SETIT 2005
3rd International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 27-31, 2005 – TUNISIA

La Reconnaissance des Phonèmes par les Réseaux de Neurones à Délais Temporels Adaptatifs

Asmaa OURDIGHI, Zehor OUKSILI et Abdelkader BENYETTOU

Laboratoire SIgnal-IMage-PArole (SIMPA)
Dépt. Informatique, Université des Sciences et de la Technologie d'Oran – USTO, Algérie

ouras2003@hotmail.com
Zeiak@hotmail.com
aek.benyettou@univ-usto.dz

Résumé: Les réseaux de neurones à délais temporels adaptatifs (ATNN) présentent une capacité d'adapter les valeurs de leurs retards temporels suivant le contexte de l'apprentissage. L'ajustement des délais assure au réseau plus de fiabilité quant à la performance du système à la reconnaissance, et donc à la connaissance de l'entrée introduite au réseau. Dans cet article, L'objectif a été d'introduire l'ATNN, pour une première contribution dans la reconnaissance de la parole, en procédant par classes phonétiques de la base TIMIT afin d'observer l'adaptation finale des délais correspondants aux taux de reconnaissance optimaux.
Mots clés: Classification phonétique, Délais temporels, Réseaux de neurones.

1 Introduction

Les réseaux de neurones à délais temporels (TDNN) inspirent, par leurs architectures, à une parfaite adaptation des données présentant des caractéristiques temporelles. Le TDNN inculque au signal entrant une sorte de spatialisation effectuée par un balayage temporel de l'entrée. Cette modélisation induit à une connaissance du signal dans le réseau lors de l'extraction.

Cependant, le choix manuel qui dirige l'extraction des caractéristiques présentées à l'entrée du réseau suscite des difficultés d'initialisation du paramètre délai propre aux connexions de ce réseau de type convolutif. Ces délais, qui représentent les décalages temporelles responsables de l'extraction du signal sur un champ de vision limité appelé fenêtre de spécialisation, sont initialement fixés sur les mêmes positionnement pendant et après l'apprentissage.

Les ATNNs proposés par (Lin, 1994) apparaissent comme une solution via cette initialisation hasardeuse des délais, qui toute au long de l'apprentissage, admettent des ajustements par le biais d'un algorithme proche de celui de la rétropropagation du gradient. Ainsi, nous obtenons une nouvelle combinaison de poids et de délais optimaux pour la tâche requise du réseau.

Dans plusieurs travaux, le TDNN a montré une fiabilité dans divers problèmes complexes de la reconnaissance des formes : reconnaissance de la parole (Weibel & al, 1989), reconnaissance de l'écriture manuscrite (Poisson & al, 2001), reconnaissance du locuteur (Bennani, 1992), reconnaissance de piétons (Wöhler & al, 1998), etc.

Par opposition, et dans des problèmes prédictifs telles que les trajectoires (Benyettou & al, 2002), ou encore les fonctions chaotiques (Lin, 1994) (Benyettou & al, 2004), l'ATNN a dévoilé des résultats bien meilleurs que ceux obtenus par un simple TDNN. Néanmoins, ses problèmes restent de simple envergure et leurs bons rendements persistent dans la nature prédictive que cachent les réseaux convolutifs par leur topologie en terme général.

Ce travail introduira l'ATNN dans une classification phonétique d'un sous corpus de la base de données TIMIT. La classification établit par catégories de phonèmes (voyelles - fricatives-plosives), sera focalisée sur les adaptations finales des délais, qui révéleront l'information retenue par le réseau vis-à-vis de chaque catégorie de phonèmes.

2 Paradigme d'apprentissage

L'adaptation des poids et des retards est basée sur la méthode de la descente du gradient. En général, les paramètres sont réadaptés d'une manière discrète, sachant que les entrées du signal sont différentiables.

Les différents paramètres participant dans un ATNN sont comme suit (Lin, 1992):

L : le nombre de couches dans le réseau.
N_h : l'ensemble de nœuds de la couche *h* {1,2,…, / N_h /}.
$\tau_{jik,h-1}$: le délai temporel de la $K_{ième}$ connexion du nœud *i* de la couche *h*-1 vers le nœud *j* de la couche *h*.
$K_{ji,h-1}$: le nombre total de connexions du nœud *i* (couche *h*-1) vers le nœud *j* (couche *h*).
$T_{ji,h-1}$: l'ensemble des délais des connexions provenant du nœud *i* (couche *h*-1) vers le nœud *j* (couche *h*).
$a^{\mu}_{i,0}(t)$: l'entrée du $i^{\text{ième}}$ canal de la séquence d'apprentissage μ au moment *t*.
$W_{ji,h-1}$: le poids synaptique de la $K_{ième}$ connexion du nœud *i* de la couche *h*-1 vers le nœud *j* de la couche *h*. *k=1,2,...,* $K_{ji,h-1}$
t_n : le $n^{\text{ième}}$ échantillon temporel.

