Reconnaissance des sons spécifiques de l'Arabe Standard par Réseaux de Neurones

Kamel FERRAT[1, 2],
Khaled BAAZI[1,2],
Mhania GUERTI[1],

[1]Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA),
Alger - Algérie
[2]Ecole Nationale Polytechnique (ENP)
El-Harrach Alger- Algérie

Résumé

Par Reconnaissance Automatique de la Parole (RAP), nous entendons la transformation automatique de la parole vers du texte écrit. Ce passage automatique de la parole vers du texte écrit doit nécessairement passer par plusieurs étapes dont la plus importante est la décomposition de la parole prononcée en un ensemble de composantes acoustiques. Une unité très petite tel que le phonème nécessite 39coefficients MFCC (Mel Frequency Cepstral Coefficients) pour représenter fidèlement la parole prononcée. Passer à la parole continue contenant des centaines de phonèmes nécessite encore une énorme quantité de composantes acoustiques à traiter. Pour remédier à ce problème, nous faisons appel à la méthode dite du Réseau de Neurones Artificiels (RNA), qui a pour but la simulation informatique du Réseau de Neurones Biologiques (RNB). En effet, le cerveau humain est constitué de milliards de neurones biologiques interconnectés les uns aux autres et permettant l'échange d'informations en des temps très courts (environ 10^{17} opérations par seconde). A travers donc le RNA, nous essayons d'exploiter cette qualité du RNB qui permet de traiter un flux très important d'informations en des temps très courts, pour réaliser un système de reconnaissance automatique de la parole spécifique à l'Arabe Standard.

Comme base de données sonores, nous avons exploité un corpus contenant 400 fichiers sonores pour l'apprentissage des phonèmes spécifiques et 320 fichiers sonores pour les tests de reconnaissance. Les résultats obtenus sont encourageants. Le taux de reconnaissance des phonèmes emphatiques est de 91.25%. Le taux global de reconnaissance des phonèmes spécifiques de l'Arabe Standard est de 89.37%.

La Reconnaissance Automatique de la Parole (RAP), doit nécessairement passer par des étapes importantes : l'extraction des paramètres acoustiques, une comparaison avec des modèles de référence préalablement enregistrés et enfin la prise de décision, c'est-à-dire la reconnaissance. En parallèle à ces étapes, un processus d'apprentissage permet d'augmenter considérablement le taux de reconnaissance.

Aujourd'hui, un état de l'art des différents travaux réalisés dans le domaine de la reconnaissance de la parole montre que de meilleures résultats sont obtenus à partir des modèles connexionnistes (réseaux de neurones) et probabilistes (modèles de Markov cachés), vu la qualité aléatoire de la parole et sa complexité. Cette méthode a donné des résultats appréciables en reconnaissance automatique de la parole en Anglais américain et en Français. Nous avons jugé utile de l'adapter pour le cas de la langue arabe. En effet, peu de travaux de recherche ont été consacrés pour le cas de cette langue (Abdelhamid2006, Azmi2008, Chouireb2008, Satori2007).

Nous avons appliqué les réseaux dynamiques TDNN (Time Delay Neural Networks) pour la reconnaissance automatique des sons spécifiques de l'Arabe. Cette méthode permet de bien classifier ces sons, car elle tient compte de l'aspect dynamique de la parole et par conséquent, des phénomènes de la coarticulation (influence d'un son sur un autre contigu), très pertinents lors d'un acte de parole.

Lors de la phase d'apprentissage, nous avons utilisé la technique de rétropropagation de l'erreur (backpropagation) basée sur l'algorithme de Levenberg-Marquardt qui minimise l'erreur quadratique d'apprentissage

Dans cette phase d'apprentissage, nous avons utilisé un ensemble de 400 fichiers sonores contenant les sons

spécifiques de l'Arabe. Ce corpus de sons a été extrait de la base de données KAPD (King Abdul aziz Arabic Phonetic Database).

Pour les tests de validation, nous avons enregistré un ensemble de 320 fichiers contenant les sons spécifiques dans les différents contextes [CV], au moyen des logiciels Praat et Matlab. Au préalable, une segmentation automatique est effectuée sur les sons enregistrés pour détecter les frontières des sons sur lesquelles, nous extrairons les coefficients MFCC (Mel Frequency Cepstral Coefficients), paramètres acoustiques d'entrée de notre système. Ces paramètres permettent de modéliser le signal vocal par des filtres conformes à notre système auditif.

