

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique



Mémoire de Master

Filière Electronique
Spécialité Instrumentation
Présenté par

BEN ACHOUR Imane

&

ZERROUK Rania

Traducteur automatique STS arabe-anglais a base de deep Learning

Proposé par :

M. ABED AHCÉNE

MCB

USD Blida

M. AMROUCHE AISSA

MRA

CRSTDLA

Année Universitaire 2021-2022

Remerciements

Tout d'abord, nous tenons à remercier le bon Dieu le tout Puissant de nous avoir donné la force et le courage de mener à bien ce modeste travail

Nous adressons nos remerciements les plus sincères à toutes les personnes qui nous ont permis d'évoluer dans la réflexion et l'élaboration de ce travail. Plus particulièrement, nous tenons à remercier :

Mr ABED, notre promoteur, qui croyait en notre capacité à réaliser quelque chose de merveilleux pour nous avoir accordé sa confiance pour la réalisation de ce projet.

Mr AMROUCHE et Mme BOUBAKEUR, chercheurs de centre de recherche scientifique et technique pour le développement de la langue arabe pour nous avoir guidées tout au long de cette étude.

Les jurys Mr YAKHLEF et Mr GUESSOUM, pour nous donnent leur temps et leur intérêt pour notre mémoire de fin d'études.

Enfin, nous tenons à remercier tous ceux qui nous ont aidés et assistés durant nos études et nous exprimons toute notre gratitude à vous

Dédicace

En premier lieu, je remercie Allah tout-puissant qui m'a donné le courage et la patience et la volonté de mener à bien ce modeste travail. Je consacre cette petite œuvre

A la meilleure des mères Nabiha

La Source de tendresse et de la force, la source de don et d'amour, ma mère qui croyait à mes compétences et m'encourageait et priait pour moi. Ces lignes ne seront pas suffisantes pour vous remercier ou vous récompenser pour vos sacrifices et votre diligence en vue d'atteindre cette étape. Tu as fait plus que ce que peut faire une mère, puisse ALLAH te protège et allonge ta vie

Au meilleur des pères Khaled

Je ne saurai jamais exprimer ma gratitude et ma reconnaissance. Père, Merci pour votre travail, votre fatigue et votre sacrifice pour ma carrière d'étudiant

A mon frère

Tu es le cadeau le plus précieux que mes parents m'aient fait, les mots seuls ne peuvent exprimer mon amour pour toi. Qu'ALLAH vous garde et vous aide à accomplir à votre désir

A mes tantes

Amel et Amira

A mon binôme Imen

Le temps passé avec vous dans ce travail m'a donné les meilleurs souvenirs et instants qui ne peuvent être effacés de ma mémoire

A tous mes enseignants

Le promoteur ABED, sa femme et Madame ZOLIKHA

Sans votre aide, vos conseils et vos encouragements, ce travail n'aurait pas donné de résultats

A mes chers amis

Hayet, Nacera et Yasmine

A mes chers collègues

Zehor, Ayoub, Amine, Mahmoud et Aymen

Rania

Dédicaces

Je dédie mon travail à

Mon père et ma mère

Ma tante Zina

Mes sœurs et Mes frères

Mes amis

Tous ceux qui ont contribué à la réalisation de ce projet

Imane

ملخص:

الهدف من هذه الأطروحة هو الترجمة الآلية للكلام من اللغة العربية للإنجليزية بناءً على الشبكات لعصبية العميقة وللقيام بذلك، استخدمت ثلاث خطوات محددة. الأولى تتمثل في التعرف التلقائي للكلام الذي تم الإدلاء به مع CNN، الخطوة الثانية هي الترجمة الآلية التي تنفذها LSTM والخطوة الأخيرة هي التركيب التلقائي بناءً على GAN. تظهر التقنية التي تمت دراستها في مشروعنا نتائج إيجابية.

كلمات مفتاحية :

الترجمة الآلية للكلام، الشبكات العصبية العميقة، التعرف التلقائي للكلام، التركيب التلقائي، CNN، LSTM، GAN.

Résumé:

L'objectif de ce mémoire est d'utiliser des réseaux de neurones profonds pour traduire automatiquement la parole entre l'arabe et l'anglais. Trois étapes distinctes ont été utilisées pour y parvenir. Le premier fait référence à la reconnaissance vocale automatique de CNN. Le LSTM est utilisé pour la traduction automatique dans la deuxième étape, et le GAN est utilisé pour la synthèse automatique dans la troisième. La technique étudiée dans notre projet donne des résultats satisfaisants.

Mots clés :

Traduction automatique de la parole, réseaux de neurones, Reconnaissance automatique de la parole, Synthèse automatique de la parole, CNN, LSTM, GAN.

Abstract :

The goal of this project is to use deep neural networks to automatically translate speech between Arabic and English. Three distinct stages have been used to accomplish this. The first refers to the CNN's automatic speech recognition. The LSTM is used for automatic translation in the second step, and the GAN is used for automatic synthesis in the third. The technique investigated in our project yields promising results.

Keywords:

Automatic translation of speech, deep neural networks, automatic recognition of speech, automatic translation, the automatic synthesis, CNN, LSTM, GAN.

Table des matières

Table des matières	vi
Liste des Abréviations	x
Liste des Figures	xii
Liste des Tableaux	xiii
Introduction Générale	1
Chapitre 1 Traitement de la parole	3
1.1 Introduction.....	4
1.2 Généralités sur la parole et la langue.....	4
1.2.1 Définition de la parole.....	4
1.2.2 Production de la parole.....	4
1.2.3 Description de la production de la parole.....	5
1.2.4 Audition et perception de la parole.....	6
1.2.5 Structure d'un système auditif.....	6
1.2.6 Classification des sons de la parole.....	8
1.2.7 Caractéristiques de la parole.....	8
1.2.8 Description de la parole.....	10
1.2.9 Définition de la langue.....	12
1.3 Traitement automatique de la parole.....	20
1.3.1 Classification des tâches de la traitement automatique de la parole.....	21
1.3.2 Domaine d'application.....	23
1.4 Conclusion.....	24
Chapitre 2 Traduction automatique de la parole	25
2.1 Introduction.....	26
2.2 Reconnaissance automatique de la parole « RAP ».....	26

2.2.1 Définition	26
2.2.2 Fonctionnement	26
2.2.3 Systèmes de la reconnaissance automatique de la parole	27
2.2.4 Approche de la reconnaissance automatique de la parole.....	29
2.2.5 Extraction des paramètres	31
2.2.6 Les Applications de la RAP.....	37
2.2.7 Les meilleurs logiciels de la RAP	38
2.2.8 Les avantages et les inconvénients de la reconnaissance de la parole.....	39
2.3 Traduction automatique de la langue « MT ».....	40
2.3.1 Définition de MT.....	40
2.3.2 Les types de la traduction automatique.....	41
2.3.3 Les modèle de traduction à base de règle	42
2.3.4 Les modèle de traduction automatique statistique	44
2.3.5 Evaluation d'un système de traduction automatique.....	45
2.3.6 Exemple des logiciels de traduction automatique	47
2.3.7 Traduction automatique à base de séquence a séquence.....	48
2.4 Synthèse automatique de la parole « SAP »	50
2.4.1 Définition de « SAP ».....	50
2.4.2 Architecture de la synthèse de parole	51
2.4.3 Les techniques de la synthèse de la parole	51
2.4.4 Avantages et inconvénients	54
2.4.5 Logiciels de synthèse de la parole	55
2.4.6 Applications de la synthèse de parole.....	56
2.5 Conclusion	56
Chapitre 3 Apprentissage profond (Deep learning)	57
3.1 Introduction	58
3.2 Intelligence Artificielle (AI)	58
3.2.1 Définition	58
3.3 Apprentissage automatique (ML).....	59
3.3.1 Définition	59
3.3.2 Types d'apprentissage automatique les plus populaire.....	59
3.3.3 Fonctionnement de l'apprentissage automatique.....	60

3.4 Apprentissage profond (DL)	60
3.4.1 Définition d'apprentissage profond	60
3.4.2 Fonctionnement de l'apprentissage profond.....	61
3.5 Pourquoi l'apprentissage profond ?.....	62
3.6 Réseaux de neurones	62
3.6.1 Neurone biologique.....	63
3.6.2 Neurone formel (artificielle).....	63
3.6.3 Réseaux de neurones artificiels « ANN »	65
3.7 L'architecture des réseaux de neurones	66
3.8 Les avantages et les inconvénients des réseaux de neurone	67
3.8.1 Avantages	67
3.8.2 Inconvénients	68
3.9 Réseaux de neurones profonds « DNN »	68
3.10 Types des réseaux de neurones profonds « DNN »	69
3.10.1 Réseau de neurone convolutive « CNN ».....	69
3.10.2 Réseaux de neurone Récurrents « RNN »	73
3.10.3 Réseaux de neurone antagonistes génératifs « GAN »	77
3.11 Conclusion	79
Chapitre 4 Implémentation et Résultats	80
4.1 Introduction.....	81
4.2 Contexte expérimentale.....	81
4.2.1 Matériel utilisé	81
4.2.2 Langage de programmation	81
4.2.3 Corpus de parole	86
4.3 Traduction automatique parole-parole.....	87
4.3.1 Reconnaissance automatique de la parole arabe	88
4.3.2 Implémentation du module RAP.....	89
4.3.3 Test d'évaluation de la reconnaissance	91
4.3.4 Implémentation du module de traduction.....	92
4.3.5 Synthèse automatique de la parole anglais	94

4.3.6 Implémentation du module de synthèse	94
4.3.7 Test d'évaluation subjectif de la qualité de la parole :	96
4.4 Conclusion.....	98
Conclusion Générale	99
Références Bibliographiques	100

Liste des Abréviations

AD	: Arabe Dialectal
STS	: Speech to Speech
RAP	: Reconnaissance de parole
HMM	: Hidden Markov Models
MFCC	: Mel Frequency Cepstral Coefficients
FFT	: Transformée de Fourier discrète
DCT	: Discret Cosinus Transforme
IHM	: interfaces homme machine
MT	: Machine Translation
RBMT	: Traduction automatique à base de règle
SMT	: Traduction automatique statistique
NMT	: Traduction automatique neuronale
RNN	: Recurrent Neural Networks
LSTM	: Long Short-Term Memory
GRU	: Gated recurrent units
BLEU	: Bilingual Evaluation Understudy
WER	: Word Error Rate
PER	: position-independent Word Error Rate
TER	: translation Error Rate
SAP	: synthèse de parole
TTS	: text to speech
SAPI	: Speech API
AI	: Artificial Intelligence
ML	: Machine Learning
DL	: Deep Learning

ANN : Artificial Neural Networks
DNN : Deep Neural Networks
CNN : Convolutional Neural Networks
RElu : Rectified Linear Units
GAN : Generative Adversarial Neural Networks

Liste des Figures

Figure 1.1 : schéma général de l'appareil vocalique.....	5
Figure 1.2: système auditif.....	6
Figure 2.1 : les principales étapes de la RAP.....	27
Figure 2.2 : Principe de fonctionnement de la reconnaissance de mot isolé.....	28
Figure 2.3: Schéma de principe d'un système de reconnaissance de la parole continue.....	29
Figure 2.4: Schéma de principe d'un système de reconnaissance analytique.....	30
Figure 2.5 : Schéma de principe d'un système de reconnaissance globale.....	30
Figure 2.6: échelle de Mel.....	32
Figure 2.7 : les étapes de calculer les coefficients de MFCCs.....	32
Figure 2.8: Fenêtre de Hamming.....	34
Figure 2.9: Banc de filtres en échelle Mel.....	36
Figure 2.10 : les caractéristique statiques MFCC.....	37
Figure 2.11: les différents modèles de traduction automatique à base de règle.....	41
Figure 2.12: architecture de la traduction automatique statistique.....	42
Figure 2.13: exemple de représentation arborescente de traduction d'une phrase français vers anglaise.....	43
Figure 2.14: la traduction automatique à base de encodeur/décodeur.....	48
Figure 2.15: fonctionnement de l'encodeur/décodeur.....	49
Figure 2.16: l'architecture classique de la synthèse de parole.....	51
Figure 2.17: Représentation de diphone dans une séquence sonore.....	52
Figure 2.18: Un système de synthèse vocale basé sur un DNN.....	54
Figure 3.1: exemple de fonctionnement de DL.....	62
Figure 3.2: Un neurone biologique et ses principaux composants.....	63
Figure 3.3: Neurone formel (artificielle).....	64
Figure 3.4: exemple d'un réseau de neurone.....	65
Figure 3.5: réseaux de neurone multicouche.....	67
Figure 3.6: Réseaux de neurones profonds « DNN ».....	68
Figure 3.7: classification de base architecture CNN.....	70
Figure 3.8: les couche de « RNN ».....	73
Figure 3.9: Réseaux de neurone Récurrents « RNN ».....	74
Figure 3.10: Représentation simplifiée d'une cellule LSTM.....	77
Figure 3.11: Réseaux de neurone antagonistes génératifs « GAN ».....	78
Figure 4.1: python logo.....	81
Figure 4.2: PyChram logo.....	82
Figure 4.3: PyScripter logo.....	83
Figure 4.4: Spyder logo.....	83
Figure 4.5: visual studio logo.....	84
Figure 4.6: google colab Logo.....	84
Figure 4.7: Occurrences de phonèmes dans le corpus.....	87
Figure 4.8: Schéma fonctionnel du système de traduction automatique Arabe-Anglais.....	88
Figure 4.9 : fonctionnement de la reconnaissance.....	89
Figure 4.10 : le graphe statistique des résultats.....	91
Figure 4.11 : Graphe de l'intelligibilité.....	97
Figure 4.12 : le graphe de la naturel.....	97

Liste des Tableaux

<i>Tableau 1.1: les phonèmes arabes</i>	12
<i>Tableau 1.2 : les consonnes de l'arabe [6]</i>	13
<i>Tableau 1.3 : Alphabets phonétiques pour la prononciation anglaise</i>	17
<i>Tableau 1.4 : Quelques symboles phonétiques pour les consonnes anglaises</i>	17
<i>Tableau 2.1 :les avantages et les inconvénients de quelque logiciels de la RAP</i>	38
<i>Tableau 2.2: les avantages et les inconvénients des technique de la synthèse de parole</i>	54
<i>Tableau 2.3 : exemple des applications de la synthèse de parole</i>	56
<i>Tableau 4.1: Phonèmes et leurs notations choisies</i>	86
<i>Tableau 4.2 : la résultats de la reconnaissance de 10 audio</i>	91
<i>Tableau 4.3: Niveau de compréhension des signaux synthétiques</i>	96

Introduction Générale

La traduction automatique de la parole, à notre avis, est l'un des défis scientifiques et économiques les plus importants de cette décennie. La nouvelle économie n'a plus de frontières, grâce aux progrès d'Internet. Néanmoins, la barrière de la langue reste entière et difficile à franchir. Nous avons désespérément besoin de systèmes de traduction automatique pour mieux connecter les gens partout dans le monde.

La traduction automatique a mieux résisté en termes d'ouverture mondiale et la langue est devenue un obstacle majeur, contribuant à sa diffusion, son développement et son innovation avec les inventions avancées les plus importantes telles que la profondeur d'apprentissage et le traitement de la parole.

La traduction automatique de la parole est le processus qui consiste à convertir automatiquement une parole donnée dans une langue source vers une parole dans une langue cible. Ce processus n'est pas une simple substitution mot à mot ou phrase à phrase, il est bien plus complexe, passant par la reconnaissance automatique de la parole qui consiste à analyser la voix humaine pour la transcrire sous la forme d'un texte exploitable par une machine. Ensuite traduisant ce dernier tout en préservant le sens du texte d'entrée et en produisant un texte fluide dans la langue cible. Un traducteur doit interpréter et analyser tous les éléments dans le texte source et savoir comment chaque mot peut influencer un autre pour générer enfin le texte cible. A la fin nous passons au dernier module qui est la synthèse de la parole, le but de la synthèse de la parole à partir du texte est de calculer automatiquement un signal de parole correspondant à un énoncé écrit.

Le but de Notre travail est de contribuer à la réalisation d'un système de traduction automatique de la parole (Arabe-Anglais) basée sur l'apprentissage en profondeur.

Aussi, dans ce mémoire, nous consacrons le premier chapitre aux généralités sur traitement automatique de la parole, un aperçu de la parole, sa production et perception chez l'être humain, son acquisition et ses traitements afin de mieux comprendre les différents traitements nécessaires à la réalisation d'un système de traduction automatique. Ensuite,

nous donnons un aperçu de l'arabe et de l'anglais, où les lettres apparaissent dans les deux langues.

Dans le second chapitre, nous présenterons les différentes étapes d'un système de traduction automatique de la parole en détails.

Le chapitre trois est consacré aux réseaux de neurones artificielles et aux différents types d'apprentissage profond.

Le chapitre quatre portera sur notre contribution dans cette étude dont le but est une meilleure compréhension du fonctionnement de la traduction automatique. Elle consiste en la mise en œuvre d'un système de traduction automatique de la parole Arabe-Anglaise.

Nous concluons ensuite notre travail par une discussion autour de notre application, des difficultés rencontrées et des possibilités d'amélioration de ce travail.

Chapitre 1

Traitement de la parole

1.1 Introduction

La parole est un moyen de communication efficace et naturel entre les humains, et l'un des plus courants dans notre société. Il est facile de parler aux gens plutôt que de leur écrire ou de leur dessiner. La parole se caractérise par la particularité d'être générée et perçue instantanément par le cerveau, c'est pourquoi le traitement de la parole tend à remplacer ces fonctions par des systèmes automatisés.

Ces dernières années, de nombreuses recherches ont été menées pour générer des systèmes de traitement automatique de la parole. Le traitement automatique de la parole est un domaine de recherche actif, Cette discipline a connu un formidable développement, ses recherches sont riches mais difficiles, elle associe de nombreuses disciplines : traitement du signal, informatique, acoustique, sciences du langage, neurosciences et même l'intelligence artificielle.

Dans ce chapitre, nous allons d'abord donner un aperçu de la parole, de sa production et de sa perception, et introduire ses principales caractéristiques. Deuxièmement, nous donnons un aperçu de l'arabe et de l'anglais, où les lettres apparaissent dans les deux langues. Enfin, nous discutons des techniques de traitement automatique de la parole et de leurs domaines d'application.

1.2 Généralités sur la parole et la langue

1.2.1 Définition de la parole

La parole se définit comme une succession de séquences sonores et de silences, ce qui est la seule façon de communiquer les idées à travers un système complexe de sons émis par les organes d'articulation. Il se distingue des autres sons par ses propriétés acoustiques qui croisent leur articulation dans les mécanismes de production.

1.2.2 Production de la parole

1.2.2.1 Description l'appareil phonatoire

L'appareil phonatoire ou appareil vocalique est un ensemble d'organes et de muscles qui permettent à l'homme d'émettre des sons et des paroles [1].

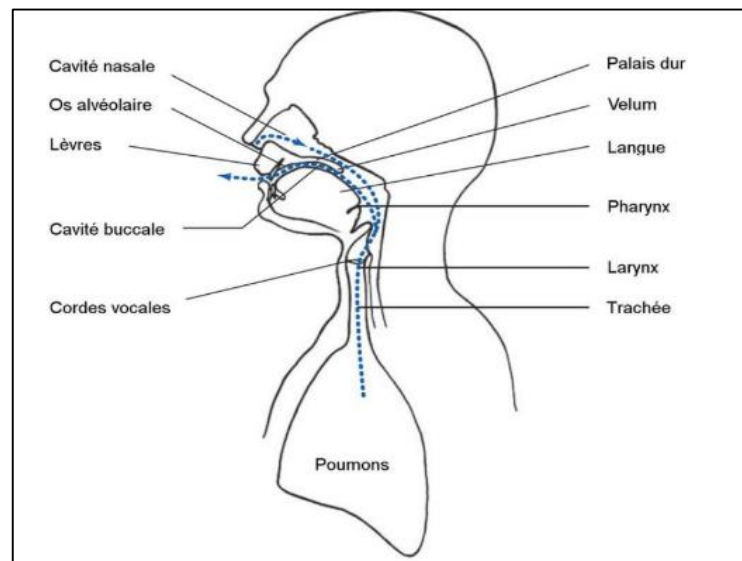


Figure 1.1 : schéma général de l'appareil vocalique [2]

- Les cordes vocales : membranes musculeuses qui coiffent l'orifice de la trachée (larynx), leur ouverture (glotte) présente une surface variable.
- Le conduit vocal : cavité de forme variable, allant des cordes vocales aux lèvres. Sa longueur est d'environ 17.5 cm chez l'adulte. Nous distinguons une cavité buccale et une cavité pharyngale.
- La cavité nasale : qui peut être mise en dérivation sur le conduit vocal, par l'abaissement du velum (comme c'est le cas dans la figure 1 .1).
- Les articulateurs : (langue, lèvres, mâchoires...) qui modifient la forme du conduit vocal.

1.2.3 Description de la production de la parole

La production de la parole est la façon dont les humains produisent la parole. C'est l'un des processus les plus complexes de la biologie humaine [1], car il dépend des interactions entre les systèmes neurologiques et physiologiques.

L'appareil respiratoire est considéré comme un générateur d'air qui fournit l'air qui est expiré par la trachée-artère. L'air passe ensuite par le larynx où sa pression est modulée grâce aux cordes vocales qui déterminent la taille de l'ouverture de (la glotte) par laquelle il peut passer. Finalement, l'air transite par le conduit vocal qui s'étend du pharynx au lèvres pour produire un message vocal [1].

1.2.4 Audition et perception de la parole

Comme il est encore difficile pour le cerveau de traiter les informations auditives, notre compréhension des informations verbales se divise en deux étapes :

La première étape consiste à transmettre l'information contenue dans le signal audio reçu au cerveau au moyen d'écouteurs. la deuxième étape consiste en la modification de l'information sur le langage dans le cerveau.

1.2.5 Structure d'un système auditif

Si les informations sont capturées et analysées par le récepteur, alors l'appareil audio qui transmet les informations sera inutile, et dans tous les récepteurs actuels, les humains ont acquis la capacité de détecter la signification cachée sous les sons prononcés par l'interlocuteur. Nous parlons ici de l'oreille humaine, qui est un modèle de l'appareil. Il est important de distinguer que la perception auditive est la fonction principale de l'oreille, tandis que la perception de la parole est la tâche du cerveau d'arriver à des concepts clairs. Nous allons maintenant présenter des illustrations d'équipements sonores [3] .

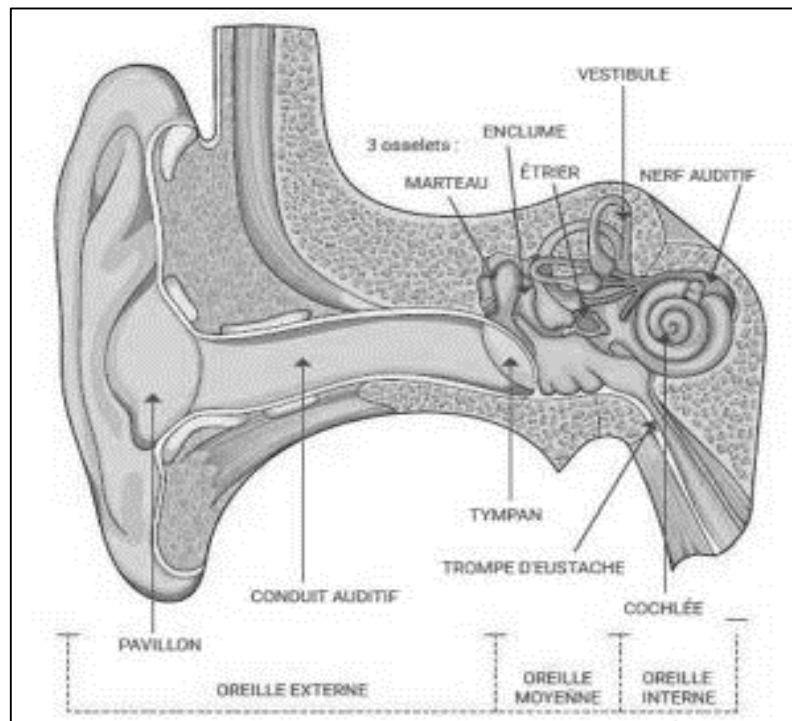


Figure 1.2: système auditif

Le système auditif ou oreille est divisé en 3 parties, et cette division est effectuée en fonction de la distance de l'air transportant le son de l'environnement :

Pour la première partie, l'oreille externe, cette partie correspond à la partie visible de l'organe, car elle est constituée de l'aile, du conduit auditif, des ailes et des lobes.

La deuxième partie, l'oreille moyenne, est constituée de 3 os (marteau, étrier et enclume) en plus du tympan.

La troisième partie, l'oreille interne, contient la cochlée et les canaux semi-circulaires.

1.2.5.1 Principes de perception

La perception de la parole est aussi importante que sa production. L'oreille externe capte les ondes sonores et le conduit auditif permet au son de se propager jusqu'au tympan, qui définit la limite entre l'oreille externe et l'oreille moyenne. Puisque les organes de l'oreille moyenne permettent la conversion du son en vibrations, une fois ces vibrations générées, la cochlée, composant principal de l'oreille interne, convertit ces vibrations en impulsions nerveuses, qui sont ensuite envoyées à la partie responsable du traitement. C'est le cerveau [3].

L'oreille humaine ne répond pas à toutes les fréquences, car la plage audible du champ auditif humain est comprise entre 20 Hz et 20 000 Hz, et l'oreille ne reçoit pas les sons d'une fréquence inférieure à 20 Hz, ni les sons d'une fréquence supérieure à 20 000 Hz..

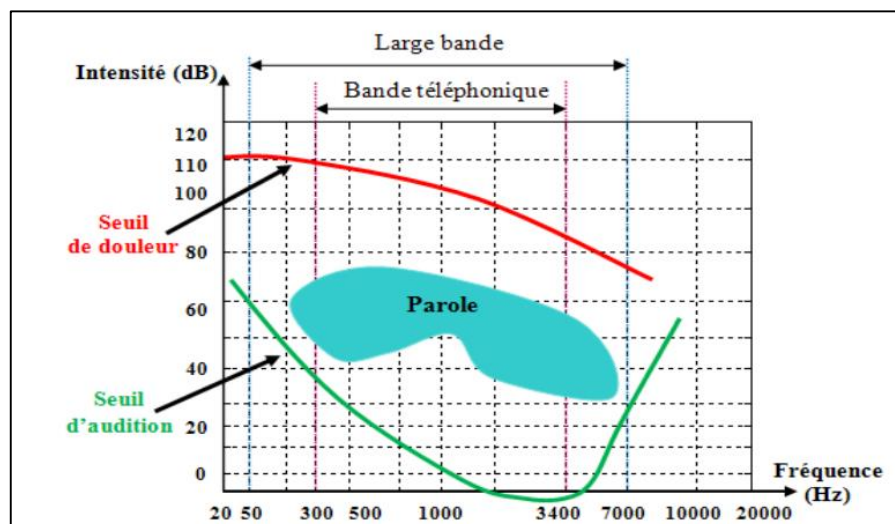


Figure 1.3 : champ auditif

1.2.6 Classification des sons de la parole

1.2.6.1 Sons non-voisés

Les sons non voisés, tels que certaines consonnes /f/et /s/..., ils peuvent être considéré comme un bruit blanc causé par le flux turbulent d'air à travers le tractus vocal. Par conséquent, les cordes vocales ne vibrent pas périodiquement. Comme le montre la figure 1.4 :

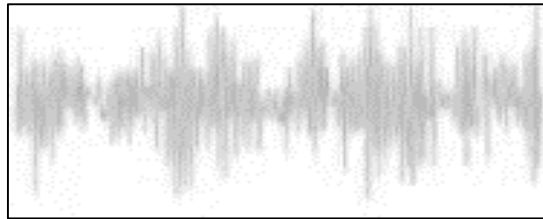


Figure 1.4: signal d'un son non-voisé

1.2.6.2 Sons voisés

Les sons voisés, telles que les voyelles /a/, /e/, /o/, les semi- voyelles /w/ et les nasales /b/, /d/, /n/..., sont produits par le passage de l'air des poumons à travers la trachée qui fait vibrer les cordes vocales, Elles sont caractérisées par une quasi-périodicité, une énergie élevée et de fréquence fondamentale, appelée « pitch ».

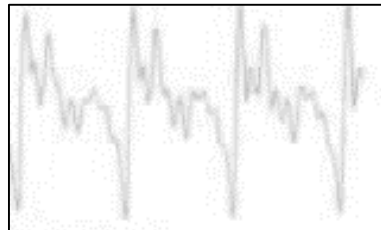


Figure 1.5 : Signal d'un son voisé

1.2.6.3 Silences

La parole est une séquence consécutive de période actives et de repos, de sorte que les silences sont des périodes pendant lesquelles les signaux utiles sont absents. En pratique il s'agit de bruits, d'origines diverses, d'énergie insignifiante par rapport au signal utile.

