République Algérienne Démocratique et Populaire.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

Université Saâd Dahlab - Blida
USDB

Faculté des Sciences
Département d'Informatique

Mémoire pour l'obtention
d'un diplôme d'Ingénieur d'Etat en Informatique
Option : Intelligence Artificielle

Sujet :

Reconnaissance Automatique des Consonnes Occlusives Orales de l'Arabe Standard par les Réseaux de Neurones Artificiels (RNA)

Présenté par : H. KECHIH
M^ed A. HADJ HAMDI

Proposé et dirigé par : Dr M. GUERTI

Organisme d'accueil : Ecole Nationale Polytechnique d'Alger-ENP

Soutenue le : 00 Septembre 2005, devant le jury composé de :

Président
Promotrice
Examinatrice
Examinateur

- N° / Promotion 2005-

Remerciements

Nous ne saurions présenter ce mémoire sans avoir à remercier Dieu tout puissant de nous avoir prodigué l'énergie pour l'élaboration de ce modeste travail.

Nous remercions tout particulièrement Mme M.GUERTI Maître de con de conférences à l'ENP- Alger, pour nous avoir offert l'occasion de nous former dans ce domaine et pour ses conseils précieux durant toute la période de la formation. Qu'elle trouve ici notre profonde gratitude.

Nous remercions aussi Mme La présidente ainsi que les membre du jury qui ont bien voulu nous faire l'honneur d'examiner et de juger ce travail.

Nous remercions aussi tout nos enseignants, qui nous a enseigner durons tout nos études.

Que Tous ceux qui, d'une façon ou d'une autre, ont contribué de près ou de loin à la réalisation de ce travail, trouvent ici notre sincère reconnaissance.

Dédicaces

Louange à Dieu

Je dédie ce modeste travail à mes chers parents :

A ma mère Ghania qui, durant des lustres, a toujours su et pu être une source intarissable de bonté, et d'affection.

A mon père Younes qui m'a toujours été une référence incontestée de bravoure, d'altruisme et de responsabilité.

A mes frères et sœurs.

A ma promotrice M .GUERTI pour les conseils, le suivi, et les suggestions nettement objectives qu'elle nous a conférées tout au long de la réalisation de ce mémoire.

A mon binôme Amine et à sa famille.

A tous mes amis :

Hakouka, Stam, Boudjemaa, Soudani, Polone, B.Hamza, Djamel, Adlen, Hatem, Mahdi et Nazime, Elmahdi, Alfa, Fouzi, Niri, Bicha, Hadjersi, Seide, Kboucha, mkfouldji, Ayachi, Farid.legraa, Amine, Badjidja, Haba Radjel, Binbo, Rouji, Anter, Elguete, A.Karima, Nabiha, Sara, Houria, Fateh et Rachid, Mostafa (haloi) et à tous mes amis que je n'ai pas cités.

KECHIH.Hamza

Dédicaces

Je dédie ce travail en signe de reconnaissance à mes très chers parents qui ont tout fait pour me donner une bonne éducation et me soutenir dans mes études.

A mes frères et mes sœurs

A ma grande mère

A M.Guerti.

A toute ma grande famille.

A mon binôme HAMZA et sa famille.

A tous mes amis.

HADJ HAMDI Mohamed Amine

Résumé :

Notre travail porte sur la Reconnaissance Automatique des Phonèmes Occlusifs Orales de l'Arabe Standard en mode monolocuteur en utilisant les Réseaux de Neurones Artificiels (RNA). Les paramètres d'entées de RNA sont obtenus après l'élaboration d'un corpus constitué de mots isolés en Arabe Standard et une étape de prétraitement basée sur une analyse sonagraphique afin d'extraire les formants qui constituent les résonances du conduit vocal.

Des algorithmes d'apprentissage supervisé par la méthode de retro-propagation sont implémentés pour avoir un tau de reconnaissance très élevé .

Mots Clés : RAP, Consonnes Occlusives Orales, Arabe Standard, Analyse Sonagraphique, Formants, RNA, PMC, retropropagation.

Abstract :

Our research is the Automatic Recognition of the oral occlusives phonemes of the standard Arabic in mono speaker using Artificial Neuron Networks (ANN). The Parameters of the ANN are the results of the development of a corpus constituted by isolated words in Standard Arabic; this is a stage of post treatment based on the sonagraphic analysis in order to extract the formants that constitute resonances of the vocal tract.

The algorithms of training supervised by the method of old-fashioned propagation are implemented in order to assess the recognition rate.

Key words: AR, Oral Occlusive Phoneme, the standard Arabic, sonagraphic analysis, Forming, ANN, MLP, retropropagation.

ملخص:

يدور عملنا حول التعرف الآلي على الفونيمات الانفجارية الشفوية للغة العربية الفصحة و ذلك باستعمال شبكة العصبونات الاصطناعية في النظام الأحادي النطق.

مقاييس دخول الشبكة العصبونية الاصطناعية المتحصل عليها باستعمال مدونة متكونة من كلمات معزولة باللغة العربية تمت في مرحلة ما قبل المعالجة التي تعتمد على التحليل المطيافي لاستخراج دوال الصوتية الناتجة عن الصدى في اجهاز الصوتية.

الخوارزميات التعلم قد تم تجسيدها لتقدير نسبة التعرف على الفونيمات.

كلمات المفاتيح : التعرف الآلي ، الشبكة العصبونية الاصطناعية، اللغة العربية الفصحة، دوال صوتية، حروف انفجارية.

Chapitre 2 : Principales Techniques et Modèle connexionniste

Chapitre 3 : Application de Réseaux de Neurones à la reconnaissance des Occlusives Orales de l'AS

Chapitre 4 : Système de RACOOAS

Liste des Figures

Liste des Tableaux

e des Abréviat

TAP : Traitement Automatique de la Parole

RAP : Reconnaissance Automatique de la Parole

DAP : Décodage Acoustique et Phonétique ou Décodage Acoustico-honétique

AS : Arabe Standard

TTS : Text-To-Speech

IA : Intelligence Artificielle

LPC : Linear Predictive Coding

PLP : Perceptual Linear Predictive

TFD : Transformée de Fourier Discrète

TFR : Transformée de Fourier Rapide

TOP : Transcription Orthographique Phonétique

DTW : Dynamic Time Warping

HMM : Hidden Markov Models

RN : Réseaux de Neurones

RNA : Réseaux de Neurones artificiels

ADALINE : ADAptiv LINeair Element

MLP : Multi Layer Perceptron

PMC : Perceptrons Multi-Couches

RNM : Réseau de Neurone Modulaire

RE : Réseau Expert

TR : Taux de Reconnaissance

NER : Nombre d'Entités Reconnues

NTE : Nombre Total d'Entités

RPOO : Reconnaissance des Phonèmes Occlusives Orales

Introduction Générale

La parole est l'un des moyens les plus naturels par lequel des personnes communiquent. Cependant, à ce jour, la commande de machines est en général effectuée par des gestes ou par le programme de langages artificiels, tels les langages de programmation ou les commandes des automatismes. Pour des raisons de facilité d'interaction, l'Homme a depuis longtemps été tenté de concevoir des machines dont les commandes seraient directement activées par la parole. Dans cette perspective, il s'agit de développer des systèmes capables de reconnaître la parole, de la comprendre et d'exécuter les actions résultant de la compréhension du message.

La Reconnaissance Automatique de la Parole (RAP) est l'un des domaines du Traitement Automatique de la Parole (TAP), les autres étant la synthèse vocale et le codage de la parole. La RAP permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots, mais aussi de phonèmes (unités sonores minimales). En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis, comme un humain le ferait. Ces deux domaines et notamment la reconnaissance vocale, font appel à des bases de connaissances de plusieurs sciences : l'anatomie (les fonctions des appareils phonatoire et auditif de l'être humain), la phonétique, le traitement du signal, la linguistique, l'informatique, l'Intelligence Artificielle, les statistiques, etc.

Ce sujet, qui a suscité un grand intérêt dans chez chercheurs du domaine, commence à avoir un impact dans la vie courante. Aujourd'hui les progrès réalisés dans le domaine de la RAP nous permettent de reconnaître la parole, de réaliser des systèmes de reconnaissance vocale et d'élaborer quelques applications interactives guidées très simples à vocabulaire limité. Cependant, nous sommes encore loin de réaliser des systèmes de dialogue Homme-Machine très performants.

La réalisation de notre application, consiste à Reconnaître les Phonèmes Occlusifs Oraux de l'Arabe Standard (AS), plusieurs étapes sont à effectuer avant d'atteindre l'étape de reconnaissance.

Après avoir enregistré un corpus de mots en Arabe Standard, comportant les Consonnes Occlusives Orales à étudier dans les différentes positions (initiale, médiane

et finale). Nous avons extrait les paramètres à utiliser comme variables d'entrées pour les Réseaux de Neurones Artificiels (RNA), qui constituent une technique utilisée dans les systèmes de RAP. Ils sont basés sur une modélisation grossière du neurone biologique. Les RNA sont des modèles, à ce titre, ils peuvent être décrits par leurs composants, leurs variables descriptives et les interactions des composants.

Notre projet de Fin d'Etudes est structuré en quatre chapitres :

- dans le premier chapitre, nous définissons les Notions fondamentales sur l'AS et la RAP, tout en commençant par le mécanisme phonatoire et auditif de l'être humain, après avoir présenté les classes des sons de l'AS. Nous présentons ensuite le système de la RAP en expliquant brièvement ses techniques ;
- le deuxième chapitre concerne une présentation des principales techniques d'analyse de la RAP, suivi d'une description de RNA ;
- le troisième chapitre concerne la mise en œuvre de notre travail, pour cela nous présentons en détail notre système de reconnaissance basé sur les réseaux Modulaires ;
- nous présentons dans le dernier chapitre les différentes expériences d'apprentissage et les résultats obtenus de la reconnaissance phonémique ainsi que la description détaillée de notre logiciel d'application.

Nous terminons ce PFE par des conclusions générales et des perspectives en interprétant les différents résultats obtenus.

Chapitre 1
Notions fondamentales sur l'AS et la RAP

1.1. Introduction

La parole est le support le plus naturel de la communication humaine, elle se distingue des autres sons par des caractéristiques acoustiques bien déterminées.

Dans ce chapitre, nous allons présenter sommairement le mécanisme de la phonation et de la perception auditive de la parole tout en exposant brièvement l'aspect phonétique et phonologique des sons du langage, avant de présenter le système phonétique de l'Arabe Standard (**AS**). Ensuite, nous donnons quelques explications concernant le Traitement **A**utomatique de la **P**arole (**TAP**), ainsi que le système de la **R**econnaissance **A**utomatique de la **P**arole (**RAP**) en expliquant et ses techniques. Nous donnons plus de détails sur le **D**écodage **A**coustique et **P**honétique (**DAP**) de l'AS.

1.2. Description de l'appareil phonatoire

Pour aborder les techniques de TAP, il faut commencer par connaître la production de la parole humaine.

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques [2].

L'ensemble du système vocal se compose des poumons, du larynx et du conduit vocal dans lequel on trouve le pharynx, la cavité buccale, la cavité labiale et en parallèle à toutes ces cavités, la cavité nasale (Figure 1.1).

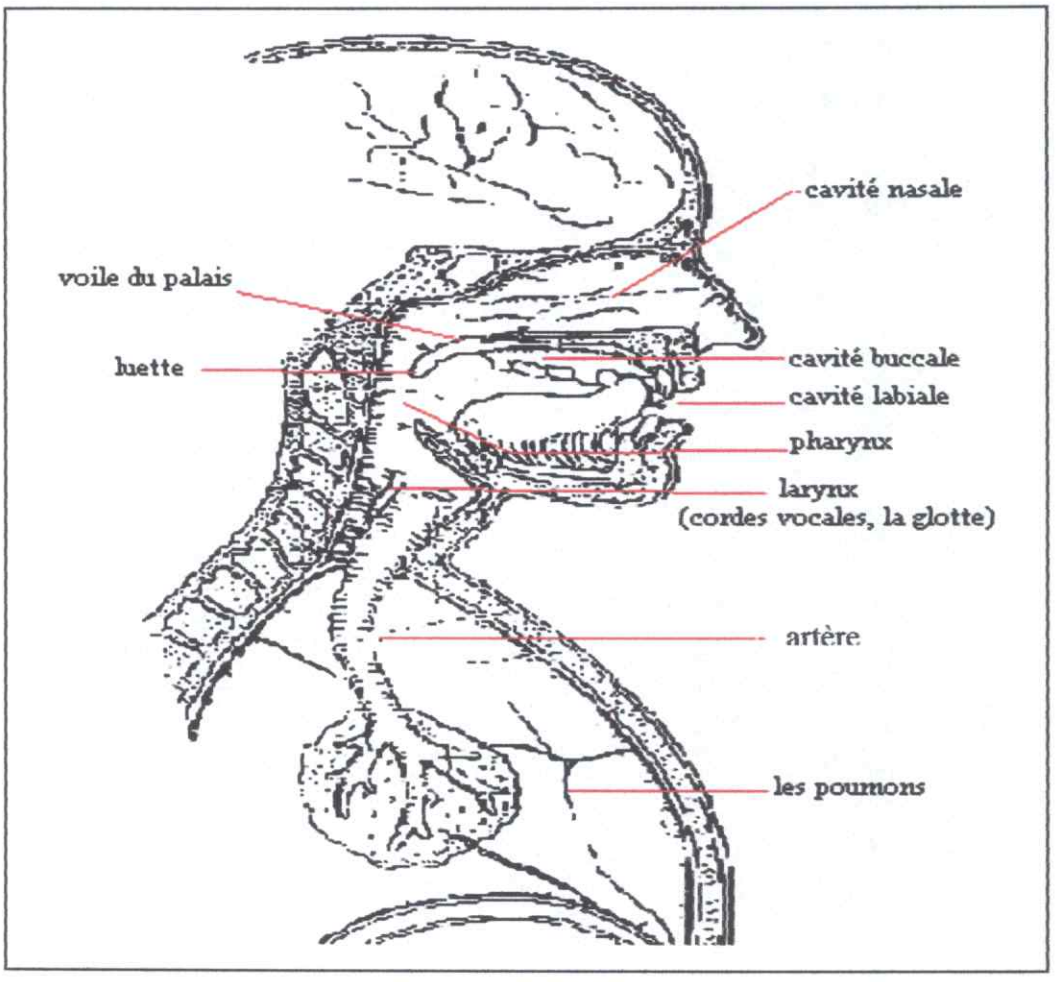


Figure 1.1 : Appareil phonatoire humain.

1.3. Production des sons

Le processus de production de la parole est un mécanisme très complexe. Il repose sur une interaction entre le système neurologique et physiologique. Des organes et des muscles entrent dans la production de sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain est basé sur l'interaction entre trois grandes classes d'organes: les poumons, le larynx, et les cavités supra-glottiques.

Les deux premières classes fournissent ce qui est essentiel pour la production de n'importe quel son, qu'il soit musical ou langagier : une source d'air et une source de bruit. La troisième classe renferme les organes qui permettent de modifier le son qui est émis par le travail conjoint des deux premières classes.

- les poumons : la fonction primordiale des poumons est évidemment de permettre au corps de s'oxygéner. Cependant, ils fournissent aussi une

source d'air qui est utilisée pour produire des sons. Lors de la phase d'inspiration, l'action conjointe du diaphragme, qui se contracte et s'abaisse, et des muscles intercostaux permet de créer un vide dans les poumons qui est rempli par la pénétration d'air. Lors de l'expiration, le diaphragme se relâche et laisse ainsi s'échapper l'air des poumons qui peut être utilisé pour produire des sons ;

- le larynx : lorsque l'air est expulsé des poumons, il passe à travers un tube formé de plusieurs cartilages appelé le larynx. Ce dernier contient des muscles et des cartilages. Les cordes vocales sont des membranes qui peuvent s'ouvrir et se refermer très rapidement (jusqu'à 400 fois par seconde chez les enfants, par exemple), produisant ainsi des variations de pressions dans l'air. Ces dernières sont perçues comme du son par l'oreille humaine (Figure 1.2) ;

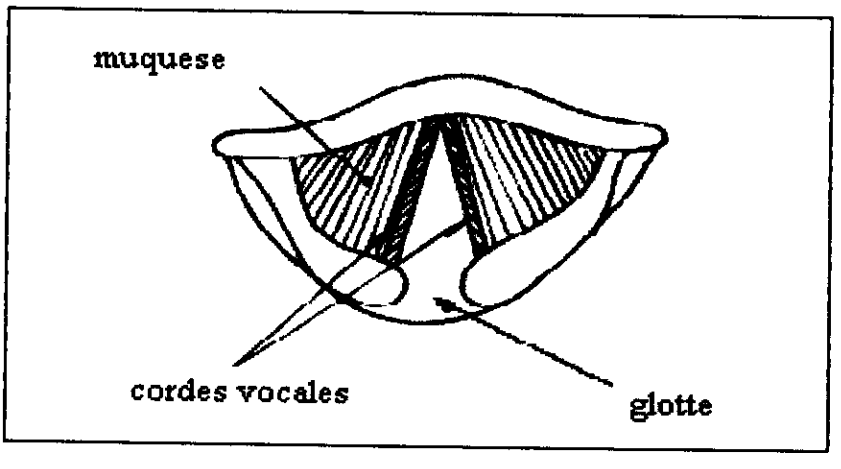


Figure 1.2 : Section du larynx, vue de haut [2].

- les cavités supra-glottiques : Lorsque le son sort de la glotte, il passe à travers les organes vocaux supérieurs appelés cavités supra-glottiques où il est modifié. Ces cavités servent à faire résonner le son et lui donner une couleur particulière qui permet de différencier les voyelles entre elles, par exemple, ou les consonnes. Cette couleur particulière donnée à chaque son provient essentiellement de la modification de la forme des résonateurs à l'aide des mouvements de la langue, des lèvres et de la mâchoire inférieure, etc.

Suivons le cheminement de l'air à travers les organes phonatoires. L'air émis par les poumons, après avoir traversé la glotte, traverse les différentes cavités :

- la pharyngo-buccale : elle est utilisée pour la production de certaines consonnes. C'est la cavité la plus importante dans le langage humain. L'utilisation de cette dernière donne lieu à des articulations orales. La forme de cette cavité peut ensuite être modifiée ;
- la nasale : lorsque la luette, reliée au palais mou, est décollée de la paroi pharyngée, le son peut passer également dans la cavité nasale, créant ainsi une articulation nasale ;
- la labiale : finalement, lorsque certaines articulations sont produites en utilisant les lèvres, on parle de sons labiaux.

1.4. Système de perception auditive

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu.

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'oreille interne qui envoie au cerveau l'influx nerveux, le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.

L'appareil auditif comprend l'oreille externe, l'oreille moyenne, et l'oreille interne. Le conduit auditif externe relie le pavillon au tympan : c'est un tube acoustique de section uniforme fermé à une extrémité. Son premier mode de résonance est situé vers 3 KHz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'oreille moyenne (marteau, étrier, enclume) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la cochlée. Celle-ci contient la membrane basilaire qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée, elle est le support de l'organe de Corti qui est constitué par plusieurs cellules *ciliées* raccordées au nerf auditif (Figure. 1.3).

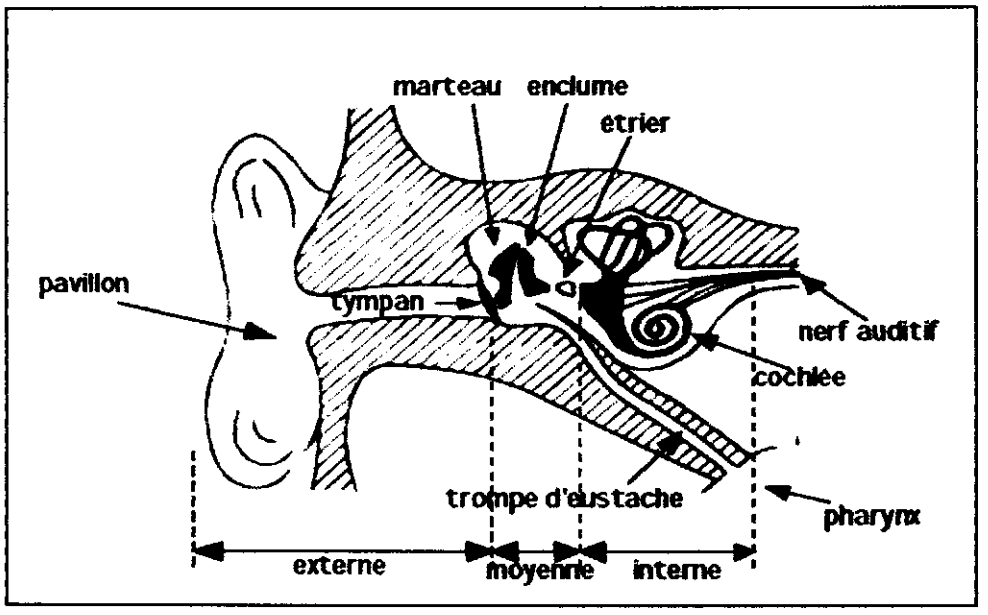


Figure 1.3 : Système auditive humain [2].

1.5. Classification des sons de l'AS

La recherche en Traitement Automatique de la Parole et notamment en Reconnaissance, dans une langue donnée doit nécessairement passer par l'étude de sa composante phonétique. Cette étude nous permet de dégager les principales caractéristiques relatives aux différents phonèmes et ainsi de cerner l'ensemble des paramètres acoustiques, en vue de les exploiter dans l'élaboration d'un système de RAP.

Le système phonétique de l'Arabe Standard (**AS**) comprend six voyelles et vingt-neuf consonnes ou [huru:t] (en incluant la hamza), avec leur Transcription Orthographique et Phonétique (Tab 1.1). La distinction entre voyelles et consonnes s'effectue de la manière suivante, si le passage de l'air se fait :

- librement à partir de la glotte, on a affaire à *une voyelle ;*
- à travers d'une glotte obstruée, complètement ou partiellement, en un ou plusieurs endroits ou lieux d'articulation on a affaire à *une consonne.*

Nous pouvons signaler que le passage de consonnes aux voyelles ne se fait pas de manière abrupte, mais sur un continuum. Nous distinguons ainsi des articulations intermédiaires (par exemples : les demi voyelles) ou les spirantes.

Modes d'articulations des consonnes	**Type de phonème**		**Phonèmes Arabes**	**Transcription Arabisante**	**Lieux d'articulation**
Occlusives	**Voisées**		ب د	**[b]** **[d]**	**bilabiale** **alvéodentale**
	Non-Voisées		ق ت ك ء	**[q]** **[t]** **[k]** **[ʔ]**	**uvulaire** **alvéodentale** **postpalatale** **glottale**
	Voisée	**Emphatiques**	ض	**[d]**	**alvéolaire**
	Non-Voisée		ط	**[t]**	**alvéodentale**
Fricatives	Voisées		ز ذ غ ع	[z] [d] [ġ] [ʿ]	sifflante dorsoalvéolaire interdentale uvulaire pharyngale
	Non-Voisées		س ث ف ش خ ه ح	[s] [t] [f] [š] [ḫ] [h] [ḥ]	sifflante dentale interdentale labiodentale chuinchante palatale vélaire glottale pharyngale
	Voisée	Emphatiques	ص	[ṣ]	dorsealvéodentale sifflante
	Non-Voisée		ض	[d]	interdentale
Nasales	Voisées		م ن	[m] [n]	bilabiale alvéodentale
Liquide	Voisée		ل	[l]	dentale
Affriquée	Voisée		ج	[ǧ]	alvéopalatale
Vibrante	Voisée		ر	[r]	apicoalvéolaire
Semi-voyelles	Voisées		و ي	[w] [y]	bilabiale palatale

Tab 1.1: TOP des consonnes de l'Arabe Standard [9].

1.5.1. Description des voyelles

La caractéristique majeure des voyelles est le passage de l'air à partir de la cavité supra-glottique. Le seul traitement que l'air peut, dès lors, subir est la résonance (c'est-à-dire le renforcement de certaines bandes de fréquences).

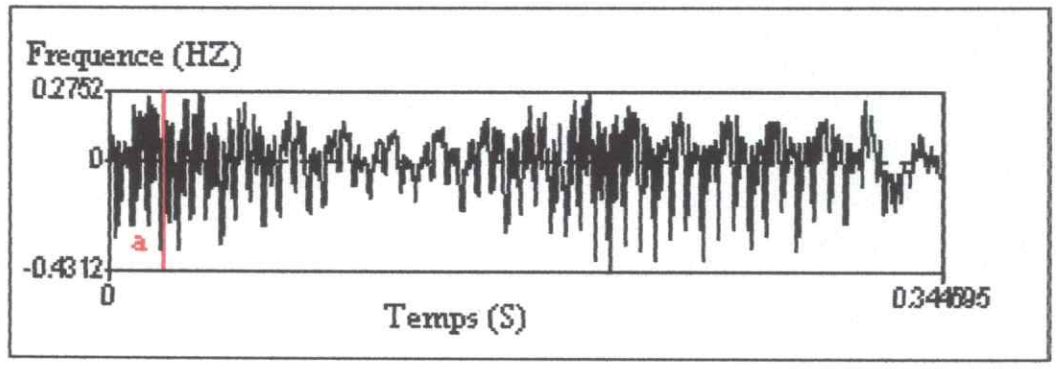


Figure 1.4 : Spectre de la voyelle [a]dans le mot [adam].

Les voyelles de la langue Arabe sont classées, en trois brèves ou courtes [*haraka:t*] et trois longues [*madd*].

1.5.2. Description des semi-voyelles

Les semi-voyelles possédant des structures formantiques similaires aux voyelles écoulement libre de l'air). Ces phonèmes ont tout de même une obstruction qui les approche des consonnes.

1.5.3. Description des consonnes

Les consonnes de l'Arabe Standard ou [*huru:t*] peuvent être classées sur le plan acoustico-physiologique, selon leurs modes de production [*sifa*]. Elles sont classées en sonores/sourdes, occlusives/spirantes, emphatiques/non-emphatiques, etc.

