République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab Blida-1

Faculté Des Sciences

Département d'informatique

Projet de fin d'étude

Pour l'obtenir du diplôme de Master

Domaine : Mathématique et Informatique

Filière: Informatique

Spécialité : Traitement Automatique de la Langue

Les réseaux de neurones pour le développement d'un système TTS en Arabe .

Réalisé par :

GHERBI Roumaissa & MELLEK Ichrak

Soutenu le 25/09/2022 devant les jurys composés de :

Mme Tebbi Hanane Maître de conférences A à USD-Blida Promotrice.

Mme Mezzi Melyara Maître de conférences B à USD-Blida Présidente.

Mr Ykhlef Faycal Maître de recherche à CDTA-Alger Examinateur.

Résumé

La synthèse vocale est la technologie qui permet l'automatisation de la production de la parole par une machine. Le rôle de la synthèse vocale à partir d'un texte donné en entrée est d'assurer la lecture de ce texte à partir d'une transformation du texte orthographique en une séquence de sons ou phonèmes. Ce travail se concentre sur l'approche de production de la voix à partir d'un texte Arabe. Notre objectif est de proposer une nouvelle approche qui intègre les avantages de l'apprentissage profond dans le domaine de la synthèse vocale en utilisant les réseaux de neurones spécialement les réseaux de neurones convolutifs (CNN).

Nous détaillons cette approche et nous décrivons les principales étapes de sa réalisation en commençant par la préparation de la base de données puis le traitement linguistique appliqué pour passer d'un graphème à un phonème et à la fin la production acoustique du texte précédemment acquis.

A la fin, nous illustrons et nous expliquons les résultats obtenus par le biais d'un rapport d'évaluation basé sur le MOS *(Mean Opinion Score :* le score moyen d'opinion) pour calculer l'intelligibilité du discours et l'aspect naturel du texte parlé.

Mot Clés : systèmes de synthèse vocale à partir de texte, Transcription orthographique Phonétique (TOP), Intelligence Artificielle (IA), Arabe Standard (AS), Apprentissage profond (DL), Réseau de neurones convolutif (CNN), Score moyen d'opinion (MOS).

Abstract

Speech synthesis is the technology that allows the automation of speech production by a machine. The role of speech synthesis from a text given as input is to ensure the reading of this text from a transformation of the orthographic text into a sequence of sounds or phonemes. This work focuses on the approach of voice production from an Arabic text. Our goal is to propose a new approach that integrates the advantages of deep learning in the field of speech synthesis using neural networks especially convolutional neural networks (CNN).

We detail this approach and describe the main steps of its realization starting with the preparation of the database, then the linguistic processing applied to go from a grapheme to a phoneme and at the end the acoustic production of the previously acquired text.

At the end, we illustrate and explain the results obtained through an evaluation report based on the MOS (Mean Opinion Score) to calculate the intelligibility of the speech and the naturalness of the spoken text.

Keywords: text-to-speech systems, Orthographic Phonetic Transcription (OPT), Artificial Intelligence (AI), Standard Arabic (SA), Deep Learning (DL), Convolutional Neural Network (CNN), Mean Opinion Score (MOS).

الملخص

تخليق الكلام هو التكنولوجيا التي تسمح بأتمتة إنتاج الكلام بواسطة الآلة. يتمثل دور تركيب الصوت من النص المعطى كمدخل في ضمان قراءة هذا النص من تحويل النص الهجائي إلى سلسلة من الأصوات أو الصوتيات. يركز هذا العمل على منهج إنتاج الصوت من نص عربي. هدفنا هو اقتراح نهج جديد يدمج مزايا التعلم العميق في مجال تخليق الكلام باستخدام الشبكات العصبية، وخاصة الشبكات العصبية التلافيفية (CNN). نقوم بتفصيل هذا النهج ونصف المراحل الرئيسية لتحقيقه، بدءًا من إعداد قاعدة البيانات، ثم المعالجة اللغوية المطبقة للانتقال من الحرف إلى الصوت، وفي النهاية، الإنتاج الصوتي للنص المكتسب سابقًا. في النهاية، نقوم بتوضيح وشرح النتائج التي تم الحصول عليها من خلال تقرير التقييم بناءً على *MOS* (متوسط درجة الرأي: متوسط درجة الآراء) لحساب وضوح الكلام وطبيعته للنص المنطوق.

الكلمات المفتاحية: أنظمة تحويل النص إلى كلام، النسخ الهجائي الصوتي (TOP)، الذكاء الاصطناعي (AI)، اللغة العربية القياسية (AS) ، التعلم العميق (DL)، الشبكة العصبية التلافيفية (CNN) ، متوسط درجة الرأي (MOS).

Remerciement

Tout d'abord, nous remercions ALLAH tout-puissant de nous avoir guidé vers le bon chemin de la connaissance, de nous avoir donné la volonté, le courage et la confiance d'affronter toutes les difficultés afin de pouvoir continuer notre étude et arriver à ce point-là, et aussi de nous avoir permis de croiser de gens prêtes à nous aider sans demander de retour.

Nous tenons à remercier sincèrement et vivement notre promotrice Madame H.Tebbi pour avoir encadré et dirigé notre recherche et qui nous a toujours redonné le moral et la confiance si indispensables dans les périodes difficiles. nous la remercions pour la patience et la gentillesse dont elle a fait preuve à notre égard.

Nous remercions sincèrement les membres du jury pour avoir accepté d'évaluer notre mémoire.

Nous tenons à remercier Monsieur Boutebali.I et toutes les personnes qui nous ont aider et qui ont contribué au succès de notre travail.

Nous tenons aussi à remercier nos familles, nos amis et nos collègues qui ont pris une dimension toute particulière entrecoupant le travail, et aussi toutes les personnes qui nous aider pour de près ou de loin

Table des matières

Liste des tableaux

Liste des figures

Abréviations

ANN : ***Artificial Neural Networks.***

API : **Alphabet Phonétique Internationale.**

AS : **Arabe Standard.**

BDS : **Base de Donnée Sonore.**

CNN : ***Convolutional Neural Network.***

IA : **Intelligence Artificielle.**

IHM : ***Interaction Homme Machine.***

KNN : ***K-Nearest Neighbors.***

MOS : ***Mean Opinion Score.***

RNN : ***Recurrent Neural Network.***

TAL : **Traitement Automatique de la Langue.**

TALN : **Traitement Automatique de la Langue Natural.**

TAP : **Traitement Automatique de la Parole.**

TOP : **Transcription Orthographique Phonétique .**

TTS : ***Text to Speech.***

2D : **Deux Dimension .**

3D : **Trois Dimension.**

Introduction générale

La parole est un outil exceptionnel pour le transfert d'informations, c'est le mode de communication le plus naturel pour l'être humain. Ce dernier a été inspiré et a intégré ces interactions aux machines. Ceci est rendu possible grâce à la technologie vocale.

La technologie vocale comprend la synthèse vocale, la reconnaissance automatique de la parole et les systèmes de dialogue. Donner la possibilité de parler aux machines comme des êtres humains, c'est le défi de la synthèse vocale.

La synthèse vocale est une tâche très importante pour assurer une meilleure interaction entre l'utilisateur et la machine ; c'est la technologie qui permet l'automatisation de la production de la parole par une machine. On peut aussi dire que c'est le processus qui assure la transformation d'un message symbolique ou d'un ensemble de paramètres de commandes, en un message acoustique utilisé pour concevoir des machines parlantes. Le rôle de la synthèse vocale à partir d'un texte donné en entrée (ou Text to speech : TTS) est d'assurer la lecture de ce texte.

Les systèmes TTS représentent une des catégories de la grande classe des systèmes de synthèse vocale. L'objectif principal de ces systèmes TTS est de fournir à l'ordinateur la capacité de lire des textes à haute voix et visant à synthétiser des sons intelligibles et naturels qui ne se distinguent pas des enregistrements humains. Mais malgré les avancées réalisées ces dernières années dans ce domaine, des progrès restent à faire afin d'augmenter le confort d'utilisation des systèmes actuels. La synthèse vocale n'est pas une technologie nouvelle, en effet il existe plusieurs travaux dans la littérature concernant la synthèse vocale à partir de texte.

Néanmoins, les travaux de recherche menés ces dernières années ont abouti à une qualité de parole acceptable mais pas parfaite (on est loin d'une voix naturelle). Au cours de la dernière décennie, la synthèse vocale est devenue si naturelle que la compréhension de la parole synthétique exige un effort supplémentaire de la part du locuteur, par rapport au cas de la parole naturelle. Au départ, l'idée était d'utiliser des modèles physiques du conduit vocal à l'aide de synthétiseurs de formants, mais après de nombreuses années de recherche, l'informatique a permis d'utiliser directement des extraits de la parole enregistrée et de les coller ensemble pour créer de nouvelles phrases sous la forme d'une

synthèse par concaténation. Aujourd'hui, avec les progrès de la technologie, il est possible de générer des voix synthétisées de très bonne qualité grâce à des technologies d'intelligence artificielle telles que les réseaux neuronaux profonds.

L'impact du développement dans ce domaine est encore tardif pour la langue Arabe. Quand il y a plusieurs applications qui ne contiennent pas la langue Arabe complètement ou elle n'est pas claire et un peu loin du son naturel, donc il est nécessaire de travailler pour développer ce domaine en langue Arabe, Cet obstacle était la raison principale pour ce travail donc le but de ce projet est de développer une voix synthétique de haute qualité pour un texte en Arabe. La voix fournie par les machines peut être utilisée dans différents domaines dont celui de l'éducation, le son obtenu précédemment doit être compréhensible par le locuteur et plus proche de la voix humaine naturelle.

Ce mémoire est réalisé en deux parties, chacune d'entre elles comprend deux chapitres, dans la première partie nous aborderons un état de l'art du traitement automatique de la parole et de la synthèse vocale tandis que dans la deuxième partie nous décrivant notre aspect pratique où on va détailler l'approche proposée et les tests obtenus.

Nous commencerons dans le premier chapitre qui est sous le titre de la parole et son traitement automatique une vue générale sur celle-ci où nous délimitons les notions les plus utilisées on Arabe Standard (AS) où nous abordons sa phonologie, son écriture ainsi que ses règles et propriétés phonologiques afin d'extraire les informations nécessaires dont nous avons besoin pour implémenter notre système. Le deuxième chapitre est consacré à la synthèse vocale et à l'application de l'apprentissage automatique à ce domaine, où nous présentons les techniques de l'intelligence artificielle les plus utilisées et spécialement les réseaux de neurone pour générer automatiquement des signes vocaux à partir de phrases écrites.

Après une brève présentation du sujet, la deuxième partie décrit le travail que nous avons effectué pour créer un texte synthétisé en Arabe. Le premier chapitre de cette deuxième partie traite l'étude conceptuelle de notre système, où nous expliquons notre choix technologique tout en présentant l'architecture générale du système et les différentes tâches que nous devons mettre en œuvre. Le dernier chapitre résume l'évaluation des

résultats obtenus par un test de la voix créée. Différentes approches ont participé afin de mesurer la qualité du son obtenu.

Nous clôturons notre travaille par une conclusion générale et quelques perspectives.

Chapitre 1 :la parole et le traitement automatique de la langue Arabe

1.1 Introduction

La parole est un outil exceptionnel pour le transfert d'informations, c'est le premier moyen de communication entre les êtres humain. De nos jours l'interaction homme machine devient de plus en plus indispensable dans notre vie quotidienne cette interaction se base sur l'utilisation de la parole ; actuellement le but des recherches est de rendre cette communication vocale la plus proche d'une parole naturelle.

Dans ce chapitre nous allons parler de manière générale sur le Traitement Automatique de la Parole (TAP), le mécanisme de la production du signal vocal, les caractéristiques de signale vocal, nous clôturons le chapitre par évoquer les particularités de l'Arabe Standard (AS).

1.2 Définition du Traitement Automatique des Langages

Le Traitement Automatique des Langues (TAL) est un ancien domaine, qui est apparu dès les débuts de l'informatique, le TAL vient à l'interaction de plusieurs domaines informatique, mathématique et linguistique. Le Traitement Automatique du Langage Naturel (TALN) est ainsi un champ de savoir et de techniques élaborées autour de problématiques diverses. Les concepts et techniques qu'il utilise se trouvent à la croisée de multiples champs disciplinaires : l'Intelligence Artificielle l'IA « traditionnelle », l'informatique théorique, la logique, la linguistique, les neurosciences, les statistiques, etc.

Le TAL est basé sur l'application des programmes et des techniques informatiques à la langage humaine dans divers activités liées au codage vocale, l'analyse de la parole, la synthèse vocale, et à la reconnaissance de la parole. Ces différents traitements peuvent se faire dans un contexte bruyant, ce qui rend le problème du TALN plus difficile [1] .

1.3 Production de la parole

Lors de la production de la parole, l'arrivé de souffle produit par les paumons qui remonte dans Trachée-artère jusqu'au larynx va générer une vibration, Lorsque la pression d'air s'accumule sous les cordes vocales, elles sont forcées de s'ouvrir partiellement, leur tension naturelle les amène ensuite à se refermer. Ce sont le débit du

flot d'air et le degré d'ouverture des 9 cordes vocales qui conditionnent l'intensité de l'onde ainsi produite. L'espace entre les cordes vocales s'appelle la glotte. [2]

Il existe deux types de son : les sons voisés qui sont les résultats de la vibration des cordes vocale, et les sons non voisées (sourds) qui sont articulée sans de telles vibration.

1.4 Caractéristiques du signal vocal

Le signal vocal est un signal périodique, il est caractérisé par :

- La fréquence fondamentale f_0 :

  La fréquence fondamentale détermine la hauteur de son en Hertz (Hz), elle correspond à la fréquence la plus base causée par la vitesse à laquelle les cordes vocales s'ouvrent et se referment durant la production du son. Cette fréquence est quasi stationnaire pour un signal de type voisé, elle varie entre [2] :

  - 80 et 200 Hz pour une voix masculine
  - 250 à 450Hz pour une voix féminine
  - Plus que 200 à 600Hz pour une voix d'enfant

  La figure 1.1 illustre que chaque son voisé correspond à une présence de fréquence c.à.d. correspond à une présence de pitch et que les sons non voisées (sourdes) sont associés à une fréquence nulle.

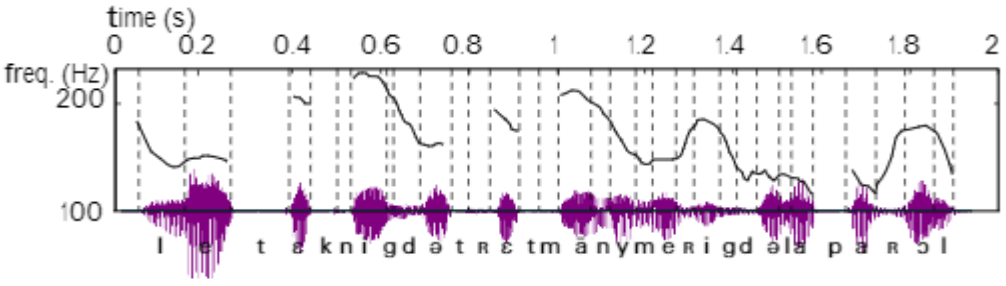

Figure 1.1: Evaluation de la fréquence de vibration des cordes vocales. *[3]*

La figure 1.1 représente l'évaluation de la fréquence de vibration des cordes vocales. dans la phrase « les techniques de traitement numérique de

la parole ». La fréquence est donnée sur une échelle logarithmique ; les sons non-voisés sont associés à une fréquence nulle. [3]

- L'Energie :

Ce paramètre caractérise l'intensité du signal vocal, il exprime le volume sonore d'un phonème(le phonème représente la plus petite unitée du son d'une langue), en d'autres termes, il représente l'amplitude des vibrations des cordes vocales, il est mesuré en décibels (Db).

- La durée :

La durée représente le temps de prononciation d'un phonème dans une phrase, pour mesurer n'importe quelle durée il serait nécessaire de spécifier les unités que nous voulons analyser avec ces frontières dans le signal vocal. Ces unités peuvent être des voyelles, des pauses, des phonèmes, etc.

1.5 Alphabet Phonétique International

Pour convertir les graphèmes (Lettre ou groupe de lettres transcrivant un phonème, par exemple [o] = o, au, eau (3 graphèmes distincts pour le même phonème)) en phonèmes, nous nous basons sur un processus de Transcription Orthographique Phonétique (TOP). L'alphabet normale ne convient pas à ce processus, puisqu'une seule lettre peut correspondre à plusieurs sons, (par exemple, pour la langue française, la prononciation du [c] dans /seconde/ et /cocher/ n'est pas la même), il est donc nécessaire d'utiliser l'Alphabet Phonétique International. Cette dernière permet d'associer à chaque symbole un son qui lui correspond.

L' Alphabet Phonétique International (API) a été développé par des phonéticiens français et britanniques sous les auspices de l'Association phonétique internationale, l'A.P.I. a été publié pour la première fois en 1888. son objectif est de transcrire les phonèmes afin de noter que ce qui est prononcé.

