République Algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de
la recherche scientifique

Centre Universitaire Chikh El-Arbi Tbessi
de Tébessa

Institut des sciences exactes et technologie

Département d'Electronique

MEMOIRE

Présenté en vue de l'obtention du diplôme de **MAGISTER**
Le : 27/06 / 2007

DEVELOPPPPEMENT D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE A BASE DES HMM

Option :

Contrôle et automatique

Par
Brai Radhia

RAPPORTEUR DE MEMOIRE :	M..M. BEDDA	Professeur	U. ANNABA

Devant le jury

PRESIDENT :	M. N. GUERFI	Maître de conférence	C.U.TEBESSA
EXAMINATEURS :	M..N. DOGHMANE	Professeur	U. ANNABA
	M. M..MAAMRI	PhD	C.U.TEBESSA
INVITE	M. A.GATTAL	C.C	C.U.TEBESSA

Je dédie cette thèse

A notre source d'amour ma mère AKILLA

Au archée de mon père MOHAMED

A mes sœurs:

Yamina Farida Amel

Et mes frères: Ali Nabil

A ma grande famille

Ainsi toutes

Mes amies

Remerciements :

Je tiens à remercier les membres du jury qui m'ont fait l'honneur de participer à l'examen de ce travail:

J'exprime ma sincère reconnaissance à Monsieur Guerfi Professeur à l'Université de Tébessa, pour avoir accepté de juger ce travail en tant que président du jury.

Je remercie respectueusement Monsieur M.Maameri, Maître de Conférences à l'Université de Tébessa, pour leur participation à ce jury. et Monsieur A. Guattal C.C au Centre universitaire de Tébessa, pour l'intérêt qu'il a témoigné à l'égard de ce travail.

J'ai été profondément honoré que Monsieur N.Doghmane, Professeur à l'Université d'Annaba d'examiner ce travail.

Je remercie Monsieur M.BEDDA Professeur à l'Université d'Annaba, qui a dirigé mes travaux, pour sa constante disponibilité, pour l'aide et les nombreux conseils et encouragements qu'il m'a toujours prodigués, pour m'avoir fait bénéficier de sa rigueur scientifique, de ses critiques objectives et de ses conseils avisés.

J'adresse également mes remerciements à Monsieur A. Guattal C.C au Centre universitaire de Tébessa, pour son soutien et ses remarques constructives tout au long de ce travail.

ملخص:

يعتبر مجال التعرف على نظام معالجة الصوت من أحد أهم مواضيع البحث فى مجال الإشارة ،وبالرغم من التطور العلمي والإكتشافات الحديثة التي صاحبت جميع العلوم وخاصة في مجال الحاسوب ، إلا أن هناك بعض المشاكل والعيوب التي مازالت قيد الاستكشاف و البحث في مجال معالجة الصوت ، مما أدى إلى تكثيف البحث والاهتمام به كعلم في حد داته ،والسعي في المستقبل إلى حل جميع ألغازه وفك رموزه ، ودلك من أجل الوصول إلى نموذج للصوت بين الإنسان والآلة.

وفي نظام معالجة الصوت هناك طريقتين: الطريقة التحليلية ، والطريقة الإجمالية.

الطريقة الأولى تعتمد على تقطيع الكلمة إلى وحدات صغيرة (phonème _syllabe) ، أما الطريقة الثانية تعتمد على الكلمة جملة .ومن خلال البحث لاحضنا أن الطريقة الاولى أحسن من الطريقة الثانية ،لأنها لا تعتمد على استعمال قاعدة معطيات ضخمة .

مع العلم ان هناك نموذجين لتمثيل هدا النظام، النموذج الأول جبري (يعتمد على الميزات الداخلية للإشارة)، والنموذج الثاني عشوائي (يعتمد على الميزات الإحصائية للإشارة خاصة الإحتمالات)،ويعطي هدين النموذجين أحسن النتائج، ولقد اعتمدنا في هدا البحث على الطريقة التحليلية ،التي تختص بالنموذج العشوائي ، وبالخصوص نمادج ماركوف الخفية (HMM) ،التي تعطينا أحسن النتائج.

Abstract:

The Field of speech recognition became one of the most interesting subjects of research in signal processing. This is due in particular to the processing performances and the calculation's speed of the current computers. This has led in fact to the availability of several commercial products.

How ever, the actual systems remain limited in terms of robustness and they have to have the ability to adapt to different environments such as shopping centres, battle fields or other scenarios where the bush ground noise is inevitable.

Usually two approaches for speech recognition are used, the analytical approach and the global approach.

The first, tries to deal with the processing of the continues speech by braking up the words, generally by proceeding to an acoustico-phonetics's decoding exploited by modules of linguistic level. the second approach consists of globally identifying a word or a sentence by composing them to recorded references.

At This stage two modélisation possibilities are proposed. the modélisation base on the determinist model and the one based on the stochastic model ;the later generally exploits few properties of the signal in view of it's modélisation , lobe for instance the form of the signal (sinusoid , same of exponentials ,amplitude …).the stochastic modélisation tries to determine the statistical caractéristics of the signal.

This is especially the case of probability distribution (Gaussan, Poisson …).

For applications of speech processing, the two models have given excellent results. We will concentrate only on one type which is the stochastic model, in fact the hidden Markov models (HMM).in addition; these models give remarkable results in practice when they are correctly applied.

Résumé :

Le domaine de la reconnaissance de la parole est devenu l'un des sujets de recherches les plus intéressants en traitement de signal. Ceci est dû en particulier aux performances de traitements et de calculs que permettent les ordinateurs actuels. Cependant les systèmes actuels restent limités en matière de robustesse, et doivent donc performer, dans le futur, leur capacité à s'adapter dans différents environnements, comme par exemple dans les surfaces commerciales, les champs de bataille ou d'autres scénarios où le bruit de fond est inévitable.

On distingue usuellement e reconnaissance de la parole l'approche analytique et l'approche globale. La première approche cherche à traiter la parole continue en décomposant le problème, le plus souvent en procédant à un décodage acoustico-phonétique exploité par des modules de niveau linguistique. La seconde consiste à identifier globalement un mot ou une phrase en les composant avec des références enregistrées.

A ce stade deux possibilités de modélisation se proposent à nous. Le modèle déterministe et le modèle stochastique ; le modèle déterministe exploite généralement quelques propriétés du signal en vue de sa modélisation, comme par exemple la forme du signal (sinusoïde, somme d'exponentielles, amplitude...). La modélisation stochastique quant à elle, elle essaie de déterminer les caractéristiques statistiques du signal. On parle dans ce cas spécialement de fonctions de distribution de probabilité (Gaussienne, Poisson …).

Pour des applications de traitement de la parole, les deux modèles ont abouti à d'excellents résultats. Nous allons nous concentrer sur un seul type de modélisation stochastique, en l'occurrence les modèles de Markov cachés (HMM).En outre, ces modèles donnent de remarquables résultats en pratique quand ils sont correctement appliqués.

SOMMAIRE

Chapitre 3 : Modèles de MARKOV CACHES

Chapitre4 : Système de reconnaissance et résultats

Listes des figures

Listes des tableaux

Notation

C	Consonne
V	Voyelle
LPCC	Cœfficient de prédiction linéaire
LFCC	Coefficient cepstraux réel (Linear Frequency Cepstral Coefficients) ;
MFCC	Mel Frequency Cepstral Coefficients).
TFD	Transformer de Fourier discret
a_p	Coefficient autorégressive d'ordre p
FFT	Fast Fourier Transform
S_k	L'énergie du signal
$e(n)$	L'écart
δ_e^2	Variance de l'écart $e(n)$
$r_{xx}(i)$	Fonction d'autocorrelation
R_{xx}	Matrice $p \times p$ d'auto-corrélation.
r_{xx}	Le vecteur $p \times 1$ d'auto-corrélation.
PLP	Perceptual Linear Prediction
Rasta PLP	RelAtive SpecTrAl
A	La probabilité de l'observation acoustique A.
P(A|W)	La probabilité de l'observation acoustique A connaissant une séquence de mots W
P(W)	La probabilité a priori de la séquence de mots W
HMM	Modèles de Markov cachés(Hidden Markov Models)
a_{ij}	La matrice de probabilité de transition sur l'ensemble des états du modèle
$b_j(x_n)$	La matrice de probabilités d'émission de l'observation x_n dans l'état q_j.
Л	La distribution initiale des états
$x=x_1 \ldots x_n$	La séquence d'observations.
$p(M / X, \lambda)$	Probabilité qu'un modèle M génère une séquence de vecteurs acoustiques X étant donné une série de paramètres λ.
MAP	Maximum a posteriori (Maximum a posteriori Probability)
MLE	Maximum de vraisemblance (Maximum Likelihood Estimation)
$\alpha_t(i)$	Fonction de probabilité « avant »
$\beta_t(i)$	Fonction de probabilité « arrière »

Introduction générale

La recherche dans le domaine de la communication parlée, notamment, celle axée sur le développement de nouvelles technologies de l'information, est une activité en pleine expansion, dont plusieurs disciplines et compétences, interagissent dans le but d'améliorer les performances des systèmes de communication homme machine.

La reconnaissance automatique de la parole (RAP) fait partie intégrante de cette discipline , et représente l'un des thèmes essentiels de la communication parlée.

Plusieurs projets visent à intégrer la parole dans des interfaces homme-machine en vue d'aboutir à des systèmes sophistiqués capables de simuler le comportement humain à tous les niveaux.

Diverses techniques dont le modèle de Markov cachés, sont mises en oeuvre afin d'aboutir à de meilleures performances. Depuis leur introduction en traitement de la parole par Baker (1975) et Jelinek (1976), les modèles de Markov cachés (*Hidden Markov Model* ou HMM) ont pris une importance considérable, au point que la quasi-totalité des systèmes de RAP utilisent cette approche. Les modèles de Markov cachés permettent de modéliser un processus aléatoire inobservable (cachés) qui se manifeste par des émissions aléatoires et observables. Ces deux niveaux donnent à l'approche markovienne une flexibilité séduisante pour modéliser un phénomène aussi complexe que la production de la parole. Toutefois, les techniques Markovienne restent encore tributaires de contraintes d'ordre algorithmique qui nécessitent une meilleure connaissance du processus , tel que par exemple, les conditions initiales liées a la convergence des algorithmes.

L'étape de reconnaissance par l'approche syllabique nécessite généralement une étape de segmentation, Cette approche consiste à considérer que le signal est une succession de segments syllabiques.

Dans le cadre de notre travail, nous nous sommes basés sur le développement d'un système de reconnaissance de chiffre arabe basé sur les HMM

Dans le premier chapitre, la parole et ses caractéristiques sont présentées ainsi que les principales techniques d'analyses du signal.

Dans le chapitre deux nous présentons les techniques de segmentation de la parole et son utilité pour la reconnaissance.

Dans le chapitre trois les modèles de Markov cachés sont exposés en explicitant les trois problèmes principaux. Et leurs utilisations pour la reconnaissance.

Dans le chapitre quatre nous présentons le système de reconnaissance développé et le résultat de reconnaissance obtenu sur une base de données composée par les digits arabes.

Et nous terminons par donner une conclusion générale.

1 Généralités et Modélisation de la Parole

Sommaire

1.2 Introduction :

La parole est le principal moyen de communication dans toute société humaine. Son apparition peut être considérée comme concomitante à l'apparition des outils, l'homme ayant alors besoin de raisonner et de communiquer pour les façonner[1] .Son abstraction par rapport à un support physique en fait un moyen de communication très simple à utiliser.

L'importance de la parole fait que toute interaction homme-machine devrait plus ou moins passer par elle. D'un point de vue humain, la parole permet de se dégager de toute obligation de contact physique avec la machine, libérant ainsi l'utilisateur qui peut alors effectuer d'autres tâches. Sans pour autant imposer la parole là où elle pourrait être un frein à l'interaction (il est par exemple difficile d'imaginer une application graphique où seule la parole serait utilisée), son utilisation permettrait de commencer à limiter l'emploi des claviers, tablettes graphiques et autres écrans
tactiles ou gants de désignation.

1.2 Le signal de la parole :

Le signal acoustique est un signal complexe. Il s'agit d'une perturbation d'air qui ne peut pas être comparée directement avec une autre dans l'échelle du temps : sa variabilité en amplitude et phase rend deux signaux différents alors qu'ils peuvent porter la même information, cette variation de pression d'air peut avoir pour origine plusieurs sources. Il peut s'agir de l'appareil phonique humain, d'un instrument musical ou de l'environnement naturel.

La parole est le résultat d'une « phonation » (source) et d'une « articulation » (filtre) selon un modèle simple de la théorie acoustique de la production de la parole
La source est un signal composé d'une partie périodique (vibration des cordes vocales) et d'une partie bruitée, utilisée séparément ou conjointement. Le conduit vocal est une cavité acoustique de forme complexe. Sa fonction est de transformer le signal de source par des phénomènes de résonance et anti-résonance. La parole est donc une alternance de sons voisés (quasi-périodiques) et de sons non voisés (bruit).

Le signal de la parole n'est pas un signal ordinaire. Il est le vecteur d'un phénomène complexe : la communication parlée, la reconnaissance de la parole pose

de nombreux problèmes aux chercheurs depuis 1950. D'un point de vue mathématiques, il est difficile de modéliser le signal de parole, compte tenu de sa variabilité. Nous allons ici tenter de mettre en évidence quelques caractéristiques importantes du signal non stationnaire afin de faire ressortir les problèmes posés lors de son traitement, et de sa modélisation.

1.2.1 Redondance du signal :

Le signal de parole est extrêmement redondant, cette grande redondance lui confère une robustesse à certains types de bruits. De nombreuses recherches sont menées afin de rendre les systèmes de reconnaissance robustes aux bruits[2], mais les performances humaines sont encore loin d'être atteintes.

1.2.2 Variabilité du signal :

Le signal de parole possède une très grande variabilité. Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution peut varier, la durée du signal est alors modifiée. Toute altération de l'appareil phonatoire peut modifier la qualité de l'émission (exemple : rhume, fatigue…). De plus, la diction évolue dans le temps. La voix est modifiée au cours des étapes de la vie d'un être humain (enfance, adolescence, âge adulte…).

La variabilité interlocuteur est encore plus accentuée. La hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. Enfin, la parole est un moyen de communication où de nombreux éléments entrent en jeu, tels que le lieu, l'émotion du locuteur, la relation qui s'établit entre les locuteurs (stressante ou amicale). Ces facteurs influencent la forme et le contenu du message. L'acoustique du lieu (milieu protégé ou environnement bruité), la qualité du microphone, les bruits de bouche, les hésitations, les mots hors vocabulaire sont autant d'interférences supplémentaires sur le signal de parole.

1.2.3 Les effets de coarticulation :

La production parfaite d'un son suppose un positionnement précis des organes phonatoires. Le déplacement de ces organes est limité par une certaine inertie

mécanique. Les sons émis subissent l'influence de ceux qui les précèdent ou les suivent. Cet effet de coarticulation est un facteur de variabilité supplémentaire important du signal de parole.

1.3 Description du signal vocal :

La parole est un signale réel, continue d'énergie finie, non stationnaire .Sa structure est complexe et variable dans le temps : tantôt pseudopériodique pour les sons voises, tantôt aléatoire pour les sons non voises, tantôt impulsionnel dans les phases explosives . (figure1.1)

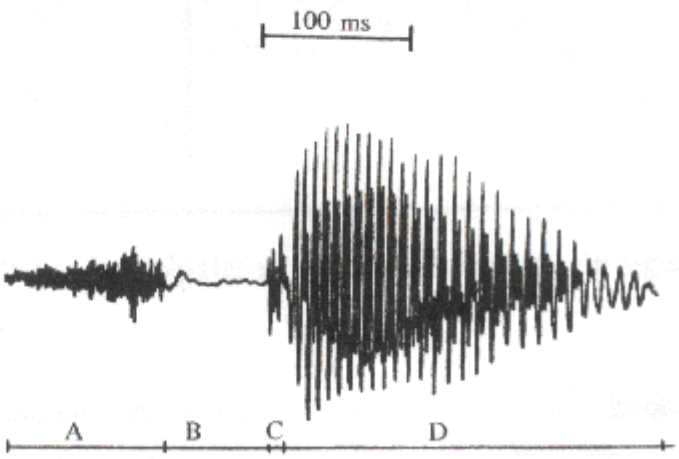

Fig.1-1 : signal vocal

Le signal de la parole est un phénomène de nature acoustique, il est caractérisé par son amplitude, sa durée et son timbre (figure1-2) :

- **L'amplitude** : elle correspond aux vibrations sonores, elle est exprimée en décibel, notant aussi, que l'amplitude peut atteindre des valeurs difficilement tolérables pour l'oreille

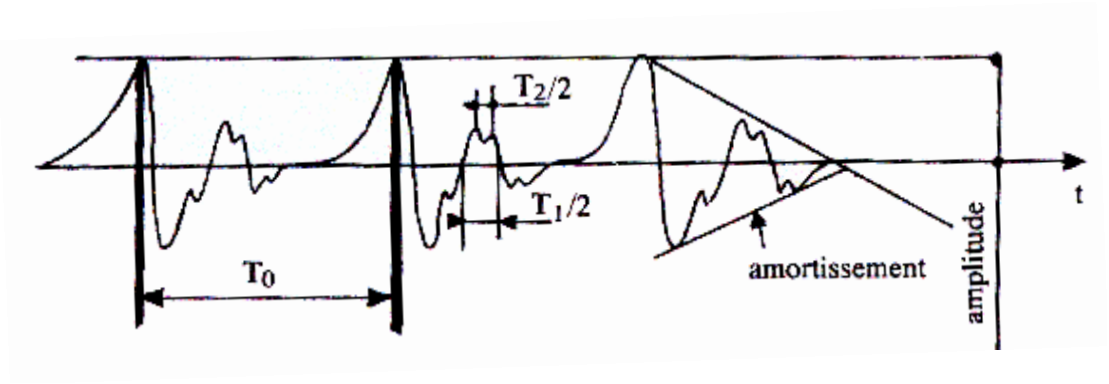

Fig.1-2 : les principales grandeurs mesurables du signal vocal

- **La durée** : elle est liée aux caractéristiques des cordes vocals, elle peut varier suivant l'âge et le sexe du locuteur :
  1. chez l'homme : de 80 à 200 Hz.
  2. chez la femme : de 150 à 450 Hz.
  3. chez l'enfant : de 200 à 600 Hz.

- **Le timbre** : il est du au audibilité des harmoniques du fondamental, il distingue deux sons de même intensité et de même hauteur :

- Pour la source : période du fondamentale $T_0 = 1/F_0$ et d'amplitude A_0 .
- Pour le conduit : période des formants $T_i = 1/F_i$, $i = 1,2,.....$ Amplitude A_i , amortissement et phases.

1.4 Modèle de production de la parole :

Une représentation fonctionnelle du modèle de production séparant sources, conduit vocal et rayonnement aux lèvres, est donnée par la (figure1-3) :

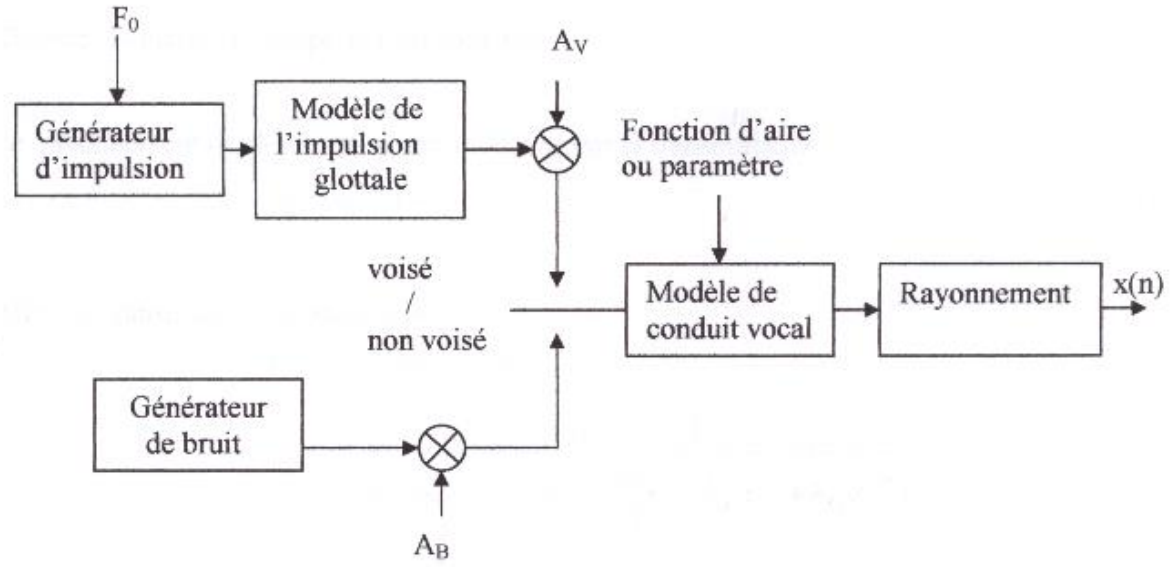

Fig.1-3 : Modèle fonctionnel de production de la parole

Une phrase est une suite de son voisés, de son non voisés, et de silences. Pour la générer, il faut connaître pour chaque intervalle de temps dT , intervalle durant le quelle le modèle est considère comme invariant (5 à 30ms).

- fréquence fondamentale F_0 .
- les amplitudes A_v et A_b .
- les coefficients des filtres modélisant le conduit vocale, l'impulsion glottale et le rayonnement aux lèvres.

1.5 Modélisation de la parole :

Pour réduire le débit tout en conservant une qualité suffisante ou pour améliorer la qualité pour un débit imposé, il faut chercher les paramètres pertinents qui constituent le signal de parole. Pour cela un modèle simplifié du système phonatoire ne retenant que les paramètres les plus significatifs du signal est recherché. Dans une grande majorité de codeurs, la production de parole est modélisée par une opération
de filtrage ,où une source sonore excite un filtre censé représenter le conduit vocal. Physiologiquement la source et le filtre ne sont pas séparés. Cependant, ils seront considérés comme tels lors de l'analyse des signaux de parole. Une modélisation précise du conduit vocal et de la source d'excitation est nécessaire pour produire respectivement un signal intelligible et naturel.

