Apprentissage et reconnaissance automatique de la parole par réseaux de neurones artificiels

Kamel FERRAT[1] & Mhania GUERTI[2]
[1]CRSTDLA, [2]ENSP-El-Harrach

Introduction

Dans le cadre de notre travail, nous avons appliqué les réseaux de neurones pour la reconnaissance des phonèmes spécifiques à la langue Arabe Standard.
Par Reconnaissance Automatique de la Parole (RAP), nous entendons la transformation automatique de séquences de parole en textes écrits ou toute autre action à posteriori, notamment dans le cadre d'interfaces homme-machine. Ce passage de la parole vers du texte écrit 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 [1, 2, 3, 4].
Nous avons appliqué les réseaux dynamiques TDNN (Time Delay Neural Networks) pour la reconnaissance automatique de ces sons spécifiques. 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 160 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 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, leur manière d'évoluer et de s'adapter tout on long de la vie d'un organisme vivant, 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. La cadence maximale de traitement des informations est d'environ 10^{17} opérations par seconde. Ce qui est impressionnant comme capacité d'action.
Le cerveau se caractérise par une organisation très complexe à analyser du fait du grand nombre de cellules, les neurones, et de liens entre cellules, les connexions synaptiques, qui le compose (Figure1). Ce grand nombre de neurones et de connexions conduit à un enchevêtrement qui est, aujourd'hui encore, très difficile à appréhender.

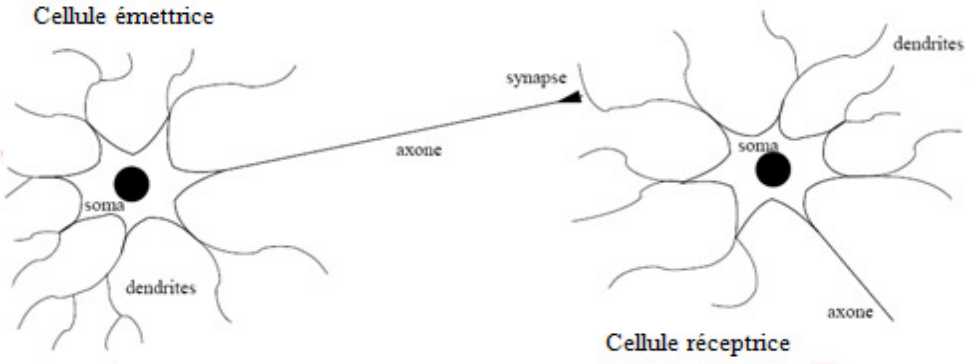


Figure 1. Schéma très simplifié d'une connexion entre deux neurones biologiques

La principale caractéristique de ces neurones est qu'ils permettent de véhiculer et de traiter des informations en faisant circuler des messages électriques dans le réseau formé par leur axone. La 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).

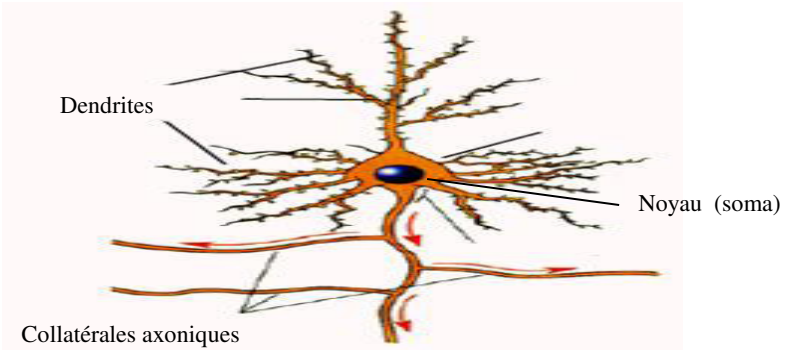


Figure 2. Représentation d'un neurone biologique

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

Un neurone formel est une représentation mathématique et informatique du neurone biologique. 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).

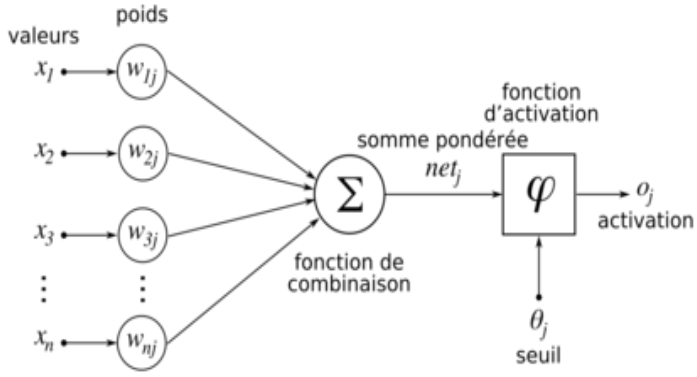


Figure 3. Représentation d'un neurone formel

Comme le cerveau humain, les réseaux de neurones artificiels (RNA) peuvent apprendre par expérience. Ainsi, suite à l'application séquentielle de plusieurs entrées à apprendre, les algorithmes d'apprentissage modifient la valeur des poids entre les neurones de façon à améliorer la performance du RNA (Figure 4). En fait, l'ajustement des poids est plus ou moins efficace tout dépendamment de la connaissance que nous avons du système à modéliser ou du problème à résoudre.

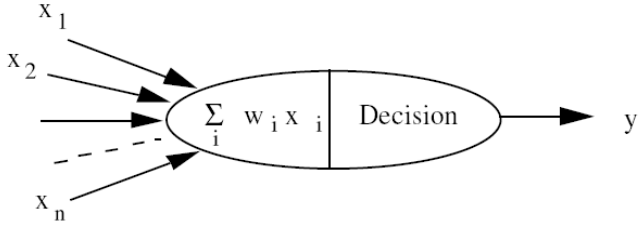


Figure 4. Expression mathématique d'un neurone formel

Ceci peut être exprimé par la fonction mathématique suivante :

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

Avec respectivement :

- y : Information de sortie obtenue ;
- f : fonction d'activation du neurone;
- w_i : poids du neurone ;
- x_i : vecteurs représentant l'information d'entrée.

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

Se trouvant à l'intersection de différents domaines (informatique, électronique, science cognitive, neurobiologie), l'étude des réseaux de neurones est une voie prometteuse de l'Intelligence Artificielle, qui a des applications dans de nombreux domaines :

- Industrie : contrôle qualité, diagnostic de panne, corrélations entre les données fournies par différents capteurs, analyse de signature ou d'écriture manuscrite...
- Finance: prévision et modélisation du marché (cours de monnaies…), attribution de crédits, sélection d'investissements,...
- Télécommunications et informatique : analyse du signal, élimination du bruit, reconnaissance de formes (bruits, images, paroles, visages), compression de données...
- Environnement : évaluation des risques, analyse chimique, prévisions et modélisation météorologiques, gestion des ressources...

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

L'idée principale des RNA est de s'inspirer de l'organisation du cerveau et des neurones biologiques humains et leurs interconnexions pour traiter l'information [5]. Tout comme les autres systèmes de RAP, nous passons nécessairement par deux étapes importantes (Figure 5) :

- Une phase d'apprentissage permettant au système de lire les paramètres de référence {R_1, R_2, …, R_n}, 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.
- Une phase de reconnaissance durant laquelle toute parole prononcée sera identifiée en comparaison avec les modèles de référence préalablement enregistrés. Le principe est de minimiser au maximum le taux d'erreur de reconnaissance qui pourra influer négativement sur la fiabilité du système.

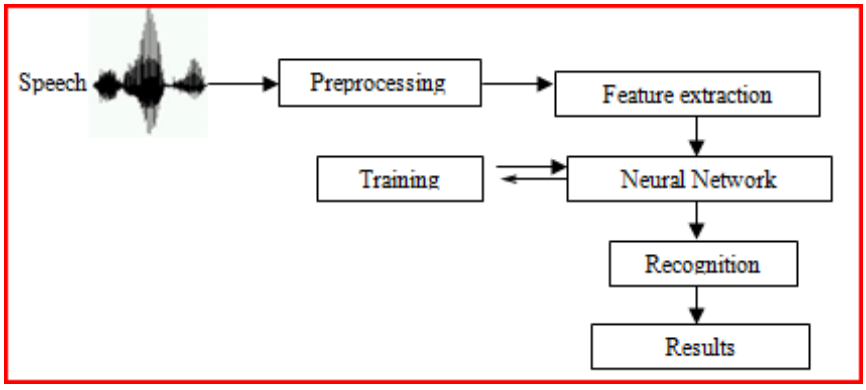


Figure 5. Structure d'un système standard de RAP, basé sur les RNA

2.1.Qu'est ce que l'apprentissage dans un système de RAP

L'objectif de la 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. Si cette phase est correctement réalisée, le réseau est capable de fournir des réponses en sortie très proches des valeurs d'origines du jeu de données d'apprentissage.
La notion d'apprentissage recouvre deux réalités souvent traitées de façon successive :
- la mémorisation, le fait d'assimiler sous une forme dense des exemples éventuellement nombreux ;
- la généralisation, le fait d'être capable, grâce aux exemples appris, de traiter des exemples distincts, encore non rencontrés, mais similaires.

2.2. L'apprentissage supervisé

Un apprentissage est dit *supervisé* lorsque l'on force le réseau à converger vers un état final précis, en même temps qu'on lui présente un motif. Lorsque le stimulus propagé atteint les cellules de sortie, il est comparé avec la réponse désirée en sortie, en calculant une valeur d'erreur qui est rétropropagée vers les couches inférieures, de façon à ajuster les poids affectés aux liens, et le seuil d'activité dans chaque cellule. Cette rétropropagation de l'erreur (error backpropagation) sera itérée jusqu'à ce que les paramètres du modèle atteignent une stabilité avec une erreur de reconnaissance minimale.

3. Reconnaissance automatique des phonèmes spécifiques de la langue arabe

Pour réaliser notre système de reconnaissance, nous avons utilisé une base de données de sons que nous avons exploitée dans les phases d'apprentissage et de reconnaissance. Pour l'extraction des vecteurs acoustiques, nous avons utilisé 39 coefficients dits MFCC.

3.1. Les sons spécifiques de l'Arabe

L'Arabe Standard comprend 34 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 ». Cette opposition temporelle brève/longue est fondamentale aux niveaux grammatical et sémantique. En effet, les deux mots [sabaqa] (devancer) et [sābaqa] (concourir) présentent deux sens différents même s'ils ne diffèrent que par la durée temporelle de la première voyelle. Les phonèmes spécifiques de l'AS sont au nombre de huit (Figure 6) :
- Quatre phonèmes occlusifs dont un est voisé et les trois autres sourds ;
- Quatre phonèmes fricatifs dont deux sont voisés et les deux autres sourds.

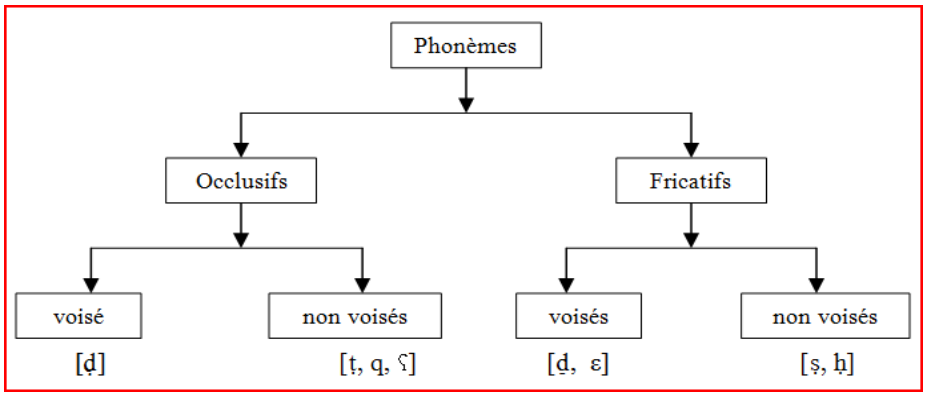


Figure 6. Classification des phonèmes spécifiques de l'Arabe Standard

La langue arabe présente quelques spécificités telles que la présence de phonèmes arrières glottales, pharyngales et vélaires, que l'on ne trouve pas dans les autres langues (Tableau 1).

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	-	-	+	-
[ḍ]	ظ	Interdentale	+	+	-	+
[ɛ]	ع	Pharyngale	+	-	-	+
[ṣ]	ص	Alvéolaire	-	+	-	+
[ḥ]	ح	Pharyngale	-	-	-	+

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

3.1.1. 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 (figure 7), en ce sens qu'il y a élargissement de la cavité buccale et une constriction du pharynx [6]. Les phonèmes emphatiques de l'Arabe Standard sont respectivement :

- l'occlusive alvéodentale voisée [†] ;
- l'occlusive apicodentale [®] ;
- la fricative interdentale [‡].
- la fricative alvéolaire [©] ;

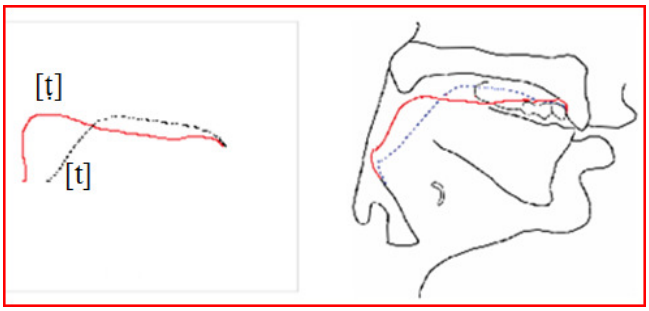
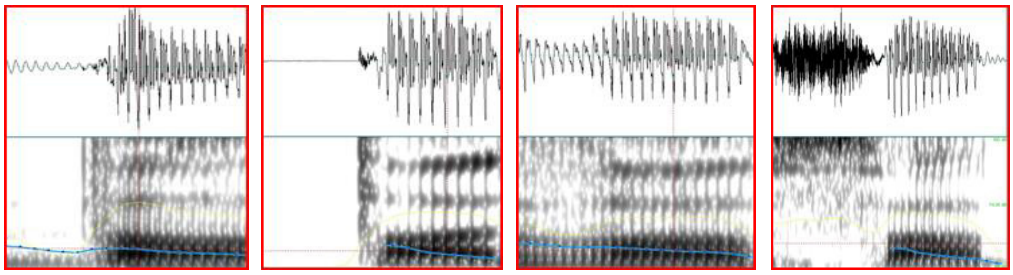


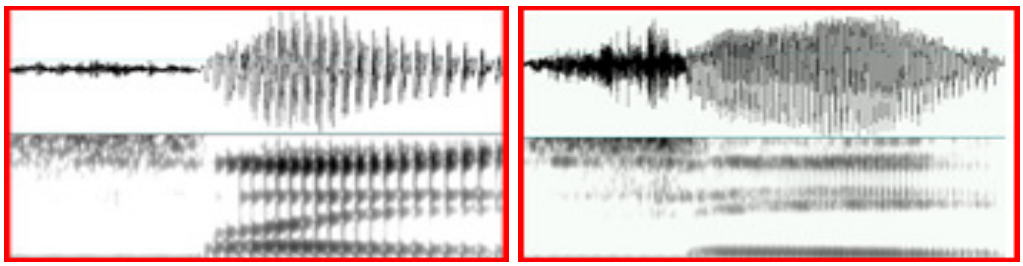
Figure 7. Articulation du phonème emphatique [t□] par rapport à son opposé non emphatique [t]

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 8, 9 et 10).



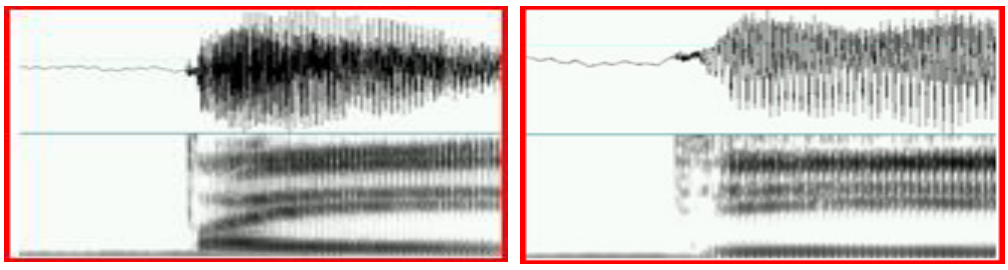
[†a] [®a] [‡a] [©a]

Figure 8. Chute de F_2 lors de la prononciation des phonèmes emphatiques, en contexte [C_ea]



[©i] [si]

Figure 9. Spectrogrammes de l'emphatique fricative [©] par rapport à son opposée non emphatique [s], en contexte [C_ei]



[®i] [ti]

Figure 10. Spectrogrammes de l'emphatique occlusive [®] 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 [7]. Ce chercheur a obtenu de très bons résultats pour la classification de trois phonèmes japonais [b], [d], [g], en partant du principe que pour une modélisation de signaux dynamiques tels que la parole, il est nécessaire d'introduire de la mémoire dans le réseau.

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.
L'approche TDNN vient pour remédier aux problèmes que l'on rencontre avec les approches statistiques utilisée en reconnaissance automatique de la parole, telles que les HMM (Hidden Markov models) qui présentent une faible résistance aux bruits et une importante quantité de données nécessaires pour l'apprentissage [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 [8].

3.3. Base de données des fichiers sons

Nous avons exploité un corpus de 160 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) [9]. 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 360 fichiers sons, avec une fréquence d'échantillonnage de 11025 Hz. 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 les paramètres MFCC, qui permettent de modéliser le signal parole par des filtres conformes à notre système auditif [10]. Nous avons complété ces coefficients MFCC par les dérivées temporelles dites premières ΔMFCC et secondes ΔΔMFCC. Ces dernières permettent de prendre en compte la variabilité temporelle de la parole, et par conséquent son aspect dynamique. Si nous utilisons 13 coefficients MFCC avec le 1er coefficient correspondant à l'énergie, en tenant compte de leurs dérivées, nous aurons 39 vecteurs acoustiques assez représentatifs du signal parole.
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 classification et reconnaissance automatique de la parole 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 [11].

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 11). Ceci consiste à minimiser l'erreur quadratique de sortie E (somme des carrés de l'erreur de chaque composante entre la sortie réelle et la sortie désirée) [12].

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

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

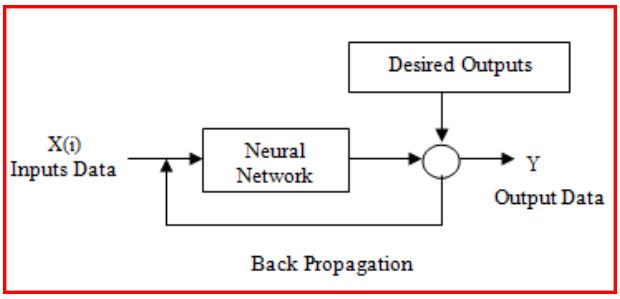


Figure 11. Rétropropagation de l'erreur avec apprentissage supervisé

Pour la technique de rétropropagation de l'erreur, nous avons utilisé l'algorithme de Levenberg-Marquardt, qui minimise l'erreur quadratique d'apprentissage. Cet algorithme donne de bonnes performances, comparé à d'autres algorithmes de rétropropagation [13].
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. Ceci a pour objectif de retrouver la classe du son à tester et ainsi prendre en compte les poids obtenus pour cette classe lors de l'apprentissage.

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

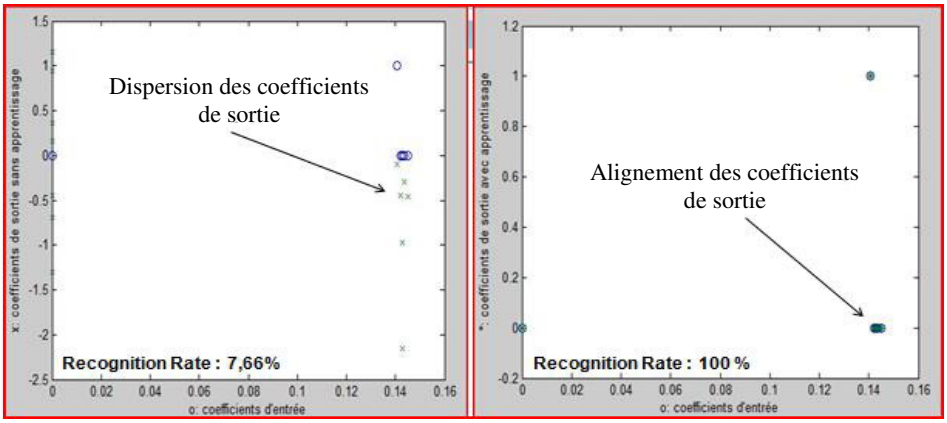
Erreur avant apprentissage

T = [1 0 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

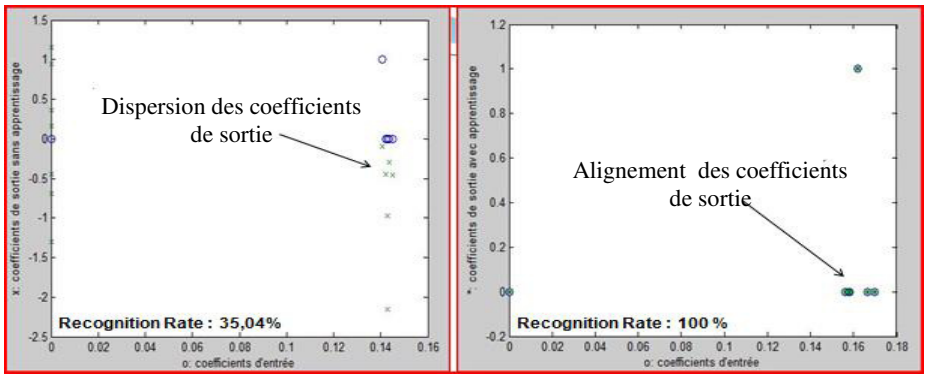
T = [1 0 0 0 0 0 0 0 0 0 0 0 0 0] (vecteur de référence)
Y_2 = 1.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000
(vecteur de sortie obtenu)

Erreur d'apprentissage = 2.3648e-013%
Taux de Reconnaissance = 100%



(a) (b)

Figure 12. Reconnaissance de la fricative emphatique [©],
(a) avant apprentissage
(b) après apprentissage.



(a) (b)

Figure 13. Reconnaissance de l'occlusive emphatique [®],
(a) avant apprentissage
(b) après apprentissage

3.6. Phases de tests de reconnaissance

Pour les tests de reconnaissance, nous avons suivi les différentes étapes qui vont de l'enregistrement online du son jusqu'au test de reconnaissance, en passant par les étapes de détection des frontières, préaccentuation et extraction des paramètres acoustiques.

- **Enregistrement de l'onde acoustique et la préaccentuation**

Nous avons échantillonné le signal à la fréquence de 11025 HZ. Nous appliquons ensuite une préaccentuation au signal (Figure 14), pour la récupération des hautes fréquences et une compensation des effets de filtrage des procédés d'acquisition du signal. Pour cela, nous avons appliqué un filtre FIR (Finite Impulse Response) passe haut, de 1er ordre.

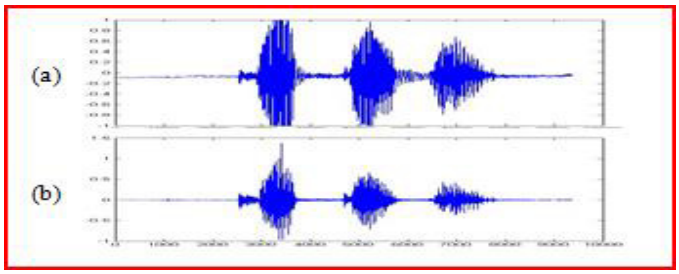


Figure 14. Représentation du mot [kataba]
(a) avant préaccentuation
(b) après préaccentuation

- **Extraction automatique des frontières des mots et fenêtrage**

L'extraction des paramètres acoustiques devrait se faire uniquement sur le signal parole. Pour cela, nous devrons éliminer toutes les trames qui ne sont pas de la parole et délimiter ainsi les débuts et fins de mots. Pour réaliser cette étape, nous avons mis au point une fonction detect_speech sous matlab. Cette fonction utilise un seuil minimal d'énergie que nous avons calculé sur la base d'enregistrements de différents bruits d'environnement (Figure 15).

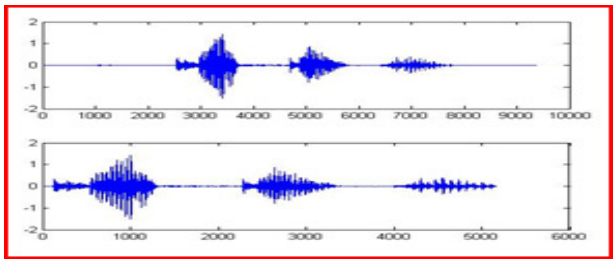


Figure 15. Détection Online des frontières du mot prononcé [kataba]

Du fait que le signal parole est non stationnaire, nous devrons extraire les paramètres acoustiques sur des portions de signal supposées stables. Pour cela, nous choisissons des fenêtres de taille de 30 ms car l'observation du signal parole montre qu'il n'évolue pas ou peu sur des durées de cette taille. Les paramètres sont extraits avec un pas de 10 ms sur toute la fenêtre. Pour cette dernière, nous utilisons une fenêtre de Hamming, qui se présente sous la forme :

$$w[m] = 0.54 - 0.46\cos\left(\frac{2\pi m}{N-1}\right) \qquad m = 0, \ldots, N-1 \qquad (2)$$

N : taille de la fenêtre
m : pas.

- **Tests de reconnaissance et commentaires**

Pour les tests de reconnaissance du vecteur de sortie, le principe est de chercher les vecteurs de référence $\{X_1, X_2, \ldots, X_n\}$ des matrices de référence les plus proches des vecteurs Test

{Y_1, Y_2, …, Y_n}. Pour cela, nous utilisons la distance euclidienne pour choisir la distance minimale, qui correspond au vecteur de référence le plus proche du vecteur test.
Soit i[1,n], un vecteur issu de la paramétrisation et appartenant au mot test, et j[1,n], un vecteur appartenant au mot du dictionnaire de référence et d(i,j) la distance euclidienne.

$$\text{Si } i=\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \quad \text{et } j=\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad \text{alors } d(i,j)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \qquad (3)$$

Il suffit de prendre en considération la valeur de d(i,j) minimale pour choisir le vecteur de référence correspondant.
Il faudra noter que pour le codage des vecteurs de référence, nous avons choisi la méthode classique qui consiste à assigner un 1 pour un seul élément du vecteur et 0 pour tous les autres, de telle sorte que tous les vecteurs possèdent une seule activation à 1 et toutes autres mises à 0.

- **Online Recognition Test**

>> Online sound recording
Cette première étape consiste à enregistrer un son online, puis à éliminer les silences de part et d'autre du signal utile. Ensuite, une préaccentuation permettra de récupérer les fréquences hautes du signal (Figure16).

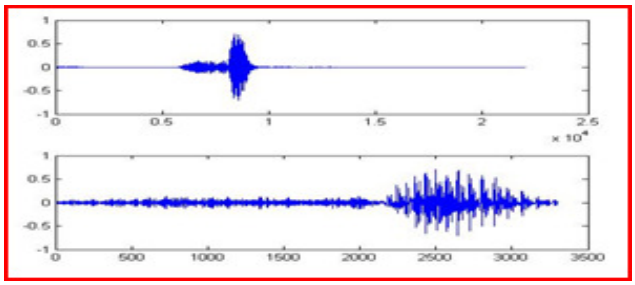


Figure 16. Détection Online des frontières puis préaccentuation du son emphatique [©] en contexte vocalique [a]

Y_test = 1.0191 -0.1314 -0.0760 0.0117 0.0630 0.1059 0.0948 0.1194 0.1044 0.0971 0.0958 0.0711 0.0639
test_error = 0.0810
test_recognition : 91.9%

3.7. 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 360 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.
Nous avons extrait les coefficients MFCC de chaque fichier de son que nous injectons à l'entrée de notre système de reconnaissance, en tenant compte des poids sauvegardés pour chaque classe de phonème lors de la phase d'apprentissage. Pour cela, nous comparons le vecteur de sortie obtenu avec l'ensemble des vecteurs de référence et nous optons pour le phonème correspondant au vecteur de référence le plus proche (tableau 2).

Confusion (%)	[ṭ]	[ṣ]	[ḍ]	[ḍ]	[q]	[ε]	[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
[ε]	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.

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

3.8. Interprétation des résultats

À partir des résultats obtenus, nous pouvons dire que :

- les phonèmes emphatiques [©] et [®] sont reconnus à 100 %. Les deux autres phonèmes emphatiques [‡] et [†] sont reconnus avec des taux respectifs de 70 %, 95 %. Une confusion a été relevée entre [‡] et [†]. 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 [Ÿ]. 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 [Ÿ] 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 donné un aperçu sur les caractéristiques acoustico-physiologiques essentielles des huit phonèmes spécifiques à la langue arabe. Ensuite, nous avons montré la contribution de la méthode des réseaux de neurones artificiels pour l'apprentissage et la reconnaissance automatique de ces phonèmes. Pour ce faire, nous avons appliqué les réseaux à délais temporels TDNN avec la technique d'apprentissage supervisé TrainBr (Bayesian Regularization Backpropagation), exploitant l'algorithme de Levenberg-Marquardt pour la minimization de l'erreur. 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 [©], [ε], [Ÿ], [®] et de 95% pour [†] 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.
Dans une perspective future, nous essayerons d'exploiter les réseaux de neurones pour la reconnaissance et la classification des voix pathologiques. Des essais sont en cours pour la classification des voix parkinsoniennes et œsophagiennes, en ajoutant de nouveaux paramètres acoustiques tels que le Jitter (perturbations du pitch), le Shimmer (perturbations de l'intensité) et le HNR (taux d'influence du bruit sur les harmoniques). Cette classification automatique permettra de situer le degré ou l'ampleur de la maladie pathologique. En d'autres termes, classifier automatiquement les voix pathologiques, d'un ensemble de patients, d'un niveau bas correspondant à une pathologie légère à un niveau haut correspondant à une pathologie profonde ou sévère.

Bibliographie

[1]- Mohamed Mostapha Azmi, Hesham Tolba, Sherif Mahdy and Mervat Fashal, Syllable-Based-Automatic Arabic Speeh Recognition, Proceedings of the 7th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA '08), ISSN: 1790-5117, University of Cambridge, UK, February 20-22, Vol. 4, N°1, pp. 246-250, 2008.
[2]- Chouireb Fatima , Guerti Mhania , Towards a high quality Arabic speech synthesis system based on neural networks and residual excited vocal tract model, Revue Signal, image and video processing, ISSN 1863-1703, vol. 2, n°1, pp. 73-87, 2008.
[3]- Hassan Satori Mostafa Harti and Noureddine Chenfour, Arabic Speech Recognition System Based on CMU Sphinx. International Symposium on Computational Intelligence and Intelligent Informatics, ISCIII'07, pp.31-35, Agadir, Morocco, 28-30 March, 2007.
[4]- Samir Abdelhamid, and Noureddine Bouguechal, SySRA, A System of a Continuous Speech Recognition in Arab Language, Proceedings of World Academy Of Science, Engineering and Technology PWASET, Volume, Vol. 11, ISSN 1307-6884, 2006.
[5]- Gérard Dreyfus & Al, Réseaux de neurones- Méthodologie et Application-, Editions Eyrolles, ISBN 2-212-11 464-8, 2004.
[6]- Kamel Ferrat, Acoustical study of the Tachdid and the Idgham in Standard Arabic. Application for speech synthesis, International Conference, Sciences of Electronic, Technologies of Information and Telecommunication, SETIT2005, Susa (Tunisia), IEEE France, ISBN: 9973-51-546-3, 17-21 March 2005.
[7]- Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin Lang, Phoneme recognition using time-delay networks, IEEE Trans.Acoustics, Speech and Signal Processing, 37(3), pp. 328–339, 1989.

[8]- Joydeep Ghosh, Brian Love, Jennifer Vining and Xuening Sun, Automatique Speaker Recognition using Neural Network, Spring 2004, in http://webspace.utexas.edu/ lovebj/ EE371D _TermProjectCode/, 2004.
[9]- Mansour Alghamdi, KACST Arabic Phonetic Database, The Fifteenth International Congress of Phonetics Science, Barcelona, pp. 3109-3112, 2003.
[10]- Mohamed Chetouani, Codage neuro-prédictif pour l'extraction des caractéristiques de signaux de parole, Thèse de Doctorat Informatique, Université Pierre & Marie Curie, France 2004.
[11]- Richard Povinelli, Michael Johnson, Andrew Lindgren and Jinjin Ye, Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces, IEEE Transactions On Knowledge And Data Engineering, Vol.16, N° 6, June 2004.
[12]- Paul Werbos, Backpropagation through time: What it does and how to do it, Proceedings of the IEEE, 78(10), pp. 1550–1560, 1990.
[13]- Meng-Hock Fun and Martin Hagan, Levenberg-Marquardt Training for Modular Networks, International Conference on Neural Networks, pp. 468-473, 1996.