REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار - عنابة

FACULTE DES SCIENCES DE L'INGENIEUR
DEPARTEMENT D'INFORMATIQUE

Thèse présentée par

Halima ABIDET-BAHI

En vue de l'obtention d'un Doctorat d'Etat en Informatique

NESSR : Un système neuro-expert pour la reconnaissance de la parole

(Neural Expert System for Speech Recognition)

jury :

Pr. M. C. Batouche	***Université de Constantine***	***Président***
Pr. M. Sellami	***Université de Annaba***	***Rapporteur***
Pr. A. Benyettou	***Université de Oran***	***Examinateur***
Pr. N. Doghmane	***Université de Annaba***	***Examinateur***
Pr. K. Smaili	***Université de Nancy 2***	***Examinateur***
Dr. Z. Zemirli	***INI-Alger OuedSmar***	***Examinateur***

Année 2005

Dédicace

Cette thèse est dédiée à la mémoire de mon père

Remerciements

Tout d'abord, je voudrais remercier Pr. *Mokhtar Sellami*, mon directeur de thèse, pour m'avoir accueillie au sein de son équipe et pour m'avoir encadrée dans ce travail de thèse. Je le remercie également pour m'avoir laissé une grande liberté tout en me faisant profiter de son expérience, et ses conseils m'ont permis d'améliorer de nombreux points de ce manuscrit.

Je voudrais également remercier tous les membres de mon jury, c'est à dire Pr. *Batouche*, Pr. *Benyettou*, Pr. *Doghmane*, Pr. *Smaili* et Dr. *Zemirli*.

Je remercie encore une fois, Pr. *Benyettou* du département d'informatique à l'université d'Oran et Pr. *Doghmane* du département d'électronique à l'université de Annaba pour leur gentillesse et leur disponibilité.

Je remercie Pr. *Smaili* du laboratoire LORIA à l'université de Nancy pour son aide précieuse.

Je remercie Dr. *Zemirli* de l'institut national d'informatique pour avoir accepté de participer à mon jury de soutenance malgré son emploi de temps chargé.

Je remercie aussi Pr. *Batouche* du département d'informatique à l'université de Constantine pour m'avoir fait l'honneur de présider mon jury de soutenance.

Je souhaite également remercier toutes les personnes que j'ai pu côtoyer tout au long de ces années de thèse. Plus particulièrement, les membres du groupe LRI : *Habiba*, *Hassina*, *Labiba*, *Abdallah*, *Nadir*, *Tarek* et *Toufik* et les membres du groupe GRIA : *Karima*, *Lynda*, *Nabila*, *Sofia*, *Yamina*, et *Adel*.

Enfin, je souhaite remercier mes parents qui m'ont toujours poussée dans mes études. Je remercie aussi ma sœur *Farida* et mon époux *Mourad* pour m'avoir aidée et soutenue pendant ces années de thèse. Je remercie également, mes enfants *Rabie* et *Rayenne* pour avoir été un grand stimulant pour moi.

Table des matières

Remerciements

Dédicace

Introduction 1

1. Introduction générale 3

2. Position du problème 4

2. 1. La reconnaissance de la parole 4

2. 2. Le connexionisme 5

3. Problématique 5

4. Objectifs 6

5. Plan de la thèse 7

Partie I : Contexte d'étude et état de l'art

Chapitre I : Contexte d'étude 11

Présentation du chapitre 13

1. La communication 13

1. 1. La communication entre humains 13

1. 2. La communication homme-machine 14

2. Le signal de la parole 15

2. 1. Le signal de la parole 15

2. 2. Caractéristiques du signal de la parole 15

2. 3. Mécanismes de phonation et sons de la parole 15

3. La langue arabe 18

3. 1. Présentation 18

3. 2. Le système d'écriture 18

3. 3. La phonologie 19
3. 4. Caractéristiques phonétiques des phonèmes arabes 21
3. 4. 1. Lieux d'articulation 21
3. 4. 2. Traits distinctifs des phonèmes arabes 21
3. 4. 3. Autres particularités 22
4. Introduction à la reconnaissance automatique de la parole 23

Chapitre II : Reconnaissance automatique de la parole 25
Présentation du chapitre 27
1. La reconnaissance de la parole 28
1. 1. Introduction à la RAP 28
1. 2. Concepts de base 28
1. 3. Quelques systèmes de RAP 29
1. 4. Reconnaissance de la parole Arabe 29
1. 4. 1. Problèmes rencontrés en reconnaissance de l'Arabe 29
1. 4. 2. Travaux antérieurs 30
2. L'analyse du signal 30
2. 1. L'échantillonnage 30
2. 2. Le fenêtrage 31
2. 3. Extraction des caractéristiques 32
2. 3. 1. Méthodes temporelles 32
2. 3. 2. Méthodes fréquentielles ou spectrales 36
2. 3. 3. Méthodes cepstrales 37
3. Les approches de reconnaissance de la parole 39
3. 1. L'approche acoustico-phonétique 39
3. 2. L'approche reconnaissance de formes 40
3. 3. L'approche intelligence artificielle 42
4. Conclusion : Vers une combinaison de méthodes 42

Partie 2 : Outils utilisés

Chapitre III : Les réseaux de neurones en RAP 47
Présentation du chapitre 49

1. Un peu d'historique 49
2. Fondements des réseaux connexionnistes 52
2. 1. Le neurone formel 52
2. 1. 1. Le modèle de McCulloch et Pitts 52
2. 1. 2. Le modèle général 53
2. 2. Les connexions 55
2. 3. Topologies des réseaux connexionnistes 55
2. 4. Taxonomie des réseaux connexionnistes 56
3. Les mécanismes d'apprentissage 57
3. 1. L'apprentissage 57
3. 2. La règle de Hebb (1949) 58
3. 3. La règle de Widrow-Hoff 59
3. 4. L'algorithme de rétropropagation 59
4. Le Perceptron multicouches 59
4. 1. Le perceptron originel 59
4. 2. Le perceptron multicouches (MLP : MultiLayer Perceptron) 60
4. 2. 1. Structure du réseau 61
4. 2. 2. L'apprentissage 61
5. Les réseaux connexionnistes en RAP 62
5. 1. Les réseaux connexionnistes et le temps 62
5. 2. L'approche statique 63
5. 3. L'approche dynamique 64
5. 3. 1. Modèles à représentation externe 65
5. 3. 2. Modèle à représentation interne implicite 65
5. 3. 3. Modèle à représentation interne explicite 66
6. Quelques architectures de réseaux connexionnistes utilisées en RAP 66
6. 1. La carte auto-organisatrice de Kohonen 66
6. 2. Le TDNN (Time Delay Neural Network) 68
6. 2. 1. La structure du réseau 68
6. 2. 2. Le fonctionnement 69
6. 2. 3. L'apprentissage 70
6. 2. 4. Le TDNN et la reconnaissance de la parole 70
6. 3. Les réseaux récurrents 71

6. 3. 1. Le modèle de Jordan 71
6. 3. 2. Le modèle de Elman 72
6. 3. 3. Les réseaux récurrents et la parole 73
7. Conclusion : Vers des systèmes hybrides 74

Chapitre IV : Les modèles neurosymboliques 75
Présentation du chapitre 77
1. Introduction 78
2. Les deux paradigmes 78
2. 1. L'IA symbolique 78
2. 1. 1. Représentation et recherche 79
2. 1. 2. Les systèmes experts 79
2. 1. 3. L'apprentissage symbolique 81
2. 1. 4. Avantages et inconvénients 81
2. 2. L'IA connexionniste 81
2. 2. 1. L'apprentissage connexionniste 82
2. 2. 2. La représentation des connaissances 82
2. 2. 3. Avantages et inconvénients 82
2. 3. L'intégration des réseaux de neurones et des systèmes experts 82
3. Les systèmes neurosymboliques 83
3. 1. Les systèmes hybrides intelligents 83
3. 2. Les systèmes neurosymboliques 84
4. Taxonomie des systèmes neurosymboliques 85
4. 1. Les modèles combinés 85
4. 2. Les modèles transformationnels 86
4. 3. Les modèles couplés 86
5. Les systèmes experts connexionnistes 87
5. 1. Introduction 87
5. 2. Principe 88
5. 3. Exemple d'un système expert connexionniste 88
5. 4. L'algorithme Pocket 90
5. 5. EXPSYS : un autre exemple de CES 90
5. 6. Conclusion sur l'approche 91

6. L'approche KBANN 91
6. 1. Introduction 91
6. 2. Construction du réseau 92
6. 3. Conclusion sur l'approche 93
7. Conclusion 93

PARTIE III : Modèle proposé

Chapitre V : *NESSR,* Un système neuro - expert pour la reconnaissance de la parole 97
Présentation du chapitre 99

Partie A: Ancrage des symboles dans une architecture connexionniste
1. Introduction 100
2. Un modèle conceptuel pour la compréhension de la parole 101
3. De la connaissance au réseau de neurones 102
3. 1. Ancrage des symboles dans le réseau connexionniste 102
3. 2. Les neurones d'entrée 104
3. 2. 1. Les traits acoustiques 104
3. 2. 2. La quantification vectorielle 104
3. 3. La syllabe: l'unité de la décision 105
3. 4. Les relations de dépendances 105
4. Exemple d'un système expert connexionniste pour la RAP 106
5. Un KBANN pour la RAP 108
5.1. Les propositions 108
5. 2. Les clauses de Horn 108
5. 3. La structure du réseau 108
6. Conclusion 110

Partie B: Proposition d'un neurone temporel spécialisé, Application à la reconnaissance de la parole
1. Introduction 111

2. Description générale du réseau 111
2. 1. Introduction au modèle 111
2. 2. Architecture générale du système 112
2. 3. Ancrage des symboles dans le réseau 113
3. Le modèle du neurone temporel spécialisé 113
3. 1. Motivations 113
3. 2. Structure des neurones *STN* 114
3. 3. Activation du *STN* 115
4. La couche sensorielle : le niveau acoustique 115
4. 1. Structure des neurones : des neurones spécialisés 115
4. 2. Détermination des classes acoustiques 116
4. 3. L'activation d'un neurone 116
5. La couche d'association : le niveau phonétique 117
5. 1. Structure des neurones : des neurones temporels spécialisés 117
5. 2. Les connexions 118
5. 2. 1. La caractérisation d'un phonème 118
5. 3. L'activation d'un neurone-phonème 119
5. 3. 1. La pré-activation d'un neurone 119
5. 3. 2. L'activation d'un neurone 119
5. 3. 3. Exemple illustratif d'activation 120
5. 4. Caractéristiques du modèle des neurones-phonème. 122
6. La couche de décision 122
6. 1. Structure des neurones 122
6. 2. Activation 122
6. 3. La reconnaissance 123
7. Conclusion 123

Chapitre VI : Evaluation du modèle 125
Présentation du chapitre 127
1. Introduction 127
2. Extraction des caractéristiques 128
2. 1. Echantillonnage 128
2. 2. Isolement du mot 128

2. 3. Pré-accentuation 129
2. 4. Fenêtrage 129
2. 5. Fenêtrage de Hamming 130
2. 6. Analyse MFCC 130
3. La quantification vectorielle 130
3. 1. Définition 130
3. 2. Etablissement des classes par la méthode de LLOYD généralisée 131
4. La reconnaissance de phonèmes 132
4. 1. La base de données 132
4. 2. Les résultats 133
5. La reconnaissance de mots 134
5. 1. Reconnaissance en mode monolocuteur 134
5. 2. Reconnaissance en mode multilocuteurs 134
5. 3. Etude comparative 135
6. Conclusion 136

Chapitre VII : Application à la détection de la dyslexie 137
Présentation du chapitre 139
1. Introduction 140
2. La dyslexie : la mal-lecture 141
3. DEDY : un système de détection de la dyslexie 142
3.1. Présentation générale 142
3.2. Batterie de test 142
3.3. Principe du test de lecture 143
3.4. Profil social 143
4. Module de reconnaissance 144
4.1. Le classifieur ***NESSR*** 144
4. 2. Le processus de reconnaissance 145
5. Module de décision 146
5.1. Structure du cas 146
5.2. Recherche de cas similaires 146
6. Résultats 148

Conclusion et Perspectives 151

1. Bilan 153
2. Perspectives du système ***NESSR*** 155
3. Perspectives du modèle ***STN*** 155
4. Perspectives d'amélioration 155

Références bibliographiques 157

Introduction

Introduction

1. Introduction générale

Aussi loin que remonte l'histoire des hommes ce sont leurs rêves et fictions qui ont porté leurs plus grandes découvertes. Créer une machine dotée de facultés sensorielles et motrices similaires à celles de l'homme, et en particulier une machine qui saurait comprendre et interagir avec un dialogue est un vieux rêve de l'humanité. Mais bien que le rêve que font les humains en ce qui est de machines qui les comprennent est relativement ancien, nous ne disposons pas aujourd'hui encore de telles machines qui puissent fonctionner dans des environnements « normaux ».

C'est cette fiction et bien d'autres qui annoncèrent l'avènement de l'intelligence artificielle, et qui portent ceux à quoi aspire cette discipline. En effet, l'intelligence artificielle est une orientation de la recherche qui tente de reproduire par des automates des facultés de l'homme. Si l'on considère une tâche cognitive telle que la reconnaissance de la parole, certains pensent (et nous en faisons partie) que l'une des voies à suivre pour doter un automate de cette faculté est de mimer les mécanismes utilisés chez l'homme à cette fin. A cet effet, il a été établi de longue date, que pour mémoriser ses connaissances ou mener un raisonnement, l'homme utilise des réseaux neuronaux (nous appelons neuronal ce qui est biologique, et connexionniste ce qui correspond à cette qualification pour les automates) ; d'autre part, ce qui fait la richesse du raisonnement humain est sa symbolique et c'est ainsi donc que dès son avènement l'intelligence artificielle se trouve à balancer entre les deux paradigmes majeurs que sont : l'IA symbolique et l'IA connexionniste [Hilario, 95].

Aujourd'hui encore, l'intelligence artificielle continue de connaître des développements importants dans le domaine de la modélisation des processus cognitifs, et un des axes intéressants de ces développements est l'orientation vers des approches

hybrides qui incorporent plusieurs paradigmes dans le même système [Sun, 97]. Parmi ces paradigmes l'intégration neurosymbolique constitue une voie principale de la complémentarité entre les deux approches connexionniste et symbolique [Boz, 95], 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.

2. Position du problème

Nous nous situons dans ce travail à la croisée de trois axes de recherche que sont les modèles connexionnistes, l'intégration neurosymbolique et la reconnaissance automatique de la parole.

La reconnaissance de la parole est un domaine de recherche qui inclut plusieurs approches et techniques, et bien que la technologie des modèles de Markov cachés (HMM pour Hidden Markov models) s'impose comme l'outil de prédilection dans cette application, les modèles connexionnistes les ont largement précédé, mais ces derniers ont été délaissés en faveur des modèles de Markov, en raison de leur aptitude à mieux prendre en charge l'aspect dynamique de la parole. Mais aujourd'hui encore les capacités discriminantes des modèles connexionnistes s'imposent et des travaux récents s'orientent vers des hybridations HMMs/ANNs (ANNs pour Artificial Neural Networks) [Trentin, 01]. Quant à nous, nous croyons qu'en dépit de la puissance indéniable qu'ont les HMMs à modéliser le signal de la parole, les modèles connexionnistes demeurent une perspective réelle, dans des applications cognitives en particulier si on peut les doter d'une sémantique qui fait largement défaut aux HMMs.

2. 1. La reconnaissance de la parole

La reconnaissance automatique de la parole (RAP) est le processus par lequel la machine tente de « décoder » le signal de la parole qui lui est destiné. Les recherches relatives à la RAP débutèrent dans les années 1950, dans une conjoncture optimiste, car on pensait que les avancées technologiques des ordinateurs rendraient la RAP une tâche aisée. Quelques dizaines d'années plus tard, on se rendait compte que c'était faux, et que la RAP, demeure un problème difficile. Aujourd'hui encore nombre de questions restent posées, les difficultés majeures étant associées à la taille du vocabulaire à

reconnaître, la reconnaissance de la parole continue, à la reconnaissance indépendamment du locuteur, la parole spontanée, …

C'est aussi, une discipline qui prend de plus en plus d'ampleur et dont les applications sont aussi nombreuses que diversifiées.

2. 2. Le connexionisme

Le connexionisme a connu trois grandes époques depuis sa création: l'ère de l'engouement à partir de 1958, date de la création du perceptron monocouche où l'on pensait à tort que les réseaux connexionnistes allaient pouvoir résoudre d'innombrables problèmes, l'ère de la désillusion suite au coup porté en 1969 par Minsky et Pappert ; ils démontrèrent que le perceptron monocouche ne pouvait réaliser une fonction aussi simple que le OU exclusif [Bishop, 95], puis l'ère du renouveau suite à la mise au point de la procédure d'adaptation des poids dans le perceptron multicouches.

Ces dernières années on ne relève pas réellement d'événement fondamental lançant le connexionisme sur une nouvelle voie, mais on constate une réelle volonté de dépasser les limitations actuelles du connexionisme. Aussi, de nouveaux types de systèmes voient le jour, inspirés de la neurobiologie, de la psychologie ou mixant des techniques connexionnistes avec d'autres symboliques ou stochastiques (modèles hybrides). Dans ce sens la question de savoir si le connexionisme est à un tournant de son existence doit être posée. Dans tous les cas, l'ère du perceptron multicouches devra évoluer pour dépasser la « simple » classification de formes. Quels sont alors les points sur lesquels nous devons nous focaliser pour dépasser les défauts actuels ?

3. Problématique

Les réseaux connexionnistes sont des associateurs, ils apprennent à apparier deux vecteurs dans deux espaces différents. Ils réalisent ainsi une fonction relativement complexe. Cette fonctionnalité est toute à fait intéressante notamment pour le traitement de bas niveau et plus précisément pour le traitement sensoriel. Les réseaux sont ainsi vus comme des boites noires et hermétiques pour lesquelles l'interprétation des constituants adaptables est difficile voire impossible. Il est claire que l'architecture et la philosophie même des structures telles que le perceptron ne vont pas dans le sens d'une

interprétation possible des résultats. Pour pouvoir traiter des problèmes de plus haut niveau que ceux de la perception, nous devons envisager des réseaux dont le fonctionnement et l'architecture soient capables « d'expliquer » leur démarche.

Ensuite, parce que le temps est une dimension très importante dans la majeure partie des problèmes que nous nous posons, il est nécessaire de prendre en compte ce paramètre. Des modèles connexionnistes temporels ont été développés, certains tel que le TDANN (Time Delay Artificial Neural Network) ont même égalé les performances des HMMs en reconnaissance de la parole [Waibel, 89]. Presque tous s'occupent de l'aspect séquentiel mais ils souffrent encore de leur manque de tolérance en particulier pour la distorsion. Qui plus est, très peu prennent en compte la notion de durée [Durand, 95].

Le traitement symbolique n'est d'autre part pas traité par la majeure partie des modèles connexionnistes. Ceux-ci sont en effet des systèmes numériques et ne sont donc pas, par nature, adaptés au traitement symbolique. Pourtant cette notion semble importante pour des tâches de haut niveau, elle constitue en particulier la base des modèles de l'IA classiques et des systèmes logiques.

Les performances du cerveau sont pour beaucoup le produit de son passé, et si l'on veut mimer son intelligence on doit s'inspirer de cela. En l'occurrence l'information doit en premier lieu pouvoir être stockée. Nous devons en outre, pouvoir récupérer les relations temporelles qui existent entre les différents éléments composant l'information.

4. Objectifs

Au delà des objectifs qui ressortent de la problématique de la thèse et qui rentrent dans une thématique très précise qui est celle des systèmes neurosymboliques, avec une application dans un domaine où le paramètre temps est omniprésent. D'autres objectifs ce sont imposés à nous dès le début de notre travail de thèse. Il s'agit d'abord de prospecter les différentes approches utilisées en reconnaissance automatique de la parole, ce qui devrait nous permettrait de justifier notre proposition. D'autre part, il s'agit d'étudier de plus près la parole arabe au travers de ses particularités acoustiques et structurelles. Ceci, permet d'enrichir les travaux sur la parole arabe qui sur le plan littérature sont très peu nombreux.

5. Plan de la thèse

Ce manuscrit est composé de trois parties. La première partie introduit le contexte de notre étude à savoir la parole arabe et un état de l'art de la reconnaissance de la parole. Dans la deuxième partie, nous présentons les outils que nous avons utilisé pour construire notre modèle. La troisième partie inclut la présentation de notre proposition ainsi que son évaluation. Une conclusion et des perspectives clôturent ce travail.

Partie I: Contexte de l'étude et état de l'art

Dans le *chapitre 1,* nous présentons le contexte de notre travail, et ce au travers de quelques rappels et notions sur la parole, qui représente notre outil de base.

Nous allons également présenter dans ce chapitre, la langue arabe qui est le cadre de nos expérimentations, et nous aborderons les problèmes spécifiques pour sa reconnaissance automatique. Et pour clore le chapitre, une introduction à la reconnaissance de la parole est faite.

Le *chapitre 2* présente un état de l'art de la reconnaissance de la parole. Il débute par la description du signal de la parole, et des paramètres que l'on en extrait. Il se poursuit par la description des trois grandes approches de reconnaissance de la parole. A la fin du chapitre, nous situons notre proposition parmi les approches existantes en RAP.

Partie II: Outils utilisés

Dans le *chapitre 3*, nous présentons une revue des réseaux connexionnistes. Après un bref historique en guise de toile de fond, nous allons revoir certains concepts fondamentaux, nous décrirons quelques architectures des réseaux connexionnistes ainsi que les procédures d'apprentissage. Dans un second volet de ce chapitre, nous aborderons les possibilités d'utilisation des réseaux connexionnistes dans le cadre de la reconnaissance de la parole, et en particulier leur prise en compte du paramètre temps.

Le *chapitre 4*, est consacré à un état de l'art des systèmes neurosymboliques. Nous y présentons d'abord les deux paradigmes de l'intelligence artificielle que sont: le symbolique et le connexionisme. Nous y proposons également une taxonomie des

systèmes neurosymboliques, ensuite nous décrivons quelques modèles avant-gardistes dans la littérature.

Partie III: Modèle proposé

Le *chapitre 5* se compose de deux parties. Dans la première partie du chapitre nous allons présenter les éléments conceptuels d'un système connexionniste expert dédié à la reconnaissance de la parole. Dans la seconde partie de ce chapitre, nous présentons le modèle d'un neurone symbolique temporel (***STN***), ce modèle représente notre apport par rapport aux modèles hybrides existants dans la littérature en particulier la conception d'un réseau connexionniste spécialisé qui implémente la notion du temps. Le système obtenu est appelé ***NESSR*** pour Neural Expert System for Speech Recognition.

Dans le *chapitre 6*, nous allons décrire les principaux choix d'implémentation que nous avons effectué, pour mettre en œuvre notre modèle, mais nous allons également y présenter des réalisations de systèmes de reconnaissance de l'Arabe que nous avons fait en utilisant des techniques classiques (MLP et HMMs) et ce, pour pouvoir évaluer les performances de notre modèle par rapport à ces techniques qui ont fait leurs preuves.

Les applications de reconnaissance de la parole arabe que nous allons décrire se situent à deux niveaux: un niveau phonème, pour pouvoir évaluer directement le modèle ***STN*** du neurone, et un niveau mot pour évaluer ***NESSR*** globalement.

Dans le *chapitre 7*, nous présentons une illustration du modèle ***NESSR***, au travers d'un système de détection de la dyslexie chez de jeunes élèves.

Conclusion et perspectives

Nous terminons ce manuscrit par une conclusion sur l'ensemble de nos travaux et la présentation des perspectives d'utilisation du modèle ***STN*** en particulier son aspect générique et son extension à d'autres applications et celles du système ***NESSR*** qui peut être étendu à la reconnaissance d'unités de langage plus grandes.

PARTIE1 :

Contexte de l’étude et état de l’art

Chapitre I:

Contexte de l’étude

Chapitre 1:
Contexte de l'étude

Présentation du chapitre

Dans ce chapitre nous présentons le contexte de notre travail, et ce au travers de quelques rappels et notions sur la parole, sujet de notre thème.
Nous présentons également dans ce chapitre, la langue arabe qui est le cadre de nos expérimentations, et nous aborderons les problèmes spécifiques pour sa reconnaissance automatique. Pour clore le chapitre, une introduction à la reconnaissance de la parole est présentée.

1. Communication

1. 1. La communication entre humains

La parole est la faculté naturelle qu'ont les êtres humains de s'exprimer et de communiquer leurs pensées, leurs idées et leurs émotions par un système de sons articulés ; c'est le moyen de communication privilégié entre les humains qui sont les seuls à utiliser un tel système structuré.

Le modèle de Kerbrat-Orrechioni [Kerbrat, 80] ci-dessous, présente la transmission d'un message d'un être humain à un autre, il est dérivé des travaux de Jacobson [Jacobson, 60]. La figure I. 1. présente l'émetteur, le canal de transmission et le récepteur. Le canal fonctionne comme le proposait Jacobson sous forme de message, il y a ensuite transmission du message puis décodage par le récepteur. Il ressort de cette représentation la complexité des mécanismes et la diversités des connaissances entrant

en jeu lors de l'émission et de la réception d'un message. Ainsi, le message émis dépend des richesses linguistiques de l'émetteur et est tributaire du passif culturel et idéologique de celui-ci. Entre également en considération l'état psychologique de l'émetteur qui influe non seulement sur son message mais également sur le décodage et l'interprétation du message par le récepteur. Du côté récepteur, ce dernier doit mettre en œuvre toutes les connaissances dont il dispose pour mettre en place un modèle d'interprétation lui permettant au mieux de comprendre le message émis [Vaufraydaz, 02].

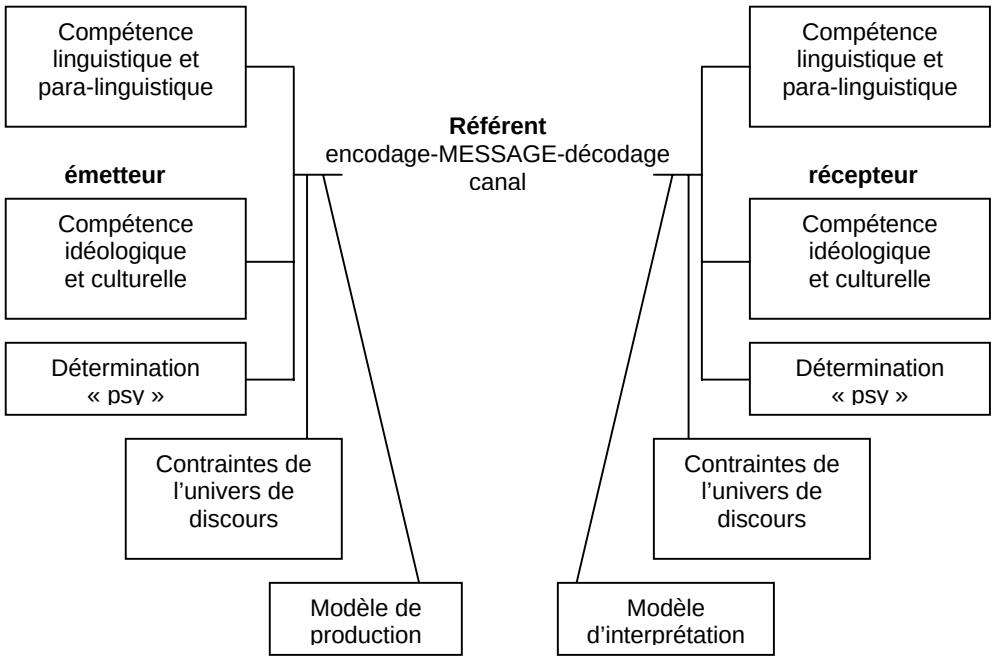


Figure I.1. Modèle de communication entre humains [Kerbrat, 80]

1. 2. La communication homme-machine

Dans le cas de la communication homme / machine, et en particulier dans un contexte de reconnaissance automatique de la parole, le schéma précédent est modifié du côté du récepteur qui est l'ordinateur. La communication homme / machine est une orientation de la recherche en informatique, qui requiert la compétence de plusieurs sciences et techniques telles que : l'intelligence artificielle, le traitement du signal, la linguistique, la phonétique …, et bien qu'elle n'a cessé de susciter un intérêt croissant dès les balbutiements du traitement automatique de l'information, cette tâche aspire toujours à réaliser un système qui puisse « comprendre » (après avoir reconnu) la parole continue.

Nous pensons que des bases intéressantes d'un système de compréhension de la parole passeraient par un modèle cognitif de la machine qui soit fortement inspiré de l'être humain.

2. Le signal de la parole

2. 1. Le signal de la parole

Le signal de la parole est un phénomène de nature acoustique porteur d'un message. L'information d'un message parlé réside dans les fluctuations de l'air, engendrées, puis émises par l'appareil phonatoire. Ces fluctuations constituent le signal vocal. Elles sont détectées par l'oreille qui procède à une certaine analyse. Les résultats sont transmis au cerveau qui les interprète.

D'autre part, le signal vocal représente la combinaison d'éléments simples et brefs du signal sonore appelés phonèmes, qui permettent de distinguer les différents mots. La parole est un signal réel, continu, d'énergie finie et non stationnaire. Sa structure est complexe et variable avec le temps [Haton, 91].

2. 2. Caractéristiques du signal de la parole

Le signal de parole n'est pas un signal ordinaire : il s'inscrit dans le cadre de la communication parlée, un phénomène des plus complexes. Afin de souligner les difficultés du problème, nous faisons ressortir essentiellement quelques caractéristiques notoires de ce signal :

1. ***Un débit intense :*** D'un point de vue mathématique, il est ardu de modéliser le signal de parole, car ses propriétés statistiques évoluent au cours du temps.
2. ***Une extrême redondance :*** Lorsqu'on a vu une représentation graphique de l'onde sonore on est certainement frappé par le caractère répétitif du signal de parole. En effet, un grossissement visuel permettrait de voir une succession de figures sonores semblant se répéter à l'excès. Un peu de recul laisse apparaître des zones moins stables qu'il convient de qualifier de transitoires. En fait, ce qui semblerait de prime abord superflu,

s'avère en réalité fort utile. Les répétitions confèrent à ce signal une robustesse car cette redondance le rend résistant au bruit.

3. ***Une grande variabilité :*** Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution en détermine la durée. Toute affection de l'appareil phonatoire peut altérer la qualité de la production. Un rhume teinte les voyelles de nasalité ; une simple fatigue et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps : l'enfance, l'adolescence, l'âge mûr, puis la vieillesse.

La variabilité inter-locuteur est encore plus flagrante. La hauteur de la voix, l'intonation et l'accent diffèrent selon le sexe, l'origine sociale, régionale ou nationale. D'ailleurs, la reconnaissance du locuteur est un axe de recherche à part entière.

Enfin, toute parole s'inscrit dans un processus de communication où entrent en jeu de nombreux éléments comme le lieu, l'émotion, l'intention, la relation qui s'établit entre les interlocuteurs. Chacun de ces facteurs détermine la situation de communication, et influe à sa manière sur la forme et le contenu du message.

4. ***Un lieu d'interférences :*** La production "parfaite" de chaque son suppose théoriquement un positionnement précis des organes phonatoires. Or, lorsque le débit de la parole s'accélère, le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis dans une même chaîne acoustique subissent l'influence de ceux qui les suivent ou les précèdent. Ces effets de co-articulation sont des interférences. Ils entraînent l'altération des formes sonores en fonction des contextes droits ou gauches, selon des règles étudiées par les acousticiens d'un point de vue articulatoire ou perceptif.

2. 3. Mécanisme de phonation et sons de la parole

La parole est le résultat de l'action volontaire et coordonnée des appareils respiratoire et masticatoire. Pendant l'élocution, un flux d'air en provenance des poumons traverse la trachée artère. Au sommet de celle-ci se trouve le larynx où les cordes vocales vibrent sous l'effet du passage de l'air à travers la glotte. Ces vibrations s'accompagnent de variations de longueur, de tension, et d'épaisseur des cordes. Cet air sera appliqué au conduit vocal qui s'étend du pharynx jusqu'au lèvres. L'onde acoustique, après avoir

parcouru le pharynx, va pouvoir être plus au moins dérivée, selon la position du voile du palais vers les fosses nasales. Le flux d'air peut être arrêté par la fermeture des lèvres.

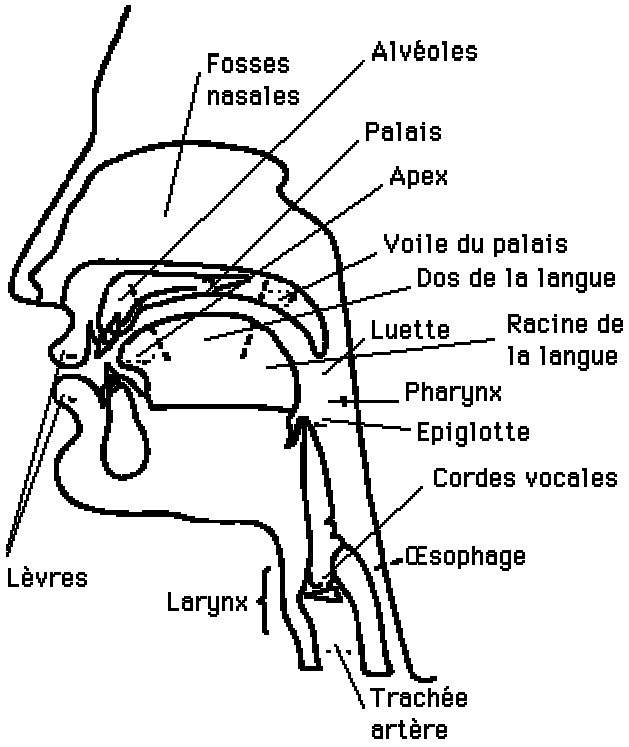


Figure I. 2. Vue d'ensemble des organes de la parole [Web 01]

En simplifiant, on peut dire que la parole est le résultat de l'excitation des cavités nasales et/ou orales par une ou deux sources acoustiques. La première, essentielle elle génère des impulsions périodiques, l'autre peut s'ajouter ou se substituer à la première : il s'agit, cette fois, de bruit d'explosion ou de friction qui peuvent naître à l'intérieur du conduit vocal (de la glotte aux lèvres). Lors de l'émission sonore, le flux d'air produit à travers le conduit vocal un son de trois manières différentes :

- En vibrant les cordes vocales de façon périodique ou quasi-périodique, produisant des sons voisés (source de voisement).
- En réduisant la dimension du conduit vocal afin de provoquer une turbulence, produisant des sons fricatifs (source de bruit).
- En libérant brusquement la pression accumulée derrière un obstacle, produisant les phases explosives des sons occlusifs.

