REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE FERHAT ABBAS-SETIF
FACULTE DES SCIENCES DE L'INGENIEUR
DEPARTEMENT D'ELECTRONIQUE

MEMOIRE

Pour obtenir le titre de **Magister**

En électronique

OPTION

Communication

Présenté et soutenu par

AZIL AHMED

THEME

AMELIORATION DES PERFORMANCES D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE A L'AIDE DE MESURES DE CONFIANCE

Date de soutenance :

Devant le jury composé de :

Président :	Mr. N. KHENFER	Prof. Université de Sétif
Rapporteur :	Mr.T.MOHAMADI	Prof. Université de Sétif
Examinateurs :	Mr. N. BOUKEZZOULA	M.C Université de Sétif
	Mr. A. BARTIL	M.C Université de Sétif

Remerciement

Je remercie tout d'abord allah notre créateur qui ma guidé dans le droit chemin et de m'avoir donné la force et la patience pour mener à bien ce travail.

J'exprime ma plus vive reconnaissance à monsieur T. MOHAMADI, professeur au département d'Electronique pour son suivi, ses conseils scientifiques et méthodiques, sa patience, ses soutiens morales et matériel et pour l'intérêt dont il a fait preuve dans l'accomplissement de ce travail, qu'il trouve ici ma profonde gratitude.

Je remercie chaleureusement monsieur N. KHENFER, professeur au département d'Electronique, qui me fait l'honneur d'accepter la présidence de jury.

Je remercie vivement monsieur B. BOUKEZZOULA, maître de conférence au département d'Electronique, pour l'honneur qu'il me fait en acceptant d'examiner mon travail et de faire partie de jury.

Je suis très reconnaissant à monsieur A. BARTIL, maître de conférence au département d'Electronique, d'avoir bien voulu me faire l'honneur de juger mon travail et de participer au jury.

Enfin, j'exprime mes vifs remerciements à mes chers parents et frères et sœurs, mon amis C. LAID qui ma encouragé et aidé.

SOMMAIRE

RESUME

La reconnaissance automatique de la parole a connu un progrès important ces dernières années, la détection des erreurs est l'une des technologies qui a aidé à améliorer les performances des systèmes de reconnaissance de la parole notamment la détection du hors vocabulaire.
L'utilisation de mesures de confiance basées sur le modèle acoustique peut se révéler très utile pour de nombreuses applications de la reconnaissance automatique de la parole.
L'efficacité d'un système de reconnaissance de la parole peut, en effet, être grandement amélioré si nous sommes capable de prédire si une hypothèse proposée par la RAP est correcte ou non.

Dans notre travail, on a conçu dans un premier temps un système de reconnaissance de mots isolés en langue arabe basé sur une approche probabiliste (l'utilisation des modèles de Markov cachés HMM).
Le processus de reconnaissance passe par deux étapes principales :
1- l'étape d'apprentissage qui consiste à ajuster les paramètres des modèles HMMs avec des densités de mélanges de multi gaussiennes en utilisant l'algorithme de Baum –Welch.
2- l'étape de reconnaissance qui consiste à calculer le maximum de vraisemblance (ML) entre l'observation et le modèle en utilisant l'algorithme de Viterbi.
Malgré le succès de cette approche, ce maximum de vraisemblance ne reflète pas la bonne liaison entre le mot prononcé et le modèle de référence.

Dans un second lieu, il s'agit d'appliquer une mesure de confiance en utilisant la technique de modélisation poubelle direct, cette technique est utilisée pour obtenir un score de référence pour le résultat de reconnaissance ,la confiance est définit comme la différence entre le score du modèle HMM et le score du modèle poubelle.
D'après les résultats obtenus, cette technique a permis de détecter un grand nombre du hors vocabulaire sans rejeter les reconnaissances correctes et ça en utilisant un petit seuil de rejet.

INTRODUCTION

Les applications pratiques de la reconnaissance automatique de la parole sont souvent confrontées à des utilisateurs peu conscients des contraintes du système et s'exprimant dans un environnement sonore bruité, comme c'est le cas, par exemple, pour les services vocaux interactifs sur le réseau téléphonique. Les systèmes de reconnaissance doivent alors être capables de rejeter les entrées incorrectes que sont les mots ou phrases ne faisant pas partie du vocabulaire de l'application ainsi que les diverses perturbations accidentelles (hésitations de l'utilisateur, bruit environnant, etc.).

De tels systèmes sont confrontés à trois grands types d'erreurs : [1]

- erreurs de suppression.
- erreurs d'insertion.
- erreurs de substitution.

Pour remédier à ce problème d'erreurs, il peut être très profitable de calculer une mesure de confiance pour chaque unité reconnue par le système (phonème, mot, etc.), c'est-à-dire d'associer à chacune de ces hypothèses de reconnaissance une mesure traduisant sa fiabilité. Cette information peut ensuite être comparée à un seuil pour permettre le rejet des hypothèses de reconnaissance les moins fiables.
Diverses méthodes ont été proposées pour calculer une mesure de confiance sur une hypothèse de reconnaissance W.

Notre travail consiste à concevoir un système de reconnaissance de mots isolés basé sur une approche probabiliste, ou on utilise les modèles HMMs, la décision de reconnaissance est basée sur un maximum de vraisemblance reliant le mot prononcé au modèle HMM de référence.
Ce maximum de vraisemblance est néanmoins relatif, il ne reflète pas la bonne liaison entre le mot prononcé et le modèle HMM. C'est pour ça, si un mot hors vocabulaire est prononcé, le système va lui attribuer un modèle HMM le plus semblant, ce qui va introduire une erreur de reconnaissance.

A fin d'améliorer la performance de notre système à savoir la détection du hors vocabulaire, on a appliqué une mesure de confiance basée sur le modèle acoustique dans le but de calculer un score de référence, cette mesure de confiance est appliquée en utilisant la modulation poubelle directe (on-line garbage modeling) [2]. Ainsi la décision de reconnaissance va être en fonction de cette mesure qui est la différence entre le score de référence et le score du modèle HMM.

Une acceptation ou un rejet du mot prononcé est déduite selon la comparaison de la confiance à un seuil prédéterminé.

Notre mémoire est constitué de quatre chapitres :

Le premier chapitre est une introduction à la reconnaissance automatique de la parole où une vision générale sur le fondement de la reconnaissance automatique de la parole est présentée ainsi qu'une explication sur l'approche probabiliste de la reconnaissance automatique de la parole.

Le deuxième chapitre aborde l'analyse acoustique du signal de la parole en vu de sa reconnaissance, ainsi que l'utilisation des coefficients MFCC pour l'extraction de l'information pertinente du signal de la parole.

Dans la première partie du troisième chapitre une étude théorique des modèles de Markov cachés (HMM) est présentée, où on a abordé les trois problèmes de Markov ainsi que les résolutions attribuées à chaque problème, à savoir les trois algorithmes de résolution.

La deuxième partie du troisième chapitre explique la mesure de confiance appliquée au système de reconnaissance automatique de la parole en vue de son amélioration. Là, on a utilisé la technique de modélisation poubelle directe.

Le quatrième chapitre donne une illustration de notre système ainsi que les expériences effectuées.

CHAPITRE I
INTRODUCTION A LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

I.1 Introduction :

Les histoires de l'homme et de la machine sont intimement liées, c'est ce qui a permis leur développement mutuel.

Tout à commencer par l'apparition de l'homme moderne, celui-ci a connu une véritable évolution grâce à l'acquisition du langage, au dix-septième siècle où l'apparition des premières sciences, l'homme a accéléré lui-même son processus d'évolution, c'est d'ailleurs à travers la maitrise de la technologie, la science et les machines qu'il a influencé lui-même le cours de son histoire. Aujourd'hui il tente de communiquer avec elles. L'utilisation de la parole comme mode de communication entre l'homme et la machine a été largement étudiée au cours des dernières années. Depuis 1945 les chercheurs commencent à s'intéresser au dialogue entre un homme et une machine, et le premier projet de recherche a véritablement commencé dans les années soixante-dix notamment avec le projet (ARPA) [Advanced Recherche Project Agency] lancé en1971 au Etats Unies d'Amérique.

Mais comment une machine pourrait elle se munir d'une " oreille" ? Par quel dispositif peut-on réaliser cette communication homme-machine ?

Comment la machine traite-elle les informations qu'elle reçoit ? En quoi cela peuvent-elles servir ?

Pour répondre à cela, nous allons tout d'abord découvrir en quoi consiste la reconnaissance de la parole et ses principales applications. Ensuite, nous verrons ses difficultés puis nous analyserons les différentes méthodes employées pour réaliser un système de reconnaissance de la parole

I.2 Qu'est que la reconnaissance automatique de la parole ?

Une définition de RAP a été proposée par Machowski [3], qui explique le rôle d'un système de reconnaissance de la parole réalisé autour d'un logiciel : " un logiciel de reconnaissance de la parole est un logiciel qui a la capacité de détecter la parole humaine et de l'analyser dans le but de générer une chaine de mots, sons ou phonèmes représentant ce que la personne a prononcé.

Donc la reconnaissance automatique de la parole est une conversion de la voix en fichier numérique, elle permet de décoder un signal acoustique de la parole en une suite de mots effectivement prononcés.

I.3 Principales applications :

Nous n'en sommes pas toujours conscients mais aujourd'hui la reconnaissance automatique de la parole nous offre une certaine autonomie, elle est de plus en plus fréquente dans notre vie quotidienne et permet dans le monde de hight-tech une certaine simplification des taches. Elle permet un gain d'autonomie en médecine. Par exemple, lorsqu' un chirurgien a les deux mains occupées, il peut parler pour demander une information technique au lieu de taper sur un clavier. Cette autonomie est aussi valable en industrie.

Mais la reconnaissance vocale est aussi utile dans différents domaines moins scientifiques, comme pour les serveurs d'informations par téléphone, ou encore la sécurité grâce à la signature vocale, et la possibilité de commande et de contrôle d'appareils à distance. Pour finir, elle peut être simplement associée à un traitement de texte : un locuteur parle et le texte s'affiche.

I.4 Les difficultés de la reconnaissance de la parole :

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

I.4.1 Le signal de la parole :

Le signal de la parole appartient à la classe des signaux acoustiques produits par des vibrations des couches d'air. Les fluctuations de la pression de l'air produisent des variations de ce signal, en fonction du temps, qui peuvent être enregistrées de façon analogique ou digitale. Ceci constitue une représentation élémentaire du signal de la parole [4]. Ce signal est le résultat du passage du flux laryngé (l'air qui est passé par nos poumons à travers nos cordes vocales) à travers le conduit vocal. Le signal de parole est donc une onde acoustique qui se propage dans un milieu donné (en général l'air) et qui est le résultat de la modulation par le conduit vocal d'une onde d'excitation.

Les phonèmes sont les éléments les plus brefs qui permettent de distinguer les différents mots. Un mot peut être considéré comme un ensemble de phonèmes. Plusieurs formes du conduit vocal peuvent produire le même phonème. La forme que le conduit vocal prend pour la production d'un phonème, dans un contexte donné, est assez variable et surtout dépendante de ce contexte. De ce fait, nous remarquons que les formes acoustiques associées à un phonème déterminé sont variables. Cette variabilité est double : d'une part la variabilité du contenu et d'autre part une variabilité de la durée du

phonème. Cette dernière variabilité résulte essentiellement du fait que le système articulatoire met en jeu des constantes mécaniques qui contrôlent les mouvements musculaires.

La parole est un signal quasi-stationnaire. Elle est formée de phonèmes et de transitions entre ces phonèmes. Plusieurs types de phonèmes existent : les voyelles, les consonnes fricatives et les consonnes plosives, les nasales et les liquides. Les voyelles sont des phonèmes voisés (l'excitation se fait par la glotte), leur production se fait généralement par un conduit vocal relativement ouvert et en absence de constrictions et leur prononciation peut être isolée et durable dans le temps. Les consonnes se caractérisent par une constriction dans le conduit vocal lors de leur production. Elles peuvent être voisées ou non voisés. Dans le cas des fricatives, la constriction génère un bruit local qui peut persister dans le temps et qui excite une partie du conduit vocal. Contrairement aux voyelles et aux fricatives, les plosives ne durent pas dans le temps, elles sont produites par un relâchement rapide d'une occlusion du conduit vocal, qui produit une perturbation locale se traduisant acoustiquement par un bruit impulsif de faible durée. Des transitions lient les phonèmes adjacents. D'une façon très simplifiée, les transitions acoustiques correspondent à des transitions dans l'appareil de production de l'état correspondant au premier phonème à l'état correspondant au suivant [5].

En conclusion la parole est un signal quasi-stationnaire formé de parties stationnaires et de transitions entre ces différentes parties. C'est un signal non déterministe, dans le sens ou deux réalisations d'un même mot auront nécessairement deux formes acoustiques différentes, même si elles sont produites par un même locuteur. Il faut distinguer deux types de variations : la variation acoustique et la variation temporelle non linéaire du signal.

I.4.2 Les difficultés liées au signal de la parole : [6] [7]

Le problème de la reconnaissance de la parole réside essentiellement dans la spécificité du signal vocal. Ce signal 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, age adulte …). La variabilité inter locuteurs est encore plus évidente. La hauteur de la voix, l'intonation, l'accent différent selon le sexe, l'origine social, régional ou national. Ainsi la parole est un moyen de communication ou de nombreux éléments entrent en jeu,

tel que le lieu, l'émotion du locuteur, la relation qui s'établit entre les locuteurs (stressantes ou amicales). Ces facteurs influencent la forme et le contenu du message. L'acoustique du milieu (milieu protégé ou environnement bruité), les mots hors- vocabulaire sont autant d'interférences supplémentaires sur le signal de parole que le système de reconnaissance doit compenser.

L'aspect continu du signal de parole complique encore la tâche de reconnaissance. En effet, lorsqu'on écoute parler une personne, on perçoit une suite de mots, alors que l'analyse du signal vocal nous permet de déceler aucun séparateur. Le même problème de segmentation se retrouve à l'intérieur du mot lui même. Celui-ci est perçu comme une suite de sons élémentaires : les phonèmes. L'analyse du signal ne permet pas aussi de découper en segments distincts le signal acoustique afin d'identifier les différents phonèmes qui le composent.

I.4.3 La variabilité intra locuteur :

Lorsque la même personne prononce deux fois de suite le même énoncé, on constate des variations sensibles sur le signal vocal. L'état physique par exemple, la fatigue ou le rhume et les conditions psychologiques, comme le stress, influent sur la production et sur la prosodie. Le niveau prosodique a un rôle primordial. Ce niveau concerne les informations extra linguistiques contenues dans la parole : la mélodie ou vibration du fondamental due aux émotions du locuteur, le rythme lié à la durée des phonèmes et l'amplitude. Ainsi seul l'examen de la mélodie permet de distinguer une phrase affirmative d'une phrase interrogative, ce qui change totalement le sens de la phrase.

I.4.4 La variabilité interlocuteur :

L'extrême diversité des réalisations acoustiques d'un même son par divers locuteurs constitue une des pierres d'achoppement majeures du décodage acoustico-phonétique. La variabilité interlocuteur est à priori la plus importante. Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonantes modifient la production acoustique. A cela s'ajoute les habitudes acquises en fonction du milieu social et géographique comme les accents régionaux.

I.4.5 Les conditions d'enregistrements :

Le système est-il capable de fonctionner proprement dans des conditions difficiles ? En effet les perturbations apportées par le microphone (selon le type , la distance et l'orientation) et l'environnement (bruit, réverbération) complique d'avantage le problème de la reconnaissance de la parole. Dés que le bruit ambiant n'est pas négligeable, il est difficile de connaitre le début et la fin des

mots même prononcés isolement. Un microphone placé trop prés de la bouche enregistre l'aspiration qui suit l'élocution, placé trop loin il entraine une baisse du rapport signal à bruit.
Ces facteurs compliquent la tache d'un système de reconnaissance automatique de la parole qui doit être capable de décider qu'une lettre prononcée par un adulte est plus proche de la même lettre prononcée par un enfant dans un environnement différent et avec un autre microphone.

I.5 Le décodage acoustico-phonétique :

Réaliser un système de reconnaissance de la parole est une tache très difficile, pour surmonter ces difficultés les chercheurs ont décomposé le processus de reconnaissance de la parole en deux étapes :

- le décodage acoustico-phonétique
- la reconnaissance

Le décodage acoustico-phonétique consiste à décrire le signal acoustique de parole en termes d'unités linguistiques discrètes. Les unités les plus utilisées sont les phonèmes, les syllabes, les mots.

Le décodage acoustico-phonétique a pour but de segmenter le signal en segments élémentaires et de les étiqueter. Le principale problème est de choisir les unités sur les quelles portera le décodage .même si les phonèmes sont souvent utilisés, il est possibles de prendre des unités plus longues tels que les mots.

Une fois la segmentation effectuée, l'identification des différents segments se fait en fonction de contraintes phonétiques, linguistiques ….

I.6 Les techniques de la reconnaissance de la parole : [6] [8]

Le problème de la reconnaissance de la parole est abordé généralement selon deux approches, l'une plus globale et l'autre plus analytique.

I.6.1 L'approche globale :

Ici, l'unité de base sera le plus souvent le mot considéré comme une unité globale, c'est à dire non décomposé. L'idée de cette méthode est de donner au système une image acoustique de chacun

de mots qu'il devra identifier par la suite. Cette opération est faite lors de l'apprentissage, où chacun des mots est prononcé une ou plusieurs fois. Le processus de reconnaissance de la méthode globale consiste à comparer le mot à reconnaitre à tous les mots de référence du vocabulaire. Le mot ressemblant le plus au mot prononcé est alors reconnu.

Cette méthode a pour avantage d'éviter les effets de coarticulation. C'est à dire l'influence réciproque des sons à l'intérieur des mots. Cette méthode représente deux inconvénients : le premier est relatif à la durée d'un mot qui est variable d'une prononciation à l'autre. Et le deuxième aux déformations qui ne sont pas linéaires en fonction du temps.

L'approche globale se révèle vite insuffisante si l'on veut traiter un grand vocabulaire ou de la parole naturelle continue, il est alors nécessaire d'adopter une nouvelle approche.

I.6.2 L'approche analytique :

Cette approche permet d'aborder le problème de la reconnaissance de la parole continue éventuellement pour plusieurs locuteurs. D'une manière très simple, l'approche analytique consiste à segmenter le message en constituants élémentaires, puis à identifier ces derniers, et enfin à reconstituer la phrase prononcée par étapes successives.

Ces constituants élémentaires peuvent être des phonèmes, des diphones ou des syllabes. Le processus de reconnaissance dans une méthode analytique peut être décomposé en deux opérations :

1- représentation du message (signal vocal) sous la forme d'une suite de segments de parole.
2- Interprétation de segments trouvés en termes d'unités phonétiques.

I.7 Les modes de reconnaissance : [6] [9]

I.7.1 Reconnaissance de mots isolés :

La segmentation d'un message parlé en ses constituants élémentaires est un sujet difficile. Pour l'éviter de nombreux projets de la reconnaissance de la parole se sont intéressés à la reconnaissance de mots prononcés isolement. La reconnaissance des mots isolés ou tous les mots prononcés sont supposés être séparés par des silences de durée supérieure à quelques dixièmes de secondes. Les systèmes spécialisés dans ce type de reconnaissance d'un nombre limité de mots peuvent être considérés comme des systèmes sur mesure ; la reconnaissance est robuste pour ces mots, elle supporte

des conditions de bruit sévères et une grande variété de prononciations. Les systèmes sont fondés aussi bien sur des algorithmes de type comparaison dynamique que sur la modélisation markovienne pour des résultats très comparables.

I.7.2 Reconnaissance de mots connectés :

Que se passe-t-il lorsqu'on passe de la reconnaissance de mots isolés à la reconnaissance de la parole continue, sans silence obligatoire entre les mots ? Evidemment, la prononciation de chaque mot peut être altérée par le phénomène de coarticulation entre les mots. En principe, il faudrait comparer la séquence d'entrée à toutes les références possibles. Il est clair qu'en générale le nombre de ces combinaisons est beaucoup trop élevé, résultant en un besoin de place mémoire et en temps de calcul prohibitif. Cependant, pour résoudre le problème des mots connectés, il faut trouver des solutions aux problèmes suivants :

1- La connaissance de nombre de mots dans une séquence.
2- La connaissance de la fin et du début de chaque mot dans la séquence.
3- La coarticulation entre mots adjacents.
4- L'accroissement de l'espace mémoire et la quantité de calcul requise.

La reconnaissance de mots connectés nécessite la définition d'un vocabulaire de base et celle d'une syntaxe rigide. L'approche privilégiée est actuellement l'approche markovienne.

I.7.3 La reconnaissance de la parole continue :

Si la plupart des applications peuvent se limiter à la reconnaissance de quelques mots-clés, il n'en est pas de même pour les applications "grand public" (serveurs vocaux, bornes d'informations) où une telle lourdeur risque d'entraîner un phénomène de rejet de la part des utilisateurs. La majorité des systèmes de dictée automatique rencontrés actuellement fonctionne en "mots isolés", du fait de la maîtrise encore très incertaine de la reconnaissance en continu. Cette technique qui apporte un confort d'utilisation indéniable est beaucoup plus complexe que la précédente en raison de phénomènes de coarticulation ou de liaisons entre mots contigus. Les systèmes les plus rudimentaires utilisent la technique du "Word Spotting". On se base sur le fait que la reconnaissance de quelques mots clés suffit généralement à reconnaître le sens de la phrase. D'autres systèmes, de complexité moyenne, réduisent la difficulté de la tâche de reconnaissance en conseillant au développeur de contraindre fortement le

dialogue. Ainsi par le choix judicieux des questions posées par le système, la variabilité des réponses peut diminuer fortement.

I.7.3.1 Système mono locuteur :

En raison de la variabilité importante du signal de parole entre locuteurs différents, de nombreux systèmes ne peuvent fonctionner qu'avec un seul locuteur. Les plus simples se contentent de stocker et de rapprocher les différentes prononciations d'un même mot, ce qui suppose de la part de l'utilisateur un entraînement préalable du système. D'autres possèdent déjà une représentation standard des unités phonétiques, réalisée par le constructeur, complétée par une phase d'apprentissage durant laquelle on améliore le modèle en fonction des caractéristiques de la voix de l'utilisateur. C'est le cas du système de dictée automatique d'IBM par exemple.

I.7.3.2 système multi locuteurs :

Les systèmes multi locuteurs peuvent être utilisés instantanément sans phase d'apprentissage. Pour cette raison, ils sont plus précisément adaptés aux applications visant un large public. Leur mise au point requiert cependant de la part du constructeur un travail important. La technique de Word-Spotting, bien qu'actuellement encore très utilisée dans les produits industriels pour des raisons à la fois techniques et économiques, va dans la mesure où les contraintes hardware (poids, volume) et coûts le permettront, progressivement céder le pas à la reconnaissance de parole continue. Celle-ci, sur des systèmes à grand vocabulaire, est d'ores et déjà une réalité. Si peu de produits sont actuellement industrialisés, de nombreux laboratoires possèdent déjà des prototypes en état de marche (pour un fonctionnement, il est vrai, rarement en temps réel). L'aspect multi locuteur ne semble pas constituer une difficulté pour ces systèmes. Le problème se pose pour le choix d'un échantillon suffisamment représentatif de la population visée pour alimenter la base de données. La puissance de calcul requise doit être très importante. Elle est cependant à la portée de quelques processeurs RISC mis en parallèle. Il ne semble pas qu'un accélérateur matériel spécifique apporte un surcroît de performance significatif. Devant la progression constante de la capacité de calcul des processeurs d'usage général, on peut penser que d'ici quelques années, un seul processeur sera suffisamment puissant pour traiter le problème. Le principal obstacle à la généralisation d'applications de reconnaissance de la parole réside

en la difficulté à constituer des bases de données réutilisables. La disponibilité de nouveaux produits fiables est également liée à la taille du marché visé pour le pays concerné. On trouve ainsi quelques systèmes intégrant le japonais, le cantonnais, le portugais "brésilien" ou même l'anglais "africaaner", bien que ces langues ne soient pas très répandues en dehors des pays concernés.

I.8 L'approche probabiliste de la reconnaissance automatique de la parole : [10]

Dans le cadre d'une application de la reconnaissance automatique de la parole, trois facteurs principaux interviennent, (Figure I.1).

Le lecteur, qui à partir d'un message M (suite de mots) qu'il veut transmettre produit un signal acoustique S (t). L'analyseur acoustique qui à partir du signal S (t) produit une paramétrisation sous ²forme d'une suite de vecteurs (séquence d'observations O) contenant l'information pertinente pour la reconnaissance.

Un décodeur dont le rôle consiste à déterminer à partir de la séquence d'observation O, la séquence des mots qui correspond au message M.

