الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

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

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière : Électronique
Spécialité : Electronique des systèmes embarqués

Présenté par

ZOUAOUI Wissam

&

BENZERROUK Imene

Vers un système de synthèse de la parole séquence à séquence pour la langue arabe

Proposé par : Dr AMROUCHE Aissa & Dr Abed Ahcene

Année Universitaire 2020-2021

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

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

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière : Électronique
Spécialité : Electronique des systèmes embarqués

Présenté par

ZOUAOUI Wissam

&

BENZERROUK Imene

Vers un système de synthèse de la parole séquence à séquence pour la langue arabe

Proposé par : Dr AMROUCHE Aissa & Dr Abed Ahcene

Année Universitaire 2020-2021

Remerciements

Tout d’abord, nous remercions ALLAH Sobhanou de nous avoir accordé la volonté et le courage d'entreprendre et d'achever ce travail.

D’emblée, nous devons une reconnaissance personnelle très profonde à notre promoteur et encadreur Dr. Aissa Amrouche , pour avoir orienté et enrichi notre travail. Nous le remercions pour sa disponibilité, ses précieux conseils, sa confiance malgré nos connaissances. Nous le remercions aussi pour son souci du détail qui a abouti à la réalisation de ce mémoire.

Nous tenons à remercier notre co-promoteur Dr. Ahcen aded , de l’Université SAAD DAHLEB, pour sa disponibilité ses conseils et son sens d’écoute et d’échange.

Nous remercions tout le personnel et les chercheurs Centre de Recherche Scientifique et Technique pour la Langue Arabe (CRSTDLA)

Nos remerciements vont également aux membres du jury pour avoir accepté d’examiner notre travail et de l’enrichir par leurs propositions.

Nous souhaitons aussi adresser nos remerciements à tous les enseignants, le corps professoral et administratif de l’Université de BLIDA qui a contribué à la réussite de nos études universitaires.

Merci…

ZOUAOUI Wissam

BENZERROUK Imen

Dédicace

Avec l'aide d'Allah le tout puissant

On a pu terminer ce modeste travail.

En témoignage de l'amour et du respect qu'on a pour eux, je dédie ce travail à :

A celui qui m'a indiqué la bonne voie en me rappelant que la volonté fait toujours les grandes personnes...

A mon cher père.

A celle qui a attendu avec patience les fruits de ses efforts et sa bonne éducation...

A ma mère d'amour.

Que Dieu vous garde.

Je dédie aussi ce modeste travail :

A mes chères sœurs

A ma famille

À M^{r} HADJAB Mohammed pour son soutien et son encouragement

Et à tous mes ami(e)s

Et à tous ceux qui nous ont soutenues de près ou de loin pour atteindre notre objectif.

BENZERROUK Imene

Dédicace

Du profond de mon cœur, je remercie Allah le tout puissant.

À Ma chère mère, que nulle dédicace ne puisse exprimer mes sincères sentiments pour sa patience illimitée, son encouragement continu, son aide, en témoignage de mon profond amour et respect pour ses sacrifices. Que dieu le tout puissant te protège et te garde à mes côtés.

À Mon père, qui a toujours prié pour moi, ma soutenu et ma épauler pour que je puisse atteindre mes objectifs. Que dieu te protège et te garde pour moi. Que ce travail soit le témoignage de ma gratitude et de mon affection.

À mes chers frères Abdellah et youcef et ma chère sœur amina , en témoignage de mon affection fraternelle, de ma profonde tendresse et reconnaissance, je vous souhaite une vie pleine de bonheur et un avenir radieux plein de réussite

Je tiens à remercier spécialement ma chère cousine Imen mes copines et aussi lekhal Merci pour tout le soutien que vous m'avez apporté.

ZOUAOUI Wissam

ملخص:

تتكون أنظمة تحويل النص إلى كلام بشكل عام من عدة مراحل، مثل تحليل النص والنموذج الصوتي ووحدة التوليف الصوتي.

غالبًا ما تتطلب هذه الخطوات خبرة واسعة في المجال وقد تحتوي على تصميم هش لوحداتها النمطية. في الآونة الأخيرة، تم تطوير نماذج تشكيل الكلام العصبي من سلسلة إلى سلسلة عن طريق نص معين وإثبات أنها تحقق طبيعية مماثلة للكلام البشري . نريد في هذا العمل أن نساهم في تحقيق مُركِّب الكلام من سلسلة إلى سلسلة خاص باللغة العربية.

كلمات المفاتيح:

تركيب الكلام، تحويل النص إلى كلام، سلسلة إلى سلسلة، اللغة العربية

Résumé :

Les systèmes de synthèse de la parole se composent généralement de plusieurs étapes, telles que, analyse de texte, un modèle acoustique et un module de synthèse audio.

Ces étapes nécessitent souvent une vaste expertise du domaine et peuvent contenir une conception fragile de ses modules. Récemment, des modèles de synthèse vocale neuronale séquence à séquence formables à partir de corpus ont été développés et démontrés pour atteindre une naturalité comparable à la parole humaine. Dans ce travail nous voulons contribuer à la réalisation d'un synthétiseur de la parole sequence-to-sequence (seq2seq) pour la langue Arabe.

Mots clés : synthèse de la parole, TTS, séquence à séquence (seq2seq), Langue Arabe.

Abstract :

Text-to-speech systems generally consist of several stages, such as, text analysis, an acoustic model and an audio synthesis module.

These steps often require extensive domain expertise and may contain a fragile design of its modules. Recently, sequence-to-sequence neural speech synthesis models formable from corpora have been developed and demonstrated to achieve naturalness comparable to human speech. In this work we want to contribute to the realization of a sequence-to-sequence (seq2seq) speech synthesizer for the Arabic language.

Keywords : speech synthesis, TTS, sequence to sequence (seq2seq), Arabic language.

Liste des abreviations :

AS (Arabe Standard)

BOS (Bull Operating System)**,**

CRF (Conditional Random Fields)

DNN (Deep Neural Model)

HMM (Hidden Markov Model)

GRU(*Gated Récurrents Unit*)

GAN (Generative Adversarial Networks)

LSTM (Les *Long Short-Term Memory)*

MBROLA (Multi-Band Resynthesis OverLapAdd)

MSA (Modern Standard Arabic)

PSOLA(Pitch SynchronousOverlapAdd)

PNL (Programmation neurolinguistique)

RNN(*Recurrent Neural Network)*

SAT (Synthèse A partir de Texte)

SEQ2SEQ (Sequence To Sequence)

TOP (Orthographique-Phonétique)

TTS (Texte To Speech)

TD-PSOLA (Time- Domain Pitch-Synchronous Overlap-and-Add)

Table des matières :

Table des matières

Table des matières

Liste des figures :

Liste des tableaux:

Introduction générale

Un système de synthèse vocale à partir du texte (Text-To-Speech synthesis) prend en entrée une forme textuelle et produit en sortie le signal de parole correspondant à une vocalisation de ce texte. Elle permet de convertir un texte donné en un signal audio de parole.

Le but de la synthèse de la parole à partir du texte est de calculer automatiquement un signal de parole correspondant à un énoncé écrit. Les sources du texte prononcé peuvent être diverses : lecture de journaux, système de réponse vocale, systèmes d'information, voire saisie au clavier de l'ordinateur. Les premières tentatives de construction de synthétiseur vocal mécanique ont commencé en 1700. Au cours de la dernière décennie, la synthèse vocale est devenue si naturelle qu'un auditeur croirait écouter une voix humaine. Il a fallu beaucoup de temps et de travaux pour obtenir ce résultat. Au début l'idée était d'utiliser des modèles physiques du tractus vocal en utilisant des synthétiseurs de formants. De nombreuses années de recherche ont permis de perfectionner l'encapsulation des propriétés acoustiques du tractus vocal. Puis avec l'aide de technologie informatique plus puissante, il est devenu viable d'utiliser directement des extraits de discours enregistrés et de le coller ensemble pour créer de nouvelles phrases sous la forme de synthétiseurs de concaténation. Maintenant, avec les progrès de la technologie, nous pouvons utiliser des méthodes probalistes et paramétriques pour générer une voix synthétisée de haute qualité avec des technologies telles que les modèles basés sur les réseaux de neurones comme le réseau de neurones récurrent (RNN) qui convertit une séquence de donnée d'un domaine en entrée vers une nouvelle séquence de donnée dans un autre domaine en sortie.

L'un des modèles qui utilise le RNN est le modèle séquence-à-séquence (sequence to sequence (seq2seq)) qui est généralement implémenté en utilisant deux RNN, un premier réseau est un encodeur et le second est un décodeur, on parle de l'architecture encodeur-décodeur.

La synthèse de la parole est un domaine pluridisciplinaire. Elle lie l'informatique à la linguistique et au traitement du signal. Elle permet de donner naissance à des systèmes TTS qui sont cependant très spécifiques de la langue utilisée.

La langue arabe est très complexe du point de vue traitement automatique, du fait des caractéristiques intrinsèques de son écriture, qui le plus souvent est dépourvue de voyelles. Ce qui explique le nombre limité des systèmes TTS pour la langue arabe.

Le but de notre travail est d'étudier le principe de cette méthode de synthèse puis de contribuer à la mise en place de quelques éléments du système. En effet, la mise en oeuvre d'un système de synthèse est une tâche très difficile qui nécessite la réalisation de plusieurs modules comme la base de données, la transcription orthographique phonétique, la sélection automatique des unités dans la base de données et bien d'autres traitements afin d'obtenir un signal de parole synthétique plus ou moins fidèle au texte d'entrée.

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

Dans le second chapitre, nous présenterons la méthode séquence-à-séquence en détails.

Le chapitre trois est consacré aux particularités de la langue arabe et aux problèmes rencontrés en traitement automatique de cette langue ;

Le chapitre quatre portera sur notre contribution dans cette étude dont le but est une meilleure compréhension du fonctionnement de la synthèse séquence-à-séquence. Elle consiste en la mise en oeuvre d'un système séquence-à-séquence.

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

1.1. Introduction :

Le signal de parole est le résultat de l'excitation du conduit vocal par un train d'impulsions ou un bruit donnant lieu respectivement aux sons voisés et non voisés (figure (1) [1]. Dans le cas des sons voisés, l'excitation est une vibration périodique des cordes vocales suite à la pression exercée par l'air provenant de l'appareil respiratoire. Ce mouvement vibratoire correspond à une succession de cycles d'ouverture et de fermeture de la glotte. Le nombre de ces cycles par seconde correspond à la fréquence fondamentale F0. Quant aux signaux non-voisés, l'air passe librement à travers la glotte (du moins pas dans tout le conduit vocal) sans provoquer la vibration des cordes vocales.

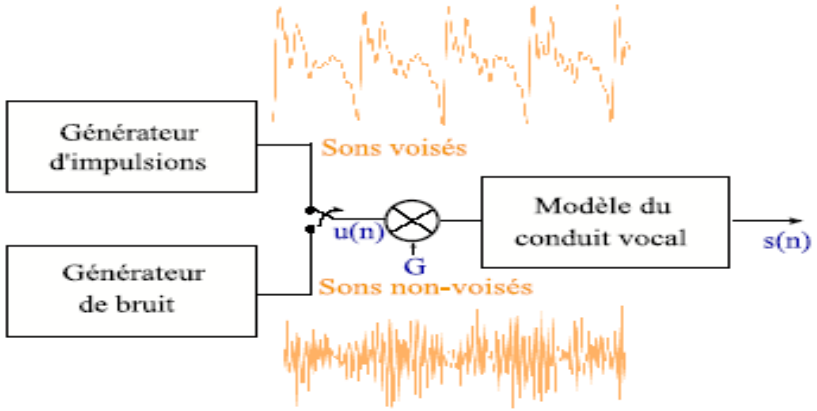

Figure 1. 1: Modèle simple de production de la parole

Le signal de parole est un vecteur acoustique porteur d'informations d'une grande complexité, variabilité et redondance. Les caractéristiques de ce signal sont appelées traits acoustiques. Chaque trait acoustique a une signification sur le plan perceptuel.

Le premier trait acoustique est l'énergie qui correspond à l'intensité sonore. Elle est habituellement plus forte pour les segments voisés de la parole que pour les segments non voisés.

Le deuxième trait est la fréquence fondamentale, fréquence de vibration des cordes vocales. Ses variations définissent le pitch qui constitue la perception de la hauteur (ou les sons s'ordonnent de grave à aigu). Seuls les sons quasi-périodiques (voisés) engendrent une sensation de hauteur tonale bien définie.

Le troisième trait est le timbre qui est une caractéristique permettant d'identifier une personne à la simple écoute de sa voix. Le timbre correspond aux renforcements d'énergies visibles sur le spectre d'un signal de parole qui sont appelés « *formants* ». Le nombre de

formants, selon les caractéristiques du conduit vocal, appelé aussi résonateur (césure, volume, forme et ouverture), est variable : d'un seul à (théoriquement) une infinité. Néanmoins, du point de vue perceptif, seuls quelques-uns d'entre eux jouent un rôle central au niveau de la parole. Par exemple, on peut caractériser toute voyelle en ne prenant en compte que ses trois premiers formants. (Pour une réalisation de la voyelle [i] par exemple, les trois premiers formants pourraient se situer respectivement à 300, 2200 et 3000 Hz.). [1]

En effet, la parole est un moyen de communication primordial chez l'homme. Elle est tellement riche en informations que les scientifiques essayent sans cesse de l'analyser afin d'en comprendre les différents aspects. [2] la reconnaissance et la synthèse de la parole sont deux domaines du traitement automatique de la parole. Le fonctionnement d'un système de reconnaissance de la parole consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composes généralement de mots mais aussi plus récemment de phonèmes en contrepartie le système de synthèse de la parole permet la prononciation d'un teste écrit numérise qui est produite par un générateur de phonèmes.

Ces deux domaines font appel aux connaissances de plusieurs sciences : l'anatomie et les fonctions de l'appareil phonatoires et de l'oreille, les signaux émis par la parole, la phonétique, le traitement du signal, la linguistique, l'informatique, l'inélégance artificielle, les statistiques etc. [3].

Dans cette étude, nous nous sommes intéressés à la synthèse de la parole séquence-à-séquence. Avant de la décrire, il convient de faire un petit état de l'art sur l'histoire de la synthèse de la parole, son principe de fonctionnement, ses applications et les méthodes de synthèse de la parole.

1.2. Histoire de la synthèse :

L'histoire de la synthèse à partir du texte est déjà longue : par exemple le premier système autonome de synthèse automatique de la parole en Français, l'Icophone V du LIMSI, date de 1974. Cependant, c'est encore un domaine de recherche très actif. Les travaux actuels portent à la fois sur la compréhension des textes et sur la restitution d'une parole naturelle, personnalisée et expressive. Les applications se multiplient. Le lecteur désireux d'approfondir les notions présentées pourra consulter les références portées. L'architecture générale d'un système de synthèse se compose ainsi de ces deux parties principales Les principaux modules correspondant à ces traitements sont décrits. Bien que tous les exemples cités dans la suite soient

tirés du français, il est important de souligner que les problèmes posés sont similaires pour toutes les langues. Cependant, des différences notables existent en fonction des Preprint, spécificités linguistiques de chaque langue particulière et notamment en ce qui concerne leur :

- Notation graphique (alphabétique (romaine cyrillique, hébraïque, sanskrite, arabe, …), syllabique (coréenne, japonaise …), idéogrammatique (chinoise, japonaise, …),
- Grammaire (agglutinante, flexionnelle, niveaux de langue, …),
- Phonologie et phonétique (système de phonèmes, langues à ton, langues à clics, …)
- Prosodie (durées, intonation, qualité vocale) Les systèmes actuellement développés pour toutes les langues s'inspirent de principes identiques, même s'ils diffèrent pour les corpus, lexiques, analyses et heuristiques linguistiques. [4]

1.3. La synthèse de la parole :

- La synthèse de la parole à partir du texte désigne l'ensemble des traitements permettant à une machine de transformer un texte écrit en message oral. Aucune restriction n'est faite sur la nature des mots à synthétiser (signal, abréviation, chiffres, date…) ni sur la taille du vocabulaire à traiter.
- Le but chercher en SAT (Synthèse A partir de Texte) est la production de la voix synthétique qui imite au mieux la voix humaine tant au niveau de l'intelligibilité des <<sons>> qu'au niveau du naturel. Cette opération fait appel à des connaissances de nature diverses : informatique (architecture logicielle, temps réel…), linguistique (analyse lexical, morphologique, syntaxique…), traitement de signal [4].