1.2.7 Caractéristiques de la parole

La parole est un signal réel d'énergie limitée, continue et non stationnaire, et la parole est le support sonore d'une information très complexe, diverse et répétitive. L'analyse d'un tel

signal est une tâche difficile car il existe de nombreux attributs associés et il existe des caractéristiques dominantes, appelées caractéristiques audios, dont chacune a une signification perceptuelle.

1.2.7.1 Fréquence fondamentale

La fréquence fondamentale est la fréquence de vibration des cordes vocales pour les sons semi-périodiques, car elle correspond à la fréquence des cycles d'ouverture et de fermeture des cordes vocales. Ses différences déterminent la hauteur du son qui constitue la perception de la hauteur lorsque les sons sont disposés de bas en haut, et il change lentement au fil du temps. La fréquence de base varie d'un locuteur à l'autre selon le sexe (homme, femme) et l'âge comme suit :

- De 80Hz à 200Hz pour une voix masculine.
- De 150Hz à 450Hz pour une voix féminine.
- De 200Hz à 600Hz pour voix d'enfant.

1.2.7.2 Spectre fréquentiel

Le spectre fréquentiel est la propriété acoustique dont dépend principalement le timbre d'un son, puisque le timbre est une propriété qui permet d'identifier un locuteur par la simple écoute de sa voix.

1.2.7.3 Energie

L'énergie de la parole est liée à la pression de l'air dans le larynx et est généralement plus forte pour les syllabes audibles (fortes) que non prononcées (basses). L'intensité sonore est généralement mesurée en dB.

1.2.7.4 Durée (rythme)

Il est exprimé en termes de durée d'émission du son, car la durée du son dépend de la pression atmosphérique, généralement de l'air expiré. Une unité de durée est mesurée par le nombre de trames qu'elle contient.

Pour calculer la durée de chaque trame, il faut mis à jour les deux événements sur le signal de parole, qui déterminent les symboles initial et final pour cette trame.

1.2.8 Description de la parole

Le signal de parole peut être traité de différentes manières selon la cible, et une fois qu'il est numérisé pour la première fois, le nombre de techniques possibles est si nombreux que nous citerons ci-dessous les outils liés aux signaux de parole :

1.2.8.1 Description temporelle

Le signal de parole est un signal quasi-stationnaire. Cependant, sur un horizon de temps supérieur, il est clair que les caractéristiques du signal évoluent significativement en fonction des sons prononcés comme illustré

La première approche pour étudier le signal de parole consiste à observer la forme temporelle du signal. Nous pouvons à partir de cette forme temporelle extraire un certain nombre de caractéristiques qui pourront être utilisées pour le traitement de la parole. Il est, par exemple, assez clair de distinguer les parties voisées, dans lesquelles on peut observer une forme d'onde quasi-périodique, des parties non voisées dans lesquelles un signal aléatoire de faible amplitude est observé. De même, on peut voir que les petites amplitudes sont beaucoup plus représentées que les grandes amplitudes ce qui pourrait justifier les choix faits pour le codage de la parole.

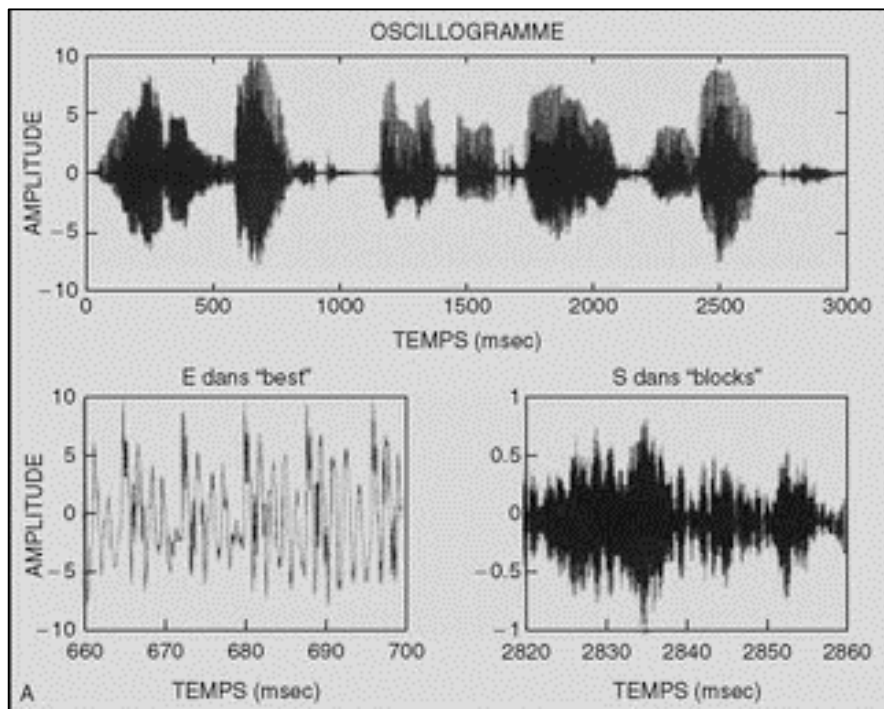


Figure 1.6: Analyse du signal de la parole

1.2.8.2 Description fréquentielle

La deuxième façon de caractériser et de représenter les signaux de parole consiste à utiliser des représentations spectrales. Les méthodes spectrales occupent une place importante dans l'analyse de la parole, où, entre autres choses, l'oreille effectue une analyse de fréquence du signal qu'elle perçoit de plus, la fréquence de la parole peut être bien décrite. La transformation de Fourier est une opération qui permet de représenter en fréquence des signaux non périodiques. Il s'agit de l'analogie des séries de Fourier pour les fonctions périodiques. Une fonction non périodique pouvant être considérée comme une fonction dont la période est infinie. Ce déplacement vers la limite nous conduit de séries à intégrales.

La transformée de Fourier :

$$F(f) = F\{f(t)\} = \int_{-\infty}^{+\infty} f(t) e^{-j2\pi ft} dt \quad (1.1)$$

1.2.8.3 Description temps / fréquence

Dans ce cas, la représentation la plus courante est le spectrogramme. Le spectrogramme pour un signal de parole ou spectrogramme est une représentation visuelle du signal acoustique dans le domaine fréquentiel il appartient à l'analyse de fréquence dépendant du temps (Figure 1.7).

Le spectrogramme représente le module de la transformée de fourrier discrète calculée sur une fenêtre temporelle.

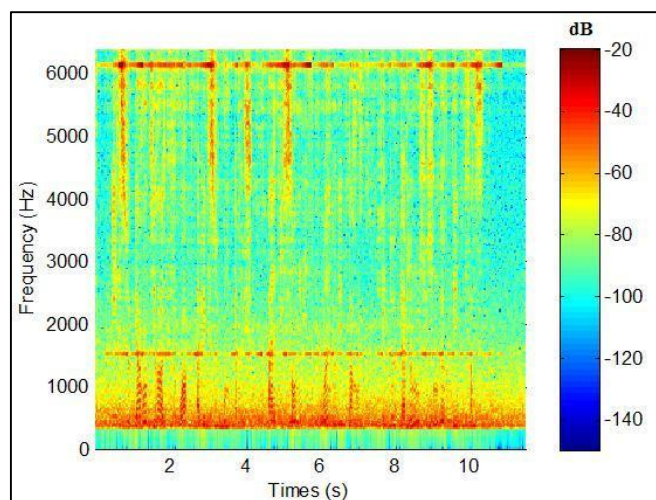


Figure 1.7: Exemple d'un spectrogramme [4]

1.2.9 Définition de la langue

Le langage en général désigne tous les moyens possibles de compréhension entre les personnes ou entre les autres êtres vivants. Ainsi, le mouvement de la main pour saluer est un langage, et le hochement de tête dans une position spécifique est un langage. Ainsi, ils servent le langage. Mêmes buts que les mots cherchent à atteindre.

Quant au langage, dans son sens propre, il renvoie aux images verbales bien connues puisqu'il est constitué de mots, de phrases et de règles. C'est ce qui distingue l'homme des autres créatures car il est le seul être qui utilise sa propre langue avec un système de parole spécial.

Environ 200 langues sont soumises à un système d'écriture spécifique soit sous forme de symboles comme le chinois ou bien sous forme alphabétique comme l'arabe et l'anglais.

1.2.9.1 La langue arabe

La langue arabe est apparue dans le passé et a occupé une place prépondérante dans le passé et le présent, ce qui la distingue des autres langues.

Aujourd'hui, la langue arabe est classée en cinquième position parmi les langues les plus utilisées au monde avec plus de 274 millions de locuteurs et une langue officielle dans 26 pays. La langue arabe adopte le système alphabétique dans sa langue, avec un nombre de lettres égal à 28. Elle s'écrit de droite à gauche comme les langues d'Asie du Sud-Est (Chine, Corée, Japon...)

La langue arabe a également suscité l'intérêt de nombreux chercheurs, ce qui a conduit à la traduction d'ouvrages de l'anglais vers l'arabe [5].

a) Phonèmes et articulation

Tableau 1.2.1: les phonèmes arabes

Isolé	Début	Milieu	Fin	Nom	Phonème	Isolé	Début	Milieu	Fin	Nom	Phonème
ا	ا	ا	ا	'alif'	[a:]	ض	ض	ض	ض	'daad'	[d]
ب	ب	ب	ب	'baa'	[b]	ط	ط	ط	ط	'taa'	[t]
ت	ت	ت	ت	'taa'	[t]	ظ	ظ	ظ	ظ	'daa'	[z]
ث	ث	ث	ث	'thaa'	[θ]	ع	ع	ع	ع	'ayn'	[ʕ]
ج	ج	ج	ج	'gym'	[dʒ]	غ	غ	غ	غ	'ghayn'	[ɣ]
ح	ح	ح	ح	'haa'	[H]	ف	ف	ف	ف	'kaaf'	[f]

خ	خ	خ	خ	'khaa'	[x]	ق	ق	ق	ق	'qaaf'	[q]
د	د	د	د	'daal'	[d]	ك	ك	ك	ك	'kaaf'	[k]
ذ	ذ	ذ	ذ	'dhaal'	[ð]	ل	ل	ل	ل	'laam'	[l]
ر	ر	ر	ر	'raa'	[z]	م	م	م	م	'mym'	[m]
ز	ز	ز	ز	'zayn'	[r]	ن	ن	ن	ن	'nuwn'	[n]
س	س	س	س	'syn'	[s]	ه	ه	ه	ه	'haa'	[h]
ش	ش	ش	ش	'shyn'	[ʃ]	و	و	و	و	'waaw'	[u :]
ص	ص	ص	ص	'saad'	[s̥]	ي	ي	ي	ي	'yaa'	[i :]

La phonétique : La phonétique étudie les propriétés physiques (articulatoires, acoustiques...) des sons. Elle s’intéresse aux sons eux-mêmes, indépendamment de leur fonctionnement les uns avec les autres. Les sons sont considérés en tant qu’unités physiologiques.

Le phonème : C’est la plus petite unité séparée ou distincte dans le domaine linguistique, et la chaîne parlée est généralement isolée par la segmentation, qui se compose généralement de consonnes et de voyelles. Le nombre de phonème varie dans chaque langue, par exemple, le français a 36 phonèmes.

La langue arabe est appelée la langue de « ض », car c'est la seule langue qui prononce cette lettre. La langue arabe se caractérise par un système différent en ce qui concerne les consonnes et les voyelles.

b) Les consonnes :

Tableau 1.2.2 : les consonnes de l’arabe [6]

Modes Lieux	Occlusives	Emphatiques	Fricatives	Nasales	Liquides	Glides (semi-voyelle)
Labiales	ب b		ف f	م m		و w
Interdentales		ظ	ذ ث			
Dentales	د ت t d	ض ط		ن n	ل ر r l	
Sifflantes		ص S	ز س z s			
Palatales	ج		ش			ي y
Vélaires	ك k		خ غ			
Uvulaire	ق q					
Pharyngales			ح ع e			

Glottales	ء		ه			
------------------	---	--	---	--	--	--

Occlusives : Les phonèmes de cette classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont donc constituées de deux parties successives : une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion.

Emphatique : On l'obtient en modifiant la forme du résonateur buccal dans sa partie arrière par rétraction et exhaussement de la racine de la langue.

Fricatives : Dans cette classe sont regroupés les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences.

Nasale : Produites de la même manière que les occlusives nasales mais l'air n'est pas, cette fois, comprimé dans le conduit vocal. Le vélum est en effet abaissé pour permettre à l'air d'être expiré

Liquides : leur durée et leur énergie sont généralement plus faibles, elles sont sonores
Glides (semi-voyelle ou les semi-consonnes) : Elles ont la structure acoustique des voyelles mais ne peuvent en jouer le rôle car elles ne sont que des transitions vers d'autres voyelles qui sont les véritables noyaux syllabiques.

c) Les voyelles

La voyelle est représentée par le signe diacritique et sa place sur les lettres (les consonnes), elle est absente à l'écrit dans la majorité des textes arabes. Les voyelles sont divisées en deux types [7] :

- **Voyelles brèves** : Divisées en trois parties,
 - Voyelle brèves simples : fatha, dama, kasra
 On les appelle aussi les signes diacritiques, Ou les voyelles courtes
 fatha (َ) → [ba], dama (ُ) → [bou], kasra (ِ) → [bi].
 - Voyelle brèves doublées (tanween)
 Elles sont représentées par la même voyelle répétée deux fois, c'est-à-dire deux mouvements simultanés (ّ) → [ban], (ُّ) → [bun], (ِّ) → [bin].
- L'absence de voyelle (soukun) :

Il est représenté dans un mouvement silencieux où lorsqu'il est placé sur la lettre, il se transforme en une lettre silencieuse, (بْ)→ [∅].

- Les voyelles longues

Elles sont représentées par les voyelles longues après la lettre en augmentant le nombre de voyelles. Ouvrir avec [ا], combiner avec [و] ou casser avec [ي], (بَا)→[bà], (بُو)→[bù], (بِي)→[bii]/[y].

- Chadda :

Il s'agit tout simplement du dédoublement de la lettre sur laquelle elle se trouve. A titre d'exemple, si on a une chadda sur la lettre arabe [b], cela signifie que l'on a deux [b] à la suite et qu'elles ont fusionnées ensemble. Concrètement pour la lecture, la règle générale est que sur le premier [b] on a une soukoun, tandis que sur le deuxième on aura la simple voyelle (fatha, dama, kasra), (بَّبْ)→[∅ba], (بُبْ)→[∅bu], (بَبِ)→ [∅bi].

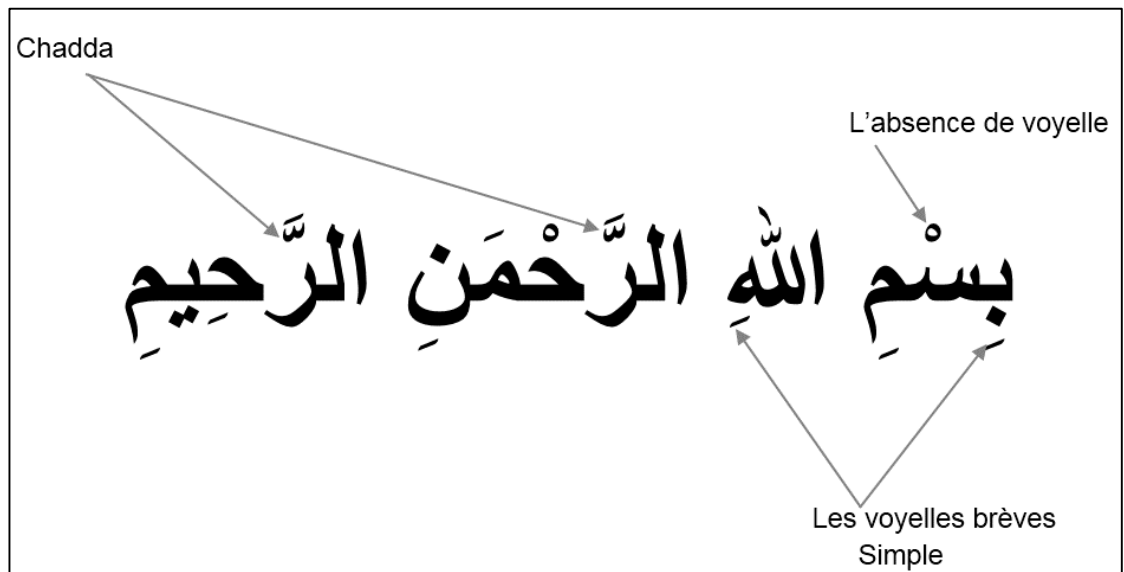


Figure 1.8 : exemple d'une phrase voyelle

L'alphabet arabe contient des lettres spéciales, ce qui signifie qu'on ne le voit pas dans d'autres langues, notamment la lettre [ض] [ق] [ع] [ظ], qui sont difficiles à prononcer, comme illustre la figure1.10

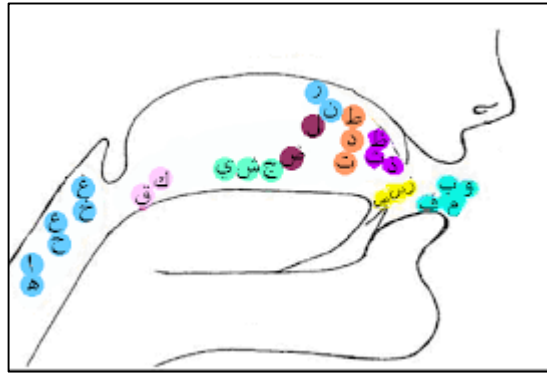


Figure 1.9 : Articulation des lettres arabes [8]

1.2.9.2 La langue anglais

La langue anglaise est originaire d'Angleterre et est la langue dominante aux États-Unis, au Royaume-Uni et dans de nombreux autres pays du reste des continents. C'est également le premier choix pour une langue étrangère dans la plupart des pays du monde en raison de sa facilité de règles et sa souplesse dans les mots et les expressions.

Le nombre d'anglophones aujourd'hui est d'environ un milliard de personnes et il est classé première dans les langues les plus parlées dans le monde [5].

La langue anglaise est basée sur un système d'écriture de gauche à droite en plus du système alphabétique.

La langue anglaise a été affectée par d'autres langues et a changé plusieurs fois au cours de l'histoire pour maintenir son existence dans le monde. Aujourd'hui, elle est considérée comme l'une des langues les plus importantes, car de nombreuses personnes du monde entier l'apprennent.

a) Phonèmes et l'articulation

Le nombre de l'alphabet est de 26. Les lettres peuvent être soit en majuscules ou en minuscule. Le début de chaque phrase est en majuscule et le reste en minuscule.

Les lettres anglaises sont divisées en voyelles et consonnes.

Tableau 1.2.3 : Alphabets phonétiques pour la prononciation anglaise

Consonnes			Voyelle	
p pill	t till	k kill	i beet	ɪ bit
b bill	d dill	g gill	e bait	ɛ bet
m mill	n nil	ŋ ring	u boot	ʊ foot
f fell	s seal	h heal	o boat	ɔ bore
v veal	z zeal	l leaf	æ bat	a pot/bar
θ thigh	ʃ chill	r reef	ʌ butt	ə sofa
ð thy	dʒ gin	j you	aɪ bite	aʊ bout
ʃ hill	ɹ which	w witch	ɔɪ boy	
ʒ measure				

b) Les consonnes :

- Les consonnes sont des sons produits avec restriction ou fermeture du tractus vocal.
- Les consonnes sont classées en partie sur la base où, dans le conduit vocal, le flux d'air est restreint (le lieu d'articulation).
- Les principaux lieux d'articulation sont : bilabiale, Labiodentales, interdentaires, alvéolaires, palatales, vélaire, uvulaire et glottale [9].

Tableau 1.2.4 : Quelques symboles phonétiques pour les consonnes anglaises

Lieux \ Modes	Bilabiales	Labiodentales	Interdentaire	Alvéolaire	Palatales	Vélaire	Glottales
Arrêt sans voix voisé	p b			t d		k g	ʔ
Nasal (voisé)	m			n		ŋ	
Fricatif sans voix voisé		f v	θ ð	s z	ʃ ʒ		h
Affriqué sans voix voisé					tʃ dʒ		
Glide sans voix voisé	ɹ w				j	ɹ w	
Liquide sans voix voisé				r l			

c) Principaux lieux d'articulation

- **Bilabiaux** [p] [b] [m][w] [ɱ] : Produits en rapprochant les deux lèvres.
- **Labio-dentaires** [f] [v] : Produits en touchant la lèvre inférieure aux dents supérieures
- **Interdentaires** [θ] [ð] : Produits en mettant le bout de la langue entre les dents
- **Alvéolaires** [t] [d] [n] [s] [z] [l] [r] : Tous ces éléments sont produits en élevant la langue jusqu'aux alvéoles crête en quelque sorte ;
 - [t, d, n] : produits par la pointe de la langue touchant l'alvéole crête (ou juste devant) ;
 - [s, z] : produits avec les côtés de l'avant de la langue relevés mais la pointe abaissée pour permettre à l'air de s'échapper ;
 - [l] : la pointe de la langue est relevée tandis que le reste de la langue reste en bas, Pour que l'air puisse s'échapper sur les côtés de la langue (ainsi le [l] est un latéral du son) ;
 - [r] : l'air s'échappe par la partie centrale de la bouche ; soit la pointe de la langue est recourbée derrière la crête alvéolaire ou le haut de la langue est repliée derrière la crête alvéolaire.
- **Palatines** [ʃ] [ʒ] [tʃ] [dʒ][j] : Produits en élevant la partie avant de la langue vers le palais.
- **Vélaires** [k] [g] [ŋ] [w] [ɰ] : Produits en élevant l'arrière de la langue au sol du palais ou du velum.
- **Uvulaires** [ʀ] [q] [ɣ] : Produits en élevant l'arrière de la langue jusqu'à la luette.
- **Glottales** [h] [ʔ] : Produits en limitant le flux d'air à travers la glotte ouverte ([h]) ou en arrêter complètement l'air au niveau de la glotte (un coup de glotte : [ʔ]).

d) Le lieu d'articulation :

- Arrêts [p] [b] [m] [t] [d] [n] [k] [g] [ŋ] [tʃ][dʒ] [ʔ] : Produits en arrêtant complètement le flux d'air dans la cavité buccale pendant une fraction de seconde. Tous les autres sons sont continus, ce qui signifie que le flux d'air est continu à travers la cavité buccale.
- **Fricatives** [f] [v] [θ] [ð] [s] [z] [ʃ] [ʒ] [x] [χ] [h] : Produits en obstruant fortement le flux d'air de manière à causer des frictions.
- **Affriqués** [tʃ] [dʒ] : Produits par une fermeture d'arrêt qui est libérée avec beaucoup de friction.

- **Liquides** [l] [r] : Produits en provoquant une certaine obstruction du flux d'air dans la bouche, mais pas assez pour provoquer de réelles frictions.
- **Glissements** [j] [w] : Produits avec très peu d'obstruction du flux d'air et sont toujours suivis d'une voyelle.

e) **Les voyelles :**

Les voyelles utilisent les cordes vocales, et elles peuvent être qualifiées de lettres vibrantes car elles vibrent lorsqu'elles sont prononcées. Les voyelles sont classées selon la hauteur ou la hauteur de la langue, si là la langue est à l'avant ou à l'arrière de la bouche, et si ou pas les lèvres sont arrondies

- Voyelles fermées : [i] [ɪ] [u] [ʊ]
- Voyelles moyennes : [e] [ɛ] [o] [ə] [ʌ] [ɔ]
- Voyelles ouvertes : [æ] [a]
- Voyelles Antérieures (avant) : [i] [ɪ] [e] [ɛ] [æ]
- Voyelles centrales : [ə] [ʌ]
- Voyelles postérieures (arrière) : [u] [ʊ] [o] [ɔ]

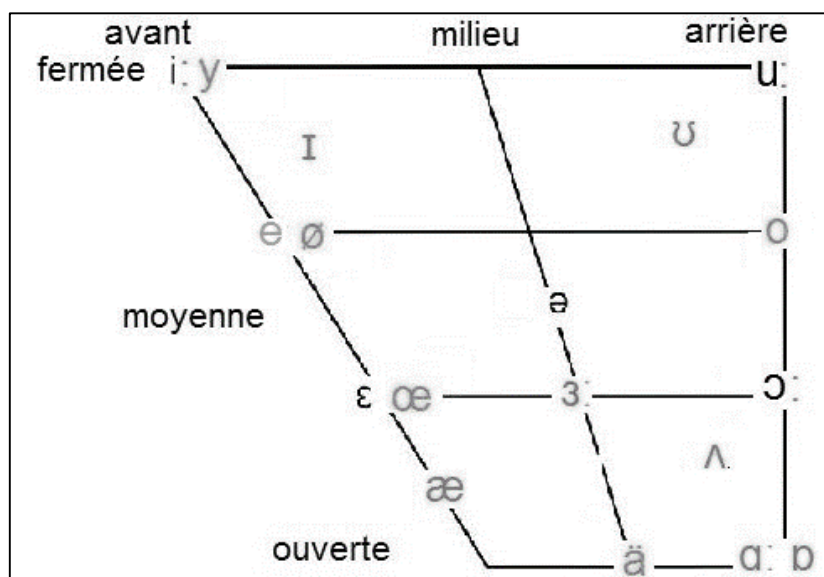


Figure 1.10 : la prononciation de voyelle [10]

f) **L'articulation :**

- **Voyelles tendues** [i] [e] [a] [u] [o] [ai] [aʊ] : Sont produits avec plus de tension dans la langue. Peut survenir à la fin de mots.

- **Voyelles relâchées** [ɪ] [ɛ] [ʊ] [ɔ] [ə] [æ] [ʌ] [ɔɪ] : Sont produits avec moins tension de la langue. Peut ne pas se produire à la fin de mots
- **Diphthongues** [aɪ] [aʊ] [ɔɪ] : Une séquence de deux voyelles (par opposition aux monophthongues que nous avons regardé jusqu'à présent)

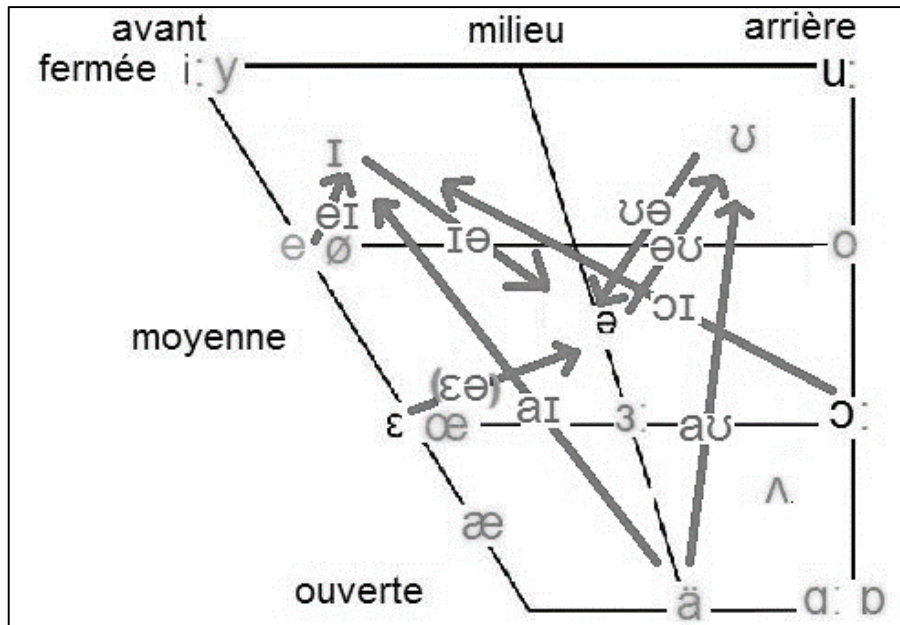


Figure 1.11 : la prononciation de diphthongues [10]

1.3 Traitement automatique de la parole

Le traitement automatique de la parole est une discipline scientifique qui vise à étudier, interpréter, comprendre et utiliser automatiquement les signaux vocaux.

Les aspects du traitement de la parole comprennent l'acquisition, la manipulation, le stockage, le transfert et la sortie de signaux vocaux. L'entrée est appelée reconnaissance vocale et la sortie est appelée synthèse vocale.