Il existe deux grands types d'articulations consonantiques soit le passage de l'air :

- est fermé et le son résulte de son ouverture subite, on a alors le passage à des consonnes *occlusives ou plosives* ;
- se rétrécit mais n'est pas interrompu, on parle dans ce cas de consonnes *continues*, dont les *constrictives* ou *fricatives* sont les plus représentatives.

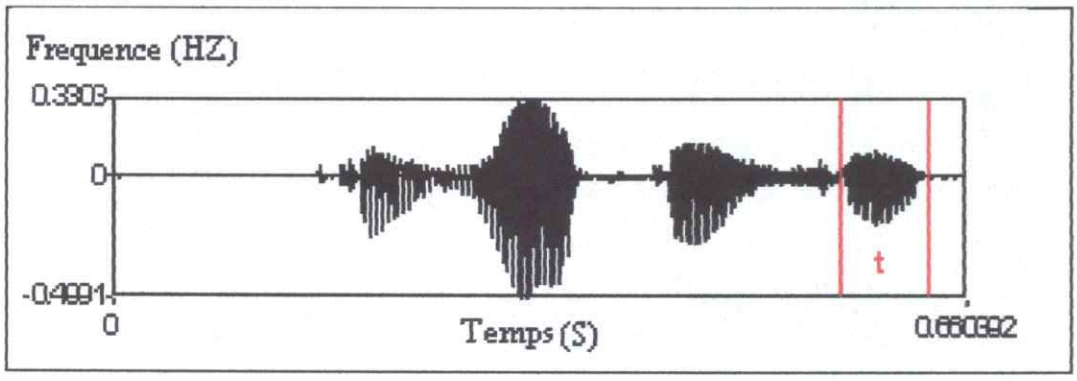


Figure 1.5 : Spectre de la consonne [t] du mot fait.

Une réalisation est dite *sourde* lorsque les cordes vocales ne vibrent pas; si celles-ci entrent en vibrations, la réalisation sera dite *sonore*. Les cordes vocales sont des replis musculaires situés au niveau de la glotte.

Les vibrations des cordes vocales est le résultat d'une obstruction de la glotte : celles-ci vibrent sous la pression de l'air interne qui force un passage entre elles.

1.5.4. Les Occlusives

Les consonnes occlusives sont produites par une fermeture complète du canal respiratoire, et non un simple rétrécissement, ce qui les différencie des autres consonnes.

L'occlusion se fait en deux temps :

- arrêt du flux d'air par la fermeture soudaine du canal expiratoire ;
- libération de l'air interne par le relâchement brusque de l'occlusion.

Il existe deux types d'occlusives : nasales et orales.

1.5.4.1. Nasales

Les occlusives nasales sont réalisées sonores dans la plupart des langues. Pendant la production des occlusives nasales, le voile du palais est plus ou moins abaissé, de manière à laisser passer une partie de l'air expiré à travers des fosses nasales. Alors que l'occlusion a lieu dans la bouche, la résonance nasale est, quant à elle, continue.

1.5.4.2. Orales

Pendant la production des occlusives orales, le voile du palais est relevé, l'accès aux fosses nasales est bloqué, et l'air ne peut traverser que la cavité buccale.

1.5.5. Point et mode d'articulation

La distinction entre le mode d'articulation et le point d'articulation est particulièrement importante pour le classement des consonnes.

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- libre passage, avec mise ou non des vibrations, de l'air au niveau de la glotte (sonore ou sourde) ;
- libre passage, ou non, en un point quelconque (le point d'articulation) des cavités supra-glotiques (voyelles ou consonnes) ;
- passage par une voie unique ou deux voies différentes (orale ou nasale) ;
- passage, dans le conduit buccal, par une voie médiane ou latéral (la plupart des articulations opposées aux latérales).

Le point d'articulation est l'endroit où se trouve, dans la cavité buccal, un obstacle au passage de l'air. De manière générale, nous pouvons dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage de l'air.

Le point d'articulation peut se situer aux endroits suivants :

- les lèvres (articulation *labiales* ou *bilabiales*) ;
- les dents (articulation *dentales*) ;
- les lèvres et les dents (articulations *labio-dentales*) ;
- les alvéoles (c'est-à-dire les gencives internes des incisives supérieures, articulations *alvéolaires*) ;
- le palais (vue sa grande surface, on peut distinguer des articulations *pré-palatales, médio-palatales* et *post-palatales*) ;
- le voile du palais (palais mou, articulations *vélaires*) ;
- la luette (articulations dites *uvulaires*) ;
- le pharynx (articulations *pharyngales*) ;
- la glotte (articulations *glottales*).

1.6. Description acoustique de la parole

L'objet d'étude de la phonétique acoustique est constitué par l'onde sonore telle que produite par les organes de la phonation. Bien que cette onde soit audible, ses propriétés physiques ne sont observables qu'à l'aide d'appareils permettant d'analyser les éléments qui la constituent.

Du point de vue phonétique acoustique, les sons du langage humain sont constitués par des ondes en mouvement. Il s'agit essentiellement d'un mouvement vibratoire régulier ou irrégulier généré par les articulateurs et les cordes vocales, mouvement qui se propage dans l'air ambiant à une vitesse de 340 m/s. La classification des sons du langage que propose la phonétique acoustique est basée sur les propriétés physiques des sons. Ces propriétés ont trait à la nature périodique ou apériodique de l'onde sonore et sont en outre responsables de la sensation de durée, de celle d'intensité et de la hauteur ou fréquence fondamentale (F_0) des sons perçus.

Le signal vocal est caractérisé par des paramètres pertinentes, ces derniers doivent représenter au mieux ce qu'ils sont censés modéliser et doivent extraire le maximum d'informations utiles pour la RAP.

1.6.1. La durée

La durée est le paramètre acoustique le plus délicat à évaluer. La difficulté de mesure réside dans sa grande variabilité qui est due au contrôle quasi impossible du système phonatoire. Chaque phonème se caractérise par ses propres durés intrinsèques et co-intrinsèques de même que le facteur de compressibilité ou expansion.

1.6.2. L'énergie du signal

L'énergie du signal correspond à la puissance du signal. Elle est souvent évaluée sur plusieurs trames successives de signal pour pouvoir mettre en évidence des variations.

1.6.3. La fréquence fondamentale (F_0)

Lors de la production de certains phonèmes (les sons voisés de la parole : les voyelles et certaines consonnes), la fréquence fondamentales (Fo) est la conséquence directe des variations de la pression sub-glottique (la tension des

cordes vocales). Sa corrélation acoustique appelée pitch est généralement linéaire aux basses fréquences d'où la supposition d'une relation linéaire entre elle et la fréquence fondamentale.

1.6.4. Les formants

Lorsqu'un excitateur entre en vibrations du spectre, il fournit un signal, dont le résonateur va amplifier certaines composantes. On obtient alors des formants qui sont un facteur fondamental dans la caractérisation du timbre. Ils servent, justement, à former ce dernier. Le nombre des formants, selon les caractéristiques du résonateur (volume, forme et ouverture), est variable : d'un seul à une infinité (théoriquement). Néanmoins, du point de vue perceptif, seuls quelques-uns d'entre eux jouent un rôle central au niveau de la parole. Par exemple, on peut caractériser toute voyelle en ne prenant en compte que ses trois premiers formants : F_1, F_2, et F_3 (pour une réalisation de la voyelle [i] par exemple, les trois premiers formants pourraient se situer respectivement à F_1 =300 Hz, F_2 =2200 Hz, F_3 =3000 Hz).

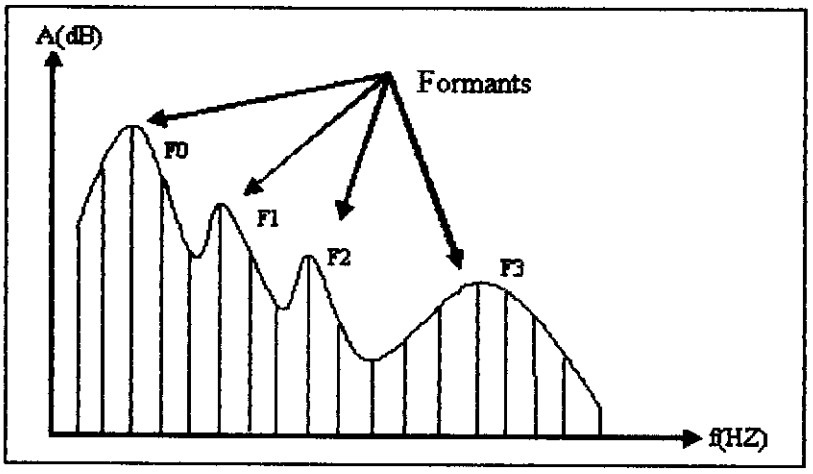


Figure 1.6 : Représentation des formants d'un son voisé [a].

1.7. Traitement automatique de la parole

L'étude de la parole est l'un des domaines dans lequel est utilisé le traitement numérique des signaux. Cette discipline connaît depuis les dernières décennies un développement considérable avec l'apparition de processeurs à hautes performances. La recherche en TAP fait appel à des disciplines de plus en plus diverses et a pour but le développement technologique dans plusieurs domaines, tels que l'analyse, la synthèse, le codage de la parole, la RAP, etc.

1.7.1. L'analyse de la parole

l'analyse de la parole cherche à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (on parle dans ce cas d'analyse perceptive), mais jamais tel qu'il est compris, ce rôle étant réservé à la reconnaissance de la parole. L'analyse est utilisée soit comme composante de la base du système de codage, de reconnaissance ou de synthèse, soit pour des applications spécialisées, comme l'aide au diagnostic médical (pour les pathologies du larynx, par analyse du signal vocal) ou l'étude des langages.

1.7.2. La synthèse de la parole

La synthèse de la parole à partir d'une représentation numérique, inverse de l'analyse, dont la mission est de produire de la parole à partir des caractéristiques numériques d'un signal vocal telles qu'obtenues par analyse :

- La synthèse par règles a principalement la faveur de la phonétique et de la phonologie. Elle permet une approche cognitive, générative du mécanisme de la phonation. Elle est basée sur l'idée que, si un phonéticien expérimenté est capable de «lire» un spectrogramme, par exemple, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel pour une suite de phonèmes donnés. Une fois le spectrogramme obtenu, il ne reste plus alors qu'à générer l'audiogramme correspondant ;
- contrairement à la synthèse par règles, la synthèse par concaténation a une connaissance très limitée du signal qu'elle met en forme. La plupart de ces connaissances se trouvent, en effet, stockée dans les unités de parole mises en oeuvre par le synthétiseur ;
- la synthèse de la parole à partir d'une représentation symbolique, inverse de la reconnaissance de la parole est capable en principe de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, on classe également la synthèse en fonction de leur mode opératoire, la synthèse à partir du texte ou Text-To-Speech (TTS), cette dernière reçoit en entrée un texte orthographique et doit en donner lecture. La synthèse à partir de concepts, appelés à être insérés dans des systèmes de dialogue Homme-Machine,

reçoit le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue.

1.7.3. Le codage de la parole

Le codage permet la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole.

1.8. La Reconnaissance Automatique de la Parole (RAP)

Reconnaître la parole, pour quoi faire ?

La réponse est que pour que l'Homme puisse communiquer avec la Machine. Avec la parole, plus de regard visé à un écran et des mains tapent sur un clavier. Grâce à la RAP, un homme, par exemple, peut se déplacer et se consacrer à sa tâche principale des divers secteurs comme, les applications grand public concernent l'automobile (commande vocale d'équipements annexes tels que la climatisation, l'essuie-glace, lève-vitres), le jouet (téléguidage vocal,...), les jeux électroniques et l'appareillage domestique (commande de téléviseur, lave-linge,...) [4].

La RAP a pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse.

On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal :

- la reconnaissance du locuteur, dont l'objectif est de reconnaître la personne qui parle ;
- la reconnaissance de la parole, où l'on s'attache plutôt à reconnaître ce qui dit. On classe également la reconnaissance en fonction des hypothèses simplificatrices sous lesquelles elles sont appelées à fonctionner. Ainsi en reconnaissance du locuteur, on fait la différence entre l'identification et la vérification du locuteur, selon que le problème est de vérifier que la voix analysée correspond bien à la personne qui est sensée la produire, ou qu'il s'agit de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé.

On sépare la reconnaissance du locuteur dépendante du texte, la reconnaissance avec un texte dicté et la reconnaissance indépendante du texte. Dans le premier cas,

la phrase à prononcer pour être reconnue est fixée dès la conception du système ; elle est fixée lors du test, dans le deuxième cas et elle n'est pas précisée dans le troisième.

On parle de la reconnaissance de parole"monolocuteur", " multilocuteur", ou indépendante du locuteur, selon qu'il a été entraîné à reconnaître la voix d'une personne, d'un groupe fini de personnes, ou qu'il est en principe capable de reconnaître n'importe qui.

On distingue enfin la reconnaissance de mots isolés, la reconnaissance des mots connectés, et la reconnaissance de la parole continue, selon que le locuteur sépare chaque mot par un silence, qu'il prononce de façon continue ; une suite de mots prédéfinis, ou qu'il prononce n'importe quelle suite de mots de façon continue.

1.8.1. Les méthodes de reconnaissance de la parole

Deux approches de la RAP existe, l'une plus globale, l'autre plus analytique permettent d'appréhender la reconnaissance des mots.

Dans l'approche globale, l'unité de base est le plus souvent le mot, considéré comme une entité globale, c'est-à-dire non décomposée. L'idée de cette méthode est de donner au système une image acoustique de chacun des mots qu'il doit identifier par la suite. Cette opération est faite lors de la phase d'apprentissage, où chacun des mots est prononcé une ou plusieurs fois. Cette méthode a pour avantage d'éviter les effets de coarticulation, (c'est-à-dire) l'influence d'un son sur un autre son contigu (voisin). Elle est cependant limitée aux petits vocabulaires prononcés par un nombre restreint de locuteurs.

L'approche analytique, qui tire partie de la structure linguistique des mots, tente de détecter et d'identifier les composantes élémentaires (phonèmes, syllabes, ...). Celles-ci sont les unités de base à reconnaître. Cette approche a un caractère plus général que la précédente : pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base.

Pour la reconnaissance de mots isolés à grand vocabulaire, la méthode globale ne convient plus car la machine nécessiterait une mémoire et une puissance considérable pour respectivement stocker les images acoustiques de tous les mots du vocabulaire et comparer un mot inconnu à l'ensemble des mots du dictionnaire.

De plus, il est impensable de faire dicter à l'utilisateur l'ensemble des mots que l'ordinateur a en mémoire. C'est donc la méthode analytique qui est utilisée : les mots ne sont pas mémorisés dans leur intégralité, mais traités en tant que suite de phonèmes.

1.9. Les problèmes de reconnaissance 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 certaines difficultés du problème, nous ferons ressortir essentiellement quelques caractéristiques notoires de ce signal :

- une grande variabilité ;
- un lieu d'interférences ;
- des ambiguïtés ;
- les effets de la co-articulation.

1.9.1. 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 diduction évolue dans le temps : l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix.

La variabilité inter-locuteur est encore plus flagrante. La hauteur de la voix, l'intonation l'accent diffère selon le sexe, l'origine sociale, régionale ou nationale.

Enfin, toute parole s'inscrit dans un processus de communication où entrent en jeu de nombreux éléments comme le milieu ambiant ou l'environnement, 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

1.9.2. Les interférences

Trois types d'interférences existent :

- l'acoustique du lieu (milieu protégé ou environnement bruyant) ;

- la qualité du microphone et sa position par rapport à la bouche ;
- les bruits venant de la bouche.

1.9.3. La redondance

Le signal vocal est caractérisé par une très grande redondance, condition nécessaire pour résister aux perturbations du milieu ambiant. Pour aborder la notion de redondance, il faut examiner la parole en tant que vecteur d'informations.

1.9.4. Les ambiguïtés

Les mots qui composent la langue Arabe entretiennent entre eux des liens complexes : les sons, les graphies, les classes grammaticales, le sens ou le "référent" ne se discernent pas de façon immédiate. C'est pour cette raison que l'ambiguïté rend la tâche de RAP très difficile.

1.9.5. Les effets de la co-articulation

La production "parfaite" de chaque son suppose théoriquement un positionnement précis des organes phonatoires. Or, lorsque la vitesse de 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, acoustique ou perceptif.

1.10. Quelques applications de la RAP

Après avoir vu le principe de fonctionnement de la reconnaissance de la parole, nous proposons une liste des applications qui paraissent intéressantes dans chaque domaine.

- l'avionique : est un domaine d'application important dans la RAP. Des systèmes de reconnaissance par mots ont été utilisés avec succès dans des avions de chasse pour permettre au pilote déjà trés occupé de commander diverses fonctions (radio, radar,...). La voix a également servi au contrôle d'un bras articulé lors de la mission de la navette spatiale. En effet, au bord d'un avion comme au bord d'une automobile, les tâches étant complexes et le tableau de bord réduit, la parole permet au pilote ou au conducteur d'avoir à

sa disposition un moyen supplémentaire d'interaction avec la machine, sans cependant gêner l'accomplissement des tâches courantes qui requièrent de sa part toute son attention visuelle [7] ;

- les Télécommunications : l'information donnée au public est aussi un domaine concerné par la numérisation de la parole. Dans les gares ou les aéroports, par exemple, on pourra bientôt voir des bornes interactives qui remplacent les agents de renseignements. Pour connaître l'horaire d'un train, il suffit de demander de vive voix à la machine où on veut aller et quand, et elle répond dans la langue de notre choix, avant de nous souhaiter un agréable voyage.

Nous citons plusieurs autres domaines d'applications : industrielles, domestiques, pour les personnes handicapées.

1.11. Décodage Acoustique et Phonétique de l'Arabe Standard

Il s'agit de décomposer les mots en unités symboliques discrètes, qui vont permettre de décrire aussi bien les mots de références contenues dans le dictionnaire que les mots que l'on cherche à reconnaître. Une unité idéale existe et permet de décrire tous les sons caractérisant une langue : c'est le phonème. La plupart des langues comportent moins d'une centaine de phonèmes. Plusieurs voies de recherche sont actuellement empruntées pour atteindre ce but. On distingue globalement trois approches :

- l'Intelligence Artificielle (I.A.) par le biais de Systèmes Experts : on utilise alors une connaissance a priori développée par les spécialistes de la phonétique ;
- les modèles statistiques qui permettent de traiter la grande variabilité du signal vocal par l'analyse préalable d'un grand nombre d'échantillons vocaux. C'est actuellement la plus répandue ;
- les modèles connexionnistes à base de réseaux de neurones [5].

Le processus de DAP, consiste à découper le signal de la parole en segments (phase de segmentation), puis à identifier ces segments en leur affectant une étiquette phonétique (phase d'identification).

La segmentation en unité élémentaire, syllabe, demi-syllabe, phonème, diphone ou triphonème, s'appuie sur la recherche des discontinuités du signal ou de son spectre au cours du temps.

L'identification consiste alors à comparer chaque spectre de ces segments à un ensemble de spectres de référence et à conserver les plus ressemblants. Les techniques de comparaison couramment employées s'appuient sur des méthodes classiques, qui tiennent compte des variations individuelles (accents, coarticulation, liaisons) et prosodiques (rythme, intensité, mélodie).

En effet, le spectre ne dépend que du faible nombre d'éléments à étiqueter, dans notre cas, nous avons les phonèmes occlusifs de l'Arabe Standard, contrairement aux très nombreux mots possibles du dictionnaire.

1.12. Conclusion

Dans ce chapitre, nous avons exposé sommairement les deux aspects : articulatoire (les modes et lieux d'articulations) et acoustique, relatifs aux phonèmes l'Arabe Standard tout en nous attardant sur les phonèmes occlusifs de cette langue. L'intérêt de ce chapitre est d'acquérir des notions fondamentales nécessaires à la reconnaissance automatique de la parole, pour le décodage acoustico-phonétique de ces phonèmes.

Chapitre 2
Principales Techniques d'Analyse vocale et Modèles connexionnistes

2.1. Introduction

Ce chapitre nous permet de présenter une des grandes techniques de reconnaissance des formes qui sont utilisées en Reconnaissance Automatique de la Parole (RAP) : les modèles connexionnistes. En premier lieu, nous allons faire une brève présentation des connaissances de la neurobiologie qui ont servi de base à l'établissement des techniques neuromimétiques après avoir présenté les principales techniques d'analyse vocale. Ensuite nous abordons les différents modèles connexionnistes tels que les perceptrons, les réseaux multicouches et leurs algorithmes d'apprentissage.

2.2. Techniques d'analyse du signal vocal

Pour résoudre les problèmes liés à la complexité de la parole, il est possible de déterminer les paramètres pertinents représentatifs du signal traité. Ces paramètres sont calculés à intervalles temporels réguliers. Le signal de parole est transformé en une série de vecteurs de coefficients. Ces derniers doivent représenter au mieux le signal qu'ils sont censés modéliser, et extraire le maximum d'informations utiles pour la reconnaissance. Pour cela, on a deux techniques d'extraction des paramètres : paramétrique et non paramétrique [10].

2.2.1. La technique paramétrique

Nous n'énumérons pas tous les types de paramètres employés dans le domaine de la recherche en parole car il y en a énormément et ce n'est pas le propos de notre PFE. Pourtant, il est à noter que d'autres approches sont plus proches de l'audition humaine. De plus, le lecteur trouve des informations sur les différents paramètres très largement utilisés dans le codage **LPC** (Linear Predictive Coding), présent pour les **PLP** (Perceptual Linear Predictive) et pour les **RASTA-PLP**, version approfondie des PLP. Cette liste n'est pas exhaustive mais permet d'avoir un aperçu des différents paramètres qu'il est possible d'extraire d'un signal de parole.

2.2.1.1. Le codage LPC

La LPC est une méthode de codage et de représentation de la parole. Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus

linéaire. Il s'agit donc de prédire le signal *s(n)* à un instant *n* à partir des *p* échantillons précédents. La parole n'étant pas un processus parfaitement linéaire, la moyenne que constitue la somme pondérée du signal vocal sur *p* pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme *e(n)*.

Le codage par LPC consiste donc à déterminer les coefficients a_k qui minimisent l'erreur *e(n)*, ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage [11].

$$S(n)=\sum_{k=1}^{p} a_k . S(n-k) + e(n) \qquad (2.1)$$

2.2.1.2. Le coefficients PLP

L'analyse PLP est une méthode inspirée du principe de la prédiction linéaire. Elle combine ce principe à une présentation du signal vocal qui suit l'échelle de l'audition humaine. Le schéma général de cette méthode est présenté dans la figure 2.1.

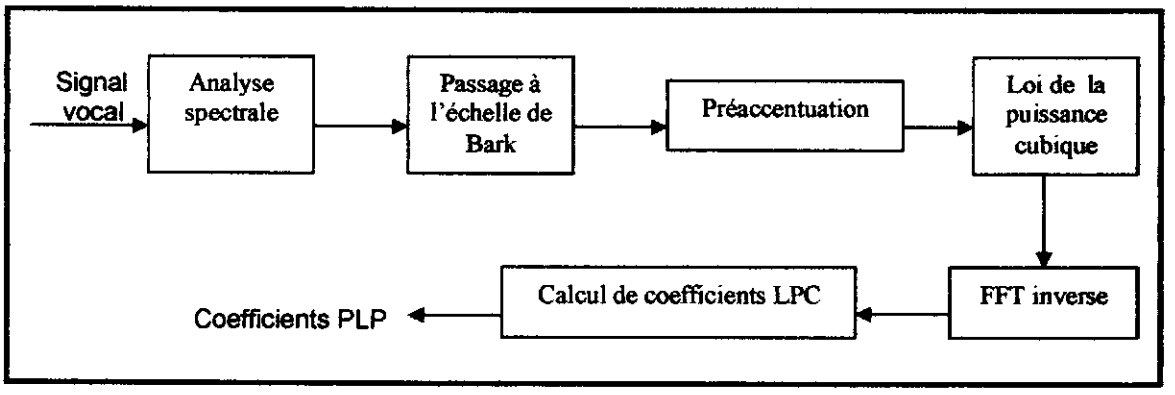


Figure 2.1 : Méthode de calcul des coefficients PLP.

On observe les différents modules de calcul de coefficients PLP, on trouve le spectre de puissance à court terme qui est calculé par (FFT). Ensuite, un passage de l'échelle de fréquence de Bark est effectué. Ce dernier, permet d'approximer la forme des filtres auditifs puis une préaccentuation est effectuée sur le signal résultant. La non-linéarité entre l'intensité d'un signal de parole et sa force de perception par l'oreille est ensuite approximée par une loi cubique. Enfin, on calcule les coefficients du filtre pour obtenir les coefficients PLP.

2.2.2. La technique non paramétrique

Ce type de paramétrisation fait appel aux méthodes classiques utilisées en traitement du signal : la transformée temps-fréquence sans connaissance a priori de sa structure fine. La transformée la plus utilisée en TAP est Transformée de Fourier Discrète (TFD) et en particulier la Transformée de Fourier Rapide (TFR ou FFT pour Fast Fourier Transform).

2.2.3. Le Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme ou sonagramme. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération (Fig. 2.2). Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms) ; ils mettent en évidence l'enveloppe spectrale du signal, et permettent, par conséquent, de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales [2].

