une largeur de bande finie,limitée à Fm , ne peut etre reconstitué fidèlement à partir de ses échantillons (n Te)que si ceux-çi ont été prélevés avec une période Te telle que :

$$(Te \leq 1/(2\,Fm)) \quad ou \quad (Fe \geq 2\,Fm) \qquad (I-1)$$

Dans notre cas,le signal de parole est limité à Fm =6 Khz tout en conservant ses caractéristiques ;par suite,le théorème de **SHANON** nous permet de choisir :

$$Fe = 1/Te = 12.8\ Khz \geq 2\,Fm \qquad (I-2)$$

où : Fe : fréquence d'échantillonnage ;

Fm : fréquence maximale du signal de parole .

2°-A-2) **Préaccentuation :**

Le signal sonore diffuse de la bouche vers l'éxtérieur ;le son étant la transmission d'une onde dans un milieu mécanique (l'air) défini par son impédance mécanique .En débouchant des levres,l'onde (son) doit attaquer un milieu mécanique plus important que celui présent dans le conduit vocal .Il ya donc désadaptation des impédances mécaniques au niveau des lèvres.Par suite , le rayonnement du son à l'éxtérieur s'accompagne d'une baisse d'énergie et d'une distorsion assimilée à une désaccentuation de 6dB/Octave sur tout le spectre [2] .

Rétablir le signal tel qu"il était,est un fait trés important surtout que les chercheurs désirent par transformation inverse estimer l'evolution de la forme du conduit vocal,les lieux d'articulation et en déduire ainsi la parole prononcée.

Soient S(n) les échantillons du signal et $S_a(n)$ ceux du signal préaccentué.La préaccentuation de 6dB/Oct que nous faisons , n'est autre n'est qu'une dérivation numérique [2] :

$$S_a(n+1) = S(n+1) - S(n) \qquad (I-3)$$

2°-A-3) Fenêtrage :

Les variables temps/fréquence sont conjuguées,liées par une relation d'incertitude du type : $\Delta T\ \Delta F = 1/\pi$ où ΔT est la largeur de la fenetre d'analyse du signal et ΔF le plus petit détail que nous puissions éspérer observer dans son spectre .Autrément dit,l'analyse temporelle d'un sinal sinusoidal pur sur une fenêtre temporelle fournit un spectre de largeur ΔF au lieu d'une seule raie.Le fenetrage que nous devons éffectuer sur le signal de parole parcequ'il limite l'amplitude des rebonds fréquentiéls dus à la limitation temporelle du signal,provoque deux inconvénients :

---les raies s'élargissent d'une part ;

---d'autre part apparaissent,autour des raies centrales,des rebons dont les niveaux maximaux peuvent etre important ,donc pratiquement une résonnance (formant) peut influer trés loin du lieu où elle se trouve et perturber certaines régions peu énergitiques .

En somme , la forme de la fenetre influence l'allure générale de la réponse spectrale.

Plusieurs fenêtres sont envisageables :

----------(Rectangulaires , triangulaires , de Hanning).

La plus appropriée reste la fenetre de **HAMMING** (cas particulier de la fenetre de Hanning)quoique donnant une lareur du pic central plus grande que dans le cas de la fenetre rectangulaire , elle ne

provoque que l'apparition du lobe central (99.96°/o de l'énergie du signal) et des raies secondaires peu énergitiques (40 dB au dessous du lobe principal) .

Cette fenêtre est définie par son expression :

$$W_H(n) = \begin{cases} 0.54-0.46\cos(2\pi(n-1)/N) & \text{pour } 0 \le n \le N-1 \\ 0 & \text{sinon} \end{cases} \qquad (I-4)$$

Par ailleurs , notons que réduire la durée de la fenêtre temprelle du signal à analyser,revient à accepter une incertitude sur la position des formants Mais en pratique cette dégradation est acceptée pour pouvoir discerner même approximativement les évolutions rapides du signal .

Dans notre cas,avec Fe = 12.8 Khz nous fixons la largeur de la fenêtre de **HAMMING** à 20ms chacune sur toute la durée d'observations . Cette durée est choisie de façon à assurer la stationnarité de notre signal pour que ses parametres restent inchangés.

B-2°-B)ANALYSE CEPSTRALE : [5]

2°-B-1)introduction :

Nous savons déja que le signal vocal est tres complexé vû sa richesse considérable en informations. Son spectre contient les informations phonétiques tandis que les informations prosodiques sont contenues dans le pitch (fréquence du fondamental).

Lorsque,comme dans le cas de nos voyelles,les sons sont voisés,la vibration des cordes vocales (ou source d'éxcitation) fournit un signal pulsé que le conduit vocal module dans sa totalité . Les cordes vocales et le conduit vocal sont alors combinés par une convolution comme deux filtres. Une déconvolution s'avère nécéssaire pour retrouver les informations phonétiques contenues dans les deux spectres.Il nous faut donc séparer le

spectre du signal de celui de la source d'excitation .

L'opération de "déconvolution" obtenue par filtrage inverse est tres délicate . Cependant,il éxiste un moyen simple pour déconvoluer les deux spectres [2] :

l'éxamen du spèctre d'un signal voisé montre que les des ondulations provenant des cordes vocales,modulent la forme du spectre du son prononçé . L'idée est de séparer les deux composantes , superposées par un produit de convolution naturel,en une somme des deux composantes dans un domaine particulier . Ceci est obtenu par des traitements homomorphiques :

Soit h(n) le signal issu de la source d'éxcitation et x(n) la fonction de transfert du conduit vocal indépendante de h(n).L'éxpression du signal sonore y(n),obtenu par convolution de h(n) et x(n) échantillonnés,s'écrit :

$$y(n) = h(n) * x(n) = \sum_{k=-\infty}^{k=+\infty} h(n-k).x(k) \qquad (I-5)$$

où * represente le produit de convolution.

Soit D l'opération de déconvolution qui vérifie les relations suivantes :

$$D(n) = D(\, h(n) * x(n) \,) = D\,(h(n)) + D(x(n)) \qquad (I-6)$$

que nous noterons :

$$y'(n) = h'(n) + x'(n) \qquad (I-7)$$

Nous savons que mathématiquement,le produit de convolution (I-5) se transforme avec les transformées en "Z" en une ultiplication :

$$Y(Z) = H(Z) \,.\, X(Z) \qquad (I-8)$$

2°-B-2) Déconvolution homomorphiques :

Le traitement homomorphique suggère simplement d'utiliser les

propriétés du logarithme sur la la transformée en "Z" pour obtenir la somme désirée :

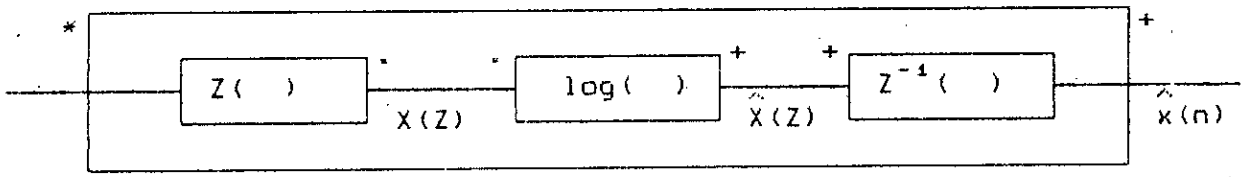


fig. I. 2. Déconvolution homomorphique.

Nous trouvons alors :

$$\begin{aligned} Y'(Z) &= \log\left[\, Y(Z) \,\right] = \log\left[\, H(z).X(Z) \,\right] \\ &= \log\left[\, H(z) \,\right] + \log\left[\, X(Z) \,\right] \\ &= H'(Z) + X'(Z) \end{aligned} \qquad (I-9)$$

la transformation ci-dessus n'est pas applicable en regle générale car le logarithme d'un nombre complexe a une partie immaginaire indéfinie:

$$Y'(Z) = \log\left[\, Y(Z) \,\right] = \log\,|Y(Z)| + j\,\arg\left[\, Y(Z) \,\right] \qquad (I-10)$$

la transformée de Fourier inverse du logarithme du spectre de notre signal nous donne son cepstre complexe :

$$y'(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} Y'(e^{jv})\, e^{jvn}\, dv \qquad (I-11)$$

dans notre cas,seule la partie réelle est utilisée ,donc le problème d'unicité ne se pose pas dans nos calculs ;le cepstre vaut alors :

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log\left|\, Y(e^{jv}) \,\right|\, e^{jvn}\, dv \qquad (I-12)$$

Les expressions (IV-10) et (IV-12) nous donnent la relation :

$$c(n) = \frac{y'(n) + y'(n)}{2} \qquad (I-13)$$

combinée à l'expression (II-11) :

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \log\left[Y(e^{jv}) \right] \cos(vn)\, dv \qquad (I-14)$$

Cette dernière relation nous montre que le cepstre peut etre calculé par la transformée en cosinus du logarithme du spectre complexe du signal considédé.

2°-B-3) Quelques propriétés des coéfficients cepstraux :

L'examen des relations (I-11,12,14) montre que les coéfficients cepstraux sont normalisés en énergie et sont insensibles à des modifications en amplitude du signal . Ces coéfficients décroissent en 1/n où n est leur rang [2] . Par suite peu de coefficients sont utilisés en pratique ,et beaucoup d'ouvrages fixent leur nombre à 12.

2°-B-4) Echelle MEL :

Des études physiologiques et perceptives de l'oreille ont montré qu'elle est sensible à une échelle presque logarithmique en fréquence . Ceci veut dire implicitement que les informations phonétiques sont réparties dans des zônes croissantesde façon exponentielle en fonction de la hauteur de la fréquence.

Une échelle quasilogarithmique du spectre [2] appelée échelle MEL, a montré son éfficacité pour le calcul cepstral mieux que l'échelle linéaire .

L'échelle MEL est linéaire sur le premier Khz et logarithmique au dessus . Sur toute la longeur de cette échelle,sont répartis des filtres triangulaires de largeurs uniformes sur le premier Khz et

variable au dessus. Ces filtres sont décalés ,les uns des autres,de la moitié de leurs largeurs.leurs débuts sont obtenus,au delà du 1^{ier}khz,en appliquant une raison géométrique supérieure à l'unité car leurs largeurs sont croissantes de façon exponentielle.

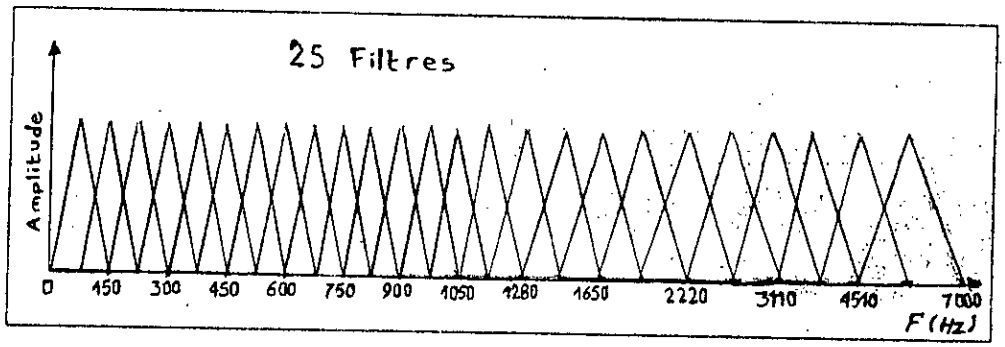


fig I -3- Echelle MEL

Avec la série des filtres,l'éxpréssion du cepstre est donnée [2] commesuit:

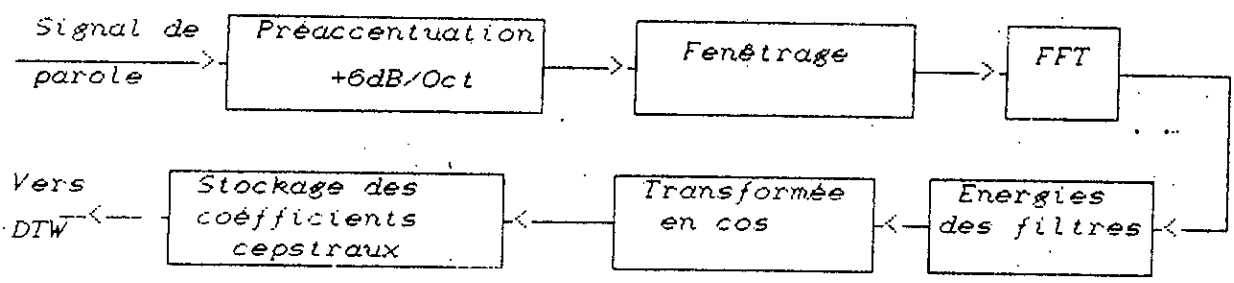
$$c(n)=\frac{1}{Nf}\left\{\sum_{k=1}^{Nf}\log\left[E(k)\right]\cos\left[n(k-1)/Nf\right]\right\} \qquad (I-15)$$

où: Nf est le nombre de filtres utilisés;

n est le rang du coéfficient cepstral;

k est le numéro du filtre d'énergie E(k).

2°-B-5) Etapes d'analyse :



La FFT sert pour le calcul du spectre du signal de parole.

2°-B-7) Organigramme : voir l'annexe .

2°-C) ANALYSE PAR PREDICTION LINEAIRE (LPC):

2°-C-1) Introduction :

La prédiction linéaire est l'une des techniques les plus appropriées au traitement de la parole.Elle nous permet d'estimer certains parametres comme le pitch,les formants et avec une grande précision et rapidité dans les calculs.Cette prédiction peut être :

progréssive ou rétrograde.

La prédiction progréssive (ou prédiction avant) est celle qui nous permet de prédire une valeur future d'un signal échantillonné x(n) à partir :

des p échantillons précédents x(n-i) , i = 1..p .

la prédiction rétrograde (ou prédiction avant) est celle qui , au lieu de prédire,vérifie une valeur passée x(n-p-1) à partir des mêmes échantillons.

2°-C-2) Modélisation du signal vocal :

L'étude du mécanisme de la phonation (3) a montré que le signal vocal est produit par un système autorégréssif (AR) qui serait soumis à une éxcitation idéalisée u(n) sous forme d'un bruit blanc pour les sons non voisés;et pour les sons voisés d'un train d'impulsions periodiques d'amplitude unité . Cette éxcitation est idéalisée car la forme réeile de l'impulsion glottique et la radiance des lèvres sont comprises dans l'expression de la transmittance de notre système AR pour les sons voisés (puisque les 6 voyelles orales que nous étudions le sont).

L'excitation u(n) est donnée par :

$$u(n) = \sum_{k} \delta(n-kP) \qquad (I-16)$$

où P est la période du pitch éxprimée en nombre de périodes

d'échantillonnage.

Dans ce modèle de production, un échantillon quelquonque x(n) d'un signal est une combinaison linéaire des p échantillons qui le précèdent en ajoutant le terme u(n) d'éxcitation pour avoir la récurrence linéaire suivante :

$$x(n) + \sum_{i=1}^{p} a(i)\, x(n-i) = \sigma\, u(n) \qquad \text{(I-17)}$$

où les coéfficients a(i) sont dits coéfficients de prédiction et σ est le gain du système .

2°-C-3) Estimation du modèle :

Nous considérons que notre signal x est à priori engendré par un système autorégréssif (AR).L'éxcitation de ce système est inaccéssible donc l'éstimation des parametres du modèle sera basée éxclusivement sur l'observation du signal.

Comme la récurrence linéaire (I-17) reste vérifiée ,nous pouvons définir une prédiction ou éstimation $\hat{x}(n)$ à partir des p échantillons qui le précèdent :

$$\hat{x}(n) = -\sum_{i=1}^{p} \hat{a}(i)\, x(n-i) \qquad \text{(I-18)}$$

Les coéfficients $\hat{a}(i)$ étant les éstimés des coéfficients a(i) .

L'érreur commise par la prédiction vaut :

$$\begin{aligned} e(n) &= x(n) - \hat{x}(n) \\ &= \sum \hat{a}(i)\, x(n-i) \quad ; \quad \hat{a}(0) = 1 \end{aligned} \qquad \text{(I-19)}$$

Si nous comparons les éxpressions (I-17) et (I-18) avec $\hat{a}(i)$ égal à a(i) pour i= 1..p ,nous trouvons :

$$e(n) = \sigma\, u(n) \qquad \text{(I-20)}$$

Dans notre suite,nous abandonons l'indice "^" pour désigner un coéfficient éstimé.

2°-C-4) Modélisation autorégréssive du signal vocal :

C-4-a) Formalisme LPC :

L'analyse LPC suppose que le signal vocal traite est approximé par un polynôme d'ordre p ,lequel ordre est choisi entre 10 et 30 .Un échantillon du signal y est représenté par les p échantillons qui le precedent par la combinaison linéaire

$$S_p(n) = \sum_{k=1}^{p} a(k)\, S(n-k) \qquad \text{(I-21)}$$

où n=1..N (indice d'échantillons par fenêtre) ;
et p l'ordre de prédiction.

C-4-b) Problème de la non-stationnarité :

Le signal vocal étant trés chaotique,ne peut être considéré quasi stationnaire que sur des intervalles de temps trés limités .

Nous sommes donc amenés à considérer des tranches successives et à éstimer un modèle AR pour chacune d'elles . La procédure usuelle consiste donc à éffectuer notre analyse successivement sur nos fenêtres de 20ms avec éxtraction des paramètres désirés au cours de chacune d'elles.

C-4-c) Energie résiduelle de prédiction :

Elle est définie come étant la variance à court terme de l'érreur de prédiction (3):

$$E_p = \sum_{n=inf}^{sup} e^2(n) \qquad \text{(I-22)}$$

Les limites inf et sup étant les bornes à définir su chaque fenêtre et selon la méthode utilisée pour l'éstimation des coéfficient LPC

C-4-d) Critère d'éstimation des coéfficients LPC :

Le critère usuel reste la minimisation de l'érreur de prédiction ou plus éxactement de l'énergie résiduelle E_p ie la somme des carrés des échantillons .

Ce critère peut être justifié par le fait que lorsque l'éxcitation est un train d'impulsions comme dans notre cas (voyelles orales : sons voisés), elle est nulle entre deux impulsions ;or l'érreur de prédiction représente l'éxcitation à un facteur prés (σ) .Il est donc bien justifié d'en minimiser l'érreur.

C-4-e) Méthode d'autocorrélation :

Dans cette méthode ,nous écrivons :

inf = 1 et sup = N+p ;

N est le nombre d'échantillons par fenêtre

$$E_p = \sum_{n=1}^{N+p} e^2(n) = \sum_{n=1}^{N+p} \left[S(n) - S_p(n) \right]^2$$

$$= \sum_{n=1}^{N+p} \left[S(n) - \sum_{k=1}^{p} a(k)\ S(n-k) \right]^2 \qquad (I-23)$$

Minimiser E_p revient à annuler sa dérivée par rapport aux coéfficients a(k),soit :

$$\frac{\partial E_p}{\partial a(i)} = 0 \qquad \text{pour} \quad i=1..p \qquad (I-24)$$

ie:

$$\sum_{n=1}^{N+p} 2 \left[S(n) - \sum_{k=1}^{p} a(k)\ S(n-k) \right] S(n-i) = 0$$

d'où:

$$\sum_{n=1}^{N+p} S(n)\ S(n-i) = \sum_{n=1}^{N+p} S(n) \left[\sum_{k=1}^{p} a(k)\ S(n-i) \right]$$

soit: $\sum_{N+p}$ $\sum_{p}$ $\sum_{N+p}$

$$\sum_{n=1} S(n-1)\ S(n) = \sum_{k=1} a(k) \left[\sum_{n=1} S(n-i)\ S(n-k) \right] \qquad (I-25)$$

Définissons la quantité $\phi(i,k)$ telle que :

$$\phi = \sum_{n=1}^{N+p} S(n-i)\ S(n-k)$$

par suite, l'équation (I-25) s'écrit plus compactement:

$$\sum_{k=1}^{p} a(k)\ \phi(i,k) = \phi(i,0) \qquad (I-26)$$

Vu que notre signal est nul en dehors de l'intervale $[1,N\ ]$, la quantité $\phi(i,k)$ peut s'écrire encore :

$$\phi(i,k) = \sum_{n=1}^{N+p} S(n-i)\ S(n-k)$$

$$= \sum_{n=1}^{N-(i-k)} S(n)S(n+i-k) = \sum_{n=1}^{N-i+k} S(n)S(n+i-k) \qquad (I-27)$$

Nous reconnaissons alors l'expression de la fonction d'autocorrelation $R(\tau)$ du signal $S(n)$ évaluée pour $\tau = i-k$.

En utilisant la parité de cette fonction, nous écrivons :

$$\phi(i,k) = R(|\, i-k \,|) \qquad \text{et} \qquad \phi(i,0) = R(|i|) = R(i)$$

La relation (II-26) devient alors :

$$\sum_{k=1}^{p} a(k)\, R(|i-k|) = R(i) \qquad (I-29)$$

avec : $i=1..p$ et $R(i) = \sum_{k=1}^{N-i} S(n)\, S(n+i)$

Cette équation est réellement une formulation matricielle tq :

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

La matrice ci-dessus est une matrice "**TOEPLITZ**" symétrique car les éléments de chaque diagonale parallèle à la diagonale principale sont égaux.

