NESSR : un système NeuroExpert pour la reconnaissance de la parole

NESSR : a Neural Expert System for Speech Recognition

Halima Bahi [1]

[1] Laboratoire LRI, Département d'Informatique, Université de Annaba, Algérie,
bahi@lri-annaba.net

Manuscrit reçu le 19 avril 2005

Résumé et mots clés

Les réseaux de neurones ont été utilisés dans une large gamme d'applications. En particulier, des résultats satisfaisants ont été observés dans le domaine de la reconnaissance de formes. Toutefois, dans le contexte de la reconnaissance de la parole, l'utilisation des réseaux de neurones est difficile vue l'absence du paramètre temps dans leur structure. D'un autre point de vue, dans une application de reconnaissance de la parole, un mot peut être reconnu et bien classé ou reconnu et mal classé, il est donc impératif de pouvoir expliquer le raisonnement qui a conduit à cette décision.

Dans cet article, nous considérons deux insuffisances des réseaux de neurones artificiels : le manque de connaissances explicites du domaine d'application et l'absence de l'aspect temps dans la structure des réseaux.

À cet effet, nous proposons un modèle de neurone temporel et spécialisé, que nous appelons STN (pour specilized temporal neuron). Ce modèle est ensuite utilisé comme élément de base dans un réseau neuro-symbolique pour la reconnaissance de la parole.

Intelligence artificielle, réseaux de neurones artificiels, reconnaissance de la parole.

Abstract and key words

Artificial neural networks (ANNs) have found applications in large spectrum of fields. Satisfactory results are obtained particularly in classification problems. In speech recognition context, the use of ANNs is hard, this is essentially due to the absence of the temporal aspect in their structure. On the other hand, assuming a speech recognition task, a word could be recognized and well categorized or recognized and badly categorized ; so the explanation of the decision is very important. In this paper, we address two limitations of ANNs : the lack of explicit knowledge and the absence of temporal aspect in their implementation. STN : is a model of a specialized temporal neuron, which includes both symbolic and temporal aspects. To illustrate the STN utility, we consider a system for speech recognition ; we underline in this paper the explanation aspect of the system.

Artificial intelligence, machine learning, connectionism, speech recognition.

1. Introduction

L'intelligence artificielle continue de connaître des développements importants dans le domaine de la modélisation des processus cognitifs. Un des axes intéressants de ces développements est l'orientation vers des approches hybrides, qui incorporent plusieurs paradigmes dans le même système. Parmi ces paradigmes l'intégration neurosymbolique constitue une voie principale de la complémentarité entre les deux approches : neuronale et symbolique [9], essayant ainsi de trouver des solutions aux inconvénients et limites de chacune d'entre elles et d'apporter des résultats satisfaisants à des problèmes complexes du monde réel.

En effet, Les réseaux connexionnistes sont de bons associateurs, ils apprennent à apparier deux vecteurs dans deux espaces différents et réalisent ainsi une fonction relativement complexe [3]. Cette fonctionnalité est tout à fait intéressante notamment pour le traitement de bas niveau et plus précisément pour le traitement sensoriel. Parallèlement, les modèles symboliques offrent un profil complémentaire car ils ont de grandes performances en raisonnement et en déduction et de bonnes qualités d'explication [9].

Dans un domaine d'application tel que la reconnaissance de la parole, l'utilisation d'une telle combinaison semble très intéressante, en particulier car la décision du système de reconnaissance quant à la reconnaissance d'un mot a souvent besoin d'être expliqué, ce qui n'est pas aisé en assumant les approches statistiques, l'introduction de la symbolique dans le système de reconnaissance ne peut qu'être bénéfique sur ce plan.

Dans ce travail, nous considérons un système expert implanté au travers d'un réseau de neurones, qui de ce fait en plus de sa puissance calculatoire, hérite de la richesse sémantique du système expert. Dans un tel système chaque neurone possède une signification symbolique, et la propagation de l'activation d'un niveau à un autre reproduit les règles de production [1]. Nous proposons ainsi, un modèle neuro-expert dédié à la reconnaissance de la parole : *NESSR* (acronyme de Neural Expert System for Speech Recognition). *NESSR* est un Perceptron multicouches où les neurones d'entrée représentent le niveau acoustique, les neurones cachés représentent le niveau phonétique et les neurones de sortie représentent le niveau lexical.

D'autre part, la reconnaissance de la parole est une application qui induit fortement la composante temporelle, il faudrait donc pourvoir ce système de la capacité de prendre en considération le paramètre temps. *STN* (pour specilized temporal neuron) : est un modèle de neurone temporel que nous proposons d'introduire dans *NESSR* afin de préserver sa sémantique et d'implanter l'aspect temps dans le réseau. De ce fait, la particularité de *NESSR*, est qu'en plus d'être une inspiration des systèmes neuro-experts existants dans la littérature, il intègre la dimension temporelle ce qui constitue la véritable nouveauté dans un tel système.

L'article est structuré comme suit, dans la deuxième section, nous donnons une brève introduction à la reconnaissance automatique de la parole et aux réseaux de neurones. Dans la section 3, nous donnons une vue générale du système de reconnaissance. Dans la section 4, nous décrivons le niveau acoustique, en section 5, le niveau phonétique et en section 6, le niveau lexical. En section 6, une discussion est faite sur la base des résultats obtenus. Finalement, une conclusion clôture l'article.