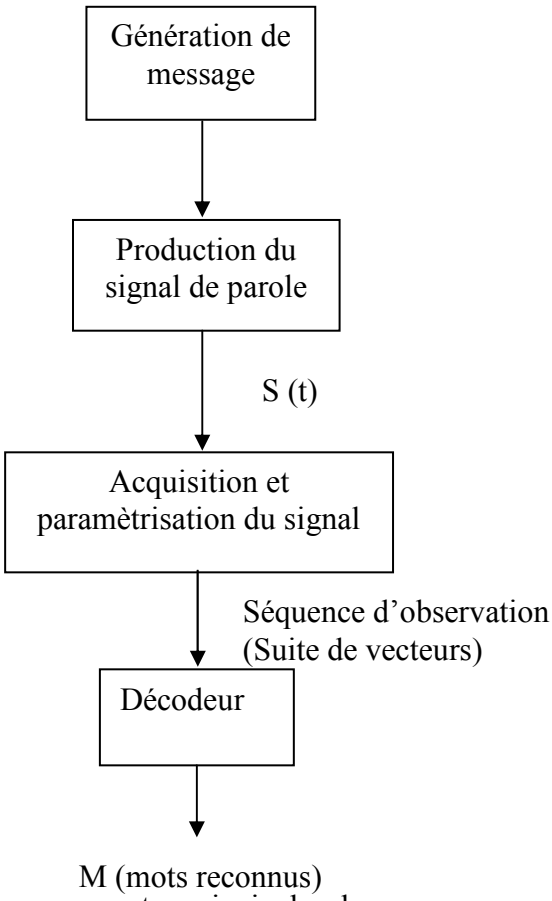


Figure 1.1 : Les composantes principales du processus de la reconnaissance de la parole

La reconstitution d'un message M inconnu à partir d'une séquence d'observation O, consiste à retrouver, parmi tous les messages possibles, celui qui selon toute vraisemblance, correspond à O, l'utilisation de la règle de Bays permet de décomposer la probabilité p(M/O) en deux composantes :

$$\hat{M} = \arg_m p(M/O) = \arg_m \max \frac{p(M)p(O/M)}{p(O)} \quad \text{(I.1)}$$

Le dénominateur est constant pour tous les messages possibles, donc on peut l'omettre, et $\hat{M}$ sera alors écrit sous la forme suivante :

$$\hat{M} = \arg_m \max p(M)p(O/M) \quad \text{(I.2)}$$

Ainsi, l'étape de reconnaissance consiste à déterminer la suite des mots $\hat{M}$ qui maximise le produit des deux termes $p(M)$ et $p(O/M)$. Le premier terme représente la probabilité à priori d'observer la suite des mots M indépendamment du signal. Cette probabilité est déterminée par le modèle de langage. Le deuxième terme indique la probabilité d'observer la séquence de vecteurs acoustiques O sachant une séquence de mots spécifiques M. Cette probabilité est estimée par le modèle acoustique. La qualité d'un tel système de reconnaissance de la parole peut être caractérisé par la précision et la robustesse des deux modèles qui permettent de calculer ces deux termes $p(M)$ et $p(O/M)$.

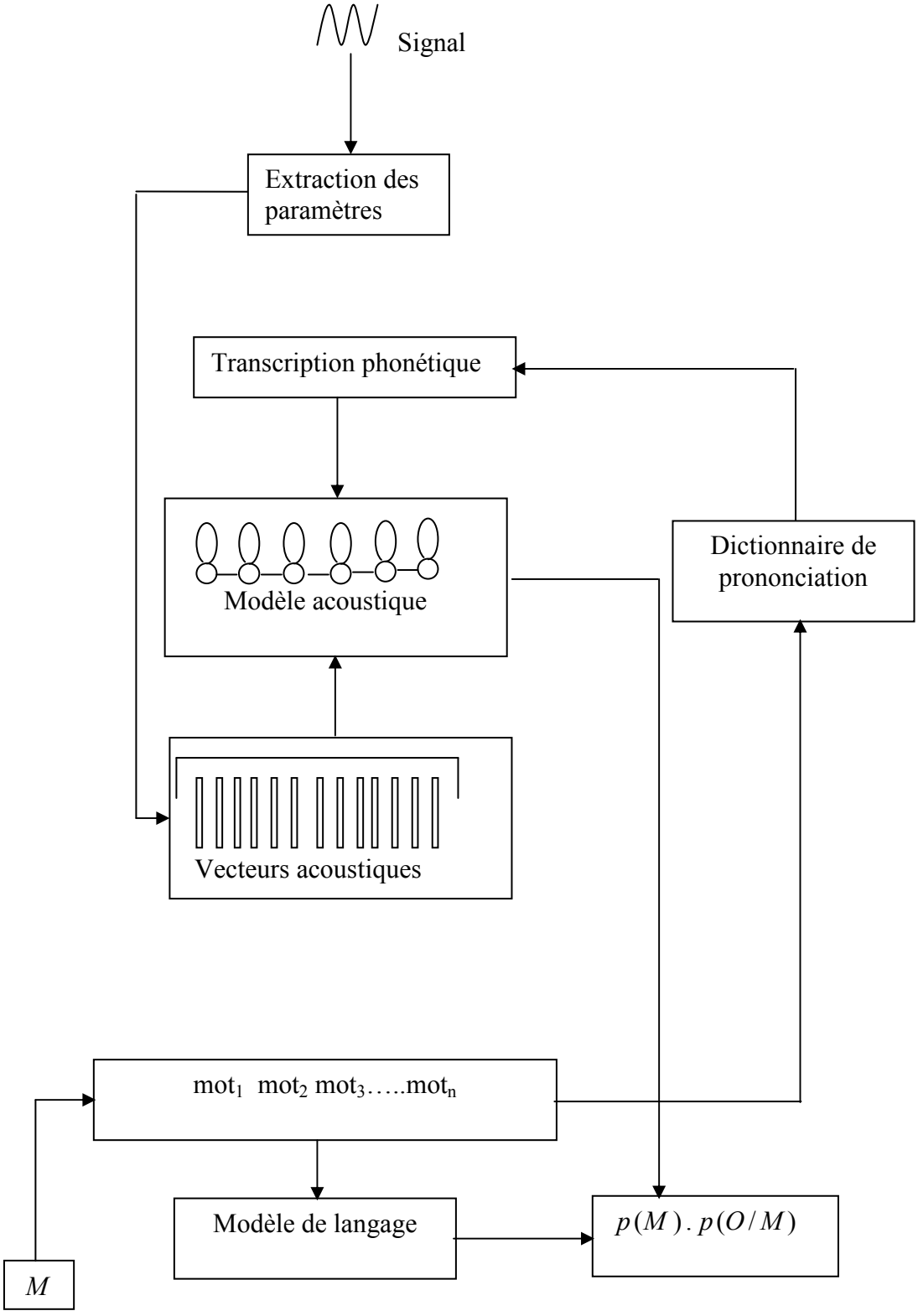


Figure I.2 : L'approche probabiliste de la reconnaissance automatique de la parole

L'outil statistique le plus utilisé et le plus performant, de nos jours, pour la modélisation acoustique est fondé sur les modèles de Markov cachés [11]. Les différentes étapes nécessaires à la reconnaissance d'une hypothèse donnée sont illustrées par la figure I.2. Tout d'abord le signal de la parole est subdivisé pour construire une séquence de vecteurs acoustiques. En utilisant ces vecteurs, le modèle acoustique se charge à partir des HMM_S de mots appris sur un corpus d'apprentissage, de construire la suite des mots hypothèses du signal prononcé. Un seul modèle HMM, représentant l'hypothèse, sera construit par la concaténation de l'ensemble des HMM_S de mots qui la compose et génère ainsi la probabilité du signal S (t), ce qui définit la probabilité $p(O/M)$. Ainsi à partir du dictionnaire des prononciations, la suite des mots hypothèses sera déterminée. Cette suite de mots sera évaluée par le modèle de langage pour estimer la probabilité $p(M)$. En principe, ce processus est répété pour toutes les hypothèses possibles. Le système donne enfin la meilleure hypothèse comme résultat de la reconnaissance.

L'espace de toutes les séquences de mots M augmente très rapidement avec la taille du vocabulaire. Il convient donc de restreindre la recherche à l'espace des séquences de mots les plus plausibles. Les applications récentes en reconnaissance de la parole utilisent souvent des modèles de langages stochastiques. Un modèle de langage est un automate à états finis dont les états représentent les mots du vocabulaire et les arcs les probabilités conditionnelles des transitions. Ces probabilités sont prises sur des corpus de textes de l'application en question.

Considérons le cas d'une séquence M constituée de la suite des mots m_i avec i $i \in \{1,2,.....L\}$

$$P(m) = P(m_1 m_2 ... m_L) = P(m_1) \prod_{i=2}^{L} P(m_i / m_1 m_2 ... m_{i-1}) \tag{1.3}$$

Dans la pratique en approxime $P(m_i / m_1 m_2 ... m_{i-1})$ par $P(m_i / m_{i-1})$

On parle dans ce cas de modèle de langage bigrammme. Ou par $P(m_i / m_{i-1} m_{i-2})$ et on parle alors du modèle de langage trigramme. Les modèles bigrammes et trigrammes sont les options les plus courantes, elles impliquent en général peu de restrictions grammaticales, puisque celles ci portent seulement sur des séquences de 2 ou 3 mots.

I.9 Conclusion :

La reconnaissance automatique de la parole est un domaine d'études très actif malgré les difficultés liées au signal vocal vue ses spécificités. Ce dernier possède des caractéristiques qui compliquent son interprétation et augmentent le nombre de données à traiter. Le problème de reconnaissance est abordé selon deux approches : l'approche globale et l'approche analytique, l'approche analytique permet d'aborder le problème de la reconnaissance de la parole continue.

Que la reconnaissance soit globale ou analytique, le processus de la reconnaissance de la parole commence par un prétraitement acoustique du signal vocal dans le but est de réduire le flux d'information et d'éliminer les redondances présentées dans celui-ci.

L'évolution rapide de la technologie informatique a permis la réalisation des systèmes de reconnaissance de la parole avec un cout raisonnable, fonctionnant au temps réel.

CHAPITRE II
ANALYSE ACOUSTIQUE DU SIGNAL DE LA PAROLE

II.1 Introduction :

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage, cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

L'importance particulière du traitement de la parole s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine.

Soit l'approche de reconnaissance de la parole est globale ou analytique, un module d'analyse acoustique de la parole est indispensable pour les deux types de reconnaissance, qui a comme objectif la transformation de la forme ondulatoire du signal de la parole en un certain type de représentation paramétrique (généralement à un taux d'information inférieur) pour un traitement ultérieur

L'objectif de cette phase de reconnaissance est d 'extraire les coefficients représentatifs du signal de la parole. Ces coefficients sont calculés à intervalles temporels réguliers. En simplifiant les choses, le signal de la parole est transformé en une série de vecteurs de coefficients, ces coefficients doivent représenter au mieux ce qu'il sont sensés modéliser et doivent extraire le maximum d'information utile pour la reconnaissance.

II.2 Mise en forme du signal de la parole : [12]

La mise en forme du signal de la parole consiste à transformer un signal analogique en un signal discrétisé. La figure II.1 illustre l'ensemble de ces opérations nécessaires à une telle mise en forme.

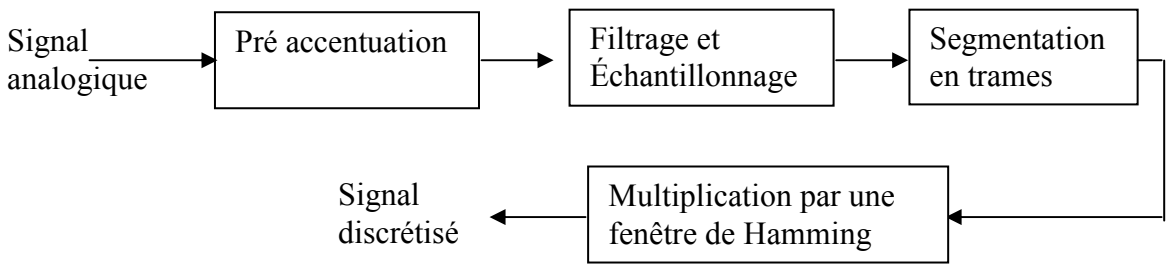


Figure II.1 : Mise en forme du signal

Une préaccentuation est effectuée sur le signal afin de relever les hautes fréquences, en suite le signal est filtré puis échantillonné à une fréquence donnée. Le nombre de filtres utilisés dans une telle analyse est choisi d'une manière empirique pour la reconnaissance de la parole. Ensuite le signal est segmenté en trames, chaque trame est constituée d'un nombre fixe de N échantillons de parole. En général N est fixé de telle manière que chaque trame corresponde à environ 30 ms de parole (durée pondant la quelle la parole peut être considérée comme stationnaire).

Enfin le fait de traiter un petit morceau de signal amène des problèmes dans le filtrage (effet de bord). Pour éviter cela, nous utilisons des fenêtres de pondérations, se sont des fonctions que l'on applique à l'ensemble des échantillons prélevées dans la fenêtre du signal original de façon à diminuer les effets de bord. Parmi les fenêtres les plus courantes, nous pouvons citer la fenêtre de Hamming.

En général, les fenêtres successives se recouvrent et elles doivent avoir une langueur suffisante. En pratique on prend 256 ou 512 échantillons, avec un recouvrement par exemple de la moitié de la taille, c'est à dire 128 à 256 échantillons respectivement.

Ce traitement implique une hypothèse importante : le signal vocal est supposé stationnaire sur une courte période.

II.3 Extraction des paramètres :

Parmi les coefficients les plus utilisés et qui représentent mieux le signal de la parole en reconnaissance de la parole, nous trouvons les coefficients cepstraux , appelés également sceptre .

Les deux méthodes les plus connus pour l'extraction de ces cepstres sont : L'analyse spectrale et l'analyse paramétrique, pour l'analyse spectrale (par exemple : Mel scale Frequency Cepstral Coefficient (MFCC)), comme pour l'analyse paramétrique (par exemple : le codage prédictif linéaire (LPC)), le signal de la parole est transformé en séries de vecteurs calculés pour chaque trame.

Il existe d'autres types de coefficients qui sont surtout utilisés dans les milieux bruités, nous citons par exemple les coefficients PLP (perceptual lineaire coefficients). Ces coefficients permettent d'estimer les paramètres d'un filtre auto régressif en modélisant au mieux le spectre auditif. Il existe plusieurs méthodes et techniques permettant l'amélioration de la qualité des coefficients.

Ces coefficients jouent un rôle capital dans les approches utilisées pour la reconnaissance de la parole, en effet, ces paramètres qui modélisent le signal de la parole seront fournis au système de la reconnaissance de la parole pour l'estimation de la probabilité p (séquence / message). Dans notre travail, nous nous sommes limités à l'utilisation des coefficients MFCC, ces paramètres ont montré une bonne représentation des aspects perceptuels du spectre de la parole [13].

II.3.1 L'analyse par prédiction linéaire (LPC) [14]

La prédiction linéaire est une technique qui s'applique directement après l'échantillonnage et la quantification du signal de la parole. C'est une méthode permettant l'approximation du signal par un modèle. Pour cela, elle considère l'appareil phonatoire comme un modèle source-filtre linéaire. Par conséquent, un échantillon de parole peut-être prédit par une combinaison linéaire d'un certain nombre d'échantillons précédents.

$$s(n)=(a_1.s(n-1)+a_2.s(n-2)+.....+a_p.s(n-p))+e(n) \quad \text{(II.1)}$$

Où : s (n) représente le signal à l'instant n.

e (n) représente un bruit blanc dû à toutes les sources d'erreurs possibles (précision des termes, arrondis de calcul, ...).

Estimation du signal à l'instant n : $\tilde{s}(n)=\sum_{i=1}^{p} a_i s(n-i)$ (II.2)

Avec p l'ordre de prédiction.

L'erreur de prédiction est égale à : $\varepsilon(n)= s(n)-\tilde{s}(n)=s(n)-\sum_{i=1}^{p} a_i.s(n-i)$ (II.3)

Les coefficients de prédiction sont calculés, afin de minimiser cette erreur.

La transformée en z de $s(n)-\tilde{s}(n)$ (II.4)

donne :

$$\varepsilon(z)=s(z)-\tilde{s}(z)=s(z)-\sum_{i=1}^{p} a_i.z^{-i}.s(z)=s(z).\left(1+\sum_{i=1}^{p} a_i.z^{-i}\right)=s(z).h(z)\Rightarrow s(z)=\frac{\varepsilon(z)}{h(z)} \quad \text{(II.5)}$$

Si on se fixe $\varepsilon(z)$ par une constante ε , on a alors $s(z)=\frac{\varepsilon}{h(z)}$ (II.6)

En posant $$z = e^{j.2.\pi\, f_n} \quad \text{(II.7)}$$

on peut alors représenter la densité spectrale de puissance (DSP) de S(z) :

$$E\left\{\left|s(f_n)\right|^2\right\} = \frac{\sigma^2}{\left|1+\sum_{i=1}^{p} a_i.e^{-i.j.2.\pi.f_n}\right|^2} \quad \text{(II.8)}$$

Où $\sigma^2 = E(\varepsilon^2)$: puissance de l'erreur.

La détermination de la densité spectrale de puissance revient donc à calculer les coefficients de prédiction a_i. Donc, l'analyse par prédiction linéaire permet de passer d'un spectre échantillonné, donc « bruité »" à une représentation spectrale continue et « lissée ».La détection des formants en est alors plus aisée. Cette méthode présente l'inconvénient du choix du nombre de coefficients (8 à 14) à prendre en fonction de la fidélité par rapport au signal analysé.

II.3.2 L'analyse par MFCC : [15] [16]

Après une mise en forme du signal (comme à la plus part des méthodes de l'analyse de la parole), une transformée de Fourier discrète (D.F.T : discret Fourier transform), en particulier FFT (Fast Fourier transform), est appliquée pour passer dans le domaine fréquentiel et pour extraire le spectre du signal. Ensuite le filtrage est effectué en multipliant le spectre obtenu par les gabarits des filtres.
Ces filtres sont en général, soit triangulaires soit sinusoïdaux. Dans nos expériences nous avons choisi d'utiliser des filtres triangulaires dont la largeur de la bande passante est relativement constante sur une échelle de Mel.

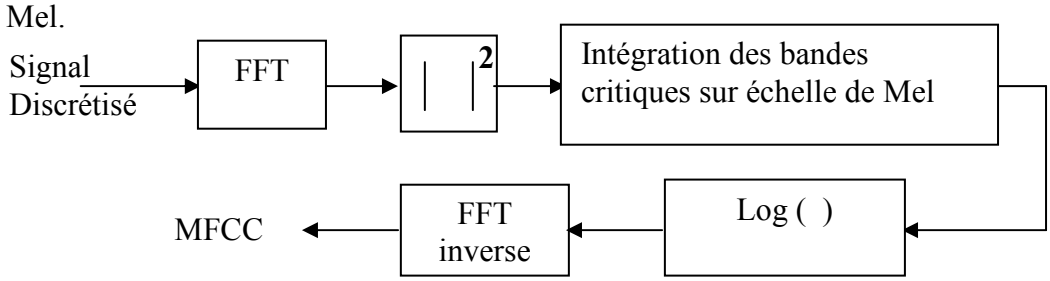


Figure II.2 : Calcul des coefficients MFCC (Mel-Scale Frequency Cepstral coefficients)

Le traitement décrit dans le paragraphe précédent permet d'obtenir une estimation de l'enveloppe (densité spectrale lissée). Il est possible d'utiliser les sorties de banc de filtres comme entrée pour le système de reconnaissance. Cependant, d'autres coefficients dérivés des sorties d'un banc de filtre sont plus discriminants, plus robustes au bruits ambiant et moins corrèles entre eux, il s'agit des coefficients cepstraux dérivés des bancs de filtres répartis linéairement sur une échelle Mel, se sont les coefficients MFCC. Le cepstre est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale. Ceci a une interprétation du point de vue de la déconvolution homomorphique : alors que le filtrage linaire permet de séparer des composantes combinées linéairement, dans le cas de composantes combinées de façon non linéaire (multiplication ou convolution), les méthodes homomorphiques permettent de se ramener au cas linéaire. Pour le signal de la parole, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal considéré comme un filtre linéaire :

$$S(t) = e(t) * h(t) \qquad \text{(II.9)}$$

Ou S(t) est le signal de parole, e(t) est la source d'excitation et h(t) la réponse impultionnelle du conduit vocal.

Soit un signal x[n] discret avec, N est le nombre d'échantillons d'une fenêtre d'analyse, la transformée de fourrier est définie par :

$$X(k) = \sum_{n=0}^{N-1} x(n).e^{\frac{-j.2\pi.n.k}{N}} \qquad \text{Avec} \quad 0=k \prec N \qquad \text{(II.10)}$$

Le spectre du signal est filtré par des filtres triangulaires (voir figure II.3 et II.4) dont les bandes passantes sont équivalentes en domaine fréquence Mel. Les points de frontière b(m) des filtres en échelle de fréquence sont calculés à partir de la formule:

$$B(m) = B(f_l) + m.\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \leq m \prec M+1 \qquad \text{(II.11)}$$

- M désigne le nombre de filtres.
- f_h désigne la fréquence la plus haute du signal
- f_l désigne la fréquence la plus basse du signal

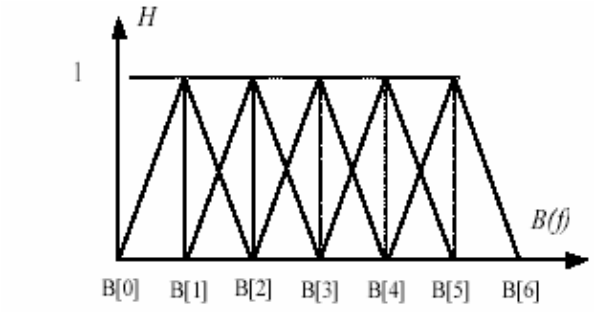


Figure II.3 : Filtre triangulaire passe bande en Mel fréquence B(f)

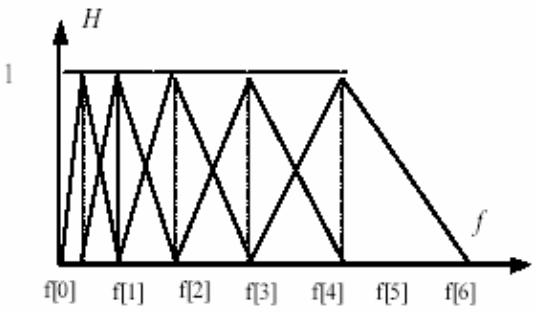


Figure II.4 : Filtre triangulaire passe bande en fréquence (f)

$$F(m) = \left(\frac{N}{F_s}\right) B^{-1}\left[B(f_l) + m.\frac{B(f_h) - B(f_l)}{M+1}\right] \quad \text{Avec} \quad 0 \le m \prec M+1 \tag{II.12}$$

$$\text{Et} \qquad B^{-1}(i) = 700.\left(10^{\frac{i}{2595}} - 1\right) \tag{II.13}$$

Les coefficients des filtres sont calculés par :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & si \quad f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & si \quad f(m) \le k \le f(m+1) \\ 0 & si \quad k \ge f(m+1) \end{cases} \tag{II.14}$$

Ensuite on multiplie les énergies de *X(k)* par les coefficients $H_m[k]$ et on calcule leur logarithme :

$$E[m] = \log\left[\sum_{k=0}^{N-1} \left|X(k)\right|^2 .H_m[k]\right] \tag{II.15}$$

Les coefficients MFCC de fréquence en échelle MEL sont obtenus par la transformée inverse des coefficients en sortie des filtres.

Remarque : Les coefficients MFCC sont obtenus par une transformée en cosinus qui génère de coefficients réels.

$$C[n]=\sum_{m=0}^{M-1} E[m]\cos\left(\frac{\pi.n.(m+\frac{1}{2})}{M}\right) \qquad \text{avec} \quad 0=m \prec M \qquad \text{(II.16)}$$

On ne peut citer toutes les études établies en utilisant les MFCC, toutefois le nombre de coefficients est aussi, en général pris égal à 13, puis réduit à 12 en considérant deux points essentiels :

- Le premier coefficient C_0 représente l'énergie de la trame et ne peut réellement contribuer à la segmentation ou la reconnaissance.
- Les coefficients de 1 à 12 représentent l'enveloppe cepstrale plus ou moins lissée, les hautes variations fréquentielles étant éliminés.

Le but final de l'extraction des paramètres est de modéliser la parole, un phénomène très variable. Par exemple, même si elle a de l'importance, la simple valeur de l'énergie n'est pas suffisante pour donner toute l'information porté par ce paramètre. Il est donc souvent nécessaire de recourir à des informations sur l'évolution dans le temps de ces paramètres. Pour cela, les dérivées première et seconde ($\Delta MFCC$ **,** $\Delta\Delta MFCC$ **)** sont calculées pour représenter la variation ainsi que l'accélération de chacun des paramètres.

II.4 Conclusion :

Soit l'approche de reconnaissance automatique de la parole globale ou analytique, un module d'analyse acoustique est indispensable pour les deux types de reconnaissance. Dans ce chapitre quelques méthodes d'analyse acoustique sont brièvement décrites, l'analyse acoustique par MFCC est détaillée, cette analyse va nous permettre d'extraire l'information pertinente du signal de la parole.

CHAPITRE III
SYSTEME DE RAP BASE SUR UNE APPROCHE PROBABILISTE ET MESURE DE CONFIANCE

III-1.Introduction :

Un problème majeur de la reconnaissance de la parole est de modéliser au mieux les unités représentatives du signal de parole, il existe en effet deux types de modélisation possibles des propriétés d'un signal donné :

- La modélisation déterministe qui exploite les propriétés intrinsèques du signal.
- La modélisation statistique, qui caractérise les propriétés statistiques du signal.

Dans ce travail nous utilisons des modèles statistiques : les models de Markov cachés qui est l'outil statistique le plus utilisé et le plus performant [11]

III.2Les modèles de Markov cachés:**[17]**

III.2.1 La chaîne de Markov:

III.2.1.1 : Définition : Une chaîne de Markov est considérée comme un processus aléatoire, c'est un automate à N états discrets.

Un processus de Markov est un système à temps discret se trouvant à chaque instant dans un état pris parmi N états distincts, les transitions entre les états se produisent entre deux instants discrets consécutifs selon une certaine loi de probabilité.