1. Neurone biologique et neurone formel

Les réseaux de neurones biologiques, de par leur multiples interconnexions, leur mécanisme d'inhibition et d'activation, ont inspiré les réseaux de neurones artificiels et continuent d'influencer le développement de nouveaux modèles tels que la reconnaissance des formes (caractères, visages, images, parole,…).

1.1. Qu'est ce qu'un neurone biologique ?

Le cerveau humain est constitué de milliards de neurones que nous pouvons assimiler grossièrement à des sommateurs, chaque neurone pouvant recevoir les entrées de dizaines ou parfois de centaines de milliers d'autres neurones. On estime généralement que l'ensemble du cerveau humain contiendrait de l'ordre du million de milliard de synapses, ramifications de neurones permettant l'échange d'informations avec d'autres neurones adjacents (Figure1). Ce grand nombre de neurones et de connexions conduit à un enchevêtrement qui est, aujourd'hui encore, très difficile à appréhender.

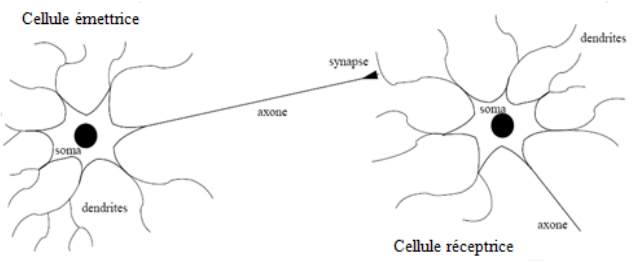


Figure 1. Schéma simplifié d'une connexion entre deux neurones biologiques.

La principale caractéristique de ces neurones est qu'ils permettent de véhiculer et de traiter des informations. Cette collecte de l'information est effectuée par les **dendrites** du neurone qui réceptionnent l'information des unités afférentes par l'intermédiaire des **connexions synaptiques**. Cette information est acheminée vers le noyau, également appelé soma. Cette information, une fois traitée, est répercutée en sortie de la cellule vers l'**axone** qui propage cette information vers d'autres cellules (figure 2).

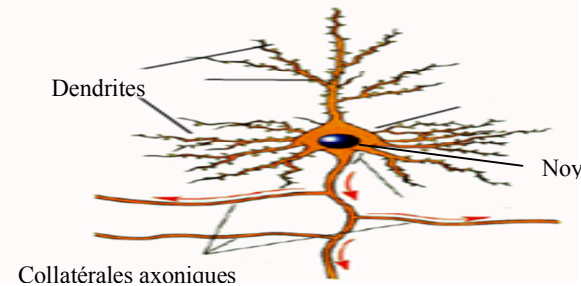


Figure 2. Représentation d'un neurone biologique.

1.2. Qu'est ce qu'un neurone formel ?

Un neurone formel est une représentation mathématique et informatique du neurone biologique (Dreyfus 2004). En d'autres termes, c'est une modélisation mathématique qui reprend les principes du fonctionnement du neurone biologique, en particulier la sommation des entrées (Figure 3).

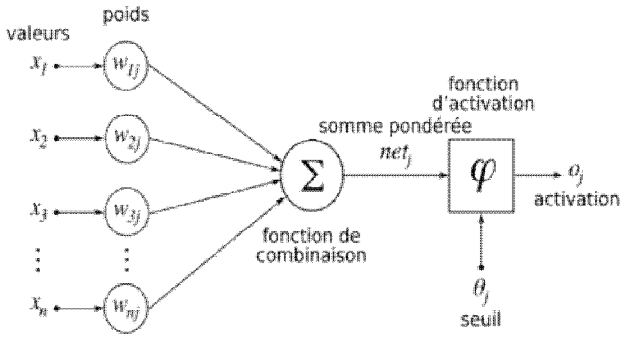


Figure 3. Représentation d'un neurone formel.

