Reconnaissance des Phonèmes Emphatiques de l'Arabe Standard par une Approche Connexionniste Modulaire

M. Kabache * et M. Guerti **

Ecole Nationale Polytechnique.10, Avenue Hacen Badi. El Harrach – Alger. Fax : 021 52 73 29.
*E-mail : mahraz_k@yahoo.fr
** E-mail : mhaniag@Yahoo.fr

Abstract

The aim of our work is the recognition of the emphatic consonant of Standard Arabic, using the Artificial Neuron Networks. To achieve this objective we used the Multi Layer Perceptron (MLP) as a technique of recognition. The phonemes which we want to recognize are taken from sentences of the Standard Arabic, these sentences recorded by a single speaker, to take in consideration the coarticulation phenomena. We analyzed the corpus of training and the test by several techniques of acoustics analysis. The PLP (**P**erceptual **L**inear **P**rediction), RASTA-PLP (**R**el**A**tive **S**pec**T**rAl-**P**erceptual **L**inear **P**rediction), LPC (**L**inear **P**redictive **C**oding), the energy and the Rate of Passage by Zeros. The objective is to determine the one that gives the best rate of recognition.

1. INTRODUCTION

La **R**econnaissance **A**utomatique de la **P**arole (RAP) a fait l'objet de plusieurs études de recherche pluridisciplinaires depuis le début des années 50 [1]. Des progrès ont été réalisés ces dernières décennies dans ce domaine. Actuellement, des logiciels commercialisés, sont capables d'effectuer une reconnaissance de la parole continue pour un vocabulaire important. Néanmoins, les systèmes de RAP dédiés à la langue Arabe et en particulièrement aux traits phonétiques qui la caractérisent sont rares. Les particularités de la langue Arabe tels que l'emphase, la gémination et l'allongement pertinent des phonèmes posent de nombreux problèmes aux développeurs des systèmes de RAP en Arabe.
Dans cet article, nous allons présenter un système de reconnaissance des phonèmes emphatiques de l'**A**rabe **S**tandard (AS) en utilisant les Réseaux de Neurones Multicouches (Multi Layer Perceptrons : MLP). Pour réaliser cet objectif, nous avons élaboré un corpus d'apprentissage et de test constitué de 84 phrases porteuses contenant des phonèmes emphatiques de la langue Arabe. Nous justifions le choix de ce type de corpus par le fait qu'il est préférable de reconnaître les phonèmes dans des contextes réels pour prendre en considération les effets de la coarticulation (l'influence d'un son sur un autre son).
Nous avons appliqué différentes analyses acoustiques pour extraire les paramètres pertinents du signal vocal dont le but de déterminer celle qui donne le meilleur Taux de Reconnaissance. Les techniques d'analyse utilisées sont : la LPC (Linear Predictive Coding), la PLP, la RASTA-PLP (RelAtive SpecTrAl-Perceptual Linear Prediction). Chacune de ces techniques est combinée avec l'Energie et le Taux de Passage par Zéros (TPZ) [2].

2. CARACTERISTIQUES ACOUSTIQUES DES SONS EMPHATIQUES DE L'ARABE STANDARD

L'**A**rabe **S**tandard (AS) est une langue consonantique composée de 28 consonnes, trois voyelles brèves représentées par des signes diacritiques placés au-dessus ou au-dessous des consonnes, trois voyelles longues (el madd) et l'absence de voyelle ou [sukūn].
La particularité de la Langue Arabe se trouve dans la présence des consonnes arrières glottales, pharyngales, vélaires, affriquée et du phénomène d'emphase et de la gémination (*La gémination se manifeste par le renforcement de l'articulation. Elle correspond à la contraction de deux consonnes identiques en une consonne dite géminée*).
La Reconnaissance Automatique de la Parole nécessite une étude acoustique des différents phonèmes à reconnaître afin de dégager les caractéristiques relatives à ces derniers, pour les exploiter lors de développement du système de reconnaissance. Le système phonétique de l'Arabe Standard est constitué de quatre phonèmes emphatiques (Table 1).
Les phonèmes [d.] et [s.] appartiennent à la classe des fricatives de l'Arabe Standard. De point de vue articulatoire, les phonèmes fricatifs sont produits par la friction de l'air dans le conduit vocal lors d'une constriction au niveau des lèvres, des dents ou de la langue. Cette friction peut être Voisée ou Non Voisée. Dans le cas de la Reconnaissance de la Parole, ils sont caractérisés par un nombre important de Taux de Passage par Zéros (TPZ) du signal temporel, ce qui permet de les identifier parmi les autres phonèmes non-fricatifs.
Les phonèmes [d-] et [t.] sont des consonnes

