UNIVERSITE SAÂD DAHLAB DE BLIDA

Faculté des Sciences de l'Ingénieur
Département d'Informatique

MEMOIRE DE MAGISTER

Spécialité : Ingénierie des systèmes et des connaissances

TRANSCRIPTION ORTHOGRAPHIQUE PHONETIQUE EN VUE DE LA SYNTHESE DE LA PAROLE Â PARTIR DU TEXTE DE L'ARABE STANDARD

Par

TEBBI Hanane

Devant le jury composé de :

A. GUESSOUM	Professeur, U. de Blida	Président
S. OUKID	Maître de Conférence, U. Blida	Examinatrice
N. BENBLIDIA	Chargé de Cours, U. Blida	Examinatrice
M. GUERTI	Maître de Conférence, ENP. Alger	Rapporteur

Blida, Juin 2007

ملخص

في أيامنا هذه و مع ظهور الانترنت، و مع كثرت استعمال النصوص في الويب، زد إلى ذلك تنوع لغات هذه النصوص, أدى هذا إلى حتمية التوسع في دراسات التقنيات الاصطناعية للغة وهذا بهدف صنع آلات قادرة على الحوار بسهولة مع الإنسان.

للحصول على هذه الآلات السهلة الاستعمال, لا بد لنا من المرور بمرحلة هامة تقوم بتحويل النص المكتوب إلى نص آخر يمثل سلسلة من الوحدات الصوتية التي توافق النطق المناسب لهذا النص، إن الدراسة التحويلية التي قمنا بها تتمحور حول إنجاز برنامج يحقق لنا عملية تحويل سلسلة من الحروف العربية إلى مجموعة من الأصوات، ولهذا الغرض استعملنا 12 قاعدة في مرحلة التحويل بواسطة القواعد، و 6 شواذ في مرحلة التحويل عن طريق الكلمات

من خلال هذا العمل نريد إنجاز جهاز نطق اصطناعي لنص مكتوب بالعربية الفصحى و هذا بهدف تسهيل قراءة النصوص للأشخاص المعاقين بصريا و المكفوفين بالإضافة إلى استعمال العربية في بعض الآلات المستعملة يوميا, بدون إن ننسى استعمال العربية في الويب.

النتائج التي توصلنا إليها تم الحصول عليها عن طر يق مجموعة جمل مختارة عشوائيا و بعض السور القرآنية, النسبة التي توصلنا إليها مرضية لان جهازنا يستطيع تحويل أية جملة عربية و بطريقة صحيحة

كلمات المفاتيح : الكلام الاصطناعي، اللغة العربية الفصحى، إنتاجية الكلام من خلال النص، تحويل نص مكتوب إلى سلسلة من الوحدات الصوتية

RESUME

De nos jours, avec l'apparition de l'Internet et grâce à l'augmentation des textes écrits en différentes langues disponibles sur le Web, le Traitement Automatique du Langage devient un champ susceptible de modifier en profondeur les technologies de l'information et de la communication. Cela donne naissance à des recherches très fines en informatique afin de construire des machines qui réagissent efficacement face au langage. Ces machines sont précisément, capables de dialoguer oralement de manière interactive avec les utilisateurs.

Pour concevoir ces machines interactives, il faut assurer un passage clé entre le monde de l'écrit et celui de l'oral. A travers cette étude de transcription des formes isolées, et des séquences de mots, nous avons élaboré un logiciel réalisant la transformation de tous les graphèmes (caractères) qui sont écrits en Arabe Standard (AS) en phonèmes (sons). Pour arriver à ce but nous avons utilisé 12 règles lors de la transcription par règles (tanwin, §adda, sukuun, ɛl madd, etc.), et 6 mots d'exceptions lors de la transcription par lexique.

A partir de cette étude nous visons à concevoir un système de synthèse à partir du texte arabe ou un système Text-To-Speech (TTS), parmi les nombreuses applications existantes dans ce domaine, notre système peut faciliter la tâche de lecture aux mal voyants et aux aveugles. Il assure aussi l'intégration de l'Arabe sur le Web et sur des applications embarquées utilisables dans notre vie quotidienne.

Nos résultats ont été évalués sur des phrases aléatoires et quelques versets coraniques / سور من ألقران / [suuarun mina alqurEani]. Le taux obtenus est satisfaisant puisque notre outil permet de transcrire n'importe quelle phrase Arabe et de manière correcte.

Mots clés : Arabe Standard, Systèmes de synthèse à partir du texte, Transcription Orthographique Phonétique, Règles de transcription Graphème Phonème, lecture pour aveugles

ABSTRACT

Nowadays with the appearance of the Internet and thanks to the increase of the texts available on the Web, with the variety of the languages of these texts; the Automatic Treatment of Language becomes a field likely to modify in-depth information technologies and communication. That gives rise to very fine research in data processing in order to build machines which react effectively and more precisely of the machines which are able to dialogue orally in an interactive way with the users.

To design these interactive machines, it is necessary to ensure a key passage between the world of the writing and that of the oral examination. With through this study of transcription of the isolated forms, and the sequences of words; we tried to conceive software carrying out the transformation of a whole of graphemes (characters) that are written in Standard Arabic (SA) in a whole of phonemes (sounds). To arrive at this goal we employed 12 rules at the time of the transcription by rules of any Arab text and 6 words of exceptions at the time of the transcription by lexicon.

From this study us minks to conceive a system of synthesis starting from the Arab text or a system Text-To- Speech (TTS), This system facilitated the task of reading to the evil indicators and the blind men, and it ensures also the integration of Arabic on the Web and embarked applications usable in our everyday life.

Key words: the speech synthesis, Arab Standard, Systems of synthesis starting from the text, Phonetic Orthographical Transcription.

Dédicaces

Ce modeste travail est dédié :

tout d'abord à la mémoire de mon père qui malgré son absence il est toujours présent dans mon cœur.

à ma mère, qui dans toute sa simplicité, savait apporter le plus tant recherché, par celui ou celle, qui désespérait de voir la lumière annonciatrice d'heureux présages. Je profite de cette occasion pour la remercier pour son aide puisque sans elle je ne serai jamais ce que je suis aujourd'hui.

à mes chers frères et toutes mes sœurs et mes nièces surtout ma chère sœur Dallal.

Je tiens à remercier chaleureusement M. Hamadouche pour son encouragement et son soutien moral.

REMERCIEMENTS

Tout d'abord je remercie Dieu pour sa faveur

Toute ma gratitude et mes vifs remerciements vont à M^{me} M. GUERTI Maître de Conférences au département d'Electronique à l'Ecole Nationale Polytechnique ENP d'Alger, ma Directrice de mémoire, pour sa patience, sa disponibilité, et ses précieux conseils, et qui a su me faire profiter de sa grande expérience.

Mes sincères remerciements à :

Mr A. GUESSOUM, Professeur au Département d'Electronique, à l'Université de Blida, d'avoir fait l'honneur de présider mon jury de mémoire ;
M^{me} S. OUKID, Maître de conférences à l'Université de Blida, et directrice du Laboratoire de Recherche et de Développement des Systèmes Informatisés LRDSI, et M^{me} N. BENBLIDIA Chargée de cours au Département d'informatique à l'Université de Blida, qui m'ont fait l'honneur d'être membres du jury d'évaluation de ce travail.

Je remercie aussi tous mes Enseignants du département d'Informatique de l'Université Saâd Dahlab de Blida.

J'adresse mes sincères remerciements à mes amis qui ont partagé avec moi les moments de doutes et d'espoir : M. Hamadouche, R. Mazari, S. Hassaine, F. Mazari Boufaresse, et M. Hammouda, D .Touahri et tous les étudiants de Magister en Informatique de l'Université de Blida.

Enfin je remercie, tous ceux qui ont contribué de près où de loin à la réalisation de ce travail.

LISTE DES FIGURES

LISTE DES TABLEAUX

.

TABLE DES MATIERES

INTRODUCTION

La communication par la voix est un des enjeux majeurs du dialogue Homme-Machine, puisque la voix véhicule à la fois un contenu linguistique explicite que l'on peut représenter sous forme écrite et un contenu non-linguistique comme le type du locuteur, son attitude, ses gestes, etc. Cela rend le Traitement Automatique de la Parole (TAP) une composante fondamentale des sciences de l'ingénieur, et un champ de recherche riche et complexe.

Le dialogue oral Homme-Machine est un sujet de recherche à multiples facettes qui nous amène à traiter l'oral, à modéliser des processus de compréhension et de reconnaissance de la parole et à étudier le processus de synthèse de la parole.

Ces dernières années, plusieurs recherches ont été développées afin d'acquérir des programmes performants de synthèse vocale. Le but est de fournir des outils pour la génération de la parole, tout en respectant une analyse très précise; et tout en assurant une synthèse de très haute qualité. Nous pouvons bien constater qu'avant d'améliorer la qualité de n'importe quel synthétiseur vocal il faut d'abord le réaliser.

Les systèmes de synthèse de la parole à partir du texte (ou Text-To-Speech : TTS) représentent une des catégories de la grande classe des systèmes de synthèse vocale. Ils sont réalisés à l'aide d'une architecture séquentielle et modulaire, classiquement divisée en un ensemble de blocs :

- un pour la construction de la base sonore, ce module représente la clé de succès pour assurer une bonne qualité de synthèse. Il permet d'identifier le processus de fabrication de notre source sonore en construisant des unités acoustiques nécessaires pour réaliser une sortie vocale associée au texte à lire ;

- un pour la conversion graphème phonème qui convertit un texte en parole. Pour assurer cette tâche des traitements grammaticaux complexes doivent être pris en considération afin d'élaborer l'ensemble des règles de transcription nécessaires à la transformation d'un texte écrit en un texte oral (TOP) ;

- et un autre pour la génération acoustique (le synthétiseur), ce dernier est chargé de générer la parole réelle qui correspond à la chaîne phonétique résultante de l'opération de transcription ;

L'objectif principal de ces systèmes TTS est de doter l'ordinateur de la capacité à lire des textes à haute voix. Malgré les avancées réalisées ces dernières années dans ces domaines, des progrès restent à faire pour accroître le confort d'utilisation des systèmes actuels [1].

L e but de ce mémoire est de mettre en oeuvre et d'évaluer un système de Transcription Orthographique Phonétique en vue de la synthèse de parole à partir d'un texte en Arabe Standard « AS ». L'AS a été moins étudiée au point de vue informatique que les autres langues, mais avec sa diffusion sur le Web son traitement devient une nécessité.
Sur le plan formel, ce mémoire de Magister est composé en 4 chapitres :

- dans le premier, nous présentons une vue générale sur le traitement de la parole où nous définissons les notions les plus utilisées, et sur quelques fondements de l'Arabe Standard ;

- le deuxième, met en évidence les techniques de synthèse de la parole ainsi que leurs variantes ;

- dans le troisième, nous abordons la phase de Transcription Orthographique Phonétique qui représente la base de n'importe quel système de génération du signal vocal. Nous exposons quelques caractéristiques de la langue Arabe Standard, ainsi que ses problématiques ;

- le dernier, concerne la simulation et l’interprétation des résultats obtenus dans le cadre de notre application ;

Enfin, des conclusions générales et des perspectives sont données pour ouvrir la voie à des travaux futurs.

CHAPITRE 1 :
GENERALITES SUR LE SIGNAL VOCAL

1.1. Introduction

La parole est le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés.

Dans ce chapitre nous allons décrire de manière générale le signal de parole et ses caractéristiques, l'appareil phonatoire qui représente l'organe principal de la génération vocale, et les différents traitements appliqués sur la parole essentiellement sur l'Arabe Standard.

1.2. Le son et la parole

Qu'est ce qu'un son ?

Par définition, le son est ce que l'oreille perçoit de la vibration d'un corps. Cette vibration est une sorte d'onde (produite par un objet, guitare, piano, tambour, marteau, etc.), qui se propage par et à travers des corps physiques (air, eau, métal, bois, etc.) (Figure 1.1) [2].

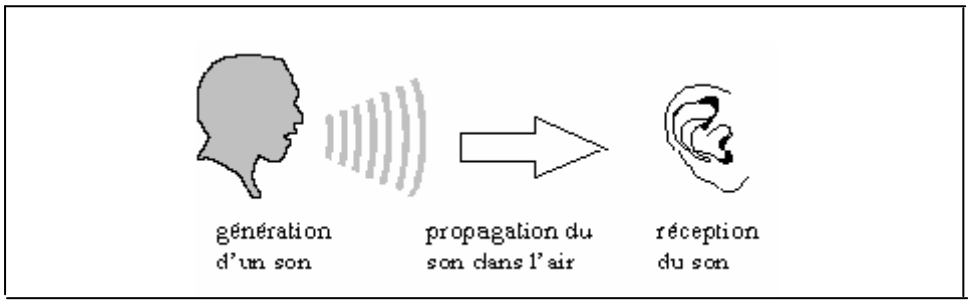


Figure 1.1 : Les trois étapes de la propagation du son [2].

La parole se distingue des autres sons par des caractéristiques acoustiques ayant leurs origines dans le mécanisme de production [3].

1.3. Le mécanisme de la production de la parole

La production de la parole réside dans les fluctuations de la pression de l'air engendrée, puis émise par l'appareil phonatoire (Figure 1.3), ces fluctuations constituent le signal vocal. Elles sont détectées par l'oreille qui procède à une certaine analyse et les résultats sont transmis au cerveau qui les interprète [4]. La génération de la voix n'est pas réalisée par un système propre, mais elle est assurée conjointement par les organes de l'appareil phonatoire et de la respiration [5]. Pour comprendre le fonctionnement de l'appareil phonatoire lors de l'émission des sons, nous devons tenir compte de ses différents composants et organes, ces derniers peuvent se résumer par :

- Les poumons qui fournissent l'énergie nécessaire pour la production des sons. Cette énergie est assurée par le biais d'un mouvement cyclique de la respiration.
  La respiration comprend deux phases : l'inspiration et l'expiration, cette dernière assure l'opération de phonation et cela grâce à un flux d'air provenant des poumons. Ce flux s'appelle *air pulmonaire (ou pulmonique) égressif ;*

- La trachée artère et son extrémité le larynx qui est un lieu important pour les mécanismes phonatoires. Il est situé dans la région moyenne du cou et il est constitué de cartilages, muscles, muqueuses et de nerfs. Il contient *les cordes vocales* qui sont un ensemble de muqueuses, de ligaments et de muscles [6]. Le principal rôle du larynx est de moduler la pression de l'air généré par les poumons avant d'être appliqué au conduit vocal ;

- Le conduit vocal est une succession de quatre cavités :
  - pharyngale, buccale, et labiale, appelées cavité pharyngo-buccale
  - et nasale.

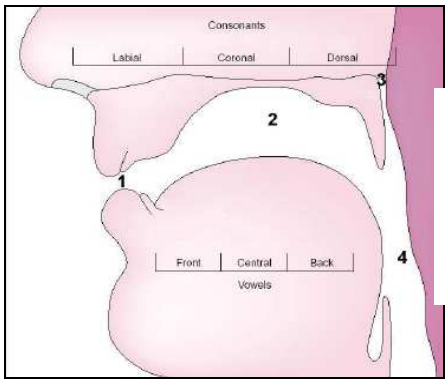


Figure 1.2 : Les principaux lieux d'articulation et résonateurs de l'appareil phonatoire humain [7]

Pour obtenir le modèle électrique de ce mécanisme de production, il suffit de remplacer chaque cavité par son circuit équivalant *« un filtre » ;*

- La langue, joue un rôle très important dans la phonation, car sa mobilité d'avant en arrière et de haut en bas et vice versa, lui permet de générer une infinité de positions (ou lieux), dont chacune correspond à l'articulation d'un son ;

- Les dents supérieures et inférieures, représentent le point d'appui contre lequel les lèvres ou la langue prennent contact pour réaliser une occlusion parfaite. La mâchoire entoure la bouche ;

- Les lèvres représentent un organe très mobile qui est utilisé pour assurer les articulations labiales. Elles sont situées à l'extrémité du conduit vocal et c'est leur écartement (et les variations de cet écartement) qui est important du point de vue acoustique [3];

- La cavité nasale contient deux cavités fixes appelées les fosses nasales et elle se termine par le nez.