1.3.1 Classification des tâches de la traitement automatique de la parole

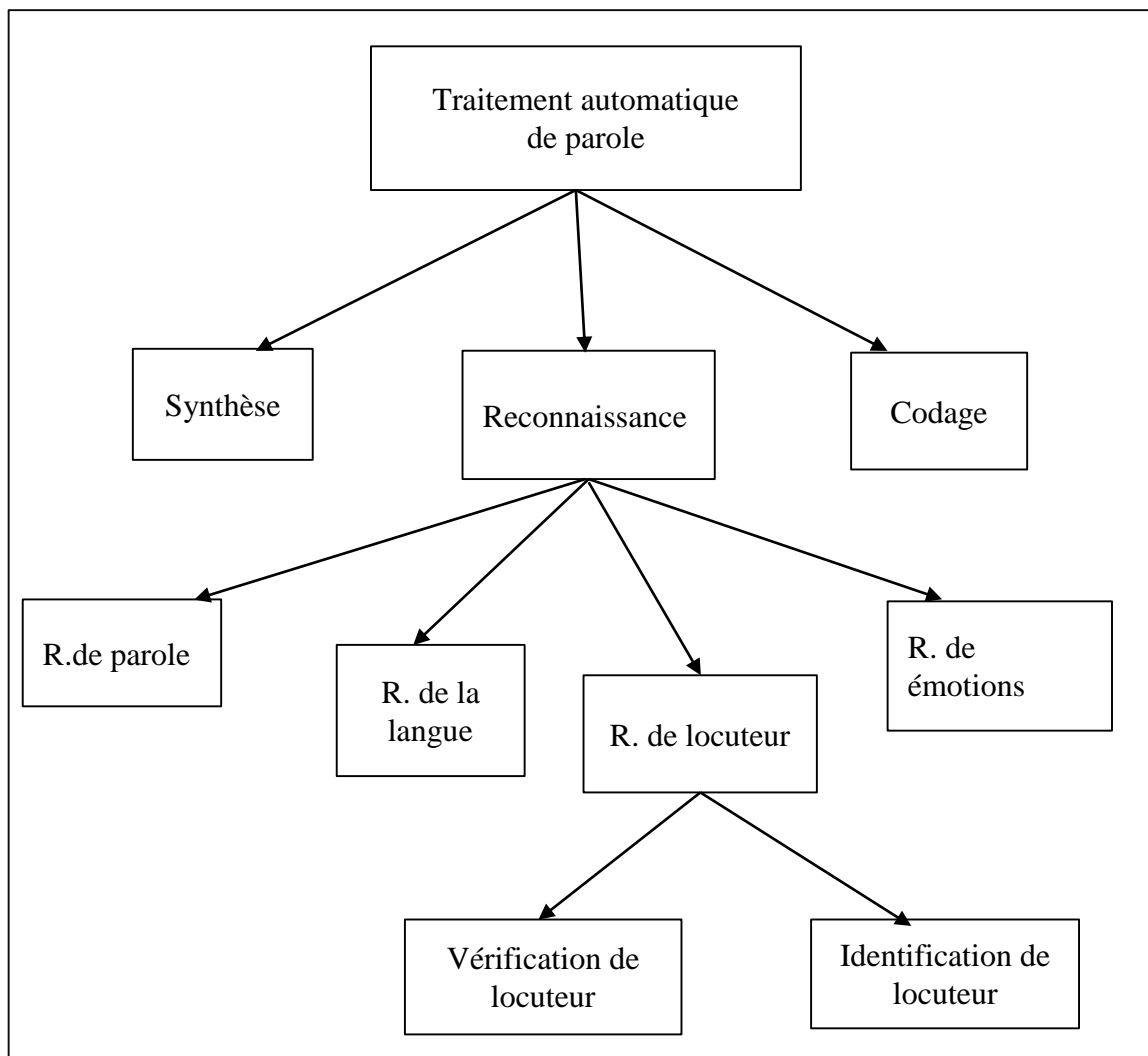


Figure 1.12 : Traitement de la parole

1.3.1.1 Synthèse automatique de la parole

La synthèse automatique de la parole est une technologie qui convertit le texte en une séquence de parole, permettant aux gens d'entendre le texte de leur choix.

1.3.1.2 Codage

Le rôle de l'encodeur est de permettre la transmission ou le stockage de la parole à un débit binaire réduit, ce qui nécessite naturellement une attention particulière aux caractéristiques de production et de perception de la parole.

1.3.1.3 La reconnaissance automatique

Le terme reconnaissance automatique de la parole fait référence à plusieurs types de systèmes dont la tâche est de décoder les informations portées par les signaux de parole. Il existe plusieurs types de reconnaissance automatique.

a) La reconnaissance automatique de la parole

La reconnaissance automatique de la parole, souvent appelée à tort reconnaissance vocale, est l'endroit où les identificateurs vocaux extraient des informations à partir de signaux vocaux.

Le mode vocal se caractérise par la façon dont vous pouvez parler au système. Il existe plusieurs modes de parole différents, dont les plus importants sont :

Identification des mots isolés, Diviser l'information parlée en ses composants de base est un sujet difficile. Pour éviter cela, de nombreux projets ARP se concentrent sur l'identification des mots prononcés isolément. Identifier des mots isolés ou supposons que tous les mots prononcés sont séparés par un silence de durée supérieure à quelques dixièmes de seconde, essentiellement réalisée par proximité Mondial.

Reconnaissance vocale continue, Bien que la méthode la plus appropriée pour la reconnaissance vocale Séquentielle est une méthode analytique, plusieurs tentatives ont été faites. Généralisation des méthodes de reconnaissance globale. Les étapes de ces systèmes Le décodage de la parole acoustique est le plus basique et dépend généralement de niveau lexical.

b) La reconnaissance automatique du locuteur

La reconnaissance automatique du locuteur est l'identification d'une personne par la parole car elle tente d'extraire les caractéristiques de la parole de chaque personne et de les utiliser pour créer une signature audio qui permet d'authentifier la parole de chaque personne. En fait, la reconnaissance du locuteur comprend deux tâches différentes :

Identification automatique du locuteur : L'identification automatique du locuteur est l'identification d'une personne qui doit comparer un message audio à un ensemble de références audio et réussir ce test pour déterminer qui parle.

Vérification automatique du locuteur : La vérification automatique du locuteur, comprenant la détermination de l'identité supposée du message vocal correspond à l'identité

réelle, et la vérification de l'adéquation du message vocal demandé à envoyer avec la référence vocale du locuteur qu'il prétend être, après quoi la réponse est binaire, après quoi la réponse est bidirectionnelle, et l'un ou l'autre des orateurs est en fait le premier orateur approuvé.

c) La reconnaissance automatique de la langue

En déterminant quelle langue parle le locuteur, la reconnaissance de la langue pour spécifier une langue d'entrée qui peut être utilisée dans la traduction.

d) La reconnaissance automatique des émotions

Les émotions sont un ensemble de sentiments qui se manifestent chez une personne à cause d'un événement, et ils entraînent également des changements physiques et psychologiques qui affectent notre comportement selon le type de sentiments. Les mouvements du corps et les expressions faciales sont un langage profond qui dépend de la psychologie, car chaque mouvement simple du corps exprime un certain sentiment.

L'âge, le genre humain, selon des expériences particulières...

La reconnaissance automatique des émotions est basée sur l'intelligence artificielle, dans laquelle tout ce qui concerne à la psychologie et au langage du corps est introduite, c'est-à-dire l'identification à l'état psychologique de locuteur (triste, nerveux, content...). Cette identification conduit à l'invention de robots dotés d'émotions similaires aux émotions humaines et d'intelligence émotionnelle humaine.

1.3.2 Domaine d'application

Le traitement de parole est un domaine vaste et a trouvé une place importante à cette époque, car elle a été très demandée et est utilisée aujourd'hui dans :

- Les serveurs d'information par téléphone
- La messagerie
- La sécurité possible grâce à la signature vocale
- Permet l'autonomie...

1.4 Conclusion

Dans ce chapitre, nous avons introduit les concepts de base de traitement automatique de la parole. Ensuite, nous avons fourni des informations générales sur l'arabe et l'anglais. Le but de ce chapitre est de définir les concepts que nous utilisons dans le reste de notre travail. Dans le chapitre suivant nous allons donner plus de détail sur la traduction automatique de la parole.

Chapitre 2

Traduction automatique de la parole

2.1 Introduction

Un système de traduction automatique parole-parole (Speech To Speech STS) est constitué de trois modules principaux : Module de reconnaissance automatique de parole, module de traduction et module de synthèse. STS est utilisé dans la plupart des domaines, y compris la traduction automatique.

Pour obtenir une traduction automatique STS arabe-anglais, il est nécessaire de convertir le signal parole arabe en un texte arabe en utilisant la reconnaissance de parole, ensuite le texte sera traduit en texte anglais, et enfin, le texte anglais est converti en parole anglaise intelligible en utilisant la synthèse de parole.

Dans ce chapitre, nous allons voir la reconnaissance de la parole, la traduction automatique de l'arabe vers l'anglais et la synthèse de parole. Par la suite nous aborderons le fonctionnement, les différents types, les avantages de chaque étape.

2.2 Reconnaissance automatique de la parole « RAP »

2.2.1 Définition

La reconnaissance automatique de la parole « Automatic speech recognition » est une technique informatique qui permet d'analyser et détecter la parole et d'extraire (le message oral) dans le but de générer une chaîne de mots ou phonèmes représentant ce que la personne a prononcé.

2.2.2 Fonctionnement

Le processus de reconnaissance vocale commence par donner à la machine un fichier audio en entrée, puis la parole est transférée à la première étape, qui passe par l'extraction de paramètres, qui est responsable de la conversion de la parole phonétique en un vecteur acoustique et est converti à l'étape suivante, qui est basée sur l'intelligence artificielle (l'apprentissage automatique ou les réseaux de neurones) pour convertir des vecteurs acoustiques en texte (figure 2.1).[11]

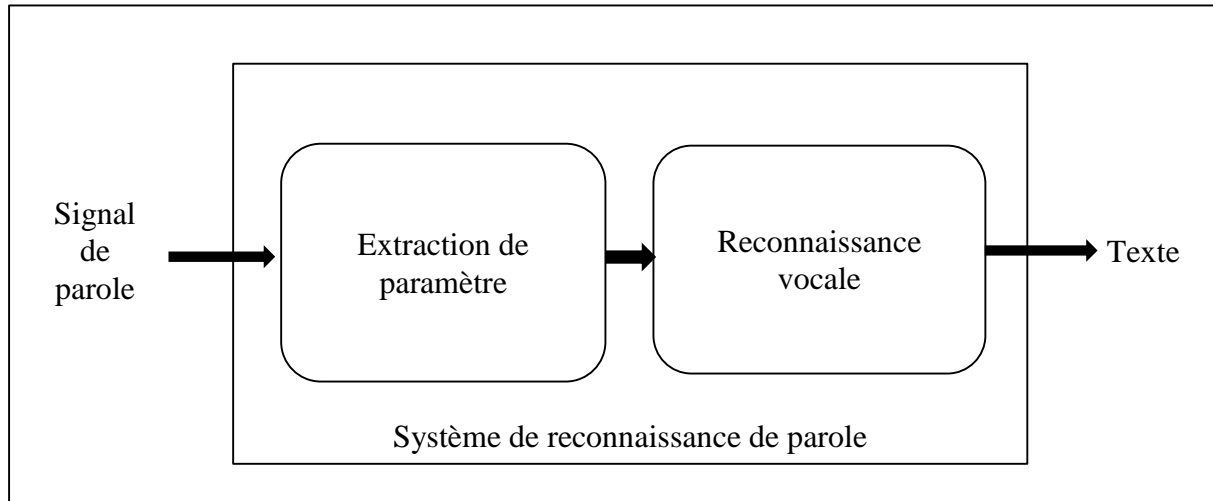


Figure 2.1 : les principales étapes de la RAP

2.2.3 Systèmes de la reconnaissance automatique de la parole

Les systèmes de la reconnaissance de la parole peuvent être séparés en plusieurs classes différentes en dérivant les types d'énoncés qu'ils ont la capacité de reconnaître. Il existe 4 types

2.2.3.1 Reconnaissance de mot isolé

De nombreux de projet de la RAP se sont intéressés à la reconnaissance de mots prononcés isolément pour éviter le problème de segmentation d'un message parlé en ses constituants élémentaires. La reconnaissance des mots isolés exige généralement que chaque mot séparé par des silences "absence de signal audio" de durée supérieure à quelques dixièmes de seconde, se fait essentiellement par l'approche globale. Comme les exemples suivants de la reconnaissance des mots isolés en la langue arabe : استمر، كفاح، نجح... [12]

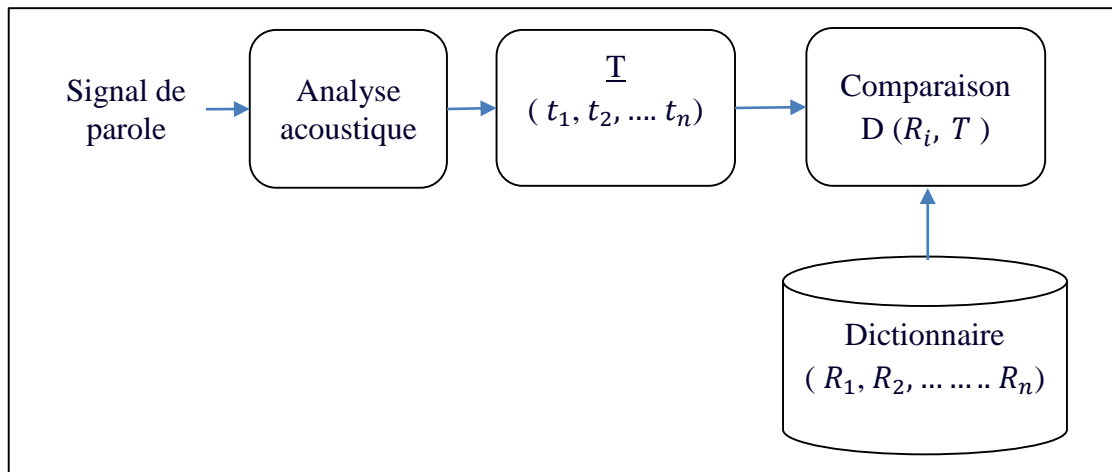


Figure 2.2 : Principe de fonctionnement de la reconnaissance de mot isolé

2.2.3.2 Reconnaissance de mot enchaîné :

Il s'agit de saisir les mots dans une séquence sans silence ni séparation. Pour parvenir à l'écriture correcte d'un texte, il est nécessaire de connaître le nombre de mots de la séquence et de connaître la fin et le début de chaque mot de la séquence et de l'interconnexion entre mots adjacents, en définissant des règles de segmentation pour certains vocabulaires en fonction de normes phonétiques. en utilisant une série de références constituées d'une séquence de mots prononcés par le locuteur de manière isolée. la méthode de partitionnement se fait par une comparaison globale de la séquence à identifier avec une série de références.

2.2.3.3 Reconnaissance de mot connecte

Les systèmes de mots connectés (ou plus précisément "déclarations connectées") sont similaires aux mots isolés, mais permettent à des instructions distinctes d'être "combinées" avec une pause minimale entre eux [12].

2.2.3.4 Reconnaissance de parole continue

La parole continue montre une série de modules réalisant les différentes étapes du processus de reconnaissance (Figure), au début, un module acoustique permet d'extraire les caractéristiques physiques du signal, destinées au module phonétique. Ce module reconnaît les sons élémentaires de la langue, en faisant appel à un dictionnaire de phonèmes. Ensuite le module lexical et le module phonologique souvent confondus, reconnaissent les mots, et utilisant pour cela un lexique des mots autorisés par l'application considérée, ainsi que des règles phonologiques décrivant les assemblages possibles de phonèmes dans la langue. Les modules syntaxiques et sémantiques reconnaissent la phrase à l'aide d'une description des

règles de grammaire et de la signification des mots admis pour l'application en question avec un niveau supplémentaire, appelé prosodique, portant sur la mélodie, le rythme, l'intensité du discours oral, intervient en parallèle avec les autres modules et fournit les informations qu'il extrait du niveau acoustique (hauteur, intensité, rythme) à tous les autres niveaux [12].

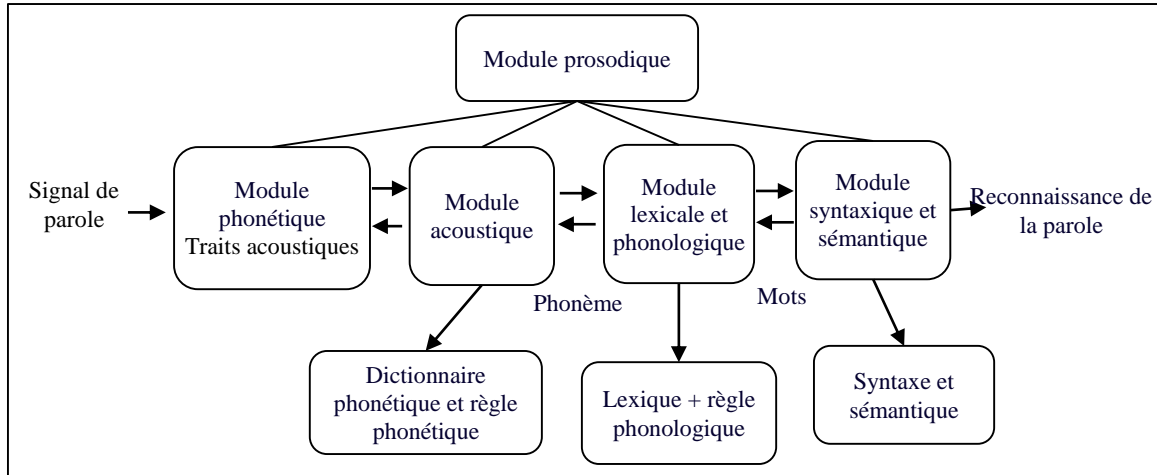


Figure 2.3: Schéma de principe d'un système de reconnaissance de la parole continue

2.2.4 Approche de la reconnaissance automatique de la parole

2.2.4.1 L'approche analytique

L'approche analytique est une méthode utilisée pour résoudre le problème de la reconnaissance du parole continu "grand vocabulaire". Cette approche consiste à segmenter le signal vocal en constituants élémentaires (mot, phonème, syllabe...) puis à identifier, et enfin régénère la phrase prononcée par étape successive ou bien mener une tâche de reconnaissance, en intégrer des modules linguistiques : niveau lexical et syntaxique (contraient liées aux règles de la grammaire), sémantique (le sens de mot), et pragmatique (la cohérence de la phrase) [13].

Le processus de l'RAP dans une telle méthode peut être décomposé en deux opérations [14] :

1. Segmentation du signal de parole sous forme d'une suite de petits segments
2. Identification des segments trouvés en termes d'unités phonétiques

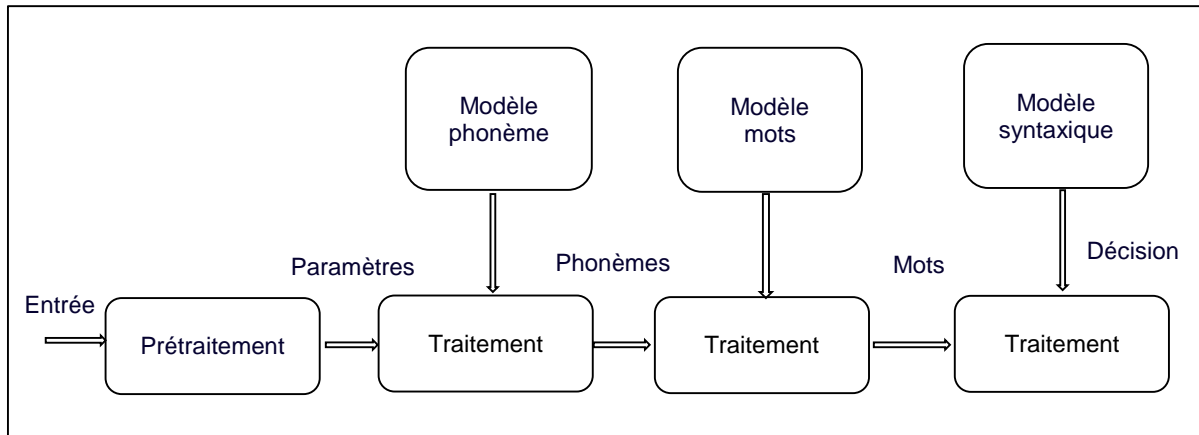


Figure 2.4: Schéma de principe d'un système de reconnaissance analytique

2.2.4.2 Approche globale

L'approche globale est une méthode qui considère le mot comme unité de base 'c'est à dire non décomposée' indépendamment de la langue. Cette méthode basée sur l'idée de donner au système au moins une image acoustique de chacun des mots qu'il devra identifier par la suite.

L'approche globale vise généralement pour la reconnaissance des mots isolés ou des mots enchaînés appartenant à des vocabulaires réduits

[13].

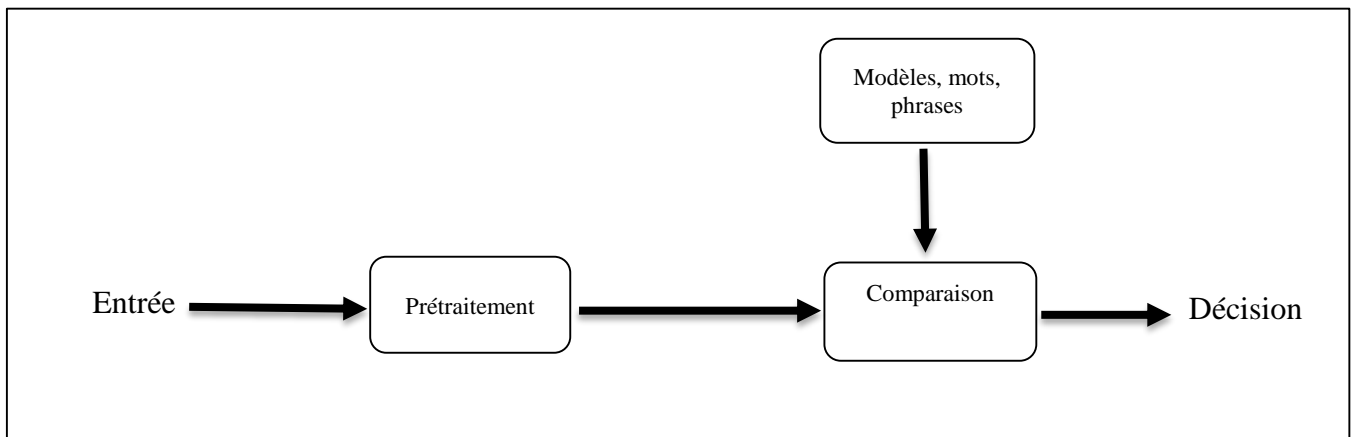


Figure 2.5 : Schéma de principe d'un système de reconnaissance globale

- L'approche statistique

L'approche statistique est une technique développée pour réaliser des systèmes RAP parmi les réseaux neuronaux, HMM... Cette approche est fondée sur le même principe de

méthodes générales, mais avec l'exploitation des niveaux linguistiques. De ce fait une analyse acoustique est nécessaire pour convertir le signal vocal en une suite de vecteurs acoustiques. Ces vecteurs sont considérés comme des exemples d'apprentissage pour construire de modèles de reconnaissance qui vont classifier les nouveaux signaux de parole

Dans notre travail, nous nous sommes penchés sur les systèmes RAP basés sur des modèles de réseau nuerons profonds spécifiant CNN [13]. Les différents composants des systèmes sont décrits dans le chapitre suivant.

2.2.5 Extraction des paramètres

Le signal vocal transporte plusieurs informations comme le message linguistique, l'identité du locuteur, ainsi que ses émotions, la langue adoptée, etc. Un système RAP consiste à récupérer seulement le message linguistique indépendamment des autres informations. Dans un tel système, l'analyse acoustique consiste à extraire du signal vocal un ensemble de paramètres pertinents dans le but de réduire la redondance du signal vocal pour une tache de reconnaissance des mots.

Le processus d'extraction des paramètres du signal de parole consiste à transformer ce signal en une suite des vecteurs acoustiques. Cette nouvelle représentation est plus compacte à la modélisation statistique et vectorielle. Pour le problème de la RAP, les paramètres les plus employés reposent sur une représentation cepstral du signal de parole.

Dans la littérature, il existe plusieurs types d'extraction des paramètres dans la RAP tel que LPC, PLP, MFCC. Les paramètres MFCC sont choisis dans ce travail.

- **Les coefficients MFCCs**

Les coefficients MFCC (Mel Frequency Cepstral Coefficients » sont des coefficients cepstraux par le passage de l'échelle fréquentielle linéaire à une échelle fréquentielle non linéaire ' l'échelle Mel'. La motivation d'extraction de coefficients MFCCs est leur correspondance avec la réponse en fréquence d'une oreille humaine.

L'échelle Mel redistribue les fréquences selon une échelle non linéaire qui simule la perception humaine des sons. Cette échelle est linéaire pour les basses fréquences (inférieures à 1000Hz) et logarithmique pour les hautes fréquences.

La conversion de la fréquence linéaire f en Mel est donnée par :

$$\text{Mel}(f) = 2595 \log_{10}(1 + f/700) \quad (2.7)$$

Où f est la fréquence en Hz

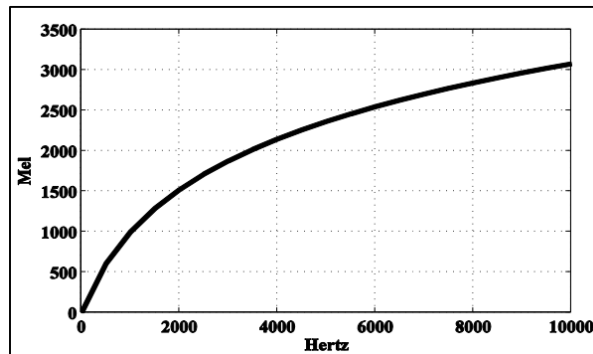


Figure 2.6: échelle de Mel

- **Les calculs des paramètres MFCCs :**

Le calcul de la MFCC est basé sur le court terme donc le calcul de ces paramètres

se réalise de la façon suivantes (voire figure)

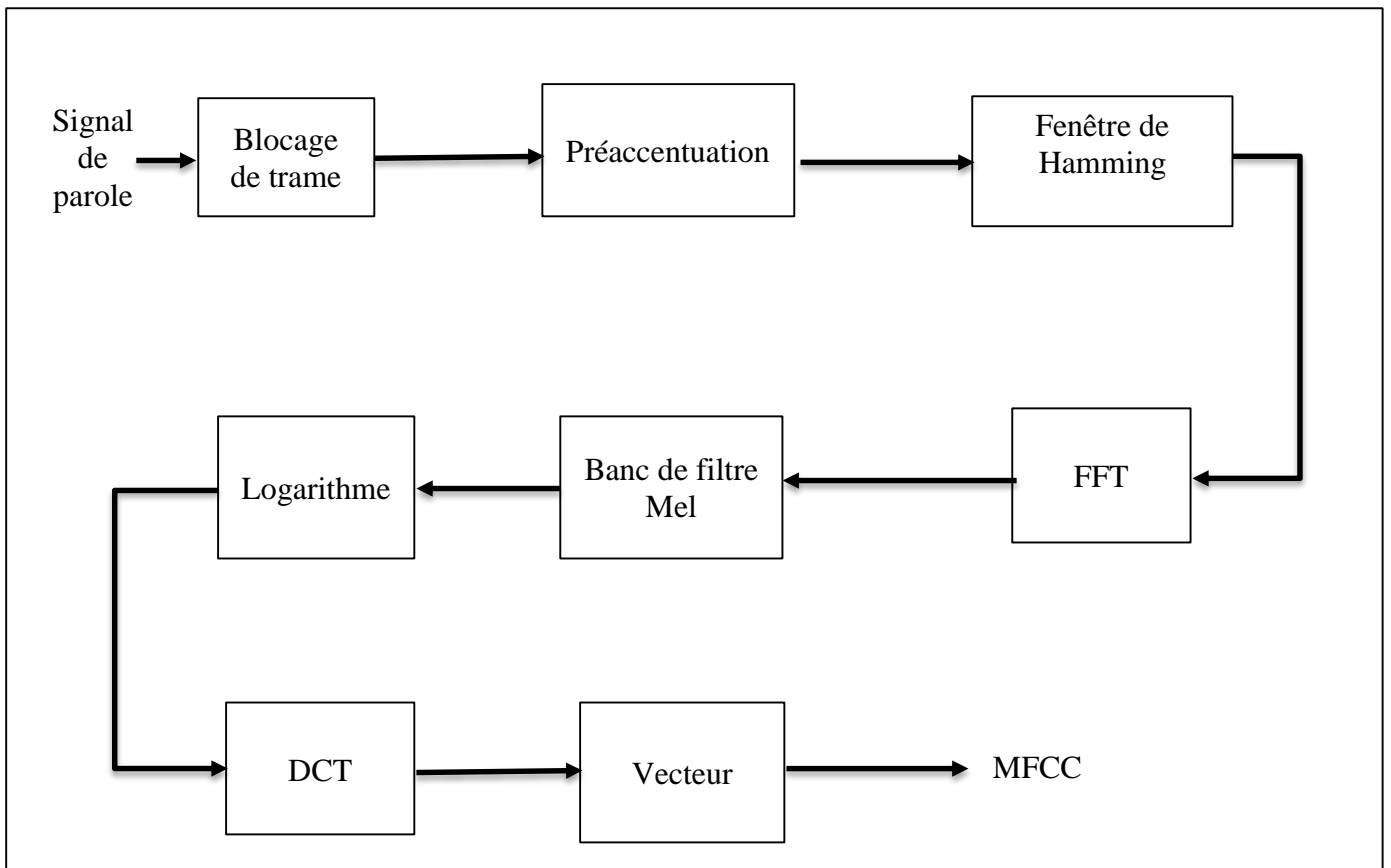


Figure 2.7 : les étapes de calculer les coefficients de MFCCs [15]

- **Segmentation en trames**

Les méthodes du traitement de signal utilisées dans l'analyse du signal vocal opèrent sur des signaux stationnaires, alors que le signal vocal est un signal non stationnaire. Afin de remédier à ce problème, l'analyse de ce signal est effectuée sur des trames successives de parole, de durée relativement courte sur lesquelles le signal peut en général être considéré comme quasi stationnaire. Dans cette étape de segmentation, le signal continue est ainsi découpé en trames de N échantillons de parole « En général N est fixé de telle manière à ce que chaque trame corresponde à environ 20 à 30 ms de parole » avec un pas d'avancement de M trames ($M < N$). C'est-à-dire que deux trames consécutives se chevauchent sur $N-M$ échantillons [15].