1.4. Schéma général d'un système de la synthèse à partir d'un texte :

Un système de SAT se compose en générale de trois parties (figure2), les deux premières parties concernent le traitement de haut niveau permettant le passage de la représentation orthographique du texte en entrée a une représentation phonétique munies d'une description prosodique. La dernière partie englobe le niveau bas du synthétiseur qui permet la génération proprement du signal acoustique [5]

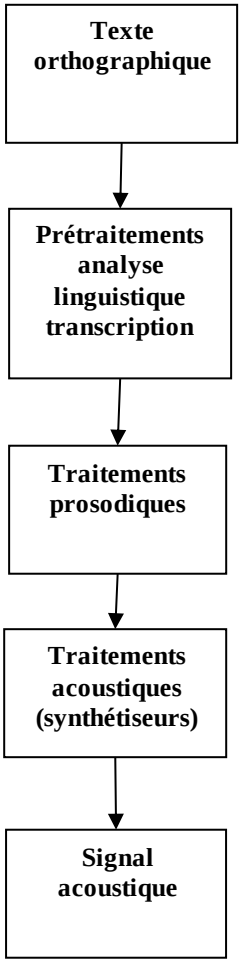

Figure 1. 2 : schéma général d'un système de synthèse à partir du texte

En amont du système, des prétraitements sont effectués pour le découpage du texte, l'élimination des caractères parasites (espace, saut de ligne, etc.) et le traitement d'unités spéciales comme les nombres, les abréviations ou bien les sigles. Une analyse syntaxique plus ou moins élaborée et ensuite appliquée pour l'étiquetage grammatical des mots suivie d'une Transcription Orthographique-Phonétique (TOP) pour définir la prononciation qui leur est associée. Cette description phonétique accompagne des informations syntaxiques et passe au module suivant de calcul des paramètres prosodiques liés à l'intonation et au rythme. Les informations calculées par analyse linguistique permettent de lever certaines des homographes hétérophones en français ou la gestion des pauses et la génération de la prosodie dans différentes langues [5]

1.5. Applications de la synthèse de parole :

De nombreuses applications commerciales intègrent des systèmes de synthèse de parole. À l'heure actuelle, le marché principal de ce type de technique est celui des services de télécommunications. Ces services constituent l'exemple typique de situations dans lesquelles la synthèse de la parole est le seul moyen par lequel un système informatique peut transmettre des informations à ses utilisateurs. Parmi les applications de la synthèse à partir du texte dans le domaine des services de télécommunications, citons :

- Les services de réservation ou de prise de commandes téléphoniques ;
- Les services d'information téléphonique pour lesquels le recours à la synthèse de parole se justifie, surtout lorsque l'information est amenée à évoluer vite, ce qui est notamment le cas pour les services bancaires (avec la fourniture, entre autres de l'état des comptes), les annonces météorologiques et routières, la lecture de mails ou de pages Internet. La synthèse de la parole est aussi utilisée dans des contextes où le nombre des réponses potentielles du système est très important comme dans les applications de renseignements téléphoniques ;
- Les majordomes, assistants personnels, pour les téléphones mobiles ou autres terminaux, qui peuvent lire des messages reçus ou des courriers électroniques ;
- Une application ambitieuse est envisagée à l'heure actuelle avec la téléphonie interprétée qui devrait permettre à deux correspondants ne parlant pas la même langue de dialoguer par téléphone. Cette application fait intervenir plusieurs des grandes problématiques du traitement de la parole – reconnaissance, synthèse –, et bien sûr traduction automatique.

La synthèse de parole est aussi couramment employée dans des situations où l'utilisateur d'un système informatique n'a pas le loisir de consulter un écran, ou bien en complément de l'écran (cabine de pilotage d'un avion, systèmes industriels de fabrication, appareillage médical, etc.). Dans ce type d'applications, le rôle de la synthèse de parole consiste principalement à faire passer des informations brèves comme les messages d'erreurs du système. Les applications dans les systèmes d'information, fixes ou mobiles sont également nombreuses :

- Portails vocaux d'application libre-service ou de sites Internet ;
- Systèmes de navigations ;
- Systèmes de renseignement ;
- Accessibilité des services ;

- Vocalisation de journaux et de livres électroniques ;
- Lecteurs d'écran ;
- Jouets, robots et autres systèmes embarqués ;
- Jeux vidéo ;
- Jeux sérieux, éducation et edutainment.

La qualité accrue des systèmes de synthèse permet maintenant de développer des applications d'apprentissage des langues étrangères qui deviendront les évolutions naturelles des applications actuelles de dictionnaire électronique de poche avec leur capacité à synthétiser des mots ou des phrases dans plusieurs langues. Un autre aspect important des applications de la synthèse de parole à partir du texte concerne les services pour personnes handicapées. Dans ce domaine, le couplage de la synthèse de la parole avec les techniques de reconnaissance automatique de caractères a permis la mise au point de véritables « machines à lire » pour les non-voyants. [4]

1.6. Les principes de la synthèse de la parole

Le processus de synthèse de parole consiste à générer automatiquement un signal de parole à partir d'un texte. Ce processus est fréquemment appelé TTS (Text To Speech).
L'entrée d'un tel système est un texte écrit et sa sortie est un signal de parole. Le mécanisme de synthèse est divisé en deux grandes étapes (figure 3) [6].

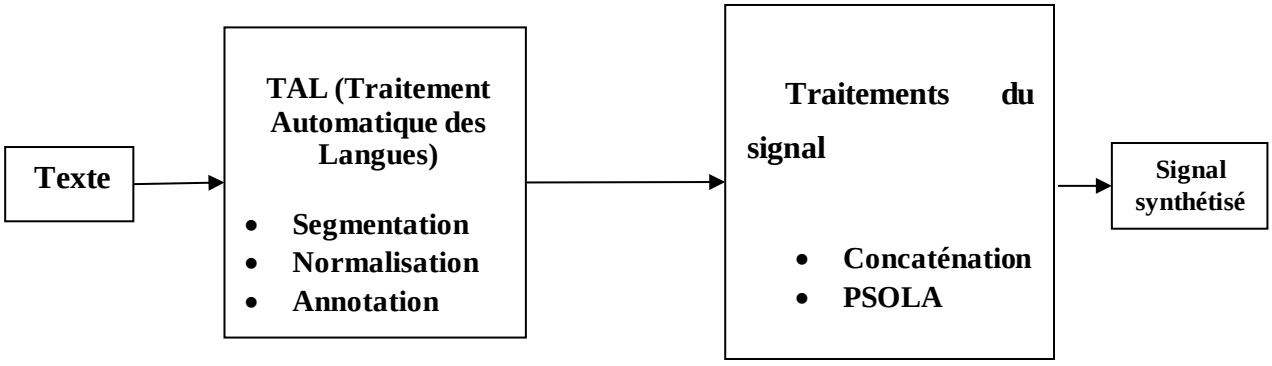

Figure 1. 3 : un aperçu d'un système TTS

La première étape consiste en une application des principes du traitement automatique de la langue naturelle sur le texte, à savoir la segmentation du texte en différents niveaux (phrases, mots, syllabes et phonèmes) [7]

En effet, le texte peut contenir différents types d'informations transcrites dont la procédure de synthèse nécessite la connaissance. Par exemple, le passage des graphèmes aux phonèmes, l'information d'accentuation qui concerne les syllabes. De même, le texte doit être normalisé (il s'agit d'une transformation du texte pour que les informations qu'il contient soient présentées sous une forme canonique pour une application en aval [8]

La deuxième étape génère le signal de parole en se basant sur les informations obtenues par la première étape de traitement du texte que ce soit par une concaténation de segments de parole (dans le cas d'une approche de synthèse par concaténation) ou en utilisant des modèles paramétriques. [6]

1.7. Méthodes de synthèse de la parole :

Différentes approches de synthèse ont été développées pour réaliser la synthèse de la parole à partir du texte. Cette section décrit les approches de synthèses les plus populaires.

1.7.1. Synthèse articulatoire

Les premières approches de synthèse vocale étaient basées sur une imitation du processus physique de la production de la parole. La parole peut être définit comme étant "la réponse du conduit vocal à une ou plusieurs sources de sons". Ainsi conformément à la modélisation source-filtre, l'approche de synthèse articulatoire est basée essentiellement sur une modélisation du conduit vocal.

La synthèse articulatoire de parole modélise le processus de production de la parole naturelle aussi précisément que possible. Ceci est accompli en créant un modèle synthétique de physiologie humaine. La modélisation géométrique du conduit vocal permet de simuler les mouvements des articulateurs

Ceci fournit un moyen pour tirer profit des propriétés du mécanisme de la production de la parole et de la phonétique. La précision d'un tel modèle est étroitement liée à l'acquisition des données et les mesures faites sur le conduit vocal. Pour ce faire, il existe différentes méthodes qui dépendent de l'application du synthétiseur et de certains facteurs de sécurité et de précision [6].

1.7.2. Méthodes statiques :

Ce sont des méthodes qui consistent en des mesures instantanées du conduit vocal. La principale caractéristique de ces méthodes est que seuls des échantillons isolés d'articulation peuvent être obtenus, mais elles sont incapables de représenter le mouvement. En outre, un problème commun pour l'acquisition des données est généralement lié à une articulation prolongée, ce qui peut entraîner des résultats non naturels. Cela peut être causé par la fatigue du participant, mais aussi par son anticipation à prolonger l'articulation et d'autres facteurs liés à la méthode. Très souvent, les articulations peuvent être vérifiées indirectement en comparant le son produit lors de l'acquisition de données avec le son produit dans des conditions plus naturelles. Pendant le traitement des données, les artefacts peuvent être supprimés en fixant le système de coordonnées sur des os plutôt que sur la position du sujet dans l'appareil de mesure [6].

1.7.3. Méthodes dynamiques :

Ce sont des méthodes capables de détecter le mouvement articulatoire. Cette capacité a un coût en résolution spatiale ou nécessite de limiter la méthode à deux dimensions ou à un ensemble de points. Certaines considérations relatives aux méthodes statiques sont également valables pour les méthodes dynamiques. En particulier, la fatigue et les mouvements indésirables du sujet pendant l'acquisition sont des problèmes universels pour l'acquisition de données. En revanche, certaines sources d'erreur sont plus courantes avec la méthode dynamique qu'avec les méthodes statiques. L'une des plus importantes est l'articulation non naturelle résultant d'un équipement, qui doit être placé dans la cavité buccale du sujet.

Le modèle géométrique du conduit vocal nécessite la modélisation de trois parties, à savoir la géométrie de base, les paramètres de mouvement et le mécanisme de génération des mouvements :

- Modèle géométrique du conduit vocal.
- Modèle des paramètres du conduit vocal.
- Modèle de mouvement.

1.7.4. Synthèse par règles

L'approche de synthèse par règles est apparue parmi les premières techniques de synthèse de parole à partir d'un texte. Cette méthode exige la compréhension des mécanismes de perception et de production de la parole. La technique de synthèse par formants est la plus utilisée des techniques de synthèse par règles.

Les formants sont les zones fréquentielles d'enveloppe maximale, ils correspondent aux fréquences de résonance de la fonction de transfert du conduit vocal. L'objectif de cette technique est de créer un signal de parole en se basant sur des caractéristiques des formants (amplitudes, fréquences centrales, largeurs de bande) et sur des règles d'évolution des formants entre les phonèmes.

Cette technique est basée sur des règles déduites à la suite de l'observation des spectrogrammes de parole naturelle. Ces règles tracent la prononciation des phonèmes et l'évolution temporelle des formants ce qui permet de générer un spectre de signal de parole.
Le processus de synthèse de la parole est décrit par une modélisation source filtre. Cette approche présente l'avantage de l'utilisation d'une taille de mémoire réduite, cependant la qualité de la parole générée ainsi que son intelligibilité sont dégradées par rapport à la parole naturelle.

1.7.5. Synthèse par concaténation

L'approche de synthèse par concaténation ne fait appel à aucune modélisation du conduit vocal. Un ensemble d'unités de parole pré-enregistrées est utilisé, le processus de synthèse est réalisé par une mise bout à bout des unités dans le bon ordre. Le choix des unités et la qualité de la parole enregistrée influent sur la qualité de la parole synthétisée (figure 1.4).

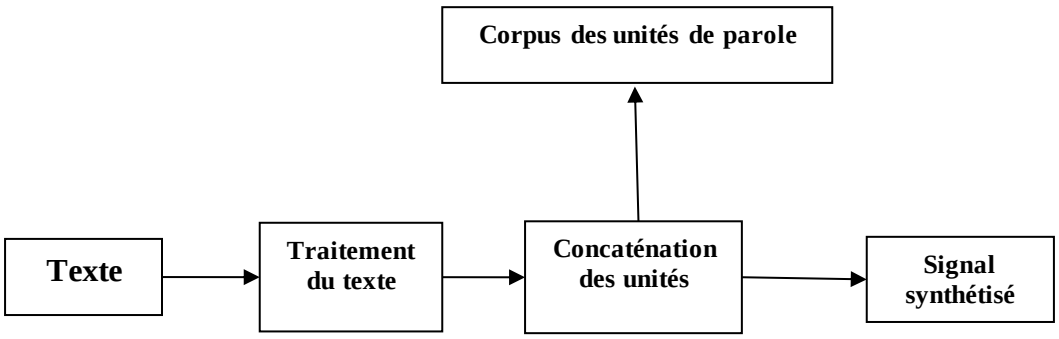

Figure 1. 4 : Principe de synthèse par concaténation

Différentes tailles d'unités de parole peuvent être utilisées. Le choix le plus fréquent correspond aux diphones. Un diphone s'étend du milieu d'un phonème (sa partie stable) au milieu du phonème suivant. La concaténation est réalisée aux frontières des diphones dans des zones stables contrairement aux frontières des phonèmes qui sont des zones instables du fait des phénomènes de coarticulation.

Pendant les années 80, une technique de traitement numérique du signal utilisée pour le traitement de la parole et plus spécifiquement la synthèse vocale basée sur la modification de la fréquence fondamentale et de la durée du signal de parole sans toucher à l'identité des segments du signal appelé PSOLA (Pitch Synchronous Overlap Add) a été développé [9]. Les modifications sont réalisées sans utiliser la modélisation source-filtre de la parole. En pratique la synthèse de parole utilise cet algorithme comme suit :

1. Isolation des périodes de la fréquence fondamentale au niveau du signal original.
2. Faire les modifications nécessaires.
3. Générer le signal final.

Il existe plusieurs variantes de cet algorithme, la plus utilisée est TD-PSOLA (Time- Domain Pitch-Synchronous Overlap-and-Add) [9]. Elle est basée sur une synchronisation du pitch, ce qui veut dire qu'il y a une seule fenêtre d'analyse par période de pitch. La qualité de la parole synthétisée par TD-PSOLA est affectée par la localisation des périodes de pitch ; à la moindre erreur de localisation, la qualité sera dégradée.

Basé sur l'algorithme TD-PSOLA, le synthétiseur MBROLA (Multi-Band Resynthesis OverLapAdd) a été conçu [10]. Le but principal était de trouver une solution au problème de localisation des périodes de pitch. MBROLA a recours à une technique de synthèse basée sur une modélisation harmonique/bruit ; il n'est pas nécessaire que ces positions soient cohérentes d'une trame à une autre. Pendant l'analyse, les trames sont retrouvées comme avant, mais lors de la synthèse, les phases sont ajustées de sorte que chaque trame de la base de données aura la phase correspondante. Cette étape permet d'ajuster efficacement toutes les périodes de pitch de manière à se trouver dans les mêmes positions relatives dans les trames [6].

1.7.6. Synthèse par sélection d'unités

À la suite des évolutions des mémoires de calculateurs, les méthodes de concaténation ont évolué. Au début, il y avait un unique exemplaire de chaque unité de parole (diphone). Par la suite, une nouvelle version de la synthèse par concaténation appelée synthèse par sélection d'unités, permet d'avoir plusieurs exemplaires de chaque unité de parole, dans le corpus enregistré, dans différents contextes phonétiques et prosodiques [11]. Le processus de synthèse est décrit dans la figure 1.5 [6]

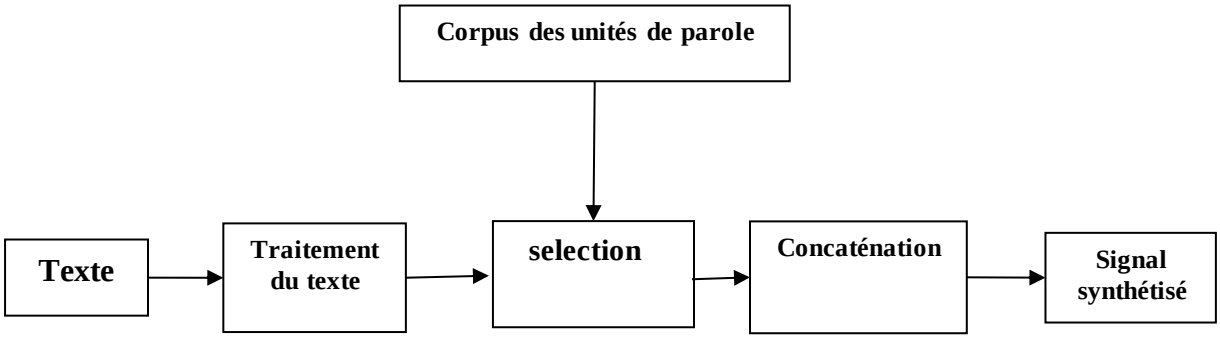

Figure 1. 5 : Principe de synthèse par sélection d'unités

Pendant la phase de sélection, la meilleure séquence des unités est sélectionnée parmi les unités candidates du corpus. Ceci en se basant sur deux critères qui sont un coût de cible pour mesurer la similarité entre les caractéristiques des unités sélectionnées et les caractéristiques désirées, et un coût de concaténation pour mesurer la qualité de la concaténation. La synthèse par sélection d'unités a permis d'obtenir un saut qualitatif de la parole générée comparée à celle obtenue avec une synthèse par concaténation de diphones ou par règles.

1.7.7. Synthèse paramétrique

Cette approche de synthèse de parole est basée sur l'utilisation de modèles paramétriques. Le signal de parole est décrit par un ensemble de paramètres acoustiques extraits à des intervalles de temps réguliers. Ces paramètres contiennent essentiellement les paramètres acoustiques (Exemple : les coefficients Mel-cepstraux et leurs dérivées temporelles, la fréquence fondamentale et la durée des phonèmes). Dans l'approche paramétrique statistique, ces paramètres sont décrits par des statistiques (les moyennes, les variances, les fonctions de densité de probabilité) pour représenter la distribution de ces paramètres. Un principe général de la synthèse par approche paramétrique est décrit dans la figure 1.6.

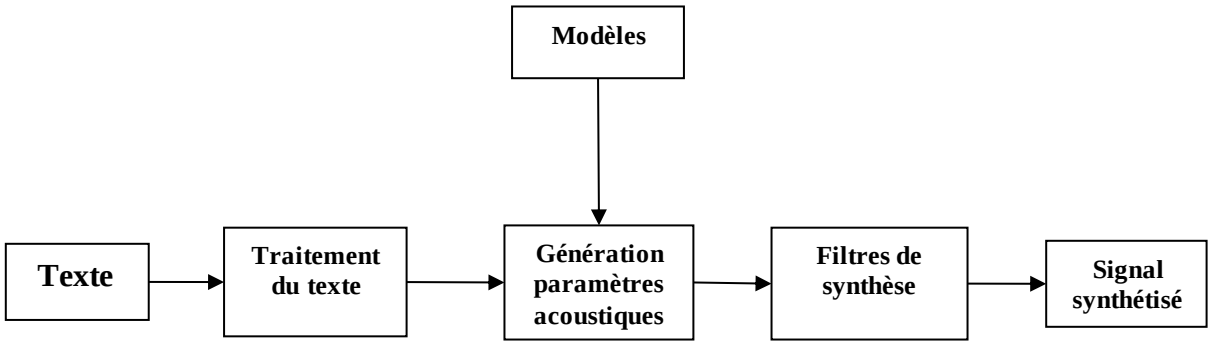

Figure 1. 6 : Principe de synthèse par approche paramétrique

Le processus commence par un traitement du texte écrit (étape commune dans toutes les approches de TTS). L'étape suivante consiste en une prédiction des paramètres acoustiques qui seront traités par un synthétiseur (filtre de synthèse ou un vocodeur) afin de générer le signal de parole correspondant au texte à synthétiser.

