REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Echahid Hamma Lakhdar – El Oued
Faculté des Sciences Exactes
Département d'informatique

N°d'ordre :.....
Série :.....

Mémoire

Présenté en vue de l'obtention du diplôme de master en Informatique
Option : **Intelligence Artificielle et Systèmes Distribués**

Proposition d'un modèle de synthétiseur de parole arabe réaliste à base de diphones

---

Par :

GUEDIRI Salah
ZEBDI Ammar

Soutenu le : 24 mai 2016

Devant le jury :

Zaiz Faouzi	Docteur	Université d'El Oued	Rapporteur
Garbi Kadour	Maître de conférences A	Université d'El Oued	Examinateur
Hamoud Meriam	Maître de conférences A	Université d'El Oued	Examinateur

Proposition d'un modèle de synthétiseur de parole arabe réaliste à base de diphones

GUEDIRI Salah
ZEBDI Ammar

28 mai 2016

Dédicace

Tout d'abord, je veux rendre grâce à Dieu, le Clément et le Très Miséricordieux pour son amour éternel. C'est ainsi que je dédie ce mémoire à :

ma mère pour sa tendresse et mon père pour sa patience et encouragement
mes très chers frères et ma chère soeur pour leurs conseils,
mes cousins et cousines,
tous ceux que j'aime,
tous mes amies.

Remerciements

Tout d'abord nous remercions le Seigneur Dieu pour la santé, l'intelligence, le courage qu'Il nous'a accordé tout au long de notre cursus en particulier durant la période de réalisation de ce projet.

Un remerciement particulier à notre promoteur Monsieur Zaiz Faouzi, pour son encadrement, sa disponibilité, sa positivité et sa motivation, et surtout pour la confiance qu'il nous'a toujours accordé.

Aux professeurs qui nous'ont marqué durant nos cinq années de master, et qui nous'ont tant appris, dans le savoir que dans la rigueur et la persévérance.

Que les membres du jury trouvent ici nous remerciements les plus vifs pour avoir accepté d'honorer par leur jugement notre travail.

Nos sincères remerciements vont à tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce projet. En particulier à nous familles et à nos amis (es).

Résumé

La synthèse de parole joue un rôle très important dans le monde actuel parcqu'il rend les machines intelligentes et capable de discuter avec l'homme et de résoudre des problèmes complexes. Malgré touts les tentatives, jusqu'aujourd'hui aucune machine n'est capable de synthétiser de la parole à 100%. Parmis les applications de synthèse de la parole nous trouvons : Acapela, ReadSpeaker et linguatec... etc.

Ces applications possèdent des résultats de synthèse beaucoup plus réaliste pour des langues latines telles que : Français, Anglais,... etc. Malheureusement, ce n'est pas le cas pour la langue Arabe parce que la langue Arabe contient beaucoup de règles (morphologique, syntaxique, sémantique et linguistique).

Dans ce travail on s'intéresse d'une part à faire une étude globale concernant le domaine de synthèse de la parole à partie du texte arabe standard voyellé . Ensuite, nous allons présenté les composants nécessaires à la synthèse de la parole arabe. puis, on affinée par un intérêt particulier à une technique de synthèse qui se base sur des diphones.

Enfin, nous allons proposes un modèle de synthétiseur vocale à base de diphones qui permet de produire de la parole arabe plus réaliste en se basant sur un nouvel algorithme.

Mots Clés: Synthèse vocale, Traitement du son , Traitement de la prosodie , Traitement de la langage arabe .

Abstract

Speech synthesis play a very important role in the actual world because it makes machines intelligent and capable to talk with human and resolve complex problems. In spite of attempts, till now no machine is capable to synthesize realistic speech 100%. Among speech synthesis applications we found :Acapela, ReadSpeaker and linguatec. . . etc.

Those applications have much more realistic results for latin laguages such as : french, enlgish,...etc. unfortunately, it is not the case for Arabic language because it has many morphological, syntactic, semantic and linguistic rules.

In this work, we are interested on the one hand to make a global study concerning speech synthesis domain from a vocalized Arabic text. On the other hand, we will present different components needed to synthesize Arabic speech. Then, a particular interest will be held on a synthesis technique based on diphones.

Finally, we propose a model of an Arabic speech synthesizer based on diphones that allows to produce much more realistic Arabic speech based on a novel algorithm.

Keywords: Speech synthesis, Sound processing, Prosodic processing, Arabic language processing.

الملخص

إن تركيب الكلام أصبح يلعب دورا هاما جدا في عالم اليوم وذلك بجعل الآلة ذكية قادرة على التحاور مع الإنسان و حل المشاكل المعقدة، فبالرغم من المحاولات الكثيرة لتحقيق هذا لا توجد لحد الآن آلة تستطيع تركيب الكلام بنسبة مئة بالمئة ، ومن بين تطبيقات تركيب الكلام الموجودة في وقتنا الحاضر نجد على سبيل المثال : Acapela ، ReadSpeaker، linguatec ... إلخ .

هذه التطبيقات لها نتائج أكثر واقعية للغات اللاتينية مثل الفرنسية والإنجليزية ولكن للأسف ليست هذه الحال بالنسبة للغة العربية لأن اللغة العربية تحتوي على عدد كبير من القواعد (المورفولوجية والنحوية، الدلالية واللغوية).

في هذا العمل ركزنا أولا على إجراء دراسة شاملة عن مجال تركيب الكلام ومن ثم المكونات و الخطوات الضرورية لتركيب الكلام ومن ثم ركزنا على وجه الخصوص بمرحلة حاسمة في تركيب الكلام و هي مرحلة المعالجة الصوتية.

و أخيرا اقترحنا نموذج وصفي للصوت العربي موضحين استقرارية جودة النتائج المقدمة.

كلمات مفتاحية:

تركيب الكلام، معالجة الصوت، معالجة العروض، معالجة اللغة العربيةّ.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

La synthèse vocale à partir du texte (Text-To-Speech) permet de convertir un texte donné en un signal audio de parole. Ainsi peut-on imaginer de multiples usages pour cette technologies, tous plus utiles les uns que les autres. Il ne s'agit pas ici de remplacer l'homme, mais de le décharger de tâches ingrates et contraignantes. Prenons par exemple un annuaire téléphonique comprenant des milliers – voir des millions d'informations.

Le but de la synthèse de la parole à partir du texte est de générer automatiquement un signal de parole correspondant à un énoncé écrit. Les sources du texte prononcé peuvent être diverses : pour les personnes aveugles ou fortement malvoyantes (lecteur d'écran), système de réponse vocale, systèmes d'information, voire saisie au clavier de l'ordinateur, lecture de journaux. Ainsi les deux pôles de la synthèse de la parole sont d'un côté l'analyse et l'interprétation du texte, et de l'autre côté la prédiction des paramètres acoustico-phonétiques du son et la synthèse du signal.

Dans ce travail, Nous allons parler sur un modèle de génération automatique de la prosodie en arabe standard, à partir de marqueurs syntaxiques, dans le cadre de la synthèse de la parole à partir du texte par diphones — on entend par là un système de lecture à haute voix de textes par un ordinateur.

A l'heure actuelle il y a plusieurs applications de synthèse de la parole se trouvent pour des différentes langues telles que : Acapela, ReadSpeaker et linguatec... etc. Ces applications possèdent des résultats de synthèse beaucoup plus réaliste pour des langues latines telles que : Français, Anglais,... etc. Malheureusement, ce n'est pas le cas pour la langue Arabe parce que le résultat de la synthèse est affecté directement par les problèmes d'algorithmes de voyellation et de prosodie utilisés.

L'objectif de ce mémoire est de proposer un modèle de synthétiseur vocale à base de diphones qui permet de produire de la parole arabe plus réaliste en se basant sur un nouvel algorithme.

Ce mémoire s'articule autour de quatre chapitres :

Le premier chapitre, présente une vue générale des systèmes de synthèse de la parole, dont on s'intéresse à introduire et présenter un état de l'art du domaine de la synthèse de parole.

Le deuxième chapitre, illustre et expose des notions générales sur le traitement du langage arabe.

Le troisième chapitre, présente la conception du système ainsi que notre contribution, il s'agit d'un algorithme de résolution d'ambigüité.

Le dernier chapitre, présente la conception et l'implémentation de système proposé Suivie des teste et résultat obtenus .

Chapitre 1

Synthèse de la parole

Introduction

La synthèse vocale (ou TTS, pour Text To Speech) est une technique informatique qui consiste à transformer de manière automatique n'importe quel texte numérique écrit dans un langage précis en une suite de sons ou parole artificielle se rapprochant autant que possible de la parole humaine. Une des applications les plus évidentes de la synthèse vocale est l'accessibilité pour les mal-voyants.

Dans ce chapitre, nous présentons une vue sur le domaine de la synthèse vocale, de traitement de signal de parole et sur les principales techniques associées au prétraitement du signal de la parole.

1.1 Historique

Depuis des décennies, l'homme a essayé de rendre les objets rigides oralisées pour l'aider dans sa vie quotidienne. derrière ces objets rigides se cache un être humain qui utilise une motte de tuyaux pour enveryer un message oralisé, (voir figure 2.2)[13].

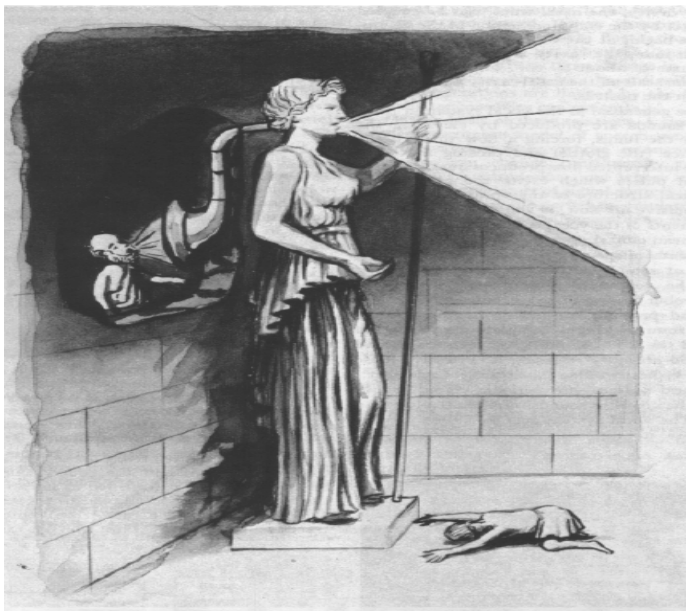

Figure 1.1: Exemple d'application de la « synthèse ».

En 1779, le savant "CHRISTIAN KRATZEUSTEU " in venta un dispositif (comprenant cinq lettre) pour la prononciation des cinq voyelles (a/e/o/u/i).

En 1791, le savant "VEN KEATSTONE " a développé l'ancien dispositif en le rendant capable de prononcer des mots et des phrases.

En 1837, le savant "CHARLES WHEATSTONE " a développé le dispositif crée par l'ancien savant " VEN KEATSTONE " le rendant capable d'écrire de la création des voyelles et des lettres, (voir figure 1.2)[13].

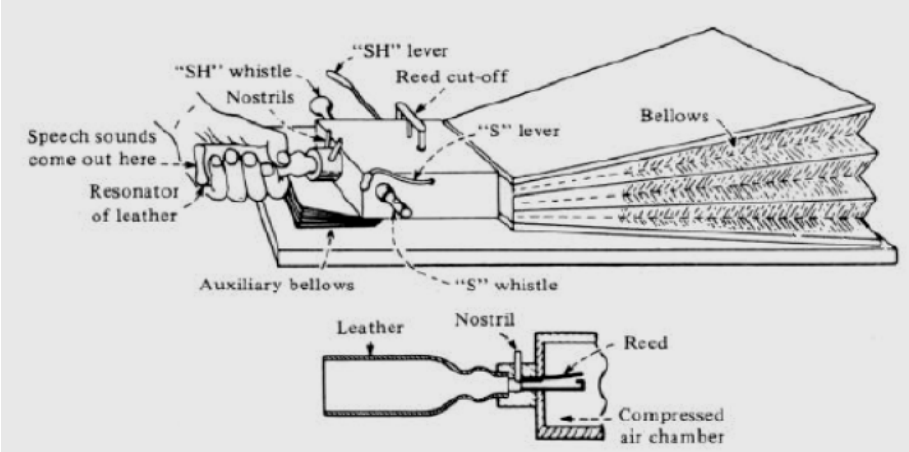

Figure 1.2: La machine de Von Kempelen (1791).

En1938 , la création d'un appareil électrique qui émet des sons en appuyant sur une pédale.
En1968, le savant japonais "NORIKO VMEDA " a mis en œuvre un appareil qui synthétise la parole et a en une reconnaissance mondiale. Et pour plus d'information sur le développement de ce dispositif et/ou l'idée de l'année 1800 jusqu'à l'an 2000, ce dispositif a connu 11 étapes, (voir figure 1.3)[13].