III.2.1.2 : Approche mathématique

Un modèle de Markov caché est défini par les données suivantes :

a) Graphe d'état:

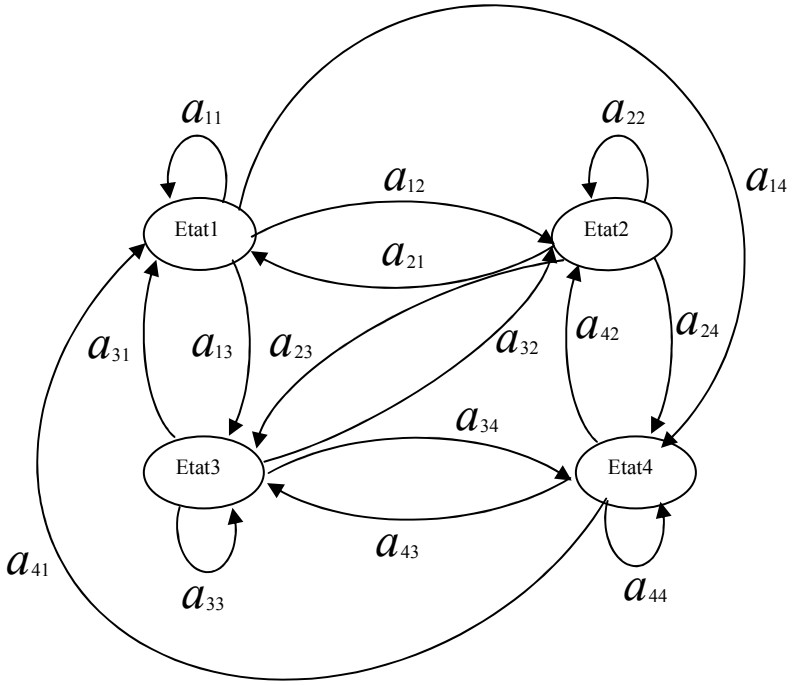


Figure III.1 : Graphe d'états d'une chaîne de Markov

b) La matrice de transition:

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$A=\{a_{ij}\} \text{ pour tout } (i,j)\in[1,N] \qquad \text{(III.1)}$$

Avec $a_{ij}=p(q_{t+1}=j/q_t=i) \quad \forall(i,j)\in[1,N]$ (III.2)

a_{ij} : La probabilité de prendre la transition de *i* vers *j*.

Avec : $a_{ij} \geq 0$ et $\sum_{j=1}^{N} a_{ij}=1 \quad \forall(i,j)\in[1,N]$ (III.3)

c) La matrice d'observation ou d'émission :

Les dimensions de cette matrice dépendent des observations; soient elle sont continues ou discrètes. Ses éléments sont définis telle qu'à chaque état une ou plusieurs observations sont associées à cette état défini par : $B(o_t / q_j)$ qui représente la probabilité d'observer ou d'émettre l'observation o_t en étant à l'état q_j.

- cas discret : les matrices contiennent des vecteurs d'observation de dimension finie N*T.
- cas continu : les matrices contiennent les paramètres de probabilités de distribution des observations par état, celles-ci peuvent être des gaussiennes, des multi-gaussiennes ou toutes autres distributions propre au contexte des données.

$B = \{bj\ (o_t)\}$ pour tout $j \in [1, N]$ et $t \in [1, T]$. (III.4)

Tel que $b_j(o_t) = b_j(o_t/q_t = j)$

$b_j(o_t)$: la probabilité d'émettre o_t en étant à l'état j.

Avec : $b_j(o_t) \geq 0$ et $\sum_{t=1}^{T} b_j(o_t) = 1$ (III.5)

d) Le vecteur d'initialisation :

Ce vecteur donne au modèle les probabilités de transitions initiales

$\pi = \{\pi_i\}$ Avec $\pi_i = p(q_1 = i) \quad \forall i \in [1, N]$. (III.6)

N.B: $\pi_i(o_t) \geq 0$ et $\sum_{i=1}^{N} \pi_i = 1$ (III.7)

Donc le modèle de chaîne de Markov est noté comme suit :

$\lambda = (A, B, \pi)$ Avec A: $N \times N$; B: $N \times T$ et $\pi = N \times 1$ (III.8)

III.3 Supposition mathématique : [17]

- les états possibles d'automate sont : $\{q_1, q_2, ..., q_N\}$
- les observations possibles sont : $\{o_1, o_2, ..., o_T\}$

III.3.1 : La règle de Bays :

$$p(q_1, q_2, ..., q_N) = p(q_1) \prod_{i=2}^{N} p(q_i / q_1^{i-1}) . \quad \text{(III.9)}$$

$$q_1^{i-1} = q_1, q_2, ..., q_{i-1} . \quad \text{(III.10)}$$

Si la variable aléatoire Q forme une chaîne de Markov de premier ordre :

$$p(q_i / q_1^{i-1}) = p(q_i / q_{i-1}) \quad \text{(III.11)}$$

$$p(q_1, q_2, ..., q_N) = p(q_1) \prod_{i=2}^{L} p(q_i / q_{i-1}). \quad \text{(III.12)}$$

III.3.2 Les suppositions de Markov:

- La probabilité de transition à l'état q_t ne dépend que de l'état au temps précèdent.

$$p(q_t / q_1^{t-1}) = p(q_t / q_{t-1}) \quad \text{(III.13)}$$

- La probabilité qu'une observation est émise au temps t dépend seulement de l'état q_t et indépendante du passé.

$$p(o_t / o_1^{t-1}, q_1^t) = p(o_t / q_t) \quad \text{(III.14)}$$

III.4 La différence entre HMM et VMM:[17]

III.4.1 Le modèle de Markov visible :

Le modèle de Markov visible est une chaîne caractérisée par :

- une séquence d'états visibles.
- Chaque état correspond à un événement observable.

Exemple : considérons trois états de chaîne de Markov pour la situation météorologique de la région de Sétif :

Etat 1 : pluie (p).

Etat 2 : froid (f).

Etat 3 : soleil(s).

La matrice de transition des états est :

$$\mathbf{A} = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Le vecteur initial des états : $\pi = \{\pi_i\} = (0.5 \quad 0.2 \quad 0.8)$.

Quelle est la probabilité pour que la situation météorologique sera pour huit jours consécutifs comme suit : {s, s, s, p, p, s, f, s} ?

Jour : {1, 2, 3, 4, 5, 6, 7,8}.

Solution :

P(o/model) = p (3,3,3,1,1,3,2,3/model)

= p (3)*p $(3/3)^2$*p (1/3) p (1/1)*p (3/1)*p (2/3)*p (3/2)

= 0.8*$(0.8)^2$(0.1)*(0.4)*(0.3)*(0.1)*(0.2).

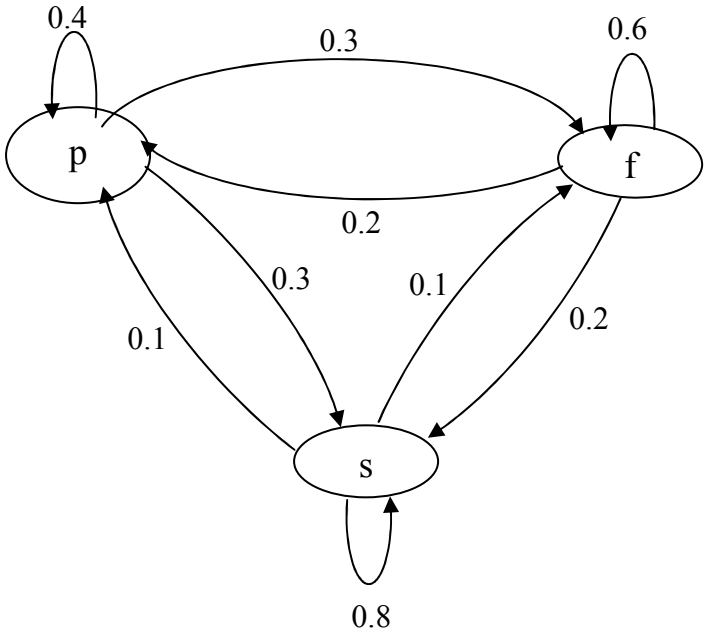


Figure III.2 : Graphe d'états de l'exemple

III.4.2 Le modèle de Markov caché :

Un modèle de Markov caché est caractérisé par :

- Une séquence d'états cachée (Hidden).
- L'état de chaque observation n'est pas observé, mais associé à une fonction de densité de probabilité.
- Le modèle de Markov est donc un processus doublement stochastique, dans lequel les observations sont une fonction aléatoire de l'état et que l'état change à chaque instant en fonction des probabilités de transition issues de l'état antérieur.
- Chaque état peut produire toutes les observations selon sa fonction *pdf* (fonction de distribution de probabilité).

Exemple : on considère trois régions (Sétif, Alger, Oran) comme les états de chaîne de Markov.

Les événements observés sont (soleil, pluie, froid).

Chaque état est défini par une distribution de probabilité {soleil, pluie, froid}.

On peut observer la séquence (soleil, pluie, froid), mais la séquence d'état correspondante est cachée.

La matrice de transition : $\mathbf{A}= \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$

Le graphe d'état de cet exemple est :

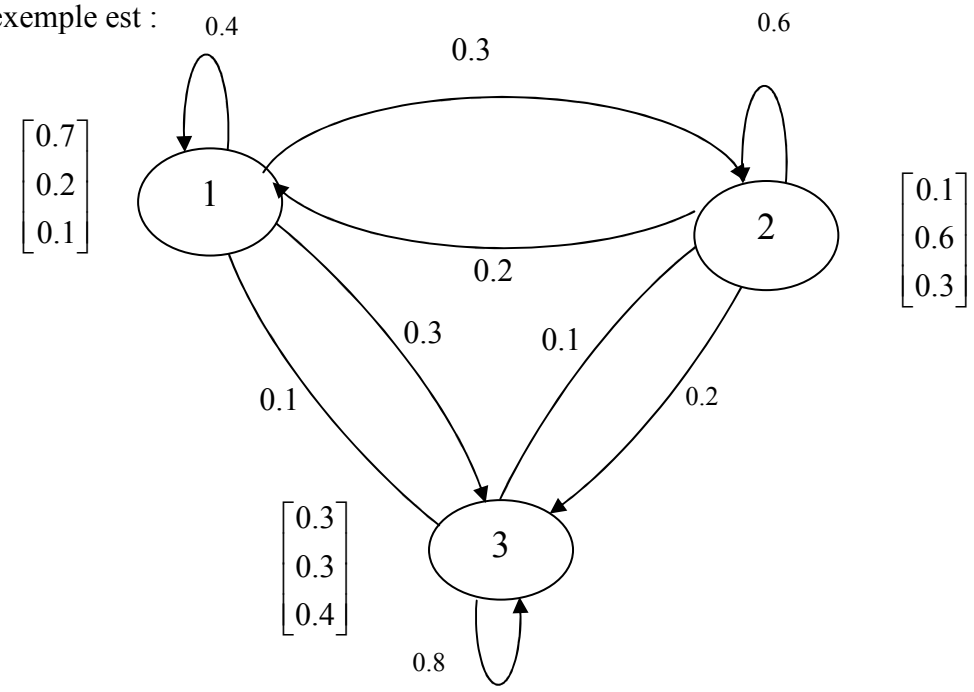


Figure III.3 : Graphe d'états de l'exemple

III.5 Les trois problèmes fondamentaux des HMMs : [17]

Soit λ un modèle de Markov caché et O une séquence d'observation, la reconnaissance de cette séquence s'effectue en trouvant le modèle λ qui maximise la probabilité $p(\lambda/O)$, cette probabilité est appelée probabilité a posteriori. Malheureusement, il n'est pas possible d'accéder directement à cette probabilité , mais on peut calculer la probabilité qu'un modèle donné génèrera une séquence d'observation $p(O/\lambda)$.

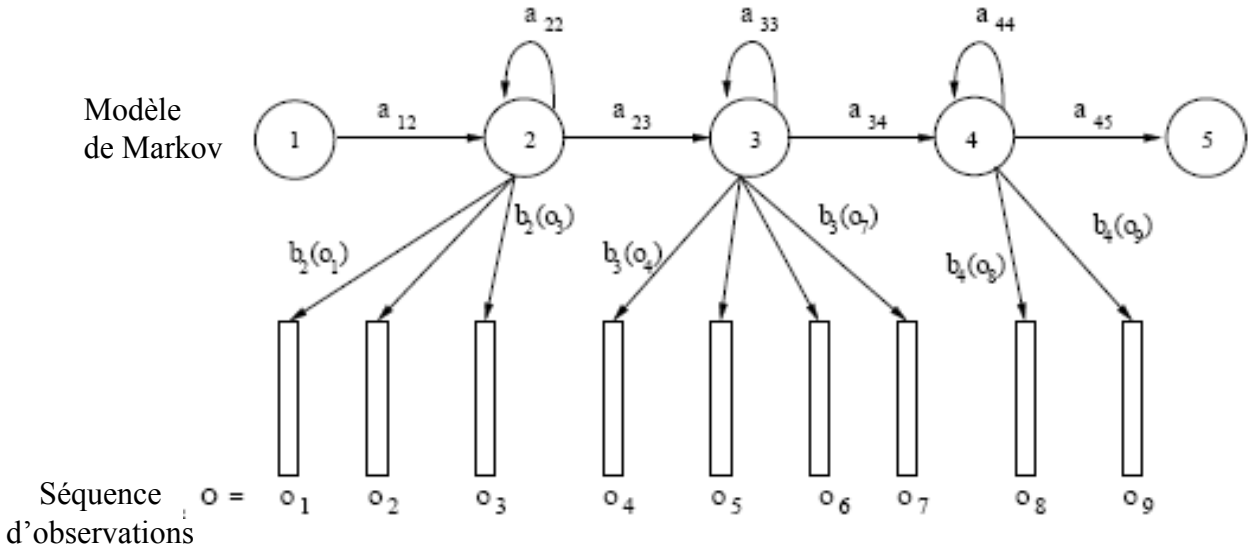


Figure III.4 : Exemple d'un HMM caractérisé par une distribution de probabilité pour chaque état associé à une observation et par des probabilités de transition entre les états

En utilisant la loi de Bayes, il est possible de lier $p(\lambda/O)$ à $p(O/\lambda)$ par :

$$p(\lambda/O)=\frac{p(O/\lambda)*p(\lambda)}{p(o)} \qquad \text{(III.15)}$$

- $p(O/\lambda)$ est la vraisemblance de la séquence d'observation O étant donné le modèle λ.
- $p(\lambda)$ est la probabilité à priori du modèle.
- $p(O)$ est la probabilité à priori de la séquence d'observation.

Etant donné $p(O)$ est une constante**,** donc maximisé $p(\lambda/O)$ revient à maximiser $p(O/\lambda)*p(\lambda)$.

Pour cela, il faut résoudre les trois problèmes fondamentaux des HMMs suivants :

III.5.1 Problème d'évaluation :

Etant donnés une séquence d'observation $O = \{o_1, o_2, ..., o_T\}$ et $\lambda = \{A, B, \pi\}$

Comment calculer efficacement $p(O/\lambda)$ la probabilité d'observer la séquence O sachant le modèle λ ?

III-5-2.Probleme de décodage :

Étant donné une séquence d'observation : $O = \{o_1, o_2, ..., o_T\}$ et $\lambda = \{A, B, \pi\}$.

Comment choisir la séquence d'états (cachée) optimale qui explique au mieux ces observations ?

III.5.3 Problème d'apprentissage:

Sachant un corpus d'entraînement *O*, comment ajuster les paramètres λ du modèle pour maximiser $p(O/\lambda)$?

Le premier problème est le problème d'évaluation, étant donné un modèle et une séquence d'observations comment faire pour calculer la probabilité d'observer une séquence étant donné le modèle ? On peut voir le problème comme un problème de score, pour quel point un modèle donné pourrait être lié à une séquence d'observation donnée ? La solution pour ce problème est de choisir parmi plusieurs modèles le modèle qui explique mieux la séquence d'observation.

Le deuxième problème est celui dans lequel on essaye de dévoiler la partie cachée du modèle, il s'agit de trouver la séquence d'état correcte, il est claire que pour tous les cas des modèles dégénérés, il n'y a pas une séquence d'état correcte à trouver. C'est pour ça pour les situations pratiques, on utilise généralement un critère optimal pour résoudre ce problème le mieux possible.

Le troisième problème est celui où on essaye d'optimiser les paramètres du modèle. La séquence d'observation utilisée pour ajuster ces paramètres est appelée la séquence d'apprentissage parce que elle est utilisée pour entraîner le HMM. Le problème d'apprentissage est le plus difficile pour la plus parts des applications des HMMs parce que on doit adapter d'une façon optimale les paramètres du modèle pour des donnés d'apprentissage, c'est à dire crée des modèles optimaux pour des phénomènes réels.

III.6 Solution aux trois problèmes : [17]

III.6.1 Solution au premier problème :

On veut calculer la probabilité d'observer la séquence $O = \{o_1, o_2, \ldots, o_T\}$ étant donné le modèle λ, $P(O/\lambda)$. La meilleure solution pour calculer cette probabilité est d'énumérer toutes les séquences d'états possibles de langueurs T. il y a N^T séquences d'états possibles.

On considère une séquence d'état fixe

$$q = (q_1 q_2 \ldots q_T) \tag{III.16}$$

Où q_1 est l'état initial. La probabilité de la séquence d'observation O étant donné la séquence d'état est donnée par l'équation (III-17)

$$p(O/q,\lambda)=\prod_{t=1}^{T} p(o_t/q_t,\lambda) \qquad \text{(III.17)}$$

Si on suppose l'indépendance statistique des observations, on aura

$$p(O/q,\lambda)=b_{q1}(o_1).b_{q2}(o_2)...b_{qT}(o_T). \qquad \text{(III.18)}$$

La probabilité pour chaque séquence d'état q peut être écrite comme suit :

$$p(q/\lambda)=\pi_{q1}a_{q1q2}a_{q2q3}...a_{qT-1qT}. \qquad \text{(III.19)}$$

La probabilité jointe de O et q est le produit des deux termes des équations (III-18) et (III-19)

$$p(O,q/\lambda)=p(O/q,\lambda)p(q/\lambda). \qquad \text{(III.20)}$$

La probabilité de l'observation O étant donné le modèle est obtenue en sommant la probabilité jointe sur toutes les séquences d'états possibles q, ce qui donne:

$$p(O/\lambda)=\sum_{tous\ les\ q} p(O/q,\lambda)p(q/\lambda) \quad = \sum_{q1,q2,...qT} \pi_{q1}b_{q1}(o_1)a_{q1q2}b_{q2}(o_2)...a_{qT-1qT}b_{qT}(o_T). \qquad \text{(III.21)}$$

L'interprétation de l'équation ci dessus est : initialement au temps t = 1 on est à l'état q_1 avec la probabilité π_{q_1}, et générant le symbole o_1 avec la probabilité $b_{q_1}(o_1)$, le temps change de t à t +1 et on fait une transition à l'état q_2 à partir de l'état q_1 avec la probabilité $a_{q_1 q_2}$ et on génère le symbole o_2 avec la probabilité $b_{q_2}(o_2)$, ce processus continu de cette manière jusqu'à la dernière transition (au temps T) de l'état q_{T-1} à l'état q_T avec la probabilité $a_{q_{T-1} q_T}$ et générant le symbole o_T avec la probabilité $b_{q_T}(o_T)$.

Le calcul de la probabilité *P (O/ λ)* en utilisant la définition de l'équation (III-21) nécessite $2T.N^T$ opérations. Ce qui rend ce calcul complexe et pratiquement infaisable, si on prend N = 5 états,T = 100 observations, il y a un ordre de $2.100.5^{100}$ opérations !

Heureusement une procédure appelée procédure 'Forward' plus efficace existe pour résoudre le premier problème.

III.6.1.1 la procédure Forward :

On considère la variable 'Avant' $\alpha_t(i)$ définit comme :

$$\alpha_t(i) = p(o_1 o_2 ... o_t, q_t = i / \lambda) \tag{III.22}$$

C'est la probabilité de la séquence d'observation partielle $o_1 o_2 ... o_t$ (jusqu'au temps t) étant donné le modèle λ , on peut calculer $\alpha_t(i)$ par récurrence comme suit :

1. initialisation

$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N. \tag{III.23}$$

2. Récurrence

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) \alpha_{ij} \right] b_j(o_{t+1}), \qquad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{III.24}$$

3. terminaison

$$p(O / \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{III.25}$$

L' Etape 1 initialise la probabilité 'Avant' comme une probabilité de l'état i et de l'observation initiale o_1, l'étape de récurrence qui est le cœur de la procédure est illustrée sur la figure III.5. Cette figure montre combien l'état j peut être atteint au temps t+1 à partir de N possibles états au temps t, le produit $\alpha_t(i) a_{ij}$ est la probabilité que la séquence $o_1 o_2 ... o_t$ est observée et l'état j est atteint au temps t+1 à partir de l'état i au temps t. En sommant ce produit sur tous les états possibles N au temps t , on obtient la probabilité de l'état j au temps t+1 avec toutes les observations partielles précédentes.

Une fois que ce calcul est fait, et l'état j est connu, il est clair de constater que $\alpha_{t+1}(j)$ est obtenu en calculant l'observation o_{t+1} à l'état j, c'est à dire en multipliant la quantité de la somme par la probabilité $b_j(o_{t+1})$. Le calcul de l'équation (III-24) est répété pour tous les états j, $1 \le j \le N$ à un temps t. Finalement l'étape 3 donne le calcul désiré de $p(O / \lambda)$ comme la somme de la variable Avant finale $\alpha_T(i)$ sur tous les états i.

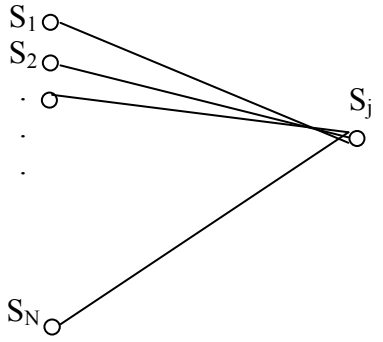


Figure III.5: Illustration de la séquence d'opérations nécessaires pour le calcul de la variable avant $\alpha_{t+1}(j)$

Remarque : les opérations nécessaires pour le calcul de $\alpha_t(j), 1 \le t \le T, 1 \le j \le N$, est de l'ordre de $N^2.T$ opérations, alors que pour la méthode directe est de $2T.N^T$. Pour N = 5 et T = 100, on a besoin de 3000 opérations alors que pour la méthode directe 10^{72} opérations sont nécessaires.

III.6.1.2 la procédure Backward:

De la même manière, on considère une variable 'Arrière' $\beta_t(i)$ définie comme :

$$\beta_t(i) = p(o_{t+1} o_{t+2} \ldots o_T / q_t = i, \lambda) \tag{III.26}$$

qui est la probabilité de la séquence d'observation partielle de $t+1$ jusqu'à T étant donné l'état i au temps t et le modèle λ. On peut résoudre le problème de $\beta_t(i)$ par récurrence comme suit :

1. initialisation

$$\beta_T(i) = 1, \qquad 1 \le i \le N \tag{III.27}$$

2. Récurrence

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \ldots, 1, \qquad 1 \le i \le N. \tag{III.28}$$

Dans l'étape 1 d'initialisation, on définit arbitrairement $\beta_t(i)$ égale à 1 pour tous les états i.

L'étape 2 qui est illustrée dans la figure III.6, montre que dans le but d'être dans l'état i au temps t et de calculer la probabilité de la séquence d'observation partielle à partir du temps $t+1$, on doit prendre en considération tous les états possibles j au temps $t+1$.

On va voir dans ce qui suit comme bien les variables 'Avant' et 'Arrière' peuvent être utilisées pour résoudre le deuxième et le troisième problème.

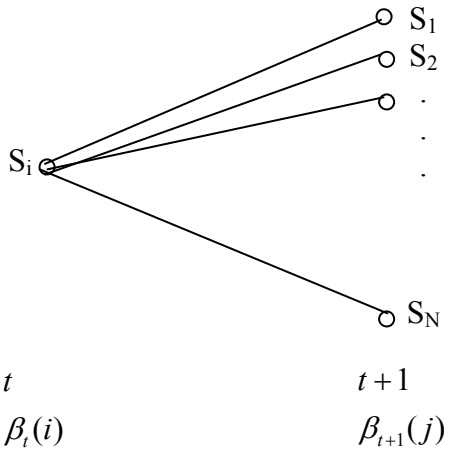


Figure III.6 : La séquence des opérations nécessaires pour le calcul de la variable arrière $\beta_t(i)$

III.6.2 Solution au deuxième problème : séquence d'état optimale (l'algorithme de Viterbi)

La solution au deuxième problème consiste à trouver la séquence d'état optimale associée avec la séquence d'observation donnée en utilisant un certain critère d'optimalité, ce dernier consiste à choisir les états q_t qui sont individuellement plus semblables à chaque instant t.

Pour implémenter cette solution au deuxième problème on doit définir une probabilité a posteriori

$$\gamma_t(i) = p(q_t = i / O, \lambda) \tag{III.29}$$

qui est la probabilité d'être dans l'état i au temps t étant donnée la séquence d'observation O et le modèle λ.

On peut exprimer $\gamma_t(i)$ comme suit :

$$\gamma_t(i) = p(q_t = i/O,\lambda)$$

$$= \frac{p(O, q_t = i/\lambda)}{p(O/\lambda)}$$

$$= \frac{p(O, q_t = i/\lambda)}{\sum_{i=1}^{N} p(O, q_t = i/\lambda)} \qquad \text{(III.30)}$$

Avec $p(O, q_t = i/\lambda)$ égale à $\alpha_t(i).\beta_t(i)$, on peut écrire $\gamma_t(i)$ comme suit :

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \qquad \text{(III.31)}$$

Ou $\alpha_t(i)$ est calculée de la séquence d'observation partiale $o_1 o_2 ... o_t$ et $\beta_t(i)$ du reste de la séquence d'observation $o_{t+1} o_{t+2} ... o_T$.

En utilisant $\gamma_t(i)$, on peut trouver les états les plus semblants q_t^* à chaque instant t comme :

$$q_t^* = \arg\max_{1 \le i \le N} [\gamma_t(i)], \qquad 1 \le t \le T. \qquad \text{(III.32)}$$

Bien que l'équation (III-32) maximise le nombre d'espérance des états correctes (en choisissant l'état le plus semblant à chaque instant t), elle peut poser quelques problèmes avec le résultat de la séquence d'état.

Par exemple, quand un HMM a une transition d'état qui a une probabilité égale à zéro ($a_{ij} = 0$), la séquence d'états optimale peut ne pas être la vraie séquence car la solution à l'équation (III.32) détermine tout simplement l'état le plus semblant à chaque instant, sans prendre en compte la probabilité d'occurrence des séquences d'états.

Une solution possible pour ce problème est de modifier le critère d'optimisation. Il s'agit de trouver une seule et meilleure séquence d'état pour maximiser $p(q/O,\lambda)$ qui est équivalente à la maximisation de $p(q,O/\lambda)$.Une technique formelle existe pour trouver la meilleure séquence d'état, elle est basée sur une méthode de programmation dynamique qui est appelée l'algorithme de Viterbi .

