الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي
جامعة وهران للعلوم و التكنولوجيا محمد بوضياف

THÈSE

En vue de l'obtention du Diplôme de Doctorat

Présenté par :MERAD-BOUDIA Nihal

Intitulé

Développement d'un système de reconnaissance de parole arabe pour des mots connectés en utilisant HTK

Faculté	***: Mathématiques et Informatique***
Département	***: Informatique***
Domaine	***: MI***
Filière	***: Informatique***
Intitulé de la Formation	***: Informatique***

Devant le Jury Composé de :

Membres de Jury	*Grade*	*Qualité*	*Domiciliation*
BENYETTOU Mohamed	*Professeur*	*Président*	*USTO-MB*
BENYETTOU Abdelkader	*Professeur*	*Encadrant*	*USTO-MB*
BOUGHANMI Nabil	*Professeur*		*USTO-MB*
TEMMAR Abdelkader	*Professeur*	*Examinateurs*	*INTIC36*
MEJDI Kaddour	*MCA*		*UNIV ORAN-1*

Année Universitaire : 2017—2018

Remerciements

Je remercie chaleureusement toutes les personnes qui m'ont aidé pendant l'élaboration de ma thèse et notamment mon directeur de thèse Monsieur le professeur BENYETTOU Abdelkader, pour son intérêt et son soutien, sa grande disponibilité et ses nombreux conseils durant la rédaction de ma thèse.

Je remercie Pr. BENYETTOU Mohamed qui nous a fait l'honneur de présider le Jury, ainsi que tous les autres membres du jury Pr. TEMMAR Abdelkader, Pr.BOUGHENMI Nabil et Dr. MEJDI Kaddour qui ont bien voulu considérer mon travail de thèse. .

Ce travail n'aurait pas été possible sans le soutien de l'Université des Sciences et de la Technologie d'Oran Mohammed Boudiaf, qui m'a permis, grâce à l'aide financière des stages de courtes durée qui m'ont aidé énormément dans mes recherches, et de me consacrer sereinement à l'élaboration de ma thèse.

Ce travail n'aurait pu être mené à bien sans la disponibilité et l'accueil chaleureux que m'a donné Monsieur RUBIO AYUSO Antonio dans son université de Grenade lors de mes stages passés en Espagne, l'aide précieuse que m'ont témoignée mes collègues du laboratoire de recherche « SIMPA » BENDAHMANE Abderrahmane et NEGGAZ Nabil enseignants et chercheurs scientifiques. Ainsi que mes collègues du laboratoire « ISI ». NAIR Ahmed Amine m'a fourni un document précieux pour avancer dans cette recherche aussi.

Au terme de ce parcours, je remercie enfin celles et ceux qui me sont chers et que j'ai quelque peu délaissés ces derniers mois pour achever cette thèse. Leurs attentions et encouragements m'ont accompagnée tout au long de ces années. Je suis redevable à mes parents, Zahira et Morsly, pour leur soutien moral et matériel et leur confiance indéfectible dans mes choix. Enfin, mon époux Ahmed EL KETROUSSI qui a accepté de me supporter ces trois dernières années, et qui m'a aidé à décompresser entre le temps de préparer cette recherche et rédiger cette thèse.

Résumé

Dans ce travail de thèse, nous décrivons et proposons une méthode efficace de reconnaissance de la parole arabe continue indépendante du locuteur, basée sur un corpus vocal phonétiquement riche. Ce corpus de parole contient deux ensembles de données : (1) les chiffres arabes parlés (SAD), enregistrés par 66 locuteurs (33 hommes et 33 femmes), prennent 6600 mots de chiffres dans lesquels chaque locuteur prononce chaque chiffre de zéro à neuf dix fois, et (2) : le second ensemble de données est restreint aux phrases coraniques de trois locuteurs célèbres avec les règles de Tajweed (une sorte de chant) des trente derniers chapitres (Sourate) du Saint Coran.

Nous traitons la problématique générale de la reconnaissance de la parole qui est la coarticulation avec les modèles dépendants du contexte : tri-phones comme modèle acoustique, et les bi-grams comme modèle de langage qui était le plus approprié dans ce texte. Nous traitons le problème particulier d'ajustement de la durée de son avec les tri-phones étendus aux modèles de mélanges Gaussiens (GMM).

Le système de reconnaissance de la parole arabe proposé est basé sur l'outil qui manipule les modèles de Markov cachés : HMM Toolkit (HTK) de l'université de Cambridge. Des tests expérimentaux montrent que l'ensemble de données de chiffres utilisant trois états émettant par phonème (son), a un excellent résultat de reconnaissance de mots qui est de 97,95 %. Le taux de reconnaissance des phrases est de 93,14 %.

Le meilleur résultat des versets coraniques obtenus est de 73,44 % de mots reconnus, et de 14,38 % de phrases reconnues. Un taux de 66,09% de précision chez le lecteur « Elmirigli ».

L'adaptation du système de base à la voix d'intonation du lecteur, puis l'application d'un GMM approprié à chaque tri-phone de liste liée, dépasse l'expérience du modèle acoustique HMM-triphone par un taux de mots reconnus de 16,93 % chez le lecteur « Alsudaissi » grâce au problème des voyelles de longue durée (appelée en arabe mudud) résolu.

Notre contribution est d'appliquer des tri-phones étendus aux modèles GMM qui est comparé à la méthode de Régression Linéaire à Ressemblance Maximum (MLLR), et de les dépasser de près de 5 % chez le lecteur « Alsudaissi ».

Mots clés : Reconnaissance de Parole Arabe, outil HTK, Modèle Acoustique, Dépendant du Contexte, Indépendant du Contexte, GMM.

Abstract

In this thesis, the authors describe and propose an efficient and effective method of speaker-independent continuous Arabic speech recognition method, based on a phonetically rich speech corpus. This speech corpus contains two datasets: (1) the Spoken Arabic Digits (SAD), recorded by 66 speakers (33 men and 33 women), holds 6600 digit words in which each speaker pronounces each digit from zero to nine ten times, and (2) the second dataset is restrained to Quranic sentences of three famous speakers with Tajweed rules (somehow of singing) of the last 30 chapters (Surat) of the Holy Quran. The problematic of co-articulation was treated in general with triphones as acoustic model and bigram as language model which was the most appropriate in this text; and the adjustment's problem of sound duration in particular was also treated, with triphones expanded to Gaussians mixtures models (GMM). The proposed Arabic speech recognition system is based on the Cambridge Hidden Markov Model (HMM) Toolkit (HTK) tools.

The Experimental tests show that the digit dataset using 3 emitting states per phone, has an excellent word recognition (WR) rate of 97.95 % and sentence recognition (SR) rate of 93.14 %. The best result of Quranic phrases obtained is 73.44 % of WR rate and 14.38 % of SR rate with 66.09 % of accuracy within "Elmirigli" reader. Adapting the basic system to the speaker's intonation voice, then applying an appropriate GMM to each tied-list triphone, it outperforms the Hidden Markov Models HMM-triphone acoustic model experiment by 16.93 % of WR rate within "Alsudaissi" reader thanks to the solved duration vowels (mudud) problem. The contribution of this work is to apply triphones expanded to Gmms method which was compared to Maximum Likelihood Linear Regression (MLLR) and beat it by almost 5 % in "Alsudaissi" reader.

Key words: Arabic Speech Recognition, Hmm Toolkit, Acoustic Model, Context Dependent, Context Independent, GMM.

Sommaire

Liste des abréviations

SAD	Spoken Arabic Digits
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTK	Hidden markov model ToolKit
WR	Word Recognition
SR	Sentence Recognition
MLLR	Maximum Likelihood Linear Regression
WER	Word Error Rate
HQ	Holy Quran
DTW	Dynamic Time Warping
PRNN	Pattern Recognition Neural Network
ANN	Artificial Neural Network
SVM	Support Vector Machine
FFBPNN	First Forward Back Propagation Neural Network
HSMM	Hidden Semi Markov Model
LVQ	Learning Vector Quantization
ML	Maximum Likelihood
DBN	Dynamic Bayesian Network
EM	Expectation Maximization
MSA	Modern Standard Arabic
WCR	Word Correct Recognition
MLP	MultiLayer Perceptron

Liste des figures

Liste des tables

Chapitre I

Introduction

Ce chapitre procède aussi bien à l'introduction de la thèse et met l'accent sur la motivation du sujet de la thèse, qu'il dévoile les problèmes de reconnaissance automatique de parole pour la langue arabe.

Dans ce chapitre, le cadre général et les objectifs souhaités de cette thèse sont clarifiés. Le chapitre revoie brièvement les solutions suggérées et développées dans cette thèse et découle le sommaire du contenu des sections restantes.

I.1 Motivation et aperçu de la thèse

La recherche dans la reconnaissance de la parole et la communication pour une partie majeure, était motivée par le désir des gens de construire des modèles mécaniques afin d'imiter les capacités de communication verbales humaines. La parole est la chose la plus naturelle de la communication de l'humain et le traitement de la parole a été l'un des domaines passionnants du traitement de signal.

La technologie de reconnaissance de parole l'a rendu possible pour la machine afin de suivre les commandes de voix humaine et comprendre les langages humains. En se basant sur des démarches majeures de la modélisation statistique de la parole, les systèmes de reconnaissance automatique de la parole ont trouvé aujourd'hui des applications très répandues dans des tâches demandant l'interface homme machine telles que : traitement de l'appel dans les réseaux téléphonique, accès aux informations : voyage, bancaire, commandes avioniques, portail automobile, transcription de la parole, personnes handicapés (personnes aveugles) dans les supermarchés etc. malgré de nombreux progrès technologiques qui ont été faits, il reste beaucoup de recherches qui ont besoin d'être abordées (Martin and Jurafsky 2000).

Même s'il y a diverses applications de RAP dans le domaine commercial, les applications dans la langue arabe sont vraiment limitées. L'arabe n'existe pas beaucoup dans des applications logicielles qui dépendent de RAP et de traitement de langage naturel (TLN). De plus, cela intéresse un grand groupe d'utilisateurs.

La motivation centrale de cette thèse est de conduire la recherche dans des aspects pratiques de la reconnaissance automatique de la parole arabe, de soutenir la langue arabe dans la nouvelle ère digitale visionnaire et de provenir les solutions de traductions les plus modernes de langue arabe qui s'appliquent à toutes les tâches opérationnelles. Lecture du coran (Yekache, Mekelleche et al. 2012).

I.2 Objectifs et questions de recherches

Cette thèse consiste à construire un système de reconnaissance de mots isolé au départ, puis des mots connectés en Arabe et les valider sous l'environnement HTK (Hidden Markov Models ToolKit). La portée de cette recherche est basée sur les propriétés suivantes (Anusuya and Katti 2010) :

Reconnaissance de mots isolés et connectés : La reconnaissance de mots isolés demande que chaque parole ait un silence (un manque de signal audio) dans les deux

extrémités de la fenêtre échantillon. L'objectif est d'accepter des mots uniques ou de parole unique à la fois. Ces systèmes ont les états « écoute / non écoute » où ils demandent que l'utilisateur attende entre les élocutions (généralement traitement durant les pauses).

Les systèmes de mots connectés (ou plus correctement des paroles connectés) sont similaires aux mots isolés, mais permettent aux paroles séparées d'être exécutées ensemble avec une pause minime entre elles.

Taille du vocabulaire moyen : la taille du vocabulaire dans les systèmes de reconnaissance de parole affecte la complexité, les demandes de traitements et l'exactitude du système. La tâche provient une meilleure indication dans l'application pratique du système performant lorsque sa complexité est limitée.

Système indépendant du locuteur : ce système est développé pour opérer n'importe quel locuteur. La reconnaissance indépendante du locuteur est plus difficile car la représentation interne de la parole doit être suffisamment globale pour couvrir tous les types de voix et toutes les prononciations possibles des mots.

Modèles statistiques pour les RAP : les modèles statistiques sont des méthodes classiques pour la reconnaissance de la parole, et sont des contributeurs majeurs et employés fondamentalement pour plusieurs applications récentes de reconnaissance de formes.

- **Contribution originale**

Dans ce premier chapitre, il est important de parler des bases de données utilisées avec la problématique associée, ensuite proposer une solution à cette problématique qui est notre contribution :

Notre travail consiste à reconnaitre des mots connectés en arabe. Nous testons la base de données de chiffres arabes « Spoken Arabic Digits » appelé (SAD). Cet ensemble de données donne de très bons résultats lors de l'utilisation de l'approche basée sur les modèles de Markov cachés pour chaque phonème du mot représenté par trois états, alors que cette approche donne des résultats vraiment médiocres sur notre deuxième base qui est les versets des 30 derniers chapitres (Surat en arabe) du saint coran.

De ce fait, des problématiques sont soulevées :

La première problématique est plus ou moins courante, celle de la coarticulation, c'est-à-dire : l'anticipation de la prononciation d'un phonème avant celui qui le précède crée ce phénomène. Une solution est donnée, c'est l'application de HMM tri-phonèmes cette fois-ci pour pallier aux défauts de coarticulations. Cette technique a malheureusement régressé les résultats, et c'est due au manque de données pour que les HMM tri-phones apprennent bien les sons.

C'est là, que nous introduisons la deuxième problématique qui est assez particulière, lorsque nous avons à faire avec de célèbres lecteurs qui lisent le coran en suivant les règles de Tajweed, nous remarquons que chez le lecteur « alsudaissi », les phonèmes sont les plus difficiles à reconnaitre, par rapport aux lecteurs « alghamidi » et « elmirigli ». En faisant la comparaison avec un autre travail similaire au notre, nous avons noté que les auteurs avaient utilisé une autre méthode pour améliorer les résultats. Leurs résultats étaient meilleurs que les notre pour tous les lecteurs, mis à part pour « alsudaissi ».

D'où notre contribution proposée : après avoir entraîné les HMMs tri-phones, il faut lier les phonèmes qui se ressemble par une gaussienne uni-variée, puis les étendre aux mélanges Gaussiens ou appelé gaussiennes multi-variées. Ainsi les résultats s'amélioraient au fur d'augmenter le nombre de GMMs. Nous avons obtenu les meilleurs résultats pour un nombre de GMM égal à 512.

Nous verrons dans le chapitre 4 l'algorithme détaillé.

I.3 Structure de la thèse

Après le chapitre de l'introduction, nous avons organisé notre thèse en deux parties essentielles.

La première étant la partie théorique, qui comporte l'état de l'art exprimé par idée. Cette partie contient deux chapitres basiques : un pour la reconnaissance de parole arabe, l'autre pour l'approche basée sur les HMMs.

La deuxième partie est celle de la pratique, comportant un seul chapitre, celui des expérimentations, ainsi que l'approche amélioré par les GMMs

Nous finissons cette thèse par la conclusion et perspectives.

Dans le deuxième chapitre, nous discuterons sur la reconnaissance de la parole en général, nous aborderons les méthodes qui ont été utilisées durant ces dix dernières années et nous nous focalisons sur les applications des systèmes de RAP. Nous décrivons dans le chapitre (2) la structure des systèmes de RAP, l'évaluation de leurs performances ainsi que les méthodes d'extractions de caractéristiques, ainsi que les travaux connexes sur les applications de la reconnaissance de parole en général. Puisque la question de recherche à laquelle s'attaque cette thèse se situe dans la langue arabe comme nous avons vu dans la section précédente, alors la langue arabe et les systèmes de RAP arabe sont discutés, puis les travaux connexes sur la reconnaissance de parole arabe sont donnés dans le chapitre (2).

Ces applications sont : la synthèse de la parole à partir du texte, la reconnaissance de locuteur, la reconnaissance de la parole enregistrée de plusieurs interlocuteurs simultanément, l'identification du locuteur pour l'amélioration de la reconnaissance de la parole, la détection magnétique de mouvements articulatoires utile pour les personnes atteintes de maladies qui affectent les cordes vocales, affichage du texte correspondant à

la parole prononcée par la machine, l'adaptation du locuteur discriminative dans la reconnaissance de parole continue Persane, l' identification des sons qui stoppent la glotte dans la parole continue Amharique, la contribution de formants et des caractéristiques prosodiques dans la reconnaissance de l'arabe.

Dans le chapitre trois, nous proposons une solution pour un problème classique tel que la reconnaissance des chiffres : les modèles de Markov cachés au niveau du phonème. De plus, différentes lectures du Coran de différents célèbres lecteurs soulèvent le problème de la coarticulation en général, mais aussi un autre problème particulier qui est la durée du son en temps due aux règles de Tajweed appliquées par ces célèbres lecteurs. Nous proposons d'utiliser tout simplement pour la difficulté de coarticulation, les modèles de Markov cachés au niveau de tri-phones, et pour le problème de durée de voyelles appelé en arabe « mudud », nous proposons d'utiliser les tri-phones liés puis étendu aux modèles de mélange de Gaussien (GMM). La raison pour laquelle cette problématique valle la peine d'être posée est la détection coranique indépendante du locuteur. Nous verrons aussi les travaux connexes sur les méthodes appliqués à la reconnaissance de la parole selon la taille du vocabulaire.

Ces modèles sont : le modèle de déformation temporelle dynamique (DTW) et le réseau neuronal de reconnaissance de forme (PRNN) pour vérifier la similarité entre les phonèmes arabes, la combinaison de modèles acoustiques graphémiques et phonétiques pour la reconnaissance de parole arabe à grand vocabulaire.

Ce chapitre quatre confirmera notre théorie présenté dans le chapitre précédent en l'évaluant selon des critères que nous verrons et validant par des chiffres en faisant une étude comparatives avec deux travaux antérieurs, sans oublier de confirmer cette hypothèse avec des statistiques ainsi que des graphiques d'évaluation. Nous introduisant aussi l'outil HTK qui nous sert à construire et manipuler les modèles HMM. HTK est un ensemble de bibliothèques et d'outils valable en source C.

Il y a une vue générale sur les principaux résultats répondant aux questions posées : sur la base des chiffres, nous avons eu un résultat de 97.95% pour la reconnaissance de mots et 93.14% pour les phrases. Sur la base du coran, le meilleur résultat obtenu est de 73,44% de mots reconnus, et de 14,38% de phrases reconnues, et c'est chez le lecteur « Elmirigli ». Le taux de reconnaissance de mot avec notre méthode chez le lecteur « Alsudaissi » dépasse de 5% de celui d'un autre travail qui a utilisé la méthode de régression linéaire à vraisemblance maximum.

Pour terminer, dans le chapitre cinq qui est la conclusion générale de cette thèse, nous verrons les limitations de notre méthode pour en décrire des idées de futures recherches. Ces limites se résument en : le taux de reconnaissance de mots du système indépendant du locuteur qui reste abaissé, pour cela il suffit d'augmenter le nombre de données ainsi que le nombre de locuteurs. La deuxième limite est que le taux de phrase de tous les systèmes est abaissé, ceci peut être résolu en essayant des modèles acoustiques basés sur les

classificateurs postérieurs tels que les réseaux de neurones et les machines à vecteur de support.

Chapitre II

Reconnaissance Automatique de la Parole Arabe

II.1. Introduction

Le but de ce chapitre est de provenir des informations générales sur la reconnaissance automatique de parole (RAP), et déclarer concisément la question de recherche à laquelle s'attaque cette thèse qui se situe sur la langue arabe. La question est comment calculer la durée de son qui se maintient dans le temps et qui change selon certaines règles dans le but de reconnaitre ce son.

Nous décrivons dans ce chapitre la structure des systèmes de RAP, l'évaluation de leurs performances ainsi que les méthodes d'extractions de caractéristiques, sans oublier les travaux connexes sur les applications de la reconnaissance de parole en général.

La langue arabe et les RAP arabe sont discutés, puis les travaux connexes sur la reconnaissance de parole arabe sont donnés en détaillant leurs résultats.

II.2. La revue de littérature de la reconnaissance de la parole

Nous présentons quelques travaux qui sont faits depuis 1920 jusqu'à aujourd'hui:

Dès les débuts des années 1920s la reconnaissance machine a connu son existence. La première machine pour reconnaitre la parole commercialement nommée Radio Rex (jouet) a été fabriquée en 1920 (Windmann & Haeb-Umbach, 2009).

Dans les années 1960s, différentes idées dans la reconnaissance de parole a fait surface et a été publié. Puisque les ordinateurs n'étaient pas assez rapides, beaucoup de buts matériels ont été construits sur un système Japonais décrit par Suzuki et Nakata du laboratoire Radio Research à Tokyo, était un reconnaisseur matériel de voyelle(Sakai & Doshita, 1962).

Dans les années 1970s, la recherche reconnaissance de la parole a achevé un nombre significatif d'étapes. La première, le domaine des mots isolés ou reconnaissance de locution discrète est devenue une technologie utilisable basée sur des études fondamentales par Velichko et Zagoruyko en Russie(Velichko & Zagoruyko, 1970). Autre domaine était couronné de succès dans la reconnaissance de parole à grand vocabulaire à IBM.

Tout comme la reconnaissance de mots isolé qui était un élément clé des années 70s, les problèmes de reconnaissance de mots connectés était l'objet de recherche dans les années 1980s(Moore, 1994). Parmi les technologies développées : l'approche modèle de Markov cachés (HMM) qui est un double processus stochastique de sorte que le processus stochastique sous-jacent n'est pas observable (d'où le terme caché) mais qui peut être observé par un autre processus stochastique qui produit une séquence d'observations, aussi les réseaux de neurones artificiels (RNA). Les réseaux de neurones (RN) ont été introduits d'abord dans les années 1950s, mais ils ne s'avéraient pas utiles à cause de problèmes pratiques. Plus tard dans les années 1980s, une profonde compréhension des forces et limites de la technologie a été achevé. Enfin les années 80s était une décennie dans laquelle un grand élan a été donné aux systèmes de reconnaissance de parole

continue par la communauté DARPA (Defense Advance Research Project Agency). Les majeures contributions de recherches ont résulté des systèmes SPHINX(Lee, Hon, & Reddy, 1990).

Dans les années 1990s, un nombre d'innovations ont pris place dans le domaine de reconnaissance des formes. Le problème de reconnaissance de forme, qui est traditionnellement suivi du système Bays et demande l'estimation de distribution pour les données, a été transformé en un problème d'optimisation impliquant la minimisation de l'erreur de reconnaissance. Ce concept a produit un nombre de techniques utilisant les critères : erreur de classification minimale (MCE) et informations mutuelles maximum (MMI)(Li & Hughes, 1974). Ces deux critères penchent vers la ressemblance maximum (ML).

Autour des années 2000, des techniques d'estimation Bayesiènne variationnelle (VB) et de clustering ont été développées. Contrairement à la ressemblance maximum (ML), l'approche VB est basée sur une distribution de paramètres postérieure. En 2007, les auteurs (Lui, 2007) ont proposé une nouvelle méthode d'optimisation « programmation semi-définie » pour résoudre le problème d'estimation de grande marge des HMMs de densité continue dans la reconnaissance de la parole.

II.3. Les approches de reconnaissance de la parole

Il existe trois approches à la reconnaissance de la parole fondamentalement, qui sont :

- L'approche phonétique acoustique

Elle est basée sur des sons de paroles étiquetés selon leurs propriétés acoustiques dans le signale de parole. Cette technique n'a pas été utilisée dans des applications commerciales.

- L'approche reconnaissance de formes

Le principe est l'apprentissage des formes puis la comparaison de celles là est basée sur la représentation de la parole soit par : le template (1970 : plusieurs répétitions du mot par le même locuteur) ou par un modèle stochastique (1980 et 1990 : qui utilise un modèle probabiliste comme les HMM à état finis (Rabiner 1989), le Déformation de Temps Dynamique (DTW), Machine à Vecteurs de Soutien (SVM) et (ANN).

- L'approche intelligence artificielle

Elle est basée sur l'extraction de connaissance à partir de règles. L'inconvénient de cette approche est la difficulté de l'analyse de l'erreur lors de tentative d'améliorer la performance de système HMM (Moore 1994).

II.4. Reconnaissance automatique de la parole (RAP)

II.4.1. Représentation mathématique de RAP

L'approche standard de reconnaissance de parole continue à grand vocabulaire est de supposer un simple modèle probabiliste de production de parole par lequel une séquence de mots spécifique W, produit une séquence d'observations acoustiques A avec la probabilité P(W,A). Le but est donc de décoder la chaine de mots, basée sur la séquence d'observation acoustique, de sorte que la chaine décodée a la probabilité maximum à postériori (MAP).

$$P(W/A) = \text{argmax}_w P(W/A) \quad \text{..... (II.1)}$$

En utilisant la règle de Bays, l'équation (1) peut être écrite ainsi :

$$P(W/A) = \frac{P(A/W)P(W)}{P(A)} \quad \text{..... (II.2)}$$

Comme P(A) est indépendant de W, la règle de décodage MAP de l'équation (II.1) est :

$$W = \text{argmax}_w P(A/W)P(W) \quad \text{.....(II.3)}$$

Le premier terme dans l'équation (II.3) P(A/W), est généralement appelé le modèle acoustique (MA) vue qu'il estime la probabilité de séquence d'observations acoustiques, conditionné sur la chaine de mots ; d'où P(A/W) est calculée. Pour des systèmes de reconnaissance de parole de grand vocabulaire, il est nécessaire de construire de modèles statistiques pour des unités de parole sous-mots, accumuler ces modèles là en utilisant un lexique pour décrire la composition de mots, puis postuler les séquences de mots et évaluer les probabilités de modèles acoustique via les méthodes de concaténation standards. Le second terme dans l'équation (II.3) P(W), est appelé modèle de langages. Il décrit la probabilité associée à la séquence de mots postulée. De tels modèles de langages peuvent incorporer des contraintes syntaxique et sémantique du langage et de la tâche de reconnaissance(Anusuya & Katti, 2010).

II.4.2. Structure de RAP

La figure (II.1) montre une représentation mathématique du système de reconnaissance de la parole dans des équations simples expliquées dans la section précédente.

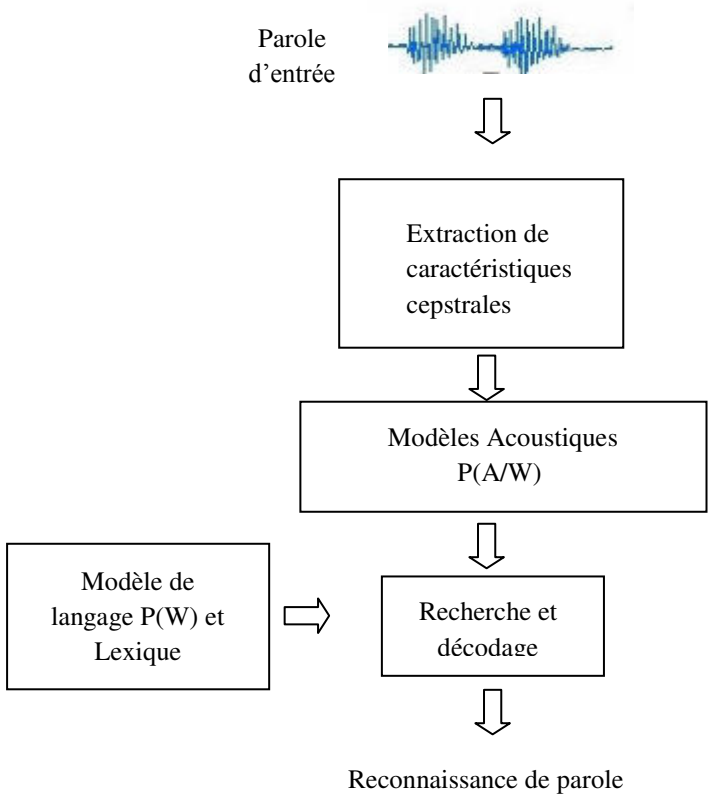


Figure II.1 Modèle basique de reconnaissance de parole.

II.4.2.1. Extraction de caractéristiques : vecteurs MFCC

Notre but dans cette sous-section est de décrire comment transformer la forme d'onde entrée en une séquence de vecteurs de caractéristiques acoustiques, chaque vecteur représentant l'information en un petit nombre de temps du signal. Comme il y a beaucoup de possibilités de représentation de caractéristique par la plus commune dans la reconnaissance de la parole est les MFCC, les Coefficients Cepstraux de Fréquences Mel. Ceux-ci sont basés sur l'idée importante du cepstre. Nous allons donner une description relativement haut-niveau du processus d'extraction de MFCC de la forme d'onde dans la figure (II.2).