Ainsi, nous pouvons définir toutes les formules de l'algorithme d'apprentissage de L'ATNN. La sortie d'un nœud *j* est générée par une fonction dérivable non décroissante '*f* ' qui est appliquée à la somme pondérée des entrées correspondantes à la séquence μ au moment t_n donnée par [1].

$$a^{\mu}{}_{j,h}(t_n) = \begin{cases} f_j(S^{\mu}{}_{j,h}(t_n) - si : h \geq 2 \\ a^{\mu}{}_{j,0}(t_n) - si : h = 1 \end{cases} \quad (1)$$

Où :

$$S^{\mu}{}_{j,h}(t_n) = \sum_{i \in N_{h-1}} \sum_{k=1}^{k_{ji,h-1}} w_{jik,h-1} * a^{\mu}{}_{i,h-1}(t_n - \tau_{jik,h-1}) \quad (2)$$

Un mécanisme d'apprentissage incrémental est utilisé dans ATNN, donc l'adaptation se fera aussi par ajustement incrémental. Une mesure d'erreur instantanée est définie par [3].

$$E(t_n) = 1/2 \sum_{j \in N_L} (d_j(t_n) - a_j(t_n))^2 \quad (3)$$

Où l indique la couche de sortie et $d_j(t_n)$ la valeur désirée du nœud de sortie *j* à l'instant t_n .La valeur des délais est modifiée étape par étape. La règle d'ajustement est par conséquent comme suit:

$$\Delta\tau_{jik,h-1} = -\eta_1(\partial E(t_n)/\partial\tau_{jik,h-1}) \quad (4)$$

Sachant que η_1 est le pas d'apprentissage.

La dérivation de l'algorithme d'apprentissage est faite suivant la série de règles [1] [2] [3] [4], le second facteur de [5] peut être exprimé comme suit dans [6].

$$\partial E(t_n)/\partial\tau_{jik,h-1} = (\partial E(t_n)/\partial S_{j,h}) * (\partial S_{j,h}(t_n)/\partial\tau_{jik,h-1}) \quad (5)$$

$$\frac{\partial S_{j,h}(t_n)}{\partial\tau_{jik,h-1}} = \frac{\partial}{\partial\tau_{jik,h-1}} \sum_{p \in N_{h-1}} \sum_{q=1}^{K_{ji,h-1}} w_{jpq\,h-1} a_{p,h-1}(t_n - \tau_{jpq,h-1})$$

$$= -w_{jik,h-1} a'_{j,h-1}(t_n - \tau_{jik,h-1}) \quad (6)$$

On définit:

$$\rho_{j,h}(t_n) = \partial E(t_n)/\partial S_{j,h} \quad (7)$$

On remplace les équations [6] et [7] dans l'équation [5] on obtient :

$$\Delta\tau_{jik,h-1} = \eta_1 \rho_{j,h}(t_n) w_{jik,h-1} a'_{j,h-1}(t_n - \tau_{jik,h-1}) \quad (9)$$

$$\frac{\partial E(t_n)}{\partial\tau_{jik,h-1}} = -\rho_{j,h}(t_n) w_{jik,h-1} a'_{j,h-1}(t_n - \tau_{jik,h-1}) \quad (8)$$

Pour dériver $\rho_{j,h}(t_n)$, nous avons besoin d'appliquer une chaîne de règles et prendre en compte deux cas différents.

$$\rho_{j,h}(t_n) = \frac{\partial E(t_n)}{\partial S_{j,h}} = \frac{\partial E(t_n)}{\partial a_{j,h}} * \frac{\partial a_{j,h}(t_n)}{\partial S_{j,h}}$$

$$= \frac{\partial E(t_n)}{\partial a_{j,h}} f'(S_{j,h}(t_n)) \quad (10)$$

À partir de [10], deux cas peuvent être considérés. Le cas où *Sij* est une unité de sortie exprimée par [11], et le cas où *Sij* est une unité cachée exprimée par [12].