occlusives, qui se produisent par la fermeture du conduit vocal (occlusion) pendant une brève durée, suivie par un brusque relâchement expirant l'air emmagasiné dans le conduit vocal. Les occlusives sont constituées de deux phases successives : une phase de silence, correspondant à l'occlusion effective, et une phase d'explosion, au moment du relâchement. Cette catégorie de sons peu être Voisées, ou Non Voisées (sourdes).
Un trait phonétique caractérisant les phonèmes [t.], [d-], [d.] et [s.] sur le plan articulatoire est appelé l'emphase. Ces phonèmes sont produit dans la partie antérieure de la cavité buccale par un report en arrière de la racine de la langue et un abaissement et creusement du dos de la langue. Au point de vue acoustique, et par opposition aux phonèmes non-emphatiques, les sons emphatiques sont caractérisés par l'élévation du premier formant (F_1) et la baisse du deuxième formant (F_2) [3].

3. DESCRIPTION DU SYSTEME

La structure du système que nous utilisons est spécialisée dans la reconnaissance des quatre phonèmes emphatiques caractérisant le système phonétique de la langue Arabe [2],[4] (Table 1).
Afin de réaliser notre système, nous avons tenu compte des étapes qui seront expliquées par la suite.

TAB. 1 : Phonèmes Emphatiques de l'Arabe Standard Avec leur durées moyennes.

Phonèmes de l'AS	Code API	Durée moyenne (ms)
[ط]	[t.]	182
[ض]	[d-]	115
[ظ]	[d.]	96
[ص]	[s.]	129

3.1. Acquisition des données

La première étape dans l'élaboration de notre système de RAP est le choix du corpus d'apprentissage et de test. Nous avons pris un corpus constitué de 84 phrases porteuses contenant tous les phonèmes emphatiques de l'Arabe Standard pris dans différentes positions (Initiale, Médiane et Finale) (Table 2). L'acquisition de ces phrases est effectuée en utilisant un matériel spécifique au traitement du signal vocal (Sonagraphe Kay 5500). Les enregistrements ont été effectués par un seul locuteur, au sein du laboratoire de la Communication Parlée du **C**entre des **R**echerches **S**cientifiques et **T**echniques pour le **D**éveloppement de la **L**angue **A**rabe (CRSTDLA) – Alger (Algérie). La fréquence d'échantillonnage choisie est de 11025 Hz, les échantillons ont été codés sur 16 bits par échantillon.

TAB. 2 : Echantillons extraits de phrases du corpus d'apprentissage et de test.

لكنه ضرب به المثل	نصر الرجل الولد بعد نهاية المقابلة
طبخ الطباخ طعاما في المطبخ الكبير	شطر اللعبة إلى نصفين
عطب الفرس بعد سقوطه	غطست غواصة في عمق المحيط
يقل جسده و ينخص	إن الوضوح نقيض الغموض
ضمى المريض ثم عافى	ألقى بنفسه فوق الفراش و بسط
عبر السنين تتغير المرء وترص	صاغ الماء بسهولة كبيرة
قضم التفاحة	صبر فعل ماضي ثلاثي

3.2. Segmentation phonémique

La phase de segmentation joue un rôle très important dans les systèmes de reconnaissance vocale et nécessite un intérêt particulier de notre part. La segmentation phonémique du corpus a été effectuée manuellement en utilisant le *speech analyzer* version 1.5 sous windows. Après chaque segmentation, nous effectuons des tests d'écoute afin de s'assurer que la segmentation a été bien faite.