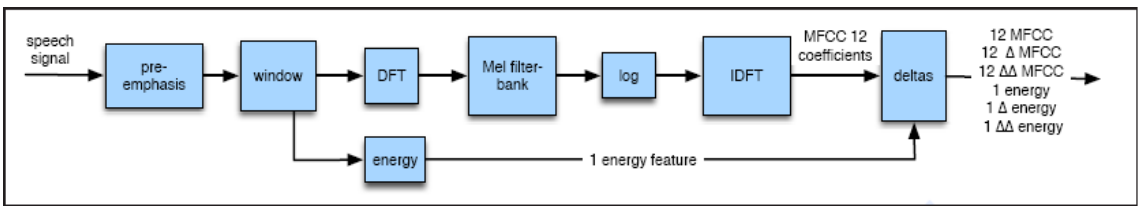


Figure II.2 Extrait de vecteur de caractéristique MFCC de dimension 39 d'une forme d'onde quantifiée et numérisée. (Martin & Jurafsky, 2000)

Le processus commence par la numérisation et la quantification de la forme d'onde de la parole. Rappelons que la première étape dans le traitement de parole est de convertir la représentation analogique à un signal numérique. Le processus de conversion *analogique-numérique* a deux étapes : *échantillonnage* et *quantification*. Un signal est échantillonnée en mesurant son amplitude à un certain temps ; le *taux d'échantillonnage* et le nombre d'échantillons pris par seconde. Dans le but de mesurer une onde avec précision, il est nécessaire d'avoir au moins deux échantillons dans un cycle : un mesurant la partie positive et l'autre mesurant la partie négative.

On référencie chaque échantillon dans la forme d'onde numérisée quantifiée par un *x[n]*, où *n* est l'indice à travers le temps. Maintenant qu'on a une représentation numérisée, quantifiée de la forme d'onde, nous sommes prêts à extraire les caractéristiques MFCC. Les sept étapes du processus sont montrées dans la figure 2 et décrit individuellement dans les sous-sections suivantes :

- **Pré-emphase**

La première phase dans l'extraction de caractéristique MFCC est de booster la quantité d'énergie dans les hautes fréquences. Cela revient à voir les segments voisés du spectre comme les voyelles, il y a plus d'énergie dans les basses fréquences que dans les hautes fréquences. Cette chute en énergie à travers les fréquences est causée par la nature de l'impulsion de la glotte. Augmenter l'énergie des hautes fréquences met l'information de ces hauts formants plus valable au modèle acoustique et améliore la justesse de détection de phone.

Cette pré-phase est faite par l'utilisation de filtre. La figure 3 montre un exemple d'une tranche du spectre de la prononciation de l'auteur (Martin & Jurafsky, 2000) de la voyelle [aa] avant et après la pré-phase.

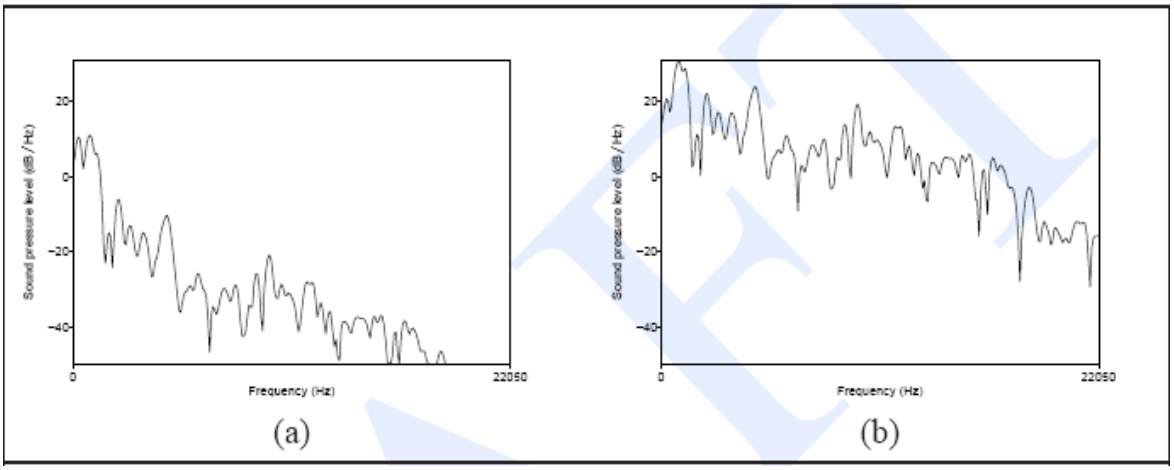


Figure II.3 Morceau spectral de la voyelle [aa] avant la pré-emphase (a), et après (b), (Martin & Jurafsky, 2000).

- **Fenêtrage**

Rappelons que le but de l'extraction de caractéristique est de fournir les caractéristiques spectrales qui peuvent nous aider à construire les classifieurs de phones ou sous-phones. Nous ne voulons pas par conséquent extraire nos caractéristiques spectrales d'une locution entière ou conversation, à cause des changements du spectre qui est vraiment rapide. Techniquement, on dit que la parole est un signal *non-stationnaire*, ce qui veut dire que ses propriétés statiques ne sont pas constantes à travers le temps. Au lieu de cela, nous voulons extraire les caractéristiques spectrales d'un petit nombre de *fenêtres* de la parole qui caractérisent un sous-phone particulier et pour lequel on peut mettre l'hypothèse approximative que le signal est *stationnaire* (i.e. ses propriétés statistiques sont constantes selon cette région).

Nous allons faire ceci par l'utilisation d'une fenêtre qui est une non-zéro dans certaines régions et zéro autre part, exécutant cette fenêtre à travers le signal de parole, et l'extraction de la forme d'onde à travers cette fenêtre.

Nous pouvons caractérisée un tel processus de fenêtrage par trois paramètres : combien *large* est la fenêtre (en millisecondes), quelle est la *compensation* entre les fenêtres successives, et quelle est la *forme* de la fenêtre. Nous appelons la parole extraite de chaque fenêtre *frame*, et on appelle le nombre de millisecondes dans le frame *taille de frame,* et le nombre de millisecondes entre les bords gauches des fenêtres successives *le décalage du frame*.

L'extraction du signal prend place par multiplier la valeur du signal à un temps *n, s[n],* avec la valeur de la fenêtre à un temps *n, w[n]* :

$$y[n]=w[n]s[n] \quad \text{(II.4)}$$

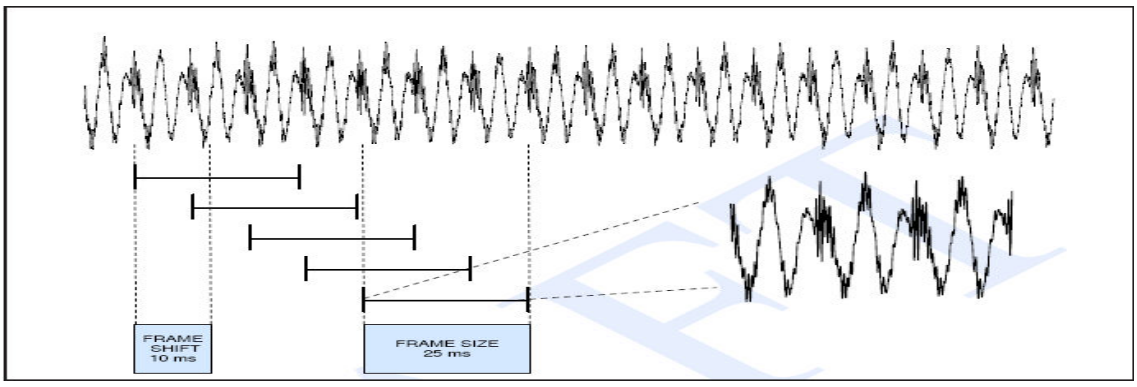


Figure II.4 Le processus fenêtrage, montrant le décalage de la frame et la taille de la frame, supposant un décalage de 10 ms, une taille de frame de 25 ms, et une fenêtre rectangulaire(Martin & Jurafsky, 2000).

La figure (II.4) suggère que ces formes de fenêtre sont rectangulaire, puisque le signal fenêtré extrait ressemble au signal original. En effet, la fenêtre la plus simple est la fenêtre *rectangulaire*. Par contre, la fenêtre rectangulaire peut causer des problèmes à cause de sa coupure du signal brusque en ses limites. Ces discontinuités créent des problèmes quand on fait une analyse Fourier. Pour cette raison, une fenêtre plus commune utilisée dans l'extraction de MFCC est la fenêtre de *hamming,* qui rétrécit les valeurs du signal vers zéro aux limites de la fenêtre, évitant les discontinuités.

Rectangulaire $\quad w[n]=\begin{cases}1 & 0 \le n \le L-1 \\ 0 & sinon\end{cases}$ …… (II.5)

La prochaine étape est d'extraire l'information spectrale pour notre signal fenêtré ; nous avons besoin de savoir combien le signal contient d'énergie aux différentes bandes de fréquences. L'outil pour l'extraction de l'information spectrale pour les bandes de fréquences discrètes pour un temps discret échantillonné est la *Trans*formée de Fourier Discrète ou TFD.

L'entrée à la TFD est un signal fenêtré $x[n]\ldots.x[m]$, et la sortie pour chaque N bande de fréquences discrètes, est un nombre complexe $X[k]$ représentant la phase d'amplitude de cette composante fréquentielle dans le signal original. Si on fait le graphe de l'amplitude par rapport à la fréquence, on peut visualiser le spectre. Par exemple, la figure (II.5) montre une portion du signal fenêtré-Hamming de 25 ms avec son spectre calculé par une TFD (avec un lissage ajoutée).

Nous n'allons pas introduire les détails mathématiques sur la TFD ici, excepté que l'analyse de Fourier en général compte sur la *formule d'Euler*

$$e^{j\theta} = \cos\theta + j\sin\theta \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(II.6)}$$

Comme un petit rappel, la TFD est définie comme suit :

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\frac{\pi}{N}kn} \quad \ldots\ldots\ldots\ldots\ldots \text{(II.7)}$$

Un algorithme communément utilisé pour le calcul de la TFD est le *Fast Fourier Transform*ou*FFT*. Cette implémentation de la TFD est vraiment efficace, mais marche seulement pour les valeurs de N qui sont des puissances de deux.

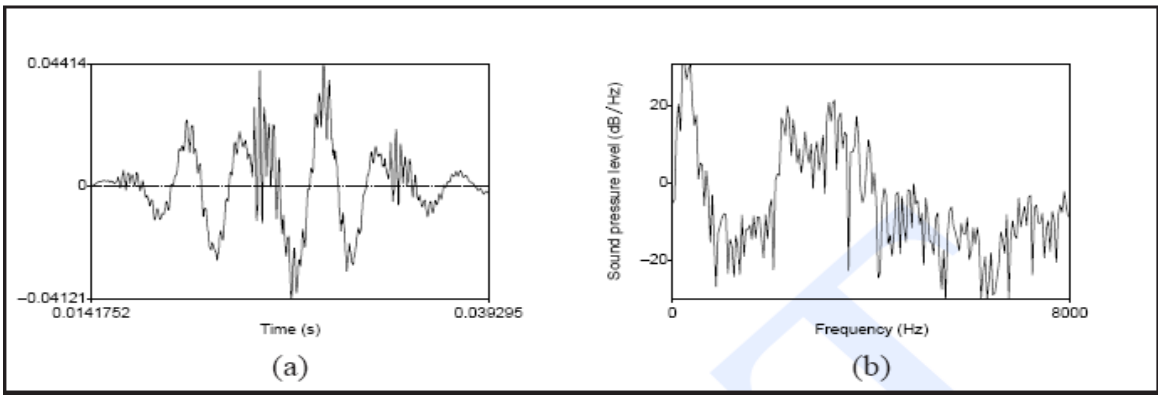


Figure II.5 (a) une portion du signal de la voyelle [iy] fenêtré-Hamming de 25 ms, et (b) est son spectre calculé par une TFD. (Martin & Jurafsky, 2000)

- **Banc de filtre Mel et log**

Le résultat de la FFT va être l'information sur la quantité d'énergie à caque bande de fréquence. L'ouïe humaine par contre, n'est pas sensitive pareillement à toutes les bandes de fréquences. Il est moins sensitif aux fréquences plus hautes, approximativement au dessus de1000 Hertz. Cela revient à modéliser cette propriété de l'ouïe humaine durant l'extraction de caractéristique à améliorer la performance de la reconnaissance de parole. La forme du modèle utilisé dans les MFCCs est de balayer les fréquences en sortie de la TFD sur l'échelle *mel.* Un *mel* (Stevens et al., 1937 ; Stevens et Volkman, 1940) est l'unité du pitch défini tel que les paires de son qui sont perceptuelle, équidistantes dans le pitch sont séparés par un même nombre de mels. La correspondance entre les fréquences en Hertz et l'échelle melesr linéaire en dessous e 1000 Hz et logarithmique au-dessus de 1000 Hz. La fréquence mel*m* peut être calculée de la fréquence acoustique brute suivante :

$$mel\ (f) = 1127\ \ln(1+\frac{f}{700}) \quad \ldots\ldots \text{(II.8)}$$

Durant le calcul MFCC, cette intuition est implémentée par la création d'un banc de filtres qui collecte l'énergie de chaque bande de fréquences, avec 10 filtres espacés linéairement en dessous de 1000 Hz, et le reste des filtres se diffuse au-dessus de 1000 Hz.

Enfin, on prend le log de chaque valeur du spectre mel. En général la réponse humaine au niveau du signal est logarithmique ; les humains sont moins sensibles aux différences négligentes dans l'amplitude dans les hautes amplitudes que dans les basses. En plus, utilisant un log fait en sorte que la caractéristique estime moins sensiblement aux variations en entrée (par exemple les variations due au mouvement de la bouche du locuteur près ou loin du microphone). (Martin & Jurafsky, 2000)

- **Le Cepstre : Tranformée de Fourier Discrète Inverse**

Tant qu'il est possible d'utiliser le spectre mel par lui-même en tant que caractéristique de représentation pour la détection de phone, le spectre a aussi ses problèmes, comme nous le verrons. Pour cette raison, la prochaine étape de l'extraction de caractéristique de MFCC est le calcul du *cepstre*. Le cepstre a un nombre d'avantages de traitement utile et aussi améliore significativement la performance de la reconnaissance des phones.

Une manière de songer au spectre est une façon utile de séparer la *source* et le *filtre*. Rappelons que la forme d'onde de la parole est crée quand la forme d'onde de la source glottale d'une certaine fréquence fondamentale qui est passée à travers le conduit vocal, qui à cause de sa forme, a une caractéristique de filtrage particulier. Mais plusieurs caractéristiques de la *source* glottale (sa fréquence fondamentale, les détails d'impulsions glottales, etc) ne sont pas importantes pour distinguer les différents phones. Plutôt, l'information la plus utile pour la détection de phone est le *filtre*, i.e. la position exacte du conduit vocal. Si nous connaissons la forme du conduit vocal, nous voudrions savoir quel phone a été produit. Ceci suggère que des caractéristiques utiles pour la détection de phone doivent trouver une façon de séparer la source et le filtre et nous montrer seulement le filtre du conduit vocal. Cela veut dire que le cepstre est le seul moyen pour le faire.

Pour faire simple, ignorons la pré-phase et le balayage-mel qui sont des parties de la définition des MFCCs, et voyons seulement la définition basique du cepstre. En bref, le cepstre peut être réfléchi tel que le spectre du log du spectre.

La figure (II.6) montre le *cepstre* (le mot *cepstre* est formée de l'inversement des premières lettres du *spectre*) est montré avec des échantillons le long de l'axe x. ceci parce qu'en prenant le spectre du log du spectre, nous avons laissé le domaine fréquentiel du spectre, et retourné au domaine temporel. Cela revient à dire que l'unité correcte du cepstre est l'échantillon.

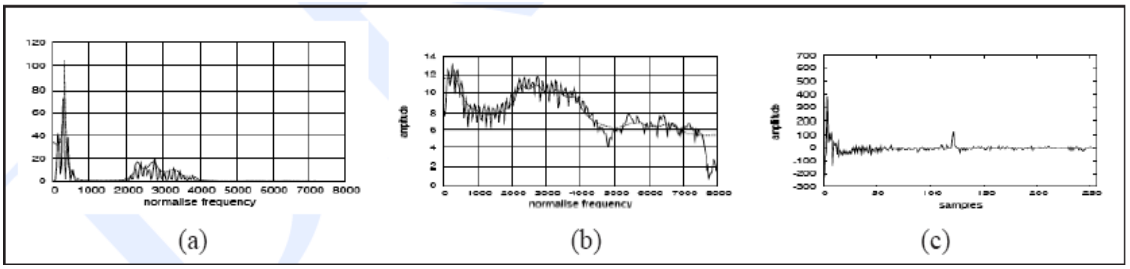


Figure II.6 Amplitude du spectre (a), le log de l'amplitude du spectre (b), et le cepstre (c). (Taylor, 2009)

Pour objet d'extraction de MFCC, on prend généralement les 12 premières valeurs cepstrales. Ces 12 coefficients vont représenter l'information uniquement à propos du filtre du conduit vocal, proprement séparé de l'information à propos de la source glottale.

Cela revient à dire que les coefficients cepstraux ont l'extrême propriété utile que la variance des différents coefficients ont tendance à être dé-corrélés. Ceci n'est pas vrai pour le spectre, où les coefficients spectraux aux différentes bandes de fréquences sont corrélés. Le fait que les caractéristiques cepstrales sont disjointes veut dire, comme nous allons le voir dans la prochaine section, que le modèle acoustique Gaussien (le modèle de mélange Gaussien ou GMM) n'a pas besoin de représenter la covariance entre toutes les caractéristiques MFCC, ce qui réduit largement le nombre de paramètres.

Pour les étudiants qui ont eu le traitement du signal, le cepstre est plus défini formellement tel qu'une *TFD inverse du log de l'amplitude de la TFD du signal*, donc pour un frame de la parole $x[n]$:

$$c(n) = \sum_{n=0}^{N-1} \log(|\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}|)\, e^{j\frac{2\pi}{N}kn} \text{.....(II.9)}$$

- **Deltas et l'énergie**

L'extraction du cepstre via la TFD inverse des résultats de la précédente sous-section dans les 12 coefficients cepstraux pour chaque frame. Nous ajoutons ensuite une treizième caractéristique : l'énergie du frame. L'énergie corrèle avec identité du phone puis c'est une réponse utile pour la détection de phone (les voyelles et les sifflantes ont plus d'énergie que les stoppantes, etc). *L'énergie* dans un frame est la somme au fil du temps de puissance des échantillons dans un frame ; ainsi pour un signal *x* dans une fenêtre d'un temps échantillon t_1 à un temps échantillon t_2, l'énergie est :

$$Energie = \sum_{t=t1}^{t2} x^2[t] \text{.....(II.10)}$$

Un autre fait important à propos du signal de parole est qu'il n'est pas constant d'un frame à un frame. Cela change, tel que le seuil du formant à ses transitions, ou la nature du changement de la clôture de la stop au relâchement de la stop, peut fournie une bonne réponse pour l'identité de phone. Pour cette raison aussi, nous ajoutons des caractéristiques liées aux caractéristiques cepstrales au fil le temps.

Nous faisons cela en ajoutant aux 13 coefficients un *delta* ou une caractéristique *vitesse*, une caractéristique *double delta* ou *accélération.* Chacune des 13 caractéristiques deltas représente le changement entre les frames dans la caractéristique cepstrale/énergie, tandis que les 13 doubles deltas représentent le changement entre les frames dans les caractéristiques deltas correspondants.

Une simple façon de calculer les deltas devrait être juste de calculer la différence entre les frames ; donc la valeur delta *d(t)* pour une valeur cepstrale particulière *c(t)* à un temps *t* peut être estimé tel que :

$$D(t) = \frac{c(t+1) - c(t-1)}{2} \text{ ..... (II.11)}$$

A la place de cette simple estimation, il est plus commun de faire des estimations plus sophistiqués du seuil, en utilisant un contexte plus large de frames(Martin & Jurafsky, 2000).

II.4.2.2. Modèles acoustiques

La sous section précédente a montré comment on peut extraire les caractéristiques MFCC représentant une information spectrale d'une forme d'onde, et produire un vecteur 39-dimensionnel chaque 10 millisecondes. Nous sommes prêts maintenant de calculer la ressemblance de ces vecteurs caractéristiques sachant un état HMM. Rappelons que [jurafsky] cette ressemblance en sortie est calculée par la fonction de probabilité B du HMM. Etant donné un état individuel qi et une observation ot, les ressemblances d'observation dans la matrice B nous a donné p(ot|qi), que nous avons appelé bt(i).

Pour l'étiquetage de partie-de-parole, chaque observation ot est un symbole discret (un mot) et on peut calculer la ressemblance d'une observation ayant une étiquette de partie-de parole juste en comptant le nombre de fois que génère une étiquette une observation donnée dans l'ensemble d'apprentissage. Mais pour la reconnaissance de parole, les vecteurs MFCC sont des nombres en réel ; on ne peut pas calculer la probabilité d'un état donné (un phone) générant un vecteur MFCC en comptant le nombre de fois où chaque vecteur se produit (puisque chacun est unique).

Dans le décodage et l'apprentissage, on a besoin d'une fonction de probabilité d'observation qui calcule p(ot|qi) sur les observations en valeurs réels. Dans le décodage, on compte tenu d'une observation ot nous avons besoin de produire la probabilité p(ot|qi) pour chaque état de HMM possible, donc on peut choisir la séquence d'états la plus probable. Une fois on a cette fonction de probabilité d'observation B, nous verrons dans le chapitre III l'algorithme de Viterbi de décodage de la parole reconnu par le système.

Nous verrons en détail comment appliquer un modèle de Markov caché à la reconnaissance de parole dans le chapitre III (Martin & Jurafsky, 2000).

II.4.2.3. Modèles de langage et lexique

Dans cette sous-section, nous allons discuter brièvement sur les modèles de langages N-gram ainsi que le lexique de prononciation.

- **N-gram**

La probabilité à priori de la séquence du mot $W=w_1 \ldots w_n$ peut être calculé tel que :

$$P(W)= p(w_1) \cdot p(w_2/ w_1) \ldots\ldots p(w_N/ w_1\ w_2 \ldots\ w_{N-1}) \ldots..(II.12)$$

Il est possible d'obtenir cette probabilité conditionnelle de mots pour toutes les séquences de mots de différente taille. Pour cette raison, les modèles N-gram surtout les bi-grams et tri-grams qui sont utilisés. Ce terme peut être approché tel que :

$$P(W)\approx P(w_j/ w_{j-N+1} \ldots\ldots\ w_{j-1}) \ldots\ldots (II.13)$$

La valeur de N est compromise entre la stabilité de son estimation et sa pertinence. Un tri-gram (N=3) est un choix commun avec un grand corpus d'apprentissage, tandis qu'un bi-gram (N=2) est souvent utilisé avec un corpus moins grand. Comme la quantité de N devient plus grande, il est plus difficile d'estimer de manière fiable la probabilité à priori. Cette probabilité peut être estimée par approche de fréquence relative :

$$P(w_i / w_{i-N+1} \ldots\ldots w_{i-1}) = \frac{F(w_{i-N+1} \ldots\ldots w_{i-1}.w_i)}{F(w_{i-N+1} \ldots\ldots w_{i-1})} \quad \ldots\ldots\ldots (II.14)$$

Où F est le nombre d'occurrence de chaines de caractère dans ses arguments dans le corpus d'apprentissage donné. Il est évident que quelques séquences de mots ne peuvent être observées dans le corpus d'apprentissage. Ceci veut dire qu'une probabilité zéro est fixé aux N-gram non visibles. En plus, la fonction de distribution des fréquences peuvent être tranchante. C'est-à-dire : certains mots peuvent apparaitre plusieurs fois, alors que d'autre se produisent uniquement quelques fois.

Au lieu d'estimer à partir des comptes, des techniques variées de lissage existent pour équilibrer les probabilités. Ceci inclue l'actualisation des estimations récursivement en baissant les N-gram et les interpoler de différents ordre.(Huang, Acero, Hon, & Foreword By-Reddy, 2001)

Les modèles de langages pour les RPCGV ont tendance à être des tri-grams ou même quatre-gram. De bons outils sont valables pour les construire et les manipuler(Stolcke, 2002), (Young et al., 2006). Les grammaires bi-gram et uni-gram sont rarement utilisées pour des applications à grand vocabulaire. Puisque les tri-grams demandent une grande quantité d'espace, les modèles de langages pour des applications à mémoire limitée comme les téléphones portables ont tendance à utiliser des techniques de compression. Quelques applications de dialogue simples prennent avantage de leur domaine limité à utiliser des états fini très simple ou des grammaires à états fini pondérés.

Le lexique appelé dictionnaire est une simple liste de mots, avec une prononciation de chaque mot telle une séquence de phones. Publiquement les lexiques valables comme le dictionnaire (Weide, 1996) peuvent être utilisés pour extraire 64000 vocabulaires de mots communément utilisés pour les RPCGV. La plupart ont une seule prononciation, même si quelques mots comme les homonymes et les mots de fonction fréquents peuvent en avoir plus (le nombre moyen de prononciations par mot dans les RPCGV est de 1 à 2.5).(Çömez, 2003)

II.4.2.4. Recherche et décodage

Maintenant, nous sommes vraiment près à avoir décrit toutes les parties d'un reconnaisseur de parole complet. Nous avons montré comment extraire les caractéristiques cepstrales pour une fenêtre. Nous savons aussi comment représenter une connaissance lexicale, que chaque mot HMM est composé d'une séquence de modèles de phones, et que chaque modèle de phone d'un ensemble d'états sous-phone.

Dans cette section, on montre comment combiner toutes ces connaissances pour résoudre le problème de décodage : combiner tous les estimateurs de probabilités pour produire la chaine de mots la plus probable. On peut poser la question de décodage ainsi : compte tenu d'une chaine d'observations acoustiques, comment devrions choisir une chaine de mots qui a la probabilité postérieure la plus élevée ?

Rappelons que selon (Martin & Jurafsky, 2000),dans le modèle de canal bruyant, on utilise la règle de bayes, comme montré dans la section (II.2) avec le résultat de la séquence de mot qui maximise le produit de deux facteurs, un modèle de langage à priori et une ressemblance acoustique :

$$\widehat{w} = arg\max_{w\epsilon\mathcal{L}} P(O/W)P(W) \ldots\ldots\ldots \text{(II.15)}$$

II.4.2.5. Evaluation : taux d'erreur de mot

La métrique d'évaluation standard selon (Martin & Jurafsky, 2000)pour les systèmes de reconnaissance de parole est le taux d'erreur mot. Le taux d'erreur mot est basé sur combien diffère la chaine de mot retourné (souvent appelée la chaine de mot hypothèse) par le reconnaisseur de la transcription correcte ou référence. Etant donné une transcription correcte, la première étape dans le calcul d'erreur mot est de calculer la distance minimum dans les mots entre les chaines hypothèses et correctes. Le résultat de ce calcul va être le nombre minimum de substitutions de mot, insertions de mot, et suppressions de mot nécessaire pour tracer entre les chaines hypothèses et correctes. Le taux d'erreur mot (TEM) est donc défini tel que suit (noter que parce que l'équation incluse les insertions, le taux d'erreur peut être plus de 100 %) :

$$\text{Taux d'Erreur Mot} = 100 * \frac{\text{Insertions + Substitutions + suppressions}}{\text{Total des mots dans la transcription}}$$

Nous parlons aussi parfois de TEP (Taux d'Erreur Phrase), qui nous dit combien de phrases a au moins une erreur :

$$\text{Taux d'Erreur Phrase} = 100 * \frac{\text{\# des phrases avec au moins un mot erroné}}{\text{Nombre total de phrases}}$$

Voici un exemple de positions entre l'élocution hypothèse et référence du corpus CALLHOME, montrant le compte utilisé afin de calculer le taux d'erreur mot :

REF:	i	***	**	UM	the	PHONE	IS		i	LEFT	THE	portable	****	PHONE	UPSTAIRS	last	night
HYP:	i	GOT	IT	TO	the	*****	FULLEST		i	LOVE	TO	portable	FORM	OF	STORES	last	night
Eval:		I	I	S		D	S			S	S		I	S	S		

Cette élocution a 6 substitutions, 3 insertions et 1 suppression :

$$\text{Taux d'Erreur Mot} = 100 * \frac{6+3+1}{13} = 76.9\ \%$$

II.5. Travaux connexes sur les applications de RAP

Des travaux connexes sont donnés afin de définir notre travail parmi les autres, de clarifier aussi les modèles utilisés dans ce domaine avec leurs limites. Certaines applications de la reconnaissance de la parole peuvent être basées sur des méthodes acoustiques ou articulatoires.

a) Méthodes articulatoires

(Heselwood) Auteur du papier « Le problème de classification et description des pharyngales ». C'est parce qu'aucun endroit de l'articulation dans le schéma alphabet phonétique internationale (API) est loin d'être aussi grand ou intrinsèquement complexe que le pharynx, qu'on fait cette classification. Son papier revoie brièvement l'histoire de classification des sons pharyngales ensuite considère comment des données instrumentales de l'Arabe pourrait informer la façon dont ils devraient être mieux classées et décrites.