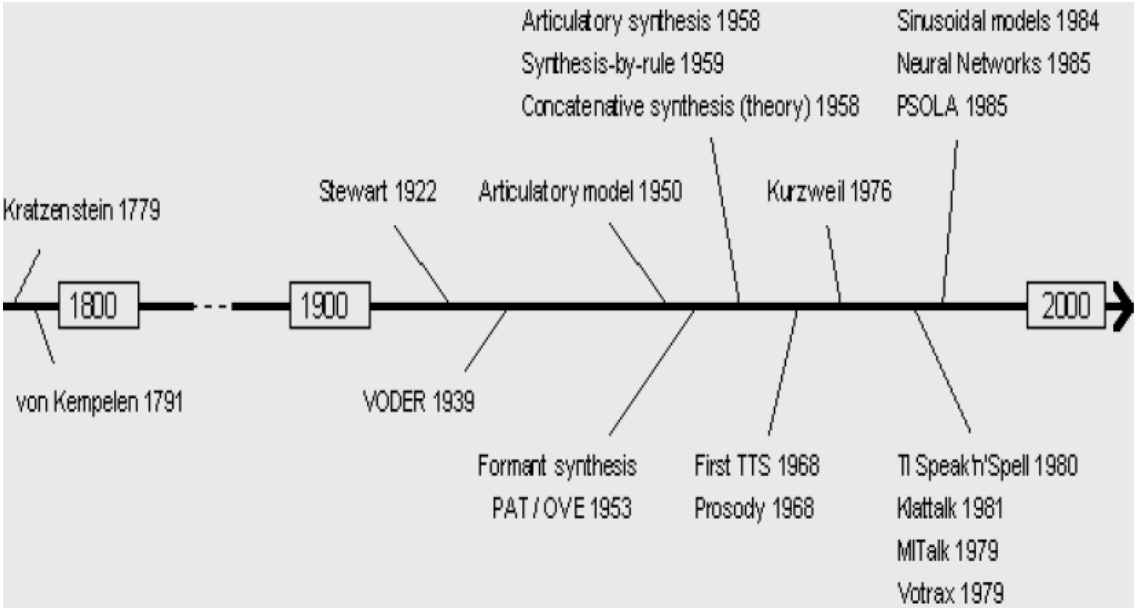

Figure 1.3: Les événements historiques les plus importants dans le (TTS) de 1800 à 2000.

1.2 Concepts de base

Dans cette section nous allons présenter quelques notions de base concernant le signal de la parole. Ces notions vont nous aider par la suite à bien délimiter le cadre de notre étude.

1.2.1 Définition de signal de parole

La parole (chose spécifique à l'être humain). est le principal moyen de communication dans toute société humaine. Elle est le produit de la pensée. D'ou, sa juste prononciation, sa juste articulation est indispensable pour une juste réception sémantique [1].

Le signal de la parole est un phénomène de nature acoustique porteur d'un message. L'information d'un message parlé réside dans les fluctuations de l'air, engendrées, puis émises par l'appareil phonatoire. Ces fluctuations constituent le signal vocal. Elles sont détectées par l'oreille qui procède à une certaine analyse. Les résultats sont transmis au cerveau qui les interprète [1].

1.2.2 Types de traitements de signal de la parole

Le signal de la parole est complexe et démontre une très grande variabilité car sa structure résulte de l'interaction entre la production des sons et leur perception par l'oreille et son traitement peut diviser à deux grands domaines principaux :

- Synthèse de la parole.
- Reconnaissance de la Parole.

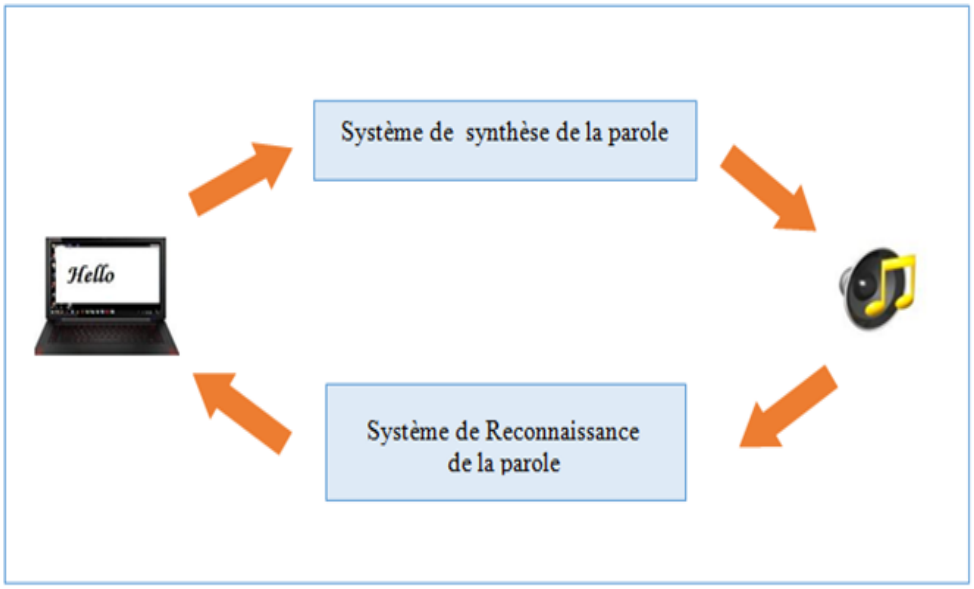

Figure 1.4: Schéma de synthèse et reconnaissance de la parole.

1.2.2.1 Synthèse de la parole

La synthèse vocale est une technique informatique de synthèse sonore qui permet à une machine de créer de la parole artificielle à partir de n'importe quel texte. Aucune restriction n'est faite sur la nature des mots à synthétiser (signale, abréviation, chiffre, date, etc.), ni sur la taille du vocabulaire à traiter. Parmi les applications, on peut citer la vocalisation d'écrans informatiques pour les personnes aveugles ou fortement malvoyantes(lecteur d'écran), ainsi que de nombreuses applications de serveurs vocaux téléphoniques, comme les annuaires vocaux de grande taille [3].

1.2.2.2 Reconnaissance de la Parole

La reconnaissance de la parole ou reconnaissance vocale est une technologie informatique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. Cette technologie utilise des méthodes informatiques des domaines du traitement du signal et de l'intelligence artificielle [8][18].

1.2.3 Caractéristiques du signal de la parole

Le signale de parole est un vecteur acoustique porteur d'informations d'une grande complexité, variabilité et redondance, dont les signaux de parole sont différencier par un ensemble des caractéristiques. Les caractéristiques de ce signal sont appelées traits acoustiques. Parmi ces caractéristiques on trouve :

- La fréquence fondamentale.
- Le spectre de fréquence.
- Le timbre.
- L'intensité.

1.2.3.1 Fréquence fondamentale

C'est le premier trait acoustique, c'est la fréquence de vibration des cordes vocales. Pour les sons voisés. Correspond à la période de l'onde. C'est la fréquence de cette onde qui nous permet d'évaluer de façon globale la hauteur du son. Les ondes qui accompagnent le fondamental sont appelées les harmoniques [16][10].

1.2.3.2 Spectre de fréquence

C'est le second trait acoustique dont dépend principalement le timbre de la voix. Il résulte de filtrage dynamique de signale en provenance du larynx ou signale glottique par le conduit vocale [16][10].

1.2.3.3 Timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix. Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son [16][10].

Les éléments physiques du timbre comprennent :

- Les relations entre les parties du spectre, harmoniques ou non.

- Les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence).
- L'évolution dynamique globale du son.
- L'évolution dynamique de chacun des éléments les uns par rapport aux autres.

1.2.3.4 Pitch

La variation de la fréquence fondamentale définit 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 des hauteurs tonales[16][10].

1.2.3.5 Intensité

L'intensité s'appelle aussi volume permet de distinguer un son fort d'un faible. L'intensité est liée à la pression de l'air en amont du larynx, qui fait varier l'amplitude des vibrations sonores [10].

1.3 Traitement de la parole

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine [6].

1.3.1 Niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur).
De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre.
L'opération de numérisation, schématisée à (la figure 1.5), requiert successivement : un filtrage de garde, un échantillonnage, et une quantification.

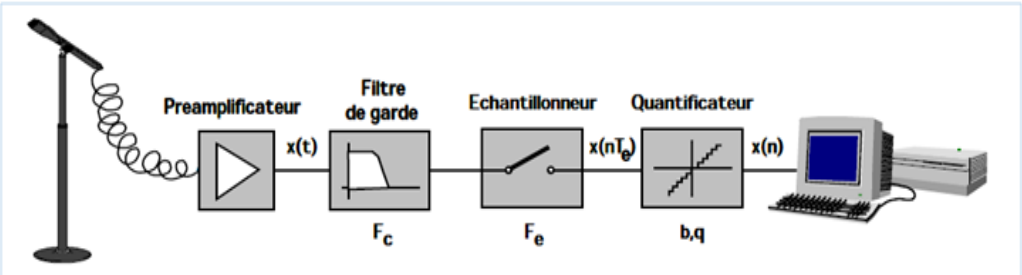

Figure 1.5: Schéma de (Enregistrement numérique d'un signal acoustique.) de la parole.

La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés fc, fe, b, et q [4].

1.3.1.1 Audiogramme

L'échantillonnage transforme le signal à temps continu x (t) en signal à temps discret X (nTe) défini aux instants d'échantillonnage, multiples, entiers de la période d'échantillonnage Te, celle-ci est elle-même l'inverse de la fréquence d'échantillonnage fe.
Pour ce qui concerne le signal vocal, le choix de fe résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz [14]. La figure 1.6 représente l'évolution temporelle, ou audiogramme du signal vocal pour les mots

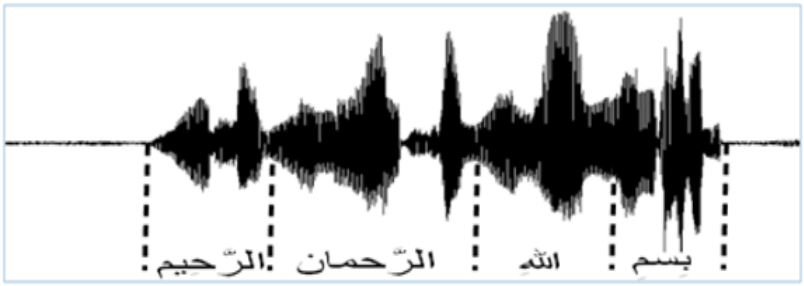

Figure 1.6: Audiogramme de signaux de parole.

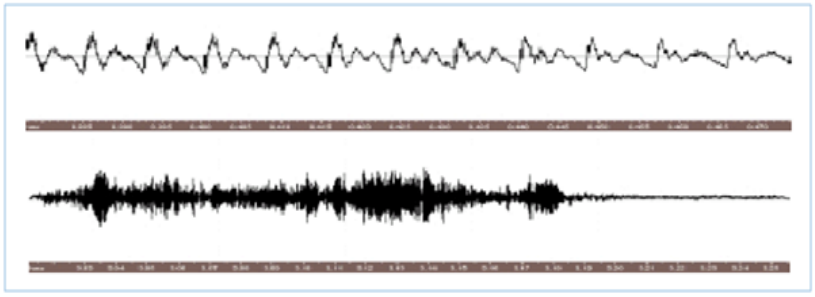

Figure 1.7: Exemples de son voisé (haut) et non voisé (bas).

1.3.1.2 Transformée de Fourier à court terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trentaine de ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant une transformée de Fourier sur ces échantillons. La figure 1.8 illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée enveloppe spectrale, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés formants et anti-formants [4].

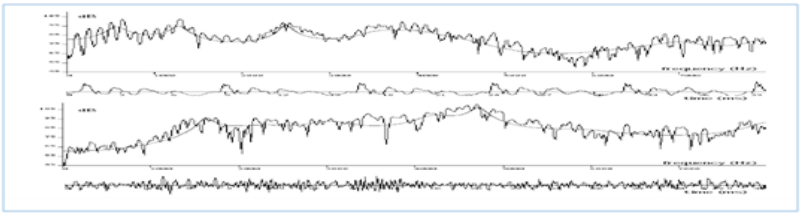

Figure 1.8: Evolution temporelle.

L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non-voisés présentent souvent une accentuation vers les hautes fréquences [4].

1.3.1.3 Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un Spectrogramme. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de Spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms), ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales[14].

1.3.1.4 Evolution de la fréquence fondamentale

Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou pitch. La figure 1.9 donne l'évolution temporelle de la fréquence fondamentale de la phrase « les techniques de traitement de la parole ». On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps.

Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [6].

