

Étude de la Paramétrisation RASTA PLP en vue de la Reconnaissance Automatique de la Parole Arabe

Houda HOSNI*, Zied SAKKA*, Abdennaceur KACHOURI*
and Mounir SAMET*

* *Laboratoire d'Électronique et des Technologies de l'Information (LETI)
École Nationale d'Ingénieurs de Sfax B.P.W, 3038 Sfax, TUNISIE*

houda.hossni@gmail.com

mounir.samet@enis.rnu.tn

zied.sakka@yahoo.fr

Abdennaceur.kachouri@enis.rnu.tn

Résumé: Aujourd'hui, la reconnaissance vocale est un domaine à forte croissance grâce au développement considérable de nouveaux outils de traitement automatique de la parole. Un système de reconnaissance acquiert sa performance particulièrement des techniques de paramétrisation du signal vocal. Dans ce cadre, notre travail consiste à évoquer le fonctionnement d'un système de reconnaissance de traits phonétiques de l'arabe en commençant par l'extraction des coefficients acoustiques via la méthode « RASTA-PLP » (Relative Spectral PLP) qui a abouti à des résultats très satisfaisants. Nous proposons un algorithme de modélisation des paramètres acoustiques basé sur les modèles de Markov cachés.

Mots clés: Reconnaissance vocale; langue arabe; RASTA-PLP; HMM.

INTRODUCTION

La parole est devenue de plus en plus partie des interfaces multimédia homme machine, sa reconnaissance constitue donc un défi à relever.

Depuis une trentaine d'années, la reconnaissance vocale fait l'objet d'un nombre croissant d'études ayant pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse, autrement elle vise à étudier le contenu linguistique du message énoncé par un locuteur donné. Le premier système pouvant être considéré comme faisant de la reconnaissance vocale date de 1952 mais ses performances se limitaient à reconnaître des chiffres isolés. En 1985 un système de reconnaissance numérique en langue Arabe a vu le jour signalant le déclenchement des recherches linguistiques fructueuses en Arabe, qui, bien que moins nombreuses en comparaison avec d'autres langues, connaissent ces dernières années un regain d'intérêt. [HAT 06]. Alors que certains travaux de recherche ont été consacrés aux caractères isolés d'autres se sont orientés à la reconnaissance des mots isolés.

La paramétrisation du signal vocal permet d'obtenir une " empreinte " caractéristique du son, sur laquelle on pourra ensuite travailler pour la reconnaissance. Ainsi, le choix d'un ensemble

pertinent de caractéristiques constituera un point critique qui aura une très grande influence sur la performance d'un tel classificateur.

L'utilisation des caractéristiques sonores d'inspiration psycho acoustique tel que les filtres RASTA ainsi que les caractéristiques cepstrales ont été montrées avantageuses pour différents problèmes de caractérisation et de reconnaissance du signal sonore. Cependant, le succès de l'utilisation des coefficients cepstraux a été freiné par des facteurs liés à la variabilité entre les données d'entraînement et de tests, le bruit ainsi que les distorsions introduites par le canal de transmission.

Dans cette perspective, et pour s'inspirer par la perception humaine des classes sonores afin de développer un espace de représentation convenable pour notre classification, nous avons opté pour le choix de la paramétrisation RASTA-PLP qui modifie l'amplitude des fréquences simulant en quelque sorte l'appareil auditif humain, et permet de garder les informations importantes du signal.

Notre travail a pour intention d'appliquer la technique RASTA-PLP sur le module d'extraction des paramètres afin d'augmenter le taux de reconnaissance.

Nous considérons ici une modélisation s'appuyant sur les Modèles de Markov Cachés qui fournissent des

outils particulièrement performants dans le domaine de la reconnaissance.

Dans la deuxième section de ce travail, nous représentons notre système de reconnaissance et nous nous concentrons sur l'étape de paramétrisation. Une description de la base utilisée sera détaillée dans la section 3. La section 4 sera consacrée pour l'interprétation des résultats obtenus.