(Esling, 1999), est l'auteur du papier « les pharyngales dans le traitement d'acquisition de langage ». Les indices auditives/ acoustique générés dans le pharynx sont les mêmes éléments de sons de production observés dans la naissance d'enfants. Pour résumer, on a identifié que le pharynx est l'origine et le site de la vocalisation primaire des manières d'articulation. Leur hypothèse initiale est que la vocalisation d'enfants montre que la qualité laryngale est primordiale, que l'acquisition de la capacité à produire des sortes de diffusion d'articulation du pharynx est un processus qui parallélise et complémente la capacité des enfants à dissocier les catégories de sons auditifs/perceptuels. Leur nouvelle hypothèse de travail est que les pharyngales sont utilisées en premier, que les sons oraux sont proliférés dans le bavardage et que les pharyngales suivent une hiérarchie de ré-acquisition dans les 12 deuxième mois de vie.

(Ouni & Laprie, 2009), a écrit le papier « étude de la pharyngalisation utilisant un articulo-graphe » cette étude articulatoire a pour but d'étudier les phonèmes emphatiques et pharyngaux dans l'Arabe et leurs effets co-articulatoires en utilisant un articulographe électromagnétique qui est une technique courante pour enregistrer les données articulatoires en les comparant avec celles prises par des images rayons-x (qu'on utilise plus à cause de son danger pour la santé). La pharyngalisation du phonème affecte la voyelle qui le précède ainsi qui le suit.

(Laufer, 2009) Auteur de « articulation pharyngale dans l'Hebrew ». L'auteur inclue des films vidéo du pharynx pour leur étude de l'articulation des consonnes [T, S, K] et l'étude a montré clairement que la production des sons pharyngales et pharyngalisées implique l'épiglotte. L'étranglement étroit semble se situer entre l'épiglotte qui s'incline vers l'arrière et la paroi pharyngée. Ils ont vu que le rétrécissement pharyngale est plus

intense et moins variable pour les sons pharyngales où c'est une articulation primaire que les sons pharyngalisés où c'est une articulation secondaire.

(Zeroual, Esling, & Hoole, 2011), les auteurs de « étude endoscopique, transillumination, EMA et ultrasons de l'emphatique, consonne arrière et labialisée de l'Arabe Marocain (AM) ». /T S/ sont généralement emphatique pas labialisés. /T/ et /K/ ont des gestuels linguales antagonistes, ce qui explique leur restriction de cooccurrence. En combinant toutes ces observations, on peut déduire que les consonnes emphatiques de l'AM sont uvulaires et secondairement pharyngalisés. L'AM a un rétrécissement aryepiglottique classique, ainsi qu'une rétraction de la langue durant ces consonnes.

b) Méthodes acoustiques

(Taqi) Auteur de « la réalisation du [s] tel que [s'] dans l'Arabe Kuwaiti (KA)». Le but est d'analyser la réalisation du (s) tel que (s') en 1er lieu, dans un contexte d'Arabe général où elle est censé apparaitre et en 2ème lieu, tel un facteur d'origine, d'âge et de genre dans l'KA. La méthode ANOVA (Analyse Of VAriance) a montré que le [s'] apparait dans la parole des vieux du Ajami de l'Iran, et il y a une faible utilisation du [s'] par les femelles par rapport aux mâles.

(Al-Tamimi, 2009) Auteur de « analyses statiques et dynamiques de l'effet de la pharyngalisation sur les voyelles de l'Arabe Marocain et Jordanien ». Le but de ce travail est d'examiner les effets de la pharyngalisation sur les F3, les formants de transition et les départs de chaque formant. Les mesures des trois premiers formants ont été extraites automatiquement en utilisant la méthode Burg proposée par PRAAT. Ils ont remarqué que toutes les voyelles deviennent plus ouvertes et plus arrière dans l'environnement [d'] par rapport à l'environnement [d]. L'analyse statique consiste à déterminer les fréquences de formants à mi point temporel qui sont considérés traditionnellement comme le point où les influences des consonnes sur les voyelles étant moins évidente. Par contre, l'analyse dynamique consiste à quantifier les seuils des départs des formants à leurs mi point temporel via l'analyse de régression linéaire et régression polynomiale de second et 3ème degré. Les résultats montrent dans les deux langues que les départs de F1 sont bas et ceux de F2 sont ascendants et ceux de F3 sont bas dans l'environnement /d/ comparé à /d'/. On conclut que l'analyse dynamique montre plus de différence entre les voyelles produites dans les deux environnements de la consonne et entre les deux dialectes aussi.

(Girgis, 2009) Auteur de « fricatives pharyngalisés dans l'Arabe Egyptien (EA) : locuteur patrimoine versus non patrimoine ». L'hypothèse nulle dit qu'il n'y a pas de différence dans la qualité de la voyelle qui suit la fricative emphatique entre les locuteurs patrimoine et non patrimoine. L'hypothèse alternative dit que l'Egyptien hérité (patrimoine) montre moins de pharyngalisation comparé au non-hérité. ANOVA a été exécuté où chaque valeur de F2 des voyelles a été traité comme un cas à part. Ses résultats ont rejeté l'hypothèse nulle du moment où l'effet est signifiant pour l'interaction de la pharyngalisation (surtout les dentales) ainsi que l'état de l'héritage.

(Bellem, 2009) a écrit « le problème de pharyngalisation et son rôle dans les systèmes de sons de langages caucasien ». Acoustiquement, l'effet de la pharyngalisation sur les voyelles avant [i, e], implique l'ascendance de F1 et la descente de F2 mais sur les voyelles arrière [a, o, u], implique l'ascendance des deux formants. La pharyngalisation dans cette langue implique aussi la rétraction de la racine de la langue et l'avancement du corps de la langue. La pharyngalisation caucasienne se contraste avec la langue Arabe où l'acoustique implique souvent les approximations de F1 et F2, et le palatal est généralement concurrent à la pharyngalisation.

Il ya une limite quand on étudie les méthodes acoustiques dans la reconnaissance des taux. Nous pensons qu'une combinaison de la méthode acoustique avec l'articulatoire on pourrait avoir un meilleur résultat.

Nous verrons maintenant les travaux faits sur les applications de la reconnaissance de la parole :

Dans le travail de (Jafri, Sobh, & Alkhairy, 2015), le système est basé sur la synthèse des formants. Lorsque les formants de la parole prononcés à partir du texte, ont une petite dimension et sont proches des signaux de perception de la parole, alors leur système hybride produit une bonne qualité de la parole. Leur système hybride est composé des modèles Hidden Semi-Markov (HSMM) et GMMs. De plus, la fréquence fondamentale des formants est extraite de la forme d'onde dans le but est l'apprentissage HSMM, mais la durée est calculée par une distribution gaussienne multi-variée.

Les auteurs (Tahiry, Mounir, Mounir, & Farchi, 2016) présentent une nouvelle étude qui dit qu'en augmentant la longueur de la production de voyelles, le comportement de ces voyelles change. Cette nouvelle méthode concerne le calcul des bandes d'énergie, des moments spectraux ainsi que l'extraction des fréquences des formants. Leurs résultats ont montré que les fréquences des formants et les bandes d'énergie peuvent différencier les courtes voyelles et longues. Les moments spectraux (centre de gravité et écart type) ont révélé que la production des voyelles se fait en deux phases. Une phase transitoire au début de la production de voyelles et une phase d'état stable lorsque la durée augmente. Les auteurs considèrent que la longueur des voyelles affecte la signification des mots en arabe.

Les auteurs (Alsulaiman, Mahmood, & Muhammad, 2017) étudient l'effet des phonèmes arabes sur la performance des systèmes de reconnaissance des locuteurs. L'étude révèle que certains phonèmes arabes ont un fort effet sur le taux de reconnaissance de tels systèmes. Les performances des systèmes de reconnaissance des locuteurs peuvent être améliorées et leur temps d'exécution peut être réduit en utilisant ce résultat. De leur recherche, ils ont constaté que les taux de reconnaissance des voyelles arabes étaient tous au-dessus de 80 %. Les taux de reconnaissance des consonnes varient de très faible (14 %) à très élevé (94 %), ces derniers obtenus par une consonne pharyngée suivie des deux phonèmes nasaux, qui ont atteint des taux de reconnaissance entre 70 % et 80 %.

Dans le travail de (Smaragdis & Raj, 2012), les auteur ont introduit un nouveau modèle de sélection Markov capable de reconnaitre la parole de plusieurs enregistrements de locuteurs simultanément. Leur travail est basé sur des représentations de spectrogrammes non négatives, qui sont démontrés très efficace en problème de séparation de sources. Contrairement aux modèles de Markov factoriel (un produit des HMMs représentant les sources individuelles), utilisé dans le passé qui sont exponentiel en nombre de sources. Leur approche caractérise un modèle de faible complexité de calcul avec un espace d'état linéaire en nombre de sources. Ils ont démontré l'utilisation de leur travail en reconnaissant la parole d'un mélange de locuteurs connus.

Les auteurs (Herbig, Gerl, & Minker, 2012) ont présenté une nouvelle approche pour la reconnaissance de parole et pour l'identification du locuteur ensemble. Des locuteurs non supervisés sont suivis et une adaptation automatique d'interface homme-machine est aboutie par l'interaction d'identification de locuteurs, la reconnaissance de parole et l'adaptation de locuteur pour un nombre limité de locuteurs fréquents. Pour cela, les auteurs ont appliqués des profils spécifiques de locuteurs qui permettent de prendre des caractéristiques individuelles de parole en considération, afin d'atteindre de grands taux de reconnaissance. En plus, la détection de locuteurs inconnus est faite dans le but d'éviter d'entrainer de nouveaux profils de locuteurs. Leur système de contrôle de la parole est idéal pour les applications automobiles tels que : la navigation contrôlée par la parole, la téléphonie main-libre ou les systèmes d'info-divertissement.

Les auteurs (Hofe et al., 2013) ont travaillé sur la reconnaissance de mots isolés et chiffres connectés en utilisant une interface de parole silencieuse basée sur la détection magnétique de mouvements articulatoires utile pour les personnes atteintes de maladies qui affectent les cordes vocales. Les auteurs ont utilisé des données de capteurs à la place de caractéristiques acoustiques pour l'apprentissage des modèles de Markov cachés. La performance de la reconnaissance de la parole d'un prototype MVOCA (aide à la communication par la voix magnétique) a été évaluée. Les résultats ont été comparés avec ceux obtenus en utilisant les MFCC d'un signal de référence acoustique. Les précisions des mots ont été en général au dessus de 90 %, même si les ensembles de données expérimentales étaient peu comparé aux quantités utilisés dans la reconnaissance de la parole acoustique. De plus, l'influence de la topologie des modèles statistiques sur la performance de la reconnaissance de la parole est discutée brièvement pour le cas de chiffres connectés. Leur travail répond aux besoins des patients qui ont perdu la capacité vocale.

Les auteurs (Bourouba, Bedda, & Djemili, 2006) ont travaillé sur une approche hybride DTW/GHMM pour la reconnaissance de mots isolés en 2006. Dans leur papier, ils ont présenté une nouvelle approche hybride pour reconnaitre les mots isolés en utilisant un HMM combiné avec un DTW (Dynamic Time warping). Dans leur travail, ils ont fait une évaluation comparative entre le HMM continu traditionnel (GHMM) et la nouvelle approche DTW/GHMM.

Les auteurs (Saon & Soltau, 2012) ont présenté dans « l'amélioration systèmes pour la reconnaissance de parole continue à grand vocabulaire » en 2012 un niveau de trame qui stimule l'apprentissage séquentiel et combine l'ensemble des modèles acoustiques de façon à ce que les derniers modèles compensent les insuffisances des premiers. Ces trames utilisent le classifieur HMM avec un faible modèle de langage et on réapprend les arbres de décisions phonétiques sur les données pondérées à chaque itération. Les poids diminuent à chaque itération de façon à ce que les trames soient décodées correctement par le système courant. Ces poids sont ensuite multipliés par les statistiques du niveau de trame pour les arbres de décision et les composants du mélange Gaussien de la prochaine itération du système.

La reconnaissance de la parole en hindi pour les mots corrigés utilisant HTK avait été réalisée par (Kumar, Aggarwal, & Jain, 2012). Le but était d'afficher le texte correspondant à la parole prononcée par la machine. Le système a été entrainé pour la reconnaissance de n'importe quelle séquence de mots parmi les 102 de leur vocabulaire. Les coefficients MFCC ont été utilisés pour l'extraction de caractéristiques des fichiers de paroles ; et le logiciel Wavesurfer a été utilisé pour les étiqueter. Les systèmes développés emploient une approche basée sur des règles de grammaire comme modèle de langage (BakusNaur). Le système a été entrainé pour estimer les paramètres HMM utilisant des modèles acoustiques au niveau mots avec le modèle du silence inclus. Un système de vocabulaire de mots utilise deux à cinq itérations pour converger. Différents nombre d'états sont sélectionnés dépendant du nombre d'unités de phones et la durée des mots. Les résultats expérimentaux montrent que le système fournie la précision de mot de 87.01 %, un taux d'erreur de 12.99 % et correction de mots de 90.93 %.

Dans le travail de (Pirhosseinloo & Ganj, 2012), les auteurs ont présenté l'utilisation de critères tels que l'erreur du phone minimum (MPE) et maximum d'informations mutuelles (MMI) qui sont étudiés pour les modèles d'apprentissage HMM discriminatifs, pour la reconnaissance de parole continue Persane. Ces critères sont attendus à améliorer l'estimation des transformations linéaires (DLT) pour l'adaptation du locuteur. Par contre, l'adaptation MLLR (Maximum Likelihood Linear Regression) est une méthode populaire qui estime les paramètres de transformations par le critère ML (Maximum Likelihood). Inopportunément, la performance des critères discriminatifs pour l'adaptation du locuteur non supervisé sont limités à cause de la sensibilité de ces critères aux erreurs dans l'hypothèse. Leur travail estime les paramètres de transformations linéaires en utilisant les critères discriminatifs sur le corpus Farsdat. Les résultats ont montré que l'apprentissage discriminatif surpasse l'apprentissage ML standard. En plus, le MPE-DLT réduit le taux d'erreur de mot par rapport au MLLR.

Les auteurs (Seid, Yegnanarayana, & Rajendran, 2012) présentent dans leur papier une technique utilisés pour identifier les régions des sons qui stoppent la glotte dans la parole continue Amharique. Ils ont utilisé quelques sources de voix tels que : les paramètres du modèle « Logarithme du Pic d'intensité d'excitation normalisée » (LPNES), les coefficients de corrélation croisé normalisé de prédiction linéaire d'un signal résiduel, et aussi une gigue normalisée. Les corrélations articulatoires acoustiques du système peuvent

être vues à partir d'un spectrogramme d'un signal. Mais si le son est produit sans effet de geste du système articulatoire, il est difficile d'observer les caractéristiques d'un tel son. C'est le cas des sons qui stoppent la glotte. La solution est de mettre une voyelle qui suit ce son afin de diffuser ce son.

Les auteurs (Amrous, Debyeche, & Amrouche, 2011) ont écrit un papier en 2011 qui étudie la contribution de formants et les caractéristiques de prosodie tel que le pitch (fréquence zéro) et l'énergie dans la reconnaissance de l'arabe sous les conditions de la vie réelle. Leur système RAP est implémenté en utilisant HTK. Les modèles sont des HMM gauche-droites avec des densités d'observation continues. Chaque modèle se constitue de 3 états tels que chaque état est modélisé par un mélange Gaussien avec des matrices de covariances diagonales. Les expériences sont performés sur le corpus ARADIGIT. Les résultats obtenus montrent que les vecteurs caractéristiques multi-variant (combinaison de MFCC et informations prosodique et formants) produit dans un environnement bruité, penchant vers une amélioration signifiante de plus de 27 % en RM correcte en comparaison avec le système basique. Ils ont concluent que les caractéristiques prosodiques et formants contiennent l'information complémentaire à celle qui provient des caractéristiques MFCC.

II.6. Reconnaissance de la parole arabe

II.6.1. La langue arabe

Dans notre thèse, la langue arabe est la langue qui reste largement vivante sémitique en fonction du nombre de locuteurs. Elle dépasse 250 millions de locuteurs de langue maternelle et l'arabe en tant que deuxième langue peut atteindre quatre fois ce nombre. L'arabe est la langue officielle dans 21 pays et classée telle la 6ème langue la plus parlée basée sur le nombre de locuteurs de langue maternelle et c'est l'une des six langues officielles des Nations Unies. Les alphabets arabes sont utilisés dans d'autres langues comme le Perse et l'Ourdou (Hyassat & Zitar, 2006).

La langue arabe a deux formes principales : Arabe Standard et Arabe Dialectale. L'arabe standard inclue l'Arabe Classique et l'Arabe Standard Moderne (ASM) tandis que l'arabe dialectal inclue toutes les formes de l'arabe parlé actuellement dans la vie de tous les jours et cela varie selon les pays et dévie de l'Arabe Standard à une certaine mesure et s'écarte même à l'intérieur du même pays de sorte qu'on peut trouver différents dialectes. Tant qu'il y a plusieurs formes de l'Arabe, il reste plein de caractéristiques en commun au niveau acoustique et au niveau de langage (Elmahdy, Gruhn, Minker, & Abdennadher, 2009).

Pour l'arabe standard nous avons choisi et la forme classique et l'ASM et pour l'arabe dialectale nous avons choisi l'Arabe colloquiale Egyptien (ACE) (l'Arabe colloquiale Egyptien veut dire habituellement le langage parlé dans le Caire) tel l'exemple typique puisque c'est le dialecte le plus populaire parmi les locuteurs arabes. Les problèmes de la reconnaissance de parole principaux pour l'arabe sont discutés : la complexité morphologique, l'existence de plusieurs formes dialectales, et les ressources textes qui ne

sont pas diacritiques, nous nous concentrons sur la langue arabe parce que c'est la plus grande langue vivante encore sémitique basée sur le nombre de locuteurs.

II.6.2. Les problèmes de RAP pour la langue arabe

II.6.2.1. Formes de la langue arabe

- **Arabe Standard Moderne (ASM)**

L'ASM est la forme standard actuelle de l'arabe. Toutes les ressources écrites en arabe sont en ASM. C'est le langage arabe parlé formel. L'ASM est utilisée dans les livres, les journaux, la diffusion de nouvelles, les discours formels, le sous-titrage de films, etc. l'ASM peut être considéré comme un langage second pour tous les locuteurs arabes. Tous les locuteurs arabes peuvent comprendre l'ASM parlé en diffusion de nouvelles. L'ASM est la seule forme acceptée partout dans les locuteurs arabes natifs, c'est pourquoi plusieurs radios et chaînes de télévision utilise l'ASM dans le but de cibler tous les locuteurs arabes. Le syllabes permises en arabe sont : CV, CVC et CVCC où C es t une consonne et V est une voyelle longue ou courte. Donc les élocutions et mots arabes ne peuvent commencer que par une consonne.

L'inventaire phonétique de l'ASM se constitue de 38 phonèmes. Ces phonèmes incluent 29 consonnes originales, et 3 consonnes étrangères, et 6 voyelles. Nous utilisons une notation SAMPA. Les consonnes étrangères sont : /g/, /p/ et /v/ et sont rarement trouvés en ASM et apparaissent seulement dans les mots d'emprunt. Le phonème /l'/ est un phonème rare puisqu'il apparait seulement dans le mot / ?al'l'a:h/ (le dieu) et ses dérivantes.

Habituellement la durée de voyelles longues double approximativement la durée de voyelles courtes. L'arabe est caractérisée par l'existence de phonèmes pharyngales et emphatiques comme : /X\/, /t'/, /d'/, /D'/, et /s'/. Ces types de phonèmes existent seulement en langues sémitiques(Holes, 2004). Les sons emphatiques ont un effet signifiant sur la totalité du mot contenant la consonne emphatique. Des travaux précédents ont montré une petite différence entre le formant F1 et F2 pour toutes les voyelles des mots contenant une consonne emphatique dans presque toutes les positions. Quelques phonéticiens ajoutent deux diphtongues qui sont : /ay/ (/a/ suivie par /y/) et /aw/ (/a/ suivie par /w/) et considérer ces diphtongues doivent demander plus d'effort dans la transcription puisqu'ils peuvent apparaitre ou pas.

Les corpus de discours ASM sont principalement valables dans le domaine de l'émission de nouvelles à un prix bas relativement et ces corpus peuvent être utilisés dans la construction de modèles acoustiques locuteurs-indépendant (Yaseen et al., 2006).

Habituellement les phonèmes étrangers dans la transcription ASM ne sont pas traités tels des sons supplémentaires et sont groupé avec le phonème le plus proche, par exemple le /f/ et le /v/ sont groupés ensemble et traités tel le même phonème, la même approche est appliquée pour le /b/ et /p/ et aussi appliquée pour les phonèmes /g/, /Z/ et /dZ/.

L'unique raison de ce groupement est essentiellement parce que les phonèmes étrangers sont rarement utilisés en ASM comparant aux sons originaux et cela est dû au fait que les lettres arabes standard n'ont aucune lettre standard affectée aux sons étrangers. Quelques efforts ont été faits afin de différencier entre les sons étrangers et les sons originaux en utilisant les lettres arabes non-standard comme utiliser la lettre پ pour le phonème /p/, utiliser la lettre ڤ pour le phonème /v/, et utiliser la lettre چ pour le phonème /g/, mais ces conventions ne sont pas standard et même la disposition de claviers arabes standard ne montrent pas ces lettres ainsi que les ensembles de caractères arabe standard tel que le ISO 8859-6 n'inclue pas ces lettres supplémentaires (Elmahdy et al., 2009).

- **Arabe Classique**

L'arabe classique est la forme la plus standard et formelle de l'arabe et c'est la langue du coran (le livre sacré pour les musulmans). Le scripte arabe classique représente presque toutes les phonétiques du monde car le script est entièrement voyellé et inclue les symboles diacritiques qui sont habituellement négligés in ASM.

Le langage arabe est caractérisé de la présence des consonnes emphatiques /d'/ et les Arabes croient que ce phonème comme il est prononcé dans l'arabe classique apparait exclusivement dans l'arabe et dans aucune autre langue, c'est pourquoi l'arabe est habituellement définie telle la langue du /d'/(Newman, 2002).

Les phonétiques du coran (selon le récit de Hafs de Assim) incluent les sons de l'ASM (excepté les phonèmes étrangers) plus quelques sons supplémentaires, ce qui suit va résumer les principaux sons supplémentaires qui peuvent apparaitre dans le coran :

- La prolongation de la voyelle avec la durée de 4, 5, ou 6 voyelles courtes.

- La prolongation nécessaire de 6 voyelles courtes.

- La nasalisation (ghunnah) avec la durée de 2 voyelles courtes.

- La prononciation emphatique de la consonne /r/ peut se faire en dépendant du contexte de la consonne /r/.

- La sonorisation du son dans les lettres d'agitation (qualquala) pour les consonnes /q/, /t'/, /b/, /dZ/, et /d/ peut se faire en dépendant du contexte de ces consonnes.

Utiliser le coran dans la reconnaissance de parole est maintenant limité aux applications d'apprentissage du récit tel que dans les travaux de (Abdou et al., 2006) et (Ibrahim et al., 2008). Dans de telles applications, le modèle acoustique est entrainé avec le récit du coran et l'utilisateur est demandé à prononcer un certain verset puis l'application identifie n'importe quelle erreur du récit.

- **Arabe Dialectale**

L'ASM n'est pas le langage parlé naturel pour les locuteurs Arabes natives tandis que l'arabe colloquiale (ou dialectale) est l'arabe parlé naturel dans la vie de tous les jours.

L'arabe colloquiale n'est pas utilisée en tant que forme standard de l'arabe dans l'écrit ou dans la publication. Il y a plusieurs dialectes arabes et presque tout pays a sa propre forme colloquiale. Aussi dans le même pays on peut trouver différents dialectes. L'arabe dialectal peut être divisé en deux groupes : l'arabe occidental et l'arabe oriental. L'arabe occidental peut être divisé en dialecte marocain, tunisien, algérien, libyen. Par contre, l'arabe oriental peut être divisé en dialecte égyptien, du Golfe, du Damas, et levantin. L'arabe du Damas est considéré comme proche de l'ASM. Les locuteurs avec différents dialectes utilisent habituellement l'ASM pour communiquer.

Parce que les dialectes arabes ne sont pas utilisés dans la forme écrite, préparer des corpus de parole adéquat pour l'arabe dialectale pour le but de modèle acoustique est vraiment couteux mais reste faisable, et préparer un grand corpus de texte pour l'arabe dialectale qui peut être utilisé dans le modèle de langage statistique large vocabulaire et encore plus difficile vue que dans le modèle de langage large vocabulaire on a besoin d'un corpus qui contient les mots dans l'ordre de millions qui reste pas réalisable. Les corpus valable pour l'arabe dialectal sont chers comparés à l'ASM et sont aussi des corpus de conversations de téléphonie à basse qualité tel que la Parole Arabe Egyptienne CALL-HOME(Canavan, Zipperlen, & Graff, 1997), ou des corpus de parole lu à haute qualité mais avec un vocabulaire vraiment limité comme dans le projet oriental(Association, 2006).

- **Arabe Colloquiale Egyptien (ACE)**

Les caractéristiques phonétiques principales des phonétiques de l'ACE comparés à l'ASM sont :

- /t/ et /s/ sont utilisés à la place de /T/. e.g. /Tala:Thah/ (trois) dans l'ASM est transformée à /tala:tah/ en ACE

- /g est utilisé à la place de /Z/ et /dZ/. e.g. /Zami:l/ (beau) en ASM est transformée à /gami:l/ en ACE.

- /?/ est utilisé à la place de /q/. e.g. /qabl/ (avant) dans l'ASMest transformée à /?abl/ en ACE.

- L'existence de la voyelle longue non arrondie milieu de devant /E:/ et la voyelle courte /E/ (ils n'existent pas en ASM). e.g. /?iTnayn/ (deux) en ASM est transformée à /?iTnE:n/ en ACE.

- L'existence de la voyelle longue non arrondie arrière ouverte /A:/et la voyelle courte/A/ (ils n'existent pas en ASM). e.g. /?arba?\ah/ (quatre) en ASM est transformée à /?ArbA?\Ah/ en ACE.

- L'existence de la voyelle longue arrondie arrière milieu /O:/ (cela n'existe pas dans l'ASM). e.g. /jawm/ (jour) dans l'ASM est transformée à /jO:m/ en ACE.

Au niveau du vocabulaire, quelques mots existent seulement dans l'ACE et non dans l'ASM comme : /t'ArAbE:zA/ qui veut dire table en ACE alors que c'est /t'awila/ en ASM. La structure de phrase en ACE a tendance à suivre un ordre VSO (Verbe Sujet Objet) tandis qu'en ASM tendent à SVO (Sujet Verbe Objet).

La transcription de l'ACE est difficile parce que les gens sont influencés par l'ASM et écrivent toujours les mots en ASM à la place, par exemple le mot ACE /tamanjah/ (huit) est généralement transcrit incorrectement tel que /tama:njah/ ou / Tama:njah / et garde l'inclusion de la voyelle longue /a:/ comme dans l'ASM tandis que dans l'ACE la transcription est remplacée par la voyelle courte /a/.

II.6.2.2. Complexité morphologique

L'arabe est une langue morphologique vraiment riche, et donc traitant l'arabe en tant que morphèmes au lieu des mots va limiter la taille du dictionnaire et va diminuer considérablement le nombre des mots hors-vocabulaire (HV). Par exemple un lexique de 65 000 mots dans le domaine de diffusion de nouvelles mène à un taux de HV de 4 % dans l'arabe tandis qu'en anglais cela mène à un taux de HV de moins d'1 % (Billa et al., 2002). La plupart des mots en arabe ont la racine qui se compose de trois consonnes appelées radicales (rarement deux et quatre). Un grand nombre d'affixes (préfixes, infixes et suffixes) peuvent être ajoutés aux trois consonnes radicales pour former des modèles. L'arabe est une langue qui est fortement conjuguée avec le genre, le nombre, le temps, les personnes et les cas. Un seul mot arabe peut représenter une phrase complète en français comme : وباستطاعتهم /wabi?stita'a:?\atihim/ qui veut dire : et avec leur capacité(Alshalabi, 2005).

II.7. Travaux connexes sur la reconnaissance de la parole arabe

Dans notre thèse, nous ciblons la langue arabe. C'est une langue sémitique parce qu'il ya 250 millions d'orateurs dont c'est la première langue, et les personnes qui parlent l'arabe comme langue seconde pourraient atteindre quatre fois ce nombre(Hyassat & Zitar, 2006). Dans ce travail, un reconnaisseur arabe basé sur SPHYNX est introduit par les auteurs pour la première fois. Les auteurs suggèrent une boîte à outils capable de produire un dictionnaire de prononciation du coran et de la langue arabe standard. Ils ont obtenu une précision de 70,813 % pour la base de données saint coran et de 98,182 % pour le corpus de commande et de contrôle.

