ANALYSE ACOUSTIQUE MULTIVARIABLE APPLIQUÉE À LA RECONNAISSANCE DES CONSONNES EMPHATIQUES DE L'ARABE STANDARD

Mahraz Kabache et Mhania Guerti
Ecole Nationale Polytechnique et CRSTDLA
mhania.guerti@enp.edu.dz
mahraz_k@yahoo.fr

Résumé

Le but de ce travail est la reconnaissance des Consonnes Emphatiques (CE) de l'Arabe Standard (AS) en appliquant une analyse acoustique multivariable afin d'enrichir les techniques d'analyse usuelles en incluant deux autres paramètres (l'énergie et le Taux de Passage par Zéro : TPZ). Nous avons analysé le corpus d'apprentissage et le corpus de test par les techniques PLP (Perceptual Linear Prediction), RASTA-PLP (RelAtive SpecTrAl-Perceptual Linear Prediction), LPC (Linear Predictive Coding), l'énergie et le TPZ. Pour atteindre cet objectif, nous avons utilisé les Réseaux de Neurones Multicouches (Multi Layer Perceptrons : MLP) comme technique de reconnaissance de formes. Les phonèmes à reconnaître sont pris dans un corpus constitué de 84 phrases porteuses en AS, enregistrées par un seul locuteur, afin de prendre en considération le phénomène de la coarticulation (l'influence contextuelle d'un son sur un son contigu). L'analyse des résultats obtenus est encourageante car elle fournit un Taux de Reconnaissance (**TR**) de 71.29 % où les phonèmes à reconnaitre sont pris dans des contextes différents (l'effet de la coarticulation a été cerné). Il faut mentionner que le développement d'un système à base de réseaux neuronaux est une tâche délicate qui nécessite beaucoup d'expériences. En effet, de nombreuses difficultés existent concernant le choix et le dimensionnement du réseau, les paramètres à ajuster, le contrôle du système, etc.

Mots-clés

Reconnaissance Automatique de la Parole - Réseaux de Neurones Artificiels - consonnes emphatiques de l'arabe standard - techniques d'analyse de la parole.

الملخص

يهدف عملنا هذا إلى التعرف الآلي على الصوامت المفخمة الخاصة باللغة العربية الفصحى، وهذا بتطبيق تحليل بنوي متعدد المتغيرات. قمنا بتحليل كل من مدونة التمرن ومدونة التعرف باستعمال مجموعة من تقنيات التحليل الصوتي مثل التشفير التنبؤي الخطي (LPC)، والتشفير التبؤي السمعي (PLP)، (RASTA-PLP)، والطاقة ونسبة القيم المنعدمة (TPZ)، وهذا بهدف معرفة التقنية التي تقدم أكبر نسبة تعرف على الصوامت. قمنا باستعمال الشبكة العصبية متعددة الطبقات(Multi Layer Perceptrons : MLP) كتقنية للتعرف. أخذت الصوامت المراد التعرّف عليها ضمن عدد من الجمل العربية الفصحى، مسجلة من طرف متحدث واحد، وذلك للأخذ بعين الاعتبار ظاهرة التداخل بين الاصوات.

الكلمات المفاتيح

التعرف الآلي للكلام - الشبكات العصبية الاصطناعية - الصوامت المفخمة الخاصة باللغة العربية الفصحى - تقنيات التحليل الصوتي.

Abstract

The aim of our work is the recognition of the emphatic consonants of standard Arabic using the Artificial Neural Networks (ANN). To achieve this objective, we have used the Multi Layer Perceptron (MLP) as a technique of recognition. The phonemes to recognize are taken from sentences in standard Arabic recorded by a single speaker purposely, to take into consideration the coarticulation phenomena. We have analyzed the corpus of training and the corpus of test by several techniques of acoustical analysis: the PLP (Perceptual Linear Prediction), RASTA-PLP (RelAtive SpecTrAl-Perceptual Linear Prediction), LPC (Linear Predictive Coding), the energy and the Zero Crossing. The objective is to determine the acoustical analysis that gives the best recognition rate. The obtained results are satisfactory (71.29 % of correct identification rate), because the phonemes to be recognized are taken in different contexts where the effect of coarticulation is taken into consideration. It is important to mention that the development of a system based on ANN is a delicate task and requires a lot of experiences. Indeed, there are many difficulties related to the choice and the dimension of the network, the parameters to adjust, the control of the system, etc.

Keywords

Automatic Speech Recognition - Artificial Neural Networks - emphatic consonants of standard Arabic - speech analysis techniques.

Introduction

L'homme a la faculté de comprendre un message vocal provenant d'un locuteur quelconque, dans des environnements souvent perturbés par le bruit, quel que soit le mode d'élocution, la syntaxe et le lexique utilisés. La machine est-elle capable d'en faire autant ? Une solution peut-elle répondre en globalité à ce problème ? Le problème de la reconnaissance vocale est un sujet en cours de développement et actuellement, seules des solutions partielles existent.

Le problème de la Reconnaissance Automatique de la Parole (RAP) consiste à extraire automatiquement l'information contenue dans le signal de parole. Une bonne technologie de la RAP permettrait aux humains d'interagir de façon plus naturelle avec les machines. Ces dernières décennies, des progrès ont été réalisés dans ce domaine. De nos jours, des logiciels commercialisés sont capables d'effectuer une reconnaissance de la parole continue pour un vocabulaire important. Néanmoins, les performances de ces systèmes sont encore largement inférieures à celles des êtres humains [1].

Parmi les nombreux modèles proposés pour résoudre le problème de la reconnaissance vocale, nous trouvons les modèles neuromimétiques ou Réseaux Neuronaux (RN). Ces derniers ont été utilisés depuis longtemps pour résoudre des problèmes difficiles de classification et de reconnaissance des formes que l'on rencontre précisément en RAP [6], [9], [10].

Le travail que nous allons présenter concerne l'élaboration d'un système en vue de la RAP, appliquée aux Consonnes Emphatiques (CE) de l'Arabe Standard (AS). Nous avons développé ce système en utilisant les RN basés sur les perceptrons multicouches modulaires, MLP (Multi Layer Perceptrons) par une analyse acoustique multivariable.

Pour cela, nous avons élaboré un corpus d'apprentissage et un corpus de test constitués de phrases porteuses comprenant les quatre CE de l'AS. Ces dernières sont prises dans les différents contextes (Initial, Médian et Final) afin de prendre en considération le phénomène de la coarticulation (l'influence progressive ou régressive d'un son sur un son contigu).

1. Caractéristiques acoustiques des emphatiques de l'Arabe Standard

L'Arabe Standard 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 (*hurūf al-madd)* et l'absence de voyelle ou *sukūn*. La particularité de la langue arabe réside dans la présence des consonnes arrières glottales, pharyngales, vélaires, affriquées, 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 RAP nécessite une étude acoustique des différents phonèmes à reconnaître pour dégager leurs caractéristiques relatives afin de les utiliser lors du développement du système de reconnaissance. Les CE caractérisant l'AS prononcées par plusieurs locuteurs avec leurs durées moyennes sont présentées dans le tableau 1 en transcription phonétique [2], [3].

L'arabe standard est caractérisée par quatre phonèmes emphatiques, qui n'ont leurs équivalents exacts dans aucune autre langue européenne : [ṭ], [ḏ], [ḍ] et [ṣ]. Le phonème [ḍ] appartient à la classe des consonnes fricatives de l'Arabe Standard. Du 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 RAP, ils sont caractérisés par un TPZ élevé du signal temporel, ce qui permet de les identifier par rapport aux autres phonèmes non fricatifs.

Les phonèmes [ḍ] et [ṭ] sont des consonnes occlusives qui se produisent par la fermeture du conduit vocal (occlusion) pendant une brève durée suivie d'un brusque relâchement expirant l'air emmagasiné dans le conduit vocal. Les occlusives sont constituées par trois phases successives : la phase d'implosion de l'air, la phase de tenue de l'air emmagasiné et une dernière phase d'explosion, au moment du relâchement. Dans cette catégorie, les sons peuvent être voisés ou non voisés (sourds).

Le trait phonétique caractérisant les phonèmes [ṭ], [ḏ], [ḍ] et [ṣ] sur le plan articulatoire est appelé emphase. Ces phonèmes sont produits dans la partie antérieure de la cavité buccale par un report en arrière de la racine de la langue, un abaissement et un creusement du dos de la langue. Du 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) [4], [5].

CE en AS	CE en transcription phonétique	Durée moyenne en ms
[ص]	[ṣ]	129
[ظ]	[ḏ]	096
[ض]	[ḍ]	115
[ط]	[ṭ]	182

Tableau 1 : *Consonnes emphatiques de l'Arabe Standard*

2. Reconnaissance Automatique de la Parole (RAP)

Quelle que soit l'approche utilisée, un système de RAP est constitué d'un ensemble de modules (figure 1). Le module d'acquisition permet la mise en forme du signal vocal avant

tout traitement. Pour cela, des opérations de prétraitement sont effectuées dans cette étape.

L'extraction des paramètres est l'une des étapes les plus importantes dans le processus de la RAP. Celle-ci peut-être réalisée par plusieurs méthodes : temporelles comme le codage prédictif linéaire (Linear Predictive Coding : LPC) ou spectrales comme le codage MFCC (Mel Frequency Cepstral Coding), le codage PLP (Perceptual Linear Predictive Coding), etc.

Le module de reconnaissance des formes peut être réalisé par deux approches différentes : globale ou analytique.

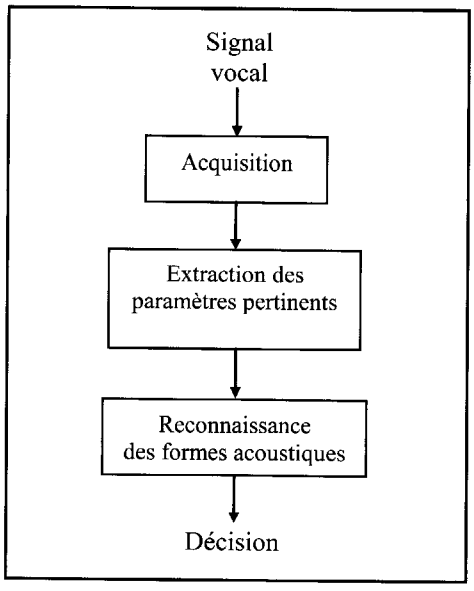

***Figure 1** : Description symbolique d'un système de RAP*

2.1. Reconnaissance par la méthode globale

Dans l'approche globale, l'unité la plus souvent utilisée se base sur le mot comme entité globale, c'est-à-dire non décomposée. L'idée de cette méthode est de donner au système une image acoustique de chaque mot qu'il doit identifier par la suite. Cette opération est faite lors de la phase d'apprentissage, où chaque mot est prononcé une ou plusieurs fois (Figure 2). Cette méthode a pour avantage d'éviter l'effet de la coarticulation ou l'influence d'un son sur un son contigu à l'intérieur des mots.
Cette méthode est utilisée dans les systèmes de reconnaissance suivants :

- de mots isolés ;
- d'unités enchaînées ;
- de parole dictée avec des pauses entre les mots.

La reconnaissance globale comprend les phases :

- d'apprentissage pendant laquelle un ou plusieurs locuteurs prononcent une ou plusieurs fois chacun des mots de l'application prévue. Ces prononciations sont

toutes prétraitées puis conservées telles quelles ou bien moyennées dans un dictionnaire de références en tant qu'images acoustiques ;

• de reconnaissance, où le signal à reconnaître subit le même prétraitement que la phase précédente. Il est ensuite comparé aux références contenues dans le dictionnaire. Le calcul d'une distance permet ou non de retenir la ou les références les plus proches.

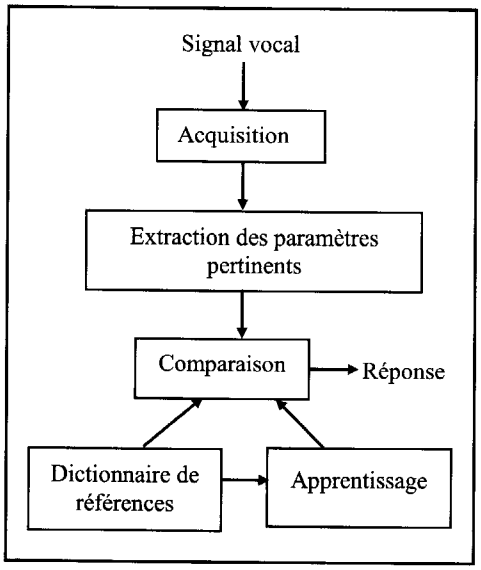

***Figure 2** : Schéma synoptique d'un système de RAP selon une approche globale*

2.2. Reconnaissance par la méthode analytique

La méthode analytique fait intervenir un modèle phonétique du langage. Plusieurs unités minimales pour la reconnaissance peuvent être choisies (phonèmes, syllabes, diphones, polysons, etc.). Parmi ces unités, le choix dépend des performances des méthodes de segmentation utilisées. Dans cette méthode, la reconnaissance passe par la segmentation du signal de parole en unités de décision puis par leur identification en utilisant des méthodes de reconnaissance des formes (classification statistique, RN, etc.). Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Dans ce type de système, les problèmes qui peuvent apparaître sont dûs en particulier aux erreurs de segmentation (insertions, substitutions, recouvrements) et d'étiquetage phonétique. C'est pourquoi le *Décodage Acoustico-Phonétique* ou DAP s'avère fondamental dans une telle approche.

Actuellement, les modèles de Markov Cachés, (HMM) pour *Hidden Markov Models,* sont les outils de modélisation les plus employés en RAP continue [1], [6]. Ils utilisent l'approche analytique comme méthode de reconnaissance (Figure 3).

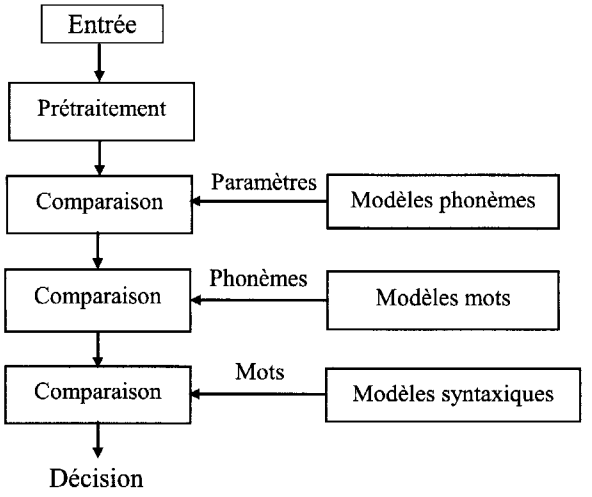

***Figure 3** : Schéma synoptique d'un système de RAP selon une approche analytique*

3. Réseaux de Neurones Artificiels (RNA)

Un RNA est une modélisation mathématique des neurones biologiques qui constituent le cerveau humain. Le premier modèle fut réalisé par Mc Culloch et Pitts en 1943. Le modèle biologique dont les modèles artificiels se sont inspirés est extrêmement simplifié par rapport à la réalité [7], [8]. Cette modélisation caractérise le comportement du cerveau par l'agrégation de cellules élémentaires, chacune effectuant une sommation pondérée des entrées dont le résultat est ensuite transformé par une fonction de transfert non linéaire. Il faut noter que cette fonction est indispensable à tout système de décision car elle permet de distinguer le neurone d'un simple système de classification linéaire.

Au cours du fonctionnement du neurone, nous pouvons distinguer deux phases. La première est habituellement le calcul de la somme pondérée *P* des entrées (Equation 1).

$$P = \Sigma (W_i * X_i) \quad (1)$$

Avec :

W_i : Poids synaptiques ;
X_i : Entrées ;
P : Somme pondérée.

Dans la deuxième phase, à partir de cette somme, une fonction de transfert calcule la valeur de l'état du neurone S (Equation 2).

$$S = F(P); \quad (2)$$

S : Sortie du neurone ;
F : Fonction de transfert.

4. Traitement de l'information par les RN

Dans un ordinateur, l'unité de mémoire passe par les phases d'écriture et de lecture. Durant la phase d'écriture, un mécanisme de stockage est utilisé pour spécifier

l'information à se rappeler. L'information stockée sera restituée durant la phase de lecture. Par analogie, deux phases existent aussi dans le traitement de l'information par les RN : la phase d'apprentissage et celle du test.

4.1. Phase d'apprentissage

L'apprentissage est vraisemblablement la propriété la plus intéressante des RN. C'est une phase du développement d'un RN durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. Au cours de la phase d'apprentissage, les poids synaptiques sont ajustés pour que le réseau remplisse une tâche définie par des exemples. L'apprentissage est défini comme tout changement dans les poids synaptiques.

Les règles d'apprentissage peuvent être divisées en deux catégories : *supervisées* et *non supervisées* :

- dans l'apprentissage supervisé, nous présentons aux RN les entrées et les sorties désirées correspondantes. Cet apprentissage se fait toujours par l'intermédiaire d'un critère à optimiser définissant la performance du réseau à chaque étape ;
- dans l'apprentissage non supervisé, seules les valeurs d'entrées sont disponibles. Les exemples présentés à l'entrée provoquent une auto-adaptation du réseau de façon à optimiser un critère de performance donné.

4.2. Phase de test

Au cours de la phase de test, nous présentons au réseau un ensemble d'exemples nouveaux (ensemble de test) mais proches des exemples appris et nous mesurons la qualité de ses réponses.

5. Reconnaissance de la parole par les RN

Dans notre travail, nous avons choisi, pour plusieurs raisons, un système de reconnaissance de la parole basé sur un réseau connexionniste de type MLP. Tout d'abord, ces réseaux ont de grandes capacités d'apprentissage à partir d'exemples de classification, leur robustesse aux données bruitées a montré leur adaptation en parole, notamment pour les mots isolés [4], [6], [8], [9], [10]. De plus, par rapport à l'ensemble des systèmes connexionnistes, ils ont l'avantage d'être basés sur des principes simples et relativement maîtrisables. Contrairement aux réseaux récurrents, leur temps de convergence peut être relativement court [10]. Il est également possible de détecter le moment où l'algorithme d'apprentissage n'est plus capable d'améliorer les performances, ce qui permet d'optimiser le temps de calcul.

L'utilisation de Réseaux Neuronaux est très répandue pour la classification de formes statiques (images, caractères écrits, etc.) et également pour la parole. Dans ce cas, l'unité à reconnaître (mot isolé ou unité sublexicale) est considérée comme une

forme acoustique globale présentée en entrée du modèle neuronal, le plus souvent un perceptron multicouche. De tels systèmes sont capables d'apprendre des fonctions de décision fortement non linéaires, ce qui est fondamental pour la reconnaissance de formes complexes telles que des mots ou des unités sublexicales.

Les performances obtenues par de tels systèmes pour de petits vocabulaires sont bonnes et sont comparables même favorablement à celles obtenues par des systèmes à base de HMM. En revanche, la méthode est difficilement adaptable à de grands vocabulaires et à la parole continue [6].

6. Description du système de reconnaissance

La structure du système que nous avons élaboré est spécialisée dans la reconnaissance des quatre phonèmes emphatiques de l'arabe standard. Afin de réaliser ce système, nous avons tenu compte des étapes suivantes :

6.1. Elaboration du corpus

La première étape à effectuer avant d'entamer les traitements est l'élaboration du corpus d'apprentissage et du corpus de test. Le choix de ce dernier afin de tester les performances de notre système de reconnaissance n'est pas arbitraire. En ce qui concerne notre travail, nous avons créé un corpus constitué de 84 phrases arbitraires porteuses contenant les phonèmes à reconnaître, pris dans les différents contextes. Nous justifions le choix de ce type de corpus par le fait qu'il est préférable de reconnaître les phonèmes dans des contextes pour prendre en considération les effets de la coarticulation (tableau 2).

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 وينخص	إن الوضوح نقيض الغموض
ضمى المريض ثم عافى	ألقى بنفسه فوق الفراش وبسط
عبر السنين تغير المرء وترص	صاغ الماء بسهولة كبيرة
قضم التفاحة	صبر فعل ماضي ثلاثي
ضربه طويلا	ظل الفاعل عن الطريق
فاظح هذا الأم	خرج للحديقة وبسط
قظم التفاحة	أخرجه من الكيس وأعط

***Tableau 2** : Echantillons extraits des phrases du corpus d'apprentissage et de test*

6.2. Acquisition des données

L'acquisition des données consiste à enregistrer les phrases du corpus choisi en utilisant un matériel spécifique au traitement du signal vocal (Sonagraphe Kay 5500). Les enregistrements ont été effectués par un seul locuteur, avec une fréquence d'échantillonnage de 11025 Hz. Les échantillons ont été codés sur 16 bits par échantillon.

6.3. 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 par écoute pour assurer que la segmentation a été correcte.

6.4. Analyse et traitement des données

Durant cette phase, le signal vocal (segments phonémiques) est préaccentué pour rehausser les hautes fréquences qui sont moins énergétiques que les basses fréquences. Cela permet de compenser le niveau le plus faible des sons. On utilise généralement un filtre passe - haut, dit de préaccentuation (équation 3).

$$H(z) = 1\text{-}az^{-1} \quad \text{Avec : } a = 0.95 \tag{3}$$

Après préaccentuation, le signal vocal qui est fortement non stationnaire est décomposé en une succession de tranches élémentaires supposées stationnaires. Ces tranches sont appelées fenêtres d'analyse ou trames. Typiquement, une analyse est appliquée toutes les 10 ms sur des fenêtres de 20 ms (par glissement et recouvrement de ces fenêtres) pour générer un vecteur acoustique. Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Pour atténuer ces effets de bord, nous appliquons une fenêtre de Hamming à chacune de ces tranches.

Après cette étape de mise en forme du signal d'entrée, une analyse acoustique multivariable est appliquée à chaque fenêtre à l'aide de la technique RASTA-PLP [11], [12], l'énergie et le TPZ, pour extraire les paramètres pertinents du signal vocal.

L'énergie et le TPZ sont des paramètres qui peuvent améliorer les performances des systèmes de reconnaissance. L'énergie correspond à la puissance du signal. Elle est évaluée souvent sur plusieurs trames successives pour pouvoir mettre en évidence la non stationnarité du signal vocal.

Le TPZ représente le nombre de fois où le signal passe par la valeur zéro. Il est fréquemment employé pour des algorithmes de détection de segment voisé/non voisé

dans un signal de parole. En effet, du fait de sa nature aléatoire, le bruit possède généralement un TPZ supérieur à celui des parties voisées.

L'introduction du paramètre énergie dans la phase d'analyse permet d'évaluer le degré d'accentuation des sons [13]. Le TPZ permet d'identifier les sons fricatifs des sons non-fricatifs [14].

6.5. Normalisation des entrées

Le même message prononcé deux fois par un même locuteur dans des conditions identiques produit deux formes spectrales différentes. Cette variabilité est dite *intra-locuteur*. La qualité de la voix, le débit de parole, le degré d'articulation sont tous des facteurs de variations acoustiques pour un signal donné. Ces variations entraînent des transformations non linéaires dans le temps du signal de parole. La non-linéarité vient du fait que les transformations affectent plus les parties stables du signal que les phases de transition [1].

Chaque segment de parole contient un nombre variable de trames, ce qui complique la gestion de la dynamique temporelle par les RN du type MLP. Pour lever cette difficulté, un alignement temporel est effectué après la phase d'analyse afin d'extraire les paramètres pertinents du signal et de garder une taille fixe pour le vecteur spectral, quelle que soit sa taille initiales. Pour cela, une procédure particulière est utilisée. qui consiste à segmenter les données sur les zones stables de chaque phonème puis à diviser chaque segment en trois intervalles sur lesquels nous effectuons une moyenne des vecteurs acoustiques. Le nombre de paramètres présentés à l'entrée de notre système est toujours fixe quelle que soit la longueur du segment [4].

7. Fonctionnement du système de reconnaissance

La structure du système que nous utilisons est basée sur la reconnaissance des phonèmes emphatiques de l'arabe standard. Le système est constitué de sous-réseaux ou modules de type MLP avec un apprentissage par rétro-propagation 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 (figure 4).

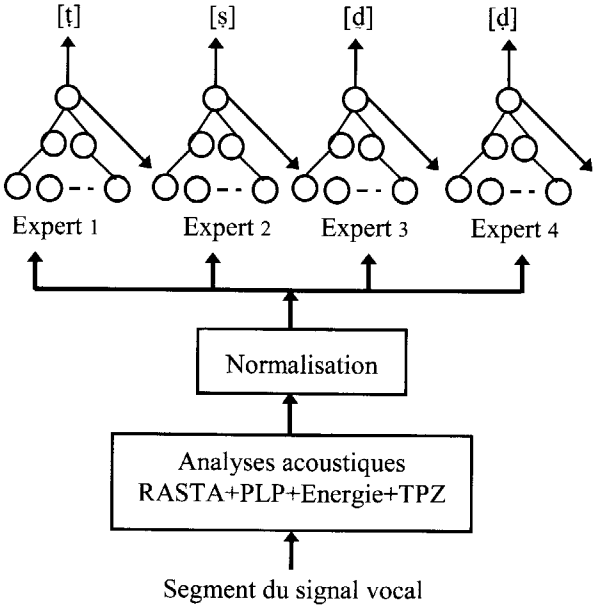

***Figure 4** : Système neuronal modulaire pour la reconnaissance des phonèmes emphatiques de l'AS*

7.1. Apprentissage

Le nombre d'unités d'entrée est fixé à 9 (7 PLP, 1 énergie, 1 TPZ), celui de la sortie à 1 pour chaque sous-réseau. Seul le nombre d'unités en couche cachée est indéterminé. Après l'extraction des paramètres pertinents des signaux acoustiques et l'initialisation des poids synaptiques par des valeurs comprises entre - 0.5 et 0.5, un apprentissage supervisé est effectué sur tout le corpus d'apprentissage avec un pas de 0.01 pour déterminer le nombre optimal d'unités cachées. Le corpus est constitué de 242 phonèmes dont 143 sont utilisés pour l'apprentissage et 99 pour le test. Le processus d'apprentissage est arrêté dès que l'algorithme d'apprentissage n'est plus en mesure d'augmenter le taux de reconnaissance.

7.2. Reconnaissance

Lors de la phase de reconnaissance, les signaux acoustiques sont traités de la même manière que lors de la phase d'apprentissage. Les vecteurs acoustiques obtenus sont injectés dans le système de test en faisant une discrimination entre les phonèmes à reconnaître. Le corpus de test est constitué de phonèmes à reconnaître en présence d'autres phonèmes pour mettre en jeu les possibilités de confusion entre les phonèmes (exemples : [ṭ] et [t], [ṣ] et [s]). 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 phonèmes à reconnaître, le processus s'arrête sans qu'il y ait discrimination.

8. Résultats et commentaires

La figure 4 donne le taux global de reconnaissance en fonction du nombre d'unités de la couche cachée. Le taux de reconnaissance atteint 71.29 % à partir de 7 unités. Une architecture plus complexe pour un intervalle d'unités compris entre 8 et 10 fournit des résultats équivalents mais il n'est pas nécessaire d'ajouter de la complexité dans l'apprentissage de notre système. Nous remarquons que les performances de notre système se dégradent à partir de 14 unités pour la couche cachée.

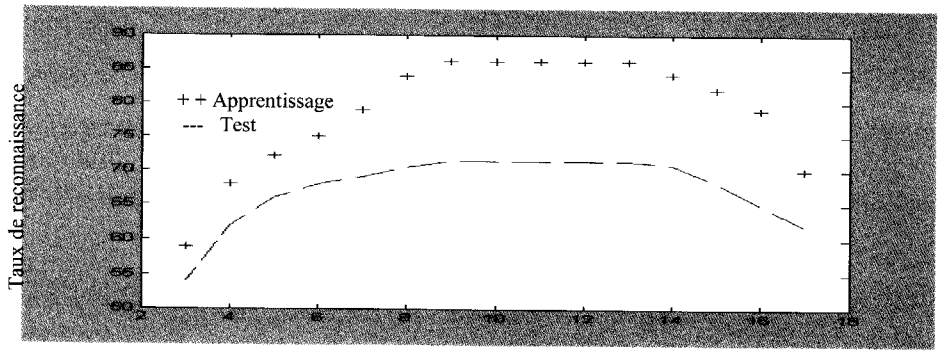

***Figure 5** : Influence de l'architecture sur les performances du système de RAP*

Comme le montre le tableau 3 qui illustre, pour chaque phonème, le TR ainsi que le Nombre optimal d'Unités de la Couche Cachée (NUC), les deux emphatiques [ṣ] et [ṭ] ont réalisé un TR intéressant comparé aux emphatiques [ḍ], [ḍ] qui présentent des problèmes avec les phonèmes qui les opposent dans le système phonétique de l'Arabe Standard (exemples [t] et [ṭ]).

Phonèmes de l'AS	Code Transcription phonétique	NUC	NTP	NPR	TR (%)
ط	[ṭ]	7	24	18	75.66
ض	[ḍ]	7	30	20	66.66
ظ	[ḍ]	6	24	16	66.66
ص	[ṣ]	7	21	16	76.19

NTP : Nombre Total de Phonèmes ; NPR : Nombre de Phonèmes Reconnus

***Tableau** 3 : TR de chaque phonème emphatique*

9. Influence du type d'analyse

Afin de tester les performances de notre système de reconnaissance pour les différentes techniques d'analyse acoustique, les mêmes conditions d'expérience 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 combinés avec l'énergie (Eng) et le TPZ, PASTA-PLP et RASTA-PLP combinés avec l'énergie et le TPZ. Le tableau 4 présente l'influence du type d'analyse acoustique sur les performances de notre système de reconnaissance. L'analyse acoustique effectuée par les coefficients RASTA-PLP combinés avec l'énergie et le TPZ est celle qui donne les meilleurs résultats.

Type d'analyse acoustique	Taux de Reconnaissance (%)
LPC	67.37
PLP	69.79
RASTA-PLP	70.58
RASTA-PLP + Eng + TPZ	71.29

***Tableau 4**: Performance du système de RAP en fonction du type d'analyse acoustique*

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 6 récapitule les TR sur un corpus de test au cours de l'apprentissage. Ainsi, nous remarquons que 280 itérations (cycles d'apprentissage) avec un pas de 0.01 sont nécessaires et suffisantes pour une bonne convergence du réseau et qu'aucune baisse de performance n'apparaît au-delà.

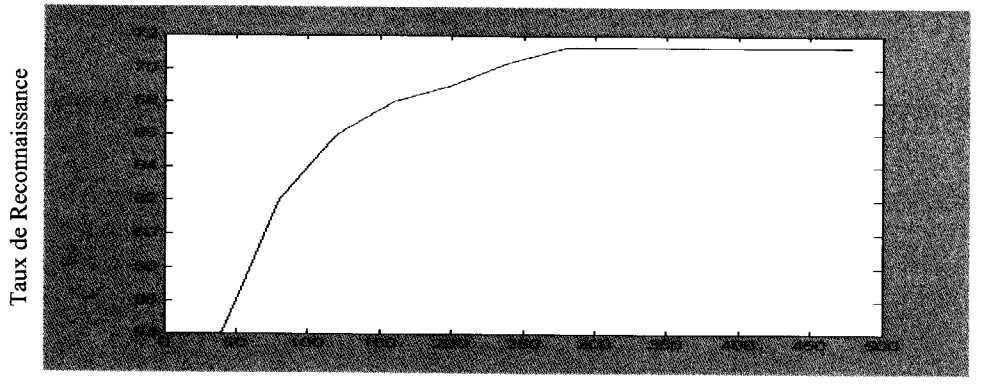

***Figure 6** : Evolution du TR lors de l'apprentissage en fonction du nombre d'itérations*

Conclusion

Dans ce travail, nous avons présenté un système de reconnaissance des phonèmes emphatiques de l'Arabe Standard en utilisant les réseaux de neurones du type MLP. Les phonèmes à reconnaître sont pris dans des phrases porteuses et situés dans différents contextes pour prendre en considération les effets de la coarticulation.

Lors de la conception de notre système, nous avons utilisé une analyse acoustique très robuste, représentative et discriminante, celle de l'analyse RASTA-PLP combinée avec l'énergie et le Taux de Passage par Zéro afin d'avoir une bonne modélisation du signal vocal. Cette dernière donne un meilleur TR par rapport aux autres techniques d'analyse.

L'évaluation des résultats obtenus est encourageante car les phonèmes à reconnaître sont pris dans des contextes différents où l'effet de la coarticulation a été cerné. Un taux de reconnaissance intéressant a été réalisé par les phonèmes [s̱] et [ṯ] comparativement aux emphatiques [d̲] et [ḍ]. Ces derniers ont des images acoustiques très proches ce qui provoque des confusions entre elles 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 celle-ci est basée 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 TPZ comme paramètres pertinents, augmente le taux de reconnaissance de notre système.

REFERENCES

[1] Calliope, 1989. La parole et son traitement automatique, Collection technique et scientifique des télécommunications, CNET/ ENST, Ed. Masson.

[2] Kabache, M. et M. Guerti, 2005. Application des réseaux de neurones à la reconnaissance des phonèmes spécifiques à l'arabe standard, Conférence internationale de IEEE : Sciences Electroniques, Technologies de l'Information et des Télécommunications, SETIT'2005, Sousse, Tunisie, p. 218, 27-31 Mars 2005.

[3] Benzaoui, M.L. et M. Guerti, 1999. Durées intrinsèques des sons spécifiques à l'arabe standard, AJOT, série B, vol.14, N°1, pp. 114-125.

[4] Selouani, S.A., 2000. Reconnaissance automatique de la parole par des techniques multi-agents, connexionnistes et hybrides : application à la langue arabe, Thèse de doctorat d'état USTHB, Alger, Algérie.

[5] Betari, A., 1993. Caractérisation des phonèmes de l'arabe standard en vue d'une reconnaissance automatique de la parole, Thèse de doctorat, Aix-En-Provence, France.

[6] Haton, J.P., 1995. Modèles neuronaux et hybrides en reconnaissance de la parole : état de recherches, fondement et perspectives en traitement automatique de la parole, Edition H. Meloni.

[7] Jodouin, J.F., 1994. Les réseaux de neurones : principe et définition, Edition Hermès.

[8] Hérault J. et C. Jutten, 1994. Réseaux neuronaux et traitement de signal, Edition Hermès.

[9] Botou, L., 1991. Une approche théorique de l'apprentissage connexionniste : application à la reconnaissance de la parole, Thèse de doctorat, Paris sud, France.

[10] Botou, L., 1988. Reconnaissance de la parole par réseaux multi-couches, Proceedings of the International Workshop on Neural Networks and Their Applications, pp. 197-217.

[11] Harmensky, H., 1990. Perceptual Linear Predictive Analyses of Speech, J. Acoust. Soc. Am. Vol. 87.

[12] Harmensky, H., 1997. Should Recognizer Have Ears ? Robust Speech Recognition for Unknown Communication Channels. Pont-à- Mousson, France.

[13] Yousfi, A. et A. Meziane, 2002. Introduction de l'énergie dans un modèle de reconnaissance automatique de la parole, XXIV^{ème} Journées d'études sur la parole, Nancy, France, pp, 317-320, 24-27 juin 2002.

[14] Aissiou, M. et M. Guerti, 2009. Genetic Supervised Classification of Standard Arabic Fricative Sounds, Int. J. Speech Technology, Vol.12. Issue 4, pp : 139- 147. Print ISSN : 1381-2416 Online ISSN : 1572-8110, décembre 2009.

DOI : 10 1007/10772-009-9061-5

http://www.citeulike.org/journal/springerlink-100275