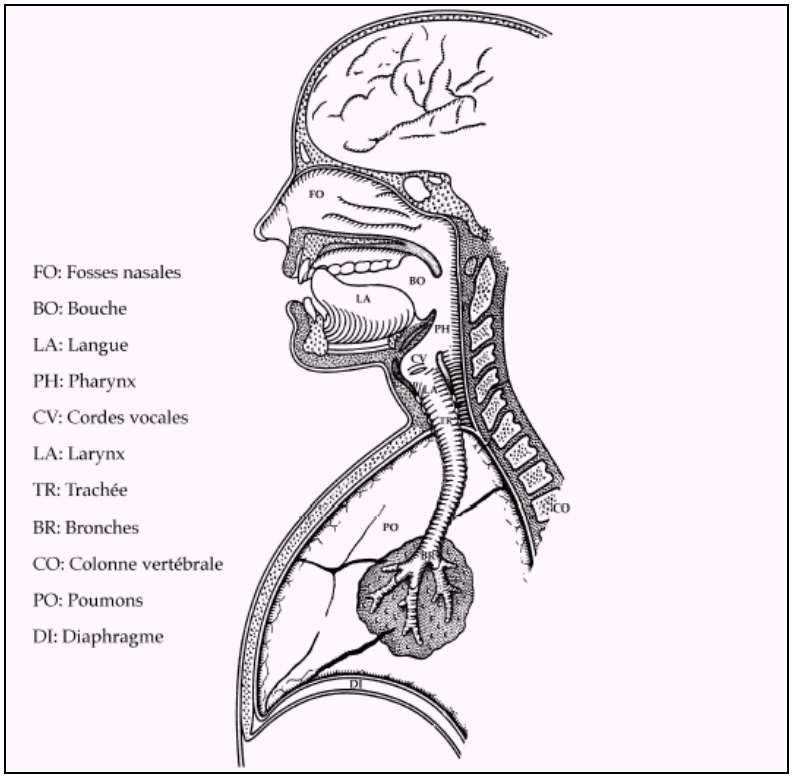


Figure 1.3 : Appareil phonatoire humain [5].

1.4. Modélisation de phénomène de production de la parole

Pour comprendre le principe utilisé en production de parole, il faut donner une représentation abstraite (un modèle généralement linéaire) de l'appareil phonatoire, et puisque ce dernier est représenté par un système résonnant qui est composé des cavités (pharyngale, buccale, labiale, et nasale), sa modélisation correspond à la combinaison des modèles associés aux ces quatre cavités.

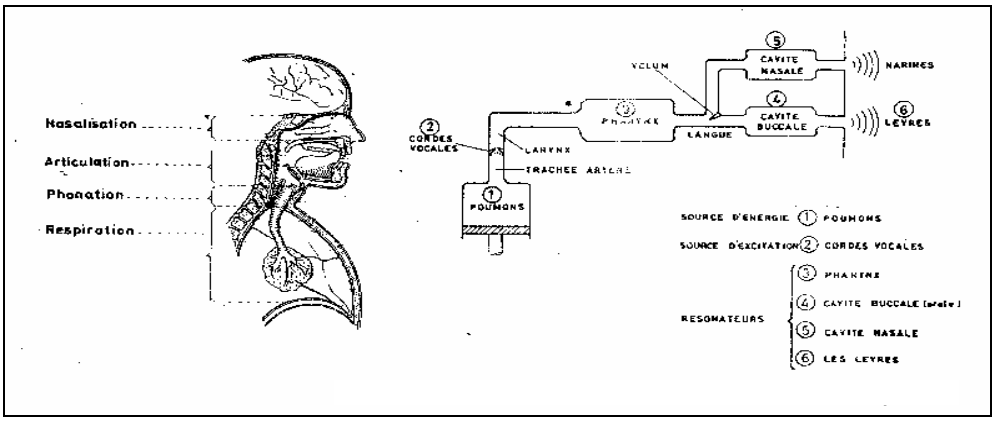


Figure 1.4 : Production et modélisation de la parole [8].

1.4.1. Modèle de la source glottique

La source d'excitation peut générer un son voisé ou un bruit; Dans le cas d'un voisement cette source peut être simulé par un train d'impulsions quasi périodiques traversant un filtre passe bas [9]; ce qui nous donne la forme de la source glottique. Ce modèle est donné par l'équation [10] :

$$G(z) = \frac{1}{(1 - e^{-ct} Z^{-1})^2} \qquad (1.1)$$

Où $ct \approx 0$ donc $e^{-ct} \approx 1$

1.4.2. Modèle du conduit vocal

Le conduit vocal peut être représenté par un tube sonore; sa fonction de transfert est donnée par la formule suivante [10] :

$$V(z) = \frac{1}{\prod_{i=1}^{K} (1 - 2 e^{-c_i T} \cos(b_i^T) Z^{-1} + e^{-2 c_i T} Z^{-1})} \qquad (1.2)$$

T : la période d'échantillonnage.

K : représente le nombre de résonances définies par le modèle.

Chaque résonance est caractérisée par sa fréquence F_i et sa bande passante B_i.

$b_i = 2\pi F_i$, $c_i = 2\pi$

1.4.3. Nasalisation

Pour certaines consonnes, le velum est abaissé et l'air passe en empruntant la Cavité Nasale (CN), pour produire un son nasal, c'est le cas de [m] et [n], par exemple. La fonction de transfert pour cette catégorie de sons est la suivante [10] :

$$V(z) = \frac{\prod_{i=1}^{i=M}(1-\frac{Z_i}{Z})(1-\frac{\overline{Z}_i}{Z})}{\prod_{i=1}^{i=N}(1-\frac{Z_i}{Z})(1-\frac{\overline{Z}_i}{Z})} \qquad (1.3)$$

L'utilisation pratique de cette fonction dans les algorithmes s'avère difficile, c'est pourquoi on augmente l'ordre N de la fonction de transfert pour approcher un zéro au numérateur [9].

1.4.4. Rayonnement aux lèvres

La radiation des lèvres transforme l'onde de vitesse volumique en une onde de pression [10]. Cette transformation est modélisée par un filtre de dérivation de la forme [10] :

$$R(z) = 1 - \frac{k}{Z} \qquad \text{avec } k \approx 1 \qquad (1.4)$$

1.5. Quelques caractéristiques de la parole

Le processus de production de la parole présente certaines caractéristiques qui sont liées au signal vocal lui-même (continuité, variabilité, conduit vocal, et encodage).

1.5.1. La Continuité

Chaque phrase est un assemblage d'une suite de mots où chaque mot est une succession d'un ensemble de lettres ou caractères. De plus, la parole est une suite continue de phonèmes malgré la séparation qui existe entre les sons des différents mots.

1.5.2. La Variabilité intra et interlocuteur

La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que ce soit pour un même ou plusieurs locuteurs [11]. On distingue trois sortes de variabilités lors de la production de la parole :

- La variabilité intra locuteur, un locuteur ne prononcera jamais deux fois de manière identique un même mot. La vitesse d'élocution, la puissance et le timbre sont extrêmement variables [12]. Cette variabilité concerne un seul individu, et elle explique sa façon de parler puisque cet individu peut prononcer la même phrase à voix faible ou forte, rapide ou lente, et chuchotée ou non ;

- La variabilité interlocuteur : les différences morphologiques et culturelles font que les paramètres vocaux sont spécifiques à chaque locuteur [12]. Cette variabilité concerne un ensemble d'individus où chacun d'eux a ses propres caractéristiques, en prononçant la même phrase, avec le même rythme, le même accent, ainsi que le même timbre.

  Cette variabilité est aussi due principalement à la différence de l'âge, du sexe, de la physiologie et de l'origine géographique de chaque individu ;

- La variabilité contextuelle, est liée au phénomène de la coarticulation des sons entre eux tels que deux sons voisins peuvent s'influencer mutuellement. Cette variabilité est appelée aussi *la variabilité due à l'environnement* et cela puisque l'environnement peut diminuer le signal vocal généré sans que le locuteur ne modifie son mode d'élocution.

1.5.3. Le conduit

Le conduit vocal est un tuyau tridimensionnel qui est excité par une ou deux sources acoustiques. La source laryngienne peut être considérée comme quasi périodique, avec une fréquence pouvant évoluer très rapidement. La seconde source génère du bruit de friction ou d'explosion [2].

1.5.4. L'Encodage

L'encodage concerne les niveaux lexicaux, syntaxiques, sémantiques, morphologiques et phonétiques (phonèmes et leurs interactions) utilisés souvent pour assurer une meilleure qualité de la parole synthétique.

1.5.5. Les caractéristiques phonétiques de la parole

Le phonème est la plus petite unité présente dans la parole, de manière générale le nombre des phonèmes est toujours limité, normalement il est inférieur à une cinquantaine, par exemple : le Français a 36 phonèmes, l'Anglais 46 phonèmes, et l'Arabe Standard 40.

Ces phonèmes sont regroupés en classes et en sous-classes. Chacune d'elles est liée à un mode articulatoire de l'appareil phonatoire.

1.6. Les paramètres prosodiques d'un signal vocal

La prosodie est une science de la linguistique qui étudie les éléments phoniques (l'accent, l'intonation, etc.) de n'importe quelle langue, et puisque la parole est un signal réel d'énergie finie, continu (Figure 1.5), et non stationnaire ; les variations des paramètres prosodiques physiques (La fréquence fondamentale, la durée, et l'intensité) influencent de manière directe sur ces éléments phoniques

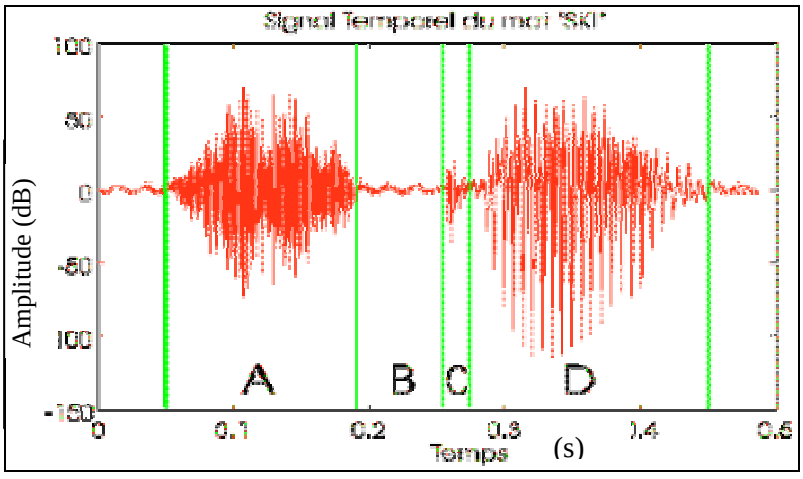


Figure 1.5 : Représentation temporelle du signal coustique de la parole [3].

Les recherches en linguistique ont montré que les caractéristiques prosodiques sont des composantes indispensables à la langue et à la fonction de communication. Puisqu'elles influencent directement sur l'intelligibilité de la parole synthétique. Il existe trois manières de définir les paramètres prosodiques, selon qu'on les considère sur les plans de la production, de l'acoustique, et le perceptif (Figure 1.6).

Production	**Acoustique**	**Perception**	
- Facteur masse-tension des cordes vocales.	- Fréquence fondamentale (ou F_0)	- Hauteur (pitch) Mélodie	Locale
- Force	- Intensité ou Energie	- Sonie	
- Débit	- La Durée	- Rythme	Globale

Figure 1.6 : Les différentes définitions des paramètres prosodiques.

De point de vue acoustique, on peut résumer les paramètres prosodiques d'un signal vocal par : la F_0, la durée et l'énergie.

1.6.1. La Fréquence Fondamentale F0 (pitch)

La Fréquence Fondamentale est la fréquence de vibrations des cordes vocales, elle varie d'une personne à une autre en fonction de la longueur et de la masse des cordes vocales de chaque personne. Elle permet de diviser l'ensemble des sons de parole en trois grandes macro classes :

- 70 -250 Hz pour les hommes ;
- 150 - 400 Hz pour les femmes ;
- 200 - 600 Hz pour les enfants.

Les variations de la fréquence au cours de la parole constituent ce qu'on appelle la *mélodie* ou *l'intonation*. Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la F_0 [11].
Chaque son voisé correspond à une présence de pitch; tandis que les sons non voisés correspondent à une fréquence fondamentale nulle (Figure 1.7).

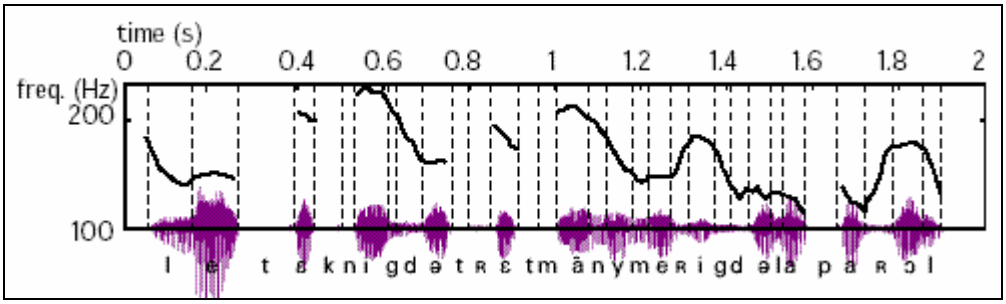


Figure 1.7 : Evolution de la F_0 de la phrase « les techniques de traitement numérique de la parole »; La F_0 est donnée sur une échelle logarithmique [13].

1.6.2. La durée

La durée est une mesure très variable. Elle représente le temps de la prononciation d'un phonème. Il existe deux types de durées :

- La durée observée, qui correspond à la mesure objective du temps de l'activation des organes de phonation ;
- La durée perçue, est liée au mécanisme de la perception et elle est fréquemment utilisée dans le cas des occlusives puisqu'elles sont caractérisées par une durée de réalisation non continue.

Généralement la durée d'une unité est mesurée par le nombre des trames qu'elle contient. Pour calculer la durée de chaque trame il faut fixer deux événements sur le signal de parole qui délimitent les repères initial et final de cette trame.

1.6.3. L'Intensité ou l'énergie

Elle est résultante de la pression sous glottique. Généralement elle exprime le volume sonore d'un phonème et dans le cas d'un voisement elle représente l'amplitude des vibrations des cordes vocales. Elle est exprimée pour un signal échantillonné x_n par :

$$E = \sum_{N=1}^{T} x_n^2 \qquad \text{tel que } n = 1,\ldots, T \tag{1.5}$$

A l'échelle perceptive elle est exprimée en déciBels (dB) par :

$$E_{dB} = 10\times \log_{10} \left(\sum_{N=1}^{T} x_n^2 \right) \tag{1.6}$$

1.7. L'Alphabet Phonétique International «API »

Pour représenter les différents sons, on se base sur l'utilisation d'un mécanisme de transcription**.** L'alphabet normal convient assez mal à ce dernier, puisqu'une seule lettre peut correspondre à plus d'un son (par exemple le [t] en Français est prononcé dans le mot /technologie/ et dans /chat/ ne sera pas prononcé) et puisqu'un seul son peut se représenter au moyen de plus d'une lettre (pensez au son [s] en Français) [14].

Transcrire phonétiquement un énoncé, c'est le noter à l'aide d'un alphabet conventionnel en général, on utilise l'API [15]. L'API (ou IPA en Anglais : International Phonetic Alphabet) est né dans le cadre de la didactique des langues étrangères : il a été créé par une association de professeurs de langues (Association Phonétique Internationale). Régulièrement révisé, sa dernière mouture date de 1993. L'objectif est clair : transcrire dans un même code de signes la prononciation de diverses langues [16].

Cet alphabet phonétique représente un système partagé par la plupart des linguistes. Elle permet d'associer pour chaque symbole un son qui lui correspond. Quand on veut représenter les prononciations dans ce système

on met la représentation entre crochets. Ainsi, pour écrire le son associé au mot *chat* on écrit [ʃa].

Il faut savoir que plusieurs niveaux doivent être pris en compte durant le passage orthographique phonétique, parmi ces niveaux on peut citer : le niveau phonétique et phonologique; le lexical; le syntaxique et même sémantique, etc. Le Tableau 1.1 représente un exemple de flexion orthographique phonétique des consonnes de l'Arabe Standard ; telle qu'il fournit pour chaque symbole arabe le code phonétique qui lui correspond.