3. La langue arabe

3. 1. Présentation

L'Arabe est la sixième langue actuellement parlée dans le monde. On estime le nombre d'Arabophone à 250 millions. C'est la langue officielle de 22 pays. Mais comme c'est aussi la langue qui porte les instructions religieuses de l'Islam dans le livre sacrée, on peut imaginer que le nombre de personnes qui parlent l'Arabe est nettement plus élevé.

Il est important de souligner que lorsqu'on évoque l'Arabe, il ne s'agit pas d'une unique variété linguistique mais bien d'un ensemble de dialectes et de sociolectes. L'Arabe classique est le plus ancien, cette forme littéraire du langage est celle qui a été utilisée dans le Quran ; l'Arabe Moderne Standard (MSA pour Modern Standard Arabic) est une version de l'Arabe classique modernisée. Le MSA est la langue officielle commune à tous les pays qui parlent l'Arabe ; c'est aussi le langage utilisé par les média (journaux, radio, TV,…), et dans les discours officiels. L'approche de reconnaissance que nous proposons a été testée sur des exemples de l'arabe standard, il en est de même pour toutes les réalisations que nous présentons dans la troisième partie du manuscrit.

3. 2. Le système d'écriture

L'Arabe s'écrit sous forme cursive et de droite à gauche. L'alphabet consiste en vingt-huit lettres, parmi lesquelles, vingt-cinq représentent des consonnes. Les trois lettres restantes représentent les voyelles longues (/a:/, (u:/,/i:/). Chaque lettre apparaît souvent en quatre formes selon qu'elle soit en début, en milieu ou en fin de mot, ou isolée. Les lettres sont le plus souvent connectées entre elles sans majuscules.

Une caractéristiques importante de l'Arabe écrit est qu'aux voyelles courtes ne correspondent pas de lettres dans l'alphabet, elles sont représentées par des diacritiques placées sur les consonnes qui les accompagnent. Il est à remarquer que les textes arabes sont rarement complètement voyellés, les diacritiques ne sont utilisées que pour prévenir les ambiguïtés.

3. 3. La phonologie

L'Arabe compte trente et un phonèmes, ceux que nous avons déjà mentionné auxquels il faut ajouter les voyelles courtes.

a. Les consonnes : Une consonne est un phonème dont la prononciation se caractérise par une obstruction totale ou partielle en un ou plusieurs points du conduit vocal. Elle est généralement précédée ou suivie d'une voyelle.

b. Les voyelles : Lors de la prononciation des voyelles, l'air émis par les vibrations des cordes vocales passe librement à travers le conduit. On distingue trois types de voyelles : les voyelles courtes, longues et les semi-voyelles.

b. 1. Les voyelles courtes : en arabe, il existe trois voyelles courtes classées d'après la position des organes de phonation qui concourent à leur émission. La première voyelle se prononce en contractant la langue au fond de la bouche et en avançant les lèvres qui s'arrondissent jusqu'à presque se joindre ‘ُ’. La deuxième voyelle se prononce en ouvrant largement la bouche et en conservant la langue dans une position horizontale ‘ َ’. La troisième voyelle se prononce en portant le devant de la langue en avant et l'étalant largement, tandis que l'arrière frôle presque le palais et que les commissures des lèvres s'étirent ‘ ِ’.

b.2. Les voyelles longues : La consonnes /w/ dépourvue de voyelle et précédée d'une ‘damma’, cesse d'être consonne et devient une voyelle longue ‘damma étendue’.

Le signe ‘ اٰ ’ dépourvue de ‘hamza’ sert à allonger la voyelle brève qui le précède, ce qui donne la voyelle longue ‘fatha étendue’. La consonne /y/ dépourvue de voyelle et précédée d'une ‘kasra’, cesse d'être consonne et devient une voyelle longue ‘kasra étendue’

b.3. Les semi-voyelles : Lorsque l'excitation glottique coexiste avec une évolution rapide du conduit vocal nous constatons qu'il existe deux semi-voyelles :

Le sekune : Nous appelons ‘sekune’ le signe ‘ ْ’, placé au dessus d'une consonne pour indiquer que cette consonne n'est pas munie d'une voyelle.

Le tanwin : Les signes qui représentent les voyelles sont quelques fois redoublés à la fin du mot. Les voyelles finales se lisent alors comme si elles étaient suivies de son /n/.

Quant au signe ' ّ ', appelé tachdid lorsqu'il est placé au dessus d'une consonne celle-ci est redoublée lors de sa prononciation ; la première portant un sukune et terminant la syllabe précédente, la seconde portant la voyelle qui accompagne le tachdid.

Letter	Name	Phoneme
ا	'alif	/ a: /
ب	baa'	/ b /
ت	taa'	/ t /
ث	thaa'	/ θ /
ج	gym	/ □ /
ح	Haa'	/ ʜ /
خ	khaa'	/ x /
د	daal	/ d /
ذ	dhaal	/ ð /
ر	zayn	/ z /
ز	raa	/ r /
س	syn	/ s /
ش	shyn	/ □ /
ص	Saad	/~~s~~ /
ض	Daad	/ ~~d~~ /
ط	Taa'	/ ~~t~~/
ظ	Zaa'	/ ~~z~~ /
ع	'ayn	/ □ /
غ	ghayn	/ □ /
ك	kaaf	/ k /
ق	qaaf	/ q /
ف	faa'	/ f /
ل	laam	/ l /
ن	nuwn	/ n /
م	mym	/ m /
ه	haa'	/ h /
و	waaw	/ u: /
ي	yaa'	/ i: /
ء	hamza	/ □ /

Tableau I.1. Lettres de l'alphabet arabe. Les noms sont donnés en Qalam romanisation [Web 02]; les phonèmes correspondants sont donnés en notation API [Web 03].

3. 4. Caractéristiques phonétiques des phonèmes arabes

3. 4. 1. Lieux d'articulation

Le lieu d'articulation est la zone du conduit vocal qui participe à la formation du son, il varie d'un phonème à un autre. Pour les phonèmes arabes, il y a plus de 28 lieux d'articulation, c'est pourquoi nombre de phonéticiens ont pris comme critère de classification, le lieu d'articulation (Tableau I.2.).

Lieux d'articulation	Phonèmes
Pharyngale	ħ ع
Laryngale	h ء
Uvulaire	x q غ
Post-palatale	k
Pré-palatale	ʃ ج ي r y
Dentale	t ط d ض s ص z n
Inter-dentale	θ ð ظ
Labio-dentale	f
Bilabiale	m b w

Tableau I.2. Classification des phonèmes Arabes selon le lieu d'articulation (source [Harkat, 89])

3. 4. 2. Traits distinctifs des phonèmes arabes

La notion de « trait » exprime une similarité aux niveaux articulatoire et acoustique. Les phonèmes de la langue arabe se regroupent en catégories naturelles dont les éléments partagent des « traits distinctifs ». Ces traits nous permettent de déterminer les phonèmes en prenant en compte leur lieux d'articulations. De tous ces traits, nous citons :

Sourd/sonore : lors de la prononciation des phonèmes sourds, les cordes vocales s'écartent mais ne vibrent pas. En revanche pour les phonèmes sonores, les cordes vocales vibrent.

Emphatique : il se traduit par la levée du dos de la langue jusqu'à ce qu'il soit superposé à la zone palatale supérieure. On obtient donc un son de moins en moins aigu.

Nasal : Un phonème est dit nasal si sa prononciation se caractérise par l'abaissement du voile du palais, et donc la mise en communication du conduit nasal avec le conduit vocal.

Occlusif / fricatif : ce trait se traduit par une obstruction de l'air dans le conduit vocal. Cette obstruction peut être totale dans le cas des sons occlusifs ou partielle pour les sons fricatifs.

Sonnantes : Ces phonèmes ne sont ni fricatifs, nui occlusifs. L'obstacle est donc le plus discret possible.

La classification selon les traits distinctifs est résumée dans le tableau (Tableau I.3).

Trait	Phonèmes
Voisé	y □ b □ d ð r z ɖ ŧ ƶ □ □ q l m n w
Sourd	k t ʜ x s □ ʂ f h
Emphatique	ʂ ɖ ŧ ƶ q
Nasal	m n
Occlusif	□ b □ d ŧ □ q k
Fricatif	θ □ ʜ x ð z s □ ʂ ɖ ƶ □ □ f h
Sonore	y d l m n w

Tableau I.3. Classification des phonèmes arabes selon les traits distinctifs (source [Harkat 89])

3. 4. 3. Autres particularités

Bien qu'en Arabe, il y'ait une correspondance un-à-un graphème-phonème, l'Arabe comporte quelques alternatives phonétiques en fonction du contexte, les plus courantes d'entre-elles sont :

- Un sous ensemble de consonnes arabes (/ŧ/, /ɖ/, /ʂ/, /ƶ/) dites emphatiques sont les formes pharyngées des consonnes (/t/, /d/, /s/, /z/). Dans leur voisinage, les sons frontaliers deviennent pharyngés.

- Taa marbuwta 'ð', est une marque grammaticale du féminin qui se place à la fin du mot, et qui est prononcée comme /a/.
- L'assimilation du /l/ dans l'article défini / ?l/ s'il est suivi par l'une des lettres (/s/, /z/, t/, /d/, /n/, /ŧ/, /ɖ/, /ʂ/, /ƶ/, /ð/, /θ/, /ʃ/, /r/) que l'on appelle lettres solaires.

4. Introduction à la 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 I. 3), d'abord l'étape d'extraction de caractéristiques ensuite l'étape de reconnaissance (ou de classification). Le module d'extraction de caractéristiques se charge de transformer le signal en entrée du système en une représentation interne de sorte qu'il soit possible de reconstituer 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.

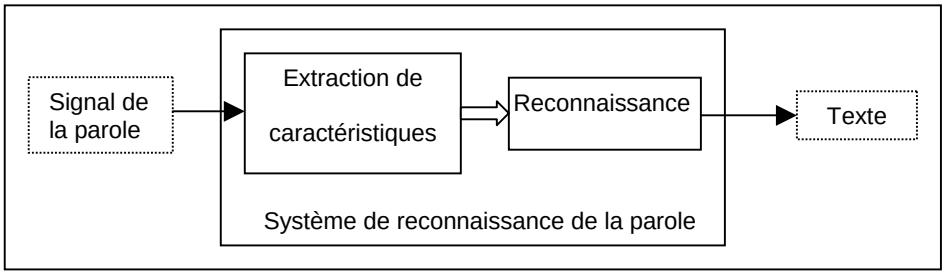


Figure I. 3. Blocs de base composant un système de RAP

Chapitre II:

Reconnaissance automatique de la parole

Chapitre II: Reconnaissance automatique de la parole

Présentation du chapitre

Le signal de la parole est un phénomène de nature acoustique porteur d'un message. L'information du message parlé réside dans les fluctuations de l'air engendrées puis émises par l'appareil phonatoire. Ces fluctuations constituent le signal vocal, elle sont détectées par l'oreille, qui procède à une certaine analyse. Les résultats sont ensuite transmis au cerveau qui les interprète.

A l'image de ce processus naturel, le processus de reconnaissance automatique de la parole inclut deux grandes phases : dans la première phase on s'attache à extraire du signal continu de la parole un certain nombre de paramètres qui le caractérisent. Ainsi, étant donné un signal en entrée du système, celui-ci va subir un pré-traitement qui consiste généralement en un filtrage et un échantillonnage qui permet de passer d'un signal continu à des valeurs discrètes, de ces valeurs dont le nombre est important seront extraites des caractéristiques qui permettent de représenter de façon compacte et pertinente le signal originel, une description plus ample de cette phase ainsi, que des différentes techniques utilisées pour l'extraction de caractéristiques constitue la première partie de ce chapitre.

La seconde partie du chapitre est consacrée aux méthodes de reconnaissance. En effet, une fois l'analyse du signal effectuée, intervient la phase de reconnaissance.

1. La reconnaissance de la parole

1. 1. Introduction à la RAP

La reconnaissance automatique de la parole (RAP) est le processus par lequel la machine tente de « décoder » le signal de la parole qui lui est destiné. Les recherches relatives à la RAP débutèrent dans les années 1950, dans une conjoncture optimiste, car on pensait que les avancées technologiques des ordinateurs rendraient la RAP une tâche aisée. Quelques dizaines d'années plus tard, on se rendait compte que c'était faux, et que la RAP, demeure un problème difficile. Aujourd'hui encore nombre de questions restent posées, les difficultés majeures étant associées à la taille du vocabulaire à reconnaître, la reconnaissance de la parole spontanée, à la reconnaissance indépendamment du locuteur, la parole bruitée, …

La reconnaissance automatique de la parole est très souvent basée sur une représentation paramétrique du signal, son but étant la communication en langue naturel avec une machine. Il s'agit là de deux objectifs différents que l'on peut assigner à un système : la reconnaissance conduisant à une application du type dictée vocale, et la compréhension, qui consiste à accéder à la signification de l'énoncé parlé.

1. 2. Concepts de base

La démarche classique suivie lors du processus de reconnaissance automatique de la parole est illustré par la figure II.1, ce schéma fait ressortir les étapes principales dans un tel processus.

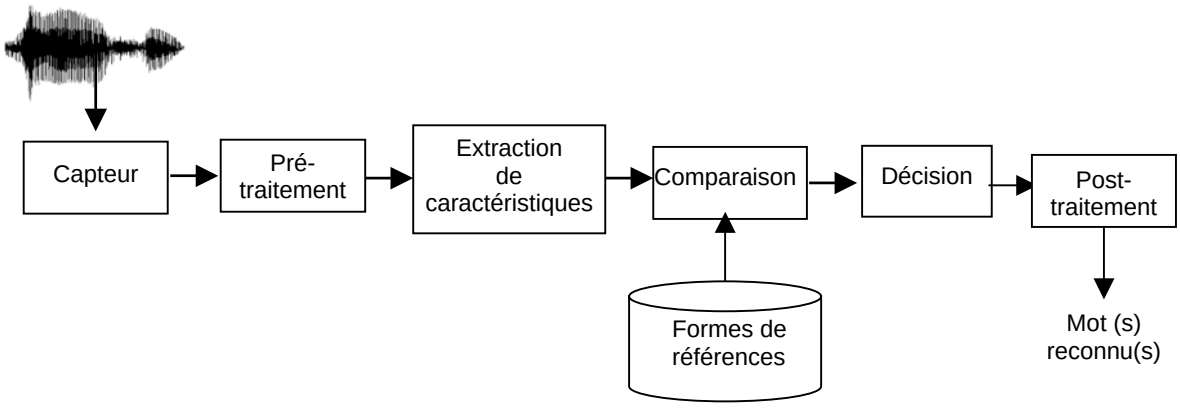


Figure II.1. Organigramme d'un système de R.A.P

Ainsi, étant donné un signal en entrée du système, celui-ci va subir un pré-traitement qui consiste généralement en un filtrage et un échantillonnage qui permet de passer d'un signal continu à des valeurs discrètes, de ces valeurs dont le nombre est important seront extraites des caractéristiques qui permettent de représenter de façon compacte et pertinente le signal originel. Cette étape permet d'avoir une première représentation du signal, ensuite et selon l'approche adoptée par le système de reconnaissance, ce modèle représentatif du signal sera comparé à des formes d'autres signaux que le système « connaît ». Sur la base du résultat de cette comparaison une décision quant au mot reconnu sera prise, celle-ci sera éventuellement validée en considérant les connaissances du domaine.

1. 3. Quelques systèmes de RAP

Les programmes de reconnaissances de la parole ont été développés principalement aux Etats-Unis dans le cadre du projet ARPA. Quatre programmes principaux sont opérationnels : Il s'agit de HARPY et HEARSAY II de CMU qui sont des programmes de reconnaissance de la parole continue, le système de BBN (Bolt, Berenek and Newman) comprend un analyseur phonétique basé sur un treillis phonétique, l'analyse syntaxique étant réalisé grâce aux ATN, et le système SDC (systems development corporation) où l'analyseur est basé sur le treillis probabilisé des syllabes.

D'autres systèmes sont apparus par la suite, en particulier : Tangora qui est un système multi-locuteurs développé par IBM. Il fonctionne en temps réel suivant une approche globale en utilisant les modèles de Markov cachés (HMM). Le logiciel DragonNaturally Speaking est un produit compétitif sur le marché, il utilise aussi une approche globale par les HMMs.

1. 4. Reconnaissance de la parole arabe

1. 4. 1. Problèmes rencontrés en reconnaissance de l'Arabe

De nombreux aspects de l'Arabe tels que la phonologie ou la syntaxe ne posent pas de problèmes particuliers en reconnaissance automatique de la parole [Kirshhoff, 02]. Les techniques standards de la modélisation acoustique et de la prononciation

indépendamment du langage peuvent tout à fait être appliquées pour la modélisation acoustique et phonétique de l'Arabe. D'autres aspects pour l'apprentissage du système de reconnaissance sont mêmes plus faciles que pour d'autres langages, en particulier la construction du lexique car il y'a une quasi correspondance un-à-un entre lettre et phonème.

Les difficultés majeures rencontrées lors de développement de systèmes performants de reconnaissance pour l'Arabe sont la prédominance de textes non voyellés, d'énormes variétés dialectales, et une complexité morphologique.

En particulier, la complexité de la morphologie de l'Arabe est bien connue pour présenter d'énormes problèmes lors de la modélisation linguistique, ceci en raison d'un nombre élevé de préfixes et de suffixes que l'on peut greffer à une racine ce qui conduit à une explosion des formes que l'on peut associer à un mot [Mrayati, 84].

1. 4. 2. Travaux antérieurs

Les travaux en RAP de l'Arabe se sont exclusivement intéressés à la reconnaissance de l'Arabe standard (MSA). On soulignera en particulier les travaux de El-Ani [El Ani, 70], qui portent sur des investigations acoustiques et structurelles des sons arabes, et ceux de Mrayati [Mrayati, 84] qui eux se penchent plus sur l'aspect syntaxique. On trouve aussi les travaux de M. Djoudi [Djoudi, 91], qui ont permis la réalisation du système de reconnaissance MARS réalisé à l'Université de Poitiers. Il s'agit d'une reconnaissance multilocuteur. Le système se compose de deux parties, un décodeur acoustico-phonétique (SAPHA) et un décodeur linguistique (SALAM), ce dernier traite les aspects morphologiques, syntaxico-sémantique et la prosodie propre à la langue arabe.

De nombreux autres travaux de moindre ampleur se penchent sur des aspects précis de la parole mais leur nombre qui augmente permettra certainement d'enrichir la littérature dans ce cadre de reconnaissance et aidera à la construction de systèmes de plus en plus robustes et performants.

2. L’analyse du signal

Le traitement numérique des signaux connaît depuis trois décennies un développement fulgurant. Une multitude de méthodes puissantes de traitement des signaux peuvent désormais être mise en œuvre grâce aux techniques numériques. L’étude de la parole a été un des domaines importants qui a bénéficié et qui continue de bénéficier du traitement numérique des signaux. Dans la suite du chapitre, nous présentons quelques unes des techniques les plus couramment utilisées en RAP.

L'étape d’analyse du signal est une opération essentielle, elle a pour but de fournir une représentation moins redondante du signal de la parole que celle obtenue par codage de l’onde temporelle tout en permettant une extraction précise des paramètres significatifs et pertinents. Le signal analogique est fourni en entrée et une suite discrète de vecteurs, appelée trame acoustique est obtenue en sortie. Mais avant tout traitement il faut discrétiser le signal continu sortant du microphone, puis le stocker en mémoire sous forme numérique.

2. 1. L’échantillonnage

Avant tout traitement, il est nécessaire de numériser le signal continu sortant du microphone ou d’un appareil d’enregistrement. Cette opération s’appelle échantillonnage du signal. L’échantillonnage procède à un découpage dans le temps du signal continu S(t). Il consiste à sélectionner au moyen de circuit de commutation, les valeurs prises par le signal en une suite d’instants t_1, t_2, …,t_n régulièrement espacés. Sachant que L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50 Hz - 8 kHz], la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon [Kunt, 91] ; mais elle peut varier en fonction du domaine d'application ou des besoins ou contraintes matériels. Il s'ensuit qu'en fonction de la fréquence d'échantillonnage choisie, un filtrage analogique passe bande est effectué, afin de réduire la bande passante correctement, et il est suivi de l'échantillonnage numérique.

Remarquons qu’un signal échantillonné à une fréquence de 11025 Hz ; est mesuré 11025 fois par seconde.

2. 2. Le fenêtrage

La quantité de points d'échantillonnage est extrêmement volumineuse, il est donc nécessaire de réduire ce nombre et d'éliminer la redondance. La trame acoustique est un ensemble de coefficients ou paramètres, calculés sur un bloc d'échantillons. Comme les techniques utilisées pour l'extraction de ces coefficients supposent que le signal sur lequel on opère est stationnaire, la plupart des algorithmes d'analyse opèrent donc sur un bloc d'échantillons de taille fixe dans lequel le signal est supposé stationnaire, il correspond à un temps de parole de 20 à 40 ms. La suite de vecteurs d'analyse est obtenue en déplaçant ce bloc de 10 à 20 ms ; il y a recouvrement de blocs.

Figure II.2. signal du mot [wahid] (Fe=11025Hz)

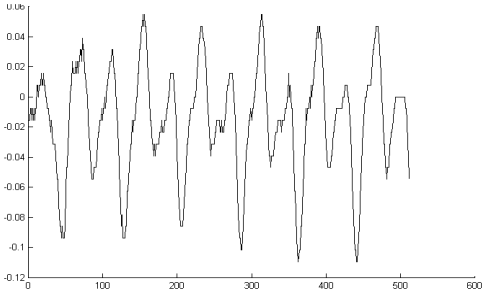


Figure II.3. la 1ere trame formée de 512 échantillons, le mot contient 56 fenêtres

Une fois le fenêtrage effectué, il est courant de pondérer ces fenêtres par des fonctions appropriées ; on cite en particulier la fenêtre de Hamming que nous avons utilisée.

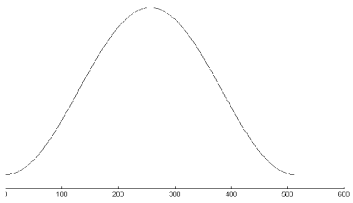


Figure II.4. Fenêtre de Hamming pour N=512

Une fois la pondération réalisée, le calcul des coefficients discriminatifs est effectué, pour ce faire il existe des méthodes classiquement répertoriées en deux catégories : les méthodes temporelles et les méthodes spectrales.

2. 3. Extraction des caractéristiques

En acoustique, un son se définit classiquement par son intensité, sa hauteur qui est fixée par la fréquence de vibration des cordes vocales, appelée fréquence du fondamental ou pitch (F_0). Deux sons de même intensité et de même hauteur se distinguent par le timbre qui est déterminé par les amplitudes relatives des harmoniques du fondamental. Ces amplitudes nommées formants se caractérisent par les maximums de la fonction de transfert (transmittance) du conduit vocal. En général, les trois premiers formants sont essentiels pour caractériser le spectre du signal vocal.

Nous avons précédemment souligné que le signal de la parole présente des particularités telle que la redondance qui justifient tout à fait la recherche d'une représentation plus compacte du signal. Pour cela, il existe différentes techniques d'analyse du signal vocal, que nous regroupons dans ce qui suit en trois catégories : les méthodes temporelles les méthodes spectrales et les méthodes cepstrales. Il y'a d'autres classification de ces techniques en particulier en méthodes spectrales, modèles d'identification et modèles d'audition (voir [Haton, 91]).

2. 3. 1. Méthodes temporelles

Les méthodes de type temporel sont basées sur l'analyse des caractéristiques temporelles du signal vocal telles que : l'énergie, le taux de passage par zéro, le calcul de la fréquence fondamental etc. Différentes techniques permettent l'analyse de l'aspect temporel du signal vocal afin de permettre de déduire ses paramètres, parmi ces méthodes nous trouvons :

- Le taux de passage par zéro (PPZ),
- L'analyse par prédiction linéaire (LPC).

2. 3. 1. 1. Le taux de passage par zéro

Cette méthode permet en comptant les passages par zéro du signal, de construire des histogrammes d'intervalles de fréquence. On ne s'intéresse pas dans cette méthode à l'amplitude du signal mais à son signe. Les résultats sont assez grossiers car la variance des passages par zéro est forte surtout dans les transitoires [Haton, 91].

Pour un signal échantillonné, il y a passage par zéro lorsque deux échantillons successifs sont de signes opposés [Boite, 87]. Le calcul du taux de passage par zéro du signal de la parole permet de faire la distinction d'une part entre le signal de la parole (information utile) et le bruit, et d'autre part entre les sons voisés et les sons non voisé.

Grâce au taux de ppz d'un signal, on peut faire ressortir trois plages de valeurs qui permettent de distinguer la nature des sons soit :

- Plage de silence : taux de ppz très faible (entre 0 et 3),
- Plage de voisement : taux de ppz moyen (entre 4 et 27),
- Plage de dévoisement : taux de ppz élevé (> 27).

2. 3. 1. 2. Analyse par prédiction linéaire (LPC)

L'étude du mécanisme de phonation montre que la production de chaque unité phonétique (phonème, syllabe, …) dépend de la position articulatoire des organes phonatoires (bouche, langue, lèvres, …). Le conduit vocal est donc considéré comme un filtre soumis à une excitation U(n). pour les sons voisés ou sonores, cette excitation est un train périodique d'impulsions ; pour les sons non voisés, l'excitation est un bruit blanc. Ce modèle de production est appelé « AutoRégressif » (AR).

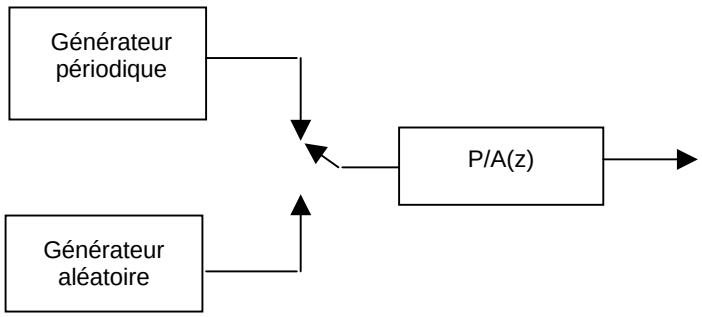


Figure II.5. Modèle AR de production de la parole

La transformée du signal peut s'écrire comme suit [Caliope, 89]:

$$S(z)=U(z)/A(z) \qquad \text{(II.1)}$$

$$\text{où} \quad A(z)=\sum_{i=0}^{p} a(i)Z^{-i} \quad avec \quad a(0)=1 \qquad \text{(II.2)}$$

Dans le domaine temporel, on aura la récursion linéaire suivante :

$$S(n)+\sum_{i=1}^{p} a(i)*S(n-i)=u(n) \qquad \text{(II.3)}$$

Cette récurrence exprime le fait qu'un échantillon quelconque S(n) peut être déterminé par une combinaison linéaire des p échantillons qui le précèdent à laquelle il faut ajouter le terme d'excitation.

Les coefficients a(i) sont appelés « coefficients de prédiction » ; en effet, si l'excitation était nulle, chaque échantillon S(n) pourrait être prédit exactement à partir des *p* échantillons qui le précédent immédiatement. Cette valeur « prédite » s'écrit :

$$S^*(n)=\sum_{i=1}^{p} a(i)*S(n-i) \qquad \text{(II.4)}$$

En rapprochant les formules II.3 et II. 4, on peut interpréter u(n) comme étant une erreur de prédiction.

$$a(n)=S(n)-S^*(n) \qquad \text{(II.5)}$$

$$u(n)=S(n)-\sum_{i=1}^{p}(a(i)\times S(n-i)) \qquad \text{(II.6)}$$

Il faut alors minimiser l'erreur quadratique totale de prédiction définie par :

$$E(p)=\sum_{n=1}^{N} u^2(n) \qquad \text{(II.7)}$$

Il existe pour cela des méthodes mathématiques. En particulier, on calcule les dérivées partielles de E(p) par rapport aux coefficients de pondération a(i) et l'on annule chacune d'entre elles. On obtient ainsi un système d'équation constitué d'un ensemble de *p* équations avec *p* inconnues a(1), a(2), …a(p).

$$\begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \cdot \\ \cdot \\ R(p) \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & \dots & R(p) \\ & R(0) & & \\ & & R(0) & \\ & & & \\ & & & R(0) \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ a_p \end{bmatrix}$$

Rs Rss

Où Rs est le vecteur d'autocorrélation et Rss, la matrice d'autocorrélation. Rss est la matrice de Toplitz symétrique et les éléments de la diagonale sont égaux. Un algorithme efficace pour la résolution de ce système d'équation a été développé par Levinson et appliqué au calcul des coefficients par Durbin [Caliope, 89].

2. 3. 2. Méthodes fréquentielles ou spectrales

Ces méthodes sont fondées sur une décomposition fréquentielle du signal sans connaissance a priori de sa structure fine ; la seule hypothèse mise en jeu concerne le choix des fonctions sur la base desquelles le signal est décomposé ou le choix des caractéristiques des filtres pour une analyse par vocodeurs. Ces méthodes dont l'ancêtre est le spectrogramme étudient le signal dans le domaine fréquentiel. Il s'agit donc de transformer le signal originel de la représentation temporelle à une représentation fréquentielle, la plus utilisée de ces transformées est la transformée de Fourier discrète DFT.

2. 3. 2. 1. La transformée de Fourier

La Transformé Discrète de Fourier TDF est une méthode d'analyse qui n'opère que sur un nombre d'échantillons qui dépasse une centaine de points d'échantillonnage. Elle utilise le fenêtrage temporel avec recouvrement donc le temps de calcul reste considérable, pour remédier à cet inconvénient, des algorithmes rapides tel que la FFT (pour Fast Fourier Transform) permettent d'obtenir des spectres en temps réel.

Principe : Cette méthode est fondée sur le théorème de Fourier qui stipule que tout signal périodique peut être décomposé en une somme de sinusoïdes harmoniques. La transformée de Fourier conduit donc à transformer un signal complexe en une combinaison de fonctions élémentaires de formes simples et bien connues. Soit s(t) le signal temporel, la transformée de Fourier ou le spectre s est donnée par la formule:

$$F(\omega)=\int_{-\infty}^{+\infty} s(t)*e^{-j\omega t}dt \qquad \text{(II.8)}$$

Le terme $|F(\omega)|^2$ représente le spectre d'énergie, il exprime la répartition fréquentielle de l'énergie du signal ; il exprime aussi un estimateur de la densité spectrale court-terme, si on procède à une pondération du signal par une fenêtre d'analyse.

Le défaut majeur de la TDF pour le calcul du spectre réside dans l'inter-modulation source/conduit qui rend difficile la mesure des formants F_i et la mesure du fondamental F_0. Le lissage ou le cepstre est une méthode qui vise à séparer leur contribution.

2. 3. 2. 2. Traitement par bancs de filtres

Cette technique d'analyse spectrale est basée sur la représentation du signal par sa transformée de Fourier pendant un intervalle de temps suffisamment court. Le signal subit ainsi une décomposition fréquentielle permettant d'isoler les informations utiles.

Le principe de cette technique est d'injecter le signal s(t) dans un banc de filtres passe bande couvrant une étendue spectrale intéressante de la voix (de 200 à 600 Hz en général). Les N filtres réalisant cette analyse doivent tous avoir un même gain unité et de fréquences centrales différentes. Les bancs de filtres se différencient entre eux par le nombre de filtres N qui varie en pratique entre 12 et 32 filtres, la distribution de la fréquence centrale et la caractéristique du filtre basse-bas à la sortie du redresseur. Cette méthode d'analyse du spectre à court terme constitue le principe de la plupart des vocodeurs à canaux (voice coders). Le schéma classique d'un vocodeur à canaux est constitué de plusieurs canaux placés en parallèle. Le schéma (figure II.6) présente l'exemple d'un canal.

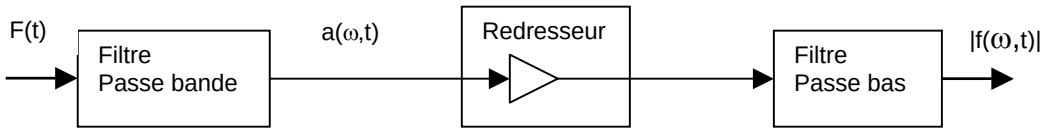


Figure II.6. Structure d'un canal de banc de filtres

L'analyse par bancs de filtres présente l'avantage d'une grande performance avec un prix de revient assez faible, son inconvénient étant un manque de souplesse car la modification des caractéristiques d'un filtre nécessite le changement de la configuration matérielle de ce filtre.

2. 3. 3. Méthodes cepstrales

Contrairement au spectrogramme qui ne fait appel à aucune connaissance a priori sur le signal acoustique, l'analyse cepstrale résulte de travaux sur le modèle de production de la parole : son but est d'effectuer la déconvolution « source / conduit » par une transformation homomorphique.

Les coefficients cepstraux sont obtenus en appliquant une transformée de Fourier numérique inverse au logarithme du spectre d'amplitude (Figure II.7). Le signal ainsi obtenu est représenté dans un domaine appelé cepstral ou quéfrentiel ; les échantillons se situant en basses quéfrences correspondent à la contribution du conduit vocal et donnent les paramètres utilisés en RAP, tandis que la contribution de la source n'apparaît qu'en hautes quéfrences.

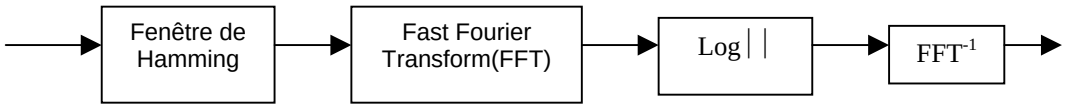


Figure II. 7. Analyse cepstrale sur une fenêtre temporelle

Les deux familles de coefficients cepstraux les plus utilisées en RAP sont issues de deux analyses différentes pour obtenir le spectre (figure II.8). Lorsque le spectre d'amplitude résulte d'une FFT sur le signal de parole pré-traité, lissé par une suite de filtres triangulaires répartis selon l'échelle Mel, les coefficients sont appelés Mel Frequency Cepstral Coefficients (MFCC). L'échelle non linéaire de Mel est donnée par la formule suivante :

$$M_{els}=\frac{1000}{\log_2} Log(1+\frac{F_{hz}}{1000}) \qquad \text{(II.9)}$$

Afin de réduire l'information, une suite de filtres (triangulaires, rectangulaires...) est appliquée dans le domaine spectral selon l'échelle précédemment décrite. Les coefficients obtenus sont alors synonymes d'énergie dans des bandes de fréquence. La figure II.9 donne un exemple de répartition d'une suite de filtres selon l'échelle Mel, couramment utilisée.