Il faut savoir que plusieurs niveaux doivent être pris en compte lors du passage phonétique orthographique, parmi ces niveaux nous pouvons citer ; le niveau phonétique et phonologique, le niveau lexical, le niveau syntaxique et même sémantique [4]

1.6 Langue Arabe

L'Arabe est l'une des langues sémitiques et parmi les langues les plus ancienne du monde. Elle s'écrit de droite à gauche avec un sens du tracé respecté (voir figure 1.2) et la construction d'un mot se fait en collant les lettres une à une, de droite à gauche. Les notions de lettre majuscule et minuscule n'existent pas (l'écriture est donc monocamérale). [4]

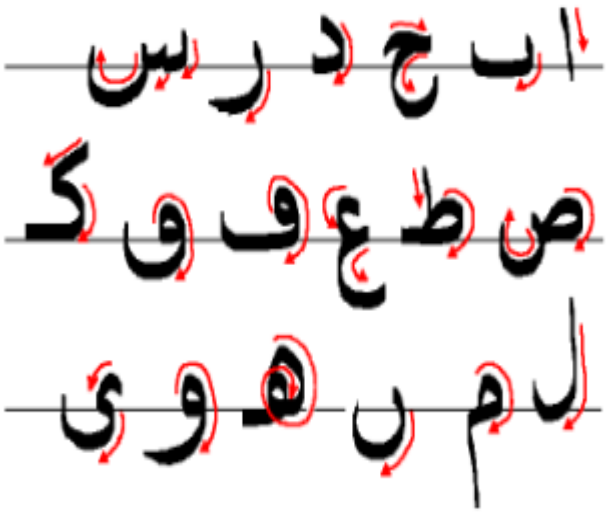

Figure 1.2:Sens d'écriture des lettres Arabe . *[4]***.**

L'Arabe comporte 29 lettres dont la forme change en fonction de la position dans le mot (voir tableau 1.1), ces lettres sont devisées en 26 consonnes et 6 voyelles (3 voyelles court et 3 voyelles langues).

Lettre	Prononciation		Phonème (API)
ء	Hamza	*همزة*	/ʔ/
ا	Alif	*ألف*	/a:/
ب	Ba	*باء*	/ b/
ت	Ta	*تاء*	/t/
ث	Tha	*ثاء*	/ θ/
ج	Jim	*جيم*	/ʒ/,/ʤ/,/g/
ح	Ha	*حاء*	/ħ/
خ	Kha	*خاء*	/x/
د	Dal	*دال*	/d/
ذ	Dhal	*ذال*	/ð/
ر	Ra	*راء*	/r/
ز	Zay	*زين*	/z/
س	Sin	*سين*	/s/
ش	Shin	*شتن*	/ʃ/
ص	Sad	*صاد*	/sˤ/
ض	Dad	*ضاد*	/dˤ/ ,/ ðˤ/
ط	Ta	*طاء*	/tˤ/
ظ	Za	*ظاء*	/zˤ/ , /ðˤ/
ع	Ayn	*عين*	/ʔˤ/
غ	Rhayn	*غين*	/ɣ/
ف	Fa	*فاء*	/f/
ق	Qaf	*قاف*	/q/
ك	Kaf	*كاف*	/k/
ل	Lam	*لام*	/l/
م	Mim	*ميم*	/m/
ن	Nun	*نون*	/n/
ه	ha	*هاء*	/h/
و	Waw	*واو*	/w /, /u:/
ي	Ya	*ياء*	/j/ , /i:/

Tableau 1.1 Lettres Arabe. *[5].*

1.6.1 Consonnes " الحروف الساكنة "

Tous les lettres de l'alphabet Arabe sont des consonnes, ces derniers s'écrivent différemment selon qu'elles sont isolées, au début, au milieu à l'intérieure de mot, ou à la fin du mot . [6]

Voici un exemple ci-dessous(tableau 1.2)

[ح]A la fin du mot		[ح] Au milieu de mot		[ح] Au début du mot	
ملح	ـح	محلل	ـحـ	حبر	حـ

Tableau 1.2: Lettre /ح/ dans les différentes positions : au début ; au milieu et à la fin du mot.

Les consonnes Arabes sont divisées en deux groupes, les consonnes solaires et les consonnes lunaires (tableau 1.3).

الحروف الشمسية	التاء	الثاء	الدال	الذال	الراء	الزاي	السين	الشين	الصاد	الضاد	الطاء	الظاء	اللام	النون
Les consonnes solaires	/t/	/ θ/	/d/	/ð/	/r/	/z/	/s/	/ʃ/	/sˤ/	/dˤ/	/tˤ/	/zˤ/	/l/	/n/
الحروف القمرية	الالف	الباء	الجيم	الحاء	الخاء	العين	الغين	فاء	قاف	الكاف	الميم	الهاء	الواو	الياء
Les consonnes lunaire	/a:/	/b/	/ʒ/	/ħ/	/x/	/ʔˤ/	/ɣ/	/f/	/q/	/k/	/m/	/h/	/w/	/j/

Tableau 1.3:Consonnes lunaires et solaires de l'alphabet Arabe.

- **Consonnes solaires " *الحروف الشمسية* "**

Elles sont limitées à 14 consonnes (voir tableau 1.3), où les linguistes attribuent la raison du nom à l'absence de prononciation de la définition ([L] : /ل/) dans le mot, c'est-à-dire lors de la prononciation on élimine le son qui correspond à cet lettre ([L] : /ل/), par exemple le mot الشمس qui correspond au soleil en français, quand -on articule ce mot en dit ([aSSamsu] : *أشَّمس*) et pas ([aLSSamsu] : *أشَّمس*).

- **Consonnes lunaires " *الحروف القمرية* "**

Elles sont limitées aussi à 14 consonnes (tableau 1.4) pendant la prononciation de ces consonnes on prononce le lam [ل: L] au contraire des consonnes solaires.

Exemple : le mot ٱلْقَمَرُ sera prononcé [alqamaru] qui signifie la lune

Il existe une autre classification des sons qui se base sur : le lieu d'articulation, le mode d'articulation et le voisement.

a) Lieu d'articulation : lieu d'articulation ou le point d'articulation est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. De manière générale, on peut dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air. [7]

- **Les consonnes bilabiales :** sont prononcées avec les deux lèvres.

  /b / /m/ /w/.

  Exemples : / b / *بسيط* [basˁi: tˁ], /w/ *ورق* [waraq], / m / *ممحاة* [mimħa:t].

- **La consonne glottale :** elles sont produites au niveau du glotte (/h/: ه) (/q /: ق /).

  Exemples : / h/ *هشام* [hiʃa:m], /q/ *إشراق* [iʃra:q]

- **Les consonnes dentales :** sont les consonnes lors de sa production en utiliser les dents supérieure et inférieure avec la langue. / ð/ذ , /ðˤ / ظ /θ/ ث.

  Exemples : /ð/ ذبابةٌ [ðuba:batun], / ðˤ / ظبي [ðˤabj], /θ / ثمن [θaman]

- **Les consonnes labio-dentales :** elles sont produites avec la lèvre inférieure et les dents supérieures, il y'a une seule consonne de ce genre c'est le /f/

  Exemple : /f/ فلفل [fulful]

- **Les consonnes pharyngales :** lors de sa production en utilise le pharynx (la gorge moyenne) /ʔˤ /ع,/ħ/ ح

  Exemples : /ʔˤ /عسل [ʔˤasal], /ħ/ حمامة [ħama:ma]

- **Les consonnes uvulaires :** ce quand appelle [حروف حلقية] il existe une seule consonne c'est le /q/ق

  Exemple : /q/ قافلة [qa:fila]

- **Les consonnes vélaires :** elles sont produites dans la partie antérieure de la gorge, il y'a trois consonnes / k/ ك,/x/ خ ,/ ɣ/ غ

  Exemples : /k/ : كراس [kura:s], /x/ : خروف [xaru:f], /ɣ/:غرفة [ɣurfa]

- **Les consonnes palatales :** elles sont produites lorsque la partie antérieure de de la langue s'élève vers le palais. /j/

  Exemple :/ j/ : يوسف [ju:suf]

- **Les consonnes alvéolaires :** sont les consonnes produites lorsque la langue est en contact avec la crête alvéolaire /s/س, / dˤ/ض, /sˤ/ص, /t/ت, /tˤ/, /d/د, /n/ن, /r/ر, /z/ز, /l/ل

Exemples : /s/ : *سيارة* [saja:ra], / dˤ / : *ضابط* [dˤa:bit], /sˤ/ : *صديق* [sˤadi:q], /t/ : *تنين* [tini:n], /tˤ/ : *طائر* [tˤa:ir], /d/ : *درج* [daraʤ], /n/ : *نعمان* [nuʔˤma:n], /r/ : *رمز* [ramz], /z/ : *زكام* [zuka:m], / *l*/ : *ليف* [lajf]

- **Les consonnes alvéolopalatales :** elles produisent lorsque la partie antérieure de la langue touche la crête alvéolaire puis le palis dur.il existe deux consonnes alvéopalatales:/ ʤ /ج,/ ʃ /ش

  Exemple : / ʤ / : *جميل* [ʤami:l],/ ʃ /:*شاب* [ʃa:bab]

b) Mode d'articulation : ce paramètre fait référence au type de constriction ou de mouvement qui se produit à n'importe quel endroit de l'articulation, comme le degré marqué de rétrécissement, ou une fermeture avec un relâchement lent . [8], selon ce mode on trouve les classes des sons suivantes :

- Les consonnes fricatives : on appelle les consonnes fricatives ou constrictive les consonnes dont l'articulation comprend une obstruction partielle du flux d'air ou ce passage d'air provoque des frottements.(/f/ف, /s/س, /z/ز, /ʃ/ش, / *ð* /ذ, / *θ* /ث , /ʤ/ج, /x/خ , /ħ/ح)

  Exemples : *فلة /f/ [*fula*]*, ثعلب /*θ*/ [*θ*aʔˤlab], سيارة/s/ [saja:ra], جمل/ʤ/[ʤamal], شجرة/ʃ/[ʃaʤara], زرع /z/[zarʔˤ]

- Les consonnes occlusives : sont les consonnes pour lesquelles le passage d'air est bloqué par une fermeture momentanée de conduit vocale, on les appelle aussi les consonnes plosives,(/q/ق,/d/د, /k/ ك,/ʤ/ج,/ b/ب)

  Exemples : */q/ قافلة [*qa:fila*], /d/ دلو [dalw]* , /ʤ/جمل [ʤamal], /k/كهف [kahf], / b/باب [ba:b].

- Les consonnes affriquées : elles sont caractérisées par une phase occlusive ou on a un blocage d'air suivie par une étape fricative ou le flux d'air est relâché pour passer.

  Exemples : جَميل /dZ/ [dZami:l]

- Les consonnes liquides : lors de production de ces sons, il y'a une combinaison d'une occlusion et une ouverture de chenal buccal c'est-à-dire il y'a une certaine obstruction du flux d'air dans la bouche.*(/l/ل,/r/ر)*

  Exemples : */l/ ليل*[lajl], */r/ رمل [*raml]

- Les consonnes vibrantes : sont les consonnes dont sont production on a une vibration entre la partie avant de la langue et la crête alvéolaire (*/r/ر*)

  Exemples : */r/ رمل [*raml]

- Les consonnes nasales : elles ont produit en abaissant le voile du palais, (*/n/ن , /m/م).*

  Exemples : */m/ منزل* [manzil] , */n/ ندى [nada]*

- Les consonnes glides : sont les semi-consonnes ou semi-voyelles, son des sons qui se trouvent à mi-chemin entre une voyelle et une consonne, elles sont produites avec ou peu ou pas d'obstruction de l'aire dans la bouche.(/w/و,/y/ي).

  Exemples : /w/*وردة*[warda], /j/*يوم*[jawm]

c) Voisement : c'est la vibration des cordes vocales lors de la production d'un son, ce qui génère des sons voisés ou sonores.

MA \ LA		Bilabial	Alvéolaire	Vélaire	Uvulaire	Glottale	Dentale	Pharyngale	Labio-dentale	Alvéolopalatale	Palatale
Occlusive	**V**	ب	ض د								
	NV		ط ت	ك	ق	ء					
Fricative	**V**		ز	غ			ظ ذ	ع			
	NV		ص س	خ		ه	ث	ح	ف	ش	
Affriqué	**V**									ج	
Nasale	**V**	م	ن								
Vibrantes	**V**		ر								
Liquide	**V**		ل								
Glides	**V**	و									ي

Tableau 1.4: Caractéristiques phonétiques des consonnes Arabes.

1.6.2 Voyelles

Contribuent à déterminer la prononciation du mot, les voyelles émergent de la gorge et des lèvres, ces lettres sont caractérisées par un écoulement relativement libre de l'aire dans le conduit vocal. Les voyelles sont nécessaires à la lecture et la compréhension des textes. Dans la langue Arabe il existe six voyelles parmi lesquelles trois sont courtes et trois sont langues.

Le tableau au-dessous (tableau 1.5) représente les voyelles courtes et le (tableau 1.6) représente les voyelles langues ce qu'on appelle en Arabe « al Mad », ce dernier c'est un allongement des voyelles courtes.

Signe	Nom	Exemple	T.O. P
◌َ	Le fatha (*الفتحة*)	حَ	/ a /
◌ُ	La damma (*الضمة*)	حُ	/ u /
◌ِ	Le kasra (*الكسرة*)	حِ	/ i /

Tableau 1.5: Voyelles courtes.

Nom	Exemple	T.O. P
Mad bil Alif	امام	/ a : /
Mad bil Waw	مؤمنون	/ u : /
Mad bil Ya	مستوطنين	/ i : /

Tableau 1.6:Voyelles longues.

1.6.3 Signe diacritique Sukun "*السكون*"

Il s'agit d'un petit cercle placé au-dessus de la lettre. Le Sukun (figure1-3)est le contraire du mouvement, il indique une consonne non suivie d'une voyelle, c'est-à-dire

qu'il s'agit de l'absence de mouvement pendant la prononciation de la lettre. Exemple la lettre (كْ) dans le mot « مَكْتَب» /maktab/

Figure 1.3: Signe diacritique Sukun "السكون"

1.6.4 Gémination " *الشّدة*"

Le signe de la gémination est une accentuation qui est placée au-dessus des lettres qui sont des lettres doivent être doublées (deux consonnes doublées) où la première consonne est muette c'est-à-dire *حرف ساكن* et la seconde liée à une voyelle, la durée de la prononciation est le double de cette consonnes singleton, certaines théories de phonologie utilise le mot double et allongement comme synonymes de gémination.

Figure 1.4:Signe de la gémination " الشّدة "

1.6.5 Diacritiques doubles "*التنوين*"

Le 'tanween' dans la langue Arabe est la rencontre de deux voyelles similaires à la fin des noms. Il est également connu sous le [nun] /ن/ de nom complémentaire. Il se prononce et non s'écrit, et cela se fait en doublant la lettre que l'on veut destiner pour aboutir à trois types de 'Tanween'(tableau1.7)

	Tanween el fath	Tanwin dhamah	Tanween kasrah
Exemple	تاً	تٌ	تٍ
Transcription	/an/	/un/	/ in /

Tableau 1.7: Diacritiques doubles.

1.7 Problème de diacritiques

Le plus gros problème confronter en Arabe dans le domaine de synthèse vocale est celui de la vocalisation. La lecture d'un texte sans vocalisation (sans signes diacritiques) peut entraîner un changement complet du sens du mot.

Exemples :

أَكَلَ: à manger أُكِلَ: a été mangé

Donc on a deux mots avec le même schème mais avec un sens différent.

Pour résoudre ce problème, les systèmes TTS *(text to speech)* pour la langue Arabe utilisent l'application "Mishkal[1]" pour vocaliser les données du texte avant de les synthétiser.

1.8 Conclusion

Dans ce chapitre nous avons présenté d'une manière générale les notions de base de traitement automatique de la parole, son fonctionnement et ses caractéristiques, nous nous sommes concentrés aussi dans ce chapitre sur les caractéristiques phonologiques et phonétiques de la langue Arabe et nous terminons en abordant le problème des diacritiques.

Le chapitre suivant traitera plus particulièrement la synthèse vocale, son processus, ses méthodes, l'application de l'apprentissage automatique sur la synthèse vocale et ces applications dans différents domaines, nous citerons aussi certains exemples des systèmes TTS *(text to speech)* avec ces avantages et ces inconvénients.

[1] https://tahadz.com/mishkal/

Chapitre 2 : La synthèse vocale et l’apprentissage automatique

2.1 Introduction

Dans ce chapitre nous nous intéressons à la synthèse vocale plus particulièrement la synthèse vocale à partir d'un texte ou *text to speech* (TTS) qui est le résultat des diverses compétences en informatique, linguistique et en traitement de signal.

Dans cette partie on va détailler le fonctionnement des systèmes TTS, les méthodes utilisées pour générer des signaux vocaux à partir de n'importe quel texte en utilisant ces systèmes, et aussi l'application de l'apprentissage automatique sur la synthèse vocale et on continuera ce chapitre par quelques exemples des applications et des systèmes TTS les plus utilisés.