Graphème	Phonème Code Informatique	E/NE	V/NV	Graphème	Phonème Code Informatique	E/NE	V/NV
Occlusives				**Fricatives**			
ب د ت ك ء ظ ط ق	[b] [d] [t] [k] [e] [^] [T] [q]	Non emphatique Emphatique	V V NV NV NV NV V NV	ز ذ ث ع غ ش ف خ ه ح ض ص	[Z] [µ] [g] [£] [s] [&] [f] [§] [x] [H] [h]	Non Emphatique Emphatique	V V V V NV NV NV NV NV NV NV NV V
Nasales				**Liquide**			
م ن	[m] [n]	NE NE	V V	ل	[l]	E	V
Vibrante				**Semi-voyelles**			
ر	[r]	NE	V	و ي	[w] [y]	NE NE	V V

Tableau 1.1 : Correspondance graphème phonème de l'AS suivant l'API avec E : Emphatique et NE : Non Emphatique V : Voisée et NV : Non Voisée.

1.8. L'Arabe Standard (AS)

La langue Arabe est une langue phonétique. Elle comporte 28 consonnes ou [huruuf] plus la hamza et 6 voyelles ou [harakaat], plus 6 variantes vocaliques en contexte emphatique.

1.8.1. Les consonnes de l'Arabe Standard

Toutes les lettres de l'alphabet Arabe sont des consonnes. Elles changent de formes de présentation suivant leurs positions (au début, au milieu, ou à la fin) à l'intérieur des mots (Tableau 1.2). A part l'ensemble / ا, و, ر, ز, د, ذ / [E, w, r, Z, d, µ] qui représente les consonnes qui ne se joignent pas à gauche, toutes les lettres se lient entre elles.

[t] au début du mot		**[t] au milieu du mot**		**[t] à la fin du mot**	
تمر	تـ	كتب	ـتـ	حياة زيت مدرسة	ـة ـت ـة

Tableau 1.2 : Exemples de variations de la lettre /ت/ [t] dans les différentes positions : initiale, médiane, et finale.

En réalité on peut diviser les 28 consonnes en deux groupes :

- 14 consonnes *solaires* /ت ث د ذ ر ز س ش ص ض ط ظ ل ن/ [n, l, ^, T, D, $, §, s, Z, r, µ, d, &, t] qui assimilent le /ل/ de l'article, c'est-à-dire lors de la prononciation on élimine le son qui correspond à la lettre /ل/.
  Exemple : le mot /ألشَمسُ/ [a§§amsu] qui signifie le soleil, sera prononcé [a§§amsu] et pas [al§§amsu] ;

- 14 consonnes *lunaires* / أ ب ج ح خ ع غ ف ق آ هـ م و ي/ [y, w, m, H, E, q, f, g, £, x, h, z, b, a] qui se prononcent /ل/ de l'article.
  Exemple : le mot /ألقَمَرُ/ sera prononcé [alqamaru] qui signifie la lune.

Suivant les organes de l'appareil phonatoire mis en jeu et leurs excitations ; Il est possible de faire une autre classification des consonnes tout en se basant sur le mode et le lieu d'articulation (Figure 1.8).

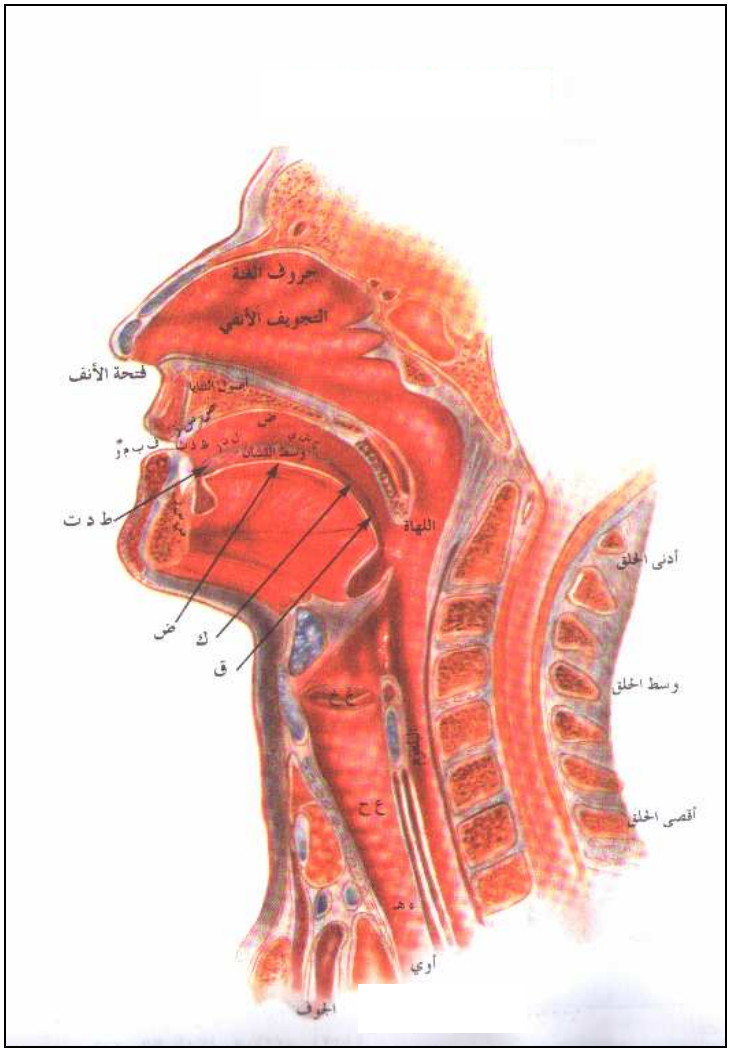


Figure 1.8 : Les lieux d'articulation des 28 consonnes de l'Arabe Standard[17].

1.8.1.1. Classification des sons selon le mode d'articulation

Suivant la classification des sons on peut distinguer plusieurs types de consonnes :

- Voisées / non voisées (sonores / sourdes), les sons voisés sont dûs à des vibrations des cordes vocales; alors que les sons non voisés correspondent à une génération de bruit.

  Le son voisé est un ensemble d'impulsions périodiques. Son spectre se compose d'une fréquence fondamentale F_0 qui représente les premières raies dans la Figure 1.9 (a), et un ensemble d'autres raies

correspondent aux harmoniques du pitch, l'enveloppe de ces derniers présente des maxima appelés *Formants* (Figure 1.9 (b)).

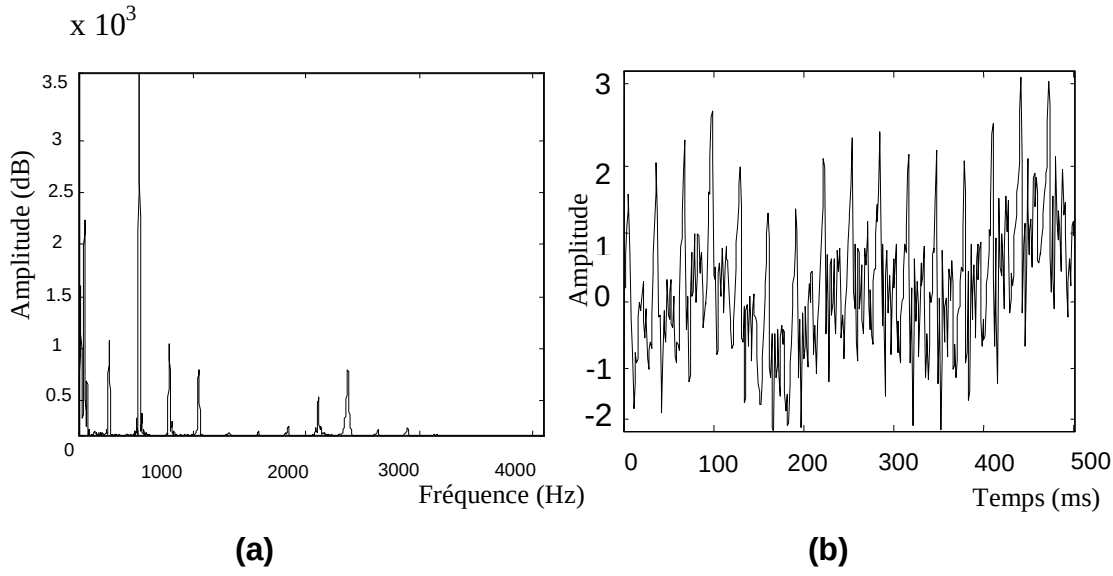


Figure 1.9:(a) Le spectre d'un son voisé (b) La forme d'onde d'un son voisé[3]

Le son non voisé peut être considéré comme un bruit blanc qui résulte d'un écoulement turbulent de l'air à travers le conduit vocal. Sa forme d'onde ne présente aucune périodicité. On remarque aussi que leur spectre ne présente pas de structure de pitch (Figure 1.10 (a, b)).

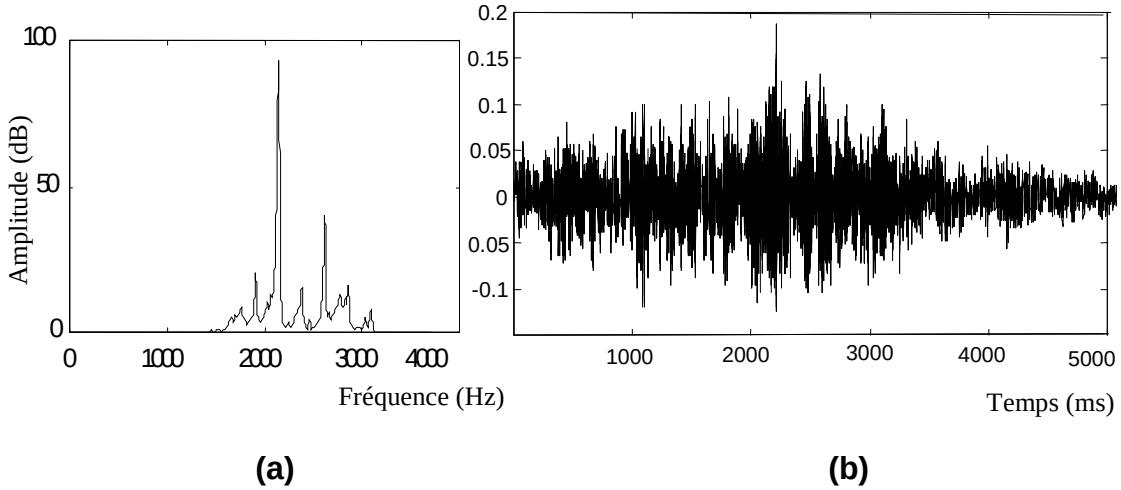


Figure 1.10 : (a) Le spectre d'un son non voisé; (b) La forme d'onde d'un son non voisé [3].

- Occlusives ou plosives / constrictives ou fricatives, le premier type de consonnes est caractérisé par une fermeture complète (occlusion) en un point du conduit vocal. La détente de cette occlusion s'accompagne d'un bruit explosif typique de la consonne occlusive [18]. Les sons du deuxième type sont générés par une constriction en un point de conduit vocal. Cette dernière est accompagnée par un passage continu de l'air ;

- L'opposition nasale /orale, dans le premier cas le son est produit à travers un couplage entre les cavités pharyngo-buccale et nasale; et dans le second l'air passe par la cavité buccale seulement (Figure 1.11) ;

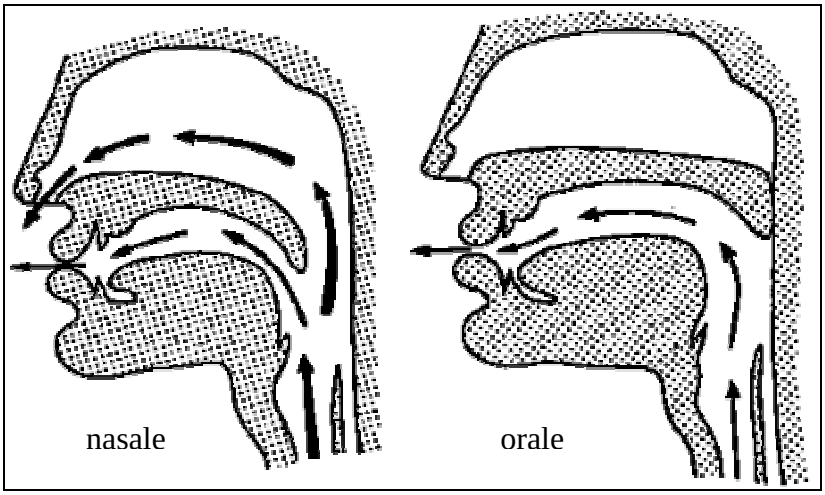


Figure 1.11 : L'opposition nasale / orale [14].

- Liquides, l'articulation des liquides ressemble à une voyelle, la seule différence réside dans la fermeture partielle de conduit vocal (c'est le cas de / ل / [l]).
  la classe des liquides est parmi les sons les plus difficiles à segmenter car elle influence les sons voisins progressivement et régressivement (phénomène de l'assimilation) ;

- Vibrantes, le passage de l'air dans une consonne vibrante est interrompu par des brèves occlusions successives.

1.8.1.2. Classification des sons selon le lieu d’articulation

Le lieu d’articulation est la zone du conduit vocal qui participe à la formation du son. Il présente la position de la constriction totale (cas des occlusives) ou partielle (cas des fricatives) d’une zone spécifique du conduit vocal lors du passage de l’air provenant des poumons. Le lieu d’articulation peut être bilabiale, glottale, labiodentale, etc. (Tableau 1.3) [11].

Semi-voyelles	Vibrantes	Affriquées	Liquides	Nasales	Fricatives		Occlusives		
sonores	sonores	sonores	sonores	sonores	sonores	sourdes	sonores	sourdes	
و				م			ب		Bilabiales
						ف			Labiodentales
			ل	ن	ز ، ص	س ، ث ، ظ			Dentales
	ر				ذ	ش	د ، ض	ت، ط	Alvéolaires
								ك	Post-alvéolaires
									Rétroflexes
ي		ج							Palatales
						خ			Vélaires
					غ			ق	Uvulaires
					ع	ح			Pharyngales
						ه		ء	Glottales

Tableau 1.3 : Les consonnes de l’Arabe Standard [19].

1.8.2. Les voyelles de l’Arabe Standard

En Arabe Standard chaque consonne (ou [harf]) est suivie par une voyelle [harakatun] pour qu’elle puisse être produite. Cette voyelle correspond au mouvement aéro-organique qui assure la réalisation de ce [harf].

Les voyelles sont ajoutées au-dessus ou au-dessous des lettres (ـَ, ـِ, ـُ, ـْ). Elles sont nécessaires à la lecture et à la compréhension correcte d’un texte, en permettant de différencier des mots ayant la même représentation. Cependant, les voyelles ne sont utilisées que pour des textes sacrés et didactiques. Les textes courants rencontrés dans les journaux et les livres n′en comportent habituellement pas [20]. En Arabe Standard il n’existe pas de voyelles nasales en tant que phonème [21]. Les 6 voyelles Arabe peuvent être divisées en deux classes (Tableau 1.4).

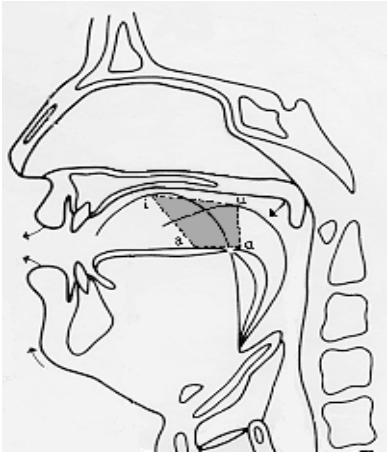
Les voyelles courtes	**Les voyelles longues (almadd)**
ـَ [a] fathatun	ا ـَ [aa]
ـُ [u] Dammatun	و ـُ [uu]
ـِ [i] kasratun	ي ـِ [ii]

Tableau 1.4 : Classification des voyelles de l’Arabe Standard.

Exemple :

/ حَنَانُ / [hanaanu] ، / كُتِبَ / [kutiba]

Il faut savoir aussi que la durée d’une voyelle longue est environ double de celle d’une voyelle courte (selon le contexte), de plus les différentes voyelles courtes se diffèrent entre elles par leurs lieux d’articulation (Figure 1.12).



	palatales (anté-rieures)	centrales	vélaires (post-érieures)
fermées	[kasratun]		[Dammatun]
semi-fermées			
semi-ouvertes			
ouvertes		[fathatun]	

Figure 1.12 : les lieux d’articulation des voyelles courtes de l’Arabe.

Si une consonne n’est liée à aucune voyelle, elle doit comporter un petit rend qu’on appelle [sukuun] ;

Exemple :

دَرْسٌ [darsun] qui signifie la leçon.