Pour trouver la meilleure séquence d'état, $q=(q_1 q_2 ... q_T)$ pour une séquence d'observation donnée $O=(o_1 o_2 ... o_T)$ on doit définir une quantité

$$\delta_t(i)=\max_{q_1, q_2, ... q_{t-1}} p\left[q_1 q_2 ... q_{t-1}, q_t=i, o_1 o_2 o_t / \lambda\right] \tag{III.33}$$

qui est le meilleur score le long d'un chemin des états au temps t .

Par récurrence on a :

$$\delta_{t+1}(j)=\left[\max_i \delta_t(i) a_{ij}\right] . b_j(o_{t+1}). \tag{III.34}$$

Pour retrouver la séquence d'états, on doit garder la trace de l'argument qui a maximisé l'équation (III-34) pour chaque instant t et chaque état j , on fait ça par un rang $\psi_t(j)$.

La procédure complète pour trouver la meilleure séquence d'états est la suivante :

1. initialisation

$$\begin{aligned} &\delta_1(i)=\pi_i b_i(o_1), \qquad 1\le i\le N \\ &\psi_1(i)=0. \end{aligned} \tag{III.35}$$

2. Récurrence

$$\begin{aligned} &\delta_t(j)=\max_{1\le i\le N}\left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) , && \begin{matrix} 2\le t\le T \\ 1\le j\le N \end{matrix} \\ &\psi_t(j)=\arg\max_{1\le i\le N}\left[\delta_{t-1}(i) a_{ij}\right], && \begin{matrix} 2\le t\le T \\ 1\le j\le N. \end{matrix} \end{aligned} \tag{III.36}$$

3. Terminaison

$$\begin{aligned} &p^*=\max_{1\le i\le N}\left[\delta_T(i)\right] \\ &q_T^*=\arg\max_{1\le i\le N}\left[\delta_T(i)\right]. \end{aligned} \tag{III.37}$$

4. Recherche

$$q_t^*=\psi_{t+1}\left(q_{t+1}^*\right) \qquad t=T-1, T-2, ..., 1 \tag{III.38}$$

Remarque : il est à noter que l'algorithme de Viterbi est similaire dans l'implémentation au calcul de la variable 'Avant' des équations (III.23) - (III-25), la seule différence est la maximisation dans l'équation (III.36).

III.6.3Solution au troisième problème : paramètres d'estimation

Le troisième problème est le problème le plus difficile, il s'agit de déterminer une méthode pour l'ajustement des paramètres du modèle λ (A, B, π) pour satisfaire un certain critère d'optimisation.
Les paramètres du modèle λ doivent être choisis de telle façon que vraisemblance $p(O/\lambda)$ soit localement maximale et ça en utilisant une méthode itérative comme la méthode de Baum – Welsh (connue aussi sous le nom de EM (expectation – maximisation) .

Pour décrire la procédure de ré estimation des paramètres du modèle λ, on doit définir une quantité $\xi_t(i,j)$ qui est la probabilité d'être dans l'état i au temps t et de rejoindre l'état *j* au temps $t+1$ étant donné le modèle et la séquence d'observation :

$$\xi_t(i,j) = p(q_t = i, q_{t+1} = j / O, \lambda). \tag{III.39}$$

A partir des définitions des variables Avant et Arrière, on peut écrire $\xi_t(i,j)$ sous la forme :

$$\begin{aligned}\xi_t(i,j) &= \frac{p(q_t = i, q_{t+1} = j, O/\lambda)}{p(O/\lambda)} \\ &= \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j)}{p(O/\lambda)} \\ &= \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j)}\end{aligned} \tag{III.40}$$

Nous avons déjà défini $\gamma_t(i)$ comme la probabilité d'être dans l'état i au temps t, étant donnée la séquence d'observation et le modèle λ, on peut relier $\gamma_t(i)$ à $\xi_t(i,j)$:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j). \tag{III.41}$$

Si on somme $\gamma_t(i)$ sur le temps t , on obtient une quantité qui peut être interpréter comme l'espérance de nombre de transitions effectuées à partir de l'état i , de la même chose la sommation sur le temps t de $\xi_t(i,j)$ peut être interpréter comme le nombre d'espérance de transition de l'état i à l'état j .

$$\sum_{t=1}^{T-1}\gamma_t(i) = \text{L'espérance de nombre de transition à partir de l'état } i \qquad \text{(III.42)}$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) = \text{ L'espérance de nombre de transition de l'état } i \text{ à l'état } j \qquad \text{(III.43)}$$

En utilisant les deux formules des équations (III.42) et (III.43), on peut donner une méthode de ré estimation des paramètres d'un HMM (l'étape E de la procédure de ré estimation). Un ensemble des formules de ré estimation pour π , A, et B est :

$$\tilde{\pi}_j = \gamma_1(i) \qquad \text{(III.44)}$$

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \qquad \text{(III.45)}$$

$$\tilde{b}_j(k) = \frac{\sum_{t=1,\, o_t = v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_{t(j)}} \qquad \text{(III.46)}$$

Si on définit un modèle $\lambda = (\text{A, B}, \pi)$ qu'on va utiliser pour calculer la partie droite des équations (III.44), (III.45), (III.46), et on définit un modèle ré estimé $\tilde{\lambda}(\tilde{A}, \tilde{B}, \tilde{\pi})$ qui est déterminé à partir de la partie gauche des équations (III.44), (III.45), (III.46), Baum et ses collègues ont prouvé que le choix entre le modèle initial λ et le modèle ré estimé $\tilde{\lambda}$ est en fonction de la vraisemblance dans le sens ou $p(O/\tilde{\lambda}) \rangle p(O/\lambda)$, dans ce cas, un nouveau modèle $\tilde{\lambda}$ est choisi pour lequel la séquence d'observation est plus probable à se produire (l'étape M de la procédure de ré estimation).

Basé sur ce choix, et si on utilise itérativement $\tilde{\lambda}$ à la place de λ et on répète la procédure de ré estimation, on va aboutir à un modèle final ayant des paramètres qui améliorent la probabilité d'observation O .

III.7 La densité d'observation continue dans un HMM : [17] [18]

Pendant toute la discussion qu'on a abordé avant, on a considéré les observations comme des symboles discrets, alors que pour les applications de la parole l'observation est donnée comme un signal continu, il est possible de convertir un signal continu en une série de symboles en utilisant la quantification vectorielle . Mais cette solution peut introduire de sérieuses dégradations, alors il est avantageux d'utiliser des HMMs avec des densités d'observations continues pour modeler le signal de la parole.

La meilleure représentation de la fonction de densité de probabilité qui peut modeler un signal continu est un mélange fini de la forme :

$$b_j(O)=\sum_{k=1}^{M} c_{jk} N(O,\mu_{jk},U_{jk}), \qquad 1 \le j \le N \tag{III.47}$$

Où O est le vecteur d'observation qui va être modelé, c_{jk} sont les coefficients de mélange dans l'état j, on assume que N est une densité Gaussienne avec une moyenne μ_{jk} et une covariance U_{jk} pour la $k^{éme}$ composante de mélange dans l'état j.

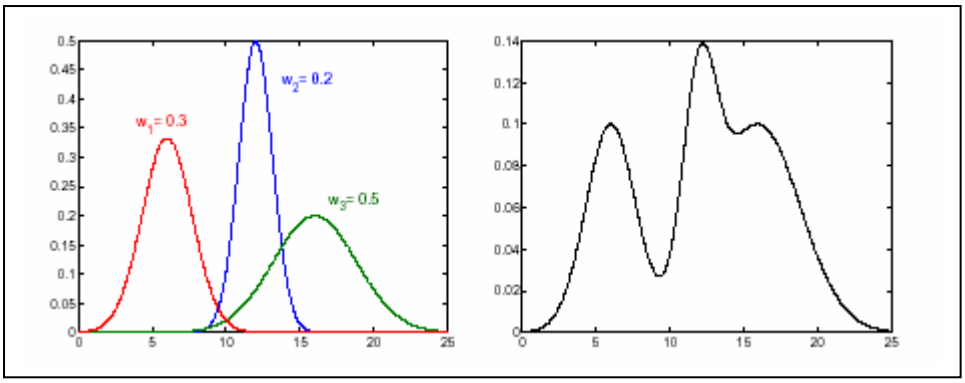


Figure III.7 : La fonction de densité de probabilité de multi gaussienne à une seule dimension

Le gain de mélange doit satisfaire la contrainte

$$\sum_{k=1}^{M} c_{jk} = 1, \qquad 1 \le j \le N \tag{III.48}$$

$$c_{jk} \ge 0, \quad 1 \le j \le N, \qquad 1 \le k \le M \tag{III.49}$$

afin que la fonction de densité de probabilité est normalisée, c'est à dire :

$$\int_{-\infty}^{+\infty} b_j(o)\,do = 1, \qquad 1 \le j \le N \tag{III.50}$$

Il a été montré que chaque état d'un modèle HMM avec une densité de mélanges de gaussiennes est équivalent à un modèle de sous états d'une seule densité mélange [18].

Considérant un état i avec une densité de M mélanges de Gaussiennes. Parce que la somme des coefficients de gain de mélanges égale à 1, ils définissent un ensemble de coefficients de transitions aux sous états i_1 (avec une probabilité de transition c_{i1}), i_2 (avec une probabilité de transition c_{i2}) jusqu'à i_M (avec une probabilité de transition c_{iM}). Dans chaque sous état i_k, il y a un seul mélange avec une moyenne μ_{ik} et une variance U_{ik}, (Figure III.8). Chaque sous état fait une transition vers l'état temporaire i_0 avec une probabilité égale à 1. La distribution composées des sous états (chacun avec une seule densité) est mathématiquement équivalent à une densité de mélanges dans chaque état.

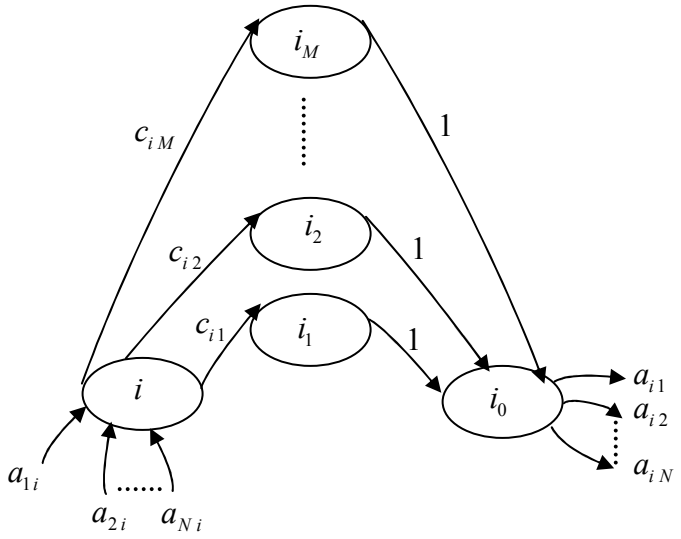


Figure III.8: l'équivalence d'un état avec une densité de mélanges à des sous états avec une seule densité [18]

Les formules de ré estimation pour les coefficients de densité de mélanges sont :

$$\widetilde{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,\mathrm{k})}{\sum_{t=1}^{T}\sum_{k=1}^{M} \gamma_t(j,k)} \tag{III.51}$$

$$\widetilde{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) * o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{III.52}$$

$$\widetilde{U}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) * (o_t - \mu_{jk}) * (o_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{III.53}$$

III.8 Résumé des quatre algorithmes :

Algorithme 1: algorithme Forward

Initialisation : $\alpha_1(i) = \pi_i \; b_i(o_1) \qquad 1 \le i \le N$ (III.54)

Récurrence : $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad t \in \{1,2,.....,T-1\} \; et \; 1 \le j \le N$ (III.55)

Terminaison : $p(o/\lambda) = \sum_{i=1}^{N} \alpha_T(i)$ (III.56)

Algorithme 2: algorithme Backward

Initialisation : $\beta_t(i) = 1 \qquad 1 \le i \le N$ (III.57)

Récurrence : $\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j), \quad t \in \{T-1, T-2, \ldots\ldots, 1\} et\ 1 \le i \le N$ (III.58)

Terminaison : $p(o/\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ (III.59)

Algorithme 3: algorithme de Viterbi

Initialisation :

$\delta_1(i) = \pi_i b_i(o_1) \quad 1 \le i \le N$ (III.60)

$\psi_1(i) = 0$ (III.61)

Récurrence :

$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{T-1}(i) a_{ij}\right] b_j(o_t) \quad 2 \le t \le T \quad 1 \le J \le N$ (III.62)

$\psi_T(J) = \arg_{1 \le i \le N} \max\left[\delta_{t-1}(i) a_{ij}\right]$ (III.63)

Terminaison :

$P^* = \max_{1 \le i \le N}\left[\delta_T(i)\right]$ (III.64)

$\psi_T^* = \arg_{1 \le i \le N} \max\left[\delta_T(i)\right]$ (III.65)

Recherche : $q_t^* = \psi_{t+1}\left(q_{t+1}^*\right) \qquad t = T-1, T-2, \ldots, 1$ (III.66)

Algorithme 4: algorithme de Baum - Welch

Estimation :

$$\tilde{\pi}_i = \gamma_1(i), \qquad 1 \le i \le N \tag{III.67}$$

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad 1 \le i \le N, \quad 1 \le j \le N \tag{III.68}$$

$$\tilde{b}_j(k) = \frac{\sum_{t=1,\, o_t = k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_{T(J)}} \qquad 1 \le J \le N \tag{III.69}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{III.70}$$

et

Avec :

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{p(o/\lambda)} \tag{III.71}$$

III.9 Types des HMMs : [19]

Un seul chemin pour classer les types des HMMs celui de la structure de la matrice de transition de la chaîne de Markov. Il existe trois types de modèles HMMs :

- HMM gauche droite.
- HMM ergodique.
- HMM gauche droite parallèle.

La figure III.9 illustre les trois types des HMMs

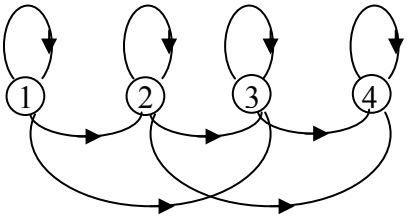


Figure III.9.a : HMM gauche droite

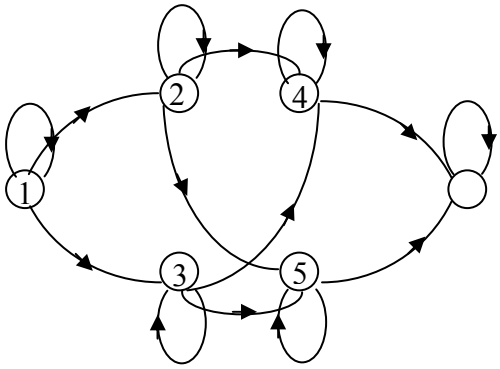


Figure III.9.b : HMM gauche droite parallèle

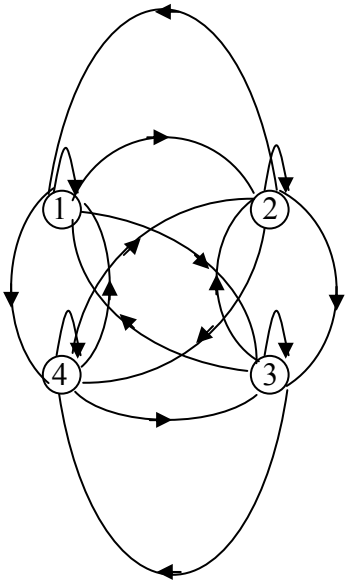


Figure III.9.c: HMM ergodique

Le modèle illustré sur la figure III.9.a est appelé le modèle type gauche –droite ou le modèle de Bakis parce que la séquence d'états associée à ce modèle a la propriété qu'à chaque fois le temps augmente, l'indice de l'état augmente, c'est-à-dire le système d'états procède de gauche à droite.

Il est clair que le type de HMM gauche –droite possède des propriétés qui peuvent facilement modeler des signaux qui ont des propriétés qui changent de la même manière (par exemple la parole).

La propriété fondamentale d'un HMM type gauche-droite est que les coefficients de la matrice de transition ont la propriété suivante :

$$a_{ij} = \text{o} \quad \text{pour} \quad \text{i} < \text{j}$$

C'est-à-dire : pas de transition permise d'un état à un autre état qui a un indice inférieur à l'état de départ. En plus la probabilité de l'état initiale a la propriété suivante :

$$\pi_i = \begin{Bmatrix} 0 & i \neq 1 \\ 1 & i = 1 \end{Bmatrix}$$

Parce que la séquence d'état doit commencer par l'état 1.

III.10 Les paramètres d'estimation initiale d'un HMM :

Dans la théorie, les équations de ré estimation doivent donner des valeurs des paramètres d'un HMM qui correspond à la fonction de probabilité locale qui doit être maximale. Une question clé est : Comment choisir les paramètres initiaux d'estimation d'un HMM de telle sorte que la fonction de probabilité locale serait maximale ?

Principalement, il n'existe pas une solution directe. D'un autre coté l'expérience a montré que les estimations initiales des paramètres A et π seraient anarchiques ou uniformes sont capables de donner des ré estimations utiles à ces paramètres dans tous les cas [17].

Pour les paramètres de B, l'expérience a montré que de bonnes estimations initiales aident dans le cas des symboles discrets et sont essentielles dans le cas de distributions continues [17].

Les estimations des paramètres initiales peuvent être obtenues de différentes manières telles que :

- La segmentation des séquences d'observations en états, et le calcul de la moyenne des observations dans chaque état.
- La segmentation avec le maximum de vraisemblance.
- La segmentation K-means avec regroupement.

III.11 Le choix du modèle :

Le choix du type de modèle est important pour l'implémentation des HMMs (modèle type gauche droite ou ergodique), la taille du modèle (nombre des états) et le choix des symboles d'observations (discret ou continu, singulier ou mélange, le choix des paramètres). Malheureusement il n'y a aucune méthode simple et théoriquement correcte de faire un tel choix. Ces choix dépendent du signal qui va être modulé.

III.12 Mesure de confiance:

III.12.1 : Introduction :

L'implémentation des systèmes de reconnaissance de la parole pratique a créé le besoin de développer des techniques pour détecter les erreurs de reconnaissance.

Les systèmes de reconnaissance des mots isolés actuels sont capables d'atteindre un haut niveau de reconnaissance dans des conditions de laboratoire. Dans la pratique, la limite de quantité des données d'apprentissage et la mauvaise liaison entre l'environnement de test et celui d'apprentissage réduit le taux de reconnaissance.

Dans cette partie de chapitre nous allons sélectionner une approche pour réduire les erreurs de reconnaissance et améliorer les performances d'un système de reconnaissance de la parole. On ne va pas directement essayer d'améliorer la précision de la reconnaissance mais on va appliquer une technique qui effectivement détecte les reconnaissances incorrectes durant la phase de test.

III.12.2 : Types d'erreurs : [1]

En général trois types d'erreurs de reconnaissance peuvent être distingués : erreurs de suppression, erreurs de substitution et erreurs d'insertion.

Dans le cas de l'erreur de suppression, l'utilisateur prononce le mot mais le système ne va pas le reconnaître. L'erreur de substitution est la reconnaissance d'un mot à la place du mot prononcé. L'erreur d'insertion correspond au cas ou un mot additionnel est reconnu.

La détection des erreurs qu'on va décrire dans cette partie de travail est basée sur la mesure de confiance.

III.12.3 Définition d'une mesure de confiance : [20]

Une mesure de confiance est définie comme une probabilité postérieure d'un mot correct étant données les valeurs des indications de confiance, elle peut être calculé à partir de la probabilité à posteriori de système de reconnaissance, les mesures de confiance sont généralement basées sur la comparaison des vraisemblances du modèle acoustique en se basant sur un autre modèle alternatif.
La mesure de confiance représente la fiabilité du résultat de reconnaissance obtenu. Pour chaque résultat de reconnaissance, une valeur de confiance qui décrit la justesse égalée entre le mot prononcé et le mot reconnu est calculée en fonction de la valeur de confiance, le résultat de reconnaissance est soit accepté ou rejeté.

Si la valeur de la mesure de confiance de reconnaissance du mot prononcé baisse au dessous d'une valeur de seuil prédéterminée, le résultat de reconnaissance est rejeté et l'utilisateur doit répéter la dernière action de la parole. Une petite valeur de confiance de reconnaissance peut être le résultat de deux différentes raisons: La première est que le mot prononcé n'appartient pas au vocabulaire actif (hors vocabulaire), et la deuxième est la liaison entre le mot prononcé et le mot reconnu est pauvre.

la sélection d'une valeur de seuil appropriée est critique pour le succès du plan du rejet. Si le seuil de rejet tend trop vers le haut, alors grand nombre de reconnaissances incorrectes vont être acceptées. Pour une valeur de seuil basse, avec additions aux erreurs de reconnaissances, des reconnaissances valides sont aussi rejetées. Comme l'exigence de la précision de reconnaissance change dépendamment de l'application, la valeur de seuil optimale doit être ajusté individuellement pour chaque système pour que un grand nombre d'erreurs de reconnaissance soit détecté et rejeté

sans supprimer les reconnaissances correctes, d'où la suggestion d'une technique de détection qui va nous aider à développer notre système de reconnaissance de la parole.
La mesure de confiance basée sur le rejet peut être vue comme un post processeur du système de reconnaissance qui contrôle la validité du résultat de reconnaissance, dans notre système on va se baser sur la détection et le rejet du hors vocabulaire.

III.12.4 L'estimation de confiance : [2]

En général, un système de reconnaissance de la parole fournit un score pour caractériser combien un HMM est lié pour une prononciation inconnue.

Le score est néanmoins relatif et il dépend beaucoup de l'utilisateur et de l'environnement, donc il ne peut pas être utilisé pour mesurer la bonne qualité de liaison.

Dans le but de détecter les mots hors vocabulaires, nous avons besoin d'un score de référence S_{ref} pour tester si le mot reconnu existe réellement dans le segment de la parole de référence. En comparant le score provenant par la reconnaissance HMM et le score de référence, nous pouvons estimer la bonne qualité de liaison entre la prononciation et le HMM-mot, la condition fondamentale d'un score de référence est qu'il doit être très proche du score produit par le HMM-mot.

Maintenant un facteur de confiance C peut être défini comme une distance entre le score HMM-mot et le score de référence S_{ref} :

$$C = S_{HMM} - S_{ref} \qquad \text{(III.72)}$$

Comme les deux scores sont déterminés de la même prononciation, nous n'avons pas besoin d'effectuer une amélioration de normalisation du temps, donc on peut effectuer une comparaison directe des scores.

Normalement, pour une reconnaissance correcte, il y a un intervalle signifiant entre le score de référence et le score HMM-mot, mais pour une mauvaise reconnaissance (hors vocabulaire) les deux scores sont très proches.

Par la sélection d'une valeur de seuil appropriée, nous serons apte à séparer les reconnaissances correctes des reconnaissances incorrectes, cette séparation est basée sur leurs scores de confiance.

III.12.5 Calcul du score de référence : [2]

Il y a beaucoup de méthodes pour calculer le score de référence, le problème de déterminer le score de référence est similaire à celui de détection des mots clés [21]. On doit définir un modèle où le score de vraisemblance, pour le cas d'une valide prononciation est inférieur que celui du HMM-mot, mais pour la présence d'un mot hors vocabulaire ce score doit être plus grand que celui du HMM-mot .

Dans (Rahim et al.1995) [22], deux approches sont présentées pour obtenir le score de référence, les deux approches sont basées sur l'utilisation des modèles acoustiques.

Dans notre travail nous allons utiliser la méthode de (Boulard et al, 1994) [21], qui s'adapte bien avec notre approche, où on a besoin d'un modèle appelé modèle poubelle qui est utilisé pour caractériser les mots hors vocabulaire.

L'idée de base de la modélisation poubelle est que le score individuel d'une trame de parole n'est jamais le meilleur mais il est souvent l'un des meilleurs candidats. Dans le modèle poubelle un score est calculé pour chaque trame de parole comme la moyenne de N meilleurs scores produits par HMM-mot.

Cette technique est appelée « modélisation poubelle directe » [01], et elle va être utilisée pour calculer le score de référence.

III.12.6 Le calcul du score de confiance en utilisant la modélisation poubelle directe [23]

Comme on l'a vu dans l'équation (III.72), le calcul du facteur de confiance nécessite deux différents scores, le score du modèle HMM-mot et le score de référence S_{ref} qui est le score du modèle poubelle qui sont déterminés au début-fin de la prononciation comme il est montré dan la figure ci-dessous :

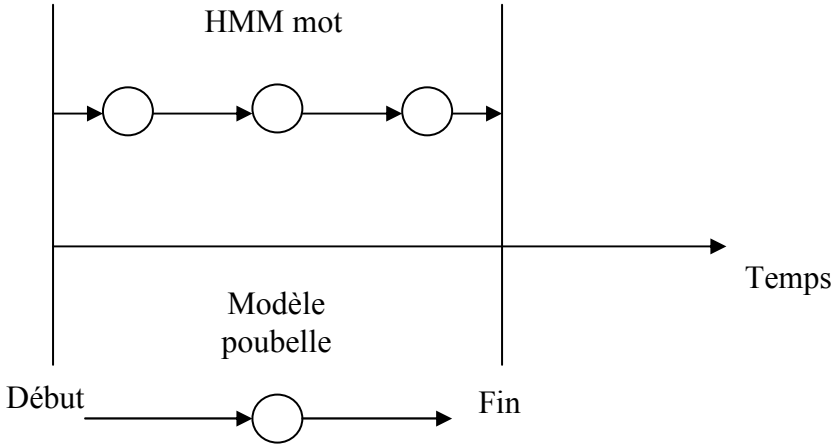


Figure III.10 : Les scores nécessaires pour le calcul de confiance

Pour chaque trame de parole, on calcule le maximum des vraisemblances S_{max} et la moyenne des vraisemblances S_{moy} , maintenant, le score du modèle poubelle local à chaque instant t avec une valeur du rang désiré r est calculé en utilisant l'interpolation linéaire entre les scores maximums et les scores moyens. La valeur du rang r est limitée de telle sorte que le score moyen correspond à la valeur du rang 0.5 et le score maximum à la valeur du rang 1.0 respectivement. En changeant la valeur du rang, on peut ajuster le score du modèle poubelle. Le score HMM-mot qui définit le début-fin de prononciation et le score du modèle poubelle local sont accumulés entre le debut et la fin de la prononciation, ainsi le score de référence nécessaire pour le calcul de confiance peut être exprimé comme suit :

$$S_{ref} = \sum_{i=debut\ temps}^{i=fin\ temps} S_{poubelle}(i) \qquad \text{(III.73)}$$

La supposition de base dans le plan de confiance est que dans les reconnaissances correctes, les états du HMM-mot reconnu produisent des scores plus élevés pour la majorités des trames de parole des autres HMM-mots, c'est pour ça que le score du HMM-mot est plus grand que celui du modèle poubelle, dans le cas des reconnaissances incorrectes, les états des HMM-mot ne produisent pas des scores plus élevés pour toutes les trames de parole, mais les scores maximums pour chaque trame de parole sont distribués sur plusieurs HMMs, dans ce cas le score du modèle poubelle est supérieur ou égale au score HMM-mot .