C-4-f) Résolution du système d'équations linéaires (S) :

De nombreuses méthodes permettent de résoudre le système d'équations d'autocorrélation (S), telles que :

- la méthode de Gauss-Seidel;
- la méthode de Jaccobi;
- la méthode de Gauss-Jordan;
- la méthode de Durbin;

Compte-tenu de la rapidité d'exécution et de la réduction de l'occu- pation mémoire, nous avons choisi, pour la résolution de notre système d'équations d'autocorrélation, la méthode de **DURBIN.**

Méthode de Durbin :

cette méthode permet de résoudre notre système par une

récursion sur l'ordre de la prédiction en considérant que la fonction d'autocorrélation est connue pour n=0..N.

Algorithme :(4)

Nous fixons l'ordre de prédiction à p =12

$$D(i)= \frac{1}{E(i-1)} \left\{ R(i) - \sum_{j=1}^{i-1} \left[a(j,i-1)R(i-j) \right] \right\}$$

$$a(i,i) = D(i)$$

$$a(j,i) = a(j,i-1) - D(i)\, a(i-j\ ;\ i-1)$$

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

avec : i =1..p ;et j =1..p

et où : E(i) : érreur quadratique;

D(i) : coéfficient de reflexion;

a(j,i): coéfficient de prédiction;

R(j) : coéfficient d'autocorrélation .

Conditions initiales :

$E(0) = R(0)$; $D(1) = R(1)/R(0)$

$a(1,1)= D(1)$; $E(1) =\left[1-D^2(1)\right] R(0)$

SOLUTIONS FINALES : $a(j) = a(j,p)$, $j= 1,p$

C-4-g) Remarques :

a) La dénomination *'autocorrélation* 'utilisée dans cette analyse ne correspond pas nécéssairement à la signification qui lui est attribuée en statistique .

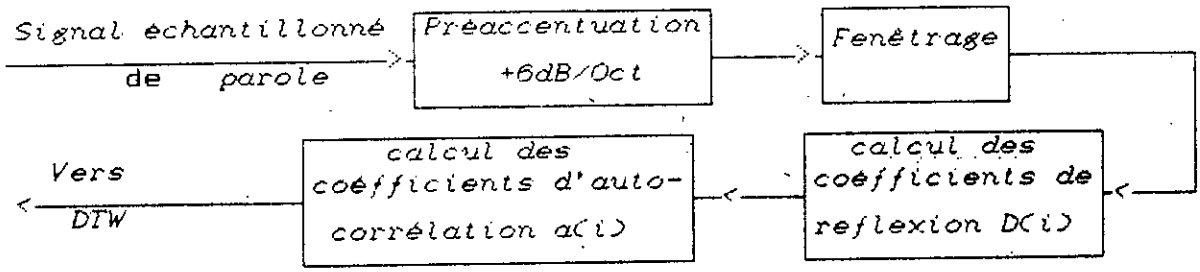
b)La *méthode d'autocorrélation* est parfois appelée *méthode stationnaire*,en effet nous étions amenés à considérer que : $\phi(i,k) = R(i,k) = R(j)$,ce qui constitue une propriété spécifique des signaux stationnaires.

c)l'érreur de prédiction doit présenter des valeurs éxagérées en

début d'intervalle car la prediction porte alors sur un nombre insuffisant d'échantillons ,idem en fin d'intervales où nous tentons de prédire des valeurs nulles pour S(n).Cependant,cet effet se trouve minimisé du fait que les échantillons du signal vocal ont déja été pondérés par une fonction fenêtre .[3].

2°-C-5) **Etapes d'analyse :**

Le synoptique suivant décrit les étapes de l'analyse LPC



2°-C-6) **Organigramme** : Voir l'annexe .

CHAP II

D.T.W

II-1°) INTRODUCTION :

Le taux de variation de la parole produit nécéssairement des fluctuations dans l'axe temporel des échantillons ou modèle considéré.Un même mot prononcé par une seule personne voit sa durée varier suivant le contexte,l'intonation,et son état personnel.L'élimination de ces fluctuations ,dite Normalisation temporelle,a été l'un des problèmes principaux dans la recherche sur la reconnaissance de la parole.

Parmi les techniques utilisées nous trouvons :

-La Nrmalisation Linéaire où les différences temporelles entre les échantillons sont éliminées par transformation linéaire de l'axe des temps;

-La Normalisation Non-Linéaire:plus conséillée que la précédente car elle consiste à moduler approximativement les fluctuations de l'axe des temps par une fonction de déformation non linéaire dotée de propriétés bien spécifiques .La différence temporelle entre deux échantillons de la parole est éliminée en déformant l'axe temporel de l'un d'entre eux de sorte que leurs maximums coinçident .La distance dans le temps sera donnée par la distance résiduèlle , entre eux , minimisée . La procédure de minimisation prouve son éfficacité avec l'utilisation de la programmation dynamique .

La comparaison des mots sera faite globalement sur leurs trames , si les mots sont différents,la distance qui les les sépare sera plus grande que celle qui découlera de la comparaison de mots identiques.

II-2°) FENETRES D'AJUSTEMENT :

La comparaison des mots se fait sur un éspace de recherche réstreint formé d'un rectangle dont chaque côté sera bordé par l'un des mots à comparer .Chaque mot est représenté par une suite

de trames dont le nombre dépent de sa longueur, et chaque trame est définie par un ensemble de coéfficients *cepstraux et/ou LPC*.

Pour optimiser les calculs , le champ de recherche de la fonction de déformation est limité par des contraintes qui éliminent les zônes trés improbables dans lesquelles toute comparaison n'a aucun sens .Les chercheurs ont donné deux fenêtres dites '*fenêtres d'ajustement*' :

-l'une est parrallepipédique (fig-II-3a) ;

-l'autre est de forme losange (fig-II-3b) .

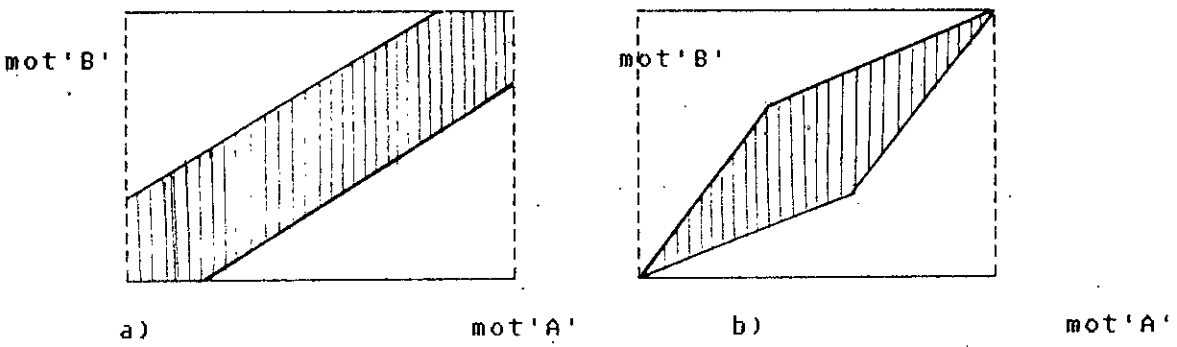


fig-V-1) Fenêtresd'ajustement

II-3°)DEFINITIONS GENERALES DE LA DISTANCE NORMALISEE DANS LE TEMPS

Sient deux vecteurs (acoustiques) A et B tq:

$A= a_1, a_2, \ldots, a_i, \ldots, a_I$ et $B=b_1, b_2, \ldots, b_j, \ldots b_J$

où a_i et b_j sont les trames d'indicesietjrespectivement des motsA'etB. Ces deux vecteurs représentent deux échantillons de la parole à comparer.

Considérons un plan (i-j) où A serait dévellopé sur l'axe i et B sur l'axe j . La distance temporelle entre A etB est représentée par une séquence de points C(i,j) pouvant donner une fonction qui réalise une correspondance approximative entre les axes temporels des échantillons A et B :

$$F = C(1), C(2), \ldots, C(k), \ldots C(K) \qquad (II-1)$$

où $C(k) = C(i(k), j(k))$

Cette fonction est appelée *fonction de déformation* .Elle coincide avec la diagonale si A et B sont identiques ,et s'en éloigne de d'autant plus que que la différence entre A et B augmente .

3° 1) Distance locale :

La mesure de la dissemblance ou de la distorsion entre les deux vecteurs A et B , est représentée par une distance D(A,B) tq:

$$D(A,B) = \sum_{k=1}^{K} d(C(k))$$

$$= \sum_{k=1}^{K} d(C_i(k), C_j(k)) \qquad (II-2)$$

où $d(C_i(k), C_j(k))$ est la distance locale où la dissemblance entre les trames a_i et b_j représentées par un nombre p de coéfficients cepstraux *et/ou LPC* $C_i(k)$ et $C_j(k)$.

Plusieurs formules de la distance locale sont envisageables :

-*Distance de Tchebytchef* :

$$d_{Tch}(a_i,b_j) = \sum_{k=1}^{p} \left|\left| C_i(k) - C_j(k) \right|\right|$$

-*distance cepstrale* :

$$d_{cep}(a_i,b_j) = \left[C_i(0) - C_j(0)\right]^2 + 2\sum_{k=1}^{p}\left[C_i(k) - C_j(k)\right]^2$$

-*Distance d'Itakura* :

$$d_{Ita}(a ,b) = \text{Log}\left[\frac{a_i\ V_i\ a_i'}{b_j\ V_j\ b_j'}\right]$$

Où V est la matrice d'autocorrélation

-*Distance cepstrale pondérée* :

$$d_{cp}(a_i,b_j) = \sum_{k=1}^{p} k^2 \left[C_i(k) - C_j(k)\right]^2$$

3° 2) Distance Globale :

La somme pondérée des distances sur la fonction de déformation F s'écrit :

$$E(F) = \sum_{k=1}^{p} d(\ C(k)\ ).w(k) \qquad (w(k) \geq 0)$$

Les coéfficients de pondération w(k) ont été introduits pour permettre une meilleure flexibilité dans la mesure de E(F) .Cette dernière atteint son minimum quand F est déterminée avec un ajustement optimal de la différence temporelle

La valeur de la distance résiduèlle minimum peut être considérée comme la distance entre les échantillons A etB . Cette distance devrait montrer une stabilité vis à vis des fuctuations des axes des temps . Par suite , la distance globale entre A et B est donnée par :

$$D(A,B) = \underset{F}{\text{Min}} \left\{ \frac{\sum_{k=1}^{K} d(\ C(k)\ ).w(k)}{\sum_{k=1}^{K} w(k)} \right\} \qquad (II-3)$$

Les caractéristiques éffectives de cette distance dépendent largement des spécifications de la fonction de déformation et de la définition des coéfficients de pondérations ainsi que des propriétés des échantillons considérés d'où deux conditions à satisfaire :

-- les échantillons sont obtenus avec une période d'échantillonnage commune et constante ;

-- les échantillons A et B sont supposés contenir la même quantité d'informations phonétiques .

II-4°) RESTRICTIONS SUR LA FONCTION DE DEFORMATION :

La fonction F est un modèle des fluctuations de l'axe des temps dans un échantillon de parole .Les caractèristiques éssentièles des échantillons de parole sont *la continuité* ,*la monotonicité* , *la limitation sur la vitesse de transition des*

parametres acoustiques....etç.

Ces conditions sont considérées comme des réstrictions sur la fonction F

1) condition de monotonicité :

$$\begin{cases} i(k-1) \leq i(k) \\ j(k-1) \leq i(k) \end{cases}$$

2)condition de continuité :

$$\begin{cases} i(k)-i(k-1) \leq 1 \\ j(k)-j(k-1) \leq 1 \end{cases}$$

Ces deux réstrictions nous permettent d'écrire :

$$C(k-1) = \begin{cases} (\ i(k)\ ,\ j(k)-1\ ) \\ \text{ou}\ (\ i(k)-1\ ,j(k)-1\ ) \\ \text{ou}\ (\ i(k)-1\ ,j(k)\ ) \end{cases}$$

3)conditionsaux limites :

$$\begin{cases} i(1)=1; j(1)=1 \\ i(K)=I; j(K)=J \end{cases}$$

4) condition sur l'ajustement de la fenêtre :

$$\left| i(k) - j(k) \right| \leq r \quad ; \text{ largeur de la fenêtre } r > 0$$

5) condition sur la contrainte de pente :

La pente de la fonction de déformation F ne doit être ni trop raide ni trop douce car de telles déviations peuvent causer des déformations indésirables de l'axe temporel. Cette contrainte a été introduite pour que sa première dérivée soit de forme discrète

Par ailleurs,la condition sur la contrainte de pente est représentée comme une restriction sur les configurations possibles des points consécutifs de'la fonction F .

L'intensité effective de cette contrainte (*fig*-II-4) peut être évaluée par : $p = n / m$

où : n est le nombre de déplacements sur la diagonale ;

et m le nombre de déplacements sur l'axe i ou l'axe j .

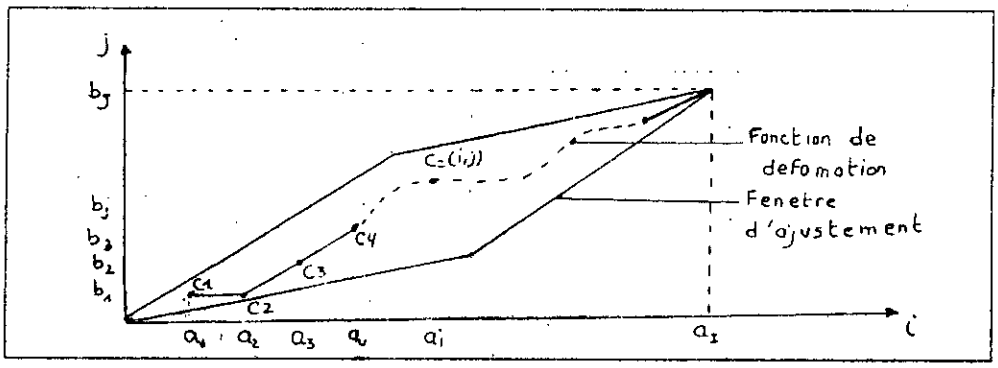
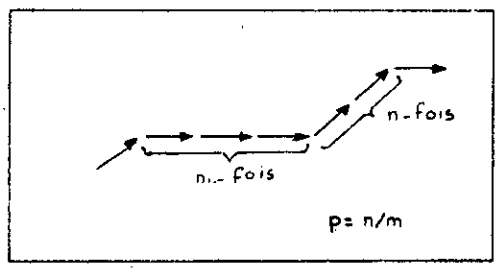


fig II-3 fonction de déformation F .



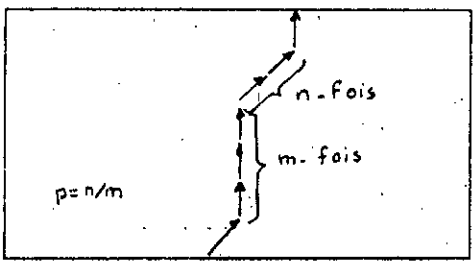


fig II-4 Intensité effective p de F .

Discussion sur l'intensité p :

- si p = 0 :il n'ya pas de réstrictins sur F ;
- si p est infinie :F est réstreinte à la diagonale i=j ;
- si p est sévère (grande) : la normalisation ne se fait pas correctement ;
- si p est douce (petite) :la discrimination entre les différentes catégories des échantillons de parole sera dégradée .

Donc p ne doit être ni trop douce ni trop sévère : $0 \leq p \leq 2$

Chemins possibles des points C(k) :

Soit une pente: p = 1

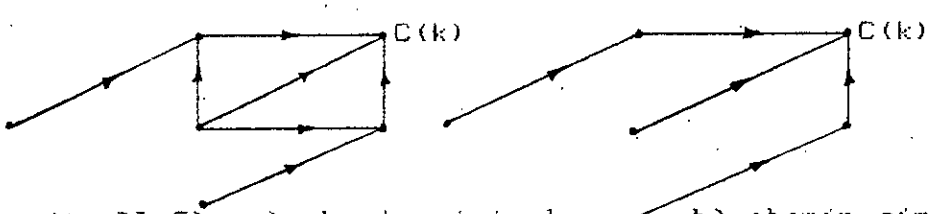


fig-II-5)- a) cheminoriginal *b) chemin simplifié*

Le chemin simplifié (b) répond à une nouvele contrainte qui est: 'la deuxième dérivée de F doit satisfaire la réstriction que le chemin du point C(k) ne doit pas changer de direction orthogonalement '.Cette nouvelle réstriction réduit donc le nombre de chemins à chercher .

II-5°) LES COEFFICIENTS DE PONDERATIONS :

Soit N le dénominateur de l'éxpression de D(A,B) (II-3):

$$N = \sum_{k=1}^{K} w(k)$$

Si N est indépendant de la fonction F de déformation, D(A,B) sera:

$$D(A,B) = \frac{1}{N} \text{Min}\left[\sum_{F} d(\ C(k)\ ).w(k) \right] \qquad \text{(II-4)}$$

Le problème de comparaison serait,alors,simplifié et facilement résolvable par la P.D. Il y a deux définitions typiques des coéfficients de pondération qui permettent cette simplification.

5°-1)Forme symétrique :

$$W(k) = (\ i(k)-i(k-1)\ ) + (\ j(k)-j(k-1)\ )$$

avec : $N = I + J$

5°-2) Forme assymétrique :

$$W(k) = \begin{cases} (\ i(k)-i(k-1)) \text{ avec : } N = I \\ \text{ou} \\ (\ j(k)-j(k-1)) \text{ avec : } N = J \end{cases}$$

Le concept de ces deux formes a été défini à l'origine par SAKOE et CHIBA. S'il est assuré que les axes du temps sont tous deux continus alors, dans la forme symétrique, la sommation dans l'expression de D(A,B) est une intégration le long de l'axe temporel : l =i+j .Par contre dans la forme assymétrique ,cette sommation est une intégration le long de l'axe i ou j .

Comme résultat de cette différence,,la distance normalisée dans

le temps est symétrique [D(A,B) = D(B,A)] pour la forme symétrique mais pas pourla forme assymétrique .

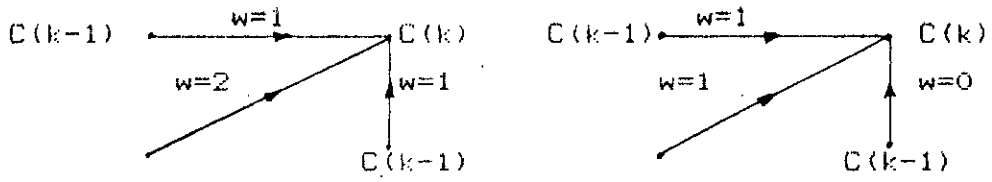


fig II-6) a)forme symétrique *b)forme assymétrique*

Un autre résultat plus important dans la différence dans les axes d'intégration est que le coéfficient de pondération w(k) est réduit à zéro dans la forme assymétrique quand le point de la fonction F se déplace dans la direction de l'axe j où : C(k) = C(k-1) + (0,1) ie que quelques vecteurs b_j peuvent être exclus de l'intégration . Par contre , dans la forme symétrique , aucune exclusion n'est faite puisque : Min(w(k)) = 1 Par suite,la forme symétrique donne de meilleurs résultats lors d'un même traitement des échantillons .