1.3. Domaines d'application des Réseaux de Neurones Artificiels (RNA)

Les RNA sont une voie prometteuse de l'Intelligence Artificielle, qui a des applications dans de nombreux domaines :

- Industrie : contrôle qualité, diagnostic de panne, analyse de signature ou d'écriture manuscrite...
- Finance: prévision et modélisation du marché (cours de monnaies…), attribution de crédits,...
- Télécommunications et informatique : analyse du signal, reconnaissance de formes (bruits, images, paroles, visages),...
- Environnement : évaluation des risques, prévisions et modélisation météorologiques, gestion des ressources...

2. Reconnaissance automatique de la parole (RAP) par Réseaux de Neurones (RNA)

Tout comme les autres systèmes de RAP, nous passons nécessairement par deux étapes importantes (Figure 4) :

- Une phase d'apprentissage permettant au système de lire les paramètres de référence, représentant les sons qui constituent le vocabulaire de l'application. Ces vecteurs de références sont obtenus à partir de modèles acoustiques qui permettent de caractériser les différents sons prononcés.

L'objectif de cette phase d'apprentissage est de permettre à un réseau de neurones "d'apprendre" à partir des exemples. Le principe est de fournir au réseau, une série d'exemples x et de résultats y. Il faudra ensuite trouver des coefficients spécifiques, appelés poids w, pour avoir un bon taux de reconnaissance et surtout une bonne généralisation. Ainsi, grâce aux exemples appris, le système est capable de traiter des exemples distincts, encore non rencontrés, mais similaires.

• Une phase de reconnaissance durant laquelle toute parole prononcée sera identifiée en comparaison avec les modèles de référence préalablement enregistrés.

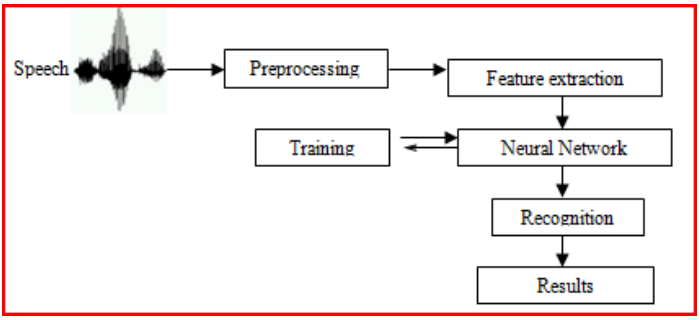


Figure 4. Structure d'un système standard de RAP, basé sur les RNA.

3. Reconnaissance automatique des phonèmes spécifiques de la langue arabe

L'Arabe Standard comprend34 phonèmes dont seulement 6 sont des voyelles. C'est une langue
consonantique contrairement à l'Anglais ou le Français qui présentent beaucoup plus de voyelles. Le système vocalique de l'Arabe Standard se compose de trois voyelles brèves [a,u,i], appelées «harakāte», et trois voyelles longues [ā,ū,ī], appelées «hurūf el-medd».

Les phonèmes spécifiques de l'Arabe Standard sont au nombre de huit (Figure 5, Table 1) :

- Quatre phonèmes occlusifs dont un est voisé et les trois autres sourds ;
- Quatre phonèmes fricatifs dont deux sont voisés et les deux autres sourds.

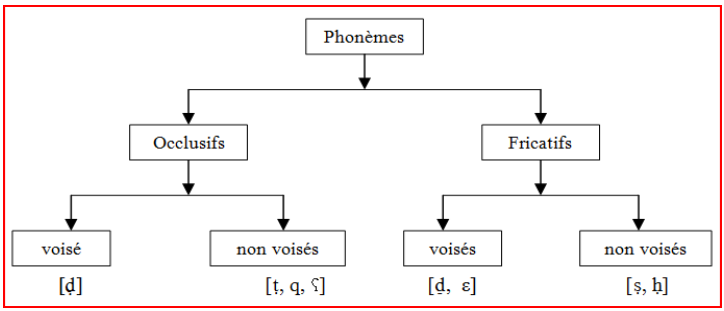


Figure 5. Classification des phonèmes spécifiques de l'Arabe Standard.