Le travail académique de(Labidi, Maraoui, & Zrigui, 2016), a donné une théorie qui dit que les dictionnaires phonétiques faits avec indépendance entre les phonèmes de voyelles et les consonnes est bénéfique pour le système d'identification de langue arabe. C'est ce qu'on appelle la première théorie des phonologues arabes. La seconde théorie est que les voyelles sont des parties de phonèmes de consonnes. Les auteurs ont obtenu une moyenne de mot de 13,41 % en utilisant le dictionnaire basé sur la première théorie et 23,09 % en utilisant le dictionnaire basé sur la deuxième théorie. Leur étude indique la faiblesse de HMM à la représentation de la variabilité dans la longueur des voyelles.

Dans la recherche de (Almisreb, Abidin, & Tahir, 2015), les auteurs ont étudié la reconnaissance des phonèmes arabes spécifiquement pour les locuteurs malais. Les méthodes proposées sont évaluées et examinées en utilisant un corpus contenant des marqueurs de phonèmes arabes. Le processus de reconnaissance est atteint par le modèle de déformation temporelle dynamique (DTW) et le réseau neuronal de reconnaissance de forme (PRNN) pour vérifier la similarité entre les phonèmes arabes. Dans leur étude, trois méthodes sont utilisées pour évaluer la phase de reconnaissance. Les résultats obtenus ont montré que le taux global de reconnaissance de DTW individuellement est de 89,92 %, 94 % pour le PRNN uniquement pour la fusion de DTW et de PRNN, le taux de reconnaissance moyen atteint est de 98,28 %.

Les auteurs (Gales et al., 2007) ont développé un système phonétique pour une reconnaissance de parole Arabe à grand vocabulaire en 2007. Le problème considéré dans leur papier, est l'apprentissage discriminatif lorsqu'il y a un grand nombre de variantes de prononciations pour chaque mot. La performance et combinaison des modèles acoustiques graphémiques et phonétiques sont ensuite comparées sur les informations de diffusion de nouvelles et aussi de diffusion de conversation. La contribution finale du papier est un schéma simple pour générer automatiquement les prononciations pour usage d'apprentissage ainsi que la réduction du taux hors vocabulaire phonétique. Le papier conclue avec une description et résultats de l'utilisation de système phonétique et graphémique dans le cadre de combinaison multi-passe. Le procédé de référence utilisé dans leur travail est l'analyseur morphologique Buckwalter (version 2.0).

II.8. Conclusion

Maintenant que nous avons défini notre motivation pour cette thèse, ainsi posé la problématique dans un premier temps : celle de la durée de son, puis un état de l'art sur les défis de la langue arabe dans un deuxième temps, nous sortons avec quelques points essentiels. Nous situons notre travail par rapport aux autres du point de vue de la langue : nous avons vu que dans l'anglais, le mandarin et beaucoup d'autres, il n'y avait pas réellement le problème de variation temporelle du phonème. Contrairement à la langue arabe ou elle peut être lu normalement, ou chanté dans le cas de la lecture du coran avec les règles de Tajweed. Dans ce cas, les voyelles longues (en arabe *mudud*) sont prononcées différemment et dépendent du contexte. Non seulement il faut voir les phonèmes qui précèdent et qui suivent cette voyelles pour définir sa prononciation selon que le phonème d'avant soit pharyngale ou emphatique, mais aussi il faut définir le temps nécessaire à prononcer cette voyelle selon les règles de Tajweed.

Il est temps de passer aux méthodes utilisées pour la reconnaissance de la parole isolée ou connectée dans le prochain chapitre, ainsi définir des solutions qui résous les problématiques posées dans le chapitre I, c'est notre contribution dans cette thèse.

Chapitre III

Outil de développement HTK

III.1. Introduction

Nous avons vu dans le chapitre de l'introduction que parmi les problèmes de la langue arabe, il y a celui de la distinction entre les phonèmes pharyngaux et les sons emphatiques qui conduisent à un grand nombre de phonèmes ou de son à reconnaître.

Nous proposons une solution pour un problème classique tel que la reconnaissance des chiffres : les modèles de Markov cachés au niveau du phonème. De plus, différentes lectures du Coran de différents célèbres lecteurs soulèvent le problème de la coarticulation en général, mais aussi un autre problème particulier qui est la durée du son en temps due aux règles de Tajweed appliquées par ces célèbres lecteurs. Nous proposons d'utiliser tout simplement pour la difficulté de coarticulation, les modèles de Markov cachés au niveau de tri-phones, et pour le problème de durée de voyelles appelé en arabe « mudud », nous proposons d'utiliser les tri-phones liés puis étendu aux modèles de mélange de Gaussien (GMM).

La raison pour laquelle cette problématique valle la peine d'être posée est la détection coranique indépendante du locuteur. Nous verrons aussi les travaux connexes sur les méthodes appliqués à la reconnaissance de la parole selon la taille du vocabulaire.

III.2. L'outil HTK en bref

Nous avons décidé de travailler avec HTK qui a été une partie fondamentale de la reconnaissance de la parole de recherche (Young et al., 2006) et en même temps, la résolution de certaines questions n'est pas si triviale. Le HTK est un ensemble d'outils logiciels pour la construction de systèmes de reconnaissance vocale. Il peut effectuer des tâches de densité continue, de densité semi-continue ou de HMM à probabilité discrète. Il est développé par un groupe de parole de l'université de Cambridge.

Il a été amélioré et on y a ajouté des propriétés au cours des années depuis le début des années 90. Principalement, HTK est conçu pour être suffisamment souple pour soutenir à la fois la recherche et le développement de systèmes HMM. Nous l'avons utilisé parce qu'il a supprimé le problème de la pharyngalisation en arabe en utilisant beaucoup de texte, et il peut mettre en œuvre des modèles triphone visant à améliorer le problème de la coarticulation (Martin and Jurafsky 2000).

HTK est un outil open source, en lui donnant des fichiers audio ".wav" contenant des phrases en plus de leurs textes associés, c'est-à-dire : avec la transcription phonétique exacte ainsi que le dictionnaire, il apprend les caractéristiques de chaque phonème. Grâce aux commandes reconnaissantes de cet outil, même avec une gaussienne de moyenne fixée à zéro, HTK apprend les caractéristiques acoustique-phonétique, c'est-à-dire : l'énergie, les formants et le spectre de chaque son. Dans le cas contraire, pour la reconnaissance de la parole continue à grand vocabulaire (LVCSR), l'emploi de modèle semi Markov (SMM), donnent de meilleurs résultats que les HMMs (Park & Yoo, 2014).

III.3. Application du modèle de Markov caché HMM

Nous appliquons l'approche HMM à la parole parce qu'ils sont des automates de Markov avec des états cachés; Il s'agit de modèles statistiques dans lesquels le système est modélisé en tant que processus markovien de paramètres inconnus (Rabiner, 1989). Ils sont adaptés à la reconnaissance des mots connectés parce que les HMM modélisent la parole sur le temps.

Voyons maintenant comment le modèle HMM est appliquée à la reconnaissance de la parole. Un modèle de Markov caché est caractérisé par les composantes suivantes :

$Q = q_1 q_2 .... q_n$	Un ensemble d'*états*
$A=a_{01}\ a_{02} \ldots a_{n1} \ldots a_{nn}$	Une *matrice de probabilité de transitionA*. chaque a_{ij}représentant la probabilité du déplacement d'un état *i* à un état *j*. $\sum_{j=1}^{n} aij = 1\ \forall i$
$O= o_1\ o_2\ \ldots\ o_n$	Un ensemble d'*observation*, chacune dessinée d'un vocabulaire V= v_1, v_2, ….v_n
$B=b_i(o_t)$	Un ensemble de *probabilités d'observation*, appelé aussi *probabilités d'émission*, chacune exprimant la probabilité d'une observation o_t étant générée d'un état *i*
q_0, q_{fin}	Un *état début et fin* spécial qui ne sont pas associés à une observation

De plus, ce chapitre introduit l'algorithme de Viterbi pour le décodage des HMMs et l'algorithme de Baum-Welch ou Forward-Backward pour l'apprentissage des HMMs.

Toutes ces facettes du paradigme de HMM jouent un rôle crucial dans la reconnaissance automatique de parole (RAP). On commence ici par discuter comment les états, transitions et observations répondent à la reconnaissance de la parole. Nous allons revenir aux applications de la RAP du décodage de Viterbi dans la section III.5 Les extensions aux algorithmes de Baum-Welch dont nous avons besoin pour faire affaire avec les langages parlés sont traitées dans la section III.5.

Pour la parole, les états cachés sont les phones, des parties de phones ou mots ; chaque observation est l'information concernant le spectre et l'énergie de la forme d'onde à un point dans le temps, et le processus de décodage marque cette séquence de l'information acoustique à des phones et mots.

La séquence d'observation pour la reconnaissance de la parole est une séquence de vecteur de caractéristiques acoustiques. Chaque vecteur de caractéristiques acoustiques représente l'information telle une somme d'énergie dans

différentes bandes de fréquences à un certain point dans le temps. Nous allons orienter la section III.5 à la nature de ces observations mais pour le moment, nous allons simplement noter que chaque observation se constitue d'un vecteur de 39 caractéristiques de valeur réel indiquant l'information spectrale. Les observations sont généralement dessinées chaque 10 millisecondes, ainsi 1 seconde de parole demande 100 vecteurs de caractéristiques spectrales. Chaque vecteur est de taille 39.

Les états cachés des modèles de Markov peuvent être utilisés pour modéliser la parole dans différentes de façons. Pour les petites taches, telle que la reconnaissance des chiffres (la reconnaissance des 10 chiffres de zéro à neuf), ou pour la reconnaissance de oui/non (la reconnaissance de deux mots oui et non), on peut construire un HMM dont les états correspondent à des mots entiers. Par contre, pour la plupart de tâches plus grandes, les états cachés de HMM correspondent à des unités ressemble-phones, et les mots sont des séquences de ces unités ressemble-phone.

Commençons par la description d'un modèle HMM dans lequel chaque état de HMM correspond à un phone unique (un phone est l'unité de base en phonétique. Dans la phonétique, la transcription est fermée par des crochets au lieu des slashs comme en phonémique). Dans un tel modèle donc, un mot HMM se forme d'une séquence d'états HMM concaténés ensemble. La figure (III.1) montre une structure schématique d'un état-phone de HMM pour le mot six en anglais.

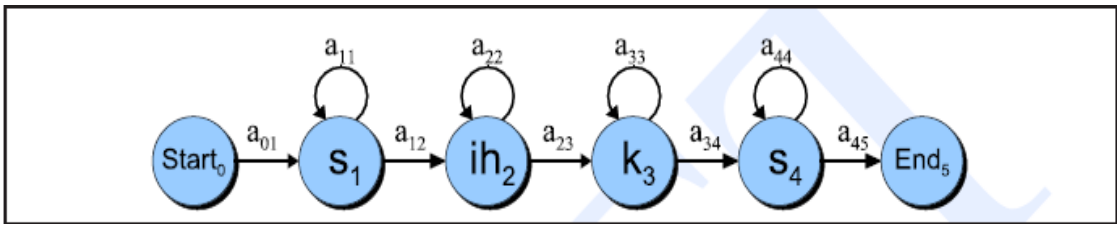


Figure III.1 un HMM pour le mot *six*, se constituant de quatre états émettant et deux non émettant, les probabilités de transitions A, les probabilités d'observations B, et une séquence d'observation échantillon.

Pour chaque tâche de parole simple (la reconnaissance d'un petit nombre de mots tel que les 10 chiffres) utilisant un état HMM pour représenter un phone est suffisant. En général, les tâches de reconnaissance de parole continue à grand vocabulaire (RPCGV), une représentation plus fine est nécessaire. Ceci parce que les phones peuvent durer plus d'une seconde, c'est-à-dire plus de 100 frames, mais les 100 frames ne sont pas acoustiquement identiques.

Les caractéristiques spectrales d'un phone, et la quantité d'énergie varient considérablement à travers un phone. Par exemple, la consonne stop [t] a une clôture qui a une petite énergie acoustique, suivie d'une ouverture brusque. Similairement, les diphtongues sont les voyelles dont les formants F1 et F2 changent significativement. La figure (III.2) montre ces grands changements dans les caractéristiques spectrales à travers le temps de chacun des deux phones dans le mot prononcé en anglai.

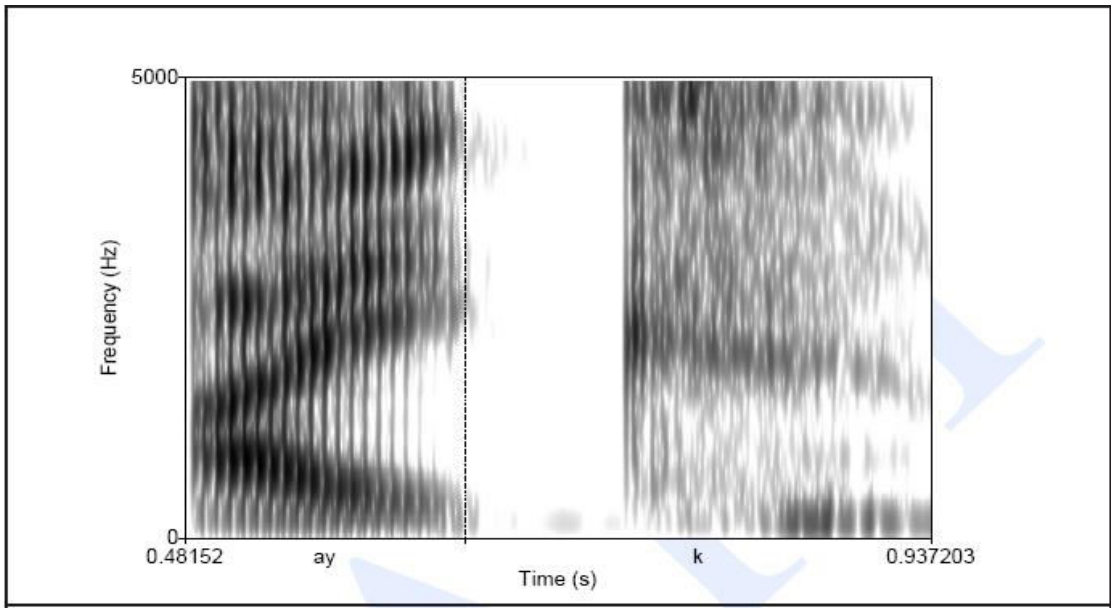


Figure III.2 les deux phones du mot « ike », prononcé [ay k]. Remarquer les changements continus dans la voyelle [ay] à gauche, vu que F2 s'élève et F1 descend, ainsi que la différence nette entre le silence et le relâchement des parties du [k] stop.

Pour capturer ce fait de nature non-homogénéité des phones à travers le temps, dans la RPCGV généralement, nous modélisions un phone avec plus d'un état HMM. La configuration la plus commune est d'utiliser 3 états HMM : un état début, milieu et fin. Chaque phone alors est formé de 3 états HMM émettant au lieu d'un seul (plus deux états non émettant).

Il est ordinaire de réserver le modèle de mot ou le modèle de phone pour soumettre au HMM un phone entier de 5-état et utiliser le mot état du HMM (ou juste état pour faire court) pour référencier chacun des 3 états des sous-phones individuels.

Pour construire un HMM pour un mot entier utilisant ces modèles de phones plus complexe, nous pouvons remplacer chaque phone du modèle de mot de la figure III.2 avec un HMM de phone 3-état. La figure (III.3) montre le mot six étendu.

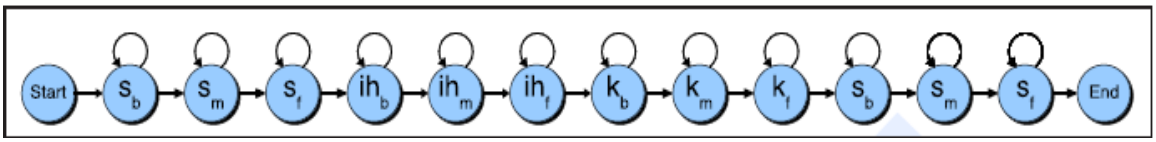


Figure III.3 une composition du modèle mot « six », [s ih k s], formée pour la concaténation de quatre modèles phone, chacun avec trois états émettant.

Nous avons maintenant traité la structure basique des états HMM pour représenter les phones et les mots dans la reconnaissance de la parole. Plus tard dans ce chapitre, plus d'augmentation du modèle HMM de mot montré dans la figure III.4

tel que l'utilisation de modèle tri-phones qui met l'utilisation du contexte de phone, et l'utilisation de phones spéciaux pour modéliser le silence.

D'abord, nous avons besoin de se retourner au prochain composant de HMM pour la reconnaissance de la parole : les probabilités d'observations et dans le but d'en discuter, nous avons besoin d'abord d'introduire les observations acoustiques actuelles : les vecteurs caractéristiques. Après avoir discuté dans la section III.2, nous revenons dans la section III.5 au modèle acoustique et les détails de calcul de ressemblance d'observations. Ensuite, on réintroduit le décodage de Viterbi et on montre comment le modèle acoustique et le modèle de langage sont combinés pour choisir la meilleure phrase.

III.4. Travaux connexes sur la reconnaissance de la parole arabe pour des mots connectés

Dans (Alotaibi, Alghamdi, & Alotaiby, 2010), un système est conçu pour reconnaître l'orthographe de l'alphabet arabe. Le HTK est utilisé pour le mettre en œuvre avec des modèles HMM basés sur des phonèmes. Puisque la plupart des alphabets arabes sont composés de plus de deux phonèmes, des modèles triphone dépendants du contexte ont été créés à partir de modèles monophoniques. Les modèles monophoniques ont été initialisés et formés par les données de formation pour les modèles triphone. Une méthode de modèle d'arbre de décision est utilisée pour aligner et lier le modèle avant la dernière étape de la phase de formation. Le modèle d'arbre de décision est la ré-estimation des paramètres HMM en utilisant l'algorithme Baum-Welsh trois fois. Le système de reconnaissance a obtenu 64,06 % de l'ensemble des alphabets corrigés en utilisant des ensembles mixtes de formation et de test collectivement.

Dans (Alotaibi, 2012), les auteurs ont comparé le réseau neuronal artificiel (ANN) aux modèles HMM dans la mise en œuvre du vocabulaire arabe limité ASR système avec des voyelles mot vecteurs. Le réseau neuronal artificiel récurrent a atteint 98,06 % tandis que la HMM a obtenu une reconnaissance de voyelle correcte de 91,6 %.

Nous notons que les ANN sont limitées aux petits vocabulaires, mais sont excellents dans ce cas. En opposition de HMM qui sont mieux dans les vocabulaires importants, mais pas dans les petits. L'outil HTK est utilisé à la fois dans les mots isolés et la reconnaissance vocale continue.

Nous appliquons l'approche HMM à la parole parce qu'ils sont des automates de Markov avec des états cachés; Il s'agit de modèles statistiques dans lesquels le système est modélisé en tant que processus markovien de paramètres inconnus(Rabiner, 1989). Ils sont adaptés à la reconnaissance des mots connectés parce que les HMM modélisent la parole sur le temps, contrairement aux modèles semi-markoviens qui conviennent à un très grand vocabulaire comme dans le travail (Park and Yoo 2014) où une performance de 90 % a été fournie.

Dans (El Moubtahij, Halli, & Satori, 2014), les auteurs présentent dans leur article un système de reconnaissance de texte en écriture arabe. Première étape, le système désintègre l'image du texte en images de ligne de texte. Deuxième étape, elle sélectionne certains attributs intelligibles à partir d'une fenêtre qui se déplace en douceur le long de la ligne de texte. Troisièmement, le système place les vecteurs d'attributs résultants dans la boîte à outils du modèle de Markov caché (HTK).

Dans (Abolohom & Omar, 2014), les auteurs établissent les caractéristiques appropriées à utiliser pour structurer une référence appelée "anaphore" en arabe. Afin d'exploiter les caractéristiques utiles, et compte tenu de leurs effets sur la présentation de la décision de l'anaphore, des caries expériences comparatives sur corpus Coran sont accompagnés. Le résultat sur l'ensemble de données de formation du Coran arabe montre que l'approche proposée est applicable à la décision de référence structurelle de l'arabe.

Dans (Hassine, Boussaid, & Messaoud, 2016), les auteurs ont classé deux dialectes maghrébins: le marocain et le tunisien, une fois par le vecteur support machine classificateur (SVM), et un autre temps par leur premier réseau de neurones de propagation en arrière (FFBPNN). Afin d'identifier dix chiffres arabes (de zéro à neuf), les résultats expérimentaux ont montré que la méthode FFBPNN dépasse celle des SVM car le taux de reconnaissance a atteint 98,3 % pour la première méthode contre 97,5 % pour la seconde respectivement.

Dans(Nahar, Shquier, Al-Khatib, Al-Muhtaseb, & Elshafei, 2016), à des fins de formation et de test, un nouveau corpus arabe enregistré en TV est réalisé. Les auteurs proposent dans leur article un système de reconnaissance hybride constitué de la HMM et de la Quantification vectorielle (LVQ). Le taux de reconnaissance des phonèmes a atteint 72 % avec LVQ seul, mais a atteint 89 % avec leur système hybride LVQ / HMM.

Dans le travail de(Adetunmbi, Obe, & Iyanda, 2016), la reconnaissance de mots isolés est enregistrée par les utilisateurs avec un système de parole-au-texte dans la langue Yoruba standard. Le système utilise une méthode basée sur la syllabe. La précision totale pour les mots bi-syllabiques était de 76 % et 84 % pour les mots trisyllabiques.

Enfin, les travaux qui se rapprochent le plus de notre travail de recherche:

Le travail de (Baig, Qazi, & Kadri, 2015), dans lequel les auteurs introduisent l'avancement de la reconnaissance de la lecture du Saint Coran. Deux techniques sont utilisées à cette fin: premièrement, le critère du maximum de vraisemblance (ML) appliqué à l'estimation des paramètres HMM. Il produit une précision de reconnaissance allant jusqu'à 83 %. La deuxième technique est l'erreur minimale du phone, ce qui réduit l'erreur de reconnaissance au niveau du phonème. Le résultat montre une amélioration de 3 à 4 % par rapport à la technique ML seule.

Dans le travail de (Zarrouk, Benayed, & Gargouri, 2015), les auteurs proposent pour le discours continu basé sur les tri-phones en arabe un système hybride de haute performance composé de SVM et Réseaux Bayésiens Dynamiques (DBN). Leur système SVM / DBN a

donné les meilleurs résultats de 78,87 % dans leur expérimentation. Le SVM / DBN WER est seulement 8,04 %, alors qu'il est de 10,58 % avec les systèmes Gaussien DBN de mélange trifonctionnel, 10,54 % avec hybridation SVM / HMM et 12,03 % dans les modèles HMM standard.

Dans (Mourtaga, Sharieh, & Abdallah, 2007), les auteurs ont fait un travail qui est proche du nôtre. Ils ont développé leur système en deux étapes: d'abord, ils ont modélisé le mot entier par treize états de gauche à droite HMM. La deuxième étape est le développement de la technique de la LRM (Maximum Linear Regression) pour l'adaptation afin d'estimer un ensemble de modèles transformés. Cela minimise la différence entre le modèle actuel et le cluster. Ils ont utilisé les 30 derniers surats du Saint Coran, avec cinq lecteurs célèbres. La précision des données testées était de 68 % à 85 % après l'adaptation.

Dans le chapitre précédent, nous avons défini l'extraction de caractéristique cepstrale : MFCC les plus utilisés, le lexique étant le dictionnaire, nous avons utilisé le modèle de langage bi-gram car il correspond à un texte moyen, pas très grand. Nous avons utilisé le HMM comme modèle acoustique dans une première expérience pour représenter le phonème. Dans une seconde expérience, le HMM représente le triphone.

Il nous reste à expliquer comment les GMM représentent les triphones de l'expérience précédente après les avoir liés dans une troisième expérience, ainsi qu'expliquer le décodage de Viterbi. Pour ce, commençons par donner comment faire le calcul de probabilités acoustiques dans notre cas qui est le calcul de fonction de densité de probabilités (FDP) qui sont les GMMs. Ensuite, nous verrons les modèles acoustiques dépendant du contexte appelés : Triphones.

III.5. Application du Modèle de mélange Gaussiens : calcul de probabilités acoustiques

Le chapitre précédent a montré comment peut-on extraire les caractéristiques MFCC représentant une information spectrale d'une forme d'onde, et produire un vecteur 39-dimensionnel chaque 10 millisecondes. Nous sommes prêts maintenant à calculer la ressemblance de ces vecteurs caractéristiques sachant un état HMM. Rappelons que cette ressemblance en sortie est calculée par la fonction de probabilité B du HMM(Martin & Jurafsky, 2000). Etant donné un état individuel q_i et une observation o_t, les ressemblances d'observation dans la matrice B nous a donné $p(o_t/q_i)$, que nous avons appelé $b_t(i)$.

Pour l'étiquetage de partie-de-parole, chaque observation o_t est un symbole discret (un mot) et on peut calculer la ressemblance d'une observation ayant une étiquette de partie-de parole juste en comptant le nombre de fois que génère une étiquette une observation donnée dans l'ensemble d'apprentissage. Mais pour la reconnaissance de parole, les vecteurs MFCC sont des nombres en réel ; on ne peut pas calculer la probabilité d'un état donné (un phone) générant un vecteur MFCC en comptant le nombre de fois où chaque vecteur se produit (puisque chacun est unique).

Dans le décodage et l'apprentissage, on a besoin d'une fonction de probabilité d'observation qui calcule $p(o_t|q_i)$ sur les observations en valeurs réels. Dans le décodage, on compte tenu d'une observation o_t nous avons besoin de produire la probabilité $p(o_t|q_i)$ pour chaque état de HMM possible, donc on peut choisir la séquence d'états la plus probable. Une fois on a cette fonction de probabilité d'observation B, nous avons besoin de réfléchir comment modifier l'algorithme de Baum-Welch pour l'entrainer dans le cadre d'apprentissage de HMMs.

III.5.1. La quantification vectorielle

Une manière de faire ressembler les vecteurs MFCC à des symboles qu'on peut compter est de construire une fonction de traçage qui marque chaque vecteur entrée en un petit nombre de symbole. Puis on peut juste calculer les probabilités de ces symboles en comptant, comme on l'a fait pour les mots dans l'étiquetage des parties-de-paroles. Cette idée de traçage des vecteurs entrés en symboles quantifiés discrets s'appelle la quantification vectorielle (QV). Bien que la quantification de vecteurs soit trop simple pour agir tel un modèle acoustique dans les systèmes de RPCGV modernes, c'est une étape pédagogique utile, et joue un rôle important dans différents domaines de RAP, donc nous l'utilisons pour commencer notre discussion de modélisation acoustique.

Dans la quantification de vecteurs, on crée un petit ensemble symboles par le marquage de chaque vecteur caractéristique d'apprentissage en un petit nombre de classes, et on représente alors chaque classe par un symbole discret. Plus formellement, un système de quantification vectoriel est caractérisé par un *livre de code*, un *algorithme de clustering* et une *métrique de distance*.

Un *livre de code* est une liste de classe possible, un ensemble de symboles constituant un vocabulaire $v=(v_1,v_2,...,v_n)$. Pour chaque symbole v_k dans le livre de code on liste un *vecteur prototype*, connu aussi tel un *mot codé*, qui est un vecteur caractéristique spécifique. Par exemple si on choisit d'utiliser un mot encodé de 256, on peut représenter chaque vecteur par une valeur de 0 à 255 (ceci est référencié à une QV 8 bit unique). Chacune de ces 256 valeurs doit être associée à un vecteur prototype.

Un livre de code est créé en utilisant un algorithme de *clustering* pour regrouper tous les vecteurs caractéristiques dans l'ensemble d'apprentissage en 256 classes. Puis on choisit un vecteur caractéristique représentatif du cluster, et faire le vecteur prototype ou le mot codé pour ce cluster. Le *clustering k-means* est souvent utilisé, mais nous n'allons pas définir ici le clustering ; voir (Huang, Acero, Hon, & Foreword By-Reddy, 2001) OU (Duda, Hart, & Stork, 2000)pour les descriptions détaillées.

L'avantage de la QV est que puisqu'il y a un nombre fini de classes, pour chaque classe v_k on peut calculer la probabilité générée par un état/sous-phone de HMM donné, simplement en comptant le nombre de fois qu'elle se produit dans quelques ensembles d'apprentissage quand elle est étiquetée par cet état et normalisé.

Le processus de clustering ainsi celui du décodage nécessite une métrique de distance ou une *métrique de distorsion* qui spécifie combien deux vecteurs caractéristiques acoustiques sont similaires. La métrique de distance est utilisée pour construire les clusters, pour trouver un vecteur prototype pour chaque cluster, et pour comparer les vecteurs entrés aux prototypes.