2. Reconnaissance automatique de la parole et réseaux de neurones

2.1. Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (RAP) est le processus qui à son entrée reçoit un signal vocal et à sa sortie le traduit sous une autre forme ; le plus souvent textuelle. Un système de reconnaissance de la parole comprend normalement deux étapes (figure 1), d'abord l'étape d'extraction de caractéristiques ensuite l'étape de reconnaissance (ou de classification). Le module d'extraction de caractéristique se charge de transformer le signal en entrée du système en une représentation interne de sorte qu'il soit possible de retrouver le signal original. Ce bloc est conçu en s'inspirant du modèle de perception de l'homme. La sortie de ce bloc est classée par le module de reconnaissance, qui en général intègre des séquences de phonèmes en des mots. En effet, ce processus qui consiste à faire correspondre à une occurrence son expression symbolique ou traduire un langage parlé en un langage écrit est appelé reconnaissance de la parole.

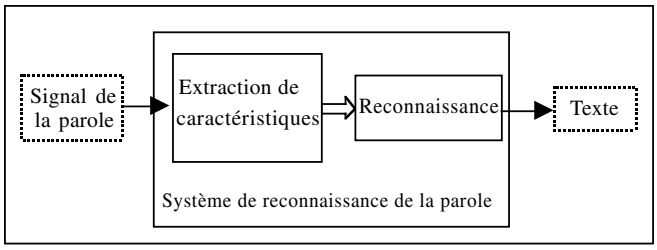

Figure 1. Blocs de base composant un système de RAP.

2.2. Les réseaux connexionnistes

Un réseau de neurones artificiel ou réseau connexionniste est fondamentalement une interconnexion dense d'éléments de calcul simples [6], non-linéaires appelés « neurones ». Il est supposé qu'un neurone possède N entrées, $x_1,x_2,..,x_N$ qui sont additionnées après avoir été pondérées par des poids w_1, w_2,... pour donner une sortie y définie par : $y = f(\sum w_i x_i - \Phi)$, où f est généralement une fonction sigmoide et Φ un seuil.

Le perceptron multicouches (MLP pour MultiLayer Perceptron) est un réseau de neurones qui se structure en couches où l'activation de la première couche est fixée à l'entrée du réseau et celle de la dernière est interprétée comme la réponse du réseau, entre les couches d'entrée et de sortie s'insère une ou plusieurs couches cachées. Dans ce type de réseau (non récurrent), la connectivité est restreinte : un neurone d'une couche inférieure ne peut être relié qu'à des neurones de couches suivantes (figure 2).

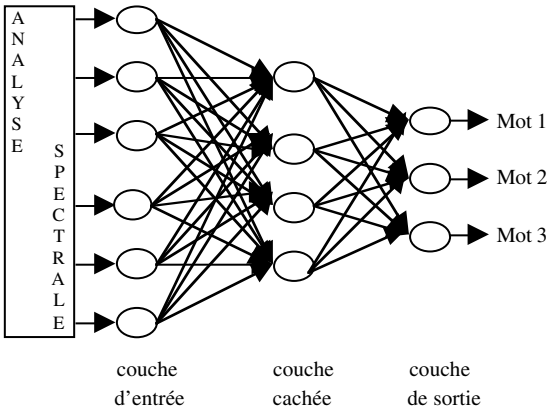

Figure 2. Structure d'un MLP à 3 couches pour la reconnaissance de mots.

2.3. Introduction des symboles dans les ANNs

Nous retiendrons les performances des ANNs en tant que classifieurs et leur grande adaptabilité [3]. Toutefois, on y relève quelques insuffisances inhérentes à leur mode d'apprentissage, à savoir :

• Les temps d'apprentissage sont importants,

• Les valeurs initiales des paramètres du réseau sont déterminants quant à l'issu de l'apprentissage,

• Le choix de la topologie du réseau relève toujours de l'empirique, bien que l'on sait que le nombre de neurones cachés influence énormément les performances du réseau,

• Après la phase d'apprentissage, le réseau connexionniste est utilisé comme une boite noire. Donc le raisonnement à partir duquel ont été obtenu les résultats est difficile à interpréter.

Une des possibilités pour dépasser ces limitations est d'utiliser les connaissances du domaine pour guider la construction du réseau et le pourvoir d'une sémantique qui permettrait d'expliquer le raisonnement mené pour arriver à une décision.

2.4. Les réseaux de neurones en RAP

Le diagramme de la figure 3 montre le modèle conceptuel du système de reconnaissance de la parole de l'être humain. Le signal acoustique en entrée est analysé par un « modèle auditif » qui fournit des informations spectrales du signal et les sauvegardent dans une mémoire sensorielle. Des informations sensorielles provenant d'autres sources (vision, toucher, ...) sont également présentes dans cette mémoire et servent à enrichir les différents niveaux de description du signal. L'analyse auditive est principalement basée sur le traitement acoustique de l'oreille. Ensuite, a lieu dans le cerveau une analyse de caractéristiques à différents niveaux. Quant aux mémoires à court et à long termes, elles offrent un contrôle externe au processus neuronal. Finalement, il est à remarquer que la configuration globale du modèle s'apparente à un réseau connexionniste « feed forward », et c'est en s'inspirant de ce schéma perceptuel du processus de reconnaissance que nous croyons que le connexionisme offre une alternative prometteuse pour la modélisation des tâches cognitives.