Phonème	Caractère arabe	Lieu d'articulation	Mode d'articulation			
			voisement	emphase	occlusive	fricative
[ḍ]	ض	alvéodentale	+	+	+	-
[ṭ]	ط	Apico-dentale	-	+	+	-
[q]	ق	Vélaire	-	-	+	-
[ʕ]	ء	Glottale	-	-	+	-
[ḏ]	ظ	Interdentale	+	+	-	+
[ɛ]	ع	Pharyngale	+	-	-	+
[ṣ]	ص	Alvéolaire	-	+	-	+
[ḥ]	ح	Pharyngale	-	-	-	+

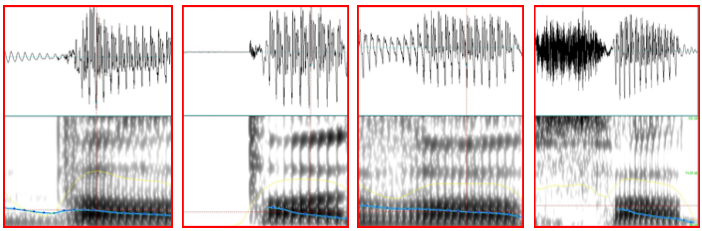
Table 1. Lieux et modes d'articulation des sons spécifiques de l'Arabe Standard.

3.1. Le phénomène d'emphase

Sur le plan articulatoire, le phénomène d'emphase consiste en un report en arrière de la racine de la langue et en un abaissement et creusement du dos de la langue, en ce sens qu'il y a élargissement de la cavité buccale et une constriction du pharynx (Ferrat 2005). Les phonèmes emphatiques de l'Arabe Standard sont respectivement :

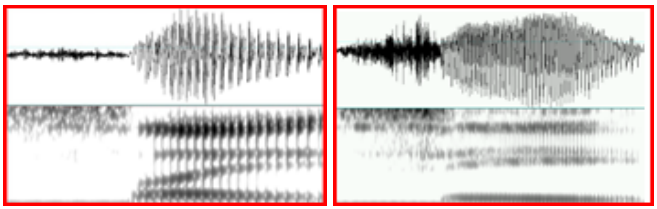
- l'occlusive alvéodentale voisée [ḍ] ;
- l'occlusive apicodentale [ṭ] ;
- la fricative interdentale [ḏ].
- la fricative alvéolaire [ṣ] ;

Sur le plan acoustique, nous remarquons une chute du formant acoustique F_2 due à l'élargissement de la cavité buccale et une montée du formant acoustique F_1 due au rétrécissement de la cavité pharyngale (Figures 6, 7 et 8).



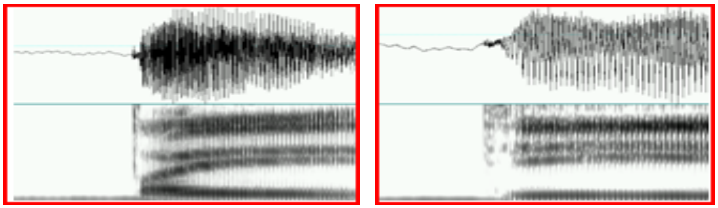
[ḍa [ṭa] [ḍa] [ṣa]

Figure 6. Chute de F_2 lors de la prononciation des phonèmes emphatiques, en contexte [C_ea].



[ṣi] [si]

Figure 7. Spectrogrammes de l'emphatique fricative [ṣ] par rapport à son opposée non emphatique [s], en contexte [C_ei].



[ṭi] [ti]

Figure 8. Spectrogrammes de l'emphatique occlusive [ṭ] par rapport à son opposée non emphatique [t], en contexte [C_ei].

3.2. Architecture de notre système de reconnaissance

Dans le cadre de notre travail, nous avons utilisé les réseaux de neurones à délais temporels TDNN. Cette architecture a été introduite pour la première fois par Alex Waibel pour la reconnaissance de la parole (Waibel 1989). Pour la phase d'apprentissage, nous avons utilisé la technique d'apprentissage supervisé TrainBr (Bayesian Regularization Backpropagation), exploitant l'algorithme de Levenberg-Marquardt. Les réseaux de neurones ainsi que la technique d'apprentissage sont implémentés avec Matlab's Neural Network Toolbox 7.5.

Les réseaux TDNN sont capables de traiter des séquences de vecteurs de parole grâce à l'introduction de délais temporels fixes sur les entrées. Ces délais visent à apprendre la structure temporelle des événements acoustiques et les relations entre ces événements (Ghosh 2004).