2.2 Définition de la synthèse vocale

La synthèse de la parole est l'ensemble des dispositifs, matériels ou algorithmes, pour générer automatiquement de la parole artificielle. C'est la technologie qui permet d'automatiser la production d'une parole artificielle par une machine. On peut dire aussi que c'est le processus qui assure la transformation d'un message symbolique ou un ensemble de paramètres de commandes, en un message acoustique utilisé pour concevoir des machines parlantes. [1]

2.3 Les méthodes de synthèse vocale

Cette partie représente les différentes méthodes utilisées dans la synthèse vocale. Deux approches sont proposées : les méthodes par connaissance explicite et les méthodes par connaissance implicites, Les premières utilisent des modèles de l'appareil vocal, qu'ils soient de type physique comme la synthèse dite articulatoire, ou de type signal comme la synthèse dite par formants. La deuxième n'utilise pas de modèle de l'appareil vocal mais manipule des segments de voix préenregistrés, c'est la méthode dite par concaténation. [9]

2.3.1 Synthèse vocale articulatoire

La synthèse articulatoire tente de modéliser le plus parfaitement possible les organes vocaux humains, on peut dire que cette méthode de synthèse descend de la fameuse machine parlante de Von Kempelen(figure2.1) pour produire une parole synthétique de haute qualité. D'autre part, c'est aussi l'une des méthodes les plus difficiles à mettre en œuvre, les processus de calcul étant tellement lourds au vu des résultats peu satisfaisants obtenus. [10]

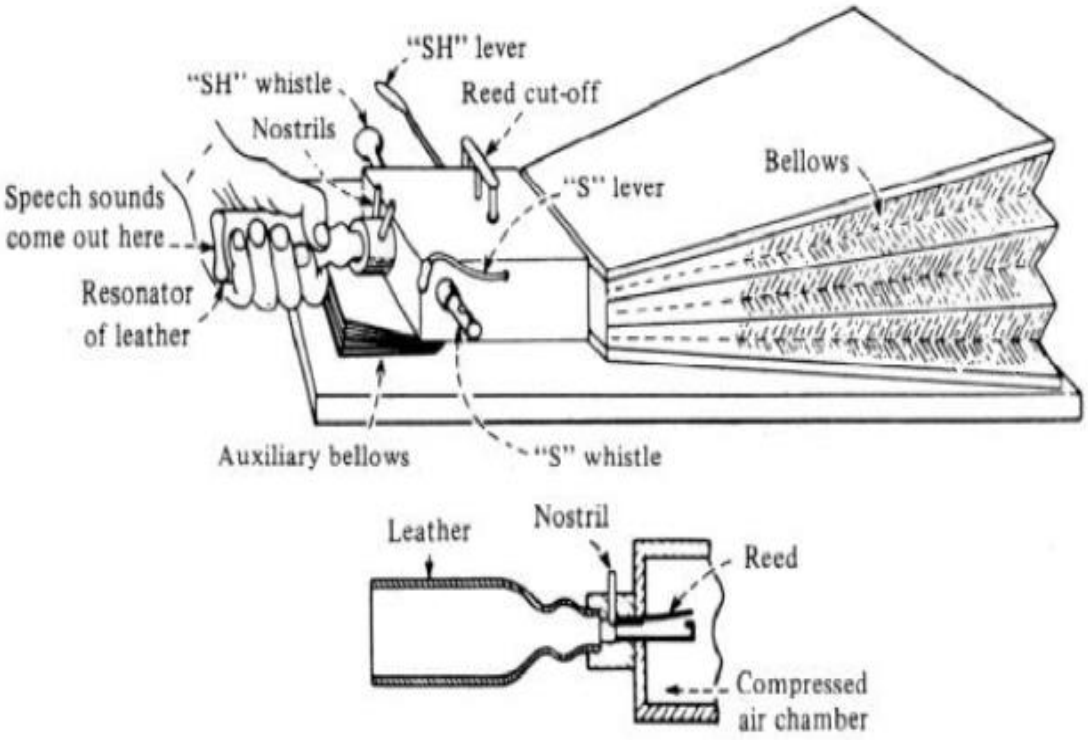

Figure 2.1: Machine parlante de von Kempelen. *[13]***.**

Les données du modèle articulatoire dérivent de l'analyse par rayons X de la parole naturelle. Cependant, ces données ne sont généralement que sur 2D, alors que le conduit vocal réel est naturellement à trois dimensions. En raison de ce manque de données sur les mouvements des articulateurs pendant la parole. Par conséquent, cette approche reçue moins d'attention et n'a pas encore atteint le même niveau de succès que d'autres méthodes de synthèse. [11]

La synthèse articulatoire repose sur une évolution dynamique des articulateurs sollicités au cours du processus phonatoire. Le système s'appuie sur la modélisation explicite du mécanisme humain de production de la parole. [12]

2.3.2 Synthèse vocale par formant

La synthèse par règles ou la synthèse par formants est basée sur un modèle linéaire source-filtre de production qui utilise des paramètres acoustiques comme entrées du synthétiseur. [9]

Cette méthode de synthèse consiste à modéliser le signale vocale par ensemble de filtres en parallèles ou en cascade. Cette technique est rarement utilisée aujourd'hui à cause des sons robotiques et non naturels qu’elle génère.

2.3.3 Synthèse vocale par concaténation

Relier des énoncés naturels préenregistrés est probablement le moyen le plus simple de produire une parole synthétique intelligible et naturelle. La synthèse par concaténation d’unités pré-stockées est la génération des sons à partir de la juxtaposition d’un ensemble d’unités préenregistrées, ces dernières sont obtenues par une opération de segmentation du signal qu’on veut produire. En réalité dans cette approche on peut trouver plusieurs types d’unités (phonèmes, diphones, syllabes, polysons, mots, etc.) . [12]

La voix obtenue en résultats par cette approche est très proche de celle générée par un être humain, alors elle est intelligible et naturelle, pour cette raison, ce type de synthèse a été adopté par un grand nombre des systèmes TTS. [12]

2.4 Définition de la synthèse vocale à partir du texte

La synthèse vocale à partir du texte est une technique informatique qui permet de générer automatiquement une parole artificielle à partir du texte en entré. On peut la définie aussi par la transformation d’un texte écrit à en un texte lu.

Littéralement cette synthèse vocale, est une technologie permettant de vocaliser n'importe quelle donnée à la seule condition qu'elle soit de nature textuelle. On passe donc du texte à un signal de parole. [13]

Nous présentons dans le point suivant les principaux modules qui compose le processus de la synthèse vocale à partir du texte.

2.5 Processus de synthèse vocale à partir du texte

Le processus de n'importe quelle synthèse vocale à partir du texte se compose de deux modules, l'analyse et la génération :

- L'analyse: c'est un module de traitement de langage consiste à analyser les données d'entrées pour les convertir en une séquence phonétique.
- La génération c'est le module de traitement acoustique (synthétisation), il s'agit de vocaliser le texte donc transforme les unités phonétiques engendrées par le précédent module en signal vocale

Une version simplifiée de ce processus est illustrée à la figure suivant (figure 2.2).

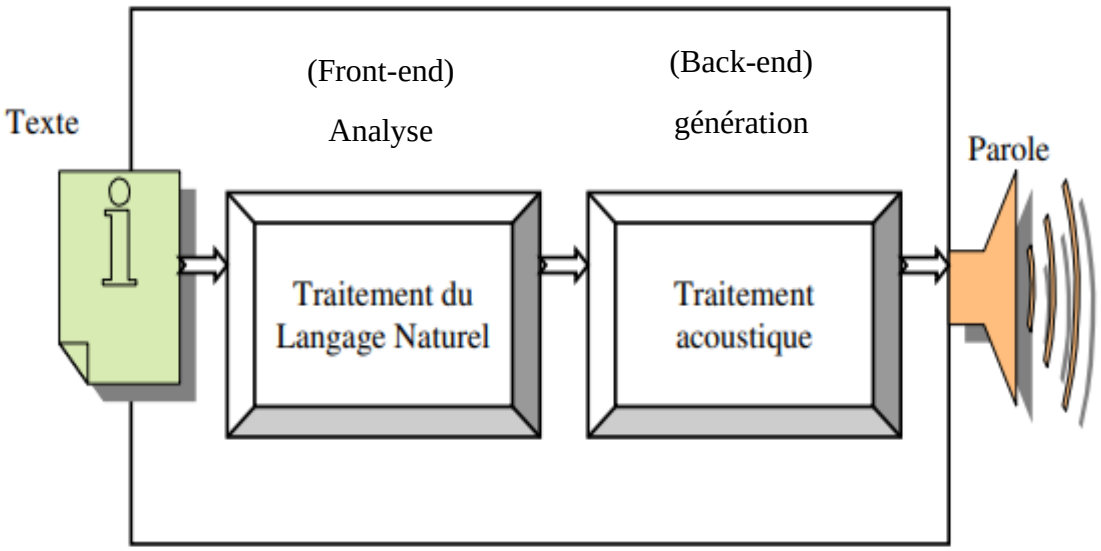

Figure 2.2:Processus de la synthèse vocale. *[14]***.**

L'entrée de processus comprend des données textuelles (e-mail, des articles, etc..), ces derniers passent par plusieurs étapes pour produit son transcription phonétique afin qu'elles puissent être convertir en parole.

2.5.1 Traitement de la langue naturel (*Front-end)*

Il contient un analyseur de texte, un générateur de prosodie (pour le traitement de la prosodie) et un module de phonétisation (pour la détermination de la transcription phonétique). [15]

a) L'analyseur de texte est composé :

- D'un module de normalisation ou de nettoyage de texte ce qu'on appelle pre-prossesing en anglais, ce dernier permet d'enlever toutes les ambigüités qui concerne les mots segmentés pour donner à la fin un texte vocalisée indiquant comment les données en entrées doivent être prononcés (les chiffres, des mots, des caractères spéciaux, des dates, heurs). Exemple « 10 » pourrait être dix, « Dr » Docteur, etc.
- D'un analyseur morphologique qui s'emploie de proposer la prononciation correcte des mots. [16]
- D'un analyseur contextuel qui sert a transformée le texte d'entrée en une suite des unités lexicales, cette phase a le but de considérer chaque mot dans son contexte afin de réduire la liste des diverses catégories possibles d'un mot.
- D'un analyseur syntaxique : l'analyse lexicale n'est pas suffisante pour lever toutes les ambigüités pour cela une analyse syntaxique s'effectué pour faciliter l'accentuation des phrases et gère les ambigüités à travers des règles grammaticales afin d'attribuer à chaque terme une étiquette morpho-lexicale.

b) Module de phonétisation : appelé aussi module **G2P** (Graphème-to-Phonème), cette étape prend en entrée les résultats de l'analyse précédente, dans le cadre de ce processus, il effectue une conversation graphème à phonème (c'est-à-dire la prononciation exacte de chaque mot de la phrase d'entré, [16]avec un dictionnaire de tous le mots (mots, abréviation, chiffre, etc) pré-étiqueté avec leurs transcriptions orthographiques phonétiques. Dans le cas d'un mot qui ne figure pas dans la lexie, les règles LTS (lettre to sounds) peuvent être utilisée, ces règles fournissent les prononciations possibles pour chaque lettre ou mots.

c) L'analyse prosodique : c'est la dernière étape dans le processus Front-end, cette analyse permet de donner à la lecture un rythme naturel et intelligibilité à travers un énoncé phonétique en créant des intonations, des modifications de la fréquence fondamentale, de la durée et de l'amplitude.

2.5.2 La synthétisation (*Back-end*)

L'entrée de synthétiseur vocale est le résultat de l'analyse prosodique, il s'agit de vocaliser le texte, on produit donc un signal vocal à partir de la séquence phonétique engendrée par les traitements linguistiques. [13]

Le backend utilise les informations fournis par le front end pour synthétiser le texte en utilisent des méthodes spécifiques, nous les détaillons dans les sections suivantes.

2.6 La synthèse vocale et l'apprentissage automatique (Machine Lerning)

2.6.1 l'apprentissage automatique (Machine Lerning)

Machine Learning ou bien apprentissage automatique est une sous-catégorie de l'intelligence artificielle (Figure2.3), qui donne à une machine la capacité d'apprendre automatiquement de données et d'expériences passées sans la programmer de façon explicite.

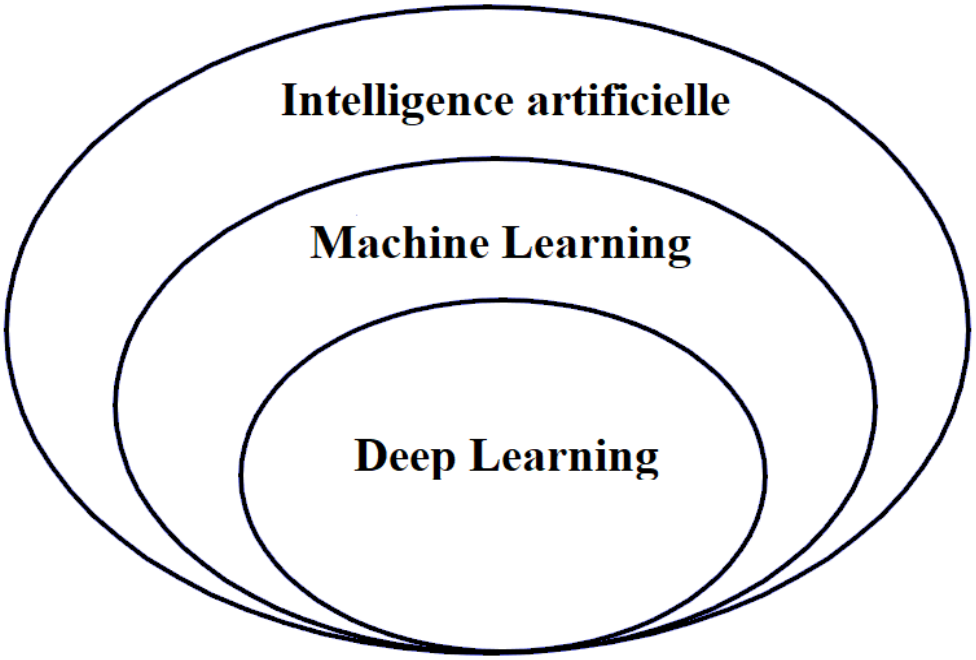

Figure 2.3: Sous-ensembles d'IA. *[17].*

2.6.2 Méthodes d'apprentissage automatique

Pour donner à un ordinateur la capacité d'apprendre, on utilise des méthodes d'apprentissage qui sont fortement inspirées de la façon dont nous, les êtres humains, apprenons à faire des choses. Parmi ces méthodes, on compte : L'apprentissage supervisé (*Supervised Learning*), l'apprentissage non supervisé (*Unsupervised Learning*) et l'apprentissage par renforcement (*Reinforcement Learning*). [18]

- **L'apprentissage supervisé (*Supervised Learning*)**

Les données utilisées pour l'entraînement sont déjà 'étiquetées'. Par conséquent, le modèle de Machine Learning sait déjà ce qu'elle doit chercher (motif, élément,etc.) dans ces données. La fin de l'apprentissage, le modèle ainsi entraîné sera capable de retrouver les mêmes éléments sur des données non étiquetées. [19]

Il existe deux types de problèmes en apprentissage supervisé : les problèmes de régression et les problèmes de classification.

Voici quelques exemples populaires d'algorithmes d'apprentissage automatique supervisé : Foret aléatoire, KNN[2], Arbre de décision, Régression logistique, Régression. [17], Réseau de neurone, etc.

- **L'apprentissage non supervisé (*Unsupervised Learning*)**

Consiste à entraîner le modèle sur des données sans étiquettes. La machine parcourt les données sans aucun indice, et tente d'y découvrir des motifs ou des tendances récurrentes. [19]

- **L'apprentissage renforcement (*Reinforcement Learning*)**

L'apprentissage par renforcement est une technique très populaire qui consiste à laisser la machine apprendre à faire une tâche en la laissant pratiquer seule. [20] Cette méthode permet de trouver par un processus d'essais et d'erreurs, l'action optimale à effectuer pour chacune des situations « états » que la machine va percevoir afin de maximiser ses récompenses. Elle est aussi non supervisée car la récompense ne donne pas l'action optimale à réaliser mais simplement une évaluation de la qualité de l'action choisie. [21]

[2]: K plus porches voisins (k-Nearest Neighbor).

2.6.3 Réseaux des neurones artificiels

2.6.3.1 Définitions

Les réseaux de neurones artificiels sont des techniques d'apprentissage automatique populaires qui simulent le mécanisme d'apprentissage dans les organismes biologiques. [22]

Le réseau de neurones artificiel *(Artificial Neural Networks ANN)* (figure 2.4) est un concept vu le jour en1943 a été inventé par 'Warren *McCullough, neurophysicien*', et le mathématicien '*Walter Pitts'*, a été inspiré par les neurones biologiques (figure 2.5). Il est constitué d'un très grand nombre de petites unités identiques de traitement appelées neurones artificiels. Qui agissent comme des chemins pour transmettre des données.

Les ANN sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau. [23]

Les Réseaux de Neurones sont des modèles bien plus complexes que tous les autres modèles de Machine Learning dans le sens où ils représentent des fonctions mathématiques avec des millions de coefficients (les paramètres). [18]

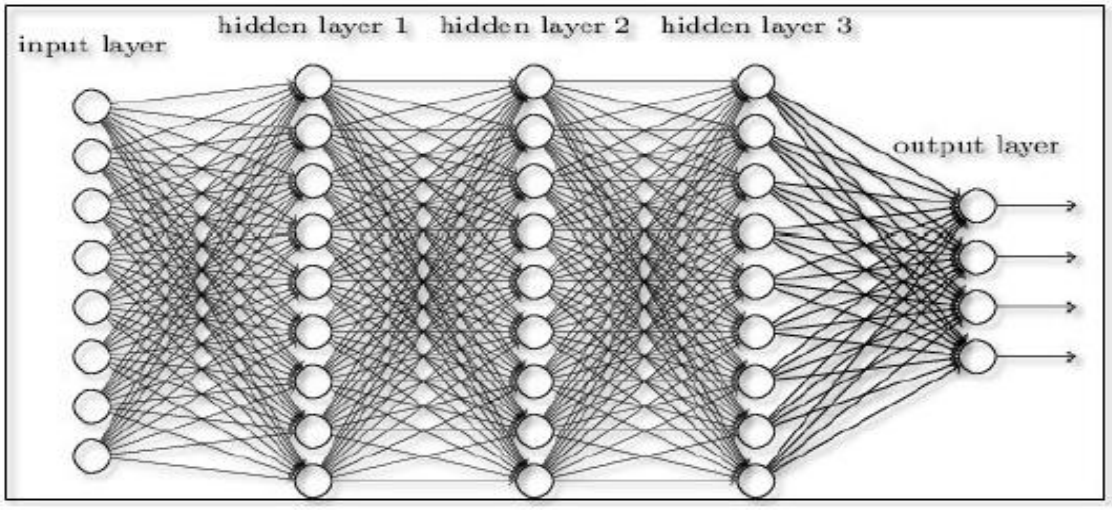

Figure 2.4: Représentation du réseau de neurones artificiel. *[18]*

Les petits nœuds sont appelés les neurones et chaque nœud de la couche cachée est une combinais de toutes les entrées. La couche cachée agit comme « entrée » pour la couche de sortie. Et chaque flèche représente un poids (nombre flottant) qui indique dans quelle mesure chaque entrée contribue à chaque étape suivante.