Lorsque le spectre correspond à une analyse LPC, les coefficients se déduisent des coefficients LPC par développement de Taylor, d'où leur nom de Linear Prediction Cepstral Coefficients (LPCC).

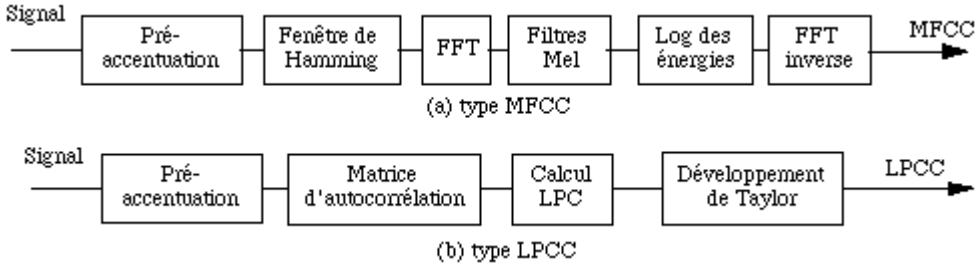


Figure II. 8. Calcul des coefficients cepstraux (source[Jacob, 95])

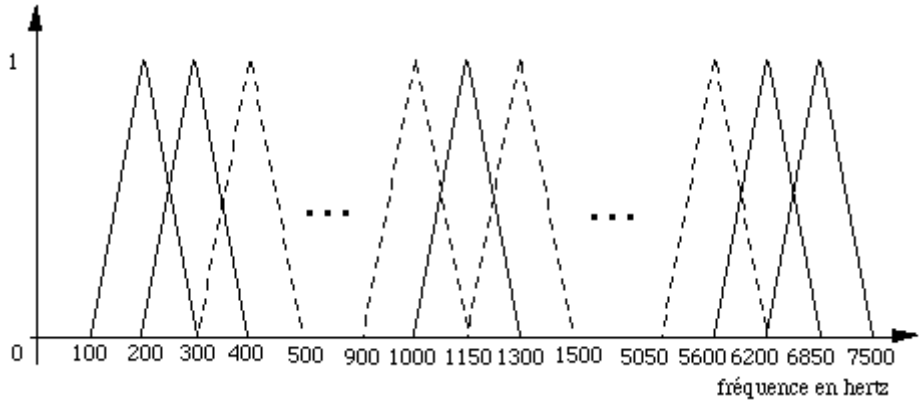


Figure II.9 Répartition fréquentielle de filtres triangulaires (source [Jacob, 95])

A l'issue de l'étape d'analyse du signal, on dispose pour chaque signal vocal en entrée d'un ensemble de vecteurs acoustiques en sortie. Ce sont ces vecteurs de caractéristiques qui seront utilisés dans la suite du processus de reconnaissance.

3. Les approches de reconnaissance de la parole

La reconnaissance automatique de la parole est le processus par lequel la machine tente de « décoder » le signal de la parole qui lui est destiné. Pour cela, on recense trois grandes approches, à savoir : l'approche acoustico-phonétique, l'approche reconnaissance de formes et l'approche intelligence artificielle.

3. 1. L'approche acoustico-phonétique

Historiquement, l'approche acoustico-phonétique (AP) est la première des approches apparues. Elle est basée sur la théorie acoustico-phonétique qui postule que dans un langage parlé, il existe un nombre fini d'unités phonétiques distinctes et que ces unités sont largement caractérisées par un ensemble de propriétés qui se manifestent au travers du signal [Rabiner, 83]. Il s'agit alors de définir une relation entre les caractéristiques spectrales du signal et les unités phonétiques du langage.

Ceci, conduit naturellement à l'émergence d'une étape particulière : l'analyse du signal. L'analyse du signal, est la première étape à effectuer dans cette approche, mais

elle est aussi commune à toutes les approches existantes de la RAP, à l'issue de cette étape, appelée aussi extraction des caractéristiques, on obtient un ensemble réduit de coefficients représentatifs du signal originel.

Le but de l'étape suivante dans l'approche A-P, est d'exploiter ces coefficients afin d'y extraire des caractéristiques qui permettraient de décrire certaines unités phonétiques. Parmi ces caractéristiques, on utilise fréquemment la nasalité (présence ou absence de résonance nasale), localisation des formants (fréquences des trois premières résonances), voisement ou non (excitation périodique ou non), …etc.

Cette phase de détection de caractéristiques consiste en général à mettre en œuvre un ensemble de détecteurs de caractéristiques qui opèrent en parallèle et qui permettent de décider sur la base d'un raisonnement logique de l'absence ou la présence d'une caractéristique.

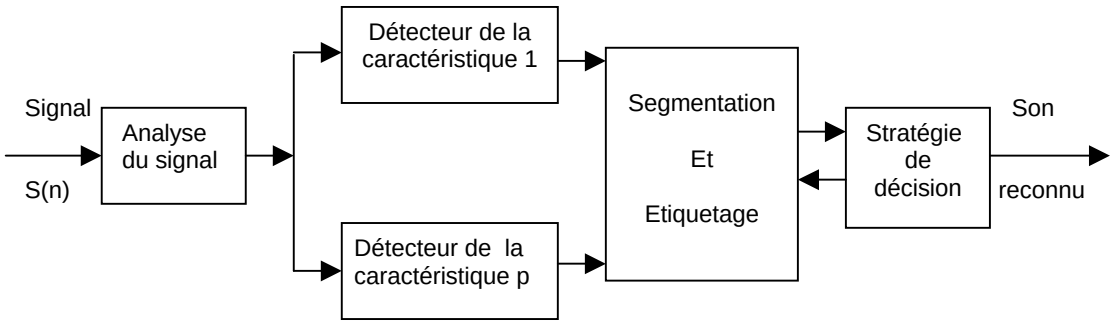


Figure II.10. Système de RAP basé sur l'approche A-P

Bien que cette approche fût étudiée pendant bien longtemps, elle ne rencontra pas sur le plan pratique, le même succès qu'ont connu les autres approches.

3. 2. L'approche reconnaissance de formes

Dans une approche de reconnaissances des formes (RF), la forme du signal (en entend par forme, les coefficients issus de l'analyse) est utilisée directement sans détermination explicite des caractéristiques dans un sens acoustique et phonétique. La plupart des approches de reconnaissance de formes, sont constituées de deux phases : d'abord l'apprentissage et ensuite, la reconnaissance de la forme présentée au système par un processus de comparaison.

Dans de tels systèmes, la connaissance est introduite via le processus d'apprentissage. Dans cette phase, on présente au système un ensemble d'exemples et à la fin de cette phase le système devrait être en mesure de les distinguer correctement. Les formes ainsi apprises par le système sont appelées formes de références.

En phase de reconnaissance, des formes qui n'existaient pas dans le corpus d'apprentissage sont présentées au système, et ce dernier devrait être capable de les caractériser correctement ; cette caractérisation est appelée : classification.

Différentes méthodes s'apparentent à l'approche RF, elles se distinguent principalement, par la manière dont les formes de références sont crées, modélisées et par la méthode qui sert à classer les formes inconnues.

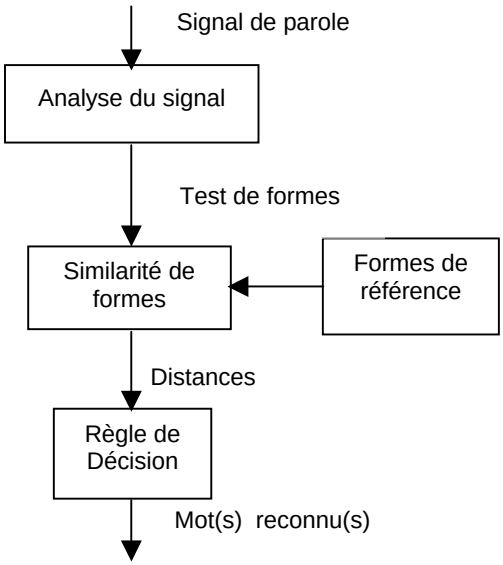


Figure II.11 Diagramme d'un système RAP basé R-F [Rabiner, 81]

On retiendra, comme méthodes de cette approche, d'une part les méthodes structurelles telles que les HMMs, et les méthodes statistiques telles que les réseaux connexionnistes. La plupart des systèmes de reconnaissances actuels sont fondés sur une approche probabiliste et bien que les réseaux connexionnistes aient été largement utilisés en RAP, les HMMs sont très puissants et constituent la base des systèmes présents sur le marché. Un travail de reconnaissance de la parole Arabe que nous avons effectué en utilisant les HMMs est présenté dans [Bahi, 01].

3. 3. L'approche intelligence artificielle

L'approche intelligence artificielle (IA) se définit comme une approche hybride combinant les approches acoustico-phonétique et reconnaissance de formes [Rabiner, 83].

Cette approche, qui inclut l'approc²he RF, tente d'automatiser le processus de reconnaissance en s'inspirant de la manière dont l'être humain utilise son intelligence en visualisant, analysant et finalement en décidant sur la base des données acoustiques dont il dispose. L'idée de base étant d'incorporer différentes techniques et d'intégrer diverses sources de connaissances pour la prise en charge d'un problème donné.

Deux concepts clefs sont inhérents à l'intelligence artificielle ce sont : l'apprentissage et l'adaptation, et l'une des voies par lesquelles ces concepts peuvent être implémentés sont les réseaux connexionnistes. D'autre part, les systèmes experts (SE) constituent l'un des outils les plus représentatifs de l'approche IA, on notera en particulier l'aspect explicabilité de la décision apporté par les SEs.

4. Conclusion : Vers une combinaison de méthodes

De la précédente présentation, il ressort une richesse inéluctable des différentes approches, et dans cet esprit, nous proposons une approche de reconnaissance de la parole qui tente d'exploiter les points forts de chacune d'entre elles. Parmi ces points on note, la capacité de généralisation obtenue grâce à l'apprentissage dans les approches RF, l'aspect symbolique et donc explicable des approches IA, et finalement l'exploitation des paramètres acoustiques des sons pour caractériser des unités phonétiques, issue de l'approche A-P, ce qui est tout à fait en accord avec le modèle biologique de la perception des sons [Caliope, 89].

De l'approche RF on retiendra les réseaux connexionnistes qui furent largement utilisés avec succès en reconnaissance de formes. Le perceptron multicouches, qui représente une architecture particulière des réseaux connexionnistes est un classifieur universel. C'est donc cette topologie des réseaux que nous retenons comme noyau pour notre système de reconnaissance de la parole. Nous lui attachons pour l'enrichir et pallier ces insuffisances, un aspect symbolique sous formes de relations de dépendances

entre objets du domaine de façon à ce que l'architecture reflète la nature du problème auquel il est destiné.

Notre motivation s'inscrit dans un esprit de réduction de la complexité rencontrée lors de la recherche de la configuration du réseau, d'une optimisation du processus d'apprentissage et d'un apport au niveau de l'explicabilité du résultat.

Dans le cadre de la reconnaissance de formes, cette explicabilité devrait aider la décision du système dans le cas d'une ambiguïté à la reconnaissance d'une forme où lorsque plus d'une forme sont candidates à la reconnaissance.

PARTIE 2:

Outils utilisés

Chapitre III:

Les réseaux connexionnistes en RAP

Chapitre III: Les réseaux connexionnistes en RAP

Présentation du chapitre

Dans ce chapitre nous présentons une revue des réseaux connexionnistes. Après un bref historique en guise de toile de fond, nous allons revoir certains concepts fondamentaux, nous décrirons quelques architectures des réseaux connexionnistes ainsi que les procédures d'apprentissage. Dans un second volet de ce chapitre, nous aborderons les possibilités d'utilisation des réseaux connexionnistes dans le cadre de la reconnaissance de la parole, et en particulier la prise en compte du paramètre temps.

1. Un peu d'historique

Au cours du développement des réseaux de neurones artificiels, les chercheurs se sont largement inspirés de leurs homologues biologiques dans leurs structures, leurs agencements et leurs stratégies de véhiculer les informations et de les traiter. En effet, les études modernes sur les réseaux connexionnistes débutèrent au 19ème siècle lorsque les premiers neurobiologistes entreprirent des études sur le système nerveux humain. En 1892, S. Cajal détermina que le système nerveux est composé de neurones qui communiquent entre eux en envoyant des signaux électriques par le biais de leur axones, ces derniers communiquent avec l'extérieur en entrant en contact avec les dendrites (zone réceptrice) de milliers d'autres neurones grâce aux synapses qui sont des points de contact (figure III.1).

Un neurone émet un signal en fonction des signaux qui lui proviennent des autres neurones. On observe en fait au niveau d'un neurone, une *intégration* des signaux reçus au cours du temps, c'est à dire une sorte de sommations des signaux. En général, quand la somme dépasse un certain seuil, le neurone émet à son tour un signal électrique.

La notion de *synapse* explique la transmission des signaux entre un axone et une dendrite. Au niveau de la jonction (c'est à dire de la synapse), il existe un espace vide à travers lequel le signal électrique ne peut pas se propager. La transmission se fait alors par l'intermédiaire de substances chimiques, les *neuro-médiateurs*. Quand un signal arrive au niveau de la synapse, il provoque l'émission de neuro-médiateurs qui vont se fixer sur des récepteurs de l'autre côté de l'espace inter-synaptique. Quand suffisamment de molécules se sont fixées, un signal électrique est émis de l'autre côté et on a donc une transmission. En fait, suivant le type de la synapse, l'activité d'un neurone peut renforcer ou diminuer l'activité de ces voisins. On parle ainsi de synapse excitatrice ou inhibitrice.

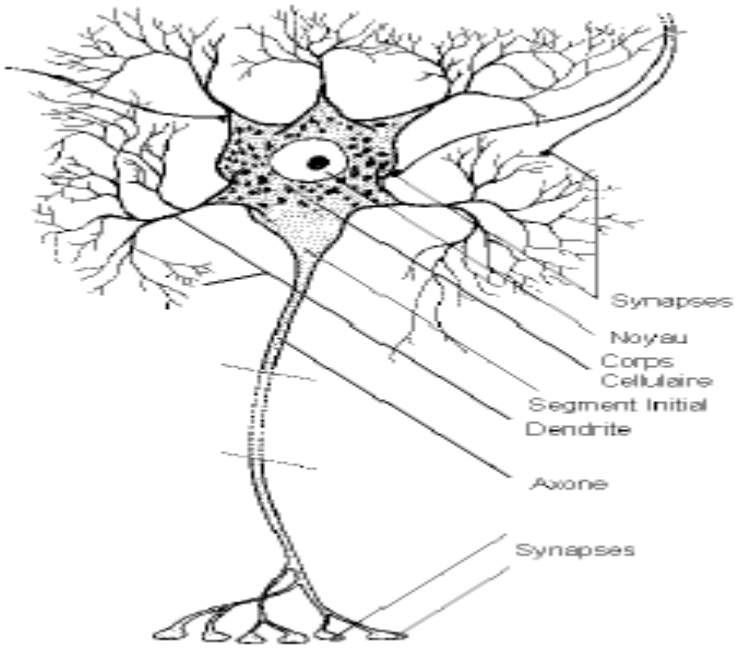


Figure III. 1. Structure d'un neurone biologique

Dans le cadre de la modélisation du neurone humain, les recherches de MacCulloch et Pitts aboutirent à un modèle formel du neurone biologique, appelé neurone binaire à seuil (binary threshold unit) [MacCulloch, 43]. Ce neurone formel est un organe de calcul inspiré du modèle biologique qui s'active si la somme pondérée des valeurs d'entrées provenant des autres neurones dépasse un certain seuil (Figure III. 2.).

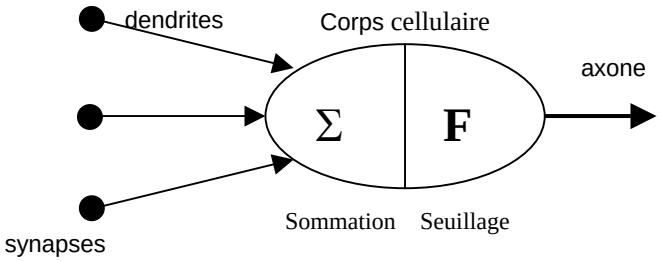


Figure III. 2. Structure d'un neurone formel

Ce modèle causa un grand engouement dès l'instant où il fut montré que de telles unités assemblées en automates d'états finis pouvaient calculer n'importe quelle fonction si l'on trouvait les valeurs adéquates pour les poids entre neurones. Aussitôt, les recherches pour mettre au point de telles procédures d'apprentissage débutèrent. En 1962, Rosenblatt mis au point une procédure d'apprentissage itérative pour un type particulier de réseaux connexionnistes: le perceptron mono couche (single layer perceptron) [Rosenblatt, 58], et il démontra que cette procédure converge toujours vers un ensemble de poids qui calcule la fonction désirée, à condition que la fonction soit potentiellement calculable par le réseau ; la communauté IA commençait alors à croire que la machine intelligente était à sa portée. Mais en 1969, Minsky et Papert montrèrent que l'ensemble des fonctions calculables par un perceptron mono couche est très restreint, et ils exprimèrent un tel pessimisme quant aux potentialités du perceptron mono couche que la conséquence directe de leurs études fût une interruption soudaine des recherches sur le connexionisme, et il en resta ainsi pendant une quinzaine d'années [Jodouin, 94].

Il y'eût un regain d'intérêt pour les réseaux connexionnistes lorsque en bon physicien John Hopfield (1982) suggéra qu'un réseau peut être analysé en termes de fonctions d'énergie [Bishop, 95]. Sa contribution fût de montrer l'applicabilité au connexionisme d'un modèle fort important en physique statistique : les verres de Spin de Ising. Ceci permis également le développement de la machine de Boltzmann (1985), un réseau stochastique qui dans sa formulation originale est fort semblable au réseau de Hopfield, mais elle se singularise entre autre par un mode d'apprentissage supervisé [Jodouin, 94].

Peu de temps après (1986), Rumelhart et al., mirent au point un algorithme beaucoup plus rapide pour l'apprentissage du perceptron multicouches pour calculer n'importe quelle fonction, cet algorithme est appelé : backpropagation (rétro propagation) [Rumelhart, 86]. Cet algorithme mis fin à la thèse pessimiste de Minsky et Papert, et offrit au connexionisme un regain d'intérêt qui perdure à ce jour.

2. Fondements des réseaux connexionnistes

Un réseau connexionniste est composé d'un grand nombre d'unités de calcul qui sont simples, et qui procèdent simultanément au calcul de leurs sorties, ce qui induit un haut degré de parallélisme. A tout moment chaque neurone calcule une fonction scalaire de ses entrées et transmet le résultat à ses voisins.

2. 1. Le neurone formel

2. 1. 1. Le modèle de McCulloch et Pitts

Ce premier modèle formel est un automate booléen, c'est à dire que ses entrées et sa sortie sont booléennes.
Soient :

- e_i (i=1..n), les entrées du neurone,
- S sa sortie,
- θ son seuil,
- w_i les paramètres de pondération,
- f, la fonction de seuillage avec : f(x)=1 si x>θ et f(x)=0 si x≤θ

La sortie S du neurone est alors activée si la somme des entrées e_i pondérée par les poids w_i dépasse le seuil θ. Soit $S = f(\sum_{i=1}^{n} w_i e_i)$ (III. 1)

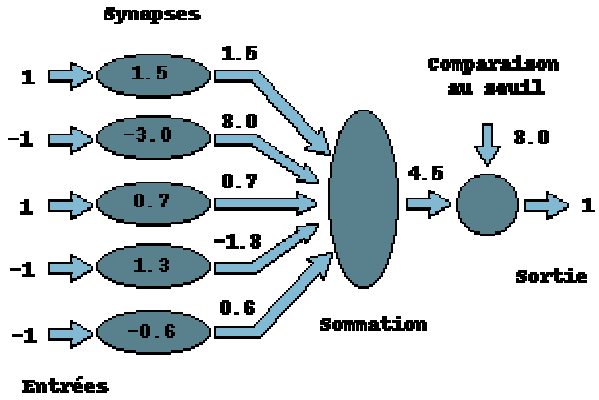


Figure III. 3. Structure du neurone de MacCulloch et Pitts

2. 1. 2. Le modèle général

De manière générale, un neurone formel peut être définit par :

- La nature de ses entrées : e_i (i = 1..n), ces entrées peuvent être binaires ou réelles.
- La fonction d'entrée totale notée h, qui définit le pré-traitement effectué sur les entrées. Cette fonction peut être booléenne, linéaire, affine ou polynomiale.
- La fonction d'activation du neurone notée f, qui définit son état interne en fonction de son entrée. f peut être une fonction binaire à seuil, une fonction linéaire à seuil ou multi-seuils, une fonction sigmoide, ou une fonction stochastique. On retiendra que toute autre fonction croissante et impaire peut être choisie.
- La fonction de sortie notée g, qui calcule la sortie en fonction de l'état interne. On notera que très souvent la fonction de sortie, est identique à la fonction d'activation.

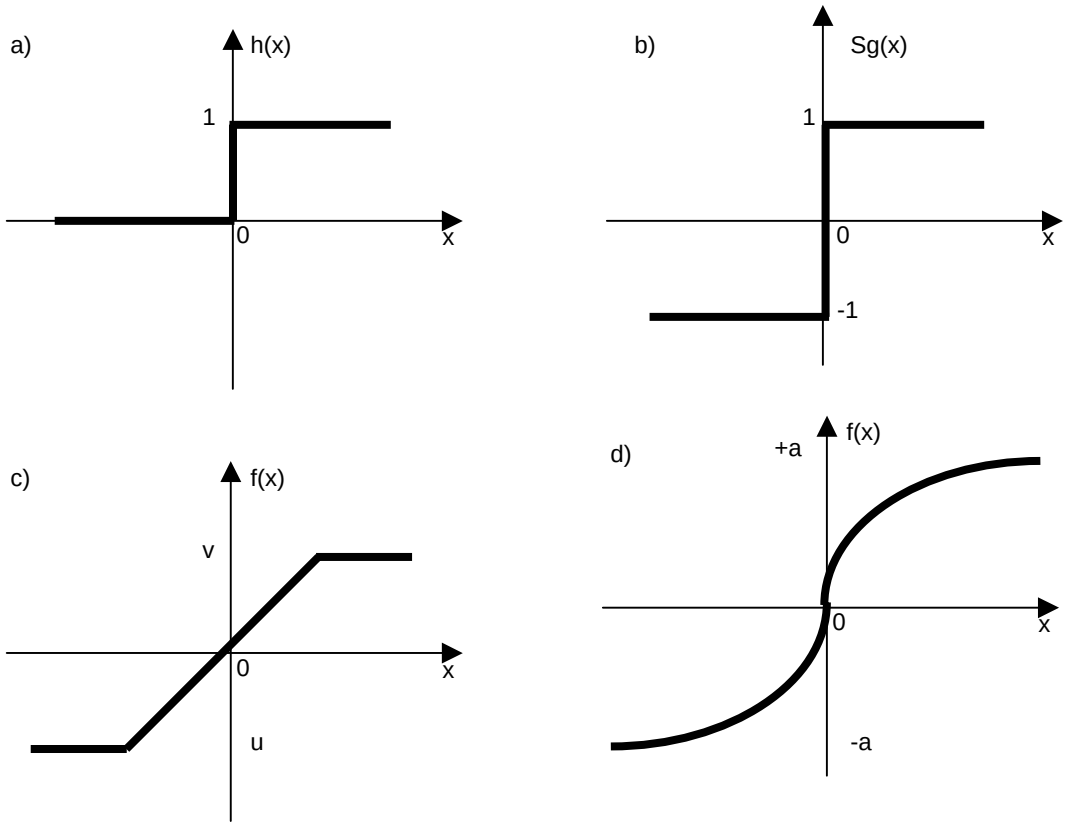


Figure III. 4. Différentes fonctions d'activation

a) fonction Heaveside
b) fonction signe
c) fonction linéaire à seuil
d) fonction sigmoïde $f(x) = a(e^{kx}-1) / (e^{kx}+1)$

Sur le plan fonctionnel, un neurone induit deux étapes de traitement, d'abord il s'agit de calculer l'entrée du neurone, ensuite sa sortie ou activation en tant que fonction des entrées. Communément pour un neurone j l'entrée est

$$x_j=\sum_i y_i w_{ji} \qquad \text{(III. 2)}$$

où y_i est la sortie des neurones émetteurs et w_{ji} le poids de la connexion du neurone i vers le neurone j.

Dans le cas général, un biais θ est ajouté à cette somme d'où :

$$x_j=\sum_i y_i w_{ji}+\theta_j \qquad \text{(III. 3)}$$

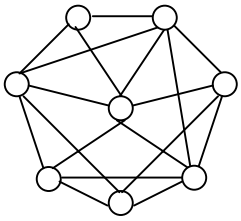
Ce biais est généralement considéré comme le poids d'un autre neurone fictif dont l'activation est $y_0=1$, il est automatiquement inclus dans l'équation précédente. Une fois x_j calculée, l'activation du neurone y_j est calculée en utilisant la fonction de transfert.

2. 2. Les connexions

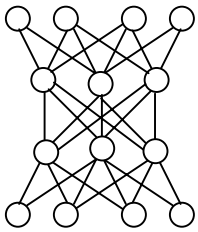
Les connexions entre neurones sont porteuses de poids qui peuvent varier de -∞ à +∞, la valeur d'un poids représente l'influence du neurone par rapport à son voisin, ainsi un poids positif traduit un lien excitateur tandis qu'un poids négatif représente un lien inhibiteur. Ces poids sont généralement unidirectionnels (d'un neurone entrée vers un neurone sortie). Les valeurs des poids du réseau déterminent la réaction du réseau à toute forme en entrée du réseau connexionniste ; ainsi ces poids représentent la mémoire à long terme ou les connaissances du réseau. Ces poids changent de valeurs sous l'effet de l'apprentissage, mais ce changement tend à être de plus en plus lent du fait de l'accumulation des connaissances.

2. 3. Topologies des réseaux connexionnistes

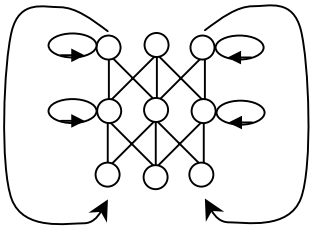
Un réseau connexionniste est constitué d'un nombre important de connexions entre les éléments de calcul simples que sont les neurones. Mais, c'est le comportement émergeant du réseau qui présente une grande complexité. Les structures qui peuvent être utilisées pour agencer les neurones dans un réseau sont très variées. Les topologies les plus représentatives étant les réseaux non structurés, multicouches, récurrents et modulaires (figure III. 5).



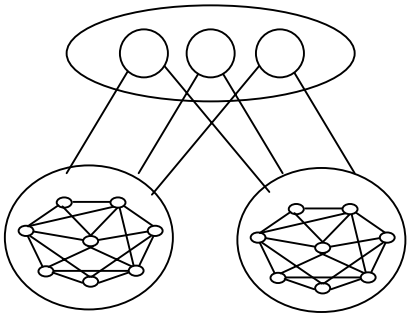
(a) réseau non structuré



(b) réseau à couches



(c) réseau récurrent



(d) réseau modulaire

Figure III. 5 Différentes topologies des réseaux connexionnistes

- Les réseaux non structurés sont très utilisés pour retrouver une forme stockée en invoquant n'importe quelle partie de la forme,

- Les réseaux à couches sont très utilisés dans les problèmes d'association,

- Les réseaux récurrents sont utilisés pour le séquencement de formes (i.e., suivre des séquences d'activation du réseau au travers du temps),

- Les réseaux modulaires sont utilisés pour la construction de systèmes complexes à partir de composants plus simples. Ces réseaux peuvent intégrer différentes topologies.

On retiendra également que des chercheurs ont montré que le cortex est divisé en plusieurs couches. Les connexions entre les neurones d'une même couche sont très grandes, mais les neurones sont également reliés aux autres couches, ce qui induit une grande complexité du réseau. De ce fait, l'une des structures classiques les plus utilisées est le réseau multicouches.

2. 4. Taxonomie des réseaux connexionnistes

Dans leur traitement de l'information les réseaux connexionnistes invoquent deux phases: une phase d'apprentissage et une phase d'exploitation. En phase d'apprentissage les données d'apprentissage sont utilisées pour déterminer les poids du réseau. Le réseau entraîné sera utilisé ultérieurement pour produire les résultats escomptés. C'est en référence au type d'apprentissage qu'est établie en général la taxonomie des réseaux connexionnistes. Il existe trois classes de procédure d'apprentissage [Tebelski, 95] :

Apprentissage supervisé : Dans ce style d'apprentissage, on fournit au réseau la sortie désirée pour chaque forme en entrée, ce qui permet de corriger explicitement l'erreur commise lors de l'activation.

Apprentissage semi-supervisé : dans ce type d'apprentissage, on ne donne pas au réseau les sorties désirées mais une évaluation de celles-ci en terme de bonne ou mauvaise approximation.

Apprentissage non supervisé : dans ce cas le réseau doit détecter par lui même les régularités dans les données en entrée. De tels réseaux auto-organisateurs sont utilisés en compression, quantification ou classification des données en entrée.

La plupart des réseaux appartiennent à l'une de ces catégories, mais il existe des réseaux hybrides, et des réseaux dynamiques dont l'architecture change au travers du temps.

3. Les mécanismes d'apprentissage

3. 1. L'apprentissage

Entraîner un réseau de neurones signifie mettre à jour ses connexions de sorte que pour toute forme en entrée, le réseau associe la sortie adéquate.

Trouver l'ensemble de poids qui permettrait à un réseau donné de calculer une fonction donnée est une procédure non triviale. Dans le cas très simple d'association de formes dans un réseau linéaire, les poids sont donnés par la formule :

$$\Delta w_{ji} = \sum_{p} \frac{y_i^p t_j^p}{\left\| y^p \right\|^2} \qquad \text{(III. 4)}$$

Malheureusement, les réseaux sont généralement non linéaires et multicouches, leurs poids peuvent alors être mis à jour en utilisant une procédure itérative telle que la déscente du gradient (Hinton 1989). Ceci nécessite plusieurs itérations ou époques sur l'ensemble des données d'apprentissage. Toutefois, comme la connaissance apprise est distribuée au travers de tous les poids, ces derniers devraient être modifiés très soigneusement de manière à ne pas perdre les acquis précédents. Une petite constante appelée vitesse d'apprentissage (ε) est ainsi introduite pour contrôler la magnitude des modifications de poids ; si la valeur ε est trop petite l'apprentissage est trop long, mais

si elle est importante les nouvelles connaissances perturbent les anciennes. Comme il n'y a pas de méthode analytique pour déterminer ε, elle est souvent choisie de manière empirique après plusieurs essais.

3. 2. La règle de Hebb (1949)

La plupart des procédures d'apprentissage, sont essentiellement des variations de la règle de Hebb (1949), qui consiste à renforcer la connexion entre deux neurones si leurs activations sont corrélées [Hebb, 49]. Cette règle se formule de la manière suivante:

« quand une cellule A excite de manière répétée une cellule B,
l'efficacité de A à exciter B est améliorée par des changements métaboliques »

Considérons un neurone formel N possédant les entrées pré-synaptiques $x_i, i = 1..n$; les poids de connexion $\omega_i, i = 1..n$; et une sortie post-synaptique *y* calculée selon l'équation :

$$y = f\left(\sum_{i=1}^{n} w_i.x_i\right) \qquad \text{(III. 5)}$$

avec *f* une fonction sigmoïde.

La règle de Hebb peut alors se formuler mathématiquement de la façon suivante :

$$w_i^{(t+1)} - w_i^{(t)} = a.y.x_i \qquad \text{(III.6)}$$

avec *a* une constante définissant le pas d'apprentissage.

On constate donc que le mécanisme d'apprentissage s'attache à mettre en relief les corrélations entre le signal de sortie et les signaux d'entrée. En effet, plus le signal de sortie *y* et le signal d'entrée x_i sont similaires, plus le poids ω_i sera augmenté ce qui renforcera l'influence de l'entrée x_i dans la réponse *y*. Toutefois la règle de Hebb n'est pas suffisante pour rendre compte des expériences de conditionnement classique [Durand, 95].

3. 3. La règle de Widrow-Hoff

Une importante variante de la règle de Hebb est la règle de Widrow-Hoff (ou la règle Delta), cette règle est appliquée lorsque l'un des deux neurones possède une sortie désirée [Widdrow, 60]. La règle consiste à renforcer la connexion entre deux neurones s'il y a corrélation entre l'activation du premier neurone et l'erreur du second (ou la possibilité de réduire l'erreur) par rapport à sa sortie désirée t_j :

$$\Delta w_{ji}=\varepsilon y_i(t_j - y_j). \qquad \text{(III.7)}$$

Cette règle permet de réduire l'erreur si y_i, y contribue. Dans le contexte des neurones booléens à seuil, en réseau monocouche, la règle Delta est connue comme le « la règle du perceptron » (perceptron learning rule), dans le contexte des réseaux multicouches (voir §4.2), cette règle est à la base de l'algorithme de rétropropagation.

3. 4. L'algorithme de rétropropagation

Cet algorithme qui est désigné souvent par « backpropagation » ou « error backpropagation », est une généralisation de la règle de Widrow-Hoff pour un réseau multicouches, il est largement utilisé lors de l'apprentissage supervisé [Rumelhart, 86].

L'idée simple qui est à la base de cet algorithme est l'utilisation d'une fonction dérivable (fonction sigmoïde) en remplacement de la fonction à seuil utilisée dans les neurones linéaires à seuil. Dans cet algorithme la propagation du signal se fait des cellules d'entrée vers la couche de sortie, mais aussi peut en suivant le chemin inverse rétro propager l'erreur commise en sortie vers les couches internes.

4. Le Perceptron multicouches

4. 1. Le Perceptron originel

En 1958, Rosenblatt décrit le premier perceptron [Rosenblatt, 58], ce modèle neuromémitique de perception était fondé sur les connaissances neurophysiologiques de l'époque.

L'architecture est une rangée de neurones de MacCulloch et Pitts qui apprennent chacun une fonction de transfert. On y trouve également une rétine composée de neurones dont le rôle est de transcrire un stimulus, et une rangée d'association qui effectuent un pré-traitement sur la rétine (figure III.6). Les neurones de sortie effectuent une somme pondérée de leurs entrées et calculent ensuite leur activité en passant cette somme par une fonction de Heaviside ou une fonction signe.