3.3. Base de données des fichiers sons

Nous avons exploité un corpus de 400 fichiers sons extraits de la base de données KAPD, conçue au laboratoire de phonétique de l'Université des Sciences et Technologies King

Abdul Aziz (Arabie Saoudite) (AlGhamdi 2003). KAPD contient plus de 46 000 fichiers de sons de l'Arabe dans les différents contextes, enregistrés par huit locuteurs. Pour la validation de nos résultats, nous avons enregistré un ensemble de 320 fichiers sons. Ces fichiers sont répartis avec un même nombre d'occurrences sur l'ensemble des phonèmes spécifiques. Les enregistrements ont été réalisés au laboratoire, en milieu naturel contenant un bruit environnant. Nous avons utilisé comme outil d'enregistrement le sonagraphe Kay CSL 4300B.

3.4. Extraction des paramètres acoustiques

L'extraction des paramètres acoustiques vise à obtenir la forme la plus représentative possible du signal afin de réduire au maximum le taux d'erreur de reconnaissance. Dans le cadre de notre travail, nous avons utilisé 39 paramètres MFCC, qui permettent de modéliser le signal parole par des filtres conformes à notre système auditif (Chetouani 2004). Pour l'extraction de ces vecteurs acoustiques, nous avons choisi une fenêtre glissante de Hamming de 30 ms, avec un pas de 10 ms. Ces vecteurs ont été ensuite normalisés sur un intervalle [-1, +1]. En effet, de meilleures performances de reconnaissance sont obtenues en choisissant la valeur moyenne des vecteurs d'entrée du système proche de 0, soit une distribution de moyenne 0 et de variance 1 (Povinelli 2004).

3.5. Phase d'apprentissage

Nous avons utilisé un apprentissage supervisé en adaptant le réseau tel que pour chaque exemple, la sortie du réseau corresponde à la sortie désirée. Ainsi, nous propageons un vecteur d'entrée, puis nous calculons l'erreur en sortie par rapport à un vecteur de sortie désirée, afin de corriger les poids en fonction de cette erreur (figure 9).

$$E = \sum_i (d_k - s_k)^2 \qquad (1)$$

Avec *dk* la sortie désirée pour le neurone d'indice *k* et *Sk* la sortie obtenue par le réseau.

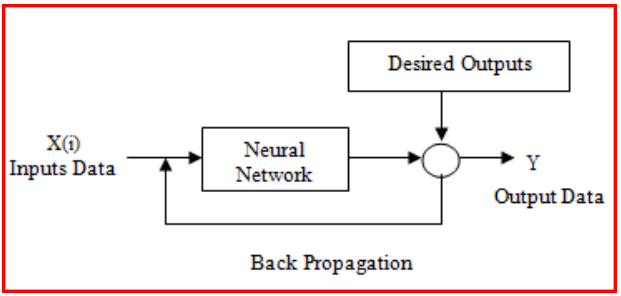


Figure 9. Rétropropagation de l'erreur avec apprentissage supervisé.

Pour la technique de rétropropagation de l'erreur, nous avons utilisé l'algorithme de Levenberg-Marquardt, qui minimise l'erreur quadratique d'apprentissage.
Pour la prise en compte des poids obtenus par apprentissage lorsque nous passons à la phase de reconnaissance, nous appliquons une DTW (Dynamic Time warping), qui permet de comparer la matrice des paramètres acoustiques du fichier test avec les matrices des paramètres acoustiques de l'ensemble des fichiers d'apprentissage.

- **Exemple d'apprentissage de l'emphatique [ṣ]**

Erreur avant apprentissage:

T = [1 0 0 0 0 0 0 0 0 0 0 0 0] (vecteur de référence)

Y_1 = 0.9527 1.6440 2.1229 -0.2251 0.0053 -0.4023 0.7578 0.6123 0.7356 0.4146 (vecteur de sortie obtenu)