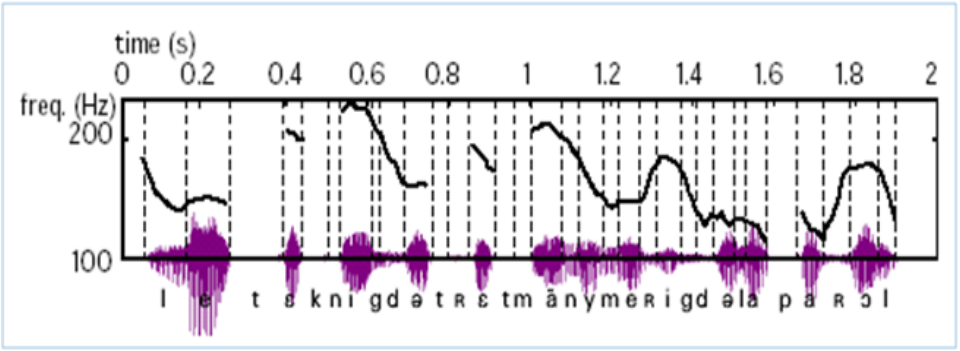

Figure 1.9: Evolution de la fréquence de vibration des cordes vocales.

1.3.2 Niveau phonétique

Le résonateur de l'appareil phonatoire est composé de quatre cavités principales en série (Figure 1.10) : le Pharynx ou arrière gorge, les deux cavités buccales délimitées par la langue et que l'on simplifiera à une seule et l'ajutage labiale situé entre les dents et les lèvres. La cavité nasale, en "parallèle" sur l'ensemble série précédent, vient compléter ce résonateur.

La source de ce résonateur est en fait décomposable en deux émissions distinctes et d'origines différentes. Les cordes vocales, en fournissant un spectre riche en harmoniques, produisent les sons voisés. Le bruit d'écoulement de l'air en provenance des poumons, dont le spectre est similaire à un bruit blanc, crée les sons non voisés.

Les sons et donc la parole naissent de l'excitation d'un résonateur et sont formés par les ouvertures et les volumes de ce dernier qui varient très rapidement. L'observation spectrale du conduit vocal laisse apparaître des pics de résonance, appelés formants. Les affaiblissements constatés dans le spectre, nommés anti-formants, sont introduits par les sons nasalisés [6].

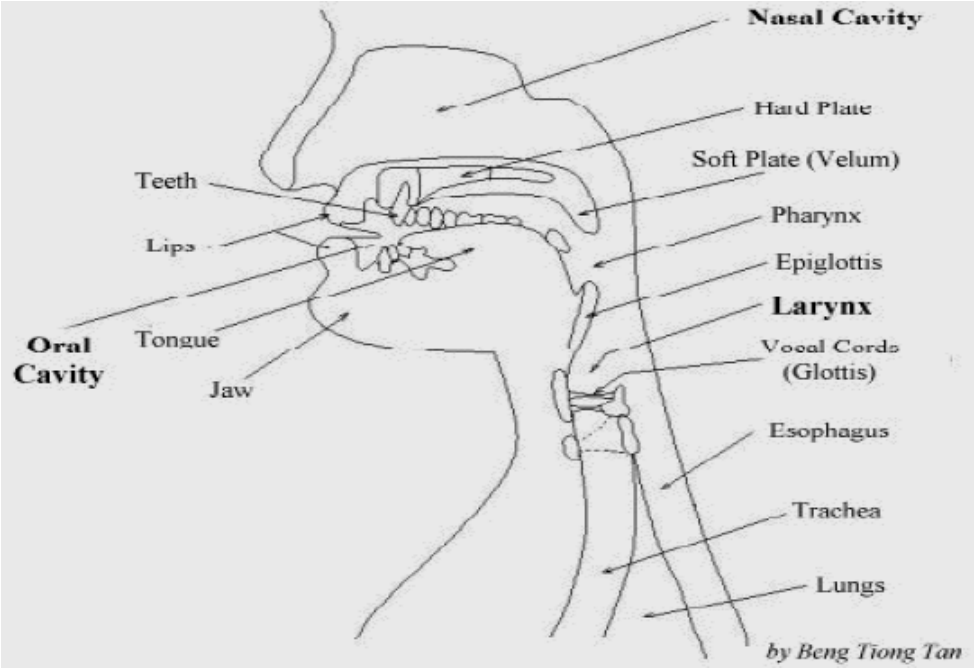

Figure 1.10: Illustration de l'appareil phonatoire.

1.3.3 Information vocale

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le timbre et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application, les phonèmes pour les machines de dictée par exemple.

La parole est surtout contenue dans les deux premiers formants, mais l'information proprement dite provient des transitions for mantiques. En général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100Hz-5KHz (300Hz-3.4KHz pour la téléphonie) [15].

1.4 Synthèse de la parole

La synthèse vocale est la génération automatique, par des dispositifs matériels et/ou des algorithmes, de parole artificielle. Il y a plusieurs types de synthèse vocale ; la plus complète est la synthèse à partir de texte (TTS, text to speech) où le but est de produire de la parole à partir d'un texte a priori inconnu » [1].

1.4.1 Principe d'un système TTS

L'objectif principale de la synthèse de parole à partir du texte donnée est de créer et de générer automatiquement une parole artificielle qui imite au mieux la voix humaine à partir de n'importe quel texte en utilisant des techniques artificielle et informatiser de traitement linguistique.

Un système de TTS se compose en général de trois parties (voir figure 1.12) . Les deux premières parties qui concernent les traitement de haut niveau permettent le passage de la représentation orthographique du texte en entrée à une représentation phonétique munie d'une description prosodique. la dernière partie englobe les traitement de bas niveau du synthétiseur qui permettent la génération proprement dite du signal acoustique [3].

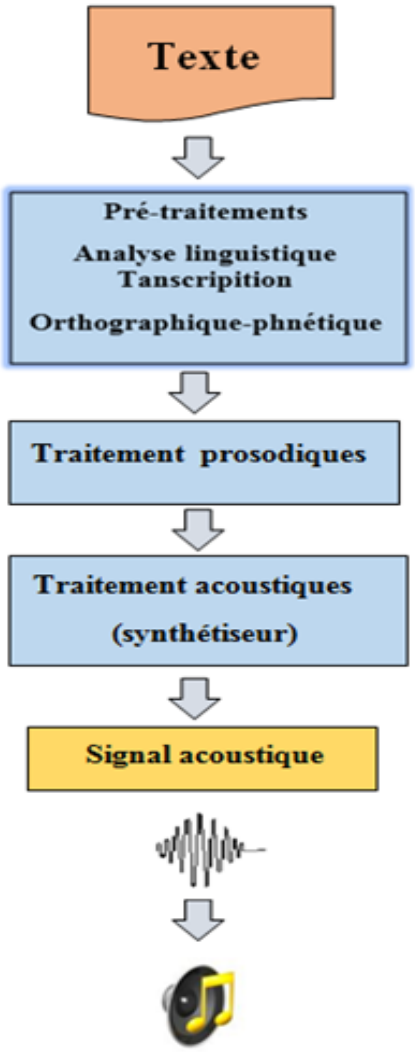

Figure 1.11: Schéma général d'un système de synthèse à partir du texte.

1.4.2 Méthodes de synthèse vocale

Afin de réaliser la tâche de synthèse vocale plusieurs méthodes ont été développées, Néanmoins, on note trois catégories principales, à savoir :

1. Synthèse par règles.
2. Synthèse par concaténation de diphones.
3. Synthèse à base de sélection des unités.

1.4.2.1 Synthèse par règles

Les synthétiseurs par règles sont basés sur l'idée que, si un phonéticien expérimenté est capable de «lire» un Spectrogramme, il doit lui être possible de produire des règles permettant de créer un Spectrogramme artificiel pour une suite de phonèmes donnée.

Une fois le Spectrogramme « dessiné », il ne reste plus alors qu'a générer le signal correspondant (à l'aide de générateurs et de résonateurs électriques) [7].

1.4.2.2 Synthèse par concaténation de diphones

a. Définition de Diphone : est l'extrait de la parole à partir du milieu d'un voix au milieu de la prochaine. Au milieu d'un voix à tendance à être sa région acoustiquement plus stable. par conséquent, diphones représentent des transitions acoustiques de la section médiane stable d'un voix à l'autre. diphone sont également distinct pour élargir le point médian de la partie de l'état d'équilibre du voix au point milieu de cette subséquence [13].

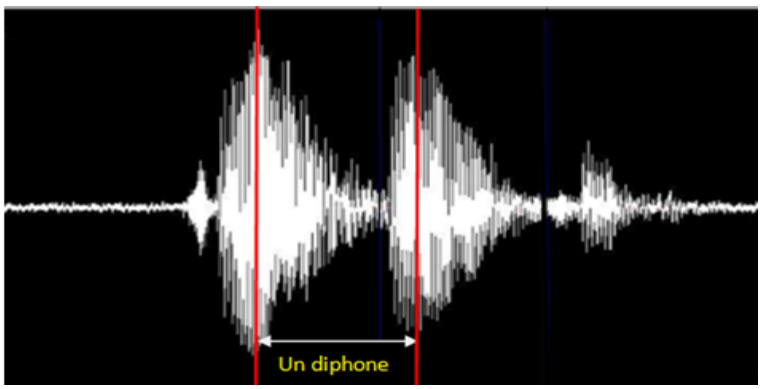

Figure 1.12: Exemple sur un diphone.

- Les synthétiseurs par concaténation de diphones sont basés sur l'idée que chaque phonème est converti à une unité acoustique adéquat, l'ensemble des unités acoustiques conformément de tous les diphones et la phonèmes possibles et candidats à utiliser dans un texte données sont enregistrés dans un grand base de données, cette base nommé "base des diphones" ou " base des unités acoustiques "et utiliser comme un entrepôt des unités acoustique candidats à utiliser dans l'étape de génération du parole synthétisé.

-Cette technique est basées sur L'idées de produire et de générer un signale du parole pour une suite de phonèmes donnée (texte phonétique) par la sélection de l'ensemble des unités acoustiques adéquats à cette suite des diphones qui sont déjà enregistrés dans la base des diphones et la concaténation ces unités acoustiques sélectionnées pour générer

le signale et le parole approprié.

Un problème supplémentaire apparaît cependant, du fait que les diphones utilisés ne respectent pas en général la prosodie que l'on cherche à produire [7].

1.4.2.3 Synthèse par sélection des unités

On assiste depuis quelques années à un important bouleversement, avec l'arrivée de techniques de sélection d'unités dans une grande base de données. Plutôt que de garder qu'un exemplaire de chaque diphone de la langue, on puise ici dans plusieurs heures de parole, préalablement segmentée phonétiquement.

Au moment de choisir les segments à mettre en œuvre (souvent des diphones), plusieurs instances d'une même unité phonétique sont alors disponibles, avec des prosodies différentes et positionnées (dans le corpus) dans des contextes phonétiques différents.

Il faut donc, pour réaliser au mieux la synthèse, choisir les segments dont le contexte est le plus proche de la chaîne phonétique à synthétiser, dont la prosodie se rapproche également le plus de la prosodie à produire, et dont les extrémités ne présentent pas trop de discontinuité spectrale l'une par rapport à l'autre. Ces techniques ont permis récemment de produire de la parole dont l'intelligibilité et le naturel rendent possible la confusion avec une prononciation humaine [7].

1.4.3 Applications de la synthèse de la parole

les applications TTS sont très diverses, nous allons exposer les plus courantes des applications du système de synthèse de la parole :[13]

Applications pour les non-voyants :
A partir d'un clavier spécifique aux non-voyants pour les aider dans leur vie quotidienne, celui-ci emmagasine les résumés de phrases (ou de mots) et en quelques secondes, les traduit en phonétique qu'on entend facilement.

Applications pour les sourde-muets :
On peut mettre en œuvre l'application (TTS) pour combler les besoins des différents handicapes (handicap mentale, sensu-articulatoires ....).

Application éducatives :
Les mots composés peuvent être employés dans différentes situations éducatives. Ils sont en effet des moyens efficaces pour l'apprentissage de nouvelles langues à l'aide d'un ordinateur. Nous pouvons l'employer dans les application didactiques inter-actives.

Recherche fondamental et les appliquées :
Les application (TTS) ont un privilège unique , ce qui le rend capable de traduire un texte écrit à l'oral (par les vibrations sonores, ressemblants à la voix humaine).

Services gouvernementaux :
Les bureaux des établissements gouvernementaux récrivent des demandes de renseignement (information sur les impôts ,les revenus ,....etc.).
Nous pouvons nous passer de ceci en ayant recours au programme auditif intégré à nos

portable, ce dernier est applicable pour ouvrir ou fermer les portes des établissements....

Application en communication :

Dans ces système , nous avons l'accès aux information (Textes) à l'aide du téléphone portable en le questionnant (besoin du demandeur) à l'aide des touches de ce dernier.

Applications Multimédia :

Pour une communication appareils-être humain ,ceux-ci aident l'être dans différents domaines (ex : le G.P.S dans la voiture aide le chauffeur en lai dictant les ordres par des phrases sonores ,rechercher fondamental et les appliquées).