1. Présentation du système de reconnaissance

La reconnaissance vocale est une succession de modules dont l'étape finale est de reconnaître le signal de parole que l'on met à l'entrée de cette chaîne. Nous avons introduit dans notre système toutes les notions de base nécessaires pour sa définition comme l'indique la figure 1. Ainsi nous disposons de 4 étages de traitement ; un module d'extraction de paramètres acoustiques suivi d'un module acoustico-phonétique gérant les HMMs, chaque phonème est représenté par une sorte d'automate à 3 états. Les modèles lexicaux fournis par un dictionnaire phonétique constituent le module d'identification des mots et les modèles de langage trigrammes associent une probabilité à toute suite de mots présents dans le lexique pour reconnaître les phrases enregistrées. [SER 06].

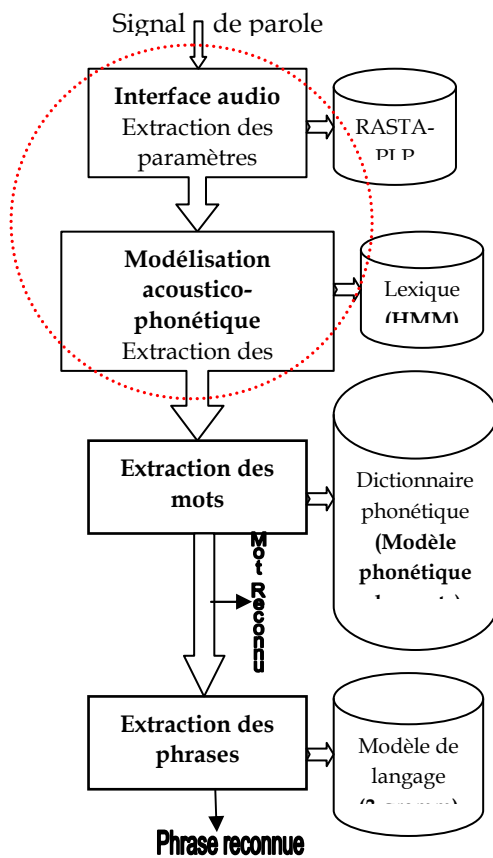


Fig.1. Description du système de reconnaissance automatique de la parole

Dans cet article, notre étude concerne essentiellement les phases de paramétrisation et la modélisation acoustico-phonétique. Ceci revient à considérer un système où la tâche est d'identifier un phonème parlé où nous devons supposer un vocabulaire de R unités phonétiques à reconnaître, chaque phonème doit être modélé par un HMM distinct. Le traitement de la figure 2 doit être alors suivi.

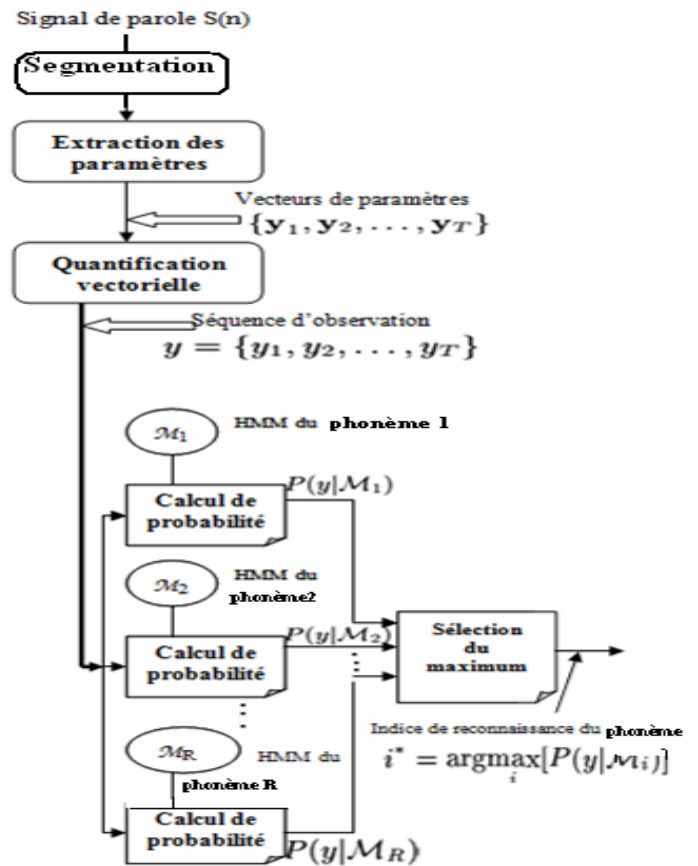


Fig.2. Schéma fonctionnel d'un système de reconnaissance de phonèmes par HMM.