Figure 2.5: Représentation du réseau de neurones humain. *[19]*

2.6.3.2 Fonctionnent des réseaux de neurones

Un réseau de neurones repose sur un grand nombre de processeurs opérant en parallèle et organisés en tiers. Le premier tiers reçoit les entrées (*input layer*) d'informations brutes, Par la suite, chaque tiers reçoit les sorties (*output layer*) d'informations du tiers précédent**.** Le dernier tiers, quant à lui, produit les résultats du système, et plus le problème est complexe, plus il faut de couches pour le traiter. [19]

Les réseaux de neurones apprennent par le biais d'un algorithme, le réseau de neurones artificiels permet à l'ordinateur d'apprendre à partir de nouvelles données. L'ordinateur doté du réseau de neurones apprend à effectuer une tâche en analysant des exemples pour s'entraîner. Ces exemples ont préalablement été étiquetés afin que le réseau puisse savoir ce dont l s'agit. [19]

2.6.3.3 Architecture des réseaux de neurones

A. Les réseaux de neurones '*Feed-Forward*'

Cette classe se distingue par l'absence de toute boucle de rétroaction de la sortie vers l'entrée. En d'autres termes, la propagation des signaux s'y fait uniquement dans le sens de l'entrée vers la sortie. Ce type de réseaux comprend deux groupes d'architectures : le perceptron monocouche et le perceptron multicouche. Il différent par l'existence ou non des neurones intermédiaires appelés neurones cachés entres les unités d'entrées et les unités de sorties. [21]

a) Les réseaux de neurones monocouches

Le perceptron contient une couche d'entrée et de sortie, dont la couche de sortie est la seule couche effectuant des calculs. La couche d'entrée transmet les données à la couche de sortie et tous les calculs sont entièrement visibles pour l'utilisateur. [22]

b) Les réseaux de neurones multicouches

Les réseaux de neurones multicouches contiennent plusieurs couches de calcul (tous les nœuds d'une couche sont connectés à ceux de la couche suivante.), les couches intermédiaires supplémentaires (entre entrée et sortie) sont appelées couches cachées car les calculs effectués ne sont pas visibles pour l'utilisateur. L'architecture spécifique des réseaux de neurones multicouches est appelée réseaux *Feed-Forward* car les couches successives s'alimentent dans le sens direct de l'entrée vers la sortie. [22]. Parmi les exemples les plus populaire de ce type on trouve le réseau neuronal convolutif.

- **Les réseaux de neurones convolutif 'CNN'**

  Le réseau neuronal convolutif (*Convolutional Neural Networks:* CNN) est un algorithme d'apprentissage en profondeur et a multicouche principalement utilisé pour les problèmes de classification, c'est un algorithme très populaire à couse sa précision révolutionnaire. [24]

  Le réseau neuronal convolutif (CNN) sont très similaires aux réseaux de neurones normaux qui peuvent être visualisés comme une collection de neurones disposés sous forme de graphe acyclique. La principale différence avec un réseau de neurones est qu'un neurone de la couche cachée n'est connecté qu'à un sous-ensemble de neurones de la couche précédente. En raison de cette connectivité clairsemée, il est capable d'apprendre implicitement des fonctionnalités.

L'architecture profonde du réseau se traduit par une extraction hiérarchique des caractéristiques (Figure 2.6). [25]

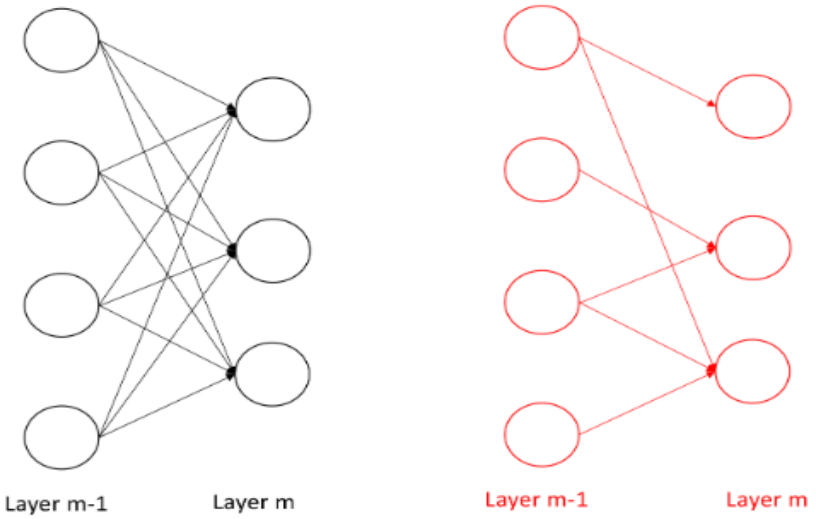

Figure 2.6:Comparaison entre l'architecture totalement connectée et partiellement connectée *[26]***.**

Un réseau de neurones convolutifs comprend successivement une couche d'entrée, plusieurs couches cachées, et une couche de sortie, la couche d'entrée sera dissemblable selon les diverses applications. Les couches cachées, qui constituent le bloc central d'une architecture CNN, consistent en une série de couches convolutives , de couches de regroupement et enfin exportent la sortie via la couche entièrement connectée (Figure 2.7) . [26]

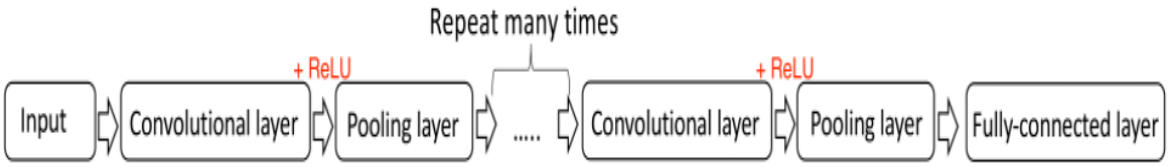

Figure 2.7:Structure de base de CNN. *[26]***.**

- **Réseaux de neurones convolutif une dimension (1D) et deux dimensions (2D)**

Les CNN pourraient être appliqués à tous les types de données. Par rapport au 2-D (principalement des données visuelles telles que des images et des

vidéos), il est possible d'utiliser des CNN pour des données 1-D (par exemple, des chaînes de caractères ou des mots). La principale différence est la dimensionnalité d'un noyau (unidimensionnel) et la façon dont il glisse sur les données (uniquement dans une direction).

B. Les réseaux de neurones récurent '*Feed-back*'

Les réseaux récurrents ou bien « *feed-back* » se distinguent des réseaux proactifs par le fait qu'ils contiennent au moins une boucle de contre-réaction des neurones de sortie vers les entrées ou au mois d'une couche vers une couche précédente, adjacente ou non. [21]

Le réseau de neurones récurrents (RNN) est un modèle puissant de la famille de l'apprentissage en profondeur, il fonctionne nettement mieux et prend moins du temps lorsque vous travaillez sur des tâches complexes avec de grandes quantités de données. Il vise à faire des prédictions sur des données séquentielles ((peut être des séries chronologiques) peuvent être sous forme de texte, audio, vidéo, etc.) en utilisant une puissante architecture basée sur la mémoire. [27]

- **Synthèse vocale basée sur les 'RNN'**

Une architecture de modèle acoustique plus populaire et efficace est une version des réseaux de neurones récurrents (RNN) qui peuvent traiter des séquences d'entrées et produire des séquences de sorties. Le RNN fonctionne non seulement sur les entrées mais également sur les états internes du réseau qui sont mis à jour en fonction de l'historique complet des entrées. Dans ce cas, les connexions récurrentes sont capables de cartographier et de mémoriser les informations dans la séquence acoustique, ce qui est important pour le traitement du signal vocal afin d'améliorer les sorties de prédiction. [28]

Les paramètres textuels et phonétiques sont d'abord convertis en une séquence de caractéristiques linguistiques en entrée, et des réseaux de neurones sont utilisés pour prédire les caractéristiques acoustiques en sortie pour synthétiser la parole. Étant donné que les RNN standard avec fonction d'activation sigmoïde souffrent à la fois de gradients de disparition et d'explosion. [28]

2.7 Applications de synthèse vocale à partir de texte

La synthèse vocale peut être utilisées dans plusieurs applications de notre vie quotidienne, ce qui la rend indispensable. Par exemple, l'augmentation de la disponibilité des systèmes TTS peut augmenter les opportunités d'emploi pour les personnes ayant des difficultés de communication.

Voici dans ce qui suit quelques exemples d'application des systèmes TTS :

- **Application pour les aveugles :** La synthèse vocale offre des nombreux services aux gens aveugle, par exemple avec l'application **voxiweb**[3] il sont capable d'accès à l'internet, là encore il existe d'autre logiciel qui fournit un lecteur d'écran grâce à la synthèse vocale.

- **Application pour les handicapés** : Les personnes sourdes et les personnes ayants des difficultés auditives ont généralement des difficultés à parler, la synthèse vocale donne aux sourds et handicapés vocaux (les malvoyants) une opportunité de communiquer avec les gens qui ne comprennent pas la langue des signes parmi ces logiciels l'application vocale presse, zommtexte[4], jaws[5].

- **Application pour les télécommunications et multimédia :** Le courrier électronique est devenu très courant ces dernières années. Cependant, il est parfois impossible de lire ces emails donc avec la synthèse ces emails peuvent être écoutés via des téléphones portable ou ordinateurs, en peut l'utilisés aussi pour prononcer les messages téléphoniques.

2.8 Quelques logiciels TTS

La synthèse vocale (la synthèse de la parole) à partir d'un texte exprime l'ensemble des traitements qui permet de transformer un texte écrit à un texte lue, son objectif est de transformer un texte en une suite de phonèmes en essayant de se rapprocher le plus possible de la parole humaine.

[3] https://www.voxiweb.com/fr/app/
[4] https://www.zoomtext.com/products/zoomtext-magnifierreader/
[5] https://www.ceciaa.com/jaws-logiciel-revue-ecran.html

Aujourd'hui des nombreux logiciels texte-to-speech sont disponibles. Ces logiciels fournissent plusieurs opportunités aux gens handicapés, sourds, ou même peuvent aider les personnes pour améliorer leur communication (personnes qui savent lire une langue et n'arrive pas à la parler), ces systèmes fournissent une description des textes à l'écran ou décrivent la scène devant eux ,il est clair qu'il est impossible de présenter tous les systèmes qui ont existé mais on va parler brièvement sur quelques systèmes les plus récemment utilisés via internet :

Voicemaker : C'est un site web de conversion de texte en parole gratuit en ligne (www.voicemaker.in), voicemaker offre plusieurs voix personnalisée (homme, femelle, enfant) selon 60 langues différentes (arabe, français, anglais, allemand, etc.), il est possible également de télécharger le texte articulé en fichier mp3.

TTSMP3[6]: Ce site web de conversation texte en parole peut synthétiser différents langue (arabe, français, turque, anglais, etc.) avec des voix différentes il est possible de passer d'un locuteur à un autre dans le texte et de télécharger le discours parlé en locale.

Natural Reader[7]: c'est un programme TTS professionnel qui convertie n'importe quel texte même des PDF, des fichier txt, des fichier Word, images. Pour télécharger le texte synthétiser il faut passer au forfait Premium pour utiliser cette fonctionnalité avec des premiums voix.

Cloud text-to-speech: ce logiciel permet d'entendre le texte écrit seulement avec une variété de choix de voix dans plusieurs langue (français, anglais, etc.)

IBM text-to-speech: pour nous c'est le seul logiciel qui produit une voix plus naturelle que les autres systèmes, il a dix personnes à choisir, il fournit un meilleur résultat de synthèse vocale par rapport aux autres logiciels de plus il fournit la possibilité de choix de dialecte des langues.

[6] : https://ttsmp3.com/
[7] https://www.naturalreaders.com/

2.9 Critères d'évaluation des systèmes TTS

L'objectif principale des recherches effectuées dans le domaine de synthèse vocale est de produire une voix la plus intelligible possible et la plus naturelle.

Le premier critère pour évaluer un système est donc la qualité de parole générée, parmi les techniques utilisées pour améliorer la voix on peut citer TD-PSOLA[8] (brevet CNECT en 1988 déposé par France Télécom) qui est souvent applicable à des systèmes de synthèse par concaténation. [12]

Le deuxième critère est la prosodie, il est important de réaliser des systèmes de conversation les plus naturellement possible sans pauses excessives, un être humain est capable d'intégrer son humeur quand il parle, réaliser plusieurs intonations dans un seul discours, changer des fréquences, ces derniers sont difficiles pour des systèmes.

Le troisième critère est la fiabilité, les systèmes TTS doit être résistants c'est-à-dire robuste pour garantie la prononciation d'une langue.

2.10 Conclusion

La synthèse vocale est un monde passionnant car elle atteint un certain niveau de qualité très proches d'une voix humaine.

Ce chapitre commence par une introduction sur la synthèse vocale à partir du texte, et une définition de la synthèse vocale et ses méthodes principales: (par formants, par concaténation, articulatoires) après on a détaillé la synthèse vocale à partir de texte. Par la suite on prend un aperçu du processus de synthèse vocale à partir du texte, en partant des deux modules principaux : le module de traitement linguistique et le module de synthétisation. On a parlé aussi sur de l'application de l'apprentissage automatique sur la synthèse vocale, puis on a cité quelques exemples des systèmes qui existent déjà. Pour clôturer le chapitre, on a mentionné quelque critère d'évaluations des systèmes TTS

Le chapitre suivant traitera les différentes étapes de l'architecture du projet pour construire une voix Arabe plus naturelle et plus compréhensible.

[8] Time Domain Pitch Synchronous Overlap Add

Chapitre 3 : Approche proposée

3.1 Introduction

Nous avons présenté dans les deux premiers chapitres le concept de la synthèse vocale et son processus pour générer une voix automatique. Pour crée une voix de synthèse, il est nécessaire d'utiliser un moteur de synthèse vocale (front- end et le back- end), ce dernier permet à façonner une voix artificielle à partir d'un ensemble du mots prononcés par une machine.

De nos jours la majorité des systèmes de synthèses vocale pour la langue Arabe sont mal articuler par rapport à d'autres langues comme l'anglais par exemple. Cette réalité nous à pousser pour penser d'approfondir dans ce domaine afin d'apporter un plus à la recherche spécifiquement pour la langue Arabe.

La plupart des travaux antérieurs utilisent l'apprentissage profond pour la reconnaissance vocale. C'est de là que vient notre idée de réaliser un système de synthèse vocale avec l'application d'apprentissage profond pour optimiser la synthèse vocale afin de rapprocher le rendu sonore de la voix humaine.

Dans ce chapitre nous détaillons notre contribution qui fonctionne sur la méthodologie et le processus utilisé pour crée un signal vocal à partir d'un texte Arabe grâce à l'inclusion de deep Learning.

3.2 Corpus utilisé

Tout système intelligent doit apprendre les connaissances pertinentes pour un meilleur fonctionnement. Pour ce faire nous avons besoin d'un corpus approprié en termes de contenu et de taille ; dans notre cas on a utilisé un corpus sonore de phonèmes doivent être bien prononcés.

Étant donné que ce domaine est en cours de développement pour la langue Arabe, nous constatons que les bases de données ne sont pas publiques, c'est pourquoi nous n'avons pas eu d'autre choix que de travailler avec la base de données de 'Nawar Halabi ARABIC SPEECH CORPUS'[9].

Cette base de données a été construite dans le cadre de la thèse de doctorat de Nawar Halabi de à l'Université de Southampton, elle a été enregistrée dans un studio

[9]http://en.arabicspeechcorpus.com/

professionnel du dialecte damascène, elle contient 1813 fichiers audio (.wav) 1813 fichiers (.lab) qui contiennent les prononciations de chaque enregistrement. Arabic Speech Corpus possède aussi 1813 Fichier (. TextGrid) il contient les moments dans le temps et les symboles qui identifient les phonèmes prononcés tout au long de la base. Ces fichiers peuvent être ouverts avec le logiciel PRAAT.

Le fichier phonetic-transcirpt.txt contient dans chaque ligne le nom du fichier audio (wav) et les phonèmes transcrits dans l'ordre.

3.3 Architecture de l'approche proposée

L'objectif principal des systèmes de synthèse vocale est de fournir à l'ordinateur la capacité de lire des textes à haute voix. Malgré les avancées réalisées dans ce domaine ces dernières années, des progrès sont encore nécessaires pour améliorer l'utilisabilité et la qualité des systèmes existants. Notre approche présente un exemple de ces progrès, qui vise à améliorer la qualité des synthétiseurs de parole en utilisant des réseaux de neurones convolutifs. L'approche proposée repose sur le développement d'un schéma (figure 3.1) basé sur les principaux blocs suivants :

- **Phase 1** : Préparation du modèle par les réseaux de neurones convolutif.
- **Phase 2** :Contient deux sous principales étapes :
  - La Génération de voix.
  - Le Prétraitement de texte.