- **Préaccentuation**

Le signal échantillonné est pré accentué en premier lieu, en appliquant un filtre numérique RIF de premier ordre. Afin de relever les hautes fréquences qui sont moins énergétiques que les basses fréquences, qui sont généralement réduites par le procédé de production de parole.

La préaccentuation du signal est obtenue en appliquant le filtre suivant [16]:

$$y(n) = x(n) - \alpha x(n-1) \quad (2.2)$$

Où :

$y(n)$: le signal de sortie

$x(n)$: la séquence d'échantillons obtenue à partir du signal temporel continu $x(t)$

α : facteur de préaccentuation prenant une valeur comprise dans $[0.9 ; 1]$

- **Fenêtrage**

Le signal de parole est un signal non stationnaire, il est analysé par une fenêtre glissante de durée court de l'ordre 25ms avec un recouvrement de 50% où le signal de parole peut être considéré quasi stationnaire.

Il existe plusieurs types de fenêtres, parmi ces fenêtres la fenêtre Hamming. Il est la plus convenable à la parole, car elle entraîne un minimum de distorsion spectrale du signal de parole, par rapport aux autres fenêtres "l'atténuation de la valeur du signal à zéro lorsqu'elle s'approche des bords de la fenêtre pour éviter les discontinuités".

Elle est définie par :

$$y(n) = s(n) * w(n) \quad (2.3)$$

où :

$y(n)$: le signal de sortie à l'instant n

$s(n)$: le signal d'entrée à l'instant n

$w(n)$: la fenêtre de pondération employée

Pour la fenêtre de Hamming :

$$w(n) = \begin{cases} 0.54 - 0.46 \cos \frac{2\pi n}{N}, & 0 \leq n \leq N - 1 \\ 0, & \text{ailleurs} \end{cases} \quad (2.4)$$

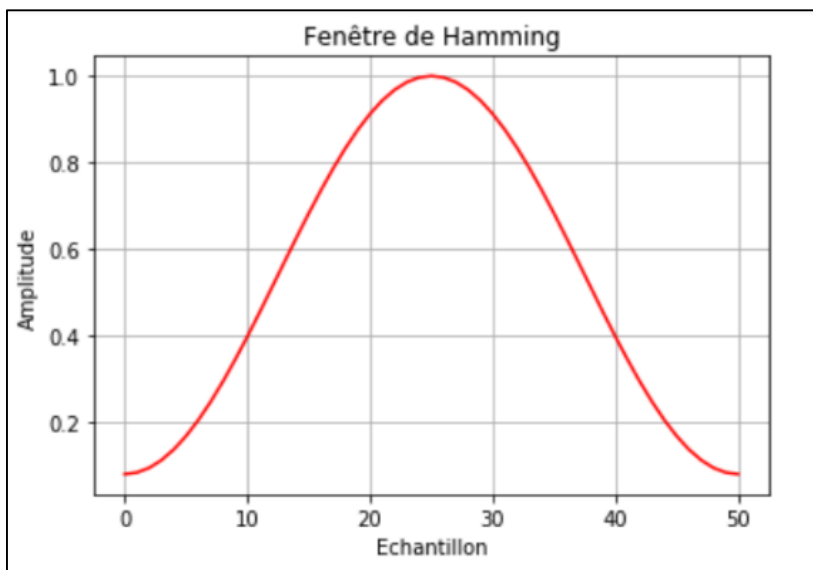


Figure 2.8: Fenêtre de Hamming [16]

- **Transformée de Fourier discrète FFT**

Au cours de cette étape chacune des trames, de N valeurs, est convertie du domaine temporel au domaine fréquentiel. On peut trouver plusieurs algorithmes pour cette transformée, ces algorithmes sont connus sous le nom de FFT.

La FFT est un algorithme rapide pour le calcul de la FFT et est définie par la forme Les valeurs obtenues sont appelées le spectre par cette relation :

$$X(k) = \sum_{n=0}^{N-1} x(n) \times e^{-\frac{j2\pi nk}{N}}; 0 < K < N-1 \quad (2.5)$$

$X(k)$: la sortie FFT

N : le nombre d'échantillons dans la trame

- **Banc de filtre Mel**

Un banc de filtre est une série de filtres à bande passante réparti d'une façon équidistante dans l'échelle Mel. Pour définir un banc de filtre, on doit définir la forme de chaque filtre ainsi que la localisation de ses fréquences'' gauche, central et droit''. Ces filtres ayant plusieurs formes, et ils peuvent être différemment placés sur l'échelle de fréquence.

Le spectre du signal (le spectre précédemment) est multiplié avec des filtres triangulaires (figure 2.9). dont les bandes passantes sont équivalentes en domaine mel-fréquence. Ces filtres possèdent la caractéristique suivante : plus la fréquence est élevée, plus la bande passante est large, et donc une meilleure résolution temporelle des hautes fréquences

Les points frontières $B[m]$ des filtres en Mel-fréquence sont calculés ainsi :

$$B[m] = B(f_1) + m \frac{B(f_h) - B(f_1)}{M+1}; 0 < m < M-1 \quad (2.6)$$

Avec :

M : le nombre de filtres.

f_h : la fréquence la plus haute du signal

f_1 : la fréquence la plus basse du signal

Dans le domaine fréquentiel, les points $f[m]$ discrets correspondants sont calculés par

$$f[m] = \left[\frac{N}{F_s} \right] B^{-1} \left[B(f_1) + m \frac{B(f_h) - B(f_1)}{M+1} \right] \quad (2.7)$$

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

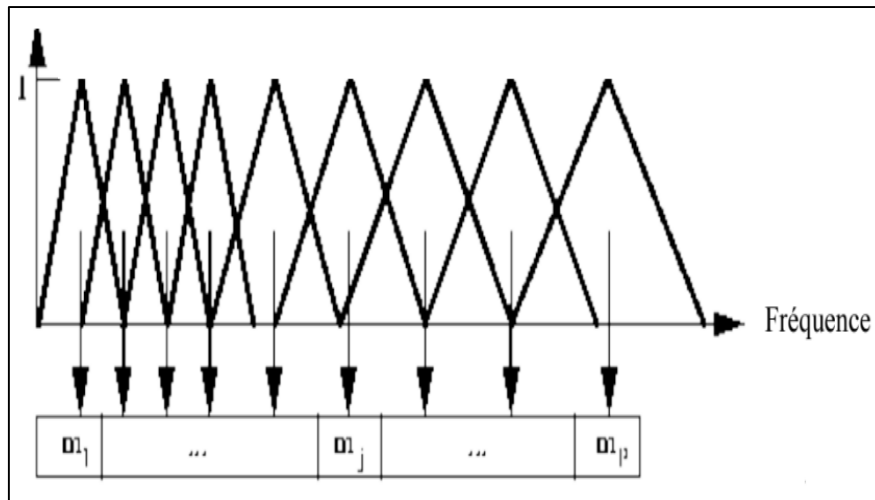


Figure 2.9: Banc de filtres en échelle Mel [16]

- **Logarithme :**

À la sortie des filtres, un logarithme d'énergie (ou un logarithme de spectre d'amplitude) est calculé pour obtenir un spectre lissé et stable. Les coefficients log du banc de filtres à l'échelle Mel FB peuvent être calculés à partir des sorties des filtres par l'équation [17] :

$$S(m) = 20 \log_{10} \left(\sum_{k=0}^{N-1} |X(k)| H_m(k) \right); \quad 0 < m < M \quad (2.8)$$

Où :

$S(m)$: spectre algorithmique

$X(k)$: est la FFT de la trame

$H_m(k)$: est la fonction de transfert du filtre Mel

M : est la fonction de transfert du filtre Mel

- **DCT Discret Cosinus Transforme**

C'est l'étape finale. Dans cette étape permet d'obtenir des coefficients peu corrélés à partir des coefficients aux sorties des filtres. Pour obtenu ces coefficients "MFCC" par une transformée en cosinus discrète « transformée l'échelle fréquentiel vers l'échelle temporelle » du logarithme des énergies issues du banc de filtre. L'expression de ces coefficients est donnée par [16] :

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left[\frac{\pi n(m - \frac{1}{2})}{M}\right]; \quad 0 \leq n \leq M \quad (2.9)$$

Où :

$C(n)$: les coefficients MFCC

$S(m)$: spectre logarithmique

N : le nombre d'échantillon dans chaque trame

M : le nombre des bancs de filtres

La figure montre Le vecteur de caractéristiques MFCCs

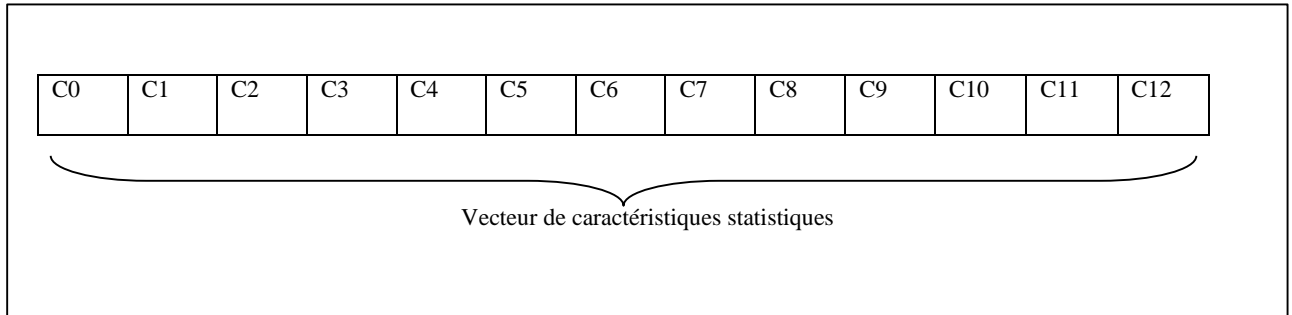


Figure 2.10 : les caractéristique statiques MFCC

Généralement les coefficients MFCC sont désignés comme des paramètres statiques, puis qu'ils contiennent seulement l'information sur une trame donnée. Afin d'améliorer la représentation de la trame, il souvent proposé d'introduire de nouveaux paramètres dans le vecteur des paramètres. Ces paramètres se compose des coefficients différentielles « les dérivées cepstraux » du premier ordre sont les coefficients Δ . Elles montrent la vitesse de variation de ces vecteurs dans le temps. Les dérivées deuxièmes sont les coefficients $\Delta\Delta$. Elles donnent des informations sur l'accélération de la parole [16].

2.2.6 Les Applications de la RAP

Les applications de la reconnaissance automatique de la parole sont multiples et peuvent variées selon leurs types. Les systèmes de la RAP est généralement utilisées dans :

- **Dicté vocale** : ou saisie vocale, permet de la transcription d'un texte dicté par un utilisateur de la meilleure manière possible. Toutefois, le texte transcrit doit respecter les règles orthographiques et grammaticales propres à la langue considérée. Ces systèmes sont utilisés pour transcrire des textes sur Word ainsi que des rapports.
- **Commande vocale** : les systèmes à commande vocales offrent une interaction entre l'utilisateur et la machine grâce à des commandes vocales, ces commandes

représentent des mots isolés. De nombreux systèmes à commande vocales sont utilisés dans des automobiles, des avions de chasse etc...

- Traduction automatique : de conversations téléphoniques avec un interlocuteur de langue étrangère

2.2.7 Les meilleurs logiciels de la RAP

Les logiciels de reconnaissance automatique de la parole apporté des changements significatifs dans le monde virtuel, où les logiciels de la RAP sont des applications qui utilise les algorithmes de reconnaissance automatique de la parole pour identifier de langage et retranscrire en texte lisible avec un degré élevé de précision grâce à l'intelligence artificielle et DL. Le tableau suivant représente certains des logiciels les plus utilisés en RAP

Tableau 2.2.1 : les avantages et les inconvénients de quelque logiciels de la RAP

Logiciel de RAP	Les avantages	Les inconvénients
Dragon	<ul style="list-style-type: none"> • Utilise l'intelligence artificielle pour une meilleure compréhension • Fait peu de fautes d'orthographe 	<ul style="list-style-type: none"> • Sans micro, il est difficile de compter sur la fiabilité de la transcription
E-speaking	<ul style="list-style-type: none"> • Plus de 100 commandes disponibles • Un tarif très abordable 	<ul style="list-style-type: none"> • N'est pas précis que les autres logiciels de RAP
Dictation.io	<ul style="list-style-type: none"> • Comprend les principales langues à travers le monde "Anglais, Français. " 	<ul style="list-style-type: none"> • Vous devez avoir Google chrome pour pouvoir l'utilisé

	<ul style="list-style-type: none"> Le nombre de commande est très satisfaisant 	
Amazone Lex	<ul style="list-style-type: none"> Idéal si vous souhaitez créer un bot répondant à des questions humaines 	<ul style="list-style-type: none"> Réservé à un usage professionnel
Trint	<ul style="list-style-type: none"> Transcription possible dans 30 langues 	<ul style="list-style-type: none"> Pas vraiment pertinent si vous avez de gros volumes à traiter

2.2.8 Les avantages et les inconvénients de la reconnaissance de la parole

- Les avantages**

La reconnaissance vocale est donc une innovation offerte de nombreux avantages, comme :

- La reconnaissance vocale est utilisée pour fournir une assistance croissante aux étudiants ayant des besoins spéciaux.
- La reconnaissance vocale offre plus de possibilités d'emploi aux personnes handicapées.
- La reconnaissance vocale permet la réalisation d'interfaces vocales, ou interfaces homme machine (IHM).
- La reconnaissance vocale peut rendre les utilisateurs d'ordinateurs plus efficaces en réduisant les erreurs humaines lors de communications.
- La reconnaissance vocale est efficace pour gagner du temps et de l'argent.
- La reconnaissance vocale permet une interaction avec les engins électroniques encore plus fluide.

- Les inconvénients**

Quelques inconvénients de la reconnaissance de la parole :

- Les techniques de reconnaissance de la parole sont insuffisantes. Certains paramètres peuvent perturber l'utilisation de cette technique biométrique (bruit autour, qualité du microphone enregistrée, bruit différent, etc.).
- La manière dont vous parlez et prononcez certains mots différents, et le programme doit apprendre ou s'adapter à cette phrase étrange.
- La reconnaissance du son exige une excellente qualité acoustique. On ne peut pas imaginer installer cette technologie là où il y a très peu de bruit.
- Faible niveau de différenciation entre deux votes, de sorte que cette technique n'est pas fiable.

2.3 Traduction automatique de la langue « MT »

La traduction en général est une application informatique à la traduction des textes d'une langue de source dans une langue cible, La traduction est divisée en deux parties : la traduction humaine et la traduction automatique. La traduction humaine, comme il est clair, est utilisée par des humains ou un traducteur compétent pour les deux langues, tandis que la machine utilise un ordinateur sans l'intervention d'un traducteur humain.

2.3.1 Définition de MT

La traduction automatique est effectuée à l'aide d'un logiciel, est basée sur un traitement de langue automatique à l'aide d'un grand nombre de données ou de ressources linguistiques pour arriver au sens correct de la langue cible.

La traduction automatique est basée sur une simple substitution de mots dans la première langue pour la seconde, ce qui à son tour affecte de manière très claire l'exactitude de la traduction et la clarté de ses défauts.

La notion de traduction automatique a émergé dans le 17ème siècle. et à l'avancement de la science. la traduction est devenue l'une des exigences fondamentales, ce qui a contribué à l'émergence de l'approche et de plusieurs façons d'obtenir un sens véritable dans la langue seconde.

2.3.2 Les types de la traduction automatique

Depuis l'arrivée de la traduction automatique est apparu plusieurs approches qui peuvent être comptés aujourd'hui dans :

2.3.2.1 Traduction automatique à base de règle (RBMT)

Le logiciel de traduction automatique fondé sur des règles repose sur l'utilisation de nombreuses grammaires et de millions d'entrées de dictionnaire par paires de langues. Cette traduction automatique fondée sur des règles fonctionne comme suit [18] :

- **Analyse** : Analyse du texte source sous forme de représentations intermédiaires dans le langage source.
- **Conversion** : Transférer ces déclarations intermédiaires aux déclarations intermédiaires dans la langue cible.
- **Génération** : Création de texte de langue Nouveau but des représentations intermédiaires du langage cible.

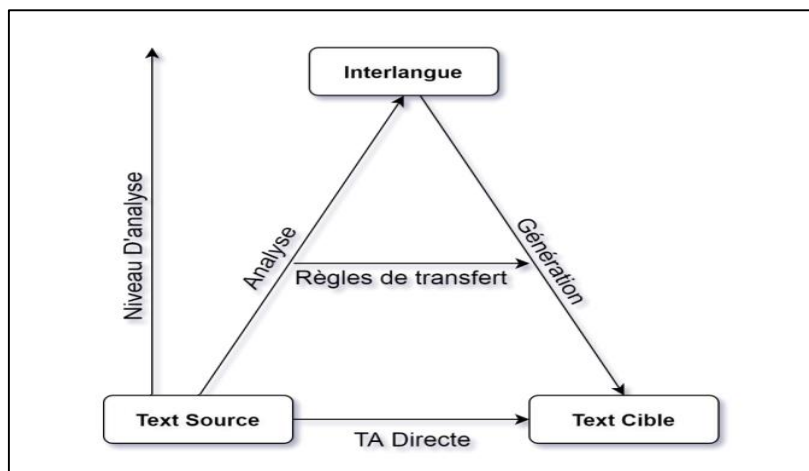


Figure 2.11: les différents modèles de traduction automatique à base de règle

2.3.2.2 Traduction automatique statistique (SMT)

Les modèles statistiques reposent sur de grandes quantités de données et sur des modèles statistiques auto construits à partir de groupes monolingues et bilingues. Le processus de traduction statistique est également un processus basé sur l'apprentissage automatique, et il utilise généralement deux types de modèles pour traduire, l'un d'eux a été formé pour travailler sur un groupe monolingue et est largement utilisé pour améliorer la traduction et

corriger les erreurs, et le second travaille sur le bilinguisme dont la tâche est d'estimer les distributions de probabilité pour traduire la phrase d'entrée en une phrase d'accès.

Bien sûr, on l'appelait la traduction statistique en raison de sa dépendance à la probabilité et aux statistiques, où le sens du mot peut différer d'une phrase à l'autre selon le contexte, et c'est là que réside l'habileté des statistiques à déterminer laquelle des traductions est corrigée.

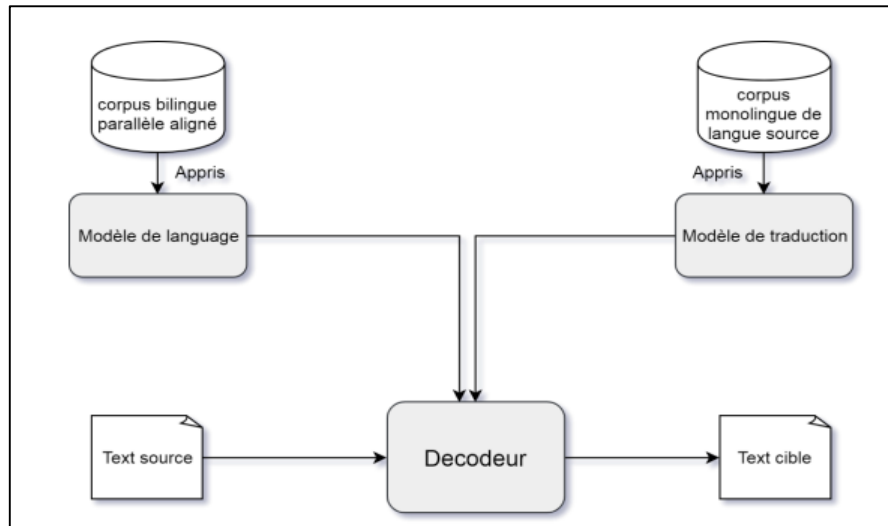


Figure 2.12: architecture de la traduction automatique statistique

2.3.2.3 Traduction automatique neuronale (NMT)

C'est une méthode moderne et nouvelle, parce qu'elle s'appuie sur des réseaux neuronaux, et en raison de son utilisation de l'intelligence artificielle, elle a contribué à sa diffusion ces dernières années par rapport au reste de la seconde approche, en raison de sa popularité croissante parmi les chercheurs et les développeurs dans l'industrie de la traduction automatisée.

Il se distingue également par une meilleure performance et une traduction humaine compétitive de haute qualité, grâce à l'apprentissage rapide des réseaux neuronaux utilisés et leur capacité à améliorer continuellement leur apprentissage.

2.3.3 Les modèle de traduction à base de règle

La traduction automatique à base de règle est divisée en 3 modèles sont :

2.3.3.1 Modèle de traduction automatique par transfert

La phrase source est analysée avec un parser et une grammaire. Cette phase d'analyse donne lieu à une représentation arborescente pour assurer le passage de là qui est ensuite converti dans la langue cible représentation intermédiaire source à la représentation cible, une table bilingue, contenant les règles de transfert entre les représentations source et cible, est requis. Comme l'exemple suivant (figure 2.4) [19] :

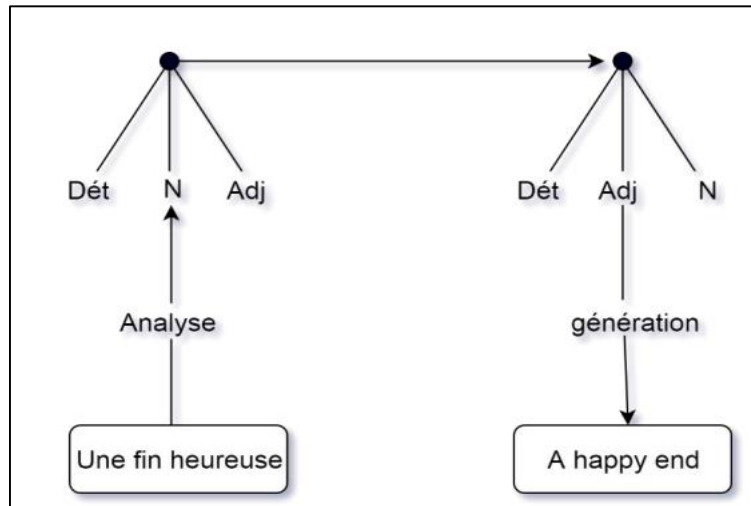


Figure 2.13: exemple de représentation arborescente de traduction d'une phrase français vers anglaise

2.3.3.2 Modèle de traduction automatique interlingue

Ce type se caractérise par son utilisation dans un environnement différent, qui est représenté dans un environnement multilingue, c'est-à-dire dans une langue, et de couvrir toutes les tendances de traduction parmi toutes les langues, d'accéder au terme de toutes les langues, le système de traduction doit construire un langage sur un langage auxiliaire sur tous les concepts communs à toutes les langues ou UNL (universel Networking Language) [19].

2.3.3.3 Modèle de traduction automatique direct

Ce type repose sur la traduction de chaque terme de la phrase seul de la langue source vers la langue cible, après découpage de la phrase en termes, suivi du processus de transfert qui consiste à placer chacun des termes dans une table bilingue, puis à relier chaque terme ou un groupe de celui-ci aux règles de traduction et enfin réorganiser l'ordre qui permet de traduire et de réorganiser les mots dans la phrase cible [19].

2.3.4 Les modèle de traduction automatique statistique

La traduction automatique statistique est divisée en 4 modèles sont :

2.3.4.1 *Modèle à base de mots (Word-Based Models)*

Le modèle basé sur les mots suivants est venu d'un candidat IBM à la fin des années 1980. Ce modèle dépend de l'utilisation de statistiques, en raison de la multiplicité des sens des mots dans les dictionnaires, par exemple, si nous recherchons un mot dans une langue spécifique dans un dictionnaire bilingue, nous remarquerons une différence dans le nombre de sens du mot.

Ce modèle nécessite la collecte de statistiques sur les traductions pour permettre à la machine de déterminer la meilleure traduction, pour y parvenir, il faut fournir des ressources sous forme de groupes de textes dans la première langue et les traduire dans la langue d'accès, puis commencer à collecter des statistiques en divisant les phrases en une série de mots, puis la recherche de chaque mot et son équivalent dans la deuxième langue commence, et c'est ce qu'on appelle le processus d'estimation de la distribution de probabilité où la fonction doit renvoyer une valeur entre 0 et 1, où 0 indique que la traduction correcte est impossible, et plus le résultat est proche de 1, il signifie que la traduction est correcte.

2.3.4.2 *Modèles bases sur les phrases (Phrase-Based Models)*

Les modèles basés sur des phrases sont plus efficaces que ceux basés sur des mots car ils utilisent des unités de traduction plus longues pour traduire simultanément de petites séquences de mots, ce qui permet de capturer davantage d'informations contextuelles dans la traduction, ce qui conduit à une meilleure sélection de phrases parmi les traductions. Une fois que le formulaire a obtenu la phrase d'entrée, il divise la phrase en mots, puis traduit chaque section et la place dans la table de traduction des phrases [19].

2.3.4.3 *Modèles bases sur les syntaxes (Syntax-Based Models)*

Ce modèle a été créé afin de gagner du temps, car l'objectif de la traduction basée sur la méthode est d'obtenir de meilleurs résultats et de réduire l'intervention humaine.

En outre, la traduction basée sur la grammaire facilite le processus de traduction, ce qui permet d'obtenir une traduction avec moins d'efforts et une meilleure précision, tout en éliminant les défauts des problèmes de traduction.

La traduction basée sur des phrases est dominante sur le marché, mais ce modèle s'est avéré être la meilleure solution pour éliminer ses lacunes.

Le processus de traduction consiste à trouver la phrase correcte ou la plus probable dans la langue d'accès. Cela se fait dans le modèle basé sur la grammaire au moyen de la distribution de probabilité et des paramètres d'apprentissage des modèles, puis la séquence cible est appelée décodage séquentiel.

2.3.4.4 Modèles bases sur des phrases Hiérarchique (Hierarchical phrase based MT)

C'est un modèle similaire à un compilateur basé sur une chaîne de caractères mais avec l'usage de SCFG qui est une grammaire sans contexte, qui permet la génération simultanée d'une paire de chaînes liées, ce qui signifie que deux structures sont définies en même temps, l'une pour le langage source et l'autre pour le langage cible.

2.3.5 Evaluation d'un système de traduction automatique

Le processus d'évaluation consiste à comparer les résultats de la traduction automatique avec la traduction humaine et à rechercher leur compatibilité L'évaluation est un processus essentiel, en particulier dans le domaine de la traduction automatique, pour déterminer l'exactitude de la traduction et améliorer les performances Il est nécessaire que ce processus soit réalisé sans contraintes de temps ou financières Pour mener à bien ce processus, plusieurs méthodes ont vu le jour, qui sont :

2.3.5.1 BLEU « Bilingual Evaluation Understudy »

Il est considéré comme l'un des premiers évaluateurs de la traduction automatique apparue en 2001 a été produit par « papineni » dépend de cette équation pour déterminer le degré de similitude entre la traduction automatique et la traduction de référence basées sur la précision de n-gramme particulière [20]:

$$\text{BLEU} = \text{BP}^* \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (2.10)$$

Avec :

- W_i : poids positifs.