$$\rho_{j,h}(t_n) = -(d_j(t_n) - a_{j,h}(t_n)) f'(S_{j,h}(t_n)) \quad (11)$$

$$\rho_{j,h}(t_n) = -\left[\sum_{p \in N_{h+1}} \sum_{q=1}^{k_{pi,h}} \rho_{p,h+1}(t_n) W_{pjq\,h}(t_n)\right] f'(S_{j,h}(t_n)) \quad (12)$$

De l'équation [9], il reste à développer $d_{i,h-1}(t_n - \tau_{jik,h-1})$. Ce terme difficile à élaborer à cause de l'apparition des mêmes dérivations dans les deux parties de l'équation, nous oblige donc à procéder à faire un rapprochement comme suit :

$$d_{i,h-1}(t_n - \tau_{jik,h-1}) \approx \begin{cases} \dfrac{a(t_n) - a(t_{n-1})}{r} - si - \tau_{jik,h-1} = 0 \\ \dfrac{a(t_{k+1}) - a(t_{n-1})}{2r} - si - (t_n - \tau_{jik,h-1}) = t_k, \tau_{jik,h-1} = 0 \end{cases} \quad (13)$$

Où $r = t_k - t_{k-1}, k \in \{0,1,.....,n\}$.

Théoriquement, $\tau_{jik,h-1}$ peut être un zéro ou n'importe quel nombre positif réel, mais en ce sens, le réseau va être entraîner dans une sorte d'un bruit généré par l'approximation [13] (Lin, 1994). Pour cette raison, on prendra la valeur entière du retard.

La règle d'ajustement d'un délai temporel est donnée par [14].

$$\Delta\tau_{jik,h-1} = \eta_1 \rho_{j,h}(t_n) w_{jik,h-1} a'_{j,h-1}(t_n - \tau_{jik,h-1}) \quad (14)$$

Où $P_{j,h}(t_n)$ s'obtient suivant les 2 cas existants.

$$\rho_{j,h}(t_n)=\begin{cases}-(d_j(t_n)-a_{j,h}(t_n))f'(S_{j,h}(t_n)) \\ -\left[\sum_{p\in N_{h+1}}\sum_{q=1}^{k_{pj,h}}\rho_{p,h+1}(t_n)w_{pjq,h}(t_n)\right]f'(S_{j,h}(t_n)\end{cases} \quad (15)$$

D'une façon similaire, la règle d'apprentissage pour les poids sera exprimée par [16].

$$\Delta w_{jik,h-1}=\eta\delta_{j,h}(t_n)a_{j,h-1}(t_n-\tau_{jik,h-1}) \quad (16)$$

Où $\delta_{j,h}(t_n)=-\rho_{j,h}(t_n)$, et $\delta_{j,h}$ s'obtient suivant les deux cas considérés :

$$\delta_{j,h}(t_n)=\begin{cases}(d_j(t_n)-a_{j,h}(t_n))f'(S_{j,h}(t_n)) \\ \left[\sum_{p\in N_{h+1}}\sum_{q=1}^{k_{pj,h}}\delta_{p,h+1}(t_n)w_{pjq,h}(t_n)\right]f'(S_{j,h}(t_n)\end{cases} \quad (17)$$

Ainsi, le réseau connaîtra deux adaptations.

3 Classification phonétique

Le but est d'aboutir à des structures neuronales à délais temporels adaptatifs capables de reconnaître les phonèmes appartenant à la même catégorie.

Pour cela nous avons procédé à l'élection d'un sous corpus de la base TIMIT qui elle-même contient 61 phonèmes constituant la phonétique de la langue anglaise, et est divisée en trois classes de phonèmes : voyelles, fricatives et plosives. Nous avons choisi de chaque classe trois phonèmes (voir table.1). Chaque phonème, est organisé sous forme d'un ensemble d'occurrences le représentant qui elles-mêmes sont tirées du corpus obtenu sous différents dialectes par plusieurs orateurs et d'un total de 6300 phrases. L'occurrence est un enchaînement de vecteurs de 13 MFCC (Mel Frequency Cepstral Coefficients).