Une première version de cette approche était basée sur l'utilisation des HMM (Hidden Markov Model)[12]. Le système résultant présente les avantages de l'utilisation d'une quantité de mémoire réduite par rapport à l'approche de synthèse par concaténation, ainsi que la possibilité de changer les caractéristiques de la voix générée [13]. Depuis quelques années, les HMM ont été remplacés par les réseaux de neurones DNN (Deep Neural Network) dans l'approche de synthèse paramétrique [14]. L'introduction de cette architecture (DNN) a permis d'améliorer la qualité de la parole générée.

1.7.8. Synthèse par le modèle Séquence à séquence

Un modèle séquence à séquence (seq2seq) est un modèle qui prend une séquence d'éléments (mots, lettres, caractéristiques d'une image …etc.) et en sort une Autre séquence en traduction automatique, une séquence est une série de mots, traites les uns après les autres. Le résultat est donc de même une série de mots.

Seq2seq est une architecture de réseau de neurones artificiels (A qui intègre des réseaux de neurones récurrents. Elle a d'abord été inventée pour faire face à la tache de la traduction automatique .[15]

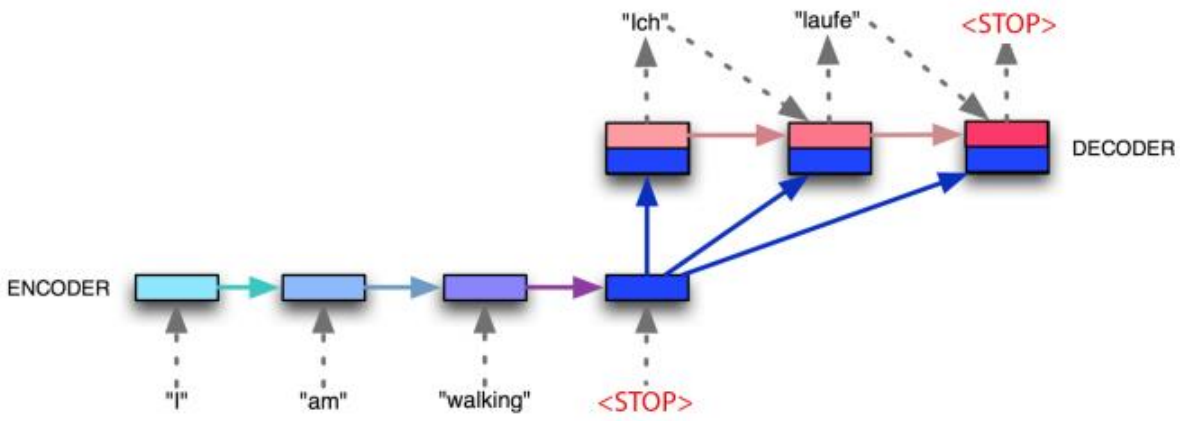

Figure 1. 7 : Un réseau seq2seq simple effectuant la traduction automatique de l'anglais vers l'allemand

Un problème avec cette architecture est que les modèles sont volumineux et nécessitent à leur tour de très grands ensembles de données sur lesquels s'entrainer. Cela a pour effet que la formation du modèle prend des jours ou des semaines et nécessite des ressources de calcul

généralement très coûteuses. En tant que tel, peu de travaux ont été effectués sur l'impact des différents choix de conception sur le modèle et leur impact sur les compétences du modèle.

1.8. Traitement numérique du signal

Pendant les années 80, une technique de traitement numérique du signal utilisée pour le traitement de la parole et plus spécifiquement la synthèse vocale basée sur la modification de la fréquence fondamentale et de la durée du signal de parole sans toucher à l'identité des segments du signal appelé PSOLA (Pitch Synchronous Overlap Add) a été développé [9]. Les modifications sont réalisées sans utiliser la modélisation source-filtre de la parole. En pratique la synthèse de parole utilise cet algorithme comme suit :

1. Isolation des périodes de la fréquence fondamentale au niveau du signal original.
2. Faire les modifications nécessaires.
3. Générer le signal final.

Il existe plusieurs variantes de cet algorithme, la plus utilisée est TD-PSOLA (Time- Domain Pitch-Synchronous Overlap-and-Add) [9]. Elle est basée sur une synchronisation du pitch, ce qui veut dire qu'il y a une seule fenêtre d'analyse par période de pitch. La qualité de la parole synthétisée par TD-PSOLA est affectée par la localisation des périodes de pitch ; à la moindre erreur de localisation, la qualité sera dégradée.

Basé sur l'algorithme TD-PSOLA, le synthétiseur MBROLA (Multi-Band Resynthesis OverLapAdd) a été conçu [10]. Le but principal était de trouver une solution au problème de localisation des périodes de pitch. MBROLA a recours à une technique de synthèse basée sur une modélisation harmonique/bruit ; il n'est pas nécessaire que ces positions soient cohérentes d'une trame à une autre. Pendant l'analyse, les trames sont retrouvées comme avant, mais lors de la synthèse, les phases sont ajustées de sorte que chaque trame de la base de données aura la phase correspondante. Cette étape permet d'ajuster efficacement toutes les périodes de pitch de manière à se trouver dans les mêmes positions relatives dans les trames [6].

1.9. Le Spectrogramme :

Un spectrogramme est une représentation visuelle du spectre de fréquences d'un signal tel qu'il varie dans le temps. Lorsqu'ils sont appliqués à un signal audio , les spectrogrammes sont parfois appelés sonographes, empreintes vocales ou vocogrammes.

Les spectrogrammes sont largement utilisés dans les domaines de la musique , de la linguistique , du sonar , du radar , du traitement de la parole , de la sismologie et autres. Les spectrogrammes audio peuvent être utilisés pour identifier phonétiquement les mots prononcés et pour analyser les différents cris des animaux.

Un spectrogramme peut être généré par un spectromètre optique , une banque de filtres passe-bande , par transformée de Fourier ou par une transformée en ondelettes (auquel cas il est également appelé scalogramme ou scalogramme).

Un spectrogramme est généralement représenté sous la forme d'une carte thermique, c'est-à-dire sous la forme d'une image dont l'intensité est indiquée en faisant varier la couleur ou la luminosité.

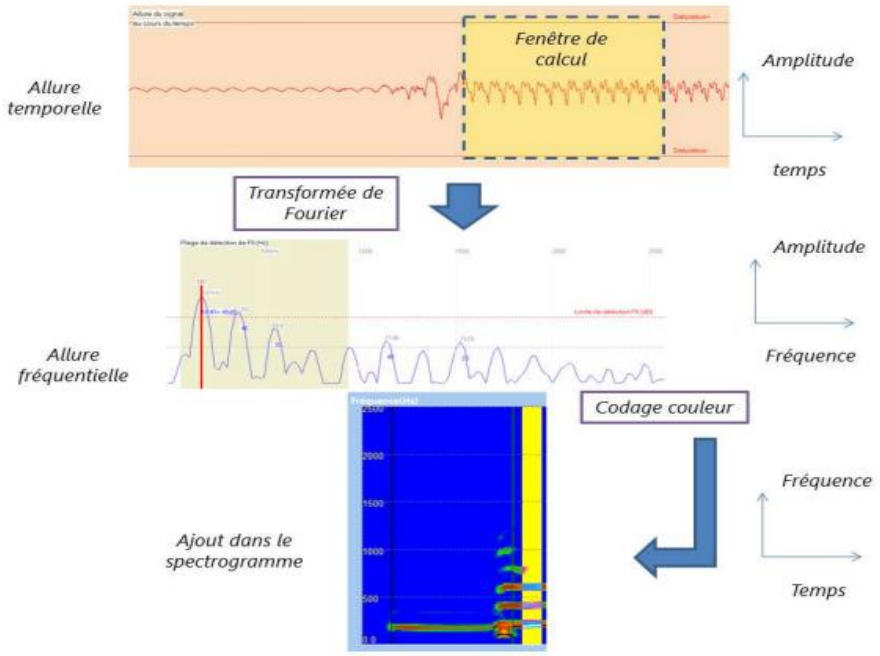

Figure 1. 8 : Exemple d'un spectrogramme

1.9. Conclusion

Dans ce chapitre, nous avons présenté le processus de production de la parole. Les différentes méthodes de la synthèse de la parole et ses applications. Ainsi qu'une brève description de la méthodes Séquence-à-séquence choisie. Dans le chapitre suivant, nous allons donner plus de détail sur cette méthode.

2.1. Introduction :

Dans ce chapitre, nous allons étudier les modèles séquence-à-séquence. Ces modèles ont pour but de générer une séquence d'éléments à partir d'une autre séquence et peuvent donc nous permettre de produire un énoncé diffluent à partir d'un énoncé fluide. Nous allons commencer par présenter différents modèles d'apprentissage automatique, notamment des modèles neuronaux, pouvant réaliser des tâches séquence-à-séquence [16].

2.2. Champs aléatoires conditionnel

Les champs aléatoires conditionnels (CRF pour *Conditional Random Fields*) sont des modèles probabilistes. Ces modèles sont très utiles en traitement automatique des langues car ils permettent, entre autres, de faire de l'annotation séquentielle [17]. En effet, à partir d'une séquence d'entrée $\mathrm{x} = (x_1, \ldots, x_T)$, ces modèles calculent la probabilité d'une séquence d'étiquettes $\mathrm{y} = (y_1, \ldots, y_T)$ selon la formule suivante :

$$\mathrm{p(y|x)} = \frac{1}{\mathrm{Z_\theta(x)}} \exp\left(\sum_{\mathrm{t=1}}^{\mathrm{T}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \theta_\mathrm{k} \mathrm{f_k}\,(\mathrm{y_{t-1}}, \mathrm{y_t}, \mathrm{x_t})\right) \qquad (1)$$

Avec $\mathrm{f_k}$ des fonctions caractéristiques, θ_k les poids associés estimés sur les données d'apprentissage de façon à minimiser le taux d'erreur d'étiquetage et $\mathrm{Z_\theta(x)}$ un facteur de normalisation. Les fonctions caractéristiques donnent généralement un résultat binaire (0 ou 1) indiquant la présence d'une caractéristique donnée parmi les paramètres fournis en entrée [18].

La séquence choisie par le modèle est celle ayant la probabilité $p(y|x)$ la plus élevée. Les CRF sont particulièrement intéressants car ils permettent de prendre en compte les dépendances entre les étiquettes[16].

2.3. Réseaux de neurones

Les réseaux de neurones sont des modèles capables d'apprendre des relations très complexes à partir de données. Comme le montre la figure 1, ce sont des modèles qui

prennent un vecteur d'entrée et le transforme successivement grâce à des couches de neurones jusqu'à produire une sortie, elle aussi sous forme vectorielle. Les neurones sont des cellules qui reçoivent des valeurs numériques en entrée, les transforment via une fonction ***f,*** et transmettent le résultat aux neurones de la couche. La transformation réalisée par un neurone consiste tout d'abord à faire une somme du vecteur d'entrée pondérée par un vecteur de poids **θ**, puis à appliquer une fonction d'activation (par exemple la fonction sigmoïde).

Le but de la phase d'apprentissage est de trouver la pondération $\boldsymbol{\theta}$ qui minimise l'erreur entrela sortie du réseau et celle souhaitée. Pour cela, la technique la plus utilisée est celle de la rétropropagation du gradient qui consiste à corriger les erreurs en fonction de l'importance des éléments quiont participé [16].

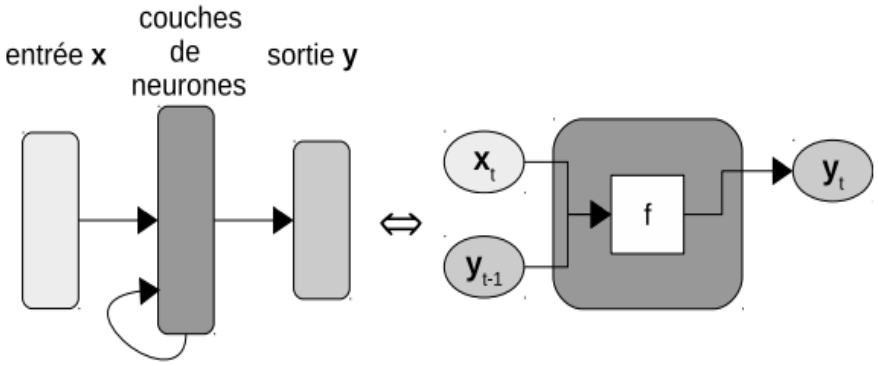

Figure 2. 1. Schéma d'un bloc RNN

On peut représenter un réseau de neurones qui admet **x** en entrée par

$$y = f_\theta(x) \qquad (2)$$

Avec **y** la sortie et $\boldsymbol{\theta}$ les paramètres du réseau. Il existe de très nombreux réseaux de neurones qui diffèrent selon la manière dont les couches sont organisées, les connexions entre les neurones ou encore le fonctionnement interne des neurones.

2.3.1. Réseaux de neurones récurrents

Les réseaux de neurones récurrents (RNN pour *Recurrent Neural Network)* sont des réseaux de neurones particulièrement adaptés aux données séquentielles.

La structure des RNN est semblable à celle des réseaux de neurones classiques. Cependant, la sortie de la couche de neurones correspondant au $(\mathrm{t}-1)$ élément de la séquence est réinjectée en tant qu'entrée de la couche pour l'élément t (figure 2.1). Ainsi, la *t*-ième sortie y_t d'un bloc RNN se calcule comme :

$$y_t = f_\theta\ (x_t\ , y_{t-1}) \qquad (3)$$

Avec : y_{t-1} la $(t-1)$ième sortie, x_t la *t*-ième entrée et $\boldsymbol{\theta}$ les paramètres du réseau.

Ces réseaux permettent donc de prendre en compte une dépendance entre les entrées d'une séquence.

Ils sont beaucoup utilisés dans les travaux de traitement du langage naturel pour plusieurs raisons [19]. Premièrement, dans le langage naturel, le sens d'un mot dépend en partie des mots précédents. De plus, les phrases n'ont pas toutes le même nombre de mots ; or, les réseaux de neurones ont un vecteur d'entrées de taille fixe. Le jeu de chaînage de l'équation (3) permet d'apporter une solution à ces deux problèmes.

La technique utilisée pour l'apprentissage est une variante de celle utilisée pour les réseaux de neurones classiques : la rétropropagation du gradient à travers le temps. Cependant, les RNN connaissent le problème de disparition du gradient [20]. À cause de ce dernier, les systèmes ont du mal à apprendre de longues dépendances. D'autres réseaux de neurones, comme les *Long Short-Term Memory (LSTM)*, permettent de contourner ce phénomène.

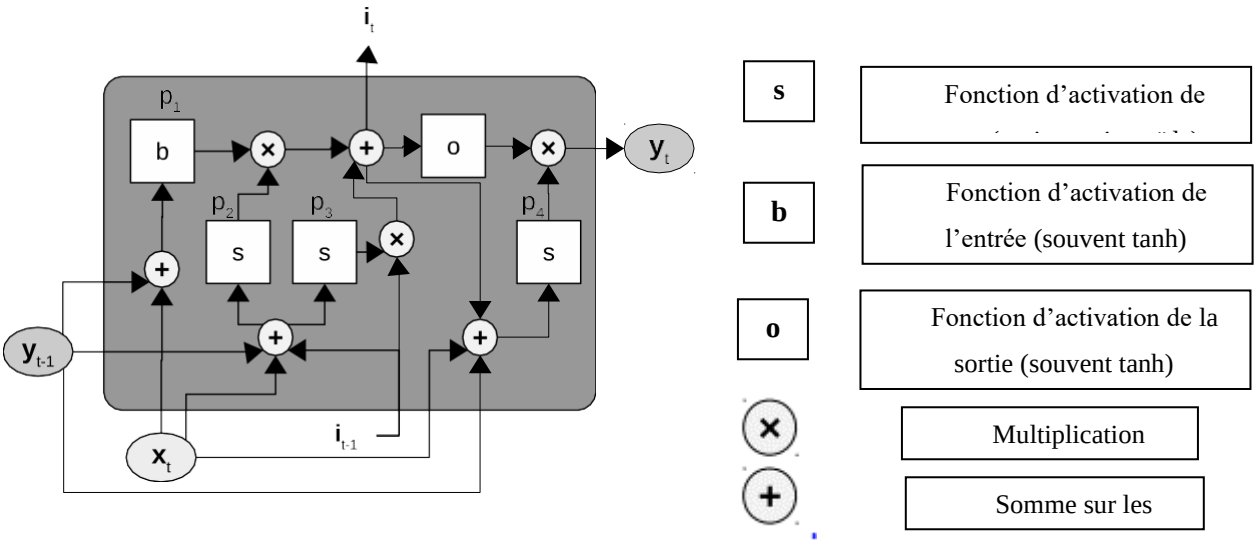

Figure 2. 2. Schéma d'un bloc LSTM. p_1 est appelé bloc d'entrée, p_2 porte d'entrée, p_3 porte d'oubli et p_4 porte de sortie.