II°-6) ALGORITHME DE COMPARAISON PAR P.D : [6]

L'algorithme de base de D(A,B) s'écrit :

---condition initiale :

$$g_1(C(1)) = d(C(1).w(1))$$

---EquationdeP.D:

$$g_k(C(k)) = \underset{c(k-1)}{\text{Min}} \left[g_{(k-1)}(C(k-1)) + d(C(k)).w(k) \right]$$

---Distance normalisée dans le temps :

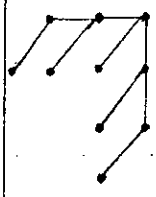
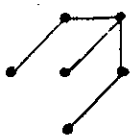
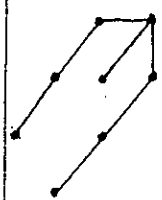
$$D(A , B) = \frac{1}{N} g_k(C(k))$$

---Suppositions implicites:

$$C(0,0) = 0 \qquad \text{et } w(1) = \begin{cases} 2 \text{ pour forme symétrique} \\ 1 \text{ pour forme assymétrique} \end{cases}$$

Dressons un tableau qui donne les équations de la P.D pour les différentes pentes et pour les deux formes proposées par SAKOE et

CHIBA .

pente	chemins	forme	Equation de P.D : g(i,j)=Min.....
p=0		symé-trique	$\begin{cases} g(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-1,j)+d(i,j) \end{cases}$
		assymé-trique	$\begin{cases} g(i,j-1) \\ g(i-1,j-1)+d(i,j) \\ g(i-1,j)+d(i,j) \end{cases}$
p=1/2		symé-trique	$\begin{cases} g(i-1,j-3)+2d(i,j-2)+d(i,j-1)+d(i,j) \\ g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \\ g(i-3,j-1)+2d(i-2,j)+d(i-1,j)+d(i,j) \end{cases}$
		assymé-trique	$g(i-1,j-3)+\left[d(i,j-2)+d(i,j-1)+d(i,j)\right]\frac{1}{3}$ $\begin{cases} g(i-1,j-2)+\left[d(i,j-1)+d(i,j)\right]\frac{1}{2} \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \end{cases}$ $g(i-3,j-1)+d(i-2,j)+d(i-1,j)+d(i,j)$
p=1		symé-trique	$\begin{cases} g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \end{cases}$
		assymé-trique	$\begin{cases} g(i-1,j-2)+\left[d(i,j-1)+d(i,j)\right]\frac{1}{2} \\ g(i-1,j-1)+d(i,j) \\ g(i-2,j-1)+d(i-1,j)+d(i,j) \end{cases}$
p=2		symé-trique	$\begin{cases} g(i-2,j-3)+2d(i-1,j-2)+2d(i,j-1)+d(ij) \\ g(i-1,j-1)+2d(i,j) \\ g(i-3,j-2)+2d(i-2,j-1)+2d(i-1,j)+d(ij) \end{cases}$
		assymé-trique	$\begin{cases} g(i-2,j-3)+2\frac{d(i-1,j-2)+d(i,j-1)+d(i,j)}{3} \\ g(i-1,j-1)+d(i,j) \\ g(i-3,j-2)+d(i-2,j-1)+d(i-1,j)+d(i,j) \end{cases}$

Autres algorithmes pour la P.D :

Algorithme	condition initiale	N	Equation de P.D : g(i,j)=
Vélichko et Zagoruyko		Max(I,J)	$\text{Max}\begin{cases} g(i,j-1) \\ g(i-1,j-1)+a(i,j) \\ g(i-1,j) \end{cases}$ avec : $a(i,j)=1-d(i,j)$
White et Neely	d(i,j)	I*J	$\text{Min}\begin{cases} g(i-1,j)+d(i,j) \\ g(i-1,j-1)+d(i,j) \\ g(i,j-1)+d(i,j) \end{cases}$
Itakura	d(1,1)	I	$\text{Min}\begin{cases} g(i-1,j)+\alpha\, d(i,j) \\ g(i-1,j-1)+d(i,j) \\ d(i-1,j-2)+d(i,j) \end{cases}$ avec : $\begin{cases} \infty & \text{si } i(k-1)=i(k-2) \\ 1 & \text{sinon} \end{cases}$

Exemple le plus simple :

Pente : p = 0 ;

Forme : symétrique ;

Ajustement de fenêtre : r =4 , $j-r \leq i \leq j \pm r$;

Equation de P.D : algorithme de Sakoe et Chiba

$$g(1,1) = 2\, d(1,1)$$

$$g(i,i) = \text{Min}\begin{cases} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{cases}$$

DistanceD(A,B): $D(A,B) = \frac{1}{N} g(I,J)$ avec N = I+J .

II°-7) ORGANIGRAMME : Voir l'annexe .

III-1 / INTRODUCTION :

Notre système de reconnaissance de la parole a pour but d'identifier une locution prononçée par n'importe quel locuteur . Le dictionnaire , que nous lui associons ,aura pour données les resultats des methodes d'analyse que nous aurons déjà éffectuées pour chaque locution issue de notre population de référence .

Plusieurs travaux en Reconnaissance de la parole dont nous citons celui de A.Mokkedem [7],ont montré qu'au sein d'une population il existe un certain nombre de classes de prononciations d'une locution donnée ;chaque classe est caractéristique d'une fraction de la population . Nous en faisons une hypothèse justifiée et nous la vérifierons lors de nos tests .

La classification que nous faisons est un ensemble d'algorithmes dont le but est ,partant d'un éspace totalement inconnu ,de definir des classes à partir de l'ensemble des éléments qui composent cet éspace [2] .

Mais pourquoi ,utilisons - nous la classification ?
Tout simplement pour la simplification de l'ensemble initial des données. La classification ,en effet ,s'attache à recouvrir la structure profonde de nos données qui sont complexes et en quantité importante donc difficiles à cerner . Par contre ,une structure sous- jacente de ces données peut etre trés simple , donc intéréssante à découvrir . C'est le but que nous recherchons .

Nous éxposons ci-aprés plusieurs algorithmes de classification que nous éxploiterons pour la détermination de nos classes et le choix du représentant de chacune d'elles . Parmi ces algorithmes, algorithmes,Nous citons :

- Algorithme de Grouprement en (CHAINMAP) .
- Algorithme des K-Moyennes .
- Algorithme ISODATA .

III-2 / TYPES DE CLASSIFICATION :

La première étape de la classification dite " d'Apprentissage " permet de choisir un représentant pour chaque classe . Ces représentants de classes constitueront le dictionnaire de référence .

On distingue 3 types de techniques d'Apprentissage [8]:

2-1 / Classification supervisée :

Dans cette technique ,le nombre et la nature des classes sont connus ainsi que la répartition des échantillons . On distingue deux méthodes :

- Méthodes statistiques .
- Méthodes géométriques .

2-2 / Classification non-supervisée :

Dans cette technique ,le nombre et la nature des classes sont inconnus . Les échantillonssont alors classés éssentiellement suivant trois méthodes que nous développerons plus tard et qui sont:

- Méthodes des centroïdes .
- Méthodes des groupements en chaines (Chainmap) .
- Méthodes du SNN (Shared Nearest Neighbor) .

2-3 / Classification adaptative :

Une Classification est dite adaptative quand des parametres ou la structure de l'algorithme sont modifiés pour optimiser la classi-classification ,comme dans le cas de l'algorithme ISODATA qui sera détaillé plus loin .

III-3 / RAPPELS MATHEMATIQUES :

Soit Ω l'ensemble des N locutions d'un Mot X tel que :

$$\Omega = \{x_1, x_2, x_3, \ldots\ldots, x_t, \ldots\ldots, x_j, \ldots\ldots, x_N\} \qquad \text{(III-1)}$$

où les x_i (i=1..N) sont les N locutions différentes du Mot X .

Nous pouvons former une matrice D carrée (NxN),où l'élement d_{ij} de rang ij est une mesure de dissimilarité entre les locutions x_i et x_j ,définie par :

$$d_{ij} = \delta(x_i, x_j) = \underset{F}{\text{Min}} \left[\frac{\sum_{k=1}^{K} d(c(k)).w(k)}{\sum_{k=1}^{K} w(k)} \right] \qquad \text{(III-2)}$$

où : K est le nombre de points de la fonction de déformation F ;

w(k): Fonction de pondération dans la $k^{iéme}$ fenêtre .

avec :

$$d(c(k)).w(k) = d(c) = d(i,j) = \| a_i - a_j \| \qquad \text{(III-3)}$$

où a_i et a_j sont les vecteurs LPC ou Cepstraux dans la $k^{iéme}$ fenêtre .

L'ensemble Ω est formé de M classes $\{w_i\}$ pas nécessairement disjointes ,de locutions du Mot X .

donc :

$$\Omega = \bigcup_{i=1}^{M} w_i \qquad \text{(III-4)}$$

Notons m_i :le cardinal de la classe w_i et $x_p^{(i)}$,le centre ou le prototype de cette méme classe . Le $k^{iéme}$ plus proche voisin de x est noté $x_{[k]}^{(i)}$ avec la relation :

$$\delta(x, x_{[1]}) \leq \delta(x, x_{[2]}) \leq \ldots \leq \delta(x, x_{[k]}) \leq \ldots \leq \delta(x, x_{[N]}) \qquad \text{(III-5)}$$

La qualité de mesure σ est donnée par le rapport de la moyenne des distances inter-classes sur la moyenne des distances intra-classes , Soit :

$$\sigma = \frac{\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \delta(x_p^{(i)}, x_p^{(j)})}{\frac{1}{M} \sum_{i=1}^{M} \frac{1}{m_i(m_i-1)} \sum_{j=1}^{mi} \sum_{k=1}^{mi} \delta(x_j^{(i)}, x_k^{(i)})} \qquad \text{(III-6)}$$

Si pour deux classes , $\sigma > 2$,cela signifie que ces deux classes ne se recouvrent pas et sont bien disjointes .

III- 4 / CLASSIFICATION :

4-1 / Choix du type de classification :

Vu que dans notre cas ,nous n'avons à priori aucune information sur l'ensemble de nos locutions . Nous devons considérer et adopter la technique d'Apprentissage non supervisée .

Les méthodes utilisées par cette technique doivent être complémentaires pour donner de meilleurs resultats . En effet ,dans la méthode des centroydes , la connaissance au préalable du nombre réel de classes assurera une convergence rapide de l'algorithme .Ce nombre peut être éstimé par la méthode du Chainmap et confirmé par la méthode du SNN . Cette dernière (SNN) éxige ,pour donner des résultats rapides ,une bonne éstimation du nombre de classes .

Ces trois méthodes peuvent être complétées par la méthode ISODATA (Interactive Self Organizing Data Analysing Technique A) qui est une méthode adaptative servant à isoler les locutions "hors communs " (the Outliers) , à fusionner ou à faire éclater des classes pour ameliorer la configuration (augmenter c) .

4-2 / Classification non-supervisée :[7]

4-2-1 / Méthode de groupement en chaine ("CHAINMAP") :[9]

C'est une méthode d'analyse simple qui lorsqu'elle est représentée graphiquement en deux dimensions,peut donner beaucoup d'informations sur la distribution des différentes locutions du mot .

En première étape , nous désignons arbitrairement une locution x_s comme début de chaine , la locution suivante est le voisin le plus proche $x_{s(1)}$ de x_s . L'étape suivante est d'ordonner toutes les locutions dans l'ordre suivant :

$$x_s \quad x_{s(1)} \quad \ldots\ldots\ldots\ldots\ldots \quad x_{s(N-1)} \qquad \text{(III-7)}$$

Au $k^{ème}$ élément de cette chaine , nous associons la distance d_k

définie par :

$$d_k = \delta(x_{s(k-1)}, x_{s(k)}) \qquad \text{(III-8)}$$

Le chainmap ou graphe de d'ordonnance est la courbe de la fonction $d_k = f(k)$. Les pics trés de cette courbe , représentent les limites entre classes qui souvent assez distinctementvisibles . Cette procédure est sensible au choix du point de commencement .

4-2-2 / Méthode du voisin comun le plus proche (SNN) : [9]

La procédure SNN est basée sur le fait que si deux locutions ont au moins Ks voisins communs les plus proches , alors elles appartiénnent à une même classe .

Soit L la liste des voisins les plus proches tq :

$$L = \begin{bmatrix} x_1 & x_{1(1)} & \cdots & \cdots & x_{1(k)} \\ x_2 & x_{2(1)} & \cdots & \cdots & x_{2(k)} \\ \vdots & \vdots & & & \vdots \\ x_N & x_{N(1)} & \cdots & \cdots & x_{N(k)} \end{bmatrix} \qquad \text{(III-9)}$$

où les N lignes correspodent aux N locutions . La ligne L_i est la liste ordonnee des locutions qui sont les k voisins les plus proches de x_i .

Si on a 2 lignes L_i et L_j tel que : $|L_i \cap L_j| \geq k_s$

k_s est le seuil fixé , c.a.d que x_i et x_j partagent au moins k_s voisins ,alors les deux listes sont assignées à une même classe .

Il est bien sûr possible qu'une "locution test" x partage $k_i > k_s$ voisins avec x_1 , et k_j k_s voisins avec : $k_i + k_j$ k alors x aux deux classes w_i et w_j .

Par ailleurs , cette procédure peut être utilisée pour identifier les recouvrements de classes .En pratique , on définit une limite lmax de classes auquelles peut appartenir la locution test .Il en est de même pour les valeurs de k_s et k .

4-2-3 / Méthode des K-moyennes (the K-Means Itération) :[9]

Cette méthode est une technique d'itération automatique qui peut trouver n'importe quel nombre spécifié de classes . Les itérations consistent en trois opérations de base :

--classification des locutions ;

--calcul des centres de classes ;

--test de convergence .

En assumant le souhait de trouver M classes w_i , nous choisissons M locutions x_i comme centres initiaux $x_p^{(i)}$ de ces classes ie :

$$x_p^{(i)} = x_i \quad \text{pour } i = 1...M \qquad \text{(III-10)}$$

-La classification est basée sur la méthode du voisin le plus proche soit pour k allant de 1 à M :

$$x_j \in w_i \quad \text{ssi} \quad \delta(x_j, x_p^{(i)}) \prec \delta(x_j, x_p^{(k)}) \qquad \text{(III-11)}$$

-Aprés avoir appliqué ce critère pour toutes les locutions du mot référence ,soit j = 1..M ,nous recalculons les nouveaux centres des classes w_i avec le critère suivant :

$$x_p^{(i)} = x_j^{(i)} \quad \text{ssi} \quad \underset{k=1..M}{\text{Max}} \left\{ \delta(x_j^{(i)}, x_k^{(i)}) \right\} \text{ est minimal .} \qquad \text{(III-12)}$$

c.à.d que la distance de x_j à l'élément le plus éloigné de la classe w_i , est la plus faible parmi tous les éléments de w_i .

-Le test de convergence consiste à voir si les locutions choisies "arbitrairement" au début comme centres de classes,sont les mêmes que celles trouvées à la fin.

Si le test n'est pas vérifié,on refait les iterations précédentes avec les nouveaux centres jusqu'à la stabilité du processus itératif.

Notons que la convergence souhaitée, n'est pas toujours garantie surtout quand le nombre M de classes diffère de la réelle structure de l'ensemble des locutions .Pour donner alors de meilleures perfor-

mances ,cette méthode est souvent accompagnée de l'ISODATA .

4-2-4 / Méthode ISODATA : [9]

Cette méthode permet d'isoler les "locutions hors communs "et de déteriner le nombre réel M de classes et cela en fusionnant des paires de classes ou en faisant éclater certaines classes selon la valeur du rapport σ donnant la qualité de mesure .

La partie principale de l'ISODATA est la procédure des "k-Moyennes" avec ajustement progréssif du nombre de classes au fil des itérations .

a / Fusion de classes : Des classes peuvent être fusionnées si l'un des trois critères suivants est vérifié :

--Le nombre présent M de classes dépasse la valeur seuil Mmax :les classes les plus rapprochées seront alors fusionnées ;

--La dimension $|w_j|$ de $j^{ème}$ classe est inférieure à une valeur seuil m_{min} soit : cette classe sera alors fusionnée avec la classe qui lui est la plus proche ;

--La distance $\delta(x_p^{(i)}, x_p^{(j)})$ entre les centres de la $i^{ème}$ et la $j^{ème}$ classes est inférieure à une distance seuil σ_{min} : ces deux deux classes seront fusionnées .

Lors d'une itération où il ya plus d'une fusion de classes , le dernier critère doit être constamment revérifié .

b / Réduction de classes :

Il ya principalement trois critères pour lesquels une classe peut être réduite ou " éclatée " :

--Le nombre présent M de classes est inférieur à une une valeur Mmin;

--La diension de la $i^{ème}$ classe $|w_j|$ dépasse une valeur seuil m_{max} ;

--La $i^{ème}$ classe est trop dense relativement aux autres classes.

Les deux premiers critères sont similaires respectivement à ceux

mances ,cette éthode est souvent accompagnée de l'ISODATA .

4-2-4 / Méthode ISODATA : [9]

Cette methode permet d'isoler les "locutions hors communs "et de déteriner le nombre reel M de classes et cela en fusionnant des paires de classes ou en faisant éclater certaines classes selon la valeur du rapport σ donnant la qualité de mesure .

La partie principale de l'ISODATA est la procédure des "k-Moyennes" avec ajustement progressif du nombre de classes au fil des itérations .

a / Fusion de classes : Des classes peuvent être fusionnées si l'un des trois critères suivants est vérifié :

--Le nombre présent M de classes dépasse la valeur seuil Mmax ;les classes les plus rapprochées seront alors fusionnées ;

--La dimension $|w_i|$ de $i^{ème}$ classe est inférieure à une valeur seuil m_{min} soit : cette classe sera alors fusionnée avec la classe qui lui est la plus proche ;

--La distance $\delta(x_p^{(i)}, x_p^{(j)})$ entre les centres de la $i^{ème}$ et la $j^{ème}$ classes est inférieure à une distance seuil θ_{min} : ces deux deux classes seront fusionnées .

Lors d'une itération où il ya plus d'une fusion de classes , le dernier critere doit être constamment reverifié .

b / Réduction de classes :

Il ya principalement trois critères pour lesquels une classe peut être réduite ou " éclatée " :

--Le nombre présent M de classes est inférieur à une une valeur Mmin;

--La diension de la $i^{ème}$ classe $|w_i|$ dépasse une valeur seuil m_{max} ;

--La $i^{ème}$ classe est trop dense relativement aux autres classes.

Les deux premiers critères sont similaires respectivement à ceux

de la fusion. Le troisième qui est plus compliqué, fait intervenir certaines procedures:

Nous calculons les distances interclasses Di :

$$D_i = \frac{1}{m_i - 1} \sum_{x \in w_i} \delta(x , x_p^{(i)}) \quad , \quad i = 1 \ldots M \qquad \text{(III-13)}$$

et nous faisons leur moyenne $\overline{D}$ donnée par :

$$\overline{D} = \frac{1}{M} \sum_{i=1}^{M} m_i \, D_i \quad , \quad m_i = \dim(w_i) \qquad \text{(III-14)}$$

Pour un seuil de réduction θ_s , w_i sara réduite ou décomposée :

$$D_i > \text{Max} \left\{ \overline{D} , \theta_s \right\}$$

Quand w_i est réduite ,elle est divisée en deux nouvelles classes w_i^+et w_i^- telles que : $w_i = w_i^+ \cup \bar{w}_i$

Cette procédure s'est révélée peu commode pour la décomposition car nous ne connaissons que les distances entre locutions . A l'opposé , nous disposons de deux méthodes simples et plus rapides :

--La plus simple consiste à choisir deux locutions x^+ et x^- dans w_i tels que la distance $\delta(x^+, x^-)$ maximum . Puis ,nous assignons aux classes w_i^+ et w_i^- les locutions de la classe w_i selon que leurs distances à x^+et x^- sont plus petites mais ces deux derniers ne sont pas les vrais centroïdes des deux nouvelles classes .

--La seconde méthode donne une plus meilleure éstimation des centres de classes . Comme précédemment , nous localisons x^+ et x^- puis nous posons :

$$\begin{cases} r^+ = \delta(x^+ , x_p^{(i)}) \\ r^- = \delta(x^- , x_p^{(i)}) \end{cases}$$

puis nous cherchons x_p^+ et x_p^- de façon à minimiser :

$$\begin{cases} \varepsilon^+ = \delta(x^+, x_p^+) + \delta(x_p^+ , x^{(i)}) - r^+ \\ \varepsilon^- = \delta(x^-, x_p^-) + \delta(x_p^- , x^{(i)}) - r^- \end{cases}$$

comme dans la première méthode , les éléments de w_i sont assignés aux classes w_p^+ et w_p^- selon leur proximité de x_p^+ et w_p^+ , ces

deux derniers sont alors les vrais centroïdes des classes w_p^+ et w_p^-.

III-5 / CONCLUSION :

Les centres de nos classes sont calculés selon l'équation (III-12) et Leurs délimitations sont faites selon la relation (III-11) ;ces deux étapes sont incluses dans la méthode des K-Moyennes .Leurs fusions et/ou réductions sont obtenues selon les les critères précédemment cités .

Nous n'utiliserons pas la méthode des SNN car elle éxige un nombre trés grand de locutions ,de l'ordre de 100 , notre temps-machine ne nous permet pas cette possibilité .

A la fin , nous faisons nos tests de convergence selon la méthode des K-Moyennes , l'arrêt est quand la configuration présente de nos locutions a un rapport de qualité de mesure σ supérieur au seuil σ_{seuil} que nous définissons .

PRECISION

IV-1 / INTRODUCTION :

La décision est la dernière étape dans la reconnaissance de la parole .Elle consiste à prendre la décision sur le mot test en éxploitant les résultats des distances globales ,issus de l'étape de comparaison dynamique où ces distances sont classées selon un ordre croissant .

Toutefois , l'ordre établi des distances globales peut ne pas être conforme à la disposition éxacte c-à-d que le mot qui ,devrait être choisi pour la décision ,ne se trouve pas à la première position et ceci pour des raisons telles qu'une mauvaise éxtraction des mots , ou une ressemblance phonétique entre les mots du dictionnaire etç.

IV-2 / TECHNIQUE DES KNN (K-Nearest Neighbors) :

Cette technique est utilisée pour remédier au problème cité çi-dessus en utilisant seulement les "K" premières références de chaque mot du dictionnaire pour établir leur classement et non plus un classement référence par référence .

Bien sûr cette technique est trés coûteuse en temps de calcul et d'éspace mémoire , c'est pourquoi nous devons veiller à optimiser ces deux contraintes surtout qu'en mode multilocuteur ,chaque mot référence est représenté par plusieurs locutions appelées prototypes et issues de la phase d'apprentissage .

Dans nos tests , nous avons opté pour la forme généralisée utilisée par Mrs Pan ,Song et Rabiner [10] .

IV-3 / REJETS :

Certains mots peuvent être "rejetés" par le système ie que ce dernier ne répond pas à l'ordre cherché lorsque la plus petite distance est plus grande qu'une distance seuil que nous lui fixons

d'avance .Cette procédure est donc utile et nécéssaire puisqu'elle nous permet d'écarter une éventuelle détection de bruit ou mauvaise prononciation d'un mot étranger au dictionnaire .

