République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université 8Mai 1945 – Guelma
Faculté des Sciences et de la Technologie
Département d'Electronique et Télécommunications

Mémoire de Fin d'Etude
Pour l'obtention du Diplôme de Master Académique

Domaine : **Sciences et Techniques**
Filière : **Electronique**
Spécialité : **Systèmes Electroniques**

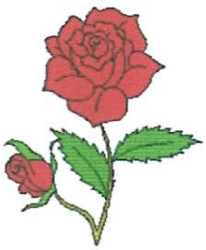
Reconnaissance vocale de chiffres par les réseaux de neurones

---

Présenté par :
- **GUEROUI Badreddine**
- **HEZIL Nabil**

Sous la direction de : **Dr. NEMISSI Mohamed**

JUIN 2011



Remerciements

Nous tenons à remercier en premier lieu Dieu le tout puissant, de nous avoir donné le courage et la patience de mener à bien notre projet.

Nos remerciements et nos profondes gratitudes à notre promoteur Dr. M.NEMISSI de nous avoir proposé ce sujet, de nous avoir encadré, pour ses remarques pertinentes et ses conseils judicieux qu'il nous a octroyé le long de notre travail.

Nous tenons à remercier chacun des membres de notre jury pour leur présence et leur participation lors de la soutenance

Nous remercions également nos professeurs qui tout au long de ces années d'étude ont transmis leur savoir sans réserve.

Sommaire

Liste des figures

Liste des tableaux

Liste des abréviations

FFT	Fast Fourier Transform (transformée de Fourier Rapide)
MLP	Multi Layered Perceptron
ZCR	Zero Crossing Rate (taux de passage par zéro)
GUI	Graphic Utilization Interface
TFD	Transformation de Fourier Discret
MFCC	Mel Frequency Cepstral Coefficients
RAP	Reconnaissance Automatique de la Parole
TZ	Transformation Z

Introduction générale

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de la nouvelle technologie. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, vient sans aucun doute du rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

Les techniques modernes de traitement de la parole tendent à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions :

- Les analyseurs de parole cherchent à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit
- Les reconnaisseurs ont 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 et la reconnaissance de la parole.
- Les synthétiseurs ont quant à eux la fonction inverse de celle des reconnaisseurs de parole : ils produisent de la parole artificielle.
- Enfin, le rôle des codeurs est de permettre 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.

On comprend aisément que, pour obtenir de bons résultats dans chacune de ces tâches, il faut tenir compte des caractéristiques du signal étudié vu la complexité de ce signal.

Ce travail rentre dans le cadre de la reconnaissance de la parole et le présent mémoire comporte trois chapitres :

Le premier chapitre constitue une introduction générale à la reconnaissance de la parole et présente les différentes étapes du traitement de la parole.

Le deuxième chapitre constitue une introduction aux réseaux de neurone et leur fonctionnement, ainsi que leurs algorithmes d'apprentissage.

Le troisième chapitre présente les résultats obtenus sur une petite base de données que nous avons élaborée l'interface graphique du système proposé.

Finalement, une conclusion générale conclue ce mémoire.

CHAPITRE 1

La reconnaissance automatique de la parole

1. Introduction

La parole est un signal réel, continu, à énergie finie et non stationnaire. Sa structure est complexe et variable avec le temps.

Le système auditif humain est surtout sensible dans une gamme de fréquence située entre 800 Hz à 8.000 Hz; les limites extrêmes sont respectivement 20 et 20.000 Hz. Par contre, le système vocal est encore plus limité, en résumé, pour des sons vocaliques à des fréquences au-dessus de 4 kHz, les hautes fréquences sont plus de 40 dB en dessous du sommet du spectre.

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique, et pragmatique

L'utilisation de la parole comme mode de communication avec une machine présente des avantage certains, notamment dans des situations qui nécessite souvent la reconnaissance et la synthèse de la parole :

- Utilisateurs ayant déjà les mains ou la vus occupées
- Utilisateurs occasionnels, non spécialistes d'un système
- Accès à distance : téléphones internet télématique vocale
- Utilisateurs handicapes [01].

2. Etude Phonétique

2.1.Production de la parole

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx ou la pression de l'air est modulée avant d'être appliquée au conduit vocal composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation.

L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou sourds). Les sons voisés (ou sonores) résultent au contraire d'une vibration périodique des cordes vocales. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire, elle détermine la hauteur du pharynx, le lieu d'articulation, ainsi que l'aperture, écartement des organes au point d'articulation; elle permet au conduit vocal d'avoir une géométrie et un volume extrêmement variable (figure 1.1) [02].

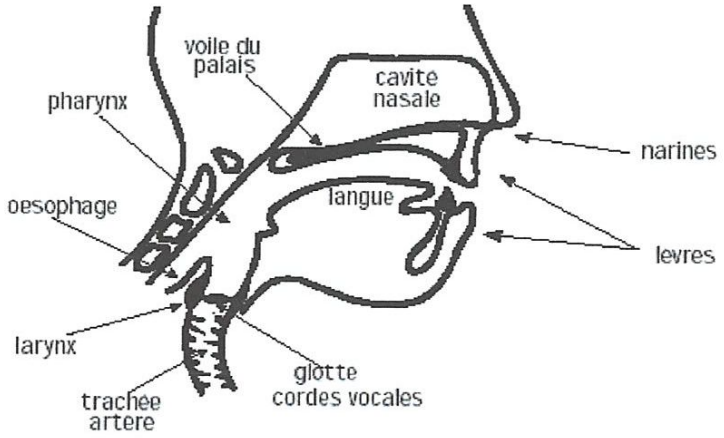


Figure 1.1: L'appareil phonatoire

2.2.Audition

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 qui en résulte.

L'appareil auditif comprend l'oreille externe, l'oreille moyenne, et l'oreille Interne.

Le mécanisme de l'oreille interne (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 (figure 1.2) [02].

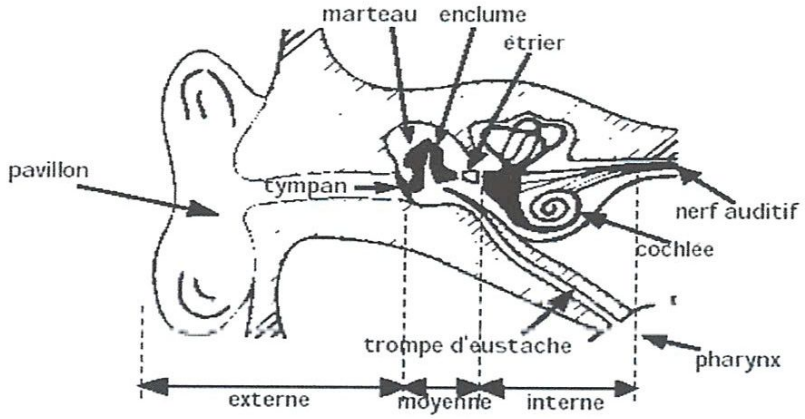


Figure 1.2: Le système auditif

Ainsi, l'oreille ne répond pas également à toutes les fréquences. La figure 1.3 présente le champ auditif humain, délimité par la courbe de seuil de l'audition et celle du seuil de la douleur. Sa limite supérieure en fréquence (16000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (32000 Hz).

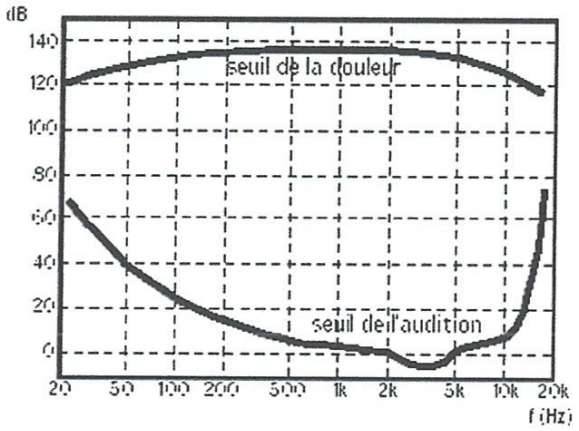


Figure 1.3: Le champ auditif humain

3. Caractéristiques du signal acoustique de la parole

Le signal acoustique de la parole est un peu particulier, il présente des caractéristiques qui rendent l'interprétation très complexe. Ce signal est variable d'un locuteur à un autre (variabilité interlocuteur), et pour le même locuteur (variabilité intra-locuteur). Nous ajoutons les variabilités dues aux conditions d'enregistrement.

3.1. Variabilité intra locuteur

La variabilité intra locuteur exprime les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie. Il existe un autre type de variabilité intra locuteur liée à la phase de production de la parole ou de préparation à la production de parole, due aux phénomènes de coarticulation (figure 1.4).

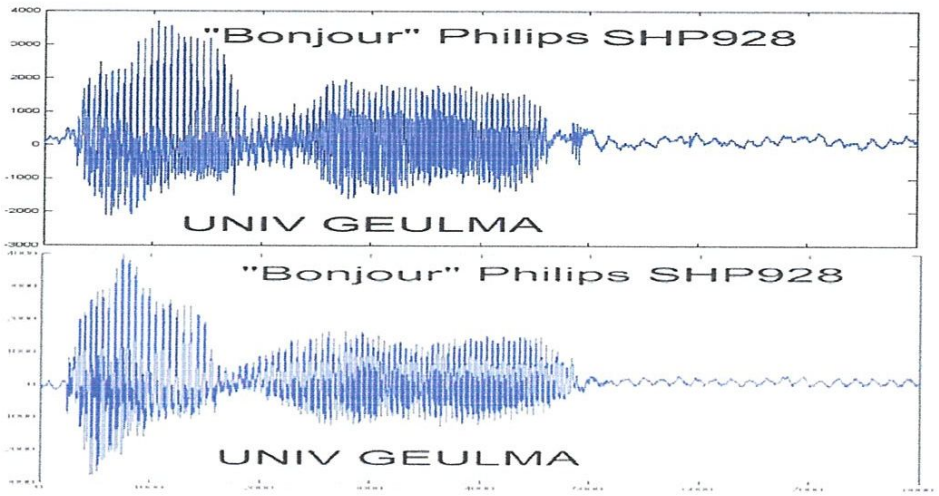


Figure 1.4: Variabilité intra locuteur

3.2.Variabilité interlocuteur

La variabilité interlocuteur est un phénomène majeur en reconnaissance de la parole.

La cause principale des différences interlocuteurs est de nature physiologique. La parole est produite par les vibrations des cordes vocales, qui déterminent l'importance et la forme du flux d'air s'échappant des poumons et amplifiées par les organes respiratoires, cette opération génère un son à une fréquence de base, le fondamental. Cette fréquence de base est différente d'un individu à l'autre et plus généralement d'un genre à l'autre ; une voix d'homme est plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible.

La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux (figure 1.5).

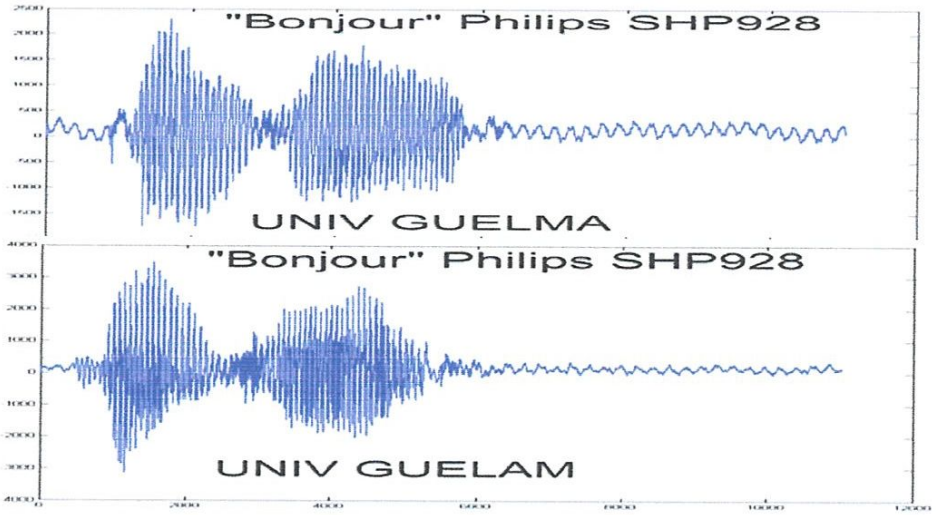


Figure 1.5: Variabilité interlocuteur

3.3.Variabilité due aux conditions d'enregistrement

Pour appliquer dans le commerce un système de reconnaissance vocal, il est important de connaître les effets de la transmission téléphonique sur un signal sonore.

La transmission de la parole par un canal téléphonique entraîne une limitation dans la gamme de fréquence, de 300 Hz à 3400 Hz de la bande passante.

La caractéristique de transfert n'est pas plate mais change de forme selon la ligne sélectionnée. Les spectres fournis par les lignes téléphoniques sont donc limités par la bande passante et également multipliés par une fonction de transfert de forme inconnue. Dans un premier stade, les études ont montré que la limitation des spectres de longue durée à la bande passante caractérisant la qualité du téléphone n'affecte pas sensiblement le taux d'identification.

Cependant, la pondération des spectres par des fonctions arbitraires du transfert, détruit la fiabilité de l'identification parce que, dans certains cas, l'effet de la fonction de transfert sur les spectres est plus important que les caractéristiques des voix (figure 1.6).

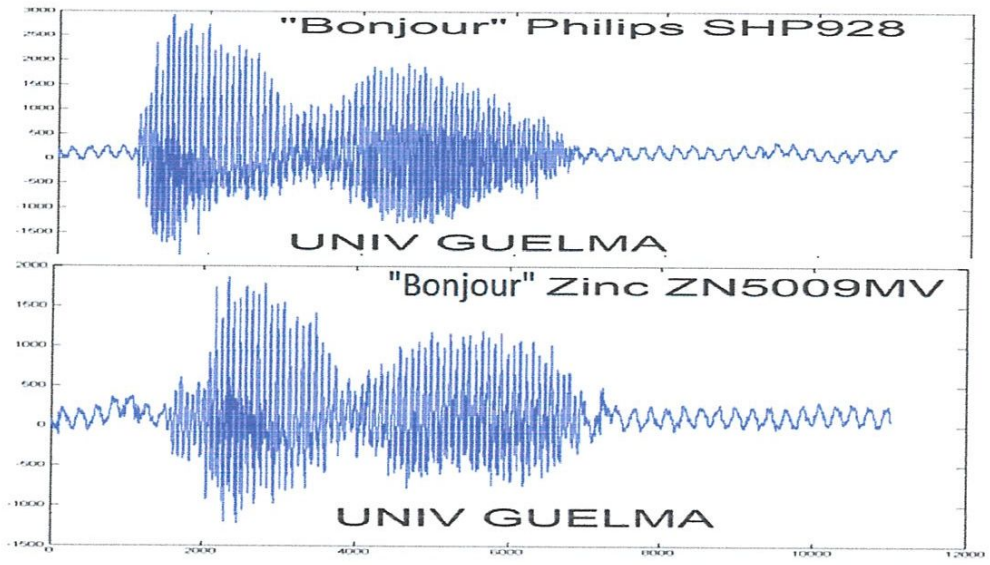


Figure 1.6: Variabilité due à l'enregistrement

4. Le Prétraitement de la parole

La reconnaissance automatique de la parole exige la réduction de la redondance du signal vocal à l'aide de traitements appropries, afin de diminuer les temps de traitement et l'encombrement en mémoire. Par ailleurs, le traitement du signal vocal permet d'extraire des paramètres pertinents, aussi invariants que possible, pour la reconnaissance [01].

L'objectif des prétraitements est de faciliter la caractérisation de la forme de l'entité à reconnaître en réduisant la quantité d'information à traiter pour ne garder que les informations les plus significatives [03].

Les dispositifs utilises ont d'abord été analogique. Avec l'évolution de l'électronique numérique et de l'informatique, les techniques numériques sont désormais généralisées. Apres numérisation du signal vocal, les traitements sont alors effectues par logiciel, initialement par des composants spécialisés et désormais par les micro-ordinateurs.

Un prétraitement d'un signal vocal est réalisé en trois étapes :

- Acquisition.
- Préaccentuation.
- Fenêtrage.

4.1.Acquisition

L'acquisition d'un signal vocal est liée avec la nature de l'application, et puisque on s'intéresse au traitement automatique (ordinateur), le signal de parole doit être numérisé.

De la sorte, on va pouvoir travailler sur une représentation spectrale du signal, décomposant ses différentes fréquences avec leurs amplitudes et leurs harmoniques, aboutissant à des « traits » qu'on appelle formants du signal.

La raison d'une analyse numérique est qu'elle est plus aisée pour un traitement sophistiqué et qu'elle est beaucoup plus fiable. Le développement rapide des ordinateurs et des circuits intégrés en conjonction avec la croissance des communications numériques a encouragé l'application des techniques numériques au traitement du signal.

La conversion analogique/numérique consiste en l'échantillonnage, la quantification et le codage.

L'échantillonnage est le processus de représentation d'un signal continûment variable comme une séquence de valeurs. La quantification conduit à représenter approximativement chaque échantillon dans un ensemble finit de valeurs.

Le codage consiste à assigner un numéro réel à chaque valeur .Avant l'échantillonnage, un filtre passe-bas de fréquence de coupure égale à la moitié de la fréquence d'échantillonnage est inséré pour éviter l'effet dénommé «repliement» ou « aliasing » postulé par le théorème de Nyquist-Shannon ; Ce filtre est donc appelé filtre « anti-repliement » ou « anti-aliasing».

Il y a deux paramètres qui affectent la qualité du son. Le premier est la fréquence d'échantillonnage (sampling rate) : On la mesure en Hertz (Hz) et des valeurs typiques pour le son sont 4 kHz, 8 kHz, 11.025 kHz, 22.05 kHz, 44.1 kHz et 48 kHz. Cependant d'après le théorème de Shannon, il faut choisir cette fréquence un peu plus grande que la moitié de la bande intéressante parce que les composants électroniques ne sont pas idéaux et qu'il est donc impossible de réaliser un filtre parfait. Le deuxième paramètre qu'affecte la qualité est la quantification en un nombre de bits fixé. Typiquement, ce nombre varie entre 8, 12, 14 ou 16 bits et détermine la dynamique et le rapport signal à bruit. Généralement, il s'agit d'une représentation uniforme mais une amélioration peut être obtenue avec des quantifications non linéaires [04].

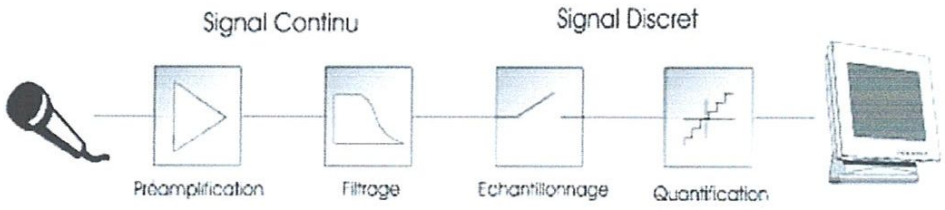


Figure 1.7: Acquisition

4.2.Préaccentuation

En général, le signal vocal se caractérise par une perte de 6 dB/Octave, due à l'influence de la source d'excitation et au rayonnement des lèvres. Une perte de 6 dB/Octave veut dire que les hautes fréquences ont une énergie plus faible que celle des basses fréquences. Pour pallier à cet inconvénient la préaccentuation permet d'égaliser les sons aigus avec les sons graves (figure 1.8).
L'opération consiste à faire passer le signal à travers un filtre de transmittance :

$$H(Z) = 1 - \mu * Z^{-1} \quad \text{où} \quad 0 \leq \mu \leq 1 \qquad (1.1)$$

Dans le domaine des signaux discrets (*échantillonnés*) *S(n)* le problème consiste habituellement à calculer :

$$Y(n) = X(n) - \mu * X(n-1) \quad \text{pour} \quad n \geq 0. \qquad (1.2)$$

Le facteur de préaccentuation est pris entre 0.9 et 1 (souvent 0.95). Comme conséquence, la préaccentuation introduit une légère distorsion spectrale.

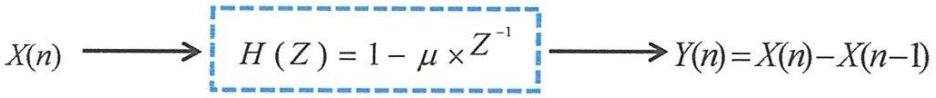


Figure 1.8: La préaccentuation

4.3.Fenêtrage

Le principe de cette phase de traitement est de minimiser la déformation du spectre dans les hautes fréquences, due au découpage en trames imposé au signal de parole, en employant une fenêtre dans le domaine temporel qui réduit progressivement l'amplitude du signal au commencement et à la fin de chaque trame [05].
Le signal vocal est un signal non stationnaire. Il présente une évolution lente dans le temps.
Le but du fenêtrage est de découper le signal de parole en petites tranches (chacune de durée 30ms environ) où il peut être considérer localement comme quasi- stationnaire.

En outre, et pour profiler de l'évolution lente du signal vocal, le fenêtrage permet le traitement en temps réel et il facilite aussi l'analyse des signaux sur la machine. Les ressources d'une machine étant limitées, le signal ne peut pas être traité dans sa globalité.

Du fait que le signal de parole est non-stationnaire il faut mettre en évidence les dépendances qui existent entre les échantillons, pour se faire, deux tranches adjacentes doivent avoir une région commune (de 5 à 12 millisecondes environ).

Il existe plusieurs types de fenêtres d'analyse. On en présente quelques-unes :

- **Fenêtre rectangulaire :**

Elle est définie par : $f(nT) = \begin{cases} 1 \text{ si } |nT| < T'. \\ 0 \text{ ailleurs.} \end{cases}$

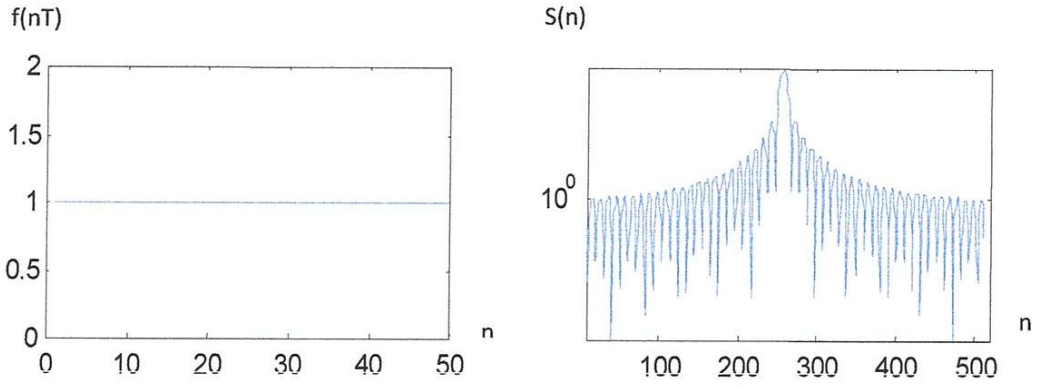


Figure 1.9: La fenêtre rectangulaire et son spectre

$T_e = 1 / f_e$ est la période d'échantillonnage.

T' est la moitié de longueur de la fenêtre.

- **La fenêtre de Hamming**

$$f(nT) = \begin{cases} 0.54\text{-}0.46\cos(\pi.nT/T') & \text{si } |nT|<T' \\ 0 & \text{ailleurs} \end{cases}$$

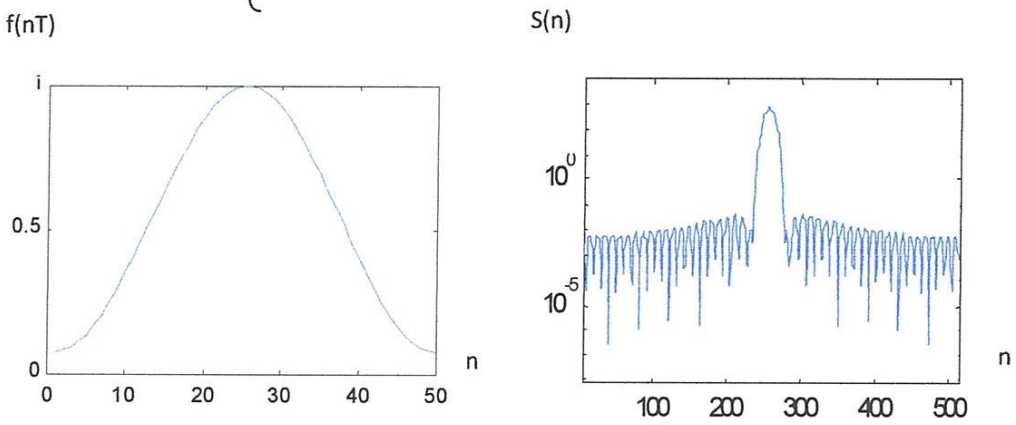


Figure 1.10: La fenêtre de Hamming et son spectre

- **La fenêtre de Hanning**

$$f(nT) = \begin{cases} 0.5 - (1 - \cos(n\pi T/T')) & \text{si } |nT| <T' \\ 0 & \text{ailleurs} \end{cases}$$

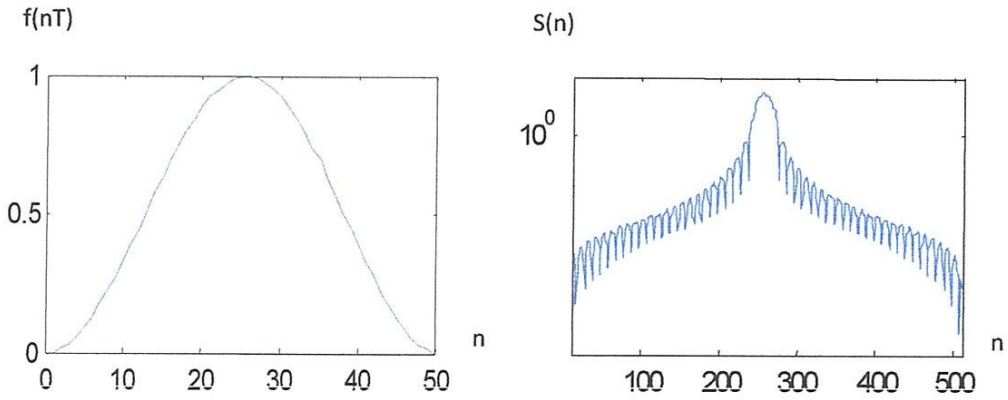


Figure 1.11: La fenêtre de Hanning et son spectre

Parmi ces fenêtres, la fenêtre de Hamming est la plus convenable à la parole, car elle entraîne un minimum de distorsion spectrale du signal de parole, par rapport aux autres fenêtres. (Atténuation du rapport du lobe principal au lobe secondaire est égale à –41 dB, c'est à dire que la concentration de l'énergie dans le lobe principal est égale à 99.96%).

5. Analyse et traitement d'un signal vocal

Plusieurs approches ont été proposées pour l'analyse de la parole, ayant toutes pour but d'extraire le minimum d'information pouvant définir complètement le signal de parole. Parmi ces approches il y a celles qui agissent dans le domaine fréquentiel, celles qui agissent dans le temps, d'autres permettant d'avoir une analyse conjointe temps-fréquence, et finalement les méthodes basées sur la modélisation du système des phonations.

5.1.La fréquence fondamentale

La période du fondamental est par définition la fréquence de vibration des cordes vocales. Elle est aussi appelée (pitch), c'est un paramètre très important dans les différentes applications de la parole [06].

L'opération d'extraction du pitch est une tâche difficile pour trois raisons principales:

- La vibration des cordes vocales n'a pas nécessairement une périodicité complète.
- Il est difficile de séparer le pitch des effets du trait vocal.
- La plage de la fréquence du fondamental est très grande. Elle s'étend approximativement de : 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes et de 200 à 600 Hz chez les enfants.

5.2.Analyse spectrale

Elle est fondée sur une décomposition du signal sans connaissance à priori de sa structure fine ou de sa source. Dans ce domaine, on trouve principalement la TFD, et le vocodeur à canaux.

5.2.1. La transformation de Fourier

Lorsqu'on veut analyser une entité complexe, une des procédures largement utilisées consiste à la décomposer en une somme d'entités plus simples. L'idée donc est d'exprimer le signal vocal par une combinaison linéaire discrète de fonctions élémentaires de forme simple [01] ; c'est le cas de la transformée de Fourier discrète pour l'estimation du spectre.

La TFD est définie par :

$$S(n) = \sum_{k=0}^{N-1} s(k) \times e^{-j\pi\frac{nk}{N}} \qquad (1.3)$$

S(k) est un signal numérique.

N est la longueur du support du signal.

Pratiquement la TFD est évaluée par un algorithme rapide appelé FFT (*Fast Fourier Trasform*). Elle s'opère sur des durées limitées du signal vocal, en prélevant les échantillons de parole à l'aide d'une fenêtre temporelle glissante, En général les fenêtres successives se recouvrent.

Ces fenêtres doivent avoir une largeur si l'on veut que la FFT ait un sens : En général, on prend 256 à 512 points, le recouvrement est par exemple la moitié soit 128 ou 256 respectivement.
La séquence des opérations est schématisée sur la figure 1.12.

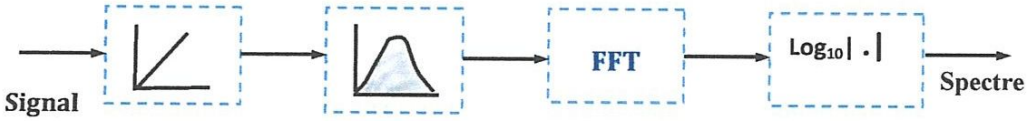


Figure 1.12: Traitement par transformé de Fourier

5.2.2. L'échelle des Mels

L'échelle des Mels est une échelle biologique permettant la modélisation de l'oreille humaine. L'échelle des Mels permet de modéliser une perception de l'oreille linéairement avant 1000 Hz puis logarithmiquement.

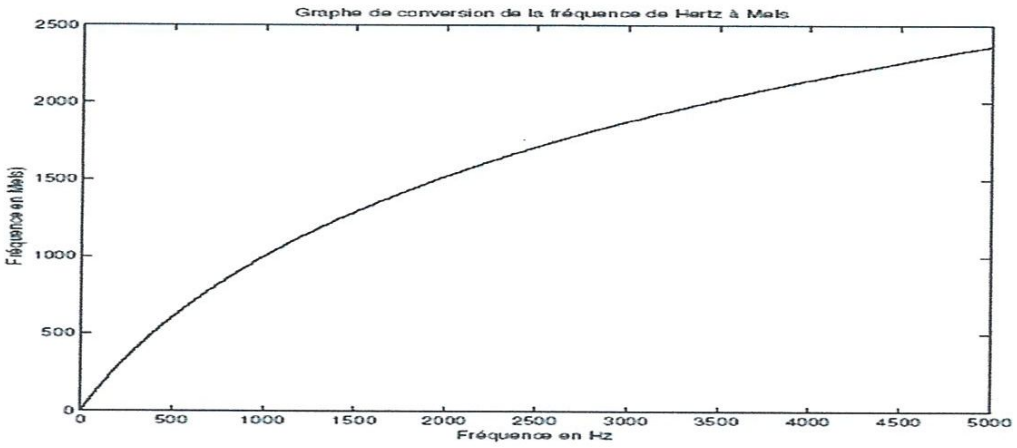


Figure 1.13: Graphe de conversion de la fréquence de Hertz à Mels

On remarque qu'avant 1000 Hz, la courbe est à peu près droite, ce qui traduit bien l'équivalence entre Hz et Mels à ces fréquences.

5.2.3. Banc de filtres Mels

Du fait que l'étendue des fréquences présentes dans le spectre est encore très large (par conséquent beaucoup de données à traiter), on a recours au banc de filtres dans l'échelle de Mels. On relie ainsi le système de reconnaissance vocale au fonctionnement de l'oreille humaine. Il s'agit de filtres passes bandes (fonction fenêtre de Hamming) centrés linéairement dans le domaine fréquentiel des Mels et de largeur telle qu'ils divisent l'espace des fréquences de manière égale dans le domaine des Mels et qu'ils se recouvrent chacun par moitié. Les filtres sont donc disposés logarithmiquement dans l'échelle des fréquences usuelles (Hz) : on a ainsi beaucoup de filtres pour les basses fréquences alors que les hautes fréquences sont disposées plus largement.

La formule donnant la fréquence en Mels à partir de la fréquence en Hz est donnée par:

$$m = 2595.\log_{10}\left(1+\frac{f}{700}\right) \qquad (1.4)$$

Ou f est la fréquence en Hertz.

Chaque filtre donne un coefficient cepstrale :

$$S_{i,k} = \sum_{n=0}^{N/2} Y_{i,n} M_{n,k}, \quad k = 0...K \qquad (1.5)$$

Avec : $Y_{i,n}$ le $n^{ème} \in [1, N]$ coefficient de la transformée de la $i^{ème} \in [1, I]$ frame, et $M_{n,k}$ le $n^{ème} \in [1, N]$ coefficient du $k^{ème} \in [1, K]$ filtre. On utilise communément 12 coefficients, on utilise alors K=13 filtres (pour obtenir 12 coefficients, il faut un filtre de plus car le $0^{ème}$ est inutile).

On a donc $S_{i,k}$ la matrice de sortie du $k^{ème}$ filtre pour la $i^{ème}$ frame. On a, à cette étape, ce qu'on appelle un Spectre Mel (Spectrum Mel).

Une représentation d'un banc de filtres dans l'échelle des Mels est donnée par la figure 1.14 : c'est une représentation de la matrice N x K des coefficients du banc de filtres (N : Nombre d'échantillons par frames, K nombres de filtres souhaités).

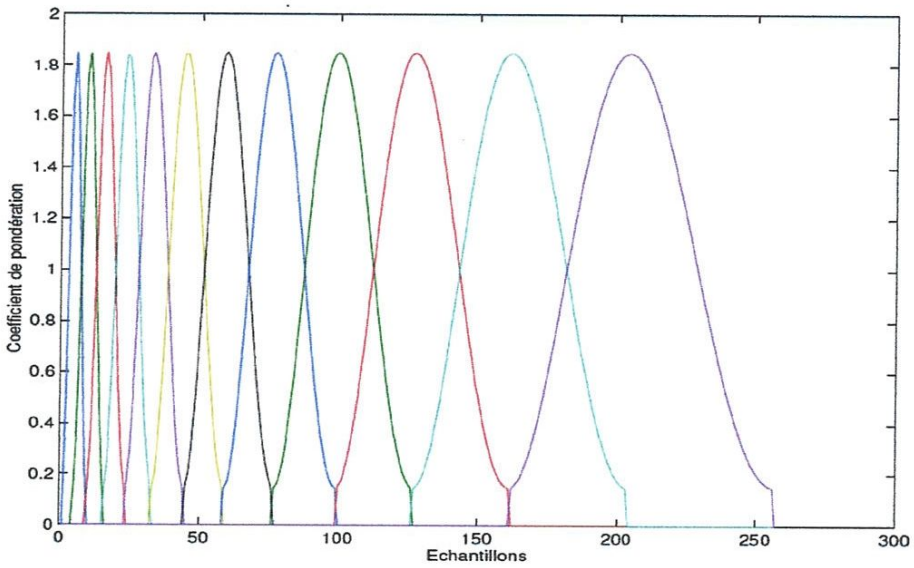


Figure 1.14: Représentation d'un banc de filtres dans l'échelle des Mels

5.3.Analyse temporelle

5.3.1. Energie totale

Elle est évaluée par :

$$E = \frac{1}{N}\sum_{k=0}^{N-1} s^2(k) \qquad (1.6)$$

N est la longueur de la fenêtre.

$S(k)$ est le signal à traité

Elle joue un rôle important pour délimiter les mots, et la localisation des voyelles qui sont caractérisées par une forte énergie par rapport aux autres phonèmes [07].

5.3.2. La densité de passage par zéro ZCR (Zero Crossing Rate)

ZCR correspond au nombre de changements de signe d'un échantillon à son successeur, dans une même trame [07].

Ce critère consiste à compter le nombre de passage par zéro pour des trames identiques à celles définies précédemment en commençant du début de l'enregistrement, respectivement de la fin. Si ce nombre dépasse un certain seuil, calculé expérimentalement à partir d'échantillons de silence, alors on est en présence de parole ; début du mot, respectivement fin du mot.

$$ZCR_x(m) = \frac{1}{L}\sum_{n=m-L-1}^{m} \left|sign\,(x(n)) - sign\,(x(n-1))\right| \qquad (1.7)$$

Avec :$sign(x(n))=\begin{cases} 1 \text{ si } x \geq 0 \\ 0 \text{ si } x < 0 \end{cases}$

L : longueur de trame.

m : indice de trame.

n : indice des échantillons.

Elle est utilisée pour distinguer le signal de parole du silence

5.4.Analyse Homomorphique (Cepstrale)

Le défaut majeur de la FFT pour le calcul du spectre vocal, réside dans l'intermodulation source/conduit qui rend difficile la mesure des formants et du fondamental. L'analyse cepstrale est une méthode qui vise à séparer leurs contributions respectives par déconvolution. Pour cela on fait l'hypothèse que le signal vocal x(n) est produit par un signal excitateur g(n) (source glottique) traversant un système linéaire passif de réponse impulsionnelle h(n) (conduit vocal).

D'après ces hypothèses, on aura le système suivant:

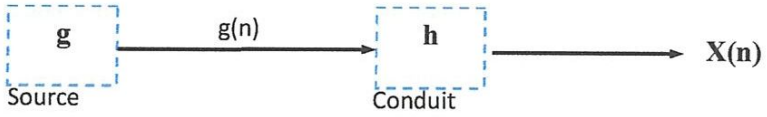


Figure 1.15: modèle de signal vocal x(n)

Donc on peut écrire pour tout n >0 :

$$x(n) = g(n) * h(n) \qquad (1.8)$$

Pour déconvoluer x(n), c'est à dire pour retrouver les deux composantes g(n) et h(n), il faut se donner une classe de fonctions admissibles pour g(n) (ou pour x(n)) : ici on suppose que g(n) est une séquence d'impulsions (périodique pour les sons voisés). Il est évident que l'ensemble de ces hypothèses est très limitatif : en toute rigueur cette analyse ne s'applique théoriquement qu'aux parties stables des sons périodiques (voyelles longues, par exemple), dans la pratique, cependant, cette méthode fournie des résultats acceptables sur l'ensemble du signal [08].

Pour déconvoluer plus aisément x(n), il suffit de transposer le problème par homomorphisme dans un espace où l'opérateur de convolution « * » correspond à un opérateur d'addition « + ». Soit D_* cet homomorphisme.

D_* est un homoimorphisme (application) qui applique l'espace vectoriel des signaux d'entrées muni de la loi « * » (convolution), sur l'espace vectoriel des signaux de sortie

muni de la loi « + » (addition), donc on est en face de la situation suivante :

$$x(n)=g(n)*h(n) \xrightarrow{D_*^+} \hat{x}(n)=\hat{g}(n)+\hat{h}(n) \qquad (1.9)$$

Après séparation de $\hat{g}(n)$ et de $\hat{x}(n)$ et si la transformation inverse D_+^* existe, on aura :

$$\begin{aligned} \hat{g}(n) &\xrightarrow{D_+^*} g(n) \\ \hat{h}(n) &\xrightarrow{D_+^*} h(n) \end{aligned} \qquad (1.10)$$

L'intérêt de la méthode réside dans le fait que $\hat{g}(n)$ et $\hat{h}(n)$ sont facilement séparables par un filtrage temporel est ceci grâce à l'hypothèse simplificatrice sur g(n).

Les homomorphismes D_*^+ et D_+^* sont inverses l'un de l'autre, et se définissent par :

$$D_*^+ = TZ(.) \circ \log(.) \circ TZ^{-1}(.) \qquad (1.11)$$

$$D_+^* = TZ(.) \circ \exp(.) \circ TZ^{-1}(.) \qquad (1.12)$$

Ce qui donne le système schématisé dans la figure 1.16 :

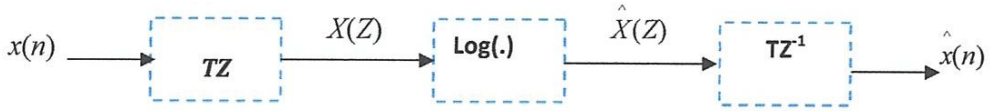


Figure 1.16: Calcul du cepstre complexe

Où :

- TZ est la transformée en Z (TZ^{-1} sa transformée inverse).
- La fonction log est utilisée pour le passage du domaine de la loi « . »(La multiplication) au domaine de la loi « + » (l'addition

5.4.1. Ambiguïté de la phase

Le problème qui se pose ici est que Arg [X(Z)] n'est défini qu'à 2π prés (la valeur principale), c'est à dire que l'on peut ajouter un multiple entier de 2π à la partie imaginaire du log complexe sans changer le résultat.

Ceci montre que l'homomorphisme tel qu'il est défini n'est pas une transformation biunivoque ; Pour contourner ce problème, on a introduit la notion du cepstre réel.

5.4.2. Définition du cepstre réel

La difficulté du logarithme complexe (à cause de la phase) peut être levée dans le cas de la parole (où l'on ne s'intéresse que rarement à l'information de phase) en prenant un log module ($\log|.|$), ce qui garantit dans 3 son inversibilité sans calcul particulier de la phase.

Soit DM_*^+ cet homomorphisme et DM_*^- son inverse :

$$DM_*^+ = TZ(.) \circ \log|.| \circ TZ^{-1}(.) \quad (1.13)$$

$$DM_*^- = TZ(.) \circ \exp|.| \circ TZ^{-1}(.) \quad (1.14)$$

Les coefficients du cepstre réel sont définis par :

$$C_n = DM_*^+ [x(n)] \quad (1.15)$$

Ces coefficients sont réels, ils conservent le spectre d'amplitude, par contre l'information de phase est perdue.

En pratique, on peut remplacer avantageusement la transformée en Z par une transformation de Fourier rapide (FFT), celle-ci possède les même propriétés de linéarité que la transformée en Z (figure 1.17).

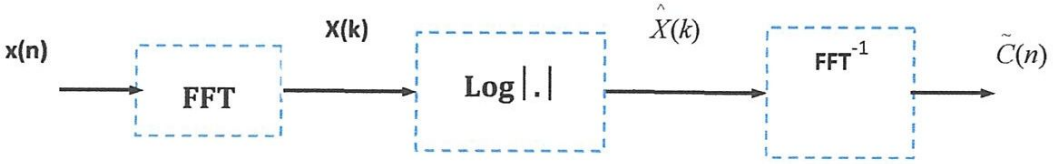


Figure 1.17: Le calcul des coefficients cepstraux réels

Dans ce cas les coefficients $\tilde{C}_n$ sont donnés par :

$$\tilde{C}_n = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)| e^{jk12\pi/N} \quad (1.16)$$

Sous certaines conditions, on peut admettre que $\tilde{C}_n \cong C_n$ donc on peut dire que les $\tilde{C}_n$ sont les coefficients cepstraux approchés prenant leurs valeurs dans un domaine pseudo temporel appelé domaine QUEFRENTIEL [09].

5.4.3. Coefficients MFCC (Mel-scaled Frequency Cepstral Coefficients)

Les paramètres MFCC sont des coefficients cepstraux obtenus à partir des énergies d'un banc de filtre en échelle de fréquence Mel. Il s'agit en fait d'un calcul classique des coefficients cepstraux auquel on a rajouté, avant le logarithme un filtre de Mel. Ces résultats sont intéressants, car le calcul d'une dizaine de coefficients cepstraux est alors suffisant pour des expériences de RAP [09].

$$MFCC_i = \sum_{k=1}^{20} Xk \cos \pi_i \frac{(k-0.5)}{20} \qquad (1.17)$$

Avec i=1,2,..... p ; 20 est le nombre de filtre et p est le nombre des coefficients.

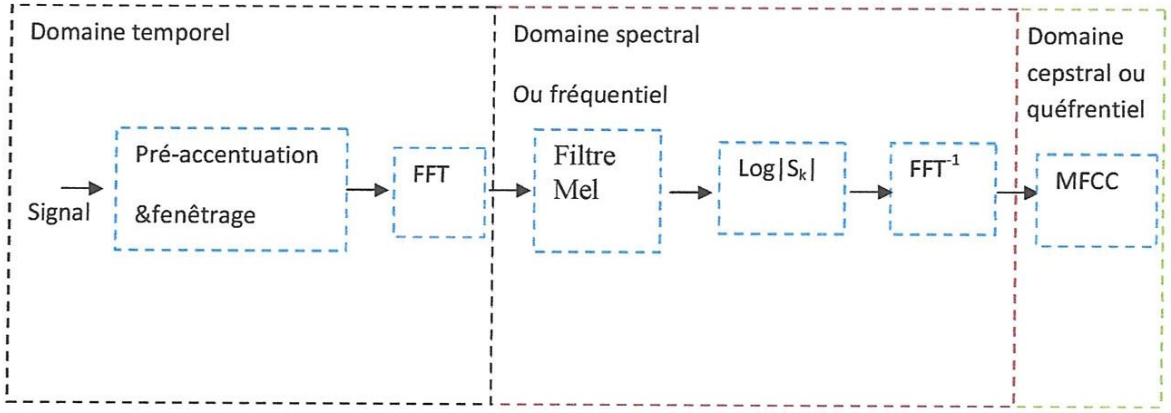


Figure 1.18: calcul des coefficients MFCC

6. Conclusion

Dans ce chapitre nous avons donné un aperçu sur la reconnaissance automatique de la parole. Dans ce travail nous réalisons une extraction des caractéristiques donnée par coefficients MFCC :(Mel-scaled Frequency Cepstral Coefficients) qui sont les plus utilisés dans ce domaine.

CHAPITRE 2

Réseaux de neurones

1. Introduction

Les dernières années ont vu un développement technologique puissant dans des domaines divers, et il y a eu un accroissement de besoin pour le contrôle et la gestion des systèmes complexes qui introduisent d'énorme calculs et un nombre de variables important ; d'où la nécessité de chercher de nouvelles méthodes pour une gestion plus souple et moins coûteuse en temps de calculs et en manipulation des variables dont le nombre ne cesse d'augmenter. Pour cela, on s'est intéressé de plus en plus aux systèmes qui apprennent, en utilisant des modélisations des neurones biologiques.

Les modèles de réseaux de neurones ou tout simplement réseaux de neurones, ont été étudiés pendant plusieurs années dans le but d'imiter les performances du cerveau de l'être vivant.

Inspirés des réseaux neuromémitiques biologiques, ils existent plusieurs modèles de réseaux de neurones artificiels, et chaque modèle se prête bien pour une application particulière (classification, reconnaissance, contrôle ...etc.) ; Mais leurs utilisations restent limitées dans quelques applications, et il reste beaucoup de domaines où les réseaux de neurones n'ont pas trouvé de solutions, telle que la planification par exemple. Les meilleurs systèmes à réseaux de neurones restent assez loin d'imiter des performances telles que celles de l'être humain.

2. Définition d'un réseau de neurones

Un réseau de neurones est un modèle mathématique qui tente de reproduire quelques fonctions du cerveau humain, telles que : le parallélisme, l'acquisition des connaissances au travers d'un processus d'apprentissage, le stockage des connaissances et la possibilité d'utilisation de ces connaissances [12].

3. Principes de modélisation des Réseaux de neurones

3.1.Les neurons

Les neurones biologiques sont des cellules neurone uses constituant la base du système nerveux central. Ils possèdent trois principales composantes : Les dendrites, le corps cellulaire et l'axiome (figure 2.1).

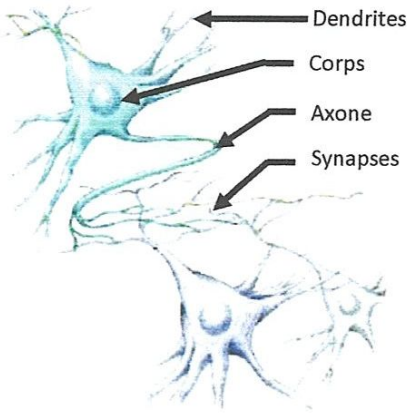


Figure 2.1: Le neurone biologique (œuvre d'artiste)

a) Dendrites

Chaque neurone possède une ''chevelure '' de dendrites. Celles-ci sont de fines extension tubulaire, de quelques dizaines de microns. Elles se ramifient, ce qui les amène à former une espèce d'arborescence autour du corps cellulaire. Elles sont les récepteurs principaux du neurone pour capter les signaux qui lui parviennent [10].

b) Corps cellulaire

Il contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie du neurone. Sa forme est pyramidale ou sphérique dans la plupart des cas. Elle dépend souvent de sa position dans le cerveau, ainsi les neurones du néo-cortex ont principalement une forme pyramidale ce corps cellulaire fait quelque microns de diamètre [10].

c) L'axiome

L'axone, qui est à proprement parler la fibre nerveuse, sert de moyen de transport pour les signaux émis par le neurone. Il se distingue des dendrites par sa forme et par les propriétés de sa membrane externe. En effet, il est généralement plus long (sa longueur varie d'un millimètre a plus d'un mètre) que les dendrites, et se ramifie à son extrémité, là où il communique avec d'autre neurones, alors que les ramifications des dendrites se produisent plutôt près du corps cellulaire [10].

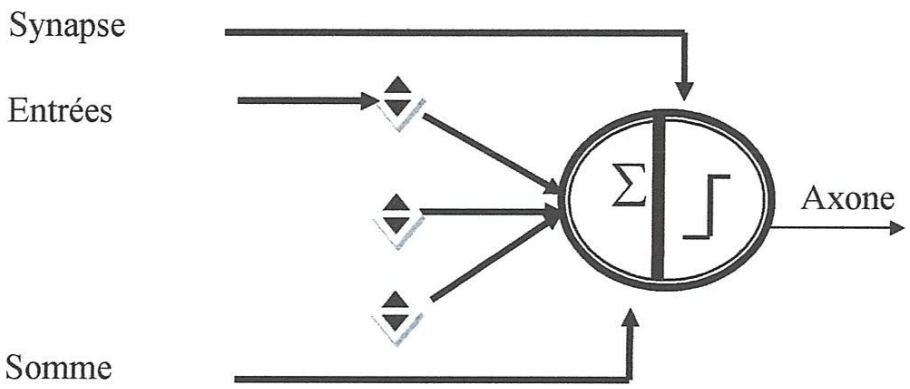


Figure 2.2: Le schéma classique présenté par les biologistes

3.3.Les neurones formels

Un neurone formel est une sorte d'automate, il se compose d'entrées, d'une fonction de sommation, d'une fonction d'activation et d'une ou plusieurs sorties. La figure 2.3 illustre un neurone de ce type.

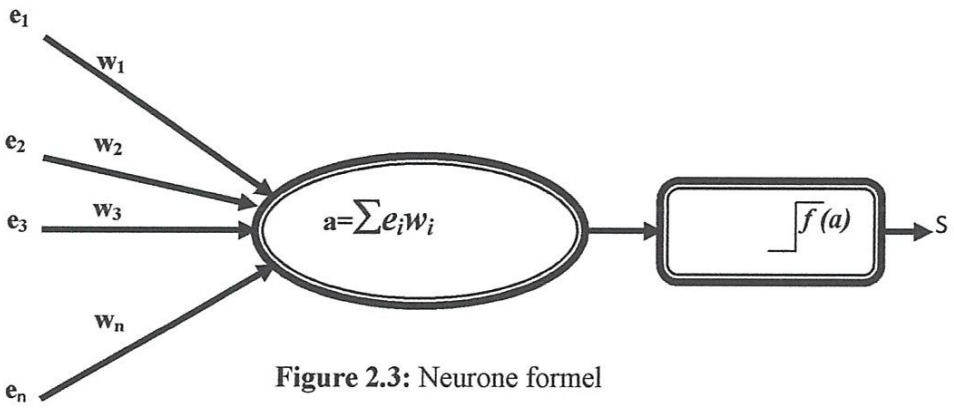


Figure 2.3: Neurone formel

3.4.Fonction d'activation

La fonction d'activation joueun rôle très important dans le comportement du neurone, elle donne une valeur représentative de l'activation du neurone. La fonction d'activation peut avoir plusieurs formes différentes, la figure 2.4 illustre quelques exemples.

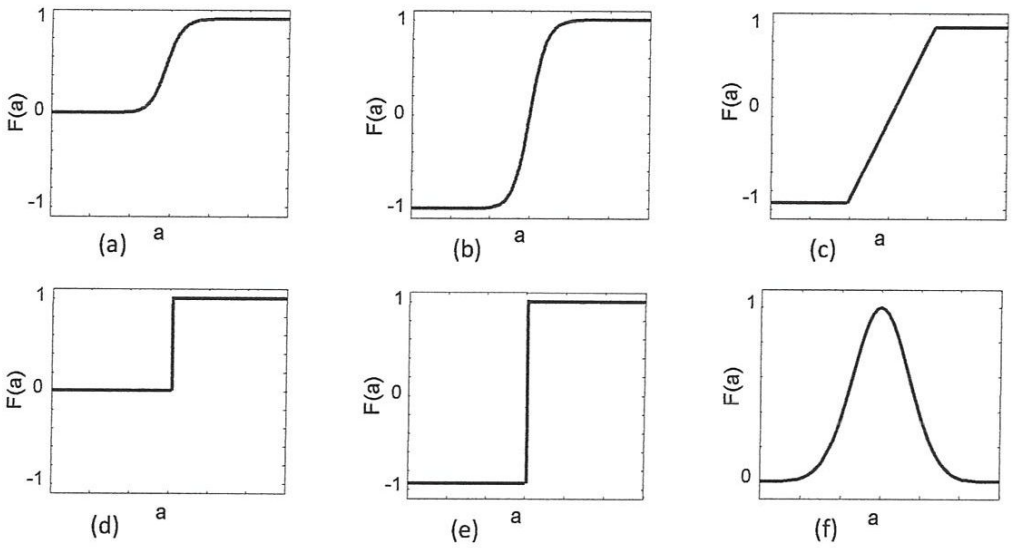


Figure 2.4: Exemples de fonctions d'activation

(a) Fonction seuil (b) Fonction seuil bipolaire

(c) Fonction RBF (d) Fonction sigmoïde

(e) Fonction sigmoïde bipolaire (f) Fonction linéaire saturée

4. Architectures de réseaux de neurones

4.1.Les réseaux non bouclés

Un réseau de neurone est non bouclé, si son graphe ne possède pas de cycle .Dans tel réseau, l'information circule de l'entrée vers la sortie sans aucun retour. Un réseau de neurone non bouclé est dit acyclique .Citons l'exemple des réseaux multicouches.

4.2.Les réseaux bouclés

Un réseau est bouclé, si son graphe possède au moins un cycle. Un réseau bouclé fait ramener une ou plusieurs valeurs à l'entrée .Un réseau de neurone bouclé est donc un système dynamique, régi par des équations différentielles ; comme l'immense majorité des applications sont réalisées par des programmes d'ordinateurs, on se place dans le cadre des systèmes à temps discret, où les équations différentielles sont remplacées par des équations aux différences [08].

4.3.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.

a) Réseau multicouche

Les neurones sont arrangés par couche. Il n'y a pas de connexion entre neurones d'une même couche et les connexions ne se font qu'avec les neurones des couches avale Habituellement, chaque neurone d'une couche est connecté à tous les neurones de la couche suivante et celle-ci seulement. Ceci nous permet d'introduire la notion de sens de parcours de l'information (de l'activation) au sein d'un réseau et donc définir les concepts de neurone d'entrée, neurone de sortie. Par extension, on appelle couche d'entrée l'ensemble des neurones d'entrée, couche de sortie l'ensemble des neurones de sortie.

Les couches intermédiaires n'ayant aucun contact avec l'extérieur sont appelés couches cachées [03].

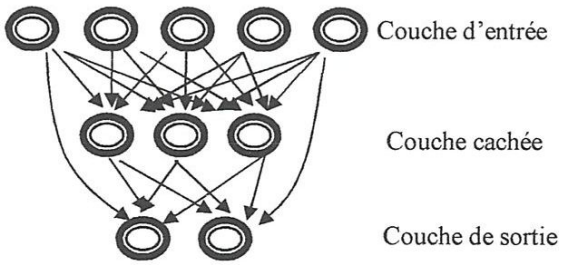


Figure 2.5: Réseau multicouche

b) 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. Les connexions sont donc moins nombreuses que dans le cas d'un réseau multicouche classique

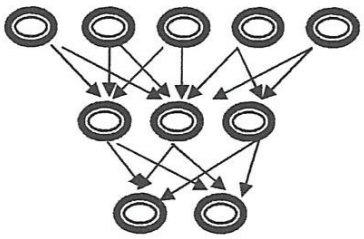


Figure 2.6: Réseau à connexions locales

c) 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

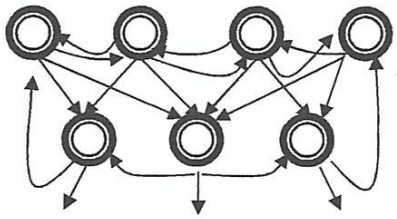


Figure 2.7: Réseau à connexions récurrentes

d) Réseau à connexion complète

C'est la structure d'interconnexion la plus générale. Chaque neurone est connecté à tous les neurones du réseau.

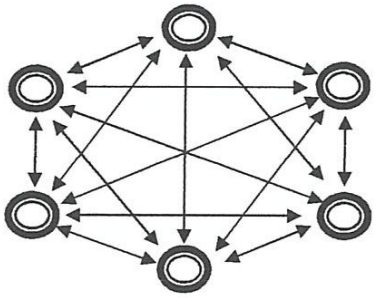


Figure 2.8: Réseau à connexion complète

5. Domaines d'applications des réseaux de neurones

L'essor des réseaux de neurones dans divers domaines actuels est certainement dû à leurs grandes capacités de calcul et à leurs hautes habilités d'apprentissage. De plus, l'estimation de leurs paramètres est indépendante de la complexité du problème traité ce qui leur permet d'être bien adaptés aux problèmes actuels qui ne cessent d'être de plus en plus complexes.

Les applications des réseaux de neurones peuvent être récapitulées en trois grands domaines:

- La modélisation : La plus part des problèmes industriels ou de recherche, que ce soit mécanique, physique, chimique ou même économique, nécessitent une représentation à l'aide d'un modèle mathématique permettant de reproduire le comportement du processus mis en œuvre. De telles tâches nécessitent des outils de calcul ayant de grandes capacités de calcul, d'apprentissage et surtout des outils dont leur conception est peu dépendante de la complexité et de la taille du problème traité. Les réseaux de neurones semblent être l'une des solutions les plus adéquates à ce type de problèmes [12].

- La commande : Commander un processus industriel consiste à concevoir un système permettant le calcul de la commande à appliquer à ce processus de manière à lui assurer un comportement dynamique désiré. Les réseaux de neurones permettent de bonnes performances en tant que partie de commande à cause de leur souplesse d'auto adaptation.

- La classification : Une autre grande catégorie de problèmes industriels consiste à attribuer, de façon automatique, un objet à une classe parmi d'autres classes possibles. La résolution de ce type de problèmes demande de représenter les exemples à classifier à l'aide d'un ensemble de caractéristiques. Il s'agit ensuite de concevoir un système capable de classifier ces exemples en se basant sur leur représentation et les réseaux de neurones sont particulièrement bien adaptés à ce type de problème. La classification est d'ailleurs le domaine privilégié des réseaux de neurones [12].

6. Apprentissage d'un réseau de neurones

6.1.Définition

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 d'apprentissage.

Dans les algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions.

6.2.Protocoles d'apprentissages

L'apprentissage des réseaux de neurones s'effectue généralement en quatre étapes :

Etape 1: Initialisation des poids synaptiques avec des petites valeurs aléatoires

Etape 2: Présentation de l'exemple d'entré et propagation de l'activation des neurones.

Etape 3: Calcule de l'erreur. Dans le cas d'un apprentissage supervisé cette erreur dépend de la différence entre l'activation des neurones et la sortie désirée.

Etape 4: Calcul du vecteur de correction à partir des valeurs des erreurs, avec lequel on effectue la correction des poids synaptiques.

6.3.Les types d'apprentissage

Les techniques d'apprentissage se subdivisent en trois grandes familles :

a) Apprentissage supervisé

Pour ce type d'apprentissage (perceptron, adaline , etc....), le réseau doit savoir qu'il a commis une erreur et il doit connaitre la réponse qu'il dû donner. Pour un stimulus présenté à la couche d'entrée du réseau, la réponse obtenue est comparée avec celle désirée et la modification et les ajustements à apporter aux poids sont déterminés en fonction de l'erreur commise par le réseau. Généralement, les règles d'apprentissage supervisé sont des formes de descente du gradient.

L'apprentissage supervisé nécessite donc la définition d'une base d'exemples d'apprentissage représentative .Chaque exemple présenté au réseau est un couple (entrée, sortie désirée). La minimisation de l'erreur entre la valeur de sortie et la valeur désirée est basée sur le principe de l'erreur quadratique [13].

b) Apprentissage non supervisé

Pour ce type d'apprentissage il s'agit d'atteindre l'ensemble des poids synaptiques pour lesquels le comportement du réseau est optimal .La modification et l'ajustement des poids se font en fonction d'un critère interne, indépendant de la relation entre le comportement du réseau et la tâche qui doit effectuer.

c) Apprentissage semi-supervisé

L'apprentissage semi-supervisé suppose qu'un comportement de référence précis n'est pas disponible, mais qu'en revanche, il est possible d'obtenir des indications qualitatives (correcte /incorrecte) sur les performances du réseau [13].

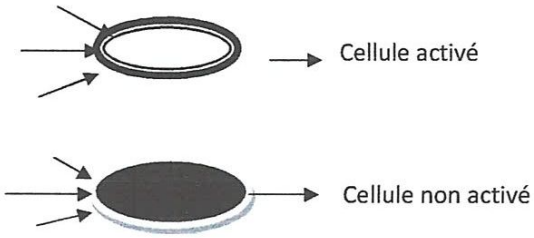
6.4.Règles d'apprentissage

a) La règle de Hebb

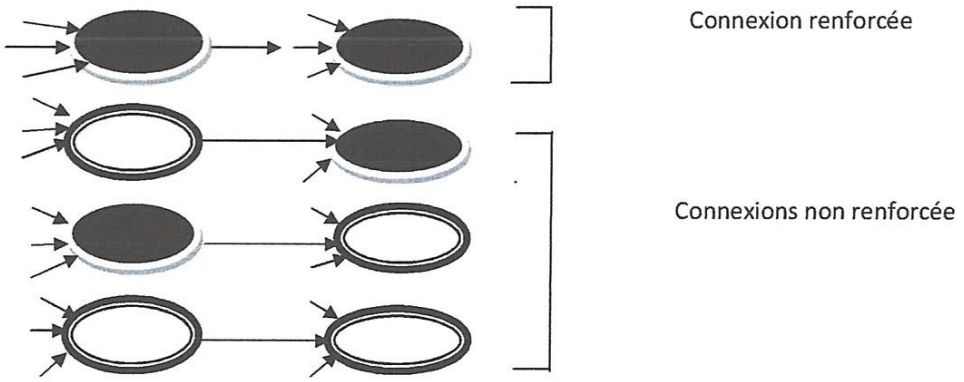
La règle de Hebb est le premier mécanisme d'évolution proposé sur les synapses. Son interprétation pour les réseaux de neurones formels est la suivante :

On considère que deux neurones connectés entre eux sont activé aux mêmes moments, la connexion qui les relie doit être renforcée et elle n'est pas modifiée, dans le cas contraire. C'est-à-dire que le poids w_{ij} d'une connexion entre un neurone i et un neurone j augmente quand les deux neurones sont activés en même temps et il n'est pas modifié, dans le cas contraire [13].

Prenons, à titre d'exemple, les conventions suivantes :



La règle de Hebb donne alors :



Quand la cellule émettrice et la cellule réceptrice s'activent en même temps, il faut augmenter le poids de cette connexion lors de l'apprentissage. La connexion entre ces deux cellules devienne alors très forte. Si la cellule émettrice s'active sans que la cellule réceptrice ne le soit, ou si la cellule réceptrice s'active alors que la cellule émettrice n'était pas activée, cela traduit bien le fait que la connexion entre les deux n'est pas prépondérante dans le comportement de la cellule réceptrice on peut donc dans la phase d'apprentissage laisser un poids faible à cette connexion [13].

En se basant sur ce principe, Hebb a donné la règle d'apprentissage suivante :

$$Wij(t + \partial t) = Wij(t) + \mu AiAj \qquad (2.1)$$

Avec Wij (t) et $Wij(t + \partial t)$: les poids de la connexion entre le neurone i et le neurone j aux instants(t) et$(t + \partial t)$.

Ai et Aj : L'activation du neurone i et l'activation du neurone j

μ : C'est le paramètre de l'intensité de l'apprentissage$(\mu > 0)$.

b) La règle de Widrow-Hoff

La règle d'apprentissage de Widrow-Hoff, ou des moindres carrés (LMS, Least Square Sum), est une règle d'apprentissage supervisé basée sur la correction d'erreurs observées en sortie. Cette règle consiste à minimiser une fonction coût caractérisée par l'erreur quadratique moyenne. Pour un ensemble d'apprentissage contenant Qpaires entrée/sortie désirée$\{(X^{(q)}/T^{(q)})\}, q = 1, \dots Q$ où $X^{(q)}$ et $T^{(q)}$ représentent respectivement la $q^{ème}$ entrée et la $q^{ème}$ sortie désirée, l'erreur $(e(r))$ à l'itération r est donnée par :

$$e(r) = T(r) - Y(r) \qquad (2.2)$$

Où :$Y(r)$ est la sortie calculée du réseau. La fonction coût est :

$$F(X) = e^2(r) \qquad (2.3)$$

L'apprentissage selon la règle LMS consiste à calculer le gradient à chaque présentation d'un exemple d'apprentissage. Le changement de poids est alors :

$$\Delta w_{ij}(t) = -\eta \nabla F(X) \quad (2.4.a)$$

$$= \eta \frac{\partial e^2(r)}{\partial w_{ij}} \quad (2.4.b)$$

Cette règle de correction permet donc aux neurones d'adapter leurs poids pour se rapprocher à une valeur désirée correspondante à chaque exemple présenté. Cette règle a été utilisée pour l'apprentissage de l'ADALINE dans lequel chaque neurone i corrige ses poids w_{ij} à l'itération$r$ selon l'équation suivante :

$$\Delta w_{ij}(r) = \Delta w_{ij}(r-1) - \eta(t_i - y_i)x \quad (2.5)$$

Où : t_i et $y_i$sont respectivement la sortie désirée et la sortie calculée correspondantes au neurone i ; x est l'entrée et η est une constante positive appelée pas d'apprentissage [12].

7. Type de réseaux de neurones

Il existe une multitude de réseaux de neurones différents les uns des autres par leurs architectures et leurs méthodes d'apprentissages.

7.1.L'Adaline

Peu de temps après que Rosenblatt proposa son perceptron, Windrow et Hoff proposèrent un autre modèle de neurone qu'ils nommèrent l'Adaline (pour Adaptative L'Inear Neurone) en 1960.

Ce modèle trouva diverses applications comme composante dans les antennes adaptatives et les modems à haute vitesse. L'Adaline est un seul neurone à valeur d'activation continue et une fonction d'activation linéaire.

$$a_i = \sum Wijaj - \theta_I \quad (2.6)$$

a_i : Activation du neurone i.

W_{ij} : Poids du neurone j vers le neurone i.

θ_i : Seuil du neurone i.

a_j : Activation du neurone précédant j.

La règle d'apprentissage utilisée est la règle de "Widrow-Hoff" ou règle Delta :

$$\Delta Wij = \eta a_j (d_i - a_i) \qquad (2.7)$$

ou :

η : Est le pas d'apprentissage.

aj : Activation du neurone j.

d_i : Activation attendue du neurone i.

a_i : Activation du neurone i.

7.2.Le Perceptron

Ce modèle fut le premier proposé en 1958 par "Frank Rosenblatt". Le perceptron était constitué de trois couches appelées :

- Rétine (neurones d'entrées)
- Aire d'association (neurones cachés)
- Neurones réponse (neurones de sortie)

8. Le Perceptron multi couche

8.1.Architecture du MLP

Le MLP est une réalisation en série de perceptrons. Ce modèle comprend en plus de la couche d'entrée et la couche de sortie, une ou plusieurs couches cachées. Dans ce réseau, chaque neurone est connecté à tous les neurones de la couche précédente et la couche suivante. Les neurones de la même couche ne sont pas connectés entre eux [13].

L'introduction de couches intermédiaires dans le réseau MLP permet de résoudre des problèmes plus complexes que la simple séparation linéaire. Lorsqu'il existe au moins une couche cachée, les états internes du réseau ne peuvent plus être donnés directement par les exemples et les sorties désirées puisque les sorties des neurones appartenant aux couches intermédiaires sont inconnues [03].

La figure (2.9) illustre un MLP, avec une seule couche cachée, comportant N neurones d'entrée,M neurones cachés et J neurones de sortie. Le $n^{ème}$ neurone d'entrée est relié avec le $m^{ème}$ neurone caché par le poids w_{nm} et le $m^{ème}$ neurone caché est relié avec le $j^{ème}$ neurone de sortie par le poidsu_{mj}. A chaque présentation d'un exemple X $(x_1, x_2, \dots, x_N)$, les composants de son vecteur caractéristique seront transmis aux neurones de la couche cachée. Les sorties de ces neurones $(y_1, y_2, \dots, y_M)$ seront à leur tour transmis aux neurones de la couche suivante (la couche de sortie). La sortie du $m^{ème}$ neurone caché est donnée par :

$$y_m = h(r_m) = h(\sum_{m=1}^{M} x_n w_{nm}) \qquad (2.8)$$

Où : h(.) est la fonction d'activation des neurones de la couche cachée.

Les neurones de la couche de sortie constituent la sortie du réseau. La sortie du j^{ème} neurone est donnée par:

$$z_j = g(s_j) = g(\sum_{j=1}^{J} y_m u_{mj}) \qquad (2.9)$$

Où : $g(.)$ est la fonction d'activation des neurones de sortie.

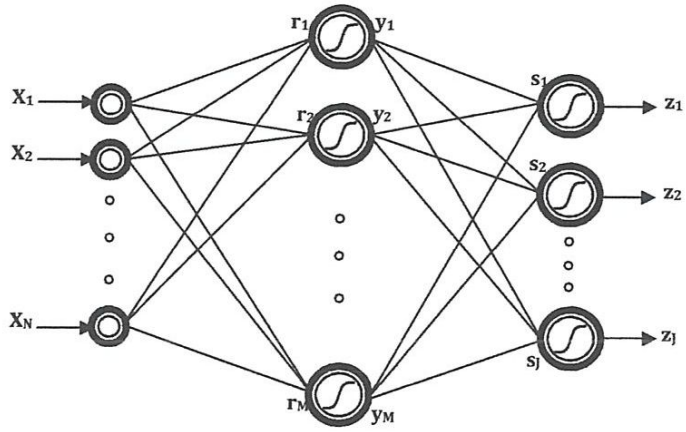


Figure 2.9: MLP avec une seule couche cachée contenant N neurones d'entrée, M neurones cachés et J neurones de sortie.

Dans les problèmes de classification, le nombre de neurones de la couche d'entrée du MLP est égal à la dimension de l'espace caractéristique parce que le rôle de cette couche est de présenter chaque

composante du vecteur caractéristique aux neurones de la couche suivante. Le nombre de neurones de la couche de sortie égale le nombre de classes, soit un neurone par classe (la sortie de ce neurone est 1 pour les exemples appartenant à la classe correspondante et les autres sont à 0).

En tant que classificateur, le MLP doit donc fournir une sortie $Z\ (z_1, z_2, \dots, z_J)$ correspondante à la classe de l'exemple $X\ (x_1, x_2, \dots, x_N)$ présenté en entrée [13].

8.2.Apprentissage du MLP par la retro-propagation

La rétro-propagation du gradient (Back Propagation) est l'une des méthodes les plus simples et les plus utilisées pour l'apprentissage des réseaux de neurones. C'est une extension de la règle d'apprentissage de Widrow appliquée aux réseaux monocouches. La RP (rétro-propagation) consiste donc à minimiser la distance entre la sortie calculée $Z^{(q)}$ et la sortie désirée $T^{(q)}$ correspondantes à chaque exemple d'apprentissage $X^{(q)}$. L'erreur quadratique est souvent employée comme étant la fonction coût de la RP. Pour un ensemble de Q exemples d'apprentissage, l'erreur quadratique totale est donnée par :

$$E = \sum_{q=1}^{Q} \sum_{j=1}^{J} \left(t_j^{(q)} - z_j^{(q)} \right)^2 \qquad (2.10)$$

L'algorithme de la RP est basé sur la modification des poids du réseau de façon à effectuer une descente de gradient sur la surface d'erreur. Au début de l'apprentissage, les poids sont initialisés avec des valeurs aléatoires et modifiés ensuite dans une direction qui réduira l'erreur. La modification Δw d'un poids w est donnés par :

$$\Delta w = -\eta \frac{\partial E}{\partial w} \qquad (2.11)$$

Où : η est une constante positive appelée pas d'apprentissage permettant de définir la taille des modifications des poids [14].

Il s'agit donc de prendre un pas dans l'espace des poids permettant de réduire la fonction coût. Chaque poids est modifié à l'itération $(r + 1)$ en fonction de sa valeur à l'itération (r) par :

$$w^{(r+1)} = w^{(r)} + \Delta w^{(r)} \qquad (2.12)$$

Pour un MLP avec une couche cachée (figure 2.9), la mise à jour des poids de la couche cachée et ceux de la couche de sortie est donnée par :

$$u_{mj}^{(r+1)} = u_{mj}^{(r)} - \eta \frac{\partial E^{(r)}}{\partial u_{mj}} \quad (2.13.a)$$

$$w_{nm}^{(r+1)} = w_{nm}^{(r)} - \eta \frac{\partial E^{(r)}}{\partial w_{nm}} \quad (2.13.b)$$

Le développement de l'Eq. (II.13.a) et l'Eq. (II.13.b) donne les équations d'adaptation suivantes :

$$u_{mj}^{(r+1)} = u_{mj}^{(r)} + \eta_1 (t_j - z_j) \dot{g}(s_j) y_m \quad (2.14.a)$$

$$w_{nm}^{(r+1)} = w_{nm}^{(r)} + \eta_2 (\Sigma_{j=1}^{J} (t_j - z_j) \dot{g}(s_j) u_{mj}) \acute{h}(r_m) x_n \quad (2.14.b)$$

Où : $\dot{g}(s_j)$et $\acute{h}(r_m)$ sont respectivement les dérivées des fonctions d'activation des neurones cachés et des neurones de sortie. Dans le cas des fonctions sigmoïde, les dérivées sont données par :

$$\dot{g}(s_j) = z_j(1 - z_j) \quad (2.15.a)$$

$$\acute{h}(r_m) = y_m(1 - y_m) \quad (2.15.b)$$

9. Conclusion

Les réseaux de neurones ont connu un essor considérable tant qu'en nouvelles architectures qu'en nouveaux algorithmes d'apprentissage, et dans ce chapitre nous avons tenté de donner un simple survol sur ces importants outils mathématiques.

Nous avons présenté avec un peu de détail le PMC ainsi que son apprentissage. Ce modèle est le plus utilisée surtout dans le domaine de la classification, et par conséquent nous l'utiliserons pour la phase de classification de notre système de reconnaissance vocale.

CHAPITRE 3

Application

1. Introduction

L'objectif de ce travail est la réalisation d'un système de reconnaissance automatique de la parole. Le système qu'on propose est à un vocabulaire limité et multi-locuteurs. Pour l'extraction des caractéristiques, on se base sur les coefficients cepstraux dans l'échelle de Mels (MFCC) qui sont les plus utilisé dans ce domaine. Dans la phase de classification, le système proposé se base sur les réseaux de neurones.

Nous réalisons une petite base de données avec dix chiffres prononcés par deux locuteurs.

2. Extraction des caractéristiques

Les paramètres MFCC sont des coefficients cepstraux obtenus à partir des énergies d'un banc de filtre en échelle de fréquence Mel.

Les différentes étapes pour l'obtention des MFCC sont :

1. Fenêtrage
2. RFFT
3. Filtre de MEL
4. Calcul du Log
5. $RFFT^{-1}$
6. MFCC

Pour le calcul de ces fonctions nous utilisons les librairies de MATLAB, les fonctions utilisées sont :

1. Enframe : découpe le signal en frames en utilisant la fonction de pondération.
2. Melbankm : création d'un banc de filtres.
3. Melcepst : création de la matrice de coefficients cepstraux.
4. RDCT : calcule de la transformée en cosinus discrète en réel.
5. RFFT : calcule de la transformée de Fourier discrète réel.

3. Exemple d'extraction de caractéristiques

Prenons un exemple d'extraction des caractéristiques du mot (trois) :

3.1.Le signale enregistré

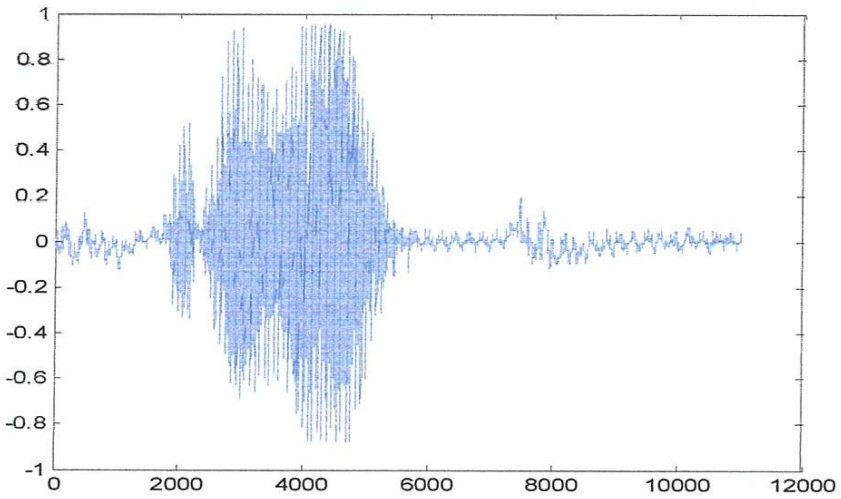


Figure 3.1: signale enregistré

3.2.La transformée rapide de Fourier réelle du signal

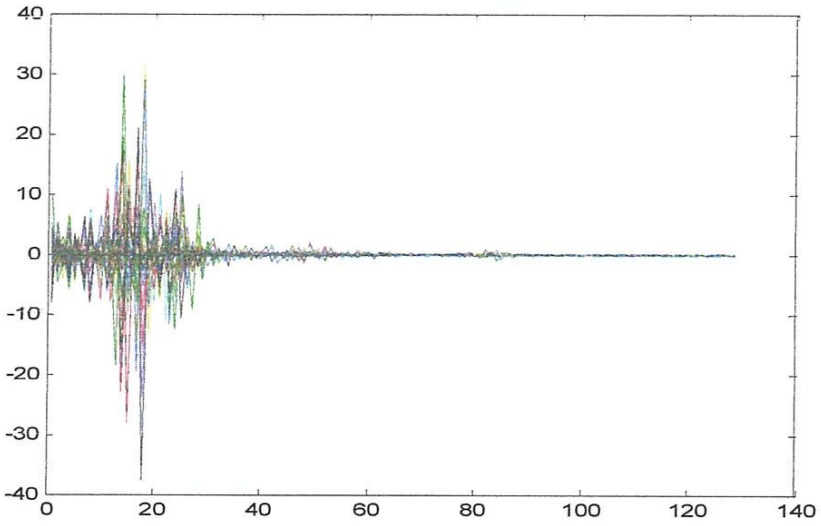


Figure 3.2: transformée rapide de Fourier réelle du signal

3.3.Le banc de filtres

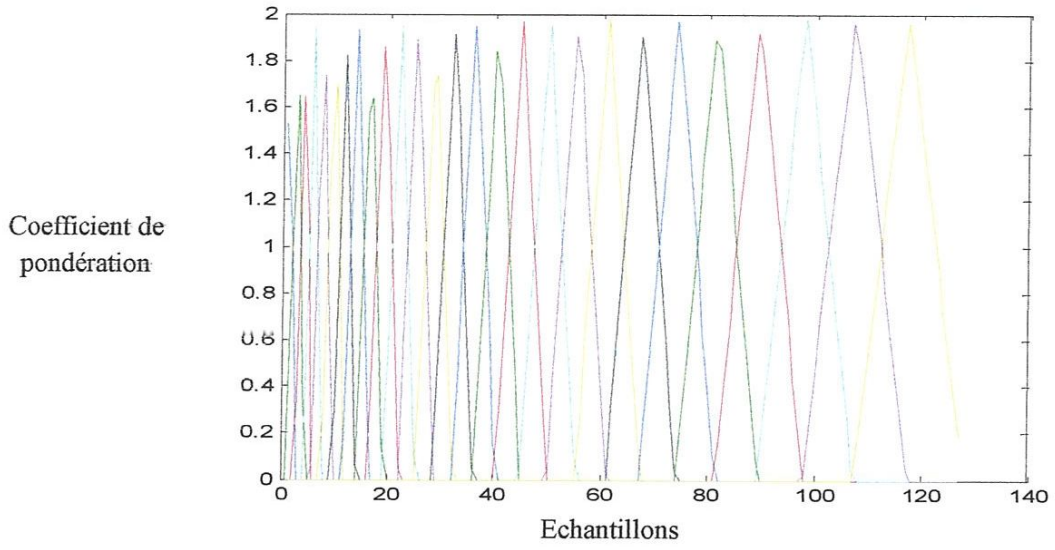


Figure 3.3: banc de filtres

4. Phase de classification

Pour la classification nous utilisons un PMC qui est le réseau de neurones le plus utilisé pour la classification. Le PMC utilisé comporte un nombre de neurones d'entrée égale au nombre de caractéristique (MFCC) donc N=12. Ce PMC comporte un nombre de neurones de sortie égale aux nombre de classe donc J=10. Pour le nombre de neurones dans la couche cachée, nous avons choisie (par expérience) M=8.

4.1.Algorithme d'apprentissage de la Back-Propagation

4.1.1. Etape 1 : introduction des données

$X^{(q)}$ (vecteurs caractéristiques)

$T^{(q)}$ (vecteurs de sorties désirées)

4.1.2. Etape 2 : détermination des dimensions de données

N (nombre de caractéristique qui sera le nombre de neurones de la couche d'entrée)

Q (nombre d'exemples d'apprentissage)

J (nombre de classes qui est le nombre de neurones de la couche de sortie)

Initialisation des paramètres

I (nombre d'itération)

η_1 et η_2 (pas d'apprentissage).

M (nombre de neurones de la couche cachée)

U_{mi} et W_{nm} (poids initiaux avec des valeurs aléatoires)

4.1.3. Etape 3 : Apprentissage

Pour i=1 à I (pour chaque itération)

Pour q =1 à Q (pour chaque exemple)

$X = X^{(q)}$ (introduction du vecteur caractéristique de l'exemple q)

$T = T^{(q)}$ (introduction du vecteur de sortie de l'exemple q)

a) calcul de sortie

$$y_m = h(r_m) = h\left(\sum_{m=1}^{M} x_n w_{nm}\right)$$

$$z_j = g(s_j) = g\left(\sum_{j=1}^{J} y_m u_{mj}\right)$$

b) la mise à jour des poids

Pour m=1 à M

Pour j=1 à J

Ajustement de chaque poids u_{mj}

$$u_{mj}^{(i+1)} = u_{mj}^{(i)} + \eta_1 \cdot (t_j - z_j) \cdot g' \cdot (s_j) y_m$$

Pour n=1 à N

Ajustement de chaque poids w_{nm}

$$w_{nm}^{(i+1)} = w_{nm}^{(i)} + \eta_2 \cdot \left(\sum_{j=1 i=1}^{J} (t_j - z_j) \cdot g'(s_j) \cdot u_{mj} \right) \cdot (h'(r_m) \cdot (x_n)$$

c) calcul de l'erreur

$$E = \sum_{q=1}^{Q} \sum_{j=1}^{J} \left(t_j^{(q)} - z_j^{(q)} \right)^2$$

4.1.4. Etape 4 : Fin

4.2.Exemple de classification

Dans cette partie nous considérons deux exemples de classification bidimensionnels à deux classes. Les exemples du premier problème sont linéairement séparable tandis que ceux du deuxième problème ne sont pas linéairement séparable (figure 3.4).

L'évolution de l'erreur au cours de l'apprentissage d'un PMC pour résoudre ces problèmessont illustrés sur la figure 3.5.Nous constant que le PMC converge plus rapidement dans le cas du premier problème.

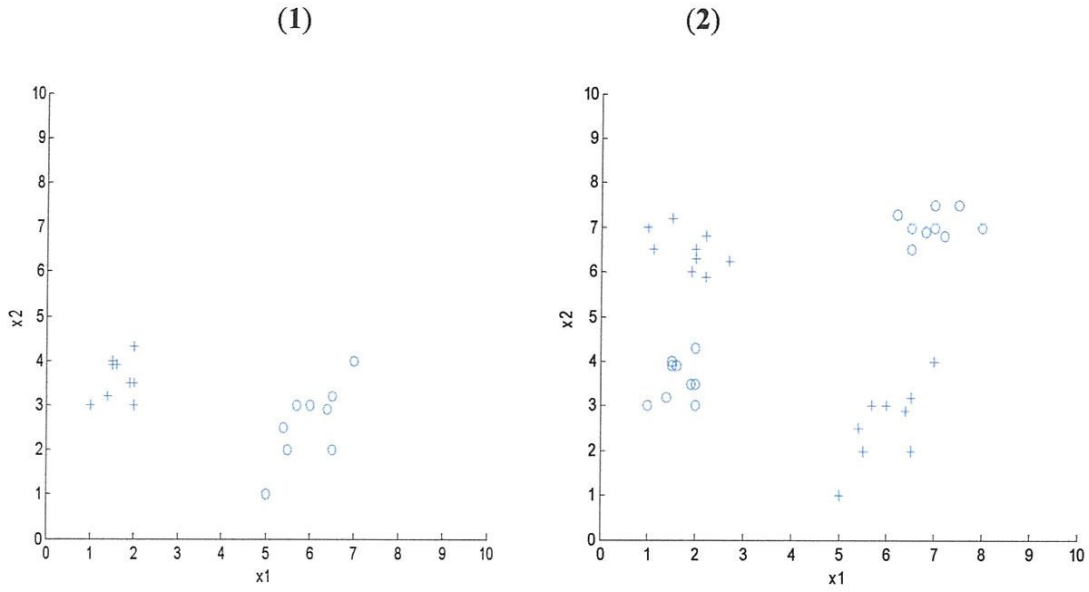


Figure 3.4: Deux problème de classification (le 1er linéairement séparable et le 2ème non linéairement séparable)

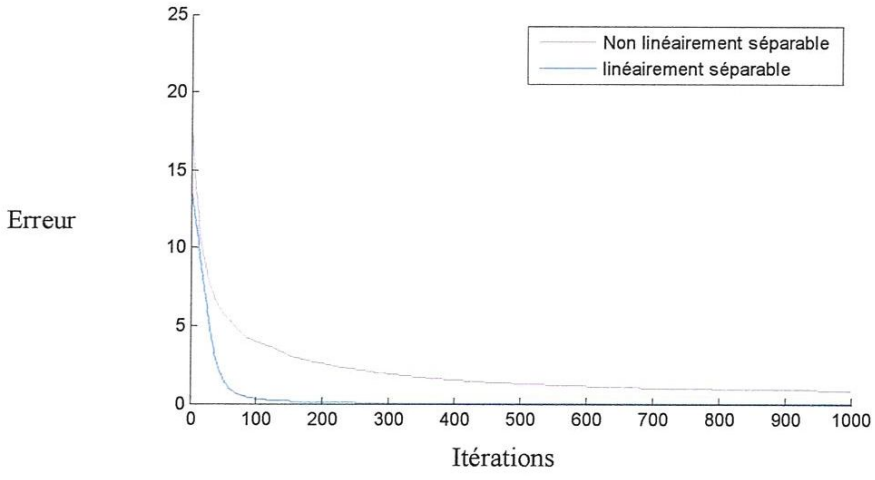


Figure 3.5: Evolution de l'erreur pour les deux problèmes

5. Résultats

Nous avons établi une base de données avec 10 chiffres, prononcés par deux locuteurs. Chacun des deux locuteurs a répété chaque chiffre8 fois.

Nous avons réalisé deux tests de classification comme suite :

1. Tests mono-locuteur (pour chaque locuteur).
2. Test pour les 2 locuteurs sur tout l'ensemble de données

Dans les deux cas, nous divisons la base de données de chaque locuteur en deux parties : une partie pour l'apprentissage et une partie pour le test. Pour évaluer les performances de généralisation de notre système, nous utilisons une validation croisée d'ordre 4. Quatre ensembles d'apprentissage, dont chacun se constitue de 6 exemples pour chaque chiffre, sont ainsi obtenus. Les ensembles de test correspondants contiennent 2 exemples.

5.1.1^{er} test : mono-locuteur

Les figures 3.6 et 3.7 représentent l'évolution de l'erreur pour chaque locuteur. Le tableau 3.1 donne les résultats obtenus: Le système parvient à bien classifier tous les exemples d'apprentissage et du test pour les deux locuteurs

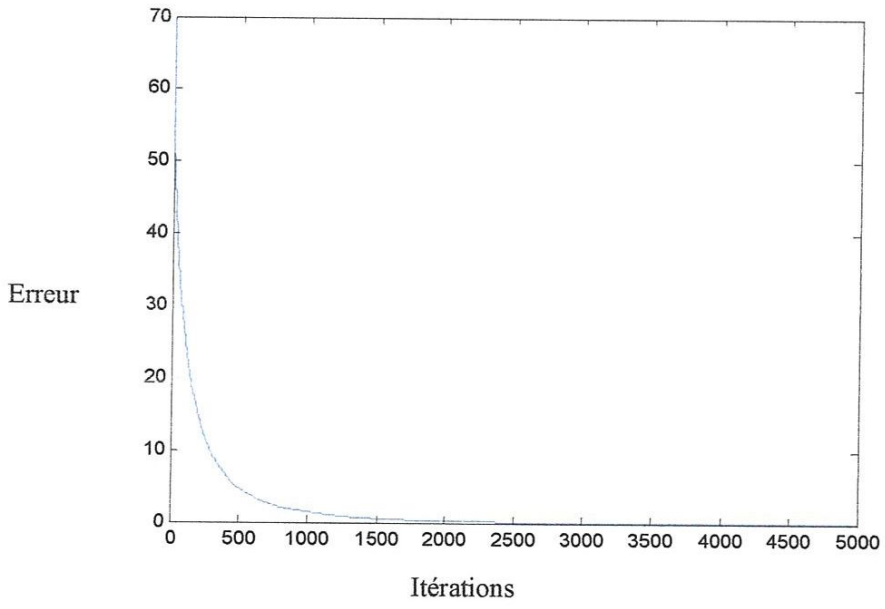


Figure 3.6: l'évolution de l'erreur de locuteur 1(base 1)

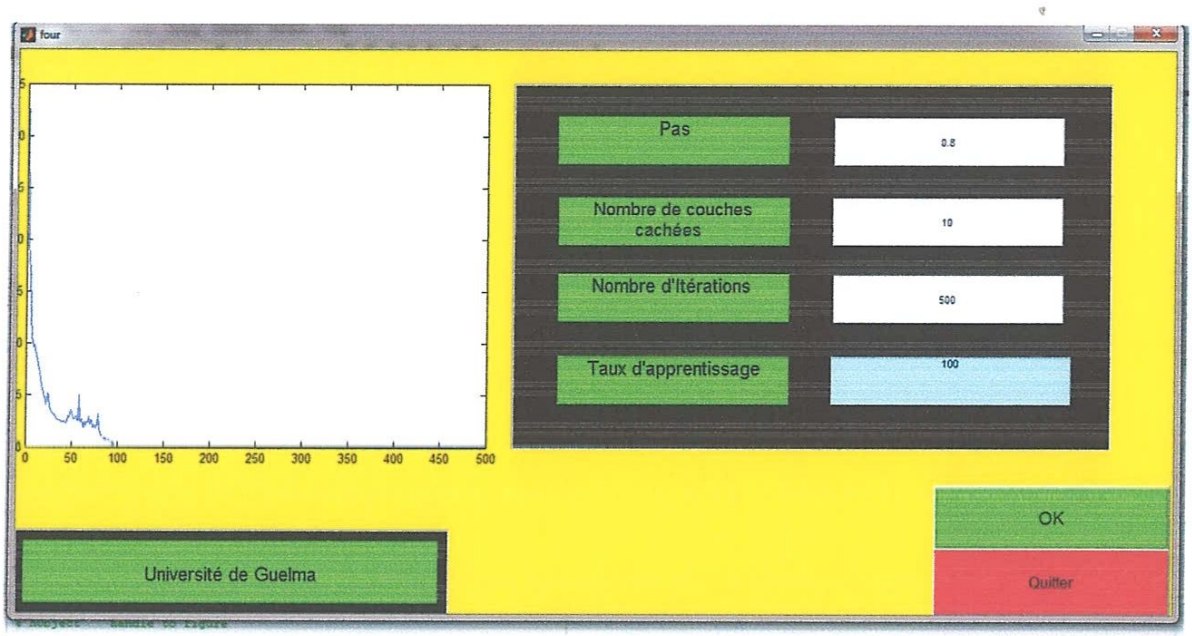


Figure 3.12: Apprentissage

Apres click sur le buttons Reconnaissance on passe à la fenêtre de la figure 3.13. Cette fenêtre donne la possibilité de reconnaitre le mot enregistré en se basant le réseau entrainé et affiche le résultat : vrai, faut ou non reconnu.

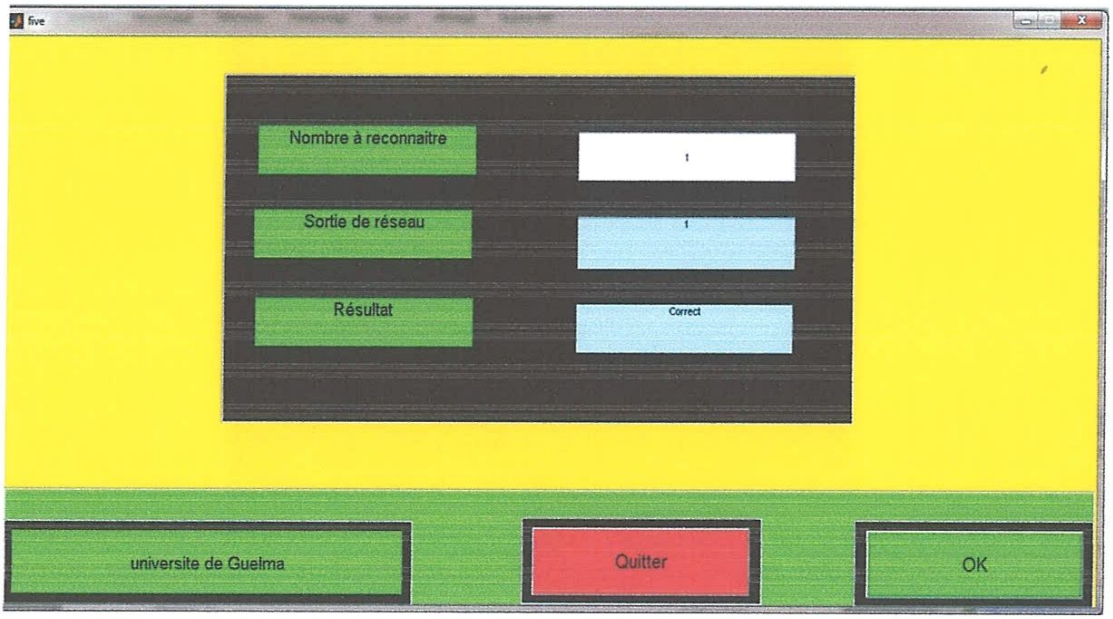


Figure 3.13: Reconnaissance

7. Conclusion

Les résultats obtenus dans les déférents tests précédant montrent que le système propose a permet de donnes des résultats satisfaisant, nous avons divisé la base de données de chaque locuteur en deux parties : une partie pour l'apprentissage et une partie pour le test. Pour évaluer les performances de généralisation de notre système, nous avons utilisé une validation croisée d'ordre 4.

Dans ce chapitre, nous avons décrit brièvement notre logiciel, développé sous l'environnement Matlab(GUI). Afin de permettre à l'utilisateur une meilleure exploitation des fonctionnalités de l'application, nous avons expliqué comment utilisé les différents fenêtres ainsi que les tâches qui leurs sont attribuées.

Conclusion générale

Dans ce travail nous avons réalisé un système de reconnaissance vocal de chiffres. Ce système est à un vocabulaire limité et à 2-locuteurs. Pour l'extraction des caractéristiques, nous avons basé sur les coefficients cepstraux dans l'échelle de Mels (MFCC) qui sont les plus utilisé dans ce domaine. Dans la phase de classification, nous avons utilisé le PMC qui est le réseau de neurones le plus utilisé dans le domaine de la classification.
Nous avons élaboré une base de données avec dix (10) chiffres prononcés par deux locuteurs.

Nous avons réalisé des tests de classification mono-locuteur et à 2 locuteurs. Les résultats obtenus sont satisfaisant surtout dans le cas mono-locuteur ce qui montre les capacités de généralisation des réseaux de neurones.

Pour simplifier l'utilisation de notre programme nous avons réalisé une interface graphique par graphic utilization interface (GUI)

Bibliographie

[01] JEAN Paul « Reconnaissance automatique de la parole » DUNOD, 2006

[02] KHEBLiI Abdelmalek « Reconnaissance vocal par mélange de Gaussiennes (GMM)», thèse Magister, ECOLE MILITAIRE POLYTECHNIQUE 2009.

[03] NADIA BENAHMED « Optimisation de réseaux de neurones pour la reconnaissance de chiffres manuscrits isoles » thèse Doctorat, UNIVERSITÉ DU QUÉBEC, 2002

[04] » RENE Boite, MURAT Kunt « Traitement de la parole » PRESSES POLYTECHNIQUE ROMANDES, 1987

[05] MEDJDOUB Bensalem « Système de vérification des locuteurs basé sur la fusion en scores des paramètres MFCC et LPCC » Mémoire Ingéniorat, ECOLE MILITAIRE POLYTECHNIQUE 2008.

[06] TAN Tien Ping « Automatic Speech Recognition for Nonnative Speakers » thèse Doctorat, UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1, 2008

[07] Rene Boite,Herve « Traitement de la parole » collection électricité, Novembre1999

[08] Mr DEGHMICHE Houari, Mr CHIBOUT Menaoeur « Utilisation des réseaux neuro-flous pour La Reconnaissance automatique e la Parole » mémoire Ingéniorat, INSTITUT DES TELECOMMUNICATIONS Oran 2003

[09] Stéphane LOISELLE « Exploration de réseaux de neurones à décharges dans un contexte de reconnaissance de parole » thèse Doctorat, UNIVERSITE DU QUEBEC 2004

[10]ERIC Davalo, PATRICK Naïm «Des réseaux de neurones » EYROLLES, 1993

[11] HLAOUI Adel « Reconnaissance de mots isoles arabes par hybridation de réseaux de neurones » mémoire Ingéniorat, ÉCOLE NATIONALE D'INGÉNIEURS DE TUNIS 2005

[12]NEMISSI Mohamed« Classification et reconnaissance des formes par algorithmes hybrides» thèse de doctorat, UNIVERSITE DE GUELMA 2009

[13]Mr. Bouyedda Hocine « Reconnaissance de l'écriture arabe imprimée par les réseaux de neurones » mémoire de magister, UNIVERSITE DE GUELMA 2007

[14]NEMISSI Mohamed « Classification automatique supervisé par les réseaux de neurones» mémoire de magister, UNIVERSITE DE GUELMA 2004

ملخص

العمل المقدم في هذه المذكرة يدخل في اطارالتعرف الألي على الصوت. وبالتحديد التعرف علي الأعداد من خلال تسجيل صوتي بالاعتماد علي الشبكات العصبية في مرحلة التصنيف

هذا العمل مكون من جزأين أساسيين الأول مخصص لاستخراجالمميزات والجزء الثاني من أجل تطوير برنامج يقوم بعملية التعرف .

كلمات مفتاحية

التعرف علي الصوت,معالجة الصوت,الشبكات العصبية

Résumé

Le travail réalisé s'inscrit dans le domaine de reconnaissance automatique de la parole et plus précisément la reconnaissance vocale de chiffres en basant sur les réseaux de neurones dans la phase de classification.

Mots clés

Reconnaissance vocal, MFCC, réseaux de neurones.

Abstract

This work relates to the Automatic Speech Recognition (ASR), specially the Vocal recognition of numbers, usingthe Neural Networks in the classification phase.

This memory has two principles parts. The first deals with caracteristic extraction and the second is reserved for developing the recognition process.

Key words

Speech Recognition, MFCC, Neural Networks