Taux de Reconnaissance = 07.66 %

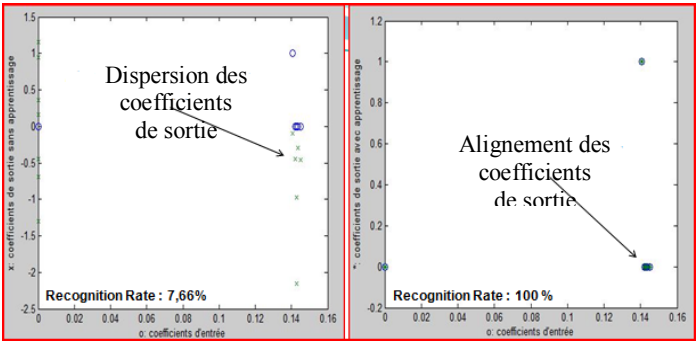
Erreur après apprentissage:

T = [1 0 0 0 0 0 0 0 0 0 0 0 0] (vecteur de référence)

Y_2 = 1.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 (vecteur de sortie obtenu)

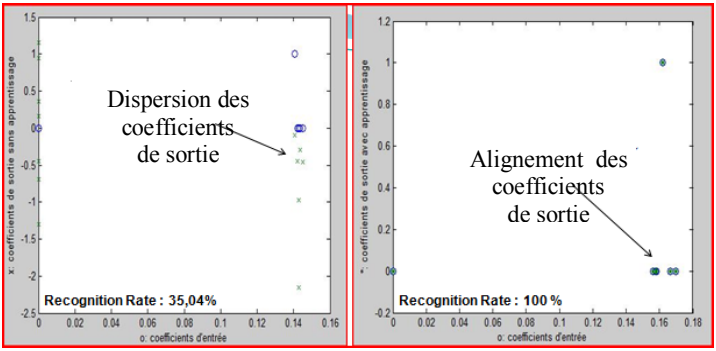
Erreur d'apprentissage = 2.3648e-013%

Taux de Reconnaissance = 100%



(a) (b)

Figure 10. Reconnaissance de la fricative emphatique [ṣ], avant apprentissage après apprentissage.



(a) (b)

Figure 11. Reconnaissance de l'occlusive emphatique [ṭ], avant apprentissage (a) et après apprentissage (b).

3.6. Généralisation pour le cas des phonèmes spécifiques prononcés en milieu bruité

Les tests de validation dans le milieu bruité (contenant un bruit d'environnement) permettent de mesurer les performances de notre système de reconnaissance. Pour cela, nous utilisons des fichiers sons qui ne sont pas connus de notre système et qui n'ont pas subit d'apprentissage. Nous avons enregistré 320 fichiers sons, répartis avec un même nombre d'occurrences sur l'ensemble des phonèmes spécifiques. Les enregistrements ont été réalisés au laboratoire, en milieu naturel contenant un bruit environnant. Nous avons utilisé comme outil d'enregistrement les logiciels Praat et Matlab.

Lors de la phase de reconnaissance, nous avons obtenu les résultats suivants (table 2):

Confusion (%)	[ṭ]	[ṣ]	[ḍ]	[ḍ]	[q]	[ɛ]	[ḥ]	[ʕ]	TR (%)
[ṭ]	**100.00**	00.00	00.00	00.00	00.00	00.00	00.00	00.00	100
[ṣ]	00.00	**100.00**	00.00	00.00	00.00	00.00	00.00	00.00	100
[ḍ]	00.00	00.00	**70.00**	**20.00**	00.00	00.00	**10.00**	00.00	70
[ḍ]	00.00	00.00	**05.00**	**95.00**	00.00	00.00	00.00	00.00	95
[q]	**10.00**	**05.00**	**10.00**	**05.00**	**55.00**	00.00	**15.00**	00.00	55
[ɛ]	00.00	00.00	00.00	00.00	00.00	**100.00**	00.00	00.00	100
[ḥ]	00.00	00.00	00.00	00.00	00.00	00.00	**100.00**	00.00	100
[ʕ]	00.00	00.00	00.00	**05.00**	00.00	00.00	00.00	**95.00**	95
								TGR	89.37

TR : Taux de Reconnaissance.
TGR : Taux Global de Reconnaissance.

Table 2. Matrice de confusion et taux de reconnaissance des phonèmes spécifiques en milieu bruité.