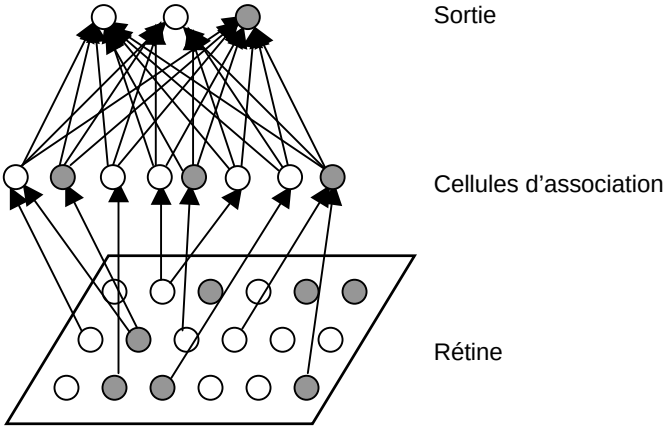


Figure III. 6. Perceptron mono-couche

La modification des poids est supervisée et s'effectue grâce à la règle de Hebb établie comme produit entre l'erreur calculée au niveau du neurone de sortie (différence entre sortie calculée et sortie désirée) et la valeur d'entrée pour le poids considéré. Le théorème de convergence du perceptron montre que s'il existe un jeu de poids capable de simuler la transformation, alors la règle d'apprentissage convergera vers une solution en un nombre de pas fini.

4. 2. Le perceptron multicouches (MLP: MultiLayer Perceptron)

Le perceptron mono-couche n'est capable de faire qu'une séparation linéaire de ses entrées. Pour obtenir des classes de formes convexes, plusieurs couches sont nécessaires (figure III.7). Cependant la procédure d'apprentissage des poids définie pour le perceptron mono-couche n'est plus valide ici puisque la sortie désirée n'est pas connue au niveau des neurones cachés.

4. 2. 1. Structure du réseau

Comme son prédécesseur, le perceptron multicouches se structure en couches. La première couche du réseau qui constitue la couche d'entrée, correspondrait aux organes sensoriels de l'homme [Bishop, 95]. La dernière couche est la couche de sortie qui fournit le résultat du traitement effectué par le réseau. Les couches intermédiaires sont appelées : couches cachées.

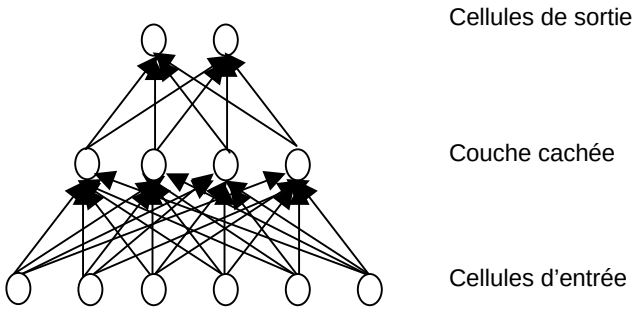


Figure III. 7. Perceptron multicouches

La connectivité de ce réseau est restreinte, ainsi pour réduire la complexité du réseau, il est convenu que les neurones d'une même couche ne sont pas connectés entre eux, et que chaque couche dans le réseau reçoit les informations de la couche précédente et les transmet à la couche suivante.

4. 2. 2. L'apprentissage

Le calcul de l'activation dans le réseau se fait en propageant l'activation initiale de la couche d'entrée jusqu'à la couche de sortie. L'erreur est calculée dans le sens inverse. Puisque le réseau ne possède pas de boucles, ces deux calculs peuvent être réalisés chacun en une passe unique dans le réseau, et ne requièrent donc qu'un temps proportionnel au nombre de liens synaptiques.

En général, les neurones du MLP sont mus par une fonction d'activation sigmoïde, exponentielle ou tangentielle.

Le perceptron multicouches utilise une règle d'apprentissage appelée la retropropagation de l'erreur, qui est proche de la règle delta de Widrow-Hoff. En effet, comme cette dernière, la rétropropagation est une technique de descente de gradient qui minimise la distance entre la sortie du réseau et la sortie désirée.

5. Les réseaux connexionnistes en RAP

La reconnaissance de la parole est fondamentalement un problème de reconnaissance de formes. Par ailleurs, les réseaux connexionnistes qui sont un outil puissant de la reconnaissance des formes, s'avérent être de bons classifieurs [Bishop, 95], de ce fait, de nombreux travaux se sont orientés vers l'utilisation des réseaux connexionnistes en RAP. Les tous premiers essais effectuaient des tâches très simples telles que : classer des segments de parole en son voisé / non voisé, ou nasal / fricatif / plosif. Le succès remporté par ces réseaux a énormément encouragé les chercheurs à considérer la classification des phonèmes ; ce qui fût effectué avec succès. Les mêmes techniques rencontrèrent quelques succès pour la reconnaissance des mots.

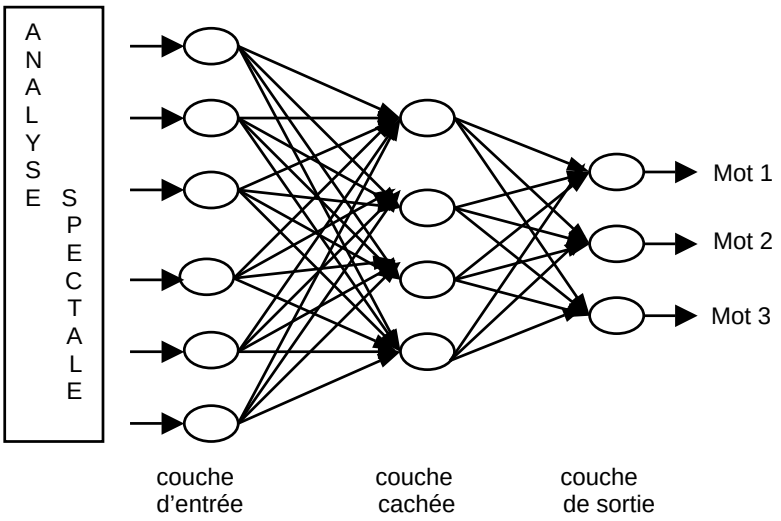


Figure III. 8. Un MLP pour la reconnaissance de mots

5. 1. Les réseaux connexionnistes et le temps

Globalement, il y'a deux approches de classification pour le signal de la parole, lorsqu'on utilise des réseaux connexionnistes : une approche statique et une approche dynamique. Dans la classification statique, le réseau connexionniste considère le signal dans sa totalité et prend une décision unique. La classification dynamique ne considère qu'une fenêtre du signal, et de ce fait prend des décisions locales, qui seront intégrées par la suite dans une décision globale. La classification statique suffit tout à fait pour la

reconnaissance des phonèmes alors que la classification dynamique est requise pour avoir de bons résultats dans la reconnaissance des mots et phrases.

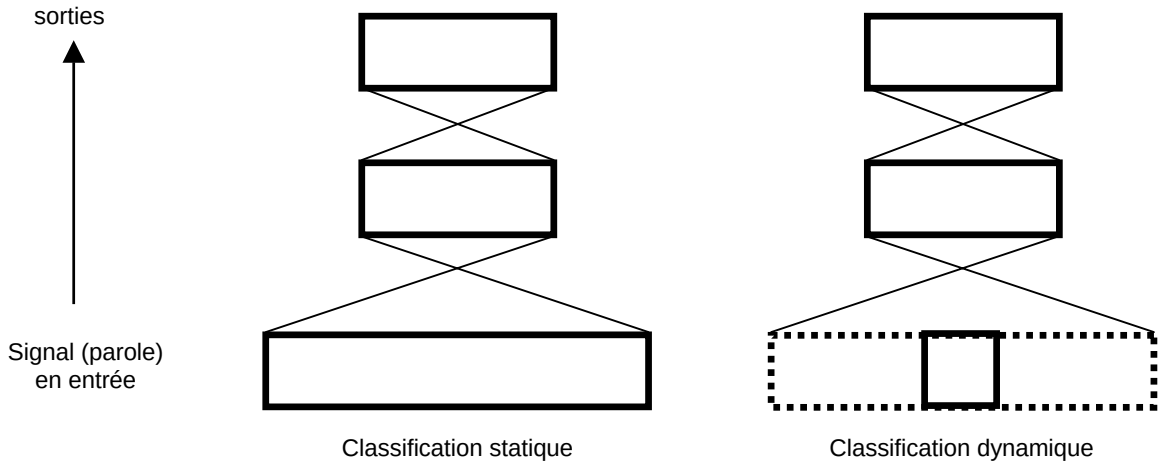


Figure III.9. Approches statique et dynamique de classification

Dans les sections suivantes, nous allons présenter une taxonomie des réseaux connexionnistes en RAP en nous basant sur la manière avec laquelle ils implantent l'aspect temps.

5. 2. L'approche statique

L'utilisation de réseaux connexionnistes pour la classification de formes statiques (images, caractères écrits, etc.) est très répandue. C'est également le cas pour la parole, dans ce cas, l'unité à reconnaître (mot isolé ou unité sub-lexicale) est considérée comme une forme acoustique globale présentée en entrée du modèle connexionniste.

Une illustration de cette approche est l'expérimentation menée par Huang et Lippman en 1988 [Huang, 88] pour la reconnaissance de phonèmes, elle montra que les réseaux connexionnistes pouvaient trouver des surfaces de décisions complexes à partir des données de parole. Ils utilisèrent pour cela un MLP avec deux (2) entrées que sont les deux premiers formants, 50 neurones cachés et 10 sorties qui représentent les 10 phonèmes à reconnaître. Après 50.000 itérations le réseau produit les surfaces de décisions de la figure III.10, ces surfaces sont quasi optimales, comparativement à une classification humaine.

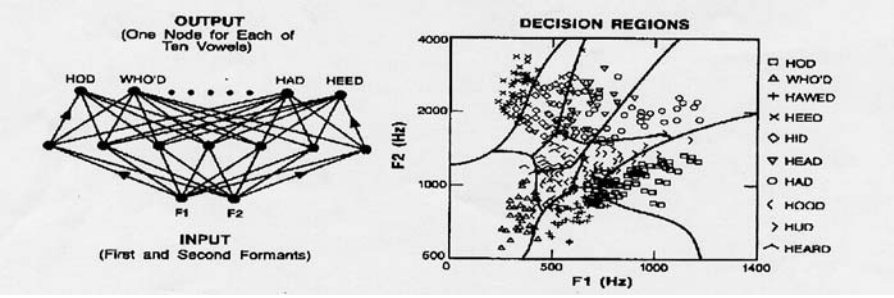


Figure III. 10 Régions de décision formées par un MLP à 2 couches (source [Huang, 88])

En 1987, Peeling et Moore ont utilisé un MLP pour la reconnaissance des chiffres avec d'excellents résultats [Peeling, 87]. Ils utilisèrent une entrée statique de 60 fenêtres (1.2 sec.) de coefficients spectraux ce qui correspond à l'occurrence la plus longue d'un mot ; les signaux de mots plus courts sont complétés par des zéros positionnés de manière aléatoire parmi les 60 fenêtres. Ils ont ensuite testé de nombreuses topologies de MLP, et ont obtenu la meilleure performance avec un réseau à une couche cachée de 50 neurones. Ce réseau réalisa des performances très proches de celles obtenues avec un système basé sur des HMMs.

Le modèle connexionniste le plus fréquemment utilisé dans l'approche statique est le perceptron multicouches. En effet, de tels systèmes sont capables d'apprendre des fonctions de décision fortement non linéaires, ce qui est fondamental pour la reconnaissance de formes complexes telles que des mots ou des unités sub-lexicales. Les performances obtenues par de tels systèmes pour de petits vocabulaires sont bonnes et se comparent même favorablement à celles obtenues par des systèmes à base de HMMs [Tebelski, 95]. En revanche, la méthode est difficilement adaptable à de grands vocabulaires et à la parole continue [Haton, 91].

Mais outre le MLP que nous venons de voir et qui est le type de réseau connexionniste le plus utilisé dans cette approche, il est aussi possible d'utiliser d'autres topologies telles que les cartes auto-organisatrices de Kohonen [Kohonen, 84].

5. 3. L'approche dynamique

Dans ce cas, le temps est vu sous l'aspect séquentiel, on considère donc l'ordonnancement de l'information sous forme de séquences. En fonction d'une telle séquence d'informations, le réseau connexionniste peut fournir ou prédire une réponse. Les informations composant les séquences sont des événements que l'on qualifiera de ponctuels. Ces évènements sont obtenus par un pré-traitement du signal.

Selon la méthode de prise en charge des séquences d'évènements par le réseau, se dégagent deux grandes classes de réseaux connexionnistes temporels [Durand, 95]. La première classe inclut les modèles pour lesquels la séquence est représentée dans sa totalité à l'extérieur du réseau, ils peuvent alors être apparentés à des classifieurs de formes statiques. La seconde classe correspond aux modèles qui intègrent le temps à l'intérieur de leurs architectures. On distingue alors deux sous-types de représentation qui distinguent le traitement temporel par la succession des états stables du réseau (représentation implicite) du traitement temporel par propagation d'activité (représentation explicite).

5. 3. 1. Modèles à représentation externe

C'est le mode de représentation le plus couramment utilisé actuellement. La technique consiste à « spatialiser » le temps, donc à considérer le paramètre temporel comme une dimension équivalente aux autres. Le modèle le plus représentatif dans ce cas serait les réseaux multicouches à retard [Waibel, 89].

5. 3. 2. Modèles à représentation interne implicite

Dans ce type de représentation, les vecteurs composant la forme à étudier sont intégrés par le réseau au cours du temps. Dans ce cas, le temps n'est pas explicitement codé dans le réseau ; le passé ou des morceaux du passé sont codés par des états internes du système.

L'exemple type de ce mode de représentation sont les réseaux récurrents. Dans le cas d'un réseau multicouches, l'état d'un neurone est fonction du vecteur d'entrée courant et de l'état du réseau à l'instant précédent. Ces réseaux apportent alors une réponse plus générale que les réseaux à retard au problème du traitement du temps.

5. 3. 3. Modèles à représentation interne explicite

L'avancement des recherches en neurophysiologie permet le développement de nouveaux modèles neuromimétiques, notamment en ce qui concerne le traitement explicite de l'information temporelle. Parmi ces modèles, on peut citer la colonne de mémorisation à court terme [Ans, 90], la triade synaptique [Dehaene, 87], ou la colonne corticale [Burnod, 88] [Guyot, 89].

Le modèle de la colonne corticale a servi de base à la conception d'une carte neuronale (TOM, *Termporal Organization Map*) qui a été testée avec succès en reconnaissance de la parole [Durand, 95].

6. Quelques architectures de réseaux connexionnistes utilisées en RAP

Outre le perceptron multicouches classique (figure III. 8), d'autres architectures de réseaux connexionnistes ont été utilisées, dans ce qui suit nous citons quelques unes.

6. 1. La carte auto-organisatrice de Kohonen

La carte auto-organisatrice de Kohonen (Self Organizing Map SOM) [Kohonen, 89] [Web 04] effectue une compression sur les données reçues en entrée en représentant des signaux de dimension *N* dans un espace de dimension moindre. La propriété la plus intéressante de cette carte est la faculté de préserver les relations topologiques des signaux d'entrées. Il est intéressant de noter que des fonctions similaires sont présentes dans les cerveaux humain et animal ce qui est un premier point en faveur de la préservation de la topologie comme un mécanisme essentiel dans le traitement du signal.

La carte se présente sous la forme d'une grille de cellules. Un signal d'entrée $\vec{S}^{(t)}$ est présenté à toutes les cellules de la grille. Chaque cellule réagit préférentiellement à un signal prototype $\overrightarrow{\mathrm{Pr}}$ spécifique. Des interactions de voisinage entre cellules de la grille sont matérialisées par des connexions latérales. Ces interactions locales assurent

un mécanisme de compétition entre les cellules pour favoriser la cellule la plus excitée par le signal d'entrée $\vec{S}^{(t)}$.

Ainsi, pour calculer les excitations consécutives à la présentation aux cellules de la grille du signal d'entrée $\vec{S}^{(t)}$, on calcule la distance δ_{ij} séparant ce signal des vecteurs prototypes $\overrightarrow{\mathrm{Pr}}_{ij}^{(t)}$ des cellules c_{ij} conformément à l'équation (III.8) :

$$\delta_{ij}(\vec{S}^{(t)}) = \left\| \vec{S}^{(t)} - \overrightarrow{\mathrm{Pr}}_{ij}^{(t)} \right\| \qquad \text{(III.8)}$$

A partir de cette distance, deux méthodes s'offrent à nous pour mettre en œuvre le mécanisme de compétition qui va favoriser les cellules les plus excitées :

- par un calcul local, en utilisant les influences exercées par les neurones les uns sur les autres (via les liens latéraux), qui fait émerger une activité en chapeau mexicain autour des neurones les plus actifs,
- par application *a posteriori* sur le voisinage du neurone le plus actif d'une activité en chapeau mexicain (algorithme du '*winnner takes all*')

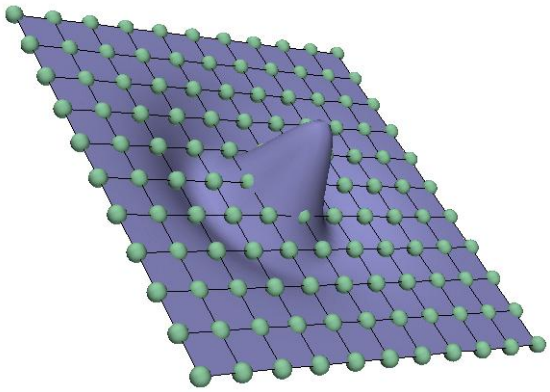


Figure III. 11 Activité en chapeau mexicain autour du neurone le plus excité (ici, le neurone central de la grille).

L'un des points forts de la carte de Kohonen (et des cartes auto-organisatrices en général) est de conserver les relations topologiques entre les différents vecteurs de l'espace d'entrée [jodouin, 94]. Ainsi, des prototypes proches sur la grille sont proches dans l'espace d'entrée ce qui autorise une certaine résistance au bruit pour les traitements qui s'appuient sur ces prototypes. Toutefois, outre le nombre d'itérations à effectuer pour obtenir un apprentissage efficace, le réglage du rayon et du voisinage en chapeau mexicain dépend fortement des données à traiter. De plus, la topologie de

l'espace d'arrivée (généralement bidimensionnelle) peut être très différente de la topologie de l'espace d'entrée du fait de la définition statique de la configuration de la grille.

6. 2. Le TDNN (Time Delay Neural Network)

Les réseaux multicouches à retard, ou *Time-Delay Neural Networks* (TDNN) [Waibel, 88], est à la base un MLP mais il se singularise d'un perceptron multicouches classique 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.

6. 2. 1. La structure du réseau

Les TDNNs sont constitués comme les MLPs d'une couche d'entrée, de couches cachées et d'une couche de sortie mais ils se différencient de part l'organisation des liaisons inter-couches. Les TDNNs introduisent des contraintes qui leurs permettent d'avoir un certain degré d'invariance par décalage temporel et déformation. Celles-ci utilisent trois idées : poids partagés, fenêtres temporelles et délais.

- Les poids partagés permettent de réduire le nombre de paramètres du réseau connexionniste et induisent ainsi une capacité de généralisation plus importante. Les poids sont partagés suivant la direction temporelle, c'est à dire que pour une caractéristique donnée, la fenêtre associée à celle-ci aura les mêmes poids selon la direction temporelle. De plus cette contrainte entraîne une capacité d'extraire les différences au fur et à mesure du balayage du signal. Ce concept de poids partagés est le comportement présumé du cerveau humain où plusieurs neurones calculent la même fonction sur des entrées différentes.
- Le concept de fenêtre temporelle implique que chaque neurone de la couche $l+1$ n'est connecté qu'à un sous ensemble de la couche l (nous n'avons plus une connectivité totale). La taille de cette fenêtre est la même entre deux couches données. Cette fenêtre temporelle permet que chaque neurone n'ait qu'une vision

locale du signal, il peut être vu comme une unité de détection d'une caractéristique locale du signal.

- En plus des deux contraintes précédentes, les TDNNs introduisent des délais entre deux fenêtres successives pour une couche donnée.

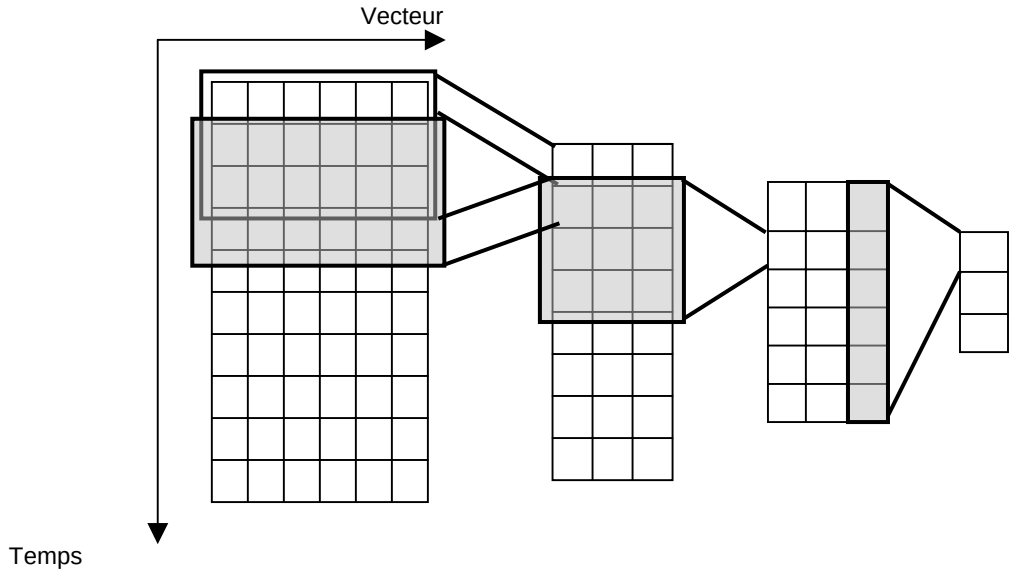


Figure III.12 Architecture du TDNN

6. 2. 2. Le fonctionnement

Le but du TDNN est non pas d'apprendre basiquement le signal mais d'extraire les caractéristique de celui-ci. La première couche acquière le signal puis une ou plusieurs couches cachées transforment le signal en des vecteurs de caractéristiques. Un neurone donné détecte une caractéristique locale de la variation de la courbe. Le champ de vision du neurone est restreint à une fenêtre temporelle limitée. Avec la contrainte des poids partagés, le même neurone est dupliqué dans la direction temps (la même matrice de poids dupliqués) pour détecter la présence ou l'absence de la même caractéristique à différents endroits le long du signal. En utilisant plusieurs neurones à chaque position temporelle, le réseau connexionniste effectue la détection de caractéristiques différentes: les sorties des différents neurones produisent un nouveau vecteur de caractéristiques pour la couche supérieure.

La composante temporelle du signal d'origine est peu à peu éliminée au fur et à mesure de sa transformation en caractéristiques par les couches supérieures, pour compenser cette perte d'informations on augmente le nombre de neurones dans la direction caractéristique.

6. 2. 3. L'apprentissage

L'apprentissage des poids de connexions du TDNN utilise l'algorithme de rétro-propagation classique. Cependant, du fait qu'un neurone pour pouvoir calculer complètement son activation, a besoin d'attendre l'information sur un certain pas de temps, il n'est pas possible de modifier les poids à tout instant entre la couche de sortie et la couche d'entrée. C'est la raison pour laquelle avant de « lancer » l'algorithme de rétropropagation, il y a mémorisation des colonnes nécessaires à l'activation des autres colonnes et duplication des poids qui sont donc partagés. La figure (III.13) montre la mémorisation de colonnes entre deux couches ainsi que la duplication associée des poids.

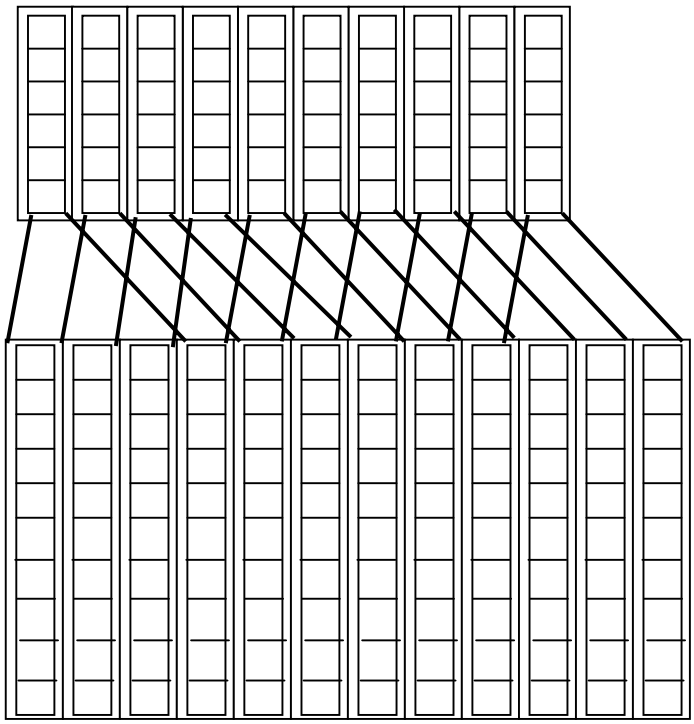


Figure III.13 stockage des informations au cours du temps pour l'apprentissage des poids du TDNN

6. 2. 4. Le TDNN et la reconnaissance de la parole

Dans le premier article concernant le TDNN [Waibel, 89], Waibel et al. décrivent une application à la reconnaissance des trois plosives voisées dans le vocabulaire japonais, /b/,/d/ et /g/. Huit cent phonèmes produits par trois locuteurs masculins constituent la base d'apprentissage et de test. Pour chaque phonème en contexte, le TDNN utilise 15

vecteurs de parole contenant chacun 16 coefficients issus de la FFT sur une fenêtre de Hamming de 256 points sur le signal de parole. Cette transformée est calculée toutes les 5 ms. Le protocole d'apprentissage et de test se situe dans le domaine mono-locuteur. Les tests effectués donnent l'avantage au TDNN avec 98.5% de reconnaissance correcte sur le corpus de test face aux HMMs conventionnels avec 93.7% dans les même conditions d'expérience. Le modèle possède plusieurs couches qui intègrent l'information de façon de plus en plus abstraite. La structure d'un phonème est différente d'un contexte à l'autre, ainsi la configuration des 15 vecteurs d'entrée est donc différente pour deux contextes différents. On constate toutefois que l'effet de contexte s'estompe dans la configuration d'activités de la dernière couche cachée du réseau.

Le TDNN a montré sa capacité de reconnaissance avec des décalages temporels au niveau du signal d'entrée. Cependant, il conserve les limitations des MLPs pour le problème de la distorsion temporelle. Pour des applications à la reconnaissance de séquences acoustiques plus longues que celles correspondant à des phonèmes, des systèmes composés de plusieurs TDNNs ont été mis en œuvre.

6. 3. Les réseaux récurrents

Différents réseaux récurrents ont été conçus et testés en reconnaissance de la parole. Ces réseaux sont le plus souvent seulement partiellement récurrents et constitués d'un perceptron multicouches avec une boucle de retard permettant de réinjecter à l'entrée la valeur de la couche de sortie et, parfois, des couches cachées. Voici quelques modèles qui rentrent dans cette classe d'architectures.

6. 3. 1. Le modèle de Jordan

L'architecture définie par Jordan [Jordan, 86] a pour base un perceptron multicouches à une seule couche cachée. Il y a donc interconnexion totale entre les couches successives. De plus, les neurones de la couche de sortie sont connectés à une partie des neurones de la couche d'entrée (figure III.14).

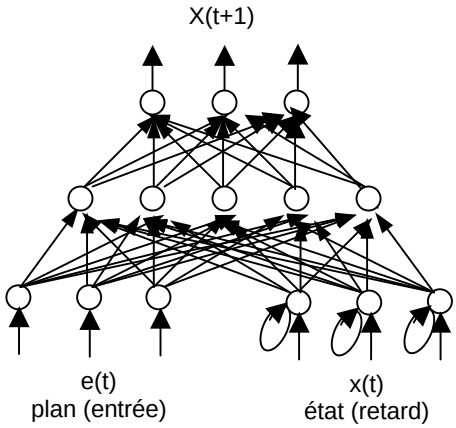


Figure III.14. Modèle de Jordan

Ces neurones correspondent aux cellules (x_t) codant les états du réseau. Les connexions récurrentes entre ces deux groupes de neurones sont pratiquement des lignes de retard permettant au réseau de disposer à un instant t de sa sortie à l'instant t-1, elles sont valuées à 1. De plus chacun des neurones d'état possède une boucle récurrente. Les poids des connexions sont fixes tout au long de l'apprentissage qui utilise une généralisation de l'algorithme de rétropropagation du gradient de l'erreur. Les neurones restant sur la couche d'entrée définissent le plan de la séquence. L'ensemble des neurones définissant ce plan reste figé durant la génération de la séquence, il constitue en quelque sorte la clef de la séquence. A partir de cette clef et de l'état courant, le réseau génère un nouvel état qui est utilisé à l'instant suivant grâce aux connexions récurrentes.

6. 3. 2. Le modèle de Elman

Le modèle de Elman [Elman, 90], dont l'architecture est donnée sur la figure III.15, fonctionne différemment du modèle de Jordan. En effet, alors que ce dernier utilise une « clef » comme générateur de séquence, le modèle de Elman intègre l'information au cours du temps pour fournir une séquence d'informations en sortie. Comme pour le modèle de Jordan, l'architecture du modèle de Elman est basée sur un perceptron à une couche cachée. La couche d'entrée se décompose en deux ensembles de neurones: un ensemble codant le vecteur d'entrée et un ensemble codant l'état passé du réseau définissant ainsi le contexte. Grâce aux connexions récurrentes entre la couche cachée et les unités de contexte, il y a recopie ou mémorisation de l'activité de la couche cachée sur une partie de la couche d'entrée. Ces connexions récurrentes sont des lignes à retard

dont les poids sont immuables au cours de l'apprentissage et sont valués à 1. Une généralisation de l'algorithme de rétropropagation est utilisée pour l'adaptation des poids.

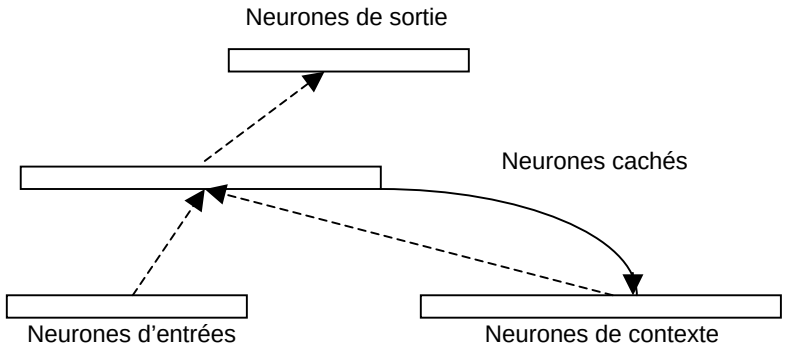


Figure III.15. Modèle de Elman (d'après [Elman, 90])

Il y'a aussi le modèle de Robinson [Robinson, 91] qui comporte une recopie pondérée de la couche cachée sur elle-même. L'activité d'une cellule de la couche cachée peut ainsi être renforcée ou affaiblie lors de la recopie, ce qui autorise une mémorisation sur de longs intervalles de temps. On peut noter que les réseaux récurrents ci-dessus peuvent être approchés par des réseaux à registre de décalage de type TDNN.

6. 3. 3. Les réseaux récurrents et la parole

Les réseaux récurrents à couches sont peu utilisés dans le traitement bas niveau de la parole. Toutefois, ils possèdent d'intéressantes propriétés à de plus hauts niveaux notamment dans le traitement éventuel du langage. Une application originale de ce genre d'architecture est donnée dans [Servan-Schreiber, 91]. Il s'agit de reconnaître des séquences de lettres de l'alphabet ordonnées selon une grammaire que le réseau devra apprendre. Si les résultats obtenus avec des séquences de lettres montrent bien la capacité des réseaux récurrents à construire des séquences suivant des règles précises, on souligne une limitation des réseaux récurrents en RAP qui est en relation avec leur pouvoir de traiter de grands vocabulaires. Une des solutions consisterait à définir une architecture hiérarchique dans laquelle des sous-réseaux récurrents sont utilisés pour identifier les différentes syllabes constituant un mot.

Pour des séquences d'entrée suffisamment courtes, le problème de l'apprentissage de tels réseaux peut être résolu par l'utilisation d'une version adaptée de la

rétropropagation de l'erreur en remplaçant le réseau récurrent par un perceptron ordinaire obtenu par « dépliage » du réseau suivant l'axe du temps, comme dans [Watrous, 88]. Des réseaux récurrents peuvent également être construits à partir d'autres modèles que le perceptron multicouches. Ainsi une machine de Boltzmann avec unités de « retenue » (*carry unit*) a été proposée dans [Prager, 86], l'apprentissage fait alors appel à des techniques de recuit simulé.

7. Conclusion: Vers des systèmes hybrides

A l'issue de ce chapitre, nous retenons la simplicité avec laquelle nous pouvons appréhender le concept des réseaux de neurones artificiels, en particulier dans un problème de perception où à un ensemble de caractéristiques en entrée (organes sensoriels), on associe une forme en sortie. Nous retiendrons également leurs performances en tant que classifieurs et leur grande adaptabilité. Toutefois, on y relève quelques insuffisances inhérentes à leur mode d'apprentissage, à savoir :

a) Le temps d'apprentissage est important ;
b) Les valeurs initiales des paramètres du réseau sont déterminantes quant à l'issue de l'apprentissage
c) 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.
d) Après la phase d'apprentissage, le réseau connexionniste est utilisé comme une boite noire. Donc le raisonnement à partir duquel ont été obtenus les résultats est inconnu.

Une des possibilités pour dépasser ces limitations est de prendre en considération 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 ; ceci rentre dans la thématique des systèmes hybrides intelligents neurosymboliques. Nous nous intéressons en particulier à l'hybridation des réseaux connexionnistes et des systèmes à base de connaissances.

Chapitre IV:

Les modèles neurosymboliques

Chapitre IV: Les modèles neurosymboliques

Présentation du chapitre

Nous avons abordé dans le chapitre précédent le commencement des modèles connexionnistes, et avions présenté quelques uns de ces modèles mais nous avons surtout clôturé ce chapitre en mettant l'accent sur la nécessité d'introduire une composante symbolique dans les systèmes connexionnistes. Cet apport permet d'améliorer grandement l'aspect représentation des connaissances dans ces systèmes et les rapprocheraient certainement un peu plus de leur inspiration biologique, ceci nous conduit à parler des systèmes neurosymboliques.

Dans ce chapitre nous allons introduire les systèmes neurosymboliques, d'abord en présentant les deux paradigmes de l'intelligence artificielle que sont : le symbolique et le connexionisme. Ces deux paradigmes sont si riches et si différents qu'il est inéluctable d'envisager leur combinaison. Effectivement leur intégration a fait l'objet de plusieurs travaux, dans ce chapitre, nous définissons ces systèmes, proposons une taxonomie de ces intégrations et présentons quelques approches combinant les systèmes symboliques et les modèles connexionnistes.