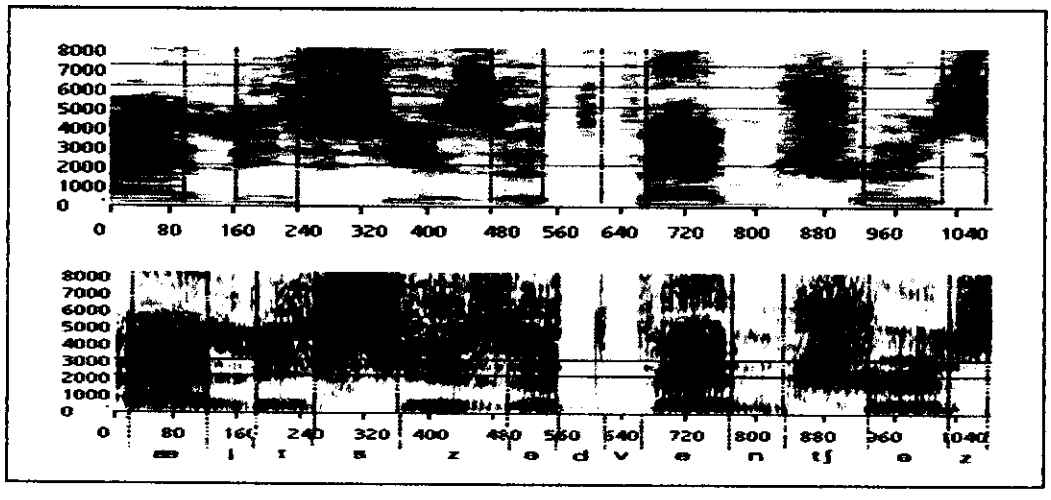


Figure 2.2 : Spectrogrammes à BE (en haut), à BL (en bas), de « Alice's dventures » [2].

2.3. Les principales techniques de la RAP

Il existe plusieurs techniques de la RAP dont les principaux sont les suivants :

- Dynamic Time Warping (DTW) ;
- les Modèles de Markov Cachés ou Hidden Markov Models (HMM) ;
- les Réseaux de Neurones (RN).

2.3.1 Dynamic Time Warping (DTW)

Dans les systèmes de reconnaissance basés sur la Dynamic Time Warping (DTW), chaque mot du lexique est représenté par une réalisation de référence. Le processus de reconnaissance consiste à évaluer la distance d'une observation à chacune des références. Toute la difficulté du décodage réside dans cette mesure d'un degré de similarité entre des formes acoustiques variables à la fois au niveau spectral et temporel. Rapide dans des tâches à petit vocabulaire, cette technique a un certain nombre d'inconvénients importants qui limitent son champ d'application. D'une part, la modélisation des mots par une instance est très peu robuste par rapport à l'ensemble des variabilités acoustiques.

2.3.2. Les Modèles de Markov Cachés (HMM)

Les Modèles de Markov Cachés ou Hidden Markov Models (HMM) est une méthode statistique puissante pour caractériser les échantillons de données.

Il faut associer, dans un premier temps, un phonème a un vecteur acoustique. Or, on n'observe pas directement les phonèmes mais bien les vecteurs. Il y a donc deux processus stochastiques imbriqués : le premier comprenant tous les vecteurs acoustiques possibles et le second tous les phonèmes à associer aux vecteurs. Les Modèles de Markov cachés permettent de calculer la suite d'états, et donc de phonèmes, la plus probable étant donnée, une suite de vecteurs acoustiques, observée. Ces modèles sont dits cachés car on n'observe pas directement les états.

Nous présentons les trois problèmes de base à résoudre pour l'application de cette méthode :

- l'évaluation : quelle est la probabilité d'un modèle générant une séquence d'observation ?

- le décodage : quelle est la séquence d'états la plus probable pour un modèle et une séquence d'observation donnée ?
- l'apprentissage : comment peut-on ajuster les paramètres du modèle pour maximiser la vraisemblance (probabilité jointe) de génération d'une séquence d'observation ?

2.3.3. Les Réseaux de Neurones (RN)

Depuis une vingtaine d'années, les réseaux neuromimétiques constituent une technique utilisée dans les systèmes de RAP. Ils sont basés sur une modélisation grossière du neurone biologique. Tout comme le neurone biologique, le neurone formel calcule sont activation S en fonction des signaux qu'il reçoit d'autres neurones X_i, pondérés par des poids synaptiques W_i et d'une fonction d'activation plus ou moins complexe $F(W_i, X_i)$. L'ensemble de ces neurones est organisé selon des architectures plus ou moins complexes matérialisés par les connexions entre ces neurones. Selon cette architecture, ainsi que le type de la fonction d'activation, les RN peuvent résoudre un certain nombre de problèmes tels que les problèmes de classification, de mémorisation et de résolution de contraintes. Une particularité des RN est qu'ils sont dotés d'algorithmes d'apprentissage qui leur permettent d'apprendre les formes et les classes à reconnaître (Figure 2.3).

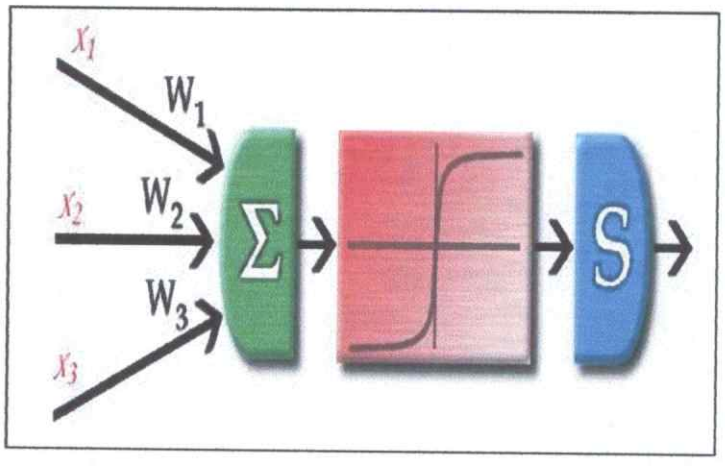


Figure 2.3 : Schéma d'un neurone formel [6].

2.3.4. Neurophysiologie

Le cerveau se compose d'environ 10^{12} neurones (mille milliards), avec 10^3 à 10^4 synapses (connexions) par neurone. De plus, tous les neurones du cerveau n'ont pas la même architecture ni le même rôle au sein de l'organisation générale.

Un neurone biologique est constitué de trois parties distinctes qui assurent :

- la collecte d'information ;
- l'intégration de cette information ;
- la restitution de l'information vers d'autres cellules.

La collecte de l'information est effectuée par les dendrites du neurone qui réceptionnent l'information des unités afférentes par l'intermédiaire des connexions synaptiques. Cette information est alors acheminée, grâce à un processus électrochimique, vers la cellule elle-même qui intègre cette information au sein de son noyau, également appelé soma. Une fois traitée cette information, est répercutée en sortie de la cellule vers l'axone qui propage cette information vers d'autres cellules via les axones terminaux et les connexions synaptiques (Figure 2.4) [12].

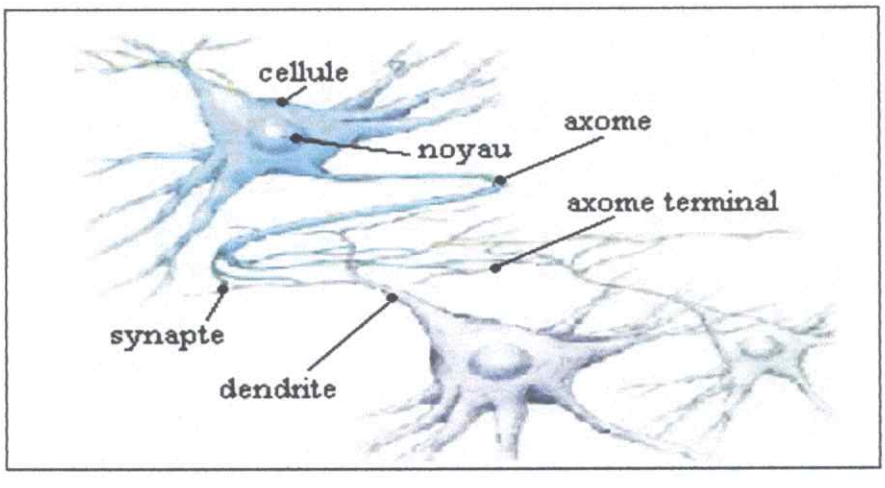


Figure 2.4 : Vue des connexions neuronales biologiques [12].

L'axone, qui permet à une cellule de propager son activité, peut être très long. Cette longueur variable permet à une cellule d'être en contact avec d'autres qui ne sont pas forcément dans son voisinage proche, de manière à répercuter une information locale dans une autre région du cerveau.

2.4. Modélisation du neurone

Les réseaux de neurones biologiques réalisent facilement un certain nombre d'applications telles que la reconnaissance de formes, le traitement du signal, l'apprentissage par exemple, la mémorisation, la généralisation. Ces applications sont pourtant, malgré tous les efforts déployés en algorithmique et en intelligence artificielle, à la limite des possibilités actuelles. C'est à partir de l'hypothèse que le comportement intelligent émerge de la structure et du comportement des éléments de base du cerveau que les RNA se sont développés. Les RNA sont des modèles, à ce titre, ils peuvent être décrits par leurs composants, leurs variables descriptives et les interactions des composants.

2.4.1. Structure du neurone formel (artificiel)

Chaque neurone formel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones en amont. A chacune de ces entrées est associée un poids «w » abréviation de weight (poids en Anglais) représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones en aval. A chaque connexion est associée un poids.

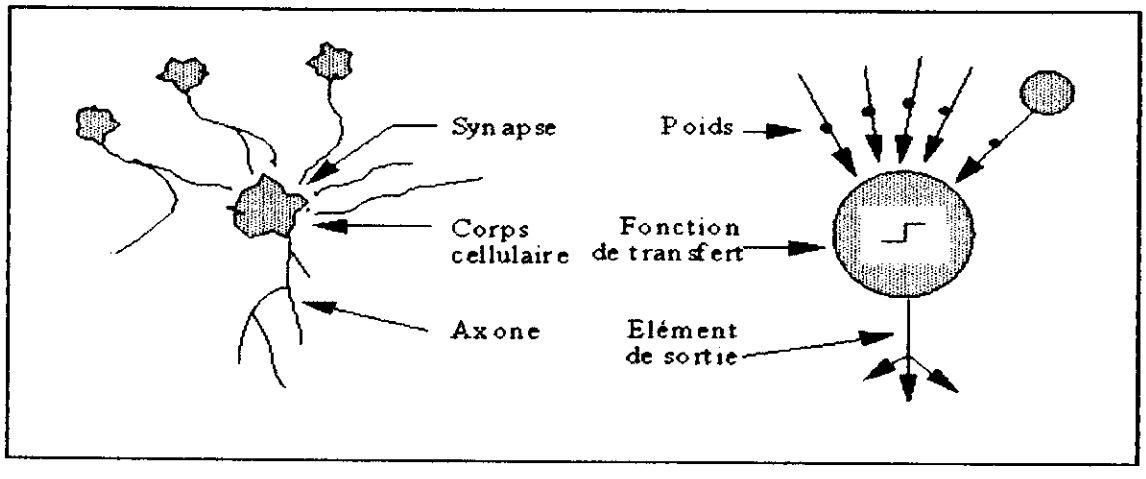


Figure 2.5 : Mise en correspondance entre le neurone biologique et le neurone artificiel [13].

2.4.2. Modèle initial

W. Mac Culloch et W. Pitts présenta la première modélisation du neurone, une formalisation du neurone conduisant à la description d'un automate à Seuil.

Cet automate effectue une somme pondérée de ses entrées (chaque entrée est une valeur numérique qui correspond à l'activation d'un autre automate) et déclenche une réponse si cette somme dépasse un certain seuil. La figure 2.6 résume la chaîne de traitement développée par l'automate.

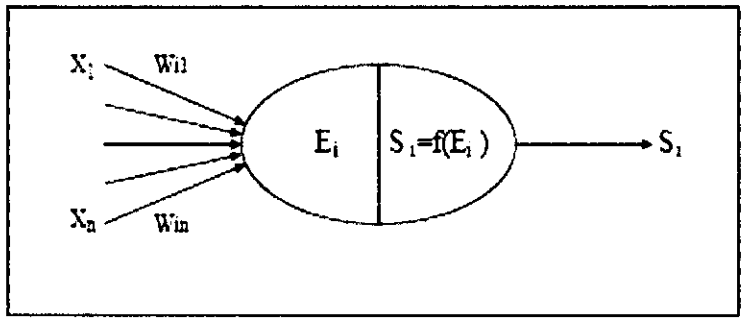


Figure 2.6 : Neurone formel.

La réponse finale est calculée selon la formule suivante :

$$S_i = f\left(\sum_{j=1}^{n} W_{ij} X_j\right) \qquad (2.2)$$

Les X_j sont les potentiels d'action binaires émanant des autres neurones, les poids W_{ij} représentent l'efficacité synaptique entre les deux neurones i et j et S_i est la sortie calculée du neurone i. La fonction de transfert f permet un seuillage de la fonction d'entrée totale et possède l'allure donnée sur la figure 2.7 ou β représente le seuil (ou biais) de déclenchement du neurone i.

2.4.3. Extension

Le modèle originel décrit un automate à seuil booléen pour lequel les entrées et la sortie sont binaires. Nous pouvons alors étendre le neurone formel en modifiant les types des entrées et sorties, la fonction d'entrée totale Ei ou encore la fonction de seuil f. Nous verrons dans la suite que les extensions de ce modèle se sont révélées nécessaires pour le dépassement de certaines limites du neurone formel.

Classiquement des entrées réelles sont utilisées et un biais β caractérisant le seuil de déclenchement du neurone est introduit. La fonction d'entrée totale devient alors une fonction affine du type :

$$E_i = \sum_{j=1}^{n} W_{ij} X_j - \beta \qquad (2.3)$$

La fonction de transfert peut prendre diverses formes comme celle décrites sur la figure 2.7.

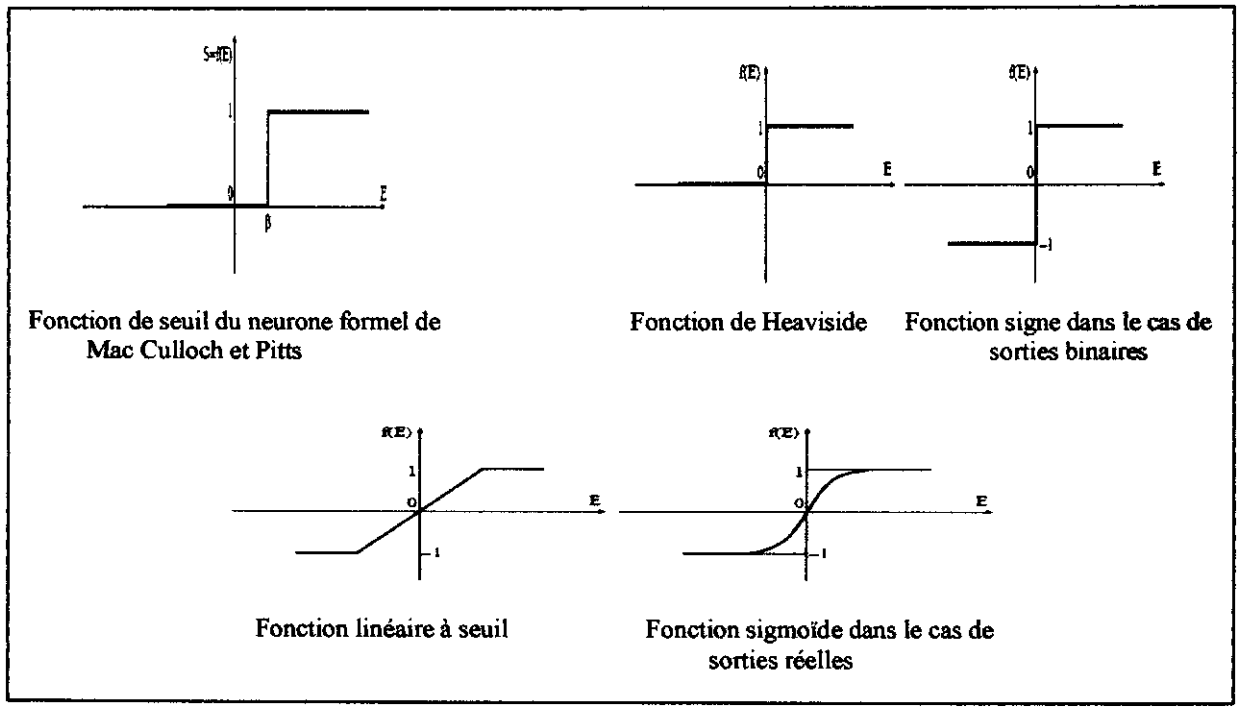


Figure 2.7 : Différentes fonctions de transfert des entrées.

Un neurone formel effectue une séparation linéaire de ses entrées. Si l'espace des entrées est linéairement séparable en deux classes, il est possible de trouver les bons poids synaptiques du neurone pour définir l'hyperplan qui sépare l'espace en deux parties. La limite de transition du neurone (sa frontière entre l'état actif et l'état inactif) est donnée par l'hyperplan d'équation :

$$\sum_{i=1}^{n} W_{ij} X_j + \beta = 0 \qquad (2.4)$$

L'algorithme d'apprentissage initialement utilisé avec ces réseaux ne permettait pas d'ajuster plus d'une couche de poids. Or, un réseau ayant cette architecture est dans l'incapacité de résoudre des problèmes non linéairement séparables. Un exemple très simple d'un tel problème est la fonction logique ou exclusive (XOR). La découverte de ce problème simple et pourtant insoluble avec la théorie a été à l'origine de l'abandon du paradigme connexionniste pendant plus d'une décennie (Figure 2.8) [12].

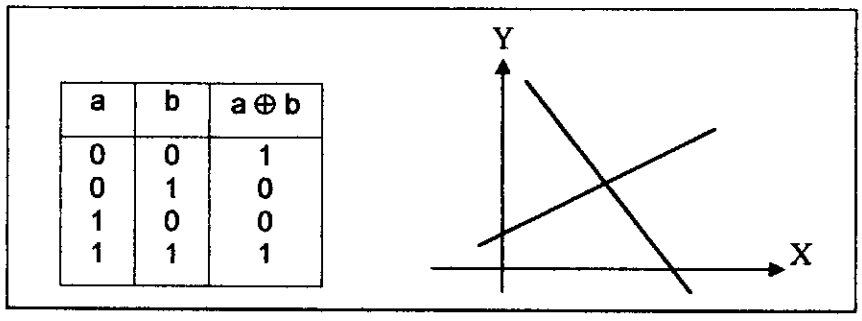


Figure 2.8 : La fonction XOR et la présentation schématique du graphe des Régions d'une fonction non linéairement séparable.

Il a alors fallu attendre une quinzaine d'années pour voir apparaître une méthode d'apprentissage capable de circonvenir ce problème. Ce type de réseau a alors connu d'importants développements et de nombreuses architectures ont été définies pour résoudre des problèmes variés.

2.5. Structure d'interconnexion

Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Elle peut être quelconque, mais le plus souvent, il est possible de distinguer une certaine régularité.

2.5.1. Réseau à connexions locales

Il s'agit d'une structure multicouche, mais qui à l'image de la rétine, conserve une certaine topologie. Chaque neurone entretien des relations avec un nombre réduit et localisé de neurones de la couche avale (Figure.2.9). Les connexions sont donc moins nombreuses que dans le cas d'un réseau multicouche classique.

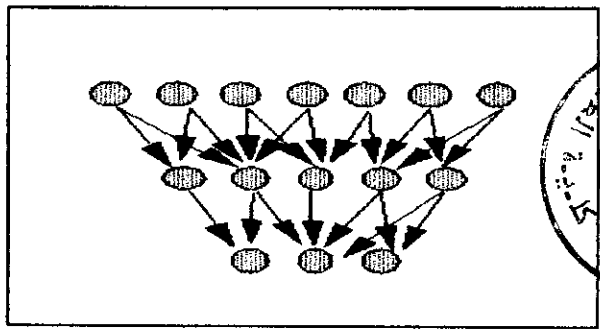


Figure 2.9 : Réseau à connexions locales.

2.5.2. Réseau à connexions récurrentes

Les connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouche. Ces connexions sont le plus souvent locales (Figure 2.10).

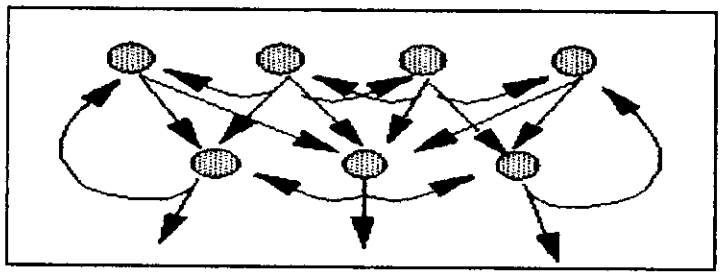


Figure 2.10 : Réseau à connexions récurrentes.

2.5.3. Réseau à connexion complète

C'est la structure d'interconnexion la plus générale. Chaque neurone est connecté à tous les neurones du réseau et à lui-même (Figure 2.11).

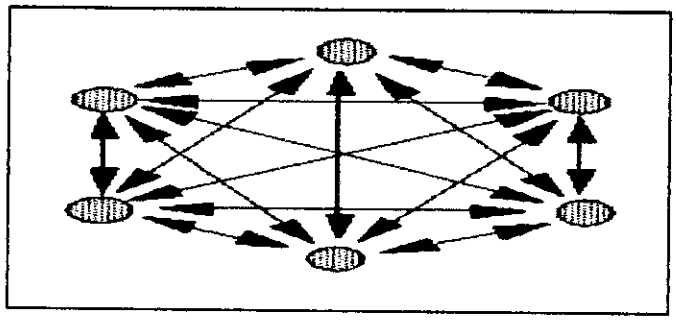


Figure 2.11 : Réseau à connexions complète.

2.6. Apprentissage

L'apprentissage est vraisemblablement la propriété la plus intéressante des RN. Cependant elle ne concerne pas tous les modèles, mais les plus utilisés.

L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. L'apprentissage neuronal fait appel à des exemples de comportement [13].

Dans le cas des RNA, on ajoute souvent à la description du modèle l'algorithme d'apprentissage. Le modèle sans apprentissage présente en effet peu d'intérêt. Dans la majorité des algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions. L'apprentissage est la modification des poids du réseau dans le but d'accorder la réponse du réseau aux exemples et à l'expérience.

Au niveau des algorithmes d'apprentissage, il a été défini deux grandes classes selon que l'apprentissage est dit supervisé ou non supervisé. Cette distinction repose sur la forme des exemples d'apprentissage. Dans le cas de l'apprentissage supervisé, les exemples sont des couples (Entrée, Sortie associée) alors que l'on ne dispose que des valeurs (Entrée) pour l'apprentissage non supervisé. Remarquons, cependant que les modèles à apprentissage non supervisé nécessitent avant la phase d'utilisation une étape de labellisation effectuée par l'opérateur, qui n'est pas autre chose qu'une part de supervision.

Comme nous venons de le voir, le neurone biologique ou formel n'a pas un comportement que l'on pourrait qualifier d'intelligent au sens courant du terme. Le neurone exécute une fonction mathématique simple (séparation linéaire) dont l'utilisation pratique se trouve très limitée.

2.7. Le perceptron originel

En 1957, Rosenblatt décrit le premier perceptron. Les propriétés du modèle ainsi que ses capacités vont rapidement le rendre célèbre.

Il s'agit d'un réseau d'automate à seuil qui est composé d'une couche d'entrée et une cellule de décision. Cette dernière est un adaptateur linéaire, de poids (w_i) $1 \leq i \leq n$ et de seuil θ variable, et réalise fonction linéaire à seuil :

$$O = \text{sign}\left(\sum_{i=1}^{n} W_i \, I_i - \theta \right) \qquad (2.5)$$

La fonction linéaire à seuil permet de séparer les exemples en deux classes (C^+, C^-), de part et d'autre de l'hyperplan :

$$\sum_{i=1}^{n} W_i \, I_i - \theta = 0 \qquad (2.6)$$

Dans la pratique, le seuil est généralement identifié à un poids supplémentaire $W_0 = \theta$, appelé biais, et qui est associé à une entrée fixé $I_0 = -1$. La couche d'entrée est composée de (n+1) unités (Figure 2.11). La fonction d'activation de ces dernières est l'identité. En posant $I = (I_0,\ldots,I_n)$ et $\mathbf{W} = (W_0,\ldots,W_n)$, on conviendra que **W.I doit être** positif pour tout exemple de C^+ et négatif pour tout exemple de C^-. L'apprentissage est supervisé et peut être décrit.

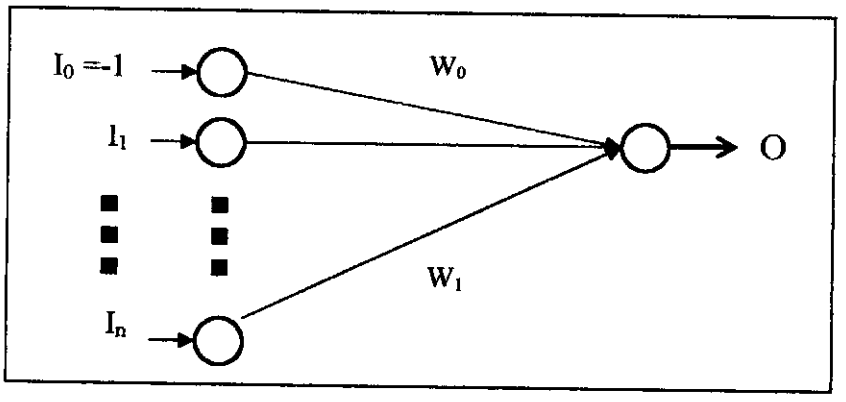


Figure 2.12 : Le perceptron originel.

2.7.1. Règle d'apprentissage du Perceptron

Soit E un ensemble d'exemples d'apprentissage de cardinalité L tel que :

$$E = \{(I^1, Z^1), (I^2, Z^2), \ldots, (I^L, Z^L)\} \qquad (2.7)$$

Où $I^k \subset C = C^+ \cup C^-$,

tel que $I^k = (1, I^k_1, \ldots, I^k_n)'$

Et $Z^k = \begin{cases} 1 \text{ si } I^k \in C^+ \\ (-1) \text{ si } I^k \in C^- \end{cases}$ pour k = 1,2,..., L.