Nous décrivons maintenant les fonctionnalités de chaque bloc, en commençant par les premiers traitements appliqués sur un texte Arabe en entré, et nous terminerons par la synthétisation automatique de ce texte.

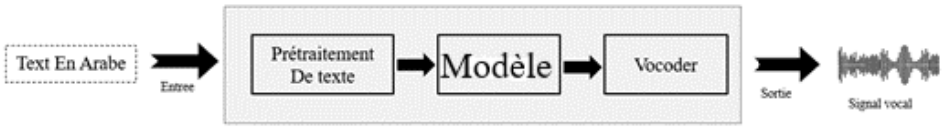

Figure 3.1:Architecture générale.

3.3.1 Phase 1 : Préparation du Modèle

Il s'agit de la phase de la création de notre modèle d'apprentissage à base de réseau de neurones, dont la réalisation passe par un ensemble d'étapes (Figure 3.2), que nous allons détailler dans le paragraphe suivant.

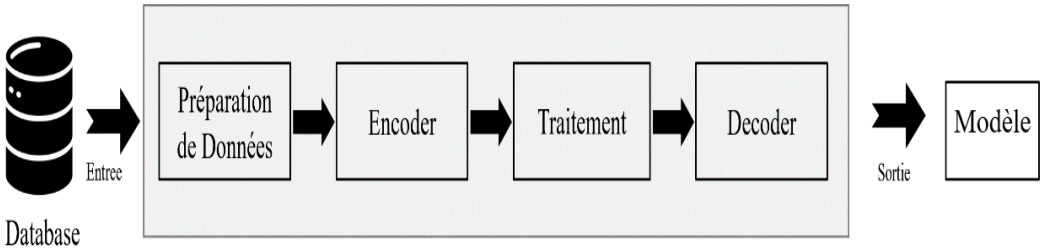

Figure 3.2:Préparation du modèle.

A. Préparation des données

L'objectif de cette première phase est de préparer les données comme entrée pour notre modèle. La figure 3.3 montre la procédure suivie pour l'extraction des données de Corpus.

Pendant l'extraction des données nous utiliserons un ensemble de fonctions(Figure 3.3). Dans un premier temps nous lisons le fichier textGride et le fichier wav (audio) pour avoir la description de chaque phrase prononcée dans la Base de Données Sonores (BDS), ensuite les deux fonctions extraire des paramètres et nettoyer wav sont utilisées pour éliminer le silence de début et de fin de chaque phrase pour éviter la coarticulation au moment de la concaténation des phonèmes lors de la génération de signal vocal associé à la phrase introduite par les utilisateurs.

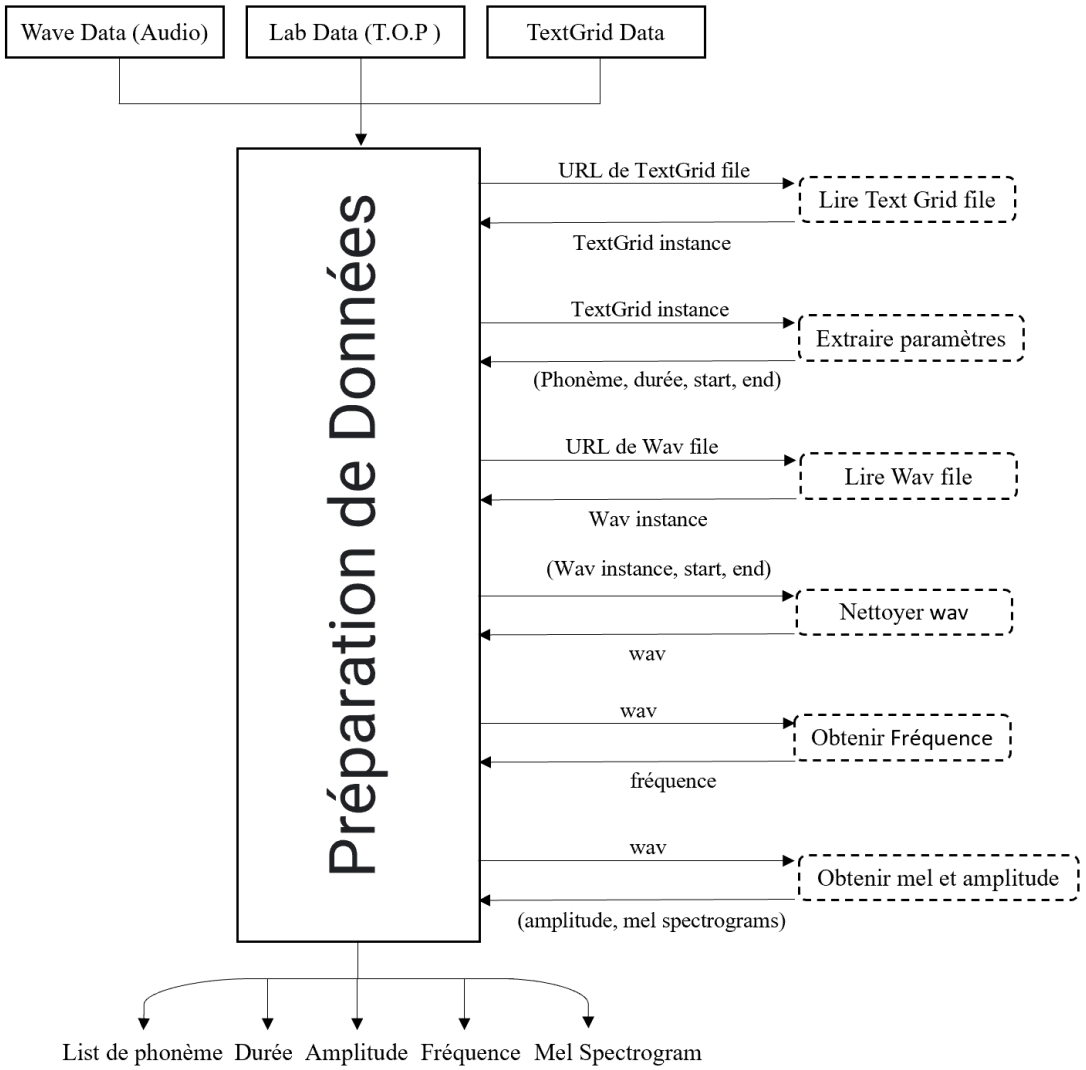

Figure 3.3: Phase de la préparation des données.

Les fonctions obtenir fréquence et obtenir mel et amplitude sont utilisées pour extraire à partir de l'audio de chaque phonème sa fréquence, son intensité et le spectrogramme mel qui lui correspond. À la fin, nous aurons les résultats suivants : une liste de phonèmes, une liste de fréquences et une liste d'intensités avec un tableau à deux dimensions qui représente le spectrogramme de la phrase.

B. Encodeur

L'encodage signifie convertir les données dans un format requis, dans notre cas l'encodeur convertir la liste des phonèmes en phonèmes embedding puis en séquence cachée de phonème.

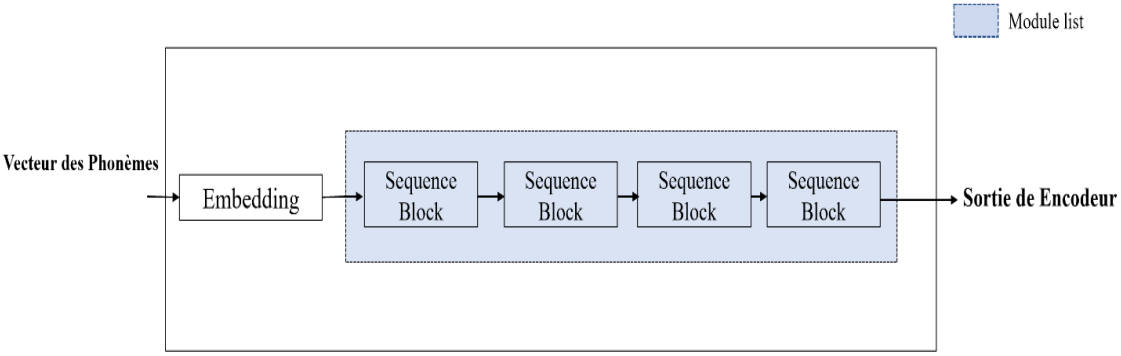

Figure 3.4:Encodeur.

a. Embedding

L'embedding indique que nous allons encoder la liste des phonèmes à une matrice de dictionnaires où chaque phonème va être encodé dans 256 Channel. Par exemple on a une phrase avec 26 phonèmes, chaque phonème on va représenter par un vecteur de taille 256.

b. Séquence Block

Un conteneur contenant deux modèles exécutés, l'un suivis par l'autre. Ce type de modèle est utilisé dans le cas de données séquentielles textuelles, les modèles utilisés dans notre cas sont :

1) Multi Head attention

Il s'est répandu en grand pourcentage récemment dans ce type de problèmes (problèmes de séquençage), comme c'est le cas pour nous, la séquence des phonèmes. Cette méthode sera endommagée en améliorant les performances de l'apprentissage automatique et en donnant de meilleurs résultats. La figure suivante (figure 3.5) exprime un exemple détaillé.

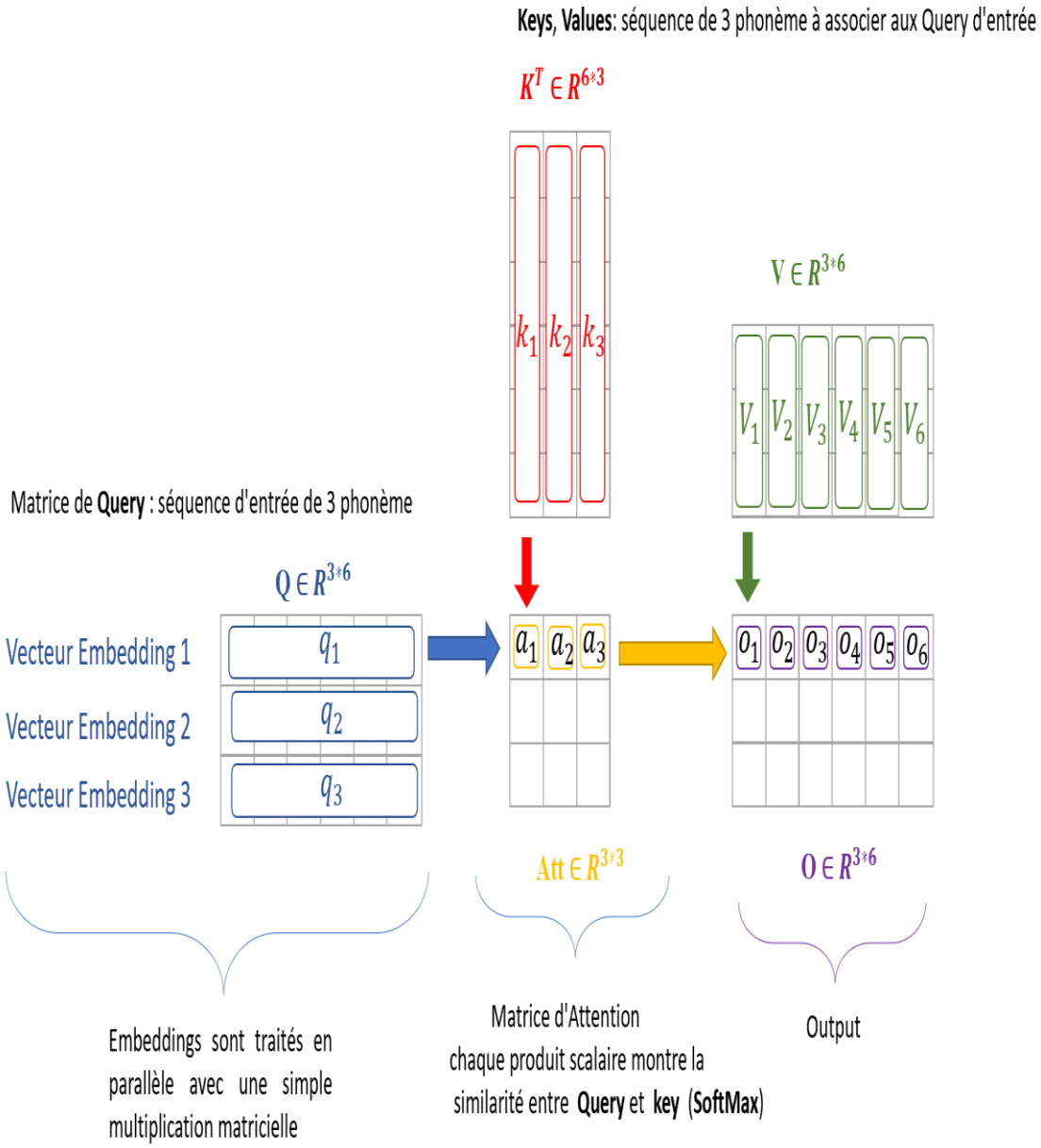

Figure 3.5:Fonctionnement de Multi-Head-Attention.

$$Attention(Q,K,V) = SoftMAx\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

2) CNN Layer

Selon [29] la convolution est une opération mathématique simple, généralement utilisée pour le traitement et la reconnaissance des images. Sur une image, son effet est similaire à une opération de filtrage comme illustre la figure 3.6 :

- Tout d'abord, on définit la taille de la fenêtre de filtrage située en haut à gauche.
- La fenêtre de filtrage, représentant la caractéristique, se déplace progressivement de gauche à droite d'un certain nombre de cellules défini au préalable (le pas) jusqu'à atteindre la fin de l'image.
- A chaque portion de l'image rencontrée, un calcul de convolution est effectué pour obtenir une carte d'activation ou carte de caractéristiques qui indique où se trouvent les caractéristiques dans l'image.

L'opération présentée dans la figure 3.6 exprime la convolution calculée dans une seule Channel, donc cette opération vas être répétée 256 fois.

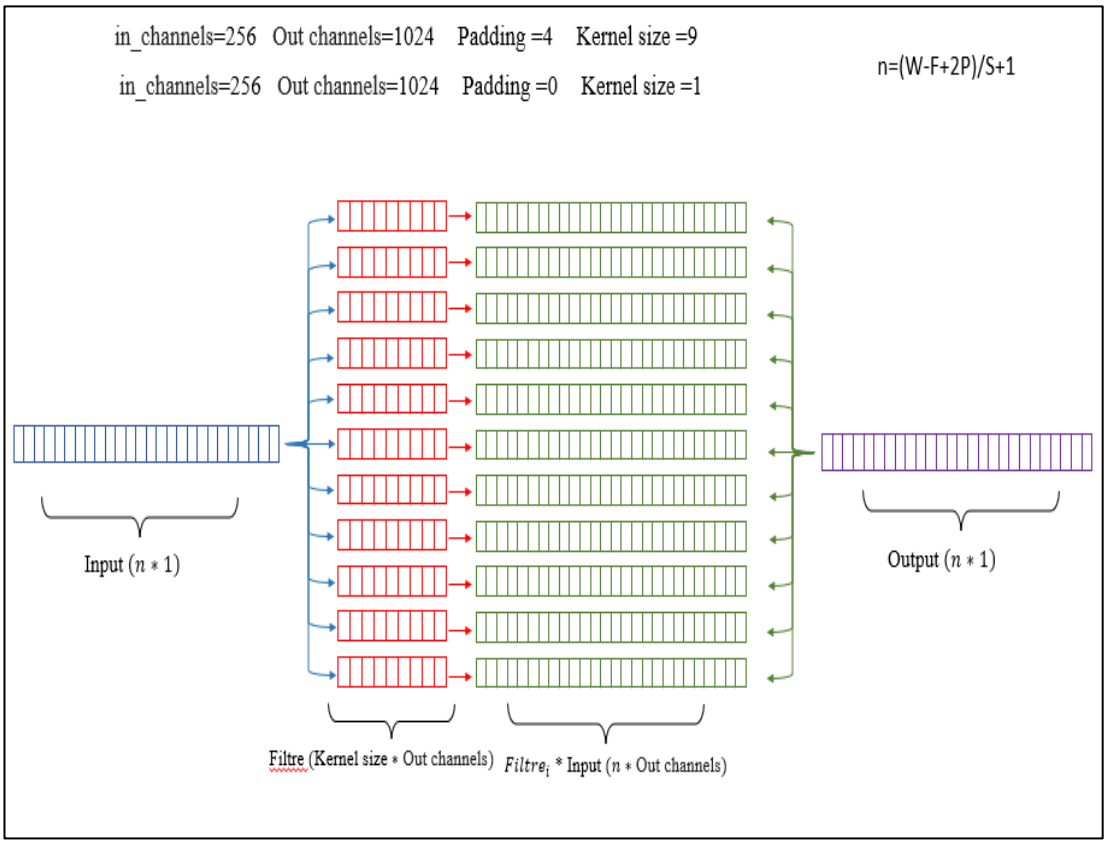

Figure 3.6: Fonctionnement de 1D convolution layer.

Dans le cas d'un texte nous avons une entrée d'une dimension, l'approche proposée utilise un filtre de taille 9 ou la sortie sera calculée par une fonction d'activation qui vas permet simplement de remplacer les

résultats négatifs par zéro, et rentre dans une autre convolution mais avec un filtre de taille 1, comme illustre la figure 3.7.