Des études menées dans le but de définir un modèle produisant un signal proche du signal de parole ont permis à Fant [3] de décrire un modèle de production linéaire.

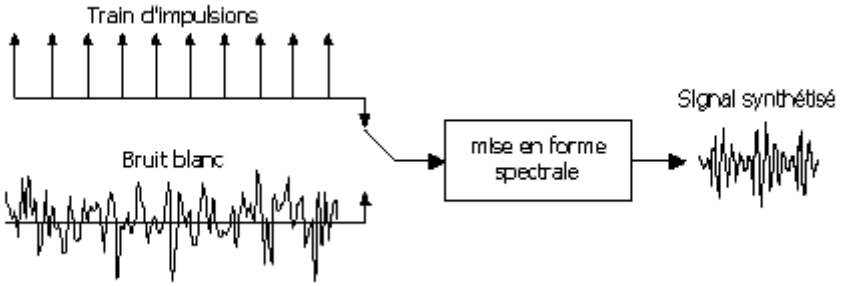

Fig.1-4: Synthèse de parole à 2 états d'excitation

Un signal d'excitation et la modélisation du conduit vocal permettent donc de caractériser la génération de la voix. Le modèle du conduit vocal H (z) est excité par un signal glottal discret U (z) pour produire un signal de parole S (z) :

$$S(z) = H(z)U(z) \quad \text{(1.1)}$$

Le signal est donc modélisé par la sortie d'un filtre excité par un train d'impulsions périodiques (peigne de Dirac) pour les sons voisés et un bruit blanc pour les sons non-voisés (figure 1-3).

La modélisation d'un signale $x(n)$ consiste à lui associer un filtre linéaire qui soumis, à une excitation particulière reproduit ce signal le plus fidèlement possible.

L'objectif essentiel de la modélisation d'un signale est de permettre la description de son spectre par un ensemble très limite de paramètres.

Un signale voisé peut être modélise par le passage d'un train d'impulsions $U(n)$ à travers un filtre numérique récursif de type tout pôle $\frac{1}{A(Z)}$.

Cette modélisation reste valable aussi pour les sons non voisé, à condition que $U(n)$,soit cette fois un bruit blanc .Le modèle finale est illustre par la (figure 1.4) il est appelé modèle autorégressif (AR).

L'estimation du modèle AR est basée sur la prédiction linéaire qu'on dispose pour celle-ci d'algorithmes rapides et efficaces.

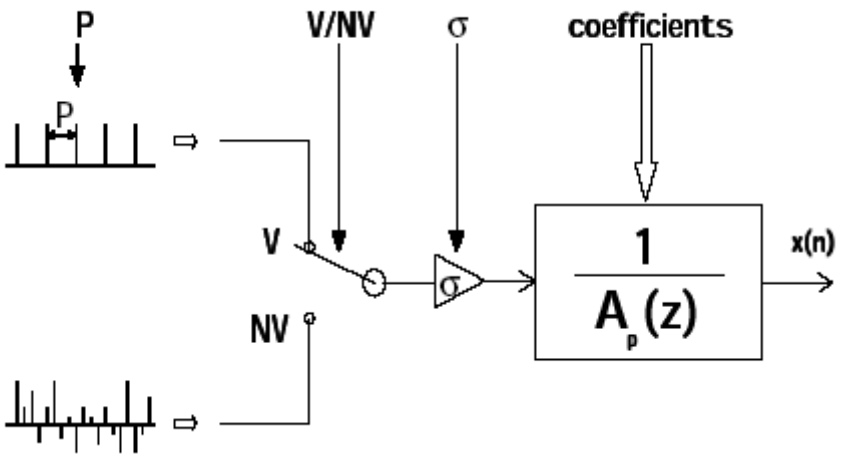

Fig.1-5 : le modèle autorégressif

La première étape pour chaque traitement dans le domaine de la parole est la paramétrisation, ou le calcul des paramètres acoustiques à partir du signal de parole. Le signal de parole n'est pas directement utilisable à cause de sa grande complexité (grande diversité d'information) et de son caractère redondant.

Le but de la paramétrisation est d'extraire l'information pertinente pour la tâche proposée. La première étape de la paramétrisation acoustique consiste à découper le signal de parole en fenêtres de taille fixe (variable de 20 ms à environ 40 ms) reparties de façon uniforme le long du signal (toutes les 10 ms). La taille des fenêtres est choisie en considérant que les propriétés du conduit vocal peuvent être considérées comme invariables sur une petite durée égale à la taille de la fenêtre. Le signal audio est donc considéré comme stationnaire sur la durée de la fenêtre.

Il existe de nombreux paramètres utilisés dans la littérature. Plusieurs classifications peuvent être envisagées mais trois catégories principales peuvent être considérées :

- les paramètres spectraux : transformée de Fourier, bancs de filtres;
- les paramètres temporels : le taux de passage par zéro, le débit d'élocution;
- les paramètres cepstraux : LPCC, LFCC, MFCC.

L'énergie est aussi un paramètre important dans le traitement de la parole. Celle-ci peut être considérée, selon la façon dont elle est calculée, comme un paramètre spectral ou temporel (le *théorème de Parceval* nous indique qu'un calcul d'énergie peut aussi bien être fait dans le domaine spectral que temporel).

Les paramètres acoustiques les plus utilisés pour la caractérisation du locuteur sont les coefficients cepstraux, notamment les coefficients MFCC. Parmi les systèmes présentés pendant les évaluations NIST [4] en reconnaissance de locuteur et en segmentation, quasiment, tous ont utilisé des coefficients cepstraux et une grande majorité [5, 6, 7] a utilisé les coefficients MFCC.

Les paramètres cepstraux [8, 9] sont utilisés pour séparer l'influence de la source d'excitation vocale et celle du conduit vocal. Cependant, elle peut être associée à d'autres paramètres pour améliorer le résultat du système de reconnaissance [10].

Les coefficients cepstraux [11] sont utilisés dans le traitement du signal et sont des coefficients d'énergie calculés dans des bancs de filtres. Selon la distribution des filtres dans la bande utile du signal l'un des deux cas suivants peut être considéré :

1. les filtres sont uniformément distribués dans la bande utile du signal. Dans ce cas les coefficients calculés sont appelés LFCC (Linear Frequency Cepstral Coefficients) ;

2. les filtres sont distribués selon une échelle non-linéaire, plus proche de la perception humaine appelée échelle Mel. Dans ce cas les coefficients calculés sont les MFCC (Mel Frequency Cepstral Coefficients).

Etant donné que les coefficients LFCC sont similaires aux MFCC, sauf pour la distribution de filtres dans la bande utile du signal.

1.5.1 Coefficients cepstraux :

a) Les coefficients LFCC [12] :

Le signal de parole est modélisé en utilisant le spectre. Il est possible de calculer directement les coefficients cepstraux en suivant le processus d'écrit sur la figure(1-6) :

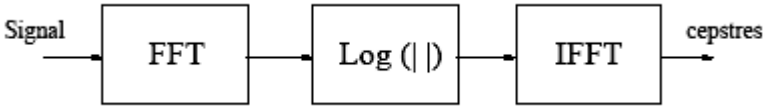

Fig 1-6 calcul du cepstre réel

Le calcul du cepstre réel par l'intermédiaire de la TFD(figure 1.6) est assez coûteux en temps de calcule. Il peut être estimé à partir des coefficients de prédiction $a_p(i)$.

La variance des coefficients cepstraux décroîts avec leur ordre :

$$C_0 = -\ln(\delta^2) \quad \text{.....(1.2)}$$

$$C_1 = -a_1 \quad \text{.....(1.3)}$$

$$pour \qquad k = 1,2,3,\ldots\ldots, p$$

$$C_k = -a_k - \frac{1}{k}\sum_{j=1}^{k-1}(k-j).C_{k-j}.a_j \quad \text{.....(1.4)}$$

b) Les coefficients MFCC :

De nombreuses études ont montré que la perception humaine des sons ne suit pas une échelle linéaire [8], d'où l'idée de définir pour chaque valeur de fréquence f (en Hz) une hauteur subjective qui est mesurée sur une échelle "mel" (en mels). Comme point de référence la fréquence de 1 kHz correspond à 1000 mels.

Pour la correspondance entre la fréquence mel et la fréquence réelle en Hz la fonction la plus utilisée est :

$$F_{MEL} = 2595.\log_{10}\left(1 + \frac{f_{hz}}{700}\right) \quad \text{.....(1.5)}$$

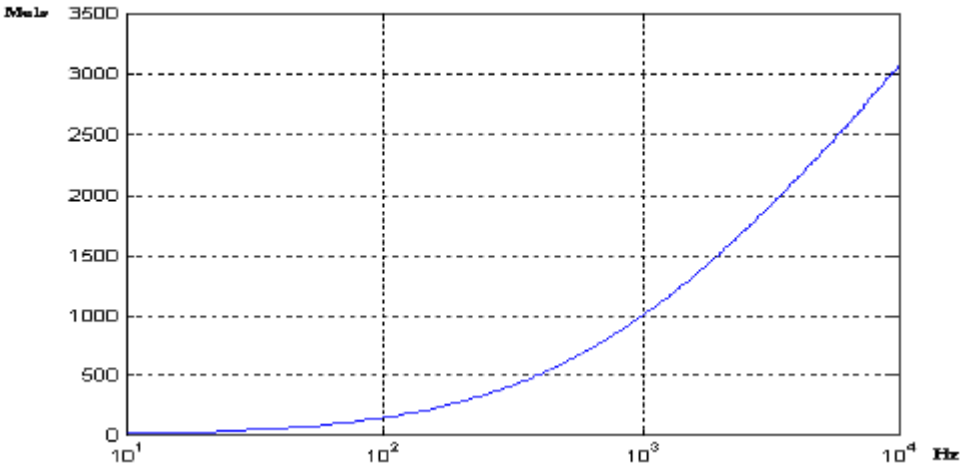

Fig.1- 7 : Correspondance mels / hertz

La Figure 1-7 montre la correspondance entre la fréquence en mels et la fréquence réelle en Hz. Pour des raisons pratiques, le spectre mel est simulé en

utilisant un banc de filtres triangulaires uniformément répartis sur l'échelle mel. La distance entre deux triangles est d'environ 150 mels et la largeur d'un triangle est d'environ 300 mels. Cependant, il existe des approches où ces deux paramètres ne sont pas respectés. Dans ce cas, les filtres sont uniformément distribués dans la bande passante du signal mesurée en Mels et pas en Hertz. Par exemple pour une bande passante de 8000 Hz ce qui fait environ 2840 Mels et pour 24 filtres la distance entre ces filtres est d'environ 120 Mels.

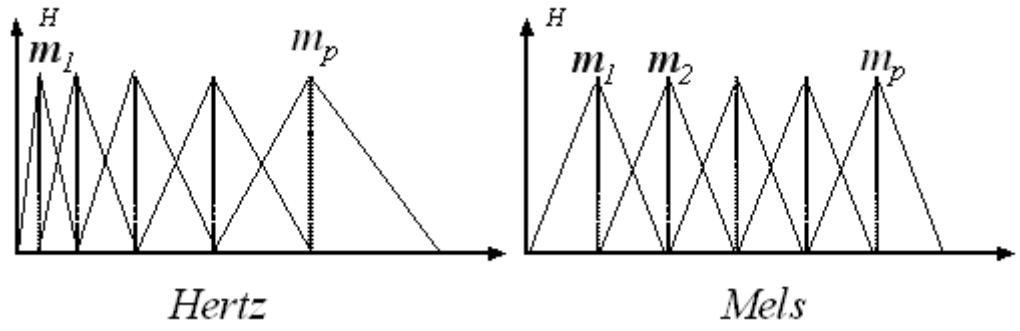

Fig.1-8 : répartition des filtres triangulaires sur les échelles fréquentielle et Mel

Selon [13] l'utilisation d'une analyse cepstrale à la place de la simple utilisation des amplitudes à la sortie des filtres produit des coefficients peu corrélés. Ceci est souhaitable si les coefficients sont ensuite utilisés par un système de reconnaissance basé sur des distributions gaussiennes avec des matrices de covariance diagonales. Les coefficients MFCC [9] d'une trame de parole sont calculés de la façon suivante :

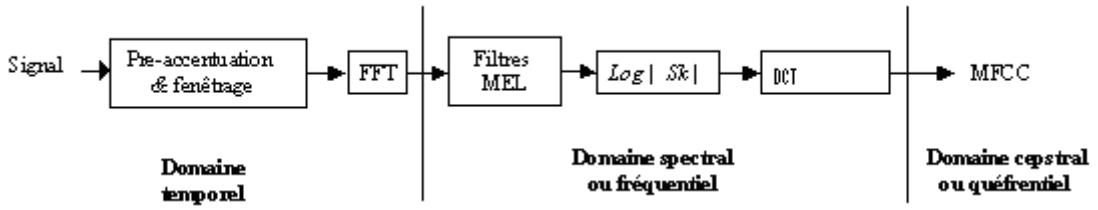

Fig.1-9 Calcul des coefficients MFCC

1. Apres le filtrage de pré-accentuation, le signal de parole est d'abord découpé en fenêtres, de taille fixe, reparties de façon uniforme le long du signal.
2. La FFT (Fast Fourier Transform) de la trame est calculée,ensuite, le calcul de l'énergie refait en élevant au carré la valeur de, la FFT, elle sera passée ensuite à travers chaque filtre Mel. Soit S_k l'énergie du signal à la sortie du filtre K , et nous aurons par la suite p_m (le nombre de filtres) paramètres S_k .

3. Le logarithme de S_k est calculé.
4. Finalement les coefficients sont calculés en utilisant la DCT (Discrete Cosine Transform)

$$C_i = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m_p} \log(S_k).\cos\left[i\left(k - \frac{1}{2}\right).\frac{\pi}{m_p} \right] \right\} \quad \text{Pour } i=1.....N \quad \ldots\ldots\ldots\ldots \quad (1.6)$$

où N est le nombre de coefficients MFCC.

Dans la littérature, le nombre de coefficients utilisés varie de 5 à plus d'une quarantaine en fonction de l'utilisation qui en est faite : reconnaissance de la parole, de la langue ou identification du locuteur par exemple. En ce qui concerne le nombre de filtres, nombreux sont ceux qui choisissent 30 pour un signal avec une bande passante de 0 à 8 KHz.

1.5.2 Prédiction linéaire (LPC) :

a) mesure de l'erreur de prédiction :

Le codage LPC consiste à estimes la valeur de l'échantillon à venir sur la base de quelques valeurs mesurées précédemment $x(n-i)$:

La valeur estimée $\tilde{x}(n)$ est calculée a partir des échantillons précèdent par des coefficients $a(i)$ qui sont généralement au nombre allant de 8à 12 :

$$\tilde{x}(n) = -\sum_{i=1}^{p} a(i)x(n-i) \quad \ldots\ldots\ldots\ldots(1.7)$$

La valeur des coefficients $a(i)$ s'obtient par minimisation de la variance δ_e^2 de l'écart $e(n)$ .Celle-ci est défini comme la différence entre la valeur estimée $\tilde{x}(n)$ et la valeur réelle $x(n)$:

$$e(n) = x(n) - \tilde{x}(n) = x(n) + \sum_{i=1}^{p} a(i)x(n-i) \ldots\ldots\ldots\ldots(1.8)$$

$$e(n) = \sum_{i=0}^{p} a(i)x(n-i). \quad \ldots\ldots\ldots\ldots(1.9)$$

$$a(0) = 1$$

La puissance on variance de l'écart de l'ensemble des N échantillons $e(n)$ à disposition avec $0 \le n \le N-1$ et alors la suivante :

$$\delta_e^2 = \frac{1}{N}\sum_{n=0}^{N-1} e^2(n) = \frac{1}{N}\sum_{n=0}^{N-1}\left(x(n) + \sum_{i=1}^{p} a(i)\, x(n-i)\right)^2 \quad \text{.....(1.10)}$$

b) Calcul de coefficients de prédiction linéaire [12] :

La procédure pour obtenir la valeur optimum des coefficients $a(i)$ consiste à rendre minimum la puissance de l'erreur commise lors de la prédiction. Un schéma fonctionnel traduisant cette démarche est présente dans la (figure1.8) :

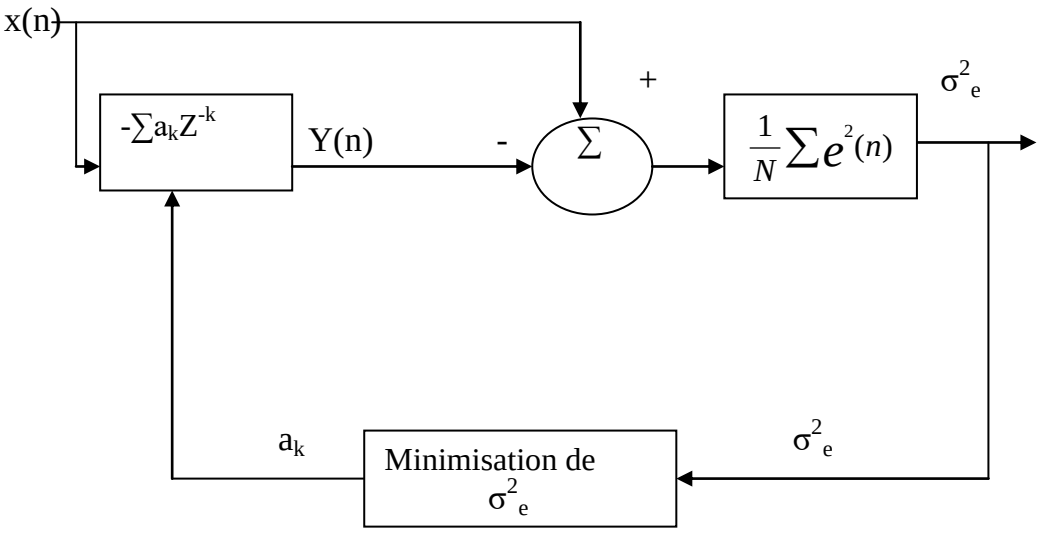

fig.1-8:schéma fonctionnelle de la prédiction linéaire

Mathématiquement, la variance est une fonction des paramètres de prédiction $a(i)$:

$$\delta_e^2 = \delta_e^2\left(a_1, a_2, \ldots\ldots\ldots a_p\right) = \delta_e^2(a_i). \quad \text{.....(1.11)}$$

Sa valeur minimale s'obtient lorsque l'ensemble des dérivées partielles de δ_e^2 par apport aux paramètres $a(i)$ sont nulles :

$$\delta_{e,\min}^2 = \frac{d\delta_e^2(a(i))}{da(i)} = 0 \qquad i = 1,2,\ldots\ldots\ldots p \quad \text{.....(1.12)}$$

Le calcule de ces p dérivés partielle conduit à p équations pour p paramètres inconnus $a(i)$:

$$a_1 r_{xx}(0) + a_2 r_{xx}(-1) + \dots\dots\dots\dots + a_p r_{xx}(1-p) = -r_{xx}(1)$$
$$a_1 r_{xx}(1) + a_2 r_{xx}(0) + \dots\dots\dots\dots + a_p r_{xx}(2-p) = -r_{xx}(2)$$

$$.a_1 r_{xx}(p-1) + a_2 r_{xx}(p-2) + \dots\dots\dots\dots + a_p r_{xx}(0) = -r_{xx}(p)$$

Avec : $r_{xx}(i) = \sum_{n=0}^{N-1} x(n)x(n-i)$ $\quad i = 1,\dots\dots\dots p$ ……………….(1.13)

Les coefficients des paramètres $a(i)$ sont les p premières valeurs de la fonction d'autocorrelation $r_{xx}(i)$ du signal $x(n)$ comportent N échantillons, cet ensemble d'équations linéaire peut s'écrire sous forme matricielle :

$R_{xx} a = -r_{xx}$ ……………………………………………………………(1.14)

R_{xx} : Matrice $p \times p$ d'auto-corrélation.

r_{xx} : Le vecteur $p \times 1$ d'auto-corrélation.

a : Le vecteur $p \times 1$de paramètres de prédiction.

On constate que la matrice R_{xx} est symétrique, et que les diagonales parallèles a principale, contiennent des éléments égaux, une telle matrice est dite Toeplitz. Il existe dans ce cas des algorithmes rapides pour la résolution appelés de Shure, burg, Levinson-Durbin.

La méthode du codage par prédiction linéaire est tout autant utilisée en RAP qu'en compression pour le transfert de la voix par téléphone ou radio. Elle n'est cependant pas parfaite puisque l'erreur de prédiction peut être importante sans qu'il soit possible, de la corriger, par cette méthode.

La méthode RELP, *Residual Excited Linear Prediction*, permet de réduire une partie de cette erreur. Le principe consiste à comparer, lors de la prédiction linéaire, le signal obtenu avec le signal original. L'erreur, obtenue par soustraction, représente la partie du signal original que le prédicteur n'arrive pas à modéliser. Dans la méthode RELP, l'erreur résiduelle est passée dans un filtre passe-bas permettant de conserver l'erreur effectuée dans la seule bande fréquentielle allant de 0 à 1000 hertz. La sortie du filtre est alors codée et passée au receveur qui peut alors reconstruire un signal à partir de la prédiction et de l'erreur observée.

Pour pallier le problème de l'erreur résiduelle, d'autres méthodes fondées sur la prédiction linéaire ont été développées. Ainsi la méthode *CELP, Code Excited Linear Prediction*, permet d'effectuer une compression de la parole par codage d'une trame vis-à-vis de références stockées dans un corpus, Ainsi, une trame de parole sera codée selon une combinaison linéaire de certaines trames du corpus et c'est cette combinaison linéaire qui sera considérée à la place de la trame dans les traitements ultérieurs. Cette méthode de codage de la parole est surtout employée pour la compression et la transmission de la parole à de faibles débits.

L'idée du codage prédictif linéaire n'a pas encore été abandonnée malgré son apparente simplicité et l'évident taux d'erreur introduit par l'hypothèse de linéarité de la production de la parole. Le groupe en charge de l'étude du GSM ("Groupe Spécial Mobile" devenu depuis "Global System for Mobile"), après avoir étudié différents systèmes de codage de la parole sur des critères de qualité subjective, de complexité algorithmique et de besoin en bande passante, a retenu le codage prédictif linéaire dit RPE-LPC (*Regular-Pulse Excited - Linear Predictive Coding*) agrémenté d'un système itératif de prédiction à long terme [scourias95]. Cet ensemble algorithmique permet de transmettre un signal de parole de bonne qualité à des taux de transfert de 13,2 kbps (kilobits par seconde). Ce choix va cependant à l'encontre des tendances actuelles de codage de la parole par des méthodes permettant de conserver une qualité objective au signal de parole lors de sa transmission.