Algorithme d'apprentissage du Perceptron

ETAPE 0 :

choisir une valeur quelconque pour **W**

poser : k = 0

t = 0

ETAPE 1 :

k = k +1

Si $Z^k = 1$ et **W'**. $I^k > 0$

alors : aller à l'ETAPE 2

Si $Z^k = 1$ et **W'**. $I^k \leq 0$

alors : $\mathbf{W} = \mathbf{W} + I^k$

t = t+1

aller à l'ETAPE 2

Si $Z^k = -1$ et **W'**. $I^k < 0$

alors : aller à l'ETAPE 2

Si $Z^k = -1$ et **W'**. $I^k \geq 0$

alors : $\mathbf{W} = \mathbf{W} - I^k$

t = t+1

aller à l'ETAPE 2

ETAPE 2 :

Si k< L

alors : aller à l'ETAPE 1

Si k = L et $t \neq 0$

alors : poser k = 0

aller à l'ETAPE 1

Si k = L et t = 0

alors : STOP

Les poids ne sont modifiés que lorsqu'un exemple est mal classé. Cet exemple tire ou repousse les poids selon qu'il s'agit d'un exemple positif ou négatif. Le principe de cet apprentissage est basé sur la correction des erreurs : pour chaque exemple présenté en entrée I^k, on connaît la sortie désirée Z^k, et on compare avec la sortie calculée par le réseau $O^k = W.I^k$, puis on exploite un signal d'erreur.

On peut étendre la structure du perceptron en disposant en sortie plusieurs cellules de décision (Figure 2.13). Cela permet de traiter des problèmes de classification à un plus grand nombre de classes.

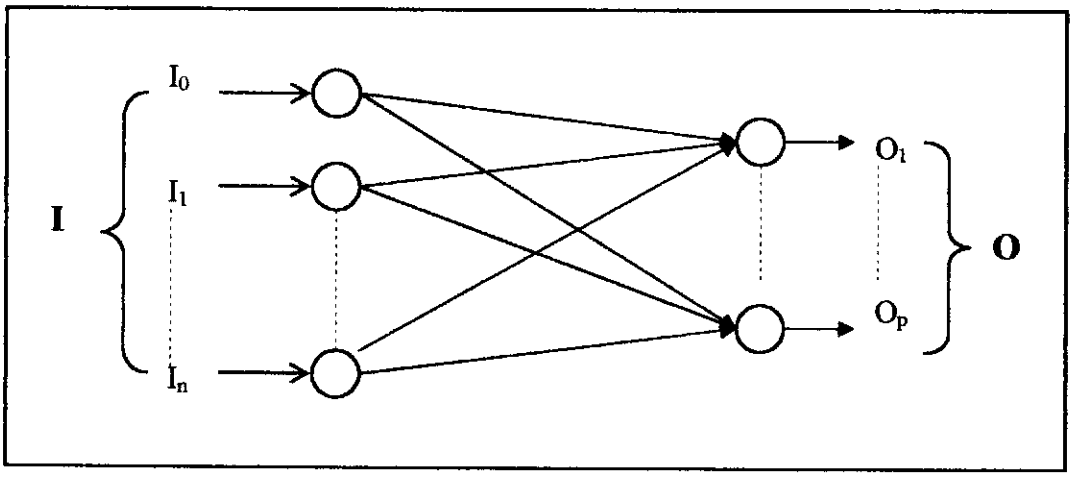


Figure 2.13 : Schéma général d'un perceptron monocouche.

2.7.2. Limites du perceptron

Lorsque les classes d'exemples ne sont pas linéairement séparables, l'algorithme d'apprentissage du perceptron ne converge pas en général, et ne garantit même pas que la fonction déterminée après un nombre fini d'étapes soit une bonne approximation de la séparation souhaitée.

2.8. L'ADALINE de Windrow

A la même époque que l'apparition du perceptron, Windrow étudiait un modèle de cellules assez proche, nommé "ADAptiv(e) LINeair Element" ou encore ADALINE. L'architecture est un simple réseau monocouche à une seule cellule de sortie.

Cette cellule est un adaptateur linéaire comme la cellule de décision du perceptron. Sa sortie peut être réelle ou binaire après seuillage. La différence essentielle avec le perceptron se situe au niveau de sa règle d'apprentissage : l'erreur est évaluée sur la sortie linéaire avant seuillage. Cette règle d'apprentissage a été déduite de la méthode des moindres carrés par Windrow et Hoff [14].

2.8.1. Règle de Windrow-Hoff, ou règle Delta

Soit E un ensemble d'exemples d'apprentissage de cardinalité L tel que :

$$E = \{(I^1, Z^1), (I^2, Z^2), \ldots, (I^L, Z^L)\} \qquad (2.8)$$

Où $I^k \subset C = C^+ \cup C^-$,

tel que $I^k = (1, I^k_1, \ldots, I^k_n)$'et $Z^k \in \Re$

Algorithme d'apprentissage de règle Delta

ETAPE 0 :

choisir une valeur quelconque pour W

poser : k = 0

t = 0

l = 0

W(t) = W

fixer ε

ETAPE 1 :

k = k +1

présentation d'un exemple I^k:

calcul de la sortie linéaire de l'ADALINE

$$O_k = \sum_{i=0}^{n} w_i(t).\ I^k_i = (W(t))'.\ I^k$$

présentation de la sortie désirée associée à l'exemple I^k : Z^k

Si $|O_k - Z^k| > \varepsilon$

alors : aller à l'ETAPE 2

Sinon aller à l'ETAPE 3

ETAPE 2 :

$wi(t+1) = wi(t) + \alpha\ .\ (Z^k - O^k).\ I^k_i(t)$ $(\forall\ i \in \{0,\ldots,n\})$

qui peut s'écrire : $\Delta W(t) = \alpha\ .\ (Z^k - (W(t))'.\ I^k).\ I^k$

t = t+1

l = l+1

ETAPE 3 :

Si k< L

alors : aller à l'ETAPE 1

Si k= L et l ≠ 0

alors : k = 0

l = 0

aller à l'ETAPE 1

Si k = L et l = 0 alors : STOP

Si l'on note E(t) l'erreur quadratique commise par la cellule sur l'exemple présenté à l'instant t, on observe que :

$$-\partial E(t)/\partial w_i = (Z_t - O_t).\ I_t(t) \qquad (\forall\ i \in \{0,\ldots,n\}) \qquad (2.9)$$

La mise à jour des poids n'a lieu que lorsqu'une erreur est commise, et elle tend à diminuer l'erreur quadratique E(t) : c'est une adaptation de la méthode des moindres carrés. Le facteur de proportionnalité α est généralement appelé pas du gradient.

Cet apprentissage est supervisé, et fondé sur une méthode de descente du gradient. En effet, si l'on part de formules semblables, mais si l'on vérifie l'échelle des temps en effectuant la mise à jour des poids après une présentation complète de la base d'exemples E (|E| = L), alors la règle de Windrow et Hoff réalise une descente en gradient sur l'erreur globale et minimise cette fonction de coût. Notons *E* l'erreur globale, c'est-à-dire la demi somme des carrés des différences entre la sortie calculée et la sortie désirée pour tous les exemples de la base *E*.

$$E = \sum_{k}^{L} E_k = (1/2).\ \sum_{k}^{L} (Z_k - O_k)^2 \qquad (2.10)$$

Si l'on considère E comme étant l'aptitude associée à un point de l'espace des poids de connexions, et puisque l'on a :

$$\partial E/\partial w_i = \sum_{k}^{L} (\ \partial E_k/\partial w_i\ ) \qquad (\forall\ i \in \{0,\ldots,n\}) \qquad (2.11)$$

2.9. Les perceptrons multicouches et la retropropagation

Les perceptrons multicouches sont les réseaux à la base des méthodes connexionnistes. Ils sont, en effet, les plus employés et les plus étudiés. Deux abréviations anglaises sont utilisées dans la littérature pour les nommer : ANN (**A**rtificial, **N**eural, **N**etworks). Ou MLP pour (**M**ulti **L**ayer **P**erceptrons) [14].

Un perceptron multicouche est composé de plusieurs couches de neurones et de connexions (Figure 2.14). Ce nombre est au moins égal à deux, signifiant ainsi que le réseau possède deux couches de poids connexionnistes, une couche de sortie et une couche cachée. Le nombre de couches cachées détermine la complexité des frontières

des différents sous-espaces que le réseau pourra représenter. La complexité de l'approximation est également déterminée par le nombre de neurones de chaque couche puisque ce nombre détermine le nombre maximal d'informations que le réseau peut extraire du signal traité.

Un réseau multicouche est une extension du modèle monocouche issu du perceptron, avec une ou plusieurs couches cachées successives entre la couche d'entrée et la couche de sortie. La connectivité la plus courante consiste à définir des connexions uniquement d'une couche vers la couche suivante. On peut aussi autoriser des connexions pondérées, par exemple de la couche d'entrée vers la seconde couche cachée.

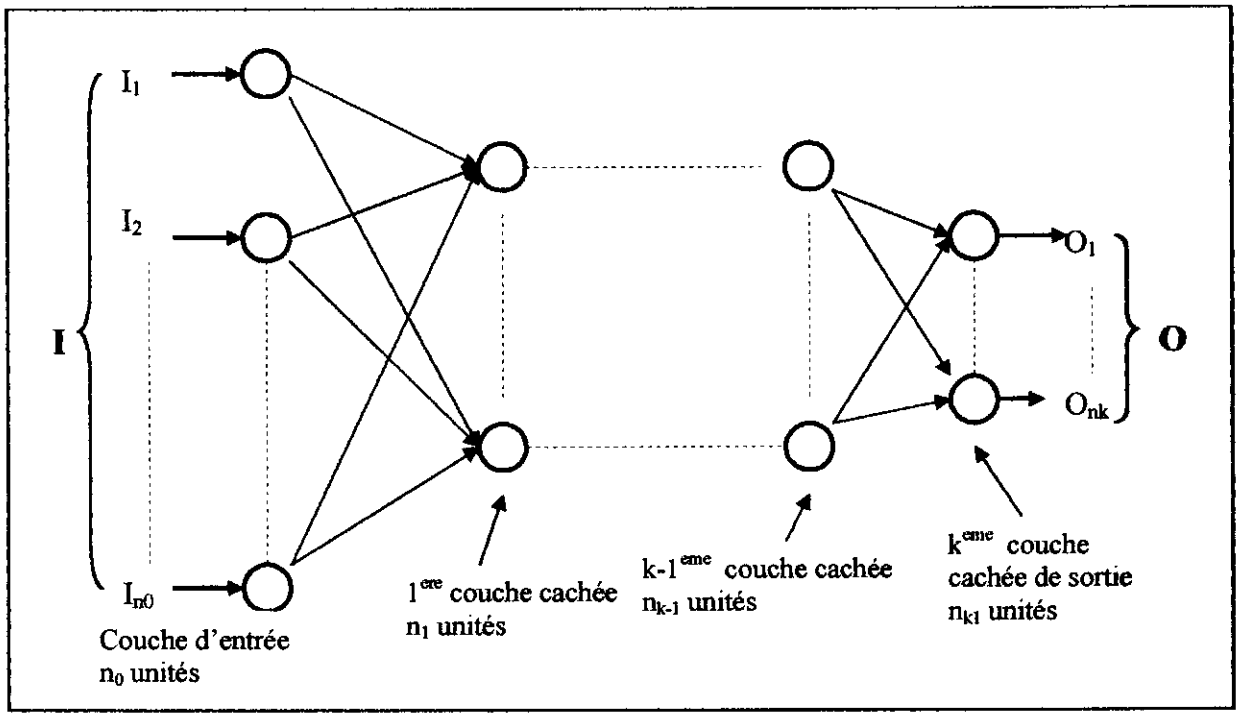


Figure 2.14 : Schéma d'un réseau neuronal multicouches avec k+1 couche.

Nous exposons ici les règles de propagation, d'activation, et d'apprentissage concernant les réseaux multicouches. La couche des entrées est parfois omise dans le décompte des couches car ces unités ont pour rôle de transmettre simplement les valeurs d'entrées vers le réseau. Les unités de toutes les autres couches sont des automates quasi-linéaires qui transmettent les activations de couche en couche jusqu'à

la couche de sortie. On désigne par :

$$w_{ijk} \; (\; 1 \leq j \leq n_{k-1} \quad 1 \leq j \leq n_k \; , \quad 1 \leq k \leq K \;) \qquad (2.13)$$

Le poids qui relie l'unité j de la couche k-1 à l'unité i de la couche k.

Ce type d'apprentissage est supervisé, et est basé sur la méthode du gradient. Plus précisément, il applique la technique de Windrow et Hoff pour minimiser une fonction de coût. Comme pour l'ADALINE, on peut définir pour chaque cellule de sortie une erreur quadratique locale :

$$E_i (I) = (Z_i (I) - O_{ik} (I))^2 \qquad (\forall \; i \in \{0,\ldots,n_k\}) \qquad (2.14)$$

Où $Z_i(I)$ est la sortie désirée de la cellule i, pour l'exemple I et O_{ik} (I) la sortie de l'unité i de la couche K (couche de sortie) qui a été calculée par le réseau multicouche sur la présentation de l'exemple I .

On en déduit un critère de coût global en calculant l'erreur moyenne sur la base d'exemple :

$$E = (1/ \text{ card } (E)). \sum_{I \in E} E (I) \qquad (2.15)$$

avec :

$$E (I) = (1/n_k). \sum_{i=0}^{n_k} E_i (I) \qquad (\forall I \in E) \qquad (2.16)$$

Comme dans le cas de l'ADALINE, on peut adapter une technique de gradient total et minimiser le coût global E, ou minimiser l'erreur E (I) sur chaque exemple, ce qui correspondra à une méthode de gradient stochastique.

Le point essentiel de la méthode est la propagation des erreurs en sens rétrograde, des cellules de sorties vers les cellules d'entrées, couche par couche, ce qui permet d'ajuster les poids. Sur le plan mathématique, c'est la relation de Chasles sur les dérivations partielles :

$$\partial f / \partial X = (\partial f / \partial U) \; . \; (\partial U / \partial f) \qquad (2.17)$$

Encore appelée " chain rule ", qui permet d'établir une fonction de récurrence entre les gradients des unités d'une couche est ceux des unités de la couche de rang supérieur.

Plus précisément :

$$\partial E(I) / \partial w_{ijk} = (\partial E(I)/ \partial a_{ik}). (\partial a_{ik} / \partial w_{ijk})$$

$$= (\partial E(I)/ \partial a_{ik}). \partial O_{j,k-1} \qquad (2.18)$$

On introduit une nouvelle variable $D_{ik}(I) = - (\partial E(I)/ \partial a_{ik})$, et l'on peut démontrer qu'il suffit en fait de calculer et de propager ces gradients de couche en couche.

On note : a_{ik} : l'activation de l'unité i de la couche k ;

$O_{j,k}$: la sortie de l'unité i de la couche k.

L'algorithme peut être défini uniquement si l'on considère des unités semi-linéaires, c'est-à-dire des unités dont la fonction de transition est différentiable. Les fonctions les plus couramment utilisées sont :

- $f(z) = 1 / (1 + e^{-\lambda z})$ qui varie entre 0 et 1;
- $\tilde{f}(z) = (e^{\lambda z} - 1) / (1 + e^{-\lambda z})$ qui varie entre -1 et 1.

Enfin, on définit un pas du signal α_k (t) qui peut être adaptatif dans le temps (dépendance de t) ou selon la couche (indexation par k). Toutes les notations étant définies, on peut maintenant décrire l'algorithme de rétropropagation, dans le cadre du gradient stochastique, de la manière suivante :

2.9.1. Règle d'apprentissage par RETROPROPAGATION

Soit E un ensemble d'exemples d'apprentissage de cardinalité L tel que :

$$E = \{(I^1 , Z^1), (I^2,Z^2), \ldots,(I^L,Z^L)\} \qquad (2.19)$$

Où $I^k = (I^k_1, I^k_2, \ldots, I^k_{n0})'$

Et $Z^k = (Z^k_1, Z^k_2, \ldots, Z^k_{nk})'$

Algorithme d'apprentissage par RETROPROPAGATION

INITIALISATION

En t = 0, initialiser les poids w_{ijk} **(0)** à de faibles valeurs aléatoires.

REPETER

Pour chaque exemple I^1 de la base E présente aléatoirement faire :

- présenter l'exemple $I^1(t) = (I^1_{i0}(t))$ aux unités d'entré ;
- calculer les états et les sorties du réseau en appliquant, couche par couche, en sens direct, la règle de propagation des activations :

$$(\forall\ i \in \{1,\ldots,n_0\}) \qquad O_{i0}(t) = a_{i0}(t) = I^1_{i0}(t)$$

$$(\forall\ k \in \{1,\ldots,k\}), (\forall\ i \in \{1,\ldots,n_k\}) \qquad O_{ik}(t) = f(a_{ik}(t))$$

avec : $a_{ik}(t) = \sum_{j=0}^{n_{k-1}} w_{ijk}(t).\ O_{ik-1}(t)$

et *f* une fonction sigmoïde

- présenter les sorties désirées aux unités de sorties, et calculer les gradients associés à ces unités :

$$(\forall\ i \in \{1,\ldots,n_k\}) \qquad D_{i,k}(t) = (\partial E(t)/\partial\, a_{i,k}(t))$$

$$= 2\ .\ f'(a_{i,k}(t).\ (Z_i(t) - O_{i,k}(t))$$

- calculer les gradients aux autres unités en appliquant, couche par couche, en sens rétrograde la formule de récurrence :

$$(\forall\ k \in \{1,\ldots,k-1\}), (\forall\ i \in \{1,\ldots,n_k\})$$

$$D_{i,k}(t) = (f'(a_{i,k}(t)) \ .\ \sum_{m=0}^{n_{k+1}} w_{m,i,k+1}(t).\ D_{m,k+1}(t)$$

- mettre à jour les poids des connexions :

$$(\forall\ i \in \{1,\ldots,k\}), (\forall\ i \in \{1,\ldots,n_k\}), (\forall\ j \in \{1,\ldots,n_{k-1}\})$$

$$w_{i,j,k}(t+1) = w_{i,j,k}(t) + \alpha_k(t).\ D_{i,k}(t).\ O_{i,k-1}(t)$$

- t = t + 1

tant que la base *E* non épuisée

jusqu'au critère d'arrêt.

La convergence de l'algorithme stochastique n'étant pas démontrée dans le cas général. On utilise en pratique divers critères d'arrêt :

- un seuil sur la fonction de coût, au- dessous duquel on stoppe l'apprentissage ;
- un seuil sur le taux de succès de la base d'exemple, au-dessus duquel on arrête ;
- un nombre de passes prédéterminés, c'est-à-dire un nombre d'interactions fixées.

2.10. Phase de généralisation

La phase de reconnaissance, ou de généralisation consiste à présenter de nouveaux motifs ou exemples et à évaluer les sorties calculées par le réseau après avoir figé les poids dans l'état obtenu après apprentissage. Il est usuel de tester les capacités de généralisation d'un réseau en évaluant le taux de succès sur une base de test, entièrement disjointe de base d'apprentissage, mais pour laquelle les sorties désirées sont connues [15].

2.11. Conclusion

Dans ce chapitre, nous avons présenté les principales méthodes d'analyse qui permettent d'extraire des caractéristiques qui seront ensuite utilisées en RAP. La présentation des ces méthodes a été suivie par les modèles connexionnistes et leurs algorithmes d'apprentissages, afin de comprendre le fonctionnement des réseaux de neurones et comment les mettre en œuvre dans une application phonémique en vue de leur Reconnaissance Automatique.

Le but de ce chapitre est d'acquérir des connaissances sur les réseaux de neurones multicouches avec un apprentissage supervisé. Ce derniers, peuvent constituer d'excellents classificateurs pour la RAP.

Chapitre 3

Application des Réseaux de Neurones à la Reconnaissance Phonétique des Occlusives Orales de l'Arabe Standard

3.1. Introduction

L'utilisation des RN dans un système de RAP implique nécessairement l'ajustement de quelques paramètres et l'optimisation de certains critères.

Le but de la première partie est de mettre en place les notions nécessaires, permettant l'élaboration d'un système de reconnaissance des phonèmes occlusifs de l'Arabe Standard en utilisant des RN statiques tels que les Perceptrons Multi-Couches (PMC). La deuxième partie est consacrée à une étude acoustique des sons spécifiques de l'AS en vue de la RAP.

3.2. L'utilisation des RN dans la RAP

Un réseau de neurones formels est constitué d'un grand nombre de cellules de base interconnectées. De nombreuses variantes sont définies selon le choix de la cellule élémentaire, de l'architecture du réseau et de la dynamique du réseau. Une cellule élémentaire peut manipuler des valeurs binaires ou réelles. Les valeurs binaires sont représentées par 0 et 1 ou -1 et 1. Différentes fonctions peuvent être utilisées pour le calcul de la sortie. Le calcul de la sortie peut être déterministe ou probabiliste.

L'architecture du réseau peut être sans rétroaction, c'est-à-dire que la sortie d'une cellule ne peut influencer son entrée. Elle peut être avec rétroaction totale ou partielle. La dynamique du réseau peut être synchrone : toutes les cellules calculent leurs sorties respectives simultanément. La dynamique peut être asynchrone. Dans ce dernier cas, on peut avoir une dynamique asynchrone séquentielle : les cellules calculent leurs sorties chacune à son tour en séquence ou avoir une dynamique asynchrone aléatoire.

Par exemple, si on considère des neurones à sortie stochastique -1 ou 1 calculée par une fonction à seuil basée sur la fonction sigmoïde, une interconnexion complète et une dynamique synchrone, on obtient le modèle de Hopfield et la notion de mémoire associative. Si on considère des neurones déterministes à sortie réelle calculée à l'aide de la fonction sigmoïde, une architecture sans rétroaction en couches successives avec une couche d'entrées et une couche de sorties, une dynamique asynchrone séquentielle, on obtient le modèle du Perceptron Multi-Couches (**PMC**), ou Multi Layer Percetron (**MLP**) [15].

Dans notre travail nous avons utilisé un système de RAP, basé sur un réseau connexionniste de type PCM pour plusieurs raisons. Tout d'abord, ces réseaux ont de grandes capacités d'apprentissage à partir d'exemples, qui résistent remarquablement aux bruits, leurs robustesses aux données manquantes, possèdent une forte capacité discriminante et ont montré leurs aptitudes en parole, notamment pour les mots isolés [12], [18]. De plus, ils possèdent une architecture flexible, des structures régulières et parallèles ce qui rendent facilement utilisables du point de vue HardWare [19]. Il est également possible de détecter le moment où l'algorithme d'apprentissage n'est plus capable d'améliorer les performances, ce qui permet l'optimisation du temps d'apprentissage.

3.3. La RAP par un Réseau de Neurone Modulaire (RNM)

Un RNM ou Réseau Expert (RE) est système neuronal composé de sous-réseaux de neurones ou modules. Dans la RAP par le RNM de type PMC, chaque module est spécialisé dans la reconnaissance d'un seul phonème par rapport aux autres (Figure 3.1).

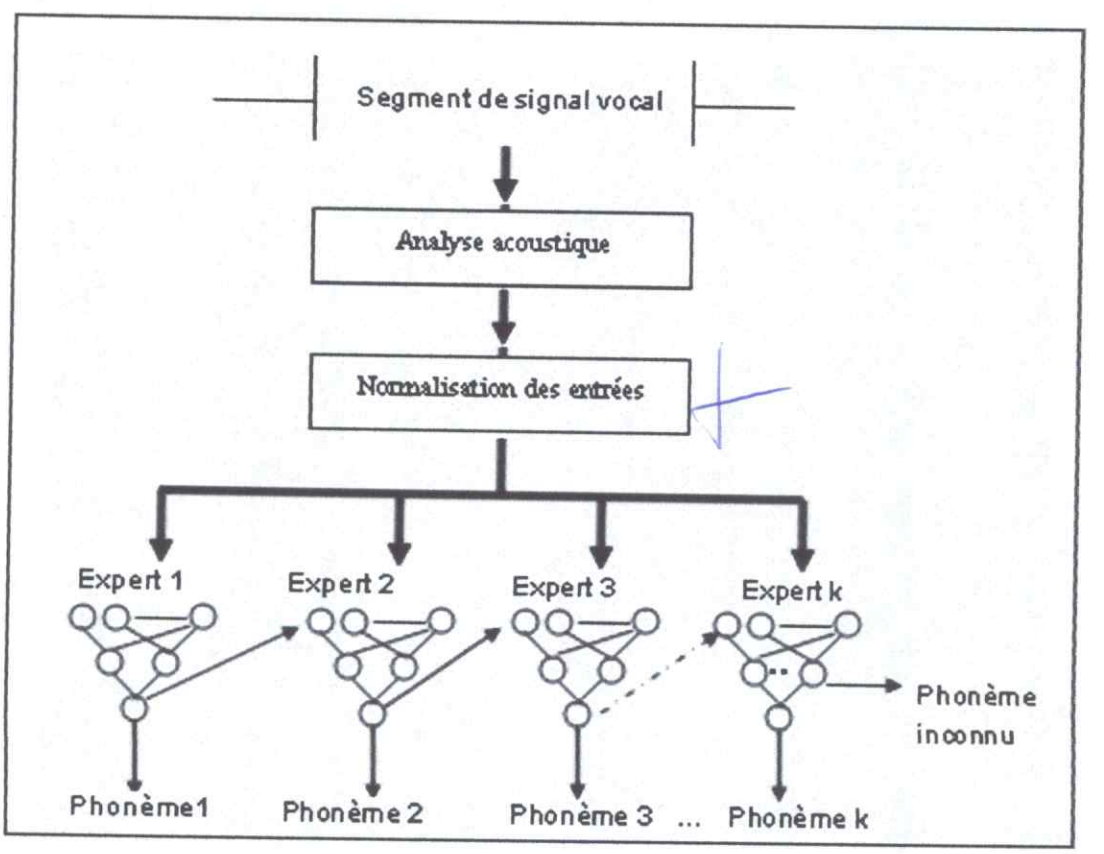


Figure 3.1 : Système neuronal modulaire pour la RAP.