- Pn : le nombre de n-grammes de traduction automatique est présent aussi dans une ou Plusieurs traductions de référence, divisée par le nombre de n-grammes totaux de traduction automatique.
- BP : la peine de brièveté est calculée sur le corpus entier et a été choisie pour être une décomposition exponentielle en "r/c", ou c la longueur de la traduction candidate et r est la longueur efficace de la traduction de référence. Comme ceci :

$$BP = \begin{cases} 1 & \text{si } c > r \\ e^{-\frac{r}{c}} & \text{si } c \leq r \end{cases} \quad (2.11)$$

2.3.5.2 WER « Word Error Rate »

Introduit en 2007 par Popovic et Ney travaille au niveau des mots au lieu des lettres Cette échelle est dérivée de la distance de Levenshtein, et son travail est divisé en 3 sections : insertion, substitution et enfin élimination. Il doit également être informé de la longueur de la phrase de référence, et la relation est [20]:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (2.12)$$

Avec :

- S : est le nombre de substitutions.
- D : est le nombre de suppressions.
- I : est le nombre d'insertions
- C : est le nombre de mots corrects.
- N : est le nombre de mots de la référence.

2.3.5.3 PER « position-independent Word Error Rate »

La PER métrique, proposé par Tillman en 1997. Comparez les mots de la traduction automatique avec les mots de référence, quel que soit leur ordre de la phrase. Le RAP est établi selon la formule ci-dessous [20] :

$$PER = \frac{1}{N_{ref}} \times d_{per}(ref, hyp) \quad (2.13)$$

Avec :

- d_{per} : calcule la différence entre les occurrences de mots dans la traduction automatique et la traduction de référence.

2.3.5.4 TER « translation Error Rate »

A été fait 2006 a été produit par « Snover » L'échelle TER et une mesure du nombre minimum de révisions nécessaire pour changer une hypothèse afin qu'elle corresponde complètement à l'un des critères.

Les modifications possibles de l'échelle sont l'insertion, la suppression et le remplacement de mots uniques. Les révisions minimales sont comptées au cours de cette relation [20]:

$$TER = \frac{Nb(op)}{AvergN_{Ref}} \quad (2.14)$$

Avec :

- Nb (op) : est le nombre minimal de révise.
- AvregN_{Ref} : la grandeur moyenne dans les références de mots.

2.3.6 Exemple des logiciels de traduction automatique

Quelques exemples des logiciels de traduction les plus populaires aujourd'hui dans ce tableau :

Tableau 2.2 : exemple de logiciels de traduction automatique

Nom	Description	Nombre de langue généré
Google traduction	Google traduction est l'un des sites de traduction les plus populaires en raison de ses nombreuses fonctionnalités Il peut conserver les traductions précédentes et lire ce qui est écrit. Il peut également traduire des fichiers ou des sites Web.	Plus de 100 langues
Bing Translator	C'est aussi l'un des programmes de traductions les plus célèbres. Il utilise également le traducteur Microsoft. Il peut	Plus de 60 langues.

	détection lui-même la langue d'entrée. Il peut également traduire la parole d'une langue vers la langue de sortie, avec la possibilité de lire le texte traduit en fonction de votre choix de la voix humaine (homme ou femme).	
Translate dict	Vous pouvez choisir d'utiliser la détection automatique de votre propre langue, soit en écrivant soit en parlant, puis rechercher la langue de sortie souhaitée et vous pouvez entendre la traduction	Plus de 50 langues
DeepL Translator	Est une petite entreprise allemande, basée à l'origine sur les données du site Web de Linguee. La qualité est vraiment proche de l'API Google, mais ils ont moins de langues disponibles	26 langues

2.3.7 Traduction automatique à base de séquence a séquence

C'est un modèle qui utilise les réseaux de neurones récurrents qui ont été proposés par Google en 2014. Le but de ce modèle est d'atteindre la traduction correcte, comme la traduction humaine, car elle est basée sur la prédiction des mots suivants.

Aujourd'hui, la modèle séquence à séquence est largement utilisé dans les applications de traduction telles que google Traduction, ainsi que dans les appareils à commande vocale sur Internet [21].

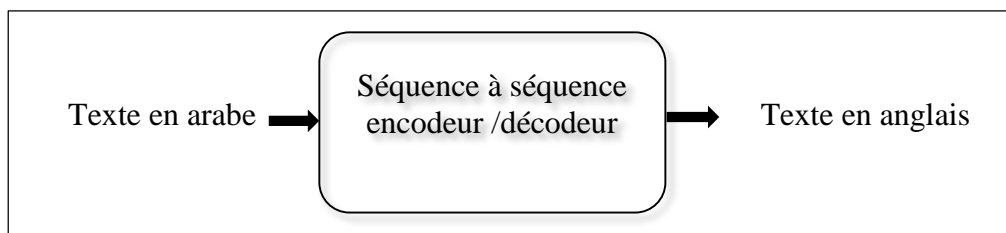


Figure 2.14: la traduction automatique à base de encodeur/décodeur

Il se compose de 3 sections : encodeur vecteur d'encodeur et décodeur (figure 2.7)

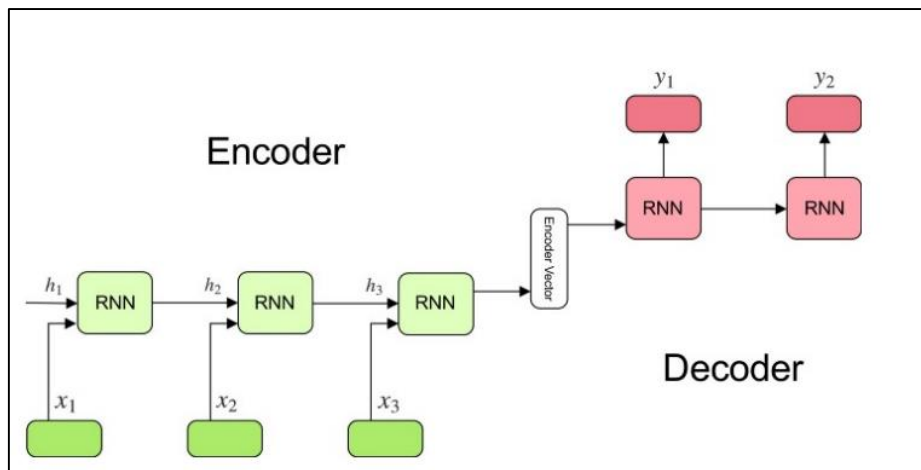


Figure 2.15: fonctionnement de l'encodeur/décodateur

2.3.7.1 Encodeur

Au début, l'entrées est donnée sous forme de vecteurs à l'encodeur qui se compose de plusieurs unités récursives qui utilisent généralement des cellules LSTM ou GRU pour de meilleures performances (chapitre 3)

Chacune cellule accepte un seul élément de la séquence d'entrée, collecte des informations pour cet élément et les propage vers l'avant. C'est-à-dire la séquence d'entrée est une collection de tous les mots d'une phrase. chaque mot est représenté par x_i où i est l'ordre de ce mot.

Les états cachés h_i sont calculés à l'aide de la formule :

$$h_t = f(w^{hh} * h_{t-1} + w^{hx} * x_t) \quad (2.15)$$

Avec :

- h_t : l'état caché ;
- h_{t-1} : l'état cache précédent ;
- x_t : l'entrée ;
- w : les poids

2.3.7.2 Vecteur d'encodeur

Ce vecteur vise à encapsuler les informations pour tous les éléments d'entrée afin d'aider le décodeur à faire des prédictions précises. Il agit comme l'état caché initial de la partie décodeur du modèle

2.3.7.3 Décodeur

Le décodeur de même structure qu'encodeur mais de manière inverse, où chaque unité répétée est supposée avoir un produit y_t dans un pas de temps t .

Chaque unité récurrente accepte un état caché de l'unité précédente et produit ainsi que son propre état caché. la séquence de sortie est une collection de tous les mots de la réponse. Chaque mot est représenté par y_t où i est l'ordre de ce mot. Tout état caché h_t est calculé à l'aide de la formule :

$$h_t = f(w^{hh} * h_{t-1}) \quad (2.16)$$

À sortie y_t au pas de temps t est calculée à l'aide de la formule :

$$y_t = \text{softmax}(w^s * h_t) \quad (2.17)$$

Nous calculons les sorties en utilisant l'état caché au pas de temps actuel avec le poids respectif $W(S)$. Soft max est utilisé pour créer un vecteur de probabilité qui nous permettra de déterminer le résultat final.

Nous utilisons le modèle encodeur/décodeur dans cette partie car il peut attribuer des séquences de longueurs différentes les unes aux autres puisque les entrées et les sorties ne sont pas corrélées et que leurs longueurs peuvent varier, il préserve le contexte de la phrase après traduction et il concurrence fortement avec traduction humaine [21].

2.4 Synthèse automatique de la parole « SAP »

Avec les progrès de la science, la tendance à fabriquer la parole est devenue une évidence en raison du temps et des efforts qu'elle représente pour l'être humain et de son grand avantage à aider les muets.

2.4.1 Définition de « SAP »

La synthèse vocale est une technique informatique de synthèse sonore qui permet de créer une parole synthétique à partir de n'importe quel texte, c'est-à-dire qu'elle convertit un texte écrit en une parole synthétique similaire à la parole humaine (TTS : Text To Speech). Ce processus est effectué à l'aide de techniques de traitement du langage. Pour arriver à ce point, il devrait y avoir une base de données des parties du discours qui y ont été enregistrées et stockées. Les volumes de parole enregistrés varient d'une base à l'autre (comme les mots ou

les sons linguistiques). Les parties de la parole sont placées séquentiellement et de manière interconnectée pour obtenir une parole facile à comprendre pour l'auditeur [22].

2.4.2 Architecture de la synthèse de parole

Comme nous l'avons mentionné précédemment, le système TTS vise à convertir le texte en parole en recevant du texte en entrée et en le sortant sous la forme d'un signal de parole. Ce processus se déroule en 3 étapes principale comme indiqué dans la figure suivante.

La première étape est liée au traitement automatique langage naturel où le texte d'entrée est épélé dans la représentation audio et la dernière étape est un traitement du signal numérique qui permet la génération réelle du signe audible [22].

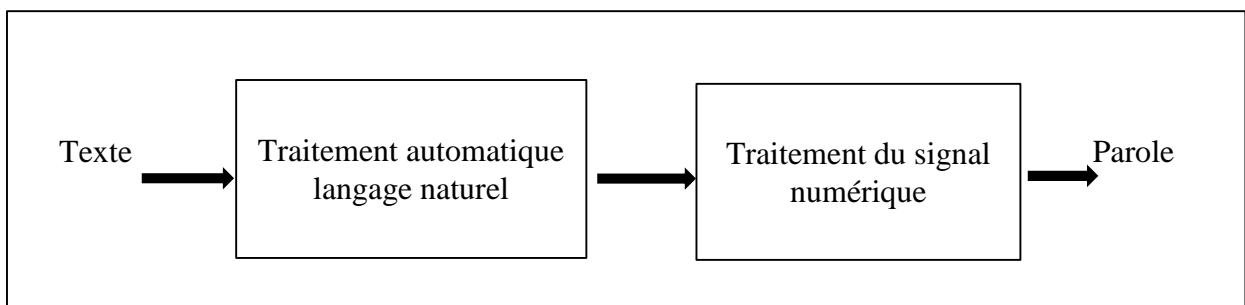


Figure 2.16: l'architecture classique de la synthèse de parole

2.4.3 Les techniques de la synthèse de la parole

La différence entre les techniques de synthèse de la parole réside dans la naturalité et l'intelligibilité de la parole où la naturalité exprime combien la voix synthétisée est proche de celle de l'homme, et pour l'intelligibilité représenter la facilité ou la clarté de compréhension de la voix. Chaque technique possède ses points forts et ses faiblesses, le choix de l'utilisation pour l'un ou l'autre dépend du type d'utilisation finale type de SAP, la puissance de calcul ainsi que les ressources nécessaires à chacun. Il existe trois techniques de synthèse de la parole [23]:

- La synthèse par concaténation
- La synthèse par règles
- La synthèse paramétrique statistique

2.4.3.1 La synthèse par concaténation

La synthèse de concaténation est une technique visant à générer un signal synthétique par concaténation d'unités acoustiques préenregistrées. Ces éléments sonores sont obtenus en segmentant les signaux vocaux. Parmi les techniques citées, celle-ci est le seul qui permet à ce jour de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain.

La synthèse concaténatoire utilise des unités de longueurs diverses. Les unités longues sont plus naturelles avec moins de points de concaténation, mais ils consomment beaucoup de souvenirs d'autre part les unités courtes sont économes en termes de mémoire, mais les procédures de collecte et d'étiquetage des échantillons sont complexes, et ils demandent beaucoup de temps. Les unités utilisées dans les systèmes d'aujourd'hui sont les mots, les phrases syllabiques, les diphones et les phonèmes.

La première idée testée était d'utiliser la segmentation en phonème mais les tests ont montré que les phrases de transition entre phonèmes entraînaient des discontinuités sur le signal reconstitué et ce l'est dû coarticulation. Ceci a conduit les chercheurs ont choisis une autre segmentation pour limiter les discontinuités aux points de concaténation est la segmentation par diphones. Le diphone est l'unité acoustique qui commence au milieu de la partie stable d'un phonème et se termine au milieu de la partie stable du phonème adjacent. Par exemple le mot [kuba] sera décomposé en 5 diphones qui sont : [#k] [ku] [ub] [ba] [a#]

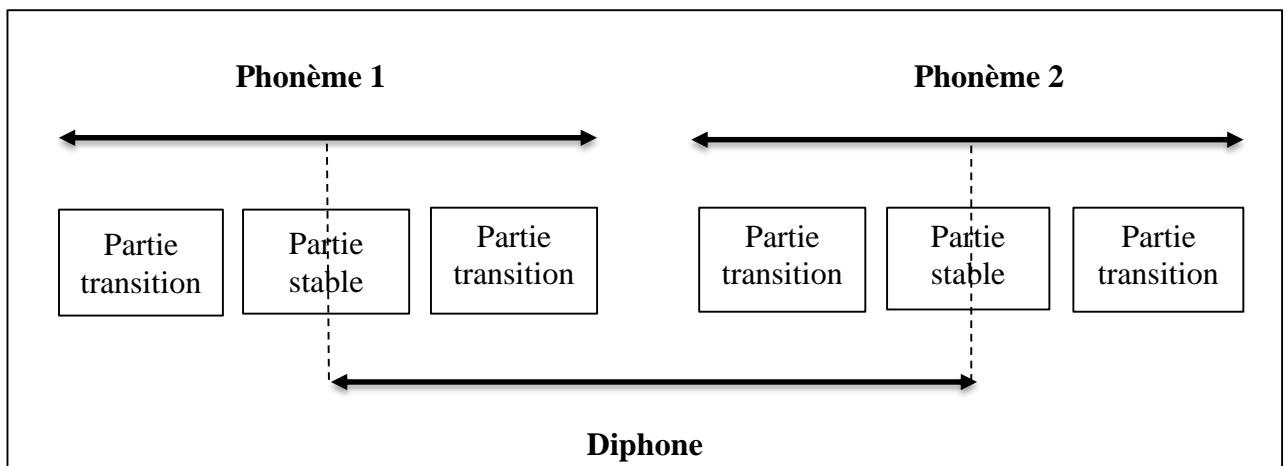


Figure 2.17: Représentation de diphone dans une séquence sonore

2.4.3.2 Synthèse par règles

Lorsque l'objectif d'obtenir une parole intelligible la synthèse des règles est un meilleur choix. Il produit une voix très facile à comprendre et fournit un nombre limité de sons, ce qui en fait une approche puissante et la plus utilisée. La synthèse de règles ou de formels est une technique qui produit de la parole en générant des signaux basés sur des règles.

2.4.3.3 Synthèse paramétrique statistique

Avec Le développement de l'intelligence artificielle, de nouvelles méthodes ont été utilisées pour la synthétisation vocale, sont des méthodes statistiques. Cette technique visant à combiner la flexibilité et la qualité de la synthèse utilisant des gros corpus de la parole. Dans la synthèse paramétrique statistique les cibles et les transitions acoustique ne sont, ici, pas fixé a priori mais apprises automatiquement à partir de grandes bases de données.

Les techniques les plus populaires de l'approche statistique, qui peuvent générer une voix compréhensible proche de celle de l'humain sont :

- Les modèles de Markov Caché
- Les Réseaux de neurones profonds
- ***Réseaux de neurones profonds***

Récemment, une nouvelle implémentation basée sur les réseaux de neurones profonds, cette technique est utilisée pour mapper les caractéristiques linguistiques aux caractéristiques acoustiques de sortie. La figure 2.18 représenté la synthèse de parole par un DNN

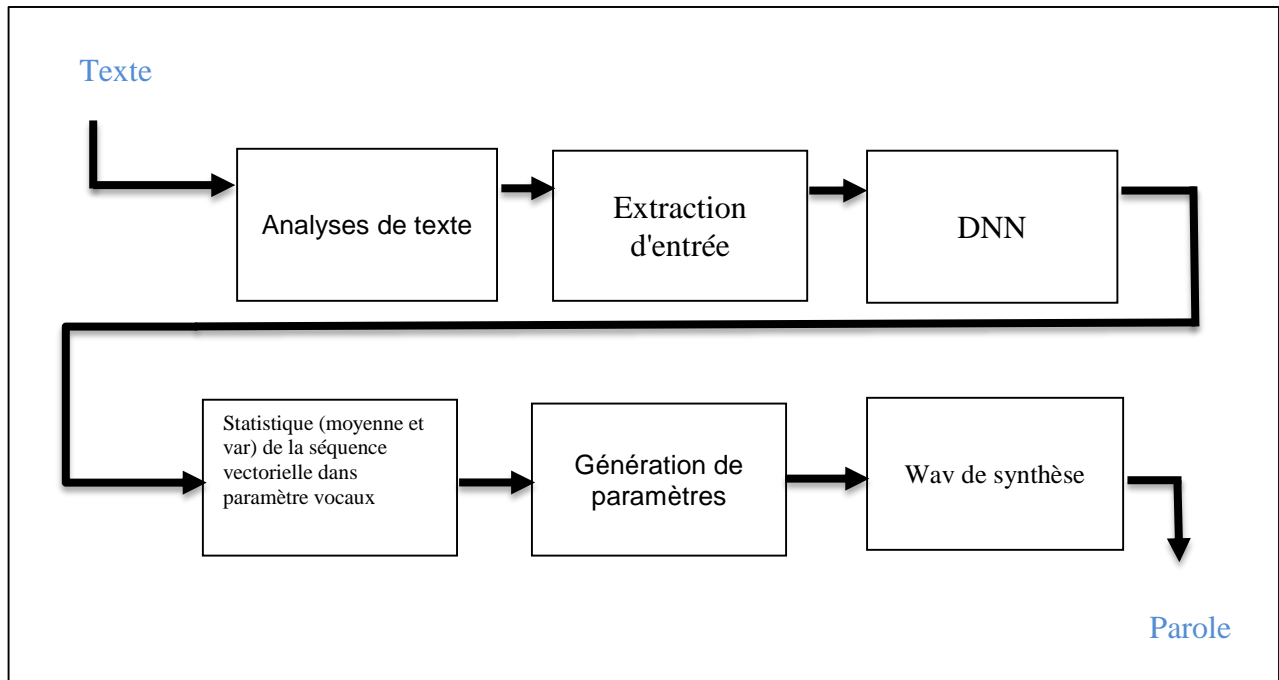


Figure 2.18: Un système de synthèse vocale basé sur un DNN

2.4.4 Avantages et inconvénients

Tel que mentionné précédemment, nous pouvons utiliser trois techniques pour synthétiser la parole. Dans le tableau suivant nous résumer les avantages et les inconvénients de chaque technique :

Tableau 2.3: les avantages et les inconvénients des technique de la synthèse de parole

Les techniques	Les avantages	Les inconvénients
La synthèse par concaténation	<ul style="list-style-type: none"> - Voix naturelle grâce a la possibilité de préserver la voix originelle. - Parole de haute qualité en termes d'intelligibilité. 	<ul style="list-style-type: none"> - Se limiter à une voix. - Nécessite une grande base de données
La synthèse par règles	<ul style="list-style-type: none"> - Parole de haute qualité en termes d'intelligibilité. - Vocabulaire illimité 	<ul style="list-style-type: none"> - Produit Une voix robotique - Le dictionnaire et les règles sont spécifiques pour chaque langue
Les DNNs	<ul style="list-style-type: none"> - Une voix intelligible qui se rapproche de celle de l'humain. 	<ul style="list-style-type: none"> - Cout de calcul élevé

2.4.5 Logiciels de synthèse de la parole

- **Balabolka** : est un logiciel qui vous permet de lire les fichiers texte. Dans le but de reproduire la voix humaine, il est possible d'utiliser n'importe quel synthétiseur installé sur l'ordinateur. La reproduction de la voix humaine peut être contrôlée à l'aide de boutons standard similaires à ceux de n'importe quel programme multimédia ("Play", "pause", "stop"). Il a la capacité de reproduire le contenu du presse-papiers, de prononcer un texte composé à partir du clavier, de vérifier l'orthographe, de diviser un fichier texte en plusieurs fichiers plus petits, de rechercher des symétries. Balabolka permet d'effacer toutes les marques de chapitre à la fin des lignes pour éviter les obstacles potentiels lors de la lecture des mots, et le texte peut être enregistré sous forme de fichier audio (compatible avec WAV, MP3, MP4). Le logiciel utilise différentes versions du progiciel Microsoft Speech API (SAPI) pour les fonctionnalités vocales. Vous permettez de modifier la vitesse et le timbre de la parole [24].
- **DSpeech** : permet d'ouvrir ou de créer un fichier texte pour jouer à voix haute, avec la possibilité de l'exporter au format MP3 ou WAV. Le programme est basé sur le moteur Windows, donc il fonctionnera uniquement avec les sons installés (SAPI 4 ou SAPI 5). Par défaut, seul le son en anglais est offert, mais vous pouvez télécharger gratuitement les voix françaises. Il peut lire l'ensemble du texte à partir de l'emplacement du curseur ou ligne par ligne. On notera également quelques options intéressantes comme la possibilité de personnaliser un son différent pour les citations, ou encore l'intégration d'un module de reconnaissance vocale qui permet, via un langage de script, de créer un dialogue avec l'utilisateur [24].
- **Notevibes** : est un logiciel de synthèse vocale qui offre une version gratuite, ainsi qu'une version payante avec de nombreuses fonctions. Les utilisateurs peuvent aussi personnaliser la prononciation. Notevibes est également doté de 177 voix uniques qui parlent 18 langues différentes [24].
- **Natural Reader** : est l'un des rares outils de synthèse de la parole à offrir des caractéristiques intéressantes bien qu'il soit complètement gratuit. Il est facile à utiliser et vous pouvez d'abord charger les documents directement dans votre bibliothèque. De plus, l'outil permet de gérer plusieurs fichiers en différents formats [24].

2.4.6 Applications de la synthèse de parole

Avec le passage du temps et les progrès de la technique, les domaines d'application de la parole artificielle se développent rapidement et la qualité des systèmes de synthèse vocale s'améliore. Parmi les machines de haute qualité les plus importantes utilisées aujourd'hui dans la parole artificielle figures :

Tableau 2.4: exemple des applications de la synthèse de parole

Application	Rôle
Interface vocale (IHM)	La synthèse vocale est utilisée pour de nombreuses interactions homme-ordinateur, telles que les systèmes d'alarme, les notifications de bureau spécifique à partir d'un ordinateur, et est utilisée pour donner des informations plus précises sur une situation.
Accessibilité	Les applications les plus utiles en synthèse de parole sont celles qui aident les aveugles à lire et à communiquer. Auparavant, ils s'appuyaient sur des livres oraux enregistrés sur une bande audio, et cela peut prendre beaucoup de temps pour en faire un en plus de l'effort humain et très cher. Aujourd'hui, les aveugles et les malvoyants disposent de systèmes de synthèse de parole simplifiés par rapport au passé et permettent également aux sourds et aux malentendants de communiquer avec des personnes ordinaires qui n'utilisent pas la langue des signes.
Applications éducatives :	Le composé peut servir à aider les enfants à épeler et à apprendre à parler. il peut aussi être utilisé pour l'apprentissage des langues et la plupart des situations d'enseignement.

2.5 Conclusion

Dans ce chapitre, nous avons détaillé le fonctionnement de la traduction automatique de la parole (STS), nous avons aussi étudié la reconnaissance automatique de la parole ainsi que la conversion parole-texte.

Dans le chapitre suivant, nous donnons étudierons l'intelligence artificielle, ses différents types et leur utilisation pour faciliter les étapes de la traduction automatique (STS).

Chapitre 3

Apprentissage profond (Deep learning)

3.1 Introduction

L'intelligence artificielle est maintenant utilisée dans une variété de domaines pour résoudre des problèmes spécifiques telles que la reconnaissance vocale, la reconnaissance du manuscrit ou la détection d'objets dans une image.

Dans ce chapitre, nous allons présenter les différents domaines de l'intelligence artificiel. En détaillant l'apprentissage en profondeur et sa relation avec les réseaux de neurones.

3.2 Intelligence Artificielle (AI)

3.2.1 Définition

L'intelligence artificielle (AI), est une discipline scientifique et technologique qu'il s'agit d'un ensemble de techniques qui permettent aux machines (ordinateurs et programmes informatiques) d'exécuter des tâches et de résoudre des problèmes généralement réservés à l'homme (selon Y. le Cun, présentation au collège de France) [25]. Les tâches liées à l'intelligence artificielle sont parfois très simples pour les humains, moins pour les machines telles que la reconnaissance et la localisation des objets dans une image, planifier les déplacements d'un robot pour attraper un objet ou conduire un véhicule. Ils nécessitent parfois une planification complexe, telle que le jeu d'échecs. Les tâches les plus compliquées requièrent beaucoup de connaissances et de sens commun, par exemple pour traduire un texte.

L'AI concerné beaucoup plus le processus et la capacité à penser et à analyser des données puissantes que tout format ou fonction particulière. Elle englobe les sous-domaines de l'apprentissage automatique et de l'apprentissage en profondeur qui sont fréquemment mentionnés en association avec l'intelligence artificielle.

Les domaines d'application et usages potentiels d'une Intelligence Artificielle sont de plus en plus divers : compréhension du langage naturel, reconnaissance visuelle, robotique, système autonome, Machine Learning.

3.3 Apprentissage automatique (ML)

3.3.1 Définition

L'apprentissage automatique est un sous-ensemble de l'intelligence artificielle (AI). Il est axé sur l'enseignement aux ordinateurs pour apprendre à partir de données et pour s'améliorer avec l'expérience au lieu d'être explicitement programmé.

En se référant au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir [26].

3.3.2 Types d'apprentissage automatique les plus populaire

L'apprentissage automatique se divise en plusieurs types se distinguant par la nature des tâches devant être apprises. Nous allons maintenant voir quatre des méthodes d'apprentissage automatique, qui sont : l'apprentissage supervisé et l'apprentissage non supervisé, apprentissage semi-supervisé, apprentissage renforcé.

3.3.2.1 Apprentissage supervisé

L'apprentissage supervisé s'appuie sur un ensemble défini de données. Les données sont étiquetées, ce qui permet au modèle de machine Learning de savoir ce qu'il doit chercher dans ces données. Le système informatique s'entraîne ainsi à classifier des données à partir de critères préalablement déterminés [27].

Parmi les algorithmes d'apprentissage supervisé, on retrouve les algorithmes de régression, les algorithmes de classification et les machines à vecteur de support.

L'apprentissage supervisé tente de répondre à deux questions :

- Classification : « quelle classe ? ».
- Régression : « combien ? ».

3.3.2.2 Apprentissage non supervisé

L'apprentissage non supervisé, à l'inverse, consiste à entraîner un modèle sur des données qui ne sont pas étiquetées. Cela signifie que le système informatique va analyser les données sans aucune indication et rechercher d'éventuels motifs récurrents. Les données sont ensuite

classées en fonction des critères que le système aura lui-même établis. Les algorithmes non supervisés sont les algorithmes de clustering, les algorithmes d'association et les algorithmes de réduction dimensionnelle [28].

3.3.3 Fonctionnement de l'apprentissage automatique

L'apprentissage automatique se compose de différents types de modèles utilisant diverses techniques algorithmiques. Selon la nature des données et le résultat souhaité, l'un des quatre modèles d'apprentissage suivants peut être utilisé : supervisé, non supervisé, semi-supervisé ou par renforcement.