1.5.3 PLP :

La méthode *PLP*, [14,15,16], *Perceptual Linear Prediction* (ou *Perceptually based Linear Prediction*), est une méthode inspirée du principe de prédiction linéaire. Elle combine ce principe à une représentation du signal qui suit l'échelle humaine de l'audition. Elle est à l'origine de toute une famille de techniques de traitement du signal de parole que nous verrons dans le paragraphe suivant.

Cette méthode peut être résumée en trois phases de traitements successifs. Le signal de parole est tout d'abord analysé pour obtenir un spectre suivant une échelle d'audition. Ce spectre est ensuite modifié par une interpolation et une transformée de Fourier inverse, le signal obtenu étant passé dans un filtre pour réduire la dimension du spectre

et augmenter la résolution fréquentielle. Une troisième étape, qui peut être omise, permet de reconstruire un signal de parole par filtrage inverse, passage dans le domaine fréquentiel hertzien et désaccentuation.

- La première étape est précisément constituée par :
  - Une analyse en bandes critiques selon une échelle Bark par un banc de filtres,
  - Une préaccentuation des valeurs obtenues selon une courbe suivant approximativement, les même principes que les traitements effectués par l'oreille, avec accentuation des basses fréquences et atténuation des hautes fréquences,
  - Une application de la loi de préaccentuation de Steven.

- La deuxième étape est, constituée des phases suivantes :

  - Une interpolation des sorties des filtres du banc pour obtenir un spectre sur Une échellefréquentielle auditive,
  - Une transformée de Fourier inverse qui permet de ramener le spectre obtenu dans le domaine temporel,
  - Une résolution d'un ensemble d'équations linéaires pour obtenir les coefficients issus d'un filtre tout pôle d'ordre 5 (ce qui permet d'obtenir au moins deux sommets caractéristiques selon [14]).

Cette méthode a pour avantage de permettre une analyse et/ou un codage de la parole qui respecte le principe de la prédiction linéaire, qui l'échelle fréquentielle observable dans l'oreille et, enfin, qui réduisent l'espace de représentation. Cette méthode a été, par la suite, améliorée pour résister à certaines conditions de bruit.

1.5.4 Rasta PLP :

La méthode PLP [15], dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de

distorsions spectrales linéaires, [17,18] propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP, RASTA étant l'acronyme de *RelAtive SpecTrAl* [18]. La mise en place de ce filtrage permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication.

1.6 Conclusion :

Nous avons brièvement parcouru dans ce chapitre les caractéristiques du signal de la parole et les méthodes utilisées pour extraire les paramètres acoustiques qui seront fournis au système de reconnaissance .Nous n'avons fait que survoler ces techniques de calcul des coefficients représentatifs du signal vocal. Nous utiliserons principalement les coefficients LPC, et MFCC dans les expériences d'écrites dans ce travail. La justification de l'utilisation des coefficients LPC, MFCC est leur indépendance vis a vis du canal de transmission, et donc leur possible utilisation dans un système de démonstration sans dégradations significatives des taux de reconnaissance par l'utilisation d'un microphone différent de ceux utilisés lors de l'enregistrement des bases de données ayant servi a l'entraînement des modèles.

2 Segmentation et Reconnaissance de la parole

Sommaire

2.1 Introduction :

Selon le Larousse, le terme segmentation désigne "la division d'un ensemble en portions bien délimitées". Autrement dit, c'est le processus de division d'une entité, généralement continue, en petites entités appelées segments. Chaque segment possède des Propriétés propres qui permettent de le différencier des autres.

La segmentation phonétique du signal de parole peut être effectuée soit manuellement par un expert humain, soit automatiquement par une méthode programmée (voir section2.2).

D'un point de vue qualitatif, l'examen de l'état de l'art de la segmentation de la parole donne la préférence à la segmentation manuelle. En effet, bien qu'il soit difficile d'évaluer la qualité d'une segmentation phonétique, il existe un large consensus sur le fait qu'une segmentation manuelle est plus précise qu'une segmentation automatique.

D'autant plus que des logiciels disponibles, dotés d'interfaces graphiques conviviales et interactives, représentant le signal de parole et ses caractéristiques temporelles et fréquentielles, avec une sortie audio pour l'écoute, permettent à l'expert de générer, d'une manière relativement aisée, la séquence phonétique alignée sur le signal de parole.

Lorsqu'il s'agit de segmenter un corpus d'une dizaine ou d'une centaine de phrases, ce processus est envisageable. Mais les besoins permanents en corpus de parole segmentée et la taille grandissante de ces corpus éliminent d'office ce type de segmentation pour son coût exorbitant. Outre cet inconvénient, la segmentation manuelle souffre tout de même d'une variabilité inter et intra-segmenteurs (inconsistante et non-reproductible). De tous ces points découle l'intérêt majeur de la segmentation automatique de la parole.

La segmentation est la première étape de système de reconnaissance.

Avant de commencer mes recherches, une étude bibliographique a été réalisée afin de bien cerner et maîtriser les différents sujets qui seront abordés. J’ai introduis d’abord la segmentation de la parole, et par la suite sera décrit le fonctionnement des

reconnaissances de la parole. Ce chapitre donne un bref aperçu des théories sur lesquelles s'appuient les travaux effectués.

2.2 Segmentation de la parole :

La segmentation est une approche pour le traitement des signaux présentant des non stationnarités « rapides » ou « ruptures », et constitue une première étape possible de traitement en vue de la reconnaissance ou du diagnostic, voire même du codage. Cette approche consiste à considérer que le signal est une succession de segments homogènes, de caractéristiques constantes ou lentement variables, séparés par des transitions brutales, où les caractéristiques du signal changent rapidement .Les algorithmes de segmentation de signaux, qui utilisent essentiellement des modèles paramétriques, sont des algorithmes qui effectuent une telle décomposition d'un signal en segments successifs, et ce avec un degré de précision variable. Certains se préoccupent uniquement de détecter (évidemment avec un retard) qu'un changement s'est produit ; d'autres raffinent la position de la frontière de segment trouvée en estimant, après détection, l'instant auquel la rupture s'est produite ; d'autres enfin vont jusqu'à estimer les paramètres caractéristiques du signal avant et après la rupture. Le formalisme sous-jacent à la construction de ces algorithmes est du type : « ***modèle paramétrique où les valeurs des paramètres changent instantanément à un instant inconnu*** ». On exclut intentionnellement ici la détection de changement d'ordre, de changement de type de distribution. .

Cependant, il est important de remarquer que les outils permettant de détecter de telles ruptures - comme les débuts et fins de voisement et de fricatives en parole - permettent aussi de détecter en pratique des « sauts » se produisant non pas instantanément mais en plusieurs échantillons, voire des transitions moins « franches » (telles que les débuts ou les ruptures de pente de certains formants en parole).

Pour cette raison, et à des fins de comparaison de performances, sont également inclus dans ce document trois algorithmes qui sont plutôt du type ***détecteurs de changement de forme*** dans un signal que du type détecteur de changement de modèle, mais qui permet de détecter ce dernier type d'événements.

L'intuition du praticien du traitement de signal conduit à distinguer, dans ce contexte paramétrique, deux types de ruptures : ***additives et spectrales***. Il est en effet visuellement Évident que les « sauts » ou changements brusques de la valeur moyenne locale dans un signal sont qualitativement différents des changements dans le « degré d'agitation locale » autour d'une moyenne.

2.2.1 Cadre général pour la structure des algorithmes :

Il convient de désigner les algorithmes présentés dans le présent document par un :

- **M** : pour les détecteurs de sauts de moyenne ;
- **D** : pour les détecteurs de sauts de dynamique ;
- **F** : pour les détecteurs de changements de forme.

Plusieurs Auteurs, comme e BASSEVILLE,Christian DONCARLI, Marie-Françoise LUCAS, Denis DE BRUCQ, Abdelahad BENHALLAM,[1] , … on divisera toutes les méthode en trois catégories : **M ,D, F**

L'exception de GLR (Generalized liklihood ratio) qui peut également traiter des signaux multi-capteurs . Les exemples d'application donnés dans chaque fiche sont ceux sur lesquels l'algorithme concerné a été effectivement Utilisé par le concepteur et/ou par nous même.

2.2.2 Conception d'algorithmes de segmentation :

Commençons par quelques remarques sur les fils directeurs qui conduisent à la conception des algorithmes de segmentation de signaux, qu'elle soit ou non fondée sur la modélisation présentée plus haut.

Sur le plan des intuitions, il paraît naturel à beaucoup de chercher à détecter un changement en surveillant de manière adéquate **l'innovation** e_n d'un filtre « qu'il soit ajusté localement une fois pour toutes ou réactualisé de temps en temps ou en permanence », en testant :

- un changement de sa moyenne (algorithmes GLR, GLRmod, EPL, DAREP),
- un changement de sa variance (algorithmes FKE-M, SNR, DAREP, DIS2f, Brandt),
- une perte de sa blancheur (algorithme FKE-D) .

Une autre idée intuitivement naturelle consiste à comparer convenablement deux grandeurs estimées respectivement à long terme (ou globalement) et à court terme (ou localement) sur le signal ; ces quantités comparées peuvent être :

- une énergie résiduelle (algorithmes FKE-M, SNR, DIS2f, Brandt),
- une corrélation résiduelle (algorithme FKE-D),
- des innovations (algorithme DIV).

Du point de vue méthodologique, l'outil de base est le ***rapport de vraisemblance généralisé*** classique en détection et dans lequel les paramètres inconnus (instant de rupture, Paramètres du modèle avant et après rupture) sont estimés par le maximum de vraisemblance . Les algorithmes de ce document qui utilisent cet outil sont tous conçus sous hypothèse gaussienne. Les autres ingrédients possibles sont, au niveau du prétraitement, des intégrateurs ou des opérateurs de dérivation, et, au niveau de la règle de décision, des tests de blancheur ou des mesures de distances spectrales.

2.2.3 Structure des algorithmes de type M [19, 20, 21,22]:

Ces algorithmes effectuent un premier traitement du signal et travaillent ensuite sur soit :

1. la sortie d'un filtre (passe-bas pour l'algorithme DF),
2. la sortie d'un intégrateur (HK),
3. l'innovation d'un filtre (algorithmes EPL, GLR, GLRmod, FKE-M) et leur appliquent, en vue de la détection de:
    1. un seuillage ((EPL), éventuellement précédé d'une dérivation (DF), ou bien adaptatif (HK), ou bien joint à un compteur (DF)),
    2. un double calcul d'énergies résiduelles à long et court terme (FKE-M),
    3. une corrélation avec une signature de rupture supposée (GLR et GLRmod) .

La détection est suivie d'au moins l'une des actions suivantes :

1. estimation de l'instant de rupture (HK, GLR, GLRmod) ; à défaut, l'instant de rupture estimée est l'instant de détection ;
2. estimation de l'amplitude de la rupture détectée (GLR, GLRmod, FKE-M) ;
3. réactualisation du filtre à l'aide de l'amplitude de saut estimée (GLR, GLRmod, FKE-M) ;
4. redémarrage de l'algorithme de détection : soit immédiatement après la détection si on a un moyen de réactualiser (GLR, GLRmod) ; soit après un temps suffisant pour permettre au filtre de « re-converger », c'est à dire au minimum de pouvoir être calculé (DF, HK, FKEM).

2.2.4 Structure des algorithmes de type D [19, 20, 22, 23,24] :

Ces algorithmes travaillent essentiellement avec une modélisation paramétrique de type autorégressive AR ou autorégressive à moyenne glissante ARMA. Ils effectuent une estimation de tel(s) filtres ajustés à court et/ou long terme selon les cas, sur une ou plusieurs fenêtres de type

1. fixe (SNR, EPL, DIS2f),
2. de taille fixe glissant de manière continue (DIS2f, DIV, Brandt) ou discontinue (DAREP, DIS2f),
3. croissante (FKE-D, DIV, Brandt)

et calculent :

a) l'innovation ou erreur de prédiction :

$$\varepsilon_n = Y_n - \sum_{i=1}^{p} \hat{a}_i \cdot Y_{n-i} - \sum_{j=1}^{q} \hat{b}_j \cdot \varepsilon_{n-j} \quad \text{.....(2.1)}$$

b) l'énergie résiduelle $\sigma^2 = \text{var}(\varepsilon)$ qui sont ensuite testées et/ou comparées à leurs homologues, à l'aide d'une mesure de distance spectrale convenable ; ce qui donne la détection.

Cette détection est éventuellement suivie d'une estimation plus fine de l'instant de rupture (DIV, Brandt, FKE-D) . Un moyen très simple d'y parvenir consiste à remarquer que la « distance » réagit à un changement dans le signal par un changement de sa dérive qui, de nulle devient positive . Il suffit alors d'adjoindre un détecteur CUSUM qui détecte ce changement de dérive et en estime l'instant d'occurrence. La détection est également souvent suivie d'une procédure de redémarrage après détection (en général, temps d'attente de l'ordre de la taille de la fenêtre pour permettre à au moins un des filtres de « converger »).

On peut définir des variantes de ces algorithmes en :

- modifiant le nombre des fenêtres,
- la position des fenêtres,
- le choix des méthodes d'identification à l'intérieur de chacune des fenêtres (car elles peuvent influencer le comportement de la règle de décision),
- le choix de la mesure de distance entre les modèles ainsi estimés.

2.2.5 Structure des algorithmes de type F :

Les algorithmes de type F ont pour objectif de détecter une rupture de forme dans une suite de signaux déterministes éventuellement présegmentées.

Par définition, la forme d'un profil est invariante par translation et changement d'échelle sur les abscisses et les ordonnées. Le terme de « forme » est parfois confondu avec celui de « gabarit » où les paramètres d'échelle, et en particulier de largeur, sont fixés. Les algorithmes de type F peuvent être paramétriques : ils dépendent alors d'hypothèses a priori et n'utilisent pas toute l'information contenue dans la forme.

Plus couramment, les algorithmes de type F sont non paramétriques : ils sont conçus initialement soit pour détecter un changement de gabarit, soit pour détecter directement un changement de forme. Cependant, une reconnaissance de gabarit peut devenir une reconnaissance de forme en adjoignant un algorithme d'ajustement des paramètres d'échelle. De la même façon, une reconnaissance de forme peut être restreinte à une reconnaissance de gabarit si l'on fixe les paramètres d'échelle.

2.2.6 Segmentation du signal de parole en mots :

Il y'a deux approches complémentaires des processus de segmentation du signal de parole en mots. La première est fondée sur des processus d'accès au lexique aboutissant, par le biais de procédures de sélection ou de compétition lexicales, à la localisation de frontières entre les mots de la chaîne parlée. La seconde accentue le rôle d'informations prélexicales comme les régularités phonologiques (liées aussi bien à des indices segmentaux que suprasegmentaux tels que la prosodie et les contraintes phonotactiques) ou distributionnelles dans la mise en place d'hypothèses sur la localisation probable de frontières de mots. Les deux dernières études présentées ont conduit leurs auteurs respectifs à interpréter les résultats obtenus comme des preuves du recours à des connaissances sur les régularités phonotactiques ou syllabiques de la langue dans les processus de segmentation de la parole en mots. Il est cependant difficile de conclure de manière définitive à un rôle des informations phonologiques sans avoir une description adéquate des concepts de contraintes phonotactiques ou de syllabation et de leur expression dans la langue en termes de régularités probabilistes.

Chaque langue du monde présente des caractéristiques qui la distinguent des autres langues. Dans le cas de la structure sonore de la langue, objet d'étude de la phonologie, ces spécificités concernent aussi bien l'inventaire des sons individuels que celui des séquences de sons qui sont attestées dans cette langue. Les théories développées en phonologie ont pour objet de décrire les mécanismes qui permettent de transformer une représentation phonologique dite sous-jacente (représentation abstraite de la séquence de phonèmes à prononcer) en une forme de surface (la forme phonétique) correspondant à ce qui est effectivement prononcé. Les modèles proposés définissent donc deux composantes essentielles : la forme des représentations sous-jacentes et les procédures permettant de transformer ces représentations en forme phonétique de surface. Ces règles peuvent être universelles -elles existent dans toutes les langues- ou paramétriques -auquel cas seules certaines langues en sont pourvues. Les représentations phonologiques et les règles qui sont décrites en phonologie peuvent être assimilées à un ensemble de 'connaissances implicites' qui serait stocké par le système cognitif du locuteur.

Nous présentons alors l'un des concepts centraux, aussi bien dans le cadre de la phonologie que dans celui des travaux sur la perception de la parole : la syllabe. Cette présentation nous permet de mettre au jour un lien étroit entre la conception hiérarchique propre au concept de syllabe et le caractère linéaire des contraintes phonotactiques qui constituent l'une des sources d'information pour la syllabation des séquences de phonèmes.

2.2.6.1 La syllabe [25] :

La syllabe est difficile à cerner, pour une bonne raison : elle varie selon la langue à analyser. Plusieurs approches sont possibles pour tenter de la définir. On peut, pour l'instant, se contenter de dire que c'est une unité phonétique plus grande que le phonème et plus petite que le mot et qu'un locuteur lambda est capable de découper un mot en syllabes dans sa langue, sans forcément savoir comment il procède. Un mot est donc composé de phonèmes, qui forment des syllabes.

2.2.6.2 Éléments d'une syllabe :

Une syllabe peut se décomposer en trois éléments (figure2.1):

- **attaque** : consonne(s) précédant le sommet
- **rime**, elle même composée de :
  - **noyau** : sommet de la syllabe ;
  - **coda** : consonne(s) suivant le noyau.

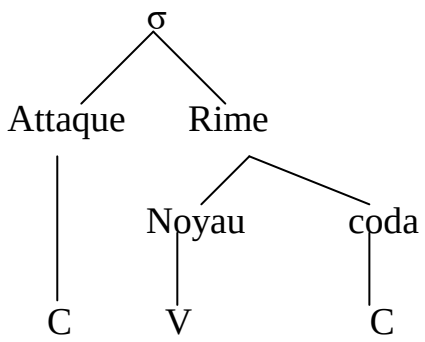

Fig. 2-1 : Représentation hautement hiérarchique de la syllabe en constituants phonémiques remplissant.

La syllabe française *dextre* [dɛkstʁ̥] s'analyse donc ainsi :

- attaque : [d] ;
- noyau : [ɛ] ;
- coda : [kstʁ̥].

On dit d'une syllabe possédant une coda qu'elle est *fermée*, sans coda qu'elle est *ouverte*.

2.2.6.3 Déterminants de la structuration syllabique :

2.2.6.3.1 Principes d'organisation indépendants des caractéristiques des Phonèmes :

Plusieurs approches ont été proposées afin de dériver une séquence syllabiquement structurée à partir d'une chaîne linéaire de phonèmes.

a) Approche par règles [26,27] :

L'une des approches concernant le mécanisme de structuration syllabique d'une chaîne phonémique consiste à proposer des règles permettant de transformer une forme d'entrée composée d'une chaîne linéaire de segments en une séquence de segments qui acquerraient un statut spécifique dans la hiérarchie syllabique.

Principe de l'Attaque Obligatoire :

Hooper (1972) propose un principe de syllabation selon lequel toute syllabe doit, tant que cette procédure ne viole pas les contraintes phonotactiques de la langue, comporter une attaque.

Dans une séquence CVCVC (par exemple /pətit/), ce principe conduit à une syllabation de la suite de phonèmes en CV#CVC car, toute syllabe devant comporter une attaque, la consonne médiane se situe nécessairement à l'attaque de la seconde syllabe. Cette opération ne s'applique cependant pas si la suite médiane CV constitue une séquence phonotactiquement illégale dans la langue. Pour des séquences contenant plus d'une consonne médiane, ce principe n'est pas suffisant. Ainsi dans une suite CVC_1C_2VC, C_2 doit nécessairement appartenir à la seconde syllabe, lui permettant de disposer d'une attaque. Par contre, le statut de C_1 est indéterminé. Elle peut tout aussi bien constituer la coda de la première syllabe que se regrouper avec C_2 à l'attaque de la

seconde syllabe. Il est par ailleurs impossible de choisir une solution définitive *a priori* qui consisterait à placer C1, quelle qu'elle soit, en coda (ou l'inverse) puisqu'en

français certaines séquences CVCCVC se syllabent en CVC # CVC (par exemple, 'taxi', /tak#si/) alors que d'autres présentent une structure CV # CCVC (par exemple, 'sacré', /sa#kre/). Des règles supplémentaires sont donc nécessaires pour améliorer la validité de ce principe.

✸ *Principe de l'Attaque Maximale :*

L'objet du Principe de l'Attaque Maximale (Selkirk, 1982; cf. Goldsmith, 1990 pour une revue) est de fournir les fondements d'une prédiction plus complète des structures syllabiques observées en fonction de la séquence linéaire de phonèmes considérée. Conformément au *Maximum Onset Principle*, l'on doit insérer un nombre maximal de consonnes en position d'attaque syllabique, ceci à condition que cette opération ne viole pas les contraintes phonotactiques de la langue. Dans une séquence CVCCCVC (par exemple, /syrpriz/), il devient possible de décider de la localisation adéquate de la frontière en fonction de la suite de consonnes médiane. L'*Obligatory Onset Principle* ne fournit pas d'information quant au nombre de consonnes à placer en attaque de syllabe, il n'est donc pas possible de choisir entre les diverses configurations possibles. Le *Maximum Onset Principle*, au contraire, contraint un maximum de consonnes à occuper cette position. Ainsi, dans la suite CVCCCVC, les trois consonnes médianes seront positionnées à l'attaque de la seconde syllabe pour autant que les contraintes phonotactiques de la langue le permettent. Si, dans une séquence $CVC_1C_2C_3VC$ provenant d'une langue donnée, les suites C_1C_2, C_2C_3 et C_3V sont phonotactiquement légales, alors la séquence de phonèmes sera découpée en CV # $C_1C_2C_3VC$. Par contre, si C_2C_3 constitue un groupe de consonnes illégal dans cette langue, la suite de segments se découpera en CVC_1C_2 # C_3VC. Le *Maximum Onset Principle* fournit donc une méthode de découpage syllabique plus contrainte que l'*Obligatory Onset Principle*. Elle ne permet cependant pas de prédire intégralement la structuration syllabique de l'ensemble des mots possibles.