Conclusion

Dans ce chapitre, nous avons fait une étude sur traitement de signal de parole et méthode de synthèse de la parole. Au terme de ce bilan rapide sur la synthèse vocale, on a pu constater que ce domaine est particulièrement vaste et qu'il n'existe pas de produit miracle capable de répondre à toutes les applications.

Chapitre 2

Traitement de la langue arabe

Introduction

L'objectif du traitement automatique des langues est la conception de programmes capables de traiter des données exprimées dans une langue naturelle pour les quels plusieurs phases d'analyse (morphologique, syntaxique, sémantique et pragmatique)sont nécessaires afin d'en extraire des informations.

Alors, pour bien prononcé ou traité une langue revient à bien reconnaître les différents composants, articulation et règles de cette langue. L'arabe est une langue sémitique de la même famille que le syriaque, l'araméen et l'hébreu. Il est parlé aujourd'hui par plus de 300 millions d'habitants dans le monde et 22 pays. Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique des langues [2].

Les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie. Dans ce chapitre, nous allons présenter les différents éléments et techniques en relation avec la langue Arabe des côtés écritures, transcription et prononciation [2].

2.1 Etude de la langue arabe

L'arabe s'écrit et se prononce de droite à gauche et lie ces différentes lettres en utilisant des règles de ligatures bien précises. Elle possède un alphabet composé de 28 lettres qui sont toutes des consonnes (voir la figure 2.1)[9].

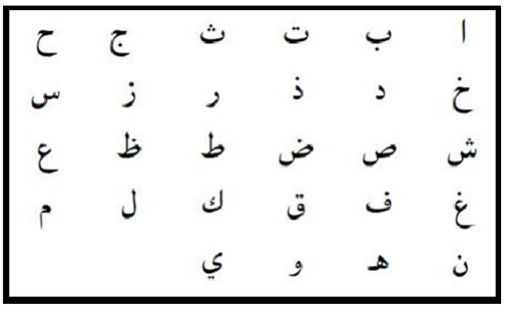

Figure 2.1: L'alphabet de la langue arabe.

— parmi les 25 consonnes et le reste (أ و ي) des voyelles longues ou consonnes selon leurs positions dans le mot, par exemple : la lettre (ي) est une voyelle longue dans le mot (فيل) → /filun/ et consonne dans le mot (يد) →/yadun/ [11][12].

— Pareillement, les lettres prennent des différentes formes selon leurs positions dans le mot [12].

Par exemple

Isolée	Finale	Milieu	Début
ع	ـع	ـعـ	عـ
ه	ـه	ـهـ	هـ

Tableau 2.1: Exemple de variation de la forme d'écriture des lettres ه et ع.

2.1.1 Signes diacritiques

Dans la langue arabe on trouve un ensemble des voyelles brèves appelés signes diacritiques. Ces symboles sont ajoutés au dessus ou au-dessous des lettres afin de résoudre le problème de complexité et d'ambigüité engendré par un texte non voyellé. Cette opération est souvent appelée vocalisation automatique ou voyellation automatique. Ils sont transcrits de la manière suivante :[5][2]

Le voyelle	transcrit	exemple
Fetha signé par « ´ »	symbolisée par un petit trait sur la consonne	بَ = transcrits comme " Ba "
Damma signé par « ’ »	symbolisée par un crochet au dessus de la consonne	بُ = transcrits comme " Bu "
kasra signé par « ¸ »	symbolisée par un petit trait au dessous de la consonne	بِ = transcrits comme "Bi "
soukoun signé par « ° »	apposé sur une consonne lorsque celle-ci n'est liée à aucune voyelle	بعْد = transcrits comme"ba$ da"

Figure 2.2: Exemple d'application de la « synthèse ».

Exemple du problème de complexité et d'ambigüité en lecture et en interprétation du mot كتب « un mot non voyellé» :[5]

Mot sans voyelles	1ère interprétation		2ème interprétation		3ème interprétation	
كتب	كتب	Il a écrit	كَتَبَ	Il a été écrit	كُتُبٌ	Des livres
مدرسة	مَدْرَسَةٌ	école	مُدَرِّسَةٌ	enseignante	مُدَرَّسَةٌ	enseignée.

Tableau 2.2: Exemple du problème de complexité du mot « كتب et مدرسة »

2.1.2 Tanwin

Les signes du tanwin sont au nombre de trois et sont ajoutés à la fin des mots indéterminés. Ils sont constitués par dédoublement des signes Fetha, Damma et Kasra. Au niveau phonétique ils se traduit par l'ajout du phonème /n/, comme suit :[3]

Signe	Phonème	Consonne	Transcription
« ً »	[an]	بًا	"Ban"
« ٌ »	[un]	بٌ	"bun"
« ٍ »	[in]	بٍ	"bin"

Tableau 2.3: Transcription des signes tanwin dans la langue Arabe.

2.1.3 Chadda

Un autre signe, la chadda peut être placé au-dessus de toutes les consonnes en position non initiale. Il consiste à doubler la prononciation de la consonne :[3]

Mot en Arabe	Prononciation	Signification
كَلَّمَ	Kallama	il a parle à

Tableau 2.4: Exemple de chadda dans le mot« كلم »

2.1.4 Gémination

Elle est présentée par la prononciation redoublée d'un phonème. En Arabe, elle est symbolisée par le signe de chadda qui signifie le doublement d'une consonne. Une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position par exemple :[3]

Mot en Arabe	Prononciation	Transcription
حضّرَ	/haDDara/	/haD :ara/

Tableau 2.5: Exemple de gémination, le cas du mot« حضر »

2.1.5 Madd

L'utilisation des voyelles longues (و /U/,ا /A/ ou ي /I/) génère un allongement dans la prononciation appelé en Arabe madd. En lecture on utilise des règles phonologiques qui ont des traits à la contraction des sons.[3]

2.1.5.1 Contraction

C'est la réunion de deux lettres, de deux syllabes ou de mots en un seul. Elle peut être obligatoire comme dans le mot قُلْ لَهُ qul/ / lahu = قُلَّهُ /qullahu/, interdite comme dans le mot مللت /,le premier /l/ ne doit pas être contracte avec le second /I/, ou permise par exemple le mot سرر /sarara/ = /sarra/[3][12].

2.1.5.2 Élision

Elle présente le changement produit par la prononciation du soukoun ou phonème /n/ devant certaines consonnes [3][12].

2.1.5.3 Assimilation homo-organique des nasales

C'est la substitution d'une consonne. elle peut se produire à l'intérieur du mot, par exemple أنبتتْ /anbatat/ = أمْبتتْ /ambatat/) ou a la frontiéres de deux mots successifs comme dans le mot من بعد / min baed/ =ممبعد /mimbaed [3][12].

2.2 Problème confronté du traitement automatique de la langue arabe

Généralement, le traitement de l'Arabe possède deux principaux problèmes :

- l'agglutination des mots.
- texte non vocalisé.

2.2.1 Agglutination des mots

La plupart des mots arabes sont composés par agglutination d'élément lexicaux de base (proclitique + base + enclitique) . Par exemple, la détermination peut s'exprimer par agglutination de l'article ال /?al/ avant le mot (الوَلَدُ /?alwaladu/ («l'enfant ») ou par agglutination d'un pronom personnel après celui-ci (ضَرَبَهُ / drabahu / («il l'a frappé»), وَلَدُهُ / waladuhu / («son enfant ») . Les particules régissant le cas indirect aux noms (كَدَارِهِ / kadArihi / (« comme sa maison ») et les conjonctions de coordination aux verbes (فَذَهَبَ / favahaha (« et il est parti »,etc[3].

2.2.2 Voyellation

Par convention la voyellation d'un mot représente l'ensemble des voyelles associées aux consonnes de ce mot. Un mot est dit ambigu s'il admet plusieurs voyellation potentielles hors-contexte. Par exemple la forme graphique (كتب) admet cinq interprétations possibles (voir le tableau 2.6). Un texte arabe complètement voyellé est donc un texte non ambigu, par contre, un texte non voyellé est un texte ambigu [5].

Signification	Prononciation	Mots en arabe
Il a écrit	/kataba/	كَتَبَ
Il a été écrit	//kutiba//	كُتِبَ
Il a fait écrire	/kattaba/	كَتَّبَ
Il A été écrit	/kutdba/	كُتِّبَ
Fais écrire	/kattib/	كَتِّبْ
Des livres	/kutub/	كُتُبٌ
Un écrit	/katb/	كَتْبٌ

Tableau 2.6: Exemple d'analyse morphologique du mot« كتب »

2.3 Analyse linguistique

Pour faire une analyse linguistique d'un texte écrit dans une langue donnée, il faux reconnaître la structure de la phrase, en identifiant les sujets, les objets des verbes, les groupes,...etc. Ce qui nécessite des connaissances syntaxiques ou grammaticales exprimant les relations entre les mots. Elle se déroule en deux phases (voir la figure 2.3) :[3]

- Une analyse morpho-lexicale qui assigne a chaque token (mot au sens large) un ensemble d'étiquettes morphologiques hors contexte.
- Un traitement syntaxique qui fournit l'ensemble des structures acceptable de la phase.

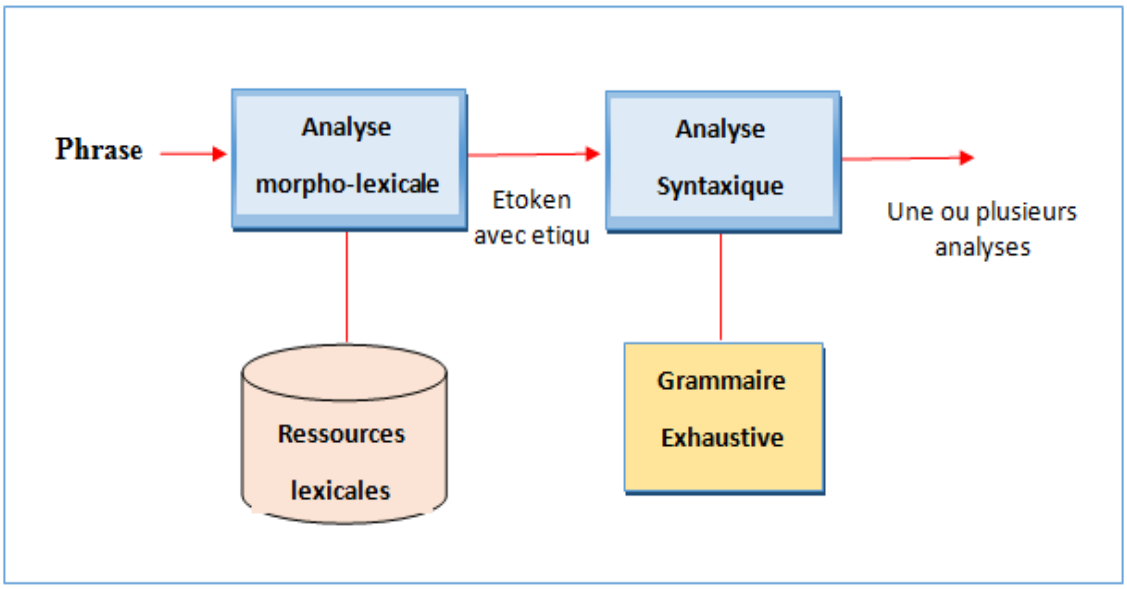

Figure 2.3: Analyse syntaxique traditionnelle.

2.3.1 Morphologie de la langue arabe

L'analyse morphologique est importante en Arabe parce que les mots sont formés dans leur majorité par assemblage d'unités lexicales élémentaires. Le rôle de l'analyse morphologique est alors de :

- Découper ou segmenter les mots en unités lexicales élémentaires.
- Attribuer a chaque unité une ou plusieurs valeurs morphologiques.

Par exemple : la segmentation du mot **فذهب** / favahaba /«puis il est parti» en **ف**/fa/ (particule de coordination) + **ذهب** /vahaba/ (verbe **ذهب**) est valide, par contre la segmentation du mot **فرح** /faraha/«il était content» en **ف** /fa/ + **رح** /raHa/ est invalide (l'élément /raHa/ n'appartient paseux lexiques de la langue). Plusieurs analyses morphologique valides peuvent être obtenues pour un mot donné : le mot **كتب** /ktb/ (sans voyelles) peut produire **ك** /k/ + /**تب** /tabba/ «comme trancher» ou **كتب** / kataba /«écrire» (ce mot non voyelle présente 9 voyellations possibles . Dans ce cas, des connaissances de niveaux supérieurs (syntaxique, Sémantique . . .) sont nécessaires pour lever l'ambigüité.

Ensuite, il faut assigner aux éléments lexicaux des valeurs morphologiques. Ces valeurs ont trait au type de l'élément (verbe, adjectif, pronom, etc.) au genre (masculin, féminin), au nombre (singulier, duel, pluriel), . . . etc. Un élément peut avoir plusieurs valeurs morphologiques ce problème est appelé l'homographie poly-catégorielle qui est traitée au niveau des modules en aval [3].