2.3.2. Mémoire longue à court terme (LSTM) :

Les *Long Short-Term Memory* (LSTM) sont un type particulier de RNN beaucoup plus complexes que ces derniers [21]. Comme le représente le système d'équations (4), la sortie $\boldsymbol{y_t}$ dépend toujours de la sortie y_{t-1} et de l'entrée $\boldsymbol{x_t}$. Toutefois, elle dépend également d'une nouvelle information récurrente i_{t-1} , appelée mémoire, qui saisit les dépendances à long terme. Pour produire la valeur i_t à l'instant *t*, le bloc LSTM est composé de plusieurs « portes » qui ont pour objectif de conserver, supprimer ou modifier de l'information (voir la figure 2.2).

$$y_t = f_\theta\,(x_t\,, y_{t-1}\,, i_{t-1}) \qquad (4)$$
$$i_t = g_\theta\,(x_t\,, y_{t-1}\,, i_{t-1})$$

Avec $\boldsymbol{\theta}$ les paramètres du réseau. Grâce à cette structure complexe, les LSTM sont efficaces pour capturer les dépendances à long terme, contrairement aux RNN classiques. De plus, ils possèdent leurs avantages (dépendances à court-terme, séquences de longueur variable).

Cependant, l'apprentissage est plus coûteux à cause de ses nombreuses connexions et donc de ses paramètres à estimer.

2.3.3. Unité récurrente fermée (réseau récurrent à portes)

Les *Gated Récurrents Unit* (GRU) sont une autre variante de RNN, introduite récemment dans [22]. Ces réseaux sont plus simples que les LSTM car ils possèdent une porte en moins.

Une étude a comparé les GRU, les LSTM et les RNN classiques [23]. Elle montre que, pour des tâches autres que le traitement du langage naturel, les réseaux GRU et LSTM sont plus performants que les RNN classiques. Toutefois, cette étude ne parvient pas à déterminer lequel des deux premiers est le meilleur. Par conséquent, de par leur plus grande simplicité, les réseaux GRU seront privilégiés lorsque la puissance de calcul disponible est limitée [24].

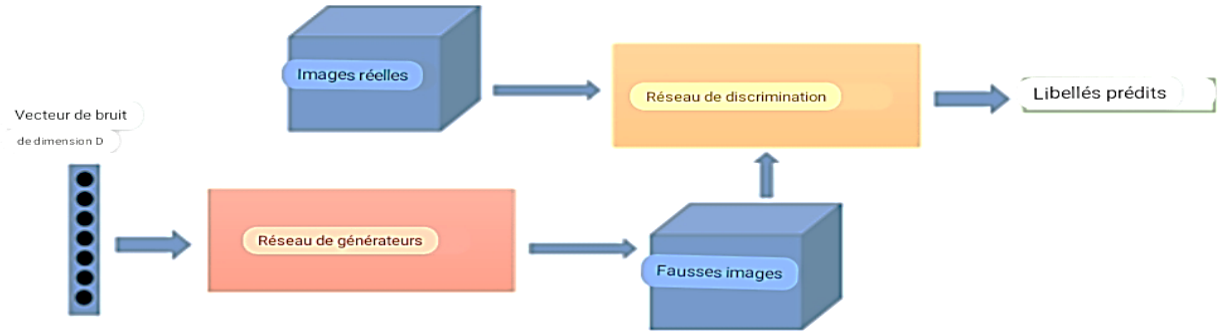

Figure 2. 3. Principe des GAN

2.3.4. Réseaux d'adversaire génératif

Les Generative Adversarial Networks (GAN), font partie de la catégorie des modèles génératifs, c'est-à-dire ayant pour but de créer des données. Comme en témoigne la figure

3, le principe des GAN est de mettre en concurrence deux modèles. Le premier est un réseau générateur qui est chargé de produire des données à partir d'un vecteur de bruit. Le second est un réseau discriminatif qui apprend à différencier les données générées de données réelles fournies par l'utilisateur. Ce modèle peut être utilisé pour produire toutes sortes de données, aussi bien des images que des textes ou des signaux audios [16].

2.4. Mécanisme d'attention

Un réseau de neurones récurrents réalise des prédictions à partir de données séquentielles. Chaque élément apporte une part d'information que le réseau utilise pour produire une sortie. Toutefois, il est possible que certains éléments d'une séquence contiennent davantage d'informations utiles à la prédiction. Le mécanisme d'attention a pour objectif de repérer quels sont les éléments les plus utiles.

Par exemple [25] propose une architecture RNN de type encodeur-décodeur avec un mécanisme d'attention. Un encodeur-décodeur est un modèle en deux composantes qui permet de faire des correspondances entre une séquence d'entrée et une séquence de sortie. La partie encodeur est chargée de représenter la séquence d'entrée en un vecteur de taille fixée, tandis que la partie décodeur utilise cette représentation afin de générer une séquence en sortie. Cependant, tous les éléments de la séquence d'entrée n'ont pas la même importance pour obtenir la séquence de sortie. Le rôle du mécanisme d'attention est donc de donner de l'importance à la position des éléments au sein de la séquence d'entrée. Pour cela, on attribue un poids à chaque élément encodé en entrée. Ces poids sont ensuite mis à jour après chaque sortie produite. Le décodeur calcule alors un vecteur de contexte en faisant une somme, pondérée par ces poids, des éléments encodés. Ce vecteur de contexte permet au décodeur de prêter davantage attention à certains éléments pour produire sa sortie. À chaque sortie produite par le décodeur, le vecteur de contexte est recalculé [16].

2.5. Le séquence -à -séquence

Le modèle séquence à séquence (seq2seq) est un modèle d'apprentissage qui convertit une séquence d'entrée en séquence de sortie. Dans ce contexte, la séquence est une liste de symboles,

correspondant aux mots d'une phrase. Le modèle seq2seq a connu un grand succès dans des domaines tels que la traduction automatique, les systèmes de dialogue, la réponse aux questions et la synthèse de texte. Toutes ces tâches peuvent être considérées comme la tâche d'apprentissage d'un modèle qui convertit une séquence d'entrée en séquence de sortie. [26]

2.5.1. Les notations de séquence

Le modèle seq2seq convertit une séquence d'entrée en séquence de sortie. Soit la séquence d'entrée et la séquence de sortie X et Y. Le i-ème élément de la séquence d'entrée est représenté par x_i, et le j-ème élément de la séquence de sortie est également représenté par y_i,. Généralement, chacun des x_i et le y_i est le vecteur à une chaude des symboles. Par exemple, dans le traitement du langage naturel (NLP), le vecteur à une chaude représente le mot et sa taille devient la taille du vocabulaire.

Réfléchissons au modèle seq2seq dans le contexte de NLP. Soit le vocabulaire des entrées et des sorties $V^{(s)}$ et $V^{(t)}$tous les éléments x_i $y_i$satisfaire $x_i \in$ R$|V^{(s)}|$ et $y_i \in$R$|V^{(t)}|$. La séquence d'entrée X et la séquence de sortie Y sont représentées par les équations suivantes :

$$X = (x_1 , \dots\dots, x_I) = (x_i)_{i=1}^{I} \qquad (5)$$

$$Y = (y_1 , \dots\dots, y_I) = (y_i)_{i=1}^{I} \qquad (6)$$

I et J sont la longueur de la séquence d'entrée et de la séquence de sortie. En utilisant la notation typique de NLP, y_0 est le vecteur à une chaleur de BOS, qui est le mot virtuel représentant le début de la phrase, et y_{J+1} est celui d'EOS, qui est le mot virtuel représentant la fin de la phrase.[17]

2.5.2. Étapes de traitement dans le modèle Seq2seq

Maintenant, pensons aux étapes de traitement dans le modèle seq2seq. La caractéristique du modèle seq2seq est qu'il se compose des deux processus [26].

1. Processus qui génère le vecteur de taille fixe z à partir de la séquence d'entrée X
2. Processus qui génère la séquence de sortie Y à partir de z

En d'autres termes, l'information de X est transmise par z, et $P_\theta(y_j|Y_{<j}), X)$ est en fait calculé par $P_\theta(y_j|Y_{<j}), z)$

Tout d'abord, nous représentons le processus qui génère z à partir de X par la fonction

$$z = \wedge(X) \qquad (7)$$

La fonction ∧ peut être le réseau neuronal récurrent tel que les LSTM.

Deuxièmement, nous représentons le processus qui génère Y à partir de z par la formule suivante :

$$P_\theta(y_j|Y_{<j}), X) = \left(\boldsymbol{h_j^{(t)}}, \boldsymbol{y_j}\right) \qquad (\mathbf{8})$$

$$\boldsymbol{h_j^{(t)}} = \psi\left(\boldsymbol{h_{j-1}^{(t)}}, \quad \boldsymbol{y_{j-1}}\right) \qquad (\mathbf{9})$$

ψ est la fonction pour générer les vecteurs cachés $\boldsymbol{h_j^{(t)}}$, et ϒ est la fonction pour calculer la probabilité générative du vecteur one-hot $\boldsymbol{y_j}$.

Lorsque j=1, $\boldsymbol{h_{j-1}^{(t)}}$ ou $\boldsymbol{h_0^{(t)}}$est z généré par ∧ (X), et $\boldsymbol{y_{j-1}}$ou $\boldsymbol{y_0}$ est le vecteur one-hot de *BOS*.

2.5.3. Architecture modèle du modèle Seq2seq

Dans cette section, nous décrivons l'architecture du modèle seq2seq. Pour simplifier l'explication, nous utilisons l'architecture la plus basique. L'architecture du modèle seq2seq peut être séparée des cinq rôles principaux [26].

a. Couche d'incorporation d'encodeur
b. Couche récurrente de l'encodeur
c. Couche d'incorporation de décodeur
d. Couche récurrente du décodeur
e. Couche de sortie du décodeur

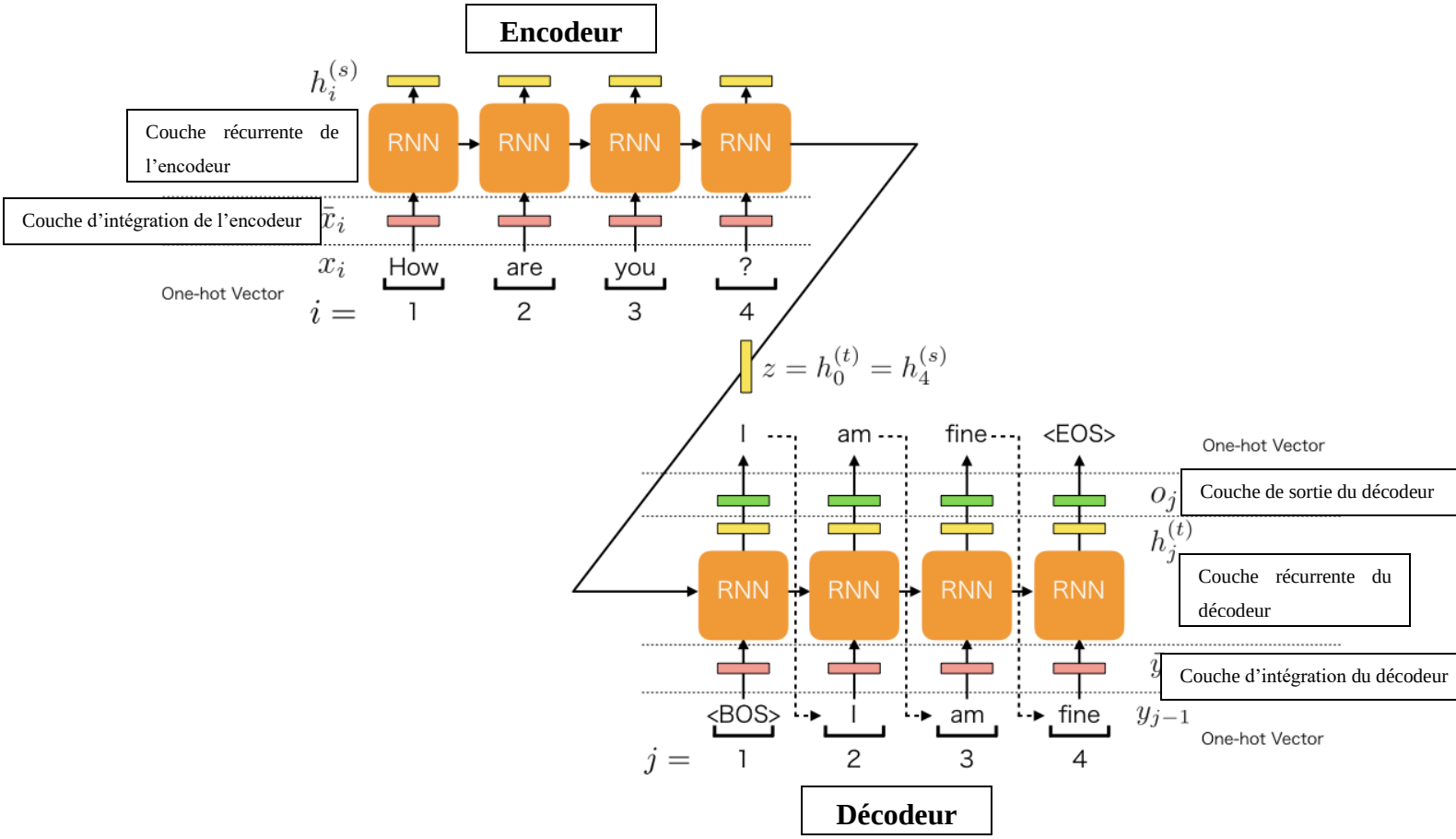

Figure 2. 4. L'architecture de seq2seq

L'encodeur se compose de deux couches : la couche d'intégration et la couche récurrente, et le décodeur se compose de trois couches : la couche d'intégration, la couche récurrente et la couche de sortie.

Dans l'explication, nous utilisons les symboles suivants :

Symbole	Définition
H	La taille du vecteur caché.
D	La taille du vecteur de plongement.
x_i	Le vecteur unique de i-ième mot dans la phrase d'entrée.
$\overline{x_i}$	Le vecteur d'enfouissement de i-ième mot dans la phrase d'entrée.
$E^{(s)}$	Matrice d'enrobage du codeur.
$h_i^{(s)}$	Le i-ième vecteur caché de l'encodeur.
y_j	Le vecteur unique de j-ième mot dans la phrase de sortie.
$\overline{y_j}$	Le vecteur d'enfouissement de j-ième mot dans la phrase de sortie.
$E^{(t)}$	Matrice d'encastrement du décodeur.
$h_j^{(t)}$	Le j-ième vecteur caché du décodeur.

Tableau 2.1. Les symboles utilisés dans les formules

a. Couche d'intégration d'encodeur

La première couche, ou la couche d'intégration du codeur, convertit chaque mot de la phrase d'entrée en vecteur d'intégration. Lors du traitement du i-ième mot dans la phrase d'entrée, l'entrée et la sortie de la couche sont les suivantes :

- L'entrée est x_i : le vecteur one-hot qui représente i-ème mot
- La sortie est $\overline{x_i}$: le vecteur plongeant qui représente i-ème mot

Chaque vecteur de plongement est calculé par l'équation suivante :

$$\overline{x_i} = E^{(s)}\, x_i \qquad (10)$$

$E^{(s)} \in \mathbb{R}^{D \times |V^{(s)}|}$ est la matrice de plongement du codeur.

b. Couche récurrente de l'encodeur

La couche récurrente du codeur génère les vecteurs cachés à partir des vecteurs d'incorporation. Lors du traitement du i-ième vecteur de plongement, l'entrée et la sortie de la couche sont les suivantes :

- L'entrée est $\overline{x_i}$: le vecteur plongeant qui représente le i-ème mot
- La sortie est $h_i^{(s)}$: le vecteur caché du i-ème position

Par exemple, lors de l'utilisation du RNN unidirectionnel d'une couche, le processus peut être représenté par la fonction suivante $\Psi^{(s)}$:

$$h_i^{(s)} = \Psi^{(s)}\left(\overline{x_i}, h_{i-1}^{(s)}\right) \qquad (11)$$

$$= \tanh\left(\mathrm{W}^{(s)} \begin{pmatrix} h_{i-1}^{(s)} \\ \overline{x_i} \end{pmatrix} + \mathrm{b}^{(s)}\right) \qquad (11*)$$

Dans ce cas, nous utilisons la **tanh** comme fonction d'activation.

c. Couche d'intégration du décodeur

La couche d'intégration du décodeur convertit chaque mot de la phrase de sortie en vecteur d'intégration. Lors du traitement de la j-ième mot dans la phrase de sortie, l'entrée et la sortie de la couche sont les suivantes :

- L'entrée est y_{j-1} : le vecteur one-hot qui représente le $(j-1)$-ième mot généré par la couche de sortie du décodeur
- La sortie est $\overline{y_i}$: le vecteur plongeant qui représente le $(j-1)$-ème mot

Chaque vecteur de plongement est calculé par l'équation suivante :

$$\overline{y_i} = E^{(t)}\, y_{j-1} \qquad (12)$$

$E^{(t)} \in R^{D\times|V(t)|}$ est la matrice de plongement du codeur.

d. Couche récurrente du décodeur

La couche récurrente du décodeur génère les vecteurs cachés à partir des vecteurs d'incorporation. Lors du traitement du j-ième vecteur de plongement, l'entrée et la sortie de la couche sont les suivantes :

- L'entrée est $\overline{y_i}$: le vecteur de plongement.
- La sortie est $h_j^{(t)}$: le vecteur caché de j-ème position.