III.13 Conclusion

Dans ce chapitre, on a présenté l'aspect théorique du modèle de Markov caché comme solution statistique à la reconnaissance automatique de la parole, ainsi que les problèmes fondamentaux de ce modèle et leurs résolutions qui sont traduites par trois algorithmes : l'algorithme Avant et Arrière pour résoudre le problème d'évaluation, l'algorithme de Viterbi pour résoudre le problème de décodage et l'algorithme de Baum Welch pour résoudre le problème d'apprentissage. Ainsi que les mesures de confiance, où on a décrit la technique de modélisation poubelle directe.

CHAPITRE IV
MISE EN OEUVRE ET RESULTATS

VI.1 Mise en œuvre du système de reconnaissance de mots isolés basé sur les Modèles HMM_s

IV.1.1 Principe général : [17] [24]

Dans cette partie on va essayer de construire un système de reconnaissance de mots isolés. Nous avons un vocabulaire de V mots à reconnaître, et chaque mot est représenté par un modèle HMM distinct, et pour chaque mot du vocabulaire, nous avons un ensemble d'apprentissage de K prononciations du mot (prononcé par plusieurs locuteurs), ou chaque prononciation représente une séquence d'observation.

La figure IV.1 représente le schéma de bloc d'un système de reconnaissance de mots isolés. Pour chaque mot v du vocabulaire, on doit construire un modèle HMM λ_v , et on doit estimer les paramètres du modèle (A, B, π) qui optimisent la vraisemblance d'apprentissage de l'ensemble des vecteurs d'observation.

Pour chaque mot à reconnaître une étape d'extraction des paramètres est indispensable, suivi par le calcul de la probabilité d'observer une séquence o étant donné un modèle λ : $p(o/\lambda_v)$, et enfin le choix du modèle qui a la plus grande vraisemblance : $v_0 = \arg_{\max}\left[p(o/\lambda_v)\right]$

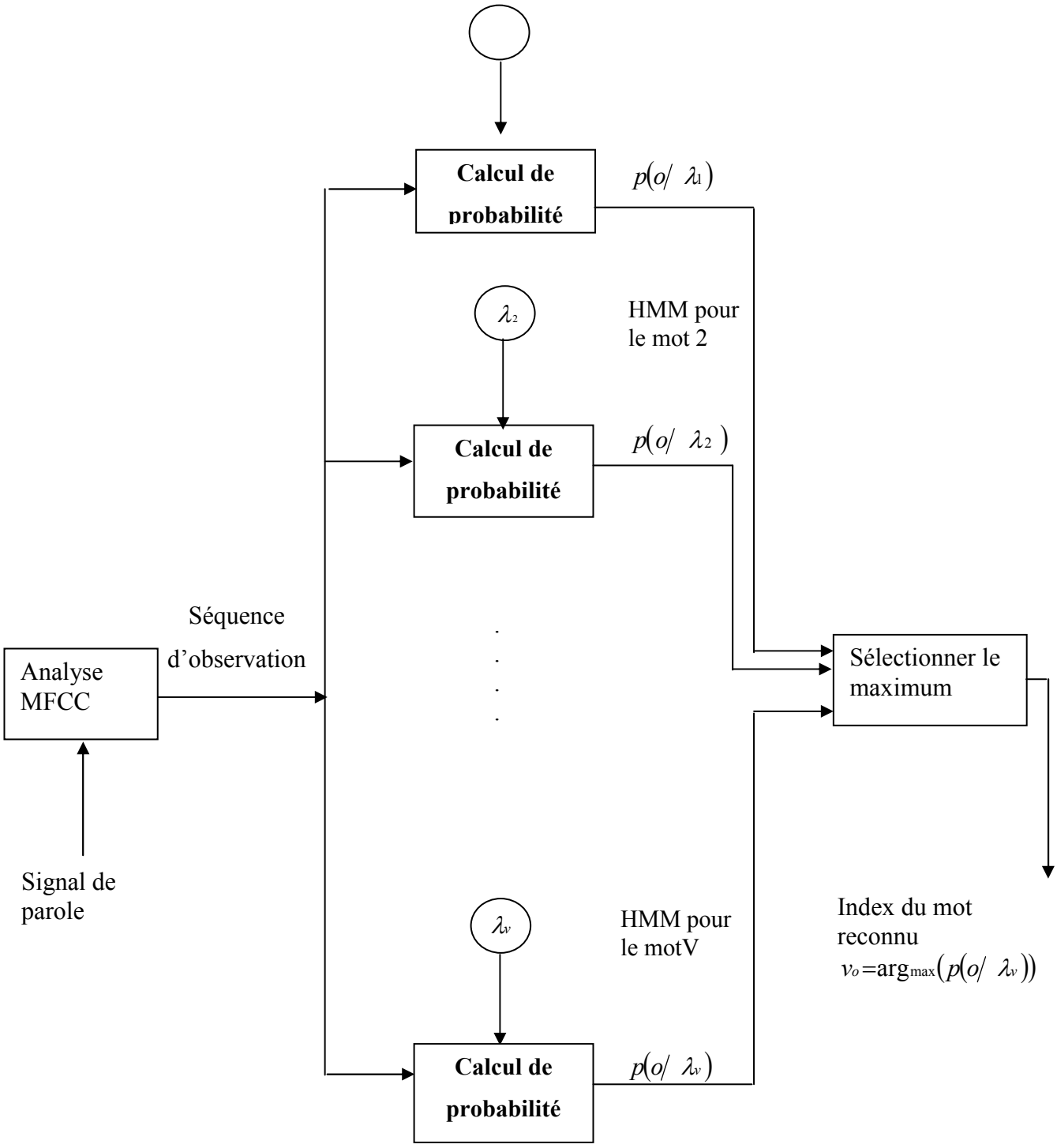


Figure IV.1 Schéma de bloc pour un système HMM de reconnaissance de mots isolés [24]

IV.1.2.Description du logiciel : [25]

Notre logiciel est composé de quatre modules principaux selon la figure IV.2

1- le module d'acquisition de données.
2- Le module d'analyse acoustique
3- le module d'apprentissage.
4- Le module de reconnaissance.

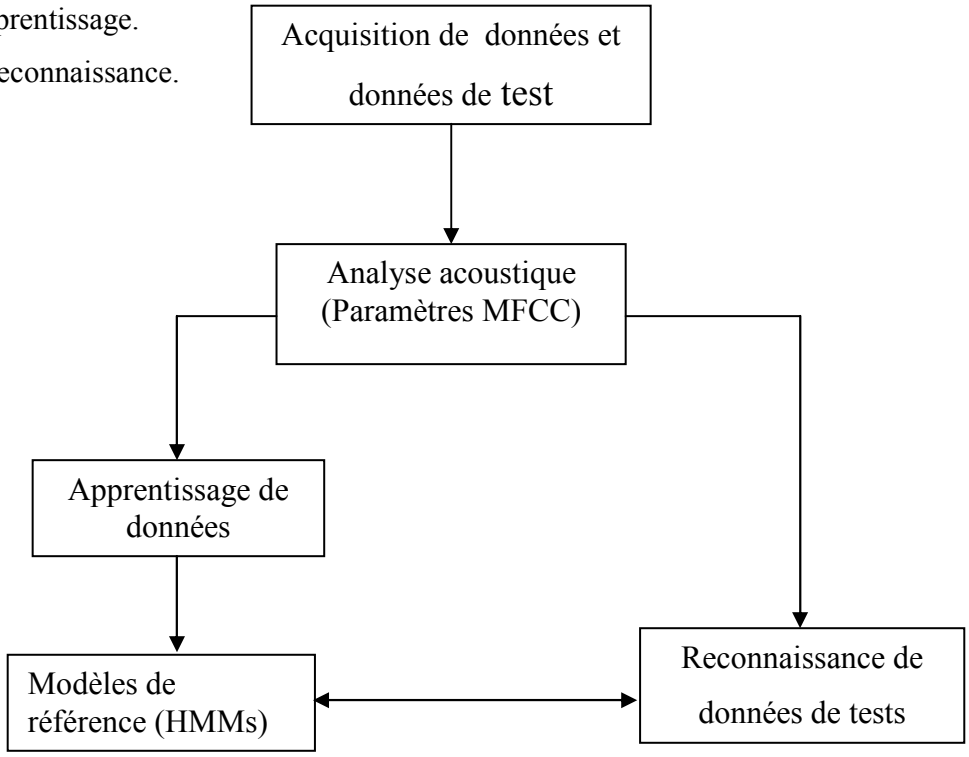
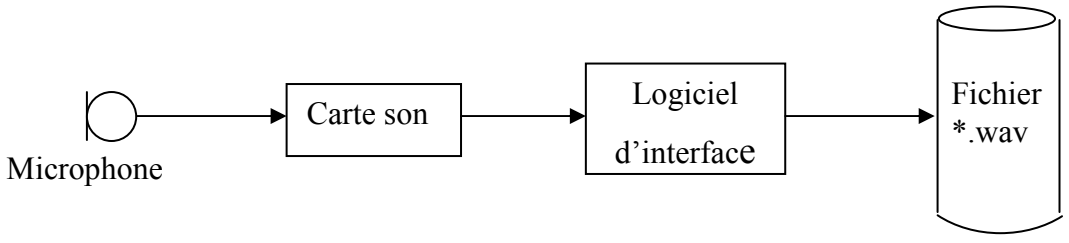


Figure IV.2 *Structure globale du logiciel*

IV.1.2.1Le module d'acquisition de données :

Le module d'acquisition de données consiste à enregistrer des sons (mots) à l'aide d'un microphone et d'une carte de son SOUND BLASTER AUDIO PCI 128 avec une fréquence d'échantillonnage de 16 KHz numérisé sur 16 bits avec un canal mono pour avoir enfin des fichiers avec extension *.wav.



FigureVI.3 Schéma descriptif de la phase d'acquisition

IV.1.2.2 Le module d'analyse acoustique :

Il consiste à extraire un ensemble de paramètres pertinents dans le signal, autrement dit c'est un module qui réduit le signal enregistré tout en conservant l'essentiel de l'information qu'il contient dans le but de réduire le temps de calcul et de stockage lors du traitement, d'apprentissage et de reconnaissance.

IV.1.2.3 Le module d'apprentissage :

Il a pour but la génération des modèles de références sous forme de modèles HMMs.

IV.1.2.4 Le module de reconnaissance :

Il consiste à reconnaître le mot prononcé (mot de test) à travers les modèles HMMs.

IV.1.3 Les organigrammes du système : [25]

IV.1.3.1 L'organigramme globale de reconnaissance :

Cet organigramme représente les différentes étapes à suivre pendant la reconnaissance.

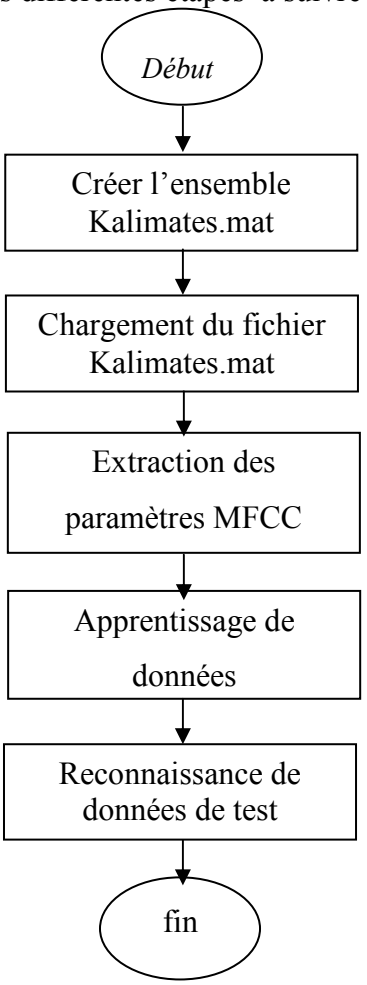


Figure IV.4 L'organigramme global de reconnaissance

IV.1.3.2 L'organigramme de la fonction Vecteurs :

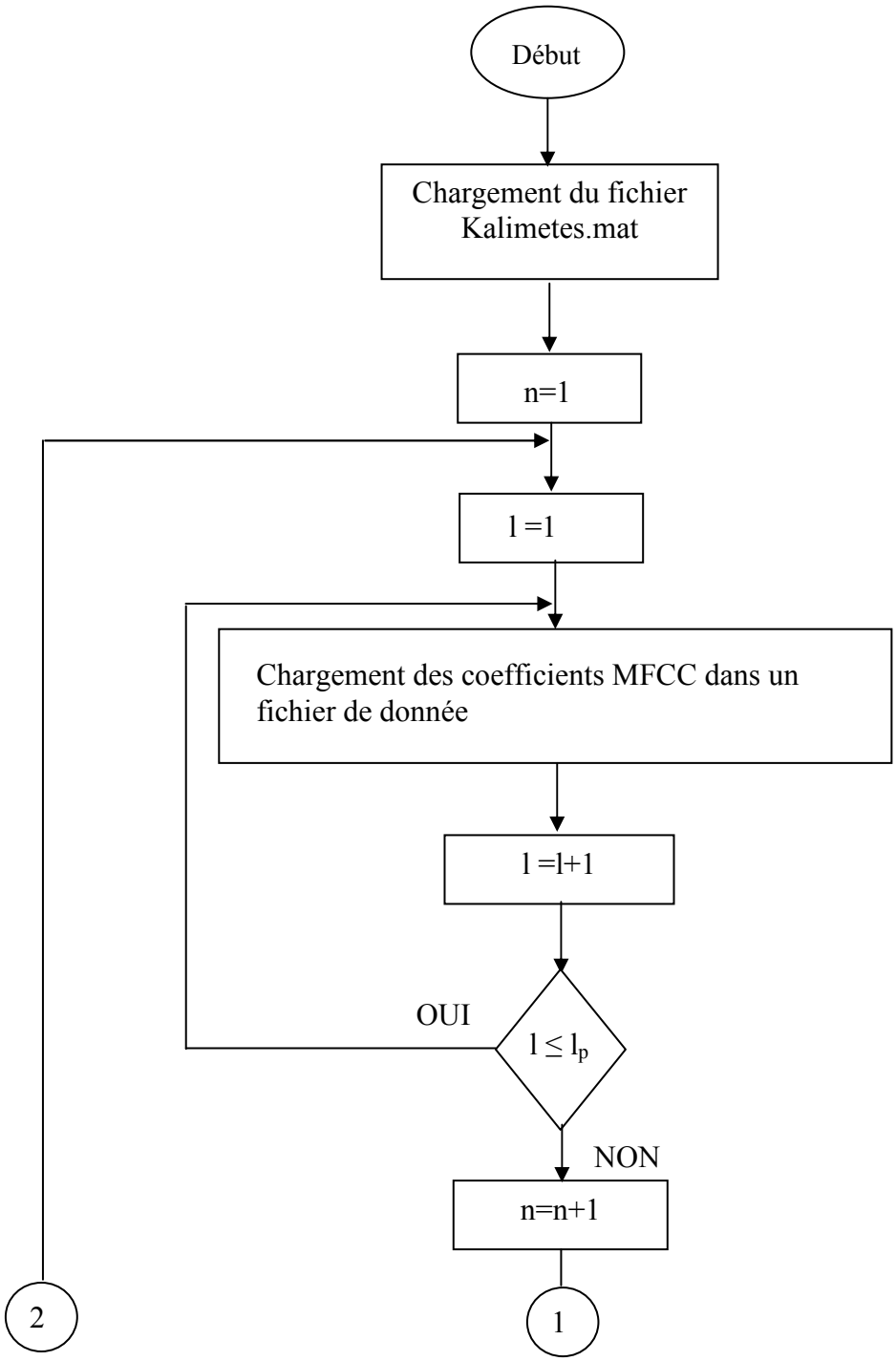
La fonction Vecteurs a comme objectif le chargement des vecteurs acoustiques sous forme de coefficients MFCC dans des fichiers de données et de test de données.

n: le nombre de mots prononcés.

l : le nombre de locuteurs pour chaque mot.

l_p: le nombre maximum de locuteurs pour les données d'apprentissage

l_t : le nombre maximum de locuteurs pour les données de test.



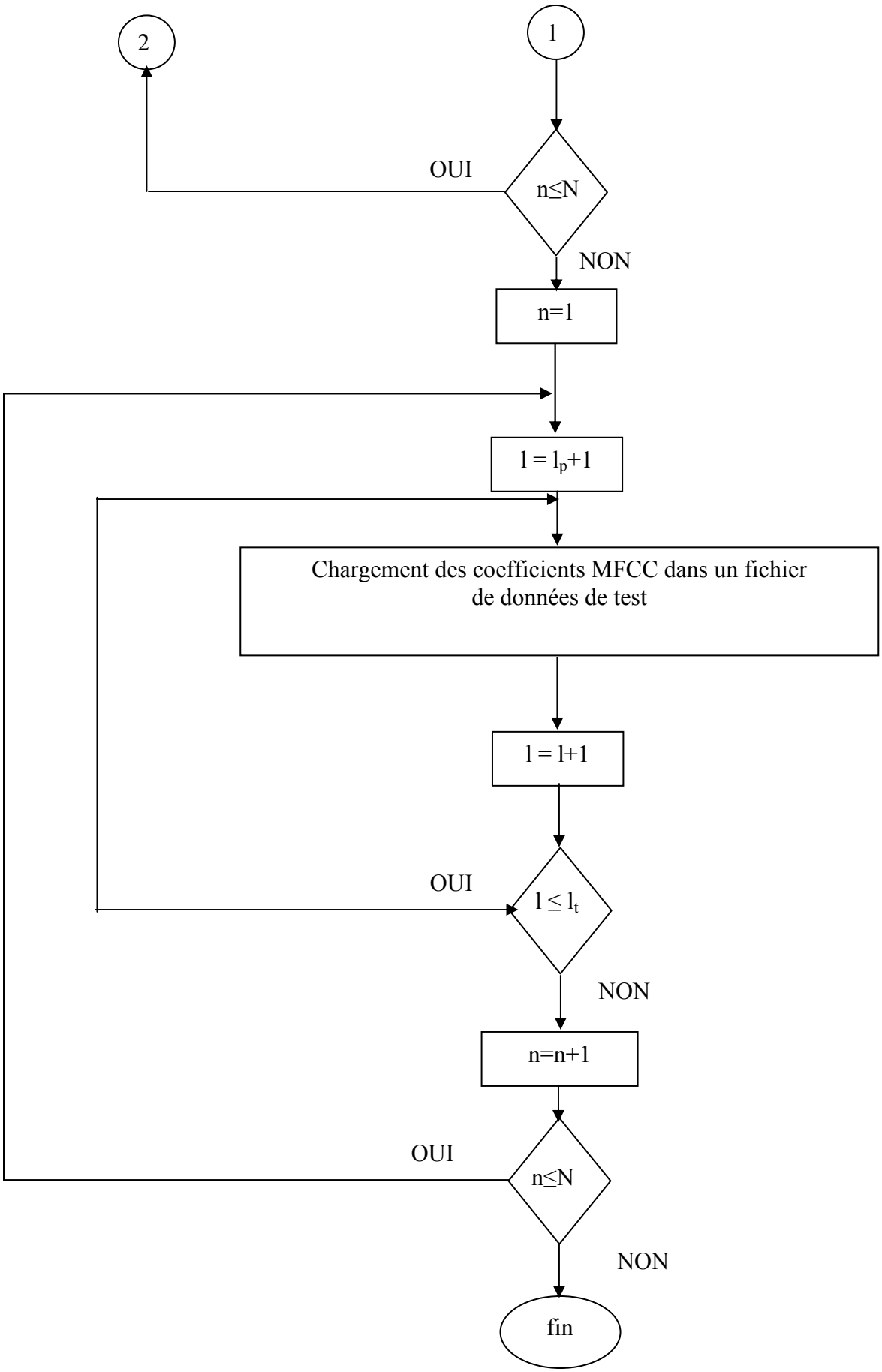
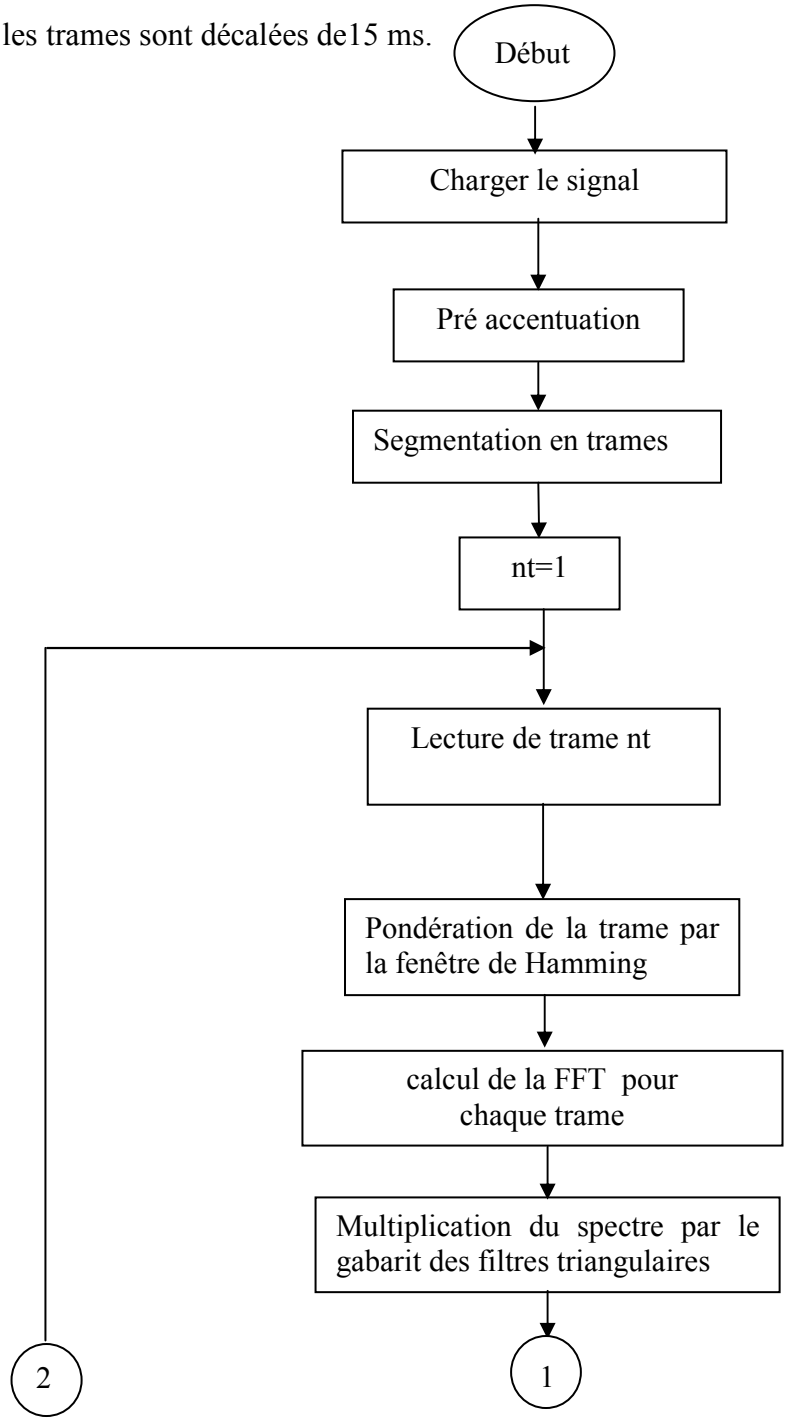


Figure IV.5 L'organigramme de la fonction Vecteurs.

IV.1.3.3 L'organigramme de la fonction Melcepst :

La fonction Melcepst a comme objectif le calcul des coefficients MFCC pour les données d'apprentissage et les données de test, les coefficients MFCC sont une représentation paramétrique du signal vocal, l'objectif de ces paramètres est de réduire la redondance du signal vocal, les coefficients MFCC ne conservent que les paramètres pertinents du signal de la parole dans le but de diminuer le temps de calcul et l'espace de stockage lors des phases d'apprentissage et de reconnaissance.

- la pré accentuation du signal échantillonné est faite à l'aide d'un filtre numérique de premier ordre.
- le signal échantillonné est segmenté en trames de durée de 30 ms (la durée pour laquelle le signal est supposé stationnaire), les trames sont décalées de15 ms.