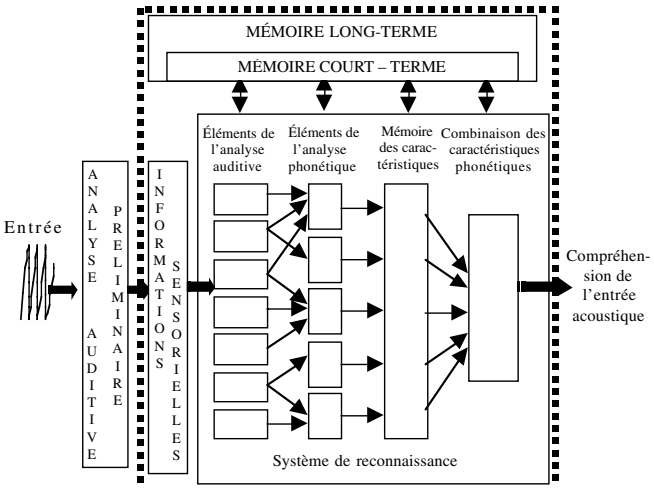

Figure 3. Diagramme conceptuel du système de reconnaissance de l'être humain (d'après [8]).

Les tous premiers essais en RAP effectuaient des tâches très simples telles que : classer des segments de parole en son voisé/non voisé, ou nasal/fricatif/plosif, en utilisant un Perceptron multicouches [10]. Le succès remporté par ces réseaux a énormément encouragé les chercheurs à considérer la classification des phonèmes ; ce qui fût effectué avec succès [7]. Les mêmes techniques rencontrèrent quelque succès pour la reconnaissance des mots, mais l'absence du paramètre temps dans de telles architectures a orienté les recherches vers d'autres alternatives telles que les réseaux multicouches à retard, ou *Time-Delay Neural Networks* (TDNN) [12]. Un TDNN est à la base un perceptron multicouches qui se singularise par le fait qu'il prend en compte une certaine notion du temps. C'est-à-dire qu'au lieu de prendre en compte tous les neurones de la couche d'entrée en même temps il va effectuer un balayage temporel. La couche d'entrée du TDNN prend une fenêtre du spectre et balaye l'empreinte. Il a été développé dans le but d'apprendre des structures spectrales spécifiques à l'intérieur de vecteurs de parole consécutifs.

Une taxonomie des systèmes connexionnistes à composante temporelle est présentée dans [5]. Mais on relèvera que ces différentes architectures sont assez compliquées à mettre en œuvre et que surtout elles n'intègrent pas une dimension symbolique du domaine d'application, c'est-à-dire les entités considérées et les relations qui existent entre-elles. D'autre part, un système expert connexionniste dédié à la parole est présenté dans [1], le modèle proposé reproduit la sémantique du domaine au travers des unités phonétiques et lexicales considérées, mais il n'inclut pas la dimension temporelle de l'application.

3. Architecture générale du modèle proposé

Le modèle de reconnaissance que nous proposons a la particularité d'intégrer dans un réseau connexionniste à la fois les connaissances du domaine ; ce qui induit une spécialisation des neurones et des connexions du réseau et une composante temporelle vu que cet aspect est indissociable de la parole. Par ailleurs, nous souhaitons souligner que ce modèle peut être adapté à d'autres applications en récupérant la structure du neurone temporel spécialisé.

3.1. Architecture générale

Le système de reconnaissance comprend trois composantes : une mémoire de reconnaissance, une mémoire court-terme et une mémoire long-terme.

• La mémoire de reconnaissance est le réseau *NESSR* qui constitue l'essentiel de notre proposition.

• La mémoire court-terme est une mémoire où sont sauvegardés des évènements temporaires qui peuvent survenir lors du raisonnement.

• La mémoire long-terme contient des informations de haut niveau qui permettent de valider une décision.

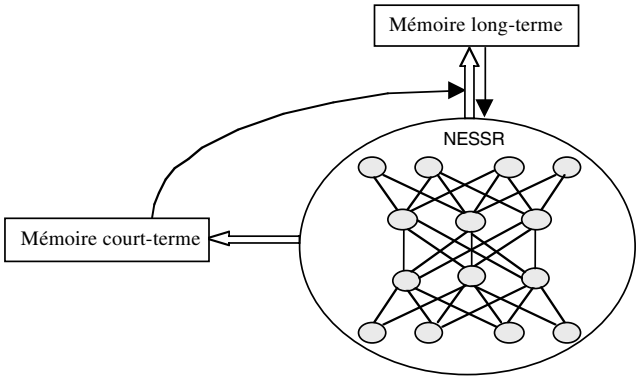

Figure 4. Architecture générale.