3.3. Analyse et traitement des données

Durant la phase de traitement des données, le signal vocal (segments phonémiques) est préaccentué pour rehausser les hautes fréquences qui sont moins énergétiques que les basses fréquences ce qui permet de compenser les niveaux faibles des sons. On utilise généralement un filtre passe-haut, dit de préaccentuation. Après la préaccentuation, le signal vocal qui est fortement non stationnaire est décomposé en trames de 20 ms avec un recouvrement de 10 ms pour générer un vecteur acoustique. Le découpage du signal en trames produit des discontinuités aux frontières de ces trames, en se manifestant par des lobes secondaires dans le spectre. Pour compenser ces effets, nous appliquons une fenêtre de Hamming à chacune de ces trames.
Pour extraire les paramètres pertinents du signal d'entrée, une analyse acoustique multi-variables est appliquée à chaque fenêtre à l'aide de la technique RASTA-PLP (**R**el**A**tive **S**pec**T**rAl-**P**erceptual **L**inear **P**rediction) [5]. Afin d'enrichir les vecteurs acoustiques, nous avons utilisé deux autres parametres : l'Energie et le **T**aux de **P**assage par **Z**éros (TPZ). Nous présentons aussi dans ce travail à titre de comparaison quelques résultats obtenus avec d'autres techniques d'analyse : les coefficients de la LPC et la PLP.

3.6. Architecture Modulaire

Le système de reconnaissance est un Réseau de Neurones Modulaire constitué de sous-réseaux de neurones ou modules (experts) de type MLP avec un

apprentissage par rétropropagation du gradient comme méthode d'apprentissage. À chacun de ces experts nous avons attribué des sous-tâches de reconnaissance des quatre phonèmes en question. Chaque expert est spécialisé dans la reconnaissance d'un seul phonème par rapport aux autres (Figure 1).

3.7. Phase d'apprentissage

Le nombre d'unités d'entrée est fixé à 9 (7 PLP, "1" énergie, "1" TPZ) [5], celui de la sortie à "1" pour chaque sous-réseaux. Seul le nombre d'unités en couches cachées est indéterminé. Après l'extraction des paramètres pertinents des signaux acoustiques, un apprentissage supervisé est effectué sur tout le corpus d'apprentissage pour déterminer le nombre optimal d'unités cachées. Le corpus est constitué de 484 phonèmes issus des phrases porteuses dont 290 ont été choisis pour l'apprentissage et 194 pour le test. Le processus d'apprentissage est arrêté dès que l'algorithme n'est plus en mesure d'augmenter le taux de reconnaissance pendant l'apprentissage.

3.8. Phase de reconnaissance

Lors de la phase de reconnaissance, les signaux acoustiques de test ont été traités de la même manière que lors de la phase d'apprentissage. Les vecteurs acoustiques obtenus sont injectés dans le système detest en faisant une discrimination entre les phonèmes à reconnaître. Le processus de reconnaissance s'arrête si le phonème est détecté, sinon le réseau expert adjacent est activé. Si la base de données de test ne porte pas de phonème à reconnaître, le processus s'arrête sans qu'il y ait discrimination.