1.1. Segmentation du signal audio

La première étape que nous voulons isoler est l'étape de segmentation, ceci est motivé par le fait que l'utilisation d'unités plus petites que le mot possède un certain nombre d'avantages.

Les unités phonétiques nécessaires pour la modélisation sont obtenus à travers un segmentation (découpage) des mots en sous unités acoustiques en se concentrant sur quelques phonèmes, ceux qui possèdent les plus grandes fréquences d'apparition dans notre base arabe décrite ultérieurement et se référant aux statistiques de certains chercheurs arabes concernant les fréquences d'occurrences des phonèmes arabes dans la langue. Les phonèmes choisis sont alors : /a/: همزة, /l/: لام & /n/: نون, /m/: ميم. Par contre, pour le mot « يسا ر » on s'est intéressé au phonème /s/.

Après une longue phase de segmentation

(découpage) des mots et des phrases arabes en phonèmes les plus fréquents, nous avons obtenu à la fin 4840 fichiers de différents phonèmes.

1.2. Extraction des paramètres RASTA-PLP

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes.

Un tel système prend un signal en entrée et retourne un vecteur de paramètres (appelé vecteur acoustique ou encore vecteur d'observations). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

Pour la réalisation de cette phase d'extraction des paramètres, nous avons utilisé la technique RASTA-PLP qui est en fait la méthode de prédiction linéaire perceptuelle (PLP) couplée à la technique spectrale relative (RASTA).

La méthode de prédiction linéaire perceptuelle (PLP), étudiée par (Hermansky 1991), consiste en un filtrage en bandes critiques du spectre à court terme suivi d'une correction de l'intensité. [HER 91]

L'amplitude du signal est alors compressée et enfin l'analyse par prédiction linéaire intervient.

Cette dernière étape est en réalité une technique de compression spectrale qui modifie le spectre de puissance à court terme avant son approximation par un modèle autorégressif.

Les zones du signal qui correspondent à des composantes linguistiques n'ont pas la même évolution temporelle que les zones qui ne correspondent pas. La technique spectrale RASTA utilise ce principe en supprimant les composantes spectrales dont l'évolution temporelle est plus rapide ou plus lente que celle du conduit vocal.

A la limite des méthodes de rehaussement de la parole, les représentations obtenues à l'aide des filtres RASTA visent à supprimer le bruit en utilisant les variations temporelles pour différencier le bruit du signal de parole. Le plus souvent, le signal clair est perturbé par un bruit ambiant de variation lente. Les filtres sont donc définis pour éliminer les signaux stationnaires. Ils consistent à calculer un banc de filtre représentant l'énergie du signal à différentes bandes de fréquence sur une échelle logarithmique, puis à filtrer chaque canal par un filtre passe haut avec une pente raide éliminant les très basses fréquences (1 Hertz), éliminant ainsi les variations lentes dues aux modifications de l'environnement. [HER 91]

Différentes études réalisées avec cette méthode ont permis de confirmer ces bonnes qualités relativement aux distorsions et ses moindres qualités face aux bruits qualifiés d'additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

Le processus de calcul des coefficients RASTA

PLP peut être décomposé en 4 étapes décrites par la figure suivante (fig.3). [BRI 97].

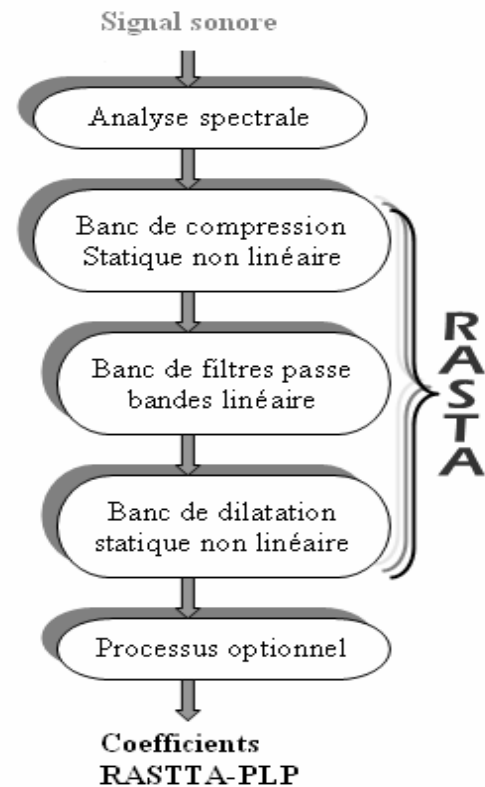


Fig.3. Chaîne de calcul des coefficients RASTA PLP.