1. Introduction

Il est fort probable que l'un des atouts majeurs dans les recherches actuelles en intelligence artificielle est la co-existence d'un grand nombre de paradigmes différents et souvent conflictuels qui se retrouvent en concurrence pour susciter et garder l'intérêt de la communauté scientifique.

En effet, l'intelligence artificielle (IA) continue de connaître des développements importants dans le domaine de la modélisation des processus cognitifs et l'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, 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.

La première concrétisation de ces systèmes neurosymboliques est le CES (Connectionnist Expert System) proposé par S. J. Gallant [Gallant, 88]. Le CES est le premier système à combiner les connaissances expertes du domaine avec l'apprentissage neuronal. Une autre approche phare dans cet axe de recherche est l'approche KBANN (Knowledge-Based Artificial Neural Network) proposée par G. Towell [Towell, 91].

2. Les deux paradigmes

Depuis son avènement l'intelligence artificielle « balançait » entre deux paradigmes majeurs que sont : l'IA symbolique et l'IA connexionniste. Il est en outre largement reconnu que ces deux paradigmes ont des points forts différents et souvent complémentaires [Sun, 97].

2. 1. L'IA symbolique

Dès son avènement, le traditionnel paradigme symbolique [Newell, 76] est un des domaines d'investigation de l'intelligence artificielle, il se définit comme le domaine de développement des modèles qui manipulent des symboles. Le traitement dans de tels

modèles est basé sur une représentation explicite qui comporte des symboles organisés de manière spécifique, et les informations de dépendance sont explicitement représentées à l'aide des symboles et des combinaisons syntaxiques à partir de ces symboles.

2. 1. 1. Représentation et recherche

Deux concepts clefs sont inhérents à l'approche symbolique : l'espace de recherche de la solution et le mode de représentation des données.

En ce qui est du problème de la recherche de la solution, il est convenu que pour résoudre un problème donné, il existe un espace de recherche, cet espace est composé d'états (nœuds), chacun d'entre eux décrit une étape dans la résolution du problème. Des opérateurs sont appliqués pour atteindre un nouvel état à partir de l'état actuel. Les stratégies de recherche adoptées aux débuts de l'IA incluent les stratégies en profondeur d'abord et en largeur d'abord [Newell, 76]. Cette notion d'espace de recherche a été introduite dans tous les champs de l'IA y compris la résolution de problèmes, le traitement de langage naturel, la robotique, …

L'autre concept est la représentation des données, ce qui suppose que la connaissance doit être exprimée sous une forme interne qui faciliterait son utilisation, selon les besoins de la tâche à effectuer.

De multiples modes de représentation ont été investigués durant ces années, la plupart furent utilisés en conjonction avec des algorithmes de recherche. L'un des modes de représentation récents invoque le raisonnement à base de règles, dans ce mode de raisonnement des règles discrètes sont utilisées pour diriger la recherche (inférence). Les règles sont composées d'une partie condition qui spécifie les conditions d'utilisation de la règle et d'une partie conclusion qui définie l'action à entreprendre. Les règles sont modulaires et donc l'ajout ou la suppression de règles n'affecte pas le reste du système.
Une des formes usuelles du raisonnement à base de règles sont les systèmes de production dont les systèmes experts qui émergèrent de quelques théories psychologiques dans les années 60 et 70.

2. 1. 2. Les systèmes experts

Un système expert est un logiciel qui reproduit le comportement d'un expert humain accomplissant une tâche intellectuelle dans un domaine précis. On peut souligner les points suivants :

- les systèmes experts sont généralement conçus pour résoudre des problèmes de *classification* ou de *décision* (diagnostic médical, prescription thérapeutique, régulation d'échanges boursiers, ...)
- les systèmes experts sont des outils de l'*intelligence artificielle*, c'est-à-dire qu'on ne les utilise que lorsque aucune méthode algorithmique exacte n'est disponible ou praticable
- un système expert n'est concevable que pour les domaines dans lesquels il existe des *experts humains*. Un expert est quelqu'un qui connaît un domaine et qui est plus ou moins capable de transmettre ce qu'il sait : ce n'est par exemple pas le cas d'un enfant par rapport à sa langue maternelle.

Un système expert est composé de deux parties indépendantes :

- une *base de connaissances* elle-même composée d'une *base de règles* qui modélise la connaissance du domaine considéré et d'une *base de faits* qui contient les informations concernant le cas que l'on est en train de traiter
- un *moteur d'inférences* capable de raisonner à partir des informations contenues dans la base de connaissance, de faire des déductions, etc.

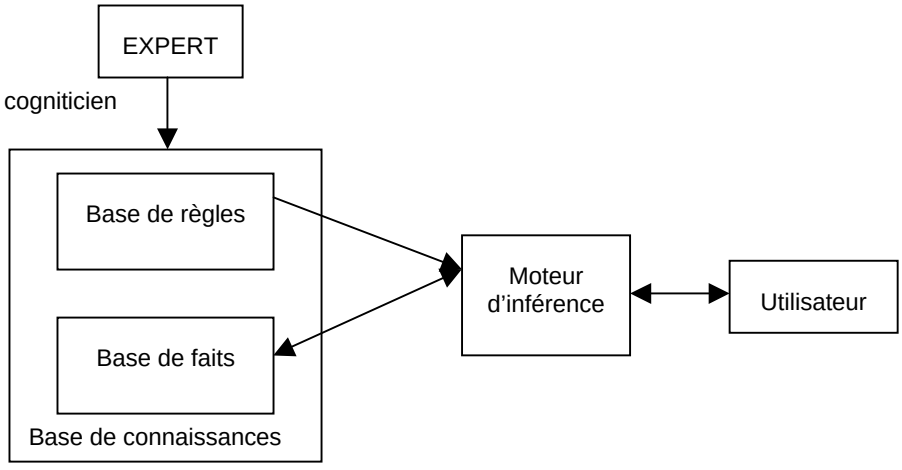


Figure IV. 1. Structure d'un système expert

2. 1. 3. L'apprentissage symbolique

L'apprentissage est une des orientations majeures en intelligence artificielle. Mais, en intelligence artificielle, l'apprentissage est difficile en dépit de modes de représentation sophistiqués, ceci est probablement dû au fait qu'à ses débuts l'intelligence artificielle symbolique s'est plutôt centrée autour de la représentation que de l'apprentissage.

Toutefois, quelques développements furent accomplis dans ce sens en particulier depuis la fin des années 80 (voir [Shavlik, 90]). Mais la majorité des algorithmes proposés implantent l'apprentissage au travers de règles simples ou d'arbres de décision [Quinlan, 86]. Ils nécessitent soit le regroupement de tous les exemples en classes distinctes ou même qui se recouvrent (i.e. clustering), soit l'induction de règles de classification qui décrivent la mise en place d'un concept [Michalski, 83]. Plus récemment des algorithmes ont été modifiés pour prendre en considération des données brouillées ou inconsistantes.

2. 1. 4. Avantages et inconvénients

Le principal avantage de l'approche symbolique est son haut niveau d'abstraction, qui permet de décrire la connaissance de manière compréhensive. En particulier les systèmes de production qui comptent parmi eux les systèmes experts, donnent des explications aux réponses que le système produit sous forme de traces des inférences effectuées.

Mais bien que le paradigme symbolique domine l'intelligence artificielle et les sciences cognitives depuis bien longtemps, il fût la cible de nombreuses critiques. Ces critiques concernent largement leur incapacité à gérer les connaissances incomplètes, incorrectes ou incertaines. De plus, un système symbolique est souvent incapable d'apprendre par lui-même.

2. 2. L'IA connexionniste

L'idée de base dans le connexionisme est d'utiliser une interconnexion dense d'éléments de calculs simples. Cette nouvelle approche rencontra à la fois un enthousiasme de la part des déçus de l'IA symbolique mais aussi une grande aversion de la part des ardents défenseurs de l'IA symbolique. Aujourd'hui encore, on peut sentir la division entre ces deux paradigmes de l'IA, toutefois, on voit aussi fleurir des hybridation entre eux.

Les modèles connexionnistes passant pour être une ultime étape vers la capture des propriétés biologiques de l'intelligence.

2. 2. 1. L'apprentissage connexionniste

Les modèles connexionnistes excèllent en apprentissage. Aussi, contrairement à l'IA symbolique qui est axée sur la représentation, les fondements des modèles connexionnistes ont toujours été leur capacité à apprendre. Plusieurs algorithmes sont alors développés le plus connu étant la rétropropagation (voir chapitre III).

2. 2. 2. La représentation des connaissances

Bien qu'il soit relativement difficile d'élaborer des représentations sophistiquées dans les modèles connexionnistes comparativement aux modèles symboliques, il y a eu des développements significatifs dans la représentation des connaissances.

2. 2. 3. Avantages et inconvénients

Les réseaux de neurones sont de bons classifieurs, en particulier, les réseaux de neurones sont des outils performants grâce à leur capacité d'apprentissage, de généralisation et de classification. Toutefois il existe encore des limitations liées à leur style d'apprentissage, c'est la raison pour laquelle plusieurs travaux se sont orientés vers les systèmes hybrides afin de remédier aux inconvénients de ce paradigme améliorant ainsi le taux de reconnaissance et la robustesse du système.

2. 3. L 'intégration des réseaux de neurones et des systèmes experts

L'évolution importante qui caractérise les systèmes experts est sans contestation leur intégration dans les modèles connexionnistes, la puissance fondamentale de cette hybridation est la complémentarité entre les deux approches. Cette complémentarité est illustrée dans le tableau suivant extrait de [Boz, 97] où les caractéristiques sont notées par une appréciation variant de 1 à 5 :

	Systèmes experts	Réseaux connexionnistes
Acquisition des connaissances	5	1
Adaptation aux nouveaux problèmes	4	1
Raisonnement de haut niveau	1	5
Raisonnement de bas niveau	3	1
Explicabilité	1	5

Tableau IV. 1 Tableau comparatif entre les réseaux de neurones et les systèmes experts

3. Les systèmes neurosymboliques

3. 1. Les systèmes hybrides intelligents

Les différentes approches d'intégration essaient d'incorporer plusieurs paradigmes dans le même système. Collectivement ces systèmes sont appelés des systèmes hybrides intelligents. Parmi ces approches l'intégration système expert / réseau de neurones constitue la principale voie de complémentarité entre les approches neuronales et symboliques.

Le concept de systèmes hybrides est très large, de tels systèmes incluent toute méthode qui intègre au moins deux approches différentes pour la solution d'un problème. Les possibilités d'hybridation entre approches étant très riches, nous nous limitons dans ce qui suit à donner un bref aperçu de ce qui se fait actuellement en terme de systèmes hybrides dont l'une des composantes est un réseau connexionniste.

Systèmes neuro-génétiques : la plupart des systèmes neuro-génétiques sont conçus afin de contourner le problème relatif au choix de l'architecture du réseau. De tels réseaux sont aussi appelés: réseaux évolutifs (evolutionary artificial neural networks). L'idée de base étant l'utilisation des algorithmes génétiques pour faire évoluer et sélectionner les architectures de réseaux connexionnistes les plus performantes par rapport à une application. On trouve aussi dans la littérature d'autres types moins répandus d'hybridations neuro-génétiques tels que: l'adaptation des poids synaptiques réalisée par un algorithme génétique, ou l'adaptation des paramètres de vitesse de convergence et d'inertie par un algorithme génétique.

Systèmes neurosymboliques : ce type de système hybride est sans doute le plus répandu parmi les approches hybrides [Orsier, 95]. On y tente de combiner une méthode symbolique et une méthode connexionniste.

3. 2. Les systèmes neurosymboliques

L'intelligence artificielle et le connexionisme sont complémentaires: les modèles connexionnistes sont mieux adaptés à la perception au sens large et l'IA à la représentation symbolique et au raisonnement conscient. On voit ainsi apparaître des systèmes hybrides de traitement de l'information fondés sur un raisonnement symbolique, dans lesquels des sous systèmes connexionnistes assurent des fonctions spécifiques et coopèrent avec d'autres entités. Le couplage entre les deux approches peut prendre diverses réalisations.

Systèmes neuro-flous: ces systèmes sont une des catégories des systèmes hybrides les plus développés car la logique floue et les réseaux connexionnistes ont beaucoup de points en communs. Les systèmes neuro-flous sont principalement de trois types: des systèmes qui intègrent des règles floues dans des réseaux connexionnistes, des systèmes qui font l'extraction de règles floues à partir des réseaux connexionnistes et des systèmes qui implantent les neurones floues.

Systèmes neuro-IDT: ce sont des systèmes qui combinent des réseaux connexionnistes avec des arbres de décision ; ils s'inscrivent dans deux approches principales: la construction et l'initialisation d'un réseau à partir d'un arbre de décision, ou l'extraction d'un arbre de décision à partir d'un réseau après la fin de son apprentissage. L'insertion d'un arbre de décision dans un réseau sert à simplifier le choix de son architecture et de ses poids initiaux. Cette approche apporte une solution alternative au problème de construction du réseau. L'autre approche adresse le problème de représentation symbolique dans les réseaux ; l'extraction d'un arbre de décision va permettre d'obtenir des explications sur le raisonnement employé.

Systèmes neuro-CBR: l'intégration d'un réseau connexionniste et d'un système de raisonnement basé sur des cas permet d'induire des connaissances de plus haut niveau, tout en gardant les avantages des systèmes CBR (Case Based Reasonning). Un tel système permettrait d'obtenir un réseau des prototypes que représentent les cas appris.

Systèmes neuro-KBS: les systèmes qui intègrent des réseaux connexionnistes et des systèmes à base de connaissances sont couramment appelés systèmes hybrides neurosymboliques.

Nous avons fait le choix d'orienter les travaux réalisés dans cette thèse sur ce dernier type de systèmes: nous les appellerons dans la suite les systèmes hybrides neurosymboliques.

4. Taxonomie des systèmes neurosymboliques

En fait, un grand nombre de travaux se situent dans le courant de l'intégration des meilleures caractéristiques de chacun des mondes symbolique et connexionniste, et plusieurs classifications des modèles d'hybridation neurosymboliques furent proposées [Medsker, 92][Hilario, 95]][Sun, 97],[Orsier, 95][Sima, 00][Web 05]. Parmi cette diversité d'approches, nous avons choisi d'adopter une classification très proche de celle proposée dans [Hilario, 95] car nous pensons que c'est la classification la plus complète, et qui nous permet surtout, de situer notre proposition au milieu de cette profusion d'approches. Cette classification assume que ces modèles hybrides sont globalement classés en trois catégories: les modèles combinés (unified approaches), les modèles transformationnels et les modèles couplés [Hilario, 95][Wermter, 00][Web 05].

4. 1. Les modèles combinés

Le but dans l'approche combinée est d'intégrer les propriétés connexionniste et symbolique en utilisant uniquement les techniques connexionnistes. Deux directions peuvent être distinguées dans ce camp neurosymbolique: l'approche basée neurone et l'approche basée symbole.

L'approche basée neurone *(neuron-to-symbol approach)**:*** Cette approche tente d'ancrer tous les traitements cognitifs dans une réalité biologique. C'est en effet, une approche biologique où l'objectif est de modéliser les fonctions de haut niveau du cerveau. Dans cette approche du type bottom-up, on débute avec des neurones et on tente de modéliser des fonctions de plus haut niveau. Un très bel exemple issu de cette approche est celui de la « theory of neural group selection » (TGNS) [Edelman, 92]. D'autre part, le

modèle de la colonne corticale en est également une illustration, il a été proposé par Guigon en 93, qui tente de modéliser des unités fonctionnelles du cortex d'un point de vue biologique [Durand, 95].

L'approche basée symbole *(symbol-to-neuron approach)**:*** Dans cette approche l'accent est mis sur la construction d'architectures connexionnistes pour le traitement symbolique. Concrètement, la conception du modèle débute avec des fonctions de haut-niveau et on arrive à la conception de l'infrastructure connexionniste appropriée. Dans cette approche top-down, le modèle de neurone utilisé est souvent le neurone formel classique. Les architectures connexionnistes basées symboles dépendent du schéma de représentation de connaissances adopté et de ce fait peuvent être classées en: des modèles localistes, où chaque nœud du réseau correspond à un concept et vice versa, et en des modèles distribués où un concept est codé sur un ensemble de nœuds, et un nœud participe au codage de plusieurs concepts. Il existe aussi des modèles combinés localistes/distribués.

4. 2. Les modèles transformationnels

Dans les modèles transformationnels (translationnels selon [Hilario, 95]), les systèmes experts peuvent être transformés en des réseaux de neurones et inversement.

La transformation du système expert en un réseau de neurones est la plus utilisée. Dans ce cas, la connaissance issue du système expert sert à produire les conditions initiales et l'ensemble d'apprentissage du réseau de neurones, cette transformation rend le développement du réseau connexionniste plus rapide.

Dans le cas de la transformation inverse, des méthodes existent qui permettent d'extraire, insérer ou raffiner des règles logiques à partir du réseau de neurones entraîné.

4. 3. Les modèles couplés

Dans les modèles couplés, l'application est constituée de deux composants séparés, qui peuvent s'échanger les connaissances. Le réseau connexionniste peut être utilisé comme composant de pré-traitement pour les systèmes experts, dans ce cas le réseau de neurones peut supprimer les données erronées ou effectuer une pré-classification sur les données à utiliser par le système expert. Les systèmes experts peuvent préparer les données pour les réseaux de neurones et contribuer à la détermination de la configuration du réseau.

Un autre exemple de modèles couplés est celui du tableau noir (blackboard model). Les agents du tableau noir peuvent être des réseaux de neurones et des systèmes experts qui s'échangent des informations par le biais du tableau noir.

Les systèmes encapsulés sont une autre variante des modèles couplés, ainsi un réseau de neurones peut être encapsulé dans un système expert pour contrôler l'inférence.

Pour les modèles couplés une classification peut être établie sur la base du degré de couplage entre les composants (fort ou faible) ou sur la base du schéma d'intégration utilisé (coopération, pré/post traitement, sous traitance ou méta traitement).

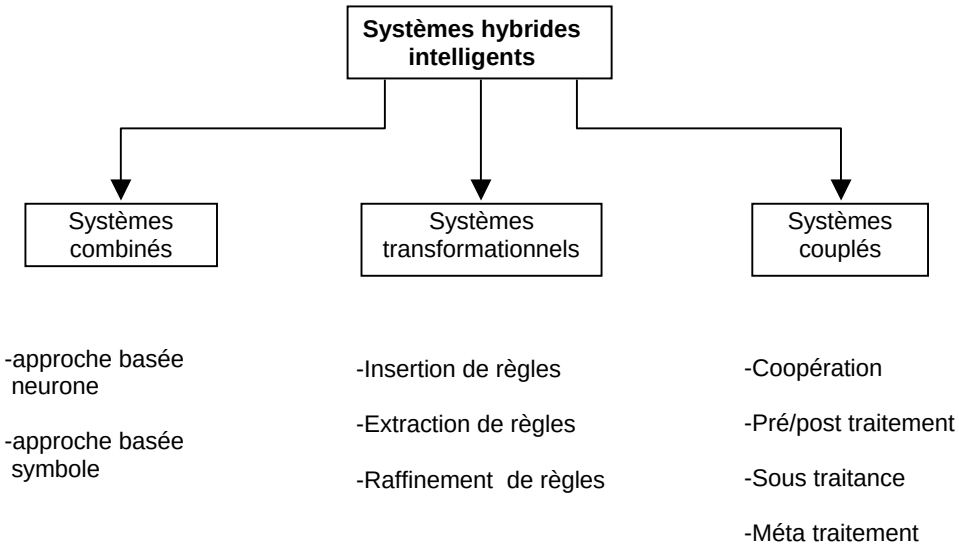


Figure IV. 2. Classification des systèmes neurosymboliques

5. Les systèmes experts connexionnistes

5. 1. Introduction

Le premier des systèmes experts connexionnistes est dû à S. Gallant qui a développé MACIE (pour MAtrix Controlled Inference Engine) [Gallant, 88]. Ce système est aussi le plus connu parmi les modèles neurosymboliques fortement couplés. MACIE est un réseau feedforward, dans lequel les neurones de sorties ont des valeurs (+1 ou –1) obtenues en appliquant un seuil à la somme des valeurs d'entrées pondérées.

5. 2. Principe

MACIE est un moteur d'inférence connexionniste qui consiste en un système expert implanté au travers d'un perceptron multicouches. Gallant fut le premier à décrire un système qui combine les connaissances expertes du domaine avec l'apprentissage neuronal. Un système expert connexionniste a communément les propriétés suivantes :

- Chaque neurone a une signification,
- Le réseau est construit en utilisant les relations de dépendances entre objets du domaine,
- Les poids positifs traduisent un renforcement, tandis que des liens négatifs reflètent une inhibition,
- Les valeurs en entrée sont discrètes, habituellement 1, -1, 0 ce qui représente respectivement vrai, faux et inconnu.

D'autre part, le réseau est du type feedforward, et comme il n'y a pas de cycles dans le réseau, les neurones peuvent être numérotés. Ainsi si un arc part du neurone u_i vers le neurone u_j, on a j > i. Alors pour représenter les poids du réseau Gallant utilise une matrice W, où w_{ij}=0 s'il n'y a pas d'arc qui relie u_i et u_j , et on utilise w_{i0} pour représenter le biais du neurone u_i. Les activations des neurones sont discrètes soit : +1, -1, 0 ; ces valeurs correspondent aux valeurs logiques vrai, faux et inconnu. Un neurone u_i calcule sa nouvelle activation selon la fonction linéaire discriminante : $S_i=\sum_{j>0} w_{ij}u_j$

Une itération dynamique du réseau consiste en la re-évaluation de chaque neurone selon son ordre et à changer son activation avant de considérer le suivant.

5. 3. Exemple d'un système expert connexionniste

Dans un système expert connexionniste, les neurones sont les symboles utilisés par le système expert. Voici à titre d'exemple, la structure d'un système expert connexionniste dédié au diagnostic médical. Le modèle proposé dans [Gallant, 88] comprend six symptômes, deux maladies dont le diagnostic se base sur les symptômes et trois traitements possibles. Les étapes suivies pour générer ce réseau sont :

a) Chaque neurone du réseau correspond à une variable du domaine d'application (symptôme, maladie, …).
b) Pour chaque variable d'entrée une question est posée à l'utilisateur (par exemple, « est-ce-que le patient a des pieds enflés ? »).
c) Déterminer les informations de dépendances pour les variables intermédiaires (maladies) et les variables de sorties (traitement). Pour chacune de ces variables, on établit une liste de variables dont les valeurs suffisent à calculer la valeur. Les informations de dépendances spécifient les dépendances du réseau, ainsi chaque fois qu'une variable u_j est dans la liste de dépendance de la variable u_i, un lien partant de u_j vers u_i est établi.
d) Finalement, un ensemble d'exemples d'apprentissage est fourni au réseau. Pour ce problème un exemple est un historique du patient.

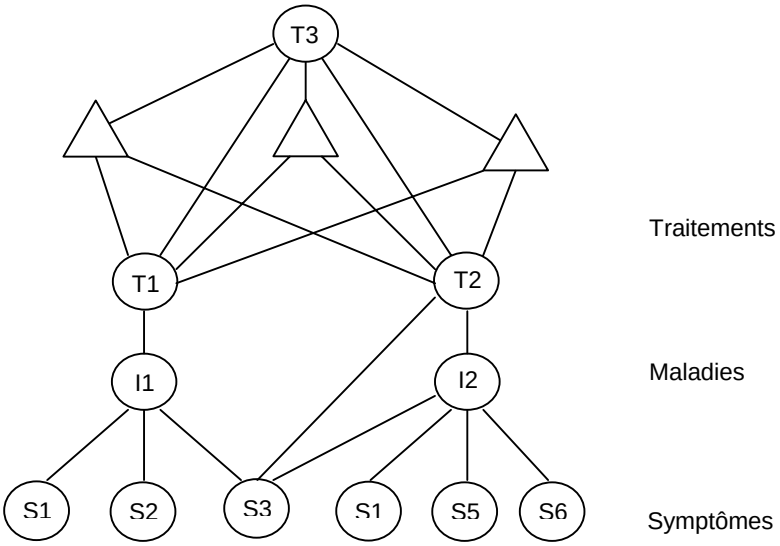


Figure. IV.3. Réseau expert pour le diagnostic médical (d'après[Gallant, 88])

Les entrées sont les symptômes représentés par {1,0,-1} selon qu'un symptôme est présent ou absent (tels que : température élevée, pression artérielle élevée, …) ou bien pas encore connu (tel que pour les analyses sanguines, …). Le système utilise des données d'apprentissage qui consistent en des symptômes avec les diagnostics connus. Des neurones peuvent être ajoutés en phase d'apprentissage pour la cohérence du réseau (c'est le cas des neurones triangulaires).

5. 4. L'algorithme Pocket

La grande majorité des algorithmes d'apprentissage pour un seul neurone sont incapables de produire pour un problème de classification un vecteur de poids qui puisse satisfaire un nombre maximum de relations entrée-sortie des données d'apprentissage. Une exception à cela est l'algorithme Pocket: il réitère l'exécution de l'algorithme du perceptron et garde (dans une poche) le vecteur poids qui demeure inchangé durant le plus grand nombre d'itérations.

C'est donc un algorithme qui considère un neurone à la fois, soit le neurone *u* et soient:

$\{E^k\}$, l'ensemble des activations des exemples d'apprentissage et

$\{C^k\}$ les activations correctes correspondantes.

$\{C^k\}$ prend les valeurs {+1, -1} et $\{E^k\}$ prend les valeurs {+1,-1,0}.

$E_0^k = +1$, et w_0 est le seuil calculé pour le neurone.

Etant donné P les vecteurs de poids, et w* les vecteurs de poids pocket, l'algorithme procède comme suit :

1. Initialiser P au vecteur 0
2. Soit P le vecteur courant de poids, choisir aléatoirement un exemple de E^k,
3. a. Si P classifie correctement E^k

   i.e. $\{P.E^k > 0$ et $C^k = +1\}$

   ou bien $\{P.E^k < 0$ et $C^k = -1\}$

   Alors

   aa. Si l'exécution de la classification avec P est meilleure qu'avec w*

   aaa. Alors remplacer les poids pocket w* par P

   b. Sinon Former un nouvel ensemble de poids $P' = P + C^k E^k$
4. Aller à 2)

5. 5. EXPSYS : un autre exemple de système expert connexionniste

EXPSYS [Sima, 95] est un exemple de système neuro-expert qui a été largement influencé par MACIE [Sima, 00]. Il en est toutefois différent en deux points essentiels :

a) les fonctions d'activation des neurones sont différentiables (fonction tangente hyperbolique),
b) un algorithme d'apprentissage tel que l'algorithme de rétro-propagation peut être utilisé pour entraîner le réseau.

Les fonctions d'activation continues améliorant la généralisation par rapport à MACIE mais le rôle des neurones cachés devient obscur [Boz, 97], et de ce fait aussi bien l'inférence que l'explication deviennent compliquées. EXPSYS prend en charge des informations incomplètes par l'introduction de la notion d'intervalle des états des neurones [Sima, 92] et les neurones peuvent avoir des valeurs d'entrées dans l'intervalle [-1,+1].

5. 6. Conclusion sur l'approche

Un CES est un réseau de neurones dont le fonctionnement peut être assimilé à un moteur d'inférence, car la progression d'une couche à une autre modélise les règles de production d'un moteur d'inférence. De plus, il implémente un algorithme d'apprentissage qui convient à son mode de représentation.

6. L'approche KBANN

6. 1. Introduction

Quelque temps après Gallant, G. G. Towell [Towell, 91][Towell, 94], propose un algorithme qu'il dénomme KBANN pour Knowledge-Based Artificial Neural Networks, où l'idée est d'insérer un ensemble de règles symboliques à l'intérieur d'un réseau de neurones. Le réseau est ensuite raffiné en utilisant des algorithmes d'apprentissage standards et un ensemble d'exemples d'apprentissage. L'application que propose l'auteur pour évaluer son algorithme est dans le domaine d'analyse de séquences d'ADN [Towell, 91].

Le KBANN consiste en deux algorithmes différents, d'abord, un traducteur des règles en un réseau, et ensuite un algorithme d'apprentissage qui sert à raffiner le réseau.

6. 2. Construction du réseau

Dans l'approche KBANN, la démarche de construction du réseau débute par un ensemble de règles approximativement correctes. Ces règles seront « traduites » en un réseau de neurones basé connaissances (KBANN-net). Les règles à traduire sont sous forme de clauses de Horn, et elles doivent en outre être acycliques. L'approche KBANN suggère sept étapes pour la construction du réseau :

1. Re-écriture: les règles définies seront transformées en des clauses de Horn. Les règles ayant la même partie conclusion seront réécrites de la manière suivante : A:-B,C,D. A:-E,F. en A:-A'. A:-A''. A':-B,C,D. A'':-E,F.

2. Passage au réseau de neurones: les règles sont organisées en un réseau de neurones. Les poids sont choisis de manière à ce que l'activation émule la fonction AND.

a) Traduction des conjonctions. Si une règle est de la forme : A:-B,C,D,not(E). elle sera transformée selon la figure IV.3.a. Toutes les connexions correspondant aux antécédents positifs sont mises à ω, et toutes les connexions correspondent aux antécédents négatifs sont mises à -ω. Le biais θ correspondant à la conclusion de la règle est mis à (P-1/2) ω. La valeur suggérée pour ω étant 4 [Towell, 94].

b) Traduction des disjonctions: les règles suivantes A:-B. A:-C. A:-D. A:-E. seront transcrites selon la figure IV.3.b. Le biais dans ce cas est mis à -ω/2, ainsi, tout neurone en entrée est capable d'activer la sortie.

a) Conjonction b) Disjonction

Figure IV. 4. Re-écriture de règles dans un KBANN

3. ***Numérotation:*** à chaque neurone du réseau, on va assigner un numéro selon son « niveau ».

4. ***Ajout de neurones cachés:*** de nouveaux neurones cachés peuvent être optionnellement ajoutés.

5. ***Ajout de neurones d'entrée:*** cet ajout s'effectue dans le cas, où on peut identifier d'autres entrées que celles déjà spécifiées par les règles.

6. ***Ajout de connexions:*** les connexions dont les poids sont nuls sont rajoutées au réseau. Un neurone du niveau n-1 est connecté à tous les neurones du niveau n. Nous obtenons alors un réseau complètement connecté.

7. ***Perturbation:*** Une petite valeur aléatoire est ajoutée à tous les poids du réseau.

6. 3. Conclusion sur l'approche

Finalement, les correspondances suivantes sont proposées par Towell [Towell, 94] pour établir une parité entre le domaine d'application et le réseau de neurones.

Domaine d'application	Réseau de neurones
Conclusions	Neurones de sortie
Conclusions intermédiaires	Neurones cachés
Faits	Neurones d'entrée
Antécédent	Connexion avec poids élevé

Tableau IV. 2. Passage des règles au réseau

Ces correspondances et compte tenu de la re-écriture des règles peuvent conduire à un accroissement important de la base de règles (risque d'explosion combinatoire). D'autre part, dans un réseau KBANN toutes les fonctions sont différentiables, donc on peut utiliser l'algorithme de rétropropagation pour l'apprentissage du réseau.

7. Conclusion

Finalement, nous ne manquerons pas de remarquer qu'il est très intéressant de se baser sur des connaissances du domaine pour guider la création du réseau donnant ainsi une signification à chaque neurone et une interprétation à chaque connexion, ceci limite le

nombre de connexions de façon significative ce qui optimise considérablement l'espace mémoire et le temps de calcul, en particulier en phase d'apprentissage.

Ayant introduits les différentes notions, algorithmes et approches nécessaires à la compréhension de notre démarche, nous nous intéressons dans la suite à la conception d'un système connexionniste expert, dédié à la reconnaissance de la parole.

En plus des points sus-cités, un autre aspect immerge dès que l'on considère l'approche basée connaissance, il est lié à la phase de reconnaissance. En effet, en phase de reconnaissance, un mot peut être reconnu et bien classé, ou bien reconnu mais mal classé, ce qui pose alors la question de fiabilité du système, et l'aspect explicabilité du raisonnement revêt alors toute son importance. Cet aspect fait défaut aux modèles connexionnistes (mais aussi aux HMMs) mais, il est parfaitement présent dans les systèmes experts.

PARTIE 3:

Modèle proposé

Chapitre V:

NESSR: un modèle neuro-expert pour la reconnaissance de la parole

Chapitre V: ***NESSR***, Un modèle neuro-expert pour la reconnaissance de la parole

Présentation du chapitre

Après avoir présenté les concepts inhérents aux réseaux connexionnistes (chap. 3) ainsi qu'un aperçu de leur utilisation en RAP, il en est ressortit une certaine tendance à doter ces réseaux d'une symbolique, nous avons alors présenté une brève introduction à de tels systèmes hybrides (chap. IV). Dans ce chapitre, nous présentons les éléments conceptuels du modèle perceptuel que nous proposons pour la reconnaissance automatique de la parole.

Le chapitre se compose de deux parties. Dans la première partie du chapitre nous allons présenter les éléments conceptuels d'un système connexionniste expert dédié à la reconnaissance de la parole. La philosophie adoptée pour mener cette conception est inspirée de celle proposée par S. Gallant [Gallant, 88] en ce qui est des grandes lignes pour l'introduction des symboles dans l'architecture connexionniste. D'autre part, le modèle conceptuel pour la compréhension de la parole proposé par Rabiner dans [Rabiner, 83] fût également une source d'inspiration en ce qui est de la topologie du réseau.

Dans la seconde partie de ce chapitre, nous présentons le modèle d'un neurone symbolique temporel (***STN***), ce modèle représente notre apport par rapport aux modèles hybrides existants dans la littérature en particulier la conception d'un réseau connexionniste spécialisé qui implémente la notion du temps. Le modèle ***STN***, est ensuite appliqué au système décrit en première partie, le système obtenu est appelé ***NESSR*** pour Neural Expert System for Speech Recognition.

Partie A: Ancrage des symboles dans une architecture connexionniste pour la reconnaissance de la parole

1. Introduction

Il est communément convenu que, la difficulté dans l'utilisation des réseaux connexionnistes réside en la manière de configurer convenablement le réseau, et d'initialiser correctement les connexions entre neurones du réseau. Pour une application en reconnaissance de la parole, le nombre de neurones dans la couche d'entrée dépend du nombre de paramètres à utiliser pour décrire les mots à reconnaître, le nombre de neurones dans la couche de sortie représente le nombre de classes d'attribution et le nombre de couches cachées ainsi que le nombre de neurones dans chacune d'entre-elles sont déterminés en utilisant des heuristiques, par exemple : la dimension de Vapnik-Chervonenkis, ou après expérimentation [Bishop, 97].

Il devient alors très intéressant de se baser sur les connaissances du domaine pour guider la création du réseau donnant ainsi une signification à chaque neurone et une interprétation à chaque connexion.

Mais, bien que les réseaux de neurones artificiels aient déjà donné de bons résultats en RAP [Tebelski, 95], et que l'orientation vers des systèmes hybrides est tout à fait d'actualité en connexionisme, les recherches ne se sont jamais intéressées aux systèmes hybrides neurosymboliques appliqués à la parole, des systèmes hybrides neuro-flous ou neuro-génétiques ayant déjà fait l'objet d'intérêt pour ce sujet.

Cette orientation nous a alors séduit et nous souhaitons au travers de ce travail apporter une contribution dans cette thématique très riche, qui nous semble une réelle perspective pour le connexionisme. Nous adoptons donc une démarche perceptuelle pour définir la topologie du réseau, en effet si l'on se réfère au schéma présenté par Rabiner dans [Rabiner 83, p : 56] (figure V.1), on peut tout à fait voir l'influence du domaine qui ressort dans l'architecture du réseau de neurones.

Dans cette première partie du chapitre, nous allons présenter nos premières tentatives pour la conception et la réalisation d'un système neurosymbolique en RAP.

Un premier essai consiste en un système expert connexionniste très proche de la philosophie de Gallant [Gallant, 88], tandis que le second système est issu de l'approche KBANN de Towell [Towell, 94]. Ces deux systèmes présentent la particularité d'encapsuler la théorie du domaine dans le réseau connexionniste qui est dans ce cas un perceptron multicouches qui représente une base de connaissances, où les neurones d'entrée représentent des traits acoustiques, les neurones cachés sont des unités phonétiques et les neurones de sortie sont les mots du vocabulaire.

2. Un modèle conceptuel pour la compréhension de la parole

Le diagramme de la figure V.1. 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 sauvegarde 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 [Rabiner, 83].

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 (et que d'autres chercheurs croient) que le connexionisme offre une alternative prometteuse pour la modélisation des tâches cognitives.

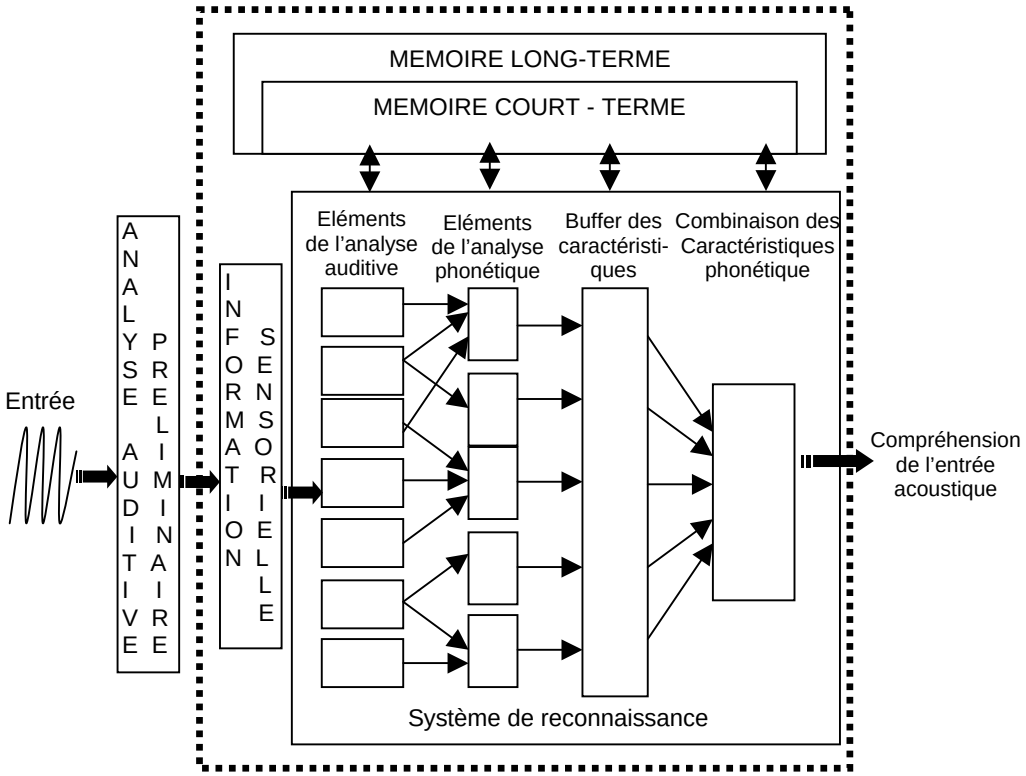


Figure V. 1 Diagramme conceptuel du système de reconnaissance de l'être humain

3. De la connaissance au réseau de neurones

3. 1. Ancrage des symboles dans le réseau connexionniste

Notre objectif est de construire un réseau connexionniste qui encapsule la sémantique du domaine d'application, en l'occurrence la reconnaissance automatique de la parole arabe. Pour générer la base de connaissances connexionniste, nous sommes amenés à définir les objets du domaine afin de les faire correspondre aux cellules du réseau. Il s'agit ensuite de déterminer la topologie adéquate pour le réseau connexionniste.

L'application considérée est la reconnaissance de mots arabes. Aussi, nous pouvons déjà avancer que parmi les objets du domaine, on retrouve les mots à reconnaître. D'autre part, tout processus de reconnaissance de formes part d'un ensemble de caractéristiques qui permet de définir la forme à reconnaître, ce qui nous amène à considérer les caractéristiques acoustiques du signal comme objets du domaine.

Ainsi, de manière très intuitive, nous pouvons proposer comme objets à modéliser : les mots du vocabulaire et un ensemble de caractéristiques acoustiques dont le nombre reste à déterminer.

Quant au choix de la topologie du réseau connexionniste, nous allons reconsidérer le schéma conceptuel du processus de compréhension de la parole proposé par Rabiner dans [Rabiner, 83]. Ce schéma suggère une architecture en couches du réseau de neurones, architecture qui est également assumée dans d'autres modèles qui sont une inspiration de modèles biologiques [Guigon, 95]. A l'instar de ces modèles, nous adoptons une architecture en couches dans notre approche. Donc les objets prédéfinis se retrouveront regroupés en deux couches: une couche d'entrée qui contient les caractéristiques et une couche de sortie qui regroupe les mots du vocabulaire.

Vu de plus près, le schéma de la figure V. 1. suggère la présence d'une couche intermédiaire entre l'entrée et la sortie qui représente le niveau phonétique, ce que nous allons également retenir. Finalement, nous obtenons un réseau connexionniste du type Percetron multicouches composé de trois couches [Bahi, 04a][Bahi, 05a] :

1) Une couche d'entrée qui représente la perception du système, elle comprend des unités qui détectent les caractéristiques acoustiques du signal ; nous l'appelons : niveau acoustique,
2) Une couche de sortie qui représente la décision du système, elle comprend les mots du vocabulaire ; nous l'appelons niveau lexical,
3) Une couche cachée, qui est le niveau intermédiaire. Cette couche représente le niveau phonétique, dans cette partie du travail nous estimons qu'une décomposition du mot en syllabes est tout à fait intéressante vu le statut bien défini de la syllabe dans la langue arabe. Et donc la couche cachée comprend des neurones qui correspondent aux syllabes qui couvrent l'ensemble des mots du vocabulaire choisi.

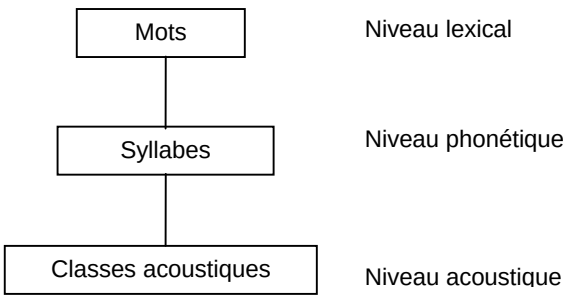


Figure V. 2. Architecture du système expert connexionniste pour la RAP

3. 2. Les neurones d'entrée

3. 2. 1. Les traits acoustiques

En s'inspirant de l'approche de Stephen J. Gallant où les neurones d'entrées du réseau expert connexionniste sont des cellules dont l'entrée est soit 0 soit1. Nous assumons que les cellules d'entrée de notre système représentent des traits acoustiques des unités phonétiques considérées.

Chaque trait par sa présence ou son absence peut servir à la reconnaissance de l'unité considérée, en l'occurrence la syllabe. Ainsi, ces cellules auront pour valeur d'entrée 0 ou 1.

Dans une première étape, nous allons déterminer ces traits acoustiques que nous avons dénommés: classes acoustiques qui servent à la détection des syllabes relatives à notre vocabulaire.

Au début du processus de reconnaissance, nous procédons à une extraction de caractéristiques qui produit une collection de vecteurs numériques, mais comme nous avons vu précédemment, dans un CES, seules les valeurs discrètes sont autorisées. Pour obtenir de telles valeurs nous avons eu recours à la quantification vectorielle. Nous retiendrons que chaque unité phonétique est définie grâce à un ensemble de caractéristiques discrètes.

3. 2. 2. La 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 leurs prototypes (voir figure V.8). 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 transformés 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.

3. 3. La syllabe: l'unité de la décision

L'Arabe a la particularité de présenter peu de voyelles, peu de consonnes et une structure régulière de syllabes[El-ani, 76]. Les syllabes pourraient être traitées facilement et ont un statut linguistique bien défini [Caliope, 89]. Ces éléments, et le fait, qu'en Arabe, on ne peut construire dans la presque totalité des cas, qu'un seul mot avec un ensemble bien défini de syllabes, ont motivé notre choix pour considérer la syllabe comme l'unité de modélisation du niveau phonétique.

Voici les cinq formes possibles d'une syllabe arabe comme présentées dans [El-Ani, 76]. Dans cette représentation, C signifie consonne, V voyelle courte et VV voyelle longue. Les quatre premiers modèles apparaissent en position initiale, médiane ou finale. La cinquième forme apparaît seulement en position finale ou isolée.

- **CV** */bi /* (avec)
- **CVC** */sin /* (dent)
- **CVV** */maa /* (pas)
- **CVVC** */baab /* (porte)
- **CVCC** */sifr /* (zéro)

Les classes, que nous dégageons à l'issu de la quantification vectorielle, caractérisent par leur présence une syllabe donnée.

3. 4. Les relations de dépendances

Une fois que les neurones du réseau (objets du domaine) ont été identifiés il faut dégager les relations de dépendances qui les relient afin de les traduire en des connexions effectives dans le réseau.

Les relations qui relient les neurones de la couche de sortie à ceux de la couche cachée sont intuitives et faciles à dégager, ce sont des règles qui permettent l'assemblage des syllabes en des unités plus larges que sont les mots. Ces règles sont de la forme:

Si ***Conjonction(Syllabes)*** **Alors** ***Mot_Reconnu***

A titre d'exemple, les syllabes [waa] et [Hid] forment le mot [waaHid] (le chiffre « un ») .

Par contre, pour extraire les relations qui relient la couche d'entrée à la couche cachée, il faut procéder à de nombreux essais. Cette phase fût encore plus longue et fastidieuse, que celle de la QV où il fallait extraire les caractéristiques les plus représentatives.

Il s'agit de dégager les classes acoustiques reliées à chacune des syllabes de notre réseau. Dans cette partie du travail l'expertise est extraite en se basant sur les observations du domaine. Les règles sont de la forme:

Si ***Conjonction(Classes)*** **Alors** ***Syllabe_Reconnue***

Voici, quelques exemples de règles que nous avons pu établir, en sélectionnant les caractéristiques qui reviennent le plus dans la structure d'une syllabe pour une application à la reconnaissance des dix chiffres arabes :

Si C_7 **et** C_8 **et** C_9 **et** C_{11} **et** C_{15} **et** C_{20} **et** C_{22} **et** C_{23} et C_{25} **et** C_{27} **et** C_{28} **et** C_{30} **alors** waa
Si C_7 **et** C_9 **et** C_{10} **et** C_{11} **et** C_{12} **et** C_{20} **et** C_{24} **et** C_{30} **alors** naan
Si C_7 **et** C_9 **et** C_{11} **et** C_{12} **et** C_{13} **et** C_{14} **et** C_{15} **alors** ?ar
Si C_1 **et** C_2 **et** C_7 **et** C_8 **et** C_{19} **et** C_{20} **et** C_{28} **et** C_{30} **alors** γam
Si C_8 **et** C_{10} **et** C_{11} **et** C_{12} **et** C_{16} **et** C_{24} **et** C_{30} **alors** tis

Où [waa], [naan], et [?ar],… sont des syllabes de l'Arabe qui représentent des neurones de la couche cachée.

Il est impératif de souligner qu'à ce stade de nos travaux l'aspect précédence dans le temps n'est pas du tout pris en considération car d'une part, nous considérons un ensemble réduit de syllabes et d'autre part, qu'un ensemble donné de syllabes (sans tenir compte de l'ordre) ne peut générer qu'un seul mot du vocabulaire.

4. Exemple d'un système expert connexionniste pour la RAP

Nous avons dans la section précédente défini les objets du domaine que nous avons traduit en des neurones, et les relations de dépendances que nous traduisons en des liens dans le réseau connexionniste. Nous présentons maintenons la topologie du réseau que nous obtenons si nous appliquons cette approche à un système de reconnaissance des

dix chiffres de l'Arabe. Le réseau de neurones est un perceptron multicouches classique, où :

- la couche d'entrée contient 32 neurones, dont les entrées sont soit 1 soit 0,
- La couche cachée contient 29 neurones, qui représentent les syllabes induites par la prononciation des chiffres,
- La couche de sortie contient 10 neurones, qui représentent les dix chiffres.

Les règles de production précédentes serviront à initialiser les connexions. Des dépendances directes entre les neurones sont concrétisées par des connexions mises à 1.

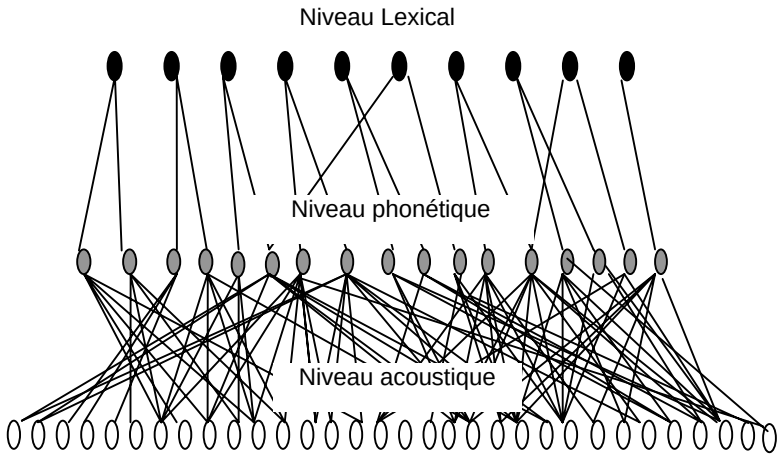


Figure V. 3. Un CES dédié à la reconnaissance de la parole

Une fois la topologie du réseau définie, nous procédons à l'apprentissage du réseau de neurones expert comme pour un MLP classique. Toutefois, au lieu d'utiliser l'algorithme de rétropropagation de l'erreur, nous allons utiliser celui préconisé et proposé par Gallant à cet effet : l'algorithme Pocket.

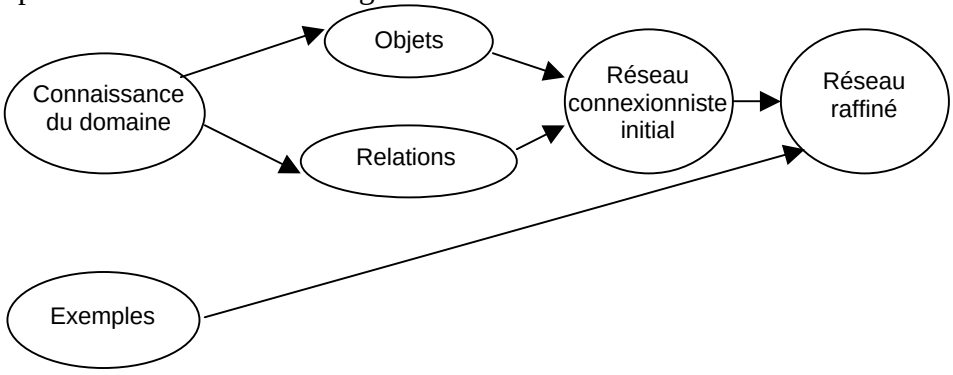


Figure V. 4. L'apprentissage du CES

5. Un KBANN pour la RAP

En suivant l'approche qui se base sur la collecte des objets du domaine, et qui nous a permis de construire le système expert connexionniste pour la reconnaissance des chiffres arabes, nous poursuivons pour la construction d'un réseau KBANN pour la reconnaissance de ces mêmes chiffres.

5. 1. Les propositions

L'approche de Towell est basée sur la logique propositionnelle, nous allons donc considérer que le vocabulaire de l'application est constitué de propositions. Ces propositions seront issues du domaine d'application et représenteront des neurones du réseau. Nous les définissons ainsi: Chaque mot du vocabulaire à reconnaître (les dix chiffres) représente une proposition, ce sont les neurones de sortie. Ces mots sont composés de syllabes ; nous rajoutons autant de propositions qu'il y'a de syllabes qui forment notre vocabulaire. De la même manière, ces syllabes sont reconnues à partir de caractéristiques acoustiques auxquelles nous associons les propositions relatives aux neurones d'entrée.

5. 2. Les clauses de Horn

Comme nous l'avons souligné, les mots sont composés de syllabes, les règles correspondantes sont donc de la forme : Mot_i:-syl_1, syl_2, etc. Les règles relevant de la formation des syllabes sont de la forme : syl_i :-c_j, c_k, etc. Remarquons que nous avons uniquement des atomes positifs. Les C_i sont des classes acoustiques issues de la quantification vectorielle.

5. 3. La structure du réseau

Le KBANN est un Perceptron multicouches, qui possède la structure suivante:

- Une couche d'entrée qui contient 32 neurones, représentant les 32 classes acoustiques issues de la quantification vectorielle. Chaque entrée recevra la valeur 0

ou 1 selon la présence ou l'absence de la caractéristique dans le signal en entrée du réseau.

- Un niveau caché qui comprend deux couches :
  - La première couche cachée contient 29 neurones qui correspondent aux syllabes composant le vocabulaire.
  - Comme nous considérons plusieurs prononciations pour un mot, ceci induit l'ajout de neurones supplémentaires que l'on regroupe dans une seconde couche cachée, ils représentent les variantes phonétiques du mot.
  - Une couche de sortie qui contient 10 neurones représentant les chiffres, car on ne fait pas de distinction à la reconnaissance des variantes phonétiques. On reconnaît le mot quelle que soit sa prononciation. Il est évident que si la transcription du mot devait être voyellée, il faudrait prévoir un neurone de sortie pour chaque prononciation.

Une fois les connexions spécialisées initiées, des connexions avec un poids faible sont rajoutées au réseau. Ensuite, le réseau est entraîné en utilisant l'algorithme de rétropropagation.

Voici quelques résultats que nous avons obtenu à titre comparatif entre les deux approches, sur l'ensemble des dix chiffres sur un ensemble de test et d'apprentissage réduit [Bahi, 04c].

Le corpus d'apprentissage comprend les occurrences de trois locuteurs. Chacun d'entre eux prononce trois fois les dix chiffres. Le corpus de test comprend quatre locuteurs qui prononcent deux fois les chiffres.

CES	KBANN
95.71 %	97.1 %

Tableau V. 1. Résultats comparatifs entre les approches CES et KBANN pour la reconnaissance de chiffres arabes

6. Conclusion

Nous venons de présenter le premier modèle neurosymbolique dédié à la reconnaissance de la parole que nous avons mis au point. Ce modèle présente l'avantage d'être le premier dans la littérature neurosymbolique à s'être intéressé à la reconnaissance de la parole. Les réflexions lors de sa conception nous ont permis d'établir une synergie entre les objets que l'on manipule dans une telle application et les cellules du réseau connexionniste. Mais ce modèle nous a permis aussi en l'étudiant de plus près de prendre conscience d'éventuelles améliorations que nous pouvions lui apporter.

Le choix de la syllabe pour modéliser le niveau phonétique est contraignant et bien qu'il nous ait permis de contourner un peu le problème de précédence temporelle, il ne peut être généralisé car si l'alphabet change ou s'agrandit, le réseau doit être reconstruit, ce qui rend toute extension future très difficile voire impossible, en particulier si on change de langage.

Cette modélisation par la syllabe n'est donc intéressante que si on considère un vocabulaire limité et dont les mots ont des réalisations phonétiques très différentes.

Nous retiendrons quand même que la décomposition en couches du réseau est tout à fait valable, nous décidons donc de la retenir dans une version ultérieure du modèle. Mais pour pouvoir prévoir une extension du modèle à d'autres langages nous décidons de remplacer la syllabe par le phonème.

Reste surtout le problème de séquence dans le temps dans un système qui se veut traiter d'un phénomène tel que la parole où la composante temporelle est omniprésente.

Finalement, nous pouvons dire que le premier modèle fût notre point de départ concret dans ce travail dans le but est d'aboutir à la mise au point d'un modèle perceptuel du processus de reconnaissance. L'apport que nous préconisons alors comparativement aux modèles hybrides existants dans la littérature est l'introduction du paramètre temps dans un modèle neurosymbolique.

Partie B: Proposition d'un neurone temporel spécialisé, Application à la reconnaissance de la parole

1. Introduction

En faisant ressortir graduellement les limitations du modèle expert connexionniste que nous avons présenté en *partie A*, nous sommes arrivés à notre actuelle modélisation en particulier l'introduction du paramètre temps.

Dans cette partie, nous reviendrons sur l'architecture générale du système de reconnaissance et ses composants. La structure du neurone temporel spécialisé (***STN***) que nous proposons sera ensuite présentée.

Enfin, nous allons détailler l'architecture du réseau connexionniste expert une fois que nous avons introduit les neurones ***STN*** ; nous appellerons alors le réseau obtenu ***NESSR*** (acronyme de **N**eural **E**xpert **S**ystem for **S**peech **R**ecognition), sa particularité, 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 système neurosymbolique.

2. Description générale du réseau

2. 1. Introduction au modèle

Le modèle de reconnaissance que nous proposons est un modèle perceptuel qui 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 que nous préconisons.

2. 2. Architecture générale du système

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

- La mémoire de reconnaissance est le réseau perceptuel qui constitue le cœur de notre proposition, il est construit sur les bases symboliques du système expert connexionniste déjà présenté (Partie A).
- La mémoire court-terme est une mémoire où l'on sauvegarde des évènements temporaires qui peuvent survenir lors du raisonnement (exemple §5.3.3).
- La mémoire long-terme contient des informations de haut niveau qui permettent de valider une décision. Ces informations peuvent être relatives à la structure grammaticale du langage ou à des informations pragmatiques sur le contexte. Mais globalement, ce sont des informations de haut niveau quelque soit l'application considérée.

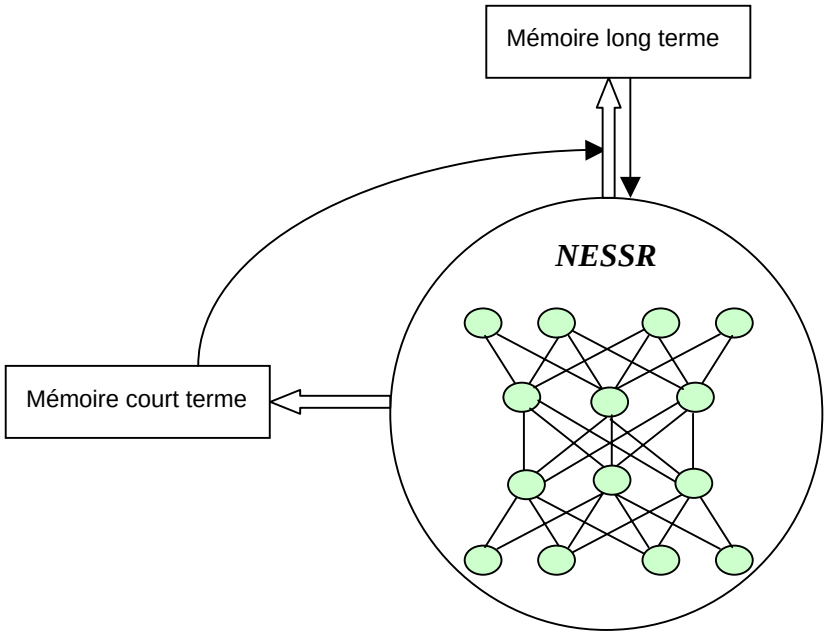


Figure V. 5. Vue générale du système

2. 3. Ancrage des symboles dans le réseau

Le choix de la syllabe comme pierre de voûte dans notre modélisation ultérieure, permet de contourner le problème de séquencement, mais si l'on veut généraliser l'approche, il nous semble impératif de passer au phonème ce qui pose alors la nécessité d'introduire des contraintes de précédence au niveau mot entre les phonèmes.

Les neurones dans le réseau sont agencés en trois couches, conformément à une terminologie perceptuelle, nous dirons que la première est la couche sensorielle, la deuxième est la couche de séparation et la dernière est la couche de décision.

La couche d'entrée est la couche sensorielle. Elle permet de détecter des caractéristiques de la forme à classer. Chaque neurone de cette couche se charge alors de détecter une caractéristique du signal, elle traduit le niveau acoustique de l'application.

La couche cachée permet de faire l'association entre des caractéristiques détectées en entrée et des unités phonétiques de la langue considérée en l'occurrence, l'Arabe. Les unités phonétiques choisies pour illustrer ce niveau sont les phonèmes ; c'est le niveau phonétique.

Les neurones de la couche de sortie représentent la décision, en l'occurrence les unités lexicales à reconnaître ; ce sont des mots.

3. Le modèle du neurone temporel spécialisé

3. 1. Motivations

Si on assume une solution connexionniste dans une application où l'aspect précédence dans le temps est crucial, il faudrait que l'architecture du réseau puisse prendre en charge cette spécificité. Pour une application telle que la reconnaissance de la parole, l'utilisation des modèles tels que ceux présentés dans le chapitre III, la prise en charge du temps émerge de la dynamique du réseau, mais si nous voulons préserver une symbolique associée au réseau, nous pensons que la prise en charge des séquences temporelles doit être considérée au niveau local.

3. 2. Structure des neurones *STN*

Le schéma de la figure V. 6., représente le neurone *n*, ne peut s'activer que si les neurones *i, j, k* et *l* sont activés dans cet ordre.

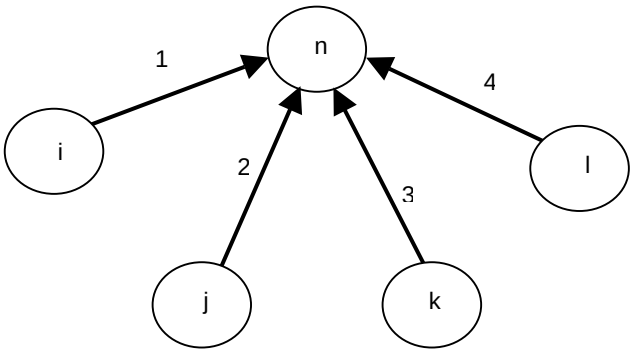


Figure V. 6. Relation d'une cellule *i* avec ses prédécesseurs

Dans un réseau connexionniste spécialisé, les neurones ont une signification, nous proposons donc de libeller chacune des entrées du neurone par sa provenance en respectant les contraintes temporelles. Pour simplifier, considérons un exemple d'ordonnancement de tâches où une tâche critique *i*, ne peut commencer que si les tâches *j, k* et *l* sont terminées dans cet ordre. Pour modéliser cela nous suggérons le modèle de neurone temporel spécialisé ***STN*** (pour **S**pecialized **T**emporal **N**euron) (Figure V. 7).

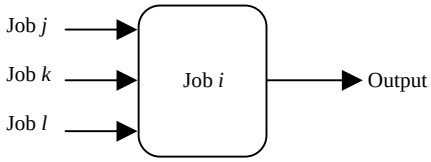


Figure V. 7. Utilisation du modèle *STN* pour un problème d'ordonnancement

Comme le neurone ***STN*** peut avoir plusieurs prédécesseurs. Il peut être considéré comme un filtre probabilisable car il ne s'active pas à chaque fois qu'un signal d'entrée arrive, mais il faut que ce signal soit immédiatement précédé par un signal représenté par la connexion précédente.

Aussi un ***STN*** peut être placé à l'entrée d'un chemin où seuls les signaux qui respectent les bonnes conditions de précédence sont autorisés à traverser.

3. 3. Activation du *STN*

Un neurone ***STN*** peut être non-activé, pré-activé (en attente) ou activé. Lorsque la première entrée du neurone est activé, il devient pré-activé. L'activation des entrées doit se faire dans un ordre bien défini (Figure V. 7), cet ordre est assuré par la structure du neurone où les entrées sont représentées dans une liste et l'accès à un élément de la liste ne peut se faire sauf si on débloque un verrou situé dans l'élément qui le précède. Lorsque toutes les entrées du neurone sont activées, le neurone cible s'active aussi.

4. La couche sensorielle : le niveau acoustique

La couche d'entrée du réseau a pour rôle de détecter les changements dans l'environnement, les neurones de la couche sensorielle 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 des traits caractéristiques dans le signal.

4. 1. Structure des neurones : des neurones spécialisés

Comme ces cellules appartiennent à un réseau symbolique ; chaque cellule se spécialise dans la détection d'une caractéristique : Nous les appelons neurones-classe.

Les caractéristiques sont déterminées en effectuant une classification automatique 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 phonème. 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. Détermination des classes acoustiques

Dans une première étape, nous allons déterminer les classes acoustiques relatives aux phonèmes du langage considéré en l'occurrence l'Arabe.

Comme nous venons de le voir, nous effectuons pour avoir des classes acoustiques une quantification vectorielle, mais en considérant ici les phonèmes comme signal de départ et non plus la syllabe.

4. 3. L'activation d'un neurone

Un neurone-classe est activé si la caractéristique qui lui est associée est détectée dans le signal. Le fonctionnement du réseau est régit par des instants discrets sur l'axe du temps ; on parle alors de top d'horloge. A un instant *t* donné, un seul neurone-classe est activé. Ceci suppose que la présentation d'un signal au réseau se passe entre les t_0 et t_n.

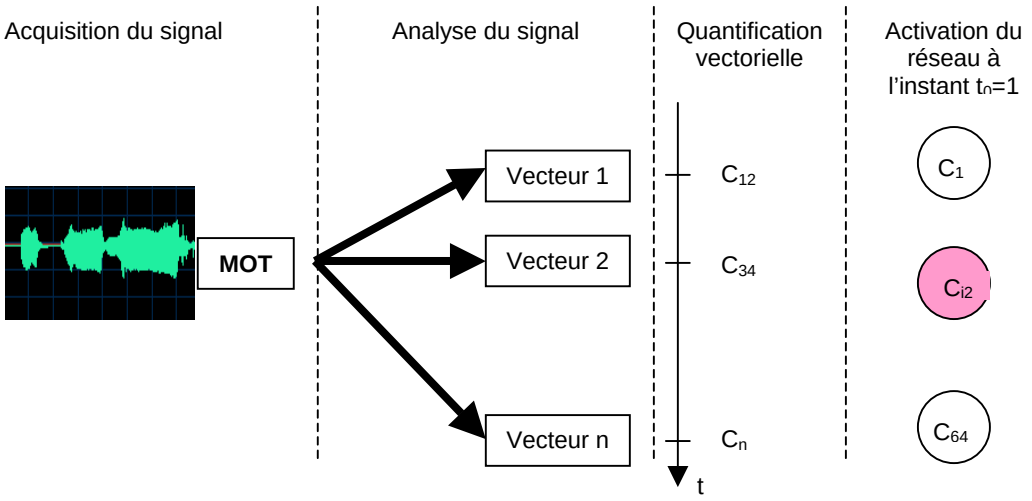


Figure V. 8. Présentation d'un signal à ***NESSR***

Dans cet intervalle de temps plusieurs neurones peuvent être activés. Considérons l'exemple ci dessus (figure V. 8) 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. En phase de quantification, chaque fenêtre sera remplacée par son prototype, on obtient une chaîne de symboles, où chaque symbole représente une classe acoustique.

A l'instant t_0=1, c'est la caractéristique acoustique C_{12} qui est détectée, ce qui induit une activation du neurone-classe correspondant, l'instant suivant c'est le 34ème neurone etc.

Nous remarquerons que l'activation successive du même neurone est tout à fait prise en charge par le réseau, c'est ce que nous verrons ultérieurement.

5. La couche d'association : le niveau phonétique

Les activations de la couche sensorielle 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.

A 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. Structure des neurones: des neurones temporels spécialisés

A l'image des neurones sensoriels, ces neurones d'association 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 phones. 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 V.9).

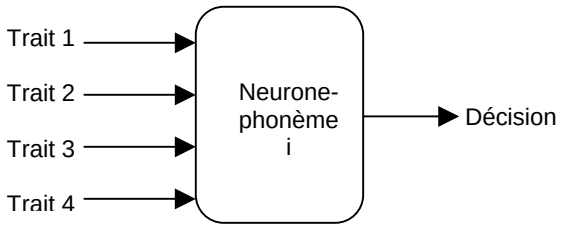


Figure V. 9. Structure d'un neurone-phonème

5. 2. Les connexions

Le réseau n'est pas complètement connecté, un neurone-phonème est connecté seulement aux neurones-classes qui le caractérisent. Ceci a pour avantage de limiter la prolifération des connexions dans le réseau.

5. 2. 1. La caractérisation d'un phonème

En phase d'apprentissage, le signal d'un mot est segmenté et libellé en phonèmes. Les signaux obtenus sont analysés puis quantifiés.

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.

Voici ci-après une ligne du tableau illustrant les relations entre les *M* classes définies par la quantification et l'apparition d'un phonème.

	C1	C2	C3	C4	C5	C6	C7	…	C32
a1_1[1]	×				×	×			

× : classe détectée

Tableau V. 2. Ligne correspondant au phonème /a/

[1] 1ère occurrence du phonème /a/ prononcée par le locuteur n°1.

Ce tableau est construit automatiquement, chaque occurrence d'un phonème dans la base d'apprentissage est analysée, puis discrétisée et à chaque fois qu'une caractéristique apparaît dans le signal, la case est cochée dans la ligne correspondant au phonème dans le tableau.

Sur l'ensemble des occurrences d'un phonème si une caractéristiques apparaît dans plus de 75% des exemples d'apprentissage, nous considérons qu'elle est constituant de base du phonème. Une fois les caractéristiques d'un neurone dégagées leur séquencement est établit.

5. 3. L'activation d'un neurone-phonème

La transmission de l'information de la couche sensorielle à la couche d'association se fait à chaque pas, c'est à dire à chaque fois qu'une caractéristique est détectée.

5. 3. 1. La pré-activation d'un neurone

Lorsqu'une caractéristique soit C_i est détectée, le neurone-classe associé s'active et toutes les connexions qui en partent sont pré-activées. Tous les neurones-phonème dont la première caractéristique est C_i se voient aussi pré-activés.

Un neurone-phonème se met en état de pré-activation 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.

5. 3. 2. L'activation d'un neurone

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 en même temps.

Mais si à un instant donné la détection d'une caractéristique peut provoquer l'activation de plus d'une cellule cible, concrètement cela signifie qu'il y avait plus d'une cellule qui était pré-activée et en attente de la dernière entrée et que cette entrée et

la même. 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 rare (eu égard au nombre de caractéristiques choisi soit 32, et leur pertinence pour la détection d'un phonème), mais il est important de le souligner car la situation reste envisageable. Nous tenterons faire ressortir cette situation au travers de l'exemple ci-dessous.

5. 3. 3. Exemple illustratif d'activation

L'exemple de la figure V. 10, n'est pas un exemple réel, nous y avons illustré des situations particulières qui constituent les conditions limites d'utilisation du modèle, et justifient son utilisation dans une application induisant fortement le paramètre temps.

Dans la réalité de notre réalisation de telles situations ne risquent pas de se produire, vu la large palette de caractéristiques définies.

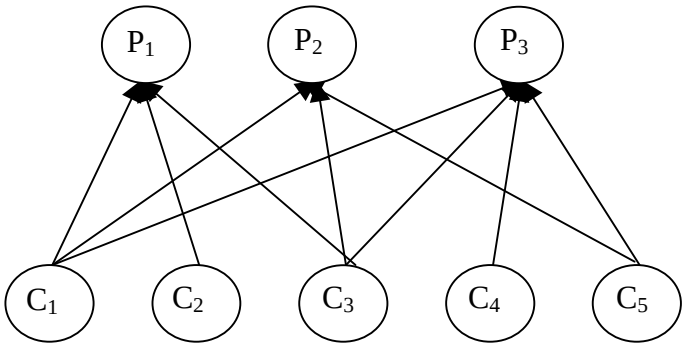


Figure V. 10. Exemple de connexions
neurones-classe / neurones-phonème

Soit le réseau du graphe ci-dessus auquel nous soumettons la séquence: …$C_1C_1C_5C_2C_3C_4$…, l'activation du réseau est transcrite dans le tableau V. 3.

A l'instant T=0, la caractéristique C_1 est détectée, donc le neurone-classe correspondant est activée, et tous les liens qui en partent sont pré-activés. Ceci induit dans le cas de notre exemple la pré-activation des trois neurones-phonèmes du réseau, car C_1 correspond à la première entrée pour tous ces neurones cibles.

A l'instant T=1, la même caractéristique est détectée, mais cette seconde (ou $n^{ème}$) activation successive n'entraîne aucun changement dans l'état du réseau.

A l'instant T=2, la caractéristique C_5 est détectée, ce qui permet la pré-activation de l'entrée C_5 du neurone-p3, mais pas celle du neurone-p2, car cette dernière ne peut l'être avant la pré-activation de C_3.

A l'instant T= 3, la caractéristique C_2 est détectée ce qui pré-active la deuxième entrée du neurone-p1.

A l'instant T= 4, la caractéristique C_3 est détectée ceci pré-active les entrées correspondantes dans les neurones cibles p1 et p2. A cette dernière pré-activation le neurone-p1 se voit toutes ses entrées pré-actives. A ce moment, il s'active. Ce qui désactive toutes les connexions ainsi que les autres neurones cibles (voir tableau V. 3).

	Neurone-p1	**Neurone-p2**	**Neurone-p3**
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
T=4	C1 C2 C3	C1 C3 C5	C1 C5 C4 C3

Connexion non activée
Connexion pré-activée
Connexion activée

Tableau V. 3. Plan d'activation des neurones-phonème

Après C_3, la séquence en entrée du réseau est segmentée et l'activation du neurone-p1 est propagée à la couche suivante. Une nouvelle session de construction d'un phonème débute par C_4.

5. 4. Caractéristiques du modèle des neurones-phonème.

La structure du neurone temporel spécialisé que nous préconisons pour modéliser les phonèmes permet de boucler sur une caractéristique acoustique du signal, comme c'est le cas pour la caractéristique C_1 dans l'exemple précédent. 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 être une classe inhérente au neurone-p1 mais non pertinente à sa détection.

Mais si le nombre de caractéristiques insérées pour un des neurones ***STN*** en compétition en phase de reconnaissance dépassent cinq (5) ce neurone est désactivé et n'est plus en compétition.

6. La couche de décision

A chaque fois qu'un phonème est reconnu, cette détection est propagée à la couche suivante.

6. 1. Structure des neurones

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.

6. 2. Activation

A 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 illustrée dans l'exemple du §5.3.3, ceci désactive définitivement les autres neurones du réseau. A la détection du deuxième phonème d'autres mots parmi les restants sont éliminés etc.

6. 3. La reconnaissance

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.

Si d'autre part, le mot reconnu n'est pas en accord avec le mot que peut attendre la mémoire à long-terme, on peut revenir sur ce choix, en invoquant les choix effectués lors de l'activation simultanée de neurones-phonème et qui sont sauvegardés dans la mémoire court-terme.

7. Conclusion

Dans ce paragraphe, nous ne posons pas un point final à ce chapitre en concluant sur un modèle totalement fini. Tout d'abord parce que le modèle n'est pas terminé dans le sens où il n'est pas fourni « clé en main » pour une application donnée. Ensuite parce que nous n'avons pas encore évalué les capacités et les possibilités de l'architecture, ceci sera fait au chapitre suivant. Nous pouvons cependant résumer d'un point de vue théorique les motivations qui nous ont conduit à ce type d'architecture connexionniste, discuter des points forts et des points faibles qui sont à nos yeux attachés au modèle.

Nous l'avons dit en introduction à cette thèse, la plupart des problèmes auxquels l'homme et la machine sont confrontés incluent une composante dynamique, un aspect temporel. Il était donc souhaitable que les réseaux connexionnistes prennent en compte cet aspect temporel dans le cadre d'applications diverses comme la reconnaissance de la parole. La communauté scientifique internationale en connexionisme s'est donc tout naturellement attaquée à ce problème. Ainsi, divers modèles ont été proposés, nous en avons vu les grandes lignes au chapitre III. Notre approche ne se situe pas dans le courant majeur des modèles connexionnistes temporels. Il s'agit en effet de s'inspirer de faits neurobiologiques pour construire une architecture perceptuelle.

Dans cette ligne droite, nos travaux gardent à l'esprit deux objectifs: introduction de la symbolique dans le réseau connexionniste et prise en compte du paramètre temps. Il est cependant difficile d'atteindre simultanément ces deux objectifs. Nous avons commencé par introduire la composante symbolique dans le réseau [Bahi, 04a][Bahi, 05a] et c'est ensuite que le modèle ***STN*** s'est imposé pour nous [Bahi, 05c]. En partant de l'application, la mise au point du modèle se trouve nécessairement influencée par celle-ci. Les résultats enregistrés sont alors bons mais le modèle a perdu de sa généricité. Dans notre travail, il est difficile de se détacher du domaine d'application lors de l'ancrage des symboles mais nous avons essayé lors de la mise au point du modèle ***STN*** de ne pas tomber dans cet écueil.

La généricitié de l'architecture semble donc être un avantage pour la conception de systèmes assez complexes traitant en particulier de plusieurs axes sensoriels.

Chapitre VI:

Evaluation du modèle

Chapitre VI: Evaluation du modèle

Présentation du chapitre

Dans ce chapitre nous présentons un plan d'évaluation de notre modèle, ainsi que les principaux choix d'implémentation que nous avons effectué, pour le mettre en œuvre.

Le plan d'évaluation que nous proposons inclut deux niveaux: un niveau phonème, pour pouvoir évaluer 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 évaluer les performances de notre modèle par rapport à des techniques classiques qui ont fait leurs preuves.

1. Introduction

Nous apportons dans ce travail une proposition d'un réseau connexionniste qui inclut la sémantique du domaine dans sa structure. Pour évaluer les performances de ce modèle et en particulier celles du modèle de neurone que nous proposons, nous suggérons une application à la reconnaissance de mots isolés.

Dans ce qui suit, nous allons présenter les différentes étapes de la mise en œuvre de ce modèle.

2. Extraction des caractéristiques

2. 1. Echantillonnage

Les données se présentent sous forme des fichiers WAV, enregistrés dans une salle fermée à une fréquence d'échantillonnage *: f_e = 11025 Hz* sur 16 bits.

En effet, la fréquence d'échantillonnage f_e doit être au moins égale à la fréquence maximale du signal à numériser (Théorème de Shanon); mais elle peut varier en fonction du domaine d'application ou des besoins ou contraintes matériels, à titre illustratif nous avons,

	Bande passante	Fréquence d'échantillonnage et précision
Téléphone	300-3500 Hz	8 KHz 8 bits
Voix	60-10000 Hz	16 KHz 16 bits
Hifi	10-18000 Hz	44 KHz 16 bits

Table VI. 1. Variation de la fréquence selon l'application

Si on se pose la question de pourquoi ne pas avoir utilisé une plus grande fréquence d'échantillonnage ou une plus grande précision. Sachant qu'un phone échantillonné à une fréquence de 8000Hz avec une précision de 8 bits préserve la plus grande partie de l'information apportée par le signal [Rabiner, 93], il apparaît ne pas être nécessaire d'aller au delà de cette fréquence ou de cette précision. D'autre part, ce choix fût adopté pour pouvoir travailler sur la même base (en particulier en ce qui est des chiffres) que celle qui a servi à établir des résultats avec d'autres travaux et pouvoir ainsi effectuer une étude comparative.

2. 2. Isolement du mot

Bien que les occurrences de mots soit séparés par des silences, il est important de déterminer les frontières d'un mot pour connaître exactement le signal qui le caractérise.

2. 3. Pré-accentuation

Comme il est commun dans les systèmes de reconnaissance de la parole, un filtre de pré-accentuation est appliqué au signal échantillonné. En effet, les caractéristiques des

cordes vocales définissent les phonèmes prononcés. Ces caractéristiques sont visibles dans le domaine fréquentiel par la localisation des formants, i.e. les pics de résonance des cordes. Toutefois, les formants de hautes fréquences ont une amplitude moindre que celles des formants de basse fréquence. Une pré-accentuation des hautes fréquences est alors requise pour avoir des amplitudes similaires pour toutes les fréquences. Ce traitement est généralement obtenu en faisant passer le signal par un filtre FIR du premier ordre (voir Chap. II), nous prenons un paramètre de pré-accentuation égal à 0.97, ce qui donne une amplification pour les hautes fréquences du signal de plus de 20 dB.

2. 4. Fenêtrage

Les méthodes classiques d'évaluation spectrales que nous avons présenté à la section II.3.3.2, supposent que l'on opère sur un signal stationnaire (i.e. un signal dont les caractéristiques statistiques sont invariantes dans le temps). Pour la parole, ceci n'est vrai que pour des intervalles de temps courts durant lesquels une analyse peut être effectuée. Il faut alors bloquer le signal x'(n) en une succession de trames $x_t(n)$ avec t=1,2,…,T. Ces trames sont appelées fenêtres.

Une fois la fréquence d'échantillonnage fixée, augmenter la résolution spectrale revient à utiliser des séquences du signal plus longues, ce qui est en contradiction avec la nécessité d'avoir des segments stables du signal. Des études ont monté qu'une longueur entre 20-40 ms constitue un bon compromis.

Dans notre application nous choisissons des fenêtres de longueur N=512 échantillons ce qui correspond à un intervalle de temps de 512/11.025 ≈ 46ms. Avec un recouvrement de 100 échantillons entre les trames.

2. 5. Fenêtrage de Hamming

Avant de calculer la transformée de Fourier du signal, chaque trame du signal est individuellement pondérée par une fenêtre (voir chap. II). Nous utilisons une fenêtre de Hamming.

2. 6. Analyse MFCC

De nos jours l'ensemble de caractéristiques le plus utilisé est celui des coefficients MFCC (voir II. 2.3.3.). Les coefficients MFCC (Mel Fequency Cepstral Coefficients) sont des coefficients cepstraux obtenus à partir des énergies d'un banc de filtres en échelle de fréquence Mel. Cette méthode présente l'avantage d'être résistante au bruit et d'avoir une plausibilité biologique puisqu'elle utilise une échelle psycho-acoustique des fréquences similaire à celle de l'oreille interne.
Une dizaine de coefficients cepstraux sont généralement considérés comme suffisants pour les expériences en reconnaissance de la parole. Nous utilisons les 12 premiers coefficients cepstraux obtenus à partir d'un banc de 26 filtres sur une échelle fréquentielle de Mel, le logarithme de l'énergie de la trame, ajoutée aux 12 coefficients cepstraux pour former un vecteur de 13 coefficients.

3. La quantification vectorielle

Une fois l'analyse du signal effectué, il faut procéder à une quantification sur les vecteurs acoustiques afin de les faire correspondre aux entrées du réseau. Chaque vecteur correspondra à une caractéristique qui sera un symbole du vocabulaire issu de la quantification vectorielle.

3. 1. Définition

La quantification vectorielle (QV), est une méthode puissante qui consiste à utiliser les propriétés statistiques des sons dans leur espace de représentation. Cette technique rentre dans la problématique plus générale de la classification automatique et connaît actuellement de nombreux développements dans le domaine de la parole. Elle part du postulat que deux formes proches dans leurs espaces de représentation sont proches en soi.

Soit $x=[x_1, x_2, \ldots, x_k]$ un vecteur réel à valeurs continues dans R^k. La quantification de x consiste à lui substituer un vecteur voisin $y_i \in R^k$ (i=1,2,..,M) choisi parmi un ensemble fini de M vecteurs. Construire un système de QV, c'est opérer une partition de R^k en classes C_i ; dans chacune d'entre-elles, on distingue un vecteur particulier y_i appelé prototype ou centroïde (code-word). Chaque vecteur $x \in C_i$ sera représenté par le centroïde associé y_i. L'ensemble des centroïdes constitue un dictionnaire (code-book).

La substitution de *x* par le vecteur *y* provoque une distorsion notée d(x,y). Pour le codage des formes d'onde, on utilise le plus souvent l'erreur quadratique:

$$d\quad(x,y)=\left[\sum_{i}(x_i-y_i)^2\right]/K \qquad \text{(VI. 1.)}$$

Le but poursuivi dans l'établissement d'un système de codage est de minimiser la distorsion moyenne intra-classe et de maximiser la distance inter-classe.

3. 2. Etablissement des classes par la méthode de LLOYD généralisée (K-means method)

On dispose d'un ensemble L de vecteurs *x* que l'on désire positionner en *M* classes. On désignera par :

- $x_j^{(i)}$, les vecteurs appartenant à la classe i
- y_i , le centroïde de la classe i
- L_j , le nombre de vecteurs de la classe i
- $d(x_j^{(i)}, y_i)$, la distance ou mesure de distorsion entre $x_j^{(i)}$ et y_i

  $D_i = \sum_{j} d(x_j^{(i)}, y_i)$

- D , la distorsion pour l'ensemble des vecteurs

  $D = \sum_{i=1}^{M} D_i$

Un nombre *M* de classes étant imposé *à priori*, le problème consiste à trouver la partition et les centroïdes de façon à minimiser la distorsion totale *D*. Une procédure itérative peut être basée sur les deux observations suivantes:

- Pour un ensemble donné de centroïdes, la partition qui minimise *D* est celle pour laquelle chaque vecteur x_j est affecté à la classe dont le centroïde est le plus rapproché.

- Pour une partition donnée, il existe pour chaque classe *i* un vecteur *j* qui minimise la distorsion totale D_i de la classe *i.*

Nous avons utilisé dans ces travaux une valeur de M égale à 32.

4. La reconnaissance de phonèmes

4. 1. La base de données

Nous considérons une base qui comprend les occurrences de quinze (15) locuteurs. Chacun des dix (10) premiers locuteurs prononce trois (3) fois chacun des 31 phonèmes, les cinq (5) restants prononcent une seule fois chaque phonème.

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) en utilisant les deux premières occurrences pour chaque phonème. Pour effectuer les tests d'évaluation nous formons les groupes:

- Le groupe TG1 comprend les deux premières occurrences pour chaque phonème des locuteurs qui ont participé à la phase d'apprentissage (base d'apprentissage).
- Le groupe TG2 comprend la troisième occurrence des locuteurs qui ont participé à la phase d'apprentissage.
- Le groupe TG3 comprend les occurrences des locuteurs qui n'ont pas participé à l'apprentissage.

4. 2. Les résultats

Dans la table ci-dessous, nous mentionnons les taux de reconnaissance (en %) que nous avons obtenu pour les phonèmes considérés (les phonèmes sont donnés en notation IPA).

	[a]	[u]	[i]	[m]	[ħ]	[w]	[□]	[q]	[s]	[d]
TG1	99.2	98.4	98.2	97.6	97.1	98.4	99	98.8	97.6	99.1
TG2	98.7	98	98.5	96.8	96	98	98.5	98.5	97.3	98
TG3	98.3	98	98.2	96	96.1	98	98.4	98	97.2	98.2

Table VI. 2. Taux de reconnaissance de quelques phonèmes

Les performances du modèle sont bonnes, en particulier si on tient compte du domaine d'application, car ces résultats sont largement tributaire de l'ensemble de caractéristiques choisies pour représenter un phonème (résultat de la QV) et du seuil de pertinence choisi pour qu'une caractéristique soit reliée à un phonème.

Et 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.

Dans la table VI. 3., nous présentons les taux de reconnaissance pour tous les phonèmes en considérant les différent groupes.

Table VI. 3. Taux de reconnaissance des phonèmes

TG1	TG2	TG3
98.6	98.2	97.5

Le fait que les groupes TG1 et TG2 aient des taux de reconnaissance proches est probablement dû au fait que nous avons les mêmes locuteurs dans TG1 et TG2. Ce taux est moindre pour TG3 car ce sont de nouveaux locuteurs.

5. La reconnaissance de 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***.

5. 1. Reconnaissance en mode mono-locuteur

Le niveau mot est d'abord évalué en mode mono-locuteur. Nous considérons le locuteur n°1, qui prononce huit (3) fois chacun des mots du tableau VI. 4.

[sifr]	[waaʜid]	[□abal]	[kataba]
99.2	99	99.3	98.1

Table VI. 4. Taux de reconnaissance de quelques mots en mode mono-locuteur

5. 2. Reconnaissance en mode multi-locuteurs

Dans le cadre de la reconnaissance multi-locuteurs, nous considérons la base des 15 locuteurs où chacun prononce deux fois chacun des chiffres, et nous formons les groupes :

- Le groupe TG4 comprend les occurrences des chiffres des locuteurs qui ont participé à la phase d'apprentissage.
- Le groupe TG5 comprend les occurrences des locuteurs qui n'ont pas participé à l'apprentissage.

Voici les résultats rencontrés en considérant les groupes TG4 et TG5.

	0	1	2	3	4	5	6	7	8	9
TG4	98.1	98.3	97.3	98.1	98.5	98.6	97.1	98.7	98.9	97.5
TG5	97.8	98.1	97.1	97.5	98.4	97.4	97.1	98.2	98.5	97.4

Table VI. 5. Taux de reconnaissance de quelques chiffres en mode multi-locuteurs

5. 3. Etude comparative

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

Pour ces tests nous avons utilisé un Perceptron multicouches 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 dix (10) gaussiennes (pour plus de détails voir [Becchitti, 99]). Le réseau *NESSR* est composé de 32 neurones-classe, 31 neurones-phonème et 10 neurones-mot (les dix chiffres).

Pour effectuer cette étude comparative, nous considérons nos 10 premiers locuteurs chacun d'entre eux prononce trois fois les dix chiffres. Nous utilisons les deux premières occurrences pour l'apprentissage du MLP et des HMMs. Et nous formons le groupe TG6 qui comprend la troisième occurrence. Dans la table VI.6., nous mentionnons les résultats obtenus avec les différentes implémentations.

	MLP	HMM	*NESSR*
TG6	98.9	99.7	98.5

Table VI. 6. Résultats comparatifs

6. Conclusion

A l'heure du bilan chiffré, nous remarquons que les taux de reconnaissance pour les phonèmes sont très prometteurs. Toutefois, l'étude comparative de résultats est évidemment en faveur des HMMs ; les HMMs pour être l'outil de modélisation le plus puissant quand il s'agit de gérer des contraintes temporelles, d'un autre côté ont eu le temps en leur faveur dans le sens où ce sont des outils qui existent depuis longtemps et qui ont bénéficié de beaucoup de mises au point. Mais bien que les taux de reconnaissance obtenus avec le modèle ***NESSR*** soient un peu moindre ils ont pour eux le temps pour améliorer le système et l'explicabilité qui se retrouve encore une fois en compétition avec les performances calculatoires.

D'un autre côté, les performances de ***NESSR*** sont proches du MLP, et pour certains mots, elles les dépassent même, ceci se justifie par la taille réduite du vocabulaire et les réalisations phonétiques des mots considérés qui sont différentes. Alors pour montrer l'intérêt d'utiliser un tel modèle nous proposons une application de la reconnaissance de la parole, dans laquelle le raisonnement mené pour la reconnaissance d'un mot est important, il s'agit de la détection de la dyslexie chez de jeunes élèves.

Chapitre VII:

Application à la détection de la dyslexie

Chapitre VII: Application à la détection de la dyslexie

Présentation du chapitre

Le but de ce chapitre est de présenter un cadre d'utilisation de ***NESSR*** de mettre en avant son aspect explicabilité, et ce au travers d'une application que nous avons développé en reconnaissance de la parole. Il s'agit de la détection automatique de la dyslexie chez de jeunes élèves. Cette application nous l'avions précédemment élaborées mais en utilisant les HMMs à la base du système de reconnaissance [Bahi, 05b]. L'utilisation du modèle ***NESSR*** va profondément affecter le processus de détection, et il a fallu repenser la conception afin d'exploiter le raisonnement déductif que l'on observe lors de la reconnaissance. Finalement, nous utiliserons en conjonction les deux classifieurs à savoir le système basé sur les HMMs et celui basé sur le modèle ***NESSR***.

Le système réalisé se compose de deux parties que nous allons développer dans ce chapitre: la première partie concerne le module d'évaluation qui inclut le module de reconnaissance de la parole et qui sert à évaluer les aptitudes de l'élève et un module de décision qui lui se base sur le raisonnement à base de cas mais qui aussi tire profit de la composante symbolique du système de reconnaissance.

1. Introduction

La dyslexie est un trouble de la parole dont l'élément révélateur est la lecture. Or, la lecture est aussi l'apprentissage le plus laborieux et le plus fondamental qui est demandé au jeune élève; il est donc évident que si cet apprentissage est déficient c'est toute la scolarité de l'enfant qui est compromise, et probablement plus tard son avenir social et professionnel.

Nous ne disposons pas de statistiques officielles sur le nombre d'enfants scolarisés et ne sachant pas lire correctement dans les petites classes, mais les échos de nos écoles ne laissent guère de doute quant à l'ampleur du problème.

Comme la dyslexie se caractérise par la mal-lecture, nous avons envisagé une détection automatique de ce trouble comme une des applications de la reconnaissance automatique de la parole. Et c'est dans une optique d'aide aux éducateurs et aux parents que nous nous sommes attelés à réaliser un logiciel convivial et attrayant pour l'enfant et pouvant aider à détecter ce trouble chez l'enfant scolarisé dans les premières classes, et envisager d'aider ultérieurement à sa rééducation. Cet outil était un but que nous poursuivions parallèlement à nos investigations dans la théorie des systèmes hybrides. Mais une fois le modèle ***NESSR*** mis en place nous souhaitons mettre en avant l'explicabilité du modèle par le biais de cette application, car la manière dont l'enfant agence les unités du langage est reproduite par le réseau de reconnaissance, ce qui peut aider à localiser les difficultés de l'enfant et décider de l'existence d'un éventuel trouble. Toutefois, même avec un système de reconnaissance performant, il est impératif de disposer d'un outil de décision, qui en se basant sur les données en entrée puisse décider de l'absence ou de la présence du trouble chez l'enfant.

L'application comporte deux composants: un module d'évaluation qui se base sur les techniques de reconnaissance automatique de la parole et un module de décision basé sur le raisonnement à base de cas. L'approche se compose donc de deux parties, d'une part, nous avons la mise en œuvre d'une plate forme constituée d'une panoplie d'outils de reconnaissance automatique de la parole arabe. D'autre part, il s'agit de mettre ces outils à la disposition du système de détection de la dyslexie (décision).

Comme, nous avions déjà développé un système de reconnaissance basé sur les HMMs [Bahi, 05b], nous souhaitons le garder comme second classifieur au côté du modèle ***NESSR***, afin de renforcer le résultat de la reconnaissance.

Quant à l'utilisation du raisonnement à bas de cas (Case Based Reasonning) dans le système de décision, elle fût motivée d'une part, par l'absence de règles systématiques permettant de poser un diagnostic irréfutable et d'autre part, par le caractère approximatif du raisonnement à mener.

Dans ce chapitre, nous allons présenter les deux modules de l'application ainsi que quelques résultats expérimentaux.

2. La dyslexie : la mal-lecture

Parmi les troubles de la communication, la dyslexie est un trouble de la parole très répandu et tout aussi méconnu dans notre société. Elle se manifeste par des lacunes profondes dans les acquisitions du langage écrit, le jeune élève a du mal à mettre en place un système de reconnaissance de mots, il ne peut de ce fait ni décoder ce qui est écrit ni accéder à son sens.

Les anomalies les plus fréquentes issues de la dyslexie se manifestent soit dans le décodage du message, soit dans sa compréhension ou dans les deux [Mucchielli, 79] [Davis, 97]. Les problèmes les plus fréquents sur le plan du décodage sont:

- Des confusions auditives ou phonétiques,
- Des inversions,
- Des omissions,
- Des adjonctions,
- Des substitutions,
- De la contamination,
- Une lecture du texte lente, hésitante, saccadée, avec un débit syllabique,
- Une difficulté à saisir le découpage des mots en syllabes, une ignorance de la ponctuation.

Sur le plan de la compréhension, le dyslexique ne saisit qu'un sens partiel ou pas de sens du tout, de ce qu'il a déchiffré; ainsi, le message du texte lui échappe totalement ou partiellement. On rencontre fréquemment des cas où il y a conjonction de ces deux types de difficultés.

3. Un système de détection de la dyslexie

3. 1. Présentation générale

Nous proposons un environnement dans le but de détecter la dyslexie chez de jeunes élèves. Ce système est composé de deux modules principaux : un module d'évaluation qui inclue le système de reconnaissance et un module de décision qui se base sur le raisonnement basé cas.

Le module d'évaluation comprend un ensemble de tests en relation avec les manifestations de la dyslexie qui sont proposés à l'enfant. La pertinence des résultats de ces tests reste largement tributaire des performances et des particularités du système de reconnaissance.

Parallèlement, un profil social de l'enfant est nécessaire pour poser un pronostic, ce profil peut être défini conjointement par l'enseignant et les parents. Ce profil est décrit par le biais d'un questionnaire portant sur les aptitudes sociales de l'enfant ainsi que son environnement. Les tests et le questionnaire sont les éléments constitutifs du module d'évaluation.

Le module de décision, opère une correspondance entre les résultats fournis par le système d'évaluation et les descripteurs du cas dans la mémoire de cas dont nous disposons. Cette recherche peut produire un ou plusieurs cas qui couvrent la pathologie du cas cible.

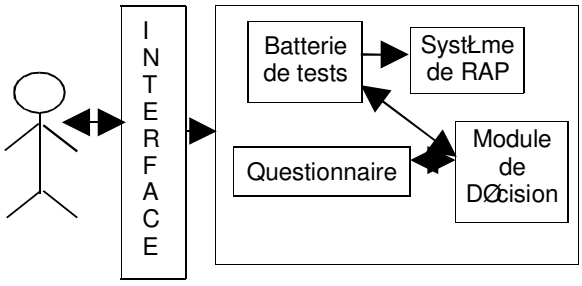


Figure VII. 1. Vue gØnØrale du systŁme [Bahi, 05b]

3. 2. Batterie de test

Comme définie précédemment, la dyslexie est un trouble oral de la lecture; nous nous sommes alors attachés au repérage des anomalies révélatrices du trouble en nous basant sur le système de reconnaissance. Ces manifestations ont été étudiées et adaptées à la langue arabe. La batterie de test comprend:

- Des exercices de calcul: les exercices sont du type « 3+4= . » ou « 8 + . =15 ». Un enfant dyslexique est globalement bon en calcul, le but de ce test est donc d'écarter d'autres déficiences de l'élève.
- La lecture des mots isolés: il s'agit de lire un certain nombre de mots présenté à l'enfant. Le but est de tester la capacité de l'enfant à reconnaître des formes qui lui étaient déjà présentées car les mots que nous lui proposons sont issus du livre de la 1ère année élémentaire. Ce test sera plus amplement décrit dans le prochain paragraphe.
- Des exercices d'association image-texte: le but de ce test est d'évaluer la capacité de l'enfant à associer un texte à l'image correspondante.

3. 3. Principe du test de lecture

Le but de ce test est de repérer des erreurs du style : inversion de syllabes, confusion entre les lettres et les sons,…..

- Des inversions de lettres ou de syllabes par exemple pour ([samʃ] pour [ʃams]) et pour ([kabata] pour [kataba])
- Des confusions de lettres et de sons, que l'on qualifie de confusions auditives ou phonétiques: / / (ʃ / s ou ɜ / z)
- Des ajouts de lettres et de sons par exemple : pour ([ba□ba□aa?] pour [baba□aa?]).
- Les omissions de lettres et de sons par exemple : pour ([mustafa] pour [mustaʃfa]).

3. 4. Profil social

En plus des tests que l'on fait passer à l'enfant, il est impératif de disposer d'une description fiable du profil social, psychologique et pathologique de l'enfant. Pour cela, un questionnaire est proposé à l'accompagnateur de l'enfant ce qui permet d'affiner le diagnostic en écartant les causes pathologiques ou sociales. Voici un exemple des questions possibles :

- Est-ce que l'enfant entend bien ?
- Est-ce que l'enfant a de la peine à maîtriser les concepts comme gros et petit, haut et bas, gauche et droite ?
- Est-ce que l'un des parents de l'enfant est malade ou décédé?
- Est-ce que l'enfant a des difficultés à penser de façon logique ?

Les réponses à ces questions sont codées par oui/non. Une amélioration possible pour ce test serait d'attribuer une note à chacune des questions ce qui peut très bien être pris en considération lors de la décision.

4. Module de reconnaissance

A l'instar de nombreux travaux et de systèmes, nous adoptons dans cette application un module de reconnaissance basé sur l'approche probabiliste que représentent les modèles de Markov cachés, mais en association avec ce classifieur nous préconisons un système de reconnaissance basé sur le modèle ***NESSR***. Dans cette application, nous avons retenu huit (8) mots. Pour chacun de ces mots nous avons construit un modèle HMM (pour plus de détails voir [Bahi, 05b]).

4. 1. Le classifieur basé *NESSR*

Le système ***NESSR*** est utilisé conjointement avec les HMMs. Dans cette application le niveau phonétique de ***NESSR*** est représenté par les syllabes, car comme nous l'avions dit ceci peut être très intéressant pour des applications où le vocabulaire est réduit et où les mots ont des réalisations phonétiques différentes. D'autres part, un dyslexique présente la particularité d'inverser les syllabes, ce qui devrait être détecté grâce à ***NESSR*** et ses neurones temporels.
A tout moment, nous connaissons le mot à lire, donc nous nous limiterons dans l'architecture de ***NESSR*** aux niveaux acoustique et phonétique.

4. 2. Le processus de reconnaissance

Voici dans ce qui suit les étapes induites par cette partie du test :

1. La calligraphie du mot à lire est affichée à l'écran, le mot à lire n'est pas voyéllé.

Figure VII.2. Exemple de lecture proposée

2. L'enfant lit le mot
3. Le système de reconnaissance est sollicité pour aligner le signal reçu avec le modèle du mot à lire (force-align approach). Le résultat renvoyé est donc une probabilité qui renseigne sur la vraisemblance entre le mot lu et le mot à lire.
4. Parallèlement, le signal est injecté dans un système du type ***NESSR***, où les neurones cachés représentent les phonèmes et les neurones de sortie représentent les syllabes.

A chaque fois qu'une syllabe est reconnue, son index est rajouté à une liste de la séquence lue, que nous sauvegardons dans le mémoire-court terme. A la fin de la lecture, cette séquence sera rajoutée au profil de l'enfant

(Meriam, Yacine, 7, ...)

(yes, yes, yes, no, no, yes, ...)

(8,6,14,23,5,....)

((0.85, [man, zi, lun]), ...)

(1,1,2,3,...)

Figure VII.3. Exemple du profil d un enfant

5. Module de décision

Comme nous l'avons précédemment souligné, le système est formé de deux composantes, une première partie regroupant les outils d'évaluation des aptitudes de l'enfant, comprenant en particulier le module de reconnaissance et une seconde composante qui consiste en le module de décision qui sur la base des éléments fournis

par cette panoplie d'outils doit décider de l'absence ou de la présence du trouble (il peut arriver que le système n'arrive pas à décider).

Vu la complexité du profil de l'enfant, incluant parfois l'absence de certaines informations, l'absence de règles systématiques pour poser un diagnostic, nous avons choisi le raisonnement à base de cas [Kolodner, 93] [Jaczynski, 94] à la base du module de décision.

5. 1. Structure du cas

La description du cas dans ce système est répartie en deux catégories de dimensions. La première concerne la partie problème et contient les descripteurs issus de la phase d'évaluation. La seconde partie est la conséquence du problème, décrite en termes de diagnostic et conduite à tenir.

- Partie problème : elle est constituée du profil de l'enfant (figure VII.3)
- Partie conséquence : elle contient le diagnostic, quant à l'existence ou l'absence du trouble. Dans le cas d'absence de dyslexie cette partie peut contenir une éventuelle explication des difficultés que présente l'enfant.

5. 2. Recherche de cas similaires

La mémoire de cas est organisée en arborescence avec des descripteurs discriminatifs. Pour organiser la mémoire nous retenons des index que nous appelons descripteurs majeurs. Ces index servent dans la recherche des cas similaires en phase de filtrage.