L'avantage du RNM par rapport à un PMC unique se trouve dans la décomposition de tâche complexe en plusieurs tâches simples ou sous-tâches, ce qui facilite la représentation des données d'entrées et diminuer la durée et la complexité de l'apprentissage (Figure 3.2). Un système de reconnaissance par PMC unique fait la tâche de reconnaissance globale de tous les problèmes avec un seul PMC, ce qui permet d'avoir des confusions inévitables entre les phonèmes [20]. En plus, la structure du RNM nous permet de mesurer le taux de reconnaissance pour chaque phonème et donc de détecter les phonèmes qui posent le plus de problèmes lors de la reconnaissance.

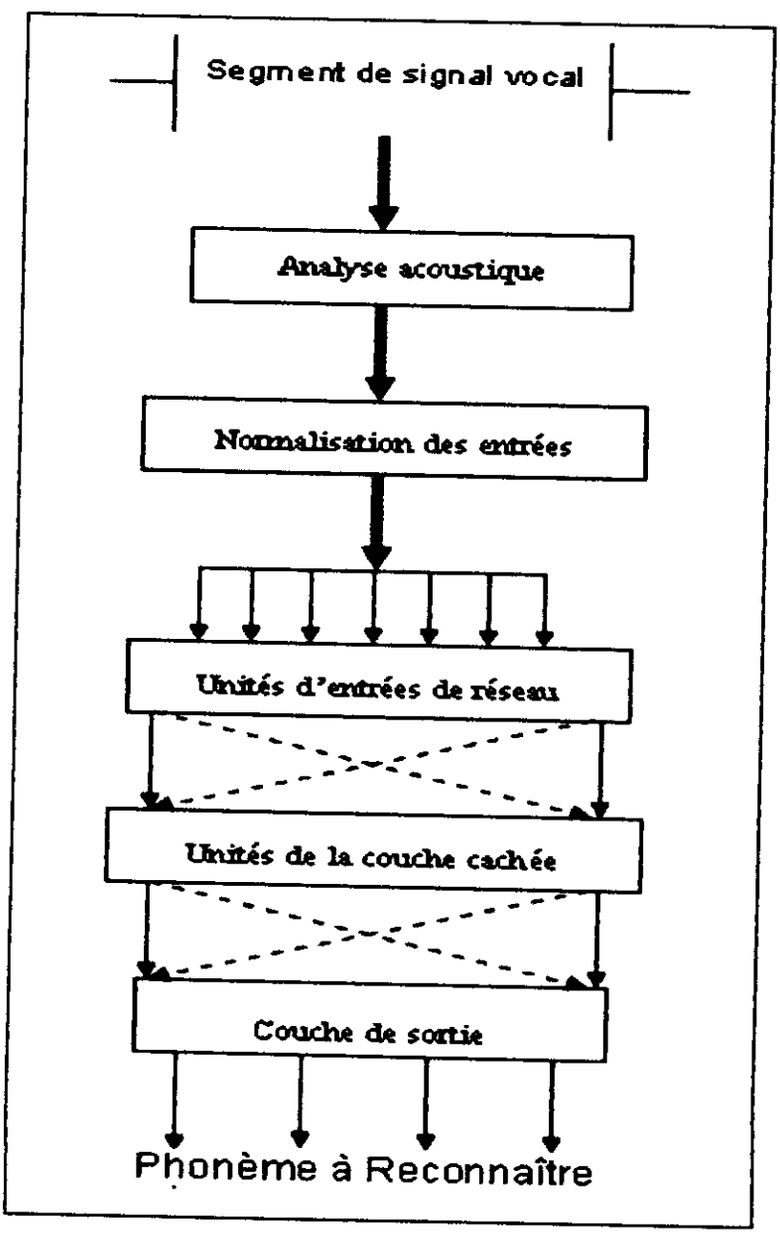


Figure 3.2 : Système basé sur un PMC unique spécialisé dans la RAP.

3.4. Mise en œuvre des RNM dans la RAP

Pour réaliser l'approximation de la fonction de régression cherchée, à partir d'échantillons généralement bruités, à l'aide d'un réseau de neurones, trois étapes successives sont nécessaires il faut :

- choisir l'architecture du réseau, c'est-à-dire les entrées externes, le nombre de neurones cachés, et l'agencement des neurones entre eux, de telle manière que le réseau soit en mesure de reproduire ce qui est déterministe dans les données. Le nombre de poids ajustables est un des facteurs fondamentaux de la réussite d'une application. Si le réseau possède un trop grand nombre de poids, c'est-à-dire si le réseau est trop «souple», il risque de s'ajuster au bruit qui est présent dans les données de l'ensemble d'apprentissage, et, même en l'absence de bruit, il risque de présenter des oscillations non significatives entre les points d'apprentissage; si ce nombre est trop petit, le réseau est trop «rigide» et ne peut reproduire la partie déterministe de la fonction. Le problème de la détermination de l'architecture optimale est resté pendant longtemps un problème ouvert, mais il existe actuellement diverses méthodes, mettant notamment en jeu des tests statistiques, qui permettent de déterminer cette architecture pour une vaste classe de réseaux ;
- calculer les poids du réseau ou, en d'autres termes, estimer les paramètres de la régression non linéaire à partir des exemples, en minimisant l'erreur d'approximation sur les points de l'ensemble d'apprentissage, de telle manière que le réseau réalise la tâche désirée ;
- enfin estimer la qualité du réseau obtenu en lui présentant des exemples qui ne font pas partie de l'ensemble d'apprentissage [21].

3.4.1. Analyse des données

Dans la RAP, il est nécessaire d'effectuer une analyse des données de manière à déterminer les caractéristiques discriminantes pour détecter ou différencier ces données. Ces caractéristiques constituent l'entrée du RN. Notons que cette étude n'est pas spécifique aux RN, quelle que soit la méthode de détection ou de reconnaissance de la forme utilisée. Il est généralement nécessaire de présenter des caractéristiques représentatives. Cette détermination des caractéristiques a des conséquences à la fois sur la taille du réseau (et donc le temps de simulation), et sur

le temps de développement (temps d'apprentissage). Une étude statistique sur les données peut permettre d'écarter celles qui sont aberrantes et redondantes. Dans le cas d'un phonème de classification, il appartient à l'expérimentateur de déterminer le nombre de classes auxquelles ces données appartiennent et de déterminer pour chaque donnée la classe à laquelle elle appartient.

3.4.2. Initialisation des poids synaptiques

L'initialisation des poids avant l'application de l'algorithme d'apprentissage par rétro-propagation du gradient dans un système de RAP est une tâche importante, car elle influe sur la vitesse de convergence du réseau [20]. Plus les poids initiaux sont proches de leur valeur finale et plus la convergence est rapide. En effet, quand ces poids sont trop faibles ceci entraîne un apprentissage très long. On peut distinguer dans la littérature deux méthodes d'initialisation : les méthodes d'initialisation aléatoires dans un intervalle choisi de manière adéquate et les celles basées sur des technique non aléatoires. En ce qui concerne notre travail, nous nous sommes limités aux algorithmes basés sur la rétro-propagation du gradient.

S. Fahlman propose de choisir des poids dans un intervalle, variant de [-0.4, 0.4] à [-0.5, 0.5], selon des données d'apprentissage [22]. L. Bottou propose d'initialiser les poids dans un intervalle $[-a\sqrt{d_{in}}, a\sqrt{d_{in}}]$, où *a* est calculé de sorte que la variance des poids corresponde au point où la pente de la tangente de la fonction d'activation est maximum, et d_{in} le nombre d'unités de la couche précédente [18]. Une autre méthode proposée par G. Burel, consiste à choisir des poids de manière uniforme dans un intervalle dépendant des données à apprendre. Ce type d'initialisation des poids n'est cependant pas facile à mettre en œuvre [22].

Dans notre cas, nous avons initialisé les poids synaptiques par des valeurs comprises entre (-0.5 et 0.5), car ces poids permettent à notre système de converger rapidement.

3.4.3. Représentation des poids

Un choix doit être fait concernant le codage des poids (utilisés par la rétro-propagation en tant que valeurs réelles). Dans les structures neuronales on est confronté dans ce paragraphe au problème du choix entre une représentation binaire ou une représentation réelle. Si l'on choisit une représentation binaire, il est

nécessaire de décoder les poids en valeurs réelles avant chaque opération de reconnaissance ou d'apprentissage des RN.

Afin de limiter le nombre d'opérations à effectuer lors de l'apprentissage, et sachant qu'il n'existe pas de règles bien définies sur la manière optimale de coder les poids, nous avons représenté des poids des connexions des réseaux par des réels sans effectuer de transformations.

3.4.4. Architecture du réseau

Trouver une architecture adéquate d'un RN à un problème donné, n'est pas une tâche simple, car le nombre optimal de couches cachées ainsi que le nombre de neurones dans chaque couche et leurs connexions se fait plus de manière empirique que par une méthode basée sur un fondement théorique. Une méthode appropriée consiste à utiliser un réseau très petit (exemple d'un neurone dans la couche cachée) puis ajouter des neurones jusqu'à l'obtention de bonnes performances [23].

Un exemple qui a été donnée sur l'influence de l'architecture sur les performances d'un système de Reconnaissance des voyelles françaises. Le corpus d'apprentissage compote 200 voyelles dont 160 sont utilisées pour l'apprentissage et 40 pour le test (Figure 3.3).

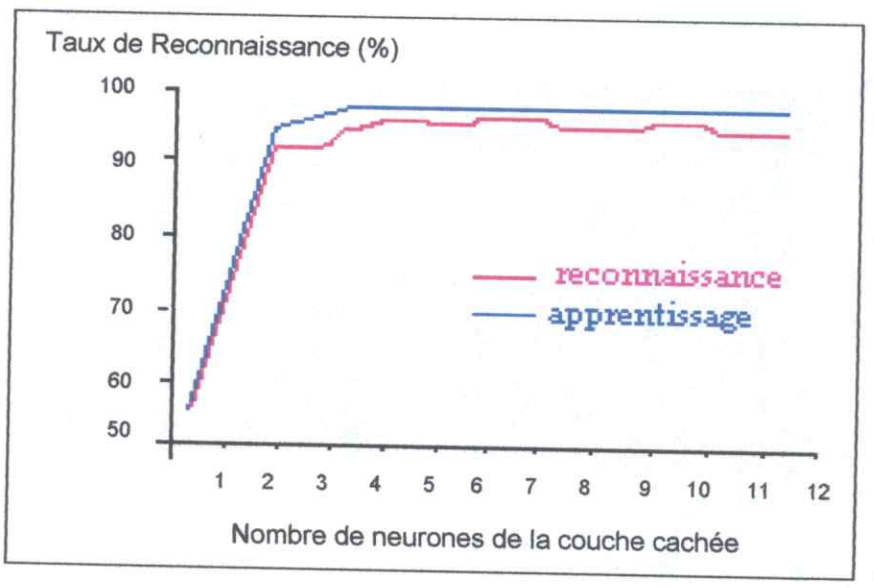


Figure 3.3 : Influence de l'architecture de RN sur les performances du système de reconnaissance des voyelles françaises [23].

3.4.5. Choix de pas d'apprentissage

Le choix des paramètres d'apprentissage influe beaucoup sur la rapidité de calcul. Dans le cas de l'algorithme de retro-propagation, le calcul du gradient consiste à définir, la direction dans laquelle doit s'effectuer la modification des poids. Le principe de descente du gradient consiste alors à effectuer de manière interactive (pas à pas) une modification des poids suivant cette direction afin d'avoir un minimum sur la fonction de coût représentant l'écart entre les sorties obtenues et celles de références. Si ce pas (ou gain d'adaptation) est trop petit, le nombre de pas nécessaire peut s'avérer relativement important et contribue donc à ralentir de manière non négligeable l'apprentissage. Notons qu'il est même possible que l'algorithme, rencontrant un minimum local, ne puisse plus en sortir. A l'inverse, si le pas est trop important, l'algorithme peut devenir instable.

Une méthode très utilisée consiste à modifier l'algorithme pour ajouter à ce gain d'adaptation un terme appelé *momentum* [22], [23]. La règle de modification des poids devient :

$$W_{ji}(t) = w_{ji}(t-1) + \eta\delta_j y_i + \alpha\,\Delta w_{ji}(t-1) \qquad (3.1)$$

Où α : le momentum varie entre 0 et 1 ;

t : un compteur du nombre d'interactions de la boucle principale.

L'intérêt de cette règle est de prendre en compte les modifications antérieures des poids dans le but d'éviter des oscillations perpétuelles (Figure 3.4).

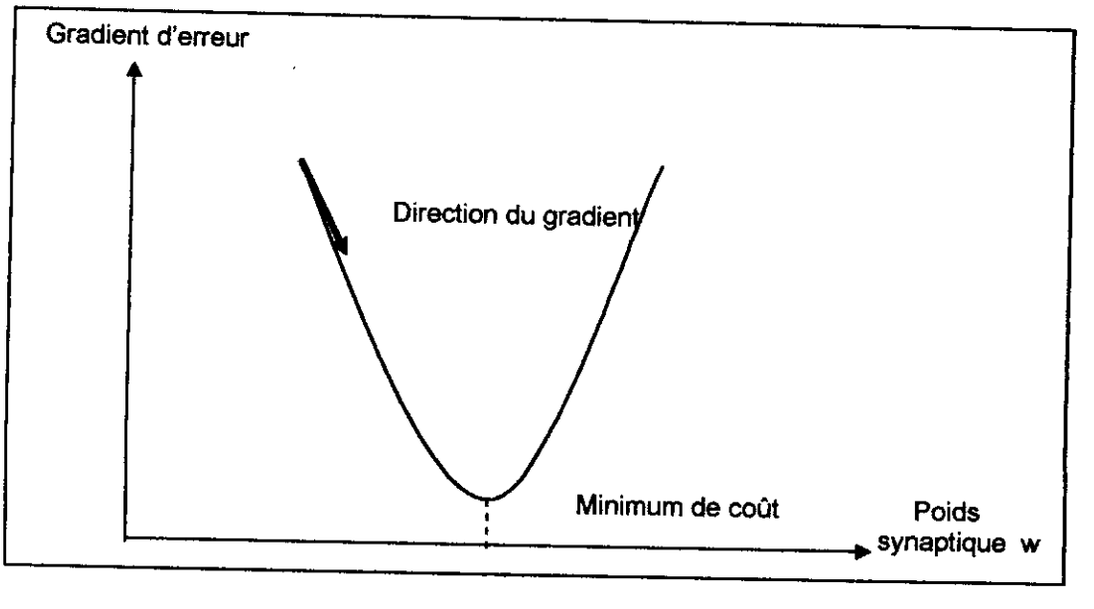


Figure 3.4 : Variations du gradient d'erreurs en fonction des poids (w) [24].

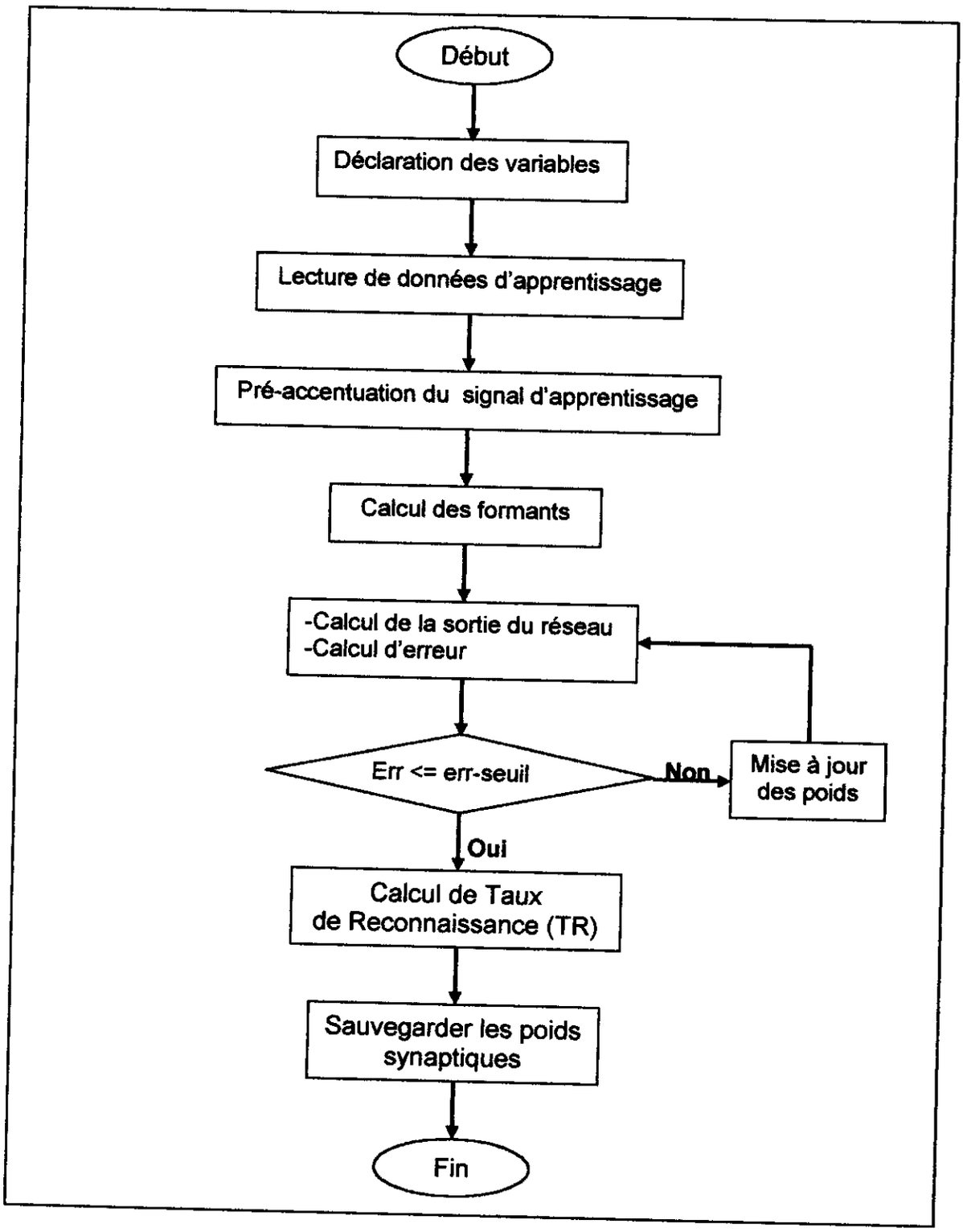


Figure 3.5 : Organigramme d'Apprentissage du système de RPOO de l'Arabe Standard.

3.4.6. Apprentissage et test de généralisation

La RAP par les RN de type PMC nécessite un apprentissage supervisé, en utilisant l'algorithme de rétro-propagation du gradient comme méthode d'apprentissage. Si un système neural possède un nombre trop grand de neurones de la couche cachée par rapport à celui des exemples de base d'apprentissage, tous les exemples seront parfaitement appris : on parle d'apprentissage par cœur ou *surparamétrisation.*

La figure 3.6 représente l'évaluation de l'erreur quadratique moyenne pour l'apprentissage et de test, en fonction des itérations d'apprentissage. Sur la base d'apprentissage, l'erreur diminue toujours, alors que la base de test, passe par un minimum. Si l'apprentissage se prolonge au-delà, les performances en test diminuent. Ce phénomène est dû à l'apprentissage par cœur des exemples de la base d'apprentissage. Afin d'arrêter l'apprentissage juste avant que ne se produise ce phénomène de surapprentissage, plusieurs méthodes ont été proposées. La plus utilisée en RAP est dite de la *validation croisée*. Cette méthodes consiste à diviser la base de données en trois bases : d'apprentissage, de test et de validation croisée. Cette dernière base est utilisée pendant l'apprentissage afin d'examiner le comportement du réseau pour des données qui lui sont inconnues. Ainsi, l'apprentissage est arrêté lorsque l'erreur sur la courbe B (base de validation croisée) atteint un minimum.

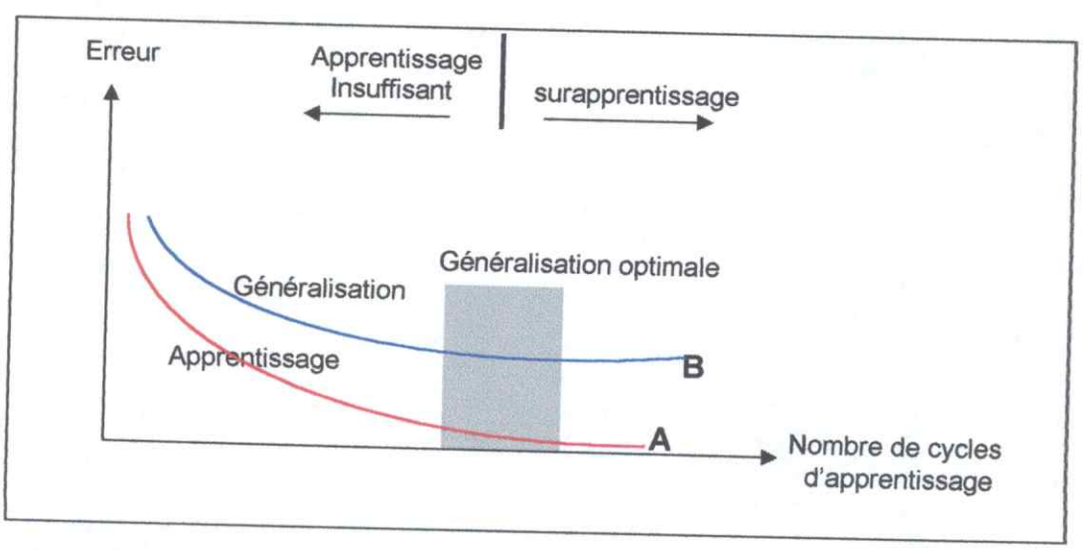


Figure 3.6 : Evolution de l'erreur d'apprentissage et de généralisation [24].

L'utilisation de la méthode de validation croisée dans les systèmes de la RAP nécessite une taille importante de données d'apprentissage et de test (Figure 3.7) pour améliorer les performances de ces systèmes [23], [24].

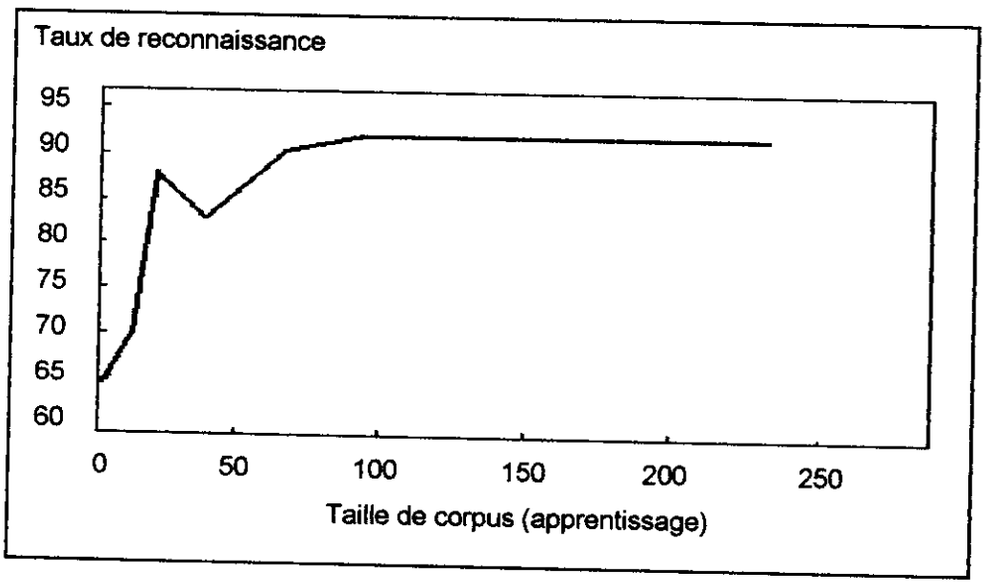


Figure 3.7 : Exemple de l'influence de la taille du corpus sur les performances du système [24].

3.4.7. Phase de reconnaissance

Dans un système de RAP avec les réseaux PMC, l'entité reconnue est celle qui correspond à la sortie ayant le potentiel maximum. Une entité est bien reconnue si la sortie procurée par le réseau est la même que celle désirée. La fonction de reconnaissance comporte trois paramètres identiques à ceux de la fonction d'apprentissage : le nombre d'entités à reconnaître, les vecteurs de caractéristiques des entrées et les sorties désirées associées.

La fonction d'apprentissage comporte trois paramètres : le nombre d'entités à reconnaître, les caractéristiques associées à ces entités et les sorties désirées pour chaque vecteur de caractéristiques. Elle fournit le Taux de Reconnaissance (TR) obtenu sur les données d'apprentissage et indique le nombre d'itérations pour obtenir ce résultat.