Dans chacun de ces modèles, une ou plusieurs techniques algorithmiques peuvent être appliquées en fonction des ensembles de données utilisés et des résultats prévus. Les algorithmes peuvent être utilisés un par un ou combinés pour obtenir la meilleure précision possible lorsque des données complexes et imprévisibles sont en jeu.

3.4 Apprentissage profond (DL)

L'apprentissage profond ou Deep Learning est un type particulier d'apprentissage automatique dans lequel la machine peut apprendre par elle-même, contrairement à la programmation où la machine ne s'exécute que selon les règles prédéfinies.

L'apprentissage profond a révolutionné l'intelligence artificielle et s'est propagé très rapidement dans plusieurs domaines d'activité. Elle est au cœur des technologies qui permettent aujourd'hui de mener à un des tâches encore inconcevables il y a quelques années, comme la conduite automatique.

3.4.1 Définition d'apprentissage profond

L'apprentissage profond est un système évolué inspiré par le cerveau humain. C'est un élément important de la science des données, qui comprend les statistiques et modélisation prévisionnelle, il est extrêmement utile que les scientifiques recueillent, analysent et interprètent un grand nombre de données, le DL facilite et accélère ce processus [29].

3.4.2 Fonctionnement de l'apprentissage profond

Le AP s'appuie sur un réseau de neurones artificiels ou réseau de neurones d'apprentissage profond s'inspirant du cerveau humain. Ce réseau consiste en des dizaines ou même quelques (couches) de neurones, chacun recevant et interprétant l'information de la couche précédente.

Les couches d'entrée et de sortie d'un réseau neural profond sont connues sous le nom de couches *visibles*. La couche d'entrée est l'endroit où le modèle d'apprentissage en profondeur assimile les données aux fins de traitement, et la couche de sortie est l'endroit où la dernière prédiction ou la classification finale est faite. Plus on augmente le nombre de couches, plus les réseaux de neurones apprennent des choses compliquées et abstraites, correspondant de plus en plus à la façon dont une raison humaine.

Nous allons étudier un exemple de reconnaissance d'image. Le système d'apprentissage en profondeur de traitement d'image reconnaîtra l'image du chat parmi les images d'autres animaux. Il recherchera le chat dans la première couche (la couche d'entrée), puis il identifiera un animal par un animal à travers différentes couches du réseau pour déterminer les caractéristiques de l'animal : silhouette, tête, oreilles Les quatre pattes... Enfin nous aurons la photo du chat.

Ensuite, la dernière couche va collecter les différentes informations pour en déduire le résultat s'il s'agit d'une photo de chat ou non.

Enfin, le système d'apprentissage en profondeur à travers le réseau de neurones fait une comparaison entre la réponse obtenue et les bonnes réponses données par les humains. Si les réponses sont les mêmes, le réseau se souvient de ce succès et l'utilise plus tard dans d'autres cas, en cas de mauvaises réponses, il est rejeté et envoyé aux niveaux initiaux pour modifier le modèle mathématique. Au fil du temps, le programme organise les données en blocs plus complexes.

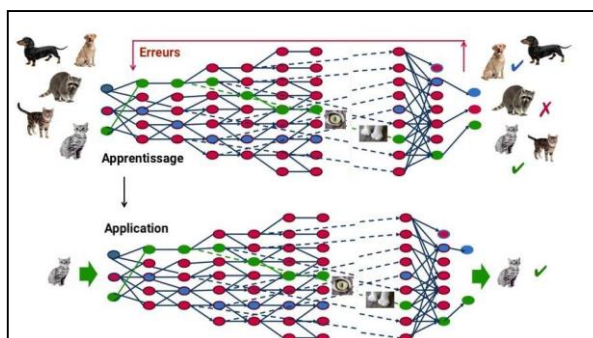


Figure 3.1: exemple de fonctionnement de DL

Au AP, les données de départ sont indispensables, plus le système accumulé de différentes expériences, plus il sera efficace.

Parmi les exemples d'applications d'apprentissage profond, citons la reconnaissance vocale, la classification d'images et la traduction automatique.

3.5 Pourquoi l'apprentissage profond ?

L'apprentissage en profondeur est principalement utilisé pour classer les données, ce qui en fait une généalogie de la tâche, et à son tour, il facilite et simplifie la tâche pour un être humain, et les résultats de son travail sont d'une grande précision.

Dans ce sujet, nous utilisons l'apprentissage en profondeur à cause des nombreuses couches et parce que le sujet est complexe, par exemple en traduction, nous devons spécifier le type de mot si c'est (adjectif, nom, verbe si c'est un verbe, nous allons faut préciser s'il s'agit d'une inflexion ou pas et si oui à quel moment et ainsi de suite...) Et c'est plus compliqué si on travaille sur une phrase ou un paragraphe, Donc il va falloir beaucoup de couches et beaucoup de temps pour obtenir le bon résultat.

3.6 Réseaux de neurones

Comme nous l'avons mentionné précédemment, l'apprentissage automatique et l'apprentissage en profondeur dépendent du système de neurones artificiels, qui à son tour s'inspire du neurone humain ou neurone biologique.

Les réseaux de neurones sont théoriquement capables d'apprendre n'importe quelle fonction mathématique avec suffisamment des données d'entraînement et ils sont devenus

des techniques d'apprentissage automatique populaires qui simulent le mécanisme d'apprentissage dans les organismes biologiques. On trouve dans le système nerveux humain des cellules appelées neurones.

3.6.1 Neurone biologique

On peut diviser le neurone biologique en trois sections principales (figure 3.2), qui sont : Récepteurs (noyau, dendrites, corps cellulaire) : cette section reçoit les signaux suivants d'une deuxième cellule et traite et collecte les signaux, s'il dépasse le total spécifié, il est envoyé à une deuxième cellule en envoyant un courant électrique à l'axone. Propagation de l'information(axone) : distribue les données reçues d'une cellule à l'autre car la cellule est interconnectée par une dendrite et un axone. Transmetteur(synapses) : permet de faire passer les signaux de l'axone vers une seconde cellule. Les cellules sont connectées par les axones et les dendrites[30].

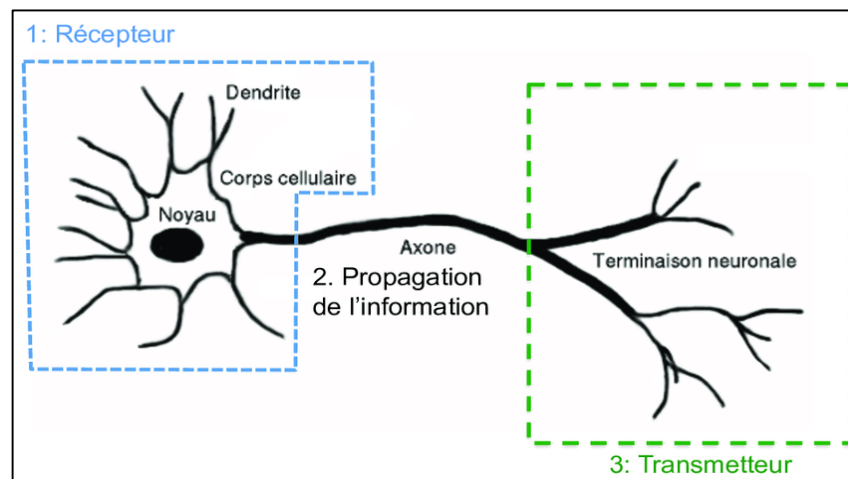


Figure 3.2: Un neurone biologique et ses principaux composants [31]

3.6.2 Neurone formel (artificielle)

Comme mentionné ci-dessus, le neurone artificiel est inspiré du neurone biologique [29], auquel le neurone artificiel est divisé sur les parties suivantes (Figure 3.3) :

3.6.2.1 Les signaux d'entrée

Entrées : elles sont représentées par un groupe d'entrées, qui sont à l'origine des sorties d'autres neurones. Elles reçoivent des signaux représentés par les valeurs E_i ($E_1, E_2, E_3 \dots E_n$) et toutes les entrées sont liées au poids w_i ($w_1, w_2, w_3 \dots w_n$),

Le signal d'entrée est le produit de chaque valeur d'entrée selon son poids, à savoir $E_i * w_i$.

3.6.2.2 Élément de traitement :

Cette partie est responsable de deux choses :

- a) La somme des signaux d'entrée pour prendre cette forme :

$$\text{Total d'entrée} = a = \sum_{i=1}^n E_i * w_i \quad (3.1)$$

- b) Fonction d'activation : Cette fonction limite la sortie du neurone, en mettant la sortie à l'état actif et 0 ou (-1) (selon la fonction) à l'état inactif.

Une des conditions de la fonction d'activation est qu'elle soit dérivable, facile à calculer, et qu'elle soit simplifiée non décroissante.

- **Signal de sortie :**

Le résultat final du neurone est 1 ou 0, et l'équation sous la forme

$$S = f\left(\sum_{i=1}^n w_i * E_i\right) \quad (3.2)$$

Les composants du rayon d'entrée entrent dans le réseau par la matrice de poids suivant :

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,R} \\ w_{2,1} & w_{2,2} & \dots & w_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{S,1} & w_{S,2} & \dots & w_{S,R} \end{bmatrix} \quad (3.3)$$

Les indicateurs de ligne des éléments de cette matrice indiquent le neurone cible, tandis que les indicateurs de colonne se réfèrent aux composants d'entrée source. C'est-à-dire que les indicateurs de l'élément $w_{1,2}$ indiquent que ce poids est relié au premier neurone, et que la composante d'entrée de ce neurone est le deuxième composant.

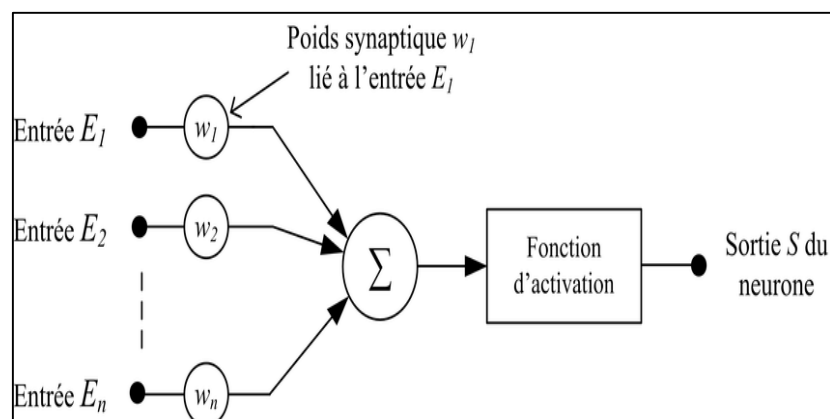


Figure 3.3: Neurone formel (artificielle)

Après avoir effectué un certain nombre d'expériences, l'homme a réalisé que la cellule nerveuse est incapable de fonctionner seule et qu'elle a besoin d'un grand nombre. D'où le principe des réseaux de neurones.

3.6.3 Réseaux de neurones artificiels « ANN »

3.6.3.1 Définitions

Ou artificiel neural networks ANN (en anglais), c'est un groupe interconnecté de neurones virtuels ou de nœuds connectés pour en former un réseau car ils sont connectés par des poids, ce réseau de neurones est créé par des programmes informatiques pour imiter le travail d'un neurone biologique et s'appuie sur un modèle mathématique pour traiter les informations.

Le réseau de neurones artificiels a été inventé afin de modéliser le mécanisme d'apprentissage et de traitement de l'information, comme ce qui se passe dans l'esprit humain, par l'expérience et se rappeler les résultats précédents, bons et mauvais.

Le réseau de neurone artificiel est divisé en trois couches : la couche d'entrée, la couche cachée et la couche de sortie.

3.6.3.2 Fonctionnement d'un réseau de neurone artificiel

Le réseau neuronal artificiel est formé d'un ensemble de neurones réunis en couches, qui fonctionnent en parallèle. Chaque couche est traitée de manière indépendante des autres et transmet le résultat de son analyse à la couche suivante jusqu'à la couche de sortie, en passant soit par une ou plusieurs couches intermédiaires dites cachés (figure 3.4).

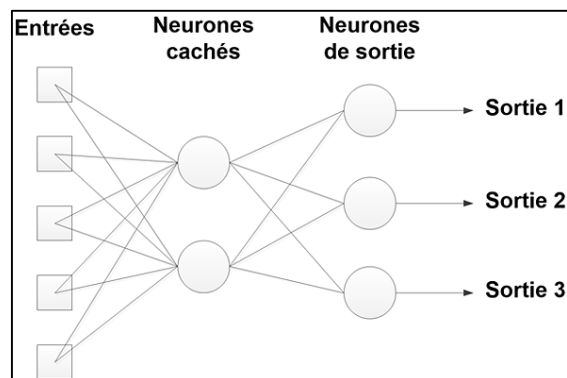


Figure 3.4: exemple d'un réseau de neurone

Chaque nœud de ces couches est connecté à tous les autres nœuds de la couche suivante qui le précède, car il prend les données de la couche précédente et les traite et donne une valeur de sortie unique qui est transmise à tous les neurones de la couche suivante. Parfois, les neurones sont connectés à une entrée fixe qui entre dans chaque processus de traitement et n'a rien à voir avec l'entrée du réseau appelée biais.

3.7 L'architecture des réseaux de neurones

En fonction de la topologie de raccordement des neurones, ils peuvent être classés en deux catégories principales : les réseaux non bouclés (statique ou feed forward et multicouche) et réseaux bouclés (dynamique, feed back ou récurrent).

3.7.1 Réseaux non bouclés (statique)

Un réseau est dit non bouclés ou statique, c'est-à-dire que ces réseaux neuronaux font circuler l'information dans une seule direction, de l'entrée à la sortie (figure 3.4).

3.7.1.1 Réseaux statiques feed forward

Les neurones sont arrangés par les couches. Il n'y a pas de connexion entre neurones d'une même couche et les connexions ne se font qu'avec les neurones des autres couches. Chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. Il se compose d'une couche d'entrée, d'une couche de sortie et d'une couche cachée (figure 3.4).

Ce type de réseaux est utilisé pour effectuer des tâches d'approximation de fonction non linéaire, de la classification ou de la modélisation de processus statiques non linéaires.

3.7.1.2 Réseaux statiques multicouche

C'est le réseau de neurones statique le plus utilisé, il se travaille comme le feed forward sauf que le multicouche se compose de plusieurs couches cachées (figure 3.5). C'est aussi le plus largement utilisé dans l'apprentissage profond. Il peut résoudre des problèmes non linéairement séparables et il suit un apprentissage supervisé avec la règle de correction de l'erreur[30].

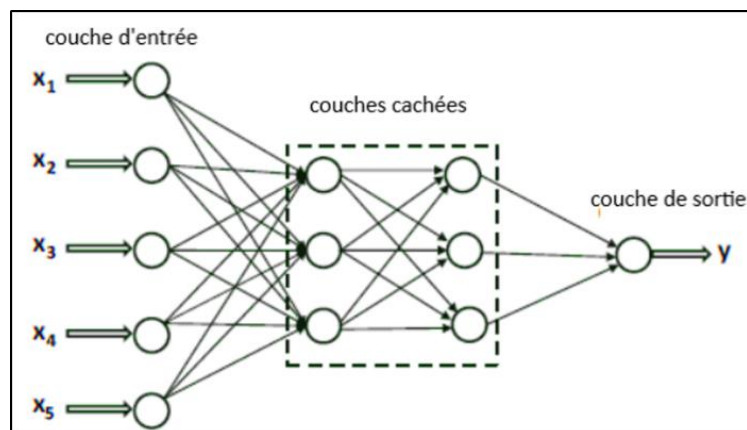


Figure 3.5: réseaux de neurone multicouche

3.7.2 Réseaux bouclés

Un réseau est bouclé si son graphe possède au moins un cycle, Ce qui conduit à renvoyer plusieurs valeurs à l'entrée, ce qui signifie qu'il s'agit d'un réseau de neurones dynamique comme la plupart des applications qui sont mises en œuvre par des programmes informatiques [3].

3.7.1.3 Réseaux dynamiques « feed back »

Ces réseaux sont assez intéressants car leur fonctionnement est séquentiel et adopte un comportement dynamique. Ils sont utilisés essentiellement pour modéliser les systèmes récurrents, comme ils permettent des applications des problèmes réels (par exemples industriels). Tout réseau de neurones bouclé est un système dynamique non linéaire que l'on peut mettre sous forme d'une représentation d'état appelée la forme canonique.

3.8 Les avantages et les inconvénients des réseaux de neurone

Nous avons vu ce que l'intelligence artificielle nous offre aujourd'hui en matière de facilitation de la vie humaine, mais il doit y avoir des points négatifs sous d'autres aspects, ainsi que le réseau de neurones artificiels.

3.8.1 Avantages

Parmi les avantages d'un réseau de neurones, citons [30] :

- Peut représenter une fonction quelconque, linéaire ou non, simple ou complexe.

- Capacité de tirer des leçons d'exemples représentatifs, par "rétropropagation des erreurs". L'apprentissage (ou la construction de modèles) est automatique.
- Résistance au bruit ou à l'absence de données fiables.
- Simple à manipuler, bien moins de travail personnel à fournir qu'en analyse statistique traditionnelle.
- Moindre comportement avec une faible quantité de données.

3.8.2 Inconvénients

Les inconvénients d'un réseau de neurones sont :

- L'absence d'une méthode systématique de définition de la meilleure topologie de réseau et du nombre de neurones à placer dans la ou les couches cachées.
- Le choix des valeurs initiales des pondérations du réseau et l'ajustement de l'étape d'apprentissage jouent un rôle important dans la rapidité de convergence.
- Le problème du sur apprentissage (apprentissage au détriment de la généralisation).
- La connaissance acquise par un réseau de neurone est codée par les valeurs des poids sont inintelligibles pour l'utilisateur.

3.9 Réseaux de neurones profonds « DNN »

Un réseau neuronal profond s'appelle également DNN "Deep Neuronal Network" est un type de réseau neuronal artificiel qui se compose de multiples couches entre ses couches d'entrée et de sortie. La structure du DNN est donnée dans la figure 11 :

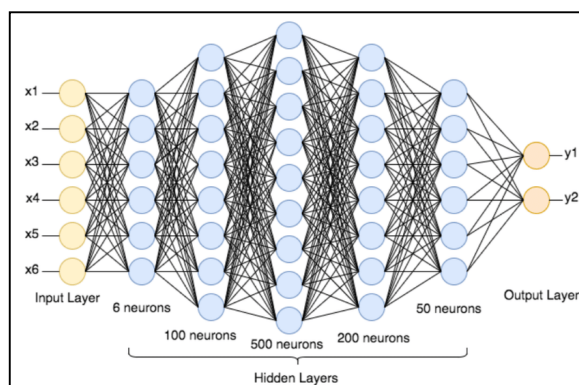


Figure 3.6: Réseaux de neurones profonds « DNN »

Les DNNs sont des versions améliorées de l'ANN conventionnelle avec un certain niveau de complexité (de nombreux neurones, couches cachées et connexions) qui utilisent l'apprentissage profond car il utilise plusieurs couches constituant l'architecture du modèle pour apprendre des données. Chaque couche cachée contient des neurones. Les neurones sont connectés les uns aux autres. Le neurone reçoit les informations et les transmet à la couche supérieure. La robustesse de l'information transmise au neurone de la couche suivante dépend de la fonction d'activation, du poids et du biais.

L'un des principaux cas d'usage de ces réseaux de neurones avancés est le traitement des données non structurées. Les réseaux de neurones profonds peuvent regrouper et classifier les données stockées sur une base de données. Ceci s'avère très utile pour organiser les données sans étiquettes ni structure.

Les réseaux neuronaux profonds se sont avérés efficaces dans de nombreuses tâches, telles que la détection d'objet, la classification d'image, la reconnaissance vocale

3.10 Types des réseaux de neurones profonds « DNN »

Les réseaux de neurones profonds ont été largement utilisés au cours des vingt dernières années dans divers domaines. Il y a plusieurs types de DNN et chaque type dispose de ses propres caractéristiques et applications. Les trois algorithmes les plus populaires sont : Réseaux neuronaux récurrents, Réseaux neuronaux convolutifs, Réseaux Adversaires Génératifs.

3.10.1 Réseau de neurone convolutive « CNN »

Un réseau de neurones convolution, appelé aussi souvent pour « convolutional network », ou encore CNN pour « Convolutional Neural Network » est un type de réseau neuronal artificiel spécialisée dans la reconnaissance et traitement d'images, en raison de sa capacité à reconnaître des motifs dans des images. C'est très utilisé dans la vision par ordinateur.

Les réseaux convolutifs sont extrêmement efficaces dans des domaines ou de grandes données comme la reconnaissance d'images et la classification. Ils sont capables de catégoriser les informations de plus simples aux plus complexes.

Les CNNs sont une forme particulière de réseau neural multicouches dont l'architecture de connexion est inspirée par l'organisation du cortex visuel animal.

CNN est un réseau neuronal performant doté d'une architecture multicouche inspirée de l'organisation et le fonctionnement du cortex visuel destiné à imiter le modèle de connectivité des neurones dans le cerveau humain [32].

3.10.1.1 Architecture

L'architecture de CNN est fondée sur plusieurs réseaux profonds formés de deux parties principales :

- **Partie convolution** : Dédiée à l'extraction automatique des caractéristiques.
- **Partie de classification** : classifier l'image.

Le bloc de construction du réseau de neurones convolutifs est formé par un empilement de couches de traitement indépendantes.

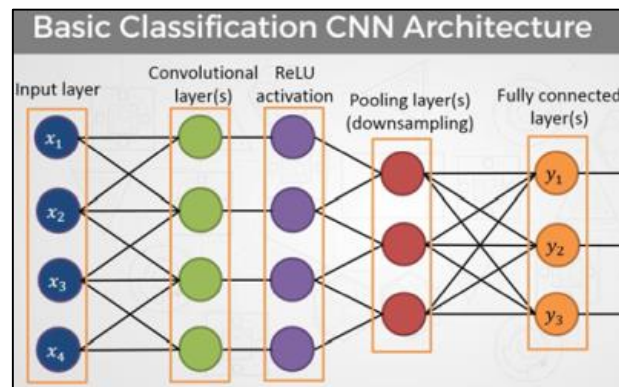


Figure 3.7: classification de base architecture CNN

3.10.1.2 Différentes couches de neurone convolutifs

Dans la partie de convolution on distingue 03 types de couches :

a) Couche de convolution

La couche de convolution est la couche la plus importante et elle est également l'élément de base d'un CNN, et c'est laque à lieu la plupart des calculs. L'objectif principal est de retirer les caractéristiques de l'image d'entrée.

Premièrement, une partie de l'image est connectée à la couche de convolution afin de réaliser une opération de convolution et calculer le produit scalaire entre le champ récepteur (il s'agit d'une zone locale de l'image d'entrée de la même taille que le filtre) et le noyau (ou filtre) comme le montre la figure.

Le résultat de l'opération est un nombre simple du volume de sortie. Ensuite, on fait glisser le filtre vers le champ de réception suivant de la même image d'entrée qui se déplace progressivement de gauche à droite d'un certain nombre de cases défini au préalable « le pas » et refaisons la même opération. Cette opération est répétée par le même processus à plusieurs reprises jusqu'à ce que l'image entière soit scannée pour produire la sortie finale de la convolution « feature map ».

b) Couche de correction RELU

Il est possible d'augmenter l'efficacité du traitement en intercalant entre les couches de traitement une couche qui actionnera la fonction d'activation sur les signaux de sortie. La fonction RELU $f(x)=\max(0, x)$, est une fonction qui brise une partie de la linéarité en supprimant une partie des valeurs et permet également d'accélérer les calculs.

RELU est très utilisé dans les réseaux de neurones à convolution, après chaque opération de convolution, le CNN applique une transformation RELU qui remplace chaque valeur négative de la caractéristique convolutive par zéro, ceci est parfois appelé activation, puisque seules les fonctionnalités activées sont rapportées dans la couche suivante.

c) Couche mise en commun (pooling layers)

Pour diminuer la complexité du calcul et obtenir une représentation d'image hiérarchique, une couche de regroupement est souvent insérée périodiquement entre deux couches de convolution successives.

Le pooling, aussi appelé regroupement de couches est le procédé de sous-échantillonnage de la taille des cartes caractéristiques (la matrice résultant de la convolution) un filtre est transmis aux résultats de la couche précédente et sélectionne un nombre à l'intérieur de chaque groupe de valeurs. « Max-pooling, mean-pooling ou le sum-pooling »,

Alors qu'une grande partie de l'information est perdue dans la couche de mise en commun, il a aussi un certain nombre d'avantages pour CNN. Elle réduit la complexité lorsqu'il s'agit de minimiser le nombre de calculs et de paramètres dans le réseau, réduire le risque de surpêche et, par conséquent, améliorer l'efficacité du système [33].

Il existe plusieurs types de pooling [34] :

- Le max pooling : qui équivaut à prendre la valeur maximum de la sélection, c'est le type le plus utilisé parce qu'il est rapide à calculer « immédiat ».

- Le mean-pooling : ou average-pooling la moyenne des pixels de la fenêtre sélectionnée, calculer la somme de toutes les valeurs, puis diviser par le nombre de valeurs. Le résultat est une valeur intermédiaire correspondant à ce lot de pixels.
- Le sum-pooling : représente la moyenne sans avoir divisé par le nombre de valeurs (on ne calcule que leur somme).

Après avoir appris la fonctionnalité dans de nombreuses couches, l'architecture d'un CNN passe au classement où à la fin des couches de convolution et de regroupement, En général, les réseaux utilisent des couches complètement connectées. Cette couche est connectée entre tous ses neurones et chaque neurone dans la couche précédente.

d) Couche entièrement connectée (Fully-Connected)

La couche fully-connected est toujours la dernière couche d'un réseau neural convolutif et est la dernière couche où la classification est réalisée. Ici, nos images filtrées et réduites passent par une étape dite "de Flattening" (ou aplatissement). Cette opération consiste à mettre l'ensemble des données dans un vecteur unique de dimensions K , où K représente le nombre de classes que le réseau pourra prévoir. Chaque élément du vecteur indique la probabilité que l'image d'entrée relève d'une classe.

3.10.1.3 Application

Les CNNs ont excellé dans le domaine d'extraction de caractéristiques et la reconnaissance des formes où il peut identifier et extraire des caractéristiques même en images de mauvaise qualité et peut également être très efficace pour classer des données autres que des images, telles que des contenus audios, des séries temporelle (vidéo, vocale ...). C'est pour cela les réseaux neuronaux convolutifs ont de larges applications, notamment :

- La reconnaissance d'image et vidéo
- Détection faciale
- Le traitement de langage naturel
- La détection d'objet

3.10.2 Réseaux de neurone Récurrents « RNN »

Le réseau neuronal récurrent ou RNN (Recurrent Neural Networks) est une sorte de réseau neuronal artificiel spécialisée dans le traitement des séries temporelles ou séquence ordinales.

Les RNN sont des réseaux avec des boucles qui permettent à l'information de rester et de la traiter plusieurs fois en la renvoyant à chaque fois dans le réseau.

Les réseaux bouclés sont plus près du fonctionnement réel du système nerveux. Ils peuvent gérer les données séquentielles et les entrées de taille variable et peuvent stocker ou mémoriser des renseignements.

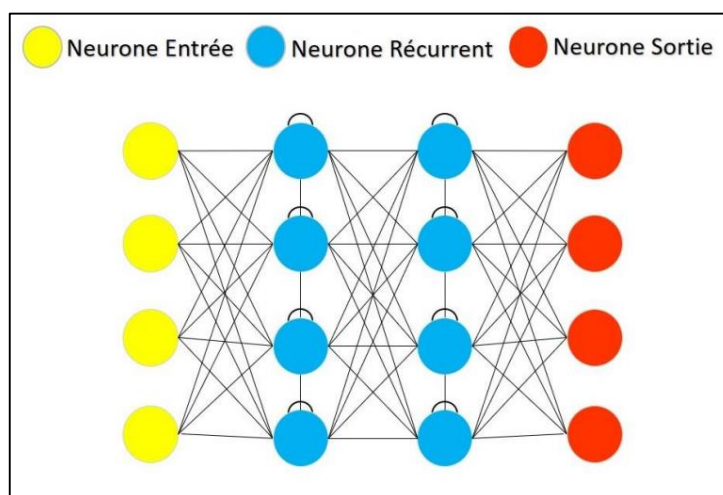


Figure 3.8: les couche de « RNN »

3.10.2.1 Architecture d'un réseau de neurone récurrent

Les réseaux de neurones récurrents utilisent une architecture similaire aux réseaux de neurones artificiels. La différence est que les RNNs sont composés de couches qui permettent de traiter l'information dans les deux sens : la propagation en avant et rétro propagation. D'autre part, les ANNs traitent l'information dans un seul sens, de l'entrée à la sortie.