1.9. Conclusion

Nous avons fait un bref tour d'horizon sur les caractéristiques de production de la parole, du processus de sa génération, et des principaux traitements appliqués sur les sons, nous avons vu aussi quelques caractéristiques de base de la langue Arabe Standard.

Les objectifs de ce chapitre sont de définir les notions que nous utiliserons dans notre travail. Cette partie théorique sera complétée dans le chapitre suivant par une étude approfondie des systèmes de synthèse de la parole et ses variantes.

CHAPITRE 2 :
TECHNIQUES ET METHODES
DE SYNTHESE DE LA PAROLE

2.1. Introduction

Notre objectif dans ce chapitre est de présenter en détail les techniques d'analyse vocale et la synthèse de la parole qui se positionne au carrefour de l'informatique (car les synthétiseurs sont des logiciels), de la linguistique (chaque système de synthèse vocale se base sur une analyse lexicale, syntaxique, morphologique et parfois sémantique, d'une langue), et de traitement de signal, puisque les sons synthétisés sont des signaux.

2.2. L'Analyse acoustique

L'Analyse acoustique est une partie importante dans le traitement que subit le signal sonore pour pouvoir réaliser un système de haute qualité de synthèse, de compréhension, ou de reconnaissance de la parole.

Cette opération consiste à tirer à partir du signal vocal un ensemble de paramètres pertinents, discriminants et robustes susceptibles de le représenter. Plusieurs techniques d'analyse sont utilisées parmi lesquelles on peut citer l'analyse par :

- Spectrogrammes ;
- Codage prédictif linéaire (Linear Predictive Coding ou LPC).

2.2.1. L'analyse par spectrogrammes

Dans l'étude du phénomène acoustique, on peut réduire la description du son à trois grandeurs physiques : la fréquence (Hz), la durée (s) et

l'amplitude (ou l'énergie) (dB). Par exemple, un son de parole simple, c'est-à-dire sinusoïdal tel que le son qui correspond à la voyelle [fathatun] de l'Arabe Standard en milieu de mot est complètement décrit par les valeurs :

F_0 = 144.4488 Hz, t = 0.0865 s, A = 83.7061 dB.

Et un autre son complexe (Bruité) sera défini, par exemple, de la manière suivante :

- t = 5 s (durée) ;
- F_0 = 100 Hz et A_0 = 70 dB (pitch) ;
- F_1 = 200 Hz et A_1 = 65 dB (2^e harmonique) ;
- F_2 = 300 Hz et A_2 = 50 dB (3^e harmonique).

Remarque : les différentes paramètres prosodiques (t, A, F_0,…etc.) dans les exemples précédents sont calculer par le logiciel PRAAT.

Cela signifie que les trois valeurs durées, fréquence et énergie sont les paramètres pertinents. Une meilleure analyse consiste à les représenter de manière claire et avec précision. L'une des représentations possibles est d'associer deux à deux ces trois grandeurs et de tracer les graphes de ces associations, on obtient les trois plans suivants :

- Dynamique (temps, amplitude) ;
- Du spectre (fréquence, amplitude) ;
- Mélodique (temps, fréquence).

Le spectrogramme est l'une des méthodes d'analyse qui assure une représentation tridimensionnelle de signal de parole tel que (Figure 2.1) :

- L'axe vertical représente la fréquence du son en Hz ;
- L'axe horizontal représente l'évolution temporaire du son ;

- Le degré de noircissement représente l'intensité (l'énergie) en dB du son.

L'objectif principal de spectrogramme est de connaître l'évolution temporelle du spectre de parole. Pour assurer cet objectif, il faut décomposer l'onde acoustique du son en ondes sinusoïdales de différentes fréquences au moyen d'une Transformée de Fourier.

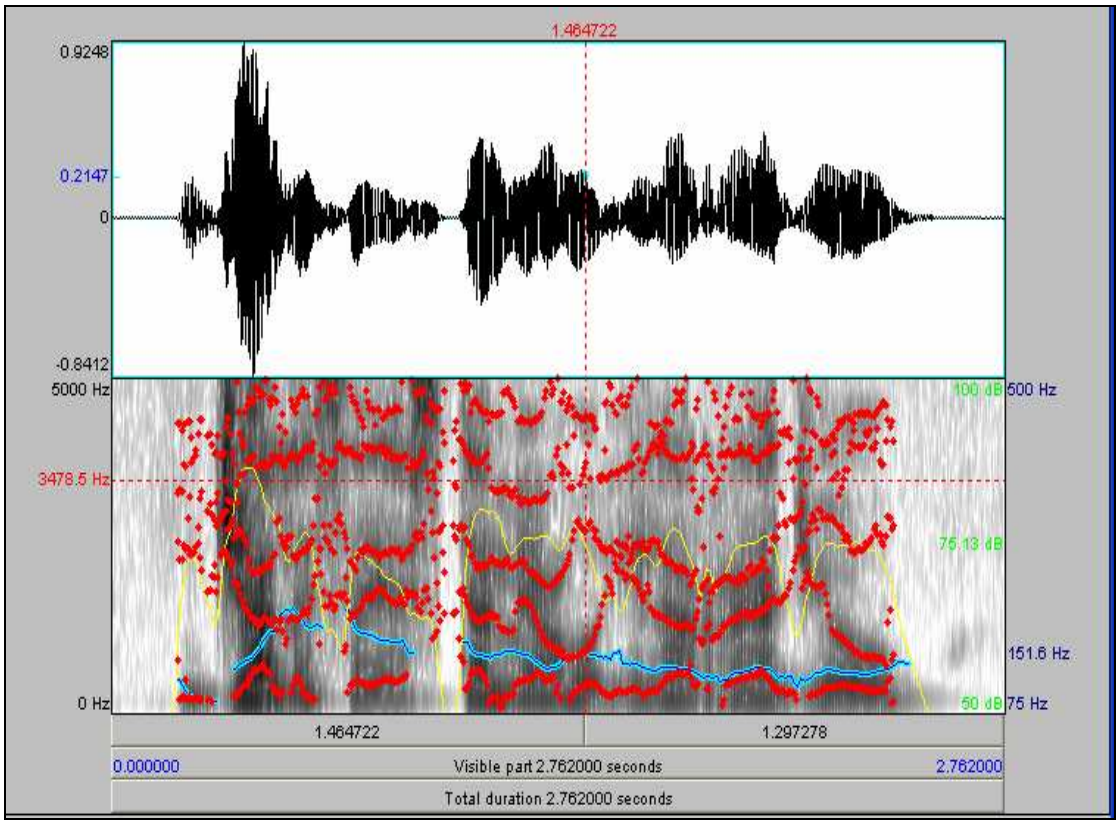


Figure 2.1 : Spectrogramme de la phrase /جلس يستمع إلى ألراديو/ [zalasa yastami£u Eilaa arraadyuu].

Il existe deux représentations possibles pour un spectrogramme, la première en *bande étroite* et la deuxième en *bande large*. La différence essentielle entre les deux réside dans le choix des paramètres qui nous intéressent :

- Un spectrogramme à Bande Large (BL) offre une meilleure résolution fréquentielle et permet de visualiser clairement l'évolution formantique des sons, mais il correspond à une mauvaise analyse temporelle (Figure 2.2. (a)). La classe des occlusives représente le meilleur exemple adapté à cette présentation ;

- Inversement, un spectrogramme à Bande Etroite (BE) offre une bonne résolution au niveau temporel, mais l'analyse fréquentielle est moins fine (Figure 2.2. (b)). Ce type de spectrogramme est souvent utilisé dans l'étude de l'intonation ainsi que dans l'analyse des fricatives.

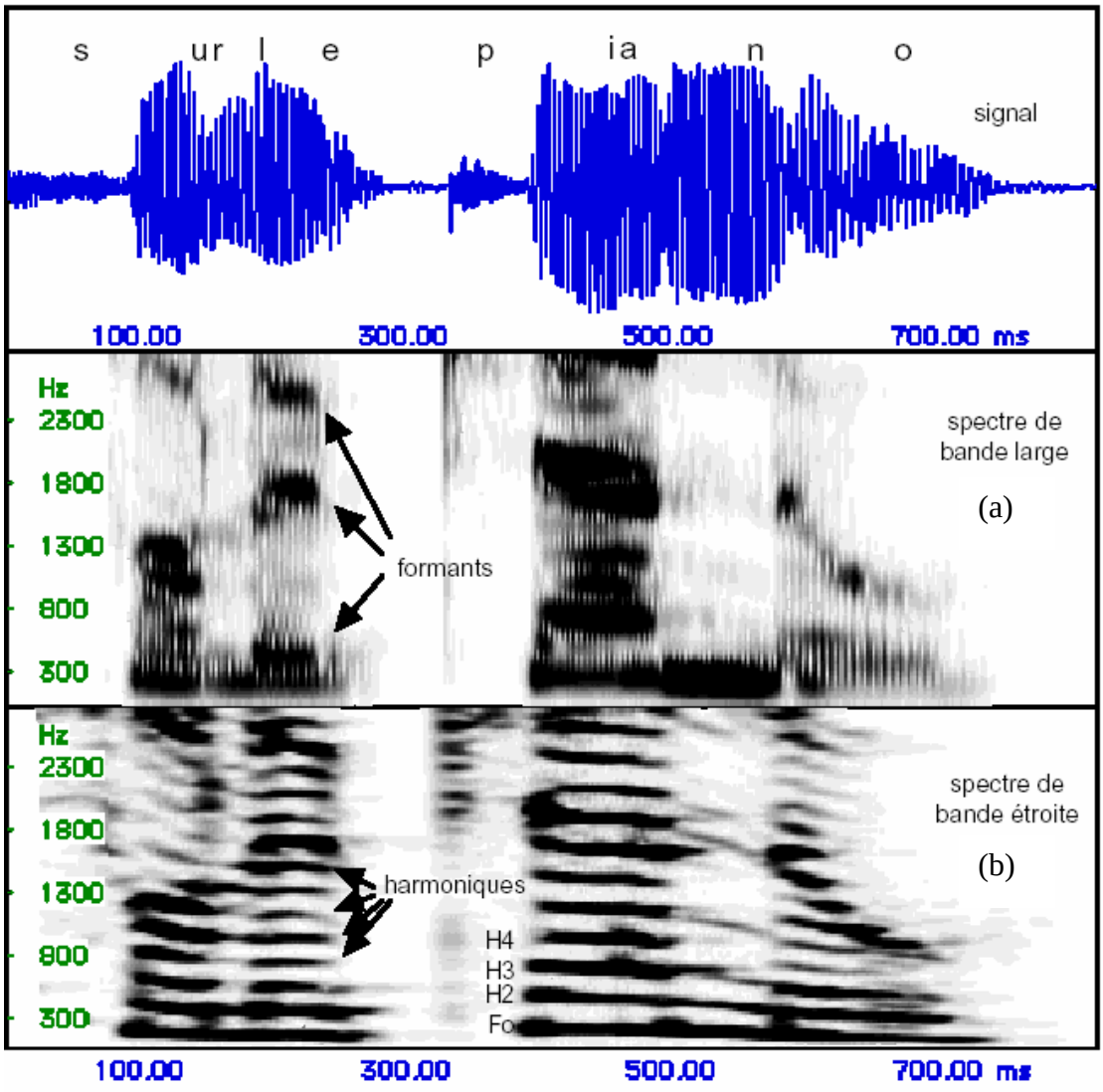


Figure 2.2 : Différentes analyses en BL et BE de l'onde sonore correspondant à l'énoncé « Sur le piano » [22].

2.2.2. L'analyse par Codage Prédictif Linéaire

De la même façon qu'un signal de parole réel créé par les poumons et les cordes vocales, et produit par le passage à travers le filtre que constitue notre conduit vocal. Une parole synthétique peut être modélisé par le passage

d'un signal d'excitation à travers un filtre numérique récursif (Figure 2.3). Cette modélisation est appelée prédictive linéaire puisqu'elle correspond à une régression linéaire entre le signal d'excitation et le signal vocal produit, et elle est représentée par la formule suivante :

$$S(n)+ \sum_{i=1}^{p} a(i)s(n-i) = \sigma\mu(n) \qquad (2.1)$$

Les coefficients de cette régression représentent les coefficients du filtre récursif.

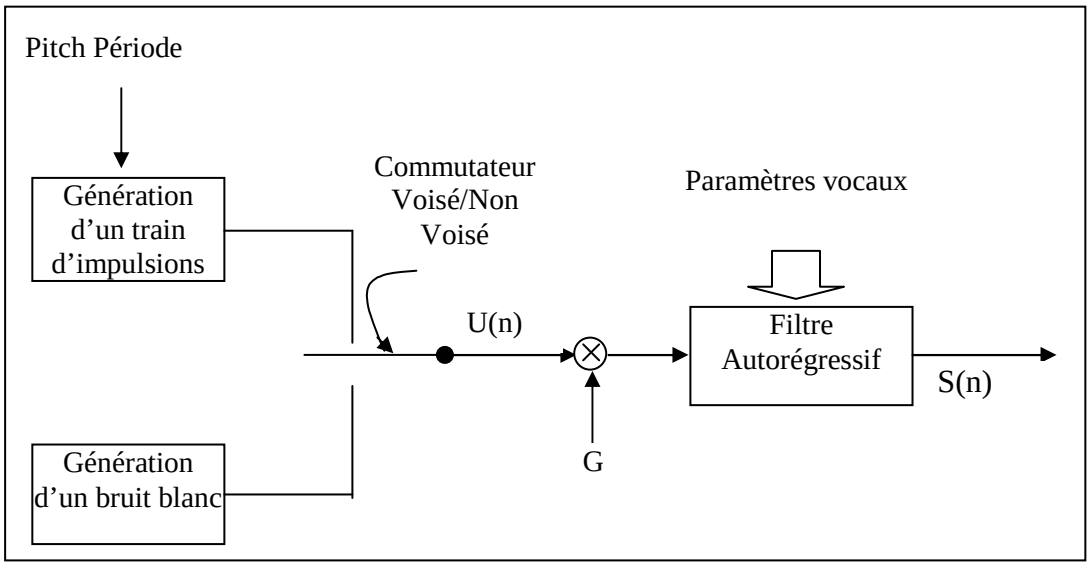


Figure 2.3 : Modèle simple de mécanisme de la génération de la parole [23].

Si x(n) est une prédiction linéaire alors sa formule générale est la suivante :

$$X(n) = \sum_{i=1}^{N} a(i)x(n-i) \qquad (2.2)$$

avec N : Ordre du filtre prédicateur
$a(i)$: Coefficients du filtre de prédiction

La parole n'étant pas un processus parfaitement linéaire, la somme ci-dessus introduit une erreur qu'il faut corriger par l'introduction du terme e (n) qui représente *l'erreur de modélisation* entre le signal original et celui qui est généré par le modèle. On obtient la formule suivante :

$$X(n) = \sum_{i=1}^{N} a(i)x(n-i) + e(n) \qquad (2.3)$$

e : Erreur de prédiction

Les principaux éléments de ce modèle de prédiction linéaire (Modèle Auto-Régressif AR) sont :

- les coefficients (a_i) i ∈ [1, N]
- l'erreur de prédiction :

$$E(n) = E\left[e^2(n) \right]$$

2.2.2.1. Quelques méthodes pour le calcul de l'erreur

Diverses techniques d'optimisation ont été développées pour le calcul de cette erreur parmi ces techniques on peut citer : la méthode des moindres carrés, le maximum de vraisemblance, le maximum d'entropie, les distances perceptives, etc.

- La Méthode des moindres carrés est une méthode qui s'applique à n'importe quelle fonction. Elle consiste à déterminer un ensemble de **r** paramètres

  a = [$a^{(1)}$, $a^{(2)}$,…, $a^{(r)}$] en minimisant un critère quadratique **J** qui correspondant à la somme des carrés des écarts entre la variable réelle y_i et la valeur correspondante de la fonction modèle optimisée $F_a(x_i)$:

$$J(a) = \sum_{i=1}^{p} (y_i - F_a(x_i))^2 \qquad (2.4)$$

  Dans notre cas cette méthode consiste à minimiser l'erreur quadratique moyenne E donnée par :

$$E = \sum_{n} e_n^2 = \sum_{n} (s_n - \hat{s}_n)^2 \qquad (2.5)$$

Où s_n est le signal original et $\hat{s}_n$ est le signal synthétique ; si on remplace $\hat{s}_n$ par sa valeur on obtient :

$$E = \sum_{n} (s_n - \sum_{i=1}^{p} a_i s_{n-i})^2 \qquad (2.6)$$

2.2.2.2. Calcul des coefficients du filtre de prédiction

Ce calcul peut être assuré par la résolution d'un système de p équations à p inconnus ; pour cela plusieurs algorithmes ont été développés tel que l'algorithme relatif à la méthode d'autocorrélation et qui a abouti aux équations de Yule- Walker suivantes :

R a = R

$$\begin{pmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & & & & \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{pmatrix} \qquad (2.7)$$

Les coefficients R_k k=1,..,p représentent les valeurs d'autocorrélation du signal Les coefficients a_i i=1,..,p représentent les valeurs du filtre de prédiction.