Par exemple, lors de l'utilisation du RNN unidirectionnel d'une couche, le processus peut être représenté par la fonction suivante **Ψ(t) :**

$$h_i^{(t)} = \Psi^{(t)}\left(\overline{y_i}, h_{i-1}^{(t)}\right) \qquad (13)$$

$$= \tanh\left(\mathrm{W}^{(t)} \begin{pmatrix} h_{i-1}^{(t)} \\ \overline{y_i} \end{pmatrix} + \mathrm{b}^{(t)}\right) \qquad (13*)$$

Dans ce cas, nous utilisons la tanh comme fonction d'activation. Et nous devons utiliser le vecteur caché de l'encodeur de la dernière position comme vecteur caché de la première position du décodeur comme suit :

$$h_0^{(t)} = z = h_I^{(t)} \qquad (14)$$

e. Couche de sortie du décodeur

La couche de sortie du décodeur génère la probabilité du j-ième mot de la phrase de sortie du vecteur caché. Lors du traitement du-ième vecteur de plongement, l'entrée et la sortie de la couche sont les suivantes : [26]

- L'entrée est $h_j^{(t)}$: le vecteur caché de j-ème position.
- La sortie est p_j : la probabilité de générer le vecteur one-hot y_j du j-ème mot.
- $p_j = P_\theta\left(y_j \middle| Y_{<j}\right) = softmax\,(\,\boldsymbol{o_j}).y_j.$

2.5.4. Principe de fonctionnement :

L'architecture séquence à séquence se compose d'un encodeur**,** qui crée une représentation interne du signal d'entrée, et d'un décodeur, qui transforme cette représentation en un spectrogramme Mel. Un élément très important du réseau est le PostNet, conçu pour améliorer le spectrogramme généré par le décodeur. [27]

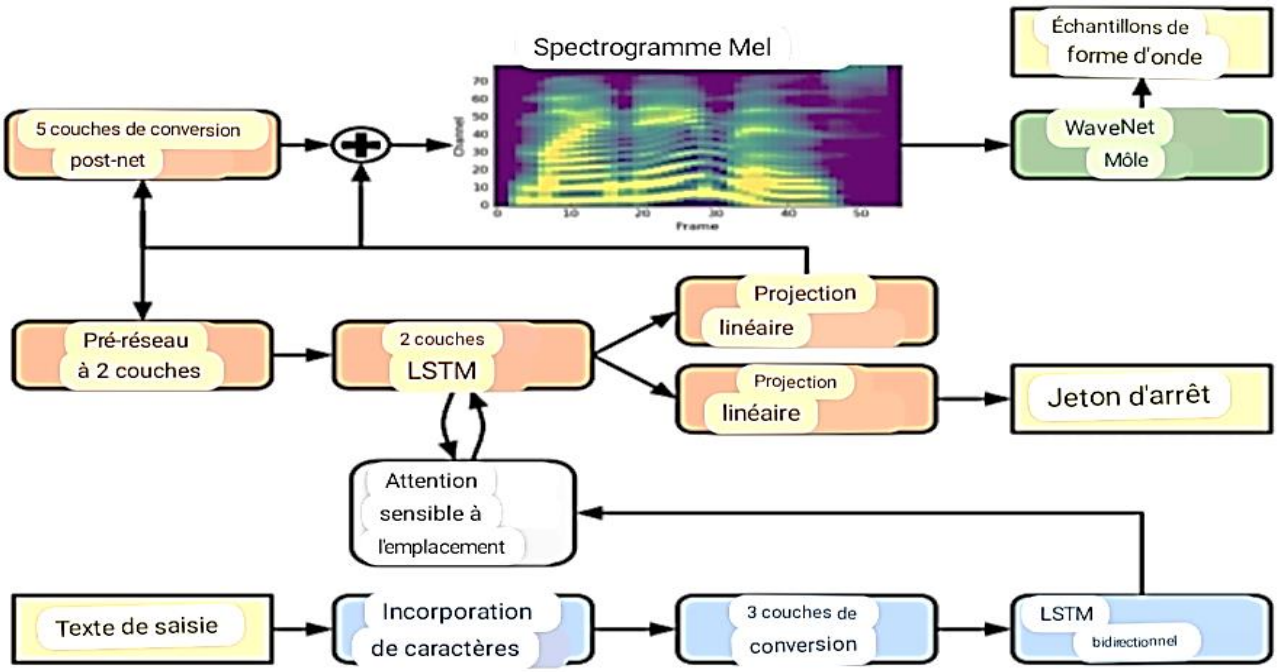

Figure 2. 5. Les blocs du modèle seq2seq

Examinons en détail les blocs réseau et leurs modules.

La première couche de l'encodeur est la couche d'intégration, qui crée des vecteurs à 512 dimensions basés sur une séquence de nombres naturels qui représentent des symboles. Plus loin, les vecteurs d'enfouissement sont dirigés dans un bloc de 3 couches convolutives unidimensionnelles. Chaque couche comprend 512 filtres d'une longueur de 5. Cette valeur est une bonne taille du filtre car elle capture un certain caractère, ainsi que deux voisins précédents et deux voisins suivants. Chaque couche convolutive est suivie d'une normalisation par mini-lot et d'une activation Relue.

Les tenseurs obtenus après le bloc convolutif sont dirigés vers des couches LSTM bidirectionnelles, chacune par 256 neurones. Les résultats avant et arrière sont concaténés. Le décodeur a une architecture récurrente. Ainsi, la sortie de l'étape précédente (une image du spectrogramme) est utilisée à chaque étape suivante.

Un autre élément crucial du système est le mécanisme de l'attention douce - une technique relativement nouvelle et populaire. À chaque étape de décodage, « attention » forme le vecteur de contexte et met à jour le poids de l'attention, en utilisant les données suivantes :

- La projection de l'état caché précédent du réseau RNN du décodeur sur la couche entièrement connectée,
- La projection de la sortie des données du codeur sur une couche entièrement connectée,
- Le poids d'attention additif (cumulé à chaque pas de temps du décodeur).

Attention définit la partie des données du codeur qui doit être utilisée au pas de décodeur en cours [27].

2.6 Conclusion

Dans ce chapitre nous avons présenté le concept du modèle séquence-à-séquence, son architecture et quelques exemples de son emploi. Pour notre cas, nous avons adopté ce modèle pour la synthèse de la parole pour la langue arabe. Le chapitre suivant, sera consacré à la langue arabe et ses particularités.

Chapitre 3 : La langue arabe

3.1. Introduction

Dans notre travail, nous parlerons de la langue arabe en référence à ce qui est communément appelé « l'arabe moderne » ou « l'Arabe standard »
La langue arabe fait partie des langues sémitiques (ougaritique, phénicien, araméen, hébreu, et arabe) et est parlée par plus de 530 millions de locuteurs dans le monde. Trois catégories sont distinguées : l'arabe classique, l'arabe standard MSA (Modern Standard Arabic) et l'arabe dialectal. L'arabe classique est défini comme la langue formelle parlée pendant l'époque de premières rédactions du Coran. L'arabe dialectal est lié à l'origine de la personne et varie selon les pays arabophones ou même selon les régions dans un pays. Dans ce présent travail, seul l'arabe standard a été considéré. Il représente une forme de langue commune à tous les locuteurs, il est enseigné à l'école. L'arabe standard s'écrit de droite à gauche, alors que les nombres sont écrits de gauche à droite. Quelques particularités de la langue arabe [6]

3.2. Historiques de la langue arabe

La langue arabe est la langue officielle, d'enseignement et de communication, de près de 22 pays avec près de 450 millions d'utilisateurs. L'arabe est une langue sémitique. Du point de vue lexical, elle est essentiellement dérivationnelle. Les règles fondamentales de la langue arabe, notamment celles morphosyntaxiques, n'ont pas changées depuis leur mise au point pour le Saint Coran. L'écriture arabe s'est développée grâce à la révélation coranique

L'écriture en arabe est plus qu'un simple code de communication, c'est un art avec une calligraphie des plus raffinées

L'écriture arabe, utilisant l'alphabet arabe, a connu tour à tour l'introduction des éléments suivants :

- La séparation des mots : l'écriture se faisait en caractères attachés. Le document se présentait sous forme d'une seule ligne dont toutes les lettres sont liées ;
- L'introduction de la césure, on la trouve dans certains anciens manuscrits écrits en Koufi ancien ;
- L'introduction de la voyellisation : l'utilisation de signes de voyelles dans une couleur (rouge ou jaune) autre que celle utilisée pour les lettres (noires), sous forme d'un point

au-dessus, au-dessous ou à gauche des lettres. Utilisation de deux points pour le tanwin ;

- L'introduction des signes diacritiques : utilisation de points, un, deux ou trois, au-dessus ou au-dessous des lettres avec la même couleur que les lettres ;
- Le changement de la voyellisation : signes de voyelles actuelles (Fatha, Damma, Kasra, . . .) ;
- L'interdiction de la césure des mots ;
- L'utilisation des signes d'arrêt et de commencement réservés au Saint Coran (Waqf, Wasl, . . .) ;
- L'adaptation restreinte et progressive des signes de ponctuation ou de numérotation.

L'alphabet arabe sert, moyennant de légers aménagements avec des points, comme alphabet d'écriture pour plusieurs langues à travers le monde, comme par exemple : le berbère, le farsi, le kirghize, le malais, le pashto, le persan, l'urdu, le sindhi, l'ouighur (est une langue appartenant au groupe des langues turques. Il est parlé en Asie centrale, principalement au Xinjiang (Chine), en Turquie et au Kazakhstan) et d'autres langues africaines. Autrefois, le turque et l'espagnol étaient écrits également à l'aide de l'alphabet arabe [28].

3.3. L'écriture arabe

Les différences entre le système alphabétique arabe, et celui à base d'un alphabet latin sont nombreuses. Citons par exemple [28]

- La langue arabe est caractérisée par un alphabet qui contient un ensemble de 28 lettres dont 25 sont des consonnes, les 3 restantes sont des voyelles longues. Des signes diacritiques indiquent l'identité des voyelles courtes [Tableau 3.1] [6][28] ;

Alphabet	API	Translittération	Alphabet	API	Translittération
ا	[ʔ]	Alif	ض	[ḍ]	ḍād
ب	[b]	bā	ط	[ṭ]	ṭā
ت	[t]	tā	ظ	[ẓ]	ẓā
ث	[θ]	thā	ع	[ʕ]	ɛayn
ج	[ʤ]	jīm	غ	[ɣ]	ghayn
ح	[ḥ]	ḥā	ف	[f]	fā
خ	[x]	khā	ق	[q]	qāf
د	[d]	dāl	ك	[k]	kāf

ذ	[ð]	dhal	ل	[l]	lām
ر	[r]	rā	م	[m]	mīm
ز	[z]	zāy	ن	[n]	nūn
س	[s]	sīn	ه	[h]	hā
ش	[ʃ]	chīn	و	[w]	wāw
ص	[ʂ]	ʂād	ي	[j]	yā

Tableau 3. 1: Liste de l'alphabet arabe et son API

- La voyelle /a/ : Elle est représentée par le signe "fatha" au-dessus de la consonne : (بَ [ba]) ;
- La voyelle /u/ : Elle est représentée par le signe "damma" au-dessus de la consonne (بُ [bu]) ;
- La voyelle /i/ : Elle est représentée par le signe "kasra" au-dessous de la consonne : (بِ [bi]) ; [6][29]

➢ Les voyelles : on distingue trois voyelles courtes et trois voyelles longues. La durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales. Elles sont représentées dans le tableau 3.2.

Courtes	Longues
(ـَ, ـُ, ـِ)	(ـِي, ـُو, ـَا)

Tableau 3. 2 : Classification des voyelles de la langue arabe

➢ Les voyelles particulières "tanwin", "sukun" et "chadda" :

3.3.1. Le tanwin :

- Le signe de tanwin est ajouté à la fin des mots indéterminés, il correspond à la prononciation du son /n/ à la fin du mot[6] : ceci consiste à doubler un des signes diacritiques déjà mentionnés: [an] : ـً [un] : ـٌ [in] : ـٍ ; [28]

3.3.2. Le "sokun" :

- : Ce signe est utilisé pour indiquer l'omission d'une voyelle. Il s'agit d'un petit cercle au-dessus de la consonne '°' [28]

3.3.3. La chadda :

- La "shadda" : Ce signe en forme qui a la lettre "w' est utilisé pour distinguer les consonnes géminées des consonnes simples : le signe de la chadda ـّـ [28] [6].
- **Example [5]**

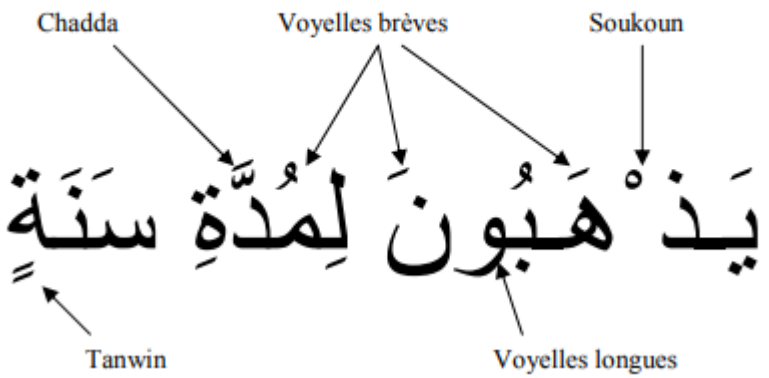

Figure 3. 1. Les voyelles particulières

Le texte arabe peut être complètement, partiellement ou non voyellé. La voyellisation pose le problème de la voyellisation de lettres en ligature ainsi que celui de la position des signes de voyelles par rapport à la lettre voyellée (juste au-dessus ou au-dessous, au même niveau ou encore à gauche).

- Le sens de déroulement : l'arabe s'écrit et se lit de droite à gauche alors que les systèmes d'écriture à base de l'alphabet latin se déroulent dans le sens inverse (d'où l'orientation de la saisie du texte, de l'affichage sur écran et du stockage en mémoire) ;
- Les lettres sont attachées entre elles avec une *kashida*. La *kashida* n'est pas une lettre en elle-même mais plutôt un allongement de certaines lettres en respectant très rigoureusement les règles de la calligraphie arabe ;
- Les lettres prennent différentes formes suivant leur position dans le mot : initiale, médiane, finale ou encore isolée (d'où les questions de la recherche syntaxique et le

codage des caractères) (Tableau 3.3), ce qui fait qu'il résulte 78 formes graphiques à partir des 28 lettres [28] ;

Isolée	Finale	médiane	Initiale
غ	ـغ	ـغـ	غـ

Tableau 3. 3: Exemple de variation de la lettre غ ghayn

- La configuration des caractères est unique : la notion de lettres majuscules, par opposition aux lettres minuscules, est inconnue ;
- La césure des mots à la fin de la ligne est également inconnue ;
- La lettre Hamza prend différentes formes suivant sa voyellisation et la voyellisation de la lettre qui la précède (ex. ء أ إ ؤ ئ لأ).
- L'absence de règles de ponctuation universellement conventionnelles ;
- La langue arabe comprend trois catégories de mots : Verbes, noms, particules, que nous allons développer plus loin.

3.4. Morphologie arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et les noms sont le plus souvent dérivés d'une racine à trois consonnes radicales [30]. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles, on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine.

3.4.1. Structure d'un mot

En arabe, un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire. La représentation suivante schématise une structure possible d'un mot (Tableau 3.4.). Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Postfixe	Suffixe	Corps schématique	Préfixe	Antéfixe

Tableau3 4: Mot arabe

Les antéfixes, les préfixes, les suffixes et les postfixes sont des morphèmes qui expriment des informations et des traits grammaticaux ;

- Antéfixes qui sont des prépositions ou des conjonctions.
- Préfixes et suffixes qui expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…)
- Postfixes qui sont des pronoms personnels.

Exemple : أَتَتَذَكَرُونَنَا, ce mot exprime la phrase en français : *"Est ce que vous vous souvenez de nous ?"*

La segmentation de ce mot donne les constituants suivants (Tableau 3.5) :

نَا	ونَ	تَذْكرُ	تَ	أ

Tableau 3.5: Application de la segmentation sur un mot arabe

Postfixe : أ **conjonction** d'interrogation ;

Préfixe : ت **préfixe** verbal du temps de l'inaccompli ;

Corps schématique : **تَذَكر** dérivé de la racine : **ذكر** selon le schème $taC_1aC_2aC_3a$:

Suffixe : **ون** suffixe verbal exprimant le pluriel ;

Enclitique : **نَا** pronom suffixe complément du nom ;

3.5. Catégories des mots

L'arabe considère 3 catégories de mots qui correspondent à chacune des huit catégories grammaticales de l'anglais et les sept du français, à savoir (Tableau 3.6) ;

Particule /Lettre حَرْفٌ	**Verbe فِعلٌ**	**Nom اِسْمٌ**
Prépositions, Conjonctions. Exemple : **إلى** (ila) signification «à» **في** (fi) signification «dans» **و** (wa) signification «et»	Verbe (comme en Français). Exemple : **كَتَبَ** (kataba) signification «écrire»	Nom, Pronom, Adjective, Adverbe, Interjection. Exemple : **رَجُلٌ** (rajulu-n), signification «Homme»

Tableau 3.6: Catégories grammaticales (catégories) de l'arabe

- Le verbe : entité exprimant un sens dépendant du temps. C'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- Le nom : élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- Les particules : entités qui servent à situer les événements et les objets par rapport au temps et à l'espace, et permettent un enchaînement cohérent du texte.
- Syntaxique, pour réduire l'ambiguïté au vu du contexte grammatical ;
- Sémantique, qui est nécessaire pour réduire l'ambiguïté au vu du sens de la phrase[28].

3.6. Transcription de la langue arabe et ses problèmes

La synthèse automatique de parole à partir du texte consiste à produire oralement un énoncé écrit. L'intelligibilité et le naturel de la voix synthétique sont les deux principaux critères de qualité de la synthèse. Les deux étapes principales de cette application sont :

- La phonétisation (transcription), ou conversion orthographique-phonétique, c'est-à-dire la création de la chaîne phonétique à partir du texte orthographique.
- La génération de la prosodie, c'est-à-dire des contours mélodiques et de la durée de chaque phonème. C'est cette étape qui apporte le naturel (voix humaine) à la voix synthétique.

La phonétisation peut se faire grâce à un lexique exhaustif qui contient la forme phonétisée de chaque forme lexicale ou grâce à un système par règles. La première méthode, plus directe, doit parfois faire appel à la seconde pour la phonétisation de mots inconnus (mots hors lexique, noms propres,...).

Le but est de phonétiser les diverses phrases composant un texte voyellisé. Nous considérons qu'une phrase peut comporter des entités comme des sigles, des abréviations, des dates, des notations diverses, etc.… Certaines langues sont plus faciles que d'autres pour effectuer cette transcription. Un compromis a dû être trouvé entre les tailles respectives du dictionnaire de règles de transcription à vocation générale et du dictionnaire de mots comportant les exceptions les plus courantes et facilitant l'analyse syntaxique souvent nécessaire de la phrase (Tableau 3.1) [31].