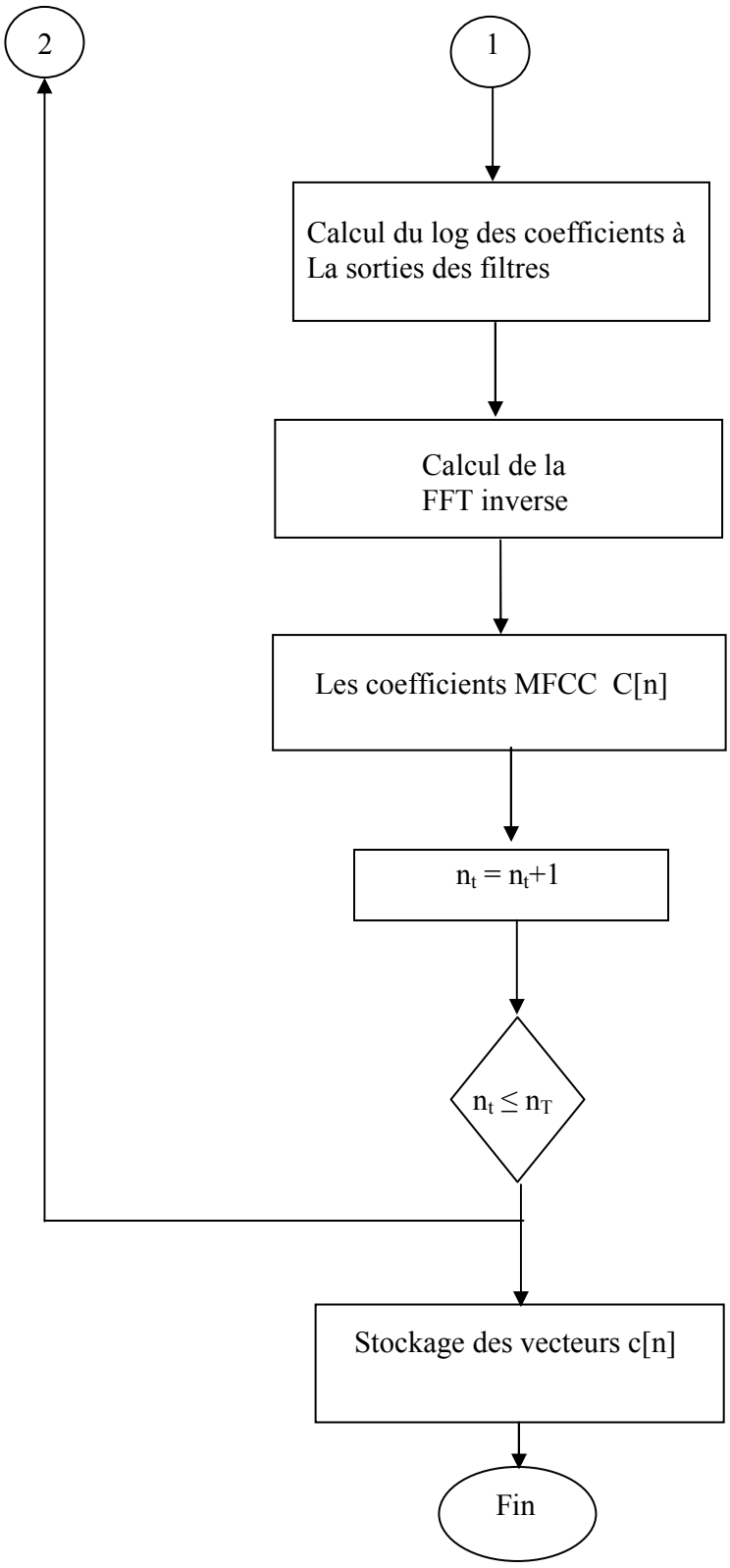


Figure IV.6 L'organigramme de la fonction Melcepst

IV.1.3.4 L'organigramme du module d'apprentissage :

Ce module correspond à toute la phase d'apprentissage, il a pour but la construction des modèles HMMs de références.

Données : Séquence des vecteurs acoustiques.

M : nombre des gaussiennes à chaque état.

Q : nombre des états

nomb_it : nombre d'itérations

diag : le type de la matrice de covariance est diagonale

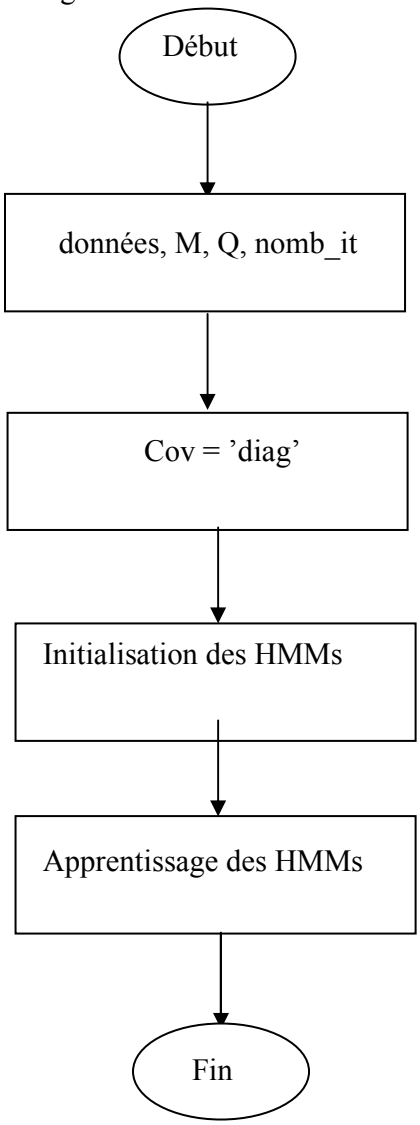


Figure IV.7 L'organigramme du module d'apprentissage

IV.1.3.4.1 L'organigramme d'initialisation des HMMs :

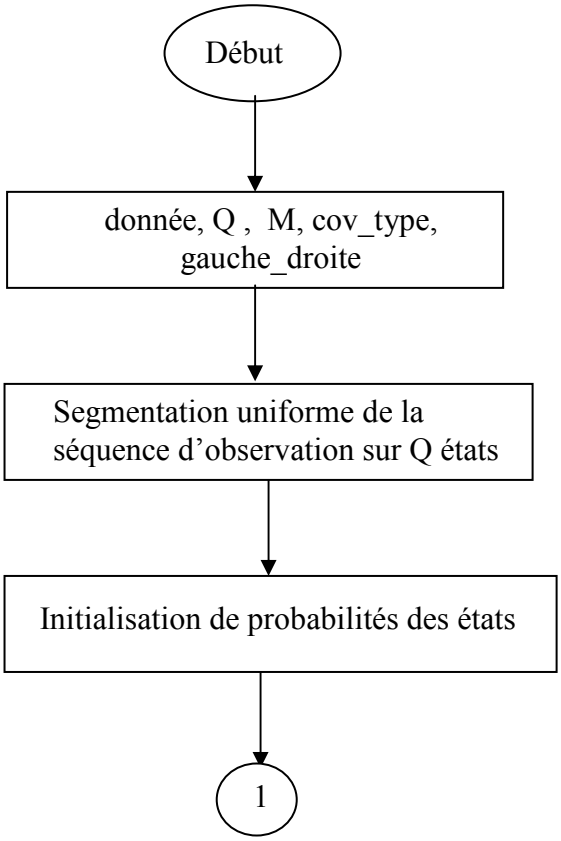
La fonction initialisation-HMM calcule les paramètres initiaux d'un modèle HMM avec des sorties multi gaussiennes.

Les entrées :

- donnée (t, 1) : Le vecteur d'observation de la première séquence au temps t
- Q : Le nombre des états de Markov.
- M : Le nombre de composantes de mélanges.
- Cov_type : Le type de la matrice de covariance.
- gauche_droite =1 : le type du modèle HMM est gauche droite.

Les sorties :

- init_state_prob(i) = Pr (Q(1) = i) : la probabilité à priori initiale.
- transmat(i, j) = Pr (Q(t+1) = j / Q(t) = i) : la matrice de transition.
- mixmat (j, k) = Pr (M(t) =k / Q(t) = j) avec M(t) la gaussienne au temps t .
- mu (j, k) = La moyenne de l'observation y(t) qui donne Q(t) = j, M(t) = k.
- sigma (j, k) = la matrice de covariance de l'observation y(t) qui donne Q(t) = j, M(t) = k.



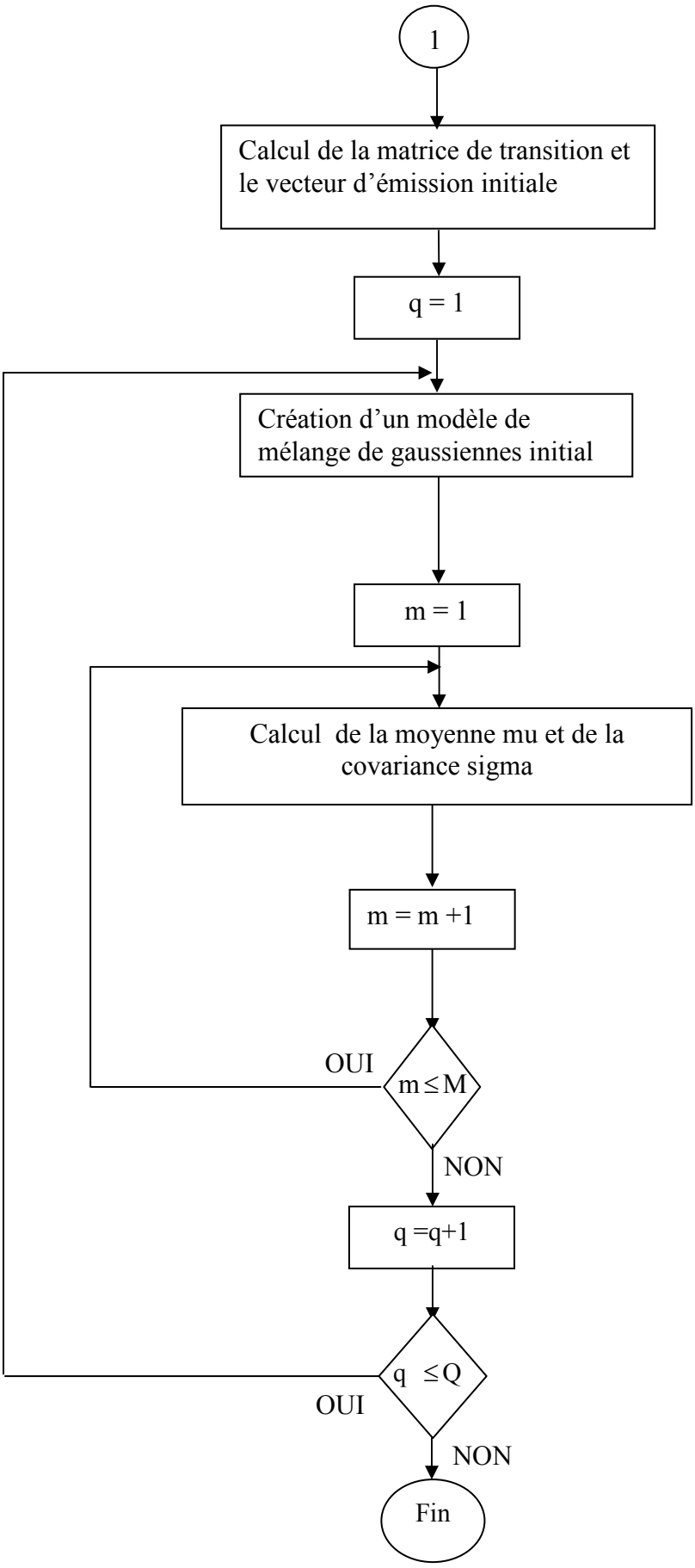


Figure IV.8 L'organigramme d'initialisation des HMMS

IV.1.3.4.2. L'organigramme d'entraînement des HMMs :

Dans cette phase d'apprentissage, nous avons défini un modèle $\lambda(A,B,\pi)$ et nous l'avons utilisé pour recalculer ses variables.

Ainsi nous avons le modèle ré estimé $\overline{\lambda}(\overline{A},\overline{B},\overline{\pi})$, nous pouvons ainsi affirmer l'une ou l'autre de ces propositions :

1- le modèle initial λ définit un point critique de la fonction de vraisemblance, dans ce cas : $\lambda = \overline{\lambda}$

2- le modèle $\overline{\lambda}$ est meilleur que le modèle λ dans le sens ou $p(o/\overline{\lambda}) \rangle p(o/\lambda)$ donc la séquence d'observation o est plus probable avec le nouveau modèle $\overline{\lambda}$. En se basant sur cette procédure, et si nous utilisons itérativement le modèle $\overline{\lambda}$ à la place de λ et si nous répétons l'étape de ré-estimation des paramètres, nous pouvons alors améliorer la probabilité que o soit observée sachant le modèle jusqu'à atteindre un certain point limite. Le résultat final de la procédure de ré estimation est appelé : l'estimation au maximum de vraisemblance du HMM (maximum liklihood estimation)

Les entrées :

$prior_0$: la probabilité d'état initiale.

$transmat_0$: la matrice de transition initiale.

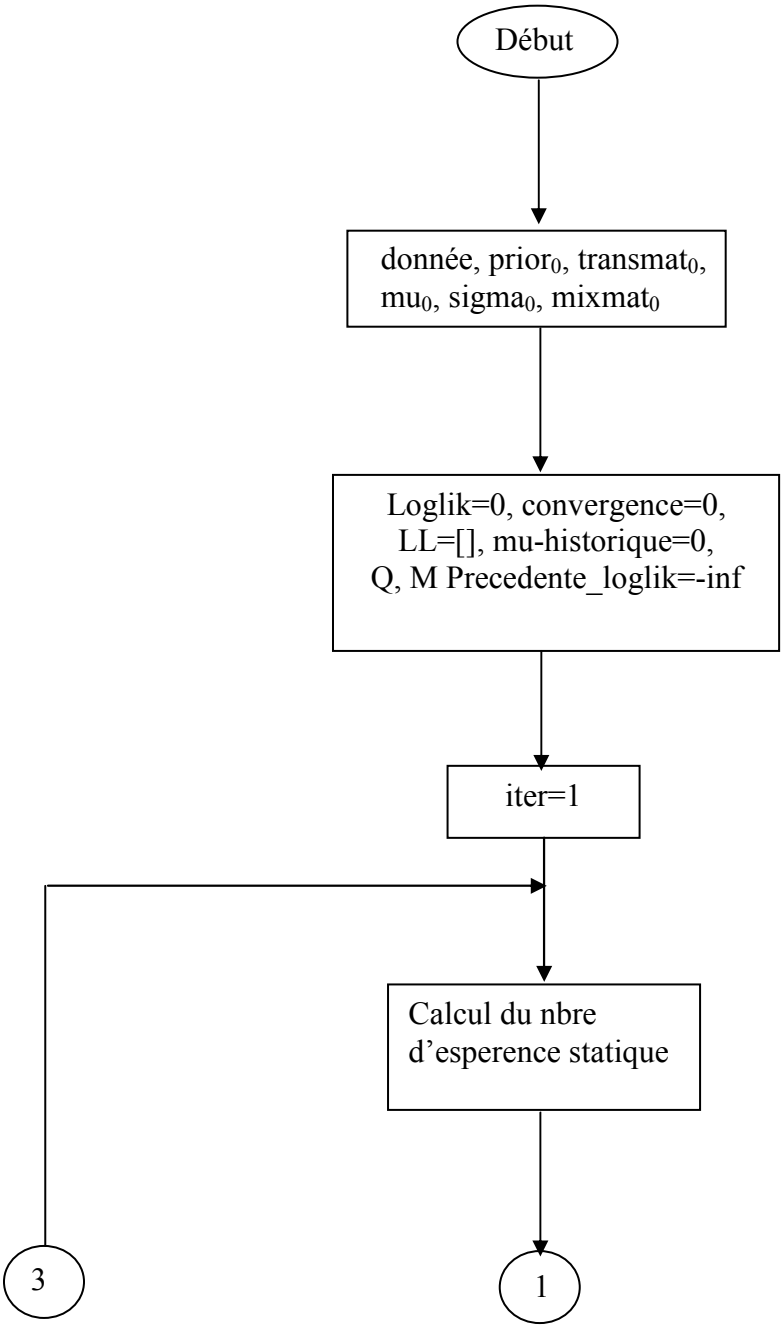
mu_0 : la moyenne initiale.

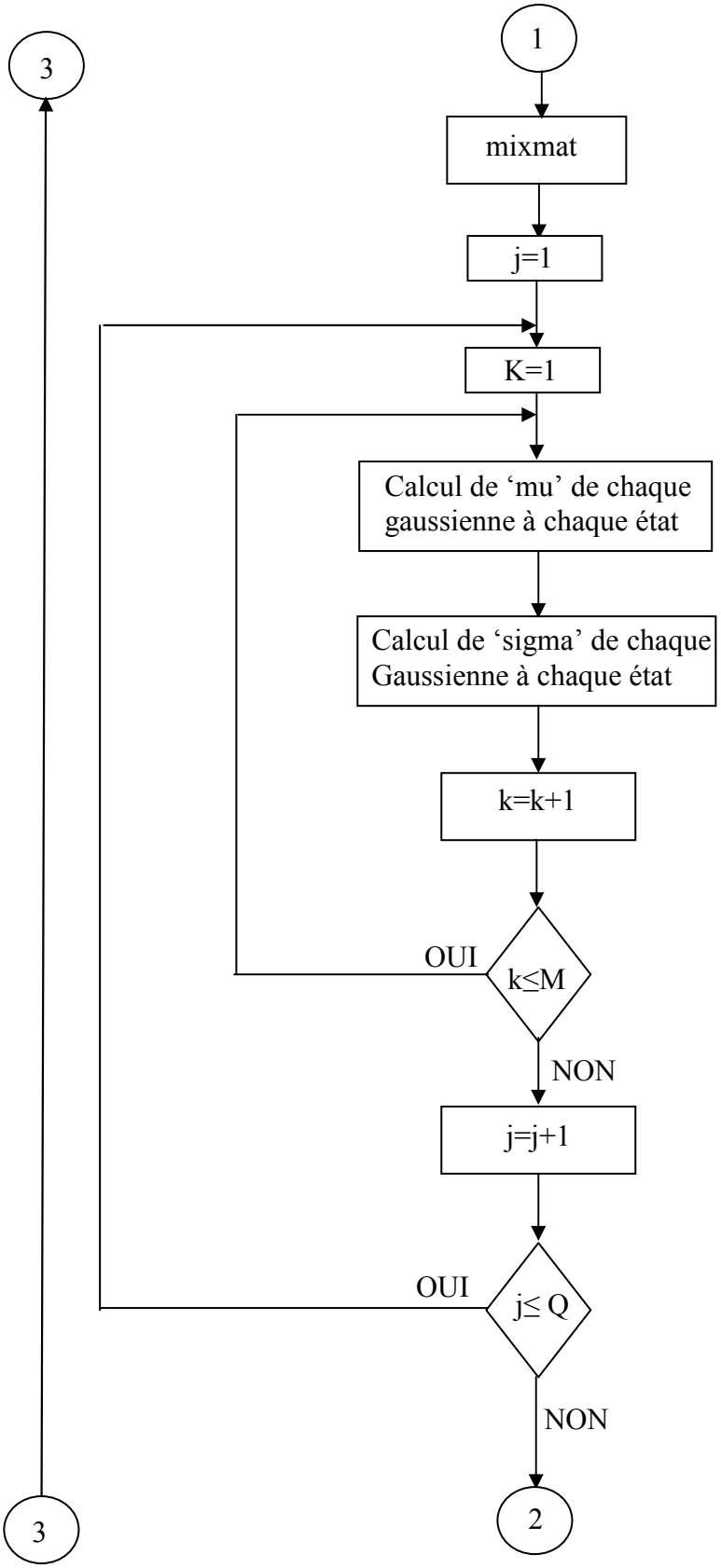
$sigma_0$: la covariance initiale.

$Mixmat_0$: la matrice de mélanges de gaussiennes initiale.

Les sorties :

LL [iter] : le vecteur de vraisemblance à chaque itération.


Début
donnée, prior0, transmat0, mu0, sigma0, mixmat0
Loglik=0, convergence=0, LL=[], mu-historique=0, Q, M Precedente_loglik=-inf
iter=1
Calcul du nbre d'esperence statique
3
1


1
mixmat
j=1
K=1
Calcul de ‘mu’ de chaque gaussienne à chaque état
Calcul de ‘sigma’ de chaque Gaussienne à chaque état
k=k+1
OUI
k≤M
NON
j=j+1
OUI
j≤ Q
NON
2
3
3

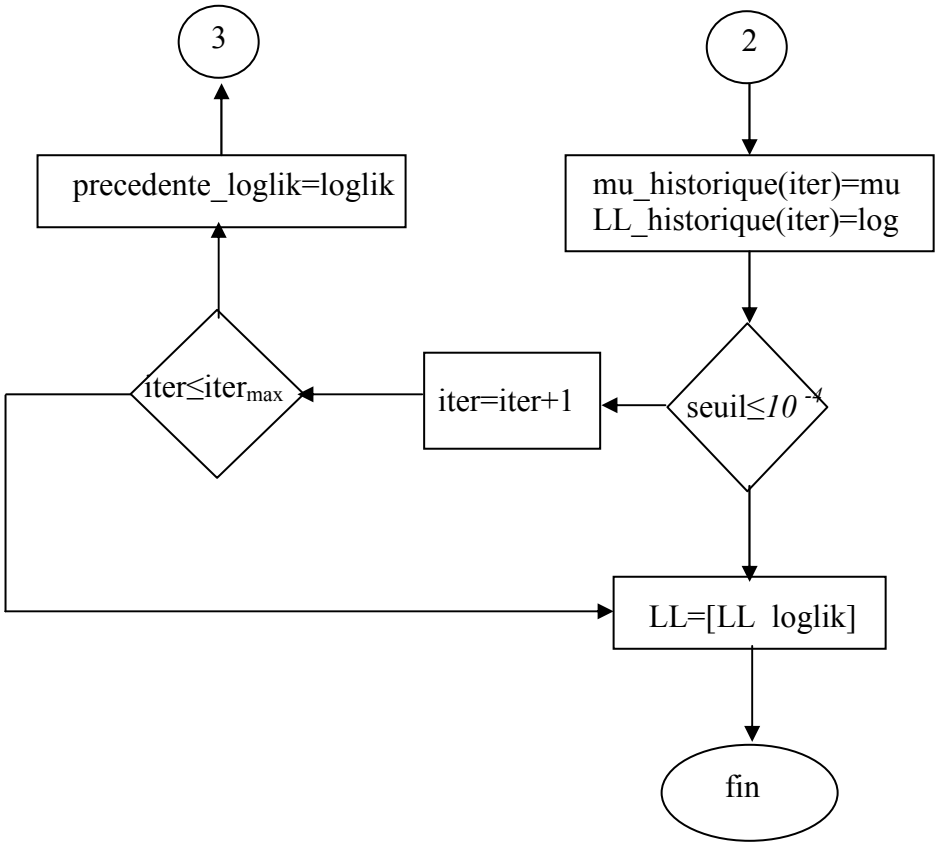


Figure IV.9 L'organigramme d'entraînement des HMMs

IV.1.3.5 L'organigramme du module de reconnaissance :

IV.1.3.5.1 L'organigramme de la fonction de reconnaissance :

Le module de reconnaissance a pour but d'attribuer aux données de tests les modèles HMMs les plus probables et cela en utilisant l'algorithme de Viterbi :

Les entrées :

Donnés (t, d) : c'est le vecteur d'observation au temps t de la séquence 1.

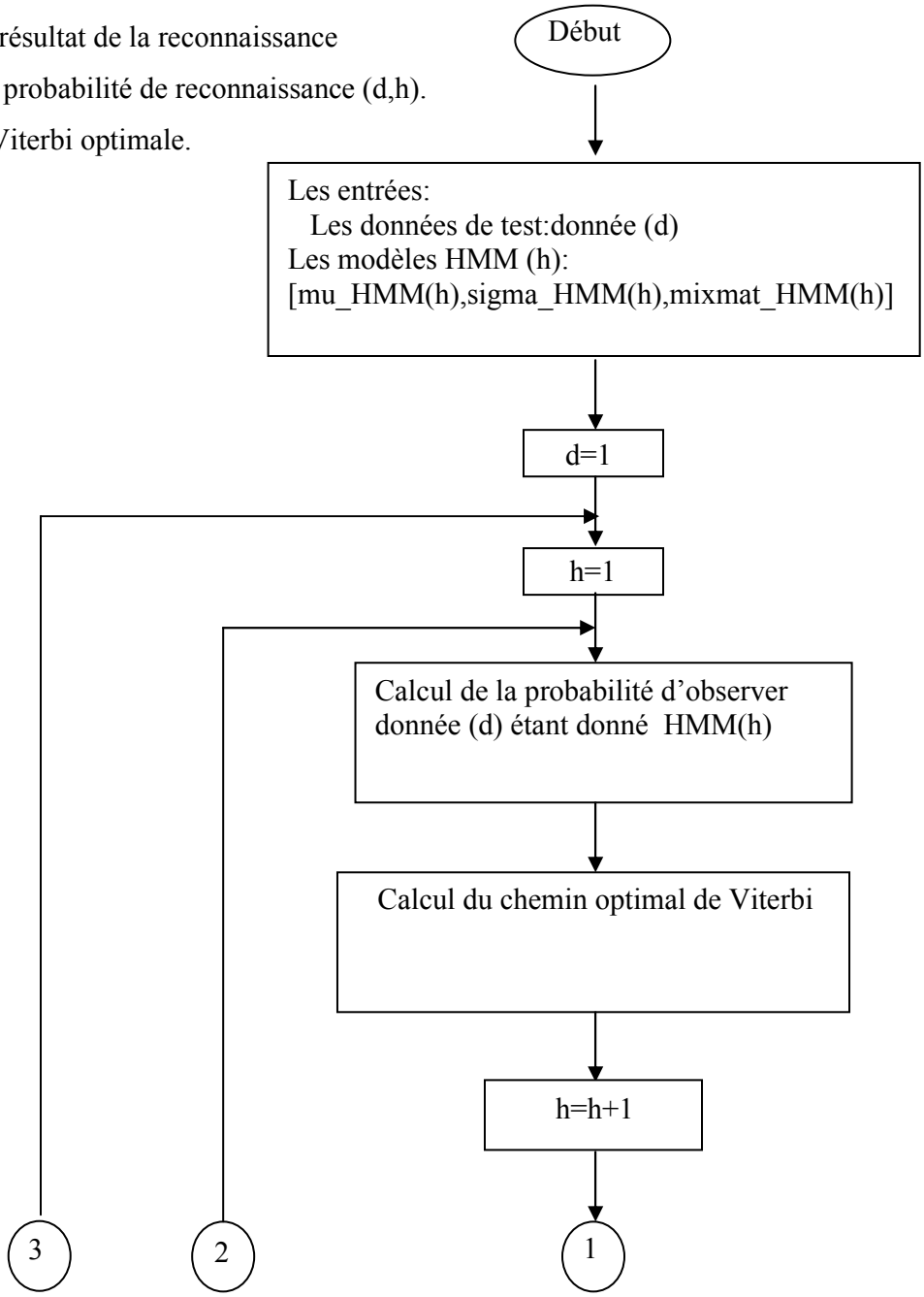
HMM_1, HMM_2, HMM_3,…: c'est les modèles de Markov.