A l'issue de l'apprentissage, l'algorithme fournit un TR correspondant au Nombre d'Entités Reconnues (NER) divisé par le Nombre Total d'Entités (NTE).

$$TR = \frac{NER}{NTE} \qquad (3.2)$$

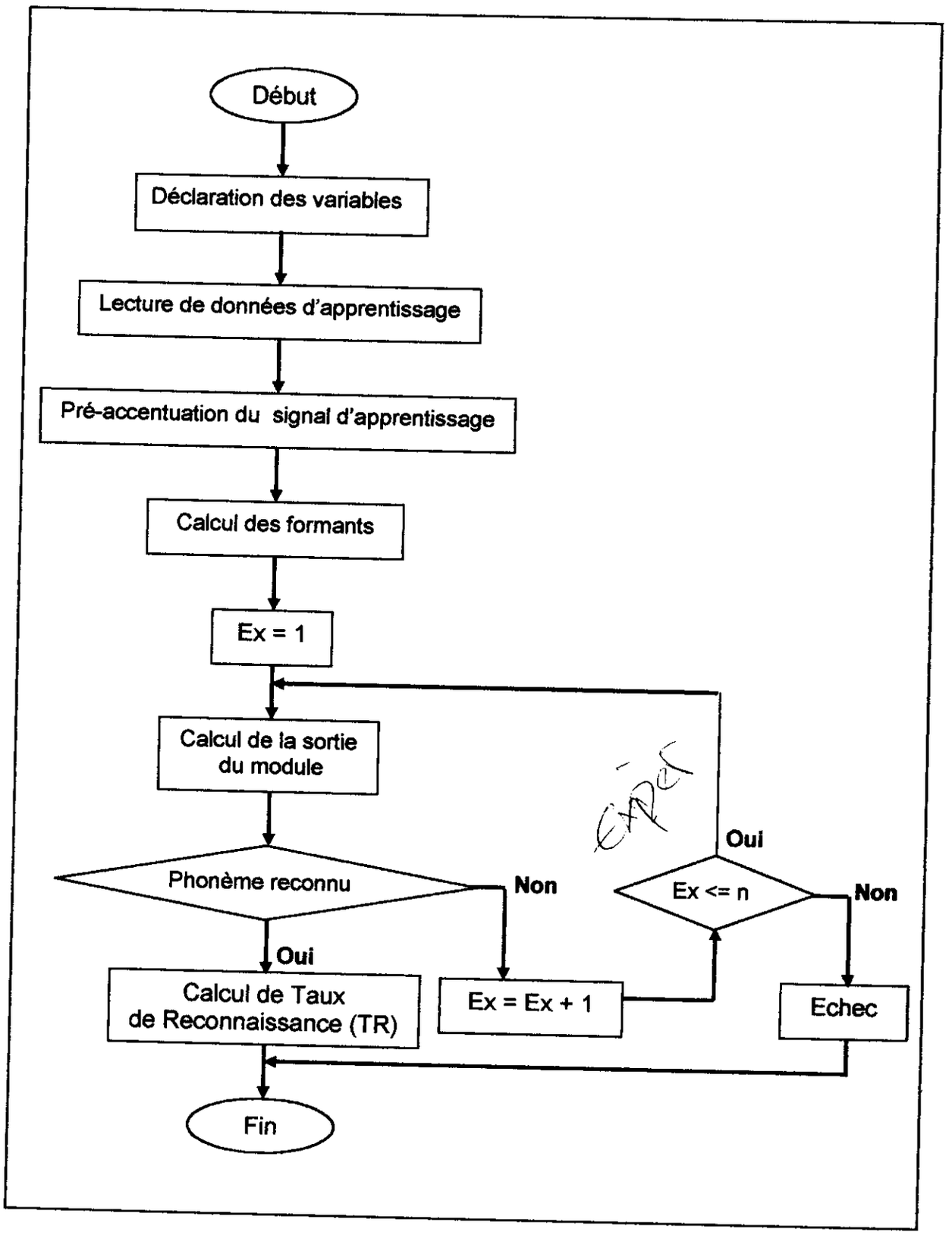


Figure 3.8 : Organigramme de Reconnaissance des Phonèmes Occlusifs de l'AS.

3.4.8. Mesure des performances

Le choix de la fonction de performance dans la RAP par RN est une étape très importante, car elle influe directement sur le TR du système. La fonction de performance définit les bases ou les manières sur lesquelles on doit arrêter l'algorithme de rétro-propagation du gradient. Le test le plus simple, mais aussi le moins performant consiste à fixer le nombre de cycles d'apprentissage. Une autre méthode souvent utilisée consiste à arrêter le calcul dès que l'erreur quadratique moyenne estimée passe sous un seuil prédéfini. Cette méthode exige une connaissance de l'erreur qui correspond à l'erreur calculée à partir de données de test.

3.5. Caractéristiques acoustiques des consonnes occlusives Orales de l'AS

La RAP nécessite une étude acoustique des différents phonèmes à reconnaître, afin de dégager les caractéristiques relatives à ces derniers pour les exploiter lors de développement du système de reconnaissance. Les consonnes occlusives de l'AS sont :

- *Bilabiale sonore* [b] : est produite par une fermeture complète du canal respiratoire, mais avec vibrations des cordes vocales. A cette occlusive correspond une articulation relâchée, qui prend la forme d'une spirante (Figure 3.9) ;

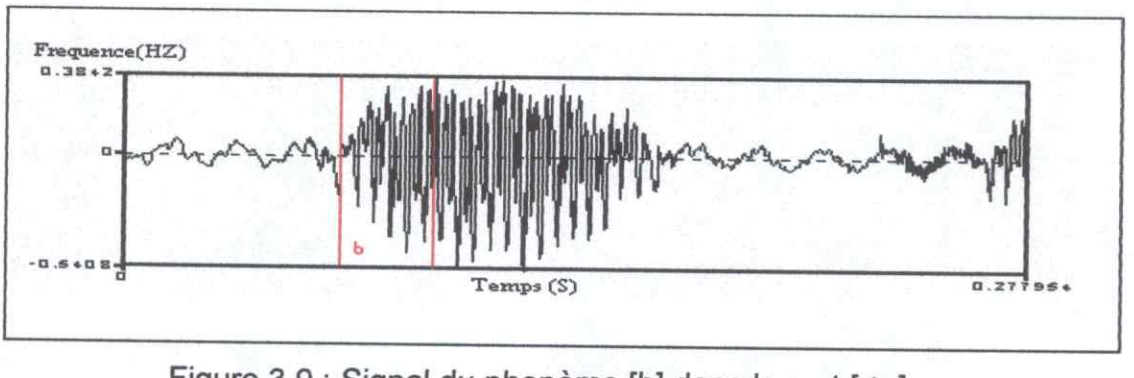


Figure 3.9 : Signal du phonème [b] dans le mot [بيت].

- *Dentale ou alvéolaire sonore* [d] : a la même articulation que la consonne *Bilabiale sonore*. L'articulation nasale dentale ou alvéolaire est également voisée. A cette occlusive correspond une articulation relâchée, qui prend la forme d'une spirante (Figure 3.10) ;

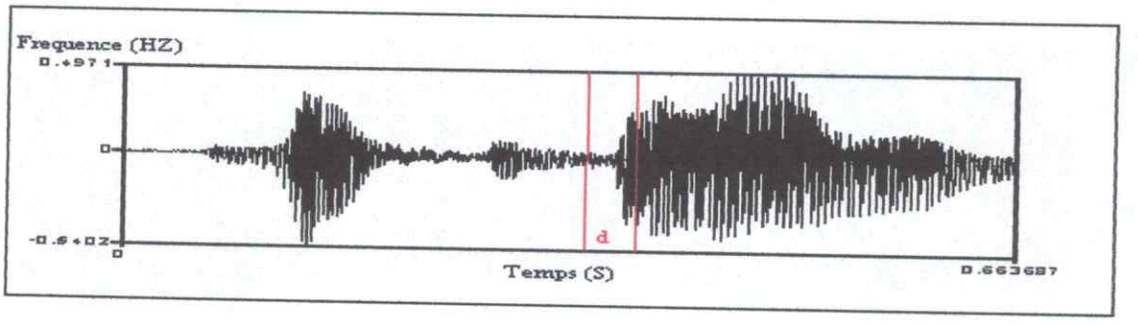


Figure 3.10 : Signal du phonème [d] dans le mot [ضفدع].

- *Dentale ou alvéolaire sourde* [t] : La langue prend contact avec le bourrelet formé par les alvéoles. A cette occlusive correspond une articulation relâchée, qui prend la forme d'une spirante (Figure 3.11) ;

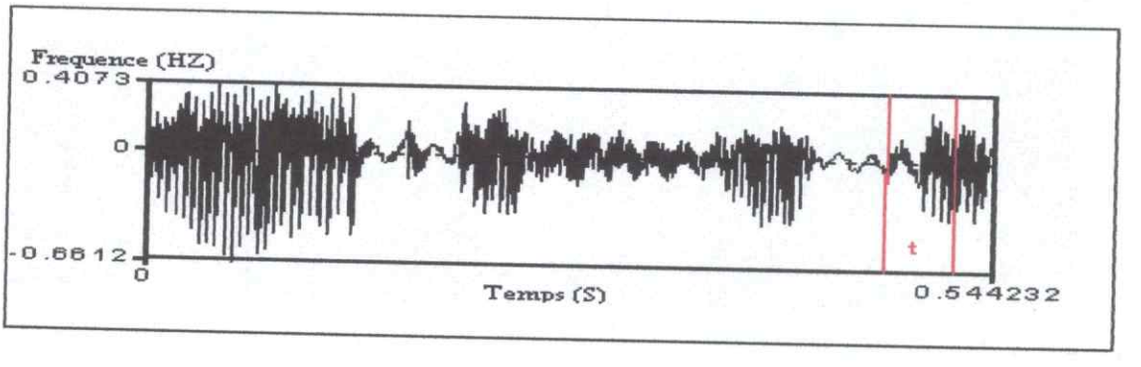


Figure 3.11 : Signal du phonème [t] dans le mot [رائحة].

- *Postpalatale sourde* [k] : Alors que la pointe de la langue est appuyée contre la face interne des dents du bas, le dos de la langue ou dorsum prend contact avec le palais mou, appelé aussi voile du palais (Figure 3.12) ;

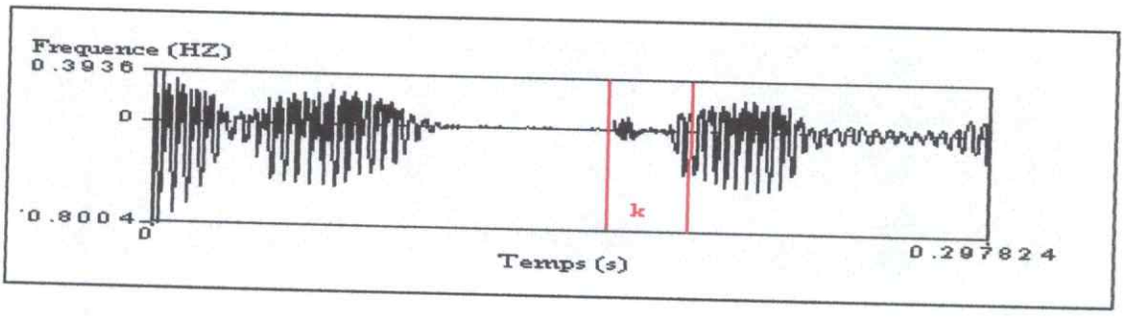


Figure 3.12 : Signal du phonème [k] dans le mot [ركب].

- *Glottale ou hamza* [ء] : l'occlusive glottale est produite, soit par l'ouverture soudaine de la glotte sous la poussée de l'air interne, soit par la fermeture

brutale du passage de l'air au niveau de la glotte. Cette occlusive se réalise toujours sourde. (L'occlusion au niveau de la glotte rend impossible toute vibration des cordes vocales (Figure 3.13) ;

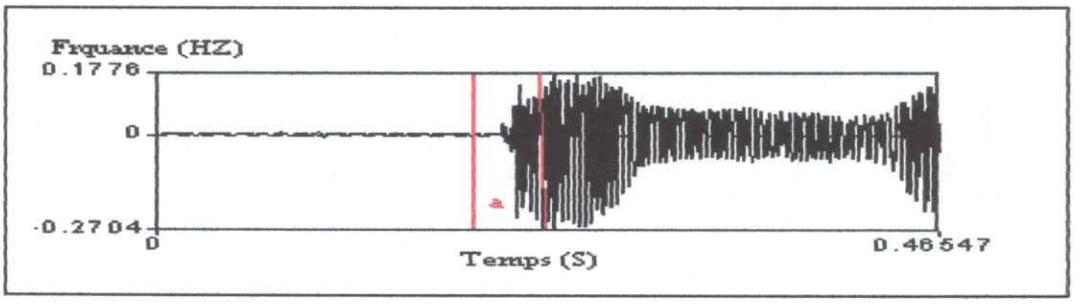


Figure 3.13 : Signal du phonème [ء] dans le mot [إن].

- *Uvulaire sourde* [t] : Pendant que la pointe de la langue demeure appliquée contre la face interne des dents du bas, le dorsum de la langue, relevé loin vers l'arrière prend contact avec le palais mou au niveau de la luette (Figure 3.14) ;
- *Rétroflexe sourde* [q] : une partie de la langue est retournée et sa pointe ou sa face intérieure prend appui sur un point de la partie antérieure du palais (Figure 3.14).

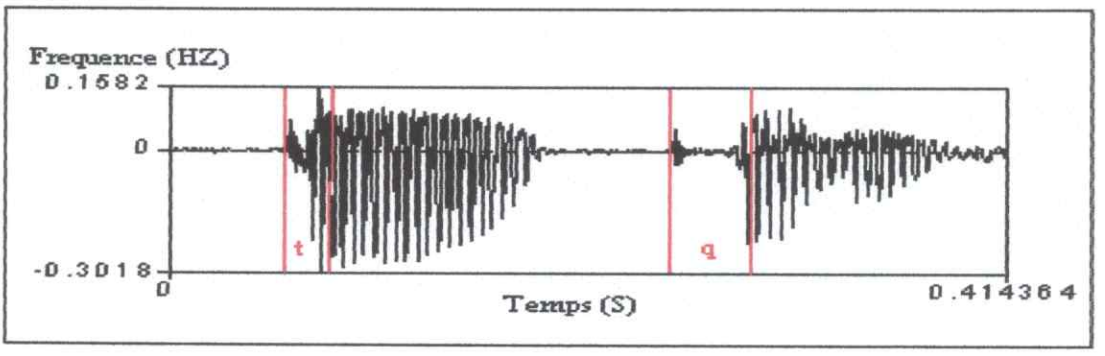


Figure 3.14 : Signal des phonèmes [t] et [q] dans le mot [طاقم].

- *Dentale ou alvéolaire sonore* [t] : Même articulation que la précédente, mais c'est une consonne emphatique (Figure 3.15).

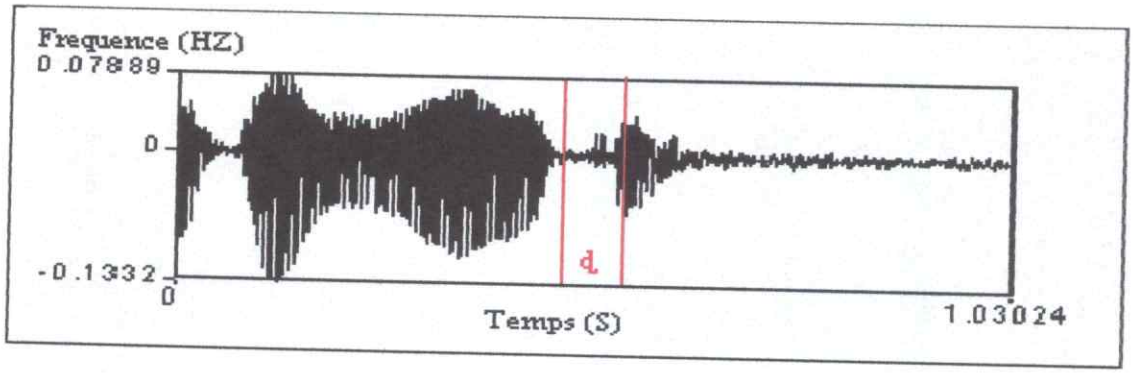


Figure 3.15 : Signal du phonème [d] dans le mot [غموض].

3.6. Problèmes liés aux particulières de l'AS

La complexité de la langue Arabe se trouve dans la particularité des traits phonétiques caractéristiques. Ces derniers peuvent aussi augmenter le degré de difficulté dans la conception d'un système de RAP. La particularité de l'AS se trouve aussi dans la présence des consonnes arrières glottales, pharyngales, vélaires, affriquée et du phénomène d'emphase et de la gémination. Toutes ces particularités rendent la RAP appliquée à la langue Arabe très délicate.

Les performances atteintes par les systèmes actuels laissent supposer qu'elles ne sont conditionnées que par la seule disponibilité d'un corpus représentatif convenablement segmenté et étiqueté. Dans le cas des occlusives de l'AS, on a deux cas particulier, les concepteurs des systèmes de RAP sont unanimes à constater que certaines caractéristiques telles que l'emphase, la gémination ou la pertinence sémantique de l'allongement temporel des phonèmes, constituent dans la majorité des cas, la source de complexité des systèmes de Reconnaissance, ajoutés aux problèmes classique de la RAP.

3.7 Conclusion

Dans ce chapitre, nous avons indiqué dans la première partie la méthode utilisée lors du développement de notre système de Reconnaissance phonémique concernant les consonnes occlusives de l'AS par les RNM. Notons que l'optimisation de certains critères tels que l'architecture de réseau, l'initialisation des poids synaptiques et le pas d'apprentissage, sont des phases importante qu'il faut prendre en considération dans notre système. Dans la seconde partie de ce chapitre, nous avons présenté les caractérisations acoustiques des consonnes occlusives de l'AS ainsi que leurs signal segmenté qu'on va utiliser dans notre système de reconnaissance.

Chapitre 4
Système de RACOOAS

4.1. Introduction

Dans ce chapitre nous réalisons un système de reconnaissance des phonèmes basé sur les RNM, en utilisant les différentes procédures présentées dans le troisième chapitre. La reconnaissance est appliquée aux phonèmes occlusifs oraux de l'Arabe Standard (AS). Ceci nécessite un corpus constitué de mots contenant les huit phonèmes à reconnaître. Après l'enregistrement de ce dernier, nous avons fait une segmentation afin d'extraire les formants qui sont utilisés comme valeurs d'entrées des réseaux. Enfin nous présentons notre logiciel avec les résultats obtenus de la reconnaissance.

4.2. Elaboration du corpus

La première étape à effectuer avant d'entamer les traitements, est l'élaboration du corpus. Dans notre cas, nous avons opté pour un corpus de parole naturelle continue en Arabe Standard constituée de mots de type énonciatif. Le nombre total de mots est de 120 mots (tab 4.1). Nous justifions le choix de ce type de corpus (parole continue au lieu de l'utilisation de logatomes) par le fait qu'il est préférable d'étudier les segments dans un continuum vocal pour pouvoir prendre en considération les effets de coarticulation existants entre les phonèmes.

Les systèmes RNM sont basés sur une modélisation grossière du neurone biologique, pour cela un bon corpus doit satisfaire les recommandations suivantes :

- nécessité d'un corpus par langue ;
- nécessité d'un corpus par type d'enregistrement ;
- afin de prendre en compte la variabilité des locuteurs, il faut des locuteurs hommes et femmes (la parité!), répartis dans toutes les tranches d'âge : enfants, adultes, appartenant à diverses catégories sociales et d'éducation variées et d'origine géographiques représentant les différents accents régionaux ;
- pour la reconnaissance phonémique il faut des milliers de phonèmes (par contexte) donc une centaine de phrases.

Les enregistrements du corpus ont été effectués au sein du laboratoire du département d'informatique de l'université Saâd Dahleb de Blida avec un matériel approprié et pour avoir de bonnes conditions on a choisi un moment de la journée assez calme.

Le locuteur retenu pour la lecture du corpus est une personne jeune, de sexe masculin jugé sans accent géographique apparent avec le débit d'élocution normal.

Nous présentons dans le tableau suivant Les mots de corpus à exploiter pour l'extraction des formants.

Phonemes	Début	Milieu	Fin
ب	بعل- باب- بشر- بكرة - برزخ	حبيب- حبر - كبير- منبر - مبراة	مجيب- لعب- حسب- عرب- لبيب
د	دخل- دب- دود- دليل- درب	مدرسة- أدب مصدر- منحدر - مدير	مهد- فهد- برد- سرد- فرد
ء	أم- أكل- ألم- أب- أسم	مؤلم- ضئيل- دائم- مؤذن- مألوف	أعداء- مليء- أقرأ- شيء- دعاء
ك	كتوم- كمال- كافر- كتاب- كرز	فكر- ركب- مكر- مكث- مكروه	برك- شريك- ديك- ركيك- مشرك
ض	ضرب- ضحك- ضفر- ضجع- ضمئ	رضي- مضغ- مضحك- نضيف- مريضة	مرض- بعض- يعض- بيض- فرض
ت	تلفزة- ترك- تمرين- توت- تبن	فتوى- أتى- حاتم- متوا- فتات	مات- حوت- سبت- ذات- موت
ق	قصير - قال- قريب- قهوة- قلم	عقاقير - فقر - مقود- حقير - بوقلقال	الريق- طليق- عميق- شهيق- عتيق
ط	طاف- طمس- طول- طار - طويل	باطن- باطل- غطس- مطر - مستطيل	محيط- سقط- لقيط- بسيط- مخطط
Nombre total des mots	120		
Nombre total des phonèmes	179		

Tab 4.1 : Corpus utilisé.

4.3 Extraction des formants

Une fois le corpus des mots porteurs, enregistré, nous passons à la phase de segmentation. Chaque mot porteur est visualisé au moyen d'un éditeur de signaux pourvu d'un sonagraphe (PRAAT) qui nous donne la représentation temps-fréquence et énergie (sonagramme) du mot porteur (Figure 4.1). De bonnes connaissances phonétiques sont nécessaires pour effectuer un étiquetage phonétique c'est-à-dire une opération qui consiste à identifier et à reconnaître sur le sonagramme chaque phonème du mot porteur.

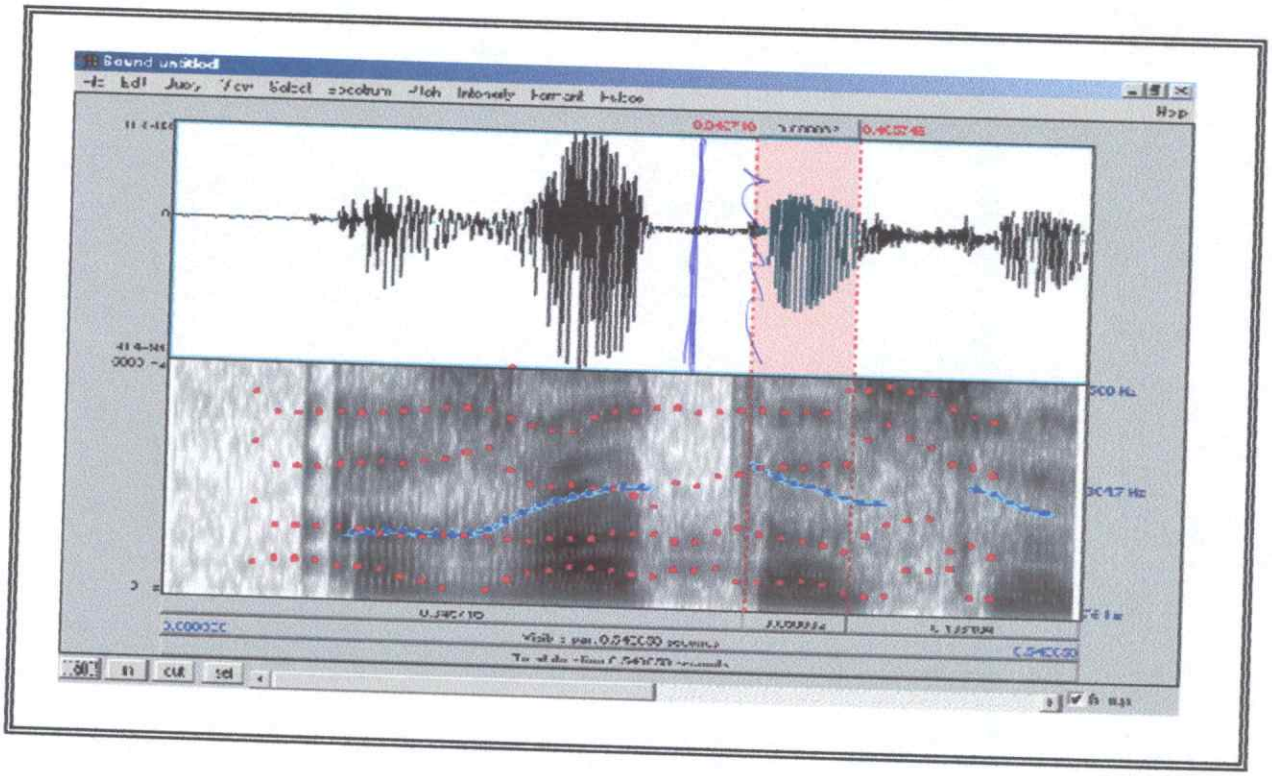


Figure 4.1 : Exemple de segmentation du phonème (ط)
dans le mot (غطست) à l'aide du logiciel PRAAT.

4.4. Notre architecture du réseau

La structure du système que nous utilisons est basée sur la reconnaissance de phonèmes occlusifs orales de l'AS. Le système est constitué de sous réseaux ou de modules de type MLP avec un apprentissage par rétro-propagation du gradient comme méthode d'apprentissage. A chacun de ces experts nous avons attribué des sous-tâches de reconnaissance des huit phonèmes en question. Chaque expert est spécialisé dans la reconnaissance d'un seul phonème par rapport aux autres (Figure 4.2).