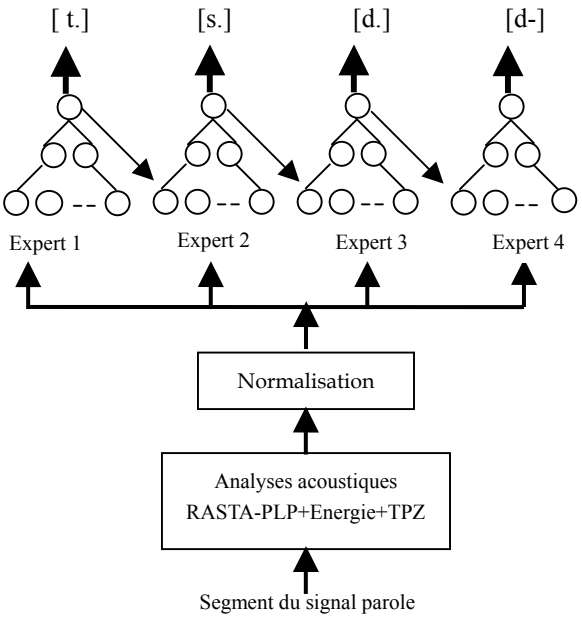

FIG. 1 : Système de Reconnaissance de phonèmes emphatiques de l'Arabe Standard.

4. RESULTATS ET COMMENTAIRES

La figure 2 présente le taux global de reconnaissance en fonction du nombre d'unités de la couche cachée. Dans la cadre de notre travail de recherche le taux de reconnaissance obtenu est de 71.29 % à partir de 7 unités de la couche cachée. Une architecture plus complexe (9 et 10 unités) fournit des résultats équivalents, cependant, il n'est pas nécessaire d'ajouter de la complexité dans l'apprentissage de notre système. Nous remarquons que ses performances se dégradent à partir de certains nombres d'unités cachées (à partir de 14 unités).

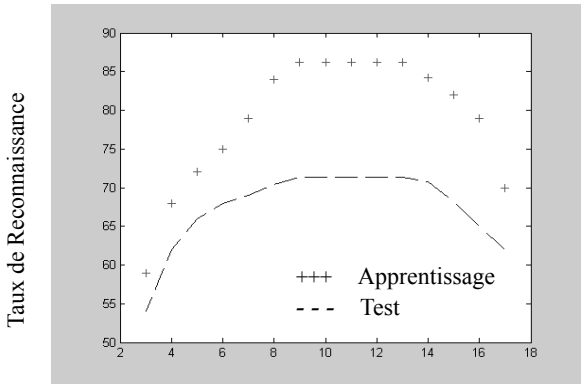

FIG. 2 : Influence de l'architecture sur les performances du système de Reconnaissance Automatique de la Parole.

Le tableau 3 montre le Taux de Reconnaissance (TR) de chaque phonème ainsi que le Nombre optimal d'Unités de la couche Cachée (NUC). Les deux emphatiques [d-] et [d.] ont des images acoustiques très proches ce qui les rend confondues dans les contextes.

TAB. 3 : Taux de Reconnaissance de chaque phonème emphatique.

Phonèmes de l'AS	Code API	NUC	TR (%)
ط	[t.]	7	75.66
ض	[d-]	7	66.66
ظ	[d.]	6	66.66
ص	[s.]	7	76.19

5. INFLUENCE DE TYPE D'ANALYSE

Afin de tester les performances de notre système de reconnaissance pour les différentes techniques d'analyse acoustique, les mêmes conditions d'expériences ont été utilisées pour les différents tests (corpus d'apprentissage, nombre d'itérations, etc.). L'apprentissage a été effectué en utilisant les

coefficients LPC, PLP, LPC, RASTA-PLP et RASTA-PLP combiné avec l'énergie et le TPZ.
Le tableau 4 présente l'influence de type d'analyse acoustique sur les performances de notre système de reconnaissance. L'analyse effectuée par les coefficients RASTA-PLP combiné avec l'énergie et le TPZ est celle qui donne de meilleurs résultats (Table 4).

TAB. 4 : Performance du système de reconnaissance en fonction du type d'analyse.