Il faut savoir que les algorithmes de *Levinson et de Schur*; dits les algorithmes rapides, représentent l'une des meilleures solutions de système précédent d'équation matricielle, puisqu'ils considèrent la matrice R comme une matrice de Toeplitz.

Une matrice de Toeplitz est une matrice carrée symétrique a les mêmes valeurs dans les lignes parallèles à la diagonale principale [10].

L'analyse de la parole par la méthode LPC présente deux avantages :

- Elle assure des résultats exacts, et elle s'adapte mieux à l'étude des phénomènes évoluant rapidement ;

- La prédiction permet d'éliminer une part importante de la redondance du signal de parole. En effet, cette redondance se traduit par l'arrivée d'éléments qui ne fournissent pas d'information nouvelle. Savoir prédire la valeur de signal de la parole à un instant t en fonction de ces valeurs passées permet de débarrasser l'ordinateur de ces pseudo informations. Bien évidement, les coefficients de prédiction linéaire devront être réajustés régulièrement [24].

L'étape préliminaire à toutes les analyses acoustiques (spectrogramme, ou LPC) est la segmentation.

2.2.3. La segmentation

Avant de réaliser n'importe quel système de synthèse vocale une opération fondamentale doit être assurée, c'est bien l'opération de segmentation. Cette dernière consiste à découper une parole naturelle et continue qui représente une source sonore en éléments acoustiques unitaires (phones, diphones, mots, etc.)

Selon le dictionnaire Larousse, le terme de segmentation désigne la division d'un ensemble en portions bien délimitées. Autrement dit, c'est le processus de division d'une entité, généralement continue, en petites entités appelées segments. Chaque segment possède des propriétés propres qui permettent de le différencier des autres [25].

En parole la segmentation consiste à couper les séquences audio enregistrées en unités de tailles variables. Tel qu'on place des marqueurs temporels aux limites de ces unités phonétiques (Figure 2.4) ; et cela tout en mettant en correspondance le texte et l'audio.

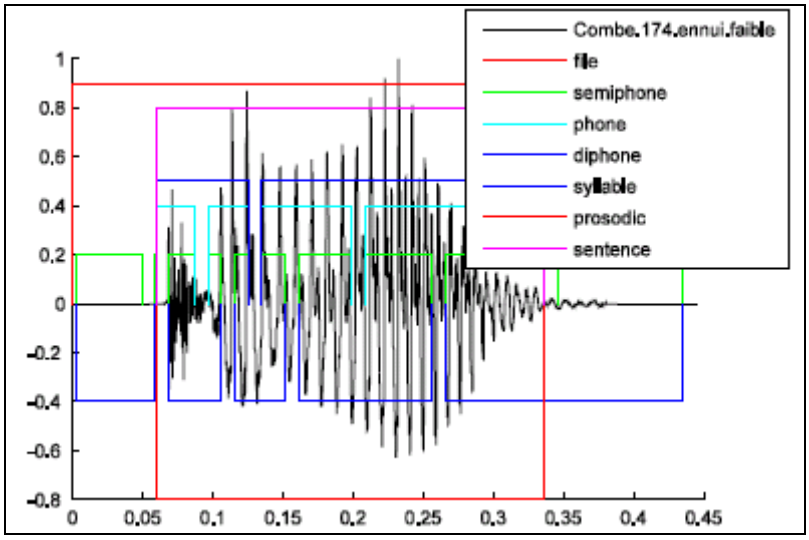


Figure 2.4 : Exemple de segmentation du mot « Comment ? »[26].

En ce qui concerne les modes de segmentation, nous distinguons deux catégories : Manuelle et Automatique

- La segmentation manuelle est assurée par des experts phonéticiens de la langue, son inconvénient majore est dû grâce à la difficulté de bien préciser les frontières des unités segmentales. De plus, une telle tâche nécessite un temps énorme et important durant l'annotation de grands corpus de parole. Il faut savoir qu'une segmentation manuelle effectuée par plusieurs experts ne fournit pas nécessairement les mêmes résultats ;

- La segmentation semi automatique / automatique, Actuellement, la segmentation complètement automatique de parole est une tâche rarement possible. En effet, étant donnée la complexité des phénomènes acoustico-phonétiques à traiter, cette tâche nécessite très souvent une intervention manuelle, que ce soit pour la préparation des données (étiquetage phonétique) du traitement automatique ou autre [25]. Malgré l'existence des outils qui assurent cette opération, ils restent toujours non fiables puisqu'ils ne garantissent pas une très bonne qualité de parole synthétique. Pour cette raison, des vérifications manuelles faites par des experts humains sont indispensables à la segmentation de la parole.

Les approches markoviennes de segmentation d'un signal de parole en phones ont montré leur efficacité à condition de disposer d'une transcription phonétique correspondant exactement à la phrase énoncée par le locuteur. L'originalité de cette étude consiste à reformuler l'hypothèse de segmentation à un niveau plus abstrait : le système de segmentation du signal de parole en phones ne dispose pas de la transcription phonétique exacte du locuteur mais celle provenant d'un système de Transcription Orthographique Phonétique automatique [27].

Lors de la détection des frontières des unités extraites à partir du signal vocal, on peut utiliser l'une des deux techniques suivantes : la première est basée sur l'utilisation de la durée phonémique et la seconde se base sur l'évolution de la fréquence fondamentale :

- L'utilisation de la durée phonémique, dans ce cas l'unité prosodique représente la portion de signal de parole délimitée par deux pauses. Cette technique est pratiquement appliquée dans le traitement de la parole spontanée, puisque cette dernière est caractérisée par un nombre très élevé de pauses.

- L'utilisation de la fréquence fondamentale, les changements importants de la fréquence fondamentale marquent les frontières des unités à extraire.

2.3. La synthèse vocale

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

Le rôle de la synthèse vocale est d'assurer la lecture d'un texte donné en entrée. Puisqu'il est impossible d'enregistrer tous les sons qui

correspondent à tous les mots de ce texte qui peut être écrit dans n'importe quelle langue; la solution est de définir le système de synthèse vocale comme étant un système de production automatique de la parole qui se base sur une transformation de ce texte orthographique en une suite de sons ou phonèmes.

2.4. Historique des systèmes de synthèse vocaux

A plusieurs reprises au cours de l'histoire, on a tenté de reproduire la voix humaine. Au 18^{éme} siècle, on met au point à cet effet des dispositifs mécaniques équipés de soufflets et d'anches vibrantes. Au 20^{éme} siècle, l'apparition de l'électricité et de l'électronique autorise des tentatives plus ambitieuses : en 1922, J. C. Stewart fabrique une machine capable de reproduire des voyelles, des diphtongues et quelques mots simples tel que « mama, Anna » ; Plusieurs années plus tard, en 1939, H. Dudley présente, à l'occasion de l'exposition universelle de New York, le VODer (Voice Operation Demonstrator) appareil mis au point par les laboratoires Bell. Mais ce n'est que dans les années cinquante que les premiers véritables synthétiseurs de la parole font leur apparition, avec par exemple le pattern Play-back, système mis au point par les laboratoires Haskins aux USA, qui se présente comme un lecteur de sonagraphe (un faisceau de lumière produit, après amplification, des sons à partir de la représentation de leur durée, de leur fréquence et de leur intensité).

Depuis les années soixante-dix, les progrès considérables ont été accomplis, avec notamment le développement de l'utilisation des calculateurs numériques, Aujourd'hui encore, ces progrès se poursuivent, dans plusieurs directions (perfectionnement des synthétiseurs à formants, des synthétiseurs à prédiction linéaire, etc.) [11].

Nous présentons dans le paragraphe suivant les principales applications actuelles, ou en cours de développement de la synthèse de parole.

2.5. Les Applications de la synthèse vocale

Les champs d'applications des systèmes de synthèse de la parole sont nombreuses nous citons quelques unes à titres d'exemples :

- Aide aux personnes handicapées : la synthèse offre également énormément de services aux personnes handicapées, par exemple elle fournit des machines d'aide au lecture pour les mal-voyants. Ces machines assurent l'accès sous forme vocale aux informations qui existent dans les ordinateurs. La synthèse permet aussi de générer des assistances aux personnes muettes, elle peut garantir des systèmes pour commander des chaises roulantes vocalement, etc ;

- Services des télécommunications, le but principal des machines de synthèse de parole est de rendre tout type d'information écrite disponible via le téléphone. Telles qu'on peut créer des serveurs vocaux diffusant les horaires des cinémas, des informations routières, ou consulter un compte bancaire vocalement, ou encore donner des explications automatisées concernant les factures de téléphone ;

- Livre et jouets parlants : le marché du jouet a déjà été touché par la synthèse vocale. De nombreux ordinateurs pour enfants possèdent une sortie vocale qui en augmente l'attrait, particulièrement chez les jeunes enfants (pour qui la voix est le seul moyen de communication avec la machine) ;

- Les appareils qui génèrent quelques mots des langues étrangères : c'est le cas des petits dictionnaires électroniques de poche, des traducteurs électroniques mot à mot qui sont apparus récemment et qui peuvent être utilisés pour lire un ouvrage dans une langue étrangère et cela par l'intermédiaire d'un stylo optique (utilisé pour sélectionner instantanément un mot inconnu et entendre à la fin la prononciation qui lui correspond). Ces appareils peuvent présenter un avantage non négligeable dans l'apprentissage des langues étrangères principalement pour apprendre la prononciation de celles-ci.

- La communication Homme-Machine tout en remplaçant les lignes de commandes qui sont utilisées pour réaliser une tâche particulière de l'ordinateur par une simple commande vocale plus naturelle ;

- La surveillance dans un centre de contrôle industriel, il est préférable de remplacer les alertes qui génèrent des sons gênants par des voix synthétiques qui de plus indiquent l'emplacement de l'anomalie ou la panne qui engendre le dysfonctionnement de la machine industrielle ;

- La recherche fondamentale et appliquée : enfin, les synthétiseurs possèdent aux yeux des phonéticiens une qualité qui nous fait défaut : ils peuvent répéter deux fois exactement la même chose. Ils sont par conséquent utiles pour la validation des théories relatives à la production, à la perception, ou à la compréhension de la parole ;

- Les systèmes embarqués représentent le plus grand marché à exploiter. Ce sont des systèmes qui combinent des parties matérielles et d'autres logiciels; le téléphone portable est un exemple de ce type des systèmes oú l'utilisation de la synthèse vocale s'impose naturellement

2.6. Quelques systèmes de synthèse vocale

Plusieurs systèmes de synthèse à partir du texte sont aujourd'hui disponibles. Citons entre autre les systèmes commerciaux d'Elan Speech, d'AT&T Labs, de Bel Labs et de Babel Technologies, les systèmes expérimentaux de France Télécom, du LIMSI , de l'ICP, et FIPSVOX développé au LATL de l'Université de Genève. En ce qui concerne l'Arabe, il existe un système expérimental développé à l'IRSIT, et celui d'IBM [1].

Concernant la synthèse vocale la plupart des solutions disponibles actuellement utilisent des logiciels qui semblent partager une déficience commune. Elles sont principalement limitées à l'Anglais, ne fournissant qu'un support très marginal pour les autres langues, ou dans la plupart des cas il n'y

a aucun support. Nous allons citer dans le Tableau 2.1 quelques applications des systèmes de synthèse de parole pour des domaines multiples.

Outils	Description
Festival	Un système de lecture à partir du texte développé au CSTR (Centre for Speech Technology Research, le centre pour la recherche en technologie de la parole) de l'université d'Édimbourg. Festival est un outil qui est développé sous linux et il supporte plus d'une langue naturelle. Il peut synthétiser l'Anglais, l'Espagnol.
EFlite	Un lecteur automatique du texte anglais seulement, il est développé à l'université de Carnegie Mellon
Mbrola	Un système de synthèse vocale de dix langues différentes. Il se base sur la concaténation de phonèmes
Speech Dispatcher	C'est une collection de synthétiseurs vocaux qui travaillent en collaboration et cela pour exploiter les avantages de chaque synthétiseur.
Infovox	C'est un système embarqué qui contient une partie câblé et une autre programmée. Il est basé sur la technique de synthèse à formants, et il utilise le diphone comme unité de concaténation. Son débit est de 400 mots par minute, ce système prononce différentes voix (hommes, femmes, enfants).
DecTalk	La voix est prononcée par plusieurs locuteurs (hommes, femmes, enfants) et selon plusieurs langues (espagnole, allemande, et anglaise). Le DecTalk peut traiter les noms propres, les mails et les liens Internet, etc.
Bell Labs	AT&T Bell laboratories, c'est le laboratoire de synthèse le plus ancien, il fut depuis l'apparition de VODER en 1939. Les systèmes actuels de ce laboratoire utilisent la concaténation par diphones, et ils prononcent plusieurs langues (française, allemande, espagnole, italienne, russe, roumaine, japonaise, chinoise, et anglaise).

CNET PSOLA	France Telecom CNET a fourni ce système qui combine la synthèse par diphones et l'algorithme PSOLA, ce dernier est souvent utilisé pour améliorer la prosodie.
Lernout & Hauspies	Il utilise la synthèse par concaténation à base de diphones et de polysons (plusieurs sons). Il supporte plusieurs langues parmi ces dernières on peut trouver l'Arabe.
Acu Voice	C'est un système de synthèse vocale par concaténation de syllabes. Sa base de données sonores contient 60 .000 segments anglais.
CyberTalk	Ce système utilise les deux méthodes de synthèse par concaténation, et par règles. Cette dernière est utilisée pour produire les voyelles, tandis que la première méthode est utilisée pour générer les nombres et les chaînes alphanumériques. Les occlusives et les fricatives sont préenregistrées.
ModelTalker	C'est un système de synthèse à partir du texte de l'Anglais. Il a été développé à l'université de Delaware (USA). Il utilise la méthode de synthèse par concaténation de diphones.
Whistler	Il contient une étape d'apprentissage qui se base sur les Modèles de Markov Cachés (Hidden Markov Models ou HMM), la génération des différents sons est effectuée à travers une segmentation virtuelle (sélectionner l'unité à partir des mots ou des phrases sans altérer ces derniers).
HADIFIX	HAlbsilben, DIphone, SufFIXe, c'est un système de synthèse par concaténation de diphones, demi syllabes et suffixes qui permet de contrôler la durée, le pitch, le rythme, la pause, etc. Sa base de données sonores contient plus de 150 diphones, 180 suffixes et 750 demi-syllabes ce qui est largement suffisant pour générer tout le vocabulaire allemand.

.

Tableau 2.1. Quelques exemples des systèmes de synthèse vocale [9].

2.7. Les techniques de la synthèse de parole

Le principe des synthétiseurs vocaux est de créer une analogie avec l'appareil phonatoire humain. Les différentes techniques offertes par ces synthétiseurs peuvent être résumées par les points suivants (Tableau 2.2) :

- Spectrale : cette technique englobe l'ensemble des vocodeurs (Voice Coder) à canaux, à formants, et à prédiction linéaire ;
- Temporelle (synthétiseur par formes d'ondes) : c'est la compression de la parole numérisée M.I.C (Modulation par Impulsions Codées) ;
- Articulatoire : c'est une simulation du conduit vocal.

Synthétiseurs Acoustiques	**Avantages**	**Inconvénients**
A Canaux	Une bonne qualité d'analyse et de synthèse Source et fonction de transfert séparées	Intégration difficile Qualité variable selon le locuteur
A Formants	Parole plus naturelle que le premier cas Grande souplesse Source et fonction de transfert séparées	Qualité variable selon le locuteur Paramètres de commandes difficiles à obtenir par une analyse automatique
A Prédiction Linéaire	Source et fonction de transfert séparées Théorie mathématique se prêtant mieux aux simulations informatiques Elimination de la redondance existant dans la forme temporelle de signal	Qualité variable selon le locuteur

Tableau 2.2 : Les avantages et les inconvénients des synthétiseurs du domaine spectral [8].