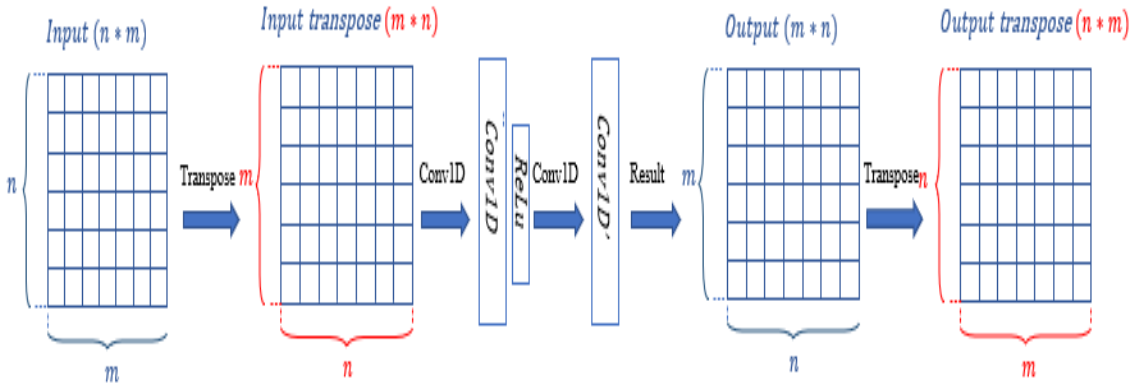

Figure 3.7 :CNN Layer.

C. Traitement

Dans cette étape nous allons entraîner notre modèle pour prédire les paramètres (fréquence, durée, énergie, spectrogramme-mel) à partir de la sortie de la phase précédente (encodeur) et les données extraites de la BDS *Arabic Speech Corpus*, pour l'entraîner nous avons utilisé séquence modele.

La fréquence prédicteur et l'énergie prédicteur sont deux fonctions utilisées pour générer l'embedding de l'entrée et sa valeur prédite, tandis que la durée prédicteur est utilisée pour calculer la fonction de perte (*loss function*). La somme de l'embedding et de la durée seront les entrées du prédicteur mel qui générera un spectrogramme prédit mel. Encoder input c'est bien une matrice qui caractérise la phrase a transforméVoici la figure suivante (figure 3.8) :

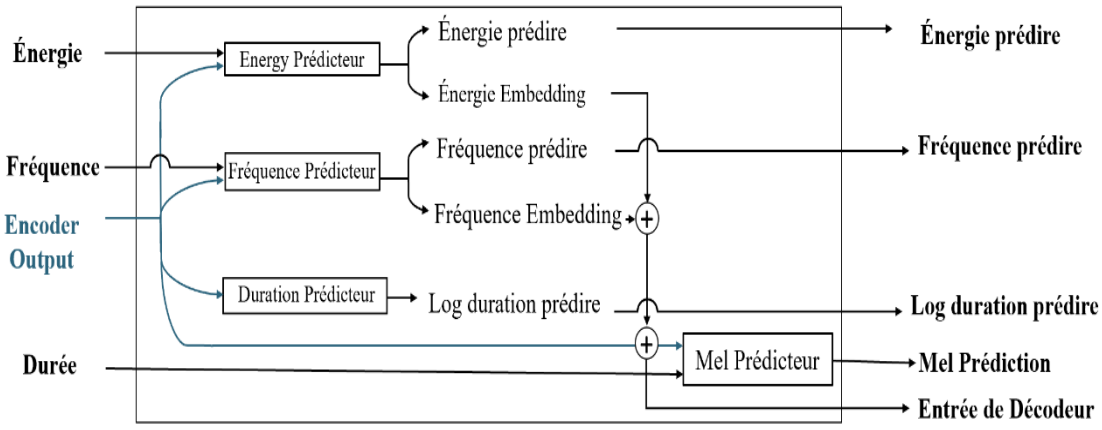

Figure 3.8: Fonctionnement de traitement.

D. Décodeur

Cette étape est spécialisée pour le spectrogramme mel (MelS), elle représente le processus inverse du l'encodeur. Le décodeur convertit la séquence cachée adaptée en une séquence de spectrogramme mel en parallèle.

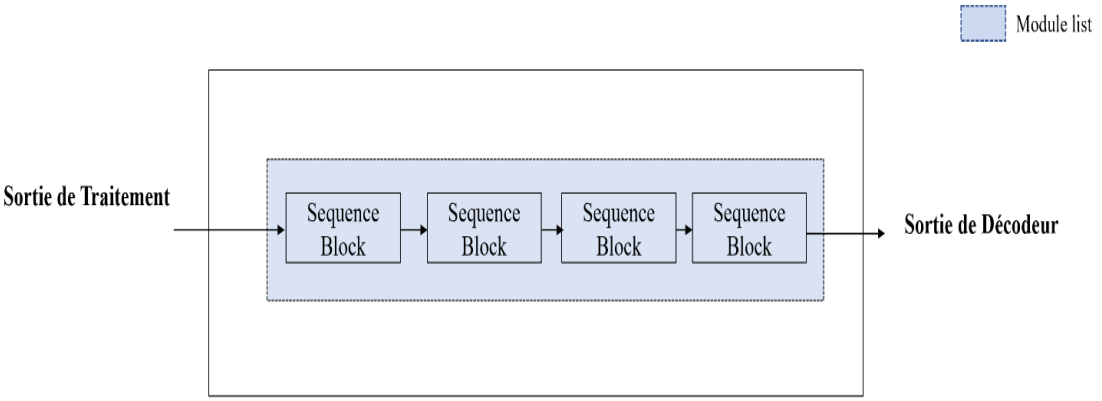

Figure 3.9: Décodeur.

3.3.2 Phase 2 : Génération de la voix et prétraitement des données :

A. La Génération de la voix

Les nouveaux systèmes de génération de voix utilisent directement le vocodeur pour produire le son afin d'adapter rapidement la génération de la phrase par la concaténation des phonèmes d'une BDS.

Comme tout système basé sur l'apprentissage, la génération est construite principalement en deux phases : une phase d'apprentissage et une phase d'inférence

(production de la voix). Dans l'approche que nous proposons, un spectrogramme est généré par le décodeur puis un vocodeur transforme cette représentation en un fichier vocal. Les entrées du vocodeur sont les paramètres acoustiques qui caractérisent la phrase à transformer et parfois aussi l'identifiant du locuteur en sélectionnant les caractéristiques vocales qui correspondent à la voix générée. Le vocodeur a pour rôle de générer un fichier vocal final à partir de la représentation compacte de l'audio généré. Dans notre cas, les données d'entrée du vocodeur sont le spectrogramme mel et le path du fichier wav associé aux paramètres. Le modèle utilisé comme vocodeur pour la génération d'une voix artificielle est le vocodeur ***hifigan universel*** [10]**.**

B. Le Prétraitement de texte

C'est l'étape où on travaille sur le texte (input) pour le préparer à être insérer dans le modèle, le résultat de cette phase sera un vecteur des chiffres, sa taille est la même que celle du texte, la figure suivante (figure 3.10) montre les étapes prises en considération pour sa réalisation.

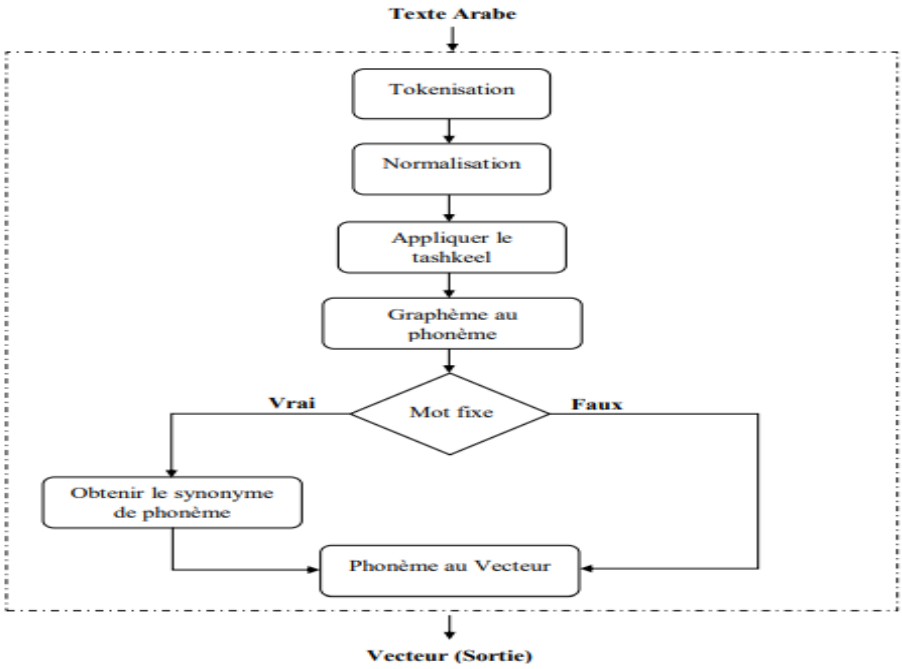

Figure 3.10:Processus de la phase de préparation de données.

[10] https://github.com/keonlee9420/Parallel-Tacotron2/blob/main/hifigan/generator_universal.pth.tar.zip

a. Tokenisation : C’est l'opération utilisée pour diviser le texte d'entrée en jeton (mots) en utilisant les espaces blancs et la ponctuation.

Exemple : *الجو جميل* devient *[جميل] [الجو]*.

b. Normalisation : Normalisation : il s'agit d'une technique qui vise à transformer des éléments non textuels en texte. Tel que les chiffres les caractères spéciaux, la ponctuation, etc.

Exemple : *يوما* 50devient *خمسون يوما*.

c. Applique le tashkeel : Comme il a déjà été mentionné dans le chapitre 1 le problème de la langue Arabe est le tashkeel, et pour éviter le cas de changement de sens des mots qui ont le même radical nous avons appliqué le tashkeel.

d. Graphème au phonème : C'est le niveau où les données d'entrée (mots) sont préparées pour la transformation des lettres en sons. C'est une étape très importante dans la synthèse vocale, elle est utilisée pour générer la prononciation correcte en fonction du texte donné, elle nécessite un mécanisme pour la conversation des mots en séquences phonétiques, pendant cette conversion, un ensemble de règles linguistiques bien définies est pris en compte et dans la plupart des cas, ces règles sont complétées par un dictionnaire de mots d’exception (mots fixe).

e. Mots fixe : A ce stade, le système vérifie si les mots sont des mots irréguliers, ces derniers sont un ensemble de mots spéciaux, dont la prononciation diffère de l'écriture. Si le mot est un mot fixe, le système obtiendra son synonyme phonémique. Si ce n'est pas le cas, il passe à l'étape suivante. Le tableau 3.1 contient quelques exemples de ces mots.

Mot fixe	Synonyme phonémique	Mot fixe	Synonyme phonémique
هذا	/ H a : d a : /	لكن	/ L a : k i n n a /
ذلك	/ ð a l I k a /	الله	/ A l l a h u /
اولئك	< u0 l aa < i0 k a	الرحمن	/ A r a h m a : n /

Tableau 3.1 : Exemple des mots fixe en Arabe.

f. Phonème au vecteur : C'est l'étape où les phonèmes sont transformés en une séquence de chiffres pour servir d'entrée à de modèle. Chaque phonème sera un chiffre spécifique diffèrent d'un autre.

3.4 Conclusion

Dans ce chapitre, nous présenterons l'aspect technique de notre travail, en déterminant les étapes nécessaires à la création d'une voix synthétique en Arabe.

On a commencé le chapitre par une description générale et globale de notre approche et de son architecture, puis on a détaillé aussi toutes les étapes qui composent cette dernière et qui sont : le prétraitement, l'entraînement, le vocodeur, la préparation des données, et le générateur de la voix.

Le chapitre suivant traite la phase d'implémentation et test, ou on va étudier et analyser les résultats obtenus de notre plateforme.

Chapitre 4 : Implémentation et résultats

4.1 Introduction

Dans ce chapitre nous allons présenter les outils utilisés pour le développement de notre système qui créé une voix synthétique à l'aide de l'apprentissage profond en utilisant des réseaux de neurones convolutifs (CNN).

On va donner une brève présentation de l'interface de notre plateforme ensuite on va décrire la méthodologie du test et ses résultats. Le test a été effectué à l'aide d'un booster de questionnaire de comparaison de trois voix Arabes, y compris celui que nous avons construit. Nous terminerons par faire le système construire open source.

4.2 Les outils utilisés

Pour assurer le bon fonctionnement de notre plateforme on a utilisé un ensemble d'outils qui peuvent être résumé par [30]:

- ***Pycharm :*** est un logiciel de type EDI (Voir environnement de développement intégré et échange de données informatisées) dédié à Python, il s'exécute dans un environnement multiplateforme, son langage de développement est Java et Python, la langue utilisée l'anglais , Il permet l'analyse du code, et il se base sur des tests unitaires, de plus il utilise un système de versionnage et permet l'utilisation des Framework Django[11], web2py[12] et Flask. Il intègre un débogueur et une fonction de refactoring.

- ***PyTorch*** est de type bibliothèque logicielle ses environnements Linux, Microsoft Windows et MacOs, langage utilisé C++, Python, C et *Compute United Device Architecture,* permet d'effectuer les calculs tensoriels dont on a besoin pour réaliser des applications d'apprentissage profond et de construire des réseaux neuronaux.

- ***TensorFlow*** est de type bibliothèque open source d'intelligence artificielle son environnement multiplateforme, langage utilisé *C++ et Python,* la langue utilisée l'anglais, Il permet de construire des réseaux de neurones complexes

[11] : Ensemble de composants logiciels destinés à construire des *applications* ou des sites web en simplifiant le travail des développeurs.
[12] : est un Framework web open source

et de les entrainer. *TensorFlow* met à disposition des utilisateurs beaucoup de modèles et d'algorithmes. Son *API*[13] *Front-end* repose sur *Python*, l'exécution des calculs s'effectue en *C++*.

- ***Sklearn :*** *encore appelé* ***Scikit-learn*** *de type* bibliothèque logicielle pour apprentissage automatique, le langage de développement utilisé *Python,C, C++* et *Cython* , Conçu pour être utilisé avec d'autres bibliothèques comme *NumPy* et *SciPy*, *Scikit-learn* implémente de nombreux algorithmes utiles pour la conception d'application d'intelligence artificielle, il permet d'effectuer des prétraitements sur des données textuelles comme la normalisation.

- ***React* :** est une bibliothèque *Javascript* open source, relativement complexe. Elle permet de créer une interface utilisateur et plus spécifiquement des applications *web* mono-page et aussi elle peut être utilisé avec un *Framework*[14] comme *AngularJS* [15] . Elle peut être utilisée aussi bien du côté client que du côté serveur. Elle utilise un DOM [16] virtuel et un flux de données unidirectionnel. Les changements de code dans le code hiérarchique de niveau inférieur ne peuvent pas influencer de code supérieur.

- ***Flask*** est un micro *Framework* open-source de développement web en *Python.* Il est classé comme micro *Framework* car il est très léger. *Flask* offre la possibilité de relier le frontend (interface de la plateforme) et le backend (code source), ce qui est appelé *REST API. REST* permet de déplacer les données entre les utilisateurs et les applications, ce qui aide les clients à communiquer avec le serveur hôte.

4.3 L'interface de notre plateforme

Le système proposé contient deux éditeurs de texte. Le premier permet à l'utilisateur de saisir son texte et le second affiche le texte écrit avec les signes diacritiques. L'utilisateur est autorisé à taper le texte qu'il veut en Arabe standard et ensuite il peut

[13] : ***A****pplication* ***P****rogramming* ***I****nterface* (Interface de programmation d'applications).
[14] Framework : Ensemble de composants logiciels destinés à construire des applications ou des sites web en simplifiant le travail des développeurs.
[15] AngularJS est un Framework Javascript.
[16] ***D****ocument* ***O****bject* ***M****odel.*

voir la transcription de son texte d'entrée et le spectrogramme de chaque texte transcrit. La figure(4.1) ci-dessous illustre l'interface de notre plateforme.

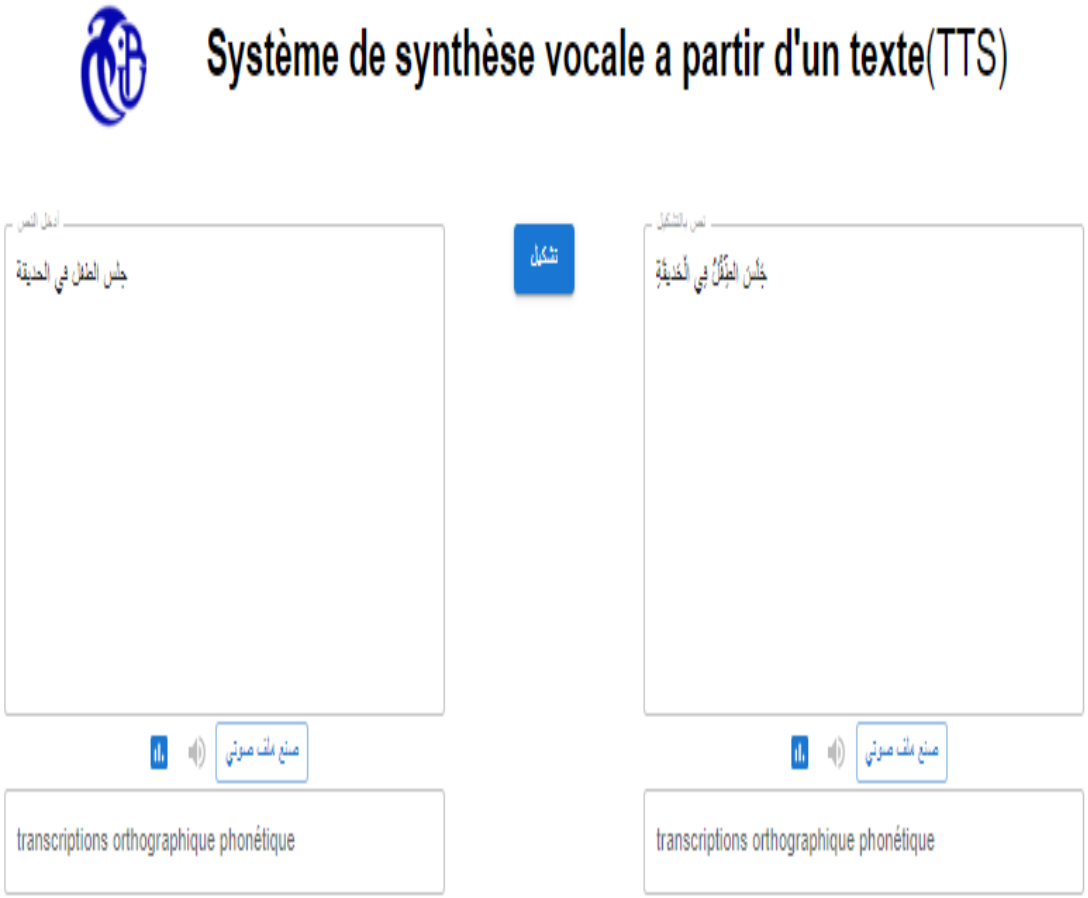

Figure 4.1:Interface de système de synthèse vocale à partir de texte (TTS-AS).