Dans les RNNs, il est possible de transmettre l'information à travers des boucles de rétroaction, et ainsi de la renvoyer à une couche précédente. Ces retours permettent au système à être constitué une mémoire.

Dans RNNs les sorties de certaines couches sont réinjectées dans les entrées d'une couche précédente à l'aide de liens dans le sens inverse. Cette réinjection permet l'analyse séquentielle des données et constitue une mémoire dans le système.

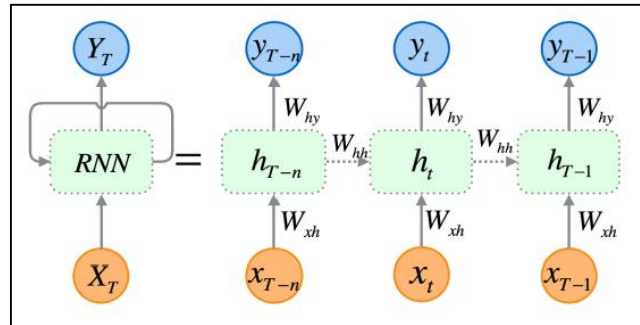


Figure 3.9: Réseaux de neurone Récurrents « RNN »

Avec :

- X_t : entrée.
- W : poids.
- Y_t : sortie.
- h_t : couche cachée.

Un réseau de neurones récurrents peut être considéré des copies multiples du même réseau, chacun transmission d'un message à un successeur comme indiqué dans la figure 3.9.

Le diagramme ci-dessus montre un RNN déroulé. En déroulant, on entend simplement montrer le réseau pour l'ensemble de la séquence. Par exemple, si la séquence que nous recherchons est une phrase de cinq mots, le réseau serait déployé dans un réseau de neurones à cinq couches, un niveau par mot. Les formules qui régissent les calculs dans une RNN sont les suivantes :

- X_t est l'entrée au moment t .
- W et R sont les poids de RNN :
- Des poids (notés W sur le schéma ci-dessous) reliant les entrées à la sortie (comme pour un réseau de neurones classique).
- Des poids (notés R) entre la sortie et l'entrée de la couche, qui sont les connexions récurrentes.

- St est l'état caché au moment t . C'est la « mémoire » du réseau. St est calculé en fonction de l'état Caché précédent et de l'entrée à l'étape actuelle :

$$St=f(UXt+WSt-1) \quad (3.4)$$

Le fonctionnement des réseaux de neurones récurrents est semblable à celui des réseaux de neurones artificiels, sauf que le neurone intermédiaire. Il reçoit les entrées et produit une sortie. Dans le détail, à chaque étape temporelle t , ce neurone récurrent reçoit l'entrée $x(t)$ ainsi que sa propre sortie produite à l'étape temporelle précédent $y(t-1)$.

a) Cellule RNN

Nous avons mentionné précédemment comment une cellule RNN a un état caché dont la tâche est de se souvenir ou de déterminer ce que la cellule doit retenir et ce qu'elle ne doit pas oublier. Il s'agit d'un graphique de ce que contient la cellule RNN (figure 3.9), car elle a une entrée et une sortie, en plus de l'état caché.

- x_t : est la donnée saisie dans la cellule courante.
- h_{t-1} : l'état caché, qui est la sortie de la cellule précédente et est utilisé dans la prédiction pour trouver le mot suivant.
- h_t : est la sortie de la cellule courante.

L'opération se fait dans la cellule RNN en parallèle avec le reste de la cellule et traite les mots comme des vecteurs d'abord, est entré h_{t-1} , puis X_t , puis ils rencontrent, se combinent en un vecteur bout à bout et s'appliquent à la fonction tanh (th) et enfin on obtient Th et le résultat est sorti et l'envoie à la deuxième couche des couches cachées ou la couche de sortie d'une part et il est conservé pour être envoyé à la cellule suivante.

Th compte les nombres de vecteurs entre -1 et 1 pour faciliter les calculs sur la cellule et gagner du temps et surtout pour déterminer ce qu'elle doit garder et ce qu'elle doit oublier. Pour apprendre, RNN utilise la méthode de descente de gradient pour mettre à jour les poids entre les neurones.

$$W = W - \alpha F_w \quad (3.5)$$

Avec :

- W : Poids.
- α : la vitesse d'apprentissage du réseau.
- F_w : le gradient du réseau par rapport à W .

Lorsque on avance à gauche la valeur de αF_w devient très petite (presque 0), c'est-à-dire que W reste constante, Cela signifie que les informations précédentes ne sont pas conservées ou perdues, ce qui nous fait prendre conscience de la faiblesse de la mémoire RNN.

Et pour cela, a conduit à l'émergence de plusieurs formes, dont LSTM, GRU, Encodeur \ décodeur, sur lesquelles la traduction automatique s'appuie aujourd'hui fortement. Les réseaux neuronaux traditionnels analysent les données étiquetées mais ne sont pas conçus pour faire des prédictions de séries chronologiques (données évoluant avec le temps). Pour réaliser ce type de calcul, il existe trois principaux types de réseaux neuronaux récurrents : le RNN simple, le LSTM et enfin le GRU [35].

3.10.2.2 Types des réseaux de neurones récurrent

Il existe 7 types de réseaux RNN : LSTM layer, GRU layer, Simple RNN layer, Time Distributed layer, Bidirectionnel layer, ConvLSTM2D layer (pour le traitement de vidéo) et le Base RNN layer.

a) LSTM et son Architecture (Long Short-Term Memory)

LSTM (long short term memory) LSTM est un cas particulier de RNN inventé par « Hochreiter » et « Schmidhuber » en (1997) conçu parce que RNN ne se souvient pas bien, LSTM contient des connexions de rétroaction (feedbacks) afin que son système ait la capacité d'apprendre en fonction des problèmes de prédiction de séquence qui est nécessaire beaucoup de traduction automatique et de reconnaissance vocale...

LSTM utilise certains types de processus de mémoire artificielle pour imiter la pensée humaine, l'unité LSTM est composée d'une cellule, d'une porte d'entrée, d'une porte de sortie et d'une porte d'oubli, ainsi que deux états : l'état de la cellule et l'état caché.

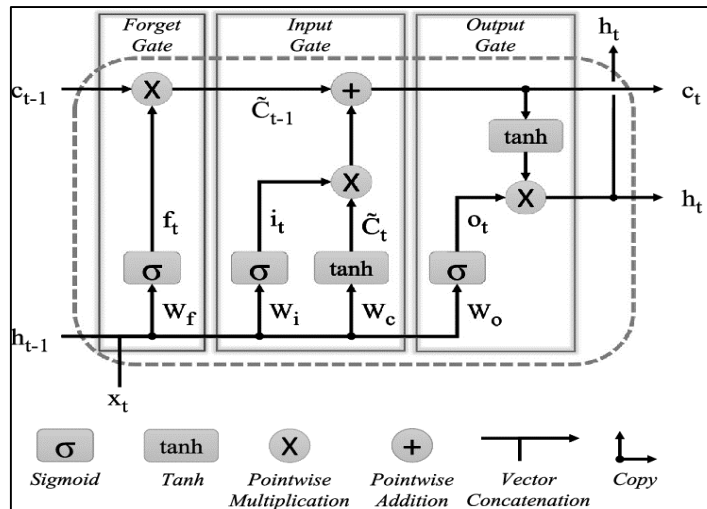


Figure 3.10: Représentation simplifiée d'une cellule LSTM.

Avec :

- C_{t-1} : État de la cellule précédent.
- C_t : Nouvel état de la cellule.
- h_{t-1} : État cachée de la cellule précédent.
- h_t : État cachée de la cellule.
- X_t : L'entrée de la cellule.

Concernant l'architecture de la cellule LSTM, chacune des portes du LSTM a ses propres poids, ces poids dépendent des dimensions de la cellule (X_t, h_{t-1}), on a :

- W_f : poids de l'entrée de la porte d'oubli.
- W_i : poids de l'entrée de la porte d'entrée.
- W_c : poids les données de la porte d'entrée pèse les données à combiner à la porte d'entrée pour mettre à jour l'état de la cellule.
- W_o : poids de l'entrée de la porte de sortie.

Le processus d'identification des informations importantes des autres dépend des poids ou de la matrice de poids, sans lesquels la cellule ne peut pas se souvenir ou apprendre [35].

3.10.3 Réseaux de neurone antagonistes génératifs « GAN »

Réseaux antagonistes génératifs ou (Generative Adversarial Network) GAN c'est une partie d'DL qui a été introduit en 2014 par I. GOODFELLOW, GAN produit des objets

réalistes comme des photos Le GAN est composé de deux composants importants : Le générateur G et le discriminateur D.

Ils fonctionnent en sens inverse, où le générateur prend un signal bruyant comme entrée et crée des images à la demande Alors que le discriminateur reçoit des données de deux sources, des données générées par le générateur et l'autre des données réelles de la base de données d'apprentissage, sa tâche consiste à vérifier les données et à déterminer lesquelles proviennent du générateur et lesquelles sont réelles.

Le discriminateur est formé pour maximiser la probabilité d'un marquage correct sur les données réelles et la génération d'échantillons, tandis que le générateur de formation est utilisé pour réduire l'enregistrement, en d'autres termes, réduire la probabilité que le discriminateur obtienne la bonne réponse. Ce processus peut être expliqué dans cette équation :

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (3.12)$$

En d'autres termes, le générateur cherche à créer des images difficiles pour le discriminateur Donc il ne pouvait pas la reconnaître, et le discriminateur devient plus intelligent pour éviter de tromper le générateur.

L'objectif initial de la conception GAN était d'éviter l'utilisation de chaînes de Markov car elles sont coûteuses en calcul.

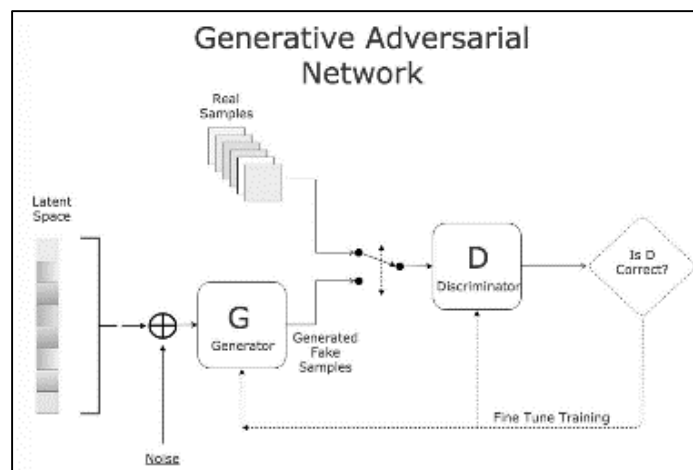


Figure 3.11: Réseaux de neurone antagonistes génératifs « GAN »

3.11 Conclusion

Dans ce chapitre, nous avons présenté un aperçu général sur l'intelligence artificielle (AI) et de l'apprentissage automatique (ML). Nous avons aussi présenté l'apprentissage profond (DL), les réseaux de neurones et les réseaux de neurones profonds ainsi que leurs différents types : DNN, CNN, RNN, GAN. Dans le chapitre suivant, nous allons montrer comment appliquer l'apprentissage profond et discuter les résultats obtenus.

Chapitre 4

Implémentation et Résultats

4.1 Introduction

Dans ce chapitre, nous allons montrer les différentes étapes pour l'implémentation de notre système de traduction automatique arabe-Anglais (Speech To Speech). Les simulations sont effectuées à l'aide du logiciel de programmation python. Les différents modules sont implémentés sous l'environnement google colab.

4.2 Contexte expérimentale

Pour évaluer et tester les performances du système proposé, des outils et des matériels sont nécessaires. Nous citons ici le matériel et le logiciel de programmation.

4.2.1 Matériel utilisé

Le matériel utilisé est résumé comme suit :

- Ordinateur (Laptop I5) avec Processeur : Intel(R) Core (TM) i5-7200U CPU, 1.70 GHz
- Logiciel de programmation Python
- Corpus de parole Arabe (pour le module de reconnaissance)
- Corpus de unités acoustiques (pour le module de synthèse)

4.2.2 Langage de programmation

De nos jours. Il existe plusieurs langages de programmation et chaque langage possède ses propres caractéristiques. Parmi ses langages, notre choix est focalisé sur python.



Figure 4.1: python logo

Python est un langage de programmation de haut niveau conçu pour sa lisibilité et sa nature moins complexe, créé par « Guido Van Rossum » et la première version de python est sortie en 1991.

C'est le langage le plus utilisée dans le domaine de l'intelligence artificielle et du deep Learning. Imaginons que tout ce qui existe autour est sous forme de données et ces données sont brutes, inadéquates, incomplètes, non structurées et volumineuses. Python peut servir de guide à l'apprentissage profond pour résoudre tous ces problèmes

Python est un langage stable, flexible et il fournit divers outils pour les développeurs. Ce qui permet de le classer en premier choix pour l'apprentissage automatique et l'apprentissage profond, à partir du développement, la mise en œuvre, et la maintenance. Il présente de nombreuses caractéristiques intéressantes telles que :

- Il est multiplateforme, c'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation : Windows, Mac OS X, Linux, Android, iOS, depuis les mini-ordinateurs Raspberry Pi jusqu'aux supercalculateurs.
- C'est un langage facile à lire, de sorte que les développeurs peuvent facilement comprendre le code
- C'est un langage interprété. Un script Python n'a pas besoin d'être compilé pour être exécuté, contrairement à des langages comme le C ou le C++. Il convient bien à des scripts d'une dizaine de lignes qu'à des projets complexes de plusieurs dizaines de milliers de lignes et il est relativement simple à prendre en main.

Pour le python plusieurs IDE sont disponibles.

4.2.2.1 PyCharm



Figure 4.2: PyCharm logo

C'est un environnement de développement intégré permettant la programmation en Python.

Il active l'analyse de code et contient un débogage graphique. Il prend également en charge la gestion des tests unitaires, l'intégration logicielle de gestion de versions et le développement Web avec Django.

Développé par la société tchèque JetBrains, c'est un logiciel multiplate-forme qui tourne sur Windows, Mac OS X et GNU/Linux. Il est disponible en version professionnelle, distribué sous licence propriétaire, et en version communautaire publiée sous licence Apache[36].

4.2.2.2 PyScripter



Figure 4.3: PyScripter logo

Est un logiciel libre Python et un Environnement de développement (IDE) sur Windows. C'est un environnement de développement intégré et une série de programmes utilisés pour le développement de logiciels et destinés à une utilisation commune [37].

4.2.2.3 Spyder



Figure 4.4: Spyder logo

Spyder (appelé Pydee au début) est un environnement de développement pour Python. Gratuit (licence MIT) et multiplateforme (Windows, Mac OS, GNU/Linux), il intègre plusieurs bibliothèques d'utilisation scientifique : Matplotlib, NumPy, SciPy et IPython.

Créé et développé par Pierre Raybaut en 2008, En comparaison avec d'autres IDE pour le développement scientifique, Spyder dispose d'un ensemble unique de caractéristiques - multiplate-forme, open-source, écrit en Python et disponible sous une licence sans copyleft. Spyder est extensible avec des plugins, comprend le support d'outils interactifs pour l'inspection des données et incorpore des instruments d'assurance de la qualité et d'introspection spécifiques au code Python, tels que Pyflakes, Pylint et Rope [38].

4.2.2.4 Microsoft Visual Studio



Figure 4.5: visual studio logo

Est une suite logicielle de développement pour Windows et Mac OS développée par Microsoft. La dernière version porte le nom de Visual Studio 2022.

Visual Studio est un ensemble complet d'outils de développement pour générer des applications web ASP.NET. Des services web XML, des applications bureautiques et des applications mobiles. VS Code (python) utilisent tous le même environnement de développement intégré (IDE) [39].

4.2.2.5 Google Colab



Figure 4.6: google colab Logo

Colaboratory ou 'Colab'. Permet d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. Offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le Cloud. Colab permet [40]:

- D'améliorer les compétences de codage en langage de programmation Python.
- De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires telles que Keras, TensorFlow, PyTorch et OpenCV.
- D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration, Mais la fonctionnalité qui distingue Colab des autres services est l'accès à un processeur graphique GPU, totalement gratuit.

Les algorithmes de l'apprentissage automatique sont très complexes, mais Python est le secours avec une large gamme de bibliothèques. Dans notre système nous avons utilisé les bibliothèques suivantes :

- **Numpy** : La bibliothèque numérique propose le type ndarray en plus de nombreuses fonctions à virgule flottante notamment. Traditionnellement, il est importé : soit directement dans l'environnement courant « from numpy import... », ou sous un nom abrégé « import numpy as np ».

La Bibliothèque numpy est spécialisée dans le traitement des matrices. On l'utilisera principalement pour les vecteurs et les matrices où les tableaux Numpy ne traitent que des objets du même type et la bibliothèque numérique fournit également un grand nombre de procédures pour un accès rapide aux données, pour des traitements divers et pour des calculs par exemple, statistiques et scientifiques arithmétique, calcul de chaque élément séparément, arithmétique matricielle.

Les tableaux Numpy sont plus efficaces que les éléments récurrents habituels en Python, et en plus des méthodes d'accès aux éléments déjà vues pour les listes, ils prennent en charge l'indexation par une liste (ou un tableau) d'entiers. Le tableau composé des cases des indicateurs spécifiés est référencé. On peut aussi indexer par un tableau de booléens de même taille. Ensuite, les endroits où se trouve la valeur réelle sont indiqués [41].

- **Matplotlib** : La bibliothèque matplotlib (et sa sous-bibliothèque pyplot) sert essentiellement à afficher des graphismes.

Son utilisation ressemble beaucoup à celle de Matlab. Traditionnellement, on l'importe Soit directement dans l'environnement courant (from matplotlib. Pyplot import *), Soit sous un nom abrégé (import matplotlib. pyplot as plt).

Dans ce qui suit, on suppose qu'on utilise le backend inline, propre aux notebooks. Si on utilise un autre backend, il pourra être nécessaire d'ajouter show () à la fin des commandes pour faire apparaître la fenêtre qui contient le dessin [42].

- **SciPy** : est utilisé en synergie avec Numpy parce qu'il fonctionne sur les données sous forme de matrice Numpy. SciPy propose un catalogue d'opérations scientifiques : algèbre linéaire, algorithmes de régression, fonctions statistiques...

SciPy vous permet de travailler sur des projets d'amélioration digitale qui visent à obtenir le plus petit (ou le plus grand) nombre possible en changeant certaines variables. SciPy permet aussi des travaux de génie avec des fonctions physiques [43].

- **TensorFlow** : est un Framework Python populaire pour le Machine Learning et le Deep Learning, qui a été développé à Google Brain. C'est le meilleur outil pour des

tâches comme l'identification d'objets, la reconnaissance vocale et bien d'autres. Il permet de travailler avec des réseaux neuronaux artificiels qui doivent gérer plusieurs ensembles de données. La bibliothèque comprend plusieurs aides de couches (tflearn, tf-slim, skflow), qui la rendent encore plus fonctionnelle. TensorFlow s'enrichit constamment de nouvelles versions, notamment en corrigeant les éventuelles failles de sécurité ou en améliorant l'intégration de TensorFlow et du GPU [44].

- **Keras** : est une excellente bibliothèque pour la construction de réseaux de neurones et la modélisation. Elle est très simple à utiliser et offre aux développeurs un bon degré d'extensibilité. La bibliothèque tire parti d'autres paquets (Theano ou TensorFlow) comme terminaux. De plus, Microsoft a intégré CNTK (Microsoft Cognitive Toolkit) pour servir d'autre backend. C'est un excellent choix si vous voulez expérimenter rapidement en utilisant des systèmes compacts – l'approche minimaliste de la conception est vraiment top [45].

4.2.3 Corpus de parole

Les signaux de parole, utilisés pour la phase de test du module de la reconnaissance automatique de parole, sont extraits à partir du corpus de parole monolocuteur "Arabic speech corpus". Ce corpus a été développé dans le cadre d'une thèse de doctorat de l'Université de Southampton. La fréquence d'échantillonnage est fixée à 16KHz et avec une quantification sur 16bits.

Ce corpus contient 1813 fichiers audios sous format "wav" et 1813 fichiers des énoncés textuels. Il contient également les fichiers de segmentation en phonème effectué grâce à une segmentation manuelle. Le tableau 3 illustre la transcription orthographique et les symboles phonétiques utilisés dans le corpus, qui contient 35 phonèmes de la langue arabe.

Tableau 4.1: Phonèmes et leurs notations choisies

Numéro	Phonème arabe	Symbole utilisé	Numéro	Phonème arabe	Symbole utilisé
1	"أ"	<	19	"غ"	g
2	"ب"	b	20	"ف"	f
3	"ت"	t	21	"ق"	q
4	"ث"	^	22	"ك"	k
5	"ج"	J	23	"ل"	l
6	"ح"	H	24	"م"	m
7	"خ"	x	25	"ن"	n

8	"د"	d	26	"ه"	h
9	"ذ"	*	27	"و"	w
10	"ر"	r	28	"ي"	y
11	"ز"	z	29	"ا"	a
12	"س"	s	30	"ا0"	u0
13	"ش"	\$	31	"ا0"	i0
14	"ص"	S	32	"آ"	aa
15	"ض"	D	33	"و"	uu
16	"ظ"	T	34	"ي"	ii
17	"ظ"	Z	35	"sil"	sil
18	"ع"	E			

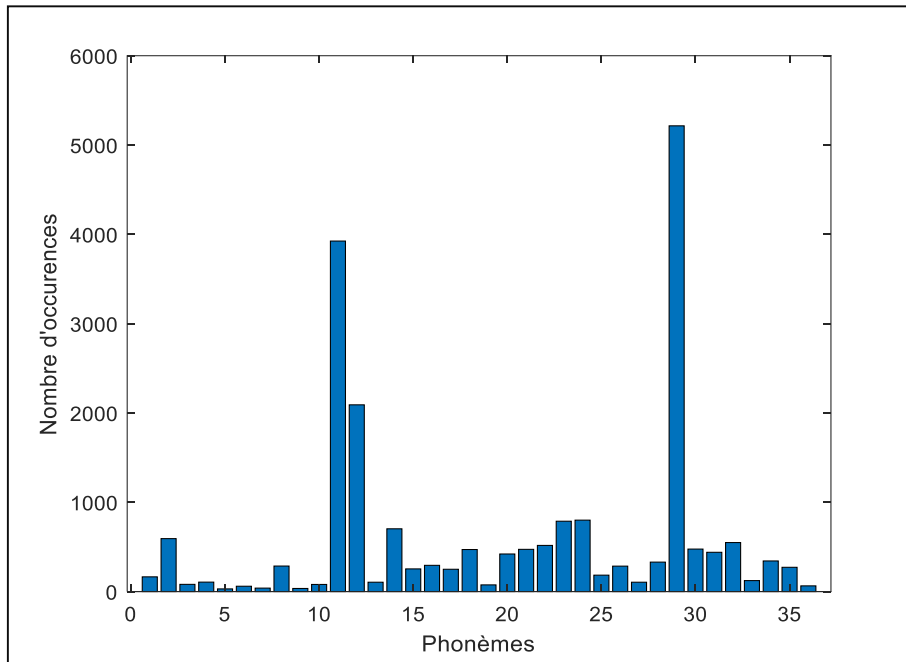


Figure 4.7: Occurrences de phonèmes dans le corpus

La figure 19 représente les 35 occurrences de phonèmes dans le corpus de parole, qui sont divisés en quatre catégories : moins de 100 Occurrences, de 102 à 400 occurrences, de 401 à 700 occurrences, et plus de 701 occurrences. Le nombre de phonèmes indiqué à la figure 19 a été extrait du tableau 3.

4.3 Traduction automatique parole-parole

Le système proposé est composé de trois modules principales : module de reconnaissance automatique de parole (parole-text), module de traduction Arabe-Anglais (text-text) et

module de la synthèse de parole (text-parole). La figure montre un schéma fonctionnel du système proposé.

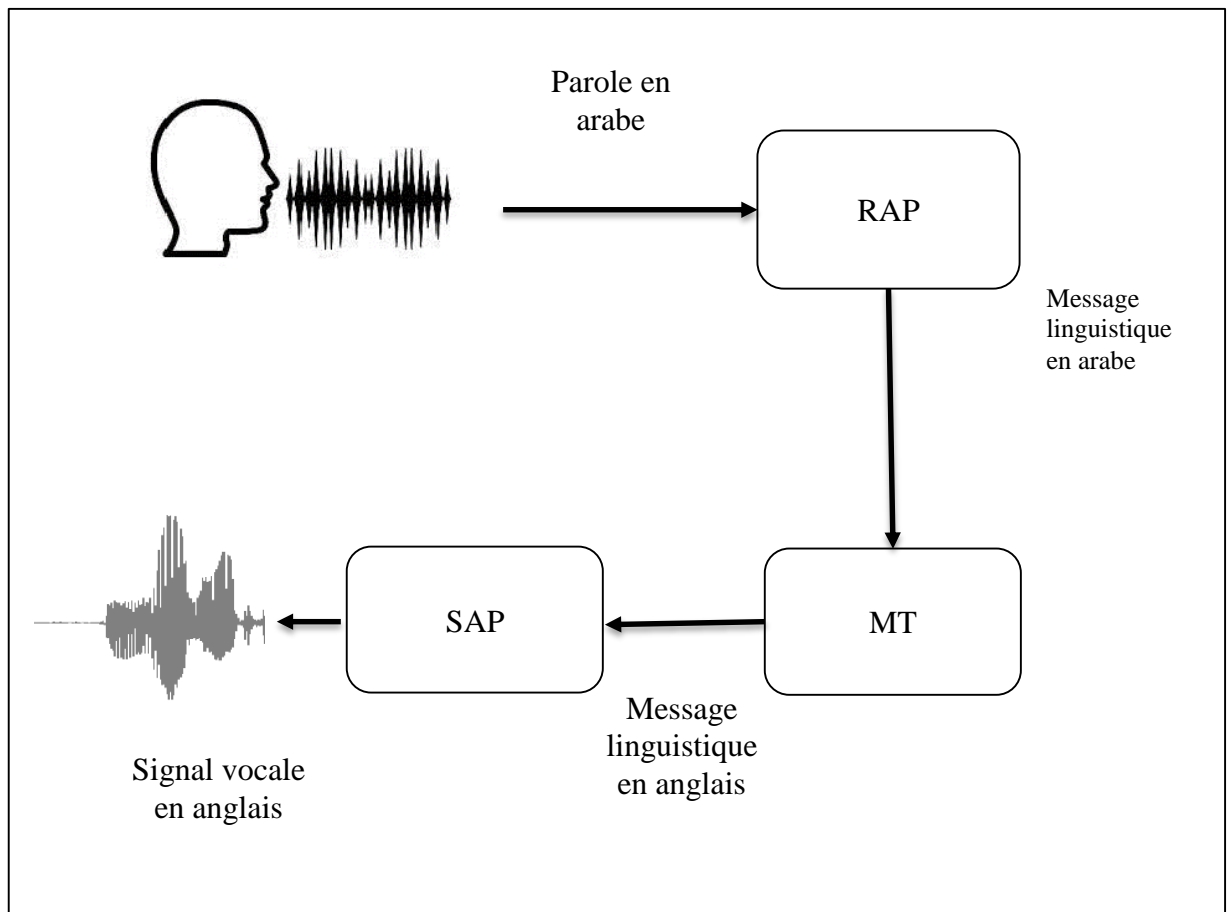


Figure 4.8: Schéma fonctionnel du système de traduction automatique Arabe-Anglais

4.3.1 Reconnaissance automatique de la parole arabe

La reconnaissance automatique de la parole est une technique qui permet à un ordinateur d'identifier les mots qu'une personne dit dans un microphone.