3.7. Interprétation des résultats

A partir des résultats obtenus, nous pouvons dire que :

- les phonèmes emphatiques [ṣ] et [ṭ] sont reconnus à 100 %. Les deux autres phonèmes emphatiques [ḍ] et [ḍ] sont reconnus avec des taux respectifs de 70 %, 95 %. Une confusion a été relevée entre [ḍ] et [ḍ]. Ceci est peut être dû au fait que ces deux phonèmes sont confondus lors de leur prononciation dans les pays du Maghreb. Cette confusion confirme les résultats de l'analyse acoustique que nous avons faite;
- le taux global de reconnaissance des phonèmes emphatiques est de 91.25 % ;
- le phonème [q] présente le plus faible taux de reconnaissance (55 %). Nous avons remarqué un taux de confusion de 15 % de ce phonème avec le [ḥ]. Il faudra noter que ces deux phonèmes présentent des caractéristiques

communes (non voisement et lieux d'articulation très proches) ;

- par contre, les deux pharyngales [ε] et [ḥ] présentent un taux de reconnaissance de 100%. En ajoutant le taux de reconnaissance de 95 % de la glottale [ʕ], nous déduisons que les phonèmes arrières de l'Arabe Standard s'adaptent bien à la méthode de reconnaissance choisie ;
- dans l'ensemble, un taux de reconnaissance appréciable de 89.37 % a été obtenu.

4. Conclusion

Dans ce travail, nous avons montré la contribution de la méthode des réseaux de neurones artificiels pour l'apprentissage et la reconnaissance automatique des phonèmes de l'Arabe. Cette méthode nous a permis d'avoir des taux de reconnaissance appréciables, en milieu bruité, des huit phonèmes spécifiques, avec notamment un taux d'identification de 100% des quatre phonèmes [ṣ], [ε], [ḥ], [ṭ] et de 95% pour [ḍ] et [ʕ]. Des confusions de reconnaissance persistent pour le cas du phonème [q] dont la prononciation présente beaucoup de caractéristiques communes avec les phonèmes emphatiques.

Ce système pourra être exploité dans plusieurs domaines futurs comme la traduction automatique et les didacticiels pour l'apprentissage de la langue arabe comme seconde langue.

Références

ABDELHAMID, S. et BOUGUECHAL, N., (2006). *SySRA, A System of a Continuous Speech Recognition in Arabic Language*. Proceedings of World Academy Of Science, Engineering and Technology, Volume, Vol. 11.

ALGHAMDI, M., (2003). *KACST Arabic Phonetic Database.* The Fifteenth International Congress of Phonetics Science, Barcelona, pp. 3109-3112.

AZMI, M.M., et Al., (2008). *Syllable-Based-Automatic Arabic Speeh Recognition*, Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA '08), University of Cambridge, UK, February 20-22, Vol. 4, N°1, pp. 246-250.

CHETOUANI, M, (2004). *Codage neuro-prédictif pour l'extraction des caractéristiques de signaux de parole*. Thèse de Doctorat Informatique, Université Pierre & Marie Curie, France.

CHOUIREB, F. et GUERTI, M. (2008). *Towards a high quality Arabic speech synthesis system based on neural networks and residual excited vocal tract model*, Revue Signal,image and video processing, vol. 2, n°1, pp. 73-87.

DREYFUS, G. et Al., (2004). *Réseaux de neurones- Méthodologie et Application-*. Editions Eyrolles, France.

FERRAT, K. (2005) *Acoustical study of the Tachdid and the Idgham in Standard Arabic*. *Application for speech synthesis*. International Conference SETIT2005, Susa (Tunisia), pp. 17-21.

GHOSH, J. et Al., (2004). *Automatique Speaker Recognition using Neural Network*. Spring 2004, in http://webspace.utexas.edu/lovebj/EE371D_TermProjectCode/

POVINELLI, R.,et Al., (2004).*Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces.* IEEE Transactions On Knowledge And Data Engineering, Vol.16, N° 6.

SATORI, H., HARTI, M., et CHENFOUR, N, (2007). *Arabic Speech Recognition System Based on CMUSphinx*. International Symposium on Computational Intelligence and Intelligent Informatics, ISCIII'07, pp.31-35, Agadir, Morocco, 28-30 March.

WAIBEL, A. et Al., (1989). *Phoneme recognition using time-delay networks.* IEEE Trans.Acoustics, Speech and Signal Processing, 37(3), pp. 328–339.