- **Éditeurs de texte :** l'interface principale de plateforme contient un éditeur de texte où l'utilisateur écrit ce qu'il veut en Arabe standard

- **Boutons de commande :** Quatre boutons de commande ont été créés pour le système que nous proposons, qui sont :'تشكيل' [tashkil], 'صنع ملف صوتي' [sune milaf sawti] (Créer un fichier audio), (écouter l'audio) et (présenter le spectrogramme) pour les deux textes avec les signes diacritiques et sons signes diacritiques.

    - **Bouton 'تشكيل'**: [tashkil] Permet de placer des signes diacritiques sur le texte que vous voulez entendre.

- **Bouton 'صنع ملف صوتي':** [sune milaf sawti] vous permet de créer des fichiers audio afin de pouvoir écouter
- **Bouton 🔊 :** (Ecouter l'audio) vous donnez la possibilité d'entendre la voix générée par notre système de la phrase d'entrée.
- **Bouton 📊 :** Lorsque vous cliquez dessus, vous obtenez le spectrogramme associé à la phrase d'entrée.

4.4 Méthodologie de test

Pour le test, nous avons choisi trois phrases écrites en Arabe standard, et nous les avons synthétisées en utilisant les voix Arabes open-source qui utilisent le même corpus pour ces systèmes :

- ***Arabic TTS (الناطق العربي)***[17] réaliser par 'Abduallah Asrajeh', génère une voix synthétique par la méthode statistique clustering et synthétisé par festival.
- **Festival-Ziad** : réaliser par 'K. Hemina' et 'O. Hemina', créé une voix en utilisant les systèmes de Markov cachée. [11]

La qualité des systèmes de synthèse vocale dépend du locuteur, il s'agit donc d'une mesure complexe à définir en raison de sa forte subjectivité. Pour l'évaluation de notre système, nous avons utilisé le *MOS 'Mean Opinion Score'*.

Le *MOS* est une mesure numérique de la qualité globale d'un événement ou d'une expérience. Dans le domaine des télécommunications, une opinion moyenne est un classement de la qualité des sessions vocales et vidéo. Chaque auditeur note l'audio selon une échelle prédéfinie (de médiocre à excellent). On fait ensuite la moyenne des notes pour définir la valeur finale de l'évaluation.

Pour la création du notre formulaire[18] on a utilisé google Forms. L'auditeur après chaque écoute notera la qualité de l'audio sur une échelle de cinq niveaux (mauvaise à excellent) . De plus les voix des trois systèmes sont anonymes afin de réaliser un test

[17] https://github.com/asrajeh/arabic-tts/issues
[18] https://docs.google.com/forms/d/e/1FAIpQLSfCJMh37wku6iaMg4lQcn1vS7jCkN-prBHEt7VyccZDE7spHQ/viewform

fiable et crédible. Voici dans ce qui suit un exemple des phrases spécifiques aux tests :

يُحْكَى بِأَنَّهُ كَانَ هُنَاكَ فَتَاةٌ فَقِيرَةٌ جَمِيلَةٌ ذَاتِ شَعْرٍ أَشْقَرٍ طَوِيلٍ

خَرَجَتْ فِي لَيْلَةِ رَأْسِ السَّنَةِ الْمِيلَادِيَّةِ مِنْ أَجْلِ بَيْعِ الْكِبْرِيتِ

وَكَانَتْ اللَّيْلَةُ شَدِيدَةَ الْبُرودَةِ

4.5 Tests et Résultats

Nous présentons dans cette partie le teste de notre plateforme qui est réalisé tout on suivant deux étapes : la première consiste a évalué notre approche proposée avec quelques systèmes TTS qu'existes et qui utilisent le même corpus, ce test est assuré suite à un questionnaire en ligne et une population aléatoire, et la seconde est de faire un Test de qualité basé sur une population ciblée pour mesurer et évaluer notre synthétiseur de parole Arabe.

- **Partie I : un test comparatif à base de formulaire en ligne à l'aide d'une population aléatoire**

Le questionnaire publié comprend les voix synthétisées par les trois systèmes. L'auditeur écoutera l'articulation de trois phrases de test par ces systèmes et les évaluera selon l'aspect de la compréhension et du naturel. Les figures suivantes présentent les résultats des votes majoritaires des auditeurs selon les deux aspects (compréhension et naturel) pour la phrase (يُحْكَى بِأَنَّهُ كَانَ هُنَاكَ فَتَاةٌ فَقِيرَةٌ جَمِيلَةٌ ذَاتِ شَعْرٍ أَشْقَرٍ طَوِيلٍ), Nous nous sommes appuyés sur la mesure de MOS pour classer les résultats obtenus.

Parmi les questions de notre formulaire en ligne on trouve :

- **Question 1 : comprenez-vous la langue Arabe ?**
  - **Oui :** 56 personnes ont répondu par oui.
  - **Non :** Une personne a répondu par non.

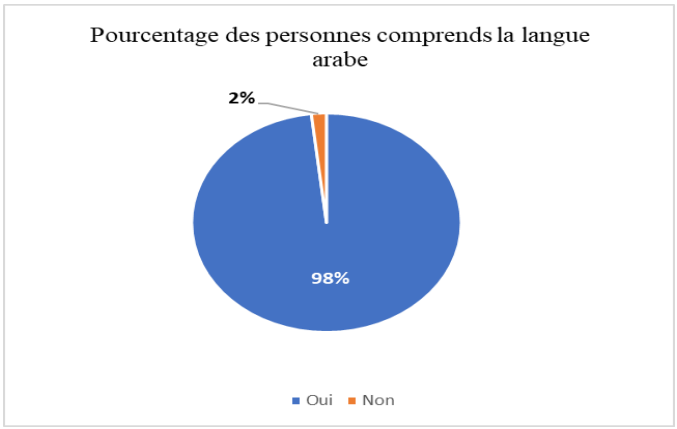

Figure 4.2:Pourcentages des personnes comprends la langue rabe

- **Question 2 : les mots d'audio, sont-ils compréhensibles ?**

Voici les résultats dans la figure(4.3) ci-dessous.

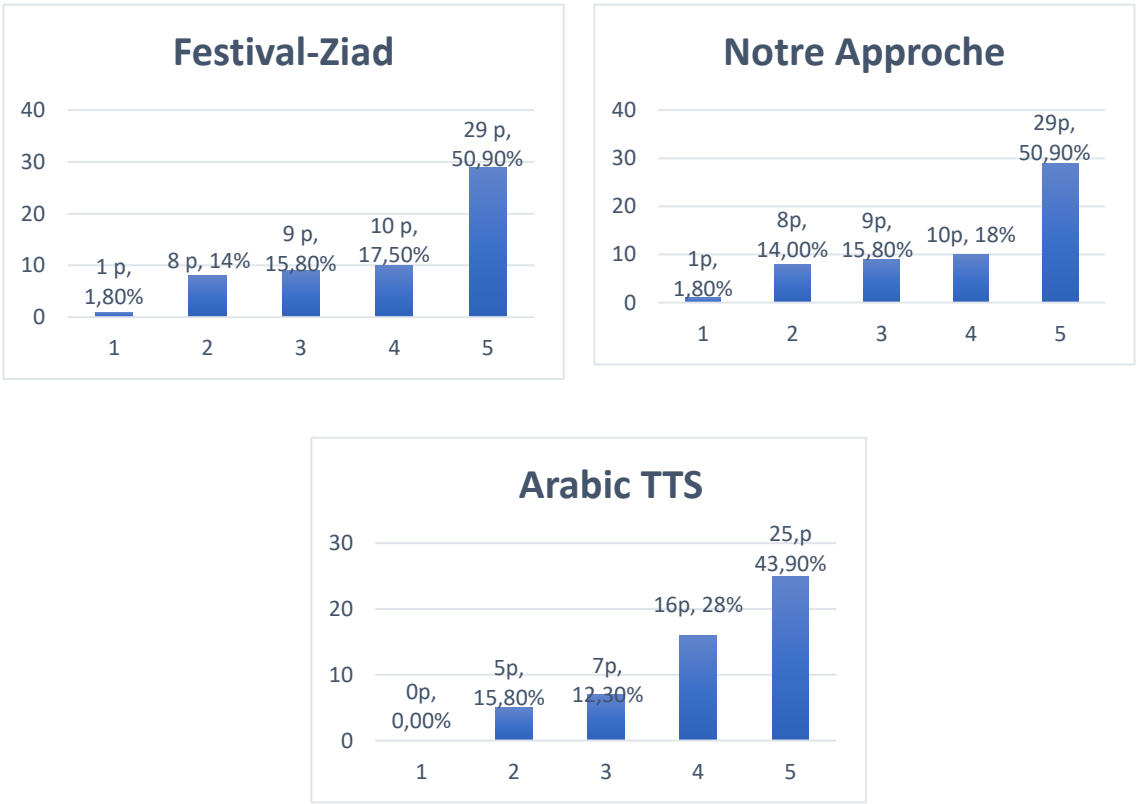

Figure 4.3:Résultats d'évaluation de compréhensibilité des trois systèmes.

- **Classement des résultats de la compréhensibilité :**

Sur la base des résultats de l'évaluation de l'aspect compréhensible de la phrase par les trois systèmes présentés dans la figure (Figure 4.3), nous

présenterons ces résultats selon la mesure MOS dans le tableau (tableau 4.1).

Le MOS calculé pour la première phrase test généré de par le système de Festival-Ziad est le suivant :

MOS : 1*1.8% +8*14% +9*15.8% +10*18% +29*50.9% = 19.07

	Festival-Ziad	Notre système	Arabic TTS
Phrase 1	19.07	35.38	17.75
Phrase2	17.36	29.09	16.63
Phrase 3	16.93	30.78	19.60

Tableau 4.1:Moyenne de test de la compréhensibilité.

- **Discussion des résultats de compréhensibilité :**

On se basant sur le tableau ci-dessus on peut conclure que la moyenne la plus élevée est celle générée par notre système, suivie par Festivale-Ziad et à la fin celle générée par l'Arabic TTS en raison de la méthode utilisée de synthèse basée sur des règles qui affecte négativement le pourcentage de phrases de mots identifiés.

- **Question 3 : le son généré par audio, est-il robotique ou bien naturel?**

Voici les résultats dans la (figure4.4) ci-dessous:

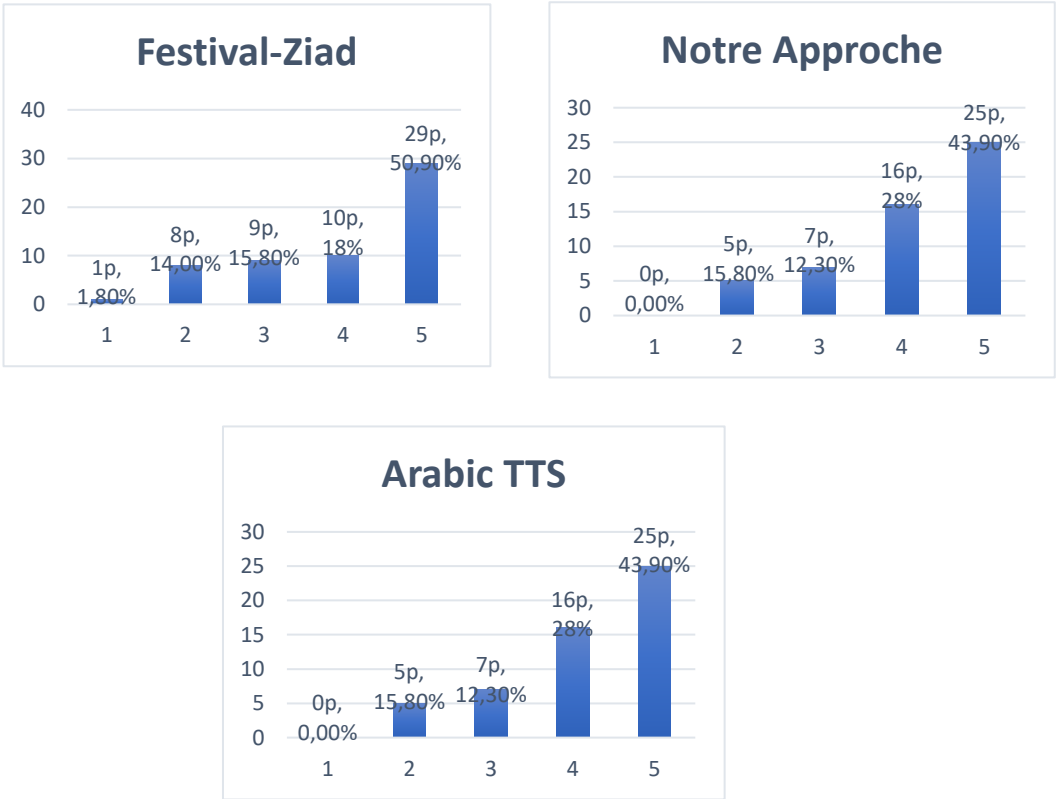

Figure 4.4:Résultats d'évaluation de la naturalité des trois systèmes.

- **Classement des résultats de la naturalité :**

A partir des résultats illustrés dans la figure 4.4, nous les présenterons dans le tableau ci-dessous(tableau 4.2) avec le calcul de MOS.

Le MOS calculé pour la deuxième phrase de test généré par notre approche est le suivant :

MOS : 2* 3.5% + 6*10.5% +5*8.8% 19*33.3% +25*43.9% = **18.44**

	Festival-Ziad	Notre système	Arabic TTS
Phrase 1	13.38	19.57	14.07
Phrase2	13.59	18.44	13.42
Phrase 3	13.82	20.05	14.29

Tableau 4.2:Moyenne de teste de naturalité.

- **Discussion des résultats de Naturalité :**

Les résultats représentés par le tableau ci-dessus montrent que notre système est classé premier puis celui de Festival-Ziad pour les 2 premières phrases et pour la troisième phrase nous avons le classement suivant : notre système puis les TTS Arabes et à la fin Festival-Ziad.

La figure 4.5 montre un résumé des résultats des deux critères proposés obtenus par le questionnaire.

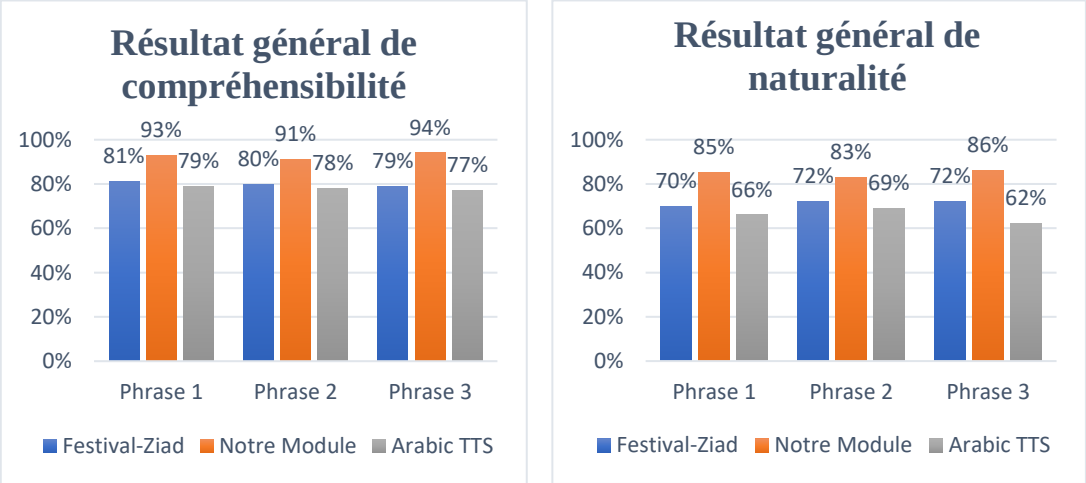

Figure 4.5:Résultats générale de l'évaluation des trois systèmes.

- **Question 4: A votre avis qui est le meilleur synthétiseur vocal ?**

Le but de cette question était de garantir et de confirmer les résultats obtenus par les échelles proposées pour chaque critère, la majorité des auditeurs ont choisi la synthèse vocale proposée par notre synthétiseur vocal et cela indique que les résultats obtenus sont fiables et crédibles voici la figure 4.6.