La complexité du problème de transcription peut être appréciée après avoir examiné le tableau 3.1 des correspondances graphèmes-phonèmes. Ces problèmes sont liés à la langue traitée elle-même et sont de différentes sortes :

- Des graphèmes qui ont plusieurs réalisations phonétiques. Le w و de بُومُ et مَوْزُ correspond à deux sons différents ayant la même graphie.
- Des phonèmes qui ont plusieurs réalisations graphémiques. Le noun dans يُقِمُونَ et dans أَنْزَلَ n'a pas la même représentation graphémique.
- Des graphèmes qui ne sont pas pris en compte. Le alif dans قَامُوا ne correspond pas à un son (silence).
- Une absence totale de correspondance graphème-phonème. Le mot هَذَا devrait être écrit هَاذَا.

Plusieurs phases sont à distinguer pour le module de conversion orthographique-phonétique [28].

3.6.1. Repérage des mots

Les séparateurs (blancs, tirets...) permettent le repérage des mots. Les séparateurs (virgules, points, deux points, points d'interrogation, d'exclamation...) sont aussi traités pour l'analyse des pauses et des arrêts.

3.6.2. Utilisation d'un lexique

Le lexique traite les exceptions qui ne peuvent être prises en compte par les règles. En effet, on ne peut pas se permettre de dépenser un nombre important de règles pour le traitement spécial de quelques mots. Un compromis est alors nécessaire.

3.6.3. Utilisation de règles

Avant d'écrire les règles de transcription, on aura soin de définir un certain nombre de catégories qui faciliteront l'écriture de ces règles. Par exemple, les consonnes solaires ou lunaires, les voyelles, etc. Ces règles traitent principalement les problèmes suivants :

- l'épellation ;
- la liaison ou non ;
- la prononciation des géminées ;
- l'élision en syllabe initiale du mot, en syllabe intérieure du mot, l'élision en finale ;

- la prononciation des graphèmes à l'intérieur d'un mot, en fonction du contexte dans lequel il est représenté par un son particulier, ou bien une catégorie de sons avec d'éventuelles exceptions, etc.

3.6.4. Conversion Orthographique-Phonétique

Cette phase concerne le passage du texte graphique au texte phonétique (après le traitement des exceptions et des règles de prononciation), suivant la table de correspondance (Tableau 3.1). Plusieurs techniques peuvent être exploitées pour faciliter les opérations décrites ci-dessus. Il est plus simple d'utiliser une grammaire contextuelle. Cette dernière peut faciliter l'écriture, puis la lecture et la modification, des règles avec un contrôle automatique de la syntaxe des règles utilisées.

3.7. Base de règles

Le rôle de la base de règles est de rendre compte des phénomènes réguliers de la langue arabe. L'ensemble des règles du système de conversion orthographique-phonétique est réparti en deux **catégories** : les règles morpho-orthographiques et les règles phonologiques.

3.7.1. Règles morpho-orthographiques

Les règles morpho-orthographiques ont pour finalité de calculer la représentation phonétique à associer à une représentation orthographique. Leur structure conceptuelle est de la forme [28] :

Séquence orthographique (S-O) → Séquence phonétique (S-P)

- **Consonnes géminées et tanwin**

La gémination est considérée comme la réalisation d'une double consonne dont la première est sans vocalisation (sukun) suivie de la même consonne avec sa voyelle. D'autres règles morpho-orthographiques opèrent en analysant la structure orthographique d'un mot. Ce type de règles morpho-orthographiques est utilisé pour phonétiser correctement des entités comme certains sigles, abréviations, nombres, etc.…

- **Nombres**

Ils peuvent être prononcés dans leurs formes isolées, chiffre par chiffre, ou dans leurs prononciations numériques. La première méthode est simple mais présente l'inconvénient

de ne pas être très explicite, Le nombre **147** sera prononcé de droite à gauche **واحد – اربعة - سبعة**.

La seconde méthode a l'avantage d'être plus naturelle, les nombres sont prononcés selon leurs formes numérales **مائة و سبعة واربعون**.

- **Caractères isolés ou sigles inconnus**

Après échec dans la consultation des différents lexiques, les caractères isolés ou sigles inconnus sont générés et prononcés dans leur forme d'épellation standard [56].

Lettre	Prononciation
م	Miim
ن	Noon

tableau.3 7: Exemple de traitement de caractères isolés

- **Abréviations et sigles**

Ils sont prédéfinis dans un dictionnaire qui peut être mis à jour par des modifications, insertions ou suppressions. Toute nouvelle acquisition obéit aux règles d'écriture des mots et de compatibilité des voyelles. Chaque usager pourra créer son propre lexique en fonction de son application.

Sigle ou abréviation	Valeur dans le dictionnaire
الخ	الى اخره
ش . و . س . ح	الشركة الوطنية للسكك الحديدية

tableau.3 8: Exemple de traitement des abréviations

Si un mot est constitué uniquement de consonnes et que les accès lexicaux échouent, nous supposons qu'il s'agit d'un sigle non prononçable pour le système de synthèse et par conséquent, nous proposons l'épellation du mot.

3.7.2. Règles phonologiques :

Il existe différentes règles nécessaires à la prise en compte des phénomènes phonologiques liés à la langue arabe. Nous pouvons citer, à titre d'illustration, les

transformations phoniques que subit une phrase avec l'élision de sa première lettre, sa seconde lettre ou des deux à la fois suivant la nature de la syllabe qui la précède ou de la consonne qui la suit. La phrase subit des transformations phoniques selon son contexte immédiat :

- Par rapport à la syllabe qui la précède :

Le أ (alif) sera élidé et le ل (lam) est prononcé comme la fermeture de la syllabe ouverte à la fin du mot précédent.

Phonèmes	Prononciation	Ecriture
Fabilghorfati (dans la pièce)	فبلغُرفة	فبِالغُرفة
Qara'telkitebe(tu as lu le livre)	قَرَاتَلْكِتابَ	قَرَاتَ الكِتَابَ

Tableau 3. 9: Exemple d'élimination du أ(alif) grâce aux règles phonologiques

- Par rapport au phonème qui suit :

Le ل (lam) s'assimilera selon la nature de ce phonème (solaire ou lunaire).[28]

Phonèmes	Prononciation	Ecriture
Annafidatu	انّنَافِذَة	النَافِذَة
Assama'u (le ciel)	اسَمَاءُ	السَماءُ

Tableau 3. 10: Exemple d'élimination du ل(lam) grâce aux règles phonologiques

3.8. Conclusion :

Ce chapitre présente des généralités sur la langue arabe. il nous a permis d'introduire la langue arabe et ses particularités. Ensuite la transcription de la langue arabe et ses problèmes.

Nous allons présenter, dans le chapitre suivant, la synthèse de la parole pour la langue arabe, basé sur la méthode seq2seq.

4.1. Introduction

Dans ce chapitre nous allons présenter la méthodologie suivie afin d'implémenter le system de synthèse de la parole pour la langue arabe, les différentes étapes d'implémentation sont détaillées. Nous commençons dans un premier temps par la description des moyens et logiciels utilisés ainsi que le corpus ensuite, nous passons à l'implémentation.

4.2. Moyens et logiciels utilisés

L'implémentation d'un système de synthèse de la parole pour la langue Arabe est réalisée de manière soft (logiciel) sous un environnement de programmation évalué. Nos moyens informatiques sont constitués d'un ordinateur HP, avec CPU core i3, à une fréquence d'horloge de 2.30 GHz et avec une RAM de 4 Go. Le logiciel utilisé est le MATLAB R2021a (Figure 1) dont le but est d'utiliser les rebriques récentes de MATLAB.

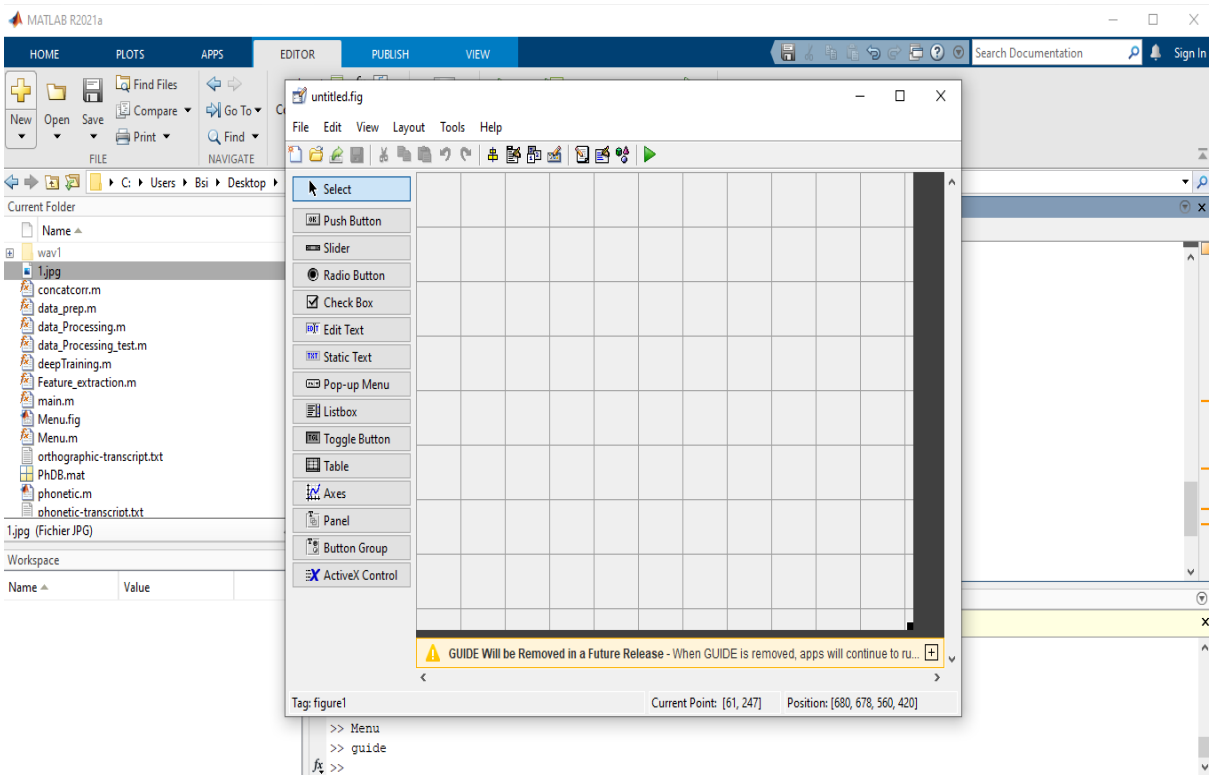

Figure 4.1. Environnement du logiciel MATLAB

4.3. L'interface graphique de l'application

L'interface graphique est organisée de telle façon à faciliter l'exploitation de cette application, cette dernière est subdivisée en sous-programmes dont chacun représente une idée de traitement comparable aux autres.

4.3.1 Phase de lancement

A partir de l'espace de travail de MATLAB, on exécute l'instruction « guide », ou bien de barre d'outils du logiciel qui contient un bouton de lancement du guide. Aussi, il est possible de taper le nom de notre script MATLAB directement sur l'espace de travail.

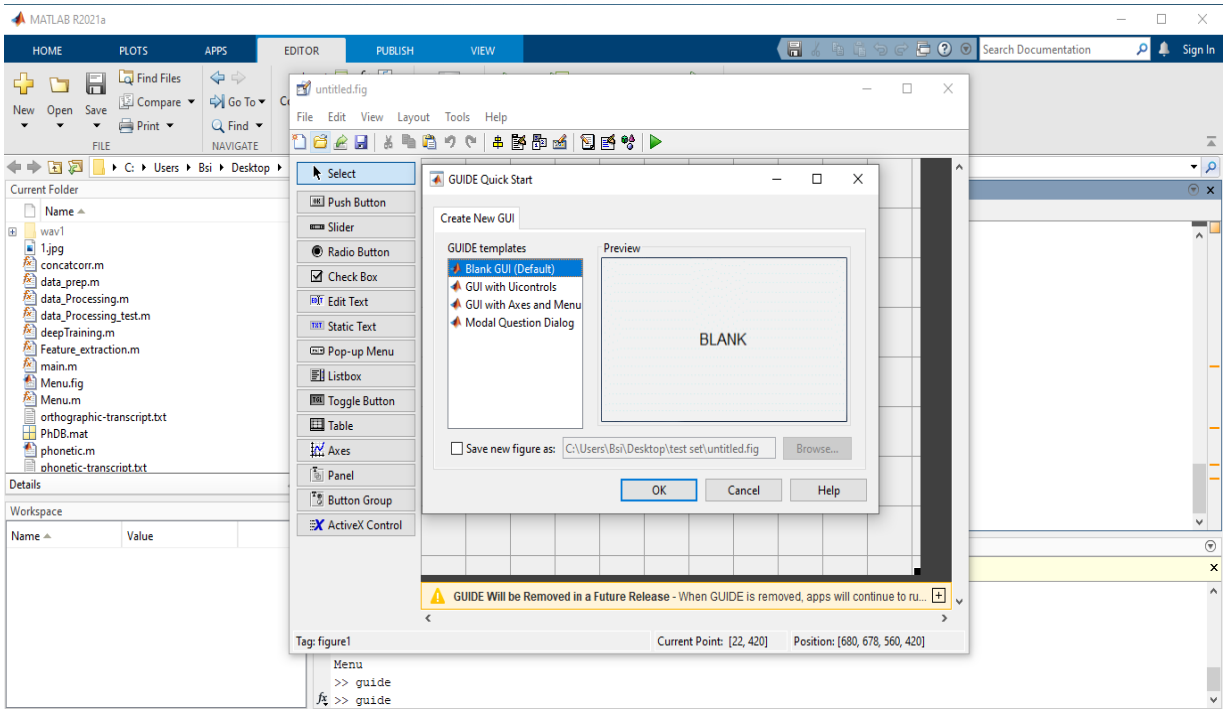

Figure 4.2. Les étapes de l'ouverture d'une fenêtre « Guide ».

4.3.2. Phase de l'exploitation

Après le lancement de la fenêtre « Guide » correspondante à la page d'accueil de notre interface, on aura la figure suivante :

Figure 4.3. Page d'accueil de l'application.

4.3.3. Interface graphique pour le système réalisé

La phase finale de notre travail a consisté à réaliser une interface graphique simple afin de permettre à des personnes non-expertes d'utiliser le système de synthèse de la parole. Les différentes étapes de l'exécution de notre système de synthèse sont présentées dans la figure 4.3.

La page d'accueil contient l'intitulé du thème traité, ainsi que deux boutons poussoirs, l'un pour quitter l'application et l'autre pour accéder à la fenêtre suivante qui est la fenêtre de la synthèse de la parole.

En cliquant sur le bouton « Suivant », la page d'accueil sera fermée et la fenêtre d'exploitation du système s'affiche (voir Figure 4.4)

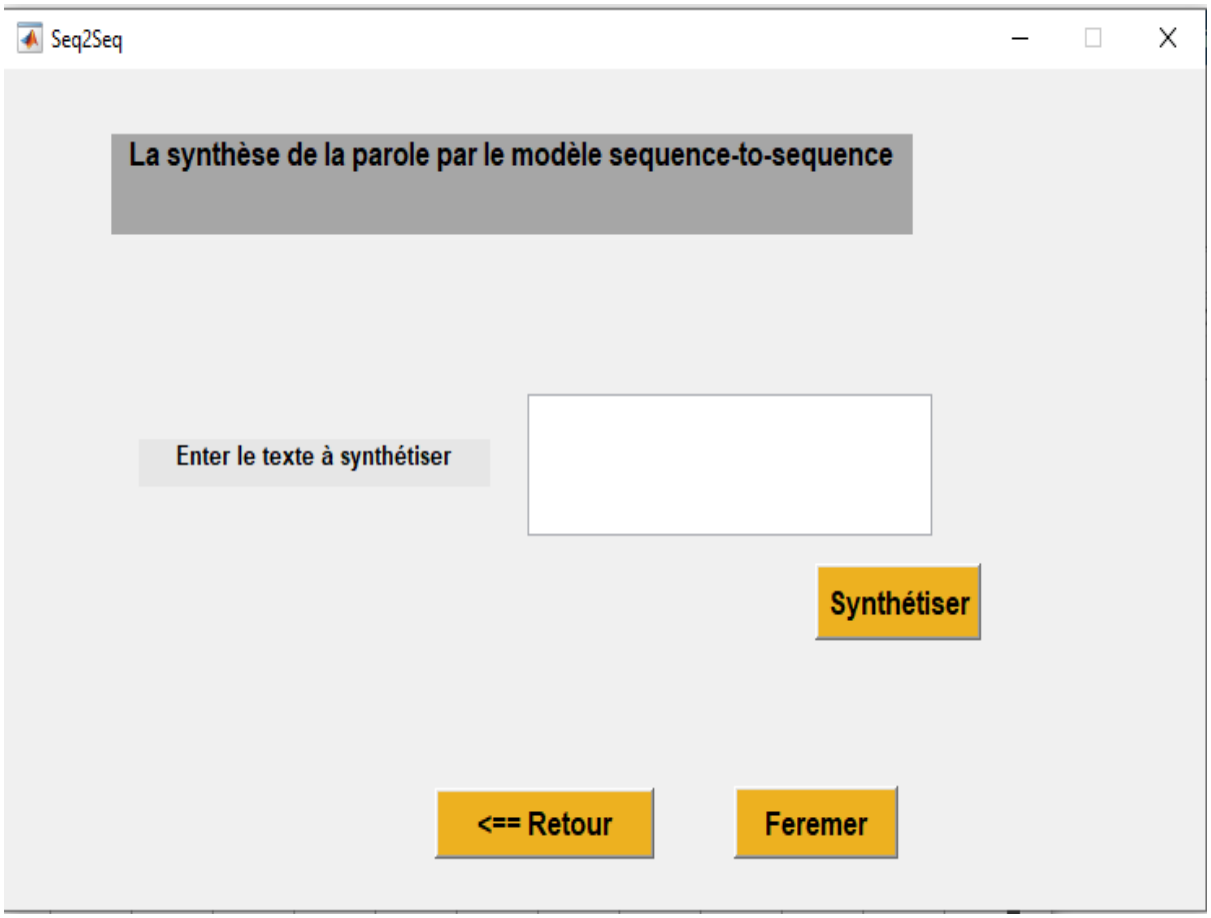

Figure 4.4. Fenêtre des techniques d'estimation.