La métrique de distance la plus simple pour les vecteurs caractéristiques acoustiques est la *distance Euclidienne*. La distance Euclidienne est la distance en l'espace de N dimensions entre les deux points définis par les deux vecteurs. Par conséquent, étant donné un vecteur *x* et un vecteur *y* de taille D, la distance Euclidienne quadrillée entre eux est défini tel que :

$$d_{euclidienne}(x,y)=\sum_{i=1}^{D}(x_i - y_i)^2 \ldots\ldots \text{(III.1)}$$

La distance Euclidienne quadrillée décrite au-dessus est aussi référencié à l'erreur de somme quadrillée, et peut être aussi exprimée en utilisant l'opérateur transposé de vecteur tel que :

$$d_{euclidienne}(x,y)=(x-y)^T(x-y) \ldots\ldots \text{(III.2)}$$

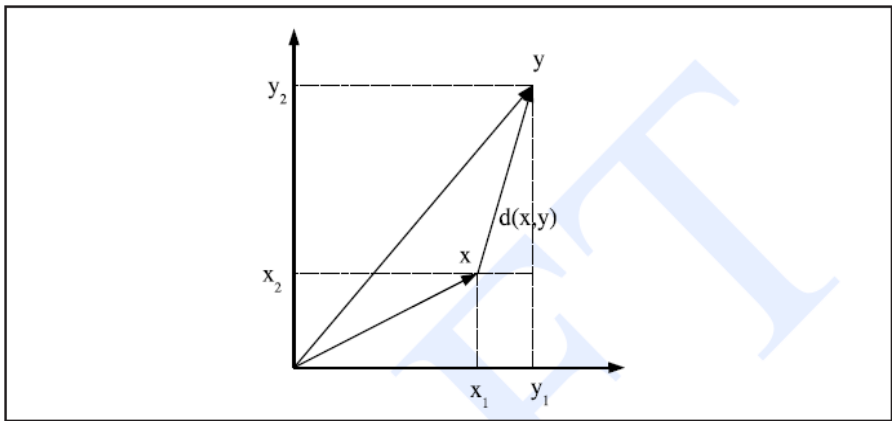


Figure III.4 La distance euclidienne dans les deux dimensions ; par le théorème de Pythagore. La distance entre les deux points dans le plan $x=(x_1,y_1)$ et $y=(x_2,y_2)$ est

$$d(x,y)=\sqrt{(x1-x2)^2+(y1-y2)^2} \ldots\ldots \text{(III.3)}$$

La métrique de distance Euclidienne suppose que chacune des dimensions du vecteur caractéristique sont tout aussi importantes. Mais chacune des dimensions ont de différentes variances. Si une variance a tendance à avoir beaucoup de variances, on voudrait qu'elle compte moins dans la métrique de distance ; une grande différence dans une dimension avec peu de variances doit compter plus qu'une grande différence dans une dimension avec beaucoup de variances. Une métrique de distance plus complexe *la distance Mahalanobis*, prend en compte les différentes variances de chacune des dimensions.

Si on suppose que chaque dimension *i* des vecteurs caractéristiques acoustiques a une variance σ_i^2, alors la distance Mahalanobis est :

$$d_{mahalanobis}(x,y)=\sum_{i=1}^{D}\frac{(x-y)^2}{\sigma_i^2}\ldots\ldots \text{(III.4)}$$

Pour les lecteurs avec plus de profondeur dans l'algèbre linéaire, voici la forme générale de la distribution Mahalanobis, qui incluse une matrice de covariances :

$$d_{mahallanobis}(x,y)=(x-y)^T\Sigma^{-1}(x-y) \ldots. \text{(III.5)}$$

Puisque la quantification vectorielle est tellement rarement utilisée, nous ne donnons pas ici les équations pour modifier l'algorithme EM (expectation Maximisation) pour faire face aux données de QV, au lieu de cela nous changeons la discussion de l'apprentissage EM des paramètres entrées continues à la prochaine section, lorsqu'on introduit les Gaussiennes(Martin & Jurafsky, 2000).

Jusqu'à maintenant, toutes les équations que nous avons données pour la modélisation acoustique ont utilisé les probabilités. Tandis qu'une *probabilité log ou (logprob)* est tellement plus facile à travailler avec qu'avec une probabilité simple. Donc, en pratique tout au long de la reconnaissance de la parole (ainsi que dans les domaines reliés), nous calculons les probabilités log au lieu des probabilités(Martin & Jurafsky, 2000).

Une raison majeur de ne pas pouvoir utiliser les probabilités est le débordement numérique.

III.5.2. Fonction de densité de probabilités Gaussiennes

Les algorithmes de la reconnaissance de parole moderne sont basés sur le calcul des probabilités d'observations sur des valeurs réelles, sur un vecteur caractéristique d'entrée continue. Les modèles acoustiques sont basés sur le calcul de la fonction de densité de probabilité ou FDP sur un espace continu.

De loin, la méthode la plus commune pour calculer les probabilités acoustiques est la FDP du Modèle de Mélange Gaussien (GMM), même si les réseaux de neurones et les machines à vecteurs de support (SVMs) sont aussi utilisés.

Commençons avec la simple utilisation de l'estimateur de probabilité gaussienne, en construisant petit à petit les modèles les plus sophistiqués qui sont utilisés.

- **Gaussienne uni-variée**

La distribution Gaussienne, aussi connue telle la distribution normale, est la fonction courbe en cloche familière des statistiques basiques. Une distribution gaussienne est une fonction paramétrée par une moyenne, ou une valeur moyenne, et une variance qui caractérise la diffusion de la moyenne ou la dispersion de la moyenne. Nous allons utiliser

μ pour indiquer la moyenne, et σ2 pour indiquer la variance, donnant la formule suivante pour la fonction Gaussienne :

$$f(x/\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \text{ ……. (III.6)}$$

- **Gaussienne multi-variée**

L'équation (III.6) montre comment utiliser une gaussienne pour calculer une ressemblance acoustique pour une caractéristique cepstrale single. Puisque une observation acoustique est un vecteur de 39 caractéristiques, nous aurons besoin de gaussiennes multi-variée, qui nous permet d'assigner une probabilité à un vecteur de 39 valeurs.

Vue qu'une gaussienne uni-variée est définie par une moyenne μ et une variance σ2, une gaussienne multi-variée est définie par un vecteur de moyenne μ ($\vec{\mu}$) de dimension D et une matrice de covariance Σ , définie en dessous. Pour un vecteur de caractéristique cepstrale typique dans les RPCGV, D est 39 :

$$f(\vec{x}/\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \text{…… (III.7)}$$

La matrice de covariances Σ capture les variances de chaque dimension aussi bien que la covariance entre deux dimensions. Rappelons que la covariance de deux variables aléatoires X et Y est la valeur attendue du produit de leurs déviations de la moyenne :

$$\Sigma = \mathrm{E}\big[\big(X-E(X)\big)(Y-E(Y))\big] = \sum_{i=1}^{N} p(X_iY_i)(X_i-E(X))(Y_i-E(Y) \text{…..(III.8)}$$

En gardant une variance séparée pour chaque dimension est équivalent à avoir une matrice de covariance qui et diagonale, c'est-à-dire des éléments non zéros seulement qui apparaissent tout au long de la principale diagonale de la matrice. La principale diagonale d'une telle matrice de covariance diagonale contient les variances de chaque dimension, σ^2_1 , σ^2_2, … σ^2_D.

Ayant une matrice de covariance non diagonale nous permet de modéliser les corrélations entre les valeurs des caractéristiques dans multiples dimensions.

- **Les modèles de mélange Gaussien**

La sous-section précédente a montré que nous pouvions utiliser un modèle de gaussienne multi-variée pour assigner un score de ressemblance à un vecteur d'observation de caractéristique acoustique. Ceci modélise chaque dimension du vecteur de caractéristique telle une distribution normale. Mais une caractéristique cepstrale particulière pourrait avoir une distribution non normale ; l'assomption d'une distribution normale peut être très forte. Pour cette raison, nous modélisons souvent l'assomption de ressemblance d'observation pas avec une gaussienne multi-variée single, mais avec un

mélange pondéré de gaussienne multi-variée. Un tel modèle est appelé un modèle de mélange gaussien GMM.

$$f(x/\mu,\Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp[(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)] \ldots. \text{(III.9)}$$

L'équation (III.9) montre la fonction de GMM ; la fonction résultante de la somme de *M* gaussiennes.

III.5.2.1. Modèles acoustiques dépendant du contexte : Triphones

Il y un problème lorsque nous utilisons un GMM fixe pour un sous-phonème. Le phonème est représenté par trois états du HMM selon le contexte de sa prononciation : au début, milieu ou à la fin du mot. Exemple : /$a_{début}$/, /a_{milieu}/, /a_{fin}/. Le problème est que ces phonèmes varient beaucoup.

Ceci est à cause du mouvement des articulateurs (langue, lèvre, palais) durant la production de parole continue est sujette aux contraintes physiques comme l'élan. Ainsi, un articulateur peut commencer le mouvement durant un phonème pour mettre en place à temps le prochain phonème. Nous définissons le mot *coarticulation* comme le mouvement d'articulateurs pour anticiper le prochain son, ou persévérer le mouvement à partir du dernier son (Martin & Jurafsky, 2000).

Dans le but de modéliser la variation marquée qu'un phonème expose dans différent contextes, la plupart des systèmes à RPGV remplace l'idée de modèle HMM indépendant du contexte (CI) par des phonèmes dépendants du contexte (CD). Le modèle dépendant du contexte le plus commun est le modèle de HMM *triphone* (Schwartz et al., 1985). Un modèle triphone représente un phonème dans un contexte droit ou gauche particulier. Par exemple le triphone/a-b+i/ veut dire le phonème /b/ précédé par le phonème /a/ et suivi par le phonème /i/. Dans les situations où nous n'avons pas un triphone complet, nous allons utiliser /a-b/ pour dire /b/ est précédé par /a/ et /b+c/ pour dire /b/ est suivi de /c/.

- **Groupement par arbre de décision**

Une propriété majeure d'arbre de décision est la capacité de produire des modèles non visibles durant l'apprentissage acoustique. Les arbres de décisions construits sont stockés afin de les utiliser dans le futur. En faisant référence à ces arbres, les modèles invisibles peuvent être construits artificiellement.

Les arbres de décisions sont construits en demandant un ensemble de questions binaires c'est-à-dire : la réponse est oui ou non, en regroupant les états. Ces questions sont demandées au branchement des nœuds de l'arbre. Initialement tous les états à regrouper appartenant au modèle, sont placés au nœud racine de l'arbre. Comme les questions sont demandées l'espace des états est divisé en deux comme montré dans la figure (III.6). Le nombre d'états dans la moitié de l'espace dépend de la réponse « oui » ou « non ».

Partageant l'espace de l'état finit lorsque le critère d'optimalité est rencontré. Ce critère dépend de deux différents seuils définis par l'utilisateur. Le seuil peut être le

minimum d'état qui doit être dans un groupe ou peut être l'augmentation dans la ressemblance totale des données d'apprentissage dans l'ensemble courant des groupes. Ensuite l'ensemble courant les états dans le groupe sont liés.

Un exemple d'arbre de décision construit pour des modèles tri-phones peut être vu dans la figure (III.6) (Young et al., 2006). Les questions utilisés dans l'exemple sont comme « est-ce le contexte droit une consonne centrale ? » ou « est-ce le contexte gauche une nasale ? ». Ces questions pour un système basé sur les tri-phones peuvent être notées comme :

QS L_nasale {*-a+*}

QS R_consonne {*-a+*}

Le groupement termine non pas suivant le nombre de questions mais suivant l'optimalité du critère.

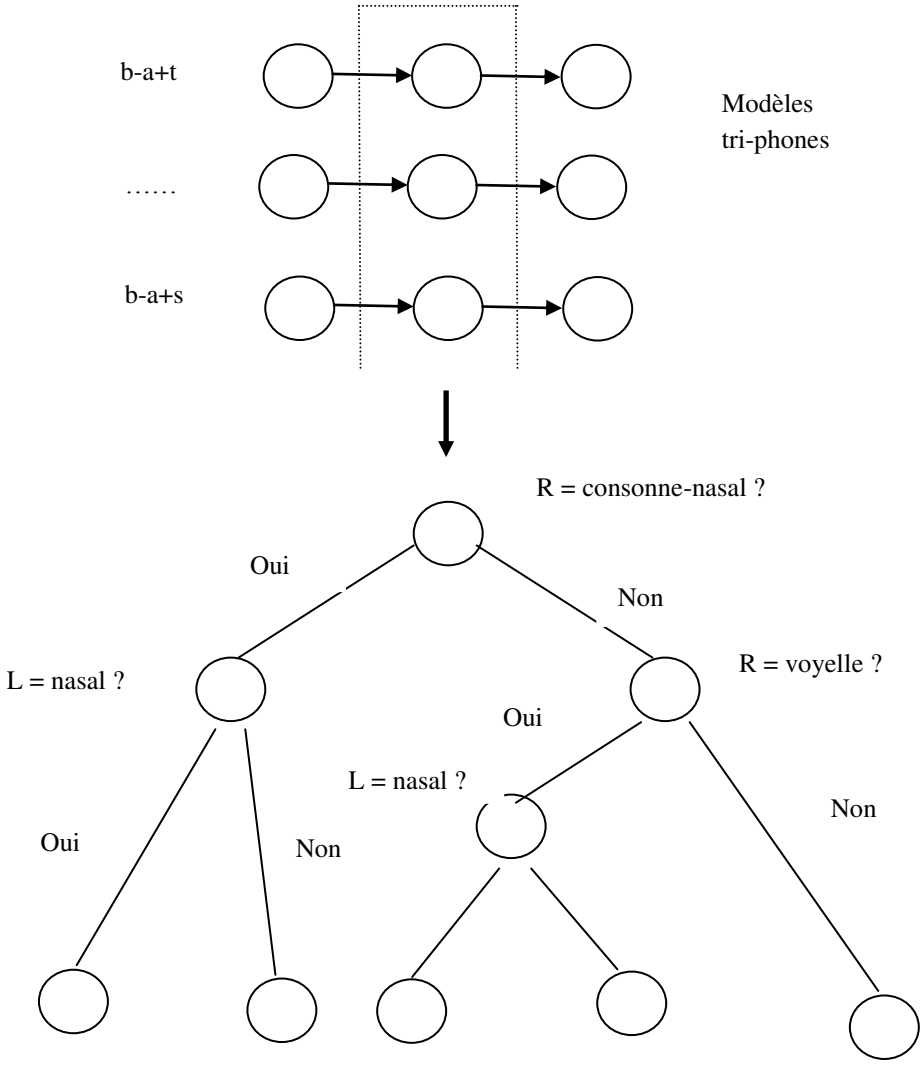


Figure III.5: groupement des états centre du phone « a »

- **Algorithme proposé : Triphones élargis aux GMMs**

Notre contribution consiste à appliquer des modèles acoustiques dépendants du contexte appelés triphones étendus aux GMMs en quatre étapes(Martin & Jurafsky, 2000), puis comparer notre algorithme avec un travail qui a utilisé un autre algorithme qui est la régression linéaire de ressemblance maximum (MLLR)(Mourtaga et al., 2007). Cela afin d'améliorer le taux de reconnaissance de la base de données coranique après l'adaptation du système à reconnaitre la parole indépendamment du locuteur.

La première étape consiste à former des mono-phones gaussiens uniques.

La deuxième étape consiste à cloner ces mono-phones en tri-phone.

Troisièmement, triphone de grappe par arbre de décision puis les lier.

Quatrièmement: les étendre aux modèles de mélange gaussien GMM.

Le pseudo-code de l'algorithme est donné ci-dessous:

1. Former des modèles monophoniques avec une seule gaussienne.

2. Clonez ces mono-phones aux tri-phones.

3. Attachez les états de ces triphones en fonction de la vraisemblance caractéristique avec les arbres de décision. Cela regroupe les états des sous-phonèmes : « début », « milieu », « fin » ensemble.

4. Développer les triphones en GMMs en remettant les trois étapes précédentes: former chaque tri-phone à état lié avec une Gaussien unique, cloner tous ses états dans deux gaussiennes identiques, perturber leurs valeurs un peu par un certain epsilon avec la commande « HHEd » en lui donnant la liste contenant tous les phonèmes représenté par le nombre de gaussienne souhaités mixés grâce à la commande « MU ». Exécuter une itération de ré-estimation pour former ces valeurs avec " HERest ".

Nous refaisons ces trois procédures : cloner tous les états à deux mélanges Gaussiens, perturber leurs valeurs, ré-entraîner et on continue à augmenter le nombre de GMM jusqu'à avoir un nombre approprié de mélanges pour la quantité d'observations dans chaque état.

La figure (III.7) clarifie cet algorithme.

Nous pouvons utiliser un modèle de mélange Gaussien pour attribuer un score de vraisemblance à une observation de vecteur caractéristique acoustique(Martin & Jurafsky, 2000). Ceci modélise chaque dimension du vecteur caractéristique en tant que distribution normale. Cependant, une caractéristique cepstrale particulière peut avoir une distribution très non normale.

L'hypothèse d'une distribution normale peut être trop forte. Pour cette raison, nous modélisons souvent la probabilité d'observation non pas avec une seule Gaussienne multi-variée, mais avec un mélange pondéré de Gaussiennes.

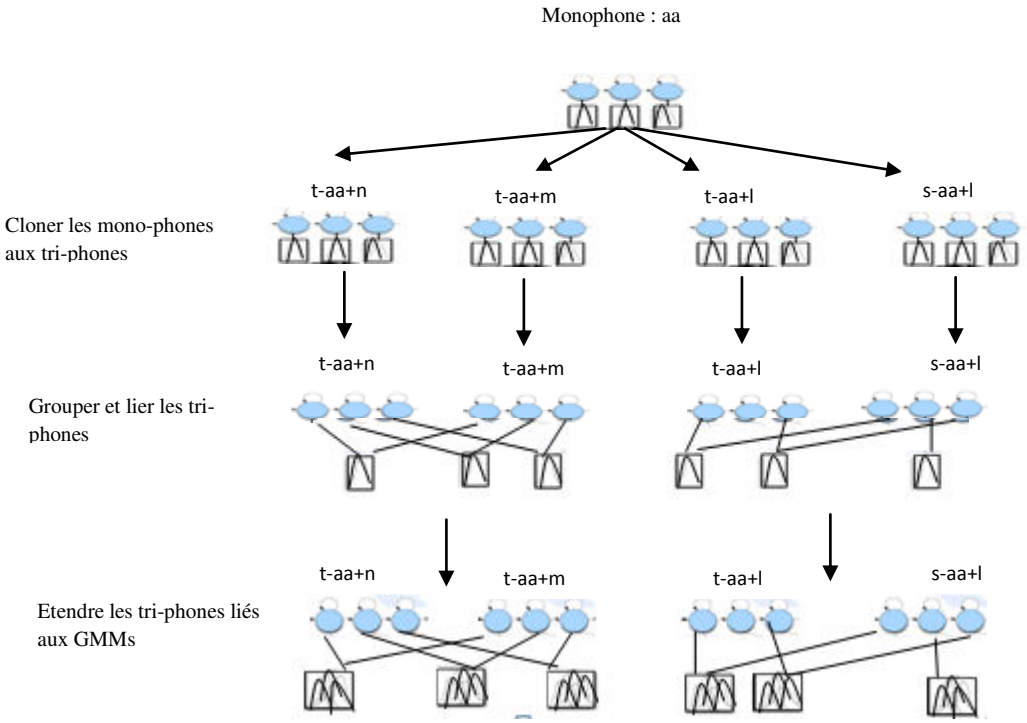


Figure IV.6 : les quatre étapes dans l'apprentissage d'un modèle acoustique de tri-phone mélangé et lié.

Nous allons voir dans la prochaine section le décodage de Viterbi. Avant cela, nous présentons les trois problèmes lorsque nous construisons les HMMs. Les voici :

Problème d'évaluation :

La question qui se pose ici est comment calculer la probabilité de la séquence d'observation en donnant le modèle P (O/λ) ? La solution à ce problème nous permet d'évaluer la probabilité de différents HMMs en générant la même séquence d'observations si nous avions un HMM différent pour chaque mot, alors le mot reconnu est celui à qui le modèle a la plus grande probabilité des données générées.

Problème de décodage :

En donnant une séquence de sortie θ et un modèle λ, comment calculons la séquence d'états Q la plus probable ? Ce problème est concerné avec la découverte de séquence d'états cachés en connaissant les symboles de sortie par le système.

Problème d'apprentissage :

Comment ajuster les paramètres du modèle A, B et Π afin de maximiser la ressemblance du modèle λ produisant la séquence de sortie ? Pour la solution de ce problème, nous avons besoin des données d'apprentissage.

Algorithme marche avant-arrière :

Le nombre total des chemins possibles conduisant au dernier état de l'automate augmente exponentiellement en augmentant le nombre d'états et les instances d'observations. La réduction du coût de calcul peut être atteinte par la procédure de marche avant-arrière. Actuellement elle est composée de deux procédures : avant et arrière. Dans le cas d'évaluation, nous avons besoin de l'une d'elle. La procédure arrière est utilisé dans la solution d'apprentissage c'est-à-dire dans l'algorithme de Baum-Welch.

Initialement, définir une nouvelle variable de probabilité avant $\alpha_t(i)$ à l'instant *t* et à l'état *i* :

$$\alpha_t(i) = P(O_1 O_2 \ldots\ldots O_t\, , q_t = S_i/\lambda) \ldots\ldots \text{(III.10)}$$

Cette fonction de probabilité pourrait être résolue pour N états et T observations itérativement :

1- Initialisations

$$\alpha_1(i) = \pi_i b_i(\theta_1) \quad 1 \leq i \leq N$$

2- Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\theta_{t+1}) 1 \leq t \leq T-1 \; et \; 1 \leq j \leq N$$

3- Terminaison

$$\mathrm{P}\,(\mathrm{O}/\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Le stage de la terminaison est juste la somme de toutes les valeurs de la fonction de probabilité α_t (i) sur tous les états à l'instant T. les résultats présentent comment le modèle donné produit probablement les observations données :

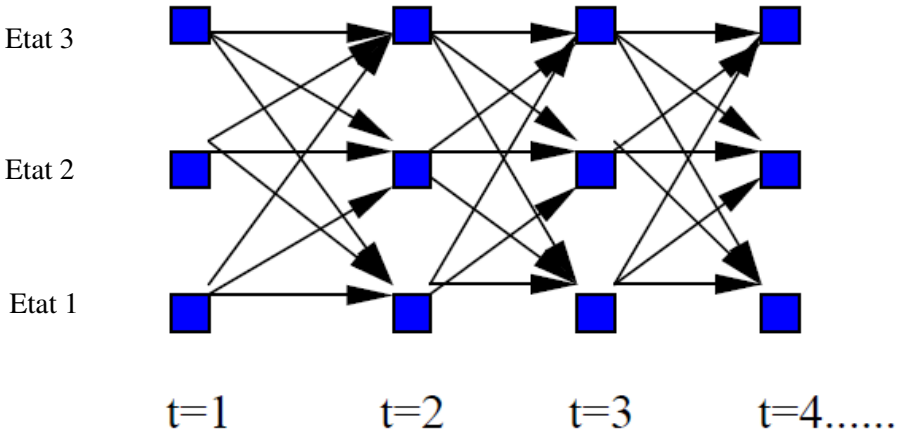


Figure III.7: le flux du processus d'un HMM dans le temps

La procédure arrière est similaire à celle d'avant. Par contre, l'état qui se déroule dans le calcul est arrière à partir de l'instant T jusqu'à l'instant 1. Définissons une fonction de probabilité arrière β_t (i) :

$$\beta_t(i) = P(O_{t+1}O_{t+2} \ldots\ldots O_T \,/q_t = S_i, \lambda) \ldots\ldots\ldots \text{(III.11)}$$

1- Initialisations

$$\beta_T(i) = 1 \quad 1 \leq i \leq N$$

Ces valeurs initiales pour les β de tous les états à l'instant T peuvent être arbitrairement sélectionnées.

2- Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(O_{t+1})\beta_{t+1}(j) \quad t = T-1, T-2..1\,;\; 1 \leq i \leq N$$

La probabilité P (O/λ) peut être calculée des deux fonctions avant et arrière. Ceci est illustré dans la figure (III.9). La probabilité avant est une probabilité de jointure alors que la probabilité arrière est une probabilité conditionnelle. La probabilité de l'occupation de l'état est déterminée en prenant compte du produit des deux probabilités.

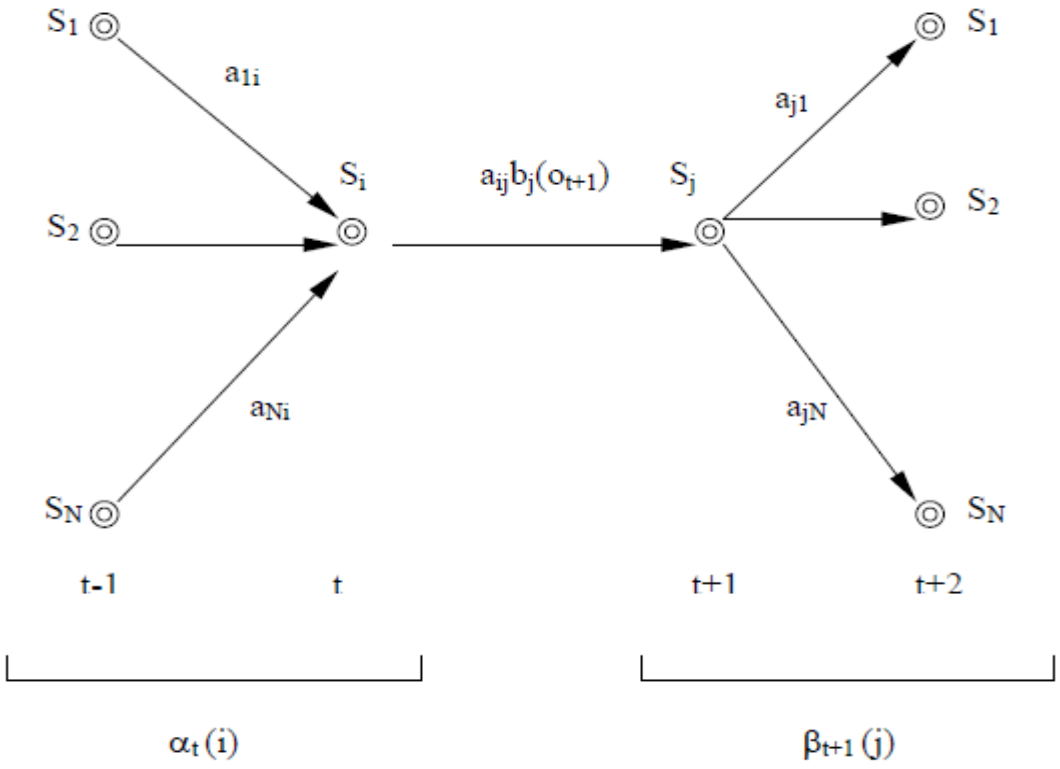


Figure III.8 : fonctions de probabilités avant-arrière pour obtenir P (O/λ)

Algorithme de Viterbi :

En solution du problème de décodage, l'algorithme populaire de Viterbi est utilisé. Le critère d'optimalité ici est de chercher pour la meilleure séquence d'états à travers la technique de programmation dynamique modifiée. L'algorithme de Viterbi est un algorithme de recherche parallèle, notamment il recherche la meilleure séquence d'états en traitant tous les états en parallèle. Nous avons besoin de maximiser P (Q/ O, λ), pour détecter la meilleure séquence d'états. Définissons une quantité de probabilité δ_t (i) qui représente la probabilité maximum tout au long du meilleur chemin de la séquence d'états à partir d'une séquence d'observations donné après t instants et en étant dans l'état i :

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \; et \; \psi_1(i) = 0 \text{.....(III.12)}$$

La meilleure séquence d'états est une marche arrière par une autre fonction ψ_t (j). Cette fonction tient l'indice de l'état à l'instant t-1, duquel la meilleure transition est faite à l'état courant. L'algorithme complet est donné tel que :

1- Initialisations

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \; et \; \psi_1(i) = 0$$

2- Récursivité

$$\delta_t(j) = max[\delta\_(t-1)\,(i)\; a\_ij\;]b_j(O_t) \quad 2 \le t \le T, 1 \le j \le N$$

$$\psi_t(j) = arg\max[\delta_{t-1}(i)a_{ij}] \quad 2 \le t \le T, 1 \le j \le N$$

3- Terminaison

$$p* = \max \delta_T(i)$$

$$q_T^* = argmax\ \delta_T(i)$$

4- Rétrogradation

$$q_T^* = \psi_{t+1}[q_{t+1}^*] \quad 1 \le t \le T-1$$

Il est clair que la récursivité Viterbi est similaire que l'induction ou incorporation de l'algorithme marche avant, excepté l'échange de somation par maximisation. Donc, il est clair que P (O/λ) peut être calculé approximativement avec l'algorithme Viterbi en prenant p* comme score. Ceci est illustré dans la figure (III.10).