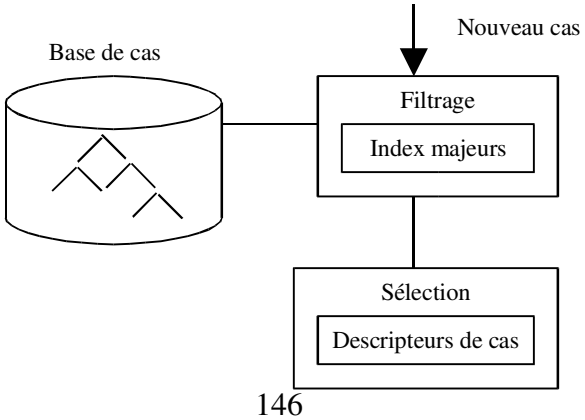


Figure VII.4. Organigramme de recherche [Bahi, 05b]

Le premier groupe de descripteurs majeurs est issu du questionnaire, ils servent dans une première étape à écarter les nouveaux cas qui présentent des profils pathologiques ou qui ont des problèmes sociaux. Un exemple typique de tels descripteurs concernerait la vision ou l'audition de l'enfant. La question correspondante dans l'arborescence est:

If questionnaire[i]=yes or questionnaire[j]=yes then inspect SB1.

Où *i* et *j* représentent les indices des questions relatives aux problèmes de vision et d'audition. Dans ce cas les cas similaires appartiennent à une sous-base de cas *SB1*. *SB1* est une sous-base qui contient des cas d'enfants qui ont d'autres pathologies ou des problèmes familiaux ce qui peut fausser le diagnostic de la dyslexie, il faut donc prendre en considération ces problèmes avant de refaire un test.

Le second groupe de descripteurs majeurs sert à partager les cas restants en deux sous-bases *SB2* et *SB3*. *SB2* contient des cas d'enfants qui ont d'autres problèmes d'apprentissage ce qui peut aussi masquer la dyslexie et *SB3* est la sous-base où la sélection doit avoir lieu. *SB3* contient à la fois des cas de dyslexiques et de non dyslexiques. La seconde question dans l'arborescence est donc :

If number-ww $\geq$ th1 and number-cc $\leq$ th2 then inspect SB2 Elsewhere inspect SB3

number-ww: est le total des prononciations fausses (on considère qu'une prononciation est erronée si sa vraisemblance est inférieure à 0.5)

th1: est un seuil correspondant au maximum d'erreurs tolérées

number-cc: est le nombre des réponses correctes aux exercices de calcul,

th2: est le seuil correspondant.

Donc le résultat du filtrage est un sous-ensemble de cas identifiés par les descripteurs majeurs. Lorsque le système détecte d'autres troubles ou problèmes qui

peuvent gêner le diagnostic (SB1 ou SB2), une explication du trouble peut être fournie, en prenant en considération les réponses au questionnaire. Dans le cas contraire, une sélection est opérée parmi les cas restant en utilisant une fonction de similarité.

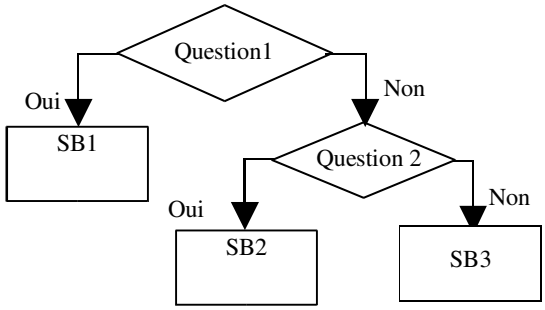


Figure. VII. 5. Organisation de la base de cas

SB3 contient des cas de dyslexique et de non dyslexique, car un enfant au profil normal (***number-ww ≤ th1 and number-cc ≥ th2)*** se trouve dans cette sous-base.

Durant le processus de sélection, nous utilisons une fonction de similarité qui utilise les tests d'évaluation. Une des possibilités simple de fonction est: deux cas sont proches s'ils ont plus de 5 réponses identiques pour le premier et le deuxième test et plus de 3 pour le dernier test.

6. Résultats

Les résultats que nous allons présenter n'utilisent pas pour la détection le classifieur ***NESSR***, le résultat de ce dernier peut être consulté afin de conforter ou de revenir sur une décision.

Nous considérons un groupe de 30 élèves 15 filles et 15 garçons. Voici les diagnostics donnés par le système en comparaison avec le diagnostic réel pour ces enfants.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
Diagnostic réel	D	**D**	ND	ND	ND	ND	ND	ND	D	ND	ND	ND	ND	ND	**ND**
Diagnostic du système	D	**ND**	ND	ND	ND	ND	ND	ND	D	ND	ND	ND	ND	ND	**D**

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15
Diagnostic réel	D	ND	ND	ND	ND	D	ND	ND	D	ND	ND	ND	ND	ND	**ND**
Diagnostic du système	D	ND	ND	ND	ND	D	ND	ND	D	ND	ND	ND	ND	ND	**D**

D : dyslexique

ND : non dyslexique

Tableau VII.1. Tests d Øvaluation

Si nous considérons la table ci-dessus, nous remarquons les erreurs commises au niveau du diagnostic de F2, F15 et G15. F2 est une fille de 8 ans qui a été dirigée vers SB2, F15 a fait des erreurs de lecture mais ceci est probablement dû à son jeune age (6 ans). G15 est un cas intéressant car il a été diagnostiqué dyslexique vu son taux d'erreurs de lecture jugé élevé par le système, mais il avait fait de bonne association texte/image ; c'est un cas tangent où la sortie de ***NESSR*** peut être exploitée pour vérifier le diagnostic.

Conclusion et perspectives

Conclusion et perspectives

1. Bilan

De part les objectifs définis au début de notre travail, notre contribution dans cette thèse revêt plusieurs aspects, il s'agit d'abord d'un travail d'investigations des particularités de la parole, ce que nous avons fait en étudiant les particularités acoustiques et structurelles de la syllabe Arabe [Bahi, 01b]. Ce travail, nous a permis de faire ressortir l'importance de l'utilisation de la syllabe dans des applications dédiées complètement à l'Arabe, ce travail nous semble être d'un grand intérêt car de telles études sont très rares voire inexistantes pour l'Arabe. Nous avons aussi construit un système de reconnaissance de mots arabes isolés, basé sur l'approche HMMs [Bahi, 00b][Bahi, 01a], ce système s'ajoute aux peu de systèmes de reconnaissance de l'Arabe, et nous a ouvert la voie pour envisager des applications plus concrètes pour l'Arabe [Bahi, 05b].

Nous avons également essayé d'appréhender la reconnaissance de la parole arabe à un niveau plus haut, il s'agit de la modélisation du langage où très peu d'études existent dans ce sens[Bahi, 04b].

D'autre part, les points évoqués dans notre problématique de thèse, nous ont conduit à définir un modèle connexionniste qui renferme une sémantique du domaine d'application, en l'occurrence la reconnaissance de la parole.

Un premier objectif de ce travail est donc d'intégrer les symboles dans une architecture connexionniste pour la reconnaissance de la parole. La langue arabe étant le cadre de notre travail, nous nous en référons pour choisir les entités symboliques du réseau incluant les caractéristiques acoustiques, phonétiques, lexicales, …

Nous soulignons que lorsque nous utilisons les informations du domaine, le réseau est construit avec beaucoup de facilité et qu'il apprend plus rapidement et généralise mieux que dans le cas classique [Bahi, 03a][Bahi, 04a][Bahi, 04c][Bahi, 05a].

Un autre aspect immerge dès que l'on considère l'approche basée connaissance, il est lié à la phase de reconnaissance. En effet, en phase de reconnaissance, si un mot est reconnu, nous pouvons suivre le raisonnement mené par le réseau en terme de propagation de l'activation à partir de l'entrée.

Il existe dans la littérature des propositions de systèmes connexionnistes symboliques, la plus ancienne remonte à 1988. Mais ce que l'on peut relever dans ce cadre c'est l'absence d'approche hybride qui intègre le paramètre temporel, or l'aspect dynamique est inhérent à la parole.

L'introduction du paramètre temps dans un modèle neurosymbolique, nous a conduit à mettre au point et à utiliser un automate connexionniste différent du neurone formel classiquement utilisé dans le domaine du neuromémitisme.

Cet automate réagit sous l'influence de quantités d'informations qu'il peut distinguer en fonction de leur provenance. Dans la version actuelle, l'automate supporte deux types d'informations. Les premiers définissent une information spatiale de bas niveau, il s'agit par exemple de coder un indice visuel, acoustique, une position, etc. Les seconds expriment les relations temporelles qui existent entre les évènements et plus particulièrement la notion d'ordre.

Cet automate est ensuite utilisé comme bloc de base dans notre modèle neurosymbolique, que nous avons appelé ***NESSR*** pour Neural Expert System for Speech Recognition [Bahi, 05b].

Globalement, notre contribution consiste en:

- La conception d'un modèle hybride neurosymbolique appliqué à la parole,
- La mise au point d'un modèle de neurone temporel spécialisé que nous avons appelé ***STN*** modèle,
- L'introduction de la composante temps dans un système neurosymbolique.

2. Perspectives du système *NESSR*

Une première perspective de ***NESSR*** consisterait à l'étendre à la reconnaissance de la parole continue. Si nous assumons qu'il y a segmentation implicite du signal à la reconnaissance d'un phonème, nous pouvons généraliser cette stratégie de segmentation à la reconnaissance d'un mot. Dans ce cas les neurones de sorties seront libellés en plus par la classe grammaticale du mot reconnu et il convient de doter la mémoire long-terme de connaissances plus amples sur les règles grammaticales et pragmatiques du langage. Ce qui doterait la mémoire à long terme d'un pouvoir d'anticipation sur le prochain mot.

Une seconde perspective qui nous semble très faisable du fait que l'unité de base est le phonème est l'extension de ***NESSR*** à d'autres langages.

3. Perspectives du modèle *STN*

En partant de l'application, la mise au point du modèle se trouve nécessairement influencée par cette application. Les résultats enregistrés sont alors assez bons mais le modèle a perdu en généricité. Nous avons tenté lors de la mise au point du ***STN*** de se détacher de l'application pour gagner en généricité, mais il reste à le tester avec d'autres applications pour pouvoir l'enrichir, nous pensons en particulier aux contrôles de processus industriels.

4. Perspectives d'amélioration

Au vu de toutes ces remarques comment peut-on envisager l'avenir d'un tel modèle ? tout d'abord précisons une fois encore qu'il ne s'agit en rien d'un modèle universel et encore moins d'un modèle du cerveau. Nous avons tenté de concevoir un système capable d'apprendre, de stocker et de reconnaître des séquences d'informations avec des propriétés de tolérances spacio-temporellele et de généralisation, tout cela dans une approche résolument nouvelle par rapport aux modèles connexionnistes temporels que l'on trouve classiquement dans la littérature. L'aspect temps était le but à court terme de

notre architecture mais il sous-tend un certain nombre de buts plus éloignés qui ont été esquissés plus haut en particulier l'introduction d'une sémantique dans le réseau. Il reste donc un travail conséquent non seulement pour perfectionner encore le système du point de vue temporel mais aussi pour pouvoir traiter des problèmes dépassant la perception.

A court terme, le modèle ***STN*** devrait être testé de façon plus extensive en parole avec d'autres tests que ceux présentés ici (avec d'autres langues) .

D'un point de vue plus théorique sur le modèle, nous devrons poursuivre les investigations dans le domaine neurobiologique afin d'étudier plus finement les mécanismes de traitement temporel modélisables dans un système tel que celui présenté dans cette thèse, à cet effet, des faits psychologiques seront également d'une aide précieuse.

Finalement, il est important de souligner que dans le domaine de la recherche concernant la modélisation du fonctionnement du cerveau humain, ***NESSR*** et le modèle ***STN***, ne sont qu'une contribution qui ne demande qu'à être poursuivie et enrichie.

Références bibliographiques

Références bibliographiques

Ouvrages

[Becchiti, 99] Becchitti C., Ricotti L. P., *speech recognition*, John Wiley, Angleterre, 1999.

[Belaid, 92] Belaid A., Belaid Y., *Reconnaissance des formes: MØthodes et applications*, InterEditions, 1992.

[Bishop, 95] Bishop C. M., *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1995.

[Boite, 87] Boite R., Kunt M., *Traitement de la parole*, Presse polytechniques romandes,1987

[Bourret, 91] Bourret P., Reggia J., Samuelides M., *RØseaux neuronaux, une approche connexionniste de l intelligence artificielle,* Teknea, 1991.

[Burnod 88] Burnod Y., *An adaptive neural network : The cerebral cortex,* Ed. Masson, Paris, 1988.

[Calliope, 89] Calliope, *La parole et son traitement automatique*, Masson, 1989.

[Davalo, 89] Davalo E., Naim P., *Des rØseaux de neurones*, Ed. Eyrolles, 1989.

[El-Ani, 87] Al-Ani S., *Abstract and concrete interaction in the arabic sound system*, Islamic and middle eastern societies, ed. Robert oslan and Salman Al-Ani, Amana Books 1987.

[Harkat, 93] Harkat M., *Les sons et la phonologie*, Ed. Dar el afaq, Alger, 1993.

[Haton, 91] Haton J. P., *Reconnaissance automatique de la parole*, Ed. Dunod, Paris, 1991.

[Hebb, 49] Hebb D. O., T*he organization of behaviour,* Wiley, New-york, 1949.

[Hecht-Nielson, 89] Hecht-Nielson R., *Neurocomputing*, Addisson-Wesley Publishing Company, 1989.

[Hérault, 94] Hérault J., Jutten C., *RØseaux neuronaux et traitement du signal,* Ed. Hermes, Paris, 1994.

[Jacobson, 60] Jacobson R., Linguistics and poetics, Style in language, Seboek, Thomas Ed, MIT Press, pp. 350-377, 1960.

[Jodouin, 94] Jodouin J. F., *Les rØseaux neuromØmitiques : modŁles et applications* , Ed. Hermès, Paris, 1994.

[Kerbrat, 80] Kerbrat-Orecchioni C., *L Ønonciation*, Paris, Armand Colin, page 19, 1ère édition, 1980.

[Kohonen, 89] Kohonen T., *Self-organization and associative Memory,* 3ème édition Springer-Verlag, , Berlin, 1989.

[Kolodner, 93] Kolodner J., *Case-based reasoning,* Morgan Kaufmann, Representation and Reasoning Series, San Mateo, 1993.

[Lewis, 81] Lewis H.R., Papadimitriou C.H, *Elements of the theory of computation*, Ed. Prentice-Hall, USA, 1981.

[Lloyd, 88] Lloyd J.W., *Fondements de la programmation logique*, Editions Eyrolles, Paris, 1988.

[Medsker, 95] Medsker L. R., *Hybrid neural network and expert systems*, Kluwer Academic publishers, 1995.

[Minsky, 69] Minsky M., Papert S., *Perceptrons,* MIT Press, Cambridge, 1969.

[Nour-eddine, 92] Nour-Eddine I., *Phonologia,* Editions Dar EL Fikr El-Loubnani, Beirut, Liban, 1992.

[Rabiner, 93] Rabiner L., Hwang B., *Fundamentals of speech recognition*, Prentice Hall, 1993.

[Rumelhart, 86] Rumelhart D.E., McClelland, et le PDP research group, *Parallel distributed processing exploration in the microstructure of cognition,* MA:MIT Press, Cambridge, 1986.

[Sun, 97] Sun, R., Alexandre, F., *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches,* Lawrence Erlbaum Associates, 1997.

[Wermter, 00] Wermter S., Sun R., *Hybrid Neural Systems*, Springer, New York, Jan. 2000.

Thèses et rapports

[Boz, 95] Boz O., *Knowledge integration and rule extraction in neural networks*, Université de Lehigh, 1995.

[Caraty, 99] Caraty M. J. , *La reconnaissance vocale et son mentor*, Rapport de recherche, LIP6 1999/011, Avril 1999.

[Djoudi, 91] Djoudi M., *Contribution à l Øtude et à la reconnaissance de la parole en arabe standard*, Thèse de l'Université de Nancy I, 1991.

[Durand, 95] Durand S, *TOM, une architecture connexioniste de traitement de sØquences. Application à la reconnaissance de la parole*. Thèse de PhD, Université Henri Poincaré, Nancy I, 1995.

[El-ani, 70] El-ani S. H., *Arabic phonology, An acoustical and physiological investigation*, Mouton & Co., The Hague, Paris, 1970.

[Frisch, 96] Fritsch J., *Modular neural networks for speech recognition*, Thèse de Master, Université de Carnegie Mellon, 1996.

[Haton, 99] Haton J. P., *Les modŁles neuronaux et hybrides en reconnaissance automatique de la parole : Øtat des recherches,* Rapport du CRIN/INRIA, 1999.

[Huang, 92] Huang X., Alleva F., Hon H., Hwang M., Rosenfeld R., *The SPHINX II-speech recognition : An overview*, Rapport CMU – CS-92-112, Jan. 1992.

[Jacob, 95] Jacob B., *Un outil informatique de gestion de ModŁles de Markov CachØs : expØrimentations en Reconnaissance Automatique de la Parole*, Thèse de doctorat, IRIT Toulouse, France, 1995.

[Jaczynski, 94] Jaczynski M., *Etude du raisonnement par cas : recherche des cas similaires et utilisation des ensembles flous*, Rapport de stage, Université de Nice Sophia-Antipolis, Sep. 1994.

[Kirshhoff, 02] Kirshhoff K., *Novel speech recognition models for Arabic,* Rapport final du workshop d'été 2002 sur la reconnaissance de la parole arabe à l'université Johns-Hopkins, 2002.

[Kirshning, 95] Kirshning I., *Continuous speech recognition using the time-sliced paradigm*, Thèse de Master, Université de Tokushima, 1995.

[Louie, 99] Louie R, *Hybrid Intelligent Systems Integration into Complex Multi-Source Information systems*, Thèse de Master, MIT, 1999.

[Orsier, 95] Orsier B., *Etude et application de systŁmes hybrides neuro-symboliques*, Thèse de Doctorat de l'université Joseph Fourier, Mars 1995.

[Peeling, 87] Peeling S., Moore R., *Experiments in isolated digit recognition using the multi-layer perceptron*, Technical report 4007, Royal speech and radar establishment, Malvern, Grande Bretagne 1987.

[Savage, 95] Savage-Carmona J., *A hybrid system with symbolic AI and statistical methods for speech recognition*, Dissertation de doctorat, Université de Washington, 1995.

[Tebelski, 95] Tebelski J., *Speech recogniton using neural networks*, Thèse PhD, Université de Carnegie Mellon, Mai 1995.

[Towell, 91] Towell G., *Symbolic knowledge and neural networks : Insertion, Refinement and extraction*, Thèse de doctorat, Université du Wisconsin, Madison, 1991.

[Vaufraydaz, 02] Vaufraydaz D., *ModØisation statistique du langage à partir d Internet pour la reconnaissance automatique de la parole continue*, Thèse de doctorat, Université Joseph Fourier-Grenoble, France, 2002.

[Watros, 88] Watros R., *Speech recognition using connectionist networks*, Thèse PhD, Université de Pennsylvanie, 1988.

[Widrow, 60] Widrow B., Hoff M.E., *adaptive switching circuits,* 1969 IRE WESON Convention Record, pp : 96-104, 1960.

Publications scientifiques

[Aamodt, 94] Aamodt A., Plaza E., *Case-based reasoning : foundational issues, Methodological variations, and system approaches*, AI Communications, Vol. 7, N° 1, pp : 39-59, Mar. 1994.

[Adnan, 99] Adnan B., Zawaydeh, K. *Stress, phonological focus, quantity, and voicing effects on vowel duration in ammani arabic*, ICPhS99, San Francisco, pp : 451454, 1999.

[Ahmed, 89] Ahmed M.S., Hagos E.M., *Implementation of an arabic digit recognition system*, The Arabian Journal for Science and Engineering, Vol.14, N°1, pp : 79-91, Jan. 1989.

[El-anani, 99] Al-Anani M., *Arabic vowel formant frequencies*, ICPhS99, San Francisco, pp : 2117-, 1999.

[Alkhairy, 99] Alkhairy A., *The arabic pharyngeal approximant*, ICPhS99, San Francisco, pp : 1029-1032, 1999.

[Amirouche, 98] Amrouche A., Debyeche M., Adoul A., Amrouch K., Rouvaen J.M., *Reconnaissance des phonŁmes par rØseaux de neurones et normalisation temporelle : Application aux consonnes pharyngales et glottale arabes*, XXIIème Journées d'études sur la parole, Martigny, pp : 397-400, Juin 1998.

[Ans, 91] Ans B., Coiton Y., Gilhodes J. C., Velay J. L., A *neural network model for temporal sequence learning and motor programming,* Neural Networks, 1994.

[Bakis, 76] Bakis R., *Continuous speech word recognition via centisecond acoustic states*, In 91st Meeting of the Acoustical Society of America, Avril 1976.

[Baldwin, 94] Baldwin J.F., Gooch R.M., Martin T.P., *Support logic for feature representation, pattern recognition and machine learning*, IEEE, pp: 421-425, 1994.

[Barkat, 98] Barkat M., *Identification dialectale des parlers arabes et dØtermination expØrimentale d indices discriminants*, XXIIème journées d'études sur la parole, Martigny, Juin 1998.

[Barkat, 99] Barkat M., *Identification of arabic dialects and experimental determination of distinctive cues*, ICPhS99, San Francisco, pp : 901-904, 1999.

[Beaugé, 94] Beaugé L. Durand S., Lallement, *De nouvelles perspectives pour le connexionisme*, Revue VALGO, (94-2), 1994.

[Béland, 01] Béland R., Mimouni Z., *Deep dyslexia in the two languages of Arabic/French bilangual patient*, Cogniton Vol. 82, ED. Elsevier, pp : 77-126, 2001.

[Ben Jemaa, 98] Ben Jemaa M., Alimi A.M., *A comparative study of neural network approaches in recognition of spoken isolated Arabic words*, Proceeding of MCSEAI'98, pp : 93-102, 1998.

[Billi, 82] Billi R., *Vector quantization and Markov Models applied to speech recognition*, IEEE, pp : 574-577, 1982

[Boudraa, 94] Boudraa B., Selouani S.A., Boudraa M., Guerin B., *Matrices phonØtiques et matrices phonologiques arabes*, XXème Journées d'études sur la parole, Trégastel, pp : 345-350, Juin 1994.

[Bourlard, 97] Bourlard H., Morgan N., *Hybrid HMM/ANN systems for speech recognition : Overview and new research directions,* Lecture Notes In Computer Science; Vol. 1387**,** Springer-Verlag, London, UK, pp : 389 – 417, 1997.

[Burgess, 92] Burgess N, *The generalization of a constructive algorithm in pattern classification problems*, International Journal of Neural Systems, Vol. 3. pp : 1-6, 1992.

[Cohen, 92] Cohen M., Rumelhart D., Morgan N., Franco H., Abrash V., Konig Y., *Combining neural networks and hidden markov models for continuous speech recognition,* www.speech.sri.com/people/ victor/papers/darpa.92.cohen.ps.

[Corredor, 98] Corredor-Ardoy C., Boula de Mareüil P., Adda-Deker M., Lamel L., Gauvain J.L., *Classement des phonŁmes dans un cadre multilingue*, XXIIème Journées d'études sur la parole, Martigny, pp : 75-78, Juin 1998.

[Dehaene, 87] Dehaene S., Changeux J. P., Nadal J. P., *Neural networks that learn temporal sequences by selection,* Proc. Natl. Acad. Sci. USA, Biophysics, N°84, pp : 2727-2731, Mai 1987.

[El-Imam, 89] El-Imam Y. A., *Unrestricted vocabulary arabic speech synthesis system*, IEEE Trans. On ASSP, Vol. 37, N°.12, pp : 1829-1845, Oct. 1989.

[Elman, 90] Elman J. L., *Finding structure in time*, Cognitive science N°14, pp : 179-211, 1990.

[Forney, 73] Forney G. D., *The viterbi algorithm*, Proceeding of the IEEE, Vol. 61, N°3, Mar. 1973.

[Gallant, 88] Gallant S. I., *Connectionnist Expert Systems*, Communications of the ACM, Vol. 31, N°2, pp : 152-169, Fev. 1988.

[Hilario, 96] Hilario M., *An overview of strategies for neuro-symbolic integration*, dans Connectionist –symbolic integration : from unified to hybrid approaches. Ed. Ron Sun, Chapitre 2, Kluwer Academic Publishers, 1996.

[Hopfield, 82] Hopfield J. J., *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the National Academy of Science, USA, pp : 2554-2558, 1982.

[Huang, 88] Huang W.M., Lippmann R., *Neural nets and traditional classifiers,* Neural information processing systems, Ed. Anderson D., pp : 387-396, New-York, 1988.

[Jelinek, 76] Jelinek F., *Continuous speech recognition by statistical methods*, Proceeding of the IEEE, Vol. 64, N. 4, Avr. 1976.

[Jordan, 86] Jordan M. I., *Attractor dynamics and parallelism in a connectionist sequential machine,* Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Erlbaum, 1986.

[Juang , 85] Juang B. H., Rabiner L. R., *Mixture autoregressive Hidden Markov Models for speech recognition*, IEEE transactions on ASSP, Vol. 33, N°6, pp : 1404-1413, Dec. 1985.

[Juang, 86] Juang B. H., Levinson S. E., Sondhi M., *Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains*, IEEE 1986, pp:307-309, 1986.

[Kohler, 01] Köhler J., *Multilingual phone models for vocabulary-independent speech recognition tasks*, Speech communication 35, Ed. Elsevier, pp : 21-30, 2001.

[Kohonen, 84] Kohonen T., Makisara K., Saramaki T., *Phonotopic maps-insightful representation of phonological features for speech recognition*, IEEE Proceedings of the 7th International Conference on Pattern Recognition, 1984.

[Lallement, 95] Lallement Y., Hilario M., Alexandre F., *Neurosymbolic Integration: Cognitive Grounds and Computational Strategies*, In M. DeGlas and Z. Pawlak, editors, World Conference on the Fundamentals of Artificial Intelligence, Paris, France, 1995.

[Lee, 90] Lee K.F., Hon HW, Reddy R. *An Overview of the SPHINX Speech Recognition System,* IEEE Trans on ASSP, 38 N° 1, pp. 35-45, Janvier 90.

[Levinson, 85] Levinson S. E., *Structural Methods in Automatic Speech recognition*, Proceeding of the IEEE, Vol. 23, N°11, Nov. 1985.

[Linde, 80] Linde Y., Buzo A., Gray R. M., *An algorithm for vector quantizer design*, IEEE Transactions on Computer, N°36, pp : 84-95, 1980.

[McCulloch, 43] McCulloch W. S, Pitts W., *A logical calculus of ideas immanent in nervous activity,* Bulletin of mathematical biophysics, Vol.5, 1943.

[Morgan, 95] Morgan N. , Bourlard H. A., *Neural networks for statistical recognition of continuous speech* , Proceeding of the IEEE, Vol. 83, N°5, may 1995.

[Mounir, 94] Mounir J., *L opposition de durØe vocalique en arabe : essai de typologie*, XXème Journées d'études sur la parole, Trégastel, pp : 395-400, Juin 1994.

[Mrayati, 84] Mrayati M., Makhoul J., *Man-machine communication and the arabbic language*, Lecture notes, Applied Arabic linguistics and signal and information processing, pp : 133-145, 1984.

[O'Neil, 97] O'Neil E.N., Jones G.W., Nye C., *Acoustic carachteristics of children who speak Arabic*, International journal of PediatricRhinoLaryngology, Ed. Elsevier, Vol. 42, pp : 117-124, 1997.

[Philips, 00] Philips C., Pellathy T., *Phonological feature representation in auditory cortex*, www.ling.udel.edu/colin/research/papers/feature_mmf.pdf , 2000.

[Poritz, 86] Poritz A. B., Richter A., *On Hidden Markov Models in Isolated word Recognition*, ICASSP 1986, Tokyo, pp : 705-708, 1986.

[Port, 80] Port R. F., Al-Ani S., Maeda S., *Temporal compensation and universal phonetics*, Phonetica 37, Ed. K. Kohler, Kiel, Swizerland, pp : 235-252, 1980.

[Prager, 86] Prager R., Harrison T., Fallside F., B*oltzmann Machines for speech recognition,* Computer speech and language, N°1, pp : 2-27, 1986.

[Rabiner, 81] Rabiner L. R., Levinson S. E., *Isolated and connected word recognition-Theory and selected applications*, IEEE Trans. On communications, Vol. Com. -29, N°.5, pp : 621-659, Mai 1981.

[Rabiner, 83] Rabiner L. R., *On the application of vector quantization and Hidden Markov Models to speaker-independent isolated word recognition*, The Bell system technical Journal, Vol. 62, N°4, USA, , Avril 1983.

[Rabiner, 86] Rabiner L. R. , Juang B. H., *An introduction to Hidden Markov Models*, IEEE ASSP Magazine, pp : 4-16, Jan. 1986.

[Rabiner, 89] Rabiner, *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proceedings of IEEE, Vol. 77, N°2, pp : 257-286, Fev. 1989.

[Reddy, 67] Reddy D. R., *Computer recognition of connected speech*, J. Acoust. Soc. Amer. Vol. 42, pp : 329-347, 1967.

[Rosenblatt, 58] Rosenblatt F., *The perceptron : a probabilistic model for information storage and organization in the brain,* Psychological review, N°65, pp : 386-408, 1958.

[Sara, 99] Sara S.I., *?Al-Hamzah the glotal stop in classical arabic* , ICPhS99, San Francisco, pp : 1317-1320, 1999.

[Selouani, 98] Selouani S. A., Caelen J., *Identification des traits phonØtiques arabes par des systŁmes connexionistes modulaires*, XXIIème Journées d'études sur la parole, Martigny, pp : 418-420, Juin 1998.

[Servan-Shreiber, 91] Servan-Shreiber D., Cleeremans C. R., McClelland J., *Graded state machines : the representation of temporal contingencies in simple recurrent networks*, Machine learning, N°7, pp:161-193,1991.

[Sima, 95] Sima J., *Neural expert system*, Journal of neural networks, Vol. 8, Number 2, pp : 261-271, 1995.

[Sima, 00] Sima J., *Review of integration strategies in neural hybrid systems*, citesser.nec.nj.com/

[Towell, 94] Towell G. G., Shavlik J. W., *Knowledge-based Artificial Neural Networks*, Artificial Intelligence 70, pp : 119-165, 1994.

[Trentin, 01] Trentin E., Gori E., *A survey of hybrid ANN/HMM models for automatic speech recognition*, Neurocomputing 37, Ed. Elsevier, pp : 91-126, 2001.

[Waibel, 89] Waibel A., Hanazawa T., Hinton G., Shiokano K., Lang K.J., *Phoneme recognition using time-delay neural networks,* IEEE Trans. On acoustics, speech, and signal processing, 37(3), mars 1989.

[Young, 96] Young S., *Large vocabulary continuous speech recognition*, IEEE Signal Processing Magazine 13(5), pp :: 45-57, 1996.

Sites web

[Web 01] http://www2.unil.ch/ling/phon/api1.html

[Web 02] Qalam http://eserver.org/laugs/qalam.txt

[Web 03] International Phonetic Alphabet, http://www.arts.gla.ac.uk/IPA/

[Web 04] http://www.cis.hut.fi/research/som-research/

[Web 05] http://www.image.ece.ntua.gr/physta/reports/hybridreview.htm

Publications personnelles

[Bahi, 98] Bahi H. , Sari T., Sellami M., *Reconnaissance des syllabes arabes basØe sur une analyse acoustique*, Journées d'études sur la communication homme/machine, pp :25-30, Alger, Dec. 1998.

[Bahi, 99] Bahi H., Sellami M., *Arabic syllables recognition using an acoustical approach*, ICCTA, Alexandrie, Egypte, 1999.

[Bahi, 00a] Bahi H., Benouareth A., Sellami M., *Un systŁme basØ MMC pour la reconnaissance de la parole Arabe*, Actes SNIB'2000, Biskra, Algérie, pp : 172-182, Mai 2000.

[Bahi, 00b] Bahi H., Sellami M., *Application des MMC pour la reconnaissance de la parole Arabe*, Proceeding of the MCSEAI'2000, Fès, Maroc, pp : 379-388, Nov. 2000.

[Bahi, 01a] Bahi H., Sellami M., *Combination of vector quantization and hidden Markov models for Arabic speech recognition*, Proceedings ACS/IEEE International conference on computer systems and applications, Beirut, Liban, pp : 96-100, Juin 2001.

[Bahi, 01b] Bahi H., Sellami M., *An acoustical based approach for arabic syllables recognition, workshop on software for the arabic language*, AICCSA'2001, Beirut, Liban, Juin 2001.

[Bahi, 01c] Bahi H., Sellami M., *Vers un systŁme de rØØducation vocale*, TAIMA'01, Hammamet, Tunisie, 2001.

[Bahi, 03a] Bahi H., Sellami M., *Hybrid approach for speech recognition*, IAPR Proceedings de la conférence ICISP'03, Agadir, Maroc, pp : 362-367, Juin 2003.

[Bahi, 03b] Bahi H., M.Sellami, *Hybrid approach for Arabic speech recognition*, proceedings ACS/IEEE de la conférence AICCSA'03-Tunis,Tunisie, Juil. 2003.

[Bahi, 04a] Bahi H., Sellami M., *SystŁme expert connexioniste pour la reconnaissance de la parole*, proceedings de RFIA, Vol 2, pp : 659-665. Toulouse, France, Jan. 2004.

[Bahi, 04b] Bahi H., Sellami M., *Approche probabiliste pour la reconnaissance de phrases parlØes de l Arabe*, IEEE Proceedings de Setit , Sousse, Tunisie, Mars 2004.

[Bahi, 04c] Bahi H., Sellami M., *Integration of rule-based systems and neural networks into speech recognition system*, WSEAS Trans. On systems, Issue 2, Volume 3, pp :778-783, Avr. 2004.

[Bahi, 04d] Bahi H., Amiar L., Sellami M., *Approche probabiliste pour la reconnaissance de phrases parlØes de l Arabe*, Génie logiciel et intelligence artificielle, MCSEAI'04, pp : 499-510, centre des publications universitaires, Tunisie, Mai 2004.

[Bahi, 05a] Bahi H., Sellami M., *Connexionnist expert system for speech recognition*, The International Arab Journal of Information Technology, Vol. 2, No. 2, pp:149-154, April 2005.

[Bahi, 05b] Bahi H., Sellami M., An ASR based tool to detect dyslexia, Proceedings de ISPS International Symposium of Programming Systems, Alger, Algérie, pp : 117-122, Mai 2005.

[Bahi, 05c] Bahi H., Sellami M., *Neural expert model applied to phonemes recognition*, Lecture Notes on Artificial Intelligence 3587, pp : 507-515, Springer Verlag, Berlin, Heidelberg, Juil. 2005.