La mémoire de reconnaissance est agencée en trois couches : la couche d'entrée représente le niveau acoustique. La couche cachée représente le niveau phonétique, tandis que la couche de sortie représente le niveau lexical [1].

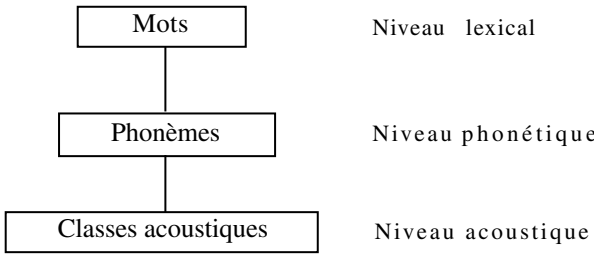

Figure 5. Architecture du réseau.

4. Le niveau acoustique

4.1. Des neurones spécialisés

Les neurones de la couche d'entrée se chargent de capter les particularités de la forme en entrée du réseau. Dans le cas de la reconnaissance de la parole, les cellules réceptrices détectent les caractéristiques du signal. Ces cellules appartiennent à un réseau symbolique ; où chaque cellule se spécialise dans la détection d'une caractéristique : nous les appellerons neurones-classe. Les caractéristiques sont déterminées en effectuant une quantification vectorielle sur tous les phones issus de l'ensemble d'apprentissage. Cette classification permet de dégager un ensemble réduit de vecteurs qu'on appelle prototypes qui représentent l'ensemble des particularités que peut présenter un signal [1]. Ces particularités n'ayant pas de signification physique particulière elles sont numérotées de 1 à n et de ce fait un neurone-classe sera appelé C_i en référence à la classe qu'il représente.

4.2. L'activation d'un neurone d'entrée

Un neurone-classe est activé si la caractéristique qui lui est associée est détectée dans le signal. À un instant t donné, un seul neurone-classe est activé. Ceci suppose que la présentation d'un signal au réseau dure de l'instant t_0 à l'instant t_n.

Ainsi dans cet intervalle de temps plusieurs neurones peuvent être activés. Considérons l'exemple ci-dessous (figure 6) où l'occurrence d'un mot, est présentée au système, le mot est d'abord analysé, on obtient un ensemble de fenêtres temporelles chacune d'entre-elles comprend un ensemble de coefficients cepstraux. Chaque fenêtre sera remplacée par son prototype, on obtient une chaîne de symboles, où chaque symbole représente une classe acoustique. À l'instant $t_0 = 1$, c'est la caractéristique acoustique C_{12} qui est détectée, ce qui induit

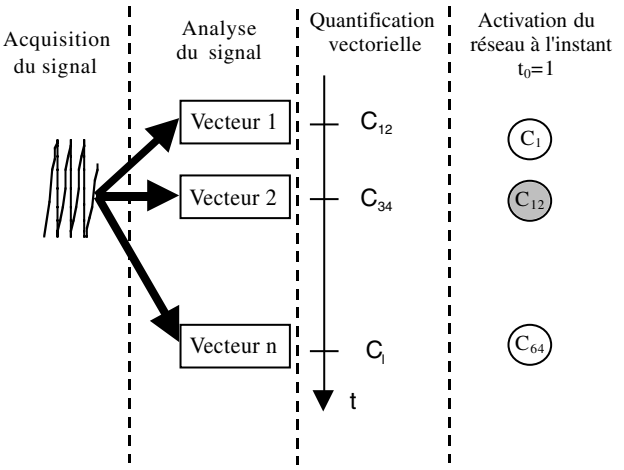

Figure 6. Présentation d'un signal au réseau.

une activation du neurone-classe correspondant, l'instant suivant c'est le 34e neurone etc. Notons que l'activation successive du même neurone est toute à fait prise en charge par le réseau (voir exemple §5. 3).

5. Le niveau phonétique

Les activations de la couche d'entrée sont transmises à la couche suivante dont le rôle est d'associer aux entrées acoustiques des unités phonétiques du langage ; en l'occurrence les phonèmes. À une certaine séquence de caractéristiques acoustiques détectée sera associé un phonème. De ce fait, la reconnaissance d'un phonème induit une segmentation implicite du signal à ce point de la structure.

5.1. Des neurones temporels spécialisés

À l'image des neurones d'entrée, ces neurones portent aussi une signification symbolique. Dans ce cas, chaque cellule représente un phonème de la langue arabe, nous les appellerons des neurones-phonème. Un phonème est définit dans ce cas par la détection d'un ensemble de caractéristiques. Ainsi, chaque fois qu'il y a corrélation entre l'apparition d'une caractéristique acoustique et la détection du phonème, une connexion entre la classe correspondante et le phonème concerné est initiée. Un neurone de ce niveau possède donc autant d'entées qu'il a de connexions initiées. La particularité que présente notre modèle de neurone est que chaque entrée est libellée au nom d'une caractéristique. De plus l'activation de ces entrées doit se faire dans un séquencement bien défini assuré par la structure du neurone, dans laquelle une entrée i ne peut être considérée que si l'entrée $i - 1$ est déjà pré-activée (figure 7).
Pour pouvoir dégager l'ensemble des classes acoustiques pertinentes pour la détection d'un phonème, on considère l'ensemble des classes obtenues après quantification vectorielle et on opère une étude de corrélation entre ces prototypes et l'ensemble des phonèmes. Ceci permet de dégager l'ensemble des caractéristiques pour chaque phonème. S'il y a une forte corrélation entre une caractéristique et un phonème nous considérons que c'est un constituant de base du phonème. Une fois les caractéristiques

