REVUE ILES D IMESLI, VOLUME 14, N°1 (2022), PP. 7-22

Reconnaissance de phonèmes spécifiques Amazigh par Réseaux de Neurones Artificiels RNA
Recognition of specific amazigh phonemes using artificial neural networks ANN

Kamel Ferrat[1]
[1]Centre de Recherche scientifique et technique pour le developpement de la langue Arabe, Alger, Algérie, k.ferrat@crstdla.dz

Article information

History of the article- Historique de l'article		
Received: 28/03/2022	Accepted : 10/05/2022	Published : 31/12/2022

Abstract
The main idea of artificial neural networks is to draw inspiration from the organization of human biological neurons and their interconnections to process information. These systems that attempt to manipulate and process information in a "similar" way to the biological neurons of the brain are particularly suited to Automatic Speech Recognition. As part of our work, we have used an artificial neural network, called TDNN (Time Delay Neural Networks). The proposed neural network allowed us the recognition of Amazigh phonemes at very appreciable rates. It provides a high percentage of recognition of emphatic phonemes (89.30%) and above all a high recognition rate for posterior phonemes (94.30%). The overall recognition accuracy of all specific phonemes is 91.44%.

Keywords: Automatic Speech Recognition, Neural Networks, Amazigh language, TDNN.

Résumé
L'idée principale des réseaux de neurones artificiels est de s'inspirer de l'organisation des neurones biologiques humains et leurs interconnexions pour traiter l'information. Ces systèmes qui tentent de manipuler et traiter l'information de manière "similaire" aux neurones biologiques du cerveau sont particulièrement adaptés à la reconnaissance automatique de la parole. Dans le cadre de notre travail, nous avons utilisé un réseau de neurones artificiels, dit TDNN (Time Delay Neural Networks). Le réseau de neurones proposé nous a permis la reconnaissance des phonèmes amazighs à des taux très appréciables. Il fournit un pourcentage important de reconnaissance des phonèmes emphatiques (89,30 %) et surtout un haut taux de reconnaissance pour les phonèmes postérieurs (94,30 %). La précision de reconnaissance globale de tous les phonèmes spécifiques est de 91,44 %.

Mots-clés : Reconnaissance Automatique de la Parole, Réseaux de Neurones, langue Amazigh, TDNN

Auteur correspondant : Kamel Ferrat, kamelferrat@yahoo.fr
ISSN: 2170-113X, E-ISSN: 2602-6449,
Published by: Mouloud Mammeri University of Tizi-Ouzou, Algeria

Introduction

Par Reconnaissance Automatique de la Parole (RAP), nous entendons la transformation automatique de séquences de parole en textes écrits ou toute autre action, dans le cadre d'interfaces homme-machine. Pour le passage de la parole vers du texte écrit, le système doit nécessairement passer par des étapes importantes : l'extraction des paramètres acoustiques, la comparaison avec des modèles de référence préalablement enregistrés et sur lesquels nous avons fait subir un apprentissage, et enfin la prise de décision, c'est-à-dire la reconnaissance.

Aujourd'hui, un état de l'art des différents travaux réalisés dans le domaine de la RAP 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 (Kamble, 2016).

Dans le cadre de notre travail, nous avons utilisé les Réseaux de Neurones Artificiels (RNA) pour reconnaitre les sons spécifiques de la langue Amazigh. Les RNA ont donné des résultats appréciables en RAP en Anglais, Français et même en Arabe (Pellegrini at al., 2016; Glackin et al., 2018; Dib, 2019). Nous avons jugé utile de les adapter pour le cas de la langue Amazigh. Pour rappel, nous avons déjà exploité ce type de modèles pour la reconnaissance de paroles pathologiques en kabyle et nous avons exploité les mêmes outils pour reconnaitre les phonèmes emphatiques de la langue arabe (Ferrat, 2015 ; Ferrat and Guerti, 2013).

1. Bref aperçu sur les Réseaux de Neurones Artificiels RNA

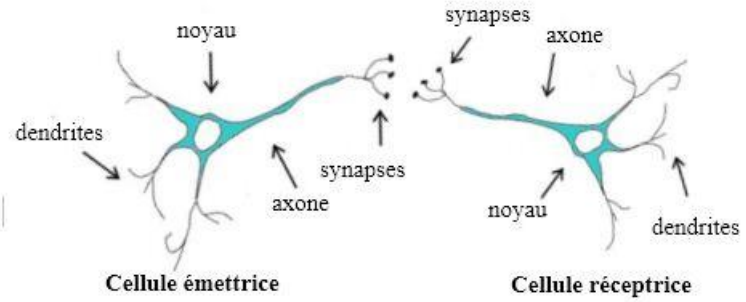
1.1. Neurone biologique et neurone formel

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 RNA et continuent d'inspirer le développement de nouveaux modèles pour la reconnaissance des formes (caractères, visages, images, parole, …).

Le cerveau se caractérise par une organisation très complexe à analyser du fait du grand nombre de cellules, les neurones, et de liens entre ces cellules, les connexions synaptiques, qui les compose. Ce grand nombre de neurones et de connexions conduit à un enchevêtrement qui est, aujourd'hui encore, très difficile à appréhender. 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 à traiter par l'intermédiaire des **connexions synaptiques**. Cette information est acheminée ainsi vers le noyau, également appelé soma, qui la traite et la répercute ensuite selon son poids et donc son importance en sortie de la

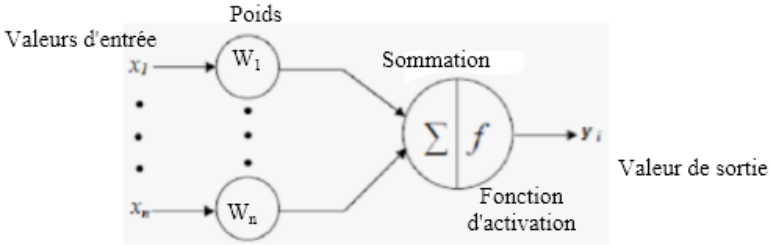
cellule vers l'**axone** qui propage cette information vers d'autres cellules (Figure N°1).

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



En RNA, nous essayons de modéliser le neurone biologique par un neurone formel, sous forme d'une représentation mathématique. En d'autres termes, nous appliquons une modélisation mathématique qui reprend les principes du fonctionnement du neurone biologique, en particulier la sommation des entrées (Figure N°2).

Figure N° 2. Représentation d'un neurone formel.



Ceci peut être exprimé par la fonction mathématique suivante :

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

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.

Plusieurs modèles de RNA sont utilisés pour reconnaitre automatiquement une parole (Jolad and Khanai, 2021). L'idée principale est de s'inspirer de l'organisation du cerveau et des neurones biologiques humains et leurs interconnexions pour traiter l'information (Jolad and Khanai, 2021 ; Dreyfus et al., 2004). Pour ce faire, nous passons nécessairement par deux étapes importantes :

• Une phase d'apprentissage permettant au système de lire les paramètres de référence $\{R_1, R_2, \ldots, R_n\}$, 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.

Dans le cadre de notre travail, nous avons appliqué les réseaux dynamiques à décalages temporels TDNN (Time Delay Neural Network) pour la reconnaissance automatique des phonèmes spécifiques de la langue Amazigh. Nous avons jugé que cette méthode permet de bien identifier et classifier ces phonèmes, 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 pour l'intelligibilité d'un acte de parole (Tebelskis, 1995).

1.2. Réseaux de neurones TDNN en RAP

L'architecture TDNN a été introduite pour la première fois par Alex Waibel pour la reconnaissance de la parole (Waibel et al., 1989). 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.

L'avantage des réseaux TDNN est qu'ils sont capables de traiter des séquences de vecteurs de parole grâce à l'introduction de délais temporels fixes sur les entrées (Figure N°3). Ces délais visent à apprendre la structure temporelle des événements acoustiques et les relations entre ces événements (Gosh et al., 2004).

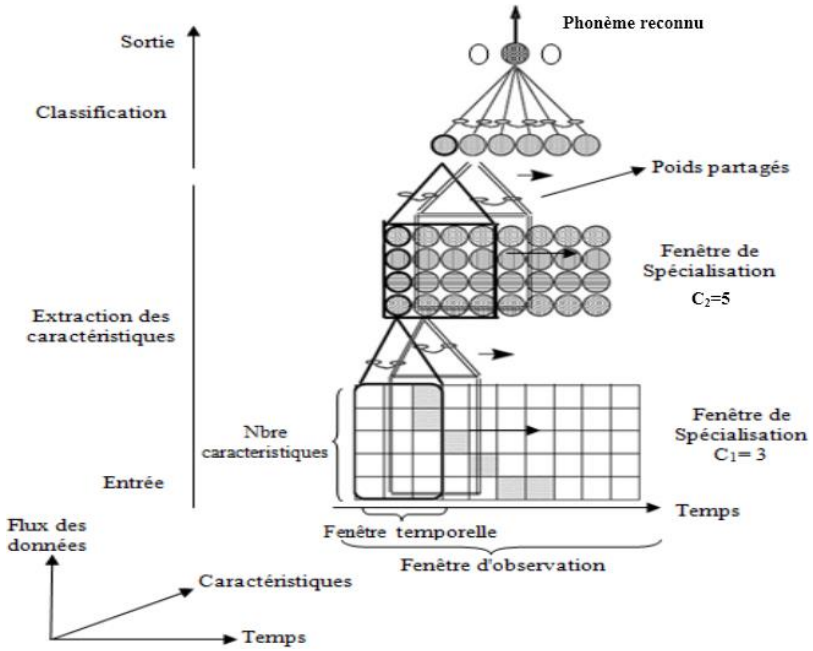
L'approche TDNN vient pour remédier aux problèmes que l'on rencontre avec les approches statistiques utilisée en reconnaissance automatique de la parole, telles que les HMM (Hidden Markov models) qui présentent une faible résistance aux bruits et une importante quantité de données nécessaires pour l'apprentissage (Dreyfus et al., 2004).

Un réseau TDNN se caractérise essentielllement par (Figure N°3) :

- Une fenêtre d'observation : nombre de neurones de chaque couche selon la direction temporelle et qui represente le nombre de neurones de la couche i suivant la caractéristique temporelle vue par un neurone de la couche i+1 ;
- Une fenêtre de spécialisation de la couche d'entrée qui représente le nombre de neurones de chaque couche selon la direction caractéristique ;
- Une fenêtre de spécialisation de la 1ere couche cachée ;
- Un délai temporel (nombre de neurones) entre deux fenêtres successives dans une couche donnée ;
- Couche d'entrée : (nombre donnée de neurones taille de la fenêtre d'observation et un nombre donnée de neurones de la taille des caractéristiques) ;
- Un nombre spécifiques de couches cachées ;

- Des fonctions d'activation (en général, une pour chaque nœud des couches cachées et une autre pour la couche de sortie.

Figure N°3. Le réseau à décalages temporels TDNN



2. Les sons spécifiques de la langue Amazigh

La langue Amazigh comprend 33 phonèmes, soit 29 consonnes pour seulement 4 voyelles [a, i, u] en ajoutant la voyelle neutre ou "ilem" et appelée "Schwa" [e] (Tigziri, 2015). C'est une langue consonantique très riche mais présente peu de voyelles contrairement à l'Anglais et le Français.

Dans le cadre de notre travail, nous avons étudié quelques consonnes spécifiques que l'on ne retrouve pas dans les langues française et anglaise. Les phonèmes spécifiques choisis de la langue Amazigh sont au nombre de sept : Quatre phonèmes emphatiques dont deux sont voisés et les deux autres sourds, et trois phonèmes postérieurs dont deux pharyngales.

Rappelant que la langue Amazigh présente quelques spécificités telles que la présence de consonnes postérieures pharyngales, et uvulaires, que l'on retrouve certes dans la langue arabe, mais pas dans les langues française et anglaise (Tableau N°1). C'est également le cas pour les consonnes dites emphatiques.

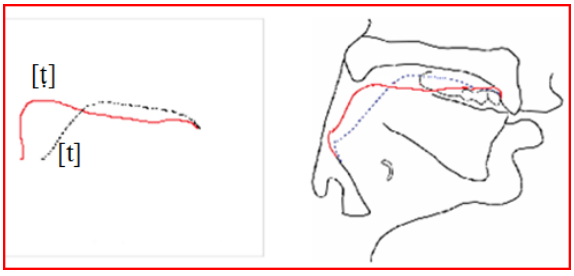
Tableau N°1. Lieux et modes d'articulation de quelques sons spécifiques de la langue Amazigh.

			Mode d'articulation			
Phonème	**Caractère Arabe**	**Lieu d'articulation**	**voisement**	**emphatique**	**occlusive**	**fricative**
[ṭ]	ط	**Apico-dentale**	-	+	+	-
[ṣ]	ص	**Alvéolaire**	-	+	-	+
[ẓ]	-	**Alvéolaire**	+	+	-	+
[ḍ]	ظ	**Interdentale**	+	+	-	+
[q]	ق	**Uvulaire**	-	-	+	-
[ḥ]	ح	**Pharyngale**	-	-	-	+
[ɛ]	ع	**Pharyngale**	+	-	-	+

Sur le plan articulatoire, les consonnes emphatiques sont prononcées avec un report en arrière de la racine de la langue et un abaissement et creusement du dos de la langue (Figure N°4), en ce sens qu'il y a élargissement de la cavité buccale et une constriction du pharynx (Ferrat & Guerti, 2013). Les consonnes emphatiques de l'Amazigh sont respectivement (Tableau 1) :

- l'occlusive apico-dentale sourde [ṭ] ;
- la fricative sifflante alvéolaire voisée [ẓ] ;
- la fricative sifflante alvéolaire sourde [ṣ] ;
- la fricative interdentale voisée [ḍ].

Figure N°4. Articulation du phonème emphatique [ṭ] 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 N°5, 6 et 7).

Figure N°5. Chute de F_2 lors de la prononciation des phonèmes emphatiques, en contexte [C_ea].

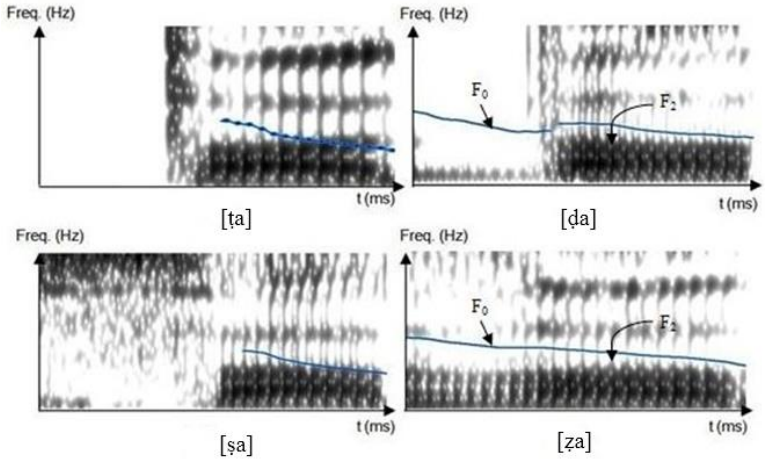


Figure N°6. Spectrogrammes de l'emphatique fricative [ṣ] par rapport à son opposée non emphatique [s], en contexte [C_ei].

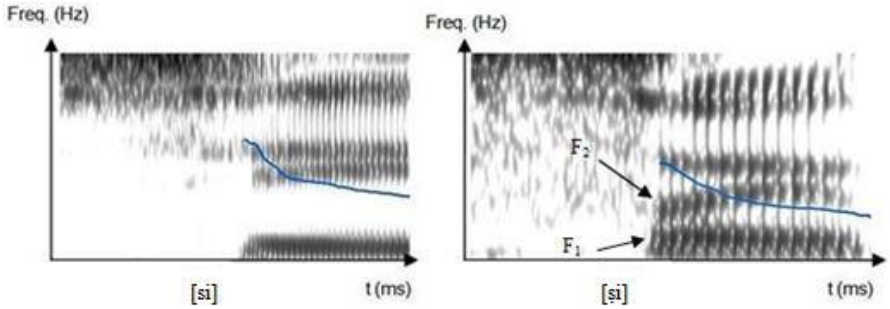
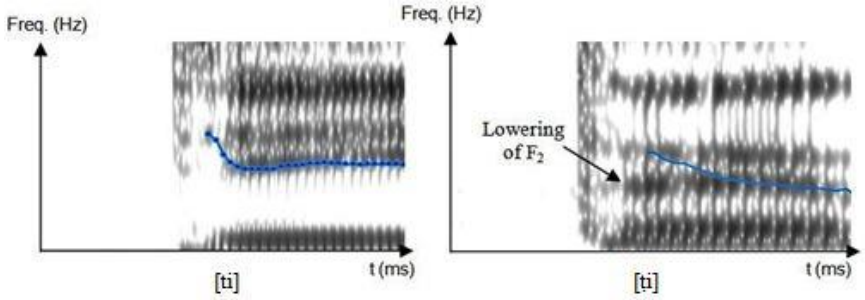


Figure N°7. Spectrogrammes de l'emphatique occlusive [ṭ] par rapport à son opposée non emphatique [t], en contexte [C_ei].



Rappelant que les formants représentent les fréquences de résonance des cavités de l'appareil phonatoire. Le formant F_1 correspond à la fréquence de résonance de la cavité pharyngale et le formant F_2 à celle de la cavité buccale.

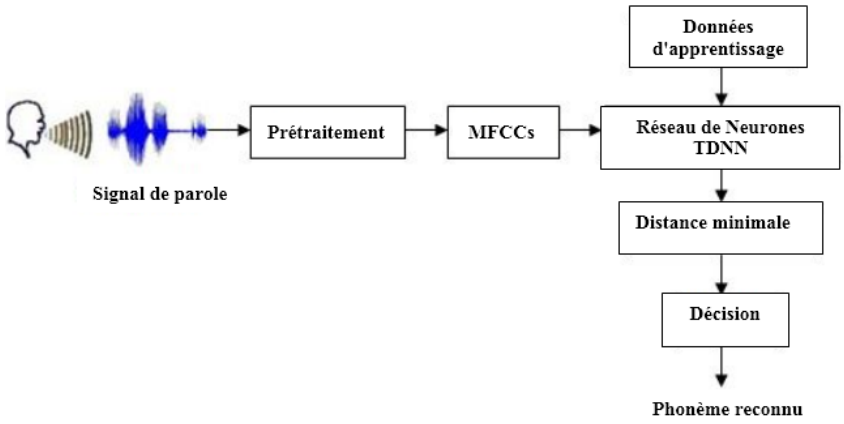
3. Reconnaissance automatique des phonèmes spécifiques Amazigh par TDNN

3.1. Architecture de notre système de reconnaissance

Pour réaliser notre système de reconnaissance, nous avons utilisé une base de données de sons que nous avons préalablement enregistrés et exploités dans les phases d'apprentissage et de reconnaissance. Pour l'extraction des vecteurs acoustiques, nous avons utilisé 39 coefficients acoustiques dits MFCC (Figure N° 8).

Lors de la phase d'apprentissage, nous avons utilisé la technique dite de rétropropagation de l'erreur (backpropagation) basée sur l'algorithme de Levenberg-Marquardt qui minimise l'erreur quadratique d'apprentissage (Werbos, 1990 ; Fun & Hagan, 1996). Dans cette phase d'apprentissage, nous avons utilisé un ensemble de 120 fichiers sonores contenant des sons spécifiques de la langue Amazigh, extraits à partir d'un corpus sonore préalablement enregistré. Les réseaux de neurones ainsi que la technique d'apprentissage sont implémentés avec Matlab's Neural Network Toolbox 7.5.

Figure N°8. Structure de notre système de RAP basé sur les RNA.



Les paramètres choisis pour modéliser notre système TDNN sont respectivement (Figure N° 2) :

• fenêtre d'observation : taille de 160 ms, soit 14 trames de 30 ms avec un pas de 10 ms
• fenêtre de spécialisation de la couche d'entrée : C1=3 ;
• fenêtre de spécialisation de la 1ere couche cachée : C2 = 5 ;
• Délai temporel entre 2 fenêtres successives, Délai=1 ;

• Couche d'entrée : (14*39) neurones (14 neurones taille de la fenêtre d'observation et 39 neurones taille des caractéristiques) ;
• 2 couches cachées : 12 et 8 trames respectivement.
• Fonctions d'activation :
• fonction tangente hyperbolique (type sigmoïde) pour chaque nœud des couches cachées ;
• fonction linéaire pure pour la couche de sortie.

3.2. Base de données des fichiers sons

Nous avons exploité un corpus de 120 fichiers sons extraits à partir de l'enregistrement préalable d'un corpus contenant un ensemble de mots et de phrases. Pour la validation de nos résultats, nous avons enregistré un ensemble de 84 fichiers sons, avec une fréquence d'échantillonnage de 11025 Hz. Les enregistrements ont été réalisés au laboratoire, en milieu naturel contenant un bruit environnant. Nous avons utilisé, comme outil d'enregistrement, le logiciel Praat (Boersma and Weenink, 2021).

3.3. 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 (Souissi and Cherif, 2016). Nous avons complété ces coefficients MFCC par les dérivées temporelles dites premières ΔMFCC et secondes ΔΔMFCC. Ces dernières permettent de prendre en compte, de façon précise, la variabilité temporelle de la parole, et par conséquent donc son aspect dynamique (Tiwari, 2010). 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 de parole.

3.4. Phase d'apprentissage

Dans le cadre de notre travail, nous avons utilisé un apprentissage dit "supervisé" qui force le réseau à converger vers un état final précis, en même temps qu'on lui présente un motif (Cunningham et al., 2008). Pour ce faire, nous adaptons 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. La méthode 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) (Werbos, 1990).

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

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

Pour minimiser l'erreur quadratique d'apprentissage, nous avons utilisé une technique de rétropropagation de l'erreur, basée sur l'algorithme de Levenberg-Marquardt. Pour la prise en compte des poids obtenus par apprentissage lorsque nous passons à la phase de reconnaissance, nous appliquons la distance euclidienne, 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 la phase de l'apprentissage.

- **Exemple d'apprentissage de l'emphatique [ẓ] (Etat isolé)**

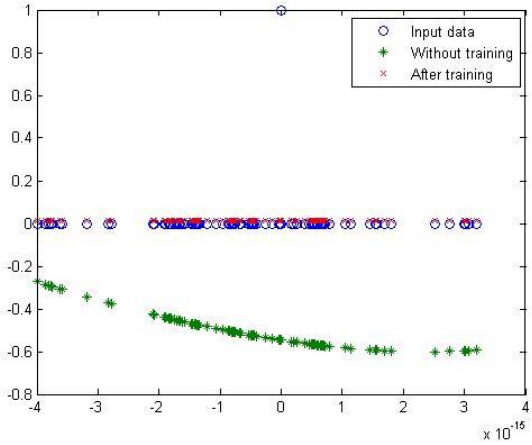
Erreur avant apprentissage :

T = [1 0 0 0 0 0 0 0 0 0 0 0 0] (vecteur de référence)
Y_1 = 0.8631 1.5640 2.2018 -0.2542 0.0074 -0.4148 0.7468 0.5031 0.8531 0.3692 (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%

Figure N°9. Reconnaissance de la fricative emphatique [ẓ], (+) avant apprentissage, (x) après apprentissage.



3.5. Phase 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.

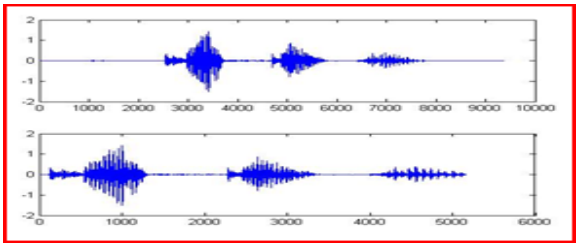
- **Enregistrement de l'onde acoustique et la préaccentuation**

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

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

L'extraction des paramètres acoustiques devrait se faire uniquement sur le signal parole. Ainsi, nous devrons éliminer toutes les trames qui ne sont pas de la parole et délimiter les débuts et fins de mots. Pour réaliser cette étape, nous avons mis au point une fonction 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 N°10).

Figure N° 10. Détection Online des frontières d'un mot prononcé.



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 Hamming 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 vecteurs acoustiques sont extraits avec un pas de 10 ms sur toute la fenêtre. Rappelant que la fenêtre de Hamming 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.

Les 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 (Povinelli et al., 2004).

- **Tests de reconnaissance et commentaires**

Pour les tests de reconnaissance du vecteur de sortie, le principe est de chercher les vecteurs $\{X_1, X_2, \ldots, X_n\}$ des matrices de référence les plus proches des vecteurs Test $\{Y_1, Y_2, \ldots, Y_n\}$. 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.

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

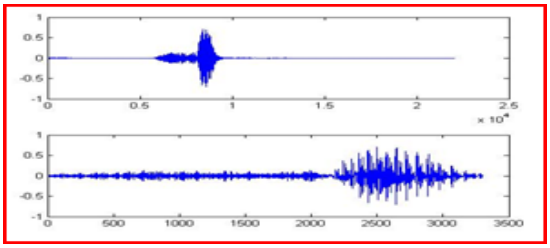
Il suffit de prendre en considération la valeur de d(i,j) minimale pour choisir le vecteur de référence correspondant.

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**

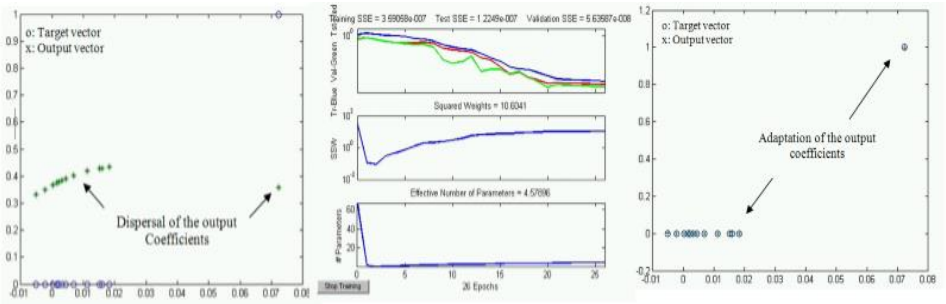
Cette première étape consiste à enregistrer un son online, puis à éliminer les silences de part et d'autre du signal utile. Ensuite, une préaccentuation permettra de récupérer les fréquences hautes du signal (Figure N°11).

Figure N° 11. Détection Online des frontières puis préaccentuation du son emphatique [ẓ] en contexte vocalique [a].



Y_test = 1.1082 -0.0936 -0.0618 0.1223 0.0541 0.24157 0.1876 0.1205 0.1189 0.0886 0.4942 0.0725 0.0645
test_recognition: 87.60%

Figure N°12. Test de reconnaissance de [ẓa] après phase d'apprentissage.



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

Les tests de validation en 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 subi d'apprentissage. Nous avons exploité 84 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 N°2).

Tableau N°2. Matrice de confusion et taux moyens de reconnaissance des phonèmes en milieu naturel.

Confusion (%)	[ṭ]	[ṣ]	[ẓ]	[ḍ]	[q]	[ε]	[ḥ]	**TR (%)**
[ṭ]	**92.30**	01.10	02.30	02.60	01.70	00.00	00.00	92.30
[ṣ]	00.10	**83.90**	15.30	00.70	00.00	00.00	00.00	83.90
[ẓ]	00.00	11.20	**87.60**	01.20	00.00	00.00	00.00	87.60
[ḍ]	00.00	00.00	3.20	**93.40**	00.00	02.30	01.10	93.40
[q]	02.80	00.00	01.70	01.90	**93.60**	00.00	00.00	93.60
[ḥ]	00.00	00.90	00.00	01.10	00.00	**95.10**	02.90	95.10
[ε]	00.00	00.10	00.10	01.70	00.00	03.90	**94.20**	94.20
TR: Taux de Reconnaissance **TGR: Taux de Reconnaissance Global**								**TGR** **91.44 %**

3.7. Interprétation des résultats

A partir des résultats obtenus, nous pouvons dire que :

- les phonèmes emphatiques [t̪] et [d̪] sont reconnus respectivement à 92,30% et 93,40%. Les deux autres phonèmes emphatiques [ṣ] et [ẓ] sont reconnus avec des taux respectifs de 83,90% et 87,60 %. Une confusion perceptible a été relevée entre [ṣ] et [ẓ]. Ceci est peut-être dû au fait que ces deux phonèmes sont très proches dans la prononciation avec une friction sous forme de bruit dans la même plage des hautes fréquences, un même lieu d'articulation et une emphase. Ces caractéristiques communes semblent voiler leur différence en mode de voisement (l'une voisée et l'autre sourde).
- le taux global de reconnaissance des phonèmes emphatiques est de 89,30% ;
- Les deux pharyngales [ε] et [ḥ] présentent les plus forts taux de reconnaissance à savoir respectivement 94,20 % et 95,10 %. Avec le taux de reconnaissance de 93,60 % pour le phonème [q], nous pouvons dire que les phonèmes postérieurs de la langue Amazigh s'adaptent bien à la méthode de reconnaissance choisie (94,30 % de taux de reconnaissance) ;
- dans l'ensemble, un taux de reconnaissance appréciable de 91,44 % a été obtenu pour les phonèmes exploités dans le cadre de ce travail.

4. Conclusion

Dans ce travail, nous avons donné un aperçu sur les caractéristiques acoustico_physiologiques essentielles de quelques phonèmes spécifiques à la langue Amazigh. 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 dite 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 naturel, avec notamment un taux d'identification de 89,30 % des quatre phonèmes emphatiques et de 94,30 % pour les phonèmes postérieurs. Des confusions de reconnaissance persistent pour les cas des deux phonèmes [ṣ] et [ẓ], dont la prononciation présente beaucoup de caractéristiques communes (sifflement, bruit dans les hautes fréquences, emphase, lieu d'articulation).

Dans une perspective future, ces résultats pourront être exploités pour la reconnaissance de la parole continue en langue Amazigh.

5. Références

Boersma Paul, Weenink David, 2021, Praat: doing phonetics by computer. Version 6.1.55, date de consultation 25 October 2021 from http://www.praat.org/

Cunningham Padraig, Cord Matthieu, Delany Sarah Jane, 2008, *Supervised Learning*. In Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg, Germany.

Dib Mohammed, 2019, "Arabic Automatic Speech Recognition", in *Automatic Speech Recognition of Arabic Phonemes with Neural*

Networks, SpringerBriefs in Applied Sciences and Technology. https://doi.org/10.1007/978-3-319-97710-2_5

Dreyfus Gérard et al., 2004, *Réseaux de neurones- Méthodologie et Application-*, Editions Eyrolles, France, ISBN 2-212-11 464-8.

Ferrat Kamel, 2015, "Parkinsonian Voice Classification of Kabyle Berber Patients Using Neural Networks", in *Journal of Rehabilitation Medicine*. Indexed SCOPUS, Thomson Reuters JCR, 47(54), p.242, Abstracts of the 9th world congress of international society of physical and rehabilitation medicine, June 19-23, 2015, Berlin, Germany.

Ferrat Kamel, Guerti Mhania, 2013, "Classification of the Arabic Emphatic Consonants using Time Delay Neural Network", in *International Journal of Computer Applications*, Vol. 80, No.10, pp: 1-6, Published by Foundation of Computer Science, New York, USA. ISSN-2250-1797. http://dx.doi.org/10.5120/13894-9341.

Fun Meng-Hock, Hagan Martin, 1996, "Levenberg-Marquardt Training for Modular Networks", in *International Conference on Neural Networks*, pp. 468-473.

Glackin Cornelius et al., 2018, "Convolutional Neural Networks for Phoneme Recognition", in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*, 1, pp. 190-195. DOI: 10.5220/0006653001900195

Jolad Bhuvaneshwari, Khanai Rajashri, 2021, "ANNs for Automatic Speech Recognition - A Survey", in *Jeena Jacob I., Gonzalez-Longatt F.M., Kolandapalayam Shanmugam S., Izonin I. (eds) Expert Clouds and Applications. Lecture Notes in Networks and Systems*, 209, 2021. Springer, Singapore. https://doi.org/10.1007/978-981-16-2126-0_4

Kamble Bhushan, 2016, "Speech recognition using artificial neural network-a review", in *International journal of computing, communication and instrumentation engineering*, 3(1), pp. 61-64, 2016.

Pellegrini Thomas, Fontan Lionel, Sahraoui Halima, 2016, "Réseau de neurones convolutif pour l'évaluation automatique de la prononciation", in *Journées d'Etudes sur la Parole (JEP2016)*, July 2016, Paris, France.

Povinelli Richard, Johnson Michael, Lindgren Andrew, Ye Jinjin, 2004, "Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces", in *IEEE Transactions On Knowledge And Data Engineering, 16(6)*.

Souissi Nawel, Cherif Adnane, 2016, "Speech recognition system based on short-term cepstral parameters, feature reduction method and Artificial Neural Networks", in *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP*), pp.667-671. doi: 10.1109/ ATSIP.2016.7523163

Tebelskis Joe, 1995, *Speech Recognition Using Neural Networks*, PhD Dissertation, School Of Computer Science, Carnegie Mellon University.

Tiwari Vibha, 2010, "MFCC and its applications in speaker recognition", in *International Journal on Emerging Technologies*, 1, pp. 19-22.

Tigziri Nora, 2015, *Phonétique Acoustique du Kabyle*, Office des Publications Universitaires OPU, Algérie.

Waibel Alex, Hanazawa Toshiyuki, Hinton Geoffrey, Shikano Kiyohiro, Lang Kevin, 1989, "Phoneme recognition using time-delay networks", in *IEEE Trans.Acoustics, Speech and Signal Processing*, 37(3), pp. 328–339.

Werbos Paul, 1990, "Backpropagation through time: What it does and how to do it", in *Proceedings of the IEEE*, 78(10), pp. 1550–1560.