2.3 Reconnaissance de la parole :

Le problème de la reconnaissance automatique de la parole consiste à extraire, à l'aide d'un ordinateur, l'information lexicale contenue dans un signal de parole,et éventuellement de l'interpréter .depuis plus de quatre décennies, de nombreux laboratoires internationaux ont mené des recherche intensives dans ce domaine et des progrès importants ont été réalisés ,notamment grâce au développement d'algorithmes puissants,alliés aux technologies de traitement numérique du signale .de nombreux systèmes de reconnaissance de la parole sont maintenant disponibles ou peuvent être développés,couvrant des domaines aussi vastes que la reconnaissance de quelques mots clés sur lignes téléphoniques,les systèmes à dicter vocaux, les systèmes de commande et contrôle sur pc,et allants jusqu'aux systèmes de compréhension du langage naturel

L'appellation "reconnaissance de la parole" (ASR pour Automatic Speech Recognition en anglais) se réfère à plusieurs types de systèmes dont la mission est de décoder l'information portée par le signal vocal. Selon l'information à extraire, on distingue deux types de reconnaissance :

- La reconnaissance du locuteur, dont le but est de reconnaître la personne qui parle permis une Population de locuteurs (identificateur) ou de vérifier son identité (vérificateur). On sépare les cas de reconnaissance dépendante du texte, avec texte dicte ou indépendante de ce dernier.
- La reconnaissance de parole, dont le but est de transcrire l'information symbolique exprimée par le locuteur. On distingue les cas de reconnaisseur monolocuteur, multilocuteur ou indépendant du locuteur. Une distinction est également faite entre reconnaisseur de mots isoles, de mots connectes et de parole continue. Ces taches demandent la contribution d'outils techniques aussi divers que puissants : traitement du signal, modèles mathématiques statistiques, algorithmique,... De plus, les cas d'applications sont toujours plus complexes que la théorie, dans la mesure où ils introduisent des bruits extérieurs, la contribution du matériel utilisé, les différences de locutions...

2.3.1 Définition [28,29]:

La reconnaissance automatique de la parole est l'un des deux domaines du traitement automatique de la parole, l'autre étant la synthèse vocale. La reconnaissance automatique de la parole permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale : voir plus loin). En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis, comme un humain le ferait. Ces deux domaines et notamment la reconnaissance vocale, font appel aux connaissances de plusieurs sciences : l'anatomie (les fonctions de l'appareil phonatoire et de l'oreille), les signaux émis par la parole, la phonétique, le traitement du signal, la linguistique, l'informatique, l'intelligence artificielle et les statistiques.

Il faut bien distinguer ces deux mondes : un système de synthèse vocale peut très bien fonctionner sans qu'un module de reconnaissance n'y soit rattaché. Evidemment le contraire est également tout à fait possible. Par contre, dans certains domaines bien précis, l'un ne va pas sans l'autre. Il est bien entendu que l'étude se portant sur la reconnaissance automatique de la parole, l'autre aspect du traitement de la parole ne sera pas traité dans cette thèse.

Le traitement automatique de la parole ouvre des perspectives nouvelles compte tenu de la différence considérable existant entre la commande manuelle et vocale. L'utilisation du langage naturel dans le dialogue personne/machine met la technologie à la portée de tous et entraîne sa vulgarisation, en réduisant les contraintes de l'usage des claviers, souris et codes de commandes à maîtriser. En simplifiant le protocole de dialogue personne/machine, le traitement automatique de la parole vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse. De plus, il rend possible l'utilisation simultanée des yeux ou des mains à une autre tâche. Il permet d'humaniser les systèmes informatiques de gestion de l'information, en axant leur conception sur les utilisateurs.

A la base, les logiciels de reconnaissance vocale servent surtout à entrer du texte en masse tout en se passant du clavier (qui offre un débit de 50 mots par minute contre plus de 150 pour la parole), le clavier reste cependant encore nécessaire aux corrections de texte et à l'utilisation de l'ordinateur.

2.3.2 Historique [30, 31,32] :

Ce chapitre va montrer l'évolution de la reconnaissance automatique de la parole depuis ses débuts jusqu'à nos jours.

a) La naissance

Les premières tentatives de création d'une machine capable de comprendre le discours humain eurent lieu aux USA à la fin des années 40, au sein du Ministère de la Défense américain. Le but était de traduire et d'interpréter des messages russes interceptés. Ces premières expériences s'appuyaient sur une approche descendante, c'est-à-dire fournissant une recherche mot à mot. Pendant ces premières années de vie de la reconnaissance vocale, il a fallu énormément de temps et de ressources informatiques pour enregistrer et emmagasiner la représentation de chaque mot dans chaque langue. Malgré tous les efforts fournis, les résultats sont médiocres et peu fiables, mais laissaient la porte ouverte à de nombreuses recherches.

b) Les premiers mots :

Vers 1950 apparaît le premier système de reconnaissance de chiffres, appareil entièrement câblé et très imparfait.

En 1951, *S.P. Smith* présente un détecteur de phonèmes ; une année après, *K.H. Davis, R Biddulph et S.Baleshek* annoncent la première machine à aborder la reconnaissance de manière globale : les dix chiffres «zero» à «nine» sont reconnus analogiquement avec un bon taux de réussite pour une seule voix.

Puis en 1960, *P.B. Denes* et *M.V. Matthews*, pour reconnaître les dix premiers chiffres, comparent globalement les représentations temps fréquence, numérisées et normalisées en durée totale : le taux d'erreur est nul pour un seul locuteur et s'élève à 6%

Pour cinq locuteurs ayant participé à un apprentissage. *H.F. Olson* et *H.Belar* envisagent, en 1961, la reconnaissance d'unités phonétiques autres que les phonèmes : leurs unités sont des «syllabes phonétiques» que le locuteur doit articuler séparément ou, du moins, avec une chute importante du niveau sonore en guise de séparation ; il s'agit donc presque d'une reconnaissance par mots, étant entendu que ces «mots» sont courts et que leur répertoire est limité : 2000 syllabes suffisent à couvrir 98% des besoins de la langue anglaise[31].

J.Dreyfus-Graf met au point en 1961 son «phonétographe», appareillage analogique composé de vingt filtres passe-bande et de circuits identificateurs de phonèmes. Le phonétographe utilise des «compresseurs sélectifs» qui augmentent l'émergence de certains sons. Obtenu en temps réel, le résultat est spectaculaire ; cependant l'appareil ne fonctionne qu'avec un seul locuteur qui doit adapter sa diction à la machine : hauteur, intensité, rythmes très faibles. Après avoir constaté que l'identification des phonèmes dans le signal de parole est un problème beaucoup plus compliqué qu'ils ne l'imaginent, les chercheurs se tournent, entre 1965 et 1970, d'une part vers la reconnaissance par mots isolés en vue d'applications pratiques comme la commande vocale, d'autre part vers l'utilisation d'informations de niveau linguistique supérieur avec lexique et syntaxe, pour compléter le message vocal reconnu au niveau phonétique. Cette seconde approche prend le nom, quelque peu abusif, de «compréhension automatique de la parole».

c) L'avancée des années 70 :

Les années 70 sont une période charnière. D'abord, elle voit la première réalisation commerciale en reconnaissance vocale : «le Voice Command system» de *J.J.W. Glenn* et *M.H. Hitchcok*, appareil autonome qui reconnaît de manière fiable 24 mots isolés après cinq cycles d'apprentissage par le même locuteur. L'analyse du message est effectuée par un banc de seize filtres ; chaque mot est représenté par huit événements prélevés aux instants de plus grande variation interne du message. Cette normalisation temporelle, ainsi que les traitements d'apprentissage et de reconnaissance, sont confiés à un mini calculateur incorporé.

Aux Etats-Unis, l'importance des recherches sur la parole a beaucoup varié au cours des dernières années. A l'effort de recherche particulièrement intensif correspondant au projet SUR (Speech Understanding Research, Recherche sur la compréhension de la parole) de l'Arpa (Advanced Research Projects Agency ou Agence de projets de
recherche avancés), succède maintenant un effort plus mesuré.

En ex-URSS, les recherches dans ce domaine ont commencé très tôt et restent à l'heure actuelle très actives. Mais à la différence des équipes américaines qui ont développé rapidement d'énormes systèmes de compréhension de la parole, les équipes soviétiques n'ont que très récemment abordé l'étude des niveaux syntaxique et sémantique ; elles sont à l'origine de l'utilisation de la technique de «programmation dynamique» dont l'emploi s'est maintenant partout généralisé.

En France, les recherches ont démarré vers 1970, et plusieurs laboratoires de recherches ont pu mettre au point différents systèmes de reconnaissance vocale avec plus ou moins de succès, ces laboratoires mettant l'accent sur le support de reconnaissance : mots isolés, syllabes, grands vocabulaires...

d) La reconnaissance du langage :

Dès lors, les recherches dans le domaine de la reconnaissance de la parole n’ont cessé de progresser dans le sens de la compréhension du langage parlé et des phrases structurées.

Aujourd’hui, le taux d’erreur ainsi que le temps d’apprentissage des systèmes de reconnaissance ne cesse de diminuer pour atteindre de nos jours des résultats proche de 95%. Ce taux est évidement variable selon la difficulté du langage. En effet la machine a parfois du mal à éviter certains pièges linguistiques.

Néanmoins nous verrons plus loin que de nos jours nous disposons d’une technologie très aboutie.

e) Dates clés :

On peut résumer en quelques dates les grandes étapes de la reconnaissance de la parole :

- 1952 : reconnaissance des 10 chiffres, par un dispositif électronique Câblé, pour un mono-locuteur
- 1960 : utilisation des méthodes numériques
- 1965 : reconnaissance de phonèmes en parole continue
- 1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots)
- 1969 : utilisation d'informations linguistiques
- 1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
- 1972 : premier appareil commercialisé de reconnaissance de mots
- 1976 : fin du projet ARPA ; les systèmes opérationnels sont HARPY, HEARSAY I et II et HWIM
- 1978 : commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés
- 1981 : utilisation de circuits intégrés VLSI (Very Large Scale Integration) spécifiques du traitement de la parole
- 1981 : système de reconnaissance de mots sur un circuit VLSI
- 1983 : première mondiale de commande vocale à bord d'un avion de chasse en France
- 1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
- 1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réel
- 1988 : apparition des premières machines à dicter par mots isolés
- 1989 : recrudescence des modèles connexionnistes neuromimétiques

- 1990 : premières véritables applications de dialogue oral homme-machine
- 1994 : IBM lance son premier système de reconnaissance vocale sur PC
- 1997 : lancement de la dictée vocale en continu par IBM.

2.3.3 Problématique

Pour bien appréhender le problème de la reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile.

- **Le système doit-il être optimisé pour un unique locuteur ou est-il destiné à devoir se confronter à plusieurs utilisateurs ?**

On peut aisément comprendre que les systèmes dépendants d'un seul locuteur sont plus faciles à développer et sont caractérisés par de meilleurs taux de reconnaissance que les systèmes indépendants du locuteur étant donné que la variabilité du signal de parole est plus limitée. Cette dépendance au locuteur est cependant acquise au prix d'un entraînement spécifique à chaque utilisateur. Ceci n'est néanmoins pas toujours possible. Par exemple, dans le cas d'applications téléphoniques, on comprend bien que les systèmes puissent être utilisés par n'importe qui et donc être indépendants du locuteur. Bien que la méthodologie de base reste la même, cette indépendance au locuteur est obtenue par l'acquisition de nombreux locuteurs (couvrant si possible les différents dialectes) qui sont utilisés simultanément pour l'entraînement de modèles susceptibles d'en extraire toutes les caractéristiques majeures. Une solution intermédiaire parfois utilisée consiste à développer des systèmes capables de s'adapter rapidement (de façon supervisée ou non) au nouveau locuteur.

- **Le système reconnaît-il des mots isolés ou de la parole en continue ?**

Evidemment, il est plus simple de reconnaître des mots isolés bien séparés par des périodes de silence que la séquence de mots constituant une phrase. En effet, dans ce dernier cas, non seulement la frontière entre les mots n'est plus connue mais les mots deviennent fortement articulés (c'est-à-dire que la prononciation de chaque mot est affectée par le mot qui précède ainsi que par celui qui suit ; un exemple simple : les liaisons du français).

Dans le cas de la parole continue, le niveau de complexité varie également selon qu'il s'agisse de texte lu, de texte parlé ou, beaucoup plus difficile, de langage naturel avec ses hésitations, phrases grammaticalement incorrectes, faux départs, etc… Un autre problème, qui commence à être bien maîtrisé, concerne la reconnaissance de mots clés en parole libre.

Dans ce dernier cas, le vocabulaire à reconnaître est relativement petit et bien défini mais le locuteur n'est pas contraint de parler en mots isolés. Par exemple, si un utilisateur est invité à répondre par «oui» ou «non», il peut répondre «oui, s'il vous plaît». Dans ce contexte, un problème qui reste particulièrement difficile est le rejet de phrases ne contenant aucun mot clé. La taille du vocabulaire et son degré de confusion sont également des facteurs importants. Les petits vocabulaires sont plus faciles à reconnaître que les grands vocabulaires, étant donné que dans ce dernier cas, les possibilités de confusion augmentent. Certains petits vocabulaires peuvent cependant s'avérer particulièrement difficiles à traiter ; ceci est le cas, par exemple, pour l'ensemble des lettres de l'alphabet, contenant surtout des mots très courts et proches au niveau acoustique.

- **Le système est-il robuste ?**

Autrement dit, le système est-il capable de fonctionner proprement dans des conditions difficiles ? En effet, de nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance ont été identifiées [29,30] :

- Bruits d'environnement (dans une rue, un bistrot etc…)
- Déformation de la voix par l'environnement (réverbérations, échos,etc…)
- Qualité du matériel utilisé (micro, carte son etc…)
- Bande passante fréquentielle limitée (fréquence limitée d'une ligne téléphonique)
- Elocution inhabituelle ou altérée (stress, émotions, fatigue, etc…)

Certains systèmes peuvent être plus robustes que d'autres à l'une ou l'autre de ces perturbations, mais en règle générale, les systèmes de reconnaissance de la parole sont encore sensibles à ces perturbations.

2.3.4 Les techniques de reconnaissances de la parole :

- Deux approches :

Dans l'approche **globale**, l'unité de base est le **mot** (donc non décomposable). Cette méthode fournit une image **acoustique** de chaque mots à identifier et permet donc d'éviter l'influence mutuelle des sons à l'intérieur des mots. Elle se limite aux petits vocabulaires prononcés par un nombre restreint de locuteurs(les mots peuvent être prononcés de manière différente suivant le locuteur).

L'approche **analytique**, qui tire parti de la structure des mots, identifie les composantes élémentaires (phonèmes, syllabes, …). Celles-ci sont les unités de base à reconnaître. Cette approche est plus générale que la précédente : pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base.

Pour la reconnaissance de mots isolés à **grand vocabulaire**, la méthode globale ne convient plus car la machine nécessiterait une mémoire et une puissance considérable pour respectivement stocker les images acoustiques de tous les mots du vocabulaire et comparer un mot inconnu à l'ensemble des mots du dictionnaire. Il est de plus impensable de faire dicter à l'utilisateur l'ensemble des mots que l'ordinateur a en mémoire.
C'est donc la méthode **analytique** qui est utilisée : les mots ne sont pas mémorisés dans leur intégralité, mais traités en tant que **suite de phonèmes.**

- **Principe général de la méthode globale et analytique**

Le principe est le même que ce soit pour l’approche analytique ou l’approche global[31], ce qui différencie ces deux méthodes est l’entité à reconnaître : pour la première il s’agit du phonème, pour l’autre du mot. On distingue deux phases:

- La **phase d'apprentissage** : un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, pour créer en machine le **dictionnaire de références** acoustiques. Pour l'approche analytique, l'ordinateur demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers.
- La **phase de reconnaissance** : un locuteur prononce un mot de vocabulaire. Ensuite la reconnaissance du mot est un problème typique de reconnaissance de formes. Tout système de reconnaissance des formes comporte toujours les trois parties suivantes:

1. Un capteur permettant d'appréhender le phénomène physique considéré (dans notre cas un microphone),
2. Un étage de paramétrisation des formes (par exemple un analyseur spectral),
3. Un étage de décision chargé de classer une forme inconnue dans l'une des catégories possibles.

2.3.5 Principe général :

Les processus réels produisent généralement des sorties en forme de signaux. Ces derniers sont, soit de nature discrète (caractères d'un alphabet fini ou des vecteurs quantifiés parmi un ensemble de valeurs...), soit de nature continue (échantillon de parole, mesure de température…).

Un problème d'intérêt fondamental est de caractériser ces signaux en terme de modèle de signaux. Ce besoin réside dans le fait qu'un modèle peut offrir les bases d'une explication théorique d'un signal réel, et par conséquent maîtriser les sorties désirées en terme de suppression de bruit par exemple. Une deuxième raison qui peut expliquer cet intérêt est le fait qu'un modèle procure une idée sur le signal sans avoir besoin de sa source. Cette propriété est importante quand le coût de l'obtention du signal à partir de sa source initiale est élevé.

A ce stade deux possibilités de modélisation se proposent à nous. Le modèle déterministe et le modèle stochastique ; le modèle déterministe exploite généralement quelques propriétés du signal en vue de sa modélisation, comme par exemple la forme du signal (sinusoïde, somme d'exponentielles, amplitude...).

La modélisation stochastique quant à elle, elle essaie de déterminer les caractéristiques statistiques du signal. On parle dans ce cas spécialement de fonctions de distribution de probabilité (gaussienne, poisson …).

Pour des applications de traitement de la parole, les deux modèles ont procuré d'excellents résultats. Dans cette thèse nous allons nous concentrer sur un seul type de modélisation stochastique, en l'occurrence les modèle de Markov cachés (HMM).Nous allons tout d'abord revoir brièvement les grandes idée sur une approche probabiliste. Les systèmes sont généralement constitués de deux unités principales, le module de décodage acoustico-phonétique et le module de modélisation du langage. Le schéma suivant présente les principales entités d'un système de reconnaissance.

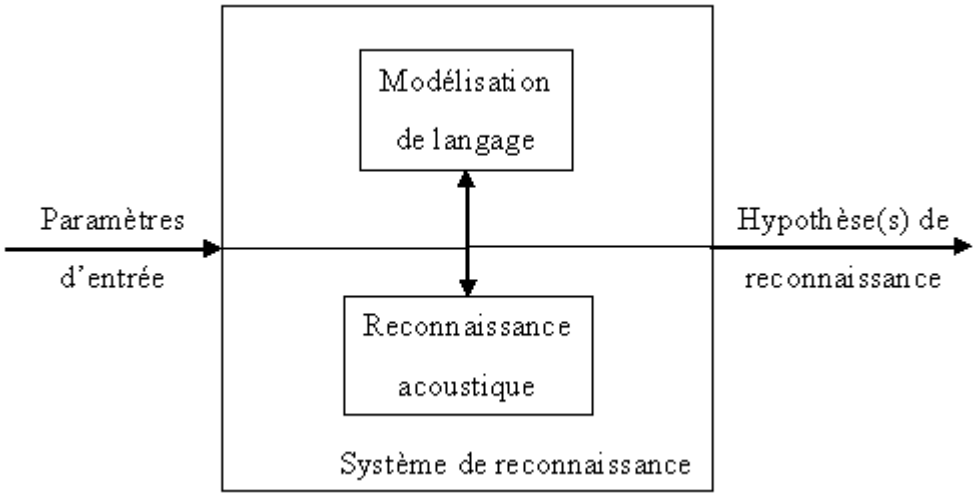

Fig.2-2 : description symbolique d'un système de reconnaissance de la parole

Le premier module permet, à partir d'une analyse paramétrique du signal à reconnaître, de définir quel est l'élément acoustique qui est le plus probablement produit. Cet élément peut être de différents types : phonèmes, diphones, syllabes, etc. Cette étape franchie, il est nécessaire de mettre en correspondance une suite d'éléments acoustiques avec une forme lexicale. C'est ici qu'intervient le second module. Il permet d'obtenir une information *a priori* sur le positionnement d'un mot dans le signal à reconnaître par différentes techniques de modélisation soit à base de grammaire, soit purement statistique, soit à base d'approches mixtes telles que les grammaires probabilistes.

La formule générale, dans le cadre d'un système entièrement probabiliste, s'exprime sous la forme d'une équation bayésienne. Elle a été énoncée dans le cadre de la reconnaissance de la parole par [32]. Le but du système est de trouver l'hypothèse W^* qui maximise pour toutes les séquences de mots *W* possibles et pour une observation acoustique *A*, l'équation suivante :

$$W^* = \arg\max_W P(W/A) = \arg\max_W \frac{P(W).P(A/W)}{P(A)} \approx \arg\max_W P(W).P(A/W) \quad \text{.....(2.1)}$$

Dans cette équation, nous pouvons identifier plusieurs facteurs :

- P(*A*) est la probabilité de l'observation acoustique *A*. Celle-ci est constante pour toutes les séquences de mot *W*, d'où l'approximation finale de l'équation précédente. Pour générer cette observation, le module de décodage acoustique doit, dans un premier temps, analyser le signal de parole, et ensuite définir quelle est la suite d'éléments acoustiques la plus probable.

- P(*A*|*W*) est la probabilité de l'observation acoustique *A* connaissant une séquence de mots *W*. Pour ce faire, on utilise un dictionnaire phonétique c'est-à-dire contenant la transcription des graphèmes avec leurs décompositions en éléments de base pour le système acoustique.

- P(*W*) est la probabilité *a priori* de la séquence de mots *W*, sans aucune notion d'acoustique, dans le langage considéré. C'est la probabilité générée par le modèle de langage. Les techniques de modélisation stochastique du langage seront exposées dans le chapitre suivant.