Les sorties :

Résultat : le vecteur de résultat de la reconnaissance

LoglikV : la matrice de probabilité de reconnaissance (d,h).

chemin : le chemin de Viterbi optimale.



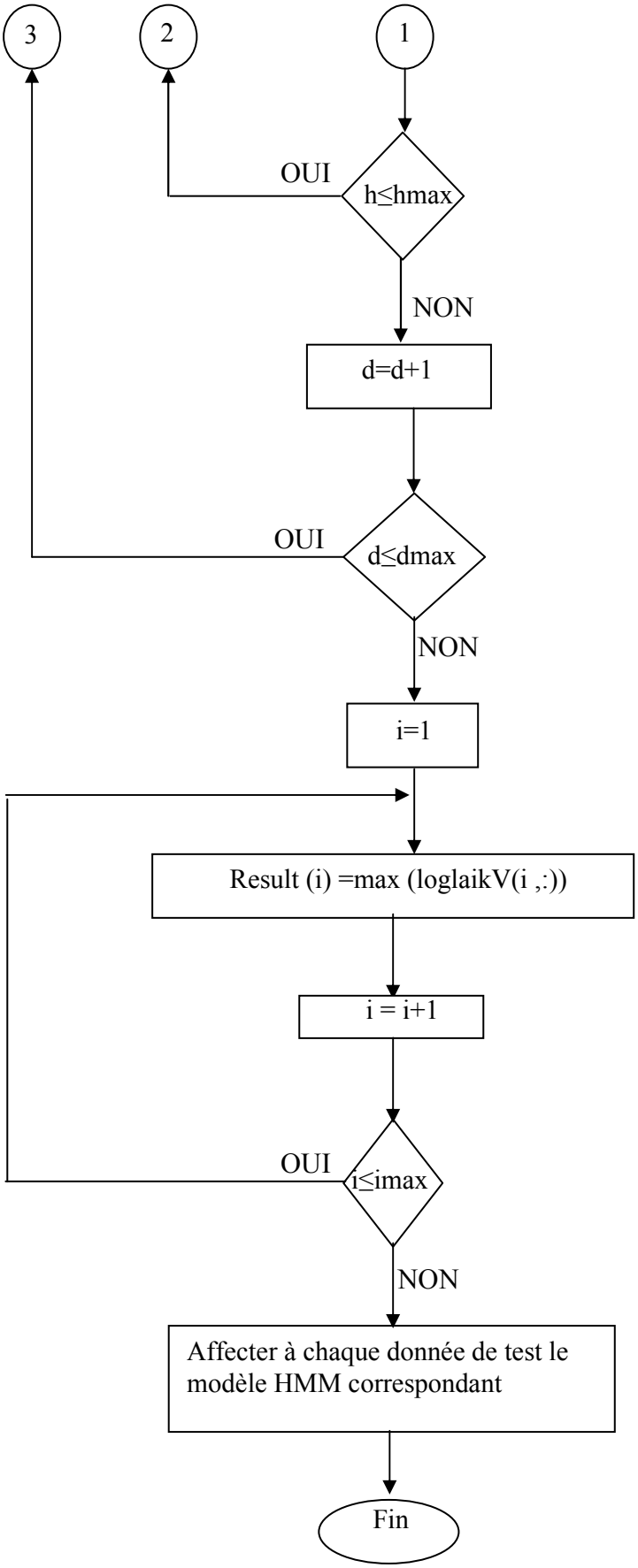


Figure IV.10 L'organigramme de la fonction de reconnaissance.

IV.1.3.5.2 L'organigramme de la fonction de Viterbi :

Les entrées :

prior (i) = pr(Q(1)=i): la probabilité d'être dans l'état i à l'Instant t =1.

transmat (i,j) = pr(Q(t+1) = j / Q(t) = i): la matrice de transition de l'état i à l'instant t vers l'état j à l'instant t+1.

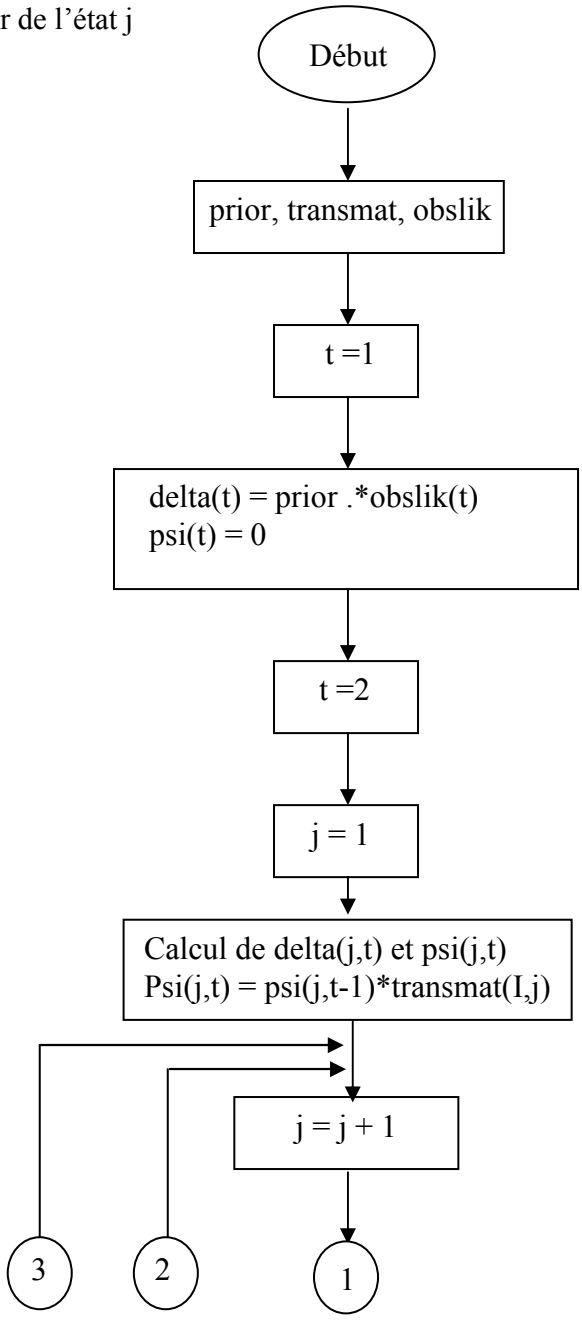
obslik (i,t) = pr(y(t) / Q(t) = i): la probabilité d'émettre l'observation y à l'instant t étant donné l'état Q.

Les sorties :

chemin (t) = q(t) : l'argument maximal des états Q à l'instant t.

delta (j,t) = la probabilité de la meilleure séquence allant à l'état j et émettant l'observation O(1 :t).

psi (j,t) = Le meilleur prédécesseur de l'état j



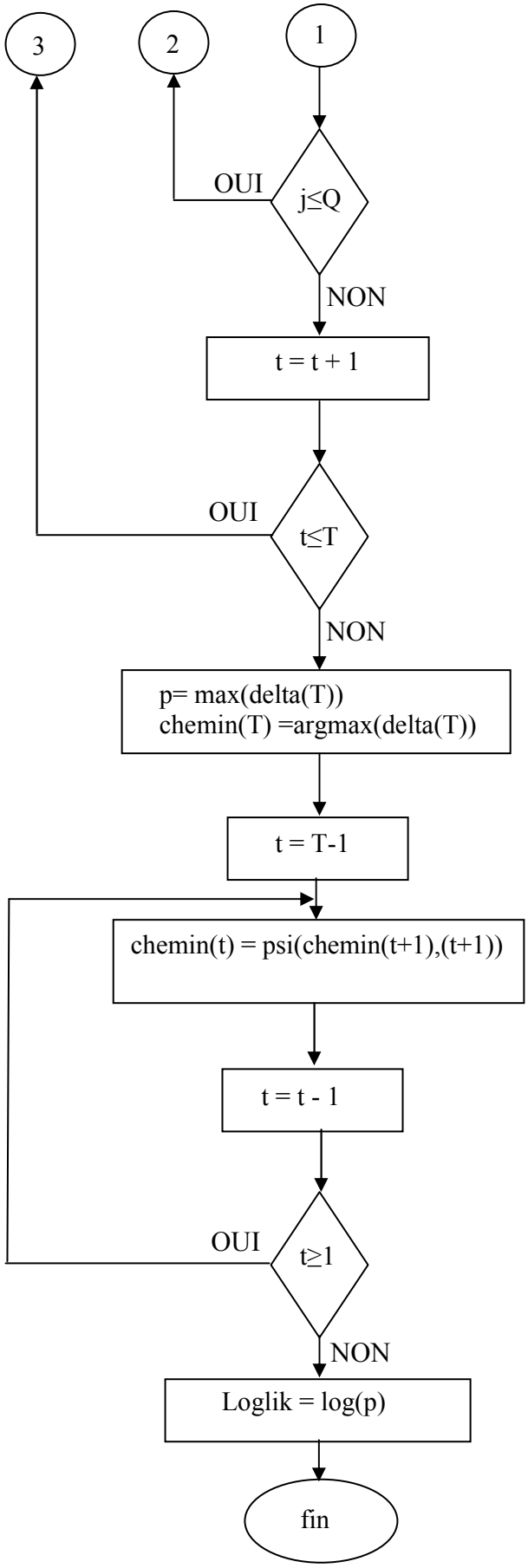
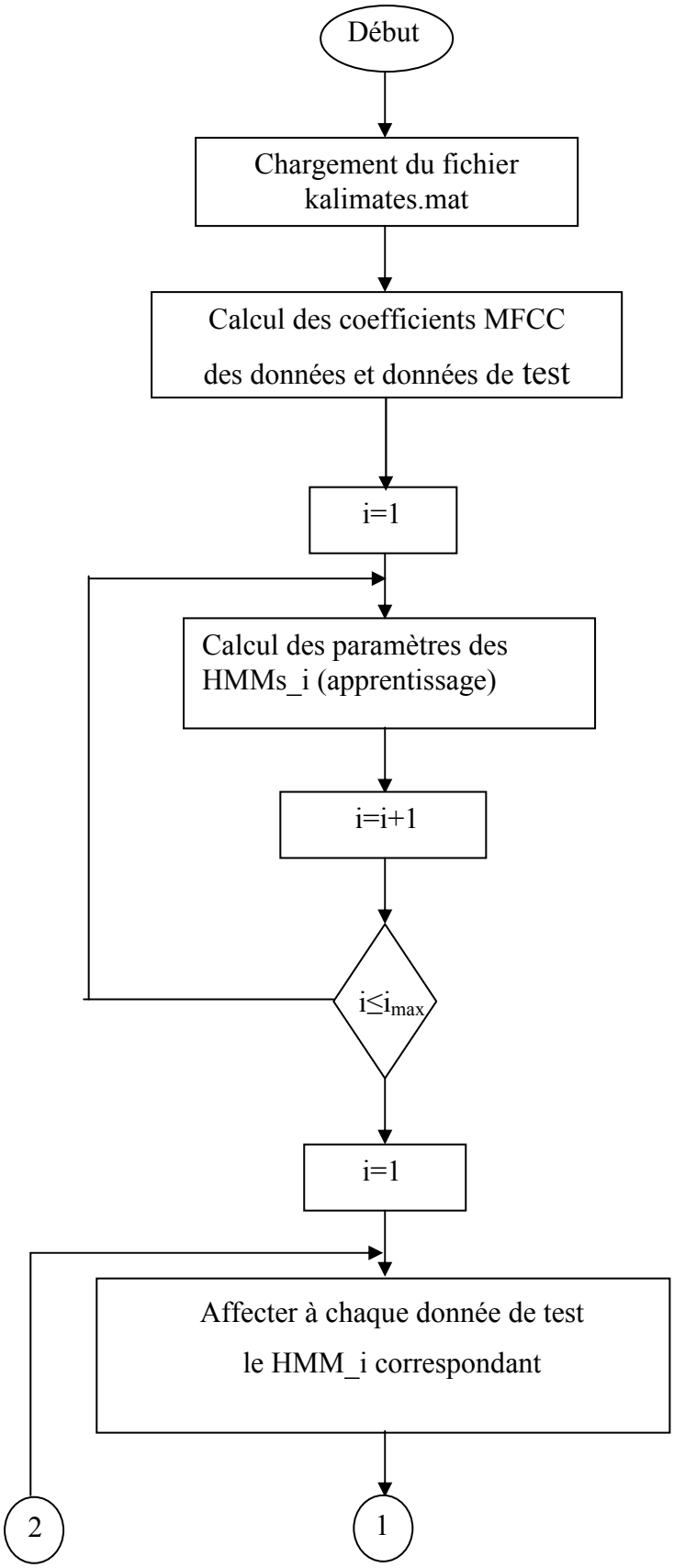


Figure IV.11 L'organigramme de la fonction de Viterbi

IV.1.3.6 L'organigramme de la fonction globale :



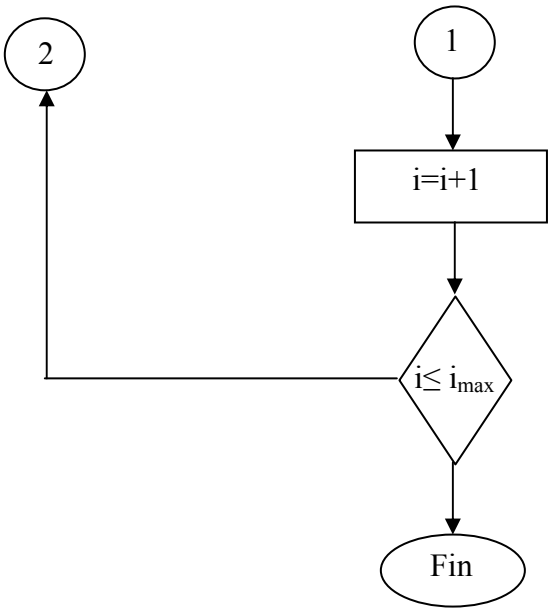


Figure IV.12 L'organigramme de la fonction globale.

VI.2 Mise en oeuvre du système de RAP avec détection du hors vocabulaire :

L'idée de base de ce système est fondée sur le même principe précèdent, utilisation des modèles de Markov cachés, c'est-à-dire que le module d'apprentissage reste le même, mais la différence réside dans la procédure de reconnaissance là où on va calculer deux scores différents : un score de reconnaissance de mots par rapport au HMM-mot qui est le même score du premier système et un score poubelle qui est l'ensemble des scores de trames de parole par rapport au HMM-mot.

En effectuant une comparaison des deux scores par rapport à un seuil on pourra juger si ce mot appartient aux vocabulaires donc on va le reconnaître si non on va le rejeter.

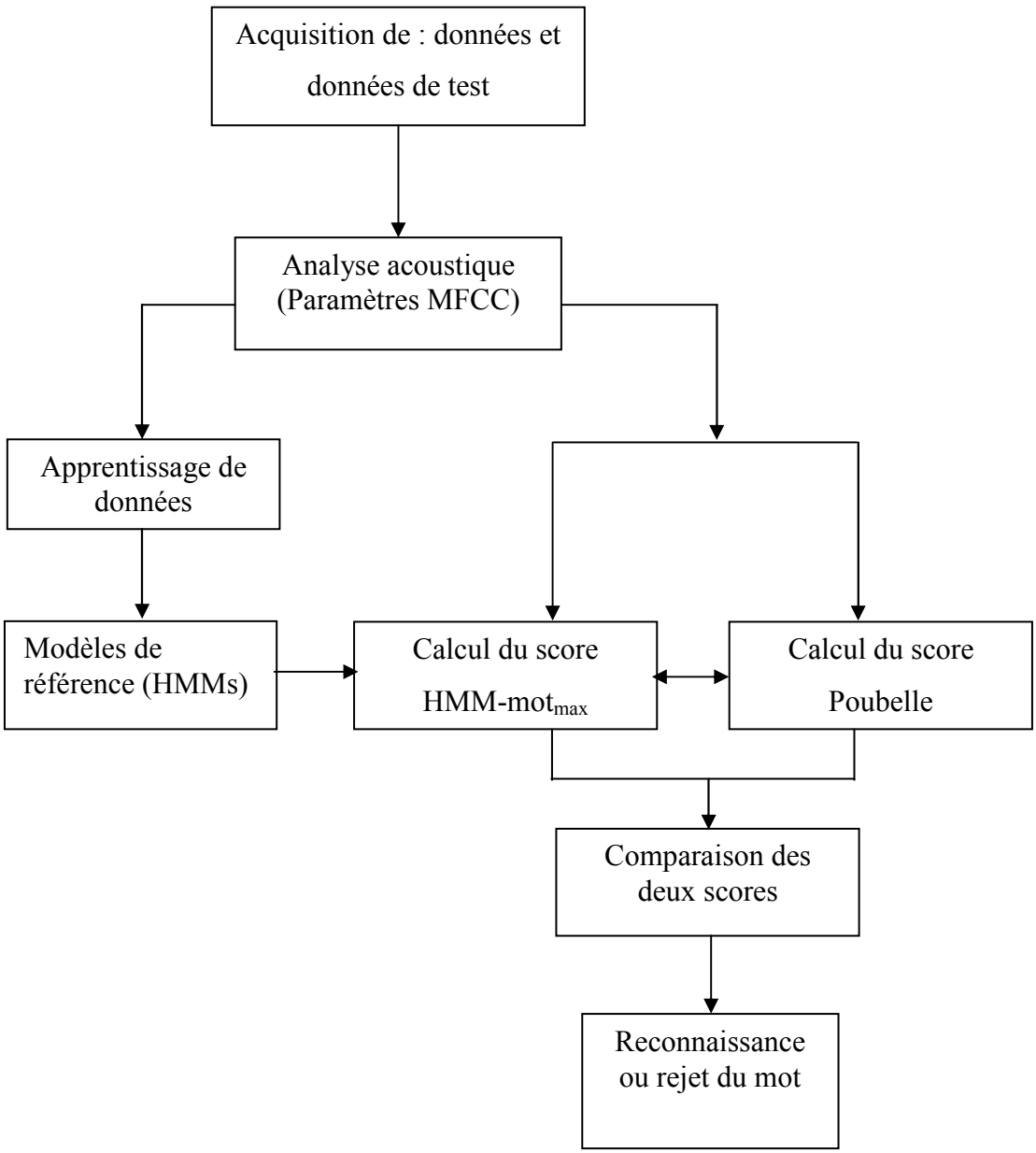
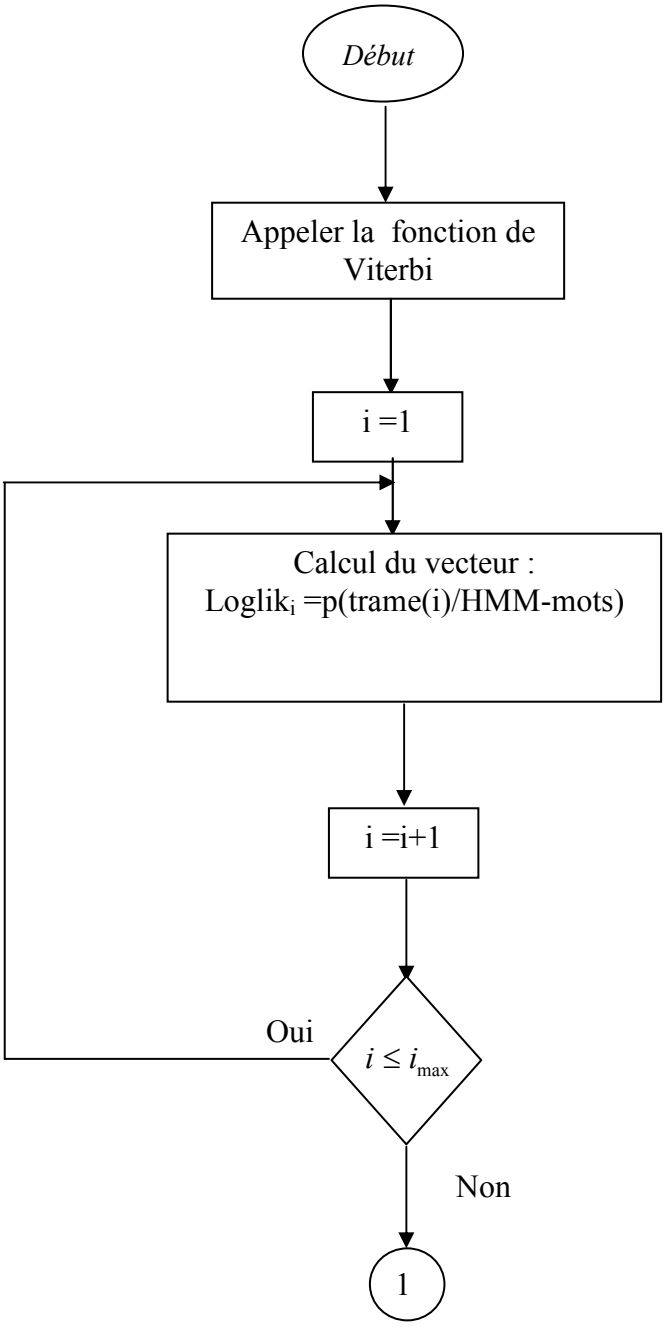


Figure IV13 Structure globale du système de RAP avec détection du hors vocabulaire

IV.2.1 L'organigramme de calcul du score de référence :

le meilleur chemin du calcul du score de référence est celui de Viterbi , dont on va utiliser pour calculer la probabilité d'observer une trame de parole étant données les modèles HMM-mots. Dans ce cas pour chaque trame de parole constituant une partie du mot prononcé une probabilité maximale $loglik_{maxt}$ et une probabilité moyenne $loglik_{moyt}$ sont cherchées. Ensuite pour chaque trame ces deux valeurs sont extraites dans un ordre descendent, puis on va calculer le modèle poubelle pour chaque trame à chaque instant i en effectuant l'interpolation linéaire entre le maximum et la moyenne du maximum de vraisemblance. Enfin le modèle poubelle ou le score de référence est tout simplement l'accumulation des scores de références de chaque trame à chaque instant i.



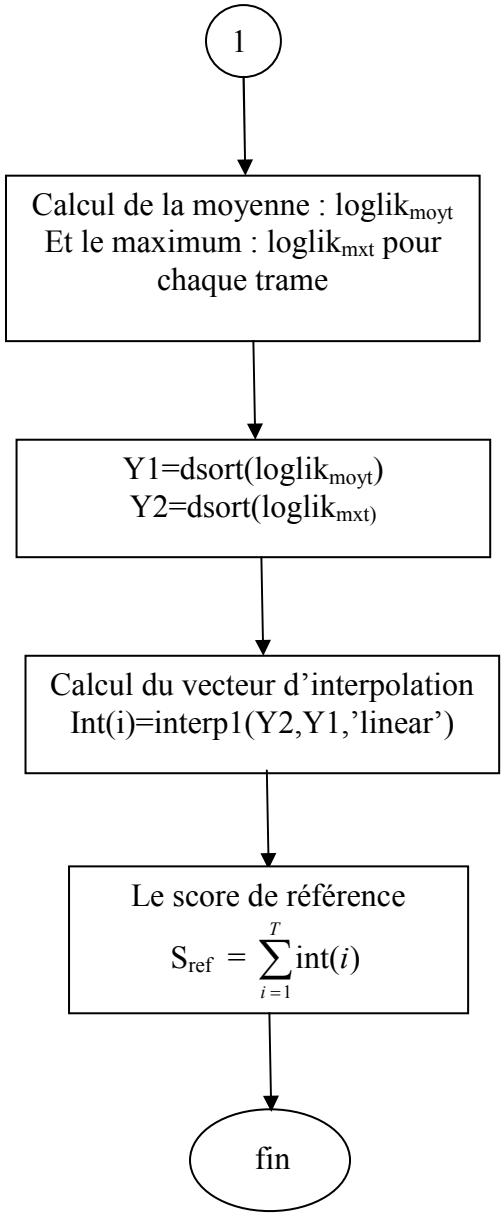


Figure IV14 l'organigramme de calcul du score de référence

IV.2.2 L'organigramme du module de reconnaissance :

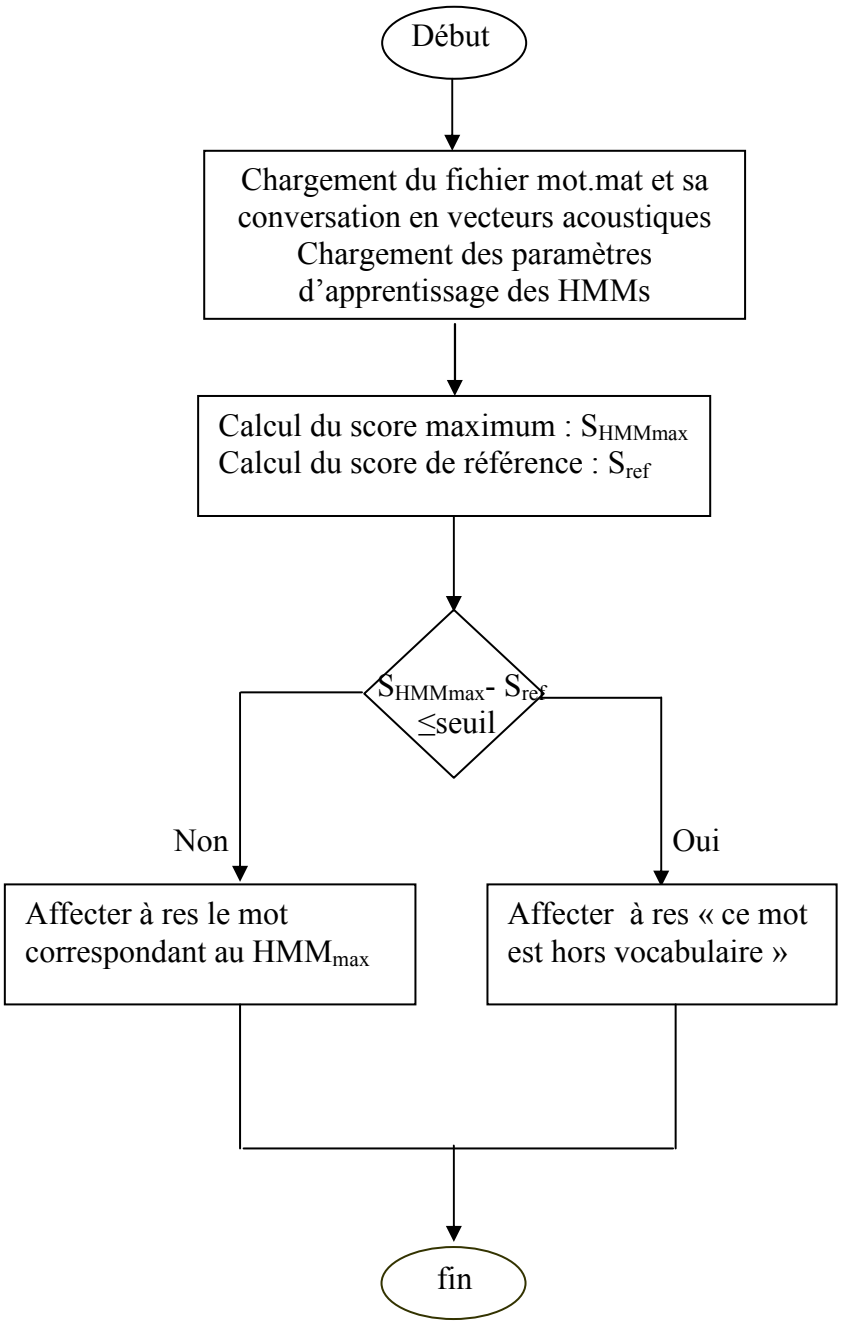


Figure IV.15 l'organigramme global de RAP avec détection du hors vocabulaire

IV.3 Tests et résultats

IV.3.1 Introduction :

Dans cette section on va décrire les tests effectués et les résultats obtenus par les deux systèmes, ainsi qu'une discussion sur les différents résultats obtenus pour chaque système.