2.4 Prosodie arabe

La prosodie, c'est l'étude des phénomènes de l'accentuation et de l'intonation (variation de hauteur, de durée et d'intensité) permettant de véhiculer de l'information liée au sens telle que la mise en relief, mais aussi l'assertion, l'interrogation, l'injonction, l'exclamation. Elle constitue un ensemble d'éléments ou facteurs qui assignent à la parole son enveloppe « musicale » et englobe tous les gestes mélodiques, dynamiques et temporels que manipulent des phénomènes tels l'intonation, l'accentuation , l'hauteur, et l'intensité et le durée qui déterminent la mélodie, le ton, la pause et le rythme .

Au niveau acoustique, elle se voie par les différents paramètres liés à la fréquence fondamentale (estimation de son laryngien a un instant donné sur le signal), la durée (intervalle de temps entre deux points sur le signal) et l'intensité (énergie contenue dans le signal). Autant que du côté perception de la parole se présente par l'étude de l'accentuation et de l'intonation (variation de hauteur, de rythme et d'intensité) permettant de véhiculer l'information liée au sens de la phrase .

Parfois, la prosodie est désignée par le terme suprasegmental. Ce qui veut dire qu'elle dépasse le cadre du phonème. S'étendant de la syllabe jusqu'à la phrase, atteignant pour certains le paragraphe. Parmi ces paramètres suprasegmentaux on note l'accent, l'intention et le rythme. La prosodie possède des traits similaires qu'on peut les considérés universels pour plusieurs langues. La tendance de l'intonation a descendre a la fin des phrases assertives et a augmenter a la fin des phrases interrogatives constitue sans doute l'exemple le plus significatif. Malgré tous ça, cette universalité ne gène pas les spécificités de chaque langue ou chaque individu dans la réalisation des énoncés : la prononciation d'un texte peut être différentes selon les caractéristiques anatomique du locuteur, sa région, sa société, son état émotif, son tempérament, ... etc [11][19].

2.4.1 Types de la prosodie

Généralement, On distingue deux sortes de prosodie, à savoir :

1. **Prosodie linguistique** est utilisée lorsque l'on accentue un mot ou un groupe de mots afin de mettre en évidence des éléments de la phrase (lorsque l'on fait un reproche par exemple : « c'est TA faute ») et aide à la compréhension de l'énoncé. Elle permet également de recourir à un mode en particulier(déclaratif, interrogatif, exclamatif, etc.) [11].
2. **Prosodie émotionnelle** appelée également prosodie affective, quant à elle, sert à véhiculer un message avec une composante émotionnelle. La tonalité de l'énoncé change afin de transmettre un message émotionnel qui peut être positif, négatif ou neutre.

La prosodie peut également être congruente ou non avec le contenu sémantique de la phrase. Dans le cas de la congruence, le ton émotionnel utilisé correspond au contenu sémantique du discours. L'humour, le sarcasme ou l'ironie amènent à des situations de non congruence car la prosodie émotionnelle et l'aspect sémantique transmettent dans ce cas un message différent [11].

2.4.2 Fonctions de la prosodie

Généralement, la prosodie possède des fonctions qui peuvent être vu en deux catégories :

1. Linguistiques, véhiculant des informations modales et structurelles sur l'énoncé (expression de la modalité, segmentation du continuum et hiérarchisation).
2. Para-linguistiques, sur le locuteur et ses rapports avec son discours et ses interlocuteurs.

Mais, on peut les étendre en quatre points :

- La fonction d'expression de la modalité permet d'avoir le mode d'inter-locution d'une phrase (assertion, question, exclamation, ordre).
- L'intonation montante sur la frontière des unités prosodiques traduit la continuation, et celle descendante traduit une finalité. En plus, une intonation haute utilisée dans les questions oui/non manifeste l'attente d'une réponse.
- La segmentation permet de distinguer et d'incorporer mentalement les unités de composition : les débuts et fins de paragraphes, phrases, des groupes syntaxiques.
- La fonction d'hiérarchisation extraire la structure d'un énoncé en un ensemble de niveaux (dérivation, emboitement des constituants), par la séparation des informations de premier et second plan,. . . etc [11][19].

Conclusion

Nous avons vu dans ce chapitre une étude sur les différentes techniques de traitement de la langue arabe, ainsi, que des informations sur ce dernier nécessaire pour un SAT arabe. De plus, une étude syntaxique et morphologique a été faite, mais cette étude restent insuffisantes à cause de la quantité vocabulaire et de significations qui rendent le processus de vocalisation d'un texte arabe très difficile.

Chapitre 3

Conception et implémentation du système

Introduction

Après avoir présenter une étude globale sur le domaine de la synthèse de la parole dans le premier chapitre, ainsi qu'un panorama sur le traitement de la langue Arabe dans le second chapitre. Dans ce chapitre, nous allons présenter l'approche de lissage proposée qui donne la voix la plus naturelle possible tout en tenant compte des particularités de la langue, ainsi que les différents modules du système de synthèse de la parole Arabe en commençant par un schéma général des différentes phases, ensuite, nous allons détailler à part chaque phase de ce processus.

3.1 Méthode de concaténation proposée

Avant d'être produit une parole synthétique passe par une succession de phase qui commence par l'acquisition d'un texte voyellé, sa normalisation, transcription et sa concaténation. Cette dernière, consiste à joindre les différentes unités acoustiques générée durant les phases qui la précèdent. Mais cette jointure génère un signal anormal parce que ces unités sont pris des différents enregistrements ou ce qui cause des discontinuités dans la parole produite. Afin de remédier à ce problème on utilise souvent des algorithmes de lissage pour améliorer la qualité du son ou de la parole générée.

Saidane et al.[17] propose un algorithme de lissage qui procède par l'application d'une technique de lissage simple qui applique d'atténuation ainsi définie a été appliqué pour un nombre de points représentants 10 % de la durée du signal de l'unité acoustique. Cette approche, améliore la qualité du signal mais un chevauchement entre les unités reste. Afin de l'éliminer une durée de silence de 10 millièmes de seconde a été introduite.

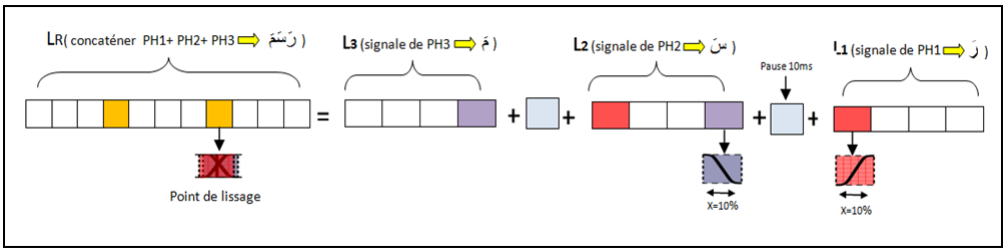

Figure 3.1: Exemple sur méthode de concaténation proposée, le cas de mot « رسم ».

Dans notre travail nous avons adopté l'idée d'utiliser une durée de silence ou pause variable. Les résultats de cette méthode ont été comparé avec celles de la méthode proposée par Saidane et al. dans la section 4.5 .

3.2 Les différents phases du système

L'objectif de notre système est la synthèse vocale à partir d'un texte voyellé donné. Il commence par l'acquisition du texte qui doit être voyellé comme entrée, ensuite, ce texte passe par trois étapes principales : 1) une phase de transcription orthographique phonétique est faite afin de donner une représentation en symbole phonétique du texte, 2) le texte est normalisé en utilisant un ensemble de règles de normalisation afin d'éliminer les abréviations des mots tel que km (كلم), 3) une succession des traitements acoustiques est réalisée pour améliorer la qualité du signal vocal produit (voir figure 3.2). Ces traitements passent généralement par quatre étapes :

- Préparation de la base de diphones à utiliser.
- La sélection (Choix) des unités acoustiques .
- La concaténation des unités acoustiques sélectionnées.
- L'amélioration de la qualité du signale à produire.

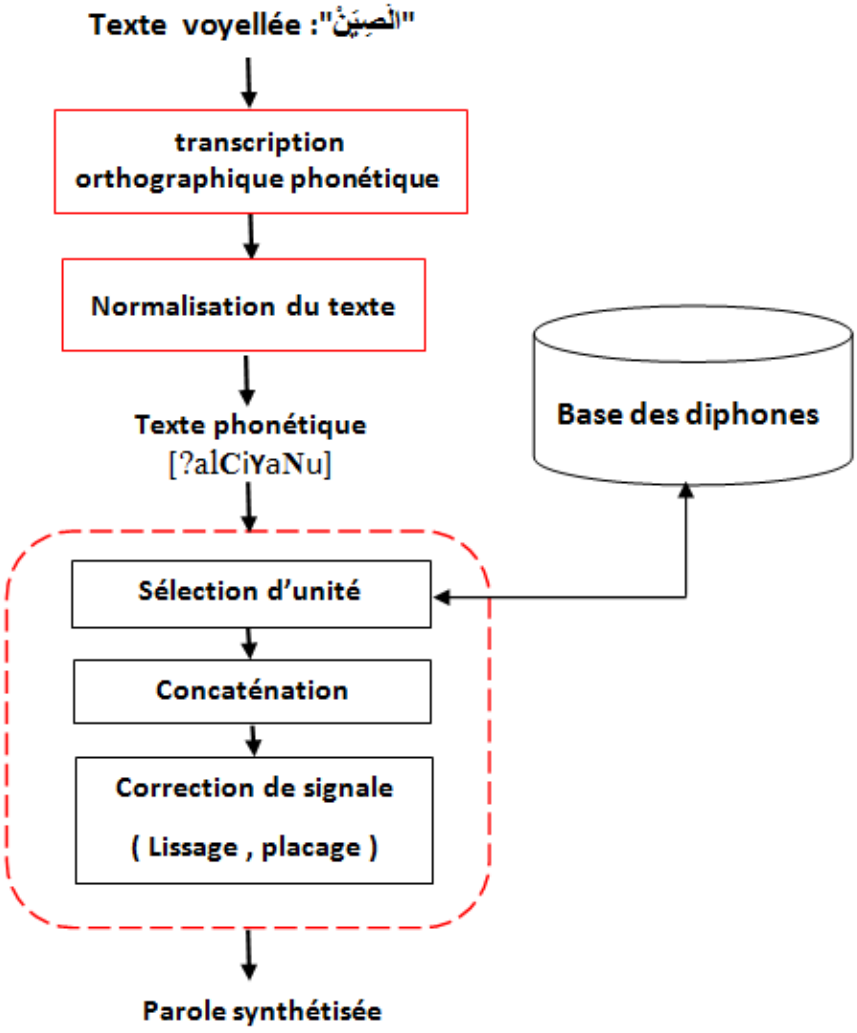

Figure 3.2: Schéma général du système.

3.2.1 Transcription orthographique phonétique

Cette phase consiste à associer des symboles phonémiques à chacun des caractères du texte voyellé généré en utilisant des règles de transcription graphème phonème comme le montre le tableau 3.2 ci-dessous.

Phrase	Prononciation
خرج عمر	/X//a//R//a//J//a//£//o//M//a//R//o/

Tableau 3.1: Exemple sur Transcription orthographique phonétique .

Le système doit appliquer une étape de prétraitement sur le texte orthographique dans le but de détecter les ponctuations, espaces et sauts de lignes. Ensuite, si le système détecte des exceptions il doit les traiter avant de commencer la transcription phonétique en utilisant une base de lexique sinon il commence l'opération de transcription en se basant sur une base de règles de transcription. Le résultat de cette phase est un autre texte avec des symboles phonétique représentant chacun le symbole adéquat à l'unité sonore ou acoustique à la prononciation d'un caractère (voir la figure 3.3).