Un exemple de calcul de quelques paramètres du signal de parole utilisant cette méthode d'extraction est illustré par la figure 4.

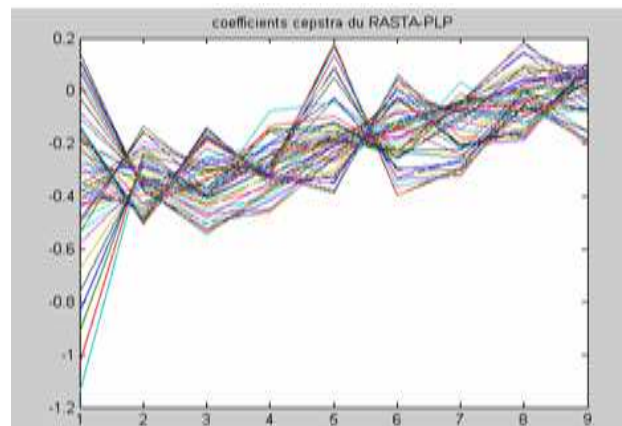


Fig.4. Coefficients cepstra du RASTA-PLP.

1.3. Quantification vectorielle

Les HMMs utilisés pour la modélisation des phonèmes arabes sont de nature discrète, ainsi que leurs densités de probabilités d'observations, ce qui justifie la nécessité d'utilisation d'un quantifieur vectoriel pour faire correspondre chaque vecteur continu (représentant une trame) à un indice discret d'un dictionnaire de référence (CodeBook), ainsi on peut assimiler le rôle de la quantification vectorielle à la conception du CodeBook. Une fois le dictionnaire

de référence obtenu, cette correspondance entre les vecteurs caractéristiques des trames et les indices du CodeBook devient un simple calcul de type plus proche voisin.

Cette procédure consiste en premier lieu à partitionner les vecteurs caractéristiques issus de l'étape de paramétrisation en K ensembles disjoints (où K est la taille du CodeBook à concevoir). En deuxième lieu, à représenter chaque ensemble par un vecteur unique (point typique) qui est généralement le centroïde des vecteurs caractéristiques de l'ensemble d'apprentissage affectés à la même région, ensuite elle optimise itérativement la partition du CodeBook. Le centroïde est le point qui minimise la distance pour l'ensemble de tous les points d'une classe. Idéalement, tous les éléments d'une classe doivent être plus proches de leur centroïde que de tout autre centroïde. Ainsi, un vecteur inconnu X , est classé comme élément de la classe c si la distance de X , au centroïde de c est plus petite que la distance entre X , et tous les autres centroïdes.

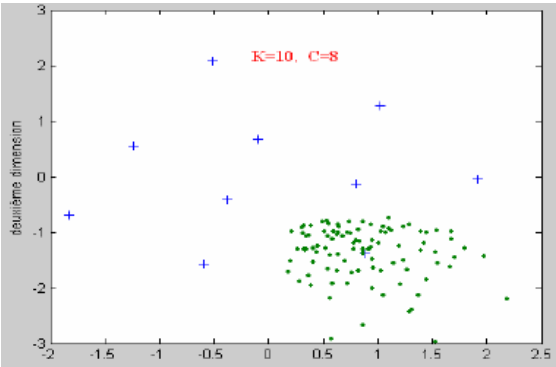
La figure 6.a illustre le processus de la reconnaissance, formé de deux dimensions de l'espace acoustique. Les '.' font référence aux vecteurs acoustiques et le centroïde du résultat est montré par des marques '+'.


Fig.6.a. Vecteurs aléatoires avec les « centroïdes » et les points d'assignement. $c=8$, $k=10$.

La taille K du CodeBook est un paramètre crucial dont la valeur affecte en grande partie les performances des HMMs utilisés pour la reconnaissance. Nous avons choisi $K=16$ afin d'obtenir une meilleure performance. Dans la figure 6.b nous représentons les éléments du vecteur du codebook pour les phonèmes arabes. [BEN 06].