IV-4 / ALGORITHME :[10]

Nous supposons que ,dans le dictionnaire de références , nous disposons de J mots notés X_j ,$j=1...J$ avec pour chacun d'eux Q locutions références ou prototypes ,soit au total (Q x J) locutions.

Pour chaque mot X_j , nous faisons un ordonnancement croissant des distances globales issues de l'étape de DTW avec le mot test ,soit :

$$D_{j(1)} \leq D_{j(2)} \leq \ldots\ldots\ldots \leq D_{j(q)} \leq \ldots\ldots\ldots \leq D_{j(Q)} \qquad \text{(IV-1)}$$

Puis nous prenons la moyenne des K premières distances pour chaque mot ie ses K-voisinslesplusproches.:

$$r_j = \frac{1}{K} \sum_{q=1}^{K} D_{j(q)} \quad , \quad j=1..J \text{ et } K \leq Q \qquad \text{(IV-2)}$$

La décision finale sur la reconnaissance du mot test repose sur la recherche du mot X_j qui donne la plus faible valeur de r_j :

$$X_{test} = X_j \quad \text{ssi} \quad r_j = \min_{j=1..J} (r_j) \qquad \text{(IV-3).}$$

N.B : Nos tests en mode multilocuteurs devraient nous renseigner sur le choix de la valeur de K pour un nombre J de prototypes donnés.

Beyond Belief

Fenton's Wonderful Electric Land

V-1/ INTRODUCTION :

Le conduit vocal favorise certaines zônes fréquentièlles où l'on obtient des résonnances .Ces zônes sont les FORMANTS , et se caractérisent par un maximum de la fonction de transfert du conduit vocal .La richesse du signal vocal en formants d'intensités relatives détermine le timbre de la voix .En général , les trois premiers formants sont éssentiels pour caractériser le spectre du signal parole . Dans le cas de nos voyeles ,les deux premiers formants ,que nous noterons F1 ET F2 ,sont suffisants et peuvent déterminer à eux seuls le timbre de la voix . Notons que les positions des fréquences de ces formants peuvent varier pour un même son chez un même locuteur et encore plus quand on passe d'un locuteur à un autre .

Par ailleurs, le premier harmonique du spectre vocal ,dit PITCH, ne présente pour nous aucun intéret particulier puisque nous nous éloignons de la reconnaissance des locuteurs .Il est donc suffisant de représenter nos voyelles par les formants F1 et F2 .

V-2/ METHODES DE SIMULATIONS :

Les sons voisés tels que les voyelles orales, que nous étudions, sont dûs à une éxcitation pseudo-périodiques où les cordes vocales vibrent sous l'action de la préssion du flux d'air envoyé par les poumons

Pour simuler nos voyeles , nous proposons deux méthodes où nous nous basons sur les valeurs moyennes de F1 et F2 [11] suivantes :

Voyelles	F1(Hz)	F2(Hz)
A	750	1350
E	400	2200
I	250	2500
O	375	750
U	250	600
Y	050	1800

2 / 1) Méthode 1 :

Cette méthode est basée sur une somme de sinusoïdes de fréquences F1-B1/2 , F1 , F1+B1/2 , F2-B2/2 , F2 , F2+B2/2 où B1 et B2 sont respectivement les bandes à 3dB des formants F1 et F2 que nous prenons respectivement égales ,içi , à 80 Hz et 90 Hz .

L'équation de génération de nos échantillons s'écrit :

$$S(k) = A1 \sin 2\pi k \left(\frac{F1}{Fe}\right) + A2 \sin 2\pi k \left(\frac{F2}{Fe}\right) + \frac{A1}{\sqrt{2}}\left[\sin 2\pi k \left(\frac{F1-40}{Fe}\right) + \sin 2\pi k \left(\frac{F1+40}{Fe}\right)\right] + \frac{A2}{\sqrt{2}}\left[\sin 2\pi k \left(\frac{F2-45}{Fe}\right) + \sin 2\pi k \left(\frac{F2+45}{Fe}\right)\right] . \quad (V-1)$$

avec :

S(k) : $k^{ème}$ échantillon de la locution considérée ($1 \leq k \leq Nech$) ;

A1 : amplitude du premier formant ;

A2 : amplitude du second frmant ;

Fe : fréquence d'échantillonnage (Fe = 12.8 Khz) ;

Nech : nombre total d'échantillons par locution .

Les limites de cette méthode se situent dans le fait que les échantillons générés sont des signaux déterministes alors que la parole réelle est un signal aléatoire et quasi-stationnaires sur de courtes durées.

Nous ne retiendrons pas alors cette méthode parce que le

spectre des signaux que nous obtenons est très différent de celui de la parole réelle et dès que nous faisons des variations en formants ou en durée temporelle , le signal devient altéré et totalement différent de celui duquel il dérive et que nous cherchons à reconnaitre .

Tous nos tests seront alors faits avec la méthode de génération suivante .

2 / 2) Methode 2 : [3]

Elle est basée sur le modèle autorégréssif du signal de la parole . Pour les voyelles non nasales ,ce modèle ne présente que des pôles .L'excitation U(Z)de ce système est une suite périodique d'impulsions de période P égale à la période du pitch

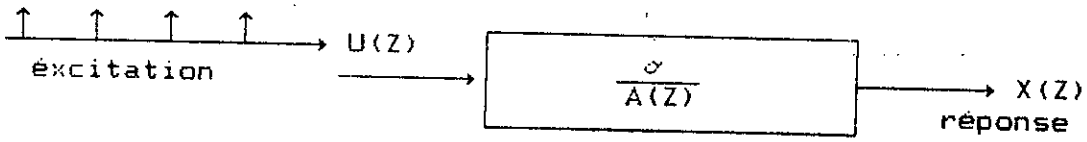


fig-V-1) Modèle autorégréssif

La fonction de transfert de ce modèle est : $\frac{X(Z)}{U(Z)} = \frac{\sigma}{A(Z)}$

avec : $A(Z) = \sum_{i=0}^{p} a(i)\ Z^{-i}$ et $a(0) = 1$

d'où : $\sigma\ U(Z) = X(Z)\ A(Z) = \sum_{i=0}^{p} a(i)\ X(Z)\ Z^{-i}$

soient u(n) et x(n) les transformées en Z inverses respectives de U(Z) et X(Z) tq : $u(n) = \sum_{k=1} \delta(n-k\ P)$

par suite : $\sigma\ u(n) = \sum_{i=0}^{p} a(i)\ x(n-i) = x(n) + \sum_{i=1}^{p} a(i)\ x(n-i)$

nus arrivons alors à :

$$x(n) = \sigma\ u(n) - \sum_{i=1}^{p} a(i)\ x(n-i) \qquad (V-2)$$

Donc chaque échantillon peut être évalué par la différence de l'excitation et des p échantillons qui le précèdent. Les seules inconnues dans cette équation sont le gain σ et les pôles a(i) du modèle .

La fonction de transfert G(Z) du conduit vocal pour un modèle tous-pôles est donnée par :

$$G(Z) = \prod_{k=1}^{\kappa} \frac{A_k}{(1 + b_{k,1} Z^{-1} + b_{k,2} Z^{-2})} \qquad (V-3)$$

où κ est le nombre de cellules de résonnances du conduit vocal .

Nous prenons une seule cellule de résonnance pour chaque formant soit :

$$g_k(Z) = \frac{A_k}{1 + b_{k,1} Z^{-1} + b_{k,2} Z^{-2}}$$

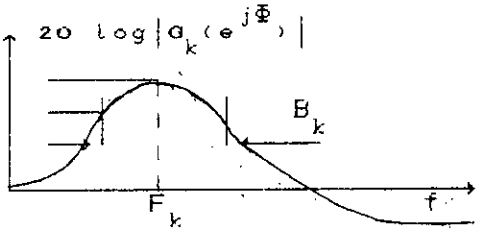


fig V-2) Module de $g_k(Z)$

avec :

$b_{k,2} = 1 - 2\pi\ B_k\ /\ Fe$

$b_{k,1} = -2\ b_{k,2}\ \cos(2\pi\ Fk\ /\ Fe)$

$Ak = 1 + b_{k,1} + b_{k,2}$

Fk : fréquence du $k^{ème}$ formant ;

Bk : bande de fréquence à 3 dB du $k^{ème}$ formant ;

Ak : amplitude (gain du modèle) du $k^{ème}$ formant ;

Fe : fréquence d'échantillonnage ;

$b_{k,1}$ et $b_{k,2}$: pôles de G(Z) .

Remarquons que dans cette méthode , c'est le modèle qui fixe l'amplitude des formants selon leurs fréquences et leurs bandes .

Puisque les deux premiers formants F1 et F2 suffisent pour caractériser nos voyelles , nous n'aurons besoin que de deux cellules de résonnance pour la simulation ,chacune pour un formant ,soit :

$$G(Z) = g_1(Z) + g_2(Z) \qquad (V-4)$$

$$= \frac{A_1 A_2}{(1+b_{1,1}Z^{-1}+b_{1,2}Z^{-2})(1+b_{2,1}Z^{-1}+b_{2,2}Z^{-2})}$$

$b_{1,1}$, $b_{1,2}$, $b_{2,1}$ et $b_{2,2}$ ainsi que A1 et A2 sont donnés par les formules précédentes .

Par identification au modèle autorégréssif , nous obtenons :

σ = A1 A2 ;

$$A(Z) = (1 + b_{1,1} Z^{-1} + b_{1,2} Z^{-2}) (1 + b_{2,1} Z^{-1} + b_{2,2} Z^{-2})$$

$$= 1 + (b_{1,1} + b_{2,1}) Z^{-1} + (b_{1,2} + b_{2,2} + b_{1,1} b_{2,1}) Z^{-2} +$$

$$+ (b_{1,1} b_{2,2} + b_{1,2} b_{2,1}) Z^{-3} + b_{1,2} b_{2,2} Z^{-4}$$

Il en résulte que notre modèle est du quatrième ordre tq:

a(0) = 1 ;

$a(1) = b_{1,1} + b_{2,1}$;

$a(2) = b_{1,2} + b_{2,2} + b_{1,1} b_{2,1}$;

$a(3) = b_{1,1} b_{2,2} + b_{1,2} b_{2,1}$;

$a(4) = b_{2,1} b_{2,2}$.

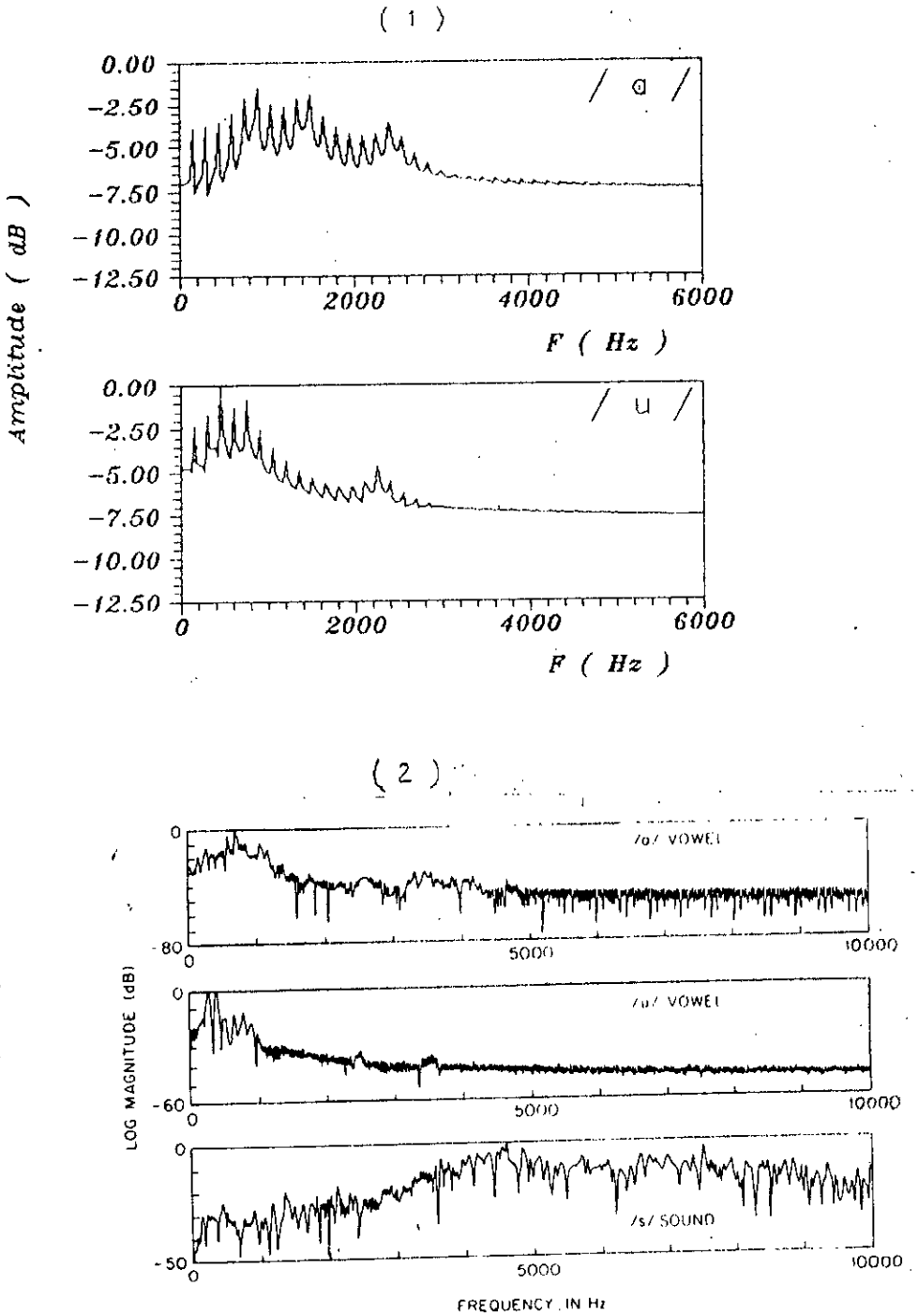


Fig V-3: Formes du spectre des voyelles (a,u)
(1) : generees par le modele AR
(2) : de la parole reelle [12]

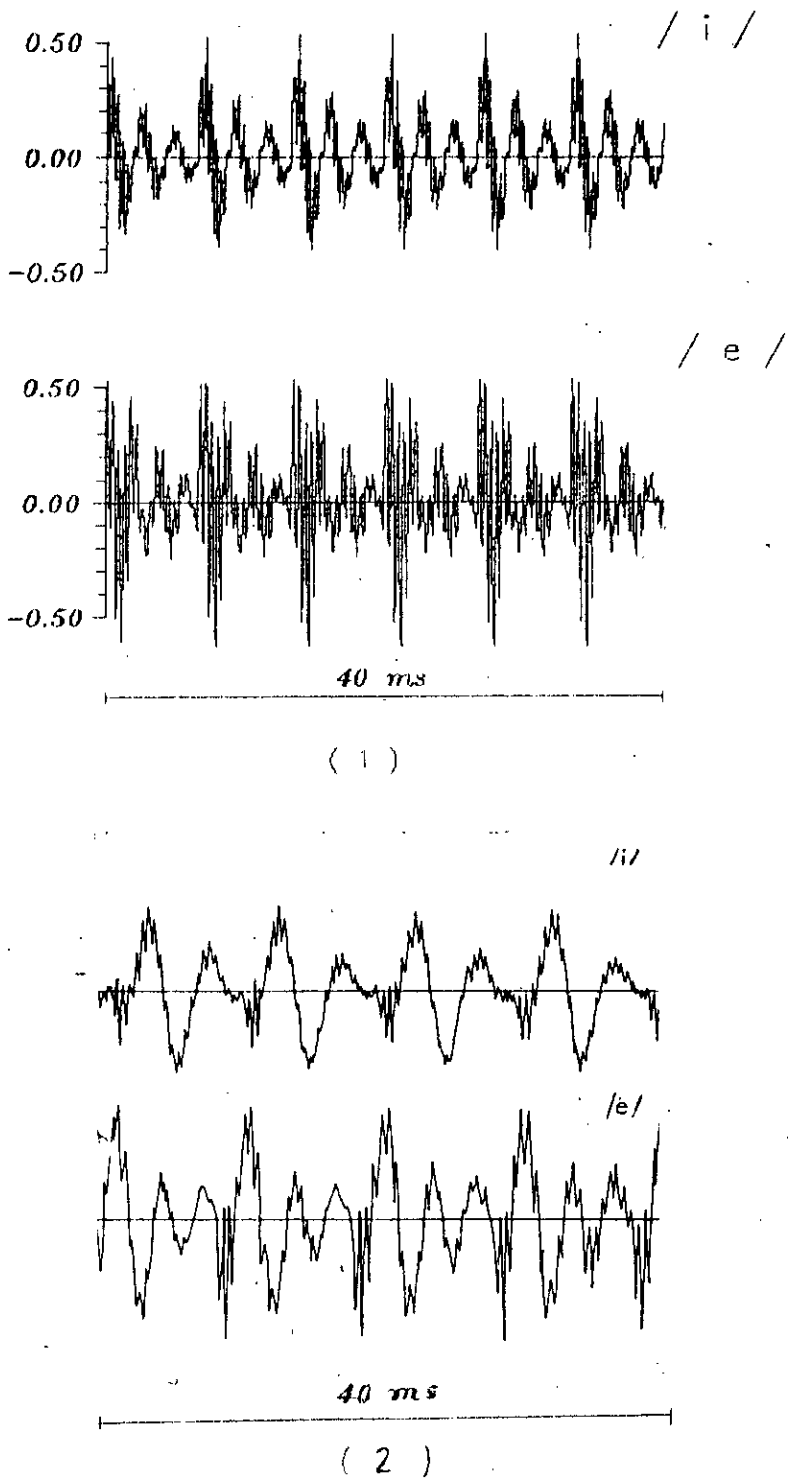


Fig V-4: Formes d'onde acoustique des voyelles (i,e)
(1) : generees par le modele AR
(2) : de la parole reelle [12]

D'où l'équatin de générations de nos échantillons :

$$x(n) = \sigma\, u(n) - \sum_{i=1}^{4} a(i)\, x(n-i) \quad , 0 \le n \le Nech \qquad (V\text{-}5)$$

ou encore

$$x(n) = \begin{cases} \sigma & , \quad n = 0 \\ \sigma\, u(n) - \sum_{i=1}^{4} a(i)\, x(n-1) & , \quad n \le 4 \\ \sigma\, u(n) - \sum_{i=1}^{4} a(i)\, x(n-i) & , \quad 4 \le n \le Nech \end{cases}$$

L'éxcitation u(n) est donnée par :

$$u(n) = \begin{cases} 1 & \text{si } n \text{ est un multiple de P ;} \\ 0 & \text{sinon .} \end{cases}$$

avec :

$$P = \frac{\text{période du pitch}}{\text{période d'échantillonnage}} = \frac{\text{F échantillonnage(Fe)}}{\text{F du pitch (F0)}}$$

V-3 / CONCLUSION :

La première méthode n'est pas retenue car comme nous l'avons déja signalé , elle nous éloigne beaucoup trop de la parole réelle .

Tous nos tests seront alors faits avec la seconde méthode de génération qui nous donne des formes d'onde et des spectres de nos voyeles trés proches de ceux de la parole réelle [12] conformément aux figures V-3 pour /a/ ,/u/ et V-4 pour /e/ ,/i/ .

IV-A / TESTS MONOLOCUTEURS :

Ces tests nous permetterons de tester l'influence de beaucoup de parametres pertinents dans nos analyses sur le taux de reconnaissance . Notre but éssentiel est de choisir par la combinaison des résultats de ces tests la meileure méthode d'analyse et l'algorithme le plus performant pour pouvoir entamer la classification avec le plus petit temps de calcul possible .

--------A-1 / Tests sur l'amplitude :

Les systèmes de reconnaissance de la parole doivent être peu sensibles aux variations énergitiques des signaux . En effet , un locuteur peut parler à plus ou moins haute voix sans cependant changer de timbre . C'est pourquoi , il y a nécéssité d'une normalisation énergitique du signal de la parole avant la comparaison dynamique des mots . Cette normalisation doit être faite dans la phase d'analyse .Par ce fait , les tests que nous faisons sont déstinés à comparer les performances de nos deux méthodes d'analyse (cepstrale et LPC) en sensibilté aux variations d'amplitude .

A-1-a / Données des tests :

--Parametres de générations :

Fpitch = 150 Hz ; Fech = 12800 Hz ; durée = 200 ms ;

Nous générons six fichiers par voyelle où nous prenons des rapports 2 ,3 ,5 ,10 ,20 fois l'amplitude de référence ;les fréquences des formants sont les valeurs moyennes données dans le chapitre V .

--Parametres d'analyse :

Préaccentuation éliminée car donnant des pics indésirables sur le spectre de nos voyelles ;la durée des fenêtres temporelles est 20 ms nombre de coéfficients :--cepstraux (8) ;

--prédicteurs (12).

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 ;

Fenêtre d'ajustement : losange (symétrique et assymétrique);

Distance locale : distance Euclidienne .

A-1-b / Résultats des tests :

Dans le programme de DTW , nous avons choisi seulement les deux algoritmes SAKOE et CHIBA : pente p = 1 , pur ses deux formes où nous calculons le taux de reconnaissance et la distance moyenne pour les voyelles reconnues pour nos deux méthodes d'analyse :LPC et cepstrale (CPS) .