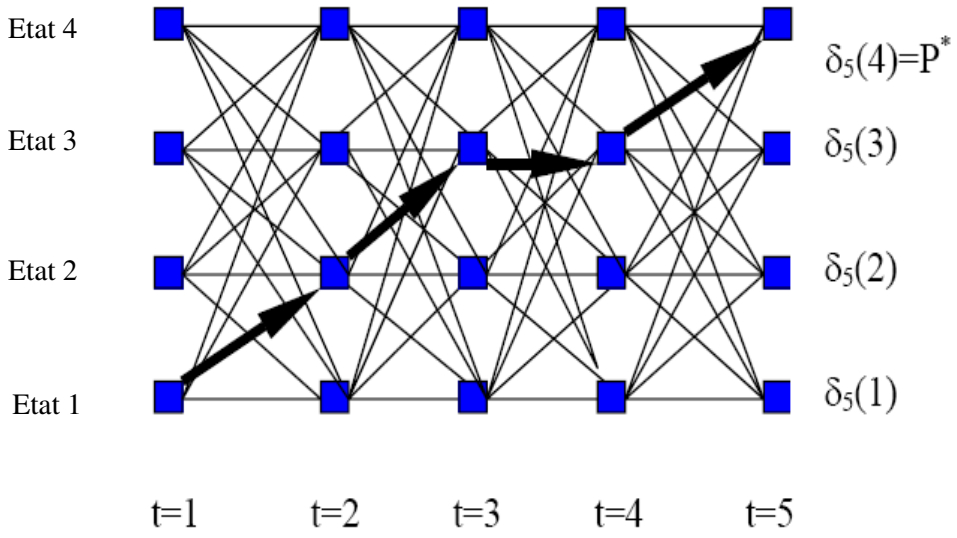


Figure III.9 : Flux de l'algorithme de Viterbi. Le meilleur chemin est en gras

Algorithme de Baum-Welch :

Cet algorithme est relié au problème d'apprentissage qui est le plus difficile. Le but est d'ajuster les paramètres du modèle selon un critère optimal. L'algorithme de Baum-Welch est strictement relié à l'algorithme avant-arrière et il essaie d'atteindre le maximum local de la probabilité de fonction P (O/λ). Le modèle converge toujours mais la maximisation globale n'est pas garantie. C'est pour cela que le point initial de la recherche est vraiment important.

Définissons la probabilité d'être à l'état Si à l'instant t, et Sj à l'instant t+1.

$$\varepsilon_t(i;j) = P(q_t = S_i, q_{t+1} = S_j/O,\lambda)\text{……..(III.13)}$$

Sa relation avec les variables avant et arrière est :

$$\varepsilon_t(i;j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}\text{……..(III.14)}$$

Le terme numérateur est égal à P(q_t=S_i,q_(t+1)=S_j,O/λ) et le terme dominateur est égal à P(O/ λ). Maintenant, définissons la probabilité à postériori d'être dans l'état Si à l'instant t, en donnant la séquence d'observation et le modèle :

$$\gamma_t\ (\mathrm{i}) = P\ (\mathrm{q_t} = \mathrm{S_i}/O,\lambda)\text{……(III.15)}$$

Sa relation avec les variables avant et arrière est:

$$\gamma_t = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}\text{……(III.16)}$$

Si on fait la somme des γ_t(i) sur l'indice t, nous avons la quantité qui peut être interprété tel que le nombre prévu de fois où l'état Si est utilisé et la somme des ε_t (i;j) sur t peut être interprété comme le nombre prévu de transitions faite de Si à Sj.

Avec l'aide de ceux-ci les formules de ré-estimations des paramètres A, B et Π sont donné tel que :

$\overline{\pi_\iota}$ = Nombre attendu de fois dans l'état Si à l'instant (t=1)= γ_1(i)

$$\overline{a_{\iota j}} = \frac{nombre\ prévu\ de\ transitions\ \text{de}\ Si\ \text{à}\ Sj}{nombre\ prévu\ de\ transitions\ depuis\ Si}$$

$$= \frac{\sum_{t=1}^{T-1}\varepsilon_t(i;j)}{\sum_{t=1}^{T-1}\gamma_t(\mathrm{i})}$$

Après le ré-estimation des paramètres du modèle, nous allons avoir un nouveau modèle qui est plus susceptible de produire la séquence d'observation O. la procédure de ré-estimation itérative continue jusqu' à ce que la non amélioration dans P(O/ λ) soit atteinte.

Il devrait être noté ici que les HMMs ont des limites :

1- La probabilité de transition dépend uniquement de l'origine et destination
2- Toutes les formes d'observations sont dépendantes seulement de l'état qui les génère, pas dans les fenêtres d'observations voisines.
3- Il n y a plus d'informations à propos de la durée de l'état.

Après avoir construit un modèle avec certains paramètres, certains raffinements doivent être faits. La solution proposée peut incrémenter le nombre de mélange de distribution de

probabilité bj(ot), ou grouper les états selon certains critères ou en liant les paramètres de différents modèles.

III.5.2.2. Décodage de Viterbi

En solution du problème de décodage, l'algorithme populaire de Viterbi est utilisé. Le critère d'optimalité ici est de chercher pour la meilleure séquence d'états à travers la technique de programmation dynamique modifiée. L'algorithme de Viterbi est un algorithme de recherche parallèle, notamment il recherche la meilleure séquence d'états en traitant tous les états en parallèle.

Nous avons besoin de maximiser P (Q/ O, λ), pour détecter la meilleure séquence d'états. Définissons une quantité de probabilité δ_t (i) qui représente la probabilité maximum tout au long du meilleur chemin de la séquence d'états à partir d'une séquence d'observations donné après t instants et en étant dans l'état i :

$$\delta_t(i) = \max \mathrm{P}[\mathrm{q}_1 \ldots. \mathrm{q}_{t-1}, \mathrm{q}_t = \mathrm{s}_i, \mathrm{o}_1 \ldots \mathrm{o}_t/\lambda] \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(III.17)}$$

La meilleure séquence d'états est une marche arrière par une autre fonction ψ_t (j). Cette fonction tient l'indice de l'état à l'instant t-1, duquel la meilleure transition est faite à l'état courant. L'algorithme complet est donné tel que :

Initialisation

$$\delta_1(i) = \pi_i b_i(\mathrm{O}_1) \quad 1 \leq i \leq N \; et \; \psi_1(i) = 0$$

Récursivité

$$\delta_t(j) = max[\delta\_(t-1)\ (i)\ a\_ij\ ] b_j(\mathrm{O}_t) \quad 2 \leq t \leq T, 1 \leq j \leq N$$

$$\psi_t(j) = arg \max\left[\delta_{t-1}(i) a_{ij}\right] \quad 2 \leq t \leq T, 1 \leq j \leq N$$

Terminaison

$$p *= \max \delta_T(i)$$

$$q_T^* = argmax\ \delta_T(i)$$

Rétrogradation

$$q_T^* = \psi_{t+1}[q_{t+1}^*] \quad 1 \leq t \leq T-1$$

Il est clair que la récursivité Viterbi est similaire que l'induction ou incorporation de l'algorithme marche avant, excepté l'échange de somation par maximisation. Donc, il est clair que P (O/λ) peut être calculé approximativement avec l'algorithme Viterbi en prenant p* comme score. Ceci est illustré dans la figure (III.11).

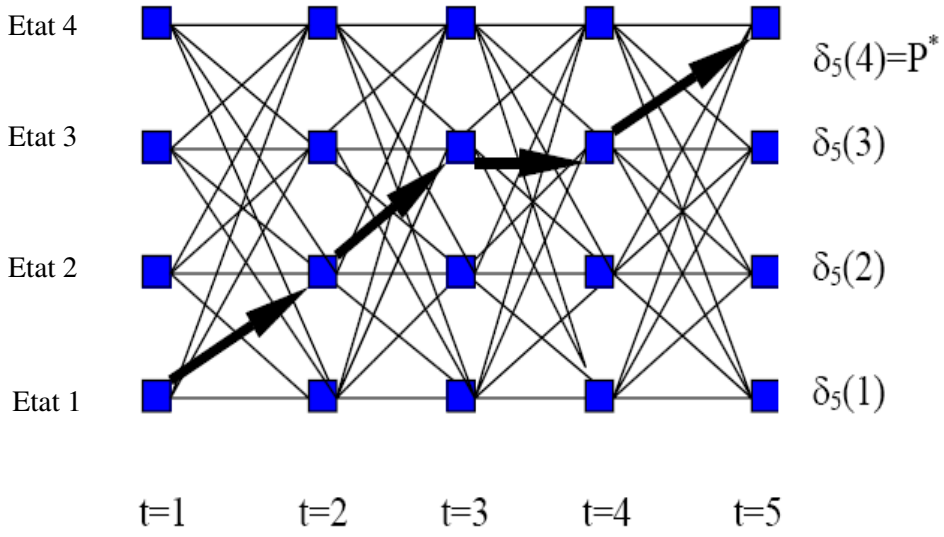


Figure III.10: Flux de l'algorithme de Viterbi. Le meilleur chemin est en gras (Çömez, 2003).

III.5.3. Apprentissage intégré

Nous retournons maintenant à voir comment un système de reconnaissance de parole basé HMM est entrainé. Nous avons déjà vu précédemment comment les modèles acoustiques GMM sont entrainés en augmentant l'algorithme EM pour faire face à l'apprentissage des moyennes, variances et poids. Nous avons vu aussi comment des classifieurs comme les SVMs et réseaux de neurones peuvent être entrainés, même si pour les réseaux de neurones nous n'avons pas encore vu comment on peut entrainer les données dans lesquelles chaque frame est étiqueté avec une identité de phone.

Dans cette section, nous complétons l'image de l'apprentissage HMM en montrant comment cet apprentissage EM s'emboîte dans le processus de modèles acoustique d'apprentissage.

Nous allons supposer que le lexique de prononciation et donc la structure de graphe état HMM basique pour chaque mot, est pré-spécifié telle une simple structure HMM linéaire avec des bouclages en chaque état, en général les systèmes de reconnaissance de parole ne tentent pas d'entrainer la structure des mots HMMs individuels. Ce qui fait que nous avons besoin de faire entrainer la matrice B seulement, et aussi faire entrainer les probabilités des transitions non-zéro (les auto-boucle et les sous-phones suivants) de la matrice A. toutes les autres probabilités dans la matrice A sont mises à zéro et ne changent et jamais.

La méthode d'apprentissage la plus simple est l'apprentissage de mot isolé étiqueté-manuellement, dans lequel on entraine les matrices A et B séparément pour les HMMs pour chaque mot basé sur l'apprentissage de donnée aligné manuellement.

Malheureusement, l'apprentissage de données segmenté manuellement est rarement utilisé dans les systèmes de parole continue. Une raison est qu'il est vraiment cher en utilisation d'humains pour étiqueter manuellement les limites phonétiques ; cela peut prendre jusqu'à 400 fois de temps réel (i.e. 400 heures d'étiquetage pour étiqueter chaque heure de parole). Une autre raison est que les êtres humains ne font pas l'étiquetage vraiment bien pour les unités plus petites que les phones, les gens ont du mal à trouver les limites des sous-phones. Les RAP ne sont pas mieux que les humains à trouver les limites, mais leurs erreurs au moins sont fixes entre les ensembles d'apprentissage et de test.

Pour cette raison, les systèmes de reconnaissance de parole entrainent chaque phone HMM intégré dans une phrase complète, et la segmentation et l'alignement de phone sont faits automatiquement en tant que procédure d'apprentissage. Ce processus d'apprentissage de modèle acoustique entier est donc appelé apprentissage intégré. Par contre, la segmentation de phone manuelle joue un certain rôle, pour l'amorçage de système initial pour les estimateurs de ressemblance discriminatifs (SVM ; non-Gaussien), ou pour les tâches de reconnaissance de phone par exemple.

En résumé, la procédure d'apprentissage intégré basique est tel que suit :

1- Construire un HMM de « phrase entière » pour chaque phrase, comme montré dans la figure (III.10).

2- Initialiser les probabilités A à 0.5 (pour les boucles-arrières ou pour le sous-phone suivant correct) ou à zéro (pour toutes les autres transitions).

3- Initialiser les probabilités B en plaçant la moyenne et la variance pour chaque Gaussienne à la moyenne globale et variance pour l'ensemble d'apprentissage entier.

4- Exécuter multiples itérations de l'algorithme Baum-welch.

III.6. Conclusion

Ce présente des solutions aux problématiques posées dans le chapitre I. la première étant les tri-phones : phonèmes dépendant du contexte, qui a résolu le problème de coarticulation. La seconde solution est une approche idéale pour la reconnaissance de parole arabe chantée, celle qui relie les modèles tri-phones par une gaussienne uni-varié dans un premier temps afin d'attacher les sons ou phonèmes identiques, puis les élargir aux modèles de mélanges gaussiens afin de reconnaître ces sons même s'ils continuent et se maintiennent dans le temps.

Dans le chapitre suivant, nous verrons l'algorithme de tri-phones élargis aux GMMs en détail, en plus les résultats de nos trois expériences afin de les analyser puis discuter leurs avantages et inconvénients.

Chapitre IV

Expérimentations et analyse des Résultats

IV.1. Introduction

Ce chapitre confirme notre théorie du chapitre précédent en évaluant selon des critères que nous verrons et en validant par des chiffres en faisant une étude comparatives avec deux travaux antérieurs, sans oublier de confirmer cette hypothèse avec des statistiques ainsi que des graphiques d'évaluation.

Pour terminer, nous verrons les limitations de notre méthode pour en proposer des perspectives qui aideraient bien de futures recherches. Nous introduisons aussi l'outil HTK qui nous sert à construire et manipuler les modèles HMM. HTK est un ensemble de bibliothèques et d'outils valable en source C.

IV.2. Base de données arabe

- **Fichier audio**

L'arabe standard a essentiellement 34 phonèmes, dont 6 sont des voyelles, et 28 sont des consonnes. La correspondance entre écriture et prononciation en arabe standard moderne (MSA). Comme l'espagnol, a une cartographie presque un à un entre les lettres et les sons. Des langues comme l'anglais et le français, présentent une cartographie du son plus complexe. MSA est largement enseigné dans les écoles, et utilisé sur les lieux de travail, le gouvernement et les médias.

Nous avons utilisé deux ensembles de données dans notre travail: le premier est de chiffres arabes que nous avons reliés en dix phrases avec "HSGen" commande de Hmm Toolkit (HTK). Nous avons téléchargé une base de données intitulée "Spoken Arabic Digits" (SAD), qui contient 6600 mots pour la formation et 2200 mots pour les tests [http://archive.ics.uci.edulml/datasets/Spoken+Arabic+Digit].

Nous choisissons le mot comme unité acoustique parce qu'il n'y en a que dix, alors chaque mot est représenté par un HMM de 12 à 15 états émettant, comme le phonème de chaque mot représenté par 3 états.

Dans la deuxième base, nous avons traité l'arabe classique: Coran parce que c'est la parole verbale de Dieu et le testament final. Il est considéré comme la plus belle pièce de littérature en arabe (Yekache, Mekelleche, & Kouninef, 2012).

Le Coran se compose de 114 chapitres de différentes longueurs, chacun connu sous le nom de Surat. Nous avons choisi de traiter avec les 30 derniers chapitres du Saint Coran. Nous avons téléchargé le fichier audio "surat_number.mp3", puis nous l'avons converti en fichier ".wav" en utilisant un taux d'échantillonnage de 16 KHz et 16 bits par échantillon, ce taux a été choisi car il fournit des informations plus précises à haute fréquence (Yekache et al., 2012).

Nous avons divisé le fichier audio en phrases séparées par un silence manuellement avec le logiciel Cool Edit Pro 2.0. Ce fichier audio est récité par trois enceintes

différentes. Le texte est divisé de 220 à 311 phrases coraniques en fonction du tajweed du narrateur.

- **Fichier de Transcription**

Le processus de transcription est fait manuellement; Qui est fait en écoutant l'enregistrement alors nous correspondons exactement ce que nous entendons dans le texte, même le silence (sil), start-sentence (<s>) et fin-phrase (</s>) ont été représentés dans la transcription.

- **Dictionnaire phonétique**

Nous avons utilisé 53 phonèmes dans le texte, y compris les voyelles courtes et longues, la double consonne (shedda) de certaines consonnes, la distinction entre sons pharyngés et emphatiques, plus les phonèmes de silence (sil) et de pause (sp); Chacun représenté par un HMM dans le dictionnaire. Nous avions 944 mots dans le dictionnaire phonétique. Nous avions mis toutes les phrases dans un système dédié à la reconnaissance indépendante du locuteur afin de tester le plus près.

Nous avons choisi une structure de dictionnaire phonétique telle qu'elle se compose de deux colonnes: une pour les mots et l'autre pour la transcription phonétique. Exemple: le mot «mashhood» est transcrit comme «m a sh &oo d sp». Tous les mots sont terminés par une courte pause, une phrase de début et une phrase de fin sont transcrits par un silence.

Voici les monophones utilisés dans le dictionnaire avec leur transcription dans la table :

Table IV.1 : table des phonèmes utilisés dans le texte arabe avec leur transcription phonétique

Monophone	**Transcription**	**Monophone**	**Transcription**
ع	AA	دّ	dd
◌َ	a	ه	&
ب	b	ي	y
د	d	ل	l
ت	t	ق	q
◌ُ	u	لّ	ll
م	m	ك	k
ن	n	ر	r

ِ	i	ش	sh
وُ	oo	ص	S
ث	th	صّ	SS
ز	z	سّ	ss
يِ	ee	كّ	kk
س	s	ثّ	thth
ط	T	ظ	TH
ف	f	هّ	&&
ح	h	بّ	bb
و	w	وّ	ww
ج	j	قٌ	qq
غ	gh	تّ	tt
آ	aa	فّ	ff
خ	kh	رّ	rr
نّ	nn	شّ	shsh
مّ	mm	زّ	zz
ض	D	طّ	TT
ش	$	ضّ	DD
يّ	yy	أ	@

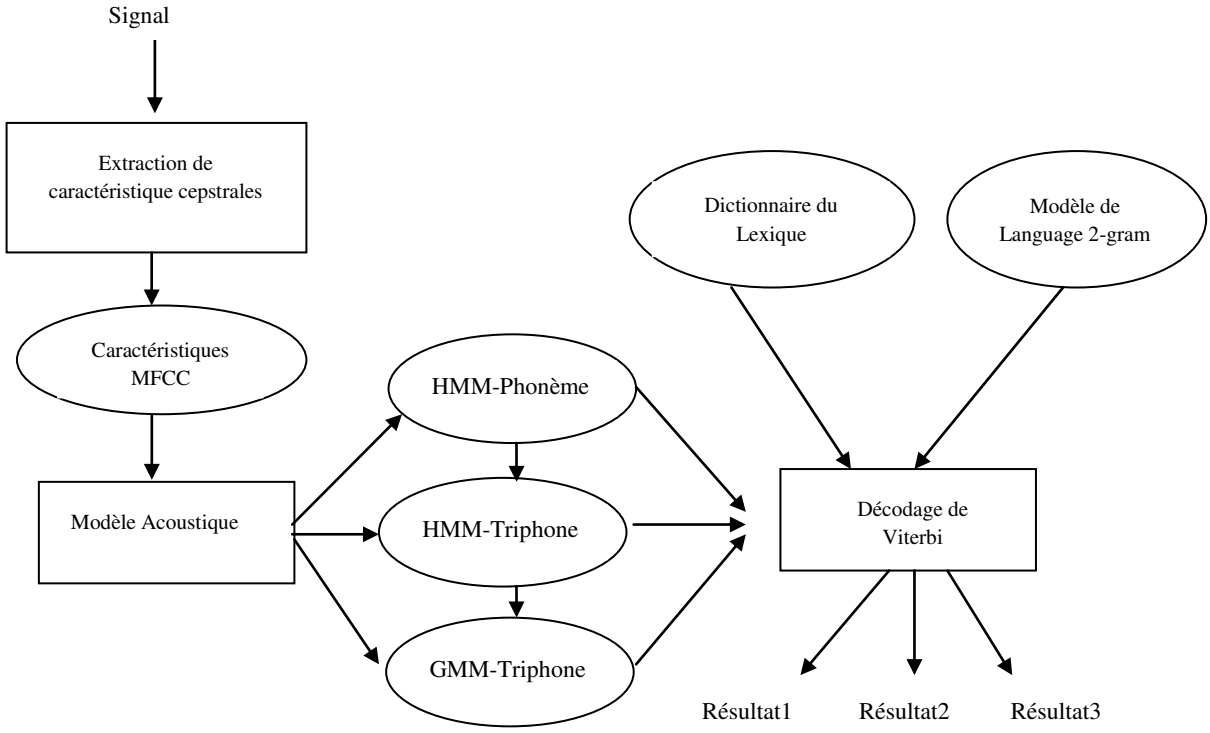
IV.3. Architecture du modèle proposé

Nous introduisons l'algorithme Baum-Welsh ou Forward-Backward pour les HMM de formation qui utilisent l'algorithme de Maximisation de l'Expectation (EM), et l'algorithme de Viterbi pour le décodage des HMM (Martin and Jurafsky 2000).

Pour la reconnaissance de chiffres (la reconnaissance de dix mots arabes de mots de zéro à neuf), nous avons construit un HMM dont les états correspondent à des mots entiers

avec 3 états émetteurs pour chaque phonème. Pour les phrases du Coran, les états cachés de l'HMM correspond à des unités semblables à un phonème, et les mots sont des conséquences de ces unités de type phonème.

Puisque la base des chiffres arabe a donné un bon résultat par rapport à la base de données coranique, nous avons résolu la confusion entre les phonèmes simplement en appliquant en premier lieu les HMMs sur les mono-phones dans le premier ensemble de données. Nous avons modifié l'algorithme d'apprentissage pour gérer les tri-phones en deuxième lieu qui surmontent le défaut de coarticulation. En troisième lieu, il faut utiliser les mélanges Gaussiens (GMM) pour remédier à la déficience de la durée sonore dans le deuxième ensemble de données. L'algorithme de Baum-welsh est utilisé à plusieurs reprises comme une composante du processus d'apprentissage intégré. Nous avons choisi d'exécuter à plusieurs reprises le chemin Viterbi (le plus probable) au lieu d'exécuter EM à chaque étape de l'apprentissage intégré. La figure (IV.1) présente le modèle d'architecture proposé.



Figue IV.1. Schéma général du système de reconnaissance de parole proposé.

IV.4. Organisation de l'espace de travail

Après être enregistré sur le site de (Young, Kershaw et al. 2000), téléchargé la version 3.4 de HTK sous Windows, ainsi que le livre HTK qui est un guide pour nous. Nous avons compilé le code source de HTK sous MS-DOS (MicroSoft Disk Operating System) pour pouvoir les utiliser en tant que commandes.

IV.4.1. Etapes de traitement des chiffres arabes parlés

Nous avons commencé à apprendre à utiliser HTK en le testant avec une première base, celle de chiffres arabes parlés enregistrés par 66 locuteurs (33 mâles et 33 femelles), dans lesquels chaque locuteur prononce chaque chiffre de zéro à neuf dix fois. Ce qui fait que nous avons 66000 mots prononcés en tout.

Nous avons organisé notre travail ainsi :

- Nous avons créé un dossier « Data », dedans il y a deux autre dossiers, un pour l'apprentissage « train » et l'autre pour le teste « test ». Dans chacun des deux dossiers, il y en a deux autres : « mfcc » converti du binaire en htk. Le dossier « lab » pour étiqueter chaque mot. Nous avons préparé la liste des mfcc.
- Nous avons créé un dossier « model », dans le quel il y a un dossier « proto » pour chaque modèle prototype HMM de chaque mot (de zéro à neuf). Un autre dossier « learning1 » pour l'initialisation des modèles HMM pour chaque mot. Puis « learningf » pour les modèles de HMM appris. Notons que le modèle prototype de chaque HMM représentant le mot est constitué de 3 états par phonèmes.
- Préparer le réseau de mot, la grammaire, le dictionnaire, la liste des hmm pour faire sortir les mots reconnus, puis les analyser pour voir le résultat. Faire de même pour l'ensemble de données de l'apprentissage ainsi que celui du teste. Nous verrons les résultats dans la section (IV.5).

Nous avons utilisé l'ensemble de données de chiffres arabes que nous avons reliés en dix phrases avec "HSGen" commande de Hmm Toolkit (HTK). Nous avons téléchargé une base de données intitulée "Spoken Arabic Digits" (SAD), qui contient 6600 mots pour la formation et 2200 mots pour les testes [http://archive.ics.uci.edulml/datasets/Spoken+Arabic+Digit].

Pour capturer ce fait de nature non-homogénéité des phones à travers le temps, dans la RPCGV généralement, nous modélisions un phone avec plus d'un état HMM. La configuration la plus commune est d'utiliser 3 états HMM : un état début (d), milieu (m) et fin (f). Chaque phone alors est formé de 3 états HMM émettant au lieu d'un seul (plus deux états non émettant : celui du début et de la fin).

Il est ordinaire de réserver le modèle de mot ou le modèle de phone pour soumettre au HMM un phone entier de 5-état et utiliser le mot état du HMM (ou juste état pour faire court) pour référencier chacun des 3 états des sous-phones individuels.

Pour construire un HMM pour un mot entier utilisant ces modèles de phones plus complexe, nous pouvons remplacer chaque phone du modèle de mot de la figure (IV.1) avec un HMM de phone 3-état.

Nous choisissons le mot comme unité acoustique parce qu'il n'y en a que dix, alors chaque mot est représenté par un HMM de 12 à 15 états émettant, car chaque phonème du mot est représenté par 3 états. La figure (IV.2) montre le mot « sitta » entendu.

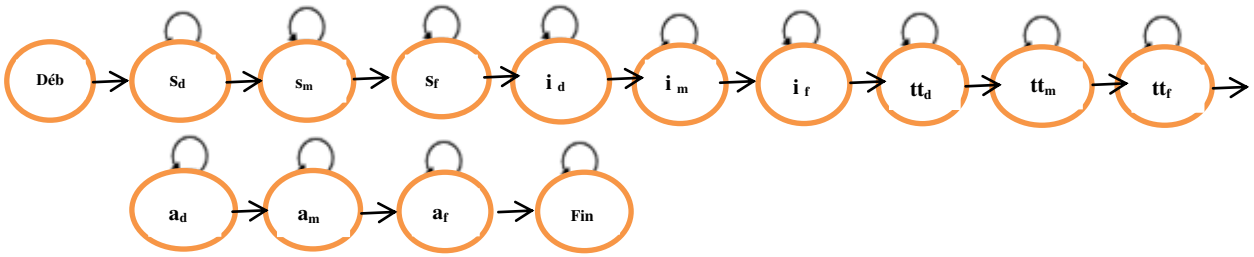


Figure IV.2 : une composition du modèle du mot « sitta », formée pour la concaténation de quatre modèles phone, chacun avec trois états émettant.

IV.4.2. Etapes de traitement de versets coranique

Dans un premier temps, nous avons exécuté les étapes basiques du traitement de reconnaissance de la parole en représentant chaque phonème par un modèle de HMM.

1- Nous avons préparé les données c'est-à-dire : diviser le texte transcription phonétique, puis les versets prononcés par chaque lecteur célèbre (nous avons testé trois) suivant le texte lu qui diffère d'un lecteur à un autre. Ce qui implique une génération d'une carte de mot vide.
2- Nous avons choisi un modèle de langage (LM) avec HTK qui est le bi-gram qui a donné le meilleur résultat, car nous avons testé les tri-grams et ça n'a pas fonctionné. Nous n'avions pas assez de données
3- Nous avons mis en place un système de reconnaissance vocale simple. Nous avons créé le réseau de mot avec la commande « HBuild », en lui donnant la liste de mots.
4- Nous avons créé le dictionnaire qui gère HTK avec la commande « HDMan ».
5- Nous avons préparé le fichier de transcription pour étiqueter le mot au niveau du phonème avec la commande « HLed ».
6- Nous avons calculé les vecteurs MFCC avec « HCopy ».
7- Puis, nous avons initialisé HMMs prototype avec « HCompV » que nous avons actualisé avec « HERest » deux fois.
8- La prochaine étape, nous avons inclus le modèle "sp" et modifié le "sil" pour estimer à nouveau les modèles deux fois avec « HHED » et « HERest ».
9- La dernière étape est de tester les phrases saisies avec la commande qui convient à l'algorithme de Viterbi « HVite », puis d'évaluer les résultats avec «

HResult ». Figue(IV.3) montre son architecture générale (Young, Evermann et al. 2006).