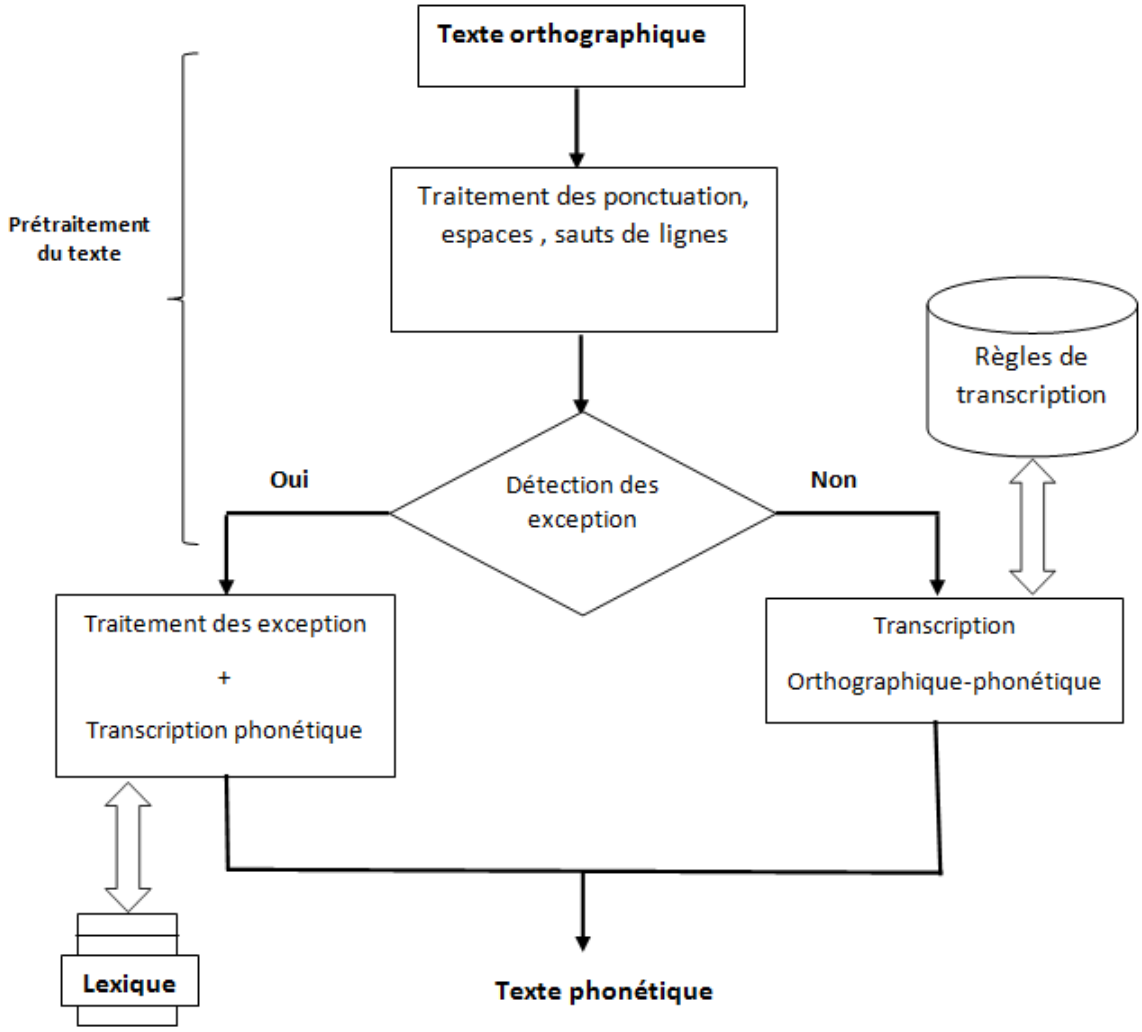

Figure 3.3: Architecture du système de phonétisation automatique.

Tableau 3.2: Symboles graphème-phonème de la langue Arabe.

Consonnes			
Graphème Arabe	Symbole phonémique	Graphème Arabe	Symbole phonémique
أ	/A/	ض	/ç/
ب	/B/	ط	/ù/
ت	/T/	ظ	/%/
ث	/€/	ع	/£/
ج	/J/	غ	/Ś/
خ	/X/	ف	/F/
ح	/H/	ق	/Q/
د	/D/	ك	/k/
ذ	/U/	ل	/L/
ر	/R/	م	/M/
ز	/Z/	ن	/N/
س	/S/	ه	/H/
ش	/$/	و	/W/
ص	/C/	ي	/Y/
ى	/¤/	ئ	/ý/
voyelles			
◌َ	/a/	◌ًا	/a :/
◌ِ	/i/	◌ٍي	/i :/
◌ُ	/u/	◌ٌو	/u :/

3.2.2 Normalisation du texte

La normalisation du texte a pour objectif de transformer un texte pouvant contenir des sigles, nombres, emails, fax,...etc en un texte normalisé écrit seulement avec des lettres. Elle consiste à faire premièrement une tokenisation du texte, ensuite une phase de près-traitement est faite dans le but de transcrire en toutes lettres l'ensemble des mots en particulier les mots extra-lexicaux comme les nombres et abréviations.

3.2.2.1 Traitement des mots spéciaux

Un sigle en arabe est forme abrégée construite en concaténant les consonnes initiales d'une séquence de mots. Par exemple : **م.خ.و** = **مكتب الخدمة الوطنية** (Bureau de services nationales). Cette opération est faite généralement en se basant sur une base de lexiques pour les abréviations. Presque de la même façon le système utilise une base similaire pour transformer les nombres, dates et heures et spéciales...etc. Le tableau 3.3 illustre des exemples de cette traitement.

Le texte arabe	**Type des mots spéciaux traités**	**Le Texte normalisé généré**
13 سيالةٌ	Un nombre	ثلاثُ عشرةَ سيالةٌ
5 م	Un abréviation	خمسةٌ أمتارٌ
1 :15	Heure	الواحدة و الربع

Tableau 3.3: les exemples sur Traitement spéciaux.

3.2.2.2 Traitement des mots irréguliers

Les mots irréguliers sont répertoriés dans un lexique d'exceptions de 20 entrées qui fournit les formes phonétiques. Ces mots ne possèdent pas une prononciation qui correspond à leur graphie. Le tableau 3.4 montre des exemples de ce traitement (voir la figure 3.4).

Le mot	Prononciation	Transcript
هذا	/havA/	/ HaUA/
ذلك	/valika/	/ UaLika /

Tableau 3.4: Les exemples sur Traitement des mots irréguliers.

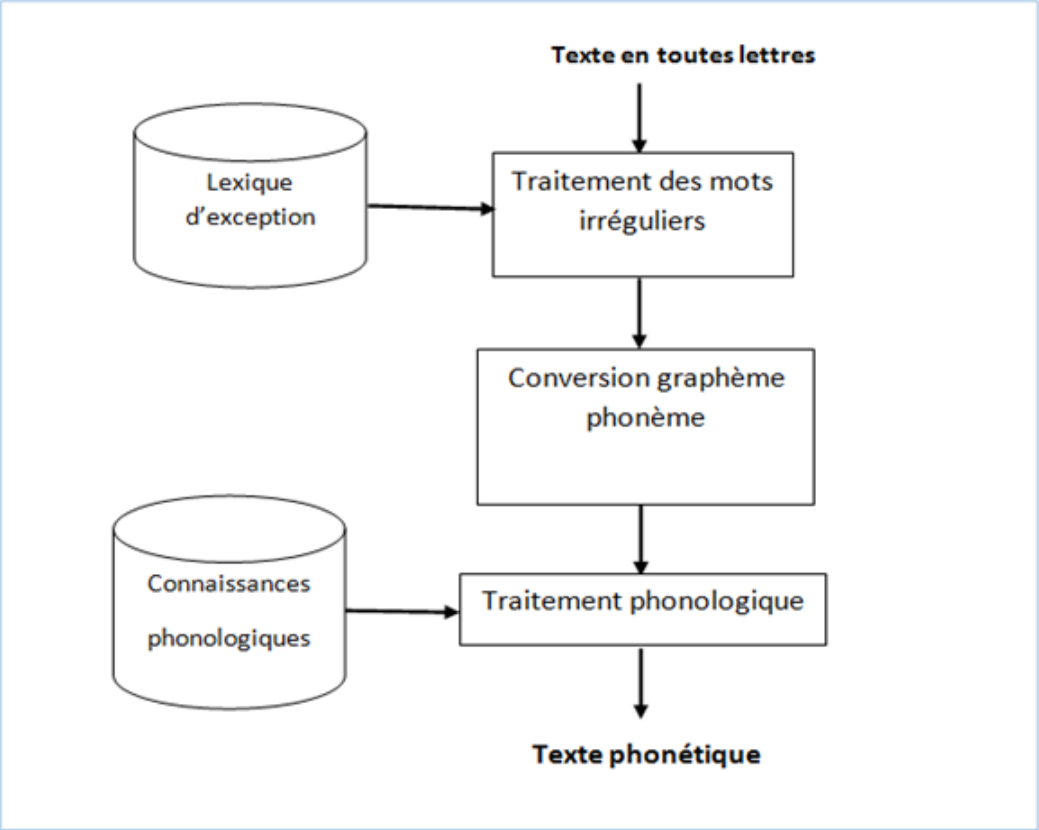

Figure 3.4: Module de transcription orthographique-phonétique.

3.2.2.3 Conversions spéciaux

Durant le processus de près-traitement il faut ajuster la conversion de quelques lettres selon leur nature ou voyellation tels que chadda, madda, tanwin. . . .etc.Le tableau 3.5 montre des exemples de ce traitement.

Lettre arabe	Le phonème	Le exemple	La Transcription	Signification
وا	[?U]	خرجوا	/XaRaJ ?U/	Ils sont sortis
آ	[?A]	آلي	/ ?ALaYi/	automatique
ىً	[An]	فـتىً	/FaTaAn/	enfant

Tableau 3.5: les exemples sur Conversions spéciaux.

3.2.2.4 Règles d'assimilation

Les règles d'assimilation rendent compte des transformations de l'article ال / ?al/ selon sa position dans la phrase et la nature du phonème (lunaire/solaire) qui le suit. Le tableau 3.6 illustre des exemples de l'application de ces règles.

Lettre arabe	Le phonème	la nature du phonème	L'exemple	La Transcription	Signification
ال	/ ?al/	Lunaire	الماء	/ ?aLMaAaýa/	Eau
ال	/ ?a/	Solaire	النهار	/ ?aNaHaRa/	lajourné
ال	/al/	Lunaire	فوق الماء	/FaWaQaMaýa/	sur de l'eau
ال	/i/	Solaire	في النهار	/FiNaHaRa/	Dans la journée

Tableau 3.6: Les exemples de Règles d'assimilation.

3.2.3 Traitements acoustiques

Dans cette phase, nous allons présenter les étapes nécessaires pour la transition de Texte phonétique à une parole synthétisée, ces étapes consistant en :

1. Préparation de la base de diphones à utiliser.
2. La sélection des unités acoustiques.
3. La concaténation des unités acoustiques sélectionnées.
4. L'amélioration de la qualité du signale à produire.

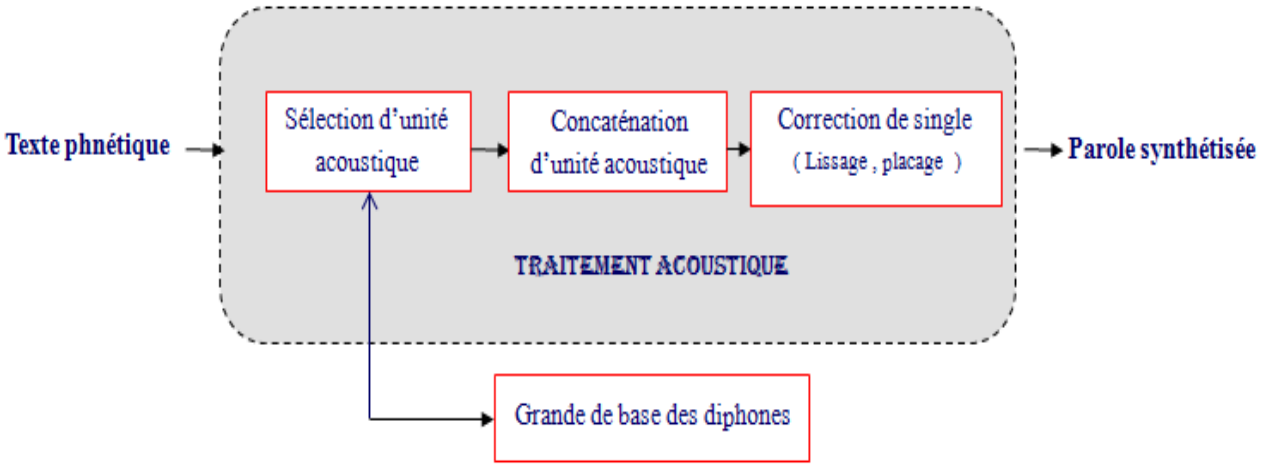

Figure 3.5: Schéma général de traitement acoustique.

3.2.3.1 Préparation de la base de diphones

Dans cette phase le système utilise une base des diphones déjà préparée avec le texte transcrit afin de générer une parole synthétisée.
La base de données doit être suffisamment grande pour contenir toutes sortes de séquences phonétiques qui peuvent apparaitre dans différents contextes linguistiques, de sorte que les segments de parole qui en résultent soient disponibles dans différent formes prosodiques

Pour préparer la base des diphones il faut premièrement enregistrer un ensemble de logatomes de base qui vont être utilisé par la suite dans une étape de segmentation permettant d'extraire l'ensemble de diphones.

A. Enregistrement des logatomes

Afin de construire une base des logatomes, nous avons utilisés un ensemble de mots dépourvu de sens contenant chacun un diphone dans un contexte phonétique neutre pour minimiser au maximum l'effet de coarticulation. L'ensemble des logatomes a été enregistré par un locuteur ayant une bonne élocution de la langue arabe, sans accent dialectal marqué.

B. Segmentation

La segmentation consiste à extraire l'ensemble de diphones de base à partir des différents logatomes enregistrés. Cette opération a été effectuée manuellement en utilisant l'application Java Sound Analyser V1.0. L'application permet de donner une représentation du signal acoustique et de faire des opérations telles que zoom, coupage et/ou sauvegarde d'une partie sélectionnée.