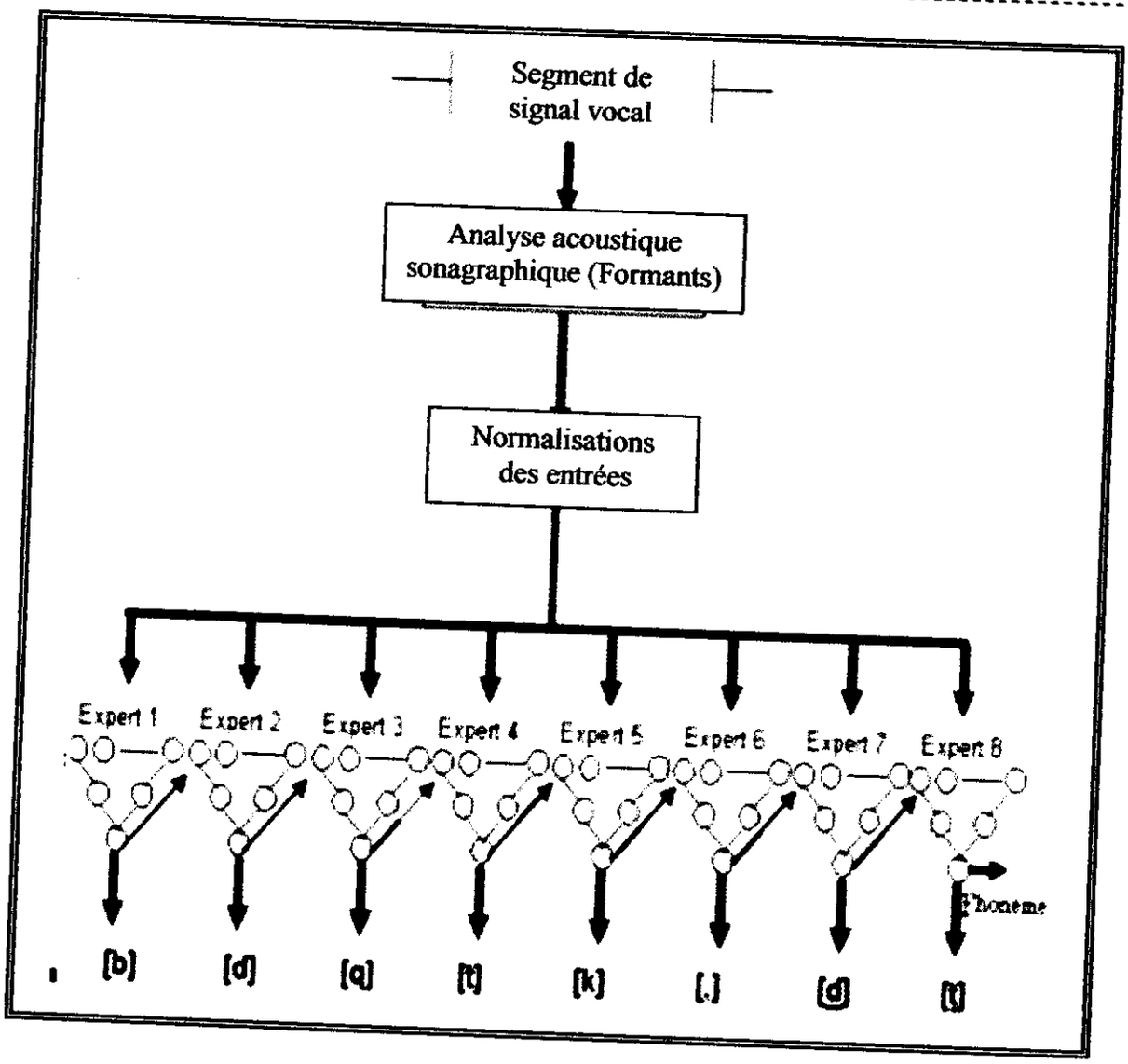


Figure 4.2 : Système neural modulaire pour la reconnaissance de phonèmes occlusifs orales de l'AS.

4.5. Phase apprentissage

Le nombre d'unités d'entrée est fixé à 4, celui de sortie à 1 pour chaque sous réseaux. Seul le nombre d'unités en couche cachée est indéterminé. Après l'extraction des paramètres pertinents des signaux acoustiques et l'initialisation des poids synaptiques par des valeurs comprises entre -0.5 et 0.5, un apprentissage supervisé est effectué sur tout le corpus d'apprentissage avec un pas de 0.01, pour déterminer le nombre optimal d'unités cachées. Le corpus est constitué de 179 phonèmes dont 80 sont utilisés pour l'apprentissage et 24 pour le test. Le processus d'apprentissage est arrêté dès que l'algorithme d'apprentissage n'est plus en mesure d'augmenter le taux de reconnaissance pendant l'apprentissage.

4.5.1. Algorithme d'apprentissage

Pour accomplir la tâche d'apprentissage il faudra passer par les étapes suivantes :

1. définir les variables : nombre de variables d'entrées, les couches cachées, pas d'apprentissage, nombre d'itérations, les poids synaptiques et la sortie désirée ;
2. lecture de données d'apprentissage ;
3. pré-accentuation des signaux ;
4. Initialisation aléatoire des poids w_i ;
5. Calcul de la sortie du réseau ;
6. Calcul d'erreur ;
7. Mise à jour des poids ;
8. calcul de Taux de Reconnaissance (TR) ;
9. arrêter l'apprentissage si le TR reste constant ;
10. sauvegarder les poids synaptiques w_1 et w_2 ;
11. fin.

4.6. Phase de reconnaissance

Lors de cette phase, les signaux acoustiques ont été traités de la même manière que lors de la phase d'apprentissage. Les vecteurs acoustiques obtenus sont injectés dans le système de test en faisant une discrimination entre les phonèmes à reconnaître. Le corpus de test est constitué de phonèmes à reconnaître en présence d'autres phonèmes pour mettre en jeu de possibilités des confusions entre phonèmes (exemple : [d] et [t]). Le processus de reconnaissance s'arrête si le phonème est détecté sinon si non le réseau (module) adjacent est activé. Si la base de données de test ne contient pas de phonèmes à reconnaître, le processus s'arrête sans qu'il y ait discrimination.

4.6.1. Algorithme de reconnaissance

Pour accomplir la tâche de reconnaissance il faudra passer par les étapes suivantes :

1. définir les variables : nombre de variables d'entrées, les couches cachées, pas d'apprentissage, nombre d'itérations, les poids synaptiques et la sortie désirée ;

2. lecture de données de reconnaissance (test) ;
3. pré-accentuation des signaux ;
4. calcul des formants ;
5. activation du premier réseau (module) ;
6. si le phonème est reconnu, afficher la réponse, sinon, activation du module adjacent ;
7. répéter l'étape 6 sur tout le corpus ;
8. calcul de TR final du phonème ;
9. fin.

4.7. Choix du langage de programmation

Dans toute branche de l'ingénierie, les outils communément disponibles jouent un rôle considérable. Parmi ces outils, le langage de programmation qui occupe une place sans doute importante dans le domaine d'informatique. D'une importance capitale pendant les phases de réalisation et la maintenance du logiciel, son choix devient par ce fait très délicat.

L'application a été développée sur un Pentium III, en utilisant le langage de programmation Builder *C++ 6.0 (sous Windows)*, ce choix est fait à base des critères suivants :

- le langage Builder *C++* est souple, modulaire et puissant ;
- programmation à base Orienté Objet ;
- permet la compilation séparée ;
- permet l'exécution rapide surtout dans le cas des calculs compliqués et itératifs le temps de réponse est réduit dans le cas de la recherche dans les grandes bases des données.

Pour quoi l'environnement Windows ?

On a développer notre application sous l'environnement Windows car il :

- offre une interface permet à l'utilisateur un usage facile de n'importe quel programme Windows ;
- propose un ensemble d'objets d'interfaces utilisateur composé de fenêtres, de menus et d'icônes, ce qui rend les applications Windows plus conviviales (se dit d'un matériel facilement utilisable par un public non spécialisé),et faciles à comprendre et à utiliser ;

- l'utilisation de logiciel de segmentation PRAAT exige la plateforme Windows dont le choix est plus convenable pour l'extraction des formants et l'exactitude des résultats de la segmentation.

4.8. Expériences et Résultats de la Reconnaissance

Nous avons traité trois cas, tel que l'un est indépendant de l'autre. Cette indépendance est en fonction dévaluations des nombres de mots utilisés pour l'apprentissage, et les mots utilisés pour la reconnaissance.

4.8.1. Première expérience

Nous utilisons pour l'apprentissage les 80 phonèmes du corpus (3 positions pour chaque phonème), et les 24 autres sont utilisés pour le test, qui sont prononcés par un même locuteur et au même endroit (tableau 4.2).

4.8.2. Deuxième expérience

Nous utilisons pour l'apprentissage 120 phonèmes du corpus (3 positions pour chaque phonème), et 24 parmi eux serons utilisés pour le test, qui sont prononcés par un même locuteur et au même endroit (tableau 4.2).

4.8.3. Troisième expérience

Les 24 mots utilisés pour le test sont pris dans la procédure d'apprentissage. En augmentant la taille de corpus de 40 phonèmes utilisés pour l'apprentissage à 80 après à 120, pour voir l'effet de cette dernière sur le Taux de Reconnaissance (TR). Les résultats sont introduits sur le tableau 4.2.

Phonèmes	Taux de reconnaissance %				
			3ème Exp		
	1ère Exp	2ème Exp	40 Phonèmes	80 Phonèmes	120 Phonèmes
[ب]	79.16	87.5	79.16	83.33	87.5
[د]	75.00	91.66	87.5	91.66	91.66
[ق]	70.83	87.5	79.16	83.33	87.5
[ت]	79.16	95.83	87.5	91.66	95.83
[ك]	75.00	95.83	87.5	95.83	95.83
[ء]	66.66	83.33	83.33	83.33	83.33
[ض]	70.83	91.66	87.5	87.5	91.66
[ط]	79.16	95.83	91.66	91.66	95.83
Taux de Reconnaissance Moyen %	74.47	91.14	85.41	88.53	91.14

Tableau 4.6 : Résultats obtenus des expérience lors de la reconnaissance.

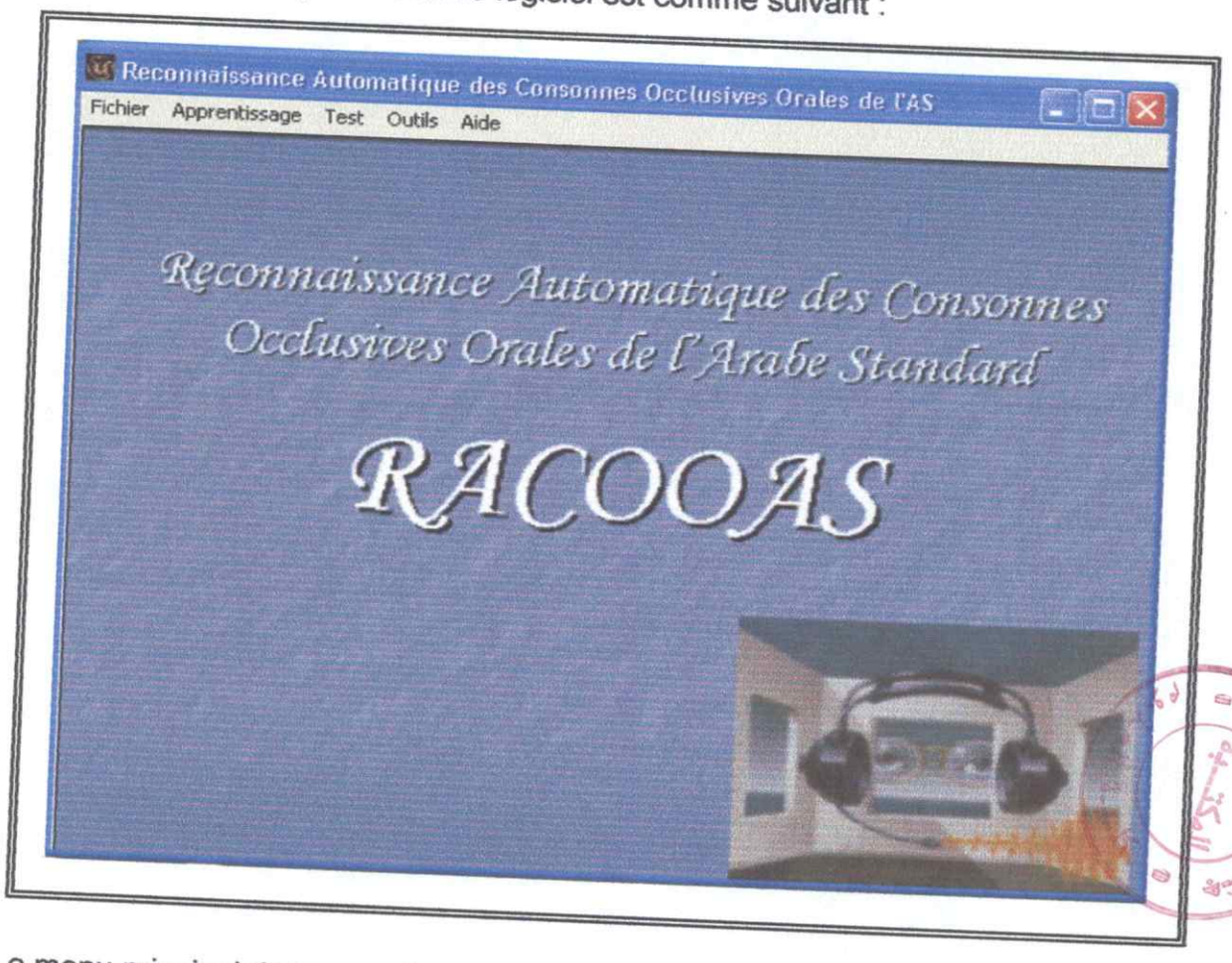
4.9. Interprétation des résultats

D'après les différentes expériences et les résultats obtenus nous avons constaté que :

- les résultats de reconnaissance obtenus montre une très nette amélioration du taux de reconnaissance qui a atteint 91.14% soit un taux d'erreur de 8.86% ;
- à partir des trois expériences, nous avons constaté que le TR augment lorsque les phonèmes utilisés pour l'apprentissage et la reconnaissance sont les mêmes, ainsi que, la taille du corpus d'apprentissage influe sur le TR. Pour 40, 80 et 120 phonèmes utilisés en apprentissage, il n'y pas le même TR, pour 120 mots est très élevé par rapport aux autres ;

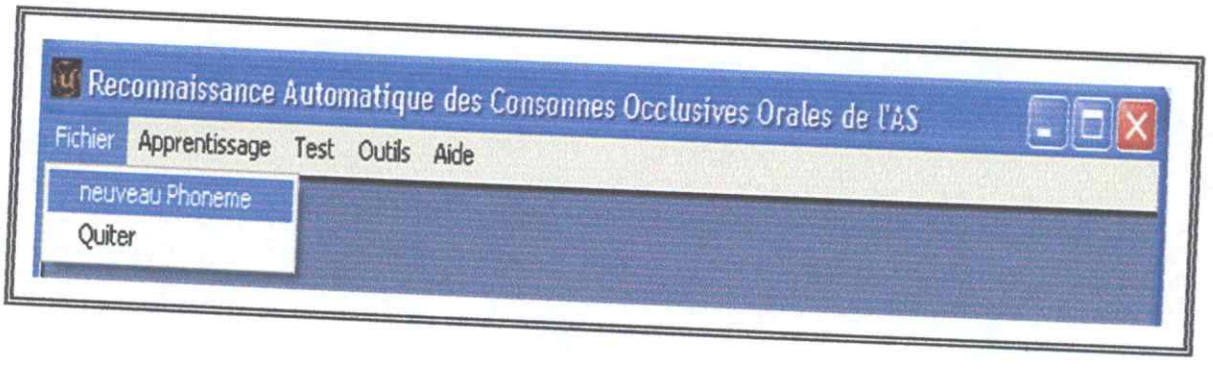
4.10. Description du logiciel RACOOAS

La fenêtre principale de notre logiciel est comme suivant :

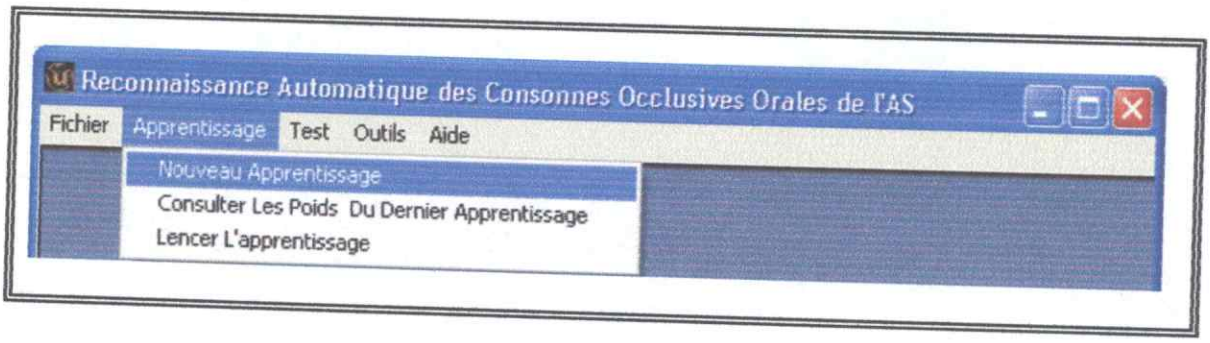


Le menu principal de la première fenêtre, permet d'accéder aux différentes tâches de notre logiciel. Par la suite nous allons détailler chaque sous menu.

Pour ouvrir l'analyseur PRAAT, enregistrer un nouveau phonème et le segmenter pour extraire les formants. Quitter RACOOAS.

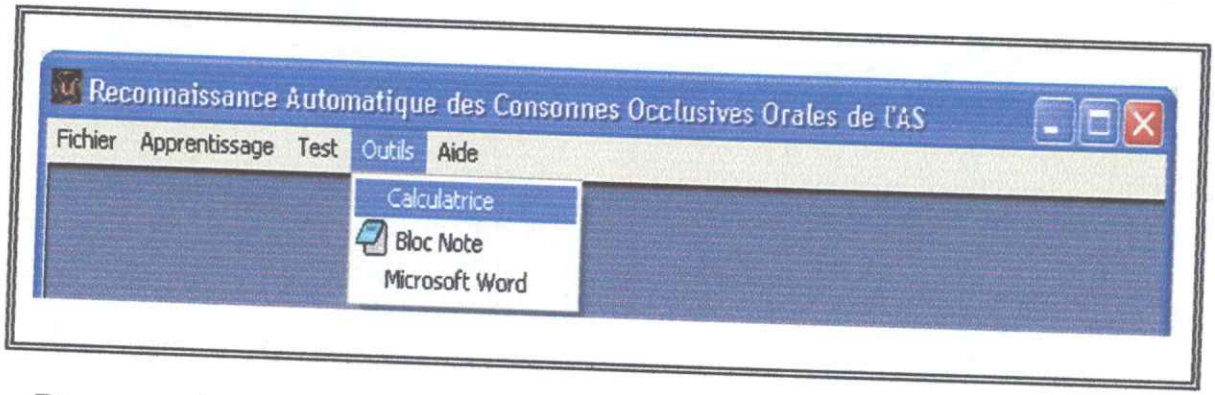


Faire lancer un nouvel apprentissage. Consulter les poids du dernier apprentissage. Lancer L'apprentissage en utilisant les anciens poids enregistrés dans la base de données.

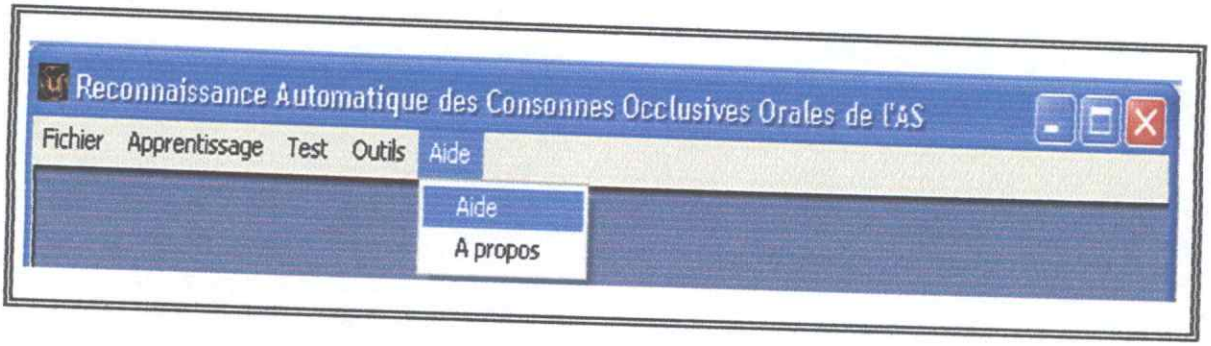


Ouvrir la page du test.

Les outils utilisés pour consulter les formant.

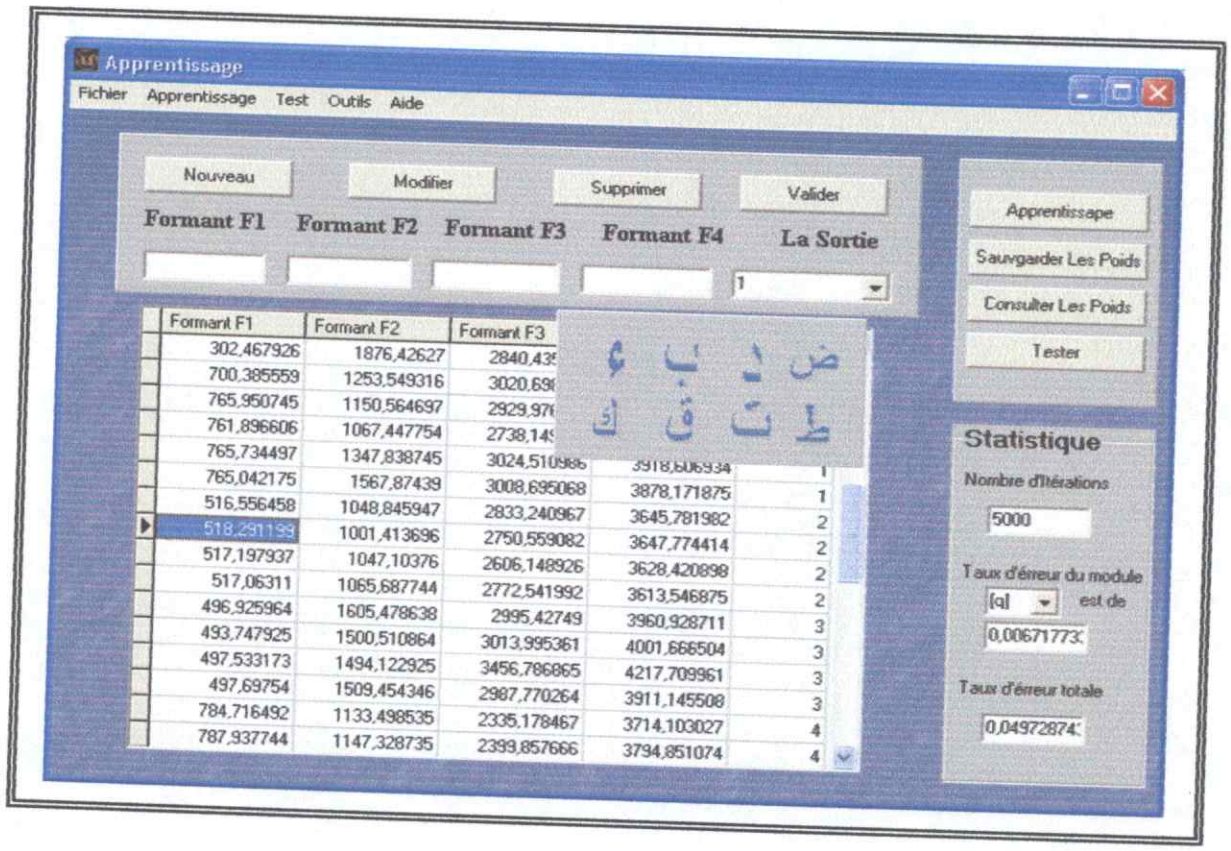


Pour connaître comment utiliser cette logiciel, et réaliser par qui.



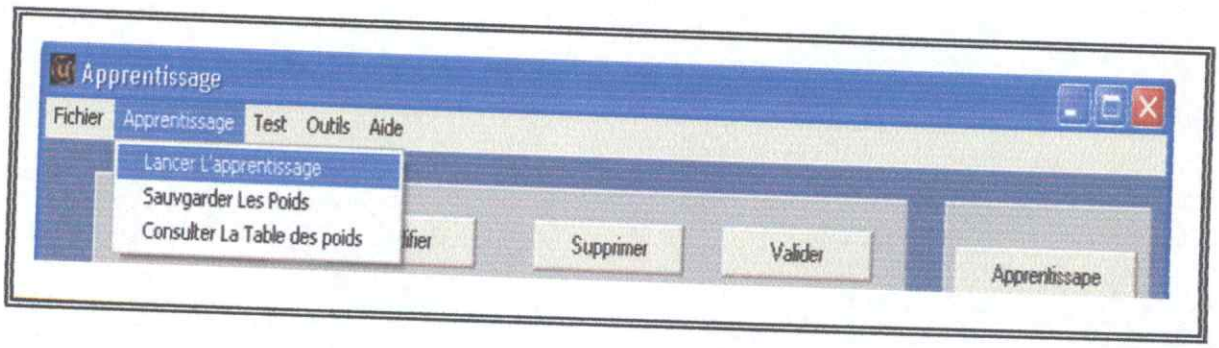
La fenêtre d'apprentissage, contient :

- la base de données de l'apprentissage avec toutes ses fonctions ;
- les statistiques concernant l'apprentissage.

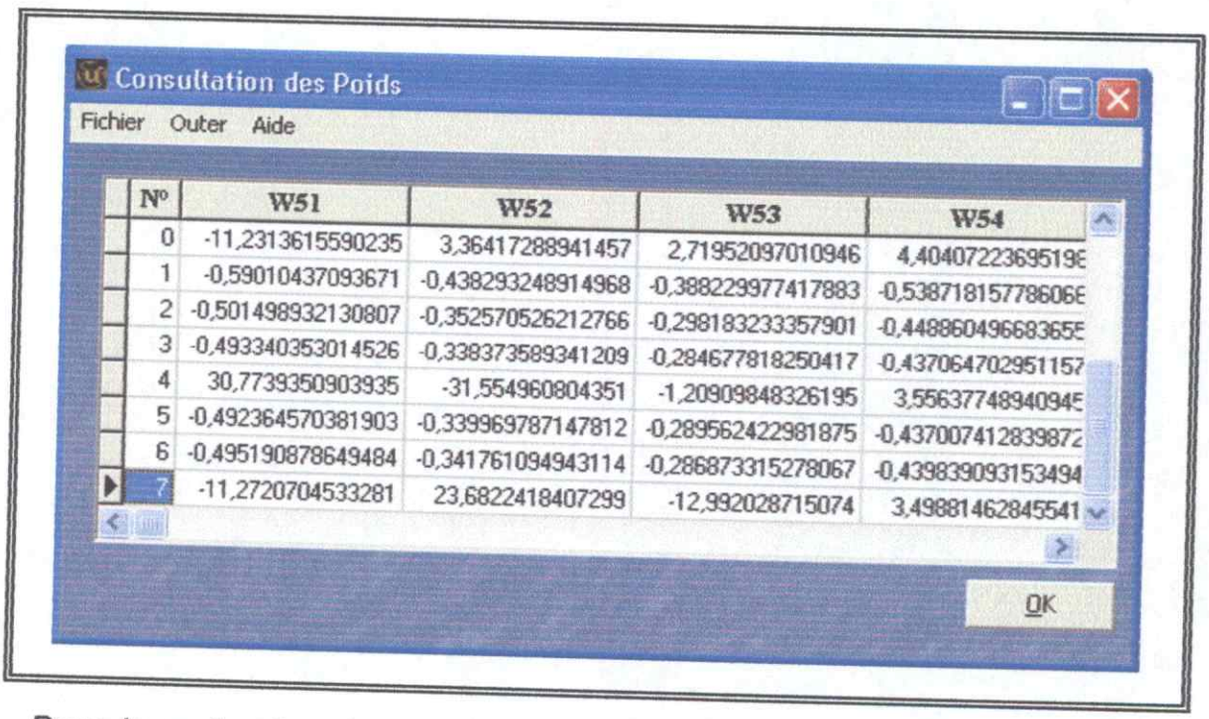


Le menu de cette page contient :

- lancer l'apprentissage à nouveau, après une modification dans la base de données des formants (ajouter, modifier, supprimer) ;
- sauvegarder les nouveaux poids obtenus de l'apprentissage ;
- consultation de la base de données des poids.

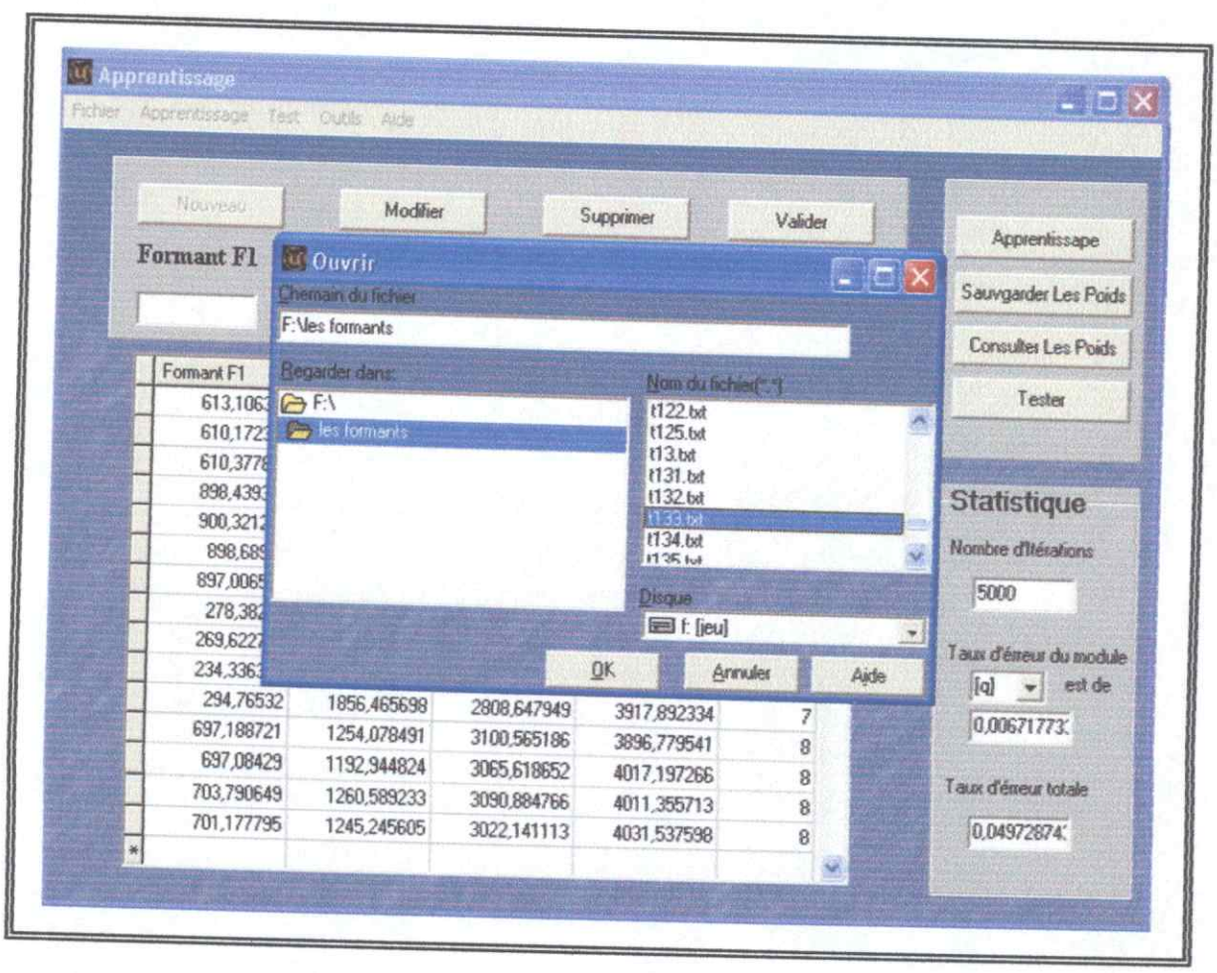


Consultation des poids.



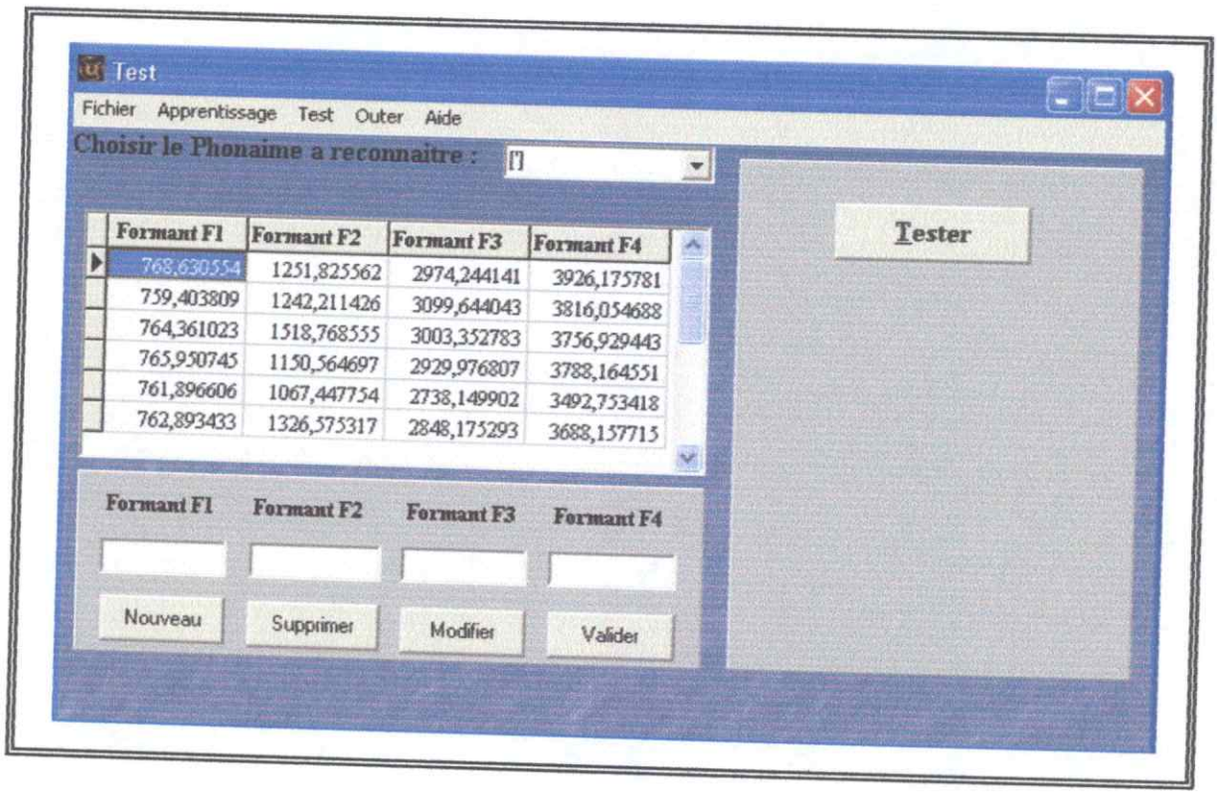
N°	W51	W52	W53	W54
0	-11,2313615590235	3,36417288941457	2,71952097010946	4,40407223695198
1	-0,59010437093671	-0,438293248914968	-0,388229977417883	-0,538718157786068
2	-0,501498932130807	-0,352570526212766	-0,298183233357901	-0,448860496683655
3	-0,493340353014526	-0,338373589341209	-0,284677818250417	-0,437064702951157
4	30,7739350903935	-31,554960804351	-1,20909848326195	3,55637748940945
5	-0,492364570381903	-0,339969787147812	-0,289562422981875	-0,437007412839872
6	-0,495190878649484	-0,341761094943114	-0,286873315278067	-0,439839093153494
7	-11,2720704533281	23,6822418407299	-12,992028715074	3,49881462845541

Pour la recherche des formants et le charger dans la base de données d'apprentissage.

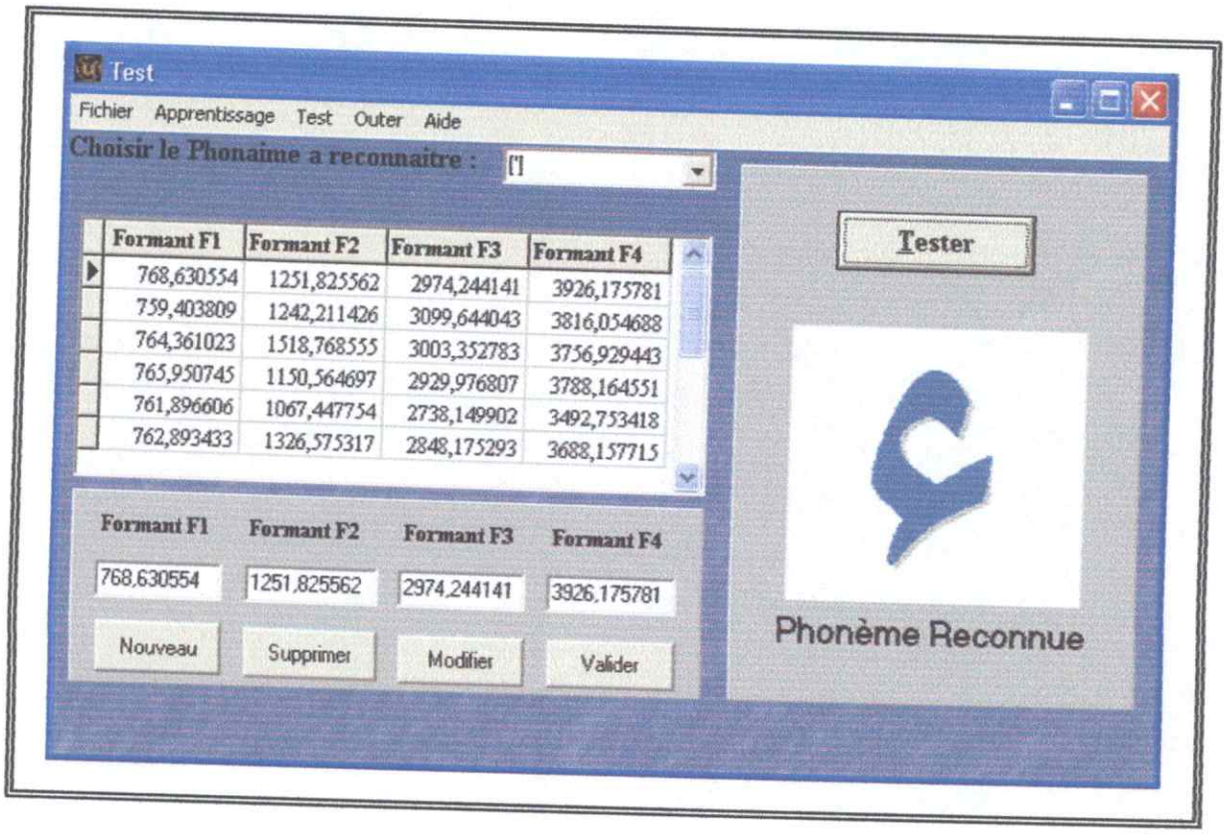


Formant F1				
613,1063				
610,1723				
610,3776				
898,4393				
900,3212				
898,685				
897,0065				
278,382				
269,6227				
234,3363				
294,76532	1856,465698	2808,647949	3917,892334	7
697,188721	1254,078491	3100,565186	3896,779541	8
697,08429	1192,944824	3065,618652	4017,197266	8
703,790649	1260,589233	3090,884766	4011,355713	8
701,177795	1245,245605	3022,141113	4031,537598	8

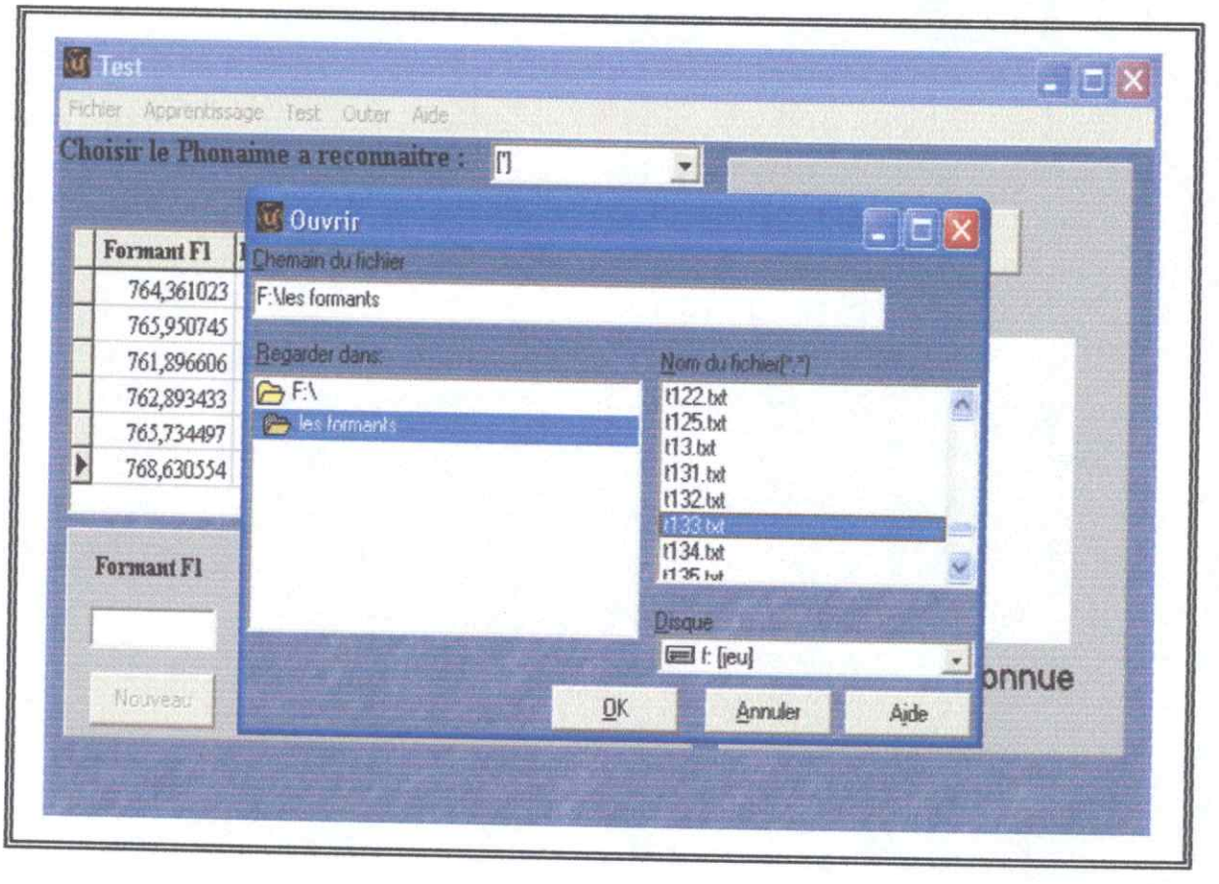
La fenêtre de test, contient la base de données des formants du test et un bouton pour le test.



Quand vous lancer le test il s'affiche le phonème s'il est reconnu, sinon il s'affiche un message « phonème non reconnu ».



Pour la recherche des formants et le charger dans la base de données du test.



4.11. Conclusion

Nous avons présenté dans ce chapitre un système de Reconnaissance des consonnes occlusives orales de l'Arabe standard en utilisant les Réseaux de Neurones de type MLP. Ce type de réseaux de neurones est caractérisé par une très grande capacité de discrimination, ce type qui a permis de l'avantager pour réaliser notre système de reconnaissance. Les phonèmes à reconnaître sont pris dans des mots enregistrés, et situés dans des positions différentes.

Lors de la conception de notre système, nous avons utilisé une analyse sonagraphique très robuste, représentative et discriminante.

Les résultats enregistrés nous ont permis de confirmer la grande souplesse des réseaux de neurones modulaires dans l'apprentissage et leur grande capacité de discrimination.

Conclusions Générales et Perspectives

Cette étude s'inscrit dans le cadre de Reconnaissance Automatique des Phonèmes Occlusifs Oraux de l'Arabe Standard.

Nous avons élaboré un corpus constitué de 120 mots isolés contenant les huit Phonèmes Occlusive Orales de l'AS dans différentes positions (initiale, médiane et finale).

A partir d'une analyse sonagraphique à l'aide du logiciel PRAAT nous avons extrait les formants qui sont considérés parmi les paramètres pertinents du signal vocal. La phase de segmentation est très difficile à cause de la variabilité du signal vocal (intra-locuteur, interlocuteur et contextuelle), cela nécessite de bonnes connaissances en phonétique et en phonologie.

Pour cela, nous avons utilisé la technique des réseaux connexionnistes modulaires, qui sont composés de sous réseaux de neurones multicouches MLP (Multi Layer Perceptron) auxquels nous avons attribué des sous tâches de reconnaissance. Le but cherché par un tel système connexionniste est la simplicité des réseaux et la facilité de leur apprentissage pour atteindre un taux de reconnaissance très élevé. Pour notre système le TR est égale à 91.14% avec un taux d'erreur d'apprentissage de 0.046%.

Enfin, il faut mentionner que le développement d'un système à base de réseaux de neurones est une tâche très délicate et qui nécessite beaucoup d'expériences. En effet de nombreux problèmes se posent concernant le choix et le dimensionnement du réseau, l'extraction des paramètres, le contrôle du système, etc.

Nous suggérons des perspectives à notre travail telles que :

- l'élaboration du corpus : corpus de grande taille en mode multi locuteurs. Le choix de ce type de corpus (parole continue) par le fait qu'il est préférable d'étudier les segments dans un continuum vocal pour pouvoir prendre en considération les effets de coarticulation existants entre les phonèmes ;
- l'utilisation de la technique d'analyse paramétrique pour extraire les paramètres acoustiques (LPC, PLP, RASTA-PLP) ;
- l'utilisation du modèle hybride pour dégager leurs performances ;
- l'implémentation de notre algorithme de reconnaissance déjà élaboré, pour une application en temps réel ;

- nous proposons de généraliser la reconnaissance pour les autres classes phonétiques de l'Arabe Standard, en tenant compte des traits phonétiques qui caractérisent cette langue telle que la gémination et l'allongement temporel.

Réferences Bibliographiques

ences Bibliograph

[01] A.Hadj Saleh, « *La notion de syllabe et la théorie ciético-impulsionnelle des phonéticiens arabes* », AL-Lissaniyat. Revue algerienne de linguistique. Vol.1, pp 63-83, I.L.P Alger, Algérie 1971.

[02] R.Boite, M.Kunt, « *Traitement de la parole, presses polytechniques Romandes* », PP 17-125, 1987.

[03] B.Mazoyer, J.Lautrey, P.Van Geert, « *Editions de la Maison des Sciences de l'Homme* », pp 193-206, Paris, France 2002.

[04] G.Marquis, « *Le codage de la parole en temps réel, à faible débit binaire* », Probatoire C.N.A.M., France 1985.

[05] A.Content, C.Meunier, R.Kearns, U.H.Frauenfelder, « *Sequence detection in pseudowords in French: where is the syllabe effect?* », Language and CognitiveProcesses, 16, 5/6, p.609-636, France 2001.

[06] K.Benbellil, « *Synthèse par polysons de l'Arabe Standard*», Mémoire de Magister en sciences du Langage et de la communication linguistique. ENSSH, Alger, ALGERIE 2003.

[07] J.Veronis, L.Khouri, « *Etiquetage grammatical multilingue : le projet ULTEXT. Traitement Automatique des Langues* », (36)1/2, 233-248.

[08] T.Moudenc, F.Emerard, « *Synthèse vocal et handicap, annales de télécommunications* », PP 928-934, 2004.

[09] H.Boudjettou, R.Djamel, « *Reconnaissance Automatique des Phonèmes Fricatives Non Emphatiques de l'Arabe Standard par les HMM* », Mémoire d'Ingénieur en Informatique, Option : IA, Université de Saâd Dahleb Blida, ALGERIE 2005.

[10] B.Jacob « *Un outil informatique de gestion de modèles de Markov caché : expérimentations en Reconnaissance Automatique de la Parole* ». Thèse de doctorat, l'Université Paul Sabatier de Toulouse III, France 1995.

[11] Calliope, « *La parole est son Traitement Automatique* », Collection technique et scientifique des Télécommunications, CNET/ENST, ed. Masson, Paris, France 1989.

[12] L.Bunet, « *Traitement Automatique de la Parole en milieu bruité : Etude de modèles Connexionnistes statique et dynamique* », Thèse du doctorat, l'Université Henri poincaré-Nancy 1, Spécialité informatique, 10 février 1997.

[13] C.Touzet, « *Les Réseaux de Neurones Artificiels. Introduction au Connexionnisme, cours, exercices plus travaux pratiques* », France 1992.

[14] M .Ait Akkache, « *Les Réseaux de Neurones séquentiels : Applications aux Modèles de Markov* », Thèse de Magister, Université de Blida, mars 1996.

[15] A.Tutin, G.Antoniadis, C. Clouzot, « *Corpus et TAL : Pour une réflexion méthodologique* », Conférence TALN 99, 12-17 Juillet 1999.

[16] K.Hornik, M.Stanchombe, H.White, « *Multilayer Feedforward Netwerk are Universal Approximators* », PP 359-366, 1989.

[17] L.Y.Boutoa, « *Reconnaissance de la Parole par Réseaux Multicouches* », Procedings of the international work stop on Neural and their applications, pp 197-217, 1988.

[18] M.Aissou, « *Application des Algorithmes Génétiques en vue de la Reconnaissance des voyelles de l'Arabe Standard* », Mémoire de Magister, ENSSH, Alger, ALGERIE 2004.

[19] L.Boutou « *Modèles Neuronaux et Hybrides Application en Reconnaissance de la Parole* ». Thèse de doctorat. Paris Sud, France 1991.

[20] F.Moutarde, « *Introduction aux Réseaux de Neurones* », Ecoles des Mines de Paris, France 2003.

[21] A.Spalanzani, « *Algorithmes évolutionnaires pour l'étude de la robustesse des systèmes de Reconnaissance Automatique de la Parole* », Thèse de Doctorat, Université Joseph Fourier – Grenoble, France 1999.

[22] R.Dugad and U.B. Desai, « *A Tutorial on Hidden Markov Models, Signal Processing and Artificial Neural Networks Laboratory Department of Electrical Engineering Indian Institute of Technology* », pp168-170, SPANN May 1996.

[23] L.Nguyen, B.Widrow, « *Approving the learning speed of Tow - Layer Neural Netwek by chousing initial values of the adaptation weights*», International Joint Conference Occidentale, 1991.

[24] M.Kabache, « Application des Réseaux de Neurones à la Reconnaissance Automatique des Phonèmes Spécifiques en Arabe Standard », Mémoire de Magister, ENSSU, Alger, ALGERIE 2005.

Références Inte

www.icp.inpg.fr/ICP/index.en/html

www.admi.net/avariste/100tc/D7-info.html

www.vieartificielle.com/articl/index.php

www.culture.gouv.fr/culture/dgle/rifal/enjeux.htm

www.grappa.univ-lille3.fr/plys/apprentissage/sortie008.htm

www.saturn.epm.ornal

www.evariste.org/100tc/1996/

www.robopolis.com/

www.irmcmaghreb.org/biblio/index.htm

www.unil.ch/

www.ph-ludwigsburg.de/franzoesisch/overmann.htm

www.sfu.ca/fren270/Phonetique/phonetique.htm

www.exuna.net/ipa.cgi

www.tsi.enst.fr/tsi/recherche/activite/2000-2001/fr/node17.html

www.hbroussais.fr/Broussais/InforMed/InforSante/Volume4/pdf4/4-8.pdf

Www3.ibm.com/software/speech/Le logiciel Via Voice d'IBM

www.speechrecognition.philips.com/index.php