2.8. Les phases fondamentales d'un système de synthèse vocale

Le schéma général d'un synthétiseur de parole peut être illustré par le dialogue entre l'homme et la machine qui se résume par la Figure 2.5 :

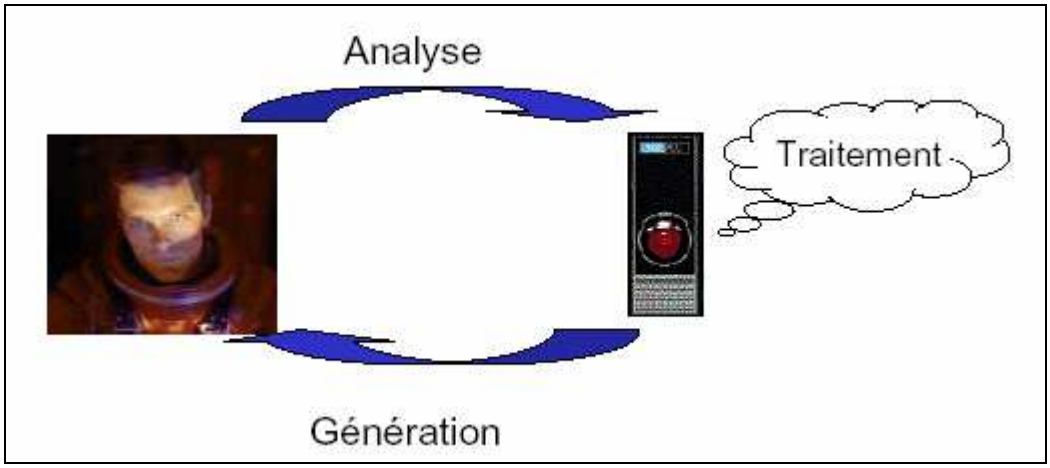


Figure 2.5 : Schéma général d'un système de dialogue [6].

Deux phases fondamentales marquent le processus de n'importe quel système de synthèse vocale, l'analyse et la génération :

- L'analyse est une reconstruction formelle du message en entrée (oral ou écrit), elle englobe l'étape de la Transcription Orthographique Phonétique ;
- La génération est la production acoustique d'un message oral à partir d'une représentation interne.

2.9. La synthèse de la parole à partir du texte

L'objectif principal des systèmes de synthèse vocale à partir du texte est la création d'un signal de parole correspondant à la prononciation d'un message écrit donné en entrée. Actuellement, l'état de l'art technologique de tels systèmes consiste à assembler des unités de parole élémentaires pour créer la matière sonore. L'ensemble des unités acoustiques, connu, pour la plupart des systèmes comme par exemple l'ensemble de diphones est fixe et déterminé par expertise phonétique et acoustique quelles que soient les phrases de synthèse à créer [27].

La synthèse vocale à partir d'une représentation textuelle est la technologie qui permet d'automatiser la production d'une parole artificielle par une machine grâce à une Transcription Graphème-Phonème des phrases à lire. Chaque système de synthèse de la parole à partir du texte est divisé en deux parties (Figure 2.6).

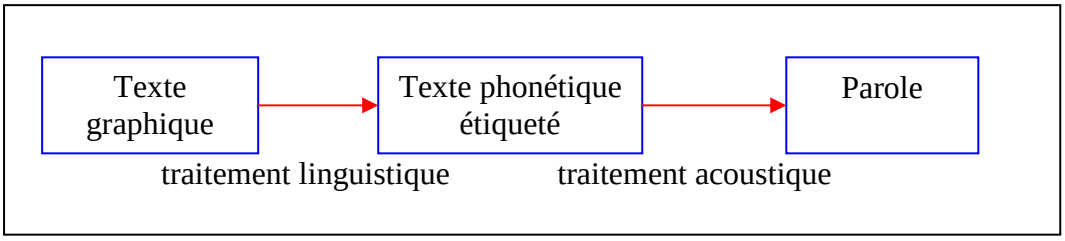


Figure 2.6 : Schéma de système de synthèse de la parole [28].

2.9.1. La synthèse linguistique ou symbolique

Cette étape consiste à analyser la phrase écrite, et à partir des règles de la phonétique, produire une phrase qui sera le reflet exact de ce que sera la phrase écrite lorsqu'elle est parlée. Cette opération représente le passage d'un texte écrit en un texte phonétique étiqueté.

Avant d'appliquer les règles phonétiques au texte écrit, il faut d'abord faire quelques transformations dans la phrase puis faire la transcription. Ces transformations se résument par les étapes suivantes :

- L'étape de prétraitement et normalisation du texte développe toute forme de texte en une forme littérale, parfaitement désambiguïsée. En particulier, ce module traite les problèmes de formats, les dates et heures, les abréviations, les nombres, les monnaies, les adresses Email, etc.

  Par exemple :

  / دنانير 5 الفستان ب / [alfustaanu bi xamsu danaaniir] qui signifie le prix de la robe est de 5 DA ;

  Cette partie du processus nécessite une connaissance approfondie des formats d'écriture dans de très nombreux contextes.

- Le remplacement de chaque caractère composé par ses équivalents ;

  Exemple : لا ⟶ ل ا

- La consultation du lexique des exceptions pour l'élimination des mots spéciaux ;

- L'application des règles de transcription établies pour la langue. Ce module doit traiter des tâches grammaticales complexes pour identifier les règles de transcription à utiliser dans un contexte donné.
  Par exemple : « Les poules du président couvent pendant qu'il préside le couvent » ; où couvent se prononce différemment si c'est un verbe ou un nom ;

- La conversion graphème phonème ou Transcription Orthographique Phonétique, cette phase associe à chaque graphème le son qui lui correspond.

A l'issue du traitement linguistique, le texte transcrit sera archivé dans une base de données.

2.9.2. La génération acoustique

C'est le traitement du signal proprement dit. Cette partie permet de recréer intégralement le signal de la parole à partir d'un modèle stocké dans la base. Ce qui correspond à la synthèse réelle. La méthode retenue dans ce cas est la concaténation directe des formes d'ondes des sons de la phrase issue de la conversion. Ces formes d'ondes sont classées en mémoire sous forme numérique [29].

Ces deux étapes sont suivies d'une Conversion Numérique/Analogique du signal vocal, puis d'une amplification audio (Figure 2.7).

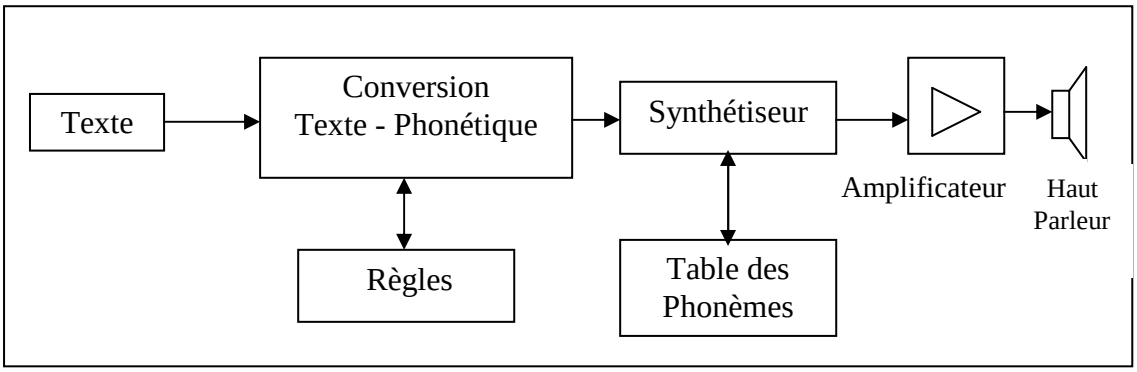


Figure 2.7 : Schéma synoptique du système de synthèse de la parole [29]

D'après la Figure 2.7 précédente on peut conclure que le synthétiseur de parole représente une étape de système de synthèse de parole et n'est pas le système lui même. Il existe deux types de synthétiseurs :

- Les synthétiseurs de parole à partir d'une représentation numérique qui permettent de générer un signal vocal à partir des caractéristiques numériques obtenues lors de son analyse. Ces synthétiseurs jouent un rôle inverse a celui des analyseurs ;

- Les synthétiseurs de parole à partir d'une représentation symbolique [13], inverse des reconnaisseurs de parole et ils sont capables, de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable.

  Dans cette seconde catégorie, on classe également les synthétiseurs en fonction de leur mode opératoire :

. - les synthétiseurs à partir du texte reçoivent en entrée un texte orthographique et doivent en donner lecture ;

  - les synthétiseurs à partir de concepts, appelés à être insérés dans des systèmes de dialogue Homme-Machine. Ils reçoivent le texte à prononcer et sa structure linguistique.

2.10. Principe de fonctionnement d'un système de synthèse à partir du texte

Les systèmes de synthèse à partir texte (les systèmes Text To Speech TTS) fait partie de la classe des synthétiseurs vocaux. L'organisation générale des opérations de traitement du langage, réalisées pour le passage d'une information écrite à un son généré est représenté par la Figure 2.8 où on remarque que chaque système de synthèse à partir de texte contient le module :

- De traitement du texte ;
- La base de données sonores ;
- Et le module de traitement de signal appelé aussi synthétiseur.

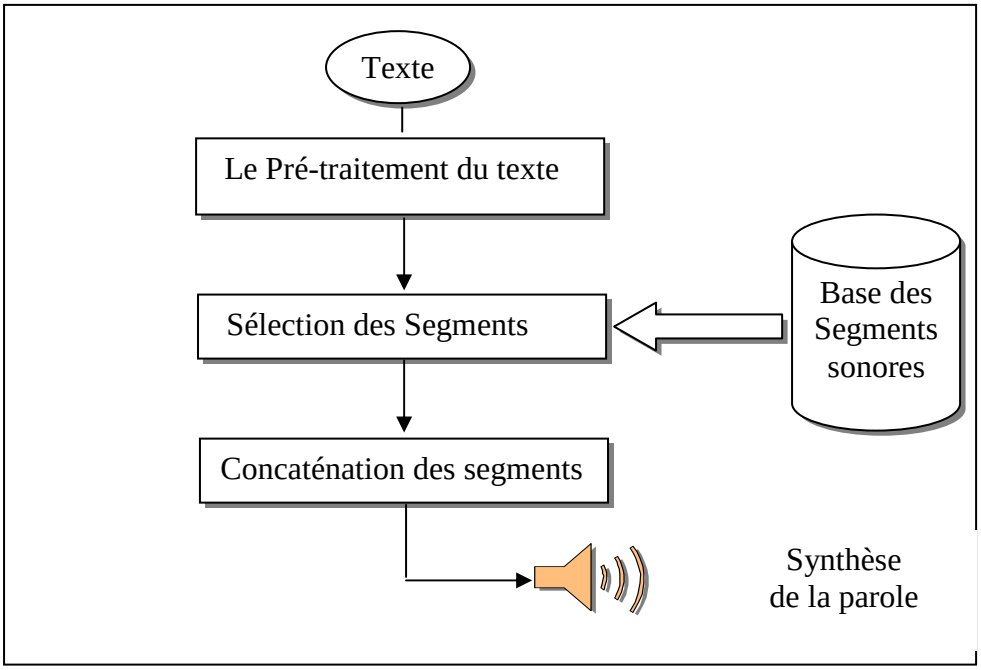


Figure 2.8 : Schéma général d'un synthétiseur à partir du texte.

2.11. Les méthodes de synthèse vocale à partir d'un texte

Il y a deux approches principales pour convertir un texte en parole : la synthèse par concaténation et la synthèse par règles.

2.11.1. Synthèse par concaténation d'unités pré-stockées

La synthèse par concaténation d'unités pré-stockées est la génération des sons à partir de la juxtaposition d'un ensemble d'unités préenregistrées, ces dernières sont obtenues par une opération d'analyse du signal qu'on veut produire. Elle consiste à choisir dans une large base de données les unités sonores les plus appropriées pour construire, par concaténation la phrase à produire (Figure 2.9). En réalité dans cette approche on peut trouver plusieurs types d'unités (phonèmes, diphones, syllabes, polysons, mots, etc.)

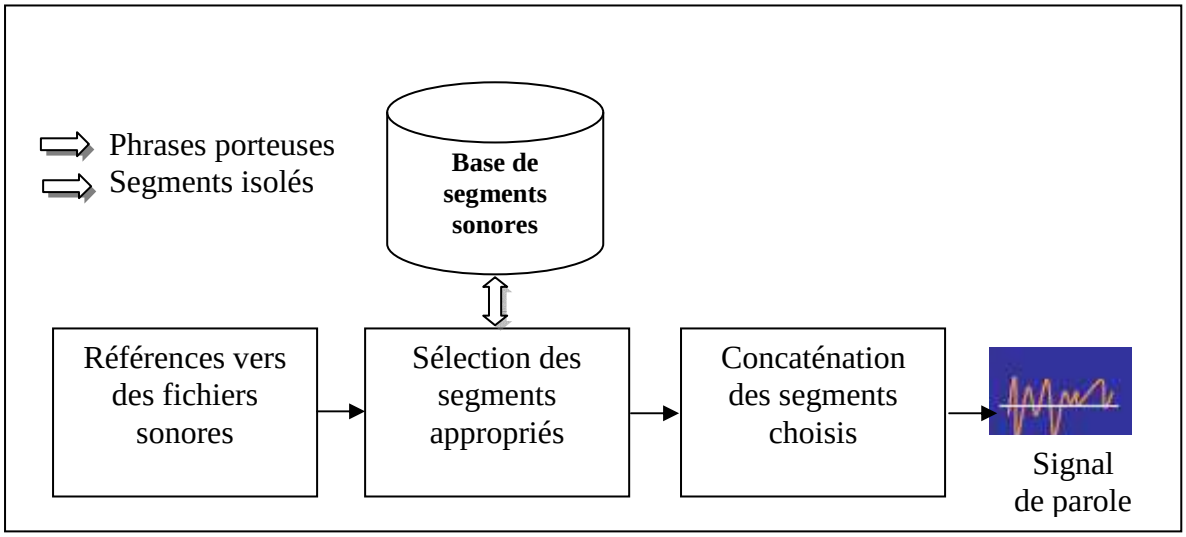


Figure 2.9 : Principe de base de la méthode de synthèse par concaténation.

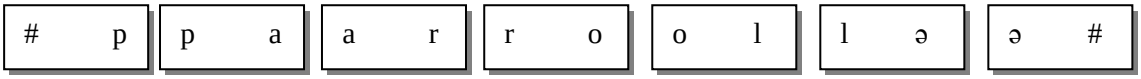
Parmi les méthodes de synthèse par concaténation on a :

- La concaténation de phrases est une simple opération d'enregistrement et de restitution des phrases à synthétiser en vue d'une réalisation bien précise. Cette précision limite leur utilisation à un petit vocabulaire, ainsi qu'à des applications très restreintes telles que les jouets pour enfants, les répondeurs téléphoniques, l'horloge parlante, etc. Cette dernière se compose de deux parties, une stable qui correspond à la phrase «il est : … heures ….minutes….secondes », et une autre variable où on trouve les nombres qui correspondent à la valeur actuelle de l'heure, des minutes, et des secondes ;

- La concaténation de mots, il s'agit de juxtaposer un ensemble de mots l'un à côté de l'autre pour générer une phrase avec une qualité moins bonne par rapport aux phrases qui sont obtenues à travers l'utilisation de type précédant de concaténation ;

- La concaténation de phonèmes, puisque les phonèmes représentent les éléments atomiques dans n'importe quelle langue, il suffit de les juxtaposer pour synthétiser un mot ou une phrase. Malgré la simplicité de cette méthode, elle présente l'inconvénient de discontinuité du signal généré et cela à cause du problème de la coarticulation qui est dû grâce à l'influence d'un son sur ses voisins. Pour résoudre ce problème la solution est de changer le phonème par une autre unité plus coûteuse en information qui est le diphone ;

- La concaténation par diphones, consiste à enregistrer dans la base de données sonores les diphones nécessaires pour produire la parole. Chaque diphone représente le segment qui est compris entre deux parties stables de deux phonémes consécutifs en prenant toute la transition. Exemple : juxtaposition de diphones pour générer [parolə], ou le symbole # correspond au silence



Malgré ça, cette synthèse présente toujours le problème des effets de la coarticulation dépassant la limite du phonème ce qui donne naissance à la synthèse par concaténation de polysons ;

- La concaténation de polysons est un phénomène qui est dû souvent à l'instabilité des phonèmes liquides, et semi-voyelles. Pour trouver une solution, il faut éviter de segmenter ce phonème instable au milieu ; c'est-à-dire il faut intégrer complètement le phonème avec ses voisins pour construire une seule unité de synthèse (Figure 2.10).

