RECONNAISSANCE DE TRAITS PHONETIQUES DE L'ARABE : COMPARAISON D'UN SYSTEME CONNEXIONNISTE MODULAIRE ET D'UN SYSTEME A BASE DE CONNAISSANCES

Sid-Ahmed SELOUANI[1] Jean CAELEN[2]

[1]LCP Institut d'Electronique, USTHB, BP 32 El Alia-Alger-Algérie
Tél: (+213) 2 515575 poste 813 - e-mail : parole@ist.cerist.dz
[2]CLIPS/IMAG, BP 53, 38041 Grenoble cedex 9 France
Tél: (+33) 4 76514627 - e-mail : jean.Caelen@imag.fr

Résumé

Nous décrivons dans ce papier une approche de reconnaissance automatique de traits phonétiques de l'arabe qui utilise des réseaux de neurones formels. L'analyse acoustique est effectuée par la technique de prédiction linéaire perceptive (PLP). Un corpus d'apprentissage et de test prononcé par 6 locuteurs algériens est établi afin d'évaluer les performances du système. Nous confrontons les taux d'identification des macro-classes aux résultats obtenus par le système basé sur des connaissances (SARPH) en nous focalisant sur les traits caractéristiques de l'arabe que sont l'emphase, la gémination et la durée. A l'issue, nous concluons sur les performances relatives des deux méthodes : les réseaux neuronaux restent supérieurs en identification pure tandis que le système à base de règles prend mieux en compte les problèmes liés à la durée phonologique.

mots clefs
Reconnaissance de la parole, réseaux de neurones, système à base de connaissances, langue arabe.

Abstract

This paper deals with a new indicative features recognition system for Arabic which uses a set of a simplified version of sub-neural networks (SNN). For the analysis of speech, the perceptual linear predictive (PLP) technique is used. The ability of the system has been tested in experiments using stimuli uttered by 6 Algerian native speakers. The identification results have been confronted to those obtained by the SARPH knowledge-based system. Our interest goes to the particularities of Arabic such as geminate and emphatic consonants and the duration. The results show that SNN achieved well in pure identification while in the case of phonologic duration the knowledge-based system performs better.

Keywords
Speech recognition, neural networks, knowledge-based system, Arabic language.

1. Introduction

Les caractéristiques essentielles des réseaux neuromimétiques sont en général, leur capacité d'apprentissage à partir d'exemples, leur adaptabilité, leur robustesse aux données bruitées ou manquantes et en reconnaissance de la parole, leur puissance de discrimination pour diviser l'espace de paramètres acoustiques en classes phonétiques. De nombreuses implémentations de ces réseaux ont été proposées dans la littérature[5][7][10][17]. La structure la plus répandue est celle du perceptron multi-couches (PMC). Ce type de réseau est capable de d'apprendre et de généraliser sur des relations complexes et non linéaires reliant l'espace des vecteurs acoustiques et les classes phonétiques que l'on désire reconnaître [17].

Dans cet article, il est question de la reconnaissance automatique de macro-classes phonétiques de l'arabe par des sous-réseaux multicouches. Nous focaliserons nos expérimentations sur des traits phonétiques spécifiques à la langue arabe. Ces Traits phonétiques sont : les voyelles longues et brèves, les fricatives emphatiques ainsi que les fricatives géminées.

Nous confrontons les résultats obtenus par cette approche purement automatique à celle faisant intervenir des connaissances phonétiques formalisées sous forme de règles gérées par le système SARPH (Système Arabe de Reconnaissance de Phones). Ce dernier est décrit en section 2. Dans la section 3, il est question des sous réseaux neuronaux et de leurs procédures d'apprentissage et d'identification. Enfin en section 4 les résultats obtenus par les deux approches sont discutés et commentés.

2. Le système SARPH (Système Arabe de reconnaissance de Phones)

SARPH [15] est un système analytique de reconnaissance (à base de règles) organisé autour d'un module de segmentation en phones homogènes et de réseaux d'états finis phonétiques pour l'identification des macro-classes de l'arabe. Celles-ci sont au nombre de 5 : les voyelles (V), les fricatives (S), les plosives (Q), les nasales (N) et les liquides (L). La structure de SARPH est analogue à celle adoptée dans le système DIRA [4] pour le français. Les sections suivantes présentent succinctement les différents modules de SARPH.

2.1 La paramétrisation

L'énergie ainsi que le taux de passage par zéro sont calculés sur des trames de 10 ms. La fréquence fondamentale est estimée et corrigée par la technique de l'ambiguïté modifiée [14]. L'extraction des formants est effectuée sur le spectre LPC. A partir du modèle d'oreille de Caelen [3], les énergies de sortie de 24 filtres couplés correspondant à une portion de la membrane basilaire sont calculées. Celles-ci permettent de déterminer les indices acoustiques statiques les plus pertinents [1]. Ils sont au nombre de 7 :aigu/grave, fermé/ouvert, bémolisé/diésé, écarté/compact, doux/strident, continu/discontinu, tendu/lâche.

2.2 La segmentation

Un codage delta des indices acoustiques est effectué afin de déceler leurs variations. Dans le but de quantifier la discontinuité entre deux trames successives, une fonction qui somme les sorties absolues des différents codeurs est estimée. Si cette somme est supérieure à un seuil variable dans le temps, une marque est attachée à la trame courante. Les trames comprises entre deux marques successives constituent un phone homogène. Pour chaque phone, les valeurs moyenne, maximale et minimale

de chaque paramètre sont calculées. Les indices acoustiques statiques et dynamiques sont codés (d'une manière non linéaire) en 5 niveaux : --, -, 0, +, ++. Une relation d'ordre existe entre ces 5 degrés de codage, ainsi, TL++ signifie très tendu et TL-- très lâche. D'autres paramètres acoustiques tels que l'indice de friction et l'indice vocalique, sont calculés. Ils constituent avec les autres paramètres, la base de données de SARPH.

2.3 L'identification

A chaque macro-classe est associé un réseau phonétique représentant la connaissance sur la macro-classe. Celui-ci est appliqué indépendamment sur la suite de phones. Un réseau est constitué d'un ensemble d'états et d'un ensemble de transitions (cf. figure 1). Les états représentent toutes les réalisations possibles des différentes phases acoustiques des macro-classes phonétiques. A chaque transition (ou arc) on associe une liste de contraintes (règles) à vérifier, une liste d'actions à effectuer en cas de succès et enfin un score à chaque passage par la transition. Un phone peut être étiqueté par un ou plusieurs réseaux comme il peut être rejeté par tous. Le nombre de phones n'est pas connu au préalable et l'étiquetage d'un phone de rang N ne se fait que si les N-1 phones le précédant ont été étiquetés.
L'accès se fait donc séquentiellement et justifie l'utilisation d'une liste chaînée pour la modélisation du réseau phonétique. Ainsi chaque maillon de la liste représente un phone et contient les informations relatives à celui-ci. La liste linéaire chaînée est bidirectionnelle afin de permettre le retour arrière lors de l'exploration du réseau en profondeur. La figure 1 donne un exemple de cheminement dans le réseau des voyelles de SARPH. Dans ce cas, si pour le phone courant, un 'établissement' est observé, on tentera d'associer au phone suivant la même phase ou bien les phases 'demie tenue' ou 'tenue orale'. On opère de la même manière pour les phones suivants en tenant compte des passages permis dans le réseau et ce jusqu'à en sortir en cas de solution. Plusieurs étiquetages sont possibles grâce au processus de retour en arrière. Seule la solution présentant le score le plus élevé sera validée.

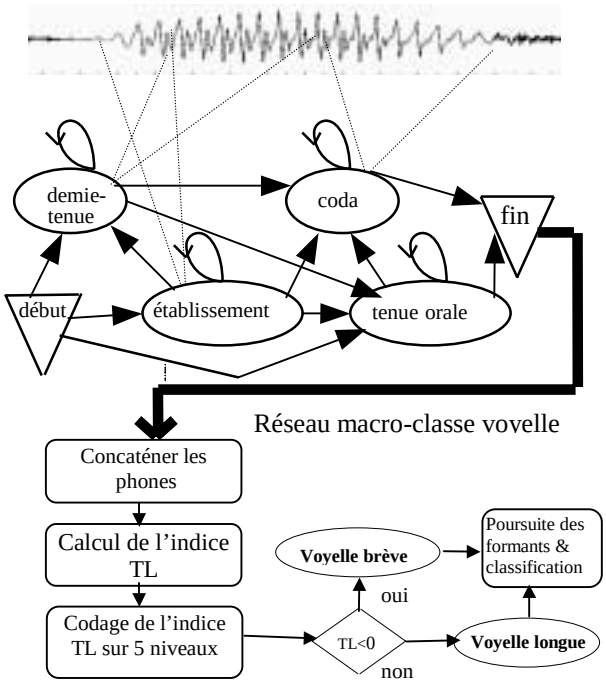


figure 1. Identification de la macro-classe «voyelle » par le réseau et discrimination voyelle longue/voyelle brève par l'indice TL (tendu-lâche)

3. Reconnaissance des macro-classes par réseaux de neurones

La structure générale du système que nous proposons, illustrée en figure 2, est constituée d'un ensemble de réseaux de neurones simplifiés auxquels nous avons attribué des sous-tâches de classification en vue de l'identification globale des macro-classes et de traits phonétiques de l'arabe.

3.1 Description générale du système

Chaque sous-réseau se spécialise dans la discrimination d'une classe (et une seule) par rapport aux autres. L'entraînement de chaque 'unité spécialisée' s'effectue sur

tout le corpus d'apprentissage. L'optimisation des paramètres tels que le nombre de cellules du réseau, la constante d'apprentissage, le nombre et la qualité des entrées ainsi que le nombre d'itérations, est opérée individuellement sur chaque sous-réseau. C'est au moyen d'une validation croisée que se réalise l'ajustement de tous ces paramètres en observant leurs taux de réussite pour différentes valeurs du paramètre à optimiser. Cette tâche est très aisée du fait de la simplicité du sous-réseau à entraîner [16].

3.1.1 L'analyse acoustique

Différentes analyses acoustiques ont été testées. Le but est de déterminer celle qui donnerait le meilleur compromis entre la rapidité d'apprentissage et la capacité de généralisation. Pour ce faire, un corpus de validation croisée a été établi. Il est constitué de 414 voyelles, 246 fricatives, 214 plosives, 106 nasales et 101 liquides. La validation des sous-réseaux a été effectuée en utilisant les coefficients LPCC (coefficients cepstraux de prédiction), les coefficients PLP (perceptuels) [11], l'énergie (En) et le taux de passage par zéro (TPZ) ainsi que leurs dérivées premières (dEn et dTPZ). L'idée de base de l'analyse PLP est d'effectuer un pré-traitement sur le signal vocal dans le but d'en extraire une représentation reproduisant les caractéristiques de la réponse du système auditif humain. Une analyse par prédiction linéaire classique (LPC) est ensuite appliquée sur le signal résultant.
La validation est effectuée sur le sous-réseau de classification des voyelles. La tâche assignée est la classification des voyelles brèves de l'arabe.
Les coefficients PLP combinés avec les dérivés de l'énergie et le TPZ sont ceux qui donnent les meilleurs résultats comme le montre le tableau 1.
Les mêmes conditions d'expériences (nombre d'itérations, constante d'apprentissage,...) ont été utilisées lors des différentes expérimentations. Les résultats de la validation montrent une nette supériorité de la modélisation auditive par rapport à la modélisation classique. Un gain en temps d'apprentissage (le nombre d'unités d'entrées est réduit) est également observé. La taille du réseau ayant également diminué, le nombre de poids et biais à mémoriser diminuera sensiblement.

Types d'entrées	unités d'entrées	validation voyelles
36 LPCC	36	89 %
15 PLP	15	87 %
36 LPCC+3TPZ+3En	42	90 %
15 PLP+3TPZ+3En	21	91 %
36 LPCC+3TPZ+3En+3dEn+3dTPZ	48	91 %
15 PLP+3TPZ+3En+3dEn+3dTPZ	27	92 %

tableau 1. Performance du sous-réseau des voyelles avec différents types d'analyses.

Après l'analyse acoustique, une gestion de l'aspect dynamique de la parole doit être effectuée car notre système opère une classification statique. La procédure de normalisation temporelle sera détaillée en §3.2.2

3.1.2 La segmentation

Lors de l'apprentissage, un flot de données segmentées en macro-classes est présenté à l'entrée des réseaux. L'apprentissage étant supervisé, la base de données est également étiquetée en macro-classes (V, S, Q, N, L). lors de la phase d'identification, nous nous assurons que les segments à classifier seraient les mêmes que ceux sur lesquels a opéré SARPH.

3.1.3 L'identification

Deux types de classification de séquences inconnues sont effectués. L'une grossière, a pour objectif la détection des macro-classes (V, S, Q, N, L), la seconde plus fine, tente de déceler le trait d'emphase et de gémination sur les plosives et fricatives. Un affinement de la détection des voyelles est également opéré par un réseau spécialisé. Après avoir subi une

normalisation temporelle les vecteurs acoustiques sont tout d'abord injectés dans le réseau V (noté 1 dans la figure 2) chargé de la discrimination grossière voyelle/consonne. Ils progressent ensuite dans les réseaux successifs S, Q, N et L (de gauche à droite et notés respectivement 2, 3, 4, 5 dans la figure 2). Selon l'activation des deux sorties d'un réseau donné, le processus s'arrête si la macro-classe est décelée sinon une activation du réseau adjacent est actionnée. Un échec du système global est comptabilisé si le dernier réseau est atteint sans qu'il y ait discrimination. Lorsque la macro-classe Q ou S est détectée les réseaux emphase (noté 2') et gémination (noté 3') sont activés. La disposition des sous-réseaux dans le système global sera discutée en § 3.3.

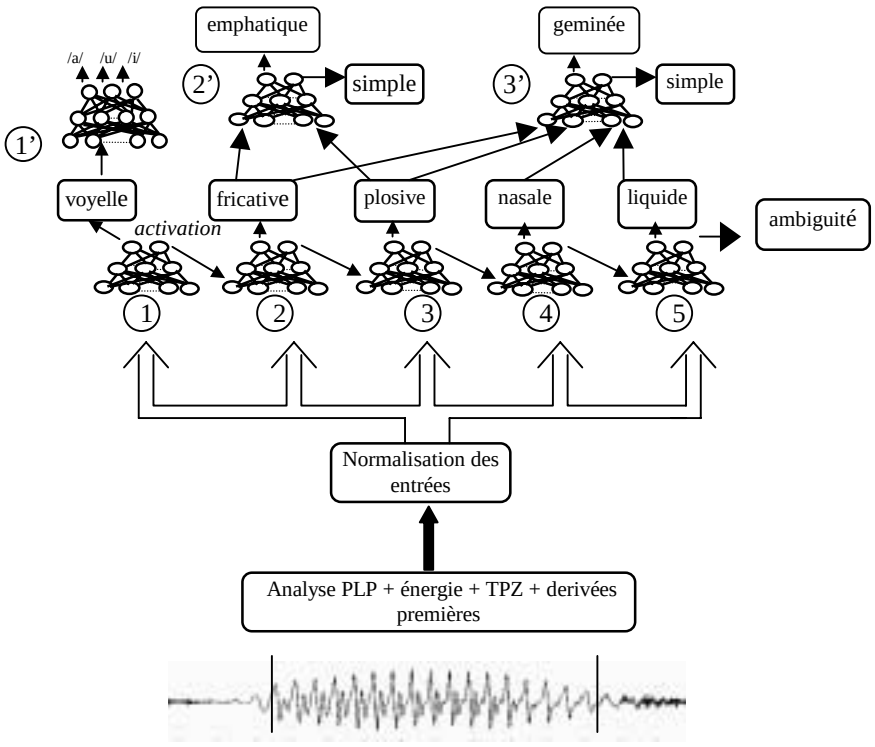


Figure 2. Structure du système neuronal de détection des macro-classes arabes.

3.2 La phase d'apprentissage

La structure des PMC utilisés comporte une seule couche cachée car il a été démontré que toute fonction continue de $\{-1, +1\}^n$ dans 3, pouvait être approchée uniformément par un réseau possédant une seule couche cachée d'unités sigmoïdes. Concernant la stratégie d'apprentissage, nous avons opté dans nos expériences pour l'utilisation d'un gradient quasi-stochastique [6]. En effet, nous avons choisi de modifier les poids et biais du réseau après un cycle d'apprentissage afin que l'ordre de présentation des exemples à l'entrée n'influe pas sur l'apprentissage. Par ailleurs, cette approche nous affranchit de

l'utilisation de facteurs tels que le moment d'inertie et la décroissance exponentielle.

3.2.1 L'initialisation

Les valeurs initiales des poids et biais sont déterminantes pour le temps et la qualité de l'apprentissage. Dans notre cas, c'est la procédure de NGUYEN-WIDROW qui a été utilisée. Celle-ci consiste à initialiser tout d'abord à des valeurs comprises entre -μ et +μ, les poids des unités cachées notés w'_{ij} , avec (i=1,.....,n) et (j=1,.....,q), n étant le nombre d'unités d'entrées, q le nombre d'unités cachées. Pour déterminer les poids et biais retenus pour l'initialisation de l'apprentissage, nous définissons un facteur d'échelle noté β, avec $\beta = 0.7\,q^{1/n}$, puis nous calculons le facteur de normalisation de la couche cachée, $\|w'_j\|$. Les poids retenus pour l'initialisation de l'apprentissage sont finalement donnés par l'expression suivante :

$$w_{ij} = \frac{\beta w'_{ij}}{\left\| w'_j \right\|}$$

Les biais sont initialisés à des valeurs aléatoires comprises entre -β et +β. Cette initialisation requiert l'utilisation de la fonction d'activation bipolaire. Nous avons constaté une amélioration du temps d'apprentissage d'un rapport de 20 par rapport à l'initialisation aléatoire.

3.2.2 Normalisation des entrées

La gestion de la dynamique temporelle par les réseaux de type PMC reste leur grand point faible. Ceux-ci ne sont pas capables de gérer les distorsions temporelles non apprises. Dans le cas de la parole, chaque segment contient un nombre variable de trames. Dans le cas d'une classification statique où l'architecture des réseaux est figée, cette difficulté doit être levée. Le système proposé ici, s'en affranchit après avoir dans un premier temps, divisé chaque segment en trois intervalles (établissement-stabilisation- fin) sur lesquels on effectue un moyennage des vecteurs acoustiques. Le premier et le dernier intervalle incluent l'information contextuelle (droite et gauche). Dans le cas ou la division n'est pas entière, l'intervalle médian est prolongé par autant de trames restantes (les phases stables sont avantagées). Par conséquent, le nombre de paramètres présentés à l'entrée est toujours fixe quelle que soit la longueur du segment. Il sera toujours égal à 3 fois la taille du vecteur acoustique de la trame. Une procédure particulière gère les cas extrêmes où un nombre inférieur à 3 trames par segment est rencontré[1].

Si m est le nombre de trames par segment et P la dimension du vecteur acoustique, alors :

$n_1 = n_3 = m/3$ et

$n_2 = m/3 + (m \equiv 3)$

n_1, n_2 et n_3 sont respectivement le nombre de trames sur le premier, deuxième et troisième intervalle sur lesquels s'effectue la moyenne des vecteurs de paramètres. Les entrées E_j présentées aux sous-réseaux sont données par l'expression ci-dessous :

Pour k={0,...,P-1},

$$E_j = \begin{cases} \dfrac{1}{n_1}\displaystyle\sum_{i=0}^{n1-1} c_{ijk} & pour \quad j = \{0,1,\ldots,P-1\} \\ \dfrac{1}{n_2}\displaystyle\sum_{n1}^{n1+n2-1} c_{ijk} & pour \quad j = \{P,\ldots,2P-1\} \\ \dfrac{1}{n_3}\displaystyle\sum_{n1+n2}^{m-1} c_{ijk} & pour \quad j = \{2P,..,3P-1\} \end{cases}$$

$\boldsymbol{C_{ijk}}$:k ième composante du vecteur de la trame i, participant au calcul de l'entrée j.
Le nombre d'unités d'entrées sera de 3P.

[1] Ce cas peut être évité en écourtant la durée de trame.

3.2.3 Paramètres des réseaux

Le même matériau (corpus d'apprentissage et de validation) est utilisé pour déterminer le nombre optimal d'unités cachées. Nous remarquons (courbe en figure 3) que les performances se dégradent à partir d'un certain seuil du nombre d'unités cachées. Par exemple, cette valeur est approximativement de 38 pour les réseaux des fricatives, lorsqu'on utilise en entrée, les coefficients LPCC, le TPZ, l'énergie et leurs dérivées.

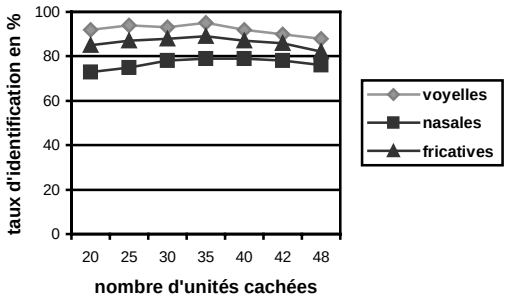


figure 3. Validation du nombre optimal d'unités cachées en utilisant 48 entrées (36LPCC + 3TPZ + 3En + 3dEn + 3dTPZ)

Le Tableau 2 donne les taux de succès en validation pour chaque sous-réseau avec une architecture optimale propre. Les entrées sont les coefficients PLP, le TPZ, l'énergie et leurs dérivées premières. Ces résultats permettent a posteriori d'hiérarchiser les sous-réseaux selon leurs performances.

Sous-réseau	Architecture	succès	Échec	Taux
Voyelle/consonne	27-15-2	634	12	98 %
Voyelles /a/,/u/,/i/	27-25-3	397	17	96 %
Fricatives	27-18-2	226	20	92 %
Plosives	27-20-2	175	39	82 %
Nasales	27-20-2	84	22	79 %
Liquides	27-15-2	78	23	77 %

tableau 2. Taux moyen d'identification des macro-classes en validation croisée (nombre d'unités d'entrée égal à 27 car P=9 en utilisant 5PLP + 1TPZ + 1En + 1dEn + 1dTPZ sur 3 intervalles) .

Un autre problème réside dans le choix de la constante d'apprentissage du gradient (notée généralement η). Une validation des sous-réseaux a été effectuée pour différentes valeurs de η avec le même nombre d'itérations. Le temps d'apprentissage dépend fortement du choix de cette constante. Pour les conditions d'initialisation énoncées précédemment, les résultats donnés en figure 4, montrent qu'une valeur de 0.4 pour η est un bon compromis entre le temps d'apprentissage et la généralisation.

Les différentes formes à apprendre sont présentées de façon alternée. L'approche que nous avons utilisée a pour avantage de faciliter l'apprentissage car la tâche de discrimination binaire ne nécessite pas un grand nombre de cycles (il est de 400) dans la majorité des expériences et chaque sous-réseau est 'initié' indépendamment des autres. Lors de la phase de reconnaissance, il n'est exigé de lui qu'une spécialisation dans le repérage d'une seule (et une seule) classe phonétique parmi les autres.

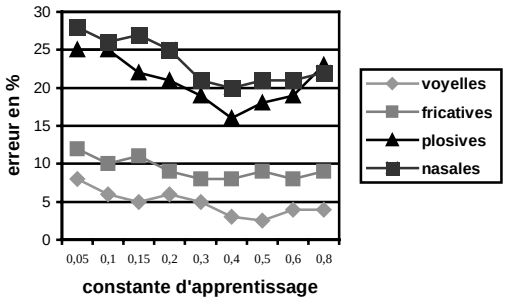


figure 4. Erreur globale en fonction de la constante d'apprentissage.

3.3 Classification par le système global

La méthode de validation croisée, utile pour affiner une architecture de réseau et déterminer les conditions optimales de fonctionnement, nous permet également de hiérarchiser les sous-réseaux selon leur compétence à déceler la macro-classe considérée. En effet, au vu des résultats de

validation obtenus par chaque sous-réseau sur une partie du corpus d'apprentissage (cf. tableau 2), il apparaît que certaines tâches sont plus aisément remplies que d'autres. Dans le système de test, les différents sous-réseaux sont sollicités par ordre décroissant de leur compétence.

Dans le système hiérarchisé que nous préconisons (cf. figure 2), une simple détection des voyelles est faite dans un premier temps. Cette classification est rendue plus aisée par le fait que la langue arabe est une langue essentiellement consonantique et ne contient que peu de voyelles (3 voyelles brèves[2] :/a/, /u/, /i/ et leurs correspondantes longues :/aa/, /uu/ et /ii/). Un deuxième argument plaide en faveur de cette discrimination et le fait que la langue arabe ne contienne pas de voyelles nasales. La confusion voyelle/consonne due à la nasalité n'a pas lieu d'être. L'architecture et la disposition de la macro-classe dans le système global dépendent des résultats de la validation croisée obtenus par chaque sous-réseau (cf. tableau 2). La structure hiérarchisée (pipe-line) de ce système peut sembler inadéquate dans la mesure où les réseaux situés en profondeur (les plus à droite dans la figure 2) sont pénalisés. Une architecture mettant au même niveau de compétence les réseaux paraît moins contraignante [17]. Cependant, dans le cas particulier de la langue arabe des arguments plaident au contraire pour la structure que nous avons adoptée. Ces arguments sont les suivants :

- les fricatives (au nombre de 14) représentent à elles seules 50 % du système consonantique arabe ;
- les plosives (au nombre de 8) se réalisent dans des lieux d'articulation dispersés (uvulaire, glottale, vélaire alvéolaires, et bilabiale). Il n'existe pas de /p/ ni de /g/ en arabe ;
- il n'existe pas de voyelles nasales. Si une nasalité est rencontrée, elle ne peut être que la réalisation d'une consonne nasale ;
- les classes liquides et nasales contiennent peu d'éléments (2 liquides, 2 nasales).

Ainsi, les tâches dévolues aux niveaux supérieurs (les plus à gauche sur la figure 2) sont caractérisées par 3 aspects importants qui justifient leur position dans le système de test et qui minimisent la pénalisation des niveaux suivants, ce sont :

- les fréquences d'apparition des macro-classes lors de l'élocution (à eux seuls les 2 premiers niveaux : voyelles et fricatives traitent environ 90 % des cas d'élocution) ;
- la simplicité de leur tache de discrimination car les cas traités sont dispersés dans le lieu d'articulation (pour les fricatives et les plosives) ;
- leur taux de réussite lors de la validation croisée.

Lorsqu'une identification arrive au dernier réseau (celui des liquides) sans possibilité de classification, une ambiguïté est décrétée. Dans nos expériences, ce dernier cas est considéré comme un échec.

Dans le cas de la discrimination voyelle brève/voyelle longue, nous pouvons préjuger des difficultés de notre système à la réaliser (préjugé qui va être d'ailleurs confirmé par l'expérience) (cf. §4). En effet, et c'est là un problème majeur des réseaux connexionnistes : l'intégration de la composante temporelle des événements acoustiques.

Des réseaux spécialisés dans la détection des traits phonétiques d'emphase et de gémination sont adjoints dans la perspective de mesurer l'aptitude des réseaux neuromimétiques à déceler ce type de traits phonétiques (très fins) spécifiques à la langue arabe. L'architecture de ces 2 sous-réseaux est analogue à celle des plosives (27-20-2). Leurs performances seront discutées en (§4).

[2]Notons qu'il s'agit ici de l'arabe standard. Pour l'arabe parlé dans les différentes régions du monde, les systèmes vocalique et consonantique peuvent être radicalement différents.

4. Résultats et commentaires

L'originalité de la phonétique arabe se fonde, pour une grande partie sur la pertinence de la durée dans le système vocalique et sur la présence de consonnes emphatiques. Une autre caractéristique déterminante est la gémination. Celle-ci joue un rôle fondamental dans le développement morphologique nominal et verbal. Ces aspects particuliers focaliseront notre intérêt dans la comparaison des deux systèmes d'identification des macro-classes.

Le corpus de test a été prononcé par 6 locuteurs (3 hommes et 3 femmes) algériens. Ces mêmes locuteurs ont participé à l'apprentissage et à la validation croisée. Les stimuli sont constitués de 40 occurrences VCV et de 20 phrases (Arabe standard) où les fréquences d'apparition des phonèmes sont respectées [13]. Le test concerne :

- les 14 fricatives (notées en API[3]) : /f/, /s/, /s̲/, /z/, /h/, /η/, /ʃ/, /θ/, /χ/, /δ/, /δ̲/, /γ/, /ε/, / /;
- les 8 plosives : /t/, /t̲/, /k/, /b/, /d/, /d̲/, /q/, /?/ ;
- les 2 liquides : /l/, /r/ ;
- les 2 nasales : /m/, /n/ ;
- les 3 voyelles brèves : /a/, /u/, /i/ ;
- les 3 voyelles longues : /aa/, /uu/, /ii/.

Au total, le test a porté sur 852 voyelles, 384 fricatives, 248 plosives, 164 nasales et 168 liquides.

Les semi-voyelles sont assimilées aux voyelles correspondantes.

Une suite supplémentaire de 108 occurrences VCV dont la consonne est une fricative géminée a été testée. Le choix de tenter la détection de la gémination indépendamment des autres consonnes est délibéré, dans la mesure où phonétiquement il n'existe pas de consonnes géminées (les utiliser dans le corpus général déséquilibrerait phonétiquement celui-ci). Le nombre de consonnes emphatiques (fricatives et plosives) testé est de 84.

L'évaluation comparative des deux systèmes se fera sur la base des particularités de la langue arabe à savoir la gémination, l'emphase, et le trait de durée des voyelles.

4.1 Evaluation de SARPH

Nous avons montré en [15] que SARPH permet de confirmer par l'expérience 3 faits concernant le système vocalique et consonantique de l'arabe, énoncé déjà théoriquement par Jakobson [12], El Ghazeli [9], Bonnot [2] et Boudraa & Selouani [1] :

- l'indice acoustique tendu/lâche, calculé puis codé par le système permet de faire la distinction voyelle brève/ voyelle longue ;
- le trait d'emphase est décelé au moyen d'une règle qui gère l'indice bémolisé/diésé ;
- une distinction entre une consonne géminée et son homologue simple est possible grâce à l'indice tendu/lâche.

Nous devons rappeler ici que SARPH a permis de valider expérimentalement (cf. figure 5) certains traits phonétiques importants de l'arabe. Cependant, il faut relever le fait que sa base de connaissance fasse appel très souvent à des seuils empiriques qui dépendent des conditions d'expériences. Les tâches de normalisation au locuteur, au rapport signal sur bruit, au débit, etc, sont très ardues et nécessitent une gestion draconienne d'un nombre important de paramètres qui compromettent l'utilisabilité du système.

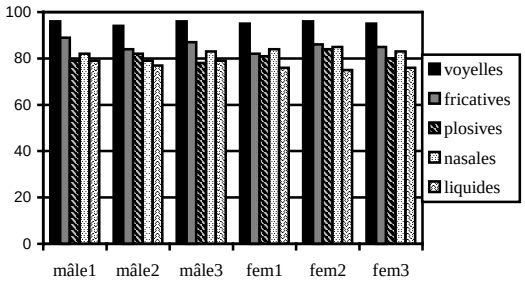


figure 5 scores moyens par locuteur obtenus par SARPH.

[3] Alphabet Phonétique International

On trouvera dans [14] une évaluation détaillée du système SARPH.

4.2 Evaluation des sous-réseaux

Pour les plosives et particulièrement / ?/ et /q/ des scores médiocres ont été réalisés. Pour les fricatives ce sont les fricatives arrières (/h/, /ħ/, /γ/, /ε/) qui posent le plus de problèmes. Leur brièveté et leur sensibilité au débit (effet de coarticulation) font qu'elles sont le plus souvent fondues dans le contexte vocalique. Les logatomes VCV constituent un matériau défavorable pour l'apprentissage de ce type de fricatives car le pourcentage d'omissions est très élevé. Il faut également, comme c'est le cas pour certaines plosives (glottales et vélaires), attacher la plus grande importance à la segmentation aux phases d'apprentissage et de test. La nasalité est détectée en moyenne dans 85 % des cas. Les cas de mauvaise détection sont souvent dus aux niveaux précédents. Les scores moyens obtenus individuellement sur chacun des six locuteurs pour les différentes macro-classes sont donnés en figure 6. Il est également possible d'affiner la détection de macro-classe en ajoutant pour les consonnes, un sous-réseau spécialisé dans la détection du voisement.

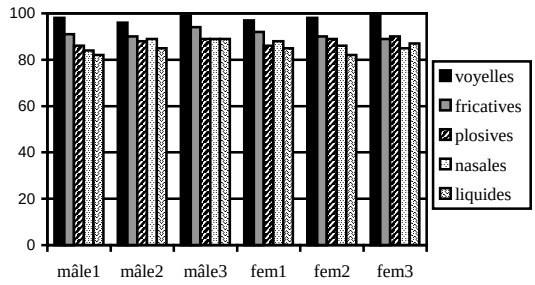


figure 6. Résultats de classification par locuteur des sous-réseaux

4.3 Evaluation des deux systèmes

Les résultats montrent (cf. figure 7) que pour l'identification de toutes les macro-classes, les sous-réseaux neuronaux surpassent le système basé sur les connaissances. Cette différence est notable pour le cas des nasales (différence de 5-10%). Dans le cas des voyelles, fricatives, plosives et liquides, la différence entre les scores est toujours en faveur des sous-réseaux avec respectivement 2%, 5%, 7% et 4%.

4.3.1 Détection de l'emphase

L'emphase est un trait phonétique caractérisant 4 consonnes, 2 plosives : /t̲/,/d̲/ et 2 fricatives : /δ̲/, /s̲/).
Ces consonnes sont articulées dans la partie antérieure de la cavité buccale, la racine de la langue est reportée en arrière contre la paroi pharyngale postérieure et un creusement de la langue est observé. Acoustiquement, elles se caractérisent par l'élévation de la transition de F1 et la baisse de la transition de F2 de la voyelle précédente et suivante.
Le taux de détection correcte est de 81 % pour SARPH et de 86 % pour les réseaux neuronaux. Notons la défaillance totale des deux systèmes dans l'identification de ce trait pour la consonne /d̲/. L'explication n'est pas dans une difficulté inhérente aux propriétés acoustiques de la consonne, mais plutôt dans la capacité des locuteurs à la prononcer correctement. En effet, dans un contexte VCV, il est très difficile de garder le caractère emphatique de /d̲/ et le plus souvent c'est son opposé par ce trait, /d/, qui est réalisée[4].

4.3.2 Détection de la gémination

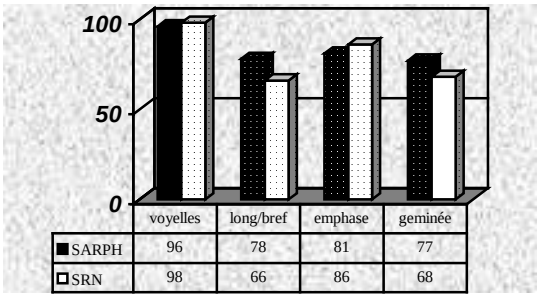
L'école traditionaliste des grammairiens arabes considère que le trait de gémination

[4] déformation qui est d'ailleurs caractéristique de l'accent régional algérois.

est un dédoublement de la consonne (prononcer une consonne d'une manière appuyée induit ce trait sur elle). Nous avons montré en [1], confirmant les thèses de Bonnot [2], que l'indice tendu/lâche peut déceler ce trait. Ceci d'ailleurs a permis à SARPH de déceler la gémination avec un taux de succès de 77%. Par contre, le système neuronal s'est avéré quelque peu défaillant avec un taux de 68 %. Nous pensons que le paramètre durée qui caractérise ce trait n'est pas intégré par ce type de système.

4.3.3 Distinction voyelle longue/brève

Dans SARPH, la distinction voyelle longue-brève est réalisée avec succès dans 78 % des cas.
Dans le cas des sous-réseaux, nous avons tenté d'effectuer cette discrimination en ajoutant au système de la figure 2, à la sortie du sous-réseau des voyelles, un réseau spécialisé dans cette classification. Moins de 68 % de taux de réussite a été atteint.



	voyelles	long/bref	emphase	geminée
SARPH	96	78	81	77
SRN	98	66	86	68

Figure 7. Taux d'identification des macro-classes de SARPH et des sous-réseaux neuronaux.

4.3.4 Problème de la durée

Le paramètre durée est très important dans la langue arabe. Il caractérise non seulement les voyelles, mais également les consonnes géminées. Cette caractéristique compense la pauvreté du système vocalique arabe. Tant au niveau grammatical qu'au niveau sémantique, ce paramètre est fondamental. Concernant ce trait, un double problème se pose en reconnaissance automatique de l'arabe : Il faut déceler les phonèmes allongés tout en s'assurant que ce prolongement est pertinent c'est à dire en le distinguant des allongements dus au débit d'élocution, à un accent particulier du locuteur etc.…Par exemple, les deux mots : /jamal/ (chameau) et /jamaal/ (beauté) ne diffèrent que par l'allongement de la voyelle finale. On exige du système de reconnaissance de déceler les 2 voyelles sans altérer la propriété temporelle. Un alignement temporel, au contraire, pénaliserait cette détection. Les réseaux d'états finis utilisés par SARPH (permettent de connaître la durée de stagnation dans un état) ainsi que les règles phonétiques concernant les voyelles longues et la gémination (indice tendu/lâche), sont à notre sens beaucoup mieux adaptés à cette tache que les réseaux neuronaux mêmes si ceux-ci intègrent la composante temporelle (TDNN, récurrents, etc,...). Ces derniers 'détruisent' le trait de longueur lorsqu'ils effectuent l'intégration temporelle.

5. Conclusions et perspectives

Nous avons présenté les résultats d'identification des macro-classes de l'arabe par deux systèmes ayant des stratégies complètement différentes : le premier basé sur des règles phonétiques et le second sur une structure hiérarchisée de sous-réseaux de neurones auxquels il a été alloué des tâches de discrimination binaire ($\in$ à la macro-classe ou $\notin$ à la macro-classe). Le souci majeur dans ce dernier système est la simplicité des réseaux, la facilité de leur apprentissage et la souplesse d'utilisation. La base de données est certes très réduite mais elle est suffisante pour le propos de cet article.
Ces expérimentations ont posé, dans le cas de la langue arabe, le problème de l'aptitude des systèmes (tout) automatiques de classification (aveugle) par rapport à la 'classification intelligente' supervisée par un expert humain, à déceler des traits aussi

subtils que la gémination, l'emphase et l'allongement pertinent des voyelles. Au regard des résultats obtenus, nous pouvons conclure que dans la détection de traits phonétiques (fins) tels que la durée phonologique (voyelles longues et gémination) en langue arabe, les systèmes experts restent plus performants. Par contre, lorsqu'une discrimination grossière est sollicitée (discrimination des macro-classes), les réseaux connexionnistes sont plus adaptés. Un compromis consiste peut être, à utiliser les méthodes connexionnistes en injectant à l'entrée non pas des données brutes mais en incluant des connaissances a priori sur les formes à classifier. Nos réseaux de PMC s'y prêtent, grâce à leur structure très souple qui permet d'ajouter en entrée des informations de nature différente. Les critiques que l'on peut porter sur notre système sont celles inhérentes aux méthodes connexionnistes en général. Celles-ci se heurtent au problème de passage d'un espace qui a subit des distorsions temporelles à un espace discret de symboles. Il s'agit d'un double problème de classification et de segmentation. Cette dernière conditionne d'une manière certaine les performances du système. Ceci a été vérifié pour le cas des consonnes glottales et vélaires Notons que les sous-réseaux mis en jeu dans notre système, sont optimisés séparément et la solution globale est sous-optimale. Cet état de fait est compensé par la simplicité de la tache exigée des sous-réseaux. L'approche proposée, affinée (en incluant des réseaux par genre de locuteur, un réseau de voisement, architecture parallèle, etc...) peut servir comme système d'appoint à d'autres techniques complémentaires, performantes dans la normalisation temporelle telles que les HMM.

Références

[1] **Boudraa B.,Selouani S.A,** 'matrices phonétiques et matrices phonologiques arabes" XXèmes JEP, Tregastel (1994).

[2] **Bonnot J.F** 'étude expérimentale de certains aspects de la gémination et de l'emphase en arabe', travaux de l'institut phonétique de strasbourg, N°11, pp 109-118, (1979).

[3] **Caelen J.,** "un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique", thèse de doc. d'état. ès Sciences, Toulouse (1979).

[4] **Caelen J., Tattegrain H**."Le décodeur acoustico-phonétique dans le projet DIRA", XIIèmes JEP, Nancy , pp 115-121, (1988).

[5] **Cook G.D, S.R. Waterhouse, A.J Robinson**, "Ensemble methods for connectionist acoustic modeling" , ESCA, Eurospeech97, Rhodes, Greece, pp 1559-1562., (1997).

[6] **Devilliers L** ., 'Reconnaissance de parole continue avec un système hybride neuronal et markovien', thèse de doctorat, Paris XI Orsay, (1992).

[7] **Dugast Ch, Devilliers L.** , 'Incorporating acoustic phonetic knowledge in hybrid TDNN/HMM frameworks', ICASSP, vol. I-421 San Francisco, (1992).

[8] **Gallinari P., Thiria S., Badran F**., 'On the relations between discriminant analysis and Multi-Layer Perceptrons'. Neural Networks, Vol 4, N°3, pp 349-360, (1991).

[9] **Ghazeli S.** "Du statut des voyelles en arabe ", analyses-théories, études arabes, N°2-3, pp 199-219, (1979).

[10] **Haton J.P,** 'Modèles neuronaux et hybrides en reconnaissance de la parole : état des recherches', in fondements et perspectives en TAP, H. Méloni éd., 1995, PP 139-154.

[11] **Hermansky H**., 'perceptual linear predictive (PLP) analysis of speech', JASA journal, 87 (4), pp 1738-1752, (1990).

[12] **Jakobson R., Fant G.M., Halle M.,** "preliminaries to speech analysis: The distinctive features and their correlates", MIT press, (1963).

[13] **Mrayati M.,**"Statistical studies of arabic roots", Applied arabic linguistics and signal and information processing, Hamshire publishing.
(1987)

[14] **Selouani S.A, Caelen J.,**' Experiments on arabic phone recognition using automatically derived indicative features', IVth ISSPA, Gold coast, Australia (1996).

[15] **Selouani S.A, Caelen J**.,'validation de traits phonétiques par un système de reconnaissance de l'arabe standard', Proc des XXI JEP, Avignon, pp 347-350, (1996).

[16] **Takuya K., Shuji T**.,'simplified sub-neural-networks for accurate phoneme recognition', ICSLP,Yokohama, Japan, pp 1571-1574, (1994)

[17] **Watrous R.L, Shastri L.,** 'learning phonetic features using connexionist networks : an experiment in speech recognition', Vol 4, IEEE, San diégo, California, june 21-24 pp 381-388, (1987).