Pour extraire un diphone par exemple /La/ du logatome /BaLaBa/ , il faut le délimiter par deux marqueurs en utilisant les boutons Add Mark (Delete Mark) et Draw. Par la suite, il faut écrire le non du diphone dans le champ du texte prés du bouton Save part. Finalement, il faut cliquer sur Save Part pour enregistrer la partie sélectionné (voir figure 3.6).

Figure 3.6: Un exemple de Traitement pour l'obtention du diphone « La ».

3.2.3.2 Sélection des unités acoustiques

Cette étape consiste à faire un balayage du texte transcrit afin d'associer à chaque représention d'un caractère une unité acoustique adéquate à partir de la base déja créée. Cette opération est réalisée par un algorithme de sélection implémenté en Java appelé <Sélection>, (voir la figure 3.7).

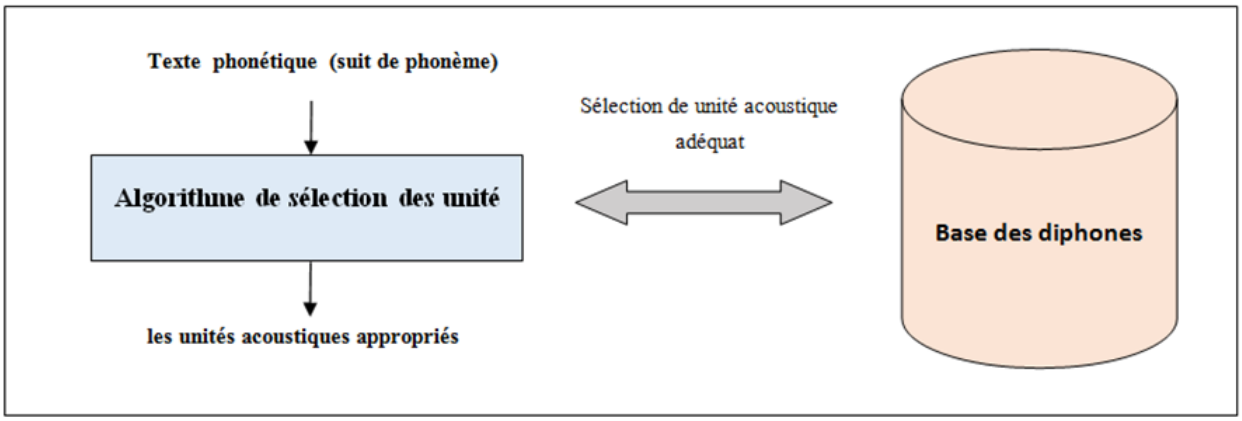

Figure 3.7: Schéma de la sélection des unités acoustiques adéquates à un texte phonétique.

3.2.3.3 Concaténation des unités acoustiques

Les différentes unités acoustiques sélectionées dans l'étape précédente doivent être concaténées pour construire la prononciation du texte. Cette concaténation a pour objectif d'éliminer les discontinuités en pronociation produite dans le cas ou elles sont prononcées de manière séparée. Un algorithme appelé <concaténation> permet de réaliser cette opération (voir la figure 3.8).

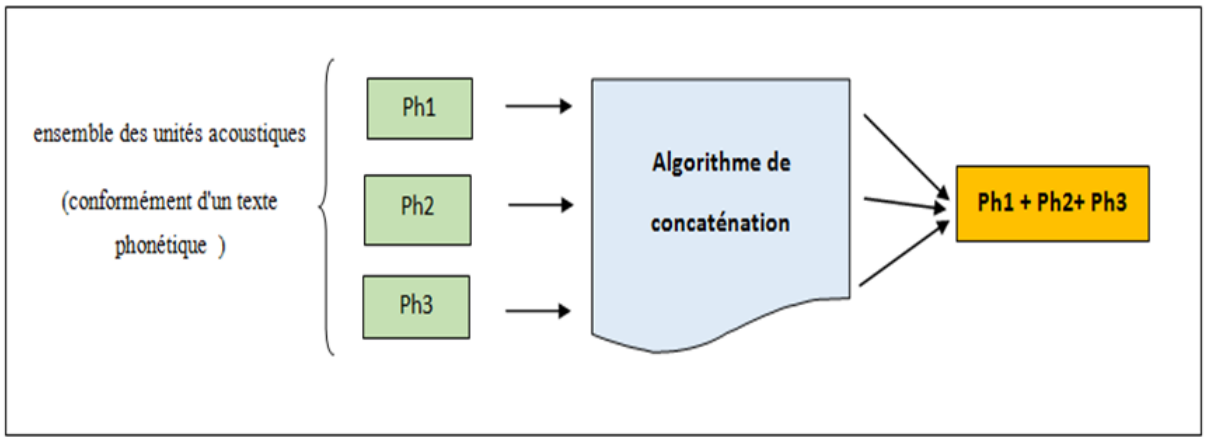

Figure 3.8: Illustration d'un exemple de concaténation.

La figure 3.9 présente un exemple dont nous appliqons les différents phases/étapes de la génération de parole synthétisée à partir d'un texte voyellée et la mise en œuvre de l'algorithme de sélection et l'algorithme de concaténation des unité acoustiques sélectionnées à partir de la base des diphones.

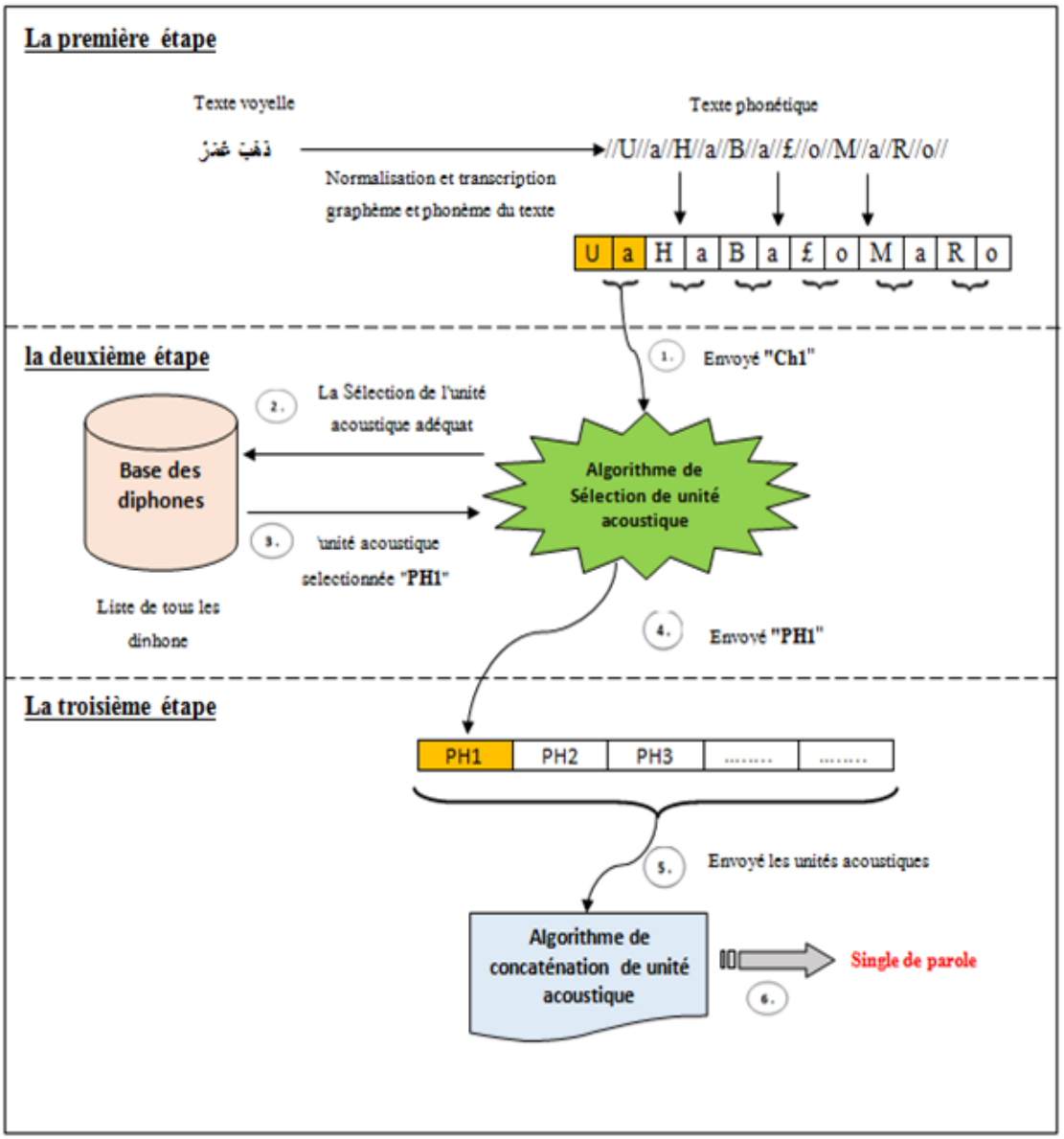

Figure 3.9: Shéma détaillé des différentes étapes de synthèse vocal de la phrese **ذهب عمر**.

3.2.3.4 Amélioration de la qualité du signale

La simple concaténation d'unités de parole extraites de contextes différents ne produit en général pas une parole de bonne qualité et souvent contient des discontinuités entre les différentes extrémités des unités acoustiques. Ces discontinuités sont le résultat des atténuations aux niveaux des extrémités des unités. Pour cela, le système commence par l'application d'un traitement temporel sur le signal produit durant les phases précédentes afin d'éliminer ce problème. Ce traitement touchera évidement la fin de la première unité et le début de la suivante (voir la section 3.1).

Conclusion

Nous avons présenté dans ce chapitre les différentes étapes qui peuvent conduire à une conception convenable d'un système de la synthèse de la parole Arabe à partir d'un texte donné en utilisant une technique basée sur les diphones. Et qui vise à résoudre des problèmes signalés dans l'étude du traitement langue arabe. En plus, nous avons proposée un algorithme de lissage à intervalle de pause variable du signale synthétisé dans le but de le rendre le plus réaliste possible que celui basé sur une technique de simple concaténation des unités acoustiques.

Chapitre 4

Résultats et bilan

Introduction

Dans le chapitre précédent nous avons présenté une conception du système en donnant une vue globale du système, en suite nous avons détaillé chaque module composant le système séparément. Dans ce chapitre nous allons voir la réalisation du système : le choix du langage , et description de environnement matériels et logiciels , l'implémentation des différents modules, quelques tests et enfin quelques résultats concernant le taux de synthèse

4.1 Choix des outils de développement

Dans ce travail, nous avons choisi comme environnement de programmation le langage JAVA qui possède une richesse et offre une grande simplicité de manipulation de son et d'images, soit en acquisition ou en génération des fichiers images. Ce langage possède des avantages très intéressants tel que :

- La portabilité des logiciels .
- La réutilisation de certaines classes déjà développées .
- La quasi-totalité de contrôle de windows (boutons, boites de saisies, listes déroulantes, menus ....etc.) qui sont représentés par classes.
- La possibilité d'ajouter à l'environnement de base des composants fournis par L'environnement soit même.

Durant la réalisation de notre travail nous avons utilisé les outils de développement Java suivants :

1. **Eclipse :**il présente une nouvelle version d'Eclipse, est un environnement de développement libre permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation (Java, C++, PHP). Eclipse Juno (4.2) est principalement écrit en Java .
2. **JDK1.8.0_25 :**Est un pack d'outils pour le développement d'application via le langage Java. Il a les composants nécessaires à la conception et au test de projets avec diverses caractéristiques.
3. **Sound Analyser 1.0 :**est une L'application permet de donner une représentation du signal acoustique et de faire des opérations telles que zoom, coupage et/ou

sauvegarde d'une partie sélectionnée.

4.2 Interfaces du système

En lançant l'application nous allons voir premièrement une image d'entrée (splash window) présentée par la figure 4.1 ci-dessous.

Figure 4.1: Interface de démarrage de notre système.

Ensuite, la fenêtre principale qui comporte les boutons principaux de l'application apparaîtra comme nous montre la figure 4.2 ci-dessous.

