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

Université Amar Thelidji- Laghouat

FACULTE: DE TECHNOLOGIE

DEPARTEMENT : D'ELECTRONIQUE

MEMOIRE DE MASTER

Réalisé par : Habchi Maroua et Baalla Soumia

DOMAINE :Science et Technologie

FILIERE : Télécommunications

OPTION : Réseaux de Télécommunications

Thème

Reconnaissance Automatique audio de la Parole Arabe par des réseaux de neurones

Jury de soutenance :

Nom et Prénom	Grade	Qualité
Mr KORIBA Mustapha	MAA	Encadrant
Mr REGUIEGUE Mourad	MCB	Co-encadrant
Mr BIRANE Mouhoub	MCA	Président
Mr CHALALI Safouane	MCB	Examinateur

Promotion : 2021/2022

Remerciement

Tous d' abord nous remercions le bon dieu de nous avoir donné du courage, force et patience pour réalisé ce modeste travail.

Nous tenons a remercie notre encadreur ***Mr. Koriba Mustapha***

et co-encadreur ***Mr. Reguiegue Moured*** *pour leur aide précieux ,leur conseils et leur remarques apportés au cour de l'élaboration de ce projet de fin d'études ainsi qu'a tous les professeurs du département d'électronique qui ont contribué à notre formation durant ce cycle.*

Nous remercions également le chef de département d'électronique monsieur ***Mr. Marah Lahcen*** *et le membre du jury d'avoir d'accepter de juger notre travail.*

Enfin nous tenons à exprimer nos gratitudes et remerciement à toute personne ayant participé de prés ou de loin à notre formation et a tous ceux qui nous ont apporté leurs soutien et encouragement durant la réalisation de ce modeste travail.

Habchi Maroua et Baalla Soumia

Dédicace

Je rends grâce à dieu de m'avoir donner le courage et la volonté ainsi que la conscience d'avoir pu terminer mes études.

Je dédie ce modeste travail :

A mon très ***cher père*** *en reconnaissance de son amour, encouragement et affection qu'ils m'ont prodigués durant mes études. Que dieu me lui garde.*

A ma très ***chère mère*** *pour toute sa tendresse et pour ses nombreux sacrifices. Que dieu me la garde.*

A ma très chèrs ma sœur: ***Mebarka****.*

A mes frères : ***Badereddine*** *et* ***Idris***

A mon binôme: ***Soumia Baalla*** *pour son esprit d'équipe et sa patience avec moi*

A tous les membres de ma famille ***Habchi*** *et* ***Kemmam****.*

A tous mes ami(e)s, tous ceux que j'aime, tous ceux qui m'aiment et tous ceux qui me sont chers .Et surtout les étudiants de la promotion de télécom, Et a toutes les personnes qui m'aine.

Habchi Maroua

Dédicace

Je rends grâce à dieu de m'avoir donner le courage et la volonté ainsi que la conscience d'avoir pu terminer mes études.

Je dédie ce modeste travail :

A mon très ***cher père*** *en reconnaissance de son amour, encouragement et affection qu'ils m'ont prodigués durant mes études. Que dieu me lui garde.*

A ma très ***chère mère*** *pour toute sa tendresse et pour ses nombreux sacrifices. Que dieu me la garde.*

A mes chères sœurs :*Khadidja .Ikram Chaima.*

, et mes frères de leur souhaite tout le succès ... tout le bonheur *A*

mon binôme: ***Maroua Habchi*** *pour son esprit d'équipe et sa patience avec moi .a la deuxième famille Kendoure*

Et tous les membres de ma famille.

Spéciale dédicace pour la personne Mohammed Abdallah qui a été très

utile pour remplire ce mémoire

A tous mes ami(e)s surtouts ***Fatima****,* ***Houda*** *,****Asma*** *et* ***Hadjer*** *et* ***Yousra*** *,tous ceux que j'aime, tous ceux qui m'aiment et tous ceux qui me sont chers .Et surtout les étudiants de la promotion de télécom , Et a toutes les personnes qui m'aine.*

Baalla Soumia

Liste des acronymes

RAP	Reconnaissance Automatique de la Parole
LP	Prédication Linéaire
LPC	Linear Prédictive Coding
MFCC	Mel-Scale Frequency Cepstral Coefficients
MLE	Maximum Likelihood
PLP	Perceptual Linear Predictive
RN	Réseaux de neurones
CNN	Réseaux de Neurones à Convolution
SAMPA	Speech Assessment Method Phonetic Alphabet
SR	Système de Reconnaissance
TF	La Transformée de Fourier
TIMIT	Texas Instruments Massachusetts Institute of Technology
TRF	Transformée de Fourier Rapide
MCC	Code de Pays Mobile
MLP	Multi-Layer Perceptron

Table des matières

Liste des Figures

Liste des tableaux

ملخص

كانت المعالجة التلقائية للكلام مجالً للدراسة النشطة منذ الخمسينيات من القرن الماضي، وهو ما يفسر ثراءها ولكن أيضًا الصعوبة. إنه يجمع العديد من التخصصات والتقنيات. تعقيد إشارة الكلام الناتج عن التفاعل بين إنتاج الأصوات وإدراكها من خلال الأذن مما يساهم في صعوبة التعرف التلقائي على الكلام، والذي أصبح موضوع بحث مثير للاهتمام للغاية.

الهدف من هذه الرسالة هو الحصول على قاعدة بيانات وتنفيذها، تتكون من مجموعة من الجمل الصحيحة نحويًا ودلالة. تم تسجيل قاعدة البيانات هذه في ظل الظروف الحقيقية للتحليل الصوتي لقاعدة البيانات هذه تم إجراؤها باستخدام طريقة MFCC وزودتنا بسلسلة من ناقلات الإدخال لنظام التعرف التلقائي على الكلام المشار إليه (RAP) .التي قمنا بتطويرها أيضًا. يعتمد ذلك على استخدام الشبكات العصبية التلافيفية (CNN)

سيسمح تقييم أداء نظام RAP لطريقة تحليل قاعدة البيانات بإبراز تأثير المعلمات.

الكلمات المفتاحية : التعرف التلقائي على الكلام ، الشبكات العصبية التلافيفية ، برنامج ماتلاب.

Résumé

Le traitement automatique de la parole est un domaine d'étude active depuis les années 50 se qui explique sa richesse mais aussi sa difficulté. Il fait collaborer plusieurs disciplines et techniques. La complexité du signal de la parole qui résulte de l'interaction entre la production des sons et leur perception par l'oreille ce qui contribue dans la difficulté de la reconnaissance automatique de la parole, qui est devenue un sujet de recherche très intéressan.

L'objectif de ce mémoire est l'acquisition et la mise en œuvre d'une base de données, constituée d'un corpus de phrases syntaxiquement et sémantiquement correctes. Cette base de données a été enregistrée dans des conditions réelles de l'analyse acoustique de cette base de données a été effectuée en utilisant la méthode MFCC et nous a fourni une série de vecteurs d'entrée pour le système référer de Reconnaissance Automatique de la Parole (RAP) que nous avons, par ailleurs, élaboré. Celui-ci est basé sur Réseaux neurones convolutionnels. L'évaluation des performances du système RAP pour la méthode d'analyse de la base de données permettra de mettre en exergue influence de la paramétrisation.

Mots clés : Reconnaissance Automatique de la Parole, Réseaux neurones

Convolutionnels, Programme Matlab.

Abstract

Automatic speech processing has been a field of active study since the 1950s, which explains its richness but also its difficulty. It brings together several disciplines and techniques. The complexity of the speech signal which results from the interaction between the production of sounds and their perception by the ear which contributes to the difficulty of automatic speech recognition, which has become a very interesting research subject.

The objective of this thesis is the acquisition and implementation of a database, consisting of corpus of syntactically and semantically correct sentences. This database was recorded in real conditions of the acoustic analysis of this database was carried out using the MFCC method and provided us with a series of input vectors for the referenced Automatic Speech Recognition system. (RAP) that we have also developed. This one is based on convolutional neural networks (CNN). The performance evaluation of the RAP system for the database analysis method will highlight the influence of the parameterization.

Key words: Automatic Speech Recognition, convolutional neural networks, Program Matlab.

Introduction générale

De nos jours, le traitement de la parole est devenu un élément fondamental de l'ingénierie. Située au carrefour du traitement numérique du signal et du traitement du langage (c'est-à-dire le traitement des données symboliques), cette discipline scientifique a connu une expansion rapide liée au développement des moyens et des technologies de télécommunication depuis les années 1960. [1]

L'objectif de la reconnaissance automatique de la parole (RAP) est de développer des technologies et des systèmes qui permettent aux ordinateurs d'accepter la parole comme entrée. Par conséquent, la reconnaissance vocale peut être considérée comme l'opération de conversion du signal vocal à partir des informations contenues dans le signal et de la connaissance préalable du domaine en texte.

Les applications RAP sont diverses et peuvent être grossièrement divisées en quatre catégories : applications téléphoniques, applications multimédias, applications industrielles, applications médicales, etc.

Ils existent là où la voix peut remplacer ou compléter des interfaces existantes pour communiquer avec des machines, par exemple pour accéder à des services ou contrôler les fonctions d'équipements. La voix est parfois le seul moyen de communication, par exemple dans les applications mains libres où l'utilisateur ne touche pas l'appareil.

Nous évaluonslesreprésentationsintermédiairesapprisesparnotre meilleur système de prédiction CNN par rapport à différents facteurs en utilisant deux approches d'évaluation : une évaluation par classification et une évaluation par visualisation. Nous essayons ensuite de tirer profit de cette analyse en proposant un apprentissage multi-tâche qui prend implicitement les informations détectées.

L'objectif de notre travail nous avons appliqué les réseaux de neuronesconvolutifs pour la reconnaissance des phonèmes spécifiques à la langue Arabe Standard. Par Reconnaissance Automatique de la Parole (RAP)

Afin de bien présenter nos travaux, nous avons organisé cette thèse en trois chapitres:

- Dans le premier chapitre, nous décrivons tout d'abord les caractéristiques du signal de la parole, et nous présenterons les méthodes d'analyse les plus adaptées à la

RAP :

- Par prédiction lineaire (LPC, Linear Prédictive Coding),
- Par extraction des coefficients cepstraux à l'echellemell (MFCC, Mel-scale Frequency Cepstral Coefficients),
- Par prédiction linéaire perceptuelle (PLP, Perceptual Linear Predictive).

Dans le deuxième chapitre, nous introduisons les notions mathématiques et les bases nécessaires pour utiliser le réseau neuronal de convolution (CNN) pour la reconnaissance de la parole.

Le dernier chapitre présente spécifiquement les différents résultats obtenus en appliquant les méthodes étudiées dans cette mémoire.

Chapitre I

La Reconnaissance Automatique De La Parole

I.1 Introduction

Dans le traitement de la parole, le signal de parole doit d'abord être transformé et compressé pour un traitement ultérieur. Il existe de nombreuses techniques d'analyse de signal qui sont utilisées pour extraire les caractéristiques importantes et compresser le signal sans perdre aucune information importante [2] [3].

Dans ce chapitre, nous expliquerons d'abord les caractéristiques particulières du signal vocal, puis présenterons les méthodes d'analyse les plus appropriées pour la reconnaissance automatique de la parole.

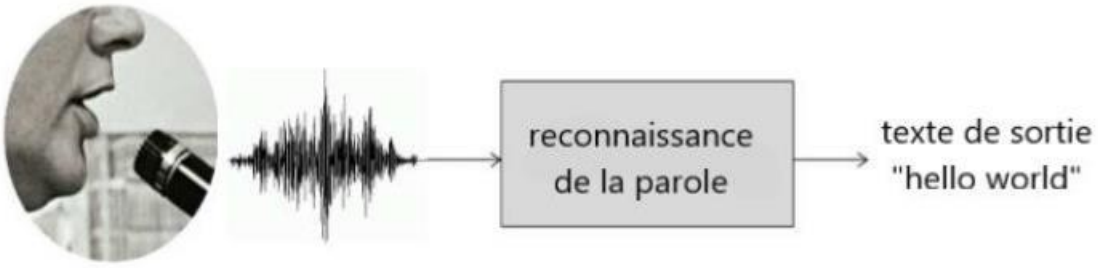

Figure I.1.Mécanisme du système de reconnaissance vocal.

I.2 Signal De La Parole

Avant de passer au processus de codage du son dans l'ordinateur ou sa synthèse, il faut d'abord bien comprendre le son lui-même. Nous commencerons donc par la définition du son.

I.2.1. Définition

Le son est une vibration de l'air. A l'origine de tout son, il y a un mouvement (par exemple, une corde qui vibre, une membrane de haut-parleur...). Il s'agit de phénomènes oscillatoires crées par une source sonore qui met en mouvement les molécules de l'air. Avant d'arriver jusqu'à notre oreille, ce mouvement se transmet entre les molécules à une vitesse de 331 m/s à travers l'air à une température de 20°C, c'est ce que l'on appelle la propagation.

I.2.2. Les caractéristiques du son

Le son est défini par trois paramètres :

➢ L'amplitude du son correspondant à la différence de pression maximale de

L'atmosphérique due aux oscillations, (volume sonore).

➢ La dynamique qui permet de mesurer la différence entre le volume sonore maximum et le bruit de fond. La dynamique se mesurer en décibels (dB). En fait, c'est le rapport entre le niveau maximum, à réduction de distorsion, et le niveau minimum acceptable, à la limite du niveau de bruit de fond.

➢ Le timbre qui est un paramètre beaucoup plus subjectif, est-ce qui distingue deux sons de même hauteur et amplitude. C'est une idée qualitative, qui dira, par exemple, que le son est clair ou profond ... [4]

I.2.3. Les avantages de la parole

Si on prend comme référence le modèle humain, les avantages de la parole semblent au premier abord déterminant :

- *Naturel:* la parole constitue le mode le plus naturel de communication entre personnes humaines, du fait que son apprentissage s'effectue dès l'enfance, ce qui est loin d'être le cas pour la maîtrise de l'écriture.

- ***Rapidité/efficacité:*** plusieurs études d'ergonomie montrent que le débit en parole spontanée est de l'ordre de 200 mots/minute à comparer aux 60 mots/minute d'un expert pour la frappe au clavier. L'efficacité de la parole ne provient pas seulement de ce qu' 'elle permet un débit d'informations plus élevé que d'autres modes de communication, mais également de ce qu'elle peut être aisément utilisée en superposition avec ceux-ci. La parole laisse l'utilisateur libre de ses mouvements, elle est donc particulièrement adaptée aux applications dans lesquelles il s'agit pour l'utilisateur de conduire plusieurs tâches simultanément, ou de contrôler des processus complexes qui monopolisent gestes et/ou vision. [5]

- ***Extension du champ d'action*****:** la parole permet d'avoir un accès immédiat à une information sans avoir à parcourir toute une arborescence hiérarchique de menus. Il est toujours possible bien sûr d'utiliser des raccourcis clavier ou des langages de commande, mais la prononciation d'un seul mot peut remplacer jusqu'à une dizaine de commandes élémentaires effectuées à l'aide de touches fonctions ou de souris et représente ainsi un effort mnémotechnique moindre. [6]

Ces avantages sont si importants qu'il existe déjà sur le marché des appareils à usage limité, mais ils sont néanmoins efficaces. Citons quelques-unes des applications qui sont déjà apparues : [7]

• Saisie de données audio.

• Donne des ordres en conduisant une voiture ou un avion.

• Aide aux handicapés.

• Chambre d'hôpital avec capacités de commande vocale pour les patients.

• Commande vocale de machines ou robots.

• Commande vocale dans une montre portable, etc.

I.2.4. Complexité du signal de parole

Pour appréhender le problème de la reconnaissance automatique de la parole, il faut

comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile.

Le signal de la parole n'est pas un signal ordinaire, il est le vecteur d'un phénomène

extrêmement complexe, la reconnaissance automatique de la parole pose de nombreuxproblèmes. D'un point de vue mathématique il est difficile de modéliser le signal de la parolecar ses propriétés statistiques varient au cours du temps.

La complexité du signal de parole provient de la combinaison de plusieurs facteurs,

principalement la redondance du signal acoustique, la grande variabilité intra-locuteurs et interlocuteurs, et les effets de la coarticulation en parole continue, qui doivent être pris en compte lors de la conception d'un système de RAP. [7]

I.2.4.1.Redondance du signal de parole

Le signal de parole est extrêmement redondant. Cette grande redondance lui confère une robustesse à certains types de bruits. De nombreuses recherches sont menées afin de rendre les systèmes de reconnaissance robustes aux bruits, mais les performances humaines sont encore loin d'être atteintes. [4]

I.2.4.2.Continuité et coarticulation

Tout discours peut être retranscrit par des mots, qui peuvent à leur tour être décrits

comme une suite de symboles élémentaires appelés phonèmes par les linguistes. Cela laisse supposer que la parole est un processus séquentiel, au cours duquel des unités indépendantes se succèdent. Malheureusement, les spécialistes de phonétique eux-mêmes ont parfois des difficultés à identifier individuellement ces unités discrètes dans le signal, même si

quelques événements acoustiques particuliers peuvent être détectés. La parole est en réalité un flux continu, et il n'existe pas de pause entre les mots qui pourrait faciliter leur localisation automatique par les systèmes de reconnaissance. De plus, les contraintes introduites par les mécanismes de production créent des phénomènes de coarticulation. La production d'un son est fortement influencée par les sons qui le précèdent mais aussi qui le suivent en raison de l'anticipation du geste articulatoire. Ces effets s'étendent sur la durée d'une syllabe, voire même au-delà, et sont amplifiés par une élocution rapide. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. La prise en compte des phénomènes de coarticulation ne suffit pourtant pas à prédire la réalisation acoustique d'une phrase en raison de la grande variabilité de la parole. . [4] [7]

I.2.4.3.Variabilité

On distingue généralement deux sources de variabilité qui peuvent rendre deux

prononciations d'un même énoncé très différentes, la variabilité interlocuteurs et la variabilité intra-locuteur. . [7]

➢***Intra-locuteurs*** : 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 nasales ;

une simple fatigue et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps : l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix de leurs sceaux. . [7]

➢***Interlocuteurs :*** Est encore plus flagrante. Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonnantes, modifient la production acoustique. En plus, il y a la hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. . [7]

I.3 Méthode D'analyse Du Signal Vocale

La parole est un signal acoustique qui contient des informations d'idée qui se forment dans l'esprit du locuteur. Le signal vocal transfère non seulement des informations sur la langue, mais donne également des informations sur les sons, la syntaxe et le locuteur, c'est-à-dire l'âge, le sexe, l'origine locale, la santé, l'état émotionnel (humeur du locuteur) etsa caractéristique unique. Une reconnaissance automatique de la parole (RAP) ne prend en

compte que les informations acoustiques contenues dans le signal vocal, d'où la nécessitéd'une analyse acoustique.

L'objectif de l'analyse acoustique est d'extraire des coefficients représentatifs du signal de parole. Ces coefficients sont calculés à intervalles de temps réguliers. En toute simplicité, le signal de parole est transformé en une série de vecteurs de coefficients, ces coefficients doivent représenter au mieux ce qu'ils sont censés modéliser et doivent extraire le maximum d'informations utiles à la reconnaissance. Cette analyse nécessite un prétraitement du signal de parole avant le calcul des paramètres. . [8]

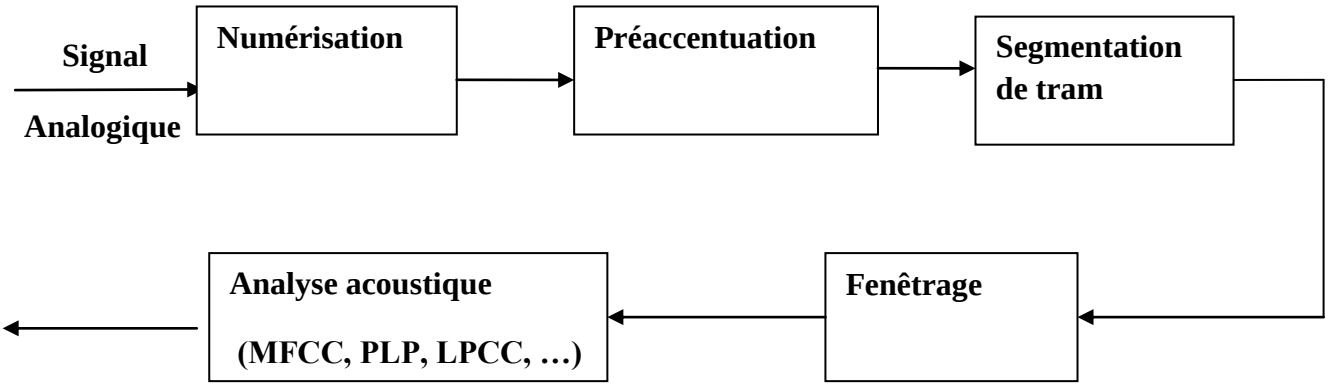

Figure I.2. Prétraitement du signal vocal.

I.3.1. Prétraitement Du Signal Vocal

I.3.1.1. Numérisation

Les signaux que nous utilisons dans le monde réel, comme notre voix, sont des signaux analogiques continus en temps et en amplitude. Pour traiter ces signaux pour la communication numérique, nous devons les convertir sous forme "numérique" (discrète en temps et en amplitude).

Pour cela, nous utilisons un processus appelé théorème d'échantillonnage de Nyquist–Shannon, la fréquence d'échantillonnage fs est supérieure ou égale à la composante de fréquence la plus élevée du signal de message:

$$Fe \geq 2Fmax \qquad \text{(I.1)}$$

Avec,

Fe: la fréquence d'échantillonnage

$Fmax$: la fréquence maximum du signal vocal

I.3.1.2. Préaccentuation

La préaccentuation est une technique utilisée dans le traitement de la parole pour améliorer les hautes fréquences du signal. Il réduit la plage dynamique spectrale élevée.

Dans cette étape, le signal passe à travers un filtre passe-haut pour aplatir spectralement le signal et le rendre moins sensible aux effets de précision finie. La fonction de transfert de ce filtre est :

$$H\,z = 1 - az^{-1} \qquad \text{(I.2)}$$

Avec : $0.9 < a < 0.98$

I.3.1.3. Segmentation de tram

Les caractéristiques statistiques d'un signal de parole sont généralement invariantes dans un court intervalle de temps. Ainsi, le signal pré-accentué est bloqué ou segmenté en trames. La longueur de trame la plus utilisée est espacée de 20 à 25 ms (millisecondes).

I.3.1.4. Fenêtrage

Une fois la procédure de segmentation de trame terminée, à chaque trame une fonction de fenêtrage est appliquée pour supprimer l'effet des discontinuités aux bords des trames.

Le fenêtrage a pour but de réduire l'effet des artefacts spectraux qui résultent du processus de cadrage.

Le fenêtrage dans le domaine temporel est une multiplication ponctuelle de la trame et de la fonction de fenêtre. Par conséquent, chaque trame est multipliée par une fonction de fenêtre w [n] de longueur N + 1, où N est la longueur de trame. La fenêtre la plus populaire est la fenêtre de Hamming donnée par :

$$W[n] \begin{cases} = & 0.54 - 0.46\cos(2\pi n/N) & ,0 \le n \le N \\ & 0 & ,\ ailleurs \end{cases} \qquad \text{(I.3)}$$

N : taille de la fenêtre

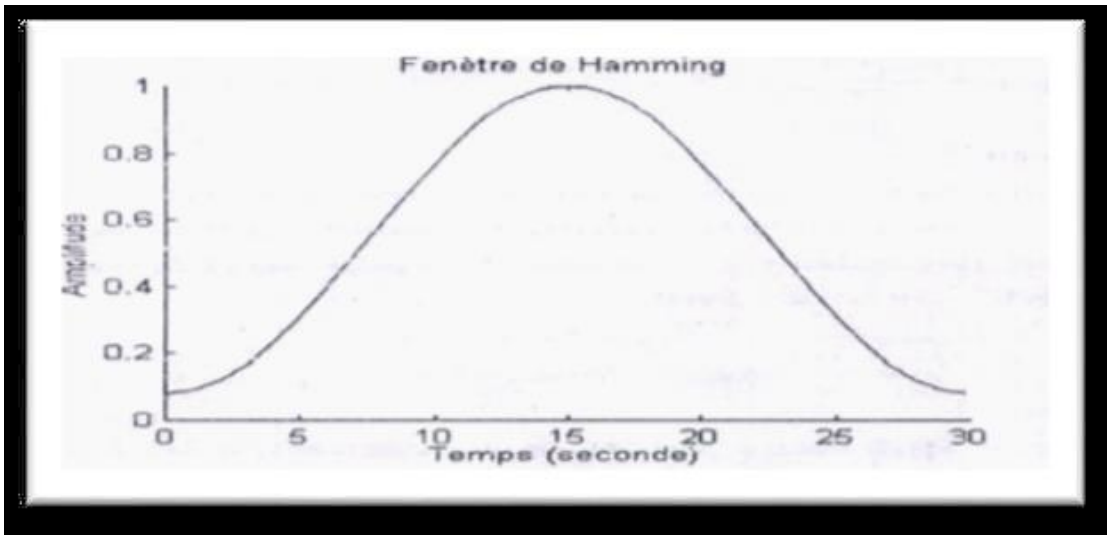

Figure I.3 : La fenêtre de Hamming.

I.3.2. Analyse par transformée de Fourier

L'analyse fréquentielle de la parole se ramène aux opérations de la transformée de Fourier (TF) et n'a d'intérêt que si elle s'applique à une période stable du signal vocal, donc sur une période assez courte. Le spectre à court terme du signal *s(n)* se calcule à partir d'une fenêtre *h(n)* qui permet d'isoler une portion du passé récent de *s(n)* :

$$\sum_{k=-\infty}^{k=+\infty} s(n)h(n-k)\exp(-jwn) \qquad (\text{I}.4)$$

La quantité |S (*w*, *n*)|² est le spectre de puissance à court terme. L'implantation algorithmique efficace associée à la TF est la transformée de Fourier rapide (TFR). Elle présente de nombreux avantages en tant que méthode d'analyse fréquentielle. La rapidité de sa mise en oeuvre l'a propulsé au rang d'élément incontournable des systèmes de traitement du signal. La TFR permet aussi une représentation fréquentielle du signal aussi fine que l'on souhaite. De plus, pour une étude qualitative de la parole, la TFR est très intéressante parce qu'elle permet une représentation par spectrogramme (évolution du spectre dans le temps) de qualité. Mais, après la naissance de la notion de représentation temps-fréquence, des études théoriques ont permis de mettre à jour quelques désavantages de la TFRquisont impossibles à éliminer et qui constituent ainsi les limites de l'exploitation de la TFR [5].

Malheureusement les limites théoriques relatives aux représentations temps fréquence ne sont pas les seuls problèmes de la TF. Le défaut majeur de la TF pour l'étude de la parole vient de l'inévitable intermodulation source/conduit présente dans le spectre qui ne permet pas de connaître précisément la hauteur du fondamental. Cette intermodulation est due à la

convolution qui est réalisée par le conduit vocal sur la fréquence fondamentale produite par les cordes vocales. La déconvolution ne pouvant pas être réalisée par une simple transformée, il a donc fallu développer une technique particulière capable de la réaliser pour fournir ces deux informations utiles à l'analyse de la parole. L'étude des représentations temps-fréquence et les limites de la TF ont donc poussé à créer des méthodes de traitements de signal plus adaptées à la parole. [9]

I.3.3. Analyse par prédictif linéaire LPC

Le codage prédictif linéaire (LPC, Linear Prédictive Coding) est une technique de codageet de représentation de la parole Markel et Gray. Elle s'appuie principalement sur l'idée que le système phonatoire peut être modélisé par un filtre linéaire .Ce filtre est excité par un train d'impulsions pour les sons voisés et aléatoires pour les sons non voisés. Il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents : [9]

$$s(w,n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n-k) \qquad (I.5)$$

u(n) étant l'entrée du filtre prédicteur suivant :

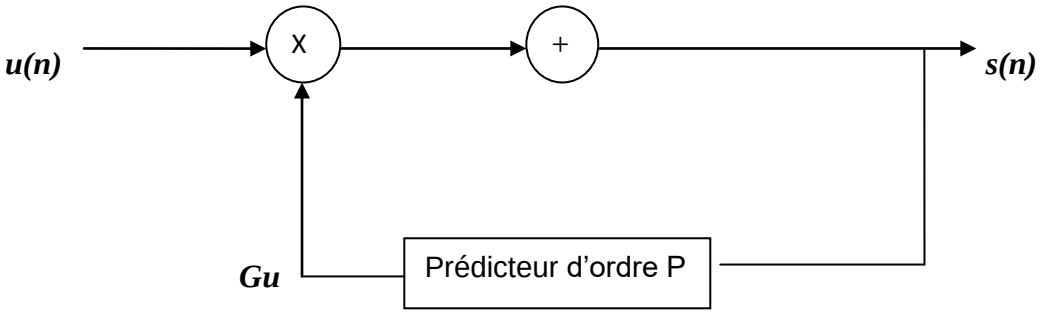

Figure I.4: Filtre prédicteur linéaire.

En supposant que cette entrée *u(n)* soit totalement inconnue, le signal de parole à un instant n peut être prédit par une combinaison linéaire des *p* échantillons précédents, soit :

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (I.6)$$

L'erreur de prédiction *e*(*n*) est définie par :

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \qquad (I.7)$$

L'énergie résiduelle de prédiction est définie par la somme :

$$E_{p=}\sum_{n1}^{n2} e^2\,(n) = \sum_{n_1}^{n_2}\left(s(n) + \sum_{k=1}^{p} a_k\, s(n-k)\right)^2 \qquad \text{(I. 8)}$$

Si les coefficients *ak*sont choisis tels qu'ils minimisent l'énergie résiduelle de prédiction, il suffît pour les obtenir de poser :

$$\frac{\partial E_p}{\partial a_i} = 0 \qquad i = 1,2,\ldots\ldots p \qquad \text{(I. 9)}$$

Le calcul de cette relation (équation I.9) conduit aux équations suivantes :

$$\sum_{k=1}^{p} a_k\, \Phi_{ik} = -\boldsymbol{\Phi_{i0}} \qquad i = 1,2,\ldots..p \qquad \text{(I. 10)}$$

$$\Phi_{ik} = \sum_{n=n_1}^{n_2} s(n-i)s(n-k) \qquad \text{(I. 11)}$$

Ces équations normales (équation I.10), dites de Yule Walker, constituent un système linéaire de *P* équations à *P* inconnues. La résolution de ce système permettra d'obtenir les coefficients*à k* du filtre. Parmi les méthodes de minimisation de l'énergie résiduelle de prédiction donc de résolution du système, on trouve principalement la méthode d'autocorrélation et la méthode de covariance.

I.3.3.1. Méthode d'autocorrélation

En raison de la nature variant dans le temps du signal de parole, les coefficients de prédiction doivent être estimés à partir de courts segments de signaux de parole (10 à 40 ms) où les caractéristiques des signaux de parole sont constantes dans cette plage [9].

L'énergie résiduelle de prédiction *Ep*, définie dans (l'équation I.8), est minimisée sur une durée infinie :

$$E_p = \sum_{n=-\infty}^{+\infty} e^2(n) \qquad \text{(I. 12)}$$

La fonction d'autocorrélation est définie par :

$$R(i) = \sum_{n=-\infty}^{+\infty} s(n)s(n-i) \qquad \text{(I. 13)}$$

R(i)=-R(i)

Le signal vocal est défini pour toutes les valeurs du temps ; il est identiquement nul en dehors d'une séquence de N échantillons ceci équivaut à multiplier le signal vocal par une fenêtre de longueur finie correspondant à N échantillons. La sommation infinie de (l'équation I.12) se ramène donc à une somme finie, soit :

$$E_p = \sum_{n=0}^{N-1+p} e^2(n) \qquad \text{(I. 14)}$$

L'équation I.11 devient alors :

$$\Phi_{ik} = \sum_{n=0}^{+\infty} s(n-i)s(n-k) = \sum_{n=-\infty}^{+\infty} s(n)s(n+i-k) \qquad \text{(I. 15)}$$

Dans ce cas *ik*n'est autre que la fonction d'autocorrélation évaluée pour (i− *k*), soit :

$$\Phi_{ik} = R(i-k) \qquad \text{(I. 16)}$$

Le système donné par l'équation I.10 s'écrira alors sous la forme suivante :

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i) \qquad i = 1,2,\ldots.,p \qquad \text{(I. 17)}$$

La méthode d'autocorrélation est largement utilisée parce que d'une part elle conduit à un système d'équations d'une structure particulière. La matrice des valeurs d'autocorrélation est une matrice Toeplitz. En effet, elle est symétrique et tous les éléments sur chaque diagonale sont identiques, ce qui facilite la résolution. D'autre part, elle assure la stabilité du modèle auto-régressif trouvé. Différents algorithmes permettent en effet la résolution du système (équation I.17) par une récursion sur l'ordre de prédiction. Parmi ces algorithmes nous pouvons citer l'algorithme de Levinson-Durbin et l'algorithme de Leroux-Gueguen [10] [11].

I.3.3.2. Méthode de covariance

Le signal est étendu par p échantillons en dehors de la plage normale de 0≤n≤N-1 pour inclure p échantillons se produisant avant n = 0 (ils sont disponibles) et élimine le besoin d'une fenêtre effilée Dans cette méthode l'énergie de prédiction *P E* est minimisée sur une durée finie :

$$E_p = \sum_{n=M_1}^{M_2} e^2(n) \qquad (I.18)$$

Soit :

$$E_p = \sum_{n=0}^{N-1} e^2(n)$$

La relation (équation I.11) devient alors :

$$\Phi_{ik} = \sum_{n=0}^{N-1} s(n-i)s(n-k) \qquad (I.19)$$

Φ_{ik} définie dans (l'équation I.18) n'est rien d'autre que la fonction de covariance σ (i, k) .

Ainsi le système de (l'équation I.10) s'écrira alors :

$$\sum_{k=1}^{p} a_k \sigma(i,k) = -\sigma(i,0), \qquad i = 1,2,\ldots\ldots,p \qquad (I.20)$$

Dans ce cas la matrice (P *x* P) des valeurs de covariance est symétrique mais non de Toeplitz. Dans cette méthode, les propriétés statistiques de l'estimateur sont meilleures que dans le cas de la méthode d'autocorrélation, mais la matrice a une structure moins simple que celle de Toelplitz ce qui rend la résolution des équations de Yule-Walker un peu plus coûteuse. Un autre inconvénient de cette méthode est qu'elle ne garantit pas la stabilité du modèle auto-régressif trouvé. Parmi les algorithmes de résolution nous pouvons citer l'algorithme de Gram-Schmit et l'algorithme de Morf [12] [13] [14].

I.3.4. Analyse cepstrale

L'intermodulation source conduit observée sur le spectre calculé par TFR rend sonutilisation pour la mesure des formants *Fi* et de la fréquence fondamentale *F0* très difficile. Le lissage cepstrale est une méthode qui vise à séparer la contribution du conduit et de la source d'excitation par déconvolution. Pour cela on fait l'hypothèse que le signal vocal *s(n)* est produit par un signal excitateur *g(n)* (source glottique) traversant un système linéaire passif de réponse impulsionnelle *b(n)* (conduit). Avec ces hypothèses on peut écrire :

$s(n) = g(n) \otimes b(n)$ (I.21)

Pour déconvoluer plus aisément *s(n)* il suffit de transposer le problème par

homomorphisme dans un espace ou l'opérateur □(convolution) correspond à un opérateur + (addition).

En pratique cette transposition par homomorphisme est réalisée par les étapes schématiséessur la figure suivante :

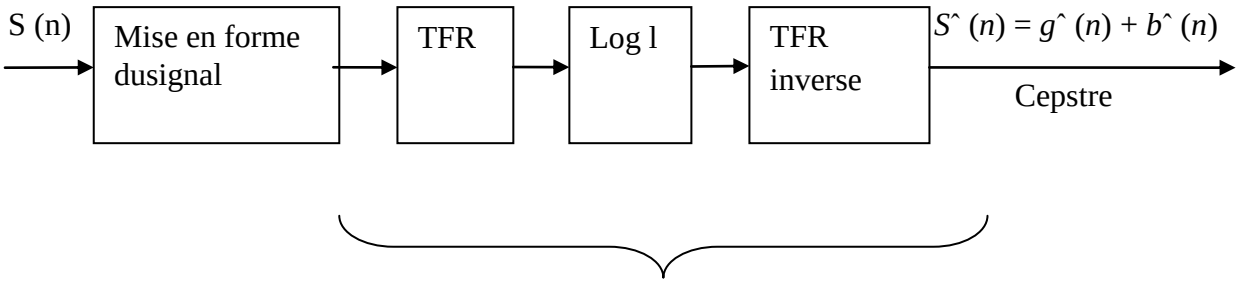

Figure I.5 : Analyse homomorphique de la parole.

Où $\hat{S}(n)$ sont les coefficients cepstraux approchés prenant leurs valeurs dans un domaine pseudo-temporel réel appelé domaine quéfrentiel. La structure de la parole et les hypothèses sur la source d'excitation et du conduit vocal permettent de dire que :

$\hat{g}(n)$ Se réduit théoriquement à une séquence d'impulsions de période *n0* (*n0* correspond a la fréquence fondamentale *F0*).

$\hat{b}(n)$ décroît rapidement (en $1/n$) avec n et devient négligeable pour $n > n0$.

Dans ces conditions, on peut admettre que la contribution du conduit est localisée dans les basses quéfrences ($n < no$) et que la séquence d'impulsions reflète la contribution de la source. Cette méthode de calcul des cepstres est élémentaire, il existe d'autres méthodes itératives effectuant un lissage, ce qui permet d'obtenir des cepstres de meilleure qualité. [12]

I.3.4.1. Calcul des cepstres à partir des coefficients LPC

Le calcul des cepstres par la méthode décrite précédemment (analyse homomorphique) est généralement moins utilisé du fait de la charge de calcul importante associée au calcul de la TFR et de la TFR inverse. On lui préfère une méthode paramétrique. Cette méthode paramétrique permet de déterminer les coefficients cepstraux des signaux de parole à partir des coefficients LPC. C'est ainsi que le signal de parole est considéré comme engendré par un filtre auto-régressif (AR) dont il faut déterminer les coefficients ai en utilisant des méthodes classiques de prédiction linéaire comme la méthode d'autocorrélation par exemple. Le calcul des cepstres est alors basé sur une procédure récursive liant les coefficients cepstraux (Cm) et

les coefficients de prédiction (ai). Cette procédure récursive est traduite par les équations suivantes :

$$\ln\left[\frac{1}{A_p(z)}\right] = \sum_{n=1}^{\infty} c_q\,(n) z^{-n} \qquad (I.22)$$

$c_0 = \ln G^2$, G^2 est le gain du modèle LPC (I.23)

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \quad \text{,pour } 1 \leq m \leq P \qquad (I.24)$$

$$c_m = + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \quad \text{,pour } m > P \qquad (I.25)$$

I.3.4.2. MFCC (Mel-scale Frequency Cepstral Coefficients)

Les coefficients MFCC (Figure I.6) sont une extension des coefficients cepstraux par le passage de l'échelle fréquentielle linéaire à une échelle fréquentielle non linéaire proche de l'audition humaine. Cette échelle non linéaire est l'échelle perceptive Mel. Celle-ci est plus exactement linéaire pour les basses fréquences (inférieures à 1000Hz) et logarithmique pour les hautes fréquences. C'est ainsi que des filtres répartis linéairement en basses fréquences et logarithmiquement en hautes fréquences (Figure I.7) sont utilisés afin de capturer les caractéristiques phonétiques importantes du signal de parole. Ces filtres possèdent la caractéristique suivante : plus la fréquence est élevée, plus la bande passante est large ce qui permet une meilleure résolution temporelle des hautes fréquences.

L'échelle *Mel* peut être définie par la relation suivante entre la fréquence en Hertz et sa correspondante en mels :

$$B(f) = x \log_{10}\left(1 + \frac{f}{y}\right) \qquad (I.26)$$

Où f est la fréquence en Hz, $B(f)$ est la fréquence en échelle mel de f.

Plusieurs valeurs sont utilisées pour x et y par exemple x =1000/log(2) et y =1000. De nos jours, les valeurs les plus couramment utilisées sont x =2595 et y =700.

Soit un signal discret $\{x[n]\}$ avec $0 \leq n \leq N-1$, N est le nombre d'échantillons d'une fenêtre analysée, Fe est la fréquence d'échantillonnage, la transformée de Fourier discrète $S[k]$ est obtenue par :

$$s[k] = \sum_{n=0}^{N-1} x[n] e^{-j 2\pi nk / N} \qquad (I.27)$$

Avec : $0 \leq k < N$

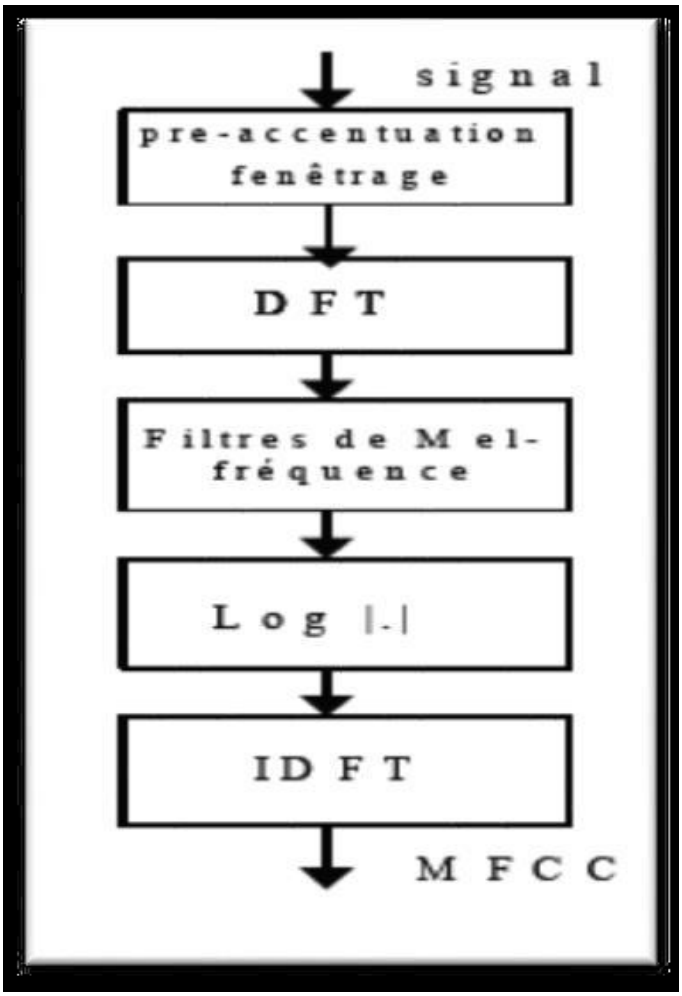

Figure I.6 : Calcul des MFCC.

Le spectre du signal est multiplié par des filtres triangulaires (Figure I.7) dont les bandes passantes sont équivalentes en domaine mel-fréqunce. Les points frontières $B[m]$ des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m \frac{B(f_h) - B(f_l)}{M + 1} \qquad (I.28)$$

Avec : $0 \leq m \leq M+1$

Où *M* est le nombre de filtres, *fh*est la fréquence la plus haute et *fl*est la fréquence la plus basse pour le traitement du signal. Dans le domaine fréquentiel, les points *f*[*m*] discrets correspondants sont calculés par l'équation :

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad (I.29)$$

Avec :$B^{-1}(b) = 700 * \left(10^{b/2595} - 1\right)$

Où B^{-1}est la transformée de mel-fréquence en fréquence.

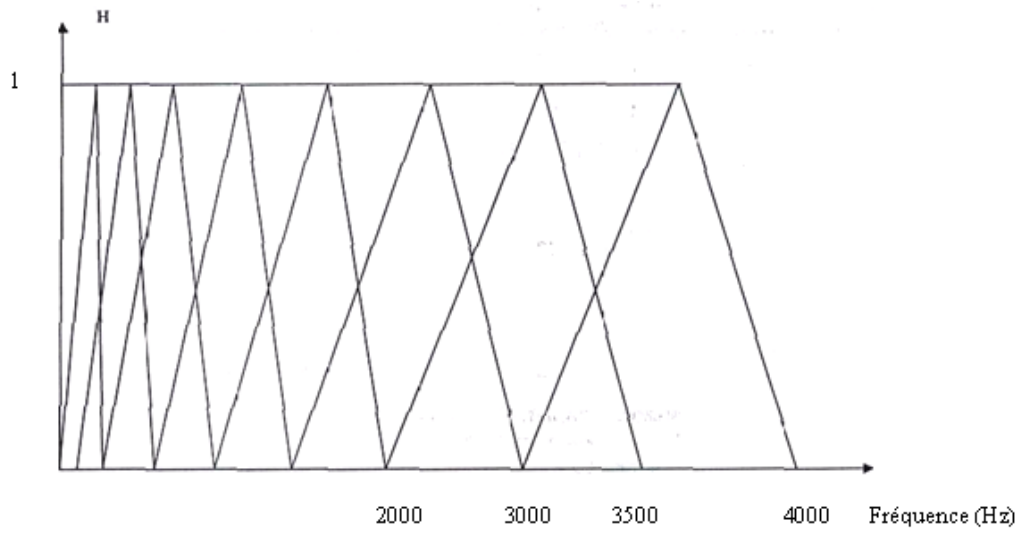

Figure I.7 : Bancs de filtres Mel.

Le coefficient *Hm*[*k*] de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \le k \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1] \end{cases} \qquad (I.30)$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme d'énergie (ou unlogarithme de spectre d'amplitude) est calculé :

$$E[m] = \log\left[\sum_{k=0}^{N-1} |s[k^2] H_m[k]\right] \qquad (I.31)$$

Avec : $0 \le m < M$

Les coefficients cepstraux de mel-fréquence (MFCCs) peuvent être obtenus par unetransformée de Fourier inverse à partir des coefficients aux sorties des filtres. Mais le

nombre de MFCCs est moins grand que le nombre des filtres, donc une transformée de cosinus discrète est plutôt utilisée :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M_s}\right) \qquad (I.32)$$

Avec : $0 \leq n < M$

I.3.4.3. Les coefficients PLP (PerceptualLinearPredictive)

PLP (PerceptualLinearPredictive) est une technique d'analyse de la parole , fondée sur la modélisation du spectre par un modèle tout pôle suivant un principe identique à la technique de prédiction linéaire (LP). Cependant, la différence réside dans le fait que les paramètres d'un filtre auto-régressif tout pôle sont estimé en modélisant au mieux le spectre auditif. Ceci est fondé sur trois effets auditifs : sélectivité spectrale de bande critique, courbe d'intensité égale et loi de puissance. La (Figure II.8) représente le processus de calcul des coefficients PLP [12].

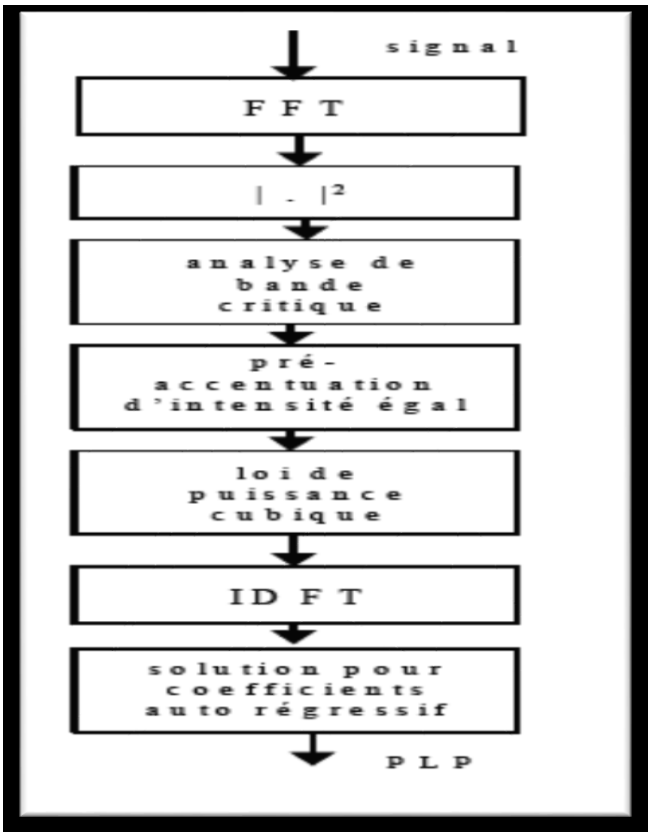

Figure I.8 : Processus de calcul des coefficients PLP

Pour obtenir un spectre auditif, la courbe de masquage$\Psi(\Omega)$ est tout d'abord utilisée

$$\Psi(\Omega) = \begin{cases} 0 & si \quad \Omega \leq 1.3 \\ 10^{2.5(\Omega+0.5)} & si -1.3 \leq \Omega \leq -0.5 \\ 1 & si -0.5 \leq \Omega \leq 0.5 \\ 10^{2.5(\Omega-0.5)} & si \ 0.5 \leq \Omega \leq 2.5 \\ 0 & si \Omega \geq 2.5 \end{cases} \quad (I.33)$$

Où **Ω**est la fréquence de Bark calculée à partir de la fréquence angulaire *w* par la définition :

$$\Omega(w) = 6 \ln\left(\frac{w}{1200\pi} + \left(\left(\frac{w}{1200\pi}\right)^2 + 1\right)^{1/2}\right) \quad (I.34)$$

Le spectre de puissance du signal $p(\omega)$ (pair et périodique) est convolué avec la courbe de masquage :

$$\Theta(\Omega_k) = \sum_{\Omega=-1.3}^{\Omega=2.3} (\Omega - \Omega_k)\psi(\Omega) \quad (I.35)$$

Puis, l'algorithme tente de faire l'approximation de la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert $E(\Omega)$:

$$\Xi(\Omega) = E(\Omega)\Theta\big(\Omega(\omega)\big) \quad (I.36)$$

La non linéarité entre l'intensité d'un son et son niveau de perception par l'oreille est réalisé en l'approchant par une loi de puissance :

$$\Phi(\Omega) = \Xi\left(\Omega^{\frac{1}{3}}\right) \quad (I.37)$$

Enfin le spectre auditif est modélisé par un modèle tout-pôle. Une transformée de Fourier inverse discrète est appliquée sur le spectre auditif$\Phi(\Omega)$pour obtenir les valeurs d'autocorrélation. M+1 premiers coefficients d'autocorrélation sont utilisés pour calculer les coefficients auto régressifs du modèle tout-pôle d'ordre M qu'on appelle les coefficients PLP.

I.4. Conclusion

Dans ce chapitre, nous avons présenté les mécanismes de production et de perception de la parole chez les humains ainsi que les caractéristiques du signal vocal responsables en grande partie des difficultés de la tâche de reconnaissance. Par la suite, nous avons décrit les différentes approches de reconnaissance avec le processus de traitement relatif à chacuned'entre elles. C'est ainsi que nous avons présenté les méthodes d'analyse du signal de parole les mieux adaptées et les plus utilisées.

Chapitre II

Les
Réseaux De Neurones

II.1. Introduction

Dans le cadre de notre travail, nous avons appliqué les réseaux de neurones pour la reconnaissance des phonèmes spécifiques à la langue Arabe Standard. Par Reconnaissance Automatique de la Parole (RAP), nous entendons la transformation automatique de séquences de parole en textes écrits ou toute autre action à posteriori, notamment dans le cadre d'interfaces homme-machine. Ce passage de la parole vers du texte écrit doit nécessairement passer par des étapes importantes : l'extraction des paramètres acoustiques, une comparaison avec des modèles de référence préalablement enregistrés et enfin la prise de décision, c'est-à-dire la reconnaissance. En parallèle à ces étapes, un processus d'apprentissage permet d'augmenter considérablement le taux de reconnaissance.

Aujourd'hui, un état de l'art des différents travaux réalisés dans le domaine de la reconnaissance de la parole montre que de meilleures résultats sont obtenus à partir des modèles connexionnistes (réseaux de neurones)

II.2. Réseaux de neurones

II.2.1. Histoire des réseaux de neurones

Les réseaux neuronaux sont composés d'une connexion massive d'éléments simples appelés neurones. La philosophie de ces réseaux neuronaux est d'imiter le cerveau humain et de comprendre son fonctionnement pour construire des systèmes intelligents artificiels.

La première apparition des réseaux neuronaux remonte à 1943, lorsque Warren McCulloch, un neurophysiologiste, et un jeune mathématicien, Walter Pitts, ont rédigé un article sur le fonctionnement des neurones. Ils ont utilisé des circuits électriques pour modéliser un réseau neuronal simple. [15]

Après cela, Donald Hebb a affirmé la dernière approche en écrivant The Organization of Behavior en 1949, un livre qui indique que les voies neuronales sont renforcées chaque fois qu'elles sont utilisées.

L'évolution des ordinateurs dans les années 1950 a permis de commencer à modéliser les rudiments de ces théories concernant la pensée humaine et la première simulation du réseau neuronal a été réalisée par Nathanial Rochester des laboratoires de recherche IBM, même si cette première tentative a échoué.

En 1959, Bernard Widrow et MarcianHoff de Stanford ont développé des modèles qu'ils ont appelés ADALINE et MADALINE. Ces modèles ont été nommés en raison de leur

utilisation d'éléments LINear multiples et adaptatifs. MADALINE a été le premier réseau neuronal à être appliqué à un problème du monde réel. Il s'agit d'un filtre adaptatif qui élimine les échos sur les lignes téléphoniques. Ce réseau neuronal est toujours utilisé dans le commerce. [15]

En 1982, plusieurs événements ont provoqué un regain d'intérêt. John Hopfield de Caltech a présenté un à l'Académie nationale des sciences. L'approche de Hopfield ne consistait pas simplement à modéliser des cerveaux, mais de créer des dispositifs utiles. Avec clarté et une analyse mathématique, il a montré comment de tels réseaux pouvaient fonctionner et ce qu'ils pouvaient faire. Pourtant, le plus grand atout de Hopfield était son charisme. Il s'exprimait bien, il était sympathique et il était le champion d'une technologie en sommeil. [15]

Aujourd'hui, les discussions sur les réseaux neuronaux se produisent partout. Leur promesse semble très prometteuse car la nature elle-même est la preuve que ce genre de chose fonctionne. Pourtant, son avenir, voire la clé même de toute la technologie, réside dans le développement du matériel. Actuellement, la plupart des développements de réseaux neuronaux consistent simplement à prouver que le principe fonctionne. Cette recherche développe des réseaux neuronaux qui, en raison des limitations de traitement, prennent des semaines à apprendre. Pour sortir ces prototypes du laboratoire et les mettre en service, il faut des puces spécialisées. Les entreprises travaillent sur trois types de puces neuro : numériques, analogiques et optiques. Certaines entreprises travaillent à la création d'un "compilateur de silicium" pour générer un circuit intégré spécifique d'application (ASIC) de réseau neuronal. Ces ASIC et ces puces numériques semblables à des neurones semblent être la vague d'un avenir proche. En définitive, les puces optiques semblent très prometteuses. Pourtant, il faudra peut-être attendre des années avant que les puces optiques ne voient le jour dans des applications commerciales. [15]

II.2.2. Réseau de neurones biologique

Un neurone (ou cellule nerveuse) est un organe biologique qui traite l'information. Elle se compose d'un noyau cellulaire, ou soma, et de deux types de ramifications arborescentes : l'axone et les dendrites.

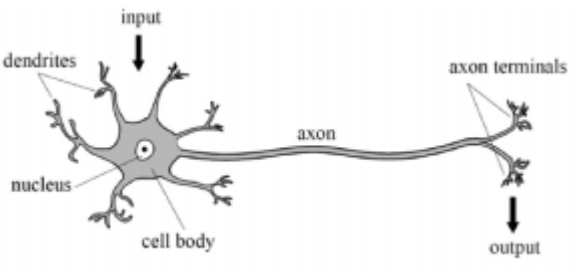

Figure II.1. Le neurone biologique

Le corps cellulaire possède un noyau contenant des informations sur les caractères génériques et un plasma qui contient l'équipement moléculaire nécessaire au neurone pour produire de la matière.

Un neurone, par l'intermédiaire de ses dendrites (récepteurs), reçoit des signaux (impulsions) d'autres neurones et transmet les signaux produits par son corps cellulaire le long de l'axone (émetteur), qui se ramifie progressivement en brins et sous-brins. Une synapse est une structure simple et une unité fonctionnelle entre deux neurones (un brin d'axone d'un neurone et l'autre dendrite). [16]

Une fois que l'impulsion atteint le terminal de la synapse, certaines substances chimiques appelées neurotransmetteurs sont produites. Les neurotransmetteurs se propagent à travers la distance synaptique pour renforcer ou supprimer la capacité du neurone récepteur à libérer des impulsions électriques, selon la forme de la synapse.

La force des synapses peut être modifiée par les signaux qui la traversent, de sorte que les synapses peuvent bénéficier des comportements qu'elles adoptent. Cette dépendance à l'égard de l'histoire sert de mémoire qui pourrait être responsable de la mémoire humaine. [16]

II.2.3. Neurone artificiel (formel)

Le premier modèle de neurone formel date des années quarante. Il a été présenté par Mc Culloch et Pitts [17]. Inspirés par leurs travaux sur les neurones biologiques, ils ont proposé le modèle suivant :

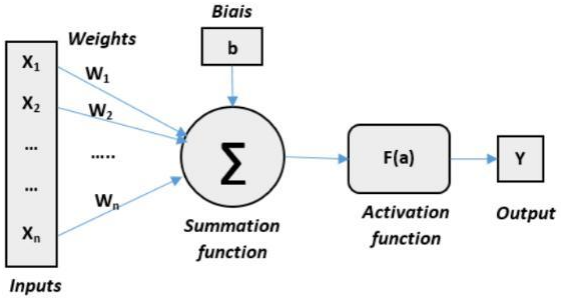

Figure II.2. Prétraitement du signal vocal

Un neurone artificiel est une fonction mathématique conçue comme un modèle grossier, ou une abstraction, des neurones biologiques.

Les neurones artificiels sont les unités constitutives d'un réseau neuronal artificiel. Selon le modèle spécifique utilisé, il peut recevoir différents noms, tels qu'unité semi-linéaire, neurone Nv, neurone binaire, fonction de seuil linéaire ou neurone McCulloch-Pitts (MCP).

Le neurone artificiel reçoit une ou plusieurs entrées (représentant une ou plusieurs dendrites) et les additionne pour produire une sortie (représentant l'axone d'un neurone biologique). Habituellement, les sommes de chaque nœud sont pondérées et la somme passe par une fonction non linéaire appelée fonction d'activation ou fonction de transfert. Les fonctions de transfert ont généralement une forme sigmoïde, mais elles peuvent également prendre la forme d'autres fonctions non linéaires, de fonctions linéaires par morceaux ou de fonctions en escalier. [16]

Plus généralement, un neurone formel est un élément de traitement (opérateur mathématique) avec n entrées (qui sont les neurones externes ou les sorties d'autres neurones), et une seule sortie.Ce modèle est décrit mathématiquement par les équations suivantes :

$$S = \sum_{i}^{n+1} w_i x_i \qquad \text{(II. 1)}$$

$$y = f(x) \qquad \text{(II. 2)}$$

Où w_i ,x_i, f, y sont respectivement, les poids synaptiques (paramètres), les entrées, la fonction d'activation et la sortie du neurone.

Nous pouvons résumer la différence entre un neurone artificiel et un neurone biologique dans la figure suivante

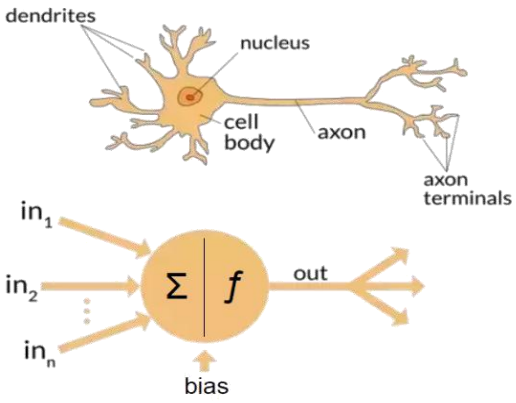

Figure II.3. Neurone artificiel et biologique

II.2.4. La fonction d'activation

Les fonctions d'activation représentent généralement certaines formes d'absence de linéarité. L'une des formes les plus simples de non-linéarité, qui convient aux réseaux discrets, est la fonction de signe II.4a.

Une autre variante de ce type de non-linéarité est la fonction échelon II.4b.Pour la majorité des algorithmes d'apprentissage, il est nécessaire d'utiliser des fonctions sigmoïdes

différentiables, telles que la fonction sigmoïde unipolaire, figure II.4c et la fonction sigmoïde bipolaire II.4d. [17]

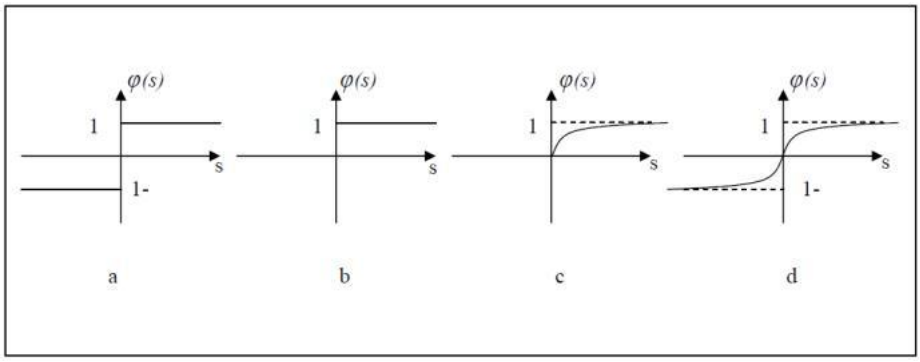

Figure II.4. Types de fonctions d'activation

II.2.5. Sélection de la fonction d'activation

La sélection de la fonction d'activation a un effet important sur les résultats du réseau neuronal. Elle constitue une transformation mathématique des unités d'entrée pondérées résumées pour générer la sortie du neurone. La fonction sigmoïde bipolaire est la fonction d'activation la plus utilisée dans le domaine de la modélisation et du contrôle des systèmes non linéaires.

Techniquement parlant, une fonction d'activation comprend deux parties : une fonction de combinaison qui intègre toutes les unités d'entrée en une seule valeur (somme pondérée des entrées) et une fonction de trans- fert qui applique une transformation non linéaire à ces unités résumées pour déclencher une unité de sortie. En général, le modèle de données détermine les formes des fonctions de transfert.

La plupart des études ont recours aux fonctions de transfert sigmoïdes. Les avantages de la fonction logistique ou sigmoïde est qu'il s'agit de fonctions continues qui augmentent ou diminuent de façon monotone, qui saturent vers les valeurs minimales et maximales, qui s'approchent très bien de la fonction et sont différentiables sur l'ensemble du domaine, ce qui permet de réduire considérablement la charge de calcul pour l'apprentissage. Charge de calculpour l'apprentissage Il est important de noter que les RN dont les neurones cachés ont une fonction d'activation sigmoïde et les neurones de sortie la fonction sigmoïde ou d'identité sont appelés Perceptrons multi-couches (Multi-Layer Perceptrons (MLP)). [16]

II.2.6. Perceptrons multi-couches (Multi-Layer Perceptron)

Nous présentons ici l'une des architectures de réseau les plus utilisées. Il s'agit d'une classe de réseaux neuronaux artificiels à anticipation qui comporte au moins trois couches de nœuds.

Il génère un ensemble de sorties $\{y_1, y_2, \ldots . y_m\}$ et un ensemble d'entrées $\{x_1, x_2, \ldots . x_n\}$. À l'exception des nœuds d'entrée, chaque nœud est un neurone qui utilise une fonction d'activation non linéaire. La diffusion de l'information se fait donc dans une seule direction, de la couche d'entrée à la couche de sortie. La fonction d'activation utilisée pour les neurones peut être n'importe quelle fonction croissante et différentiable.

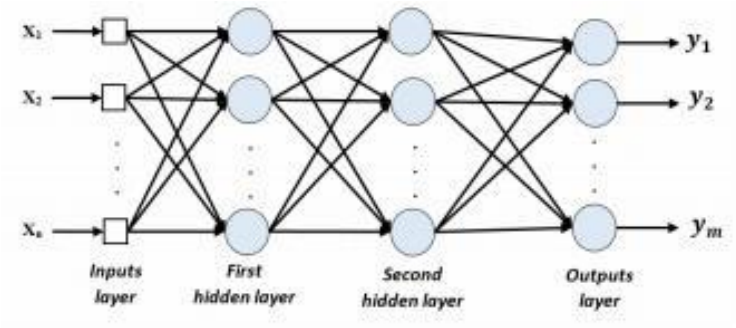

Figure II.5. Modèle MLP

Un réseau neuronal est formé avec des modèles de paires d'entrée et de cible avec la capacité d'apprendre. MLP peut séparer des données qui ne sont pas linéairement distinguables. Il est notamment entraîné à l'aide d'une technique d'apprentissage supervisée appelée algorithme de back-propagation (BP), qui vise à minimiser l'erreur globale mesurée au niveau de la couche de sortie par la relation suivante [18]

$$e(t) = y_d(t) - y_m(t) \qquad \text{(II. 3)}$$

Où $y_d(t)$désigne la sortie désirée, et $y_m(t)$ la sortie mesurée du neurone.

L'algorithme BP utilise une procédure d'apprentissage itérative supervisée, où le MLP est entraîné avec un ensemble d'entrées et de sorties prédéfinies. L'erreur globale $E_d(t)$ calculée par l'équation ci-dessous, cette erreur peut être minimisée par la technique de descente de gradient. [19]

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(\mathrm{y_{d,i}(t)} - \mathrm{y_{m,i}(t)})^2 \qquad \text{(II. 4)}$$

II.2.7. Architecture de réseau

Les réseaux neuronaux composés de seulement deux couches, la couche d'entrée et la couche de sortie, sont les plus simples, c'est-à-dire qu'ils ne comportent pas de couches cachées. C'est ce qu'on appelle souvent la couche de saut qui constitue essentiellement une modélisation de régression linéaire conventionnelle dans une conception de RN où les couches d'entrée et de sortie sont connectées directement, contournant ainsi la couche cachée.

Comme tous les autres réseaux, cette forme de RN dépend du poids comme connexion entre une entrée et la sortie ; le poids représentant l'importance relative d'une entrée spécifique dans le calcul de la sortie.

Il est important de garder à l'esprit que la sortie générée dépendra fortement du type de fonction d'activation que nous allons utiliser. Cependant, comme la couche cachée confère une forte capacité d'apprentissage au RN, dans les applications pratiques, une architecture RN à trois et plus de trois est utilisée.

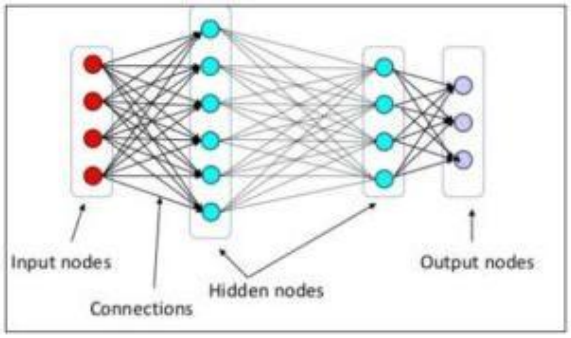

Figure II.6. Couches des réseaux neuronaux

La figureII.6 montre les couches des réseaux neuronaux. Il est important de différencier deux classes de poids dans RN ; ceux qui relient les entrées à la couche cachée et ceux qui relient la couche cachée à la couche de sortie. Sinon, il existe deux classes de fonctions d'activation ; l'une se trouve dans la couche cachée et l'autre dans la couche de sortie.

Il existe un nombre infini de façons de construire un RN. Il y a donc deux termes utilisés pour décrire la façon dont un RN est organisé. Tout d'abord la neurodynamique (donne fondamentalement les propriétés d'un neurone individuel telles que sa fonction de transfert et la combinaison des entrées) et l'architecture (définit la structure du RN incluant le nombre de neurones dans chaque couche et le nombre de types d'interconnexions) Les éléments suivants doivent être pris en considération lors de la construction d'un réseau. - Meilleures valeurs de départ (initialisation du poids)

- Nombre de couches cachées

- Nombre de neurones dans chaque couche cachée

- Nombre de variables d'entrée ou combinaison de variables d'entrée

- Taux d'apprentissage

- Taux d'élan

- Temps de formation ou quantité de formation (c'est-à-dire le nombre d'itérations à employer)

- Type de fonction d'activation à utiliser dans les couches cachées et de sortie.

II.2.8. Réseaux neuronaux Apprentissage

Le besoin d'apprentissage apparaît lorsque l'information a priori est incomplète, et son type dépend du degré de complétude de cette information, Comme l'information qui peut acquérir un réseau de neurones est représentée dans les poids des connexions entre les neurones, l'apprentissage consiste donc à ajuster ces poids de manière à ce que le réseau présente certains comportements désirés. [17] Mathématiquement, l'apprentissage est défini par:

$$\frac{\partial y}{\partial x} \neq 0 \qquad \text{(II. 5)}$$

Où w est la matrice de poids.

Nous pouvons distinguer trois types d'apprentissage, l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

Apprentissage supervisé

Le réseau reçoit une réponse correcte (sortie) pour chaque modèle d'entrée (présence d'un maître qui fournit la réponse souhaitée). En outre, les poids sont déterminés pour permettre le réseau de produire des réponses aussi proches que possible des réponses correctes connues. L'algorithme (The back-propagation) appartient à cette catégorie. [16]

Cette procédure est répétée jusqu'à ce qu'un critère de performance soit satisfait. Une fois que la procédure d'apprentissage est terminée, les coefficients synaptiques prennent des valeurs optimales par rapport aux configurations stockées et le réseau peut être opérationnel

Apprentissage non supervisé

Ce type d'apprentissage ne nécessite pas de réponse correcte associée à chaque modèle d'entrée dans l'ensemble d'apprentissage et la procédure d'apprentissage est basée uniquement

sur les valeurs d'entrée. Il explore la structure sous-jacente des données, ou les corrélations entre les modèles dans les données et organise les modèles en catégories à partir de ces corrélations.

Le réseau s'auto-organise de manière à optimiser une certaine fonction de coût, sans qu'on lui donne la réponse souhaitée. Cette propriété est appelée auto-organisation. L'algorithme de Kohonen appartient à cette catégorie. [17][16]

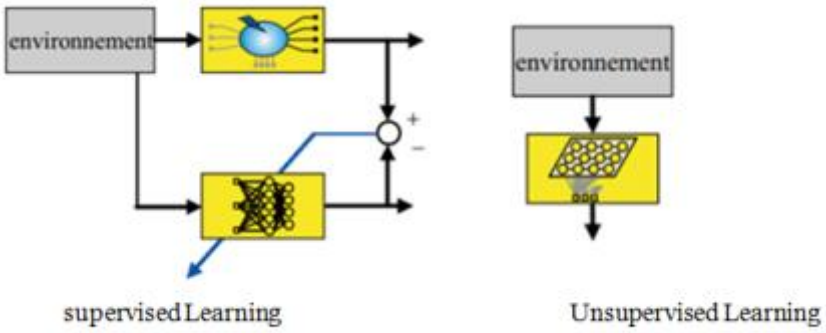

Figure II.7. Principe de l'apprentissage supervisé et non supervisé

Apprentissage renforcé

Combine l'apprentissage supervisé et non supervisé ; une partie des poids est déterminée par l'apprentissage supervisé et les autres sont obtenus par un apprentissage non supervisé.

L'idée de base de l'apprentissage par renforcement s'inspire des mécanismes d'apprentissage chez les animaux. Dans ce type d'apprentissage, nous supposons qu'il n'y a pas de maître (superviseur) qui peut fournir la bonne réponse, mais le système à former est indirectement informé de l'effet de l'action choisie. Cette action est renforcée si elle conduit à une amélioration de la performance du système piloté, et les éléments qui contribuent à la génération de cette action sont soit récompensés, soit punis.

II.3. Les Réseaux de neurones en RAP

Le diagramme de la figure montre le modèle conceptuel du système de reconnaissance de la parole de l'être humain. Le signal acoustique en entrée est analysé par un « modèle auditif » qui fournit des informations spectrales du signal et les sauvegardent dans une mémoire sensorielle. Des informations sensorielles provenant d'autres sources (vision, toucher, ...) sont également présentes dans cette mémoire et servent à enrichir les différents niveaux de description du signal. L'analyse auditive est principalement basée sur le traitement

acoustique de l'oreille. Ensuite, a lieu dans le cerveau une analyse de caractéristiques à différents niveaux. Quant aux mémoires à court et à long termes, elles offrent un contrôle externe au processus neuronal. Finalement, il est à remarquer que la configuration globale du modèle s'apparente à un réseau connexionniste « feedforward », et c'est en s'inspirant de ce schéma perceptuel du processus de reconnaissance que nous croyons que le connexionnisme offre une alternative prometteuse pour la modélisation des tâches cognitives.

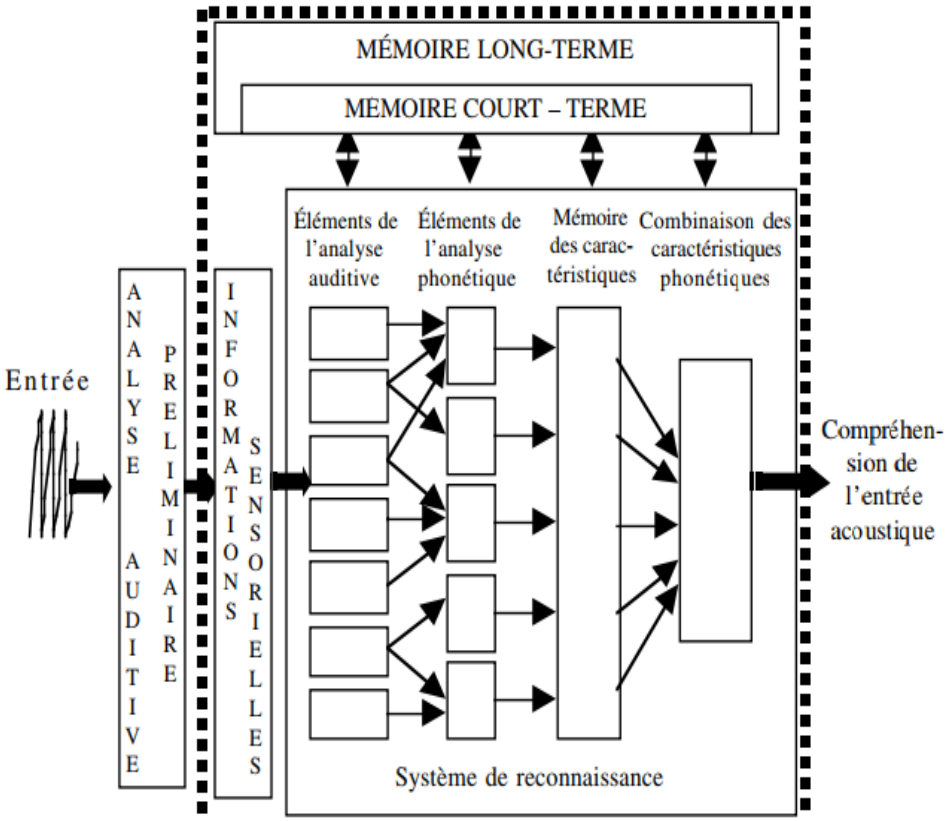

Figure II.8. Diagramme conceptuel du système de reconnaissance de l'être humain (d'après [20]).

Les tous premiers essais en RAP effectuaient des tâches très simples telles que: classer des segments de parole en son voisé/non voisé, ou nasal/fricatif/plosif, en utilisant un Perceptron multicouches [21]. Le succès remporté par ces réseaux a énormément encouragé les chercheurs à considérer la classification des phonèmes; ce qui fût effectué avec succès [22]. Les mêmes techniques rencontrèrent quelque succès pour la reconnaissance des mots, mais l'absence du paramètre tempsdans de telles architectures a orienté les recherches vers d'autres alternatives telles que les réseaux multicouches à retard, ou *Time-Delay Neural Networks* (TDNN) [23]. Un TDNN est à la base un perceptron multicouche qui se singularise par le fait

qu'il prend en compte une certaine notion du temps. C'est-à-dire qu'au lieu de prendre en compte tous les neurones de la couche d'entrée en même temps il va effectuer un balayage temporel. La couche d'entrée du TDNN prend une fenêtre duspectre et balaye l'empreinte. Il a été développé dans le but d'apprendre des structures spectrales spécifiques à l'intérieur de vecteurs de parole consécutifs.

Une taxonomie des systèmes connexionnistes à composante temporelle est présentée dans [24]. Mais on relèvera que ces différentes architectures sont assez compliquées à mettre en œuvre et que surtout elles n'intègrent pas une dimension symbolique du domaine d'application, c'est-à-dire les entités considérées et les relations qui existent entre-elles. D'autre part, un système expert connexionniste dédié à la parole est présenté dans [25], le modèle proposé reproduit la sémantique du domaine au travers des unités phonétiques et lexicales considérées, mais il n'inclut pas la dimension temporelle de l'application.

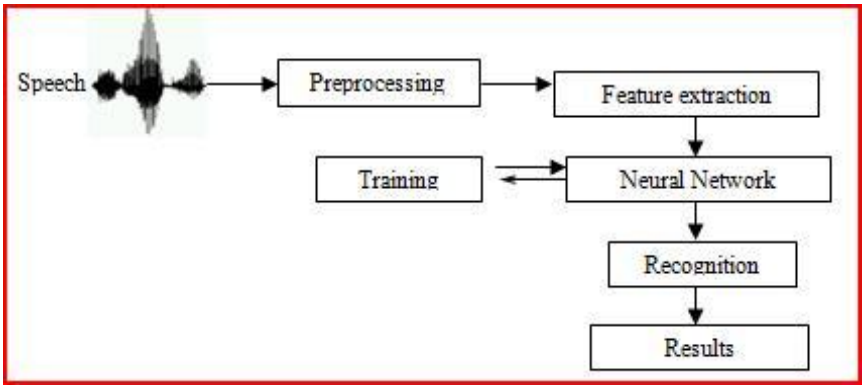

Figure II.9. Structure d'un système standard de RAP, basé sur les RNA

II.4. Réseaux de neurones convolutifs

Convolutional Neural Network (CNN) (réseaux de neurones convolutifs) sont un type de réseau de neurones spécialisés pour le traitement de données ayant une topologie semblable à une grille. Qui se sont avérés très efficaces dans des domaines tels que la reconnaissance et la classification d'images et vidéos. CNN a réussi à identifier les visages, les objets, panneaux de circulation et auto-conduite des voitures. Récemment, les CNN ont été efficaces dans plusieurs tâches de traitement du langage naturel (telles que la classification des phrases). Dans le ML, un réseau convolutif est un type de réseau de neurones feed-forward, il a été inspiré par des processus biologiques [26]. Il existe quatre (4) principales opérations illustrées dans le CNN à savoir :

- La couche convolution.
- La couheRectifiedLinear Unit.

- La couche Pooling.
- La couche entièrement connectée.

II.4.1. Architecture du système

Nous commencerons par entraîner un réseau de neurones convolutif le plus classique. Il a une base de données et des paramètres entraînables en entrée. Il se compose de couches convolutives. Chaque couche convolutive est placée avec une couche de pooling maximale, et enfin une couche entièrement connectée. Afin d'apprendre plus rapidement, Relu sera utilisé comme fonction d'activation.

Figure II.10. Architecture générale d'un CNN.

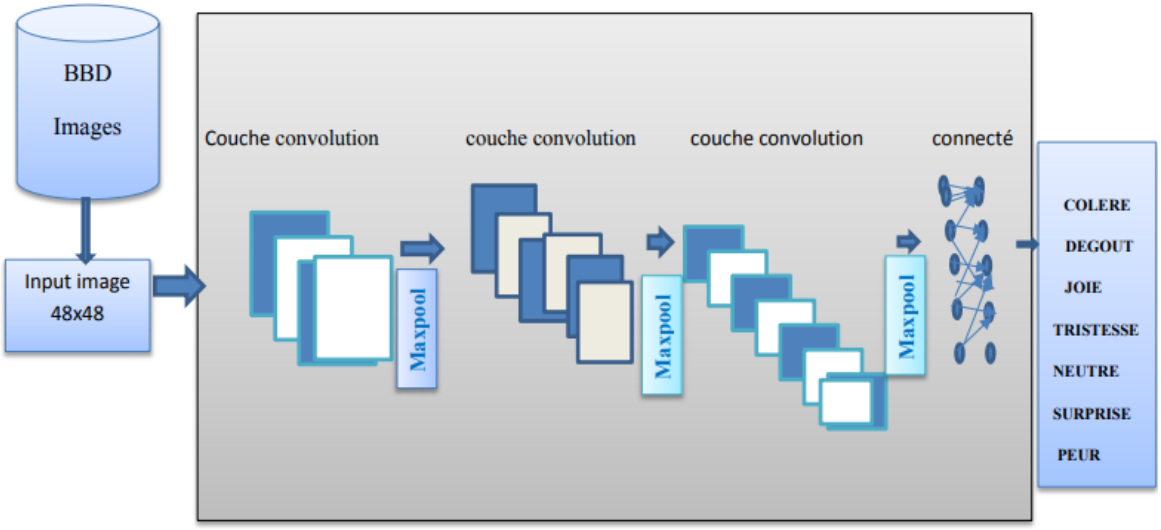

Figure II.11. Architecture proposée.

II .5.Conclusion

Ce chapitre nous a permis d'exposer quelques notions de base des réseaux de neurones, puis a brièvement résumé l'aspect général de l'identification Reconnaissance automatique de la parole (RAP) par Réseaux de Neurones, puis nous définissons d'abord l'architecture utilisée, et essayons également d'expliquer les principaux concepts des réseaux de neurones convolutifs (CNN) en termes simples.

Chapitre III

Résultats et interprétations

III.1. Introduction

Ce chapitre est consacré aux différents résultats obtenus en appliquant les méthodes qui font l'objet de ce mémoire. Notre objectif est d'abord de construire une base de données sous la forme d'un fichier wav, puis de lui appliquer des commandes Matlab (réseau de neurones convolutifs) pour calculer le taux de reconnaissance pour des phrases dans la base de données. Par conséquent, nous utilisons les fonctions fournies par Matlab pour concevoir notre système de reconnaissance. Pour cela, nous avons suivi les étapes suivantes :

• Créer un groupe d'apprentissage : chaque élément de vocabulaire est enregistré plusieurs fois et a été nommé le mot correspondant sur le son.

• Analyse audio : les signaux enregistrés sont convertis en une série de caractéristiques vectorielles (coefficients MFCC).

• Définition de modèles CNN : formation d'un modèle d'apprentissage en profondeur qui détecte la présence de commandes vocales dans une voix.

Modèles d'entraînement : entraîner un réseau de neurones convolutifs à reconnaître un ensemble donné de commandes.

• Définir la tâche de reconnaissance : les règles grammaticales à suivre sont définies.

• Reconnaissance et évaluation des performances dans le groupe test.

III.2. Organisation de la base de données

III.2.1.Organisation du corpus

Le corpus comprend 19phrasessyntaxiquement et sémantiquement correctes. Les mots de notre corpus ont été vérifiées par des linguistes de l'Université de Laghouat.

1 - نَشْكُرُكَ عَلَى استِعْمَالِكَ خِدْمَة الدَّفْعِ المُسَبَّقْ,إلى اللِّقَاءْ.

2 - إجَابَتُكَ خَاطِئَةْ.

3 - هَذِهِ الخِدْمَةُ غَيْرُ مُتَوَفِّرَةٍ فِي الحِينْ,يُرْجَى إعَادَةُ المُحَاوَلَهْ.

4 - لاَ يُمْكِنُكَ الحُصولُ عَلَى هَذِهِ الخِدْمَةِ انْطِلاَقاً مِنْ رَقْمِ هَاتِفِكْ.

5 - لِلُّغَةِ العَرَبِيَّةِ اضْغَطْ عَلى الرَّقْمِ وَاحِدْ.

6 - مَرْحَبًا بِكَ فِي خِدْمَةِ الدَّفْعِ المُسْبَقْ.

7 - لإِعَادَةِ تَعْبِئَةِ حِسَابِكَ اضْغَطْ عَلَى الرَّقْمِ وَاحِدْ.

8 - لِمَعْرِفَةِ رَصِيدِكَ اضْغَطْ عَلَى الرَّقْمِ اثْنَيْنْ.

9 - يُرْجَى الانْتِظارُ, مُكَالَمَتُكَحُوِّلَتْ إِلى مَصْلَحَةِ الزَّبَائِنْ.

10 - يُرْجَى الانْتِظَارُ,سَيَتِمُّ البَحْثُ عَنْ رَصِيدِكْ.

11 -لَمْ يَبْقَى لَدَيْكَ رَصِيدٌ فِي الحِسَابْ.

12 - لَقَدْ أَدْخَلْتَ عِدَّةَ أَرْقَامٍ خَاطِئَهْ.

13 - لَقَدْ أَدْخَلْتَ رَقْمًا خَاطِئا.

14 - حِسَابُكَ نَاقِصْ.

15 - لَيْسَ لَدَيْكَ رَصِيدٌ فِي الحِسَابْ.

16 - الرَّمْزُ الَّذِي أَدْخَلْتَهُ خَاطِئْ.

17 -لاَ يُمْكِنُكَ تَعْبِئَةُ بِطَاقَتِكَ الآنْ.

18 - يُرْجَى المُحَاوَلةُ لاَحِقاً.

19 -لَدَيْكَ رَصِيدٌ فِي الحِسَابْ

III.2.2.Identification et critères de choix du locuteur

Suivant le protocole TIMIT (Texas Instruments Massachusetts Institute of Technology) pour mieux gérer les locuteurs et pour une meilleure organisation de la base de données, chaque locuteur doit avoir un code unique et ce selon plusieurs critères (sexe W / M, adulte / enfant, niveau de l'éducation). Les locuteurs doivent avoir une bonne prononciation de l'arabe standard et une bonne qualité vocale. Nos locuteurs sont des étudiants de l'Université de Laghouat et d'autres locuteurs.

III.3. Phase d'enregistrement du corpus

III.3.1. Conditions d'inscription

Nous devons définir correctement tous les paramètres :

• Le lieu d'enregistrement doit être un lieu normal.

• Chaque enregistrement doit commencer et se terminer par un silence

•Recommencez l'enregistrement s'il a été interrompu.

•La distance entre le locuteur et le capteur doit être ajustée en fonction de la voix de locuteur.

• Vérifiez chaque fois l'enregistrement pour éviter la saturation ou que le signal est faible avec un éditeur de signal.

III.3.2. Manipulation d'enregistrement

Le débit de lecture doit être normal (ni rapide ni lent). Il est également nécessaire de corriger le locuteur si le niveau d'émission tend à s'affaiblir (tendance à parler de moins en moins fort).

Donc :

- Nos locuteurs sont des étudiants de l'Université de Laghouat et d'autres locuteurs.
- Paramètres d'entrée: fréquence d'échantillonnage (8 KHz), nombre de canaux (1), codés sur 8 bits.
- Le lieu d'enregistrement est un endroit normal avec du bruit.

III.3.3. Acquisition des fichiers sons

On utilise le logiciel Cool Edit Pro 2.1pour enregistrer nos fichiers sons. Cool Edit Pro 2.1est un outil très efficace qui nous a permis de lire un fichier son (le visualiser, l'écouter, le

découper…) ou même en créer un nouveau, de faire une analyse acoustique (durées, Fo, intensité)

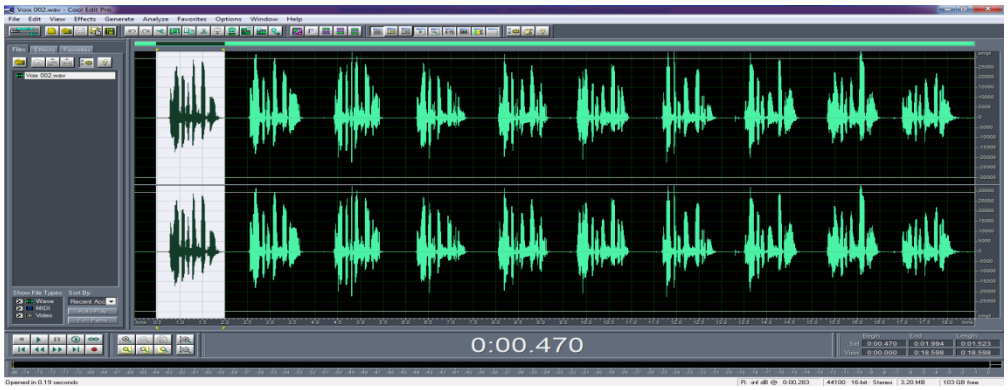

Figure III.1. Cool Edit Pro 2.1

Chaque élément du vocabulaire est enregistré, et étiqueté avec le nom correspondant. Le résultat de cette phase sont les fichiers sons (.wav) représentants le vocabulaire.

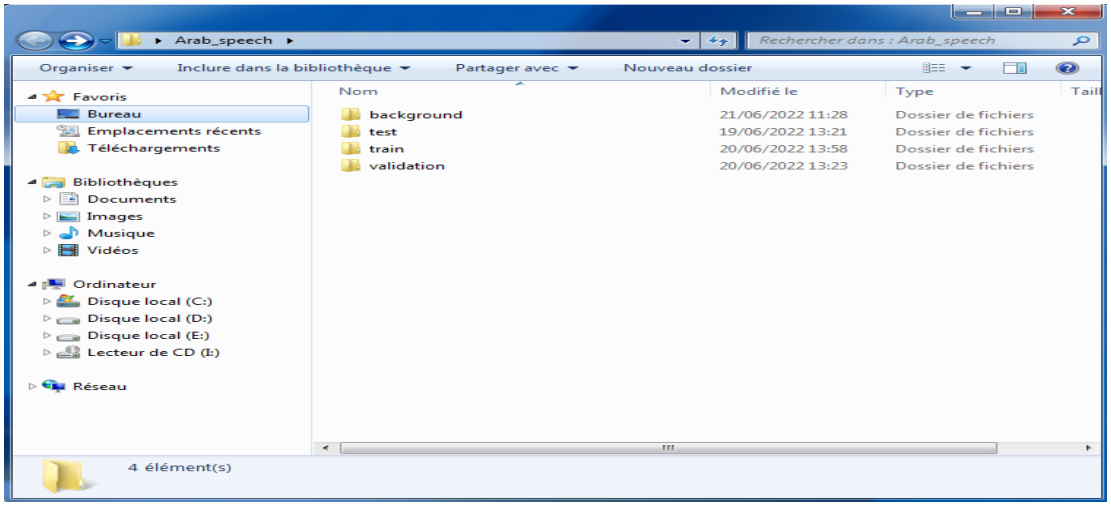

Figure III.2. Les fichiers de quelques enregistrements.

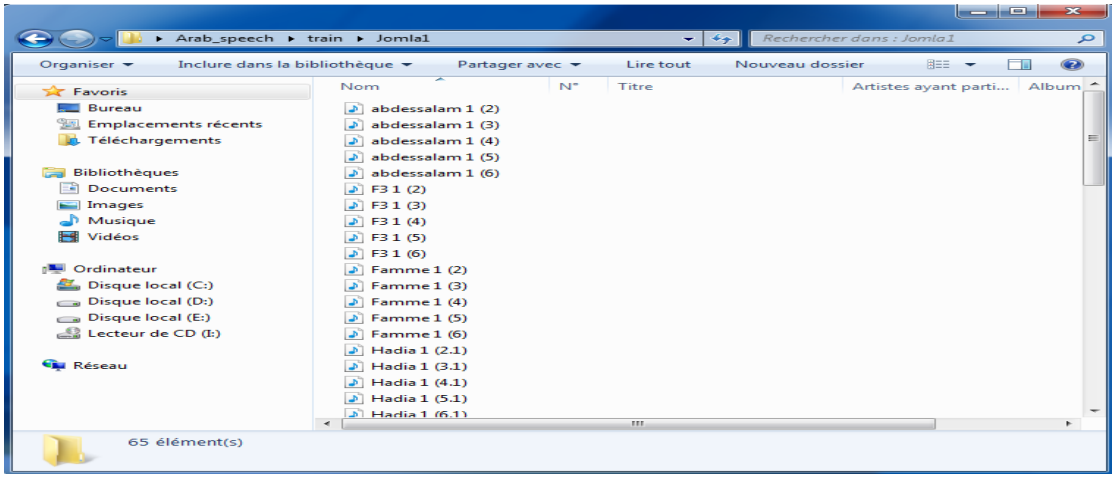

Figure III.3. Les fichiers sons de quelques enregistrements.

III.3.4. Segmentation et étiquetage

Nous avons utilisé l'outil Cool Edit Pro 2.1pour faire l'étiquetage manuel du corpus et le système pour la transcription phonétique SAMPA (Speech AssessmentMethodsPhonetic Alphabet). L'identification des différentes unités a été faite en utilisant l'outilCool Edit Pro 2.1, nous avons ainsi contrôlé, à chaque enregistrement la forme temporelle de l'onde acoustique correspondant à l'enregistrement, son spectrogramme, l'onde temporelle du pitch F0 ainsi que son énergie en s'appuyant toujours sur l'écoute.

III.3.5. Transcription de corpus

La transcription orthographique et phonétique des phrases de notre corpus est donnée ci-dessous :

Tableau III.1 : Transcription orthographique et phonétique des phrases de notre corpus

PHRASES DE NOTRE CORPUS	LA TRANSCRIPTION ORTHOGRAPHIQUE	LA TRANSCRIPTION PHONETIQUE
نشْكُرُكَ عَلَى استِعْمَالِكَ خِدْمَة الدَّفْعِ المُسَبَّقْ,إلى اللِّقَاءْ.	NACHKOROKA ALA ISTAMALIK KHIDMATI ALDAFAI ALMOBAKI ILA ALLIKA	na sheku ru ka Eala Ai se ti Ee ma: li ka khi de ma ti Aa le da feEi Aa le mu se baqe Ai la li qa:
إِجَابَتُكَخَاطِئَةْ.	IJABATOUKA KHATIA	Ai ja: ba tu ka kha Ti Aa
هَذِهِ الخِدْمَةُ غَيْرُ مُتَوَفِّرَةٍ فِي الحِينْ,يُرْجَى إِعَادَةُ المُحَاوَلَهْ.	HADIHI AL KHIDMATO GHYRO MOUTWAFIRA FI ALHIN YORJA IADAT ALMOHAWALA	ha dhi hi Aa le khi de ma tu Ga ye ru mu ta wa fi ra fi Aa le Hi : ne y ure ja Ai Ea : da tu Aa le mu Ha : wa la
لاَ يُمْكِنُكَ الحُصولُ عَلَى هَذِهِ الخِدْمَةِ انْطِلاَقاً مِنْ رَقْمِ هَاتِفِكْ.	LA YOMKINOKA ALHOSOUL ALA HADIHI ALKHDMAYI INTILAK MIN RAKMI HATIFIK	la: yu me ki nu ka Aa le Hu Su: Ea la ha di hi Aa le khi de ma ti Ai ne Ti la : qaa mi ne ra qe mi ha ti fi ka
لِلْغَةِ العَرَبِيَّةِ اضْغَطْ عَلَى الرَّقْمِ وَاحِدْ.	LILOUGHATI ALARABIYA IDGHAT ALA ALRAKMI WAHID	li lu Ga ti Aa le Ea ra bi ya Ai De Ga Te Ea la Aa le ra qe mi wa: Hi de

مَرْحَبًا بِكَ فِي خِدْمَةِ الدَّفْعِ المُسْبَقْ.	MARHABA BIKA FI ALDAFAI ALMOSBAK	ma re Ha baa bi ka fi: khi de ma ti Aa le da feEi Aa le mu se baqe
لِإِعَادَةِ تَعْبِئَةِ حِسَابِكَ اضْغَطْ عَلَى الرَّقْمِ وَاحِدْ.	LIADAT TABIATI HISABAK IDGHAT ALA ALRAKMI WAHID	li Ai Ea: da tit a Ee bi Aa ti Hi sa: bi ka Ai De Ga Te Ea la Aa le ra qe mi wa : Hi de
لِمَعْرِفَةِ رَصِيدِكَ اضْغَطْ عَلَى الرَّقْمِ اثْنَيْنْ.	LIMARIFATI RASIDAK IDGHAT ALA ALRAKMI ITNAN	li maEe ri fa ti ra Si: di qe Ai De Ga Te Ea la Aa le ra ke mi Ai the na ye ne
يُرْجَى الانْتِظَارُ, مُكَالَمَتُكَ حُوِّلَتْ إِلى مَصْلَحَةِ الزَّبَائِنْ.	YORJA ALINTIDAR MOKLAMTOKA HOUWILAT ILA MASLAHATI ALZABAIN	y ure ja Aa le Ai ne ti Dha : re , mu ka : la ma tu ka Hu wi la te Ai la ma Sela Ha ti Aa le zaba Ai ne
يُرْجَى الانْتِظَارُ,سَيَتِمُّ البَحْثُ عَنْ رَصِيدِكْ.	YORJA ALINTIDARSAYATIM ALBAHTO ANRASIDOKA	yureja: Aa le Ai ne ti Dha: resay a ti me Aa le ba He thuEa n era Si: di ke
لَمْ يَبْقَى لَدَيْكَ رَصِيدٌ فِي الحِسَابْ.	LAM YABKA LADAYKA RASID FI ALHISAB	la me yabeqa la da yi ka ra Si : duu fi : Aa le Hi sa be
لَقَدْ أَدْخَلْتَ عِدَّةَ أَرْقَامٍ خَاطِئَهْ.	LKAD ADKHALTA AIDATA ARKAM KHATIA	la ka de Aa de kha le ta Ei da ta Aa reqa : me kha Ti Aa
لَقَدْ أَدْخَلْتَ رَقْمًا خَاطِئا.	LKAD ADKHALTA RAKAM KHATIA	la ka de Aa de kha le ta ra qe maa kha : Ti Aa
حِسَابُكَ نَاقِصْ.	HISABOKA NAKIS	Hi sa : bu ka na : qi Se
لَيْسَ لَدَيْكَ رَصِيدٌ فِي الحِسَابْ.	LAYSA LADYKA RASID FI ALHISAB	La y sa la da ye ka ra si : de fi : Aa le Hi sa : be
الرَّمْزُ الَّذِي أَدْخَلْتَهُ خَاطِئْ.	ALRAMZO ALADI ADKHALTAHO KHATI	Aa le ra me zu Aa le la dhi Aa de kha le ta hu kha : Ti Ae
لاَ يُمْكِنُكَ تَعْبِئَةُ بِطَاقَتِكَ الآنْ.	LA YOMKINOKA TABIATO BITAKATIKA ALAN	La: yu me ki nu ka taEe bi Aa tu bi Ta : qat i ka Aa le Aa ne
يُرْجَى المُحَاوَلَةُ لاَحِقاً.	Y OURJAALMOUHAWALATOU LAHIKA	yureja: Aa le mu Ha: wa la tu la: Hi qa:

لَدَيْكَ رَصِيدٌ فِي الحِسَابْ	LADYKA RASID FI ALHISAB	la da ye ka ra si : de fi : Aa le Hi sa : be

III.4.Visualisation des données

Pour préparer les données pour une formation efficace d'un réseau de neurones convolutifs, nous avons transformé les formes d'onde de la parole en spectrogrammes auditifs.

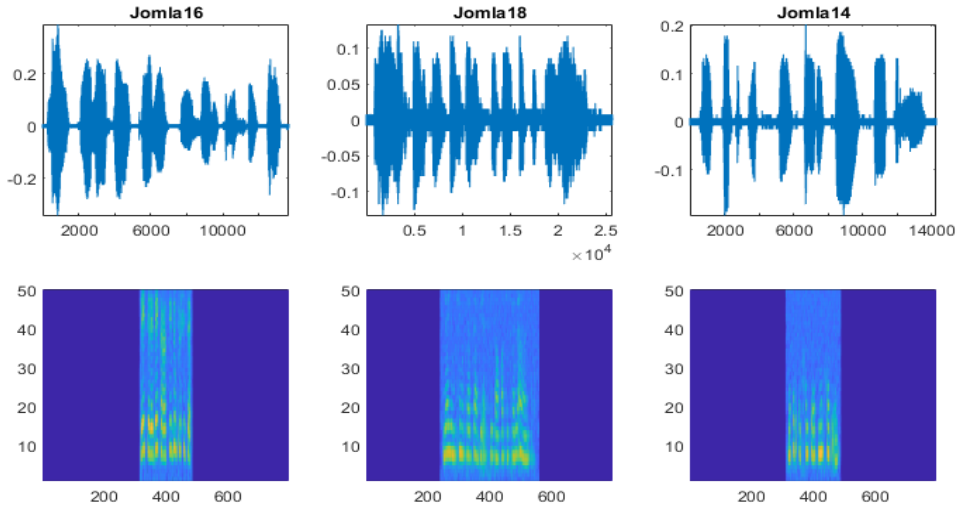

Figure III.4. Les spectrogrammes auditifs de quelques échantillons d'apprentissage

La figure III.5 illustre la répartition des tracés du spectre entre les ensembles d'apprentissage (65 enregistrements pour chaque phrases‘Jomla’concernant l’entrainement, 26 pour la validation, et avec 13 locuteurs).

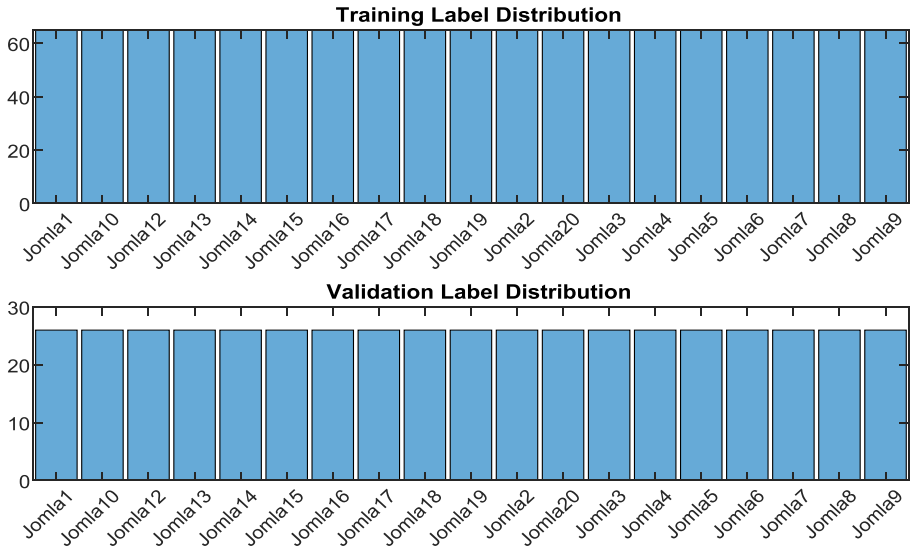

Figure III.5. La distribution des différentes étiquettes de classe dans les ensembles d'apprentissage et de validation.

III.5.Création de réseau de neuronal

Nous avons créé une structure de réseau simple sous la forme d'un tableau de couches. Nous avons utilisé des couches d'aplatissement convolutif et par lots, et des cartes de caractéristiques réduites « spatialement » (c'est-à-dire en temps et en fréquence) en utilisant des couches de pooling max. Nous avons ajouté une couche d'agrégation maximale finale qui agrège globalement la carte des entités en entrée au fil du temps.

Cela renforce (approximativement) l'invariance de la traduction temporelle dans les tracés d'entrée spectrale, permettant au réseau d'effectuer la même classification indépendamment de la position exacte de la parole dans le temps. L'agrégation globale réduit également considérablement le nombre de paramètres dans la couche finale entièrement connectée. Pour réduire la capacité du réseau à mémoriser des caractéristiques spécifiques des données d'apprentissage, nous avons ajouté une petite quantité de fuite d'entrée à la dernière couche entièrement connectée.

```
%CNN
numF = 12;
layers = [
imageInputLayer([numHopsnumBands])

convolution2dLayer(3,numF,'Padding','same')
batchNormalizationLayer
```

Figure III.6. Fichier de Matlab du modèle CNN

Le réseau est petit, avec seulement cinq couches convolutives et quelques filtres. numF contrôle le nombre de filtres dans les couches convolutives. Pour augmenter la résolution du réseau, nous avons essayé d'augmenter la profondeur du réseau en ajoutant des blocs identiques de couches convolutives, de normalisation par lots et ReLU. On peut aussi essayer d'augmenter le nombre de filtres convolutifs en augmentant numF.

```
reluLayer


maxPooling2dLayer(3,'Stride',2,'Padding','same')


convolution2dLayer(3,2*numF,'Padding','same')
```

```
batchNormalizationLayer
reluLayer
maxPooling2dLayer(3,'Stride',2,'Padding','same')
convolution2dLayer(3,4*numF,'Padding','same')
batchNormalizationLayer
reluLayer
maxPooling2dLayer(3,'Stride',2,'Padding','same')
convolution2dLayer(3,4*numF,'Padding','same')
batchNormalizationLayer
reluLayer
```

```
convolution2dLayer(3,4*numF,'Padding','same')
batchNormalizationLayer
reluLayer
maxPooling2dLayer([timePoolSize,1])
dropoutLayer(dropoutProb)
fullyConnectedLayer(numClasses)
softmaxLayer];
```

Figure III.7. Fichier de Matlab qui présente les couches de CNN

a. Réseau d'entrainement

Nous avons sélectionné des options de formation. Nous avons utilisé l'optimiseur Adam avec une petite taille de lot de 20 . Il s'est entraîné pendant 120 époques.

```
miniBatchSize = 20;
validationFrequency = floor(numel(YTrain)/miniBatchSize);
options = trainingOptions('adam',...
'InitialLearnRate',3e-4, ...
'MaxEpochs',120, ...
'MiniBatchSize',miniBatchSize, ...
'Shuffle','every-epoch', ...
'Plots','training-progress', ...
'Verbose',false, ...
'ValidationData',{XValidation,YValidation}, ...
'ValidationFrequency',validationFrequency, ...
'LearnRateSchedule','piecewise', ...
'LearnRateDropFactor',0.1, ...
'LearnRateDropPeriod',20);
```

Figure III.8. Options d'entrainement du modèle CNN

Pour faire l'entrainement du réseau, on doit exécuter l'instruction suivante :

```
trainedNet = trainNetwork(XTrain,YTrain,layers,options);
```

L'exécution de notre application et pour bénéficier du calcul en parallèle, a été réalisée par une station de calcul avec les propriétés illustré par la figure suivante :

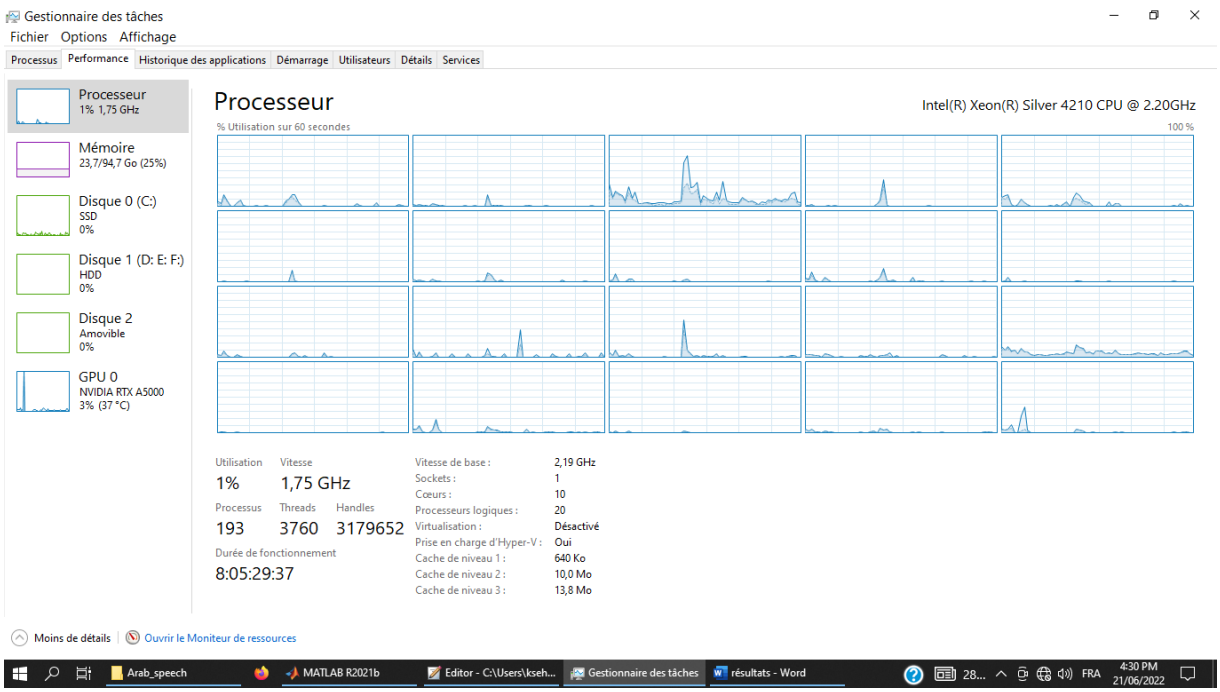

Figure III.9. Station de calcul

La figure III.10 montre les étapes d'entrainement.

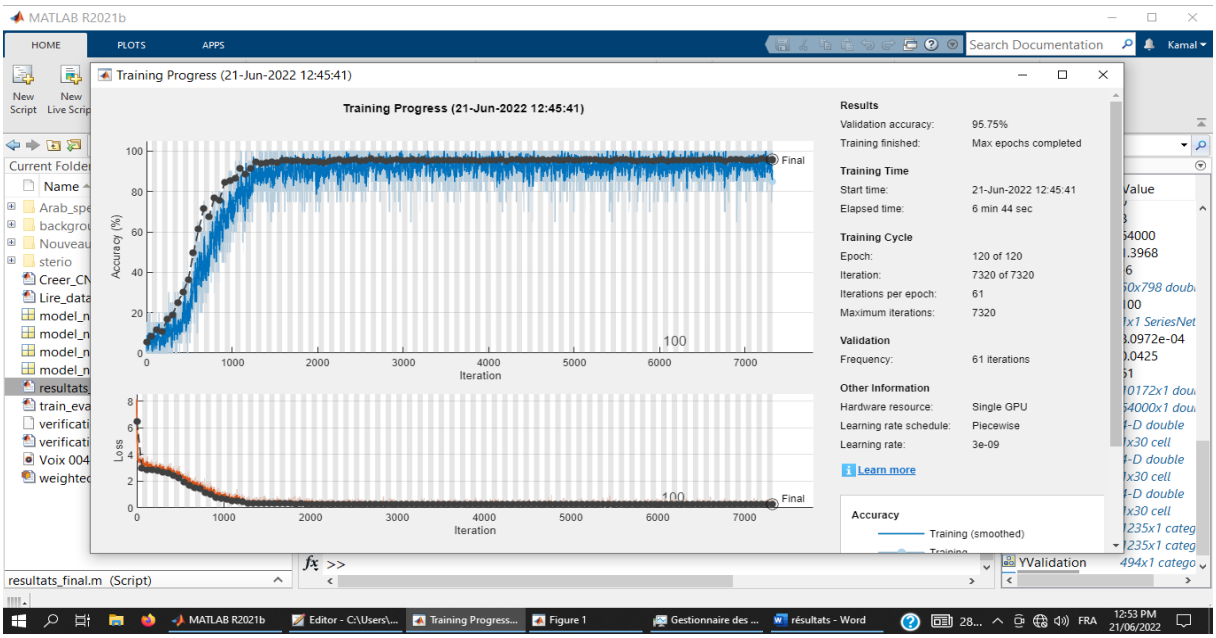

Figure III.10. Fenêtre d'entrainement

Après la finalisation d'apprentissage de notre modèle CNN, on a pu obtenu un tauxde précision de validation de 95.75% et un taux d'erreur de 4.25 %.

On peut dire que la performance obtenue est très bonne (voir figure III.11).

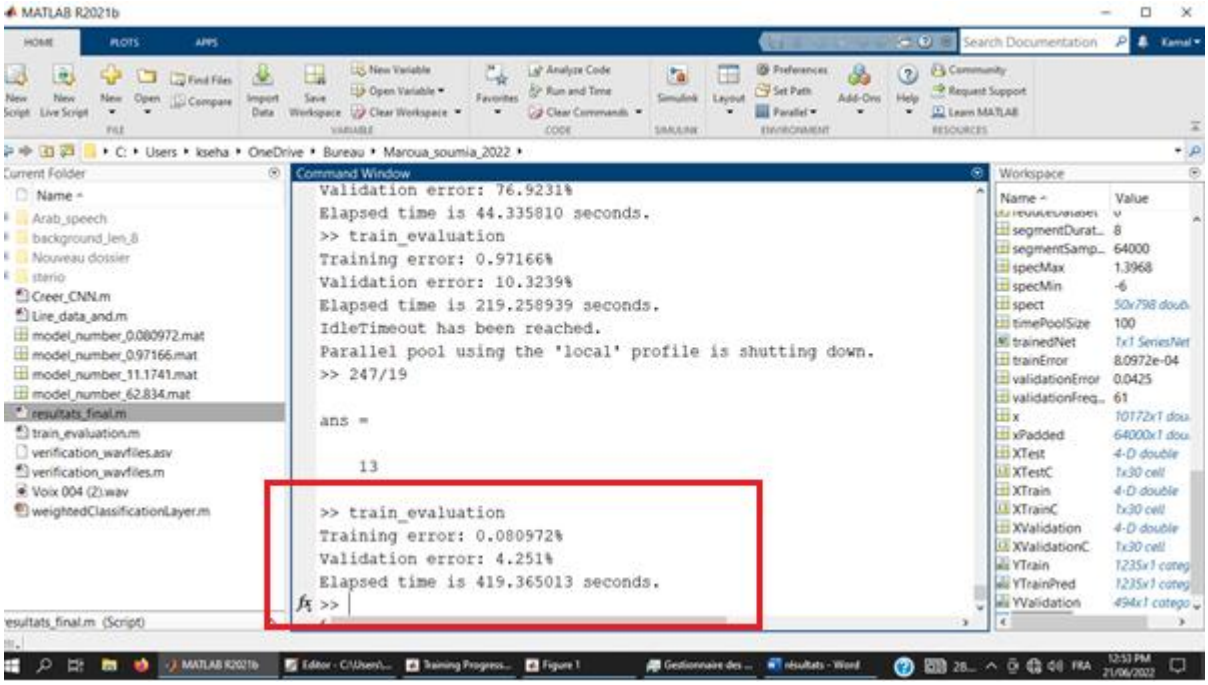

Figure III.11.Résultats d'entrainement

En plus nous avons tracé la matrice de confusion afin de montrer l'efficacité de notre modèle, le résultat obtenu est donnée par la figure suivante :

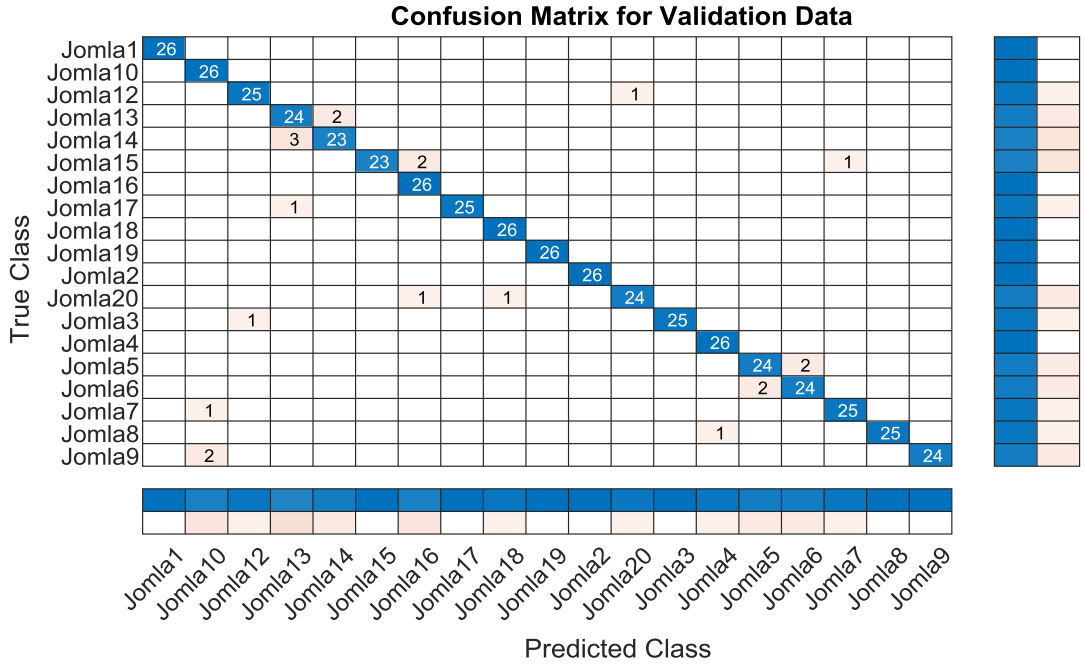

Figure III.12. Matrice de confusion de validation

b. Évaluation du réseau

Dans cette étape on a utilisé une base de données de test qui contient des nouveaux enregistrements pour 3 cas : 2 locuteurs, 6 locuteurs et 8 locuteurs. Les résultats obtenus sont donnée par les figures ci-dessous :

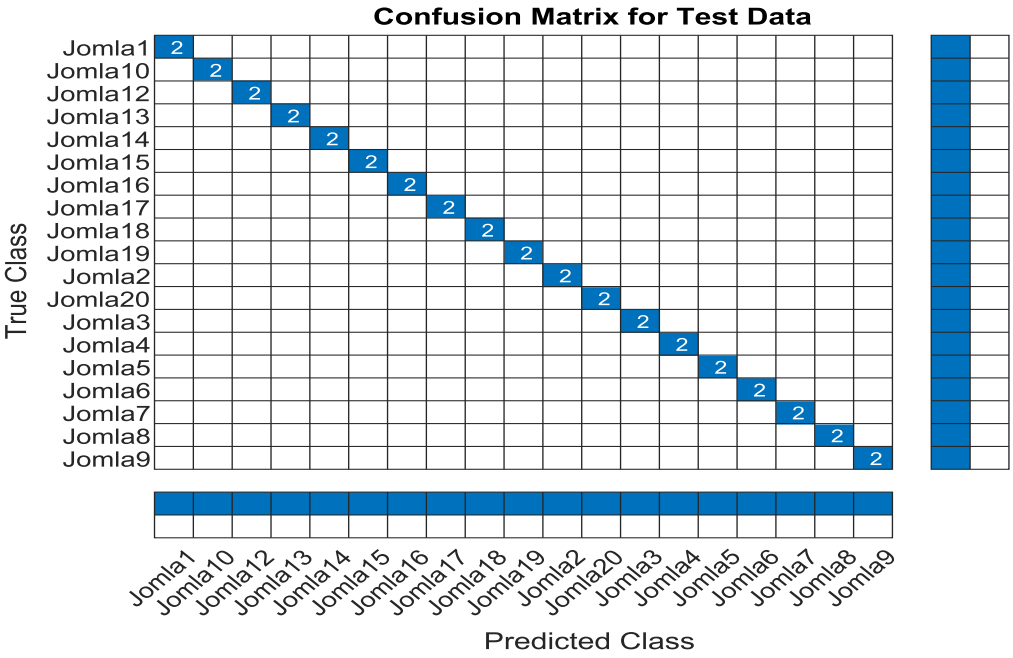

Figure III.13. Matrice de confusion de test, 2 locuteurs

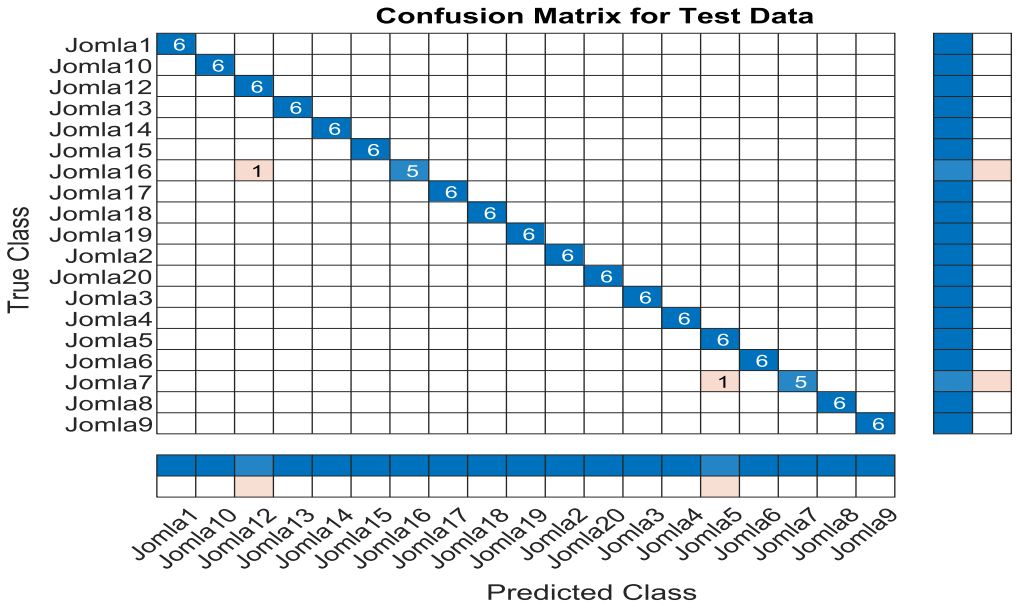

Figure III.14. Matrice de confusion de test, 6 locuteurs

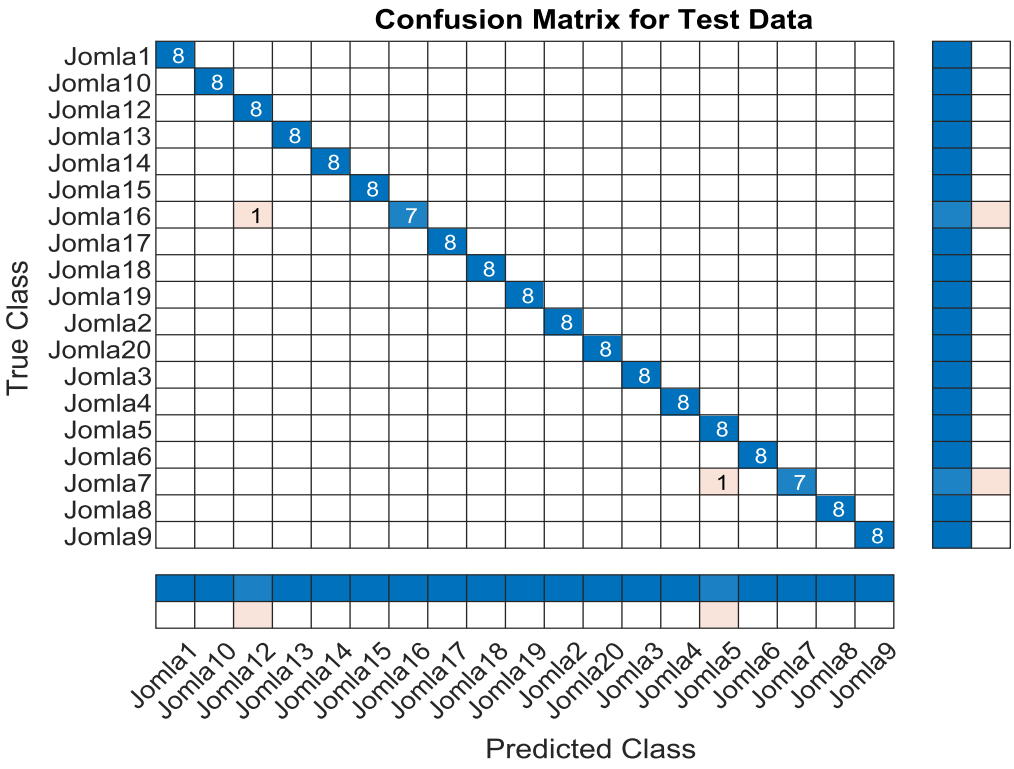

Figure III.15. Matrice de confusion de test, 8 locuteurs

Le taux de reconnaissance obtenus à partir des matrices de confusion pour les différents locuteurs qui ont été utilisés dans le test est presque de bons résultats.

III.6. Conclusion

Le présent document vise à mettre en évidence la pratique de CNN pour construire un système de reconnaissance des phrases.

Pour évaluer la fiabilité de notre système de reconnaissance, nous avons effectué le test pour la méthode d'analyse MFCC.

Le test consiste à reconnaître différents enregistrements pour plusieurs locuteurs qui ont été utilisés dans l'apprentissage.

Les résultats obtenus nous permettent de conclure que la méthode CNN donne de bons résultats, cela confirme la robustesse et l'efficacité de notre système de reconnaissance.

Conclusion Générale

La parole est la principale méthode de communication de toute société humaine, et bien sûr c'est aussi la méthode de communication la plus naturelle. Cependant, il est également certain que c'est sémantiquement plus facile. Pour les chercheurs qui développent des systèmes de reconnaissance vocale, faciliter la communication homme-machine et permettre aux machines de fonctionner en langage naturel est un défi.

Le travail que nous développons ici s'inscrit dans l'objectif d'établir un système de dialogue homme-machine. Il s'agit de la reconnaissance de la parole dans le domaine des communications.

Dans le cadre de cette mémoire, nous nous sommes particulièrement intéressés à la reconnaissance automatique de la parole (RAP).

Nous avons utilisé les techniques d'analyse MFCC pour la reconnaissance automatique. Notre choix s'est également porté sur les outils les plus appropriés (CNN, MATLAB,..) pour l'RAP.

Notre contribution à ce sujet peut être résumée en plusieurs points :

- L'acquisition et le développement d'une base de données composée d'un corpus de mots syntaxiquement et sémantiquement correctes de l'arabe classique. Ils ont été évalués par des linguistes de l'Université de Laghouat.
- Modélisation statistique utilisant le réseau neuronal de convolution (CNN) pour la reconnaissance de la parole.
- Développement d'un système de référence pour RAP basé sur la modélisation CNN sous MATLAB .

Les résultats obtenus nous amènent à conclure : La méthode CNN est une méthode très efficace, car elle nous donne de très bons résultats. Cela confirme la robustesse de notre système. Ce présent travail nous a permet d'apprendre et d'entrer en contact avec de nombreux domaines tels que le traitement du signal, la programmation, le traitement du langage, etc.

Références

[1] H.delassi, F.benharzellah, Automatic recognition of the speech by LPC, MFCC Application to GSM signals, Université Amar Telidji- Laghouat.2019.

[2] Vimala C., Dr. V. Radha, "A review on speech recognition challenges and approaches",

World of Computer Science and Information Technology Journal (WCSIT), ISSN: 2221-0741

Vol. 2, No. 1, 1-7, 2012.

[3] Lalimasingh, "Speech Signal Analysis using FFT and LPC ", International Journal of

Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 4, 04.2015.

[4] A.Souadkia , Reconnaissance automatique de la parole arabe: Approche évolutionniste, mémoire de magister, Présenté à l'Université de Guelma Faculté des sciences et sciences de l'ingénierie, 2010.

[5] Chapanis. A "Interactive communication: A few research answers for a technological explosion". Communication présentée au cours de la CEE. Orsay, F .1979.

.[6] Bellik. Yacine "Multimodal text editor interface including speech for the blind".Speech Communication, 23, 319-332. 1997.

[7] w.douib, Reconnaissance Automatique De La Parole Arabe Par CMU SPHINX 4, Mémoire Magister, Université Ferhat Abbas, Faculté de Technologie, 2013.

[8] H.delassi, F.benharzellah, Automatic recognition of the speech by LPC, MFCC Application to GSM signals, Université Amar Telidji- Laghouat.2019.

[9] M.benberech, W.ouazene. reconnaissance de la langue arabe par les méthodes d'intelegences artificielles, Université Amar Telidji- Laghouat, 2018.

[10] M. Kunt et R.Boite , traitement de la parole , édition masson 1987.

[11] J. Tisal , le réseau GSM, édition Dunod,2003.

[12]Calliope , la parole et son traitement automatique , édition masson,1999.

[13]M. Debyche, reconnaissance Automatique de la parole appliquée à la langue Arabe, thèse de doctorat d'état USTHB , 2007.

[14] R. Boite et M. Kunt, traitement de la parole édition presses polytechniques ROMANDES, 1997.

[15] Artificial Neural Networks Technology. url: http://www2.psych.utoronto.ca/ users/reingold/courses/ai/cache/neural4.html. (accessed: 25.03.2020).

[16] Artificial Neural Networks. url: http://freakysquare.com/artificial-neuralnetwork-simplified. (accessed: 27.03.2020).

[17] ABB IRB 6600. url: https://www.robots.com/robots/abb-irb-6600.(accessed: 21.03.2020).

[18] Mohammed Saber et al. "Artificial neural networks, support vector machine and energy detection for spectrum sensing based on real signals".In: International Journalof Communication Networks and Information Security 11.1 (2019), pp. 52–60.

[19] StigMoberg and Sven Hanssen."On feedback linearization for robust tracking control of flexible joint robots". In: IFAC Proceedings Volumes 41.2 (2008), pp. 12218–12223.

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

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

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

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

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

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

[26] Matsugu, M., Mori, K., Mitari, Y., &Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network.Neural Networks, 16(5-6), 555-559.