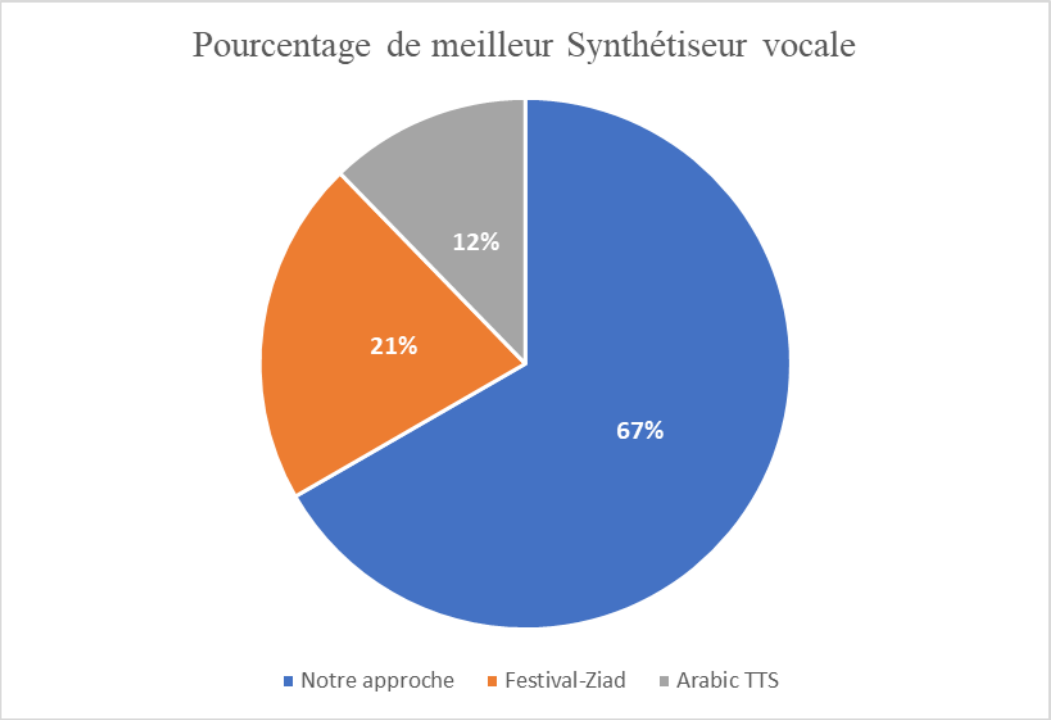

Figure 4.6:Pourcentage de meilleur synthétiseur vocal.

- **Question 5 : vous pouvez laisser votre avis sur l'évaluation effectuée.**

Nous avons terminé le questionnaire en laissant l'auditeur faire des commentaires constructifs pour montrer l'importance du projet et de connaître les critiques dans le domaine de la synthèse vocale afin d'améliorer notre travail et d'assurer une approche qui répond mieux aux besoins des utilisateurs, cette question n'est pas obligatoire à la suite nous avons seulement deux réponses :

- Ajouter des voix féminines
- بالتوفيق

Partie II : Test de qualité basé sur une population ciblée.

L'un des critères d'évaluation les plus subjectifs est le degré de naturel et d'intelligibilité. Cette fois-ci, nous avons choisi 20 participants (âgés de 18 à 35 ans) ce test est effectué pour savoir si le son généré par notre système est de bonne qualité, le test est effectué sur deux textes, l'un écrit avec des signes diacritiques et l'autre sans

signes diacritiques pour s'assurer que les signes diacritiques ont un très grand impact sur les voix générées Nous résumons les résultats comme suit :

- **Naturalité :**

	Naturel	Acceptable	Non-naturel	Totale
Participants	0	1	19	20
Pourcentage	0%	5%	95%	100%

Tableau 4.3: Evaluation de mesure de naturalité de texte sans signes diacritiques.

MOS **=** 0*0% + 1*5% + 19*95% = **18.95 (Robotique)**

	Naturel	Acceptable	Non-naturel	Totale
Participants	9	7	4	20
Pourcentage	45%	35%	20%	100%

Tableau 4.4: Evaluation de mesure de naturalité de texte avec signes diacritiques.

MOS **=** 9*45% + 7*35% +4*20% = **7.3 (Naturel)**

- **Intelligibilité :**

	Intelligible	Acceptable	Non-intelligible	Totale
Participants	0	0	20	20
Pourcentage	0	0	100%	100%

Tableau 4.5: Evaluation de mesure d'intelligibilité de texte sans signes diacritiques.

MOS **=** 20*100% = **20 (Non intelligible)**

	Intelligible	Acceptable	Non-intelligible	Totale
Participants	13	7	0	20
Pourcentage	65%	35%	0%	100%

Tableau 4.6: Evaluation de mesure d'intelligibilité de texte avec signes diacritiques.

MOS **=** 13*65% + 7*35% + 0*0% =**10.9 (Acceptable)**

La figure 4.7 montre un résultat général des deux critères proposés

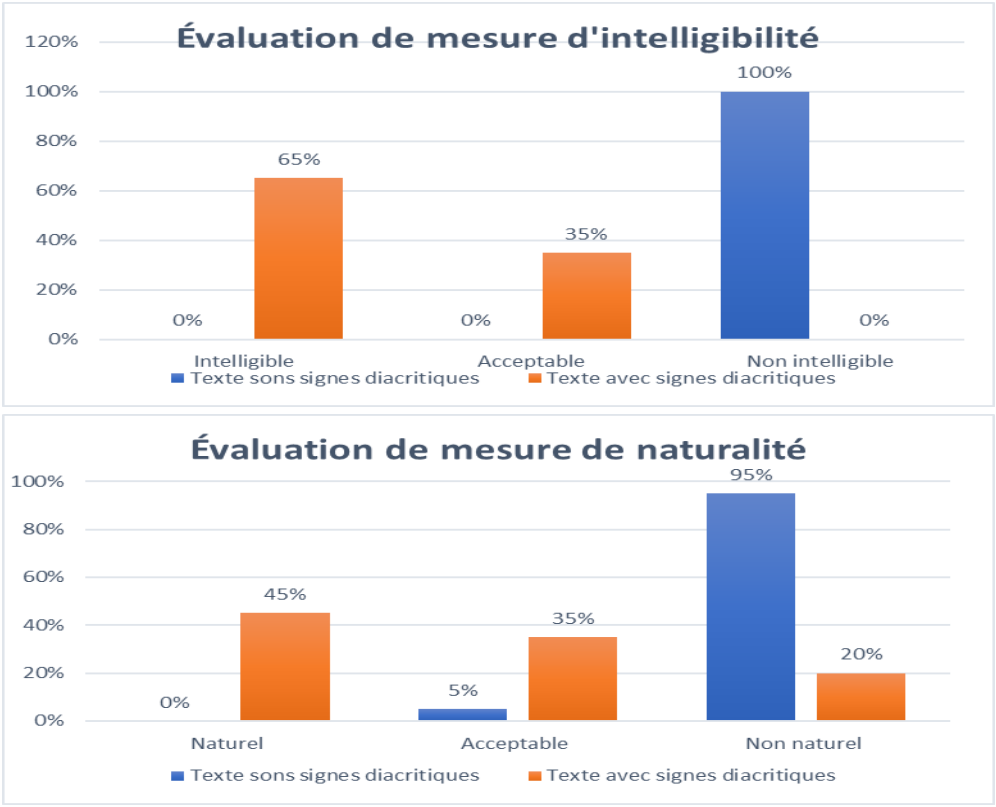

Figure 4.7:Résultats générale de test de qualité.

- **Discussion des résultats**

Sur la base des résultats obtenus lors des tests des phrases, nous pouvons conclure que l'approche proposée a eu une influence très positive sur le caractère naturel et l'intelligibilité des sons vocaux synthétiques. Près de 64% des sujets du test ont estimé que les mots générés étaient suffisamment ou même très intelligibles et naturels, par conséquent, le système présente une intelligibilité et une qualité acceptables de la parole synthétisée. En général, le résultat a été acceptable pour la plupart des auditeurs et les résultats obtenus sont prometteurs par rapport aux travaux précédents sur la synthèse de la parole Arabe.

4.6 Conclusion :

Dans ce chapitre, nous allons présenter une évaluation de la voix créée par notre plateforme en la testant avec deux systèmes open source afin d'améliorer la qualité, et de répondre au mieux aux besoins des utilisateurs.

Tout d'abord, nous présenterons l'interface de notre système et ensuite la méthodologie de test, y compris les systèmes à évaluer dans le formulaire, la mesure de qualité utilisée (MOS), les phrases spécifiques pour le test et les résultats obtenus à partir du questionnaire présenté. Ensuite, nous testerons la qualité du projet selon 20 participants et nous terminerons par la procédure suivie pour rendre notre projet open source.

Conclusion générale

La parole est l'un des moyens de communication les plus naturels pour les humains, c'est un outil exceptionnel pour le transfert d'informations.

Dans ce projet, nous avons présenté le développement d'une voix synthétique open source. Notre objectif est de créer un système de synthèse vocale selon deux facteurs, à savoir : la génération d'un signal vocal de bonne qualité, et le développement d'un système de synthèse open source pour la langue Arabe.

Ces dernières années, l'apport de l'apprentissage profond a connu un grand engouement, pour cela nous allons proposer une approche pour lier le domaine de la synthèse vocale avec l'intelligence artificielle pour la génération d'une voix avec une qualité proche de celle d'un humain. Nous avons également fait une étude bibliographique sur l'état de l'art de la synthèse vocale, ces techniques existantes, sur l'Arabe standard pour obtenir toutes les informations phonologiques nécessaires à l'implémentation du système. Nous commencerons la phase d'implémentation en extrayant les données des fichiers audio de la BDS de Nawar Halabi[19], puis nous terminerons par la phase d'entraînement du modèle qui comprend différentes étapes que nous présenterons dans l'architecture proposée.

Pour l'étape de synthèse, nous avons utilisé dans notre modèle une technique d'apprentissage profond basée sur les réseaux de neurones convolutifs (CNN).

En fin, nous présenterons un rapport détaillé sur les résultats obtenus à partir de toutes les phrases de test. Par conséquent, notre modèle étudie la possibilité d'intégrer l'apprentissage profond pour améliorer la qualité de notre synthétiseur de parole.

Les résultats obtenus lors du test d'évaluation de la perception témoignent de l'intelligibilité, le naturel et la bonne qualité générée . La majorité des participants étaient satisfaits de la qualité de la voix générée par notre plateforme. Ce fait nous encourage à poursuivre nos recherches dans ce domaine.

Notre synthétiseur vocal n'est pas le premier dans le domaine de la synthèse vocale, néanmoins il est disponible en libre téléchargement et ne nécessite pas de ressources importantes pour son utilisation. La version actuelle est destinée aux programmeurs et aux développeurs. Cette version ouvrira la porte à tous les systèmes avec sortie vocale.

[19] http://en.arabicspeechcorpus.com/

Nous avons rencontré quelques difficultés lors de la réalisation du projet, notamment le manque de corpus Arabe nous empêche de construire d'autres voix synthétiques. De plus il n'est pas facile d'utiliser ces systèmes à cause du manque de documents, dans ce cadre nous avons fait recours à des travaux similaires appliqués pour d'autres langues pour accomplir notre projet.

Comme perspectives nous envisageons de poursuivre et d'améliorer ce travail. Nous pensons créer une bande dessinée sonore avec plusieurs locuteurs, dont des femmes, et avec différents dialectes, notamment le dialecte Algérien. Utiliser la voix générée dans des environnements mobiles. Donner à l'utilisateur la possibilité d'introduire sa propre voix.

Enfin, nous pouvons dire que ce pont entre le monde écrit et le monde oral joue un rôle important dans l'amélioration de la technologie d'aujourd'hui, et que notre plate-forme est toujours porteuse d'espoir pour les malvoyants.

Bibliographie

[1] H. Tebbi, *modélisation de la synthèse vocale par un système expert,* Bab Ezzouar,Alger, Algérie: thèse doctorat en science, Université des Sciences et de la Technologie Houari Boumediene., 2019.

[2] K. Zaabi, *"Implémentation d'une méthode de reconnaissance de la parole sur le processeur de traitement numérique du signal.",* Mémoire de maîtrise électronique, Montréal, École de technologie supérieure., 2004.

[3] T. Duroit, *introduction au traitement automatique de la parole notes de cours / DEC2,* Collection électronique, Faculté polytechnique de Mons., 2000.

[4] T. Redouane, *"La reconnaissance en-ligne du manuscript arabe.",* Oran: Ph.D. dissertation, Université des schiences et de la technologie d'Oran -Mohamed Boudiaf,, 2012.

[5] N. Halabi, *Modern standard arabic phonetics for speech synthesis,,* Angleterre: Ph.D,thèse de doctorat,université de Southampton,, 2016.

[6] N. Mezghani, N. Mitiche et M. Cheriet, «A new representation of shape and its use for high performance in online Arabic character recognition by an associative memory,,» *Journal international sur l'analyse et la reconnaissance de documents (IJDAR),* vol. 7, n° 14, pp. 201-210, 2005.

[7] «phonétique française - fle,,» [En ligne]. Available: http://flenet.unileon.es/phon/phoncours2.html. [Accès le 15 05 2022].

[8] A. Abdulrasoul Salman, «An Acoustic Phonetic Analysis Of Fortis-Lenis Consonants In English And Arabic,» *Journal al-Adab,* n° 108, pp. 33-76, 2014.

[9] F. Lionel, *Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales,* thèse de doctorat, université pierre et marie curie., 2013., 2013.

[10] S. Lemmetty, *Review of speech synthesis technology,* thèse de master, Département de génie électrique et des communications,Université Helsinki de technologie, 1999.

[11] H. Oussama et H. Karim, *développement d'une voix arabe open synthétique source,* mémoire de master, département Informatique, université Saad Dahlab, Blida., 2020.

[12] H. Tebbi, *La transcription orthographique phonétique en de la symthèse de la parole à pafiir d'un texte en Arabe Standard .*, Mémoire de Magister, département d'informatique, Université Saad Dahleb de Blida, Algérie,, 2007.

[13] C. Lecorgne, *Validation d'un processus de création de voix contextuelle et intégration de nouvelles langues à une application de synthèse vocale grand public,* Université FR Langage lettres et arts du spectacle, information et communication (LLASIC): Mémoire de master 2 professionnel, Spécialité Industries de la langue, France, 2015.

[14] D. Touahri, *synthèse polyphonique de l'arabe standard.,* Blida: mémoire de magistère,Université Saad Dahleb , 2008.

[15] A. Charbel, *Réalisation d'un système de synthèse vocale de la langue Fon,* Bénin: mémoire de fin de formation,Université d'abomey-calavi (UAC), 2012.

[16] J. Bensesty, M. Sondhi et Y. Huang, Springer handbook of Speech Processing, 1st ed.Secaucus, New Jersy: Spring-Verlag,2008, 2008.

[17] S. Mehtouk et S. Rekkis, *Conception et réalisation d'un assistant vocal pour les maisons intelligentes,* Mémoire de master, Département d'informatique, Université Saad Dahleb de Blida,Algérie., 2021.

[18] S.-C. Guillaume, Apprendre_le_ML_en_une_semaine, 2019.

[19] H. Addaoud, *Classification des défauts multiples détectés par les ultrasons en utilisant les réseaux de neurones convolutifs,* Mémoire de master, Département d'électronique, Université Saad Dahleb de Blida1, Algérie., 2020.

[20] G. Saint-Cirgue., «Introduction au Machine Learning,» 2019. [En ligne]. Available: https://machinelearnia.com/machine-learning-introduction/. [Accès le 11 05 2022].

[21] H. Mezaache, *Les réseaux de Neurones formels Et Les systèmes Neuro-Flous Pour l'apprentissage par renforcement,* Mémoire de magister, Département d'électronique, Université Hadj Lakhdar,Batna,Algérie., 2008.

[22] C. C. Aggarwal, Neural Networks and Deep Learning, Springer, 2018.

[23] T. Claude, *les reseaux de neurones artificiels introduction au connexionnisme cours, exercices et travaux pratiques,* N. Giambiasi. ffhal-01338010f EC2, Collection de l'EERIE,, juillet 1992.

[24] S. Khadimally, Applications of Machine Learning and Artificial Intelligence in Education”,, Progrès des technologies éducatives et de la conception pédagogique, Université nord-américaine,USA, 2022.

[25] A. Neena et M. Geetha, «A Review on Deep Convolutional Neural Networks,» chez *Conférence internationale sur la communication et le traitement du signal*, India, 2017.

[26] B. Carolin, H. Nico, H. Bailan, J. Haris, P. Marianna, S. Vectoria, T. Xiao-Yin, Y. Rui, W. Joshua, A. Mattias, H. Christian et S. Daniel, Approches modernes en traitement du langage naturel,, LMU Munich, Séminaire étudiant, semestre d'été,, 2020.

[27] S. Kostadinov, *Recurrent Neural Networks with Python Quick Start Guide,* Éditions Packt,, 2018.

[28] K. Alexey, P. Rodmonga et M. Iosif, *Speech and Computer: 19th Conférence International , SPECOM 2017, Hatfield, UK, Septembre 12-16, 2017,,* 2017.

[29] «DataScientest,» [En ligne]. Available: https://datascientest.com/convolutional-neural-network. [Accès le 20 09 2022].

[30] «https://ma-petite-encyclopedie.org/accueil,» [En ligne]. Available: https://ma-petite-encyclopedie.org/accueil. [Accès le 10 09 2022].