Type d'analyse acoustique	TR (%)
LPC	65.56
PLP	68.84
RASTA-PLP	70.04
RASTA-PLP + Eng + TPZ	71.29

Afin de lever toute ambiguïté sur la possibilité d'apprentissage par coeur, nous avons vérifié que les performances ne chutent pas au cours de cette phase. La Figure 3 récapitule les taux de reconnaissance sur le corpus de test au cours de l'apprentissage. Ainsi, nous remarquons que 280 itérations (cycles d'apprentissage) avec un pas d'apprentissage de 0.01 sont nécessaires et suffisantes pour une bonne convergence du réseau et qu'aucune baisse de performance n'apparaît au-delà.

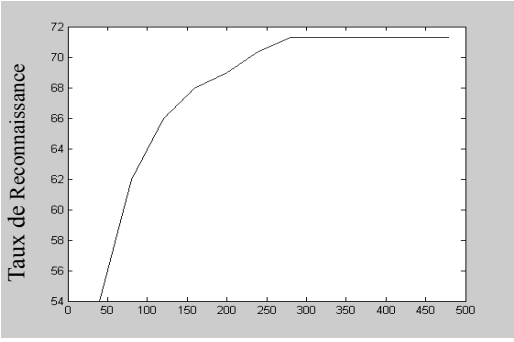

FIG. 3 : Evolution du taux de reconnaissance lors de l'apprentissage en fonction du nombre d'itérations.

6. CONCLUSION

Nous avons présenté dans ce travail un système de reconnaissance des phonèmes emphatiques de l'Arabe Standard en utilisant les Réseaux de Neurones du type MLP (Multi Layer Perceptrons). Les phonèmes à reconnaître sont pris dans des phrases porteuses situés dans différentes positions afin de prendre en considération les effets de la coarticulation.

Lors de la conception de notre système, nous avons utilisé une analyse acoustique multi-varaiables très robuste, représentative et discriminante, celle de l'analyse RASTA-PLP combinée avec l'énergie et le Taux de Passage par Zéros, afin d'avoir une bonne modélisation du signal vocal.
L'évaluation des résultats obtenus est motivante (71.29 %) car les phonèmes à reconnaître sont pris dans des contextes réels différents où l'effet de la coarticulation a été cerné. Un Taux de Reconnaissance intéressant a été réalisé pour les phonèmes [t.] et [s.] comparé aux autres emphatiques [d.] et [d-]. Ces derniers ont des images acoustiques très proches ce qui les rend confondues dans les contextes.
Dans ce travail, nous avons remarqué aussi que l'analyse acoustique qui utilise la PLP et RASTA-PLP donne de meilleurs résultats par rapport à la technique d'analyse classique LPC, car celles-ci sont basées sur des modèles auditifs, ce qui permet d'avoir une meilleure représentation des vecteurs acoustiques à reconnaître. Ainsi, l'introduction d'autres variables dans la phase d'analyse tels que l'énergie et le Taux de Passage par Zéros, comme paramètres pertinents, augmentent considérablement le Taux de Reconnaissance de notre système.

6. REFERENCES

[1] Calliope. La parole et son traitement automatique. Collection Technique et scientifique des télécommunication, CENT/ ENST, Ed. Masson, 1989.

[2] M. Kabache et M. Guerti. Application des Réseaux de Neurones à la reconnaissance des phonèmes spécifiques de l'Arabe Standard ", *Conférence Internationale de IEEE : Sciences Electronique, Technologie de l'information et des Télécommunications*, SETIT'2005, Sousse, Tunisie, 27-31 Mars 2005,.

[3] A. Betari. Caractérisation des phonèmes de l'Arabe Standard en vue d'une reconnaissance automatique de la parole. Thèse de Doctorat, Aix-en-provence (France), 1993.

[4] M. L. BenZaoui et M. Guerti. durées intrinsèques des sons spécifiques à l'Arabe Standard. *A.J.O.T., série B*, vol.14, N°1, 1999.

[5] H. Harmensky, N. Morganm A. Bayya et P. Kohnm. RASTA-PLP speech analysis technique . *In IEEE. ICASSP-92*, I.121–124, San Francisco (USA), CA, 1992.