Classes phonétiques	Phonemes	Train	Test
voyelles	/ah/	2200	879
	/ax/	3352	1323
	/uh/	502	221
fricatives	/dh/	2058	822
	/f/	2093	911
	/sh/	2144	796
plosives	/b/	399	182
	/g/	1337	546
	/p/	2056	779

Table1. *Le sous-corpus de TIMIT*

4 Structure de l'ATNN

Vu que l'architecture de l'ATNN est modélisée pour des données temporelles, effectuer une quantification vectorielle sur les phonèmes est dérisoire, à cause de son traitement qui nous fait perdre la notion du temps présente dans l'enchaînement des vecteurs MFCC à l'intérieur des occurrences de chaque phonème. Donc, travailler sur le sous corpus sans compression est la façon la plus avérée pour ce type de réseau, mais sans doute la plus coûteuse en temps d'exécution. Ce raisonnement conduit à une architecture ayant comme entrée 13 neurones représentant les 13 coefficients d'un vecteur MFCC.

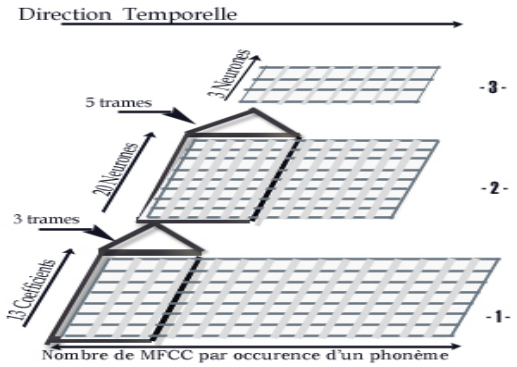

Figure 1. *Architecture du réseau*

Après plusieurs essais l'architecture finale optée pour les trois réseaux chargés de la classification par catégorie est un réseau à deux niveaux. C'est-à-dire, un réseau à une seule couche cachée de 20 neurones, chaque neurone possède une vision de 3 trames sur la couche d'entrée. La couche de sortie est organisée en 3 neurones représentant les phonèmes extraits de chaque classe. La fenêtre de spécialisation propre à ces neurones a été estimée à 5 trames (voir figure.1)

5 Apprentissage

Nous avons lancé l'apprentissage de trois réseaux ayant les mêmes paramètres initiaux. Chacun d'eux est chargé de classifier trois phonèmes d'une même classe. Pendant toute la phase d'apprentissage, les taux ont été observés afin de spéculer sur les comportements des réseaux vis-à-vis de la reconnaissance.

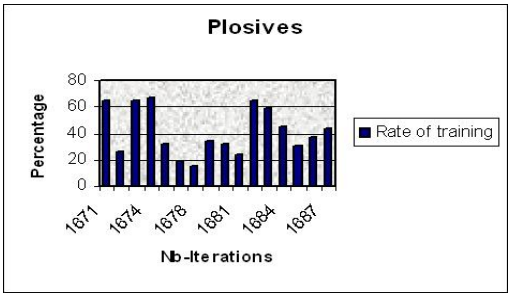

Figure 2. *Le taux d'apprentissage à travers un nombre d'itérations (plosives)*

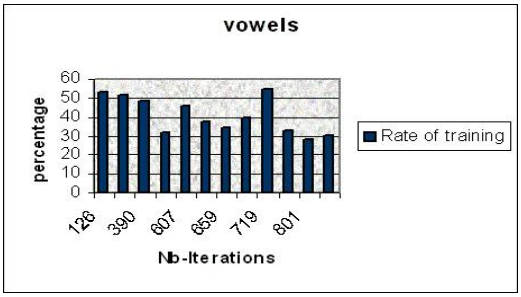

Figure 3. *Le taux d'apprentissage à travers un nombre d'itérations (voyelles)*

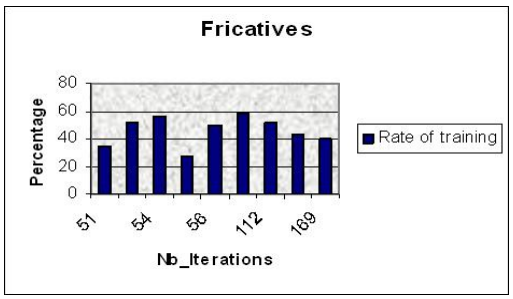

Figure 4. *Le taux d'apprentissage à travers un nombre d'itérations (fricatives)*

D'une façon générale, les fricatives sont les mieux reconnues dés les premières itérations, et un par un. Le taux d'apprentissage total peut atteindre jusqu'à 60%(voir Figure.4).