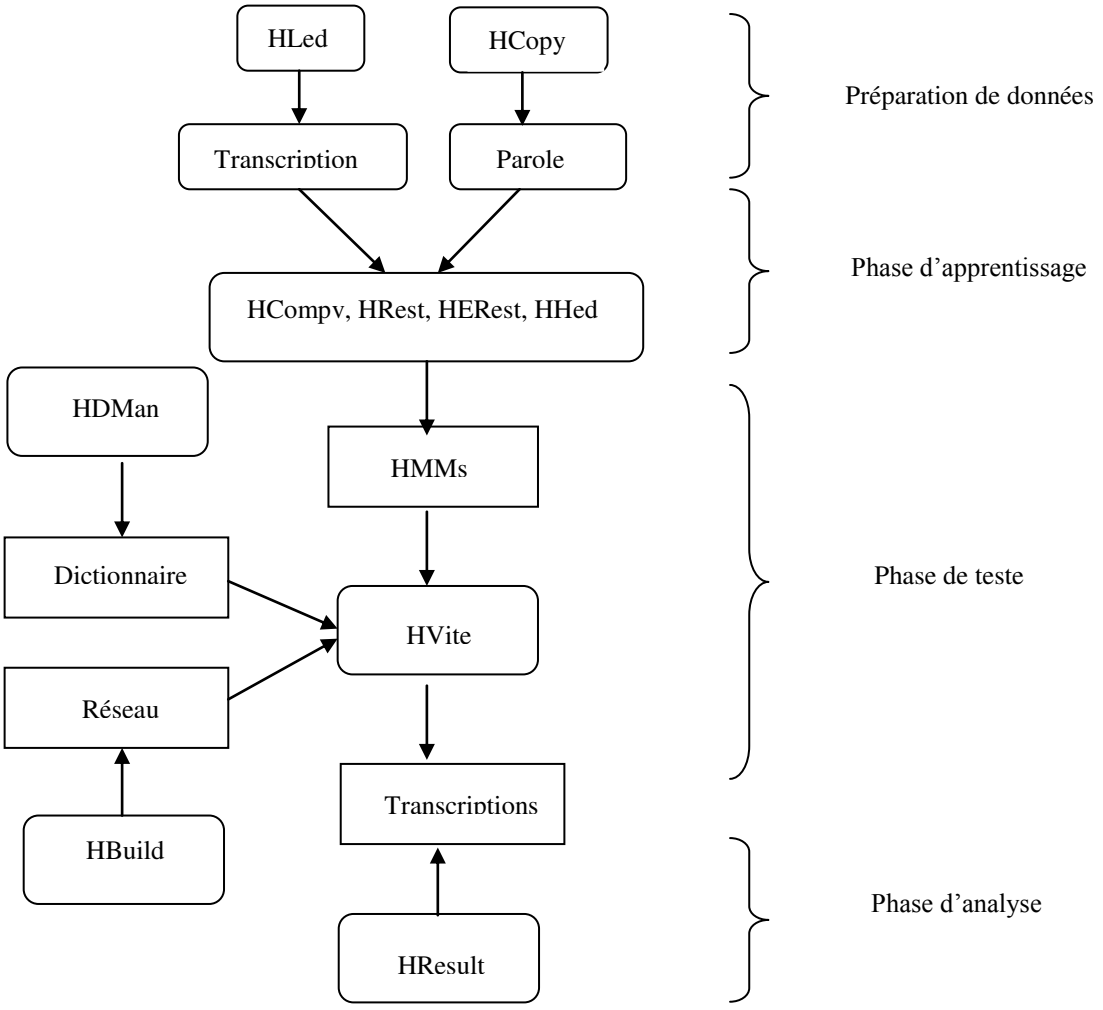


Figure IV.3 : Etapes basiques de traitement avec l'outil HTK

Dans un deuxième temps, nous avons représenté chaque phonème par trois modèles de HMM selon son contexte. Soit que le phonème est au début, au milieu ou à la fin du mot.

10- Nous avons réaligné les données d'apprentissage toujours avec « HVite ».
11- Nous avons construit des tri-phones à partir des mono-phones avec la commande « CL ». Pour cela, nous avons converti les mono-phones en un ensemble de tri-phone équivalents avec la commande « mktri.led ». Nous avons crée la commande mktri.hed avec le script écrit en perl téléchargé pour éliminer les tri-phones inutiles. Nous créons trois nouveaux fichier dans chaque étape pour ré-estimer les HMMs.
12- Nous avons construit des tri-phones à état liés avec la commande « TI ».

Dans un troisième temps, nous avons élargi ces tri-phones liés selon la ressemblance des phonèmes selon leur contexte en un mélange de gaussiennes multi-variées. Nous verrons dans la section 3 de ce chapitre, les étapes détaillées de l'algorithme associé qui est la contribution de cette thèse.

Maintenant, regardons les caractéristiques spectrales de la phrase en phonétique « wassama i thati l burooji walyawmi l mawAAoodi» prononcé par le célèbre lecteur « Alsudaissi » selon les régles de Tajweed montrée dans la figure (IV.4), ensuite prononcé par moi-même, d'une lecture normale montrée dans la figure (IV.5).

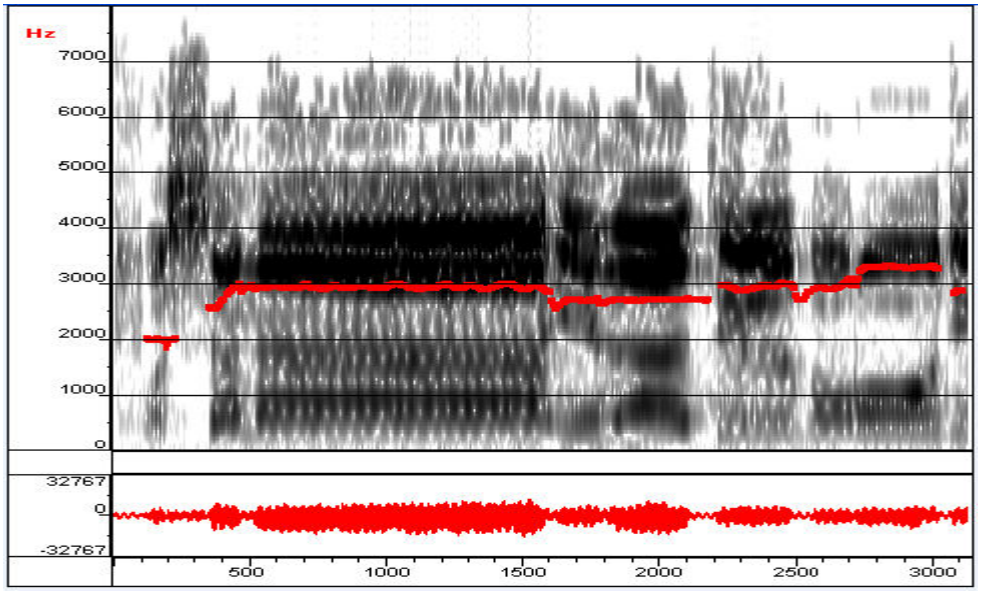


Figure IV.4 : représentation temporelle (en bas oscilloscope) et fréquentielle (en haut spectrogramme), et la fréquence fondamentale marquée en rouge d'une phrase prononcée selon les règles de Tajweed par le lecteur « Alsudaissi »

Nous voyons que l'énergie s'étend de 0 à 7000 Hertz (Hz), cela veut dire que les harmoniques sont renforcées dans cet intervalle de fréquences. Le temps est de 3000 milliseconde (ms).

Nous remarquons que la fréquence fondamentale F0 qui véhicule une grande partie de l'information prosodique, s'étend de 2000 à 3300 Hz dans le spectrogramme chez « alsudaissi », cela est dû à la voix masculine.

Ce qu'il nous faut avec le spectrogramme de notre voix normale, c'est une bande large car un spectrogramme à bande large contrairement à la précédente offre une bonne résolution au niveau fréquentiel de la structure harmonique du signal. L'analyse temporelle est moins fine ici, ce qui nous importe peu.

Ce qui nous importe chez « alsudaissi » c'est les voyelles. Regardons dans les deux spectrogrammes de la figure (4) et (5) : les voyelles *a* (al-fatha) noté en phonétique par [oe] ; *i* (al-kasra) noté [i] et la voyelle *ou* (al-ddamma) noté [u].

La structure harmonique est claire pour ces voyelles (voir figure IV.4), pour [u] c'est surtout les harmoniques en basse fréquences qui sont renforcées (0-1000) Hz, pour [i] les harmoniques entre (1000-3000) Hz sont affaiblies, pour [oe] les harmoniques sont maintenues à peu près partout entre (0-4000) Hz, ce qui laisse penser que les résonateurs du conduit vocal agissent comme un filtre sur une source pour la production des voyelles.

Certes, mais les voyelles ne sont pas les seules à présenter un spectre riche en harmonique, les consonnes sonores tel que [j], [n], [r] selon l'alphabet phonétique international (API) qui montrent que leur production est accompagnés d'une vibration des cordes vocales, source de l'onde périodique de période f0 et riche en harmonique, mais leur durée en général est moins que importante. Ce n'est certainement pas le cas pour le [s] qui ne présente pas de structure harmonique et une dispersion d'énergie instable située dans les hautes fréquences comme du bruit. Enfin pour [t] ainsi que pour la plupart des occlusives sourdes, une barre d'explosion de bruit dont la durée dans le temps est vraiment très petite.

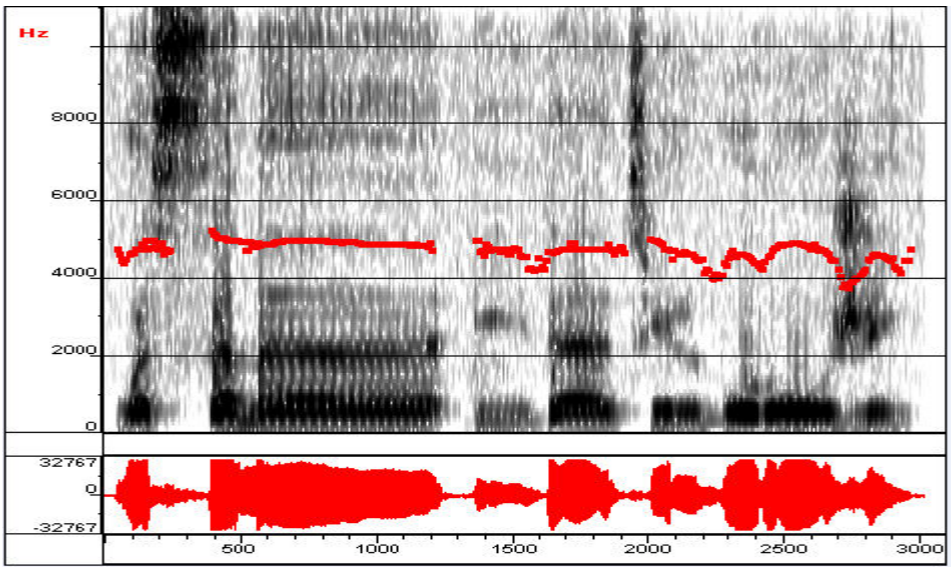


Figure IV.5: représentation temporelle (en bas oscilloscope) et fréquentielle (en haut spectrogramme), et la fréquence fondamentale marquée en rouge de la même phrase d'une lecture normale prononcée par moi même

Nous voyons que l'énergie s'étend de 0 à 10000 Hertz (Hz). Le temps est de 3000 milliseconde (ms). Nous pouvons voir les formants de chaque phonème : chez les semi voyelles /wa/, /ya/.

La différence entre la phrase prononcé selon les règles de Tajweed et la lecture normale se voit seulement lorsque nous dessinons la fréquence zéro (F0 ou appelé le pitch, nous donne l'information sur la prosodie de la phrase prononcée).

Nous remarquons que la fréquence fondamentale F0 qui véhicule une grande partie de l'information prosodique, s'étend de 3800 à 5500 Hz dans le spectrogramme chez « moi-même », cela est dû à la voix féminine.

Ce qu'il nous faut maintenant avec le spectrogramme de « alsudaissi », c'est une bande étroite car un spectrogramme à bande étroite offre une meilleure résolution temporelle et donc nous permettra de dégager les formants vocaliques que nous pourrons voir comme étudier le phénomène de courte durée tel que le phonème *t* dans notre exemple de phrase 'une petite barre d'explosion). L'analyse fréquentielle est moins fine avec la disparition harmonique du signal dans la bande étroite, ce qui nous importe peu justement chez « alsudassi ».

Regardons le cas du phonème *t :* consonne occlusive sourde dentale non nasale :

Consonne : un son qui est produit grâce à un obstacle.

Occlusive : l'obstacle doit être total.

Sourde : et ne pas générer de vibrations.

Dentale : on l'obtient en mettant la langue sur les dents.

Non nasale : sans qu'aucun air ne passe dans le nez.

IV.5. Expérimentations et résultats sur les HMM

- **Application des HMMs sur les mono-phones**

Dans cette expérience, nous avons appliqué HMM-phonème qui convient à notre corpus (les deux bases de données), cela signifie pour les mots liés. Les résultats sur l'ensemble de données de test des chiffres connectés sans courte pause sont présentés ci-dessous dans la Table IV.2:

Table IV.2 : Matrice de confusion des chiffres connectes de l'ensemble de donnée teste.

Matrice de confusion	0	1	2	3	4	5	6	7	8	9	Word recognition %
0	215	0	0	0	0	0	1	0	0	4	98
1	0	220	0	0	0	0	0	0	0	0	100
2	0	0	220	0	0	0	0	0	0	0	100
3	2	1	0	214	0	0	0	3	0	0	97
4	0	0	0	0	212	1	0	7	0	0	96
5	0	0	0	0	0	220	0	0	0	0	100
6	1	0	0	1	0	0	216	0	0	2	98
7	2	0	0	0	7	0	0	207	0	4	94
8	3	0	0	2	0	1	1	1	212	0	96
9	0	2	1	0	0	0	0	0	0	219	99

On note que les mots étiquetés correspondant à "1, 2, 5" ont donné le meilleur résultat et sont reconnus à 100%, il a confondu beaucoup entre "0, 6" chiffres ("sifr, sitta" prononcé en arabe).

Comme nous l'avons montré, les mots corrects testés sont excellents avec 97,95% de reconnaissance de mot correcte (WCR) qui influe sur le taux de phrases correctes 93,14% de SCR, ce qui montre que HMM peut représenter le phonème comme unité acoustique donnant assez bon résultat. (WR) et (SR) sont définis pour le taux de reconnaissance de mot et taux de reconnaissance de phrases respectivement

Ci-dessous est la Table IV.3 montrant la reconnaissance de taux de données de chiffres, plus les trois haut-parleurs chacun seul dans un système, en plus du système indépendant de l'orateur.

Table IV.3: Résultats d'expérience HMM-phoneme sur les deux ensembles de données.

Base de données	Chiffres	Alghamidi	Alsudaissi	Elmirigli	Indépendant du locuteur
WR %	97.95	70.05	64.11	71.24	53.70
SR %	93.14	12.79	9.73	15.16	8.97

Les résultats sur le corpus du Coran sont médiocres par rapport au corpus de chiffres, car nous avons d'abord reconnu la conférence tajweed du Coran par trois célèbres narrateurs (chacun avec son rythme et non une lecture normale du texte) qui justifient les résultats. Nous notons une lacune dans le narrateur "Alsudaissi" par rapport aux autres en raison de "mudud", ce qui signifie la longue durée temporelle de la prononciation des voyelles.

Par contre le narrateur "Elmirigli" qui est le plus rapide et il répète beaucoup ses phrases: nous avons obtenu 311 phrases courtes dans ce système.

En conséquence de la lecture de tajweed, il ya une confusion entre les phonèmes dans leur prononciation, en plus de la durée sonore qui crée un WER élevé, dans ce cas le taux de phrases correctes diminuer puisque si il ya un mauvais mot reconnu, les phrases contenant ce mot est fausse aussi. Si les phrases sont courtes et énorme, ce qui est le cas de "Elmirigli", alors le taux de phrase correcte dans ce cas est plus élevé que les autres.

- **Application des HMMs sur les tri-phones**

Nous améliorerions ces résultats en utilisant des triphones afin d'apprendre au système les nuances entre les phonèmes dont leurs caractéristiques se ressemblent. Comme /ح/ et /ه /transcrit respectivement comme / h / et /& /, en plus des sons pharyngés et emphatiques / س / et / ص / transcrits respectivement comme / s / et / S /, / د / et / ض / transcrits respectivement comme / d / et / D /, qui sont présentés en arabe. Lorsque nous appliquons l'unité de triphones comme modèle acoustique (AM) considéré comme dépendante du contexte (CD) au lieu d'être indépendant du contexte (CI) pour la reconnaissance, le système englobera la variation allophonique des phonèmes: / $a_{début}$ /, / a_{milieu} /, / a_{fin} / Diffèrent selon le contexte.

Ci-dessous la table IV.4, montrant les résultats de la deuxième expérience sur les trois locuteurs en plus du système indépendant du locuteur.

Table IV.4 : Résultats de l'expérience HMM-Tri-phone sur la base coranique.

Base de données	**Alghamidi**	**Alsudaissi**	**Elmirigli**	**Indépendant du locuteur**
WR %	59.59	55.48	58.65	36.22
SR %	5.53	2.65	6.15	1.54

C'est notre deuxième expérience qui a fait diminuer les résultats, parce qu'il ya peu de phrases en entrées. C'est pour cela, les tri-phones ne fonctionnent pas aussi bien que comme dans (Hyassat & Zitar, 2006), ils ont utilisé tout le texte du Coran.

IV.6. Expérimentations et résultats sur les GMMs

L'utilisation de mélanges gaussiens aide à séparer et à neutraliser l'effet des sons liés et le problème de co-articulation en raison de l'anticipation de la prononciation du phonème suivant, le GMM facilite également l'ajustement de la durée du son de parole qui diffère d'un narrateur à un autre.

Ainsi, le système reconnaîtra l'élocution indépendamment du locuteur, comme dans (Elhadj, Alghamdi, & Alkanhal, 2014). Nous relançons une troisième expérience en introduisant les GMM dans les résultats améliorés de la formation monophonique comme valeurs initiales lors de la formation des tri-phones. Nous avons vu en détail l'algorithme proposé dans le chapitre III.

Les résultats de cette expérience sont exposés dans la table IV.5.

Table IV.5 : Résultats sur l'expérience des tri-phones à états liés étendus aux GMMs.

Base de données	**Alghamidi**		**Alsudaissi**		**Elmirigli**		**Indépendant du locuteur**	
Nombre GMM	WR	SR	WR	SR	WR	SR	WR	SR
16 GMM(%)	63.69	6.42	60.72	2.65	62.5	6.47	38.11	2.29
32 GMM(%)	64.59	4.61	60.50	3.10	63.6	7.12	41.04	2.76
64 GMM(%)	67.24	5.96	64.89	3.14	64.9	7.59	42.99	3.13
128 GMM(%)	69.35	7.18	70.38	4.09	68.1	9.12	45.39	3.53
256 GMM(%)	71.39	9.48	71.88	6.22	73.10	13.36	48.09	4.34
512 GMM(%)	73.17	11.11	72.41	6.67	73.44	14.38	49.90	4.37

Pour surmonter les défauts de la langue arabe en particulier sur le corpus coranique qui est une tâche un peu délicate, par rapport à la prononciation des chiffres, nous suggérons d'appliquer un modèle de tri-phone à états liés élargi aux GMM en ajoutant une quatrième étape à l'algorithme.

En conséquence, nous avons 512 GMM qui convient aux trois systèmes (chaque locuteur dans un système), et même pour le système indépendant du locuteur. Nous validons les résultats en dépassant nos propres résultats de la première expérience par la troisième. Surtout en comparant le résultat de nos narrateurs :

Alsudaissi par notre système proposé avec le système de régression linéaire à maximum de vraisemblance (MLLR) (Mourtaga, Sharieh, & Abdallah, 2007).

Notre taux de mots correctes est de 72,41, tandis que celle de la méthode MLLR est de 68%, faisant une différence de près de 5% que nous verrons dans la section suivante.

- **Discussion**

D'après ces résultats, il y a encore du travail d'où l'intérêt d'annoncer clairement une contribution originale à ces connaissances dans cette thèse est.

Dans cette thèse nous avons utilisé l'outil de développement HTK qui comprend les HMM comme les GMM aussi pour pouvoir démontrer qu'il apporte une contribution substantielle et innovatrice à la reconnaissance de parole continue pour un système indépendant du locuteur dans la langue arabe.

Nous avons commencé à comparer nos résultats par nos propres expériences : résultats des HMM avec ceux des GMM. Ensuite, nous comparerons dans la section suivante nos résultats avec d'autres travaux.

Table IV.6 : récapitulation des chiffres des deux expérimentations

Base de données	**Alghamidi**	**Alsudaissi**	**Elmirigli**	**Indépendant du locuteur**
WR(%) HMM	70.05	64.11	71.24	53.70
WR(%) GMM	73.17	72.41	73.44	49.90

D'après la table IV.6 récapitulative des résultats des HMM et ceux de GMM, nous voyons la différence entre ces deux approches. Les chiffres des résultats de GMM ont dépassé ceux de HMM, ceci est dû au manque des HMM qui ne peuvent modéliser la variation temporelle du son qui est présente dans les règles de Tajweed, c'est pour cela que nous avons utilisé les modèles de mélange Gaussien qui sont parvenus à remédier à cette lacune. Nous remarquons que la différence entre ces deux approches se remarque surtout chez le lecteur « Alsudaissi » car c'est celui qui varie beaucoup ses voyelles (mudud), et elle est notée par un grand pourcentage 8.3 %. La différence la moins remarquable est chez le lecteur « Elmirigli » car c'est celui qui utilisent des petite phrases et qui les répètent en plus, il était le meilleur par rapport aux autres lecteurs au départ lors de l'application des HMMs. Cette petite différence est seulement notée par un pourcentage de 2.2 %.

Effectivement, grâce aux trois procédures mises en œuvre combinées: le clonage de triphones, les valeurs perturbatrices puis en les ré-estimant, les résultats précédents sont améliorés, en particulier sur le lecteurAlsudaissi qui prend beaucoup

de temps pour prononcer des voyelles (mudud) et varie dans le temps, ce qui est résolu suite du papier (Mourtaga et al., 2007). Ils ont travaillé avec MLLR et ont été très bien pour tous les narrateurs, qui dépassent les résultats de nos lecteurs à l'exception du locuteur « Alsudaissi ».

C'est pourquoi nous l'avons amélioré avec des tri-phones élargis au processus de GMM et par conséquent notre méthode a surpassé leur technique de la régression linéaire de maximum de vraisemblance de 5% chez le narrateur « Alsudaissi ».

La table IV.7 est faite pour comparer entre notre travail et celui du (Mourtaga et al., 2007).

IV.7. Etude comparative

Après avoir amélioré nos résultats avec l'algorithme proposé et présenté, nous avons établie une étude comparative avec un autre qui s'en rapproche, et donc la table (IV.6) dévoile les différences et similitudes de chacun d'eux selon certain critères d'évaluation.

Table IV.7 : Différences entre notre travail et celui de « Mourtaga » selon des critères d'évaluation suivant :

Processus du système	**Notre travail**	**Travail de (Mourtaga et al., 2007)**
Nombre d'états du modèle HMM	Trois par phonème	Treize par mot
Entre les mots	Existence de courte pause (sp)	Très courte pause ou pas existante
Nombre de mono-phones	53 + sp+ sil	33 + sp + sil
Le bruit	Pas modélisé	Géré par quelques modèles
Transition entre état début et état fin	N'existe pas	Existe
Nombre de mots	944	2000
Nombre de phrases	757	2431
Processus du système d'apprentissage	Apprentissage de mono-phone, la fixation du modèle de silence, la fabrication de modèles tri-phone à partir de mono-phone, groupement des Tri-phones.	
Processus ajouté pour l'adaptation	Grouper les tri-phones avec gaussienne un-varié, puis les étendre aux GMMs	Arbre de la classe de régression linéairepuis adaptation à maximum de vraisemblance.
Taux de mot reconnus après amélioration	Est de presque 73 % chez le lecteur « Alsudaissi »	Est de 68 % chez le lecteur « Alsudaissi »

Pour confirmer cette différence, nous montrons un tableau récapitulatif de nos résultats de taux de reconnaissance de mot en (%) sur nos différents célèbres lecteurs.

Nous avons décidé d'élaborer cette comparaison avec un critère d'évaluation qui est le taux de mot reconnu après l'amélioration de l'algorithme proposé en tenant compte de quelques éléments de comparaison qui sont : le nombre d'état du modèle Hmm, le nombre de monophone, la transition entre l'état début et l'état fin, le processus du système d'apprentissage et le processus ajouté pour l'adaptation.

Table IV.8 : comparaison entre notre travail avec celui de « Mourtaga » suivant le critère taux de reconnaissance de mots.

Célèbre lecteurs	**Travail du groupe**	**Travail de Mourtaga**
Lecteur 1	73.17 %	85 %
Lecteur 2	73.44 %	78 %
Lecteur 3	-	80 %
Lecteur 4	-	82 %
Alsudaissi	72.41 %	68 %

Nous voyons bien que le travail de « Mourtaga » a de très bons résultats chez tous les lecteurs. Ces résultats dépassent largement nos deux lecteurs de coran, mis à part le lecteur « alsudaissi » où ils ont eu un résultat pas très fameux à cause de la longue durée des voyelles qu'utilise ce fameux lecteur come les auteurs l'ont précisé dans leur article.

Nous avons profité de cette insuffisance, d'où l'intérêt des modèles de GMMs qui modélisent bien les voyelles, et qui compense ou comble le déficit de ce manque.

La figure (IV.6) montre la comparaison des résultats du travail de groupe avec ceux de « Mourtaga » en graphe.

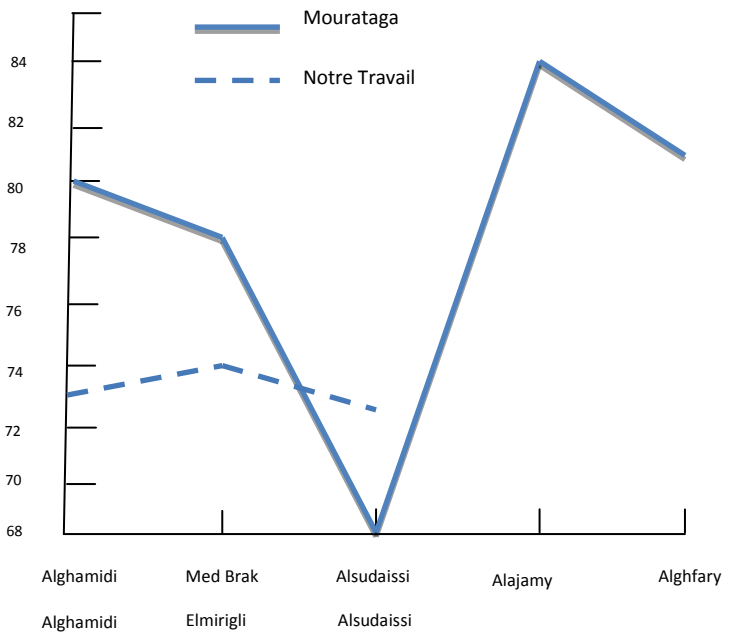


Figure IV.6 : comparaison des résultats du travail de groupe avec ceux de Mourtaga

Après avoir vu le critère d'évaluation selon les éléments de comparaison de notre travail avec celui de « Mourtaga », vu les statistiques, contemplons maintenant le graphique d'évaluation.

Bien que nous avons utilisé qu'un seul critère d'évaluation et avons utilisé que trois lecteurs par rapport à cinq lecteurs du travail de « Mourtaga », nous voyons clairement dans le graphe que tous les lecteurs de l'auteur « Mourtaga » dépasse le taux de mots reconnu chez les lecteurs du travail de groupe, mis à part chez notre lecteur « alsudaissi » qui dépasse de presque 5% celui de « Mourtaga ».

Nous présentons notre travail comme une évolution des travaux antérieurs (Hyassat & Zitar, 2006), (Mourtaga et al., 2007). Nous avons donc comparé notre système qui a utilisé HTK avec un autre qui a utilisé le moteur Sphynx (Hyassat & Zitar, 2006), et qui s'est arrêté à 256 Gmms.

Voici le deuxième travail avec qui nous avons comparé nos résultats :

Table IV.9 : comparaison entre notre travail avec celui de « Hyassat » suivant le critère taux de reconnaissance de mots, de phrase, le mode de lecture et de nombre de GMM.

-	**Travail de groupe**	**Travail de Hyassat**
Ensemble de base	Les 30 derniers chapitres	Tout le coran
Mode de lecture	Avec Tajweed	Normale
Nombre de GMM	256 GMM	256 GMM
WR	73.10 %	70.813 %
SR	13.36 %	20 %

Nous avons noté que notre résultat de 256 Gmms surpassé leur taux de 2,39% de WR vu qu'ils ont utilisé un seul narrateur, et cela avec 70,813% d'exactitude.

Le meilleur taux de reconnaissance de mots est chez « Elmirigli » avec un taux de : 73,10%. Leur taux de SR était de 20,879% qui nous ont surpassés parce que notre meilleur taux de reconnaissance des phrases est chez « Elmirigli », avec 14,38% lors des 512 Gmm, mais 13.36 % lors des 256 Gmm appliqués.