Dans cette étape, nous utilisons un système RAP basé sur DNN qui s'appuie sur l'architecture CNN pour reconnaître la parole en langue arabe et conversion en texte écrit

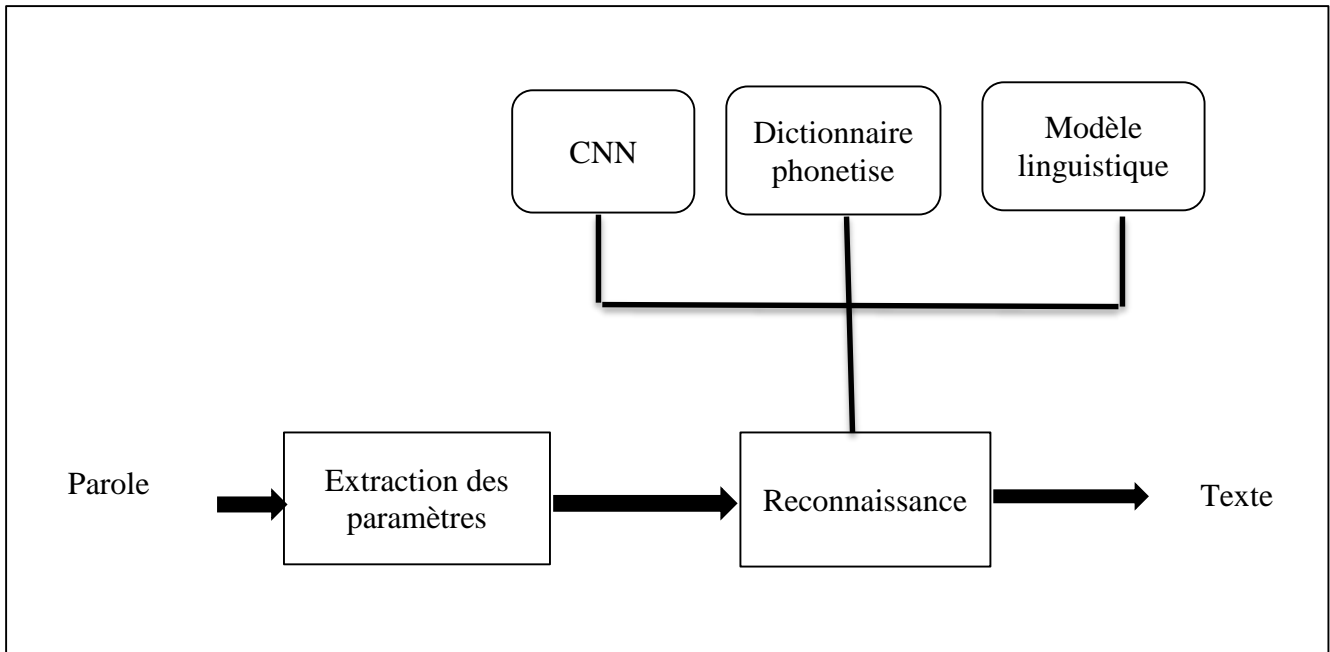


Figure 4.9 : fonctionnement de la reconnaissance

- **La base de données d'évaluation**

La base de données utilisée dans la partie de la reconnaissance automatique de la parole "RAP" est la base de données Arabe Speech corpus. L'arabic speech corpus est un corpus de parole en Arabe standard moderne "MSA" à un seul locuteur. Ce corpus a été développé dans le cadre du travail de doctorat effectué par Nawar Halabi à l'université Southampton. Le corpus était enregistré dans un studio professionnel en arabe Levantin du sud (Accent damasien). Il contient :

- 1813 fichiers audio sous format "wav"
- 1813 fichiers des énoncés textuels
- 1813 dossiers de Lab

4.3.2 Implémentation du module RAP

Dans cette section, nous présentons les différentes étapes pour l'implémentation du module de la reconnaissance automatique de parole. Ce module est basé sur la bibliothèque KLAAM

Dans cette section, nous présentons les différentes étapes pour l'implémentation du module de la reconnaissance automatique de la parole. Ce module est basé sur la bibliothèque KLAAM. Cette bibliothèque est conçue à l'aide du deep learning (CNN). Elle est implémentée sous colab

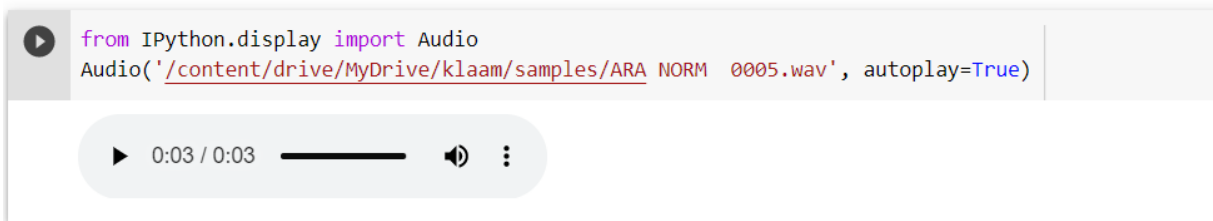
L'installation de cette bibliothèque est effectuée à l'aide du script

Pip install requirements.txt

Une fois que la bibliothèque sera installée il est possible d'importer un exemple de fichier audio :

```
from IPython.display import Audio
Audio('./samples/demo.wav', autoplay=True)
```

Pour le fichier audio copier le chemin d'accès d'un import fichier audio existe dans le dossier 'samples'



importer un objet de la reconnaissance de la parole

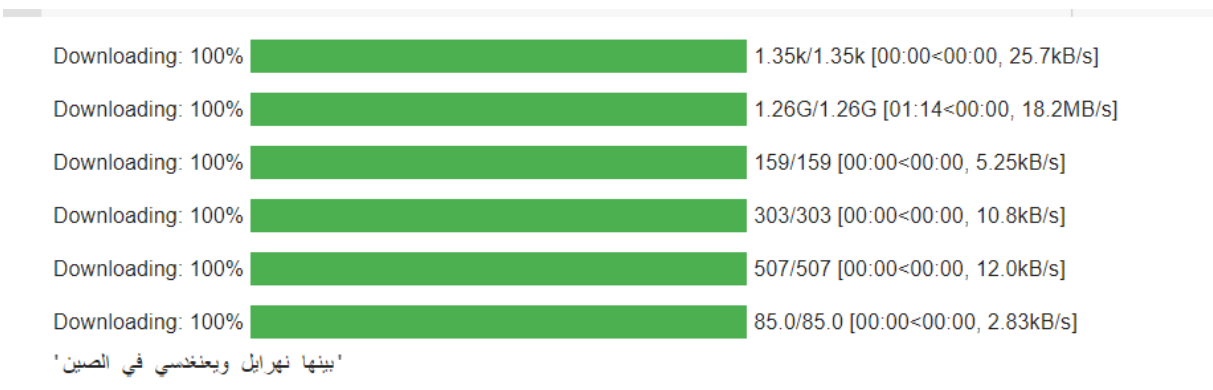
```
from klaam import SpeechRecognition
```

Il existe deux modèles de reconnaissance disponibles pour le MSA et le dialecte égyptien. Vous pouvez définir n'importe lequel d'entre eux à l'aide de l'Lang attribut `model = SpeechRecognition(lang='msa')`

ainsi, transcrire le même chemin d'accès d'un fichier audio précédemment

```
model.transcribe('/content/drive/MyDrive/klaam/samples/ARA NORM 0005.wav')
```

main tenant, exécutez à Colab pour obtenir le résultats "le message linguistique"



4.3.3 Test d'évaluation de la reconnaissance

La mesure d'évaluation la plus répandue de la précision d'un système de reconnaissance automatique de la parole est le taux d'erreur mot "Word Error Rate en Anglais". Le WER compare la sortie prédite et la transcription de référence mot par mot pour déterminer le nombre de différence entre elles. Il est défini comme suit :

$$\text{WER} = \frac{I+D+S}{W} \cdot 100$$

Où :

- I : le nombre d'insertion
- D : le nombre de suppression
- S : le nombre de substitutions
- W : le nombre de mots dans la référence

Tableau 4.2 : le résultat de la reconnaissance de 10 audio

Waves	WER
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	7.69
9	11.11
10	0

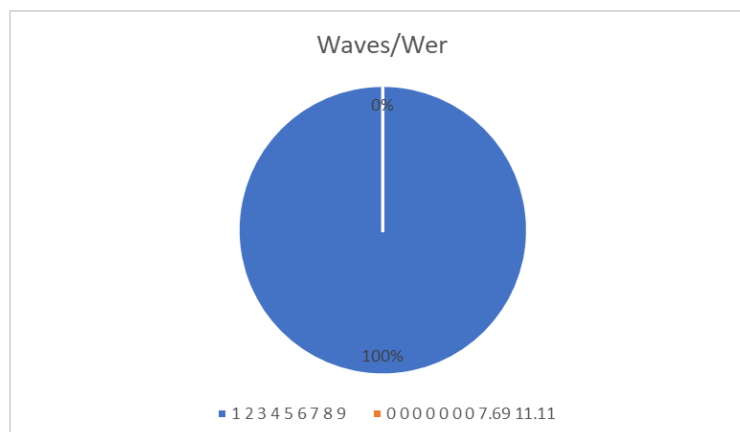


Figure 4.10 : le graphe statistique des résultats

On remarqué que la persentage des erreurs dans la partie reconnaissance elle tres faible presque 0 c'est-à-dire que la reconnaiossance a base de deeplearning est de bon qualité

4.3.4 Implémentation du module de traduction

Le module de traduction arabe-Anglais est basé sur la bibliothèque dl-translate. Cette bibliothèque est concu à laide du deep learning (LSTM), le tronsfomrer, tensorflow et keras. Elle est implémenté sous python et offert la possiblité de faire la traduction bidirectionnelle entre 50 differentes langues. L'installation de cette bilbiothèque est effecté à l'aide du script

```
pip install dl-translate
```

Une foix que la bibliothèque sera installé il est possible de l'importer comme un objet *dlt*:

```
import dl_translate as dlt
```

```
mt = dlt.TranslationModel()
```

```
Ci-dessus, text_ar = "الأمين العام للأمم المتحدة يقول إنه لا يوجد حل عسكري في سوريا".  
mt.translate(text_ar, source="Arabic", target="fr")
```

On peut voir que `dlt.lang` contient des variables représentant chacune des 50 langues disponibles avec prise en charge de la saisie semi-automatique. Alternativement, vous pouvez spécifier la langue (par exemple "arabe") ou le code de langue (par exemple "fr" pour le français) :

Nous pouvons vérifier si une langue est disponible, vous pouvez la vérifier :

```
print(mt.available_languages())  
print(mt.available_codes())  
print(mt.get_lang_code_map())
```

4.3.4.1 Sélection d'un appareil

Lorsque vous chargez le modèle, on peut spécifier le périphérique à l'aide de `device` argument. Par défaut, la valeur sera `device="auto"`, ce qui signifie qu'il utilisera un GPU si possible. Vous pouvez également définir explicitement `device="cpu"` ou `device="gpu"`. En général, il est recommandé d'utiliser un GPU pour un temps de traitement raisonnable.

```
mt = dlt.TranslationModel(device="auto")  
mt = dlt.TranslationModel(device="cpu")  
mt = dlt.TranslationModel(device="gpu")  
mt = dlt.TranslationModel(device="cuda:2")
```

4.3.4.2 Modification du modèle que vous chargez

Deux familles de modèles sont disponibles pour le moment : [m2m100](#) et [mBART-50 Large](#) , qui permettent respectivement la traduction dans plus de 100 langues et 50 langues. Par défaut, le modèle sélectionnera m2m100, mais vous pouvez également choisir explicitement le modèle en spécifiant le raccourci ("m2m100"ou "mbart50") ou le nom complet du référentiel (par exemple "facebook/m2m100_418M"). Par exemple:

```
mt = dlt.TranslationModel("m2m100")
mt = dlt.TranslationModel("facebook/m2m100_418M")

mt = dlt.TranslationModel("mbart50")
mt = dlt.TranslationModel("facebook/mbart-large-50-many-to-many-mmt")
```

4.3.4.3 Utilisation hors ligne

Contrairement aux API Google translate ou MSFT Translator, cette bibliothèque peut être entièrement utilisée hors ligne. Cependant, vous devrez d'abord télécharger les packages et les modèles, puis les déplacer vers votre environnement hors ligne pour les installer et les charger dans un fichier venv.

Tout d'abord, exécutez dans votre terminal :

```
mkdir dlt
cd dlt
mkdir libraries
pip download -d libraries/ dl-translate
```

Une fois tous les packages requis téléchargés, vous devrez utiliser le hub huggingface pour télécharger les fichiers. Installez-le avec pip install huggingface-hub. Ensuite, exécutez à l'intérieur de Python :

```
import os
import huggingface_hub as hub

dirname = hub.snapshot_download("facebook/m2m100_418M")
os.rename(dirname, "cached_model_m2m100")
```

Maintenant, déplacez tout dans le dlt répertoire vers votre environnement hors ligne. Créez un environnement virtuel et exécutez ce qui suit dans le terminal :

```
pip install --no-index --find-links libraries/ dl-translate
```

Maintenant, exécutez à l'intérieur de Python :

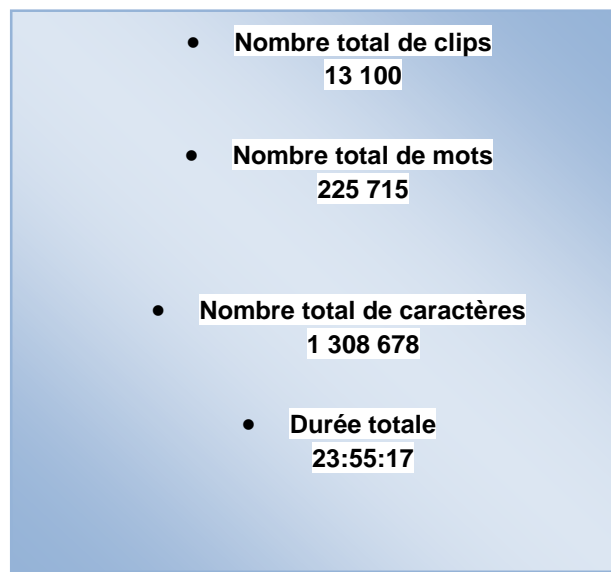
```
import dl_translate as dlt

mt = dlt.TranslationModel("cached_model_m2m100", model_family="m2m100")
```

4.3.5 Synthèse automatique de la parole anglais

4.3.5.1 La base de données d'évaluation pour mel GAN

La base de données utilisée dans la partie de la synthèse automatique de la parole "SAP" est la base de données LJSpeech . Il s'agit d'un ensemble de données vocales du domaine public composé de 13 100 courts extraits audio d'un seul locuteur lisant des passages de 7 livres de non fiction. Une transcription est fournie pour chaque clip. Les clips varient en longueur de 1 à 10 secondes et ont une durée totale d'environ 24 heures



- **Nombre total de clips**
13 100
- **Nombre total de mots**
225 715
- **Nombre total de caractères**
1 308 678
- **Durée totale**
23:55:17

4.3.6 Implémentation du module de synthèse

Pour la dernière section de la traduction automatique de STS, nous présentons les étapes d'implémentation du module de la synthèse automatique de la parole. Il est basé sur la bibliothèque Transformer TTS. Transformateur TTS est Un transformateur texte-parole dans TensorFlow 2 Synthèse audio avec transformateur direct TTS et Vocodeur MelGAN. Elle est implémentée sous colab. L'installation de cette bibliothèque est effectué

```
apt-get install -y espeak
```

```
pip install -r TransformerTTS/requirements.txt
```

si la bibliothèque sera installé il est possible de l'importer comme un objet path et sys

```
from pathlib import Path  
MelGAN_path = 'melgan/'  
TTS_path = "TransformerTTS/"
```

```
import sys
```

```
sys.path.append(TTS_path)
```

puis Charger le modèle pré-entraîné

```
from model.factory import tts_ljspeech
from data.audio import Audio
```

```
model, config = tts_ljspeech()
audio = Audio(config)
```

une fois tous les exécution juste, nous Synthétiser le texte

```
sentence = 'Hello professors.'
```

```
out_normal = model.predict(sentence)
```

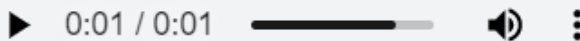
ainsi, Convertir le spectrogramme en wav

```
wav = audio.reconstruct_waveform(out_normal['mel'].numpy().T)
```

en fin, nous trouve un signal vocal.

```
import IPython.display as ipd

ipd.display(ipd.Audio(wav, rate=config['sampling_rate']))
```



Le vocal obtenu est n'est pas clair, Faites un peu de nettoyage sy par MelGAN où MelGAN est un modèle de réseau contradictoire génératif (GAN) qui génère de l'audio à partir du spectrogramme Mel. Nous pouvons importer torch et numpy

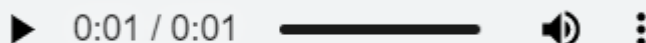
```
sys.path.append(MelGAN_path)
import torch
import numpy as np
```

```
vocoder = torch.hub.load('seungwonpark/melgan', 'melgan')
vocoder.eval()
```

```
mel = torch.tensor(out_normal['mel'].numpy().T[np.newaxis,:,:])
```

finalement , exécutez pour Affichage audio très clair

```
# Display audio
ipd.display(ipd.Audio(audio.cpu().numpy(), rate=22050))
```



4.3.7 Test d'évaluation subjectif de la qualité de la parole :

Les systèmes de SAP ont été évalués sous différents aspects, tels que l'intelligibilité, la compréhension le naturel et la préférence de la parole synthétique. La qualité de la voix de la voix générée diffère d'un système à l'autre selon la technique utilisée.

Tableau 4.3: Niveau de compréhension des signaux synthétiques

	Intelligibilité	Naturel
1	Très bien	Suffisant
2	Excellente	Modéré
3	Excellente	Très insuffisant
4	Très Bien	Très mauvais
5	Excellente	Très mauvais
6	Très Bien	Modéré
7	Excellente	Bien
8	Excellente	Modéré
9	Bien	Insuffisant
10	Excellente	Modéré
	Taux= 60%	Taux= 40%

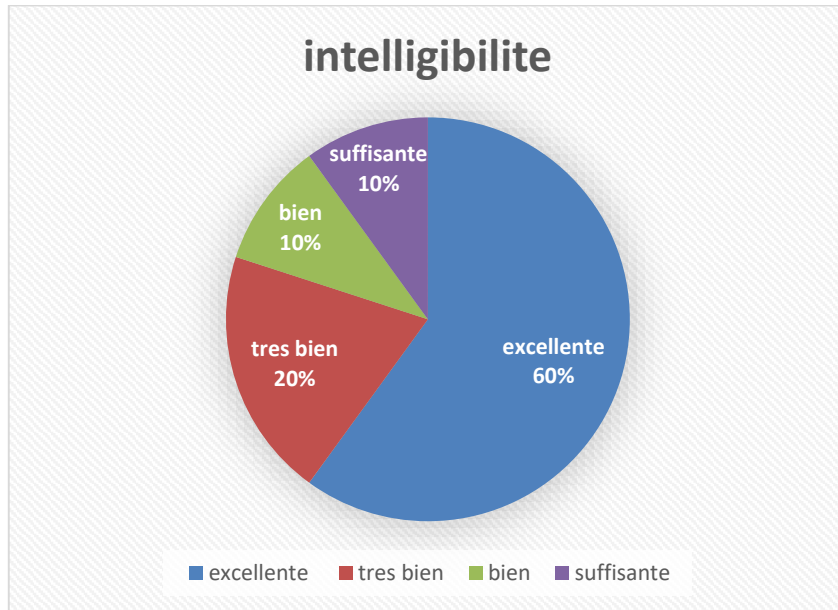


Figure 4.11 : Graphe de l'intelligibilité

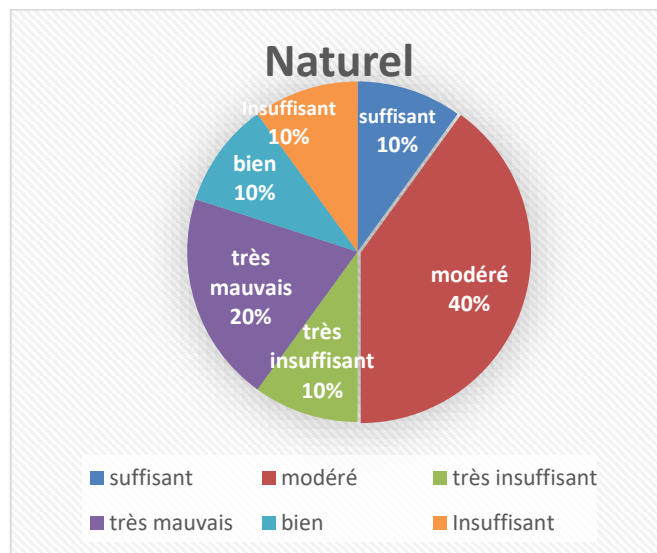


Figure 4.12 : le graphe du naturel

Afin d'évaluer la qualité de la parole synthétisée obtenue à l'aide de Mel GAN, nous avons fait recours aux tests d'évaluation subjectif. Il s'agit de faire entendre de la parole synthétisée par 10 personnes qui donneront leurs avis sur sa qualité du point de vue de son intelligibilité et son naturel

4.4 Conclusion

Dans ce chapitre on a présenté la traduction automatique Arabe-Anglais avec Deep Learning sous l'environnement de programmation python (google colab). Nous avons présenté les étapes d'implémentation des trois modules : reconnaissance automatique de parole, traduction Arabe-Anglais et synthèse de parole. Le système résultant montre des performances importantes pour le problème de la traduction automatique Arabe-Anglais (Speech To Speech).

Conclusion Générale

Notre projet est de créer une application éducative pour la langue anglaise, basée sur la traduction vocale automatique de l'arabe vers l'anglais, basée sur le deep learning et les réseaux de neurone

Dans ce mémoire, nous avons abordé la généralité sur la langue et la parole dans le premier chapitre, le deuxième chapitre nous avons expliqué la nature de la traduction automatique de la voix et ses étapes, tandis que le troisième chapitre nous avons étudié le sujet de l'apprentissage en profondeur et traité les plus importants. Neurones utilisés dans notre projet et le quatrième chapitre nous avons traité du résultat final et évalué la qualité et la précision du système

Nous avons travaillé sur l'évaluation de chaque étape du projet unitaire pour estimer et réduire le niveau de la dernière erreur dans le résultat final et nous sommes rendu compte que :

La reconnaissance vocale sur laquelle nous avons travaillé en nous appuyant sur la base de données Nawar Al-Halabi, nous nous sommes rendu compte que le taux d'erreur dans ce système est quasi inexistant, mais parfois il a du mal à convertir la parole dans l'enregistrement audio moyen ou long en texte, car nous avons obtenu un résultat incomplet

Pour la synthèse de la parole, nous nous sommes entièrement appuyés sur GAN qui contient une bonne base de données, en plus du son robotique, avec intelligibilité bien mais les personnes âgées peuvent avoir quelques problèmes pour comprendre la parole en raison de son.

Références Bibliographiques

- [1] <http://www.claudegabriel.be/Cine%20acoustique%209.pdf>
- [2] <https://flenantes.org/lappareil-phonatoire-echauffement-detente-musculation/>
- [3] M. H. Rahmouni – M. A. H. Benouared, « Utilisation des réseaux de neurone pour la reconnaissance automatique de la parole-, », mémoire de master, université de Saad Dahleb département d’Aéronautique Blida 1, Algérie, 2011.
- [4] https://www.researchgate.net/figure/Spectrogrammes-experimentaux-a-Spectrogramme-du-signal-vocal-b-Spectrogramme-du_fig2_266183651
- [5] <https://blog.lingoda.com/fr/les-langues-les-plus-parlees-dans-le-monde/>
- [6] <https://docplayer.fr/20829888-Le-systeme-phonetique-de-l-arabe.html>
- [7] M. M. Ferrougha, « Reconnaissance des formes phonétiques en arabe-, » mémoire de master, département de l’électronique université Saad Dahleb Blida1, Algérie,2010.
- [8] <https://journals.openedition.org/cerri/docannexe/image/2900/img-7.png>
- [9] M. A. Szczegielniak, « Phonetics : The Sounds of Language-, », cours de doctorat, université de Harvard, Etats-Unis.
- [10] M. N. Musk, « The Vowels & Consonants of English Lecture Notes », cour de doctorat, Département de Institution Culture et Communication, université Linköpings, suède
- [11] <https://intelligence-artificielle-robotique.weebly.com/application-agrave-la-reconnaissance-vocale-et-de-caractegravere.html>
- [12] M. A. H. Hammadeche - M. Taki, « Reconnaissance automatique de la parole arabe continue-, », mémoire, département d’informatique université de Saad Dahleb, Blida1, 2019.
- [13] M. A. BENDAHMANE, « Reconnaissance de la parole par distance DTW exemple d’application pour la reconnaissance de chiffres isolé dans la langue arabe-, », Article, Laboratoire SIMPA, Oran, Algérie, 2014.
- [14] M. A. LOUNI - M. A. BENYETTOU, « Un codage neuro-prédictif pour l’extraction des traits distinctifs appliqué à la reconnaissance des phonèmes arabe-, », Oran, Algérie, 2008.
- [15] https://www.researchgate.net/figure/steps-of-MFCC-Feature-Extraction-The-spectrum-produced-from-the-human-vocal-tract-and-it_fig1_338006282

- [16] M. N. Zerari, « Intégration d'un module de reconnaissance de la parole au niveau d'un système au diovisuel- Application téléviseur-, », thèse de doctorat en Génie industriel, université Batna 2, Algérie, Avril 2021.
- [17] M. N. Asbai, « Identification et authentification de locuteurs, par les techniques de fusion des paramètres et les modèles dans un environnement réel-, », thèse de doctorat en télécommunication université de Houari Boumediene, Alger, Algérie,2015.
- [18] https://stringfixer.com/fr/Machine_translation
- [19] M. R. Benzeghioa – M. A. Sifi, « Proposition d'un système de traduction automatique anglais arabe -, » mémoire de master en Traitement automatique de la langue, université Saad Dahleb Blida 1, Algérie ,2019.
- [20] M. H. Abdelli - A. Echikr, « Traduction automatique de la parole de l'anglais vers l'arabe-, », mémoire de master en Traitement automatique de la langage, université Saad Dahleb Blida 1, Algérie,2021.
- [21] <https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346>
- [22] M. A. Amrouche « Contribution à l'amélioration du signal de synthèse dans un système TTS pour la langue arabe-, », thèse de doctorat en sciences en Communication parlée, université Houari Boumediene, Alger, Algérie, 2017.
- [23] M. K. Hemina- M. O. Heminna, « Développement d'une voix arabe synthétique open source-, » mémoire de master en Ingénierie de logiciels, université Saad Dahleb Blida 1, Algérie,2020.
- [24] <https://fr.gadget-info.com/46230-9-best-text-to-speech-tts-software>
- [25] <https://www.ibm.com/cloud/learn/what-is-artificial-intelligence>
- [26] M. A. ALLAL. « Utilisation du deep learning dans la radio cognitive-, » mémoire de Master professionnel en Réseaux et Systèmes Distribués (R.S.D), Université Abou Bakr Belkaid, Tlemcen, 2018.
- [27] M. S. K. Chadalawada, « Real Time Object Detection and Recognition Using Deep Learning Methods-, », mémoire de Master professionnel en Sciences Informatique, Institut de technologie de Blekinge, Suède, 2020.
- [28] M. H. Trad, « La détection d'objet avec OpenCV et deep learning-, », mémoire de Master professionnel en Réseaux et Télécommunication, Université Mohamed Khider, Biskra, 2020.
- [29] <https://www.ibm.com/cloud/learn/deep-learning> .
- [30] M. L. Amiar, « Un système Hybride AG/PML pour la reconnaissance de la parole arabe-, », mémoire de master en intelligence artificielle distribuée (IAD), université Badji Mokhtar, Annaba, Algérie,2005.
- [31] https://www.researchgate.net/figure/Schema-dun-neurone_fig1_280792237
- [32] M. A. B. Louam, « Deep learning base sur les méthodes de réduction pour la reconnaissance de visage-, », mémoire de master, département de génie électrique université Mohamed khaidar, Biskra, Algérie, 2019.
- [33] M. M. A. Djaballah, « Système de prédiction consommation d'énergie base deep learning-, », mémoire de master département d'informatique université de 8 Mai 1945, Guelma, Algérie, 2021.

- [34] M. K. Ben Kaddour, « Reconnaissance des formes et classification automatique : Application à l'identification biométrique- », thèse de doctorat en traitement du signal et de l'image, Université Djillali Liabès, Sidi Bel Abbès, Algérie, 2020.
- [35] <https://penseeartificielle.fr/comprendre-lstm-gru-fonctionnement-schema/>
- [36] <https://www.jetbrains.com/pycharm/download/other.html>
- [37] <https://www.techworld.com/review/programming-software/pyscripter-review-3238828/>
- [38] <https://github.com/spyder-ide/spyder/tree/v1.0.0>
- [39] https://fr.wikipedia.org/wiki/Microsoft_Visual_Studio
- [40] <https://www.zebnet.co.uk/freeware/tools-and-utilities/duplicate-line-remover>
- [41] <https://cedric.cnam.fr/vertigo/Cours/NFE205/tpPython.html>
- [42] <https://zestedesavoir.com/tutoriels/469/introduction-aux-graphiques-en-python-avec-matplotlib-pyplot/>
- [43] <https://github.com/scipy/scipy/releases/tag/v1.9.0>
- [44] <https://www.kernix.com/tensorflow-un-ecosysteme-open-source-dedie-a-la-creation-des-modeles-de-deep-learning/>
- [45] https://docs.aws.amazon.com/fr_fr/dlami/latest/devguide/tutorial-keras.html