Les plosives nécessitent un grand nombre d'itérations pour arriver à une bonne reconnaissance (voir Figure.2).

Cependant, Les voyelles réagissent mal au réseau, où elles n'arrivent pas à reconnaître séparément les phonèmes (voir Figure.3), même si le taux d'apprentissage total atteint des fois 54%.

5 Tests et résultats

Les tests sont opérés par une simple propagation des occurrences des phonèmes de la base de test de TIMIT. À travers cette propagation, le calcul des activités des neurones de la couche cachée et ceux de la couche de sortie est automatiquement accompli. Pour la sélection du phonème reconnu, nous calculons les scores de la couche de sortie en sommant les activités des neurones appartenant au même phonème selon la direction temporelle (voir Figure.5).

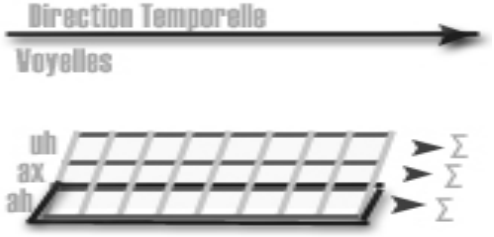

Figure 5. *Le principe du calcul de l'activation des neurones de la couche de sortie dans le test (voyelles)*

Le score le plus élevé correspondra donc au phonème reconnu.

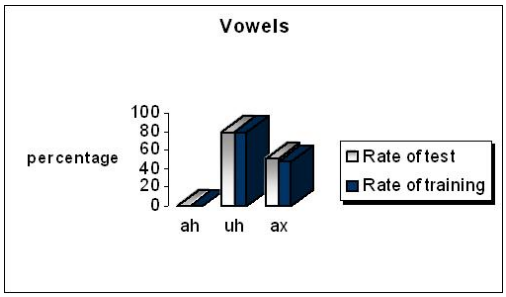

Figure 6. *Les taux de reconnaissance et d'apprentissage pour les voyelles à la 773 ième itération*

La Figure 6 montre qu'à un taux de reconnaissance global qui atteint 35.28%, le réseau est incapable de reconnaître /ah/, sachant que pour ce phonème le taux d'apprentissage a été quasiment nul tout au long de l'apprentissage. Ceci confirme l'échec de ce type d'architecture dans la classification de la catégorie des voyelles.

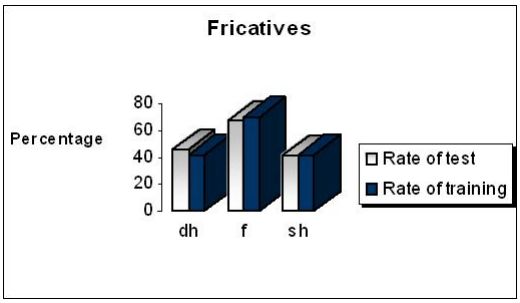

Figure 7. *Les taux de reconnaissance et d'apprentissage pour les fricatives à la 112 ième itération*

Les résultats du test sur les fricatives reflètent la capacité du réseau à gérer une bonne reconnaissance globale de cette classe et une meilleure répartition quant à la reconnaissance isolée par phonème.

Le réseau manifeste une reconnaissance relative pour les plosives. Le taux de reconnaissance global optimal correspond à un état de faiblesse du réseau à reconnaître indépendamment le /b/.

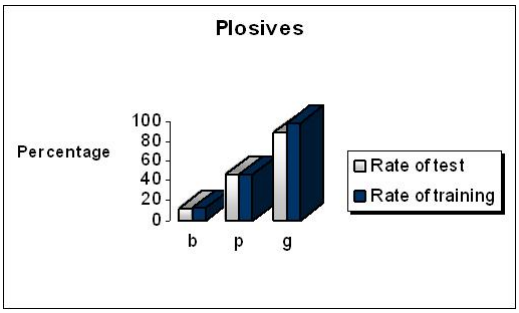

Figure 8. *Les taux de reconnaissance et d'apprentissage pour les plosives à la 1674 ième itération*

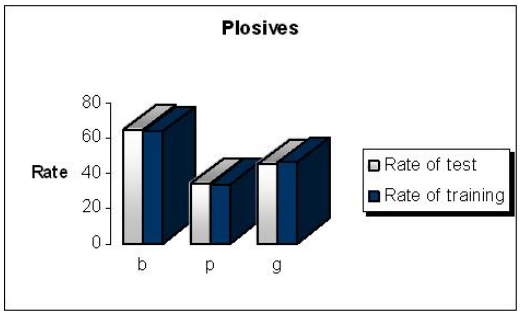

Figure 9. *Les taux de reconnaissance et d'apprentissage pour les plosives à la 1687 ième itération*

Cependant, il existe des itérations où le taux de reconnaissance global est moins bon, mais la répartition de la reconnaissance est bien établie (voir Figure 9).

6 Délais temporels

Arriver à des taux optimaux mène à des combinaisons des délais et poids reflétant les architectures finales optimales à la reconnaissance. A cause du grand nombre de neurones dans la couche cachée, il nous est impossible de représenter toute la configuration finale des retards des connexions reliant la couche d'entrée à la couche cachée. Sachant que c'est par le biais de ces connexions que s'effectue l'extraction des caractéristiques des phonèmes présentés à l'entrée du réseau. La figue 10 illustre l'initialisation de départ des retards des connexions des neurones de la couche cachée.

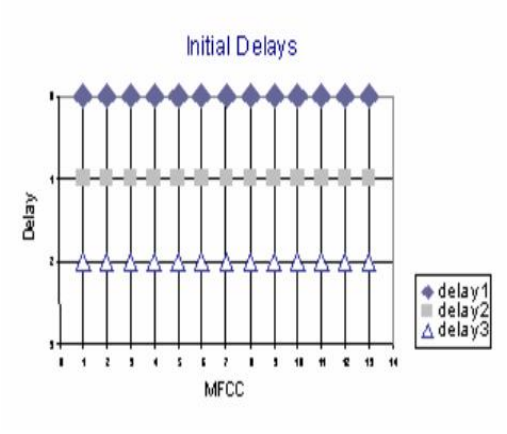

Figure 10. *Les retards initiaux des connexions des neurones de la couche cachée*

Les figures 11, 12 et 13 correspondent aux combinaisons des délais optimales du premier neurone de la couche cachée déduit après l'arrêt des apprentissages. Nos inductions se sont faites par l'adaptation de ses délais de connexions, vu que les autres neurones suivaient le même comportement d'ajustement observé par ce neurone durant tout l'apprentissage.

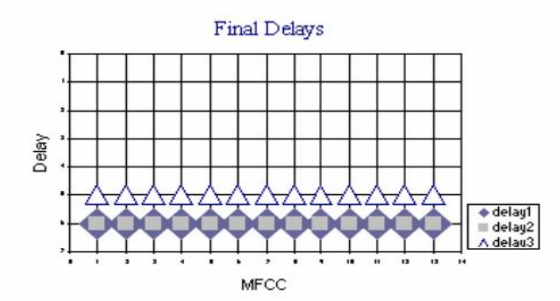

Figure 10. *Les retards finaux des connexions du 1er neurone de la couche cachée (voyelles)*

Pour les voyelles, on peut clairement constater que l'ajustement des délais ramène à un alignement dans l'extraction des coefficients MFCC, de telle sorte que tout le vecteur en entrée est extrait en sa totalité à l'instant $t_n - \tau_m$ (τ_m représente le retard de la connexion). Ce comportement a été observé durant toute la phase d'apprentissage, ce qui a permis au réseau d'avoir un certain comportement stable par rapport aux autres (fricatives et plosives).

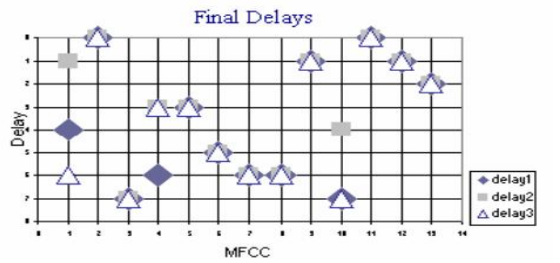

Figure 12. *Les retards finaux des connexions du 1er neurone de la couche cachée (fricatives)*

Contrairement aux voyelles, les fricatives eux présentent un comportement aléatoire dans l'extraction des coefficients MFCC, de telle sorte que les neurones de la couche cachée repartissent leurs visions, qui laisse l'extraction basée sur les caractéristiques du présent et entrées ultérieurs à l'instant de l'exaction.

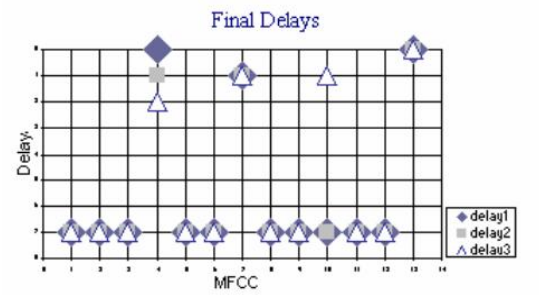

Figure 13. *Les retards finaux des connexions du 1er neurone de la couche cachée (plosives)*

Cette répartition est légèrement présente dans les plosives, où l'information est presque toujours extraite des vecteurs inférieurs, qui laissent à croire que les plosives n'ont besoin que des premiers vecteurs MFCC pour atteindre une bonne reconnaissance.

7 Conclusion et perspectives

Il est essentiel de savoir que l'utilisation des réseaux de neurones à délais temporels nécessite des moyens matériels considérables, surtout dans les problèmes complexes telle que la reconnaissance de la parole, où les bases sont plutôt de grandes tailles.

Dans un contexte de comparaison entre le TDNN et l'ATNN, nous classons ce travail comme un résultat préliminaire d'une application de l'ATNN dans la reconnaissance de la parole, renfermant une idée de base, qui est le rôle des délais temporels dans la gestion de l'extraction des caractéristiques. Le comportement de l'ajustement de ces derniers pendant des phases d'apprentissage a révélé des spécificités d'adaptation des délais concernant chaque catégorie de phonèmes de la base TIMIT.

Ce qui porte à croire que l'on pourra exploiter ce type d'architecture dans les classifications des phonèmes par catégorie selon de nouveaux critères ciblés sur une détection de la nature des phonèmes dans le domaine de la reconnaissance de la parole continue.

Références

- Y. Bennani, Approches Connexionnistes pour la Reconnaissance Automatique du Locuteur : Modélisation et Identification, *Thèse de Doctorat en Sciences- Spécialité informatique*, ORSAY, N° d'ordre : 1948, 1992.

- A. Benyettou, A. Mesbahi, H. Abdoune, A. Ait-ouali, La reconnaissance de formes spatio-temporelles par les réseaux de neurones à délais temporels, Conf. Nationale sur l'Ingénierie de l'Electronique –CNIE'02, pp. 159-163, univ. USTOran, Algérie, 15-16 Décembre 2002.

- A. Benyettou, T. Hamza, A. lotfi, K. Mezzoug, Classification phonétique à base de cartes auto-organisatrices temporelles, Conf. Nationale sur l'Ingénierie de l'Electronique –CNIE'02, pp. 164-168, univ. USTOran, Algérie, 15-16 Décembre 2002.

- D.T. Lin, The Adaptable Time Delay Neural Network Characterisation and Application to Pattern Recognition, Prediction and Signal Processing, *Thesis Report, ISR*, 1994.

- L. Mesbahi, A. Benyettou, A New Contribution towards a temporal radial Basis Function Applied to Mackey-Glass, *The 3rd Inter. Conf. On Neural Networks and Artificial Intelligence –ICNNAI'2003*, pp.72-77, Minsk, Belarus, 12-14 November, 2003.

- L. Mesbahi, A. Benyettou, A New Look to Adaptable Temporal Radial Basis Function Applied in Speech Recognition, *Journal of Computer Science 1(1): 1-6*, 2005.

- A. Ourdighi, Z. Ouksili, Etude et comparaison de deux architectures neuronales à délais temporels pour la classification phonétique, *Rapport de Recherche SIMPA-Parole, N°PAR15/04/SIMPA, juin 2004*, Dept. Informatique, Université USTOran- Algérie

- I. Poisson, Réseaux de neurones à convolution, Reconnaissance de l'écriture manuscrite non contrainte, Ecole polytechnique de l'université de Nantes, France, Equipe Image et Vidéo Communication, 2001.

- A. Weibel, T. Hanazawa, G. Hinton, K, Shinkano, Phoneme recognition using time-delay neural networks, IEEETrans. on ASSP, 1989.

- C. Wöhler C., J.K. Anlauf, A Time Delay Neural Network Algorithm For Real-Time Pedestrian Recognition, IEEE In. Conf. On Intelligent Vehicles, pp. 247-252, Stuttgard, 1998.

- D. Goldberg, *Algorithmes Génétiques*. Addison Wesley 1994.