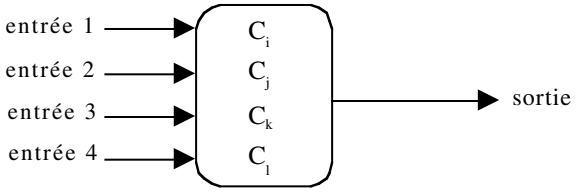

Figure 7. Le modèle STN.

d'un neurone dégagées leur séquencement est établit.

5.2. L'activation d'un *STN*

Au début d'une session de reconnaissance lorsqu'une caractéristique, soit C_i, est détectée le neurone-classe associé s'active et toutes les connexions qui en partent sont pré-activées, donc tous les neurones-phonème dont la première caractéristique est C_i se voient pré-activés. Ainsi, un neurone-phonème est mis en état pré-activé dès que sa première entrée est pré-active, ceci suppose que plusieurs neurones-phonème peuvent être pré-activés simultanément. Mais un neurone ne s'active que si toutes ses entrées sont activées. Concrètement, cela signifie que plus d'un phonème sont candidats à la reconnaissance mais au final le neurone gagnant prendra toutes les activations (winner takes all). En effet, lorsqu'un neurone-phonème s'active toutes les connexions provenant de la couche précédente sont désactivées, il en est de même pour tous les neurones concurrents, *i.e.* ceux qui étaient pré-activés simultanément. Si à un instant donné la détection d'une caractéristique peut provoquer l'activation de plus d'une cellule cible, une seule cellule est activée mais l'information est sauvegardée dans la mémoire à court terme au cas où il faudrait revenir sur ce choix. Remarquons que cette situation est très rare (eu égard au nombre de caractéristiques choisies soit 32), mais il est important de le souligner pour une éventuelle utilisation de ce modèle de neurone pour une application autre que la RAP.

5.3. Exemple illustratif

L'exemple de la figure 8, n'est pas un exemple réel, nous y avons illustré des situations particulières qui constituent les conditions limites du modèle, et justifient son utilisation dans une application induisant fortement le paramètre temps.
Soit le réseau du graphe ci-dessus auquel nous soumettons la séquence : $...C_1C_1C_5C_2C_3C_4...$, l'activation du réseau est trans-

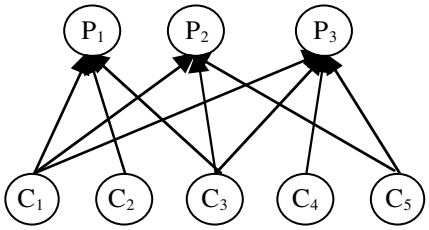

Figure 8. Exemple de connexions neurone-classe / neurone-phonème

Table 1. Exemple d'activation de neurones.

	Neurone-p_1	***Neurone-p_2***	***Neurone-p_3***
***T* = 0**	**C1** C2 C3	**C1** C3 C5	**C1** C5 C4 C3
***T* = 1**	**C1** C2 C3	**C1** C3 C5	**C1** **C5** C4 C3
***T* = 2**	**C1** C2 C3	**C1** C3 C5	**C1** **C5** C4 C3
***T* = 3**	**C1** **C2** C3	**C1** C3 C5	**C1** **C5** C4 C3
***T* = 4**	**C1** **C2** **C3**	**C1** **C3** C5	**C1** **C5** C4 C3

Ci : entrée activée
Ci : entrée non activée

crite dans le tableau suivant, notons que l'ordre reproduit dans le tableau indique l'ordre dans lequel doivent être détectées les caractéristiques.
À l'instant $T = 0$, la caractéristique C_1 est détectée donc le neurone-classe correspondant est activé, et tous les liens qui en partent sont pré-activés. Ceci induit dans le cas de notre exemple la pré-activation des trois neurones-phonème du réseau, car C_1 correspond à la première entrée pour tous ces neurones cibles.
À l'instant $T = 1$, la même caractéristique est détectée, mais cette deuxième (ou $n^{ème}$) activation successive n'entraîne aucun changement dans l'état du réseau. À l'instant $T = 2$, la caractéristique C_5 est détectée, ce qui permet la pré-activation de l'entrée C_5 du neurone-p_3, mais pas celle du neurone-p_2, car cette dernière ne peut l'être avant la pré-activation de C_3.
À l'instant $T = 3$, la caractéristique C_2 est détectée ce qui pré-active la deuxième entrée du neurone-p_2. À l'instant $T = 4$, la caractéristique C_3 est détectée ceci pré-active les entrées correspondantes dans les neurones cibles p_1 et p_2.
À cette dernière pré-activation, le neurone-p_1 voit toutes ses entrées pré-actives. À ce moment, il s'active. Ce qui désactive toutes les connexions ainsi que les autres neurones cibles.
Après C_3, la séquence en entrée du réseau est segmentée et l'activation du neurone-p_1 est propagée à la couche suivante. Une nouvelle session de reconnaissance de phonème débute par C_4.
D'autre part, la séquence en entrée jusqu'à la reconnaissance de p_1, soit $S_i = C_1C_1C_5C_2C_3$, est temporairement sauvegardée dans la mémoire court-terme sous l'étiquette p_1, elle y demeure jusqu'à la détection d'un mot.

5.4. Caractéristiques du modèle STN

La structure du neurone temporel spécialisé que nous proposons pour modéliser les phonèmes permet de boucler sur une caractéristique acoustique du signal, comme c'est la cas pour la caractéristique C_1 dans l'exemple précédent. Mais cette structure permet surtout l'insertion de caractéristiques moins répandues dans le phonème parmi les classes pertinentes. Si on se réfère à l'exemple précédent la caractéristique C_5 peut apparaître dans le neurone-p_1 mais elle n'est pas essentielle dans sa structure.

6. La couche de décision

À chaque fois qu'un phonème est reconnu, cette détection est propagée à la couche suivante. Les cellules de cette couche représentent la décision du réseau ; en l'occurrence les mots du vocabulaire. Ces cellules sont appelées neurones-mot. De ce fait, chaque neurone-mot a autant d'entrés qu'il y a de phonèmes qui le composent. À chaque fois qu'un neurone-phonème est effectivement activé, cette activation est transmise à la couche suivante.
L'activation du neurone-phonème provoque la pré-activation de tous les neurones-mot du réseau dont la première entrée est le phonème reconnu, et contrairement à la politique suivie dans le niveau précédent et illustré dans l'exemple du 5.3, ceci désactive définitivement les autres neurones du réseau. À la détection du deuxième phonème d'autres mots parmi les restants sont éliminés etc. Lorsqu'un mot est reconnu par ce système, nous pouvons retracer facilement l'historique du processus de reconnaissance en suivant le chemin inverse des cellules activées, les cellules ayant une symbolique, nous pouvons expliquer le résultat. Dans l'exemple de la figure 9, on considère quatre mots qui représentent la sortie du réseau. À la détection du phonème p_1,

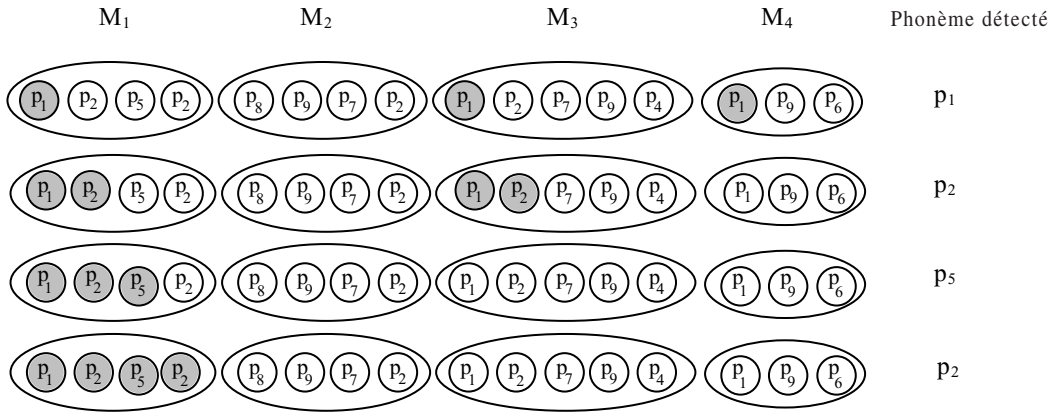

Figure 9. Exemple d'activation de neurones-mot.

les mots M_1, M_3, et M_4 sont pré-activés et le mot M_2 est complètement désactivé. À la détection du phonème p_2 le mot M_4 est également désactivé. Cette politique de sélection continue jusqu'à la reconnaissance d'un mot. Dans cet exemple la reconnaissance du mot M1 suppose la détection des phonèmes p_1, p_2, p_5, et p_2. Par ailleurs, la détection du phonème p_1, suppose la détection des caractéristiques inhérentes à ce phonème.

7. Résultats expérimentaux

Le plan d'évaluation de notre proposition inclut deux niveaux : un niveau phonème, pour pouvoir tester directement le modèle ***STN*** du neurone, et un niveau mot pour évaluer ***NESSR*** globalement. Dans le cas de la reconnaissance de mots, nous allons également utiliser d'autres techniques de reconnaissance et ce pour pouvoir comparer les performances de notre modèle par rapport à des techniques classiques qui ont fait leurs preuves, en l'occurrence le Perceptron multi-couches et les modèles de Markov cachés.

Pour effectuer nos tests nous utilisons d'abord une base de données qui comprend un ensemble de vingt cinq (25) mots regroupant tous les phonèmes de l'Arabe, parmi ces mots nous retrouvons les dix chiffres de l'Arabe. Chaque mot est prononcé huit fois par quinze (15) locuteurs différents.

7.1. Extraction des caractéristiques

Le signal acquis est échantillonné à une fréquence de 11025 Hz avec une précision de 8 bits. Le signal numérisé, est ensuite filtré de sorte que les hautes fréquences aient des amplitudes similaires à celles des basses fréquences [2].

Le signal obtenu est bloqué en une succession de trames de N échantillons chacune. Ces trames sont appelées fenêtres. Dans notre application nous choisissons des fenêtres de longueur N = 400 échantillons ce qui correspond à un intervalle de temps de 400/11.025 $\approx$ 36 ms, avec un recouvrement de 100 échantillons entre les trames. Le calcul de la transformée de Fourier sur ces fenêtres provoque la distorsion du spectre estimé du signal. Pour réduire ces effets, on utilise une fenêtre de pondération qui est souvent la fenêtre de Hamming en reconnaissance de la parole [2], aussi, chaque trame du signal est individuellement pondérée par une fenêtre de Hamming. Finalement, pour chaque fenêtre du signal nous calculons un ensemble de 13 coefficients MFCCs. Lorsque le spectre d'amplitude résulte d'une FFT (Fast Fourier Transform) sur le signal de parole prétraité, lissé par une suite de filtres triangulaires répartis selon l'échelle Mel, les coefficients obtenus sont appelés Mel Frequency Cepstral Coefficients (MFCC) [2].

7.2. Quantification vectorielle

La quantification vectorielle est une opération qui permet de regrouper des vecteurs proches au sens d'une distance (par exemple la distance euclidienne) dans la même région de l'espace et leur assigne un représentant qu'on appelle prototype. Ceci permet de réduire la dimension de l'espace de représentation. Concrètement, ces vecteurs proches se verront remplacés en phase de discrétisation par leur prototypes (voir figure 6). Les prototypes sont regroupés dans un dictionnaire (code-book) où chaque indice d'entrée au dictionnaire correspond à un prototype. Dans le cas de notre travail, les vecteurs initiaux sont ceux issus de l'analyse du signal et nous appellerons les prototypes des classes acoustiques. Finalement, un signal composé d'une suite de vecteurs acoustiques se voit transformé après quantification en une suite de symboles C_i où i représente l'entrée du dictionnaire qui correspond au prototype le plus proche du vecteur considéré.

7.3. Reconnaissance des phonèmes

Pour l'extraction de l'expertise sur les phonèmes nous considérons la base de données précédemment présentée où dix (10) parmi les locuteurs ont participé à l'étape d'apprentissage (lors de la définition des classes acoustiques et des caractéristiques d'un phonème). Pour effectuer les premiers tests d'évaluation nous formons deux groupes : le groupe TG1, inclut de nouvelles occurrences des locuteurs qui ont participé à l'apprentissage. Le groupe TG2, inclut des occurrences de nouveaux locuteurs, ainsi que des anciens locuteurs. Dans la table ci-dessous, nous mentionnons les taux de reconnaissance (en %) que nous avons obtenu pour les phonèmes considérés (les phonèmes sont données en notation API de l'association internationale de phonétique, avec [a], [u] et [i] qui représentent les voyelles de l'Arabe).

Table 2. Taux de reconnaissance de quelques phonèmes.

	[a]	[u]	[i]	[m]	[H]
TG1	99.2	98.4	98.2	97.6	97.1
TG2	97.7	97	97	95.8	95

Bien que la palette de mots choisie couvre l'ensemble des phonèmes arabes, des tests de caractérisation avancés ont particulièrement concerné les phonèmes impliqués dans la prononciation des chiffres arabes.

En outre, il est à remarquer que cette étape est la plus délicate et fastidieuse dans la construction du système ; elle correspond à la capitalisation des connaissances du domaine.

7.4. Reconnaissance des mots

Une fois le modèle ***STN*** testé par le biais de la reconnaissance de phonèmes, nous effectuons quelques tests au niveau reconnaissance de mots pour évaluer les performances du modèle ***NESSR***. Dans ce cadre, nous formons les groupes :

• Le groupe TG3 correspond aux occurrences des vingt cinq (25) mots prononcés par les dix (10) locuteurs qui ont participé à l'apprentissage.

• Le groupe TG4 correspond aux occurrences des vingt cinq (25) mots qui ont servi à l'apprentissage prononcé par les cinq (5) locuteurs qui n'ont pas participé à la phase d'apprentissage.

• Le groupe TG5 comprend huit (8) nouveaux mots prononcés par de nouveaux locuteurs.

Table 3. Taux de reconnaissance des quelques mots.

TG3	98.3 %
TG4	97.9 %
TG5	97.4 %

Ces premiers résultats, en particulier ceux relatifs à TG5, sont très encourageants car ils vont dans le sens d'une bonne généralisation du système.

7.5. Étude comparative

D'autres tests ont été effectués pour évaluer les performances du modèle ***NESSR*** comparativement aux autres approches en reconnaissance de la parole, en particulier le Perceptron multi-couches classique et les modèles de Markov cachés (HMMs).

Pour ces tests nous avons utilisé un Perceptron multi-couches avec un nombre de fenêtres statiques (le nombre de fenêtres choisi correspond à la plus longue des occurrences, on rajoute aléatoirement des zéros pour les autres occurrences), pour les HMMs nous avons utilisé des HMMs continus, avec autant d'états par mot qu'il y a de phonèmes, la fonction de densité relative à l'observation est formée par une mixture de cinq gaussiennes (pour plus de détails voir [2]). Le réseau *NESSR* est composé de 32 neurones-classe, 31 neurones-phonème et 10 neurones-mot (les dix chiffres).

En phase de test nous utilisons les groupes TG6 et TG7, qui correspondent aux groupes TG1 et TG2 précédemment décrits mais nous n'y considérons que les occurrences relatives aux dix chiffres de l'Arabe. Nous formons également le groupe TG8 qui compte des occurrences de nouveaux locuteurs. Dans la table 4, nous mentionnons les taux de reconnaissance obtenus avec les différentes implémentations. Le taux de reconnaissance (en %) représente le nombre de mots bien classés sur l'ensemble des mots à reconnaître.

L'étude comparative des résultats montre un avantage en faveur des autres approches, avantage qu'il est possible d'inverser en améliorant la caractérisation des phonèmes. D'autre part, un autre paramètre est à souligner, qui est le temps de calcul. Le temps de calcul en phase de reconnaissance est largement en faveur du système *NESSR*, vu la simplicité du traitement effectué comparativement au MLP et aux HMMs.

Table 4. Résultats comparatifs.

	MLP	HMM	*NESSR*
TG6	98.9	99.7	97
TG7	98	99.2	96.8
TG8	98	99	96.8

8. Conclusion

Dans ce papier nous avons présenté notre contribution qui s'inscrit aussi bien dans le domaine des systèmes neuro-symboliques que dans celui des modèles connexionnistes à composante temporelle. Notre suggestion consiste à combiner dans le même réseau les deux composantes, symbolique et temporelle, de par la proposition d'une nouvelle structure du neurone, que nous avons appelé le modèle ***STN***. Une utilisation de ce modèle est proposée dans la cadre de la reconnaissance de la parole. Nous retiendrons également que les modèles experts connexionnistes sont une tendance prometteuse dans la résolution de problèmes de la perception, car cette classe de problèmes implique à la fois les deux modes de raisonnement connexionniste et symbolique. Une perspective de ce travail est la reconnaissance de la parole continue où la segmentation se fera au niveau mot. La mémoire long-terme servira alors à valider une décision en se basant par exemple sur des règles grammaticales du langage, dans le cas où une décision est écartée, nous pourrons revenir sur le choix en faisant référence à la mémoire court-terme.

Plus généralement, le but dans la construction de ce système n'est pas d'augmenter le taux de reconnaissance déjà obtenu par le biais de méthodes statistiques, mais d'arpenter d'autres sentiers en RAP dont la finalité est d'apporter une explication au raisonnement mené lors de la reconnaissance.

Halima **Bahi**

Halima Bahi a reçu son doctorat d'état de l'université de Annaba (Algérie) dans la spécialité de l'intelligence artificielle en 2005. Elle est actuellement maître de conférence à l'université de Annaba et fait partie de l'équipe RADAR du laboratoire LRI. Ces recherches sont principalement en reconnaissance de la parole avec un intérêt particulier pour les modèles hybrides et la fouille de données audio.

Références

[1] H. BAHI, M. SELLAMI, Système expert connexionniste pour la reconnaissance de la parole, *proceedings de RFIA*, Vol 2, pp. 659-665, Toulouse, France, 2004.

[2] C. BECCHITTI, L. p. RICOTTI, Speech recognition : theory and C++ implementation, John Wiley, Angleterre, 1999.

[3] C. M. BISHOP, Neural networks for pattern recognition, Clarendon Press, Oxford, 1995.

[4] H. BOURLARD, N. MORGAN, Hybrid HMM/ANN systems for speech recognition : Overview and new research directions, Lecture Notes In Computer Science ; Vol. 1387, pp. 389-417, Springer-Verlag, London, UK, 1997.

[5] S. DURAND, TOM, une architecture connexionniste de traitement de séquences. Application à la reconnaissance de la parole. PhD thesis, Université Henri Poincaré, Nancy I, 1995.

[6] R. HECHT-NIELSON, Neurocomputing, Addisson-Wesley Publishing Company, 1989.

[7] W. M. HUANG, R. LIPPMANN, Neural nets and traditional classifiers, Neural information processing systems, Ed. Anderson D., pp. 387-396, New-York, 1988.

[8] L. RABINER, B. HWANG, Fundamentals of speech recognition, Prentice Hall, 1993.

[9] R. SUN, F. ALEXANDRE, Connectionist-Symbolic Integration : From Unified to Hybrid Approaches, Lawrence Erlbaum Associates, 1997.

[10] J. TEBELSKI, Speech recognition using neural networks, PhD thesis, Carnegie Mellon University, 1995.

[11] E. TRENTIN, E. GORI, A survey of hybrid ANN/HMM models for automatic speech recognition, Neurocomputing 37, Ed. Elsevier, pp. 91-126, 2001.

[12] A. WAIBEL, T. HANAZAWA, G. HINTON, K. SHIOKANO, K. J. LANG, Phoneme recognition using time-delay neural networks, IEEE Trans. On acoustics, speech, and signal processing, 37(3), 1989.