2.3.6 Du signal de parole à l'observation acoustique :

Modules acoustiques :

Les premiers modules de traitement dans un système de reconnaissance de la parole sont les suivants :

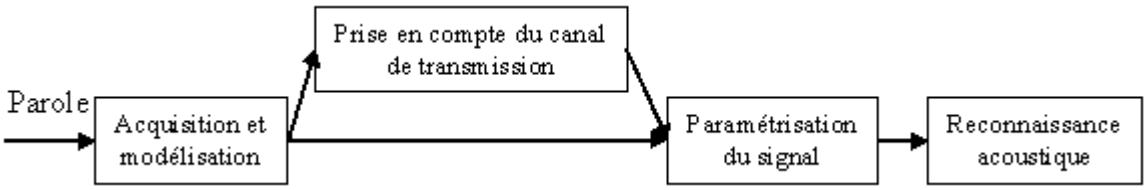

Fig. 2-3 : chaîne de traitement acoustique d'un système de reconnaissance de la parole

Comme le montre la figure 2.2, le signal de parole est d'abord numérisé puis modélisé sous une forme généralement fréquentielle. Pourtant, avant d'obtenir ces mesures, le signal a subi des modifications dues à l'environnement dans lequel se trouve le locuteur, à l'influence du système d'acquisition, et à une éventuelle transmission par le biais d'un média informatique, par exemple un réseau. Ces modifications sont souvent regroupées sous le terme générique de « *canal de transmission* ». Certains systèmes de reconnaissance dispose d'un module de prise en compte de ce canal pour tenter d'éliminer son influence sur le signal de parole. Le module suivant, dans la chaîne de traitement acoustique, est celui qui extrait des paramètres pertinents pour la reconnaissance de la parole. Ces paramètres sont ensuite envoyés au module de reconnaissance acoustique qui identifie les sons présents dans le signal.

2.3.7 Elocution :

Le mode d'élocution caractérise la façon dont on peut parler au système. Il existe quatre modes d'élocution distincts :

- **Mots isolés :**

Chaque mot doit être prononcé isolément, c'est à dire précédé et suivi d'une pause.

- **Mots connectés :**

Le système reconnaît des séquences de quelques mots sans pause volontaire pour les séparer (exemple : reconnaissance de chiffres connectés ou de nombres quelconques…).

- **Parole continue lue :**

C'est le discours usuel, si ce n'est que les textes sont lus.

- **Parole continue spontanée :**

C'est le discours usuel, sans aucune contrainte.

La reconnaissance de mots isolés fonctionne relativement bien de nos jours pour différentes langues. De bons résultats ont été publiés par de nombreux laboratoires. Généralement, de tels outils de reconnaissance de parole sont utilisés pour un vocabulaire de commande correspondant à des actions spécifiques et simples (gestion de menus…).Le premier mode d'élocution sera abordé lors de cette étude. Les expériences décrites dans ce travail ont été effectuées sur de la parole claire (absence de bruit, de réverbération …).

2.4 Conclusion :

Dans ce chapitre, nous avons décrit les premiers modules de la segmentation de la parole en vue de sa reconnaissance. Nous avons abordé les outils et les algorithmes les plus répandus de nos jours. Nous avons présenté ces données pour caractériser notre système d'expérimentation, décrit en détail plus loin dans ce manuscrit. Nous avons jusqu'ici volontairement éludé les modèles de langage. Ceux-ci étant la base de notre travail, le chapitre suivant leur seront entièrement consacrés

3 Modèles de MARKOV CACHES

Sommaire

3.1 Introduction :

Le problème de la reconnaissance de parole peut être formulé dans ces termes :

Comment modéliser au mieux des unités représentatives du signal de parole ?

Il existe en fait deux types de modélisation possible des propriétés d'un signal donné :

- Les modèles déterministes, qui exploitent les propriétés intrinsèques du signal,
- Les modèles statistiques, qui caractérisent les propriétés statistiques du signal.

Dans ce travail, nous avons opté pour des modèles statistiques : les modèles de Markov cachés, appelés aussi HMM, l'abréviation anglaise HMM sera utilisée couramment pour parler des modèles de Markov cachés. Les HMM se sont imposés comme une technique prédominante en reconnaissance de la parole ces dernières années. Nous allons présenter dans ce qui suit les bases nécessaires à l'utilisation de ce type de modèle pour la reconnaissance automatique de la parole. Ces modèles se sont avérés les mieux adaptés aux problèmes de la reconnaissance de la parole et la quasi-totalité des outils de reconnaissance de la parole disponibles actuellement sur le marché sont basés sur cette technologie. Pour bien comprendre le fonctionnement de ces derniers, il est bon discuté au préalable des modèles de Markov discrets à nombre d'états fini.

3.2 Des Modèles de Markov Discrets aux Modèles de Markov Cachés :

Les Modèles de Markov Discrets sont basés sur une suite (ou boucle) d'états dans lesquels on navigue par des probabilités de transition et suivant des observations. Nous les illustrerons par l'exemple de la météo. Le modèle créé est une boucle d'état de temps : pluies nuageux ensoleillé. Ils sont représentés figure 3.1.

Grâce à ces modèles, on peut connaitre, sachant l'observation passée (La météo de la veille), quel va être la probabilité des observations futures.

Dans l'exemple que nous avons pris, ce serait par exemple la probabilité

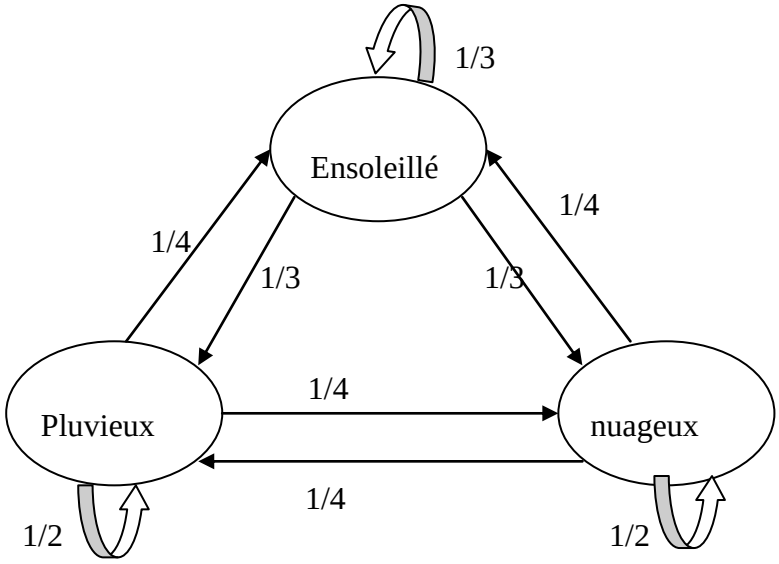

Fig.3-1 – Modelés de Markov Discrets

Qu'il fasse beau pendant 10 jours. La nouveauté qu'apportent les Modèles de Markov Cachés par rapport aux Modèles de Markov Discrets, réside dans le fait que les états sont caractérises par des distributions de probabilité sur l'espace des observations possibles, pluie, nuageux ensoleillé. Ces états correspondent alors à une variable qui n'est plus observée directement. Ceci pourrait par exemple correspondre au fait d'être ou non dans une dépression atmosphérique. On a alors deux modèles stochastiques liés : le temps et la dépression atmosphérique.

On peut aussi illustrer ces modèles par le lancement de plusieurs pièces de monnaie biaisées les unes après les autres dont on ne retient que le résultat des lancements. Les deux séquences stochastiques liées seraient alors, la suite des états pile ou face et la suite des choix des pièces. Le modèle caché permet alors de connaître la probabilité d'avoir une séquence de pièces connaissant la séquence d'état pile ou face. Les Modèles de Markov Cachés (MMC), dont nous préférerons employer l'acronyme anglais HMM pour Hidden Markov Models, sont, à l'heure actuelle, les outils de modélisation les plus employés en reconnaissance de la parole continue. Les HMM ont montré leur adéquation à traiter la parole.

3.3 Présentation des modèles de Markov Cachés :

3.3.2 Définition :

Un modèle de Markov caché est un automate stochastique capable, après apprentissage, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle. Idéalement, il faudrait pouvoir associer à chaque entité un modèle. Il va de soi que ceci est irréalisable en pratique car le nombre de modèles serait beaucoup trop élevé. Des sous-unités lexicales comme le mot, la syllabe, ou le phonème sont utilisées afin de réduire le nombre de paramètres à entraîner. A chacune de ces unités est associé un modèle de Markov caché constitué d'un nombre fini d'état prédéterminé. Formellement, un modèle de Markov caché peut être défini par l'ensemble des paramètres λ: **λ=L, A, B, π**

1. L : est le nombre d'états du modèle,
2. $A = a_{ij} = p(q_j / q_i)$: La matrice de probabilité de transition sur l'ensemble des états du modèle (on définit ainsi une topologie par l'intermédiaire de cette matrice A).
3. $B = b_j(x_n) = p(x_n / q_i)$: La matrice de probabilités d'émission de l'observation x_n dans l'état q_j.

- Dans le cas d'entrées discrètes (après quantification vectorielle des observations), les probabilités d'émission peuvent être décrites comme des fonctions de densité de probabilité discrète) $p(y_i / q_j)$
- Dans le cas d'entrées continues $x^n \in \Re^d$), $p(x_n / q_j)$ est :
    - ♦ Supposée être de la forme d'une distribution multivariable gaussienne, entièrement définie par le vecteur moyenne et la matrice de covariance ou d'une distribution de type multigausienne (somme pondérée de gaussienne s, voir hypothèses HMM paragraphe 3.3.4.4).

4. Л : la distribution initiale des états, $\forall j \in [1, L], p(q_j / q_I)$.

En reconnaissance de la parole, des modèles de Markov gauche-droite d'ordre 1 sont les plus souvent utilisés du fait de l'aspect séquentiel du signal de la parole [36]. Un modèle de Markov à 3 états est visible sur la figure 3-2.

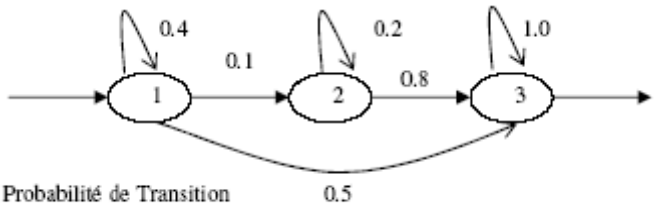

Fig.3-2 Exemple d'un modèle de Markov caché à trois états

Les modèles de Markov cachés (HMM) supposent que la séquence de vecteurs acoustiques représentatifs du signal de parole soit une succession de segments stationnaires. Ainsi la parole est modélisée par une succession d'états, avec des transitions instantanées possibles entre ces états) $p(q_j / q_k)$. Chaque observation est supposée être une fonction probabiliste de l'état.

Deux processus stochastiques concurrents sont observés :

- Un premier processus qui est la séquence d'observations, $X = X_1,\ldots,X_N$
- Un second processus, la séquence d'états non directement observable.

C'est pourquoi ces modèles sont dits 'cachés'. ***La séquence d'états n'est pas directement observable.***

3.3.3. Problème à résoudre [12, 33, 34,35] :

Soit M le modèle de Markov caché associé à la phrase X et constitué d'une concaténation de sous unités lexicales. La reconnaissance de la séquence de vecteurs acoustiques X s'effectue en trouvant le modèle M qui maximise la probabilité $p(M / X, \lambda)$ (probabilité qu'un modèle M génère une séquence de vecteurs acoustiques X étant donné une série de paramètres λ). Cette probabilité est aussi appelée probabilité a posteriori. Malheureusement, il n'est pas possible d'accéder directement à cette probabilité par le processus d'entraînement des modèles de

Markov, mais seulement à la probabilité qu'un modèle donné générera une certaine séquence de vecteurs acoustiques p(X/M).

En utilisant la loi de bayes (3.1), il est possible de lier ces deux probabilités selon

$$p(M/X) = \frac{p(X/M).p(M)}{p(M)} \quad \text{.....(3.1)}$$

Où

- p(X/M) est la vraisemblance de la séquence d'observations X étant donné le modèle M,
- p(M) est la probabilité a priori du modèle,
- p(X) La probabilité a priori de la séquence de vecteurs acoustiques.

Nous verrons un peu plus tard qu'il est nécessaire de choisir un critère pour l'entraînement des paramètres $\lambda = \{A, B, \prod\}$:

- Critère MAP (Maximum a posteriori Probability): maximum a posteriori.
- Critère MLE (Maximum Likelihood Estimation) : maximum de vraisemblance.

Les trois propriétés élémentaires suivantes seront couramment utilisées dans ce chapitre et le reste du document :

$$p(a,b) = p(a/b).p(a) = p((b/a).p(a) \text{ .....(3.2)}$$

$$p(a,b/c) = p(a/b,c).p(b/c) \text{ .....(3.3)}$$

Si les événements b_k sont mutuellement exclusifs et collectivement exhaustifs :

$$p(a) = \sum_{\forall k} p(a, b_k) \text{ .....(3.4)}$$

- **Hypothèses :**
  - **H1** On suppose que P(M) peut être calculé indépendamment des observations. Cette probabilité est en effet indépendant de X et peut être estimée à partir du modèle de langage.
  - **H2** Pour une séquence d'observations connue, P(X) peut être considéré constant, puisqu'il est indépendant du modèle, si les paramètres de ces modèles sont fixés. Ainsi maximiser $p(M/X) = \frac{p(X/M).p(M)}{p(X)}$ revient à maximiser p(X/M).p(M) .

Il faut alors résoudre 3 problèmes liés à ces modèles [12 ,37].

1. **L'estimation des probabilités :** comment calculer p(X|M) et quelles sont les hypothèses nécessaires à propos du modèle pour se définir une série de paramètres utiles pour la reconnaissance ?
2. **L'entraînement :** étant donné une séquence d'observation X_j associée à leurs modèles de Markov respectifs, comment déterminer les paramètres des modèles afin que chacun ait la probabilité la plus grande possible de générer les séquences d'observations associées ? Comment trouver l'ensemble des paramètres λ qui maximisent p(M/X,λ) pour l'ensemble des séquences de vecteurs acoustiques X associé au modèle M ? Cette probabilité n'étant pas directement accessible, on préfère maximiser p(X/M,λ) (soit utiliser le critère MLE plutôt que MAP).
3. **Le décodage** : étant donné une séquence de modèles de Markov avec leurs paramètres entraînés et une séquence d'observations X, comment trouver la meilleure séquence M_k de modèles de Markov élémentaires pour maximiser la probabilité que M_k ait généré les observations ?

3.3.4 Problème 1 : Estimation des probabilités:

Le problème de l'estimation des probabilités peut être énoncé de la façon suivante :

Étant donné un modèle de Markov M, comment calculer la probabilité P(X|M) qu'il génère la séquence de vecteurs acoustiques X ?

De manière générale, le calcul de cette probabilité s'effectue selon :

$$p(X/M) = \sum p(C, X/M) = \sum_{\Gamma} p(Q_1^N, X/M) \quad \text{.....(3.5)}$$

Où

- Γ= ensemble des chemins autorisés C dans le modèle M,
- $Q_1^N = \{q_1 = q^0, q^1, ..., q^N, q^{N+1} = q_F\}$ est une séquence ordonnée de N états,
- q^n représentant l'état du HMM à l'instant n,
- q_k^n signifie que l'état q_k du HMM est visité à l'instant n.

Il existe deux procédures récurrentes de calcul de cette probabilité que nous proposons de décrire :

- **L'algorithme Forward -Backward** qui fournit une solution exacte à ce problème faisant intervenir tous les chemins dans le modèle HMM.
- **L'algorithme Viterbi** fournissant une solution approximative faisant intervenir uniquement le meilleur chemin dans le modèle HMM.

a) L'algorithme Avant – Arrière (Forward – Backward) :

L'algorithme Forward -Backward peut être utilisé pour calculer P(X|M) de manière récursive en posant :

$$\forall l \in [1, L], \alpha_n(l/M) = p(q_l^n, X_1^n/M) \quad \text{.....(3.6)}$$

Il vient alors :

$$\forall l \in [1, L], \alpha_{n+1}(l/M) = [\sum_{k=1}^{L} \alpha_n(k/M).p(q_l/q_k, M)].p(\chi_{n+1}/q_l) \quad \text{..... (3.7)}$$

Avec : $\alpha_l(l) = p(\chi_l/q_l).p(q_l/q_1)$

De même posons :

$$\forall l \in [1, L], \beta_n(l/M) = p(X_{n+1}^N/q_l^n, X_1^n, M) \quad \text{.....(3.8)}$$

Il vient alors :

$$\forall l \in [1, L], \beta_n(l/M) = \sum_{k=1}^{L} p(q_k/q_l, M).p(\chi_{n+1}/q_k).\beta_{n+1}(k/M) \quad \text{..... (3.9)}$$

Avec : $\forall l \in [1, L], \beta_N(l) = 1, ou, \beta_N(l) = 0$

En utilisant ces deux dernières formulations il est possible de calculer P(X|M) selon :

$$\forall l \in [1, L], p(X/M) = \sum_{l=1}^{L} \alpha_n(l/M).\beta_n(l/M) \quad \text{.....(3.10)}$$

Où les termes $\alpha_n(l/M), et, \beta_n(l/M)$ peuvent être calculés de manières récurrentes.

De plus en posant n=N, q_f état final et q_I état initial :

$$p(X/M) = \sum_{\forall l \in \Re} \alpha_n(l/M) = \alpha_{n+1}(q_F/M) = \beta_0(q_I/M) \quad \text{.....(3.11)}$$

Remarque : $\Re$ représente l'ensemble des états finaux.

C'est souvent cette dernière formulation qui est utilisée en reconnaissance de la parole, car elle ne nécessite que le calcul des termes a pour déterminer la probabilité P(X|M).

Il faut noter que les équations décrites ici ont nécessité une série d'hypothèses simplificatrices et limitatrices. Les équations présentées ici sont basées sur le critère du maximum de vraisemblance. Il est possible d'utiliser un autre critère permettant de réduire sensiblement la charge de calcul : le critère de Viterbi.

b) L'algorithme de Viterbi :

Au lieu de prendre en compte tous les chemins autorisés, seul le plus probable est gardé. Ainsi, il suffit de remplacer dans les équations précédentes l'opérateur $\sum$ par max. Ce critère est largement utilisé en reconnaissance de la parole du fait du faible coût qui lui est associé (en effet, il est évident que l'opérateur max est moins coûteux en temps de calcul que l'opérateur $\sum$ sur tous les états).

L'algorithme de Viterbi est donc une simplification de la récurrence avant qui devient :

$$p(q_l^n/M_j) = \max_k \left\lfloor p(q_k^{n-1}, X_1^{n-1}/M_j).p(q_l^n/q_k^{n-1}, M_j) \right\rfloor . p(x_n/q_l^n \cdot M_j) \quad \text{...(3.12)}$$

Soit en passant au log :

$$-\log(p(q_l^n, X_l^n/M_j) = \min_k \left\lfloor -\log(p(q_k^{n-1}, X_1^{n-1}/M_j)) - \log(p(q_l^n/q_k^{n-1}, M_j)) \right\rfloor - \log.p(x_n/q_l^n \cdot M_j)$$

Cette dernière équation montre que le calcul de la probabilité P(X|M) est très semblable à une récurrence du type DTW (Dynamic Time Warpping).

3.3.5. Hypothèses simplificatrices :

L'utilisation pratique de ces modèles ne peut se faire qu'après avoir fixé quelques hypothèses qu'il est important de rappeler [38]:

- **H1** On a supposé que P(M) peut être calculé indépendamment. Cette probabilité est en effet indépendante de X et peut être estimée à partir du modèle de langage.
- **H2** Pour une séquence d'observations connue, P(X) peut être considéré constant, puisqu'il e st indépendant du modèle, si les paramètres de ces modèles sont fixés.
- **H3** Les modèles de Markov sont supposés du premier ordre ; ainsi la probabilité que la chaîne de Markov soit dans l'état q l au temps n dépend uniquement de l'état de la chaîne de Markov au temps n-1 et est indépendante du passé.
- **H4** Les vecteurs acoustiques ne sont pas corrélés, la probabilité qu'un vecteur acoustique soit émis au temps n dépend uniquement de la transition de l'état q_k^{n-1} à q_l^n et est indépendante du passé.
- **H5** la probabilité d'émission est supposée dépendante uniquement de l'état courant pour réduire le nombre de paramètres.

Pour calculer la probabilité d'émission, chaque état q l doit être associé à une densité de probabilité $p(x_n/q_l)$. Il faut donc émettre des hypothèses supplémentaires à propos de cette densité de probabilité.

- **H6** dans le cas d'entrées continues,$p(x_n/q_l)$ est supposée être de la forme d'une distribution multivariable gaussienne. Ainsi la probabilité peut être exprimée selon :

$$p(x/q_l) = \sum_{j=1}^{N} c_{lj} \cdot N_{lj}(x) \quad \text{.....(3.13)}$$

Où $N_{lj}(x)$ désigne la valeur au point x (trame acoustique) de la gaussienne.

- **H7** dans le cas d'entrées discrètes, l'hypothèse **H6** n'est plus nécessaire. La séquence de vecteurs acoustiques X est quantifiée. Chaque vecteur acoustique x_n est remplacé par un centroïde y_i (le plus proche au sens d'une distance Eucludienne) sélectionné dans un dictionnaire prédéterminé (codebook). Ainsi les probabilités d'émission peuvent être décrites comme des fonctions de densité de probabilité discrète $p(y_i/q_l)$.

Ce type de système est couramment appelé système discret. Le temps de calcul nécessaire à l'utilisation de ces modèles est beaucoup moins important que pour les systèmes continus, mais les performances restent limitées.