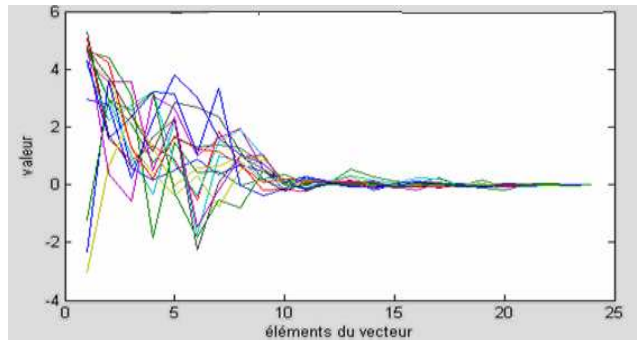


Fig.6.b. Vecteurs du codebook pour les phonèmes arabes, k (taille du CodeBook)=16

Les principaux avantages de la quantification vectorielle sont : espace réduit nécessaire pour le stockage du dictionnaire et puissance de calcul réduite.

1.4. Modélisation acoustico-phonétique

La modélisation acoustico-phonétique doit prendre en compte les sources de variabilité dans la production du signal vocal. L'indépendance par rapport au locuteur est obtenue en estimant les paramètres des modèles acoustiques à l'aide des corpus de parole contenant les enregistrements de plusieurs locuteurs, et la variabilité contextuelle se traduit par l'utilisation d'un grand nombre d'unités phonétiques dépendantes du contexte phonétique local. [BAU 72]. Ces unités sont les phonèmes issus de notre base arabe suite à un découpage qui sera détaillé dans la section suivante.

Les techniques stochastiques, tel que les HMMs, sont actuellement les plus utilisées pour la modélisation acoustique de la parole permettant de déterminer la probabilité d'une suite d'observations.

Un tel classifieur est basé sur un critère de maximum de vraisemblance, il prend le mot à reconnaître comme étant une séquence d'observations discrètes (codes) produites par analyse et quantification vectorielle de la séquence de vecteurs de caractéristiques. Ce classifieur calcule la probabilité qui correspond à la probabilité d'obtenir la séquence par le modèle. Ces probabilités sont évaluées par la version logarithmique de l'algorithme de Viterbi. Finalement, le mot testé est affecté à la classe du mot K du lexique L pour laquelle le modèle maximise la probabilité d'émission. [FAR 94]

Dans notre système et pour la base arabe utilisée, un codebook est produit d'une séquence d'entraînement de 40 occurrences de chaque mot, $T = 3849$ vecteurs sont employés pour trouver les 16 centroïdes désirés de cluster. Dans ce cas-ci, la déformation moyenne résultante est de 0.045983.

En utilisant les mêmes fichiers son d'entraînement, un ensemble de 40 HMMs sont spécialisés en détectant un mot spécifique, c.-à-d., les mots parlés de la base Arabe. La figure 7 montre les log-likelihoods (log de vraisemblance) résultants pour chaque HMM comme fonction du nombre d'itération dans

l'algorithme de re-estimation de FB. Les maximums locaux obtenus dépendent de l'initialisation aléatoire de l'algorithme de FB.

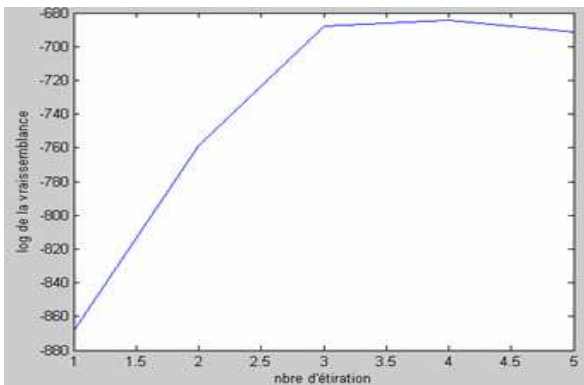


Fig.7. Représentation du log de la vraisemblance en fonction du nbre d'itération.

2. Base arabe

Suite aux recherches fructueuses concernant la langue arabe que certains sont dernièrement orientés pour, nous avons opté pour l'implémentation d'un système de reconnaissance de la parole en arabe. Afin d'évaluer l'approche proposée dans cet article, Nous avons employé un corpus préparé au sein de notre laboratoire de recherche.

Le travail effectué a concerné tout d'abord l'enregistrement, dans des conditions favorables correspondantes à un échantillonnage de 8 bits à une fréquence d'échantillonnage de 16 KHz, d'une base de données comportant 4 mots et 2 phrases arabes cités ci-dessous étant prononcées par 22 locuteurs de différents sexes et ages (14hommes, 5 femmes et 3enfants).

« خلف », « أمام », « يسار », « يمين »,
« دع عنك لومي فان اللوم إغراء »

Et

« أهدى لنا جارنا ما صاده في الغاب »

Après la phase de numérisation du signal vocal enregistré, nous avons été amenés à effectuer une phase d'isolation (découpage) entre les mots et les phrases prononcées par un locuteur quelconque plusieurs fois de suite afin d'éliminer les silences et les parasites intermédiaires (glissement, respiration, hésitation ...etc.).

3. Résultats obtenues par la méthode de paramétrisation RASTA PLP

Notre objectif est d'évaluer la méthode de paramétrisation RASTA PLP en vue de construire un système de reconnaissance arabe de phonèmes et de phrases robuste.

Nous avons considéré deux modes d'apprentissage.

*Pour l'apprentissage locuteur inconnu le nombre

d'occurrence pour l'entraînement est 60 (du locuteur 1 au locuteur 6) et le nombre d'occurrence pour le test est 142 (du locuteur 7 au locuteur 22).

*Pour l'apprentissage multi locuteur le nombre d'occurrence pour l'entraînement est 60 (du locuteur 1 au locuteur 6) et le nombre d'occurrence pour le test est 182 (du locuteur 4 au locuteur 22)

La performance du système totale était de 76.05 % pour un système utilisant un modèle HMM à trois états et 83.56 % avec un modèle HMM à cinq états qui est raisonnablement haute et cela pour un apprentissage locuteur inconnu (tableaux 1 et 2).

La performance totale du système était de 84.03 % pour un système utilisant un modèle HMM à trois états et 85.91 % avec un modèle HMM à cinq états qui est raisonnablement haute, pour un apprentissage multi locuteur (tableaux 3 et 4).

Les résultats obtenus sont satisfaisants surtout qu'ils sont comparables à d'autres résultats existants. En effet, la technique 'RASTA-PLP' a prouvé sa supériorité par rapport à d'autres techniques.

Apprentissage locuteur inconnu*Tableau 1.** *Matrice de confusion relative à un système de reconnaissance de phonèmes Arabes utilisant Un modèle HMM à trois états.*

	Phonème 'm' du mot 1	Phonème 's' du mot 2	Phonème 'a' du mot 3	Phonème 'l' du mot 4	Phonème 'a' de la phrase 5	Phonème 'n' de la phrase 6	Taux de reconnaissance
Phonème 'm' du mot 1	121	–	9	10	–		85.21%
Phonème 's' du mot 2	8	115	14	3	8	9	80.98%
Phonème 'a' du mot 3	5	2	104	7	24	–	73.26%
Phonème 'l' du mot 4	–	21	1	114	6	–	80.28%
Phonème 'a' de la phrase 5	4	2	16	2	87	21	61.26%
Phonème 'n' de la phrase 6	2	–	32	1	–	107	75.35%
Total							76.05%

Tableau 2. *Matrice de confusion relative à un système de reconnaissance de phonèmes Arabes utilisant un modèle HMM à cinq états.*

	Phonème 'm' du mot 1	Phonème 's' du mot 2	Phonème 'a' du mot 3	Phonème 'l' du mot 4	Phonème 'a' de la phrase 5	Phonème 'n' de la phrase 6	Taux de reconnaissance
Phonème 'm' du mot 1	127	–	11	4	–	–	89.43%
Phonème 's' du mot 2	–	125	4	3	–	10	88.02%
Phonème 'a' du mot 3	8	–	110	1	1	22	77.46%
Phonème 'l' du mot 4	–	–	–	137	5	–	96.47%
Phonème 'a' de la phrase 5	6	2	15	1	97	21	68.3%
Phonème 'n' de la phrase 6	4	–	21	–	–	116	81.69%
Total							83.56%

Apprentissage multi locuteur*Tableau 3.** *Matrice de confusion relative à un système de reconnaissance de phonèmes Arabes utilisant un modèle HMM à trois états.*

	Phonème 'm' du mot 1	Phonème 's' du mot 2	Phonème 'a' du mot 3	Phonème 'l' du mot 4	Phonème 'a' de la phrase 5	Phonème 'n' de la phrase 6	Taux de reconnaissance
Phonème 'm' du mot 1	126	–	1	8	3	4	88.73%
Phonème 's' du mot 2	3	117	13	3	4	2	82.39%
Phonème 'a' du mot 3	9	–	116	–	–	17	81.69%
Phonème 'l' du mot 4	1	–	–	135	6	–	95.07%
Phonème 'a' de la phrase 5	2	1	11	7	103	18	72.53%
Phonème 'n' de la phrase 6	3	–	20	–	–	119	83.8%
Total							84.03%

Tableau 4. Matrice de confusion relative à un système de reconnaissance de phonèmes Arabes utilisant un modèle HMM à cinq états.

	Phonème 'm' du mot 1	Phonème 's' du mot 2	Phonème 'a' du mot 3	Phonème 'l' du mot 4	Phonème 'a' de la phrase 5	Phonème 'n' de la phrase 6	Taux de reconnaissance
Phonème 'm' du mot 1	128	–	3	7	2	2	90.14%
Phonème 's' du mot 2	1	130	10	–	–	1	91.54 %
Phonème 'a' du mot 3	7	2	114	–	–	19	80.28%
Phonème 'l' du mot 4	1	–	2	132	7	–	92.95 % %
Phonème 'a' de la phrase 5	–	4	10	4	107	17	75.35%
Phonème 'n' de la phrase 6	3	–	18	–	–	121	85.21%
Total							85.91%

* Notez que le phonème noté par « a » n'est pas la voyelle courte « a » de la langue arabe mais nous désignons par cette notation la consonne « hamza » transcrite « ء ».

4. Conclusion

Dans cet article, nous avons achevé une mise en œuvre d'algorithmes de reconnaissance des unités phonétiques, mots et phrases arabes dans lesquels nous avons appliqué la méthode de paramétrisation RASTA-PLP. Cette méthode a donné des bons taux de reconnaissances et a prouvé sa supériorité par rapport à d'autres techniques existantes telles que la paramétrisation LPCC que nous avons aussi testé. Notre système était basé sur une approche bayésienne grâce à l'utilisation des états de Markov cachés pour les phases d'apprentissage et de reconnaissance ce qui a amélioré davantage ses performances.

Notre stratégie de recherche dans le domaine de reconnaissance consiste à concevoir des systèmes dont la robustesse est liée à la fidélité de la modélisation de la parole. Ainsi, on est amené à penser dans nos recherches futurs d'améliorer davantage notre système de reconnaissance par la conception des modèles de détection des traits phonétiques permettant d'améliorer la reconnaissance des mots arabes à partir d'une pré classification du signal sonore en grande classes phonétiques et ainsi pouvoir étendre notre .

RÉFÉRENCES

- [BAU 72] Baum, L. E. An inequality and associated maximization technique in statistical estimation of probabilistic functions of finite state Markov chains. *Inequalities* 3, 1 (1972).
- [BEN 06] Benouareth, A. Ennaji, A. Sellami, M. Utilisation des HMMs de Durée d'Etat Explicite pour la reconnaissance des Mots Arabes Manuscrits. RFIA 2006, 15ème congrès francophone Reconnaissance des formes et intelligence artificielle, Tours:37-41.

[BRI 97] Brian, K et Morgan, N. Recognizing reverberant speech with RASTA-PLP. In ICASSP, volume 2, pages 1259{1262, Munich, Germany, April 1997.IEEE.

[FAR 94] Farrel, K. Mammone ,R.J. et Assaleh, K.T. "Speaker using neural networks and conventional classifiers", IEEE Transactions on Speech and Audio Processing, vol.2, N°1, pp.194-205, Janv. 1994.

[HAT 06] Haton J.P, Reconnaissance automatique de la parole : Du signal à son interprétation, Dunod Paris, 2006.

[HER 91] Hermansky, H., Morgan, N., Bayya, A, and Kohn, P. RASTA-PLP Speech Analysis, ICSI Technical Report TR-91-069, Berkeley, California.

[SER 06] Serignat .J, Système de Reconnaissance Automatique de la Parole RAPHAEL. (2006).