La figure (IV.7) montre les résultats en graphe.

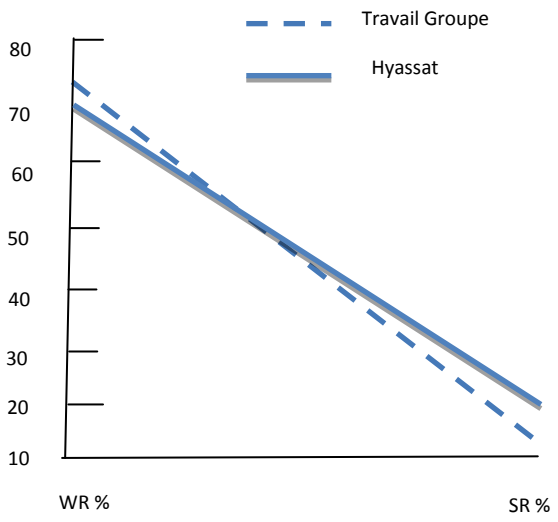


Figure IV.7 : comparaison des résultats du travail de groupe avec ceux de Hyassat

Cette fois-ci encore, nous avons choisi comme critère d'évaluation le taux de mot reconnu. Nous voyons clairement que notre travail du groupe dépasse celui de « Hyassat » en 256 GMMs, mais si nous voyons le critère taux de reconnaissance de phrases, c'est le leur qui nous devance.

Ceci veut dire que c'est l'une des limites de notre travail où nous verrons dans le chapitre de la conclusion et perspectives ce qui peut compenser ce déficit justement.

Ci-dessous, la Figure (IV.8) montre les taux de reconnaissance des mots des quatre systèmes sur les trois expériences :

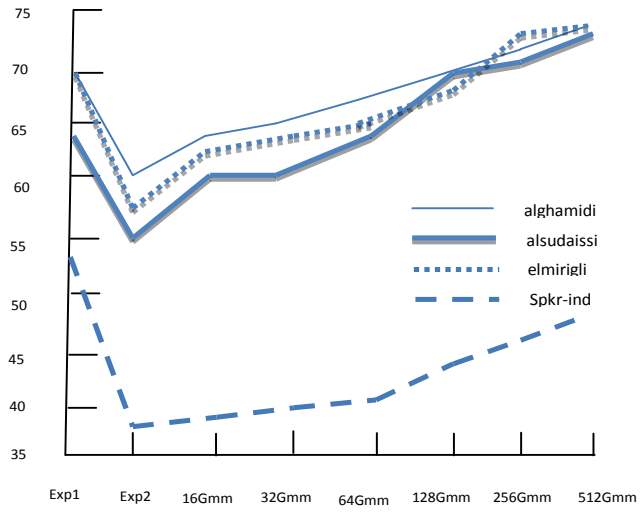


Figure IV.8: Courbes des quatre systèmes sur les trois expériences 1,2,3 en taux de mots reconnus (%).

Nous voyons bien que l'expérience de hmm-monophone a donné un bon résultat au début, puis chuté dans la deuxième expérience de hmm-triphone, puis remonté petit à petit qu'on ajoute le nombre de gaussiens dans l'expérience trois car plus nous ajoutons les coefficients, plus ça s'améliore : 512 GMM veut dire 512 coefficients.

Nous voyons que la courbe du système indépendant du locuteur est trop basse par rapport aux autres courbes. Ceci est dû au manque de lecteurs pour pouvoir faire l'apprentissage du système, ce qui coute énormément en temps et en espace, vu que trois lecteurs dans un seul système nous ont déjà couté en temps. Nous voyons que durant l'expérience GMM-Triphone, la courbe a bien augmenté, mais elle n'a pas dépassé l'expérience une comme c'est le cas chez les trois autres systèmes. Ce qui est notre première limitation, même si elle a atteint 49.90 % lors des 512 Gmms.

La figure (IV.9) nous montre les taux de reconnaissance de phrases des quatre systèmes en graphe.

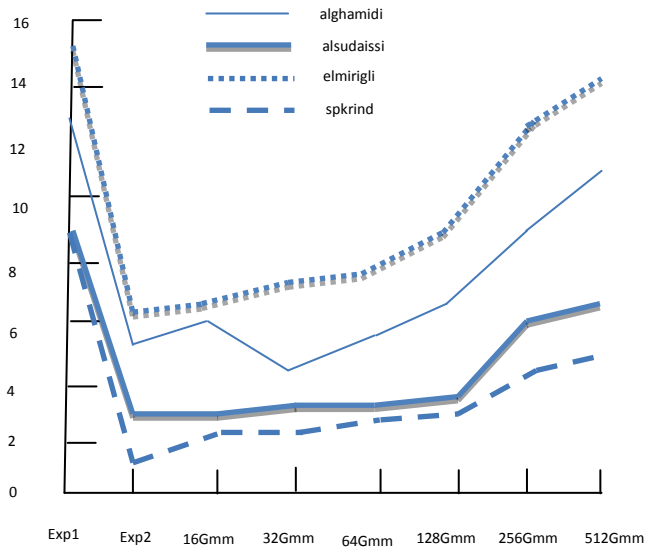


Figure IV.9 : Courbes des quatre systèmes sur les trois expériences 1,2,3 en taux de phrases reconnues (%).

Nous voyons bien que l'expérience de hmm-monophone a donné un bon résultat au début, puis chuté dans la deuxième expérience de hmm-triphone, puis remonté petit à petit qu'on ajoute le nombre de gaussiens dans l'expérience trois car plus nous ajoutons les coefficients, plus ça s'améliore : 512 GMM veut dire 512 coefficients.

Nous voyons que la courbe du système indépendant du locuteur est trop basse par rapport aux autres courbes en taux de reconnaissance de phrase tout comme en taux de reconnaissance de mots.

La différence entre le système indépendant du locuteur et les trois aux autres systèmes est moins flagrante en taux de reconnaissance de phrase qu'en taux de reconnaissance de mots vu que les quatre systèmes n'ont pas un bon taux de reconnaissance de phrase. Ce qui nous cause une deuxième limitation.

- **Critiques et limites**

Nous voyons clairement dans les deux courbes que ce soit en taux de reconnaissance de mots ou de phrases que les courbes de « Alghamidi », « Alsudassi » et « Elmirigli » ont augmenté et surpassé l'expérience hmm-phonème par l'algorithme du tri-phone étendu aux GMMs, même jusqu'à 16,93% chez le narrateur « Alsudaissi ». Tout cela montre que notre algorithme fonctionne aussi bien que nous l'avions prédit.

Le système indépendant du locuteur n'a pas battu l'expérience Hmm-monophone mêmesi il a été augmenté,il reste encore très bas par rapport aux autres systèmes. Ceci est dû à l'apprentissage qui coûte en temps et en espace si on ajoutait d'autres lecteurs

afin d'améliorer ce taux. Cela est notre première limitation à noter pour voir dans le chapitre qui suit comment améliorer ou compenser ce déficit.

La deuxième limitation est que nos quatre systèmes en courbe de taux de reconnaissance de phrases comme le montre si bien la figure 8, sont restés encore bas malgré l'ajout de nombre de GMM, cela veut dire qu'il faut bien proposer d'autres approches qui relèveront ces courbes jusqu'à atteindre les 20 % comme le travail de « Hyassat ».

La table (IV.10) montre les phrases en entrées avant la reconnaissance, puis après la reconnaissance en sortie.

Table IV.10 : différence entre les phrases avant et après leur reconnaissance.

Phrases en entrée	**Phrases en sortie**
<s> qulhuwa al llahuahadun al llahu ssamad </s>	al llahu ssamad
<s> lam yalidwa lam yooladuwa lam yakun lahu kufuwan ahad </s>	lam yalidwa lam yooladuwa lam yakun lahu kufuwan ahad
<s> qul aAAoothu bi rabbi al falaq </s> <s> min sharri ma khalaq </s>	qul aAAoothu bi rabbi al falaqi min sharri ma khalaq
<s> wa min sharri ghasiqin itha waqab </s>	wa min sharri ghasiqin itha waqab
<s> wa min sharri al nnaffathaati fee al AAuqad </s> <s> wa min sharri hasidin itha hasad </s>	wa min sharri al nnaffathaati fee al AAuqadi wa min sharri hasidin itha hasad
<s> qul aAAoothu bi rabbi nnasi maliki nnasi ilahi nnas</s>	qul aAAoothu bi rabbi nnasi maliki nnasi ilahi nnas
<s> min sharri al waswasi al khannas</s>	min sharri al waswasi al khannas
<s> al lathee yuwaswisu fee sudoori nnas</s> <s> mina al jinnati wa nnas</s>	al lathee yuwaswisu fee sudoori nnas mina al jinnatiwa nnas

Nous notons qu'il n y a pas une grande marge de différence entre les phrases entrée avant leur reconnaissance et après leur reconnaissance en sortie.

Nous notons juste une légère différence dans le 3ème résultat où nous avions deux phrases en entrées, et en sortie nous avons eu la prononciation des deux phrases en une seule liées par le « i » en rouge qui est une diacritization en arabe appelée « alkasra ».

Pour le 5ème résultat, pareil que le 3ème. Nous notons que les deux phrases en entrées sont liées en sortie par le mot « AAuqadi » au lieu de « AAuqad ».

Par contre dans le dernier résultat, malgré les deux phrases entrée, elles sont sorties telles quelles en une seule phrase.

Ci-dessous, quelques exemples pour illustrer la pertinence des expérimentations et des résultats de la figure (IV.10).

```
Aligning File: data/train/mfc/219.mfc
Created lattice with 3 nodes / 2 arcs from label file
al llahu ssamad  ==  [254 frames] -82.1606 [Ac=-20868.8 LM=0.0] (Act=8.4)
Aligning File: data/train/mfc/220.mfc
Created lattice with 11 nodes / 10 arcs from label file
lam yalid wa lam yooladu wa lam yakun lahu kufuwan ahad  ==  [744 frames] -79.0171 [Ac=-587
Aligning File: data/train/mfc/221.mfc
Created lattice with 10 nodes / 9 arcs from label file
qul aAAoothu bi rabbi al falaqi min sharri ma khalaq  ==  [590 frames] -80.0065 [Ac=-47203.
Aligning File: data/train/mfc/222.mfc
Created lattice with 6 nodes / 5 arcs from label file
wamin sharri ghasiqin itha wa qab  ==  [445 frames] -82.7460 [Ac=-36822.0 LM=0.0] (Act=9.6)
Aligning File: data/train/mfc/223.mfc
Created lattice with 12 nodes / 11 arcs from label file
wamin sharri al nnaffathaati fee al AAuqadi wamin sharri hasidin itha hasad  ==  [1100 fram
Aligning File: data/train/mfc/224.mfc
Created lattice with 9 nodes / 8 arcs from label file
qul aAAoothu bi rabbi nnasi maliki nnasi ilahi nnas  ==  [895 frames] -75.8720 [Ac=-67905.5
Aligning File: data/train/mfc/225.mfc
Created lattice with 6 nodes / 5 arcs from label file
min sharri al waswasi al khannas  ==  [594 frames] -76.4397 [Ac=-45405.2 LM=0.0] (Act=8.7)
Aligning File: data/train/mfc/226.mfc
Created lattice with 11 nodes / 10 arcs from label file
al lathee yuwaswisu fee sudoori nnas mina al jinnati wa nnas  ==  [1174 frames] -77.8622 [A

No HTK Configuration Parameters Set
```

Figure IV.10 : Phrases du coran en sortie pour le teste de performance.

Nous jugeons notre travail comme étant un petit plus. Après la construction d'un reconnaisseur de parole arabe que nous avons amélioré par l'approche proposée, nous remarquons qu'il ya toujours une épine qui est les sons pharyngaux qui ne sont pas présents dans d'autre langues comme l'anglais ou l'espagnol, ne sont pas distingués complètement par notre système encore plus de phonèmes et de réaliser une véritable amélioration de la reconnaissance de la parole arabe.

IV.8. Conclusion

Nous avons appuyé nos résultats de l'hypothèse proposée dans le chapitre trois, celle de l'ajout du nombre de Gaussiennes mélangés augmente le taux de reconnaissance de mots lorsqu'il s'agit de longue durée de sons tel que les voyelles comme celles employées par le célèbre lecteur « alsudaissi » lors de l'énoncé de la problématique dans le chapitre I. Contrairement au HMM qui est très satisfaisant dans le cas des chiffres connectés, mais qui n'arrive pas à compenser ce manque lorsqu'il s'agit de mots connectés arabe plus compliqués, et qui comprend une sorte de chant accompagnant la parole, comme le cas du saint coran.

Sauf qu'il reste à proposer les perspectives dans le chapitre qui suit pour pallier aux deux limitations de notre travail qui sont le taux de reconnaissance de phrases qui reste bas ; ainsi que les résultats du système indépendant du locuteur qui reste en dessous des chiffres des trois autres systèmes en taux de reconnaissance de mots.

Nos travaux ont été sanctionnés par une publication en décembre 2016 (Merad, Benyettou & Rubio, 2016).

Chapitre V

Conclusions et Perspectives

Nous avons défini notre motivation de cette thèse dans le chapitre 1, nous avons déterminé les objectifs, ainsi posé les problématiques toujours dans le chapitre de l'introduction pour pouvoir discuter sur les contributions de cette thèse par rapport aux autres travaux que nous avons vu dans la première partie de la thèse qui est celle de « l'état de l'art ».

Nous avons procédé par idée en établissant l'état de l'art. A commencer par les travaux connexes sur les applications reconnaissance de la parole dans un premier temps, puis sur la langue arabe dans un deuxième temps après avoir vu ses défis dans le chapitre (2).Il était grand temps de définir une solution qui résous la problématique posé dans le chapitre (1), dite : contribution de la thèse qui sera détaillé dans le chapitre (4).

Avant cela, le chapitre (3) présente une nouvelle approche à la reconnaissance de parole arabe, qui relie les modèles tri-phones par une gaussienne uni-varié dans un premier temps afin d'attacher les sons identiques, puis les élargir aux modèles de mélanges gaussiens afin de reconnaître ces sons même s'ils continuent et se maintiennent dans le temps.

Dans le chapitre (4), nous verrons l'algorithme de tri-phones élargis aux GMMs en détail, en plus les résultats de nos trois expériences afin de les analyser puis discuter leurs avantages et inconvénients.

Dans le chapitre (5), nous avons appuyé nos résultats de l'hypothèse proposée dans le chapitre trois lors de l'énoncé de la problématique dans le chapitre un, et prouvé dans le chapitre quatre, sauf qu'il reste à proposer les perspectives pour pallier aux limites de notre travail.

1 Résumé des contributions

Le premier apport de cette thèse est plus ou moins courant, celui de la coarticulation. En utilisant le HMM pour représenter le phonème, on parle de *modèles indépendant du contexte,* cela engendre des problèmes de reconnaissance de sons vue que chaque son ou phonème varie énormément dépendamment des mouvements des articulateurs (la longue, les lèvres.. etc) des sons ou phonèmes prononcés avant ou après le phonème étudié. Une solution est donnée, c'est l'application de *modèles dépendant du contexte* appelés tri-phones : chaque phonème est représenté par la concaténation de trois HMM., pour pallier aux défauts de coarticulation.

Cette technique exécutée seule a régressé les résultats, ce qui est normale, c'est due au manque de données pour que les phonèmes soient appris selon leur contexte de début, de milieu ou de fin du mot par les modèles HMM tri-phones.

Lorsque nous avions à faire avec de célèbres lecteurs qui lisent le coran en suivant les règles de Tajweed, nous avons remarqué que chez le lecteur « alsudaissi », les phonèmes étaient plus difficiles à reconnaitre, par rapport aux lecteurs « alghamidi » et « elmirigli ». Nous avions du hybrider la technique des tri-phones avec celle de mélange gaussiens, après avoir attaché les tri-phones qui ont des phonèmes ressemblant caractéristiquement.

C'est pour cela notre deuxième apport de cette thèse est les *modèles de tri-phonesà états liés élargis aux mélange gaussiens* appelé les gaussiennes multi-variée

En faisant la comparaison avec un autre travail similaire au notre, nous avons noté que les auteurs avaient utilisé une autre méthode pour améliorer les résultats. Leurs résultats étaient meilleurs que les notre pour tous les lecteurs, mis à part pour « alsudaissi » qui les a surpassé de 5%.

D'où notre contribution proposée : après avoir entraîné les HMMs tri-phones, il faut lier les phonèmes qui se ressemble par une gaussienne uni-variée, puis les étendre aux mélanges Gaussiens ou appelé gaussiennes multi-variées. Ainsi les résultats s'amélioraient au fur d'augmenter le nombre de GMMs. Nous avons obtenu les meilleurs résultats pour un nombre de GMM égal à 512.

En comparant notre travail avec ceux qui ont utilisé tout le coran mais seulement la lecture normale du coran, pas la lecture dans les règles de Tajweed, nous les avons dépassés par 2.39 % de taux de mots reconnus.

Ce projet de recherche a pour intérêt de servir aux personnes malvoyantes qui ne peuvent qu'écouter le texte. Que ce texte soit en arabe ou en une autre langue, ce qui importe c'est que l'intérêt de cette recherche est d'être appliqué sur du texte lu d'une manière rythmique, comme du chant. C'est pour cela nous l'avons appliqué sur la lecture du coran indépendante du locuteur.

2 Future recherche

Ainsi, nous avons pu apporter un petit plus aux problèmes de coarticulation et de variation des sons arabes, ceci dit il reste à améliorer davantage la reconnaissance dans le système indépendant du locuteur qui pourrait être résolu en augmentant encore les données et en testant plus de locuteurs. Aussi, et surtout il reste à améliorer le taux de phrases reconnues.

Pour cela, nous proposons, en tant que perspective, d'essayer des modèles acoustiques basés sur des classificateurs postérieurs afin d'améliorer le taux de SR et le système indépendant du locuteur.

Comme le classificateur gaussien est loin du classificateur de vraisemblance acoustique le plus couramment utilisé, il est possible d'utiliser des classificateurs qui sont des estimateurs postérieurs, tels que les réseaux neuronaux (NN) ou SVM en les intégrant à une architecture HMM.

Ces approches hybrides (HMM-SVM et HMM-MLP) estimeront la probabilité d'un vecteur caractéristique cepstral en une seule temps*t* exactement comme GMM-Triphone.

En outre, les approches postérieures utilisent une grande fenêtre d'information acoustique parce qu'elles ont un contexte important, ce qui convient à l'amélioration du taux SR.

Contrairement à ce que nous avons fait dans notre travail, SVM et NN utilisent des phonèmes plutôt que des sous-phonèmes ou de tri-phones, et calculent un postérieur pour chaque phonème. C'est pourquoi nous encourageons vivement les tests dans ce but.

Références bibliographiques

Abdou, S. M., Hamid, S. E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., & Nazih, W. (2006). Computer aided pronunciation learning system using speech recognition techniques. Paper presented at the Interspeech.

Abolohom, A., & Omar, N. (2014). A machine learning approach to anaphora resolution in Arabic. International Review on Computers and Software, 9(12), 1956-1963.

Adetunmbi, O., Obe, O., & Iyanda, J. (2016). Development of Standard Yorùbá speech-to-text system using HTK. International Journal of Speech Technology, 19(4), 929-944.

Al-Tamimi, J. (2009). Effect of pharyngealisation on vowels revisited: Static and Dynamic analyses of vowels in Moroccan and Jordanian Arabic. Paper presented at the Workshop on Pharyngeals and Pharyngealisation.

Almisreb, A. A., Abidin, A. F., & Tahir, N. M. (2015). Investigation of Dynamic Time Warping and Neural Network for Arabic phonemes recognition based Malay speakers. Paper presented at the System Engineering and Technology (ICSET), 2015 5th IEEE International Conference on.

Alotaibi, Y. A. (2012). Comparing ANN to HMM in implementing limited Arabic vocabulary ASR systems. International Journal of Speech Technology, 15(1), 25-32.

Alotaibi, Y. A., Alghamdi, M., & Alotaiby, F. (2010). Speech recognition system of Arabic alphabet based on a telephony Arabic corpus. Paper presented at the International Conference on Image and Signal Processing.

Alshalabi, R. (2005). Pattern-based stemmer for finding Arabic roots. Information Technology Journal, 4(1), 38-43.

Alsulaiman, M., Mahmood, A., & Muhammad, G. (2017). Speaker recognition based on Arabic phonemes. Speech communication, 86, 42-51.

Amrous, A. I., Debyeche, M., & Amrouche, A. (2011). Prosodic Features and Formant Contribution for Arabic Speech Recognition in Noisy Environments. Paper presented at the Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011.

Anusuya, M., & Katti, S. K. (2010). Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267.

Association, E. L. R. (2006). On WWW at http://www. elra. info: Accessed.

Baig, M. M. A., Qazi, S. A., & Kadri, M. B. (2015). Discriminative Training for Phonetic Recognition of the Holy Quran. Arabian Journal for Science and Engineering, 40(9), 2629-2640.

Bellem, A. (2009). The “problem” of pharyngealization and its role in the sound systems of North-East Caucasian languages. Paper presented at the poster presented at the International Workshop on Pharyngeals and Pharyngealization, Newcastle.

Billa, J., Noamany, M., Srivastava, A., Liu, D., Stone, R., Xu, J., . . . Kubala, F. (2002). Audio indexing of Arabic broadcast news. Paper presented at the Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on.

Bourouba, E.-H., Bedda, M., & Djemili, R. (2006). DTW/GHMM.

Canavan, A., Zipperlen, G., & Graff, D. (1997). Callhome egyptian arabic speech. Linguistic Data Consortium.

Çömez, M. A. (2003). Large vocabulary continuous speech recognition for Turkish using HTK. MIDDLE EAST TECHNICAL UNIVERSITY.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern Classification and Scene Analysis Part 1: Pattern Classification. Wiley, Chichester.

El Moubtahij, H., Halli, A., & Satori, K. (2014). Arabic Handwriting Text Offline Recognition Using the HMM Toolkit (HTK). Computers and Software, 1214.

Elhadj, Y. O. M., Alghamdi, M., & Alkanhal, M. (2014). Phoneme-based recognizer to assist reading the Holy Quran Recent Advances in Intelligent Informatics (pp. 141-152): Springer.

Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2009). Survey on common Arabic language forms from a speech recognition point of view. Paper presented at the proceeding of International conference on Acoustics (NAG-DAGA).

Esling, J. H. (1999). The IPA Categories “Pharyngeal” and “Epiglottal” Laryngoscopic Observations of Pharyngeal Articulations and Larynx Height. Language and Speech, 42(4), 349-372.

Gales, M. J., Diehl, F., Raut, C. K., Tomalin, M., Woodland, P. C., & Yu, K. (2007). Development of a phonetic system for large vocabulary Arabic speech recognition. Paper presented at the Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.

Girgis, N. (2009). Pharyngealized fricatives in Egyptian Arabic: heritage vs. non-heritage speakers. Paper presented at the Newcastle University International Workshop on Pharygeals and Pharyngealisation. Or
http://www. ncl. ac.uk/linguistics/assets/documents/NoahGirgis_000. pdf.

Hassine, M., Boussaid, L., & Messaoud, H. (2016). Maghrebian dialect recognition based on support vector machines and neural network classifiers. International Journal of Speech Technology, 19(4), 687-695.

Herbig, T., Gerl, F., & Minker, W. (2012). Self-learning speaker identification for enhanced speech recognition. Computer Speech & Language, 26(3), 210-227.

Heselwood, B. The problem of classifying and describing pharyngeals.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., & Rybchenko, S. I. (2013). Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. Speech communication, 55(1), 22-32.

Holes, C. (2004). Modern Arabic: Structures, functions, and varieties: Georgetown University Press.

Huang, X., Acero, A., Hon, H.-W., & Foreword By-Reddy, R. (2001). Spoken language processing: A guide to theory, algorithm, and system development: Prentice hall PTR.

Hyassat, H., & Zitar, R. A. (2006). Arabic speech recognition using SPHINX engine. International Journal of Speech Technology, 9(3-4), 133-150.

Ibrahim, N. J., Razak, Z., Yusoff, Z. M., Idris, M. Y. I., Tamil, E. M., Noor, N. M., & Rahman, N. N. A. (2008). Quranic verse recitation recognition module for support in j-QAF learning: A review. International Journal of Computer Science and Network Security (IJCSNS), 8(8).

Jafri, A., Sobh, I., & Alkhairy, A. (2015). Statistical Formant Speech Synthesis for Arabic. Arabian Journal for Science & Engineering (Springer Science & Business Media BV), 40(11).

Kumar, K., Aggarwal, R., & Jain, A. (2012). A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering, 1(1), 25-32.

Labidi, M., Maraoui, M., & Zrigui, M. (2016). New birth of the Arabic phonetic dictionary. Paper presented at the Engineering & MIS (ICEMIS), International Conference on.

Laufer, A. (2009). Pharyngeal articulation in Hebrew. Paper presented at the International workshop on pharyngeals and pharyngealization, Newcastle University.

Lee, K.-F., Hon, H.-W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1), 35-45.

Li, K. P., & Hughes, G. (1974). Talker differences as they appear in correlation matrices of continuous speech spectra. The Journal of the Acoustical Society of America, 55(4), 833-837.

Lui, X. (2007). A study of variable parameter Gaussian mixture HMM modeling fro Noisy speech recognition. IEEE Transactions on Audio, Speech and Language processing, 15(1).

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition, 710.

Moore, R. K. (1994). Twenty things we still don't know about speech. Paper presented at the Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology.

Mourtaga, E., Sharieh, A., & Abdallah, M. (2007). Speaker independent Quranic recognizer based on maximum likelihood linear regression. Paper presented at the Proceedings of world academy of science, engineering and technology.

Nahar, K. M., Shquier, M. A., Al-Khatib, W. G., Al-Muhtaseb, H., & Elshafei, M. (2016). Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. International Journal of Speech Technology, 19(3), 495-508.

Newman, D. (2002). The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad. Antwerp papers in linguistics., 100, 65-75.

Ouni, S., & Laprie, Y. (2009). Studying pharyngealisation using an articulograph. Paper presented at the International Workshop on Pharyngeals and Pharyngealisation.

Park, H., & Yoo, C. D. (2014). Tied triphone semi-Markov model for large vocabulary continuous speech recognition. Paper presented at the The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014).

Pirhosseinloo, S., & Ganj, F. A. (2012). Discriminative speaker adaptation in Persian continuous speech recognition systems. Procedia-Social and Behavioral Sciences, 32, 296-301.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Sakai, T., & Doshita, S. (1962). The Phonetic Typewriter. Paper presented at the IFIP Congress.

Saon, G., & Soltau, H. (2012). Boosting systems for large vocabulary continuous speech recognition. Speech communication, 54(2), 212-218.

Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., & Makhoul, J. (1985). Context-dependent modeling for acoustic-phonetic recognition of continuous speech. Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.

Seid, H., Yegnanarayana, B., & Rajendran, S. (2012). Spotting glottal stop in Amharic in continuous speech. Computer Speech & Language, 26(4), 293-305.

Smaragdis, P., & Raj, B. (2012). The Markov selection model for concurrent speech recognition. Neurocomputing, 80, 64-72.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. Paper presented at the Interspeech.

Tahiry, K., Mounir, B., Mounir, I., & Farchi, A. (2016). Energy bands and spectral cues for Arabic vowels recognition. International Journal of Speech Technology, 19(4), 707-716.

Taqi, H. The realization of (s) as [s ʕ] in Kuwaiti Arabic.

Taylor, P. (2009). Text-to-speech synthesis: Cambridge university press.

Velichko, V., & Zagoruyko, N. (1970). Automatic recognition of 200 words. International Journal of Man-Machine Studies, 2(3), 223-234.

Weide, R. (1996). The Carnegie Mellon Pronouncing Dictionary v0. 4: Carnegie Mellon University.

Windmann, S., & Haeb-Umbach, R. (2009). Approaches to iterative speech feature enhancement and recognition. IEEE Transactions on audio, speech, and Language processing, 17(5), 974-984.

Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., . . . Rashwan, M. (2006). Building annotated written and spoken Arabic LR's in NEMLAR project. Paper presented at the Proceedings of LREC.

Yekache, Y., Mekelleche, Y., & Kouninef, B. (2012). Towards Quranic reader controlled by speech. arXiv preprint arXiv:1204.1566.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Povey, D. (2006). The HTK book (for HTK version 3.4). Cambridge university engineering department, 2(2), 2-3.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Povey, D. (2006). The HTK book (v3. 4). Cambridge University.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). HTK. World Wide Web, http://www. htk. eng. cam. ac. uk.

Zarrouk, E., Benayed, Y., & Gargouri, F. (2015). Graphical models for the recognition of Arabic continuous speech based triphones modeling. Paper presented at the Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on.

Zeroual, C., Esling, J., & Hoole, P. (2011). EMA, endoscopic, ultrasound and acoustic study of two secondary articulations in Moroccan Arabic. Instrumental studies in Arabic phonetics, 319, 277.