3.3.6 Problème 2 : Estimation des paramètres et entraînement des Modèles :

Le but de l'entraînement des modèles acoustiques est de trouver l'ensemble de s paramètres λ maximisant sur l'ensemble des données d'entraînement X_j la vraisemblance des données étant donné les modèles associés M_j, soit :

$$\arg\max_{\lambda} \prod_{j=1}^{j} p(X_j / M_j, \lambda_j) \quad \text{.....(3.14)}$$

Il est nécessaire d'estimer deux ensembles de paramètres :

- Les probabilités de transitions entre les états : a_{ij}.
- Les probabilités d'émission des observations pour chaque état : $b_j(x_n)$

Les approches les plus utilisées sont basées sur des adaptations de l'algorithme EM (Expectation-Maximisation) appelées :

- **Algorithme de Baum-Welch** : P(X|M) est estimée en tenant compte de tous les chemins possibles (voir paragraphe précédent), implémentation de l'algorithme Expectation-maximisation (EM).
- **Algorithme de Viterbi** : P(X|M) est estimée en tenant compte du meilleur chemin uniquement (approximation de l'algorithme EM).

a) Apprentissage Baum-Welch :

L'algorithme de Baum Welch est un processus itératif où, à chaque itération, de nouvelles valeurs des paramètres λ des modèles sont estimées à partir des anciennes valeurs. L'entraînement des modèles est effectué à partir de l'estimation de P(X|M) en tenant compte de tous les chemins possibles. On utilise les formules de récurrence 'avant' et 'arrière' définies dans ce chapitre. Nous avons déjà montré en posant :

$$\forall l \in [1, L). \alpha_n(l / M) = p(q_l^n, X_1^n / M) \quad \text{.....(3.15)}$$

$$\forall l \in [1, L). \beta_n(l / M) = p(X_{n+1}^N / q_l^n, X_1^n, M) \quad \text{.....(3.16)}$$

Que :

$$\forall n \in [1, N],\ p(X/M) = \sum_{l=1}^{L} \alpha_n(l/M).\beta_n(l/M) \quad \text{.....(3.17)}$$

Et

$$\gamma_n(k/M) = p(q_k^n / X, M) = \frac{\alpha_n(k/M).\beta_n(k/M)}{p(X/M)} \quad \text{.....(3.18)}$$

Le processus itératif mis en oeuvre consiste donc en deux phases :

- Une phase d'estimation où les récurrences avant et arrière nous permettent d'obtenir $p(q_k^n / X, M, \lambda)$ à partir d'un ensemble de paramètres λ.

- Une étape de maximisation où les paramètres λ sont mis à jour en utilisant les formules décrites dans ce paragraphe.

b) Apprentissage Viterbi :

Les paramètres sont optimisés de façon à maximiser la vraisemblance du meilleur chemin. Cela revient à supposer que les probabilités $\gamma_n(k) = P(q_k^n / X, M)$ soient égales à 0 et 1. Comme pour l'algorithme EM classique, on part d'un ensemble de paramètres initiaux λ^0 et les paramètres optimaux l sont obtenus de manière itérative. Le processus d'entraînement est composé d'une étape d'estimation (E) qui sert à trouver la segmentation qui maximise la vraisemblance à partir des paramètres, et d'une étape de maximisation (M), qui effectue une mise à jour des paramètres étant donnés une segmentation. L'ensemble des paramètres initiaux λ^0 peut être estimé à partir de modèles déjà entraînés par un corpus ou par exemple par l'intermédiaire d'un corpus déjà segmenté comme TIMIT. Il est ensuite possible à partir de segmentation optimale trouvée de calculer les paramètres des fonctions de vraisemblance en considérant tous les vecteurs associés à chacune des classes. Ce processus de réalignement des données acoustiques à l'aide d'un modèle et de réentraînement d'un nouveau modèle est effectué jusqu'à ce qu'une certaine convergence soit atteinte (la segmentation ne varie plus ou l'accroissement relatif de

la vraisemblance pour l'ensemble des données d'entraînement est inférieur à un seuil fixé).

3.3.7. Problème 3 : Le décodage

Après l'entraînement des paramètres des modèles HMM, la reconnaissance d'une séquence acoustique X correspondant à une phrase s'effectue par le calcul de la probabilité $p(X / M_j)$ pour tous les modèles ou séquence de modèles M_j. Ce calcul peut être effectué en utilisant les récurrences déjà décrites dans ce chapitre. Il faut toutefois noter que dans le cas de grands vocabulaires, ces récurrences sont coûteuses en temps de calcul. De plus, dans le cas de la parole continue, le nombre de combinaison possible des modèles (séquences de M_j) est très important. Ainsi, des techniques d'élagage (pruning) ont été développées pour permettre l'utilisation des ces récurrences.

La procédure de reconnaissance consiste à trouver le modèle (ou la séquence de modèles) M_k pour lequel :

$$k = \arg\max_{\forall j} p(M_j / X) = \arg\max_{\forall j} p(X / M_j).p(M_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.19)$$

Pour une reconnaissance de la parole continue, le score du modèle acoustique $p(X / M_j)$ doit être multiplié par la probabilité de la séquence de mots associée $P(M_j)$.

3.4 HMM en reconnaissance de la parole :

Le signal associé à un mot isolé peut être considère comme une suite de sons de base agissant comme un alphabet. Un mot est alors caractérise par une suite caractéristique d'éléments de cet alphabet. En absence de mémoire, il suffirait de comparer élément par élément les constituants d'un mot. Toutefois en pratique on constate que la probabilité d'un élément, dans la suite considérée, dépend des éléments qui ont précède. D'où l'idée d'une modélisation markovienne de la suite de ces éléments.

Un premier traitement, dont la valeur est essentielle, est donc d'extraire du signal un nombre faible d'éléments pertinents qui sont supposés suivre un modèle HMM. Deux approches sont fréquemment utilisées :

- **approche temporelle**: on effectue une prédiction linéaire (LPC: linear prediction coding) sur le signal. Cela consiste à approximer le signal par une combinaison linéaire des p valeurs précédentes. On peut donc écrire que $s(s) = a_1.s(n - 1) + \ldots + a_p.s(n - p) + b(n)$ où b(n) est un processus aléatoire qui prend en compte les erreurs du modèle. On le modélise par un bruit blanc de puissance σ^2. Le vecteur $\theta = (\sigma^2; a_1; \ldots ; a_p)$ est choisi de façon à minimiser σ^2.

La solution est donnée par les équations de Yule-Walker :
$R(1; a_1; \ldots ; a_p)^T = (\sigma^2; 0; \ldots ; 0)^T$ qui lie θ à la matrice de covariance R du signal s(n).Une estimation de R permet d'estimer le paramétrique θ caractérisant le bloc.

- **approche fréquentielle**: le signal est passé par un banc de filtres. En général les filtres couvrent la bande de Nyquist, avec un chevauchement adéquate, suivant une échelle non linéaire. Les plus fréquemment utilisées sont des échelles logarithmiques analogues à celle de l'oreille humaine.

Les points de FFT de l'ensemble des sous-bandes représentent le vecteur de paramètres du bloc. Il est parfois intéressant de prendre le cepstre plutôt que le spectre dans la mesure où les coefficients cepstraux sont généralement décorrélées. Cela justifie l'utilisation de matrice diagonale dans le traitement HMM.

3.4.1 Utilisation des modèles HMM en reconnaissance de mots isolés :

Il existe différentes manières d'utiliser les HMM en reconnaissance automatique de la parole, selon l'application visée :

a)Reconnaissance de mots isolés en nombre limité(<100 mots) :

Chaque mot m^v du vocabulaire V est modélisé par un HMM^v. La phase de reconnaissance consiste, pour une suite d'observations acoustiques données, à calculer la vraisemblance de ces observations par rapport à chacun des modèles et à considérer comme mot reconnu le mot correspondant à la vraisemblance maximum. Cette approche n'est qu'une version améliorée de l'approche de type programmation dynamique.

Cette approche n'est réalisable qu'avec un vocabulaire limité dans la mesure où l'algorithme de reconnaissance considère tous les chemins modélisant chaque mot m^v, ce qui entraîne un important coût de calcul.

b) Reconnaissance de mots isolés en nombre inférieur à 1000 :

Chaque mot m^v du vocabulaire V est modélisé par un HMM^v. Tous les HMM^v sont liés entre eux par une entrée et une fin commune dans un réseau global. L'entrée correspond à un HMM représentant un silence précédant le mot, tandis que la fin correspond à un HMM modélisant un silence suivant le mot.

Chaque chemin du réseau global correspond à la prononciation d'un mot. En phase de reconnaissance, le mot reconnu correspond alors au chemin le plus probable étant donné la suite d'observations.

Cette approche se différencie de la précédente par l'algorithme de reconnaissance: dans ce cas, celui-ci ne prend en compte que le chemin le plus vraisemblable parmi tous ceux modélisant les mots du vocabulaire V, réduisant ainsi considérablement le coût de calcul.

c) Reconnaissance de mots connectés :

L'utilisation de modèles globaux pour tous les mots du dictionnaire soulève quelques problèmes :

- Le stockage de tous les mots du vocabulaire devient très important.
- Une grande quantité de parole est nécessaire pour réaliser l'estimation statistique de tous les paramètres (probabilité de transitions, lois d'émissions d'observations).

C'est pourquoi on préfère représenter les mots à partir d'unités phonétiques. Ce qui nous ramène à représenter chaque unité phonétique par un MMC élémentaire.

Le choix de l'unité phonétique reste un problème ouvert. On peut choisir des unités de l'ordre du phonème, mais on peut alors difficilement traduire les variations dues aux contextes.

Pour pallier cet inconvénient, sont apparus les allophones qui modélisent chaque unité de l'ordre du phonème dans un contexte phonétique précis.

A partir de ces modèles est construit par concaténation un modèle de Markov caché pour chaque mot du vocabulaire, en tenant compte des différentes

variantes de prononciations. Le modèle global est obtenu en reliant les modèles de mots selon la syntaxe de l'application. Ce type de construct ion permet de prendre en compte, lors des différentes substitutions et concaténations, des connaissances de type phonétique et phonologique.

Au niveau phonétique, il est ainsi possible de rendre compte de phénomènes spécifiques tels que la décomposition d'un son en phases élémentaires.
Au niveau phonologique sont introduites des règles inter-mot et intra-mot afin de prendre en compte les phénomènes de coarticulation et d'assimilation.
Tout chemin du graphe ainsi obtenu correspond à une phase du langage à reconnaître et la phase de reconnaissance consiste alors, comme précédemment, à rechercher au sein du graphe le chemin le plus probable pour la suite d'observation donnée.

3.4.2 Conception du système de reconnaissance :

Depuis leur introduction en traitement de la parole [39,40], Les modèles de Markov cachés (*Hidden Markov Models* ou HMM) ont pris une importance considérable, au point que la quasi-totalité des systèmes de RAP utilisent cette modélisation.

Les modèles de Markov cachés supposent que le phénomène modélisé est un processus aléatoire et inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Ces deux niveaux donnent à l'approche markovienne une flexibilité qui est séduisante pour modéliser un phénomène aussi complexe que la production de la parole.

Un système de reconnaissance de mots isolés à base d'une modélisation Markovienne est réalisé en deux étapes. Une première étape, est consacrée à l'apprentissage du modèle, à l'issue de cette étape, un modèle Markovien est généré pour chaque mot du corpus, une deuxième étape s'intéresse à la reconnaissance des mots de tests.

Toutefois, il faut noter qu'une phase de quantification vectorielle est réalisée, afin d'avoir une classification des différents vecteurs acoustiques.

a) Quantification vectorielle [12,41] :

Chaque trame d'analyse ou vecteur acoustique est un point dans un espace vectoriel (figure 3-3) de dimension 12 (12 coefficients cepstraux sont générés par analyse MFCC (voir chapitre1) de chaque fenêtre d'analyse) normé par la distance Euclidienne. L'algorithme de classification effectue une partition géographique d'un nuage de points (vecteurs acoustiques) en différant classes en minimisant la distorsion moyenne de l'ensemble.

Chaque nuage ou famille, sera représenté par son centre de gravité. Cette famille de représentants forme un dictionnaire (codebook) de classes ou prototypes.
Le centre de gravité d'une classe est alors défini comme la moyenne de l'ensemble de vecteurs de cette classe et est utilisée comme vecteur prototype.

Donc, chaque vecteur acoustique sera représenté par le code de la classe dont il est membre.

re.

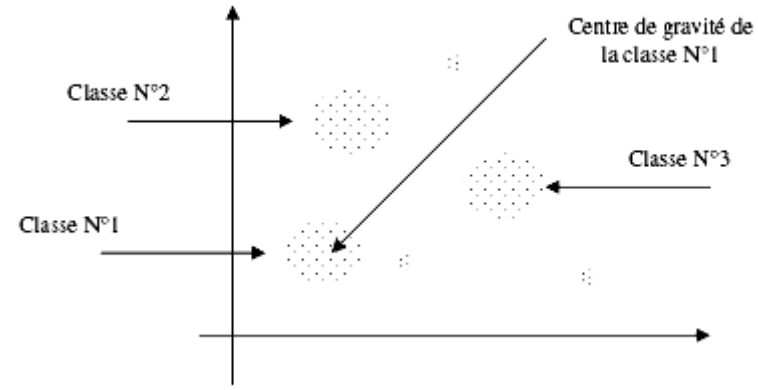

Fig.3-3 Technique de quantification vectorielle

b) Topologie du modèle :

La topologie du modèle définie la structure du modèle HMM utilisée dans l'algorithme d'apprentissage et de reconnaissance. Les structures sont nombreuses et diversifiées, nous allons citer, par la suite, trois types de structures intéressantes du point de vue des applications :

1. Le modèle général où aucune transition n'est absente. Ce modèle, dans lequel nous pouvons aller de n'importe quel état à n'importe quel autre en une seule transition, est appelé modèle ergodique. Un tel modèle est illustré, pour 4 états, dans la figure 3-4.

2. Le modèle à branches parallèles. Ce type de modèle peut être utile pour certaines applications par la symétrie qu'il présente dans sa structure. Un exemple de ce modèle, à 4 états, est illustré dans la figure 3-5.

3. Le modèle Gauche-Droite défini par les trois propriétés suivantes :
   - La première observation est produite alors que la chaîne de Markov se trouve dans le premier état,
   - La dernière observation est générée alors que la chaîne arrive sur l'état final,
   - Une fois que la chaîne quitte un état, elle ne peut y revenir, d'où le nom de modèle Gauche-Droite :

     $\Pi_i = 1$ si $i=1$, $\Pi_i = 1$ si $i \# 1$ et $a_{ij} = 0$ si $j<1$

Un tel modèle, à 4 états, est illustré dans la figure 3-6

- Dans le cas de modèles de mots,ceux-ci sont souvent représentés par de modèles HMM à plusieurs états(typiquement entre 5et 10),dont le nombre est parfois proportionnel au nombre de phonèmes dans le mot ou à longueur de celui/ci.
- Dans le cas de modèles de phonèmes,les modèles HMM sont souvent à 3états,comme représenté à la figure 3-7,le but initial étant d'avoir l'état central modélisant la partie stable du phonème alors que les deux états extrêmes modélisent la partie transitoire.

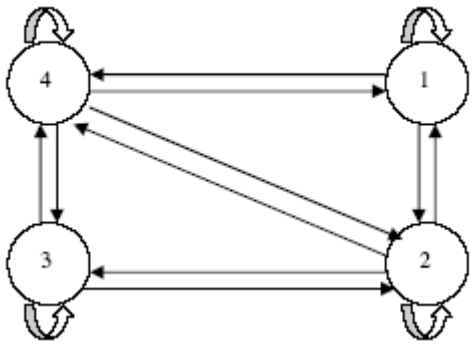

Fig.3-4 : Exemple de Modèle ergodique

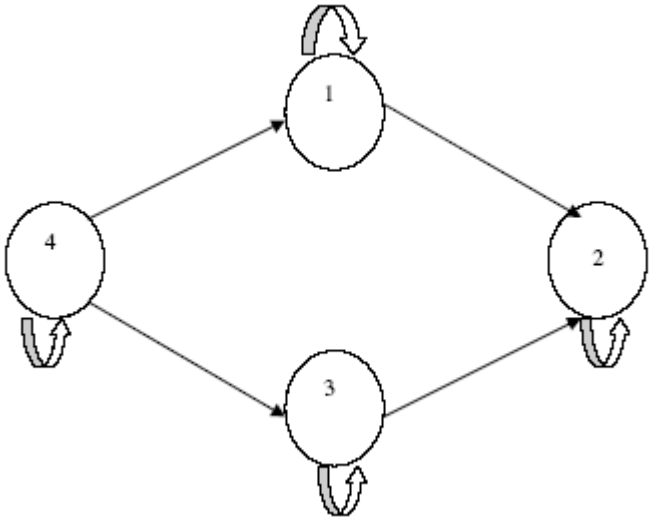

Fig3-5 : Exemple de Modèle à branches parallèles

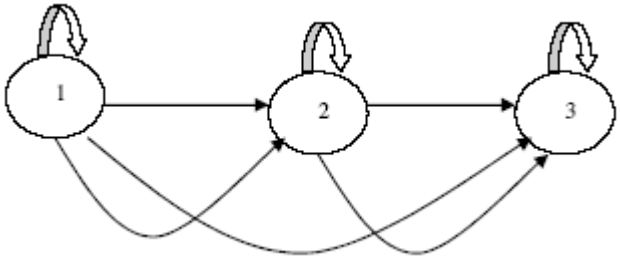

Fig.3-6 : Exemple de Modèle Gauche – Droite

Fig. 3-7 : illustration d'un modèle de phonème à une seule classe q_k répétée plusieurs fois de façon à introduire des contraintes de durée minimale

Nous avons choisi pour notre modèle HMM la topologie Gauche-Droite, cette dernière est bien adaptée au problème de la reconnaissance de la parole. L'avantage d'utiliser tels modèles réside dans le fait qu'ils introduisent des contraintes temporelles fortes sur les chaînes de Markov, ce qui est bien adapté à la reconnaissance de la parole.

3.5 Conclusion :

Les modèles de Markov cachés, présentés dans ce chapitre sont des techniques largement utilisées en reconnaissance de formes, et sont les plus utilisés en reconnaissance de la parole. Ils bénéficient d'algorithmes d'entraînement et décodage performants.

Toutefois, les hypothèses nécessaires à la mise en oeuvre de ces algorithmes peuvent pénaliser les performances de ces modèles.

Les principales hypothèses les plus contraignantes sont :

- Entraînement non discriminant (maximisation de la vraisemblance au lieu des probabilités a posteriori).
- Forme des densités de probabilité fixée (multigaussiennes ou discrète).
- Les composantes des vecteurs acoustiques sont supposées non corrélées.
- La séquence des états est un processus de Markov du premier ordre.
- Le formalisme est rigide, l'intégration d'autres sources de connaissance (syntaxe, sémantique …) est difficile.

Notons que la plupart des systèmes de reconnaissance proposés sur le marché actuellement sont basés sur ce type de technique (Watson d'AT&T, VoiceType d'IBM et Easy Speaking de Dragon Dictate…).

4 Système de Reconnaissance et Résultats

Sommaire

4.1 Introduction :

L'objectif principale de notre travail est le développement d'un système de reconnaissance des chiffres arabe à base de modèle de Markov Caché ; le travail est subdiviser en deux phases (Voir figure 4.1) :

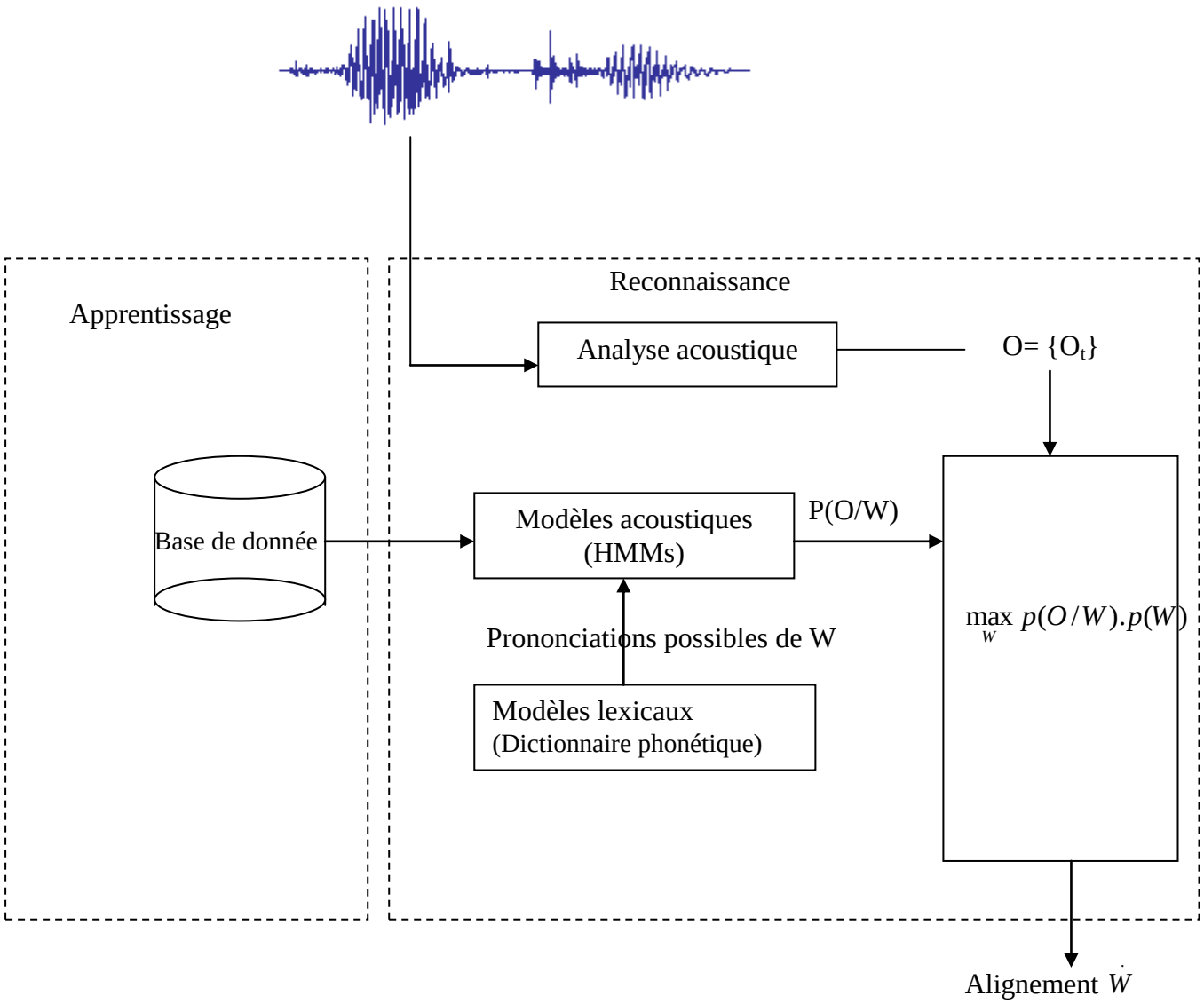

Fig. 4-1 : Schéma général d'un système de reconnaissance de la parole.

4.1.1 Phase d'apprentissage : Comme nous avons l'approche analytique pour le système proposé en reconnaissance, l'unité de base choisie et la syllabe .un modèle de Markov est calculé pour chaque syllabe. elle consiste à attribuer à chaque syllabe du vocabulaire un modèle : λ_i **=(A,B,Л).(l'algorithme I)**

Une fois la phase d'apprentissage (Modélisation par **HMM**) est achevée, un représentant pour chaque groupe de syllabes est déterminé par la méthode du centre de gravité .

Algorithme I : Algorithme d'apprentissage :
algorithme de Baum-Welch

Etape 1 : Chargement du modèle initial. A, B, Л, N et M .où

N : le nombre d'états.

M : le nombre de symboles distincts.

Etape 2 : Charger les observations de toutes les occurrences d'une syllabe *(Matrice OB)*.

Etape 3 : Calculer les nouvelles matrices A, B et Л. En utilisant les fonctions Forward $\alpha_t(i)$ et Backward $\beta_t(i)$.

Etape 4 : Calcul de la probabilité totale à partir de la matrice alpha.

Etape 5 : Retour à l'étape 4, si cette dernière probabilité est supérieure à un seuil fixé déjà et on n'a pas dépassé le nombre d'itération maximale, sinon retour à l'étape 1 pour traiter un autre syllabe.

4.1.2 Phase de reconnaissance : la reconnaissance d'un syllabe $\mathbf{X_T}$ du vocabulaire. Se réduit à une simple comparaison de probabilité $\mathbf{p(X_T/\lambda_i)}$,où **i=1,2,3** sont les modèles correspondant au représentant de chaque groupe, le syllabe inconnu est identifié au modèle $\boldsymbol{\lambda_i}$ si la probabilité est la plus élevée et dépasse un certain seuil empirique(la méthode qu'on à utilisé pour reconnaître un syllabe est basée sur la Maximisation de vraisemblance « ML :maximum like lihood »)

Le syllabe à reconnaître est pris comme étant une séquence d'observations $\mathbf{O^T=(O_1,O_2,\ldots,O_T)}$. **(L'algorithme II)**

Algorithme II : Algorithme de reconnaissance :

Etape 1 : Chargement des signaux de test (mot)

Etape 2 : Calcul de maximum de vraisemblance pour chaque fenêtre du signal

Etape 3 : Comparaison des probabilités $\mathbf{p_j(X_T/\lambda_i)}$,où **i=1,2,3** ,.. j=1, 2,3… sont les modèles Correspondant au représentant de chaque groupe.

Etape4 : Retour à l'étape 1, pour traiter un autre segment.

Etape 5 : traitant des résultats à l'étape 4, par un filtre logique.

4.2 présentation des étapes de traitement :

Avant de présenter les étapes de traitement nous rappelons les caractéristique de la base des données qui est une base dont la collecte des données a été effectuée au centre universitaire de Tébessa .sous des conditions d'enregistrement favorable; nous avons choisi 6 locuteurs masculin et 6 locuteurs féminin chacun a prononcé le mot dix fois.

la base de donnée est enregistrée sous format (wav), avec une fréquence d'échantillonnage de 11.025khz.chaque échantillon est codé sur 8bits, le codage choisi est le PCM ce qui nous donne un débit binaire de 11.025*8bits=88.2 kb/s.

4.2.1 La segmentation :

Comme nous avons opté pour l'approche analytique les mots doivent passer par une étape de segmentation que nous avons réalisée manuellement à l'aide du logiciel **wavsurfer (figure (4.2))** ce dernier permet de représenter le signal en 3-D (amplitude/fréquence/temps). Le temps et la fréquence respectivement sur l'axe horizontal et vertical, l'amplitude étant suggérée par la noirceur de l'affichage. Un spectre d'amplitude est obtenu par calcul d'une FFT sur une fenêtre à chaque instant du signal.

La *segmentation* du signal de parole vise à extraire la trace acoustique des unités sur lesquelles portera la "décision", le décodage nécessite de:

- Choisir les unités de décision;
- Opérer une segmentation correcte.

Plusieurs essais ont été effectués afin d'aboutir à une segmentation en syllabe raisonnable.

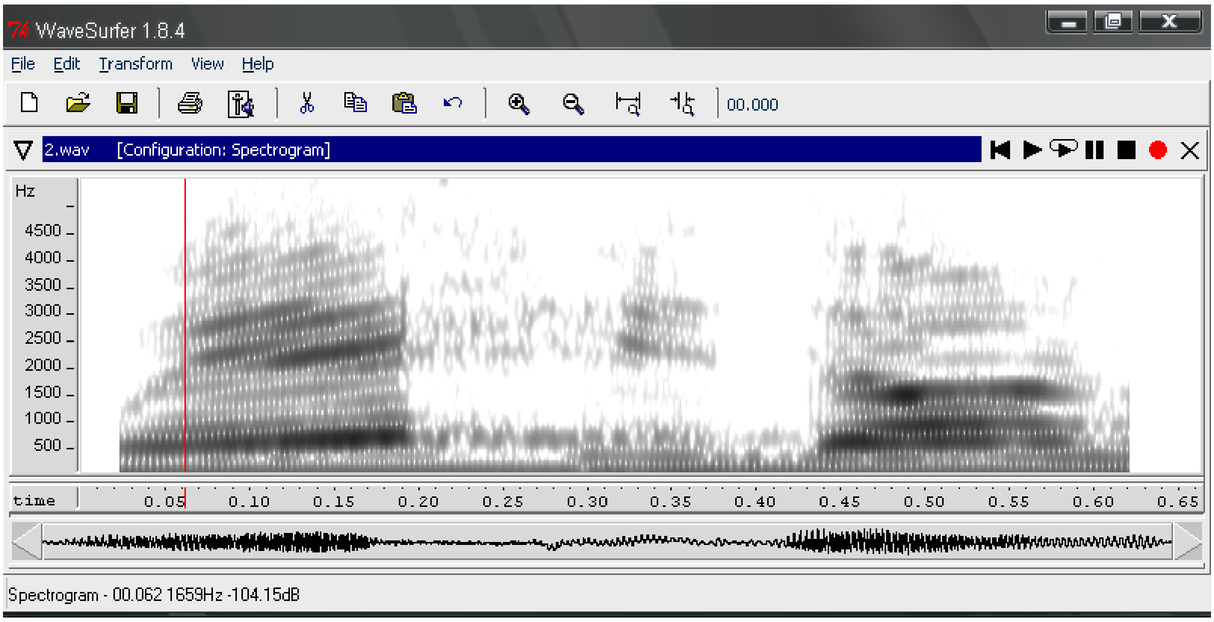

Fig. 4-2 : Logiciel Wavesurfer

Etapes de Traitement:

4.2.2 Prétraitement :

Une fois numérisé, le signal subit une opération de préaccentuation, qui consiste en un filtrage de type passe-haut qui relève le niveau des aigus, pour naturaliser l'atténuation des glottes qui est de 12db/octave les lèvres crée une accentuation de 6db/octave et il reste a ajouter par un filtre adéquat une autre augmentation de 6db/octave. En pratique, on utilise simplement un filtre de réponse impulsionnelle finie (1, a) avec a = -0,95. Si s (n) désigne le signal de parole et $s_p(n)$ le signal pré- accentué on a :

$$s_p(n) = s(n) - 0{,}95 s(n-1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

D'autre part on est conduit, dans la suite, à traiter les données en travaillant sur des trames de valeurs consécutives pondérés par la fenêtre de Hamming:

$$s_w(n) = s_p(n) w(n) \text{ où } w(n) = 0{,}54 - 0{,}46 \cos(2\pi n/(N-1)) \text{ avec } 0 \leq n \leq N-1 \ldots\ldots (4.2)$$

La longueur N d'une trame est choisie de façon à avoir des trames dont la durée est de l'ordre de 20ms. Enfin l'opération de découpage en trames de longueur N comporte un recouvrement de 50% entre trames successifs. Dans notre cas on a choisi des tanches successives de 30ms avec un recouvrement de 10 ms.

4.2.3 Paramétrisation du signal :

Une fois le signal est préaccentué, Pour résoudre les problèmes liés à la complexité de la parole, il est possible de calculer des coefficients représentatifs du signal traité. Ces coefficients sont calculés à intervalles temporels réguliers. En simplifiant les choses, le signal de parole est transformé en une série de vecteurs de coefficients.

Ces coefficients doivent représenter au mieux le signal a modéliser, et extraire le Maximum d'informations nécessaires à la reconnaissance. L'étude bibliographique que nous avons faite au chapitre1 sur les méthodes de modélisation, nous à permis de choisir 2 méthodes :le codage prédiction linéaire LPC et les coefficients cepstraux aussi appelés cepstres, qui seront utilisés dans notre système de reconnaissance.

Pour mener à bien une analyse LPC il faut choisir la méthode d'analyse et l'algorithme correspondant, ainsi que l'ordre de l'analyse :

- La méthode d'analyse choisie est la méthode de l'autocorrelation et l'algorithme correspondant est celui de burg pour des raisons de stabilité.
- L'ordre d'analyse conditionne le nombre de formants que l'analyse est capable de prendre en compte. On estime en général que la parole présente un formant par kHz de bande passante, ce qui correspond à une paire de pôles pour $A_p(z)$. Si on y ajoute une paire de pôles pour la modélisation de l'excitation glottique, on obtient les valeurs classiques de p=10, 12, et 18 pour fe=8, 10 et 16 kHz respectivement. Elles trouvent d'ailleurs une justification expérimentale dans le fait que l'énergie de l'erreur de prédiction diminue rapidement lorsqu'on augmente *p* à partir de 1, pour tendre vers une asymptote autour de ces valeurs : il devient inutile d'encore augmenter l'ordre, puisqu'on ne prédit rien de plus.

4.3 Construction des modèles de Markov cachés :

Modèles de mots ou modèles de syllabe ?

Dans la partie précédente, nous avons vu la paramétrisation du signal de la parole. Dans cette partie, nous allons montrer que des formes phonétiques différentes peuvent être choisies pour la reconnaissance de la parole : on peut décider de modéliser les mots ou les phonèmes.

Dans le cadre de la reconnaissance de la parole continue, le système acoustique doit être fondé sur des phonèmes afin de limiter le nombre de modèles (le nombre de modèles « exploserait » si on utilisait des modèles de mots avec un vocabulaire de grande taille). Il faut cependant obtenir, pour chaque entrée du dictionnaire phonétique, un modèle qui lui est propre. Ces modèles sont alors obtenus par concaténation de HMMs de phonèmes. Dans notre système de reconnaissance, le vocabulaire se limitant à une dizaine de mots, on peut décider d'utiliser soit des modèles de mots, soit des modèles de phonèmes.

Pour les modèles de phonèmes (voir figure 4.3), on utilise le dictionnaire phonétique suivant, qui décrit la suite de phonèmes constituant chaque mot, avec éventuellement des variantes de prononciation pour chaque mot :

	MOT	LES PHONEMES ET SYLLABES	ARTICULATION
0	صِـفْـرٌ	ٌر - ف - ص	Si , ef, ron
1	وَاحِـــدٌ	دٌ - ح - و	Wa, hi, don
2	إِثْـــنَـانِ	نِ - ن - ثْ - إِ	Ii, eth ,na ,ni
3	ثَـلا ثَـة	ة - لا - ثَ	Tha, la, ten
4	أَرْبَـعَــة	َع - بَ - رْ - أَ	A, ar, ba, aa
5	خَـمْـسَـة	سَ - مَ - خ	Kha, em, ssa
6	سِــتَــة	تَ - سِ	Ssi, ta,
7	سَـبْـعَـة	عَ - بْ - سَ	Ssa, eb, aa
8	ثَـمَـانِيَة	ي - ن - م - ث	Tha, ma, ni, ya
9	تِـسْعَـة	عْ - س - تِ	Ti, ess, aa,

Tableau 4.1: Le dictionnaire acoustique pour des modèles de phonèmes

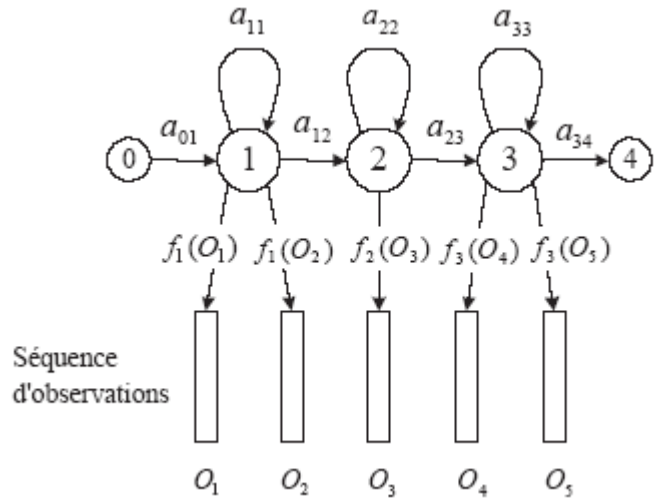

Fig. 4-3 : HMM de type 'gauche-droite' modélisant un phonème.

À l'aide de ce dictionnaire et des HMMs de chaque phonème, il est possible de construire, par concaténation, un ou plusieurs modèles pour chaque mot.

Fig. 4-5 : Concaténation de modèles de phonèmes pour le mot"خَمْسَةٌ "

Dans le cas d'utilisation de modèles de mots, chaque mot est considéré comme une unité phonétique.

Lorsque le vocabulaire est limité, les modèles de mots sont plus simples que les modèles de phonèmes: le nombre de modèles acoustiques (modèles de Markov Cachés) est alors plus petit que dans le cas de modèles de phonèmes. Le temps d'apprentissage et le temps de test sont plus faibles aussi.

L'inconvénient principal des modèles de mots est l'augmentation du nombre de modèles acoustiques quand la taille du dictionnaire (c'est à dire le nombre de mots que l'on veut pouvoir reconnaître) augmente. Par conséquent, on utilise toujours des modèles de phonèmes pour les systèmes de reconnaissance à grand vocabulaire.

Dans nos implantations, nous avons utilisé des modèles de phonème.

Choix de nombre d'état N :

- Dans le cas de modèles de phonèmes, les modèles HMM sont souvent à 3états.
- Deux méthodes dans le cas de modèles de mot se présente :
  - Le nombre d'état correspond au nombre de phonèmes dans le mot (5à10).
  - Le nombre d'état correspond au nombre de trames dans le mot, dans ce cas chaque état correspond à un intervalle de temps de 20 à30ms d'observation du signal.

Initialisation des HMM :

Comme dans tout algorithme de maximisation, le problème de choix de valeur initiale est crucial .Ici une façon simple d'opérer consiste à découper le signal correspondant de chaque mot en portions égales, on estime alors, dans chaque portion, la moyenne et pour chaque composante la covariance, cela donne le vecteur de moyenne initial et la matrice de covariance.

En ce qui concerne la matrice de transition, on choisit comme valeur initiale, une matrice de dimension N*N. (N : nombre d'états).

Avec :

$a_{ij}=0$ pour $i>j+\Delta$ (HMM gauche-droite) ……..(4.3)

Et

$a_{ij}=a_{ii+1}=a_{ii+2}=1/3$ (équipartition des 2états à droite)…….(4.4)

Et que le vecteur de probabilité de l'état initial est suppose égale à :Л=(1,0,…,0).

Le problème de l'underflow :

L'algorithme de Baum-Welch, contient à la fois des produits et des sommes dans les quantités qu'il calcule . En effet, cet algorithme d'apprentissage nécessite le calcul des probabilités partielles Forward, $\alpha_t(i)$ et Backward $\beta_t(i)$ pour $(1 \le t \le M) et (1 \le i \le N)$ D'après les formules récursives de calcul de a et de b, il est clair que lorsque M croit, a et b tendent vers 0. Pour un grand nombre d'observations, la fonction de probabilité aura ainsi une valeur trop petite pour être représentée dans un ordinateur.

Une première solution consiste à transposer tous les calculs dans le domaine Logarithmique. La multiplication est transformée en addition et l'addition de deux valeurs a et b à partir de leur logarithme est réalisé par :

$$Log(a+b)=loga+log(1+exp^{logb-loga}) \quad si\ loga>=logb \quad \text{……..(4.5)}$$

Ce qui limite considérablement les erreurs numériques.

Une autre solution propose d'introduire un facteur d'échelle qu'on multiplie par la Probabilité $\alpha_t(i)$ pour permettre à cette dernière de rester dans une échelle convenable. Une opération similaire est alors faisable pour le calcul de $\beta_t(i)$. Il suffit d'effacer toute trace de ce coefficient à la fin du calcul pour retrouver les valeurs réelles. Le facteur d'échelle utilisé est :

$$C_1 \frac{1}{\sum_{i=1}^{N} \alpha_t(i)} \text{……(4.6)}$$

Dans notre développement, nous avons utilisé cette dernière opération.

4.4 Phase de reconnaissance :

Le signal inconnu subit le même prétraitement c'est à dire préaccentuation ; Segmentation avec entrelacement ; la taille de chaque segment et égal à la taille du phonème le plus petit dans notre base ; élimination des effets de bord par l'application d'une fenêtre de Hamming.

Puis on compare chaque segment à tout les modèles HMM, avec la méthode du maximisation de vraisemblance on identifie le modèle correspondant au phonème inconnu (les résultat sont donnés dans le tableau ((4.2)).

- **Filtre logique :** Certain phonèmes ne sont pas reconnu par le système à cause de la segmentation d'où la nécessité d'un filtre logique pour apurer les résultats. Il permet de procéder à un vote majoritaire pour prendre une décision sur le mot.

  Exemple si le système reconnaît les trois syllabes don ef ron.

  On remarque dans le résultat ef ron correspondent dans le même ordre que les deux derniers syllabes du SIFR par contre

don est un syllabe final du wahidon il est clair le filtre logique va nous permettre d'éviter ce type d'erreur.

Finalement, pour tester l'efficacité du système le calcul du taux de reconnaissance est nécessaire qui est défini par:

$$\text{Taux de reconnaissance} = \frac{\text{(Nombre d'occurrences réussies)}}{\text{(Nombre d'occurrences totales)}}$$

4.5 Résultats et discutions :

4.5.1 l'influence des vecteurs des paramètres sur le taux de Reconnaissance :

Notre étude consiste à faire une comparaison entre deux représentations du signal :

- Cepstre en sortie d'un banc de filtres en échelle MEL (MFCC)
- Coefficients de prédiction linéaire (LPC).

Les résultats obtenus sont résumés sur le tableau suivant :

COEFFICIENT / MOT		LPC	MFCC
0	صفـر	20	80
1	واحـد	53	86
2	اثـنـان	80	93
3	ثـلاثـة	3	6
4	أربـعـة	13	30
5	خـمـسـة	53	86
6	سـتـة	26	73
7	سـبـعـة	60	13
8	ثـمـانـيـة	6	80
9	تـسـعـة	13	60

Tableau 4. 2 : influence des paramètres sur le taux de reconnaissance

Selon les résultats de tableaux (4. 2) on remarque que le taux de reconnaissance obtenu par les paramètres MFCC est plus grand que ce lui de LPC.

Pour la suite nous avons opté pour l'usage des paramètres MFCC.

4.5.2 Les résultas de reconnaissance des mots par rapport une base Des données des modèles HMM des syllabes :

On fait la reconnaissance de quelques mots (de 0 à 9) avec la base de données des modèles HMM des phonèmes qu'on à enregistrer.

Les résultats sont mentionnés sur le tableau (4. 4)

A titre d'exemple :

On prend le mot (صفـــر) avec numéro de répétition (N^0=1) :

Qui nous donne la séquence de segment comme suit :

6- 2-3-3

Chaque numéro de séquence présente un HMM donc on a :

HMM6-HMM2-HMM3-HMM3

Puis on compare chaque numéro de HMM avec la base de données des modèles HMM Dans le tableau (4. 3), qui nous donne : HMM6 : don

HMM2 : ef

HMM3 : ron

Le résultat donc est le mot : donefron et à l'aide du filtre logique on peut dire que ce mot est bien (صفـــر).

MOT	PHONEME	LES MODELES HMM
0	SI	HMM1
	EF	HMM2
	RON	HMM3
1	WA	HMM4
	HI	HMM5
	DON	HMM6
2	I	HMM7
	ITH	HMM8
	NA	HMM9
	NI	HMM10
3	THA	HMM11
	LA	HMM12
	A	HMM13
4	ERRE	HMM14
	BA	HMM15
	AA	HMM16
5	KHA	HMM17
	EMME	HMM18
	SSA	HMM19
6	SSI	HMM20
	TA	HMM21
	TOUN	HMM22
7	SA	HMM29
	EBB	HMM23
8	MA	HMM24
	NI	HMM25
	YA	HMM26
9	TEE	HMM27
	ISS	HMM28

Tableau 4. 3 : les différent modèles des HMM de la base de données

N^0 MOT	صفر	واحد	اثنان	ثلاثة	أربعة	خمسة	ستة	سبعة	ثمانية	تسعة
1	6-2-3-3	4-4-9-9-11	7-7-10-10-9-9-25-25	20-1-20-20-1029-20-11-10-12	16-14-6-16-19-15-15-22-1	17-22-18-23-23-2-6-18	20-8-7-11-18-23-22	29-22-16-19-18-6-1	11-1-12-10-10-3-5-23	27-28-4-21-17-28
2	1-2-2-3-3	4-4-5-7-6-11	7-7-10-9-9-25-25	7-7-11-1-12-7-22-5-6-10-28-28	16-22-15-16-19-18-15-1	17-19-24-18-23-18-22	20-8-7-2-6-6-6-18	19-23-15-21-18-6-22	11-24-9-10-26-12-6-17-22	27-2-4-21-21-6
3	1-2-23-3-18-2-28-28	4-4-5-7-6-11	7-7-10-9-9-25-25	1-12-12-12-5-23-1-6-18-17-1	16-22-15-15-19-14-6	17-22-24-18-23-2-6-19	20-8-7-6-7-7-6-22	1-22-19-16-23-6-22	11-1-9-10-10-26-5-28-22	27-2-4-21-28-28
4	1-27-7-7-3-10	8-21-26-29-7-11	7-2-25-9-25-25	20-26-8-12-21-11-28-18-25-7-7-7-1-20	13-28-23-16-19-21-11	17-18-28-11-28-4	8-25-2-11-28-11-25-7	29-11-17-19-2-11	11-25-24-25-25-9-2-11	27-28-19-21-7-11-22
5	1-20-8-7-3-3	4-12-5-7-8	7-7-25-9-9-25	1-20-29-11-11-19-2-1128-3	13-14-8-16-16-21-11-18	17-18-28-21-8-18-25	20-25-7-11-2-11-25	1-12-16-19-18-6	11-18-24-25-10-9-2-22-8	27-28-21-21-2-18
6	20-25-7-3-3	21-12-7-5-8-18	7-7-8-12-9-8-25	1-1-29-11-12-21-5-21-28-18-8-7-7-1-1	13-3-8-16-16-19-28-18	21-25-7-11-22-18	20-25-7-28 -11-28-4	1-12-23-16-23-11-8	11-18-26-25-10-9-2-18	27-2821-19-5-4
7	1-1-2-14-3-14	4-11-5-8-6-6	7-7-8-26-11-10-27-7	11-12-12-11-29-8-6-25-8	13-14-14-13-16-23-14-22	17-22-18-29-21-28-22	20-2-2-11-2-6-22	20-19-23-16-19-28-3-22	29-11-11-12-19-24-10-9-11-2-3-18	24-1-2926-26-26-19-2-22-22
8	1-1-1-12-29-18	4-11-5-18-6-11	7-29-8-9-9-8-7	29-29-21-12-11-11-2-22-8-7	13-14-14-13-13-23-15-4-22	17-22-28-28-11-6-22-8	7-20-2-7-11-2-6-22	20-19-11-16-19-21-11-18	29-11-11-19-12-10-10-11-2-11-22	24-1-2-20-17-17-19-18-3-22-22

Tableau 4.4 : montre les différents résultats de reconnaissance des mots par rapport une base Des données des modèles HMM des syllabes

N^0	صفر	واحد	إثنان	ثلاثة	أربعة	خمسة	ستة	سبعة	ثمانية	تسعة

MOT										
9	1-1-2-1-3-1	4-11-11-8-6-22	7-20-8-9-9-8-7	29-11-12-21-8-11-2-18-8-7	13-3-14-13-16-15-28-22	17-22-18-29-11-2-22-8	20-1-2-28-11-2-6-8	20-19-11-17-17-21-22-22	11-18-12-12-256-10-9-11-2-22	12-1-2-29-19-16-21-16-29-22
10	1-2-1-1-1-2	4-12-12-12-25-22-19	7-7-8-9-9-9-25-25	20-1-20-1-11-12-11-5-11-5-6-6-22-2-23	15-3-23-15-15-15	17-17-18-23-23-18-6-6	1-20-28-28-28-21-11-6-22-23	20-1-17-29-16-15-19-6-6-14-14	23-23-11-11-10-10-24-11-10-23-23	20-28-28-17-17-14-6-22-23
11	1-1-2-3-3-3-1-14	4-12-11-5-23-22-22	7-7-8-9-9-9-25-25	1-1-29-11-11-11-6-22	13-19-6-14-13-13-16-15-15-13	13-3-18-14-18-18	20-27-28-28-26-21-6-6	1-17-19-16-16-23-15-6-23	11-23-21-10-10-25-11-22-6	28-28-28-17-17-18-17-22
12	12-2-2-23	4-12-21-5-5-22-14	7-7-8-9-9-9-25-25	20-20-20-11-12-12-11-11-6-14-19-18-18-18	15-15-14-15-16-16-15-16-22-19-14	17-14-18-18-23-6-22-18	20-27-28-28-21-11-6-14	1-13-14-16-17-19-6-23	11-23-21-21-10-9-11-5-2	28-28-28-16-19-19-28
13	1-2-2-6-3	4-11-12-12-5-5-26-8-22-22	7-7-8-24-24-5-26-25	29-1-1-20-11-10-12-12-5-19-11-6-6-18-18-18	13-19-6-22-13-13-13-15-19-17-23	17-13-17-23-18-21-23-6-22-22	27-27-28-2-11-11-22-22-23	20-1-29-11-21-17-11-19-22-22-22	11-11-11-11-12-24-9-11-11-6-6-24	27-28-2-19-16-19-23-6-22
14	1-2-2-3	4-29-21-11-7-26-3-6-22-22	25-7-8-12-24-24-12-26	1-20-1-29-11-11-12-12-12-21-11-6-6-22-18-7-29-29	13-22-22-2-13-16-15-19-19-22-14	17-13-17-19-29-11-23-22-22	20-27-10-28-11-11-22-22	20-1-29-23-21-21-21-19-6-22-24	11-24-11-12-12-9-11-2-6-22-22	27-28-28-17-17-18-18-23
15	1-1-2-6-13-6	4-4-12-5-5-5-6	7-7-7-12-24-24-12-26-7	29-20-1-21-11-24-12-12-21-6-6-2-18	13-15-6-14-13-16-21-19-6	17-13-18-18-21-23-6-22-24	20-28-28-2-21-11-22	20-1-29-23-21-17-21-19-22-19-23	11-24-11-11-12-12-9-11-11-6-22-23	27-28-28-28-17-23-18-22-22

Tableau 4.4 : montre les différents résultats de reconnaissance des mots par rapport une base Des données des modèles HMM des syllabes

On peut montrer quelques résultats de tableau (4. 4) sous forme graphique :

F : numéros de segment

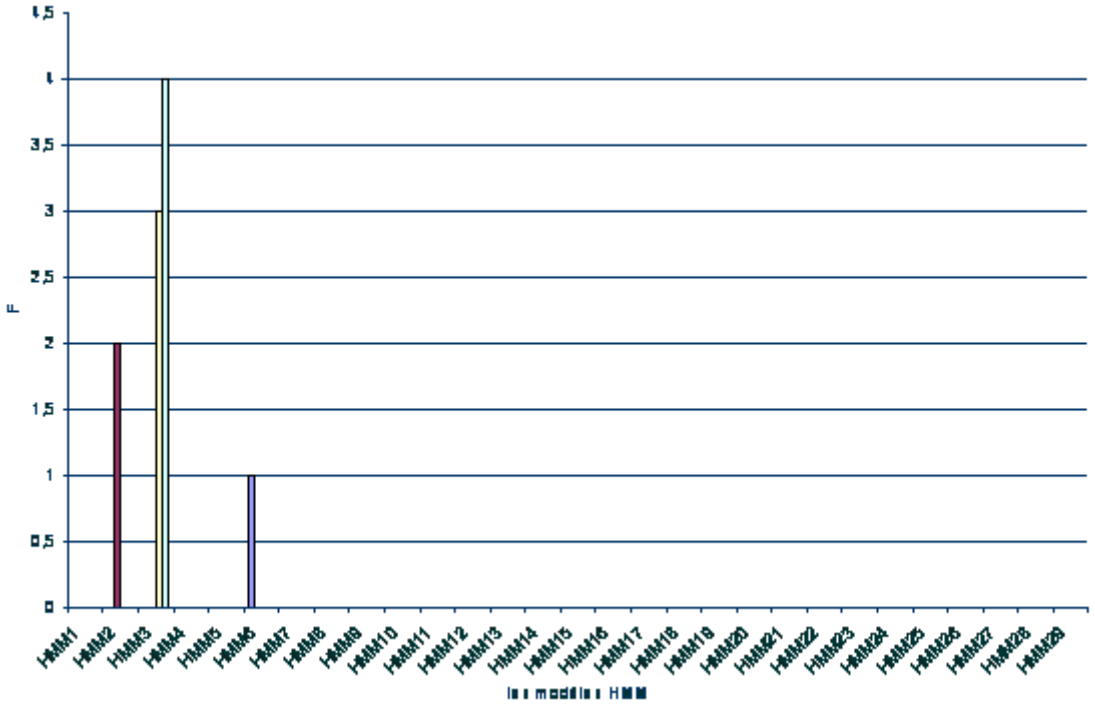

Fig. 4-6 : représentation du mot (صفـــر) avec numéro de répétition (N^0=1) , de séquence de segment(6-2-3-3).

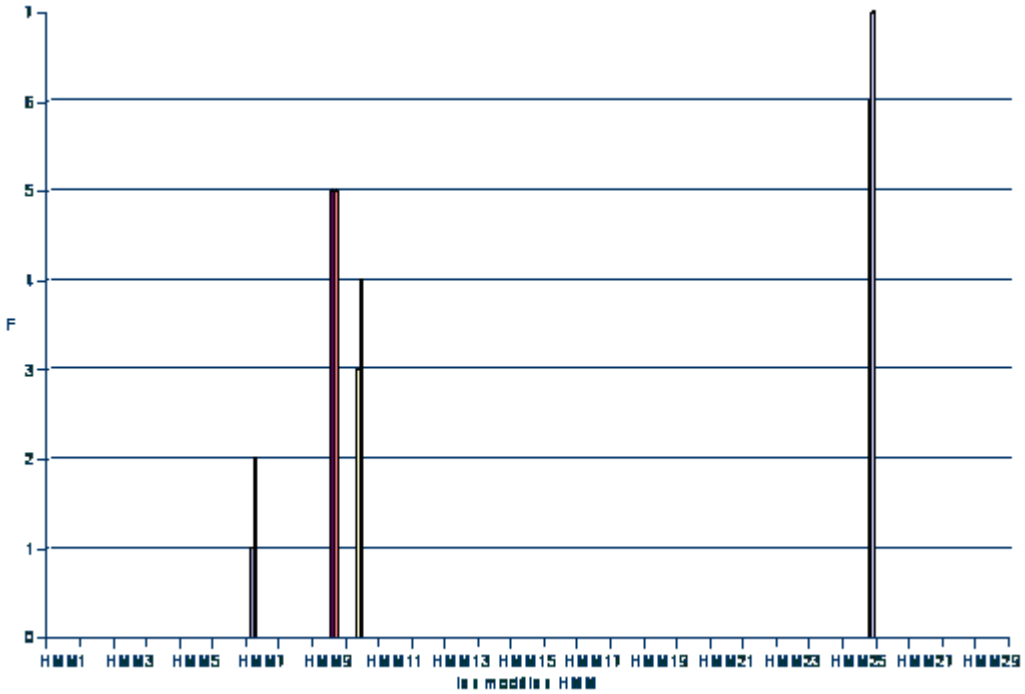

Figure 4.7 : représentation du mot (اثـنـان) avec numéro de répétition (N^0=1) , de séquence de segment(7-7-10-10-9-9-25).

4.5.3 Les résultats de comparaison entre modèles mot et modèle syllabe :

MOT		TAUX DE RECONNAISSACE DE MODELE MOT	TAUX DE RECONNAISSACE DE MODELE PHONEME
0	صفـر	60	80
1	واحـــد	100	86
2	اثـنـان	20	93
3	ثـلاثـة	40	66
4	أربـعـة	30	30
5	خـمـسـة	80	86
6	ســتـة	40	73
7	ســبـعـة	80	13
8	ثـمـانـيـة	20	80
9	تـسـعـــة	100	60

Tableau 4. 5 : le taux de reconnaissance des modèles mot et les modèles phonème.

On remarque que la comparaison de taux de reconnaissance du modèle de mot avec le modèle de phonème qui montre que le taux de reconnaissance des modèles de syllabe et plus élevé que ce lui du modèle de mot, et aussi le temps de calcul.

L'inconvénient principal des modèles de mots est l'augmentation du nombre de Modèles acoustiques ainsi que taille du dictionnaire et par la suite le temps de calcul.

4.5.4 l'influence de filtre logique :

Les résultats obtenus par l'application de filtre logique sur la reconnaissance de la parole sont résumés sur le tableau suivant :

MOT		TAUX DE RECONNAISSACE	TAUX DE RECONNAISSACE AVEC FILTRE LOGIQUE
0	صفـر	80	86
1	واحـد	86	86
2	اثـنـان	93	93
3	ثـلاثـة	6	26
4	أربـعـة	30	80
5	خـمـسـة	86	93
6	سـتـة	73	73
7	سـبـعـة	13	20
8	ثـمـانـيـة	80	80
9	تـسـعـة	60	80

Tableau 4.6 : le taux de reconnaissance avant est après filtre logique

D'après les résultas du tableau (4.6), on remarque que le taux de reconnaissance s'améliore par l'application du filtre logique.

4.6 Conclusion :

Nous avons présenté dans ce chapitre un système de reconnaissance basé sur la segmentation du signal vocal en syllabe (approche analytique) les résultats obtenus sont comparés à ceux obtenus par l'approche globale les résultats montent que l'approche analytique donne des meilleurs résultats en associant un filtre logique au système de reconnaissance.

Conclusion Générale

Conclusion Générale

- Dans Ce travail, nous avons présenté les bases théoriques et pratiques des systèmes modernes de reconnaissance de la parole, et plus particulièrement les modèles Markov cachées (HMM). Après plus de 25ans de recherche, ces modèles HMM offrent maintenant un très bon formalisme mathématique robuste.

Dans le cadre de ces modèles, nous avons alors une définition claire du problème de la reconnaissance de la parole qui consiste à recherches l'interprétation la plus probable d'une séquence de vecteurs acoustique. Ce problème peut alors être résolu à l'aide d'algorithmes efficaces (en termes de mémoire et CPU) dont la mise en œuvre a été fortement optimisée durant ces derniers années. Les algorithmes de recherche actuels, utilisant notamment des méthodes d'élagage sophistiquées, permettent en effet de développer des systèmes de reconnaissance de la parole continue grands lexiques tournant en temps réel sur la plupart des machines modernes. Ces algorithmes de reconnaissance ont également la propriété de traiter les entrées et sorties de façon strictement séquentielle dans le temps (entrées et sorties séquentielles et continues).

L'approche HMM est cependant limitée à l'apprentissage de paramètres statistique sur base de grands corpus d'apprentissage et peut donc être considérée comme un type particulier de modélisation « basée sur l'ignorance ».les performances du système dépendront donc uniquement des propriétés de généralisation du modèle sur de nouvelles observations, cette généralisation dépendant notamment de :

- La représentation des données d'entraînement, faisant intervenir non seulement la taille des bases de données mais aussi la qualité du prétraitement acoustique.
- La qualité des modèles HMM sous jacents, incluant notamment les hypothèses relatives à la topologie des modèles et hypothèse concernant les densités de probabilités.

Par conséquent, pendant la phase de reconnaissance ,le système calculera simplement l'explication la plus probable de toute séquence présentée à son entrée, et aura donc tendance à « halluciner » si cette séquence ne correspond à aucun mot lexique ou n'est même pas un signal de parole.

Finalement, les modèles HMM présentent encore de nombreuses propriétés intéressantes telles que :

- Les unîtes et niveaux linguistiques sont implicites dans la structure du modèle.
- Les paramètres peuvent facilement être partagés entre plusieurs modèles.
- On peut modéliser des signaux composés.
- Finalement, une segmentation explicite (en unités phonétiques)

N'est jamais requise mais peut être obtenue comme un sous-produit de la reconnaissance.

Malheureusement, malgré les progrès importants réalisés dans le domaine de la reconnaissance automatique de la parole ces dix dernières années, les performances des système existants laissent encore bien souvent à désirer du fait de leurs lexiques trop limités, de leur taux de reconnaissance trop faibles ou trop sensibles aux perturbations diverses, et des contraintes imposées à l'utilisateur.

Ce travail que nous avons mené concerne le développement d'un outil informatique pour la reconnaissance de la parole des syllabes arabes basé sur Les modèles de Markov cachés (HMM).

Le choix des paramètres initiaux des modèles a été effectué d'une façon aléatoire tout en respectant les conditions nécessaires.

Le corpus utilisé est un ensemble de prononciations de 0 à 9 en arabe, enregistrés en mono-locuteur et multi-locuteur, l'ordonnancement de la base de donnée nous à aidé à résourdre deux problèmes de la reconnaissance de la parole

- Le changement de locuteur.
- Ainsi la réduction du temps de calcul dans la phase de reconnaissance

Les tests que nous avons effectués sur le système de reconnaissance de la parole Qui utilisent des modèles de syllabes a montré leur fiabilité .ainsi, ils peuvent être utilisés pour les systèmes de reconnaissance à grand vocabulaire.

Nous remarquons que le choix adopter (reconnaissance par l'approche analytique) est bien justifié en faisant des résultats de reconnaissance obtenus par l'approche globale. le nombre d'états du modèle, le taux de reconnaissance peut être amélioré par l'utilisation d'une segmentation automatique.

Bibliographie

Bibliographie

[1] A. Leroi-Gourhan. Le geste et la parole ; tome 1 : technique et langage. Collection Sciences d'aujourd'hui, 324 pp, Éditions Albin Michel, 1992.

[2] NIST, The NIST 2002 Speaker Recognition Evaluation

[3] Liu M., Chen T., Ma C., Li X., Chang E., "MSRA NIST 2002 SRE System", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[4] Navratil J., Ramaswamy G., Chaudhari U., Zilca R., "IBM 1-Sp Detection Systems", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[5] Reynolds D., Campbell J., Dunn B., Jones D., Sturim D., Quatieri T., "MIT Lincoln Laboratory System: 1sp, 2sp and Segmentation", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002.

[6] Calliope, *"La parole et son traitement automatique"*, Masson et CENT-ENST, Paris, 1989, ISBN 2-225-81516-X

[7] Rabiner L. R., Juang B.H., *"Fundamentals of Speech Recognition"*, Prentice Hall, 1993, ISBN 0-13-015157-2

[8] Furui S., "An Overview of Speaker Recognition Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994

[9] Furui S., "Cepstral Analysis Technique for Speaker Verification", IEEE Transactions on Acoustic Speech and Signal Processing, vol 29(2), pages 254-272, 1981

[10] H. Hermansky, B. A. Hanson et H. Wakita. Low-dimensional representations of Vowels based on all-pole modeling in the psychophysical domain. Speech Communication, vol. 4, pp 181-187, 1985

[11] Liu M., Chen T., Ma C., Li X., Chang E., "MSRA NIST 2002 SRE System", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[12] R.Boite, h. bourlard, T. dutoit, j. hancqu et H.leich, « traitement de la parole » Novembre 1999.

[13] H. Hermansky. Perceptual linear predictive analysis of speech. Journal of the Acoustical Society of America, vol. 87, no 4, pp 1738-1752, 1990

[14] N. Morgan et H. Hermansky. RASTA extensions: robustness to additive and Convolutional noise. ESCA Technical Research Workshop: Speech processing in adverse conditions, pp 115-118, 1992.

[15] N. Morgan, H. Hermansky, H. Bourlard, P. Kohn et C. Wooters. Continuous speech recognition using PLP analysis with multilayer perceptrons. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 49-52,1991

[16] D. P. Morgan, C. L. Scofield et J. E. Adcock. Multiple neural network topologies Applied to keyword spotting. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 313–316, 1991.

[17] H. Hermansky et N. Morgan. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, vol. 2, no 4, pp 578–589, 1994.

[18] H. Hermansky. Perceptual linear predictive analysis of speech. Journal of the Acoustical Society of America, vol. 87, no 4, pp 1738-1752, 1990.

[19] M . BASSEVILLE (1982). A survey of statistical failure detection techniques . In Contribution à la détection séquentielle de ruptures de modèles statistiques . Thèse d'État, Université de Rennes 1, France.

[20] J . DESHAYES et D . PICARD (1983) . Ruptures de modèles en statistique . Thèses d'État, Université de Paris-Sud, Orsay, France

[21] M . BASSEVILLE and A. BENVENISTE, eds. (1986) . Detection of Abrupt Changes in Signais and Dynamical Systems . Lecture Notes in Control and Information Sciences, LNCIS 77, Springer-Verlag, New York

[22] M. BASSEVILLE and I. V . NIKIFOROV (1992) . Detection of abrupt changes : theory and applications . International Series in Information and Systems Science, Prentice-Hall, Englewood Cliffs, N .J . à paraître.

[23] R . Di FRANCESCO (1990) . Real-time speech segmentation using pitch and convexity jump models : application to variable rate speech coding . IEEE Trans . Acoustics, Speech and Signal Processing, vol . ASSP 38, n' 5, pp . 741-748 .

[24] A. Gattal, « Aquisition et segmentation de la parole » Thèse de Magistére de l'université de Annaba,juin 1994.

[25] A. Ganapathiraju et al. Syllable-Based Large Vocabulary Continuous Speech Recognition. IEEE transactions on speech and audio processing, vol.9, N°4, May 2001.

[26] C. Schrumpf et al. Syllable-based language models in speech recognition for English spoken document retrieval. Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib, 2005.

[27] C. Pallier. Syllabation des représentations phonétiques de Brulex et de Lexique. 2004.

[28] V. Bac « Reconnaissance automatique de digits en anglais en conditions bruitées » thèse de doctorat juin 2002

[29] L. BUNIET « Traitement automatique de la parole en milieu bruité :étude de modèles connexionnistes statiques et dynamiques » Thèse de doctorat, 1997

[30] P. HANNA : « Modélisation statistique de sons bruités : tude de la densité Spectrale, analyse, transformation musicale et synthèse » Thèse de doctorat Décembre 2003

[31] J. DANIEL : « Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia » Thèse de doctorat de l'Université Paris 6 . 2001

[32] A. Spalanzani : « Algorithmes évolutionnaires pour l'étude de la robustesse des systèmes de reconnaissance automatique de la parole ».
Thése de doctorat. l'Université Joseph Fourier - Grenoble I 1999.

[33] J-P Haton, J-M Pierrel, G. Perennou, J-L Gauvain, "Reconnaissance automatique de la parole", Edition Dunod, 1991.

[34] R. Vallet : « Application de l'identification des chaînes de Markov cachées aux communications numériques » Thèse de doctorat 1991

[35] C.ajols L.Miclet: « l'apprentissage de modèle de Markov caché » eyrolles 2002

[36] A. Kriouile, "La reconnaissance automatique de la parole et les modèles markoviens cachés modèles du second ordre et distance de Viterbi à optimalité locale", Thèse de doctorat de l'université de Nancy I, octobre 1990.

[37] Olivier Deroo, "Modèles dépendants du contexte et méthodes de fusion de Données à la reconnaissance de la parole par modèles hybrides HMM/MLP", Thèse De Doctorat de la Faculté Polytechnique de Mons, Laboratoire TCTS Mons, Décembre 1998.

[38] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol. 77, No. 2, Février 1989

[39] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP MAGAZINE JANUARY 1986.

[40] Mohamed Ali MAHJOUB, "Application des modèles de Markov cachés stationnaires et non stationnaires à la reconnaissance en ligne de l'écriture arabes", Thèse de Doctorat ,1999.