IV-3-2Condition expérimentale :

Les deux systèmes sont réalisés avec un micro_ ordinateur Pentium IV équipé d'une RAM de 128 Mo de mémoire, un disque dur de 20 Go et une carte son (SOUND BLASTER AUDIOPCI 128) avec un logiciel d'acquisition et de traitement (CREATIVE WAVE STUDIO).

Le corpus d'apprentissage utilisé pour les deux systèmes est constitué de 12 mots dont chaque mot est prononcé respectivement par sept locuteurs des deux sexes, en langue arabe (4 garçons , 3 filles) dont l'age varie entre 14 à 45 ans.

L'enregistrement a été effectué dans une chambre ordinaire.

Notre système se classe en mode indépendant du locuteur .Nous avons pris comme application, un sous ensemble de commandes de l'explorateur Windows version arabe.

IV-3-3Le dictionnaire de référence : Le dictionnaire de référence est constitué de12 mots.

Tableau IV.1 : Le dictionnaire de référence

أدوات	1
عرض	2
بحث	3
حذف	4
إلغاء	5
إزالة	6
لصق	7
ملف	8
نسخ	9
تشغيل	10
تعيين	11
طباعة	12

Les mots de notre corpus sont enregistrés avec une fréquence d'échantillonnage Fs=16 KHZ et codés sur 16 bits. La séparation signal – silence a été faite manuellement à l'aide du logiciel CREATIVE WAVE STUDIO. Après cette phase du traitement, le signal sera transformé en un ensemble de paramètres qui sont les coefficients MFCC. Les figures ci-dessous montrent les différentes étapes pour l'analyse acoustique exemples des mots : **(أدوات .عرض).**

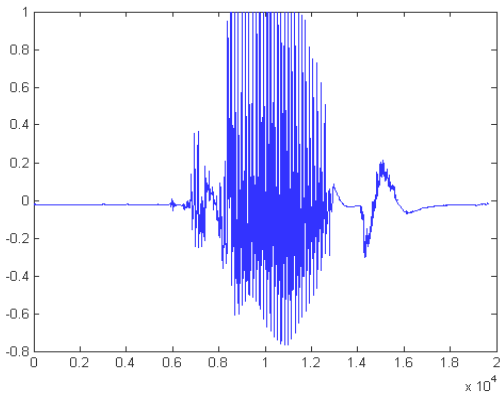


Figure IV.16 Le mot **' أدوات '** original

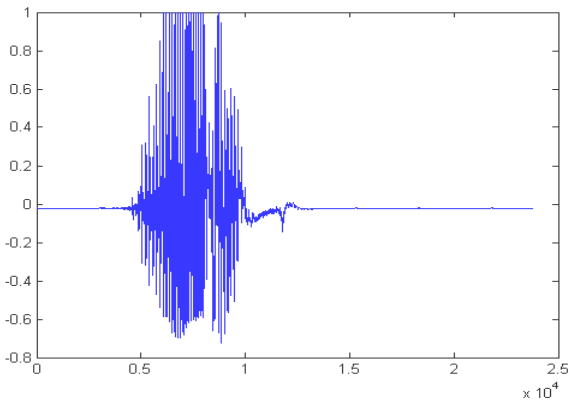


Figure IV.17 Le mot**' عرض '** original

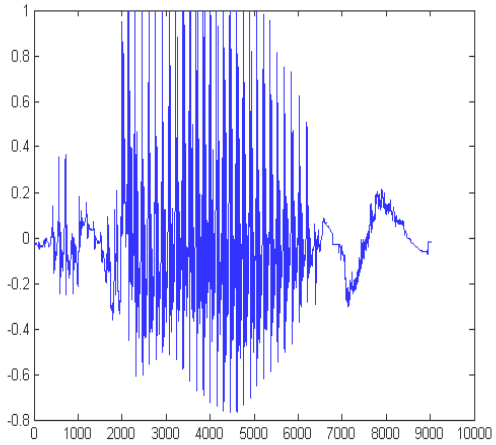


Figure IV.18 :Séparation : signal – silence pour le mot : **ادوات**

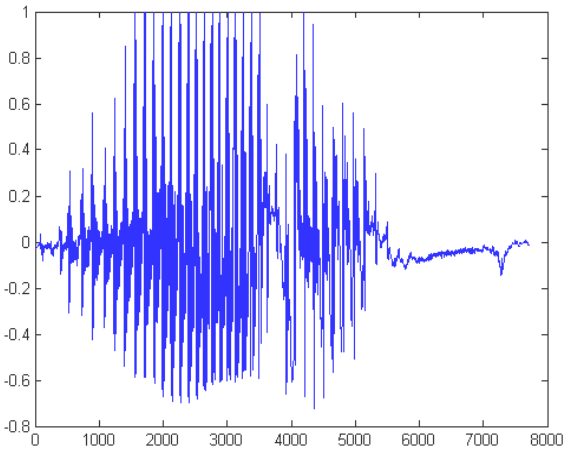


Figure IV.19: Séparation : signal – silence pour le mot : **عرض**

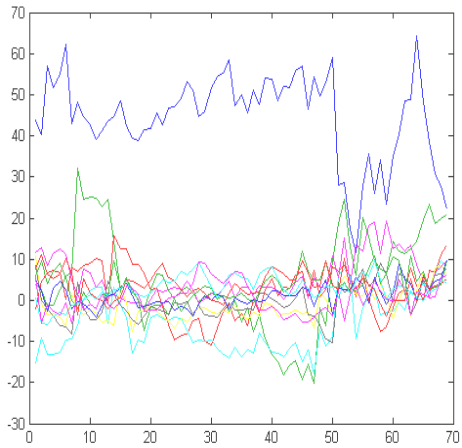


Figure IV.20: Cœfficients MFCC pour le mot :**أدوات**

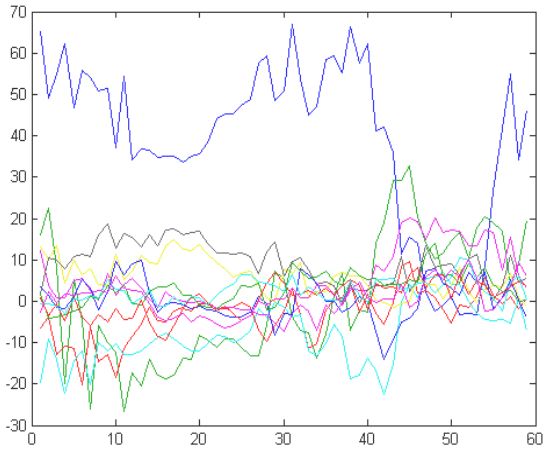


Figure IV.21: Cœfficients MFCC pour le mot :**عرض**

IV.3.4 Expérience 1 :

Le but de cette expérience est de tester la performance de notre système à savoir le calcul du taux de reconnaissance.

La base de données de test utilisée pour cette expérience est constituée des mêmes mots utilisés pour la phase d'apprentissage, mais prononcés par des locuteurs différents de ceux qui ont participé à la phase d'apprentissage, ce qui permet de tester l'indépendance de notre système.

Le modèle HMM utilisé pour notre système est un HMM type gauche – droite constitué de trois états. Pendant la phase d'apprentissage, trois composantes gaussiennes sont utilisées par état pour modéliser les données d'apprentissage, et une seule itération pour l'apprentissage.

Pendant la phase de reconnaissance, le système attribue au mot à tester le modèle HMM qui a le plus haut score.

Le tableau ci-dessous donne le taux de reconnaissance des mots prononcés.

Tableau IV.2: Résultats de test pour le premier système

Mots		HMM_{mot}
1	**أدوات**	reconnu
2	**عرض**	reconnu
3	**بحث**	reconnu
4	**حذف**	reconnu
5	**إلغاء**	reconnu
6	**إزالة**	reconnu
7	**لصق**	reconnu
8	**ملف**	reconnu
9	**نسخ**	reconnu
10	**تشغيل**	reconnu
11	**تعيين**	reconnu
12	**طباعة**	reconnu
Taux de reconnaissance : 100 %		

IV.3.5 Commentaires :

Malgré le succès de notre test, mais cette implantation reste limitée à un petit vocabulaire, notre système fournit un score qui caractérise la liaison entre HMM et le mot à reconnaître, ce score est néanmoins relatif, et il dépend du locuteur et de l'environnement d'utilisation, c'est pour ça il ne peut pas être utilisé comme mesure de la justesse de liaison.

IV.3.6 Expérience 2 :

Le but de cette expérience est de vérifier la performance de notre système à savoir la détection du hors vocabulaire. Le test a été effectué par des locuteurs indépendants de ceux ayant participés à la phase d'apprentissage autrement dit mode indépendant.

Le corpus de test comporte vingt et un mots, parmi eux douze mots qui ont été utilisés pour la première expérience et que le système doit reconnaître, et les neufs autres mots sont des mots hors vocabulaire, et que le système doit rejeter.

En tous le corpus de test comporte 114 mots dont 60 mots appartenant au vocabulaire et 54 sont des mots hors vocabulaire.

Pour ce test, la valeur du rang pour le calcul du score de référence égale à 0.90.

Le test a été effectué pour différentes valeurs de seuil :

Le tableau ci-dessous représente le corpus de test utilisé pour notre système.

Tableau IV.3: Corpus de test pour le deuxième système

Mots du vocabulaire		Mots hors vocabulaire	
01	أدوات	01	إيقاف
02	عرض	02	إلقاء
03	بحث	03	إنهاء
04	حذف	04	قص
05	إلغاء	05	خصائص
06	إزالة	06	مسح
07	لصق	07	موافق
08	ملف	08	تجاهل
09	نسخ	09	تكبير
10	تشغيل		
11	تعيين		
12	طباعة		

Les résultats obtenus sont présentés dans le tableau ci-dessous :

Tableau IV.4: Différents résultats de reconnaissance

Seuil	10	9	8	7	6	5	4	3	2	1	-1
Taux du rejet du HV (%)	69	73	78	88	89	89	91	97	98	98	99
Taux du rejet du V (%)	0	0	0	0	0	0	0	0	0.033	0.05	0.13
Taux de reconnaissance (%)	85	87	89.5	94	95	95	96	98.5	99.04	99.03	99

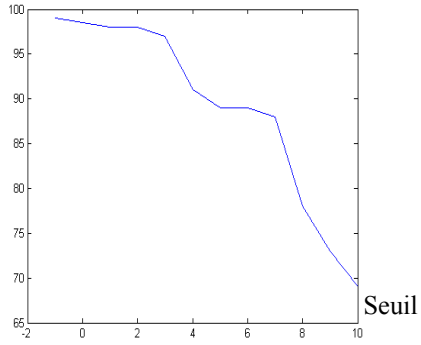


Figure IV.22: Taux du rejet du HV (%)

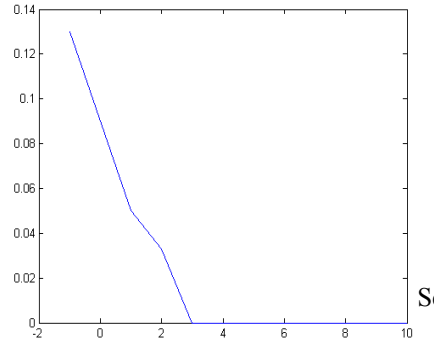


Figure IV.23: Taux du rejet du V (%)

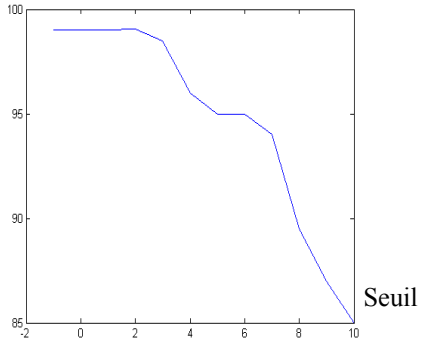


Figure IV.24: Taux de reconnaissance (%)

IV.3.5 Commentaires :

D'après les résultats obtenus ci-dessus, on remarque que le taux du rejet du hors vocabulaire change en fonction du seuil, et il atteint sa valeur maximale pour un seuil égale à -1 (un taux de rejet égale à 99 (%)), alors que le taux du rejet du vocabulaire égale à zéro pour une valeur de seuil supérieure à 2. Le taux de reconnaissance atteint sa valeur maximale pour une valeur de seuil égale à 2. On constate aussi qu'en utilisant un petit seuil positif, on peut détecter 97% du hors vocabulaire sans rejeter aucune bonne reconnaissance.

En plus de ça, le système ne rejette pas les mots qui sont acoustiquement proche des mots du vocabulaire qu'à une petite valeurs de seuil, dans notre cas se sont les mots (**إلقاء،قص،مسح**) qui sont acoustiquement proche des mots (**إلغاء،لصق،نسخ**).

IV-4.Conclusion :

Dans ce travail, on a mis en œuvre deux systèmes de RAP basé sur les modèles de Markov cachés (HMM).

Le premier système est un système qui permet de reconnaître le mot prononcé et ça en calculant le maximum de vraisemblance entre le mot prononcé et le modèle HMM comme score de liaison.

Ce score est relatif et ne reflète jamais la bonne liaison entre le mot prononcé et le modèle HMM.

Autrement dit, si on prononce un mot hors vocabulaire, le système va lui attribuer un mot du dictionnaire le plus semblant, ce qui va introduire une erreur de reconnaissance,

Pour remédier à ce problème, on a implanté un deuxième système qui est basé sur le même principe que le premier, et la différence réside dans la décision de reconnaissance, où on a appliqué une mesure de confiance basée sur le modèle acoustique, cette mesure de confiance accepte le mot prononcé, donc le mot va être reconnu si non elle le rejette.

CONCLUSION GENERALE

Le travail que nous avons fait vise à résoudre un des problèmes primordiaux de la reconnaissance automatique de la parole. Nous avons réalisé un système de reconnaissance de la parole des mots isoles de la langue arabe en mode indépendant du locuteur basé sur les modèles de Markov cachés (HMM)

Pendant la phase d'apprentissage, des locuteurs des différents sexes, âges et dialectes prononcent un ensemble de mots choisit, les mots prononcés sont enregistrés en qualité de 16 KHZ 16 bits mono.

Une fois le mot prononcé et enregistré, il passe par la phase d'analyse acoustique à savoir l'extraction de l'information pertinente (les coefficients MFCC).

La phase d'apprentissage consiste à construire des modèles HMMs ainsi que l'optimisation de chacun des paramètres de chaque modèle HMM et ceci est réalisé en utilisant l'algorithme de Baum Welsh.

A la fin de la phase d'apprentissage, on obtient les modèles HMMs constituants le dictionnaire de référence.

Pendant la phase de reconnaissance les mots de tests passent par la même analyse acoustique que ceux d'apprentissage, la phase d'apprentissage consiste à trouver la liaison qui lie le mot prononcé au modèle HMM en utilisant l'algorithme de Viterbi.

A la fin de cette phase, le système va attribuer au mot prononcé le modèle HMM qui est le plus semblant.

A fin d'améliorer notre système de telle façon qu'il soit capable de détecter le hors vocabulaire, on a appliqué une mesure de confiance basée sur le modèle acoustique en utilisant la technique de modélisation poubelle directe (on-line garbage modeling) qui consiste à calculer un score de référence qui est l'accumulation des scores des trames constituantes le mot prononcé, ainsi la confiance n'est que la différence entre le score de référence et celui fournit par le modèle HMM.

Selon un seuil prédéterminé, le système doit décider si le mot prononcé appartient au dictionnaire de référence, donc il va lui attribuer le modèle HMM le plus semblant, si non le mot prononcé sera rejeté.

Les tests que nous avons effectué ont fournit de bon résultats, ce qui implique la compatibilité de la technique utilisée (modélisation poubelle directe) avec les modèles HMMs, où nous avons réalisé un système qui permet de détecter le hors vocabulaire sans rejeter le vocabulaire.

Malgré le succès de cette technique pour notre système, elle reste dépendante de l'application, où la valeur de seuil optimale doit être ajuster individuellement pour chaque système.

BIBLIOGRAPHIE

[1] Nicolas Moreau, Denis Jouvet, « Détermination d'une Mesure de Confiance pour le Rejet des Entrées Incorrectes », XXIIIèmes Journées d'Etude sur la Parole, Aussois, 19-23 juin 2000

[2] Olli Vikki,Kari Laurila, petri Haavisto, « A confidence measure for detecting recognition errors in isolated words recognition », Speech and audio systems, Tampere, Finland.

[3] M. Machowski, « Speech recognition and natural language processing as a highly effective means of human computer iteration », http://www.coefaculty.valdosta.edu/icellis/project speech recognition .doc .

[4] R. Boite et M. Kunt, «Traitement de la parole», press polytechniques Romandes, 1987.

[5] J. Koreman, B. Andreeva, et H. Strik, « Acoustic parameters versus phonetic features in ASR », In proceeding of international congress of phonetic sciences, pages 549-553, 1999.

[6] Hacine Gharbi Abdenour, « Reconnaissance automatique de mots isolés arabes par différentes variantes de la DTW », mémoire de magistère, institut d'électronique, université de Sétif, 2002.

[7] Foret Cathrine , « La reconnaissance vocale », http://www.informed.free.fr.

[8] http://www.geneve.ch/heg/campus/travaux/igs/sites/2002_04/reconnaissancevocale.html

[9] http://www.easytel.fr/info/dicteevocale.html.

[10] Yassine Ben Ayed , « détection de mots clés dans un flux de parole », thèse de doctorat, ENST Paris, 2003.

[11] L. Rabiner et B. H. Juang, « Fundamentals of speech recognition, Prentice-Hall », 1993.

[12] Christophe Gérard « Etude de la paramétrisation du signal de parole à partir de représentation en ondeléttes », Thèse de Doctorat de l'université de paris 11,1995

[13] S. B. Davis et P. Mermelstein, « Comparaison of parametric representations for monosyllabic word recognition in continuously spoken sentences », IEEE Transactions on Acoustic, Speech and Signal Processing, pages : 357-366,1980.

[14] Calliope, « La parole et son traitement automatique ». Editions Masson, 1989.

[15] .Alexy Ozerov, « Représentations robustes pour la reconnaissance automatique de la Parole », http://www.ee.kth.se/~Ozerov/publications/ozerov_DESS_report.ps.

[16] Technologie de l'information et de la communication, « Reconnaissance de la parole », (Fiche technologie N 59).

[17] L. Rabiner et B. H. Juang , « A tutorial on hidden Markov models and selected applications in speech recognition, processing of IEEE, 77 : 257-285,1989.

[18] B. H. Juang S. E. Levinson et M. M. Sondhi, « Maximum likelihood estimation for multivariate mixture observations of Markov chains », IEEE Trans. Information Theory. IT-32(2) : 307-309, 1986

[19] R.Bakis, « continuous speech word recognition via centisecond acoustic states », in proc.ASA meeting (Washington, DC).April 1976.

[20] Erhan Mengusoglu, Olivier Deroo, « Confidence measures in HMM/MLP hybrid speech recognition for Turkish language », Proceeding of the ProRISC/ IEEE workshop, pp 413-417

[21] Boolard. H. D'hoore. B. Boite. J-M, (1994), « Optimizing recognition and rejection performance in wordspotting systems, proceeding of the IEEE international conference of acoustics », speech and signal processing, pp.I-373 –I-376 , Adelaide , Australia.

[22] Rahim, M,G ,Lee C.H ,Juang, B.H (1995), « Robust utterance verification for connected digits recognition », proceeding of the IEEE international conference on acoustics ,speech ,and signal processing , pp.285-288,detroit.USA.

[23] Iso –Sipila, J, Laurila, K,Haavisto, P(1996), «Optimal adaptive garbage modelling in speech recognition », proceeding of the IEEE Nordic signal processing symposium, Helsinki,Finland.

[24] L.R.Rabiner,B.H.Juang,S.E.Levinson and M.M.Sondhi, « recognition of isolated digits using Hiden Markov Models with continuous mixtures densities ». AT.and T tech.J,64(6):1211-1234, July-Aug 1985.

[25] Barbara Resch, « Automatic speech recognition with HTK, a tutorial for the course computational intelligence », http: //www.igi.tugraz.at/lehre/ CI

ANNEXE

Liste des figures **Pages**

ملخص:
شهد التعرف الآلي على النطق في هذه السنوات الأخيرة تطورا كبيرا، كشف الأخطاء هو احد التكنولوجيات التي ساعدت على تحسين أداء أنظمة التعرف على النطق، و خاصة كشف المفردات الخارجة.
إن استعمال قياسات ثقة مؤسسة على النموذج الصوتي يمكن أن تكون جد مفيدة في عدة تطبيقات التعرف الآلي على النطق.
إن فعالية نظام التعرف الآلي على النطق يمكن أن تطور بشكل كبير إذا كنا قادرين على التنبؤ إذا كان الاقتراح المقدم من طرف نظام التعرف الآلي صحيح أو لا.

في عملنا هذا أنشانا أولا نظام التعرف على الكلمات المعزولة باللغة العربية الذي يعتمد على منهج احتمالي (استعمال نماذج ماركوف المخفية « HMM_S »)
تمر آلية التعرف بمرحلتين أساسيتين :
1- مرحلة التمرن التي تقوم بضبط مقاييس نماذج « HMM_S » بواسطة كثافات خليط قوس و ذلك باستعمال خوارزمية بوم ولش.
2- مرحلة التعرف تقوم بحساب الاحتمال الأعظمي بين المشاهدة و النموذج و ذلك باستعمال خوارزمية فيتربي.
رغم نجاح هذا المنهج، إلا أن هذا الاحتمال الأعظمي لا يعبر عن حقيقة الصلة بين الكلمة المنطوقة والنموذج المرجعي.

ثانيا : يتعلق الأمر بتطبيق قياس ثقة باستعمال تقنية نمذجة القمامة المباشرة، هذه التقنية استعملت للحصول على نتيجة مرجعية لنتيجة التعرف.
الثقة معرفة على أنها الفرق بين نتيجة النموذج « HMM » و نتيجة نموذج القمامة .
من خلال النتائج المحصل عليها، هذه التقنية سمحت بكشف عدد كبير من خارج المفردات بدون رفض المفردات الصحيحة و ذلك باستعمال عتبة رفض صغيرة.
الكلمات المفتاح : التعرف على النطق – قياسات الثقة – نموذج القمامة.

Abstract :

The automatic speech recognition has known an important progress these last years, the errors detection is one of the technologies that has helped to improve the speech recognition systems performances notably the detection of out-of-vocabulary.
The use of confidence measures based on the acoustic model can be very useful for numerous applications of automatic speech recognition. The efficiency of a system of speech recognition can, in fact, be greatly improved if we are capable to predict if a proposed hypothesis by the RAP is correct or not.

In our work we have realized first of all a recognition system of isolated words in Arabic language based on a probabilistic approach (the use of MARKOV's hidden models HHM).
The process of recognition goes through two principle stages:

1- The training stage that consists to adjust the parameters of HMMs models with the mixtures Gaussians density using the algorithm of Baum-Welch.
2- The recognition stage that consists to calculate the maximum of likelihood (ML) between the observation and the model using the Viterbi algorithm.

Though this approach is successful, this maximum likelihood does not reflect the goodness of the match between the uttered word and the reference model.

Secondly, it deals with the application of a confidence measure using the on-line garbage modeling technique, this technique is used to obtain a reference score for the result of recognition, the confidence is defined as the difference between the score of HMM model and that of garbage model.
According to the obtained results, this technique has permitted to detect a large number of out-of-vocabularies without rejecting the correct recognitions and this using a small threshold of rejection.
Key words : speech recognition – confidence measures – garbage model..