Cette dernière donne la possibilité d'accéder au synthétiseur, ainsi un bouton « Retour » de retour vers la page d'accueil et le bouton « Fermer » pour sortir de l'application. La Figure 4.4 présente les accès possibles.

4.4. Organigramme

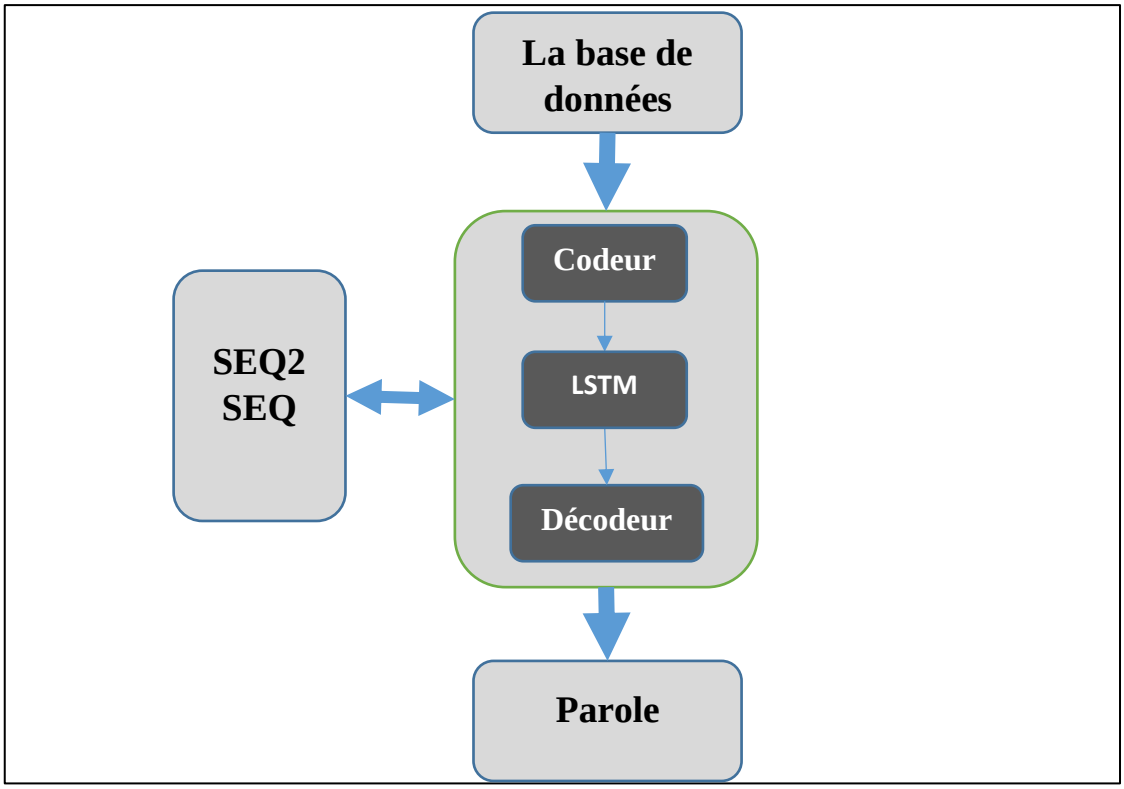

Figure 4.5. Organigramme de modèle seq2seq.

- ***Préparation de données :***

Base de données de la parole arabe

Le corpus de la parole arabe de qui est un corpus à un seul locuteur, est utilisé comme matériau textuel pour construire notre système de synthèse vocale. Ce corpus a été développé dans le cadre des travaux de doctorat de Nawar Halabi à l'Université de Southampton. Le corpus a été enregistré dans un studio professionnel en arabe levantin du sud (accent damascien). Ce corpus de parole contient 1813 fichiers wav avec des énoncés parlés, 1813 fichiers avec des énoncés textuels et 1813 fichiers avec des étiquettes de phonèmes et des horodatages des limites.

Le tableau 4.1 illustre la transcription orthographique et les symboles phonétiques utilisés dans le corpus, qui contient 44 phonèmes en langue arabe.

Nombre	1	2	3	4	5	6	7	8	9
Orthographique	" Un"	"B"	"T"	"w"	"C"	"h"	"X"	"Dr"	"ou"
phonétique utilisé	"<"	"b"	"t"	"^"	"j"	"H"	"X"	"ré"	"*"
Nombre	10	11	12	13	14	15	16	17	18
orthographique	"ر"	"ز"	"س"	"ش"	"ص"	"ض"	"ط"	"ظ"	"ع"
phonétique utilisé	"r"	"z"	"s"	"$"	"S"	"RÉ"	"T"	"Z"	"E"
Nombre	19	20	21	22	23	24	25	26	27
orthographique	"G"	"q"	"s"	"K"	"à"	"M"	"n"	"e"	"Et"
phonétique utilisé	"g"	"F"	"q"	"k"	"je"	"m"	"n"	"h"	"w"
Nombre	28	29	30	31	32	33	34	35	36
orthographique	"ي"	"ا"	"و"	"ي"	"ٱ"	"ٱ"	"ِ"	"(ا)"	"(و)"
phonétique utilisé	"ou"	"aa"	"uu0"	"ii0"	"une"	"u0"	"i0"	"AA"	"UU0"
Nombre	37	38	39	40	41	42	43	44	
orthographique	" (ي)"	"(ً) "	"(ُ) "	"(ِ) "	"[ُ] "	"[ِ] "	"([ُ])"	"([ِ])"	
phonétique utilisé	"II0"	"UNE"	"U0"	"I0"	"u1"	"i1"	"U1"	"I1"	

Tableau 4.1. Les phonèmes et leurs notations choisies

La figure 4.6 représente les 44 occurrences de phonèmes de la langue arabe dans le corpus de parole, qui sont divisées en cinq catégories :

- Moins de 100 occurrences,
- De 102 à 400 occurrences,
- De 401 à 700 occurrences,
- De 701 à 1200 occurrences
- Plus de 1200 occurrences.

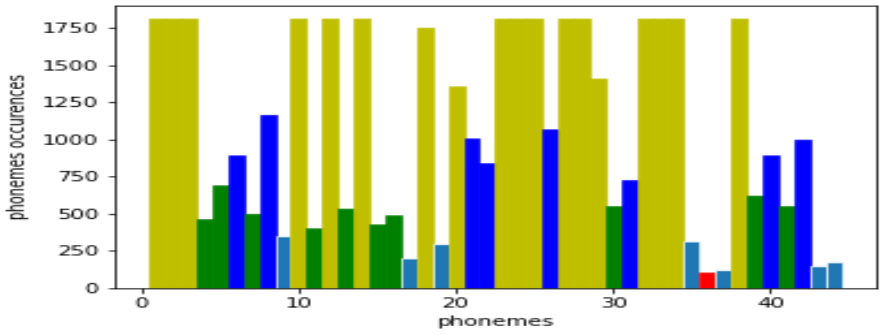

Figure 4.6. Occurrences de phonèmes dans le corpus

Où le numéro du phonème indique sa position dans le tableau 4.1.

- ***Le codeur :***

Le codage signifie convertir les données dans un format requis. Dans notre cas, nous convertissons un mot (texte) en voix (parole). Dans le contexte de l'apprentissage automatique, nous convertissons une séquence de mots arabes en un vecteur bidimensionnel, ce vecteur bidimensionnel est également appelé état caché. L'encodeur est construit en empilant un réseau neuronal récurrent. Nous utilisons ce type de couche car sa structure permet au modèle de comprendre le contexte et les dépendances temporelles des séquences. La sortie du codeur est un vecteur bidimensionnel qui encapsule toute la signification de la séquence d'entrée et l'état du dernier pas de temps RNN.

- ***LSTM :***

Les composants centraux d'un réseau LSTM sont une couche d'entrée de séquence et une couche LSTM. Une couche d'entrée de séquence entre des données de séquence ou de série temporelle dans le réseau. Une couche LSTM apprend les dépendances à long terme entre les pas de temps des données de séquence.

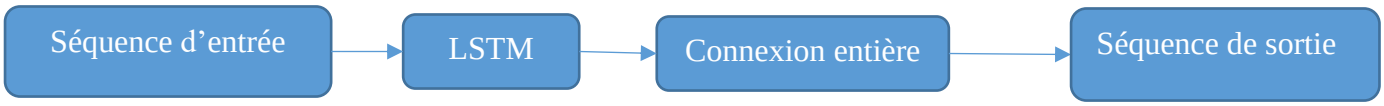

Figure 4.7. Blocs d'un réseau LSTM.

Ce schéma illustre l'architecture d'un réseau LSTM simple. Le réseau commence par une couche d'entrée de séquence suivie d'une couche LSTM. Le réseau se termine par une couche entièrement connectée et une couche de sortie.

- ***Le décodeur:***

Décoder signifie convertir un message codé en langage intelligible. Dans le modèle d'apprentissage automatique, le rôle du décodeur sera de convertir le vecteur bidimensionnel en une séquence de sortie, la voix. Il est également construit avec des couches RNN et une couche dense pour prédire le texte en arabe.

4.5. Méthodologie

Pour transformer l'information, former le niveau de forme d'onde au niveau linguistique et vice versa, les systèmes de production et de perception de la parole humaine utilisent des structures hiérarchiques clairement stratifiées. Pour résoudre le problème de la

synthèse vocale, cette étude a utilisé une architecture de réseau de neurones profonds. La figure 4.8 illustre l'ensemble de l'approche méthodologique qui a été utilisée.

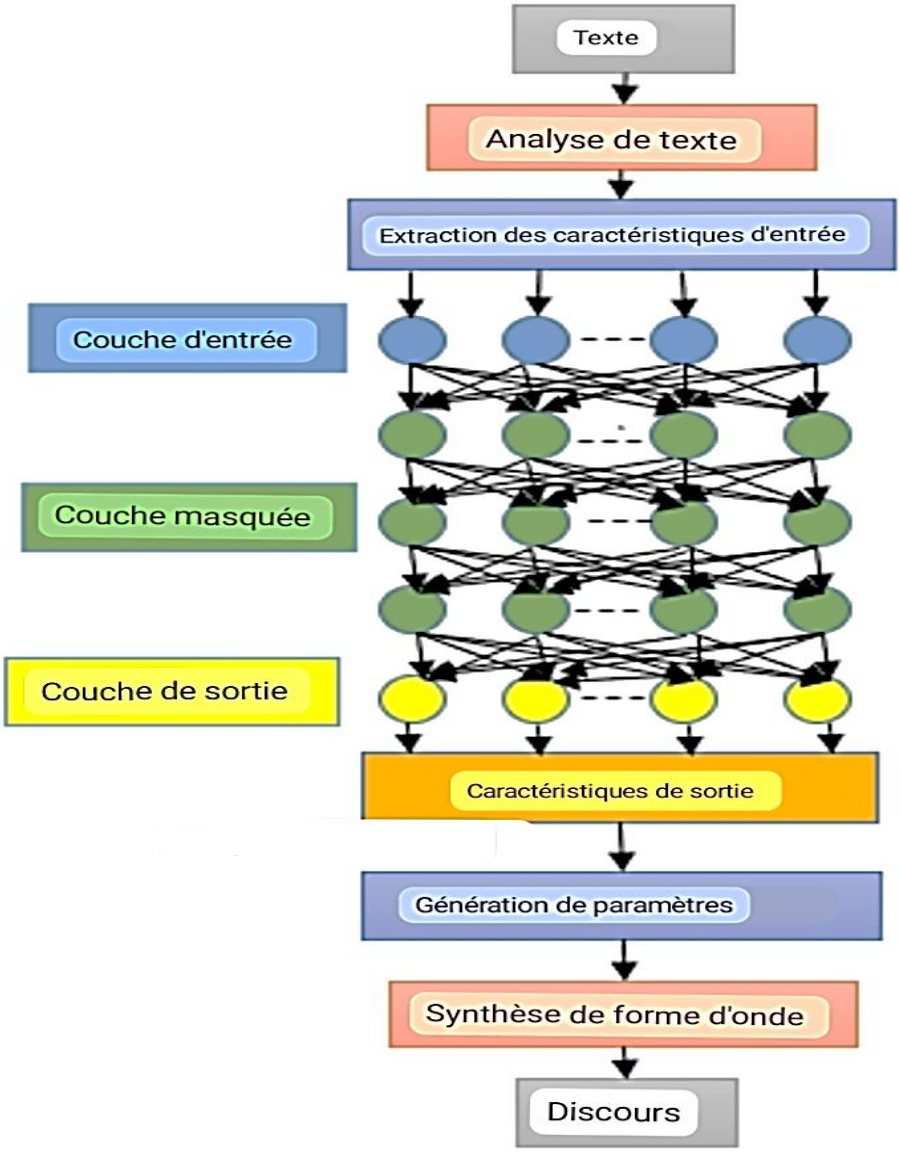

Figure 4.8. Synthèse vocale de la méthode proposée

Pour synthétiser un texte donné, il faut d'abord le décomposer en états de mots. Il s'agit simplement d'une étape de pré-traitement du texte fourni. Chaque mot sera converti en symbole phonétiques (conversion graphème-phonème). Le texte fourni sera ensuite transformé en une série de caractéristiques d'entrée. Ces caractéristiques sont les contextes linguistiques et les valeurs numériques telles que le nombre de mots dans la phrase, la

position relative de la trame actuelle dans le phonème actuel et les durées des phonèmes en tant que réponses binaires.

Ensuite, un réseau de neurone profond entraîné utilise la propagation vers l'avant pour mapper les caractéristiques d'entrée aux caractéristiques de sortie, qui incluent les paramètres spectraux et d'excitation ainsi que leurs dérivées temporelles (caractéristiques dynamiques). Les poids de réseau de neurone profond peuvent être entraînés en extrayant des paires de caractéristiques d'entrée et de sortie à partir des données d'entraînement.

L'algorithme de génération de paramètres vocaux peut générer des trajectoires fluides des caractéristiques des paramètres vocaux qui satisferont à la fois les statistiques des caractéristiques statiques et dynamiques en définissant les caractéristiques de sortie prédites du réseau en tant que vecteurs moyens et les variances précalculées des caractéristiques de sortie de toutes les données d'apprentissage en tant que matrices de covariance. Enfin, sur la base des paramètres de parole, un module de synthèse génère une parole synthétisée.

4.6. L’évaluation de la synthèse de la parole :

L’évaluation de la synthèse de la parole est un domaine incontournable tant que la qualité de la parole produite sera inférieure à la qualité de la parole naturelle. En effet, l’évaluation de la qualité de la synthèse permet de déterminer certains défauts du système et d’y remédier. A ce jour, la grande majorité des tests utilisent les impressions subjectives de plusieurs sujets humains. Cette évaluation est appelée "globale" car elle juge la sortie du système de synthèse sans se préoccuper de son fonctionnement interne. L’autre type d’évaluation consiste à évaluer la qualité de chaque module du système synthèse. En effet, le système de synthèse est une chaîne de traitements. Dans cette chaîne, le traitement qui fonctionne mal est celui qui va limiter les performances du système.

4.7. Qualité et intelligibilité de la parole :

L’intelligibilité de la parole correspond à la capacité de comprendre un message linguistique contenu dans un signal de parole. L’intelligibilité est donc une mesure objective définie par le nombre de mots prononcés est correctement identifiés par l’auditeur. Chaque mesure d’intelligibilité est une interaction entre le locuteur, l’environnement de transmission et l’auditeur. Le meilleur moyen de juger l’intelligibilité est d’effectuer des tests d’écoute avec des sujets, dont la capacité d’écoute est normale.

La qualité d'un signal de parole permet de prendre en compte la présence d'agents extérieurs "perturbateurs" (environnement bruyant, distorsions, . . .). La clarté du message peut, en effet, être affectée par le bruit environnemental, ce qui nuit au confort d'écoute. C'est donc une mesure subjective liée à l'aspect agréable de l'écoute du signal de parole par l'auditeur. Cependant, même après le débruitage, la qualité de la parole n'est pas totalement restituée ; elle est même parfois encore plus dégradée. Les éléments fondamentaux qui influent sur la qualité de la parole après débruitage sont les distorsions du signal et le bruit résiduel communément appelé "bruit musical". Les tests de jugement de la qualité par des auditeurs sont les seuls moyens d'évaluation valables et fiables d'un système de débruitage de la parole. Mais comme pour l'intelligibilité, il existe des critères objectifs d'évaluation de la qualité, tels que le PESQ, MBSD, etc. Ces critères ont un caractère perceptuel parce qu'ils sont fondés sur des notions psycho acoustiques pour simuler notre perception vis-à-vis du signal de parole.

Pour conclure, l'intelligibilité est donc une notion à ne pas confondre avec la qualité de la parole. Une amélioration de la qualité de la parole n'implique pas une amélioration en termes d'intelligibilité. Dans les environnements bruyants, améliorer l'intelligibilité de la parole s'avère une tâche plus difficile qu'améliorer la qualité de la parole.

La qualité de la parole synthétisée peut être évaluée par plusieurs méthodes comme : MOS (Mean Opinion Score), le DMOS (Degradation Mean Opinion Score) et le CMOS (Comparison Mean Opinion Score)

4.7.1. MOS :

Le MOS est la méthode la plus fréquente pour les mesures de qualité. C'est le résultat de l'analyse par catégories absolues ACR (Absolute Category Rating), dans laquelle un groupe d'auditeurs écoute un ensemble de fichiers audio et les évaluent indépendamment, en donnant une note/5 selon une échelle de notation sur la qualité perçue.

Score MOS	***Qualité MOS***
5	Exelent
4	Bon
3	Passable
2	Mauvais
1	Mediocre

Tableau 4.2. Échelle MOS

Lors de ce sondage, les auditeurs sont invités à écouter et à juger la séquence du signal à évaluer. Le jugement se fait à travers l'attribution d'une note sanctionnant la qualité perçuedu signal de parole qu'ils ont écouté. La moyenne des notes attribuées constitue donc le MOS.

L'avantage du MOS est qu'il quantifie la qualité perçue par les auditeurs participant aux tests.

C'est donc une évaluation réelle, fiable et correcte de la qualité des signaux mis en jeu.

Cependant, ce test est souvent écarté du fait qu'il requiert :

- Un grand nombre d'auditeurs
- Un équipement audio adapté
- Une formation des auditeurs à la bonne façon d'attribuer des notes pour que celles-ci soient exploitables
- Une collecte d'informations et de traitements statistiques pour réduire l'aléa.

4.7.2. Evaluation objective (PESQ):

L'évaluation objective est effectuée à l'aide du modèle perceptif 'PESQ' programmé en langage C par l'Unité Internationale de Télécommunication (L'UIT 2001). La figure 4.8 présente l'interface de l'exécutable réalisée, nous donnons en entrée les deux signaux (original et synthétique), la valeur de la fréquence d'échantillonnage ainsi que la bande passante. Le test a été effectué en utilisant les mêmes signaux que les tests précédents. Après la comparaison entre le signal original et celui synthétisé, le PESQ donne deux notes : la première allant de -0.5 à 4.5 et l'autre (MOSLQO) est obtenue après transformation à l'échelle MOS de 1 à 5.

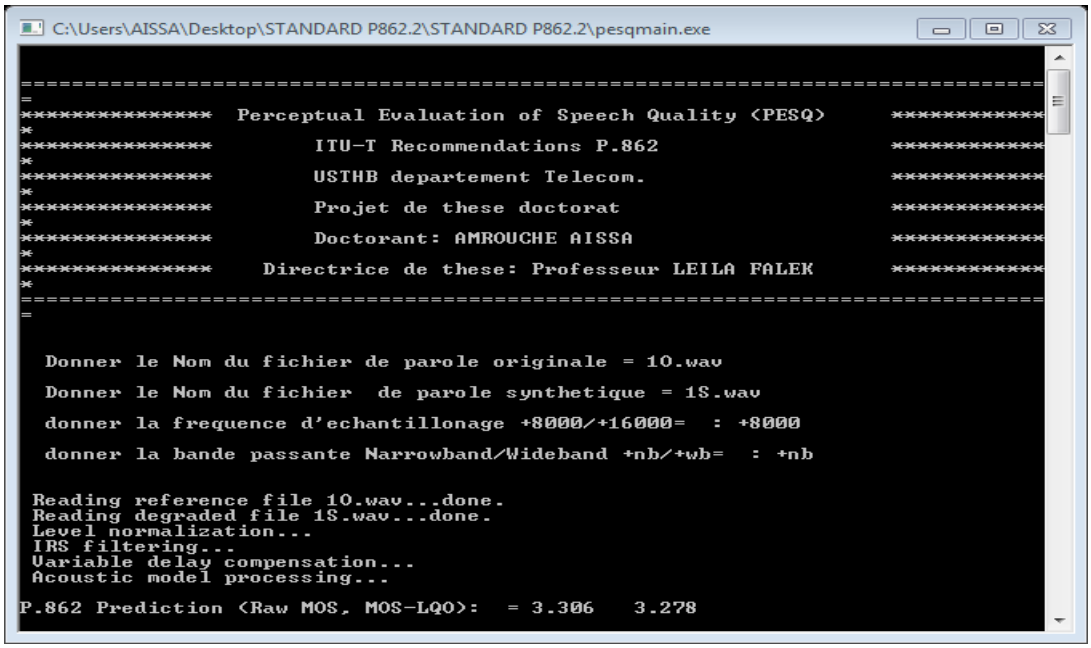

Figure 4.9 : Interface principale du PESQ.

4.8. Tests, résultats et discussions :

Pour déterminer la qualité du synthétiseur vocal, nous avons utilisé des tests d'évaluation à la fois objectifs et subjectifs. Nous utilisons le score d'évaluation perceptive de la qualité de la parole (PESQ) pour une évaluation objective [28, 29]. Nous avons utilisé 100 phrases aléatoires de l'ensemble des tests et les enregistrements correspondants comme données vocales d'origine pour calculer les scores PESQ de la synthèse vocale. Nous avons ensuite généré 100 formes d'onde de sortie à partir des phrases sélectionnées. Enfin, en envoyant les discours originaux et synthétiques par paires, l'algorithme PESQ a été utilisé pour calculer le score PESQ.

Pour l'évaluation subjective, nous avons utilisé le score d'opinion moyen (MOS) afin d'évaluer la qualité globale et le naturel des signaux vocaux produits. La qualité globale des signaux générés est appelée qualité globale. L'intonation et le rythme des signaux vocaux synthétisés sont utilisés pour déterminer le naturel. Nous avons demandé l'aide de quelques locuteurs natifs de l'arabe pour effectuer le test. Vingt personnes ont participé au test MOS en tant qu'auditeurs. Le public était composé de dix hommes et dix femmes âgés de 18 à 45 ans.

Nous avons pris 22 phrases aléatoires du corpus et les avons synthétisées à l'aide du système développé. Les auditeurs ont écouté les phrases et attribué à chaque système un score de naturel allant de 1 à 5. Un score plus élevé indique un plus grand naturel. Le score moyen du système a été calculé en faisant la moyenne des scores.

La figure 4 illustre les scores MOS et PESQ pour les phrases sélectionnées.

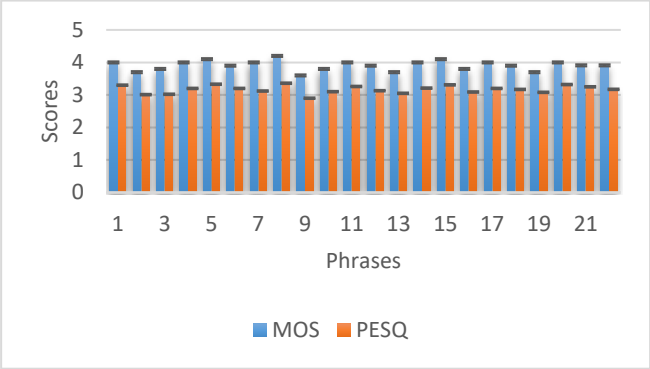

Figure 4.9. Test de scores MOS et PESQ.

Pour tester l'intelligibilité de certaines consonnes et voyelles, le Diagnostic Rhyme Test DRT est l'un des meilleurs moyens de le juger. Le DRT permettait de tester aussi bien les voyelles que les consonnes quelle que soit leur position dans un mot ou une phrase. Le nombre de mots et de phrases doit être suffisant pour tester systématiquement toutes les consonnes, associées à au moins plusieurs voyelles différentes. Nous avons choisi 20 mots et 15 phrases pour ce test. De petits changements ont été apportés aux voyelles, consonnes ou particules dans les mots et les phrases.

La figure 4.10 et 4.11 illustrent un exemple des tests de résultats DRT pour les mots et les phrases respectivement.

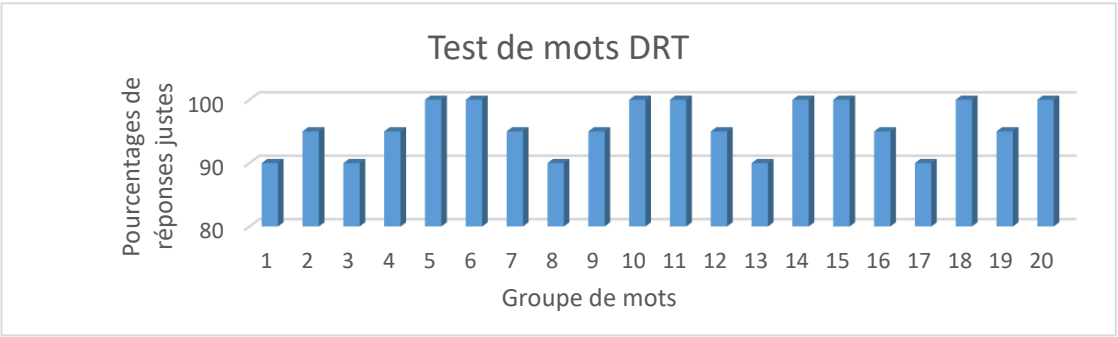

Figure 4.10. Résultats du test DRT pour les mots

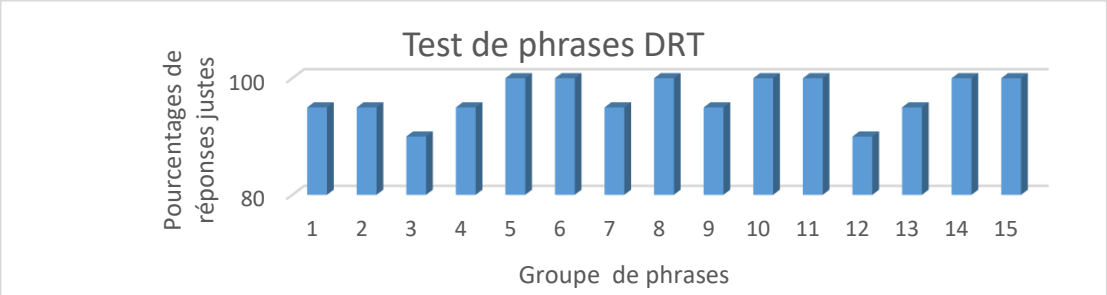

Figure 4.11. Résultats du test DRT pour les phrases

La majorité des changements de consonnes et de voyelles dans les mots et les phrases ont été reconnus avec les pourcentages de 94,23 %, 97,43 % respectivement les résultats montrent que les mots en contexte sont plus intelligibles que les mots isolés. La figure 4.12 représente la moyenne des résultats de tests obtenus du système.

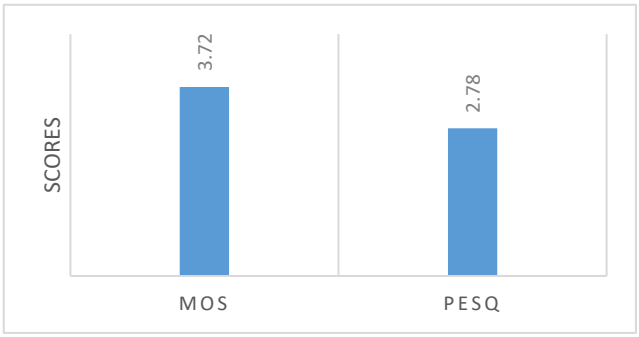

Figure 4.12. Les résultats d'une évaluation globale de la qualité.

Les signaux produits avec la synthèse vocale basée sur DNN ont un MOS et des scores élevés en termes de qualité globale et de naturel, selon les résultats. Tous les participants ont évalué la qualité globale du système comme satisfaisante, avec un score moyen de 3,72.

4.9. Conclusion

Dans ce chapitre, nous avons présenté le system de synthèse de la parole pour la langue arabe développée en évaluant sa qualité de la parole synthétique. Nous avons, par la suite testé et expliqué les méthodes d'évaluations de la qualité de parole. Les résultats obtenus sont satisfaisants et les auditeurs ont jugé le system d'une qualité acceptable.

Conclusion générale :

Le travail réalisé au cours de cette étude a porté sur la contribution à l'élaboration d'un système de synthèse de la parole partir du texte à base de la méthode séquence-à-séquence, l'objectif de notre travail a été de mener une étude dans le domaine de la synthèse de la parole à partir du texte Arabe et de comprendre le fonctionnement du système de synthèse séquence-à-séquence.

Tout au long de ce travail, nous avons abordé différents aspects tous aussi importants les uns que les autres. Nous avons commencé par comprendre le processus de génération de la parole par l'être humain et les différents moyens utilisés pour analyser et traiter ce signal. En effet, l'analyse du signal de parole montre la présence de plusieurs informations (linguistiques, paralinguistiques et extralinguistiques) qui entrainent une grande variabilité du signal. Par la suite nous avons abordé la synthèse de la parole, et présenté les différentes étapes du processus de cette synthèse à partir de la représentation textuelle et les différentes méthodes de génération sonore.

La synthèse par le modèle séquence à séquence (qu'on notera seq2seq) est en effet l'un des méthodes qui donne la meilleure qualité de parole synthétique.

Le modèle seq2seq se compose de deux sous-réseaux, le codeur et le décodeur. L'encodeur, sur la gauche, reçoit des séquences du langage source en tant qu'entrées et produit, en conséquence, une représentation compacte de la séquence d'entrée, essayant de résumer ou condenser toutes ses informations. Ensuite, cette sortie devient une entrée ou un état initial pour le décodeur, qui peut également recevoir une autre entrée externe.

A chaque pas de temps, le décodeur génère un élément de sa séquence de sortie en fonction de l'entrée reçue et de son état courant, ainsi que de la mise à jour de son propre état pour le pas de temps suivant.

Les séquences d'entrée et de sortie sont de taille fixe, mais elles ne doivent pas nécessairement correspondre - la longueur de la séquence d'entrée peut différer de celle de la séquence de sortie.

Le point critique de ce modèle est de savoir comment amener le codeur à fournir la représentation la plus complète et la plus significative de sa séquence d'entrée dans un seul élément de sortie au décodeur, car ce vecteur ou état est la seule information que le décodeur

recevra de l'entrée pour générer la sortie correspondante. Plus l'entrée est longue, plus il sera difficile de compresser en un seul vecteur.

C’est à n’en pas douter là l’enjeu de la synthèse vocale pour le futur, presque aussi intelligible que la voix humaine, mais encore si peu naturelle.

Dans notre cas l’intelligibilité a été jugée de qualité par les auditeurs.

Notons que cette étape d’évaluations noua a permis de mettre en évidence les insuffisances de notre système car en effet, la synthèse vocale n’aura de finalité que dans un traitement temps réel et une prosodie réellement vivante, qui ne serait plus générée dans l’optique de rendre le discours supportable, mais qui serait vecteur d’informations cruciales, non syntaxiques, ni phonétiques. C’est alors que le véritable potentiel de cette technologie pourra enfin se révéler dans des projets de recherche ambitieux et aidant au développement humain.

Ce projet nous a permis aussi d’apprendre et surtout de toucher à plusieurs domaines tels que le traitement de signal, la programmation, le traitement automatique du signal de parole et bien d’autres domaines.

BIBLIOGRAPHIE

[1] L. R. Rabiner and R. W. Schafer. Digital Processing of Speech. Prentice-Hall, Englewood Cliffs, N.J., 1978.

[2] https://www.aquiladata.fr/insights/voix-ia-reconnaissance-automatique-de-la-parole/

[3] https://www.reconnaisance et synthèse de la parole.com

[4] Synthèse de la parole à partir du texte par Christophe D'ALESSANDRO Directeur de recherches LIMSI-CNRS, Orsay, France et Gaël RICHARD Professeur Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI, Paris France

[5] Sofiane Baloul :'Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé', thèse de doctorat, l'université du MAINE, le Mans, France,2003

[6] Amal HOUIDHEK:'Synthèse paramétrique de la parole Arabe', thèse de doctorat, l'université de LORRAINE,2020

[7] Taylor, P. (2009). Text-to-speech synthesis. Cambridge university press.

[8] Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M. et Richards, C. (2001). Normalization of non-standard words. Computer speech & language

[9] Moulines, E. et Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech communication,

[10] Dutoit, T. et Pagel, V. (1996). Le projet MBROLA : vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale.

[11] A. Amrouche, L. Falek, H. Teffahi. Design and Implementation of a Diacritic Arabic Text-To-Speech System International Arab Journal of Information Technology (IAJIT).

[12] Black, A. W., Zen, H. et Tokuda, K. (2007). Statistical parametric speech synthesis. In ICASSP, International Conference on Acoustics, Speech and Signal Processing,

[13] Yamagishi, J., Masuko, T. et Kobayashi, T. (2004). HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions. In Proc. of Special Workshop in Maui (SWIM).

BIBLIOGRAPHIE

[14] Zen, H. et Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In ICASSP, International Conference on Acoustics, Speech and Signal Processing,

[15] Hemina Karim, Heminna Oussama : 'Développement d'une voix arabe synthétique open source',Mémoire de fin d'étude, l'université de Blida 1.2020

[16] Lasselin, Gwénolé Lecorvé. 'Make text look like speech : disfluency generation using sequence-to-sequence neural networks'. [Rapport de recherche] Univ Rennes, CNRS, IRISA, France ; IRISA, équipe EXPRESSION. 2018.

[17] Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., and Billot, S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. TALN.

[18] Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning.

[19] Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017). Recent trends in deep learning based natural language processing. Computation and Language.

[20] Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. Proceedings of the 30-th International Conference on Machine Learning.

[21] Greff, K., Srivastava, R., Koutnik, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM : A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems.

[22] Cho, K., Merriënboer, B. V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.

[23] Chung, J., Cho, K., Gülçehre, Ç., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

[24] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems.

BIBLIOGRAPHIE

[25] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate.

[26] https://docs.chainer.org/en/v7.8.0/examples/seq2seq.html

[27]https://nix-united.com/blog/neural-network-speech-synthesis-using-the-tacotron-2-architecture-or-get-alignment-or-die-tryin/

[28] Amrouch Aissa : Contribution à l'amélioration du signal de synthèse dans un système TTS pour la langue arabe ,thèse de doctorat, l'université des sciences de la technologies Houari Boumediene,2017

[29] A.Chentir, Etude de la Microprosodie en vue de la Synthèse de la parole en Arabe Standard. Thèse de Doctorat, Ecole Nationale Supérieure Polytechnique, Algérie, 2009.

[30] S. Baloul, M. Alissali, M. Baudry et P. Boula de Mareuil : Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe, 24es Journées d'Étude sur la Parole, 24-27 juin 2002 Nancy, pp.329-332.

[31] Z. Zemirli, M. Sellami et N. Vigouroux, Modélisation des règles phonologiquesdans un système de génération automatique de la langue arabe, France,