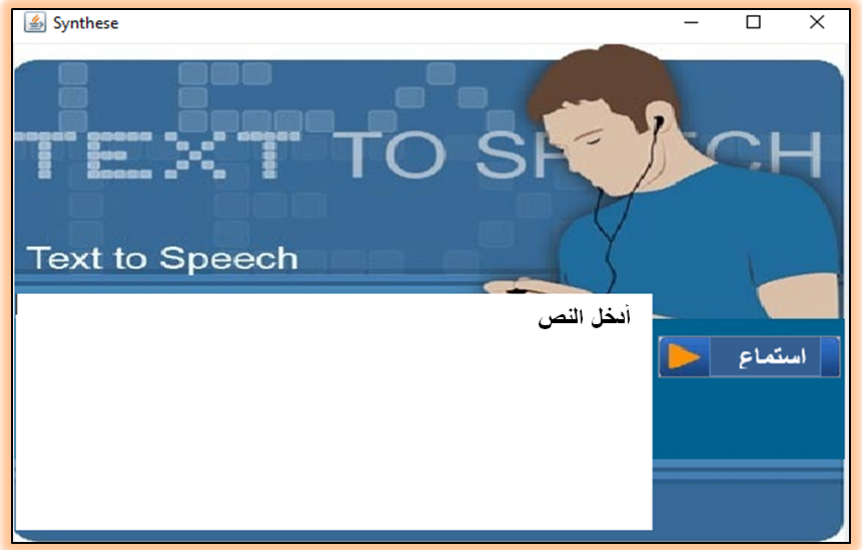

Figure 4.2: Fenêtre principale de l'application.

Notre application permet à l'utilisateur d'entrer un texte dans la zone de texte ou mémo à gauche du label **(أدخل النص)**, comme le montre la figure 4.3.

Figure 4.3: Exemple de saisie de texte.

4.3 Test et résultats

L'évaluation du système a été faite à l'aide d'une méthode de test basée sur l'écoute et l'identification des mots des phrases synthétisées. Pendant les tests nous avons utilisé 12 phrases, soit 53 mots, 211 unités acoustiques dont 73 différentes ce qui constitue 37.2 % de la totalité des unités acoustiques qu'utilise notre système. Nous les avons fait écouter à 4 personnes (2 femmes et 2 hommes) ce qui a permis une évaluation statistique réaliste des résultats. Chaque personne écoute trois fois toutes les phrases. Ensuite, pour chaque passage le sujet doit orthographier ce qu'il entend. La figure 4.5 montre les résultats obtenus à chaque essais.

T de R / N de PH	1er Essai	2éme Essai	3éme Essai
01	40%	60%	60 %
02	80%	92 %	85%
03	65%	80%	79%
04	65%	70%	75%
05	80%	75%	78%
06	92%	82%	85%
07	85%	80%	83%
08	75%	78%	80%
09	80%	92%	82%
10	80%	80%	80%
11	81%	79%	79%
12	83%	81%	78%

Tableau 4.1: Illustration des résultats obtenus à chaque essais .

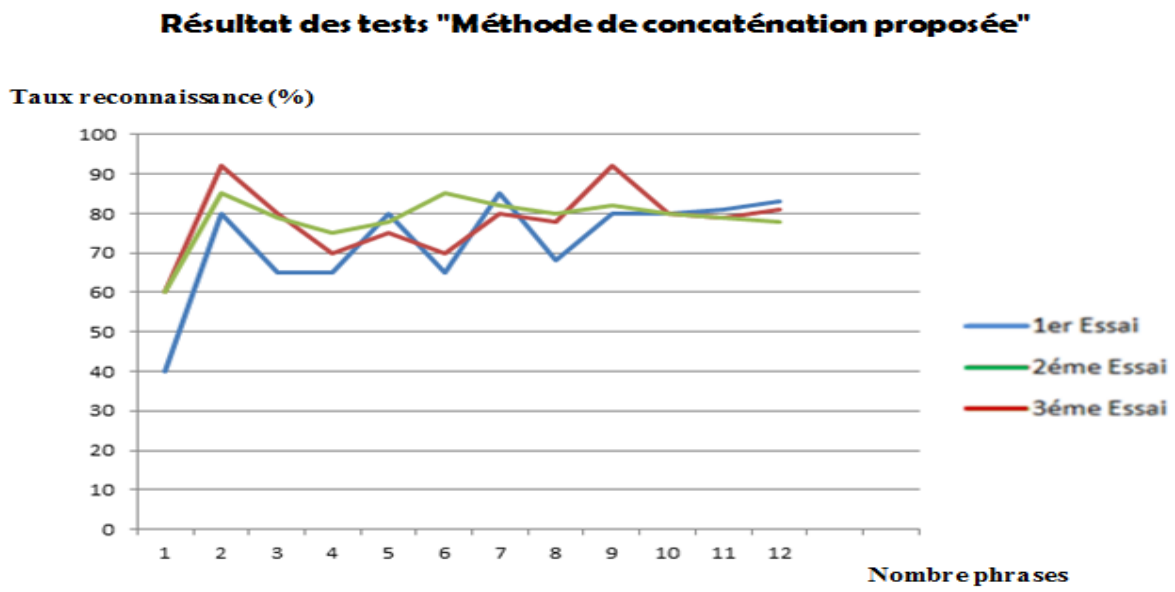

Figure 4.4: Les résultats de la phase de test " méthode de concaténation proposée "

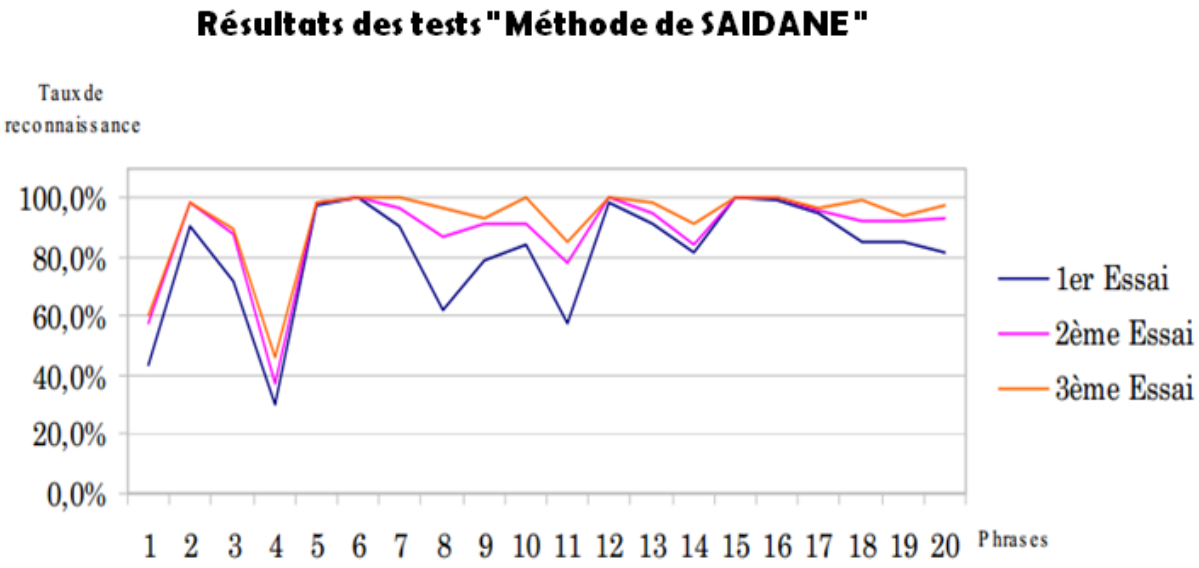

Figure 4.5: Les résultats de la phase de test " méthode de SAIDANE "

la comparison :

Après avoir vu les résultats de chaque méthode, nous notons que la proportion de la reconnaissance vocale dans Méthode proposée de stable et régulière dans chaque les phrases dans trois essais, par contre avec la méthode de " SAIDANE ", nous notons que la proportion de la reconnaissance vocale est instable et volatile, Car il y a quelques phrases et les mots ne sont pas identifiés les bien dans les trois essais telles que la quatrième phrase et huitième .... etc, Et à cause de l'algorithme utilisé dans le lissage pour améliorer la qualité du son ou de la parole générée.

conclusion

Nous avons présenté dans ce chapitre les outils de développement utilisés : l'environnement de développement Eclipse et java SE Développement Kit (JDK)), en plus, nous avons expliqué les différentes fenêtres de notre système. puis nous avons présenté les résultats obtenus par notre système avec une comparaison avec celles obtenus par Saidane et al.[12]

Mais ces résultats peuvent être améliorés par ce qui suit :

- L'ajout d'un nouveau algorithme de lissage tel que cel de Savitzky et Golay au lieu de celui utilisé.
- Le contrôle des caractéristiques des différentes unités acoustiques tel que le volume, fréquence,...etc. pour les rendre les plus proches possible.

Conclusion générale

L'objectif de notre mémoire est de concevoir et mettre en œuvre une applications de synthèse de la parole à partir du texte arabe standard voyellé , telle que nous avons commencé par l'étude Quelque concept de base du sujet en général ,en suite dans La Deuxième chapitre nous présentons Comment traiter la langue arabe.

Puis nous sommes allés ensuite, à troisième étape la phase de conception et de modélisation du système réalisé, en suite dans le dernier chapitre nous présentons la réalisation du système et quelques tests et enfin quelques résultats concernant le taux de synthèse .

Mais Malgré les efforts et le travail intensif que nous avons obtenus dans la présente La programme, Il n'y a pas encore de système TTS fiable 100 , Dans ce travail été intéressé par l'offre un modèle de synthétiseur de la voie Arabe à base des diphones, En noyant d'améliorer la voyellation des mots ambigus .

Nous avons été confrontés de Plusieurs ambigüité et problèmes durant ont notre étude, parmi les quelles nous citons :

- Les conditions d'enregistrement ne répondent pas aux contraintes d'applications (bruit, position et sensibilité du microphone. . . etc.).
- Les états variés des locuteurs (le tempérament du locuteur, état émotif, état de fatigue. . . etc.).
- Une grande base de données de diphones.
- La langue Arabe contient beaucoup de règles (morphologique, syntaxique, sémantique et linguistique).

Nous espérons que nous avons réussi la création de ce projet et que nous avons atteint les objectifs que nous avons tracés à ce projet

Bibliographie

[1] L. Amiar – *« un systeme hybride ag/pmc pour la reconnaissance de la parole arabe »*, Mémoire, Université Badji Mokhtar Annaba, 2005.

[2] A. Amrouche1 – « Contribution à la réalisation d'un synthétiseur de la parole pour la langue arabe », Tech. report, Université de Algérie., 2000.

[3] S. Baloul – « Développement d'un système automatique de synthèse de la parole à partir de texte arabe standard voyelle », Thèse, Université du Maine France, 2003.

[4] R. Boite et M. Kunt – « Traitement numérique du signal théorie et pratique », Tech. report, Université de Paris, 1998-2002.

[5] F. S. Douzidia – *« résumé automatique de texte arabe »*, Mémoire, Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique, Université de Montréal, 2004.

[6] T. Dutoit – « Introduction au traitement automatique de la parole », Tech. report, Faculté Polytechnique de Mons, 2000.

[7] T. Dutoit, L. Couvreur, F. Malfrère, V. Pagel et C. Ris – « Synthèse vocale et reconnaissance de la parole : Droites gauches et mondes parallèles », Tech. report, Faculté Polytechnique de Mons, 2005.

[8] J. L. Grand – *« amélioration des systèmes de reconnaissance de la parole des personnes âgées »*, Mémoire, LaboratoireLIG, Equipe : GETALP BP 53, 2011/2012.

[9] F. Imedjdouben1 et A. Houacine – « Outil de transcription phonétique à partir du texte arabe », Tech. report, Université des Sciences et de la Technologie Houari Boumediene Alger, Faculté d'Electronique et d'Informatique., 2001.

[10] B. Kamal – *« modèle de markov cachés : Application à la reconnaissance automatique de la parole »*, Mémoire, Université de Lager, 2014.

[11] C. Martin – *« la reconnaissance de la prosodie émotionnelle après un traumatisme crânien »*, Mémoire, Université d'Angers, 2012/2013.

[12] Z. Mounir – « Traitement automatique de la langue arabe », Tech. report, faculté des science de Monastir ,Tunisie, 2007.

[13] A. Oasim et M. A. Jayousi – *« araic.text-to-speech synthesizer »*, Mémoire, Faculy of computer science and information echnology univerity of malaya kualumpur, 2007.

[14] B. René et K. Murat – *«conception et réalisation d'un système de reconnaissance de locuteur par réseau de neurones artificiels»*, Mémoire, Université de Biskra, 2005.

[15] B. Rodolphe – *« la reconnaissance vocale, techniques utilisées, applications actuelles et futures »*, Mémoire, Université De Paris, 1998.

[16] M. Safa, N. Khellat et K. Badra – *« la reconnaissance automatique de la maladie de parkinson »*, Mémoire, Université Des Sciences Et De La Technologie D'Oran, 2012.

[17] T. Saidane, M. Zrigui et M. B. Ahmed – « La transcription orthographique phonétique de la langue », Tech. report, Faculté des Sciences de Monastir, Tunisie, 2004.

[18] T. Somaia, K. WafaaEl, T. Hesham et M. Eman – « The effect of using integrated signal processing hearing aids on the speech recognition abilities of hearing impaired arabic-speaking children », *Arab Research Institute in Sciences & Engineering* **3** (2014), p. 215–224.

[19] S. Wilhelm – « Prosodie et correction phonétique , », Tech. report, Université de Dijon, 2000.