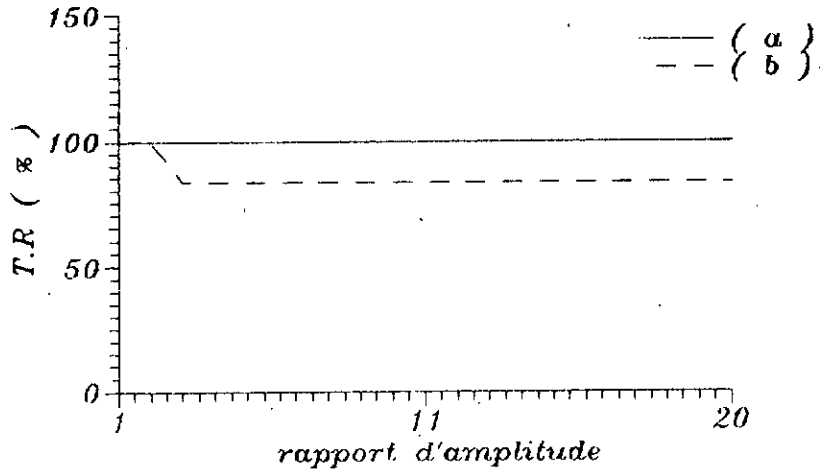
Ce choix est bien justifié car les tests portent sur les méthodes d'analyse que nous avons faites et non sur les algorithmes et les pentes .

Dans le tableau des résultats suivant , ρ est le rapport d'amplitude

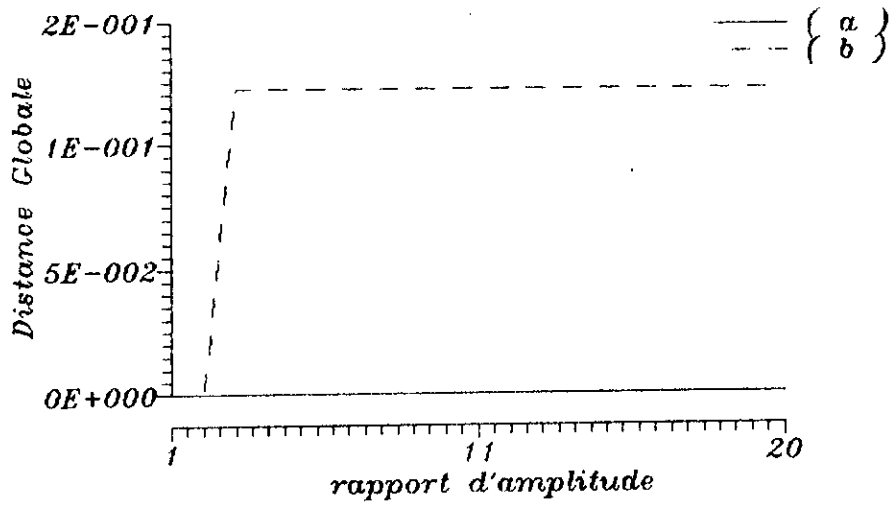
Rapport ρ	Algorithme de SAKOE et CHIBA (p = 1)			
	Forme symétrique		Forme assymétrique	
	CPS	LPC	CPS	LPC
2	100% $9.85346\ 10^{-7}$	100% 0.0000	100% $9.85346\ 10^{-7}$	100% 0.0000
3	100% $2.8054\ 10^{-5}$	83.33% $1.226657\ 10^{-1}$	100% $2.80540\ 10^{-5}$	83.33% $1.22665\ 10^{-1}$
5	100% $9.2976\ 10^{-5}$	83.33% $1.229812\ 10^{-1}$	100% $9.29760\ 10^{-5}$	83.33% $1.52776\ 10^{-1}$
10	100% $9.9244\ 10^{-5}$	83.33% $1.229810\ 10^{-1}$	100% $9.92440\ 10^{-5}$	83.33% $1.52776\ 10^{-1}$
20	100% $9.9590\ 10^{-5}$	83.33% $1.229816\ 10^{-1}$	100% $9.9590\ 10^{-5}$	83.33% $1.52776\ 10^{-1}$

A-1-c / Intèrprétations :

Sur le tableau çi-dessus ou les courbes fig VI-*1-1 et -2* , nous remarquons que l'analyse cepstrale (CPS) est insensible aux variations d'amplitude puisque gardant un taux de reconnaissance de 100% même pour ρ = 20 , de plus la distance globale moyenne se



(1) : Effets sur le taux de reconnaissance



(2) : Effets sur la distance globale

Fig VI-1 : Resultats des tests sur l'amplitude
(a) : Analyse Cepstrale
(b) : Analyse LPC

stabilise autour de $3.3\ 10^{-5}$.

Quant à l'analyse LPC , dés que le rapport ρ dépasse 2 , le taux de reconnaissance diminue et se stabilise à 83.33% , la distance globale moyenne varie entre $1.2\ 10^{-1}$ et $1.5\ 10^{-1}$.

A-1-d / Conclusion :

Nous avons pu vérifier que l'analyse LPC est plus sensible aux variations énergitique que l'analyse cepstrale . Ce résultat est conforme à la théorie puisque cette dernière (CPS) réalise une normalisation en énergie à l'oppose de la première (LPC) .

----------A-2 / Tests sur la durée :

La distance entre deux locutions doit être peu sensible aux petites distorsions des axes temporels . En effet nous savons qu'un même locuteur ne prononce pas toujours un mot donné avec le même rythme . Les algorithes de normalisation temporelle de la distance entre deux mots ont pour but de réaliser cette condition .

Les tests que nous faisons , portent sur tous les algorithmes de comparaison dynamique que nous avons définis dans le chapitre III et ce pour nos deux méthodes d'analyse .

A-2-a / Données des tests :

--*Paramètres de générations :*

Fpitch = 150 Hz ; Fech = 12800 Hz ;

Nous générons 12 fichiers par voyelle où nous prenons des rapports 0.37 ,0.43 ,0.57 ,0.71 ,0.86 ,1.2 ,1.6 ,2 ,2.4 ,2.5 et 2.8 fois la durée de référence (100 ms);les fréquences des formants sont les valeurs moyennes données dans le chapitre V .

--*Paramètres d'analyse :*

Préaccentuation éliminée car donnant des pics indésirables sur le spectre de nos voyelles ;la durée des fenêtres temporelles est 20 ms nombre de coéfficients :--cepstraux (8) ;

--prédicteurs (12).

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 ;

Fenêtre d'ajustement : losange ;

Distance locale : distance Euclidienne .

A-2-b / Résultats des tests :

Dans les programmes de DTW , nous calculons le taux de reconnaissance et la distance globale moyenne pour les voyelles reconnues pour nos deux méthodes d'analyse :LPC et cepstrale (CPS) .

Nous ne donnons pas les tableaux des résultats obtenus car ils sont trés grands et trés denses , nous préférons les remplacer par les courbes çi-aprés qui sont d'ailleurs plus éxplicites .

Le rapport de durée est noté : α

A-2-c / Intèrprétations :

Algorithme SAKOE et CHIBA :

--Courbe (VI-2-9) : p =0 (LPC ,CPS) pour les 2 formes :

Le taux de reconnaissance atteint 100% pour α =0.57 et reste constant jusqu'à α =2 au delà duquel il diminue pour s'annuler à 2.4 pour cette courbe les deux méthodes donnent donc les mêmes résultats.

--Courbes : p= 1/2 ,*CPS (VI-2-3 les 2 formes)*,

LPC (VI-2-3 f-sym- et -2-4 f-assym-) :

Pour la méthode CPS ,le taux de reconnaissance atteint 100% pour α =0.57 et reste constant jusqu'à α =2.4 au delà duquel il diminue pour s'annuler à 2.9 pour les deux formes de la fenêtre d'ajustement. Quant à la méthode LPC ,la forme assymétrique donne de meilleurs résultats mais le rapport de durée reste compris dans le même intervalle que pour la méthode CPS qui est encore la plus performante .

--Courbes : p =1 ,*CPS (VI-2-5 f-sym et -2-6 f-assym)*

LPC (VI-2-7 f-assym et -2-8 f-sym)

Pour la méthode CPS ,le taux de reconnaissance atteint 100% pour

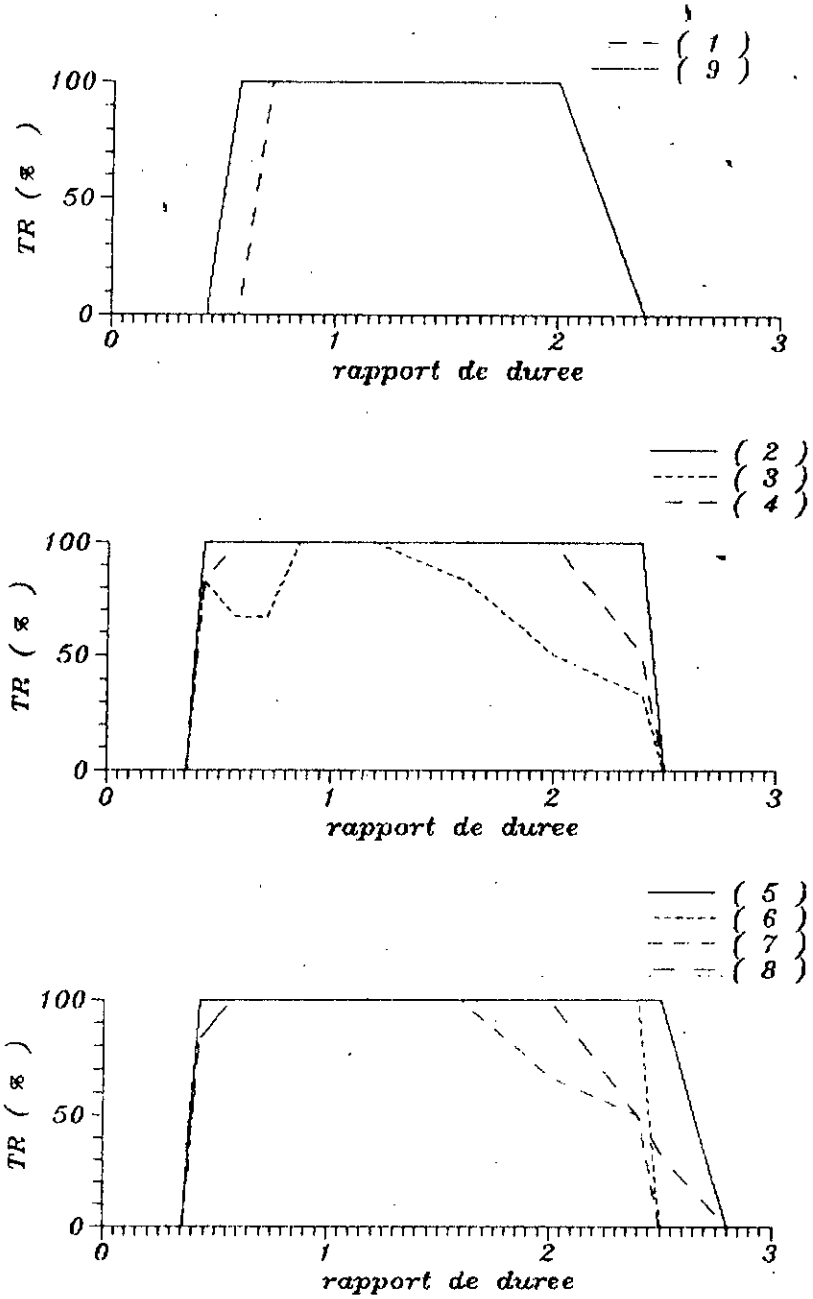


Fig VI-2: Resultats des tests sur la duree pour les algorithmes de SAKOE et CHIBA pente=(0,1/2,1,2), les deux formes pour les deux methodes d'analyse

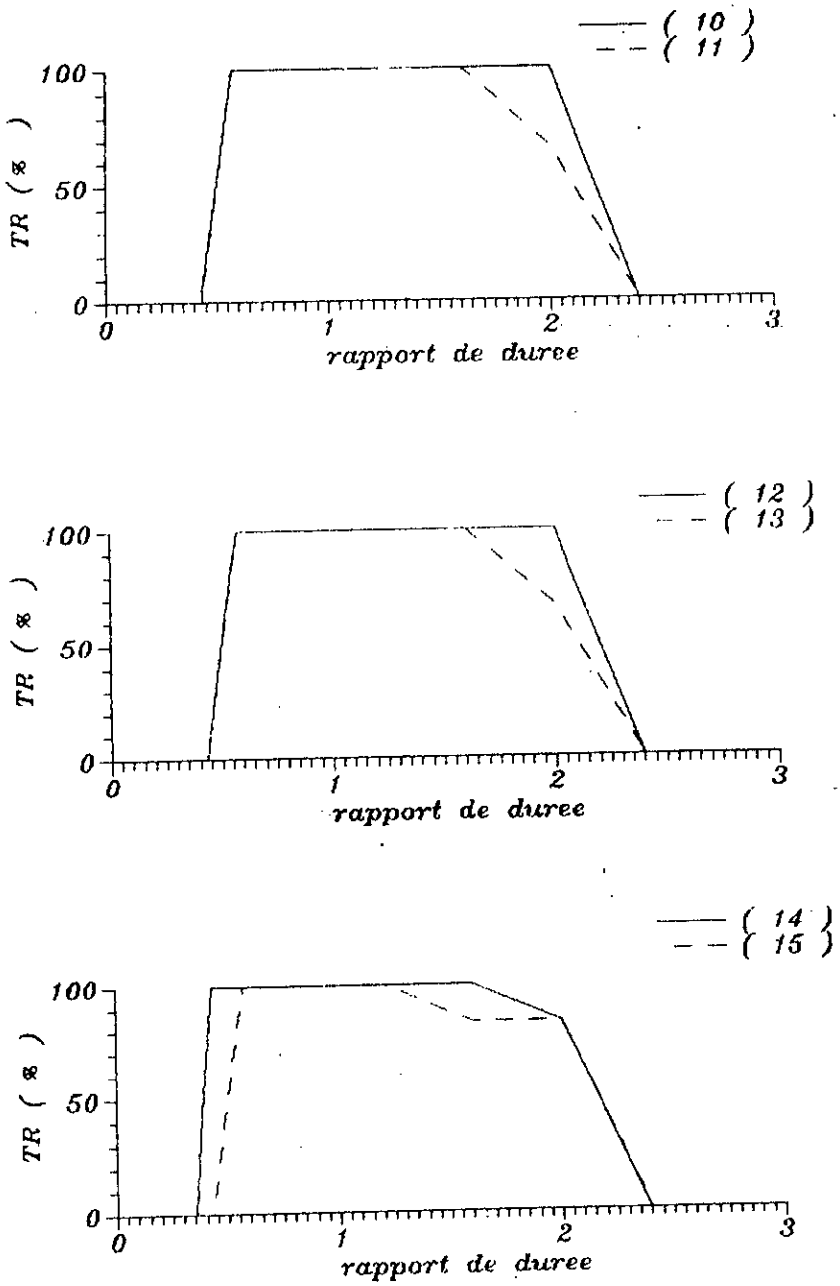


Fig VI-3: Resultats des tests sur la duree pour les algorithmes de : ITAKURA ,WHITE et NEELY VELICHKO et ZAGORUYKO

α =0.43 et reste constant jusqu'à α =2.4 pour la forme assymétrique (nul à 2.5)et jusqu'à α =2.5 pour la forme symétrique (nul à 2.8) . Quant à la méthode LPC ,la forme symétrique donne de meilleurs taux de reconnaissance sur un meilleur intervalle temporel (α =0.43 jusqu'à 2.8) .

--*Courbe (VI-2-1)* : p =2 (LPC ,CPS) pour les 2 formes :

Pour les deux méthodes et les deux formes , le taux de reconnaissance atteint 100% de α =0.7 jusqu'à 2 et s'annule à 2.4 .

Algorithme d'ITAKURA : *Courbes (VI--3-10 et -11)*

La méthode CPS donne de meilleures reconnaissances et sur un meilleur intervalle de temps (de 0.43 à 2.4) ,par contre la méthode LPC , le taux diminue et atteint parfois 66.66% .L'intervalle temporel reste le même .

Algorithme de WHITE et NEELY : *Courbes (VI-3-12 et -13)*

Les deux méthodes donnent des reconnaissances sur un même intervalle de temps (de 0.43 à 2.4) mais la méthode CPS est plus performante .

Algorithme de VELICHKO et ZAGORUYKO : *Courbe (VI-3-14 et -15)*

La méthode CPS est plus performante mais l'intervalle de temps est le même (de 0.32 à 2.4) .

N.B : Le calcul des distances globales moyennes nous a servis içi à rejeter certaines valeurs incohérentes .

A-2-d / Conclusions :

Notre but étant d'obtenir une normalisation temporelle soit des taux de reconnaissance dans le plus grand intervalle de temps possible , c'est pourquoi nous concluons que l'algorithme le plus performant parmi ceux que nous' avons adoptés est celui de SAKOE et CHIBA obtenu pour une pente p =1 avec la forme symétrique et cela pour nos deux méthodes d'analyse .

Cette conclusion est obtenue aprés comparaison de nos courbes pour les différents algorithmes , pentes ,et formes .

----------A-3 / Tests sur les formants :

Nous savons que le spectre du signal de la parole n'a pas de forme fixe puisque le signal lui-même est aléatoire . Ce spectre varie d'une personne à une personne et aussi chez une même personne selon son état . Aussi,nous faisons des tests sur les fréqueuces des formants pour mesurer la sensibilité du système de reconnaissance de la parole à ces variations mais en mode monolocuteur ie chez une même personne .

Les résultats de nos tests nous permetterons d'opter pour une méthode d'analyse et un algorithme de DTW .

A-3-a / Données des tests :

--Paramètres de générations :

Fpitch = 150 Hz ; Fech = 12800 Hz ;

Nous générons cinq fichiers par voyelle où nous prenons les fréquences des formants ,données dans le tableau suivant :

Voyelle et formants		Fichiers				
		XP1	XP2	XP3	XP4	XP5
A	F_1	720	740	750	760	770
	F_2	1280	1900	1950	1400	1420
E	F_1	370	390	400	410	490
	F_2	2120	2190	2200	2210	2280
I	F_1	290	240	250	260	270
	F_2	2450	2470	2500	2590	2550
O	F_1	955	965	975	985	995
	F_2	720	790	750	770	780
U	F_1	290	240	250	260	270
	F_2	570	580	600	620	690
Y	F_1	290	240	250	260	270
	F_2	1750	1770	1800	1830	1850

Notons que XP1....XP5 sont des fichiers générés avec les formants correspondants pour les six voyelles considérées .

--Paramètres d'analyse :

Préaccentuation éliminée car donnant des pics indésirables sur le spectre de nos voyelles ;la durée des fenêtres temporelles est 20 ms sans recouvrement ;

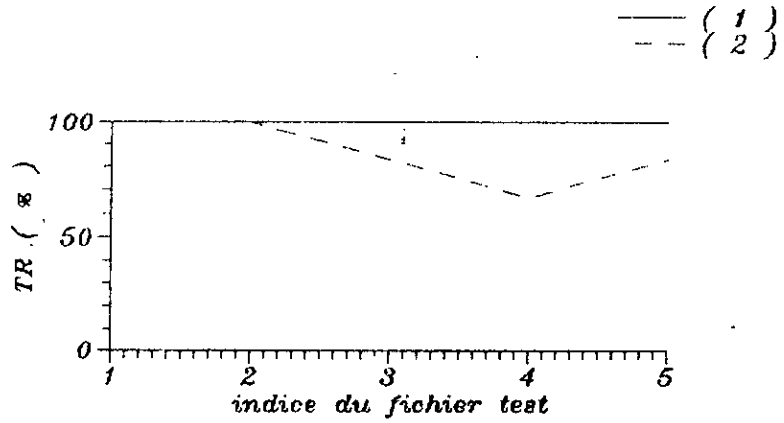


Fig VI-4: Resultats des tests sur les formants pour les algorithmes SAKOE et CHIBA analyse Cepstrale

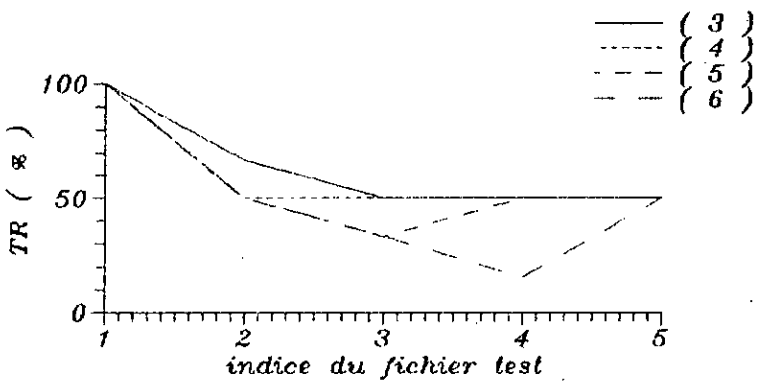


Fig VI-5: Resultats des test sur les formants pour les algorithmes SAKOE et CHIBA analyse LPC

nombre de coéfficients :--cepstraux (8) ;

--prédicteurs (12).

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 ;

Fenêtre d'ajustement : losange ;

Distance locale : distance Euclidienne .

A-3-b / Résultats des tests :

Dans les programmes de DTW , nous calculons le taux de reconnaissance et la distance globale moyenne pour les voyelles reconnues pour nos deux méthodes d'analyse :LPC et cepstrale (CPS) .

Nous ne donnons pas les tableaux des résultats obtenus car ils sont trés grands et trés denses , nous préférons les remplacer aussi par les courbes çi-aprés qui sont plus éxplicites .

A-3-c / Interprétations :

Algorithme SAKOE et CHIBA :

--Courbes VI-4: -1, -2, et VI-5: -3, -4, -5, -6) :

- Méthode CPS

Le taux de reconnaissance et de 100% pour les cinq fichiers sauf pour la pente p =1/2 forme assymétrique qui donne des taux de 83.33% et 66.67% .

- Méthode LPC

Le taux de reconnaissance ne dépasse pas 66.67% et atteint même 15.33% pour la pente p =1/2 forme assymétrique ; pour les autres pentes , ce taux est souvent de 50% pour les deux formes .

Les autres algorithmes :

--Courbes VI-6: -7 ITAK, -8 WelN, -9 VelZ) : Méthode CPS

Le taux de reconnaissance est de 100% pour l'algorithme d'ITAKURA pour les cinq fichiers mais pour les algorithmes de WHITE et NEELY et de VELICHKO et ZAGORUYKO ce taux atteint parfois 83.33% .

--Courbe (VI-7 : -10 ITAK , -11 WelN , -12VelZ) : Méthode LPC

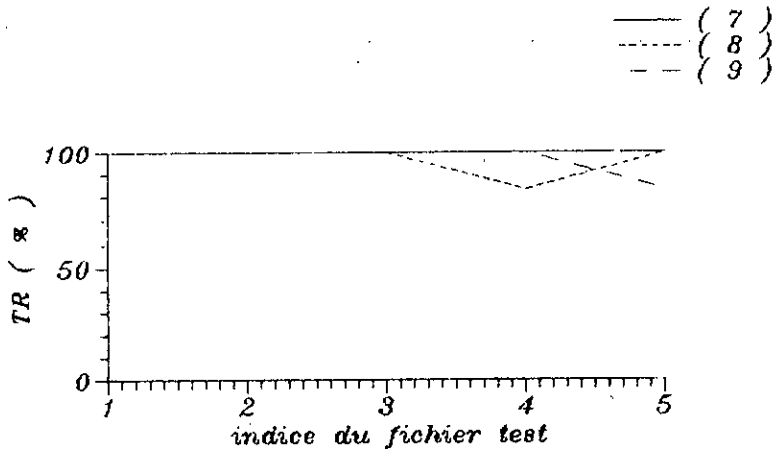


Fig VI-6: Resultats des tests sur les formants pour les algorithmes de ITAKURA ,WHITE et NEELY VELICHKO et ZAGORUYKO analyse Cepstrale

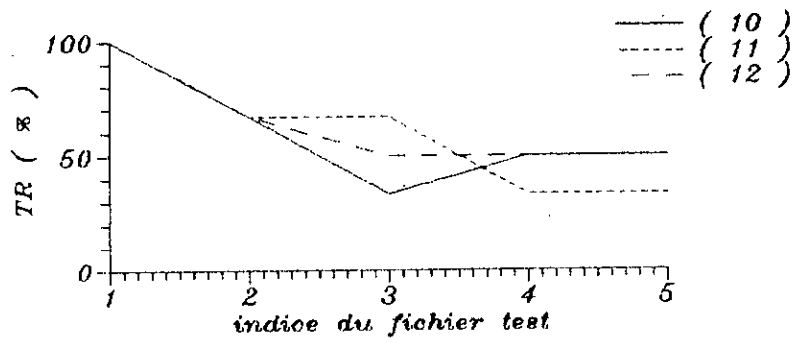


Fig VI-7: Resultats des tests sur les formants pour les algorithmes de ITAKURA ,WHITE et NEELY VELICHKO et ZAGORUYKO analyse LPC

Pour cette méthode ,l'algorithme de WHITE et NEELY donne des taux de reconnaissance relativement meilleurs sans atteindre toutefois 100% , le plus faible taux est obtenu avec l'algorithme d'ITAKURA (33.33%) .

A-3-d / Conclusions :

D'aprés nos courbes , tous les algorithmes DTW ,sauf celui de SAKOE et CHIBA pour la fore assymétrique et p =1/2 , donnent de trés bons taux de reconnaissance pour la méthode cepstrale et des mauvais taux pour la méthode LPC (autour de 50%) . Par conséquent nous ne pouvons pas ,dans ce test , déceler le plus performant des algorithmes que nous avons utilisés . Par contre ,nous constatons tout de même que la méthode LPC est plus sensible aux variations des formants que la méthode cepstrale . Cela est dû réellement à l'utilisation dans cette dernière d'une analyse spectrale et des filtres MEL qui prennent en considération les déplacements fréquentiels qui sont d'ailleurs inférieurs à la largeur des filtres que nous avons pris (150 Hz) .En effet si ces filtres contiènnent les déplacements des formants , l'énergie du signal dans la bande de ces filtre est trés peu sensible à ces variations

----------*A-4* **/ Tests sur le nombre des coéfficients :**

Il a deja été démontré [13] que le taux de reconnaissance de la parole dépend étroitement du nombre de paramètres pertinents d'analyse qui renseignent le système de reconnaissance sur la nature des mots .C'est pourquoi , nous faisons des tests pour évaluer, dans notre cas ,le nombre optimal de coéfficients cepstraux et LPC .

A-4-a / Données des tests :

--*Paramètres de générations :*

Fpitch = 150 Hz ; Fech = 12800 Hz ; durée = 200 ms ;

Nous générons six fichiers par voyelle où nous prenons les différentes fréquences des formants données dans le test précédent

nous répétons cette opération cinq fois pour la méthode cepstrale en prenant successivement comme nombre Nc de coéfficients :6 ,8 ,10 ,13 et 18 .Pour la méthode LPC , les six fichiers sont regénérés avec les nombres Np de coéfficients suivants :10 ,12 ,15 , 18 et 24 .

--Paramètres d'analyse :

Préaccentuation éliminée car donnant des pics indésirables sur le spectre de nos voyelles ;la durée des fenêtres temporelles est 20 ms

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 pour chaque valeur de Np ou Nc;

Fenêtre d'ajustement : losange ;

Distance locale : distance Euclidienne .

A-4-b / Résultats des tests :

Dans le programme de DTW , nous avons choisi seulement un des algoritmes SAKOE et CHIBA (pente p = 1 , forme symétrique) où nous calculons le taux de reconnaissance et la distance globale moyenne pour les voyelles reconnues .

Nos résultats sont regroupés dans les deux tableaux suivants:

Nbre de Coéff. Nc	ANALYSE CEPSTRALE			
	Reconnaissance par/ au fichier1 ≡ fichier-référence			
	fichier2	fichier3	fichier4	fichier5
6	100% $2.28973\ 10^{-1}$	100% $3.415416\ 10^{-1}$	33.33% 2.048710	100% $6.1717706\ 10^{-1}$
8	100% $2.758098\ 10^{-1}$	100% $4.136498\ 10^{-1}$	100% $5.875295\ 10^{-5}$	100% $7.0797\ 44\ 10^{-1}$
10	100% $2.966696\ 10^{-1}$	100% $4.4919864\ 10^{-1}$	100% $6.955885\ 10^{-1}$	100% $7.898694\ 10^{-1}$
13	100% $3.701723\ 10^{-1}$	100% $5.9888941\ 10^{-1}$	100% $8.178472\ 10^{-1}$	100% $9.782918\ 10^{-1}$
18	100% $4.606299\ 10^{-1}$	83.33% $7.3022007\ 10^{-1}$	100% $9.650166\ 10^{-1}$	83.33% 1.1859969

Nbre de Coéff. Np	ANALYSE LPC			
	Reconnaissance par/ au fichier1 ≡ fichier-référence			
	fichier2	fichier3	fichier4	fichier5
10	50% $6.36619\ 10^{-1}$	50% $9.892454\ 10^{-1}$	50% 1.1813231	50% 1.3667058
12	66.66% $6.683981\ 10^{-1}$	50% 1.0469720	50% 1.1406252	50% 1.2411747
15	66.66% $8.905961\ 10^{-1}$	50% 1.2422756	33.33% $9.955885\ 10^{-1}$	50% $1.499187\ 10^{-1}$
18	66.66% $9.9059612\ 10^{-1}$	50% 1.3419981	50% 1.242818	33.33% 1.5475521
24	50% $8.124924\ 10^{-1}$	50% $3.2568947\ 10^{-1}$	50% $5.326411\ 10^{-1}$	33.33% 1.4562839

A-4-c / Intèrprétations :

Sur nos deux tableaux , nous remarquons que l'analyse cepstrale donne toujours de meilleures reconnaissances ,le taux est de 100% pour toutes les valeurs du nombre de coéfficients sauf pour Nc =6 où il atteint 33.33% . Quant à la méthode LPC , les reconnaissanses obtenues ne dépassent pas moyennement 50% pour toutes les valeurs du Np ,le taux n'atteint 66.66% que trés peu de fois .

A-4-d / Conclusion :

Si nous nous basons sur la comparaison des distances globales moyennes données dans nos tableaux , nous pouvons constater que le nombre optimal de coéfficients est 8 pour l'analyse cepstrale et 12 pour l'analyse LPC . Cependant ce critère n'est pas trés justifié puisque la comparaison devrait être faite sur le taux de reconnaissance lui-méme surtout que nous avons plusieurs expréssions pour les distances locales qui changeraient les resultats des distances plobales moyennes .

Nous éstimons que cela est dû au fait que nous ne travaillons pas sur de la parole réelle puisque nous ne faisons que la simuler .

----------A-5 / Tests sur la fenêtre d'ajustement :

Nous avons déja précisé que pour optimiser les calculs dans la comparaison dynamique des mots ,le champ de recherche de la fonction de deformation est restreint à la fenêtre d'ajustement qui peut être en forme de losange avec une largeur constante ,ou de parallepipède de largeur 'r 'variable .

Pour trouver la fenêtre adéquate [14], nous faisons des tets en lui changeant de forme et de largeur . Pour ce ,nous avons préféré procéder avec la méthode LPC puisque la cepstrale donne des taux de reconnaissance pas trop différents pour pouvoir tirer des conclusions . Nous ne donnons alors que les résultats des tests obtenus avec la méthode LPC

A-5-a / Données des tests :

--Paramètres de générations :

Fpitch = 150 Hz ; Fech = 12800 Hz ;

Nous générons quatre fichiers par voyelle où nous changeons les fréquences des formants .Nous répétons cette opération pour la fenêtre parallelipipédique pour les largeurs suivantes :2 ,4 ,6 .

--Paramètres d'analyse :

Préaccentuation éliminée ;la durée des fenêtres temporelles est 20ms

nombre de coëfficients :--cepstraux (8) ;

--prédicteurs (12).

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 ;

Forme symétrique ;

Distance locale : distance Euclidienne .

A-5-b / Résultats des tests :

Dans les programmes de DTW , nous faisons les même calculs que pour les tests précédents avec la méthode LPC et les algorithmes de SAKOE et CHIBA

La largeur de la fenêtre est notée : r

Algorithme Sakoe et Chiba (forme symétrique)					
Fenêtre		Parallépipede			Losange
Pente	Fichiers	r =2	r =4	r =6	
p =0	Fich2	66.66% 0.62903291	66.66% 0.59806692	66.66% 0.59806692	66.66% 0.6119034
	Fich3	50% 0.681374	50% 0.7022228	50% 0.58954240	50% 0.75918193
	Fich4	50% 0.8720877	50% 0.86352628	50% 0.66772007	50% 0.88659483
p=1/2	Fich2	66.66% 0.9817649	50% 0.85181671	50% 0.21527691	66.66% 0.7766248
	Fich3	50% 1.1469481	33.33% 0.81111449	33.33% 0.3492508	50% 1.0252062
	Fich4	50% 1.1496208	50% 0.99577087	50% 0.48715198	50% 1.1613997
p =1	Fich2	66.66% 0.8731417	50% 0.68131981	50% 0.6813181	50% 0.68131981
	Fich3	50% 0.99650651	50% 0.94095116	50% 0.94095116	50% 0.94095116
	Fich4	50% 1.1024755	50% 1.0175938	50% 1.0175938	50% 1.0175938
p =2	Fich2	66.66% 1.3127420	66.66% 1.3127420	66.66% 1.31274202	66.66% 1.3127420
	Fich3	50% 1.5507787	50% 1.5507787	50% 1.55077870	50% 1.5507787
	Fich4	50% 1.2444672	50% 1.24446728	50% 1.24446728	50% 1.2444672

A-5-c / Intèrprétations :

Nous remarquons qu'en élargissant la largeur de la fenêtre parallelipédique , le taux de reconnaissance diminue notament pour les pentes p =0 et P =1 qui donnent ,quand r =2 ,les mêmes taux qu'avec la fenêtre losange . La seule différence entre ces deux fenêtres n'est perçue que pour la pente p =1 où nous trouvos moins de reconnaissance avec la deuxième forme (losange) .

A-5-d / Conclusions :

Nos tests nous prouvent qu'en réduisant l'éspace de recherche de la fonction de déformation dans la comparaison dynamique des mots, le taux de reconnaissance augmente .Cela est évident puisque si des

locutions contiennent les mêmes quantités d'informations phonétiques les distances qui les séparent sont trés réduites .Par suite , nous avons tout intérêt à contraindre notre fonction de déformation à ne pas s'éloigner de la diagonale du plan où sont développées nos locutions.

Par ailleurs , nous concluons que la fenêtre losange est moins performante pour nos tests de reconnaissance qu'une fenêtre parallélépipédique de largeur r =2 .

----------A-6 / Tests sur les distances locales :

Dans le but d'affiner les résultats de nos tests ,nous allons appliquer plusieurs formules pour le calcul des distances locales avec les algorithmes de Sakoe et Chiba ,forme symétrique avec a méthode cepstrae puisque ce sont ces derniers qui nous ont donné les meilleurs reconnaissances .

A-6-a / Données des tests :

--Parametres de générations :

Fpitch = 150 Hz ; Fech = 12800 Hz ;

Nous génerons quatre fichiers par voyelle où nous changeons les fréquences des formants .Nous répétons cette opération pour chaque formue de la distance locale .

--Paramètres d'analyse :

Préaccentuation éliminée ;la durée des fenêtres temporelles est 20ms nombre de coéfficients :--cepstraux (8) ;

--Paramètres de DTW :

Nombre de fichier dictionnaire par voyelle : 1 pour chaque formule de la distance locale ;

Forme symétrique ;

A-6-b / Résultats des tests :

Dans le tableau suivant , nous présentons les même calculs que

pour les tests précédents ,la distance locale est notée Dloc .

Algorithme Sakoe et Chiba (forme symétrique)				
Pente	Dloc	Fichiers-test (fichier-référence :Fich1)		
		Fich2	Fich3	Fich4
p=0	Euclidienne	66.66% 0.62963231	50% 0.68137460	50% 0.87208771
	Tchebytchef	66.66% 0.16776743	50% 0.18801691	50% 0.22691955
	Cepstrale	66.66% 0.17091140	50% 0.20298499	50% 0.25921174
	Cepstrale pondérée	66.66% 3.717994	50% 4.662708	50% 4.845945
p=1/2	Euclidienne	66.66% 0.91867645	50% 1.146948160	50% 1.1496208
	Tchebytchef	66.66% 0.23714978	50% 0.30109247	50% 0.3071036
	Cepstrale	66.66% 0.2905199	33.33% 0.30557212	50% 0.49028622
	Cepstrale pondérée	66.66% 4.896574	33.33% 6.474652	50% 5.984176
p=1	Euclidienne	50% 0.68131981	50% 0.94095116	50% 1.0175098
	Tchebytchef	50% 0.19686461	50% 0.28803709	50% 0.30768615
	Cepstrale	50% 0.2853900	50% 0.37274703	50% 0.38400127
	Cepstrale pondérée	66.66% 5.378224	50% 6.904998	50% 5.675147
p=2	Euclidienne	66.66% 1.3127420	50% 1.5507787	50% 1.2444672
	Tchebytchef	50% 0.49295525	66.66% 0697075967	33.33% 0.364789742
	Cepstrale	66.66% 1.0412498	50% 0.4228728	50% 0.42513751
	Cepstrale pondérée	66.66% 8.791913	66.66% 7.094632	50% 6.670096

A-6-c / Interprétations :

Nous remarquons que les taux de reconnaissance sont presque les mêmes pour toutes les distances locales considérées . Néaumoins pour les pentes p=1 et p=2 ,la distance cepstrale pondérée semble donner en moyenne c-à-d sur les trois fichiers (trois variations des

formants) de meilleurs résultats que les autres .

Notons que le calcul des distances globales ne nous sert ici qu'à vérifier s'il y a des rejets .

A-6-d / Conclusions :

Malgré que nous ne travaillons pas sur de la parole réelle , nous avons pu approcher les résultats des travaux menés par Mr.S.Furui [15] qui a démontré que la distance locale cepstrale pondérée permet de meilleurs reconnaissances .

---------- A-7 / Conclusions générales des tests monolocuteurs :

La combinaison de nos résultats nous permettent de dire que :

--la meilleure méthode d'analyse dans la reconnaissance de la parole,dont nous ne faisons qu'une simulation ,est la méthode cepstrale avec un nombre de coéfficients égal à 8;

--si nous utilisons une fenêtre d'ajustement parallépipédique ,sa largeur doit être trés réduite ;

--la distance locale à utiliser est la distance cepstrale pondérée ;

--le meilleur algorithme pour la DTW est celui de Sakoe et Chiba forme symétrique pour les pentes p=1 et p=1/2 .

Par ailleurs , nous devons prendre en considération que la comparaison dynamique des mots est mieux accomplie dans une fenêtre losange car la fonction de déformation y est contrainte à arriver au point final (Imax,Jmax) des locutions considérées à l'oppoosé de la forme parallépipédique où la comparaison s'achève quand Imax ou Jmax est atteint (selon leur ordre de grandeur); de plus si nous comparons les deux pentes que nous avons retenues ,la pente p=1 nous a offert un meilleur temps de calcul et car elle éxige beaucoup moins d'opérations dans la comparaison dynamique .

Nous choisissons alors pour l'étape de classification la méthode d'analyse cepstrale avec pour la DTW ,l'algorithme de Sakoe et Chiba forme symétritrique avec la pente p =1 dans une fenêtre d'ajustement

de forme losange .

IV-B / RESULTATS DE LA CLASSIFICATION :

B-1 / INTRODUCTION :

La répartition des locutions d'un mot en différentes classes n'est pas aisée du tout à cause de la forte variabilité de prononciation interlocuteurs . Certaines fois ,il faut attendre plusieurs itérations pour espérer trouver une bonne classification surtout que nous devons ,à chaque fois , faire des fusions et/ou éclatements de classes pour une meilleure valeur de la qualité de mesure ($\alpha \geq 2$) .

En pratique , les chercheurs travaillant sur la parole réelle préparent des centaines de locutions d'un même mot du vocabulaire du système de reconnaissance . Dans notre cas et pour essayer d'approcher le plus possible la parole réelle ,nous nous sommes basées pour la générations (simulation) des références multilocuteurs sur le graphe de distributions des voyelles (formant 2 en fonction du formant ---*fig VI-8*) donné par Shafer et Rabinner (14).

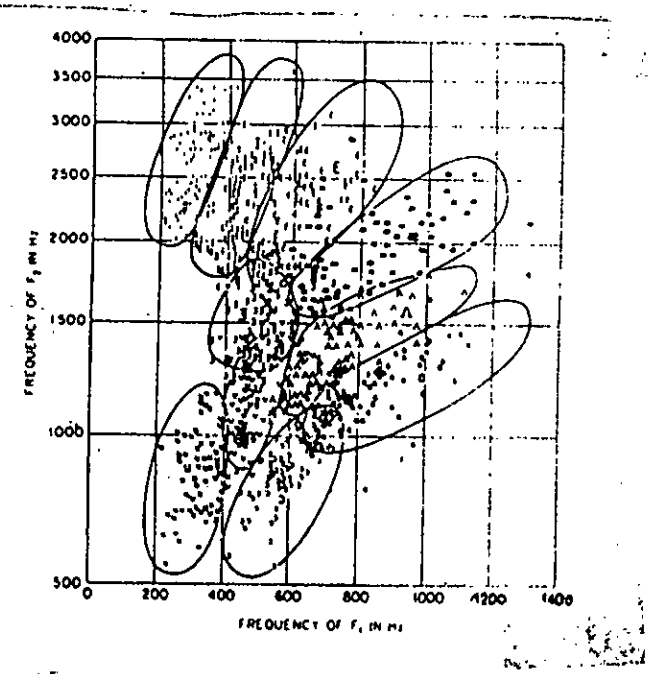


fig VI-8 formant F2 en f(F1) pour plusieurs locuteurs .

Nous faisons alors des variations des formants et du pitch mais

dans un nombre limité de locutions (26 pour chaque voyelle) faute de temps et de la panne du Micro Vax .

B-2 / GROUPEMENT EN CHAINE :

La méthode "CHAINMAP" nous permet de représenter graphiquement la distribution des différentes locutions d'un mot .Les pics obtenus sur nos graphes d'ordonnancement (voir *fig VI-9*) permettent de de délimiter les classes de locutions . Notons que c'est à cause de l'aspect 'simulation' de la parole que les limites des classes ne sont pas trés nettes . Néaumoins , nous pouvons compter le nombre de locutions "hors-communs" ou "outliers" et éstimer plus ou moins le nombre de classes .

B-3 /CLASSIFICATION :

Les fichiers de chaque voyelle sont numérotés de 1 à 26 .Pour une bonne classification ,nous donnons dans les tableaux de résultats la distance moyenne maximale intraclasse notée D_{Cmax} ,la distance minimale entre les prototypes notée D_{cmin} , le facteur de qualité de mesure σ , le nombre Nc de classes ,ainsi que la dimension Dim de chaque classe et dans le même ordre le numéro N°p de sa locution prototype.

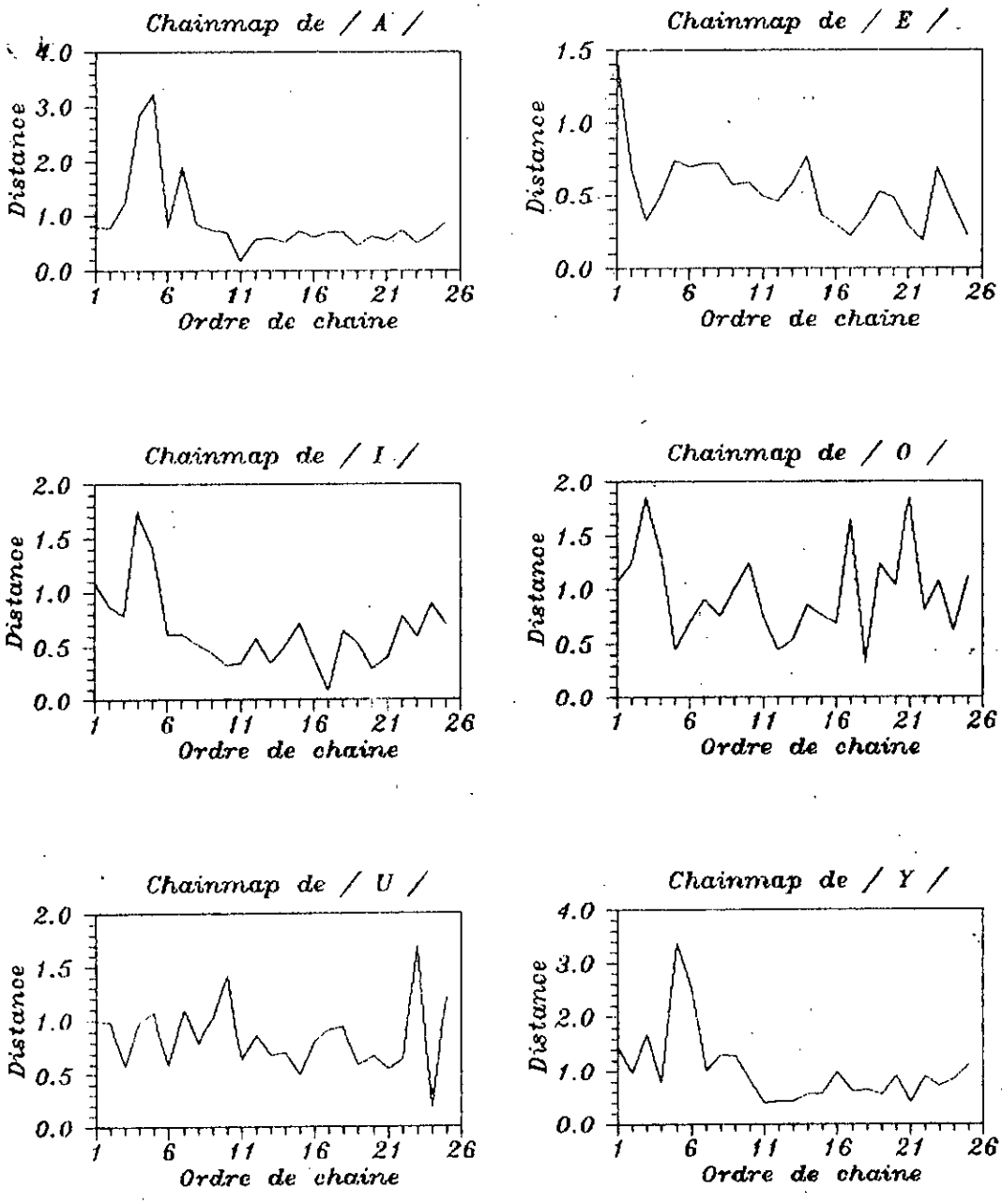


Fig VI-9: Graphes des Groupements en chaine des voyelles (a,e,i,o,u,y) ,avec 26 locutions pour chaque voyelle

VOYELLE "A"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	11 ; 15	29 ; 01	3.5281	1.3023	3.5850
3	7 ;11;08	21;13 ;01	2.3802	1.0997	1.5452
4	7 ;9 ;4 ;6	29;17;01;10	2.5744	0.8965	0.9653
5	5;6;9;10;2	29;17;01;15;26	2.6996	0.9657	0.6263
6	4;5;9;3;2;9	24;16;1;18 26;10	2.5601	0.7717	0.7960
7	4;5;9;3; 2;7;2	24;16;01;18 26;10;09	2.4035	0.7717	0.1675
8	4;5;3;3; 1;7;2;1	24;16;01;18 26;10;09;25	3.4206	0.7321	0.1675

Interprétations :

Nous remarquons qu'une bonne répartition des échantillons de la voyelle"A" n'est obtenue qu'avec 8 classes avec une qualité de mesure σ =3.42 et l'apparition de 2 locutions "outliers" (hors-communs) .

VOYELLE "E"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	11 ; 15	26 ; 05	2.4733	0.7052	1.4568
3	5 ;18;03	25;15 ;26	1.8220	0.5811	0.6881
4	4 ;14;3 ;5	25;01;26;17	2.7443	0.9480	0.6881
5	6;14;3;5;1	25;01;26;17;20	3.2093	0.5480	0.6327
6	3;3;1;4;14;1	26;25;20;15 01;21	3.6241	0.4691	0.6327
7	3;3;3;1; 1;14;1	18;26;25;21 20;01;17	4.1881	0.4585	0.5568

Interprétations :

Il faut au moins 5 classes (σ =3.20) pour pouvoir bien répartir

les échantillons de la voyelle "E"; encore mieux avec 7 classes car donnant 3 "outliers" et σ =4.2 .

VOYELLE "U"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	18 ; 12	12 ; 24	1.7252	0.7767	1.4926
3	6 ;3 ;17	2 ;29 ;19	1.7057	0.7316	0.8095
4	5 ;3 ;9 ;9	22;29;5;15	1.6921	0.7081	0.7787
5	4;2;9;2;9	25;29;5;26;17	1.7835	0.7845	0.6971
6	5;2;5;6;2;6	25;29;9;15 26;17	1.9616	0.6637	0.6188
7	2;2;4;5 5;1;7	26;29;25;17 10;18;05	2.4933	0.6201	0.6971
8	2;1;4;8 5;1;4;1	26;29;25;17 10;18;9;20	3.2392	0.6131	0.6620

Interprétatins :

La qualité de mesure n'est bonne qu'au delà de Nc=6 (σ <2) et il faut 8 classes pour trouver un très bon σ (3.23) et faire apparaitre 3 locutions "outliers" .

VOYELLE "I"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	14 ; 12	24 ; 10	2.0326	0.6365	1.1681
3	2;12;12	26;01 ;20	1.9330	0.8654	1.2154
4	1;4 ;10;11	25;25;20;01	2.7269	0.9601	0.8054
5	8;1;1;08;08	20;26;25;17;01	3.9578	0.5896	0.8654
6	09;1;1;9;9;9	24;26;25;01 22;14	3.4110	0.6428	0.5476

Interprétation :

Il faut 5 classes pour avoir la plus bonne configuration des

échantillons avec σ =3.95 et l'apparition de 2 locutions "outliers".

VOYELLE "O"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	19 ; 19	22 ; 01	2.0026	0.9508	1.9033
3	10;11;05	22;01 ;19	2.1008	0.9005	1.1357
4	2;11;15;8	26;01;19;22	2.1298	0.8845	1.0701
5	2;5;11;5;9	26;19;01;19;22	2.1768	0.8821	0.8644
6	2;5;4;4;9;8	26;21;01;09 25;09	2.1532	0.6328	0.7227
7	2;5;4;5; 9;6;1	26;21;12;19 22;11;01	2.3347	0.6623	0.7369
8	2;5;4;4 9;5;1;2	26;21;12;09 25;15;01;09	2.4779	0.6563	0.6164
9	2;5;4;4;5 1;1;9;9	26;21;12;09 25;15;1;11;9	2.5885	0.6098	0.4462

Interprétation :

Il est difficile d'obtenir une bonne répartition des échantillons de la voyelle "O",même avec 9 classe la qualité de mesure n'atteint pas 3 .

VOYELLE "Y"					
Nc	Dim	N°p	σ	Dicmax	Dicmin
2	16 : 10	01 ; 22	2.3977	1.4091	3.0323
3	14;06;06	01;26 ;10	2.4799	1.2597	1.9989
4	12;6 ;5 ;3	01;17;22;26	2.4573	1.1450	1.4456
5	2;4;5;09;12	01;17;22;26;12	2.6268	0.8855	1.4456
6	10;3;3;3;4;3	01;19;22;25 12;15	2.6802	1.0449	0.9785
7	9;3;3;1; 11;3;2	24;16;01;18 26;10;09	3.3777	0.8855	0.9667

Interprétations :

Il faut arriver à 7 classes pour avoir une très bonne valeur de σ (3.37) où apparait 1 locution "outlier".

B-4 / CONCLUSOINS :

Nos résultats attestent que nous avons bien appliqué les méthodes de classification à nos fichiers . Leur nombre réduit (inssufisance d'espace mémoire) , n'a pas été un inconvénient majeur pour prouver que la qualité de mesure (σ) augmente avec le nombre de classes et que les meilleures répartitions sont obtenues avec des valeurs de σ proches de 3 . Par conséquent, nous pouvons entamer les tests multilocuteurs sans difficulté .

En résumé , les meilleures configurations obtenues sont les suivantes :

--Voyelle "A" : Nc =8 classes , σ =3.42 ;
--Voyelle "E" : Nc =7 classes , σ =4.18 ;
--Voyelle "I" : Nc =5 classes , σ =3.41;
--Voyelle "O" : Nc =9 classes , σ =2.56 ;
--Voyelle "U" : Nc =8 classes , σ =3.03 ;
--Voyelle "Y" : Nc =7 classes , σ =3.37 ;

VI-C / TESTS MULTILOCUTEURS .

Les tests multiocuteurs nous permetteront de finaliser notre travail et de vérifier si les taux de reconnaissances éstompés en théorie s'accordent avec la réalité pratique de notre système .

Notons encore une fois que les contraintes de temps de calcul et de disponibilité des micros ne nous ont pas permis de multiplier nos tests pour tirer plus de conclusions sur les performances du système.

C-1 / TESTS :

Les tets , que nous faisons , doivent nous renseigner sur les

valeurs optimales du nombre "K" des techniques KNN dans l'étape de décision . Les taux que nous donnons dans les tableaux suivants sont les moyennes de plusieurs tests c-à-d le nombre de tests positifs sur le nombre total de tests .

Test 1) Les mots tests appartiènnent au dictionnaire (M.T.A.D):

Nombre de classes	Taux de reconnaissance			
	Valeur de K			
	1	2	3	4
2	100%	83.33%	------	------
3	100%	94.44%	88.89%	------
4	100%	100%	91.67%	87.50%
5	100%	96.67%	90%	86.67%
6	100%	100%	94.44%	94.44%

interprétations : courbe VI-10 (MTAD)

Les taux obtenus pour k =1 sont évidents puisque les distances globales calculées sont nulles ; pour les autres valeurs de K ,les taux sont toujours très bons et augmentent d'autant plus que le nombre de classes augmente .Ceci est attendu parceque les configurations des échantillons deviennent plus meilleures .

L'ordre élevé de nos résultats est dû au fait que les mots test sont comparés à eux mêmes .

Test 2) Les mots tests ont pris part à l'apprentissage (M.T.P.A) :

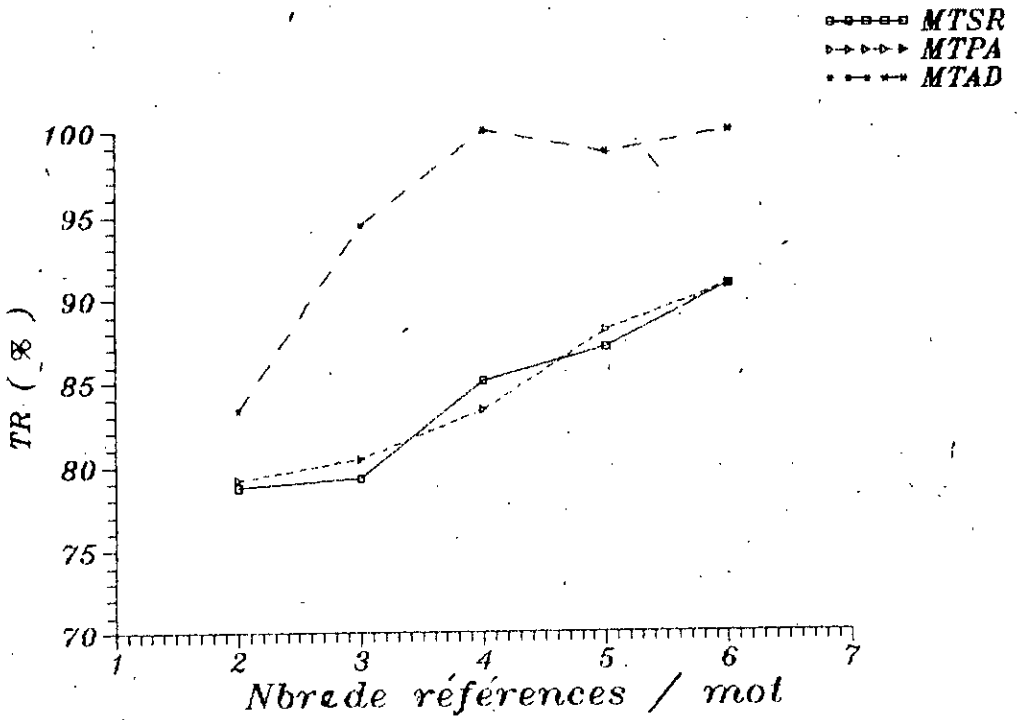


Fig VI-10: Influence du nombre de références par mot sur le taux de reconnaissance pour KNN=2

Nombre de classes	Taux de reconnaissance			
	Valeur de K			
	1	2	3	4
2	84.72%	79.17%	------	-----
3	77.54%	80.43%	79.91%	-----
4	89.33%	89.99%	79.55%	68.18%
5	86.51%	88.10%	85.71%	80.95%
6	92.50%	90.89%	86.67%	85.60%

Interprétation : courbe VI-10 (MTPA)

Les taux obtenus sont toujours satisfaisants et augmentent suivant le nombre de classes vû leur appartenance à ces mêmes classes . La valeur ptimale de K est 2 à partir de 3 classes .

.Test-3-) Les mots tests (M.T.S.R) n'ont pas pris part à l'apprentissage :

Nombre de classes	Taux de reconnaissance			
	Valeur de K			
	1	2	3	4
2	89.99%	78.74%	------	-----
3	78161%	79.91%	79.56%	-----
4	89.91%	85.06%	80.46%	70.11%
5	86.21%	87.06%	84.48%	81.09%
6	90.80%	90.80%	86.78%	86.21%

Interprétation : courbe VI-10 (MTSR)

Les taux obtenus sont trés satisfaisants et augmentent suivant le nombre de classes . le nombre K optimal des "K" premières références est toujours 2 quand le nombre de classes est plus de 2 .

C-2 / CONCLUSION :

Nos tests ont montré que les reconnaissances sont d'autant

meilleures que le nombre de classes est élevé et que des valeurs optimales de "K" , existent réellement (K =2) indépendament configuration des échantillons .Ceci prouve que nous avons bien appliqué les méthodes de classification et les techniques de décision puisque la valeur de K obtenue avec de la parole réelle par Rabiner [16] est la même .

CHAP VII

CONCLUSIONS

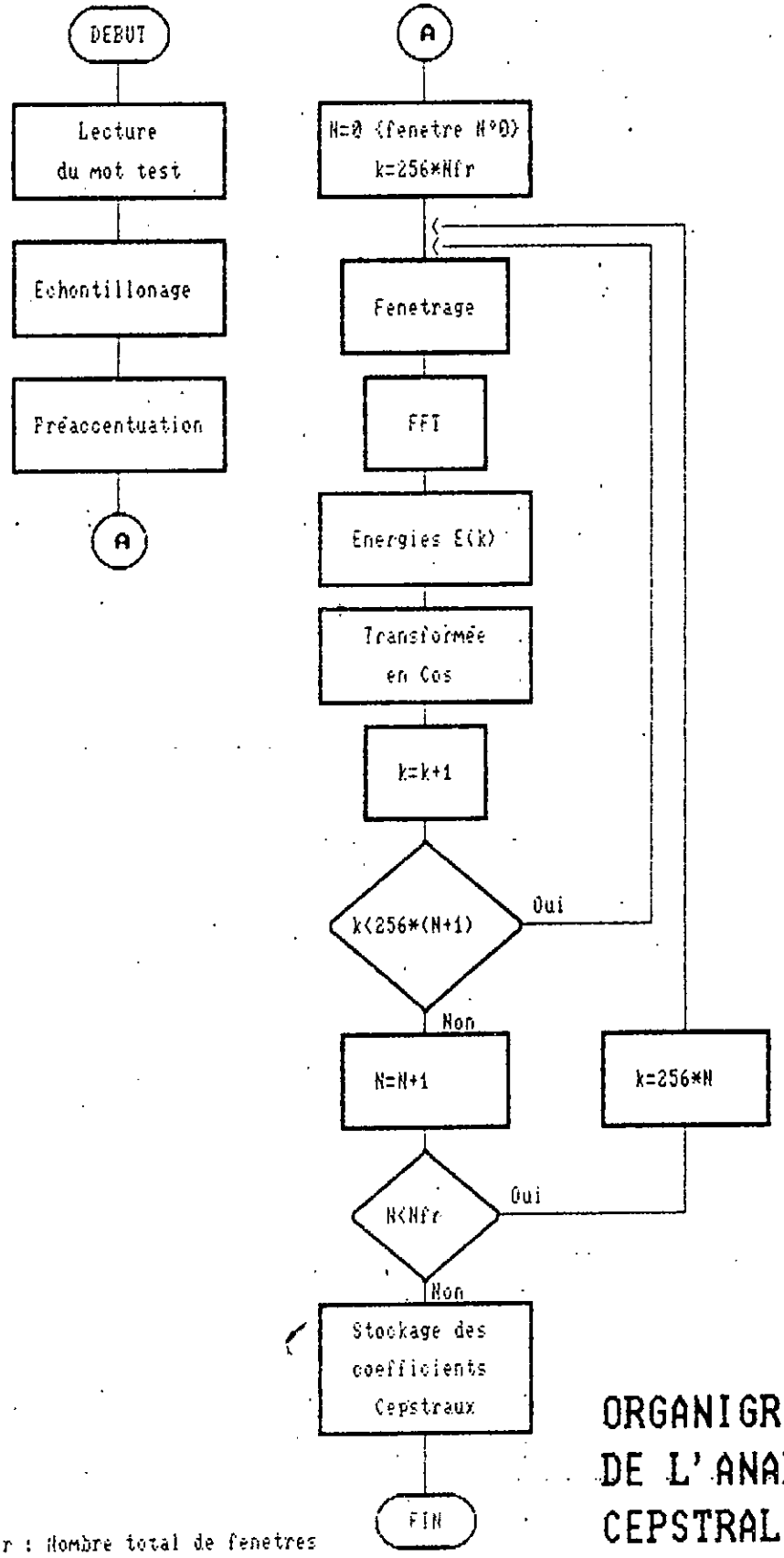
L'objectif initial de notre travail était au départ , l'étude d'un système de reconnaissance de la parole par les méthodes glabales en mode multilocuteurs en vue d'une éventuèlle implantation sur le microprocésseur TMS 320 . Cependant ,beaucoup de contraintes nous ont empéché de concrétiser ce but . Nous étions donc forçés de nous limiter uniquement à la partie étude où nous avons mis au point des programmations trés élaborées et trés opérationnelles .

Le manque de temps et la panne prolongée du Vax nous ont crée énormémment de lassitudes . Néaumoins ,nous avons pu générer assez de fichiers pour nos différents tests où nous avons pu approché ,plus au moins ,beaucoup de résultats publiés par les chercheurs travaillant ,contrairement à nous ,sur la parole réelle .

Par ailleurs , nous insistons sur le fait que la recherche sur la reconnaissance de la parole est devenue ,de nos jours ,presque totalement éxpérimentale et éxige beaucoup de moyens pour pouvoir arriver à un quelconque résultat . Des données théoriques relatives à la parole ne peuvent plus servir de base dés lors que les procédures actuelles des chercheurs ,commencent par émettre des hypothèses puis procéder aux éxpériences pour les valider ou bien les infirmer surtout avec les performances de l'intélligence artificièlle .

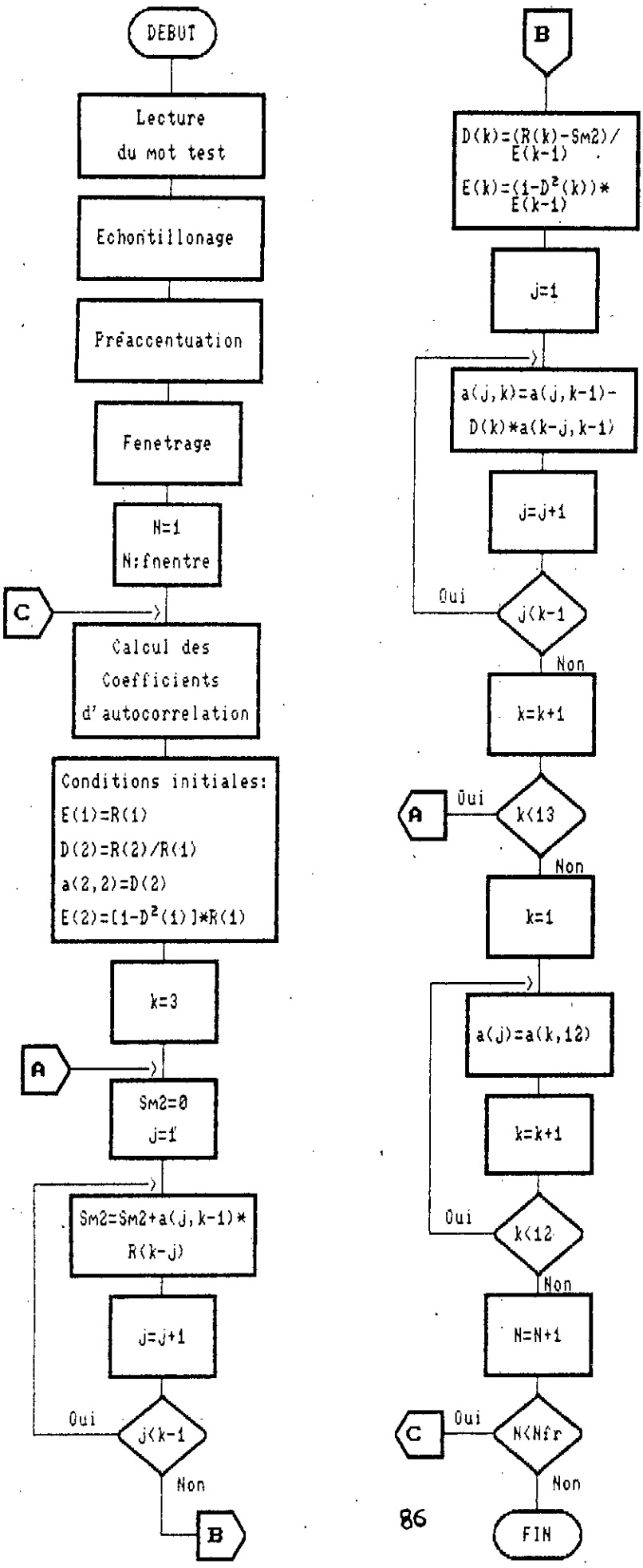
Enfin , il est intuitivement évident qu'avec ce sujet ,nous avons développé de beaucoup nos connaissances en programmation , en algorithmique et en traitement du signal .

ANNEXE

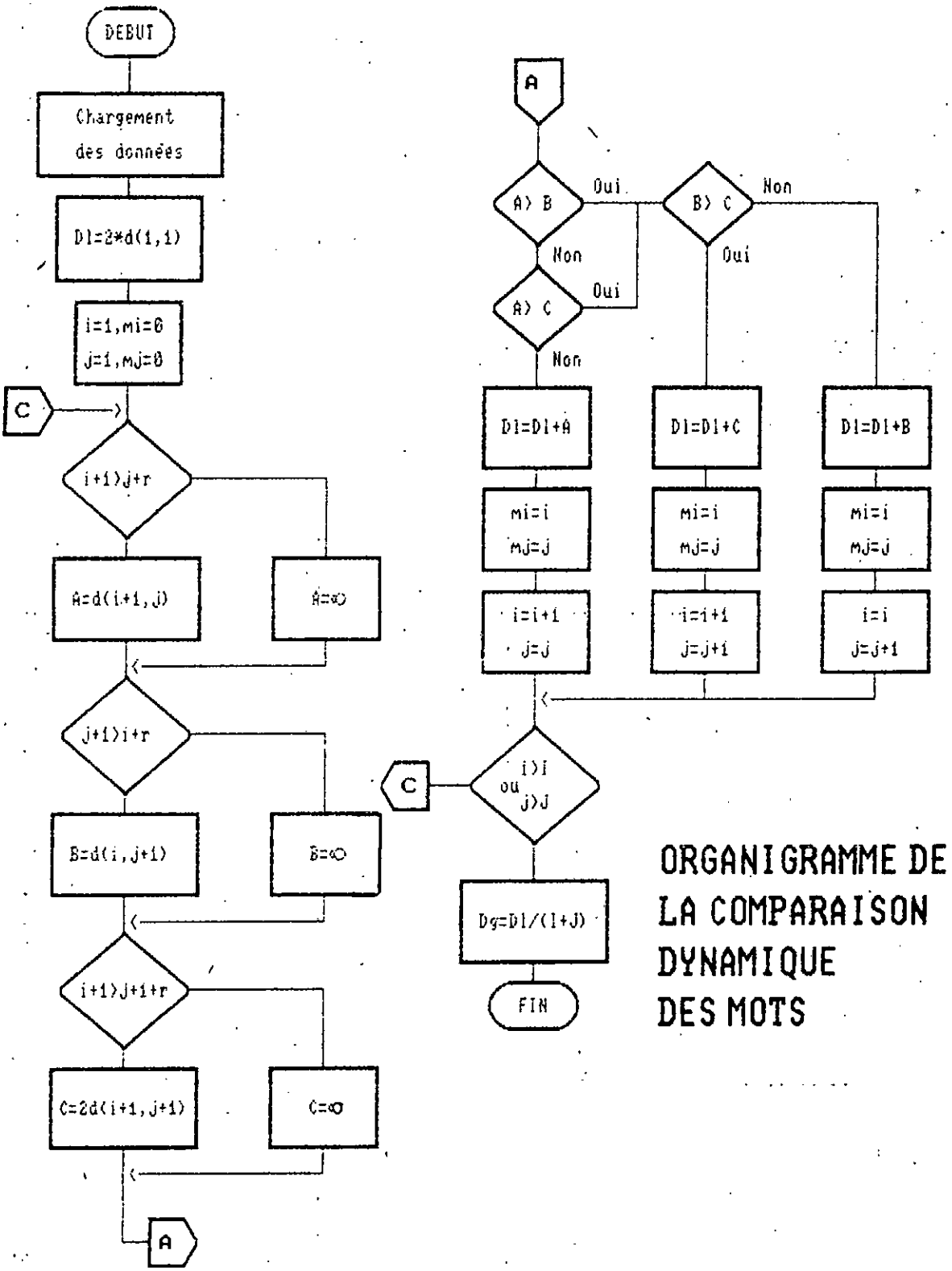


Nfr : Nombre total de fenetres

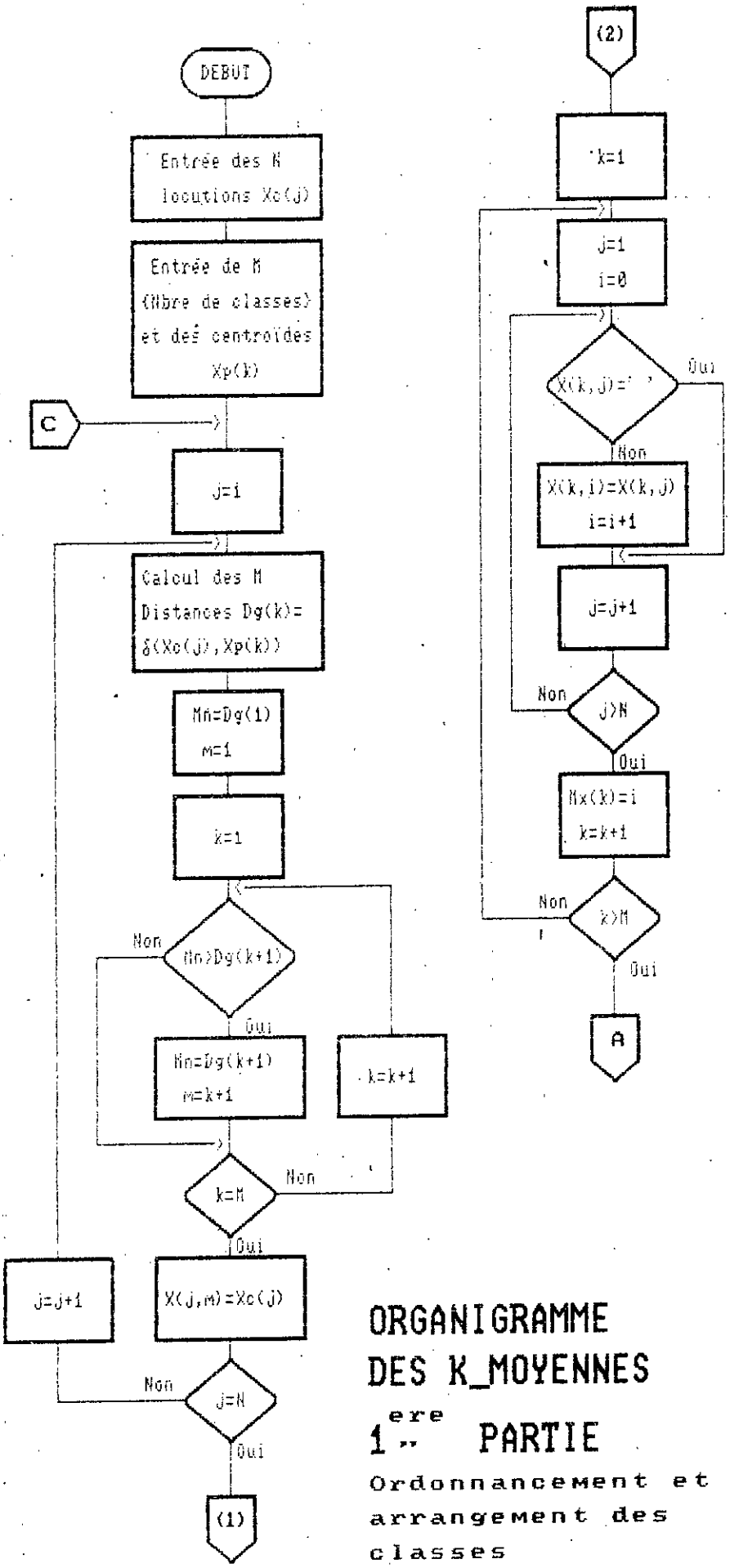
ORGANIGRAMME DE L'ANALYSE CEPSTRALE



ORGANIGRAMME DE L'ANALYSE LPC



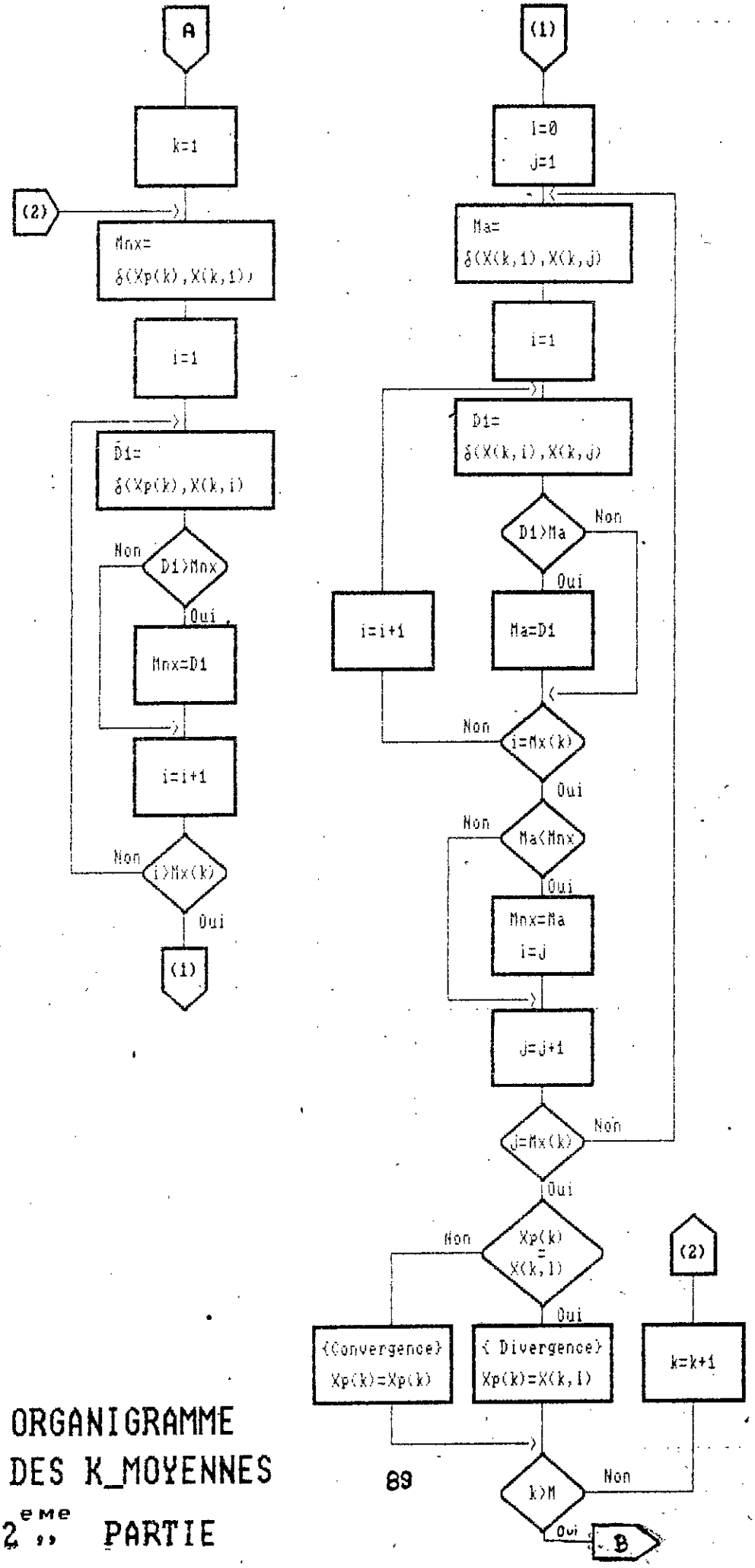
ORGANIGRAMME DE LA COMPARAISON DYNAMIQUE DES MOTS



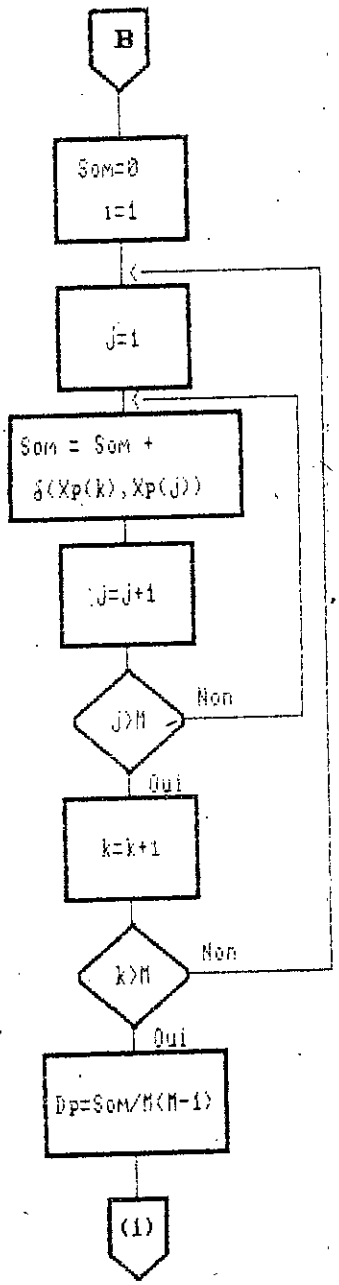
ORGANIGRAMME DES K_MOYENNES

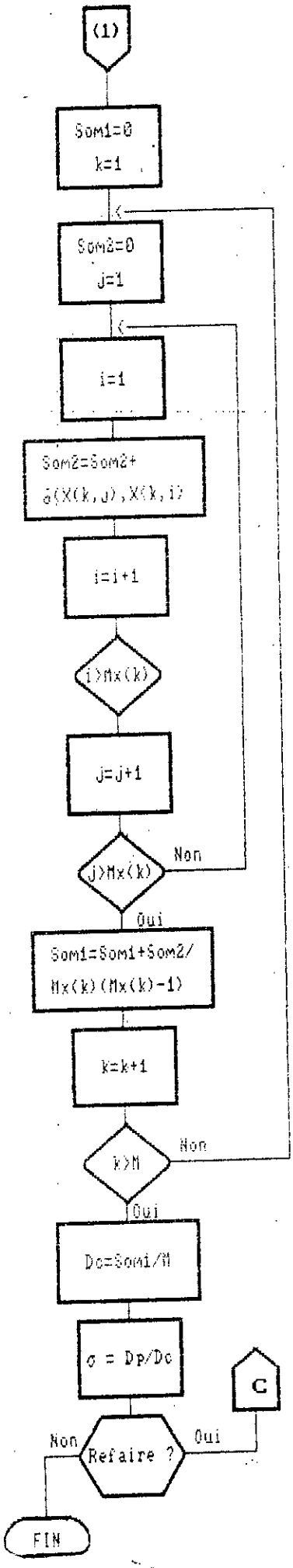
1ere PARTIE

Ordonnancement et arrangement des classes



ORGANIGRAMME DES K_MOYENNES 2ème PARTIE





ORGANIGRAMME DES K_MOYENNES 3eme PARTIE

Calcul de σ la qualite de mesure

BIBLIOGRAPHIE

[1]: Dictionnaire "Petit Robert" 1972 ;

[2]: " Reconnaissane de la parole en mode multilocuteur " ,1984
A Menacer ,thèse Docteur-Ingénieur ,Université de Neutchatel ;

[3]: " Traitement de la parole "
par M.Kunt ,R.Boite ,Presses Polytechniques Romandes, 1987 ;

[4]: "Linear Prédiction :A totorial review proceedings "
par J.Makhol , IEEE ,vol.63 ,Avril 1975 ;

[5]: " Traitement Numérique des signaux "
par M.Kunt ,édition Dunod , 1987 ;

[6]: " Dynamic Programming Algorith optimization for Spoken word Recognition " par H.Sakoe ,S.Chiba
IEEE ,ASSP ,vol.35 ,N°10 ,Octobre 1987 ;

[7]: " Reconnaissance de la parole en mode multilocuteur de mots isolés par les systèmes miniaturisés " , 1985
A.Mokkedem ,thèse Docteur es-Science ,université Neutchatel

[8]: " La parole , compréhension et synthèse par les ordinateurs "
par J.Guibert ,Presses Universitaires ,1979 ;

[9]: " Interactive Clustring Techniques for Selecting Speaker Independent References Templates for Isolated word and Recognition " ,
par S.Levinson ,L.Rabinner ,J.Wilpon ,A.Rosenberg
IEEE ,vol.27 , N°2 ,Avril 1979 ;

[10]: "A vector Quantization based Preprocessor for Speaken Indépendent Isolated word Recognition "
par K.Pan ,R.Chafer ,F.K.Soong ,L.Rabiner
IEEE , ASSP, v.33 ,N°3 ,Juin 1985 ;

[11]: " Cours de Phonétique " par E.Emerit
SNED , 1977 ;

[12]: " Digital Processing of Speech Signals "
par L.Rabiner ,R.Chafer ,
Printice Hal , Alan Openheim , New Jersey , 1978 ;

[13]: "A Weighted Cepstral Distance Measure for Speech Recognition"
par Y.Tohkura ,
IEEE , ASSP , vol.35 ,N°10 ,Octobre 1987 ;

[14]: -" Considérations in Dynamic Time Worping Algorithms for Discrete Word Recognition "
par L.Rabiner ,S.Levinson ,A.Rosenberg ,
IEEE , ASSP ,Vol26.34 ,N°6 , Decembre 1978 ;

[15]: "Speaker Independent Isolated word Recognition Using Dynamic Features of Speech Spectrum " ,
par S.Furui ,IEEE,ASSP ,vol.34 ,N°1 ,Fevrier 1986 ;

[16]: "Speaker Independent Isolated Word Recognition a Moderate Size Vocabulary " ,
par L.Rabiner ,IEEE , vol.27 ,N°6 ,Decembre 1979