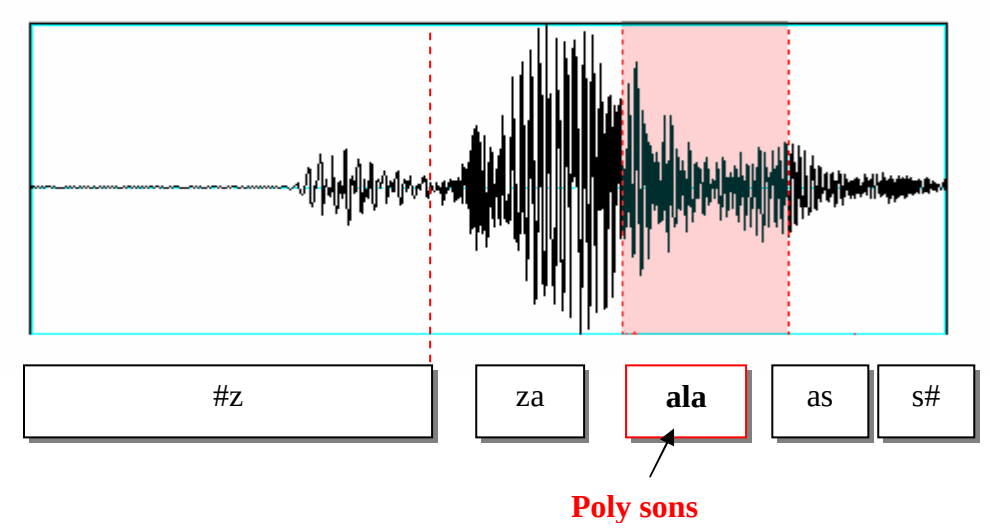


Figure 2.10 : Décomposition en polysons du mot / جلس / [zalas]

2.11.2. La synthèse par règles

La synthèse par règles suppose la connaissance des mécanismes de production et de perception de la parole. Le signal acoustique est d'abord analysé pour extraire une représentation simplifiée du phonème sous forme de valeurs cibles. La transition entre ces valeurs cibles est ensuite modélisée à l'aide de règles contextuelles. L'ensemble des valeurs cibles et de règles de transition représente alors les paramètres de commande d'un synthétiseur [1]. Le principal but de cette méthode est d'assurer la modélisation des transitions qui existent entre phonèmes par le biais des règles qui aboutissent à un ensemble de paramètres situés en entrée d'un modèle acoustique (Figure 2.11).

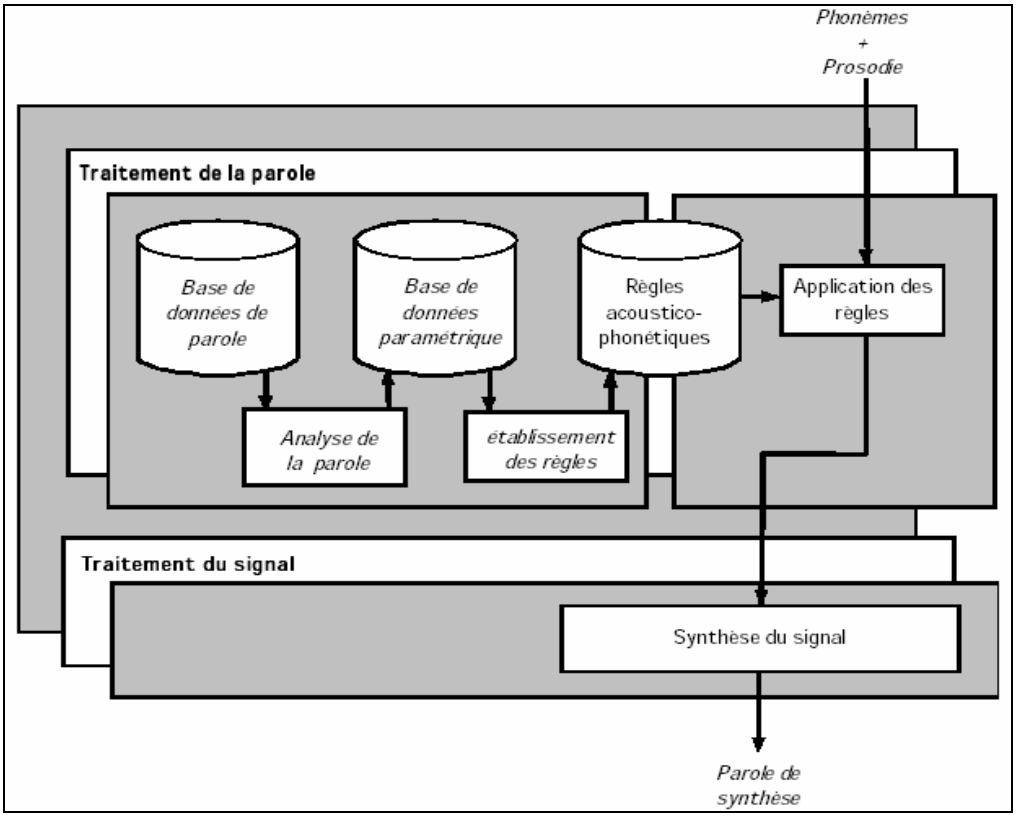


Figure 2.11 : Schéma de conception et fonctionnement typique d'un système de synthèse par règles [13].

2.11.3. La synthèse à partir d'unités sélectionnées

De nos jours, la synthèse à partir d'unités sélectionnées où synthèse par sélection dynamique d'unités non uniformes, est considérée comme la méthode la plus efficace pour générer une voix intelligible. Cette synthèse concatène des unités de taille variable. Ces dernières sont obtenues par une opération de sélection dans un grand corpus sonore. A chaque fois qu'on augmente la taille de l'unité sonore la synthèse sera meilleure c'est-à-dire que le phénomène de la coarticulation est réduit mais cette approche a l'inconvénient de la combinatoire tel que par exemple :

En Arabe Standard si on utilise les diphones comme des unités sonores on doit stocker 1188 unités, et si on se base sur les di-syllabes on doit stocker 8832 unités, tandis que si on utilise les mots on doit stocker tout le vocabulaire de la langue Arabe.

Cette méthode utilise une large base de données hétérogènes de sons choisis et des caractéristiques classées et segmentées suivant des paramètres estimés sur le signal sonore. Le segment qui ressemble le mieux - au sens d'un critère donné - au résultat désiré est trouvé par des méthodes efficaces de recherche et d'extraction utilisées par l'algorithme de sélection d'unités. Pour répondre aux exigences concernant les paramètres de synthèse nécessaires, le segment sonore trouvé est transformé par des techniques temporelles ou fréquentielles de re-synthèse telles que la re-synthèse additive, filtrage, etc. [30].

Actuellement la technique la plus utilisée dans les systèmes de synthèse à partir du texte c'est la synthèse par concaténation; mais il y a une autre technique de traitement de signal qui est appliquée souvent au signal de parole généré, c'est la méthode PSOLA (Pitch Synchronious Overlap and Add). Cette dernière se base sur la modification des paramètres prosodiques du son ce qui donne naissance à deux autres techniques : la première TD PSOLA (Time Domain PSOLA) qui consiste à modifier le paramètre prosodique Temps (Figure 2.12) et la seconde FD PSOLA (Frequency Domain PSOLA) qui modifie la fréquence.

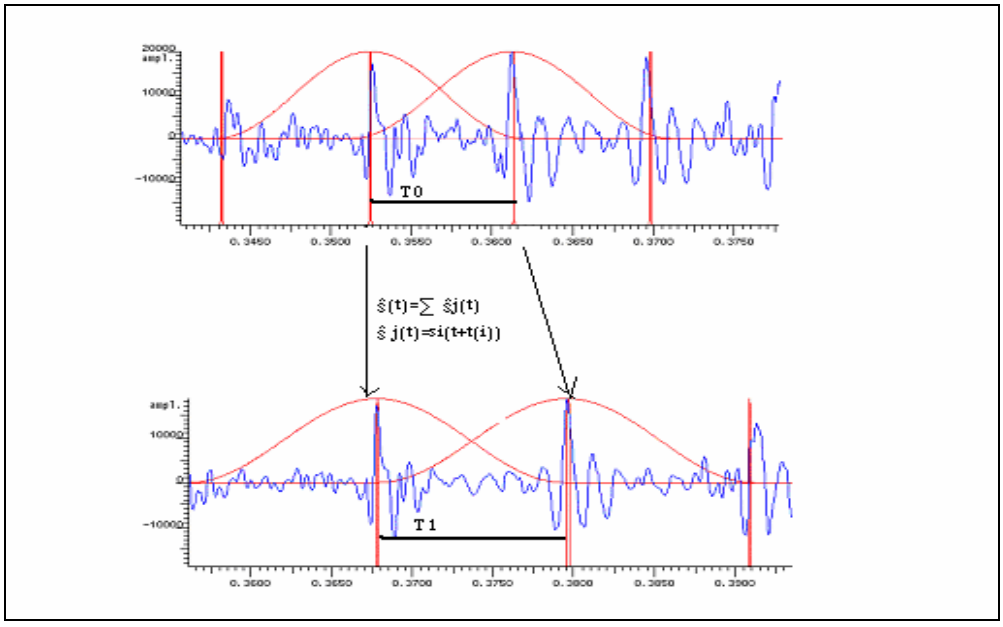


Figure 2.12 : Principe de fonctionnement de la technique TD PSOLA (superposer ou ajouter des segments dans le paramètre durée).

2.12. Quelques critères d'évaluation des systèmes TTS

Les systèmes d'évaluation des synthétiseurs TTS représentent un domaine de recherche très vivant qui évolue côte à côte avec la synthèse vocale. En effet, si l'on savait évaluer précisément et diagnostiquer les défauts de qualité des synthétiseurs, on saurait aussi comment y remédier, ou au moins comment chercher les solutions [11]. Ces systèmes se basent sur les critères suivants :

- La qualité de la parole générée, parmi les systèmes de synthèse qui utilise ce critère on peut citer TD-PSOLA (brevet CNET en 1988 déposé par France Télécom) qui est souvent applicable à des systèmes de synthèse par concaténations ;

- L'enregistrement ou synthèse à partir du texte ?
  Malgré ces progrès manifestes, le naturel de la parole de synthèse reste encore aujourd'hui nettement inférieur à celui de la parole des êtres humains. Cet écart de qualité n'est accepté par les usagers que pour des services nouveaux, qui ne pourraient pas leur être fournis d'une autre manière (lecture de FAX ou des E-mails, par exemple). Pour réduire cet écart de qualité, certains prototypes récents d'application combinent la souplesse de la synthèse à partir du texte pour produire les parties variables des messages avec l'utilisation de patrons prosodiques naturels, spécifiques aux messages de l'application [31] ;

- L'intelligibilité est un facteur crucial qui permet de vérifier si la phrase générée a été bien perçue, par rapport à son niveau linguistique (phrase affirmative, négative, interrogative, etc.) ;

- La fiabilité, de nos jours les systèmes de synthèse vocale sont utilisés dans des services grand public. Il est clair qu'ils doivent être robustes pour assurer une très grande durée de vie, et une meilleure publicité de synthétiseur vocal lui-même ;

- L' Interface Homme machine (l'interactivité), un système de synthèse de bonne qualité doit assurer une meilleure interaction entre l'utilisateur et le système (la machine de synthèse). Cette nouvelle notion intègre la reconnaissance de parole, la communication intelligente et la synthèse vocale sophistiquée qui utilise des fondements de langage parlé, connus a priori.

2.13. Les avantages et les inconvénients des systèmes TTS

Cet ensemble peut être résumé par le Tableau 2.3 :

Avantages	**inconvénients**
Cette technique est utilisée pour une taille illimitée de vocabulaire	La parole synthétique est un peu loin de la parole naturelle
Elle est souvent employée dans le domaine des handicapés visuels à cause de sa rapidité que la lecture en Braille	Elle produit pour quelques textes des erreurs de prononciations et cela surtout de certains mots étrangers
Elle ne pose aucune contrainte pour les textes à lire	Une bonne qualité de voix nécessite une bonne segmentation
Elle assure une qualité acceptable des phrases synthétisées.	La difficulté d'élaborer l'ensemble des règles phonologiques et les modèles de grammaire qui sont utilisés dans la phase de transcription.
La synthèse à partir texte est une méthode flexible puisqu'elle peut être intégrée dans n'importe quel système interactif oú la voix est un moyen de communication avec l'utilisateur	Le groupe des mots d'exceptions qui est utilisé en transcription par lexique n'est pas complet (non finalisé) tel qu'à tout moment on peut insérer de nouvelles exceptions.
Elle est évolutive et moins coûteuse	
Elle est plus rapide qu'un message écrit.	

Tableau 2.3 : Avantages et inconvénients des systèmes TTS

2.14. Conclusion

Actuellement la synthèse vocale représente un domaine très ouvert pour la recherche. Celle-ci est orientée pratiquement vers l'amélioration de la qualité des synthétiseurs qui existent.

Il faut savoir qu'en réalité la technologie de la synthèse vocale à partir du texte (TTS) n'est pas un concept voué à rester sur les tablettes des laboratoires de recherche puisque leurs applications existent déjà. Dans le chapitre suivant nous allons essayer de détailler une partie fondamentale de ces synthétiseurs, c'est bien la phase de Transcription Orthographique Phonétique appliquée sur un texte en Arabe Standard.

CHAPITRE 3 :
TRANSCRIPTION ORTHOGRAPHIQUE PHONETIQUE D'UN TEXTE EN ARABE STANDARD

3.1. Introduction

Ce troisième chapitre représente une étude plus approfondie des relations entre flexions orthographiques et phonétiques à travers le passage de la modalité graphique vers la modalité phonétique. Dans notre étude ce passage est appliqué sur la langue Arabe Standard.

3.2. Définition de la Transcription Orthographique Phonétique (TOP)

La Transcription Orthographique Phonétique ou phonétisation est une étape clé dans tout système de synthèse de la parole à partir du texte. Elle correspond au passage d'un texte écrit vers un texte lu. Elle fournit la prononciation associée au texte qu'on veut entendre.

La TOP est une tâche complexe, puisque les conventions adoptées lors de cette opération représentent souvent un compromis entre des choix théoriques et pratiques et cela pour traiter l'ensemble des exceptions et pour élaborer les règles de transcription.

3.3. Les ressources utilisées en TOP

La TOP est une étape qui peut être assurée sans se préoccuper du sens, ni de la signification. Cette étape se base sur l'utilisation de deux ressources qui sont la phonologie et la phonétique.

3.3.1. La phonologie

La phonologie est une science qui étudie les sons du langage du point de vue de leur fonction dans le système de la communication linguistique [32]. Cette science se base sur l'étude de l'écriture.

Exemple : la différence fonctionnelle (dans l'écriture et dans le sens) entre les deux mots :

نحل [nahlun] qui signifie abeilles

نخل [naxlun] qui signifie palmiers.

En Arabe, la graphie (phonologie ou aussi l'écriture) des lettres est différente selon leur position dans le mot. Par exemple la lettre / ع / [£] est transcrite /عَادَ / [£aada] (il est revenu) en début de mot, /لَعِبَ / [la£iba] (il a joué) en milieu de mot, /مَعَ / [ma£a] (avec) en fin de mot, et isolé en fin de mot / وَدَّعَ / [wadda£a] (il a quitté). Il résulte 78 formes graphiques à partir des 28 lettres arabes. Par ailleurs, la distinction minuscule/majuscule n'existe pas [1].

3.3.2. La phonétique

La phonétique est une science qui étudie les sons du langage dans leur réalisation concrète, indépendamment de leur fonction linguistique [32]. Quand on fait de la phonétique on doit laisser à côté l'écriture de la langue, car ce n'est pas la forme orthographique qui influe sur la prononciation, mais plutôt le contraire.

Exemple de l'assimilation des sons :

أنبئهم [a**n**biehum] ⟶ أمبئهم [a**m**biehum] n ⟶ m qui signifie informer

En phonétique on peut distinguer entre trois classes (Figure 3.1) :

- La phonétique articulatoire considère le phonème comme étant le résultat de l'activation des cordes vocales, de la cavité buccale, de la

mâchoire et des lèvres, etc. donc elle étudie la configuration des organes phonatoires lors de la production d'un son ;

Exemple : Selon ce type on distingue deux classes des sons :

- Les voyelles [a, i, u, y, e, o, ɛ, ə,...] qui correspondent à un passage libre de l'air, et un conduit vocal suffisamment ouvert ;
- Les consonnes [p, t, b, g, m,…] qui sont caractérisées par une fermeture momentanée ou complète du conduit vocal, et par un passage de l'air obstrué dans des endroits bien précis. Ces derniers sont appelés lieux d'articulation.

o La phonétique acoustique, dans cette classe le phonème correspond à l'onde résultante de la vibration des cordes vocales ;
Cette approche se base sur l'étude de la nature physique des sons (la durée, l'intensité, la fréquence).

Exemple : selon cette classe, on peut établir deux catégories de sons :

- Voisés [b, d, g, v, z,…] : présence de la fréquence fondamentale ;
- Non voisés [p, t, k, f, s, ʃ] : absence de la fréquence fondamentale.

o La phonétique auditive, étudie la manière dont un son est perçu par l'oreille humaine.

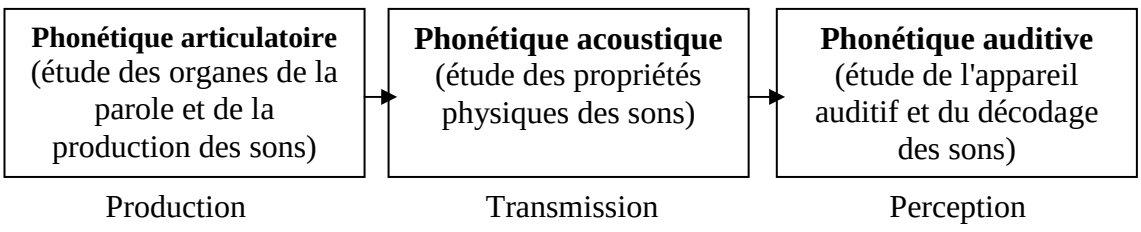


Figure 3.1 : Les principales branches de la phonétique.

3.4. L’étiquetage

L’étiquetage est l’opération qui consiste à choisir e meilleur code pour bien représenter la partie phonétique résultante de la transcription. Il existe deux types d’étiquetage phonémique et prosodique :

- L’étiquetage phonémique se base sur le niveau phonétique des sons de la parole, pour cela il utilise l’Alphabet Phonétique Internationale (API, 1982). Il faut savoir que écrire avec l'alphabet phonétique n'est jamais une fin en soi, mais un moyen pour indiquer les faits phoniques pertinents associés à un son ;

- L’étiquetage prosodique, bien que des efforts d’harmonisation aient été accomplis dans cette voie, il n'existe à ce jour aucun standard pour l’étiquetage de la prosodie, comme il en existe un pour l’étiquetage phonémique, sous forme de l'API.

  La dernière version de l’API propose d'inclure des symboles destinés à la notation des faits prosodiques (API, 1989). Cependant, ces nouvelles propositions n'ont pas reçu un accord unanime et n'ont pas fait l'objet d'une évaluation systématique (Bruce, 1989). Il n'en demeure pas moins que la nécessité de disposer d'un système de notation des sons, donc de la prosodie se fait de plus en plus pressante, à la fois pour les linguistes et les spécialistes du Traitement Automatique de la Parole. L’étiquetage prosodique soulève de multiples problèmes qui concernent plus particulièrement [33] :

  - Le choix des événements à étiqueter le support servant de référence à l'étiquetage (acoustique ou impression auditive) ;
  - L’inventaire des symboles de notation proprement dits.

3.5. Exemples de Transcription

Nous illustrons la transcription à l’aide d’exemples : le signal et la représentation phonétique résultante de la transcription de mot Anglais [phonetician] (Figure 3.2).

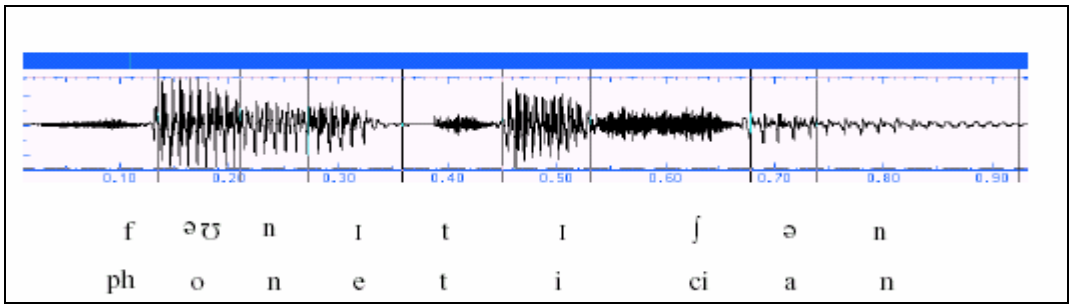


Figure 3.2 : Signal de parole et phonèmes (mot Anglais *phonetician*) [6]

La liste ci-dessus représente d'autres exemples concrets de transcription en API des phrases écrites en Français :

- C'est une question de style
  [sɛtynə kɛstjõ də stil]

- Le chef de gare ne trouve plus sa casquette
  [lə ʃɛf də gar nə truvə ply sa kaskɛtə]

- Notation
  [notasjõ]

- Exemple de transcription
  [ɛgzãmplə də trãskripsjõ]

- kasra est une voyelle arabe
  [kasra ɛtynə vwajɛl arabə] .

3.6. Les approches de la TOP

La TOP peut se faire grâce à l'utilisation de lexiques et/ou de règles de réécriture [1]. Cette possibilité donne naissance à deux grandes façons ou approches pour transformer un texte orthographique en un texte phonétique : la TOP à base de lexique et à base de règles.

3.6.1. L'utilisation de la lexique

Dans ce cas on doit attribuer pour chaque mot en entrée la prononciation qui lui correspond sans se préoccuper de son contexte.

La rapidité, la souplesse et la simplicité représentent les principaux avantages de cette approche.

3.6.2. L'utilisation de règles

Dans cette approche chaque graphème est converti en phonème selon le contexte et cela grâce à l'utilisation d'un ensemble de règles de réécriture. Ces dernières sont bien détaillées dans les paragraphes suivants.

Le principal avantage de cette approche à base de règles réside dans la possibilité de modéliser les connaissances linguistiques des êtres humains par un ensemble de règles qui peuvent être intégrées dans des systèmes experts. *Flex* et *COMPOST* (qui est appliqué à la langue française et qui est développé à l'ICP-INP Grenoble : France) sont les langages de programmation par règles de réécriture les plus utilisées. Par exemple pour *COMPOST* chaque règle est représentée comme suit :

C $\longrightarrow$ R /G+D

C'est-à-dire transformer le caractère (le graphème) G en phonème R, s'il a comme contexte gauche G et comme contexte droit D.

L'outil de base pour ce type de transcription par règles qui assurent les relations existant entre le code orthographique et le code phonétique est *la grammaire contextuelle*. Chaque règle de cette grammaire a la forme suivante :

[Phonème] = {CG (Contexte Gauche)} + {C (Caractère)} + {CD (Contexte Droit)}

Le problème de cette approche réside dans la détermination et la gestion de nombre énorme de règles de transcription.

3.7. La TOP de la parole spontanée

C’est un autre type de transcription. Il est souvent utilisé pour assurer quelques objectifs d'analyse scientifique comme la possibilité de transcrire des conversations, des réunions de travail, des débats, des discours, etc. et cela, en un temps réel. Dans ce type de transcription il faut faire attention aux différents aspects spécifiques de l'oral. Ces derniers peuvent être représenté :

- Pour les réalisations segmentales elles-mêmes, le transcripteur a le choix entre une transcription en API et en Orthographe Standard. Le premier type de transcription peut être plus précis et exact que le second mais, il présente quelques inconvénients puisqu’il représente une tâche plus longue à écrire et à lire que l’utilisation de l’Orthographe Standard ;

  En revanche une transcription de second type pose aussi quelques problèmes et doit parfois être précisée tel qu’on doit considérer les segments qui n’ont pas été réalisés lors de la prononciation orale. Par exemple :

  [ɛfo pa ɛt dificil] $_{oral}$ ⟶ / Il ne faut pas être difficile / $_{écrit}$

- Quelques phrases orales peuvent faire l'objet d'un marquage spécifique comme :
  - Les séquences incompréhensibles ;
  - Les hésitations, les reprises, les corrections, etc. par exemple la phrase : un logiciel qui soit plus [plys] plus [plys] plus [ply] convivial ;
  - La longueur variable des pauses ;
  - Les chevauchements qui existent dans les dialogues entre les différents locuteurs.

Le but de notre travail est de réaliser un outil en vue de la synthèse vocale des textes Arabes, pour cela nous devons exploiter cette Transcription Orthographique Phonétique sur la langue Arabe Standard.

3.8. La TOP des textes en Arabe Standard

Avant d'effectuer n'importe quel traitement (Transcription Orthographique Phonétique, analyse, synthèse, ou reconnaissance) sur n'importe quelle langue, il faut toujours passer par l'étude de la phonétique de cette langue. Cette dernière nous a permis de tirer les principales caractéristiques associées aux différents phonèmes.

3.8.1. Caractéristiques phonétiques de l'AS

Le système phonétique de l'Arabe fait partie de la classe des langues plurielles; sa pluralité est due grâce aux différents dialectes parlés dans le monde arabe lui-même. La particularité de l'Arabe réside dans son écriture qui va de droite vers la gauche, de plus de la présence des caractéristiques phonétiques suivantes (Figure 3.4) :

- La vocalisation, en Arabe Standard les textes sont dépourvus des signes dits de vocalisation [§akl] ou signes diacritiques. Ces derniers assurent le passage d'une consonne [harf] à une autre, et ils jouent un rôle important dans le sens du texte.

  Exemple :
  كَتَبَ مُحَمَدُ الدرسَ [kataba muhammadu addarsa] qui signifie Mohamed a écrit le cours

  كُتُبُ مُحَمَّدٍ [kutubu muhammadin] qui signifie les livres de Mohamed

- [almad], les voyelles longues sont caractérisées par une partie stable plus allongée que la partie stable des voyelles courtes ou brèves et cela sur le plan acoustique (Figure 3.3).

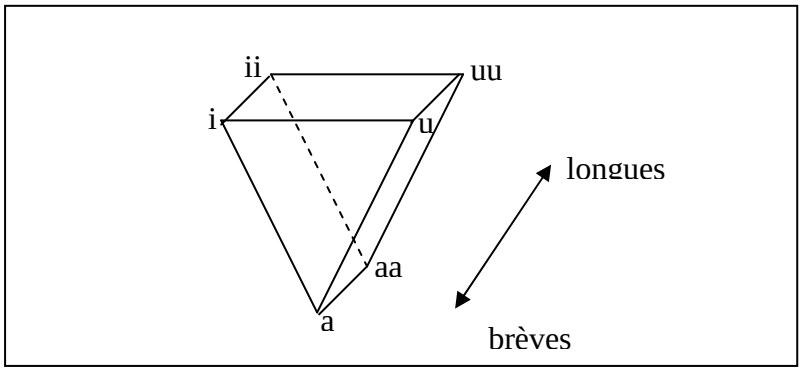


Figure 3.3 : Système vocalique de l'Arabe [21].

Sur le plan articulatoire la spécificité de [almad] réside dans la similitude qui existe entre les voyelles longues et quelques voyelles françaises. Ce phénomène est souvent réalisé dans la cas de /تفخيم/ [tafxiim] ou la présence des consonnes emphatiques reportent en arrière le point d'articulation des voyelles [a, aa, u, uu, i, ii] de sorte qu'elles deviennent /a, o, e/ [21].

En ce qui concerne la sémantique, / ألمد / [almad] peut changer totalement le sens des mots, exemples :

جَمَل [zam**a**l] qui signifie chameau

جَمَال [zam**aa**l] qui signifie beauté.

- La gémination « ou dédoublement de consonnes, ou [a§§addatu] en Arabe » est l'opération qui consiste à répéter une consonne. Elle sert à renforcer l'intensité de la consonne géminée. On peut dire aussi que c'est la succession de deux consonnes identiques. Elle apparaît souvent lors de [Edgaam]. Ce dernier correspond au phénomène de l'assimilation de deux sons, ou le premier son est masqué tandis que le second est doublé.

  Exemple de Edgaam

  قُلْ لَهُ [qul lahu] ⟺ قُلَّهُ [qu**ll**ahu]

une gémination succédant toujours une voyelle. Elle permet la différenciation entre deux mots, l'exemple de حَمَامّ [hamaamun] qui signifie pigeon et حَمَّامّ [hammaamun] qui signifie bain. Elle peut être placée sur toutes les consonnes exceptionnellement sur la consonne glottale hamza, exemple :

وَدّعَ ⟶ [waddaɛa]

- Le signe de [tanwiin] est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination (ال) placé en début de mot. Les symboles de [tanwiin] sont au nombre de trois et sont constitués par un dédoublement des signes diacritiques ci-dessus, ce qui se traduit par l'ajout du phonème [n] au niveau phonétique [1].

  [an] : signe ً (بًا [ban])

  [un] : signe ٌ (بٌ [bun])

  [in] : signe ٍ (بٍ [bin])

- L'emphase, l'Arabe c'est la langue dans laquelle est écrit le saint Coran. Ce dernier utilise des mots dits de [zalala], où on doit utiliser des phonèmes emphatisés, exemple :

  [الله] qui signifie le Bon Dieu, on prononce ce mot [allah] et non pas [alleh]. Dans cet exemple le « ل » représente le [harf] qui est emphatisé.

  Les phonèmes emphatiques sont caractérisés par une tonalité plus pleine et grave car ils exigent la dépense d'un volume d'air important et une tension organique supérieure par rapport aux autres consonnes. L'intérêt portée par ce phénomène remonte jusqu'aux Grammairiens Arabes du moyen âge attirés par le système phonétique de leur langue [11].

  Il faut savoir que la notion d'emphase est applicable sur les consonnes ainsi que sur les voyelles telles que :

- Une consonne est emphatique si et seulement si elle appartient à l'ensemble /ظ, ص, ط, ض / et elle est emphatisée selon le contexte si elle appartient à /ل , ق /
- Une voyelle est emphatisée si et seulement si elle est voisine d'une consonne emphatique.

o Les phonèmes arrières [34]

Le système phonétique de l'Arabe Standard possède quatre phonèmes arrières spécifiques à cette langue et ils n'ont pas leurs équivalents exacts dans aucune autre langue européenne :

- Les spirantes pharyngales /ح/ [h, ε] qui ont comme point d'articulation la partie médiane du pharynx ;
- L'occlusion uvulaire /ق / [q] qui a pour point d'articulation la partie la plus reculée de la langue et la région du palais supérieur ;
- L'occlusion glottale / ء / [hamza], les Grammairiens Arabes indiquent pour ce phonème la partie la plus reculée du pharynx.

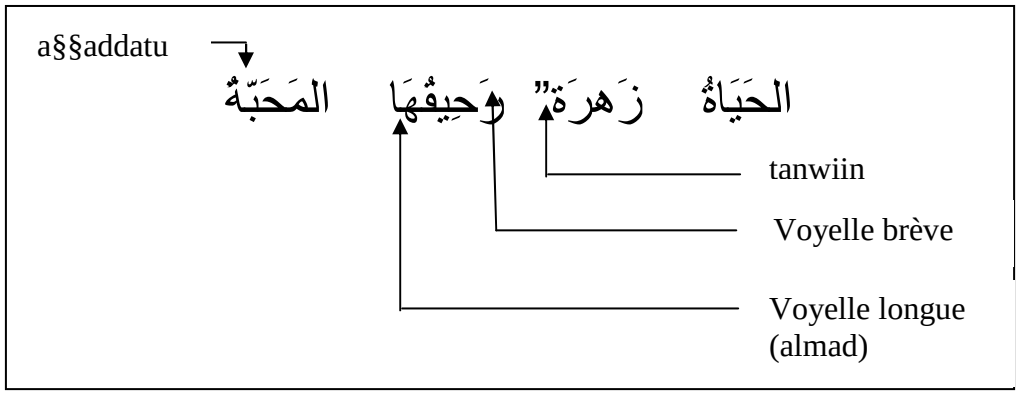


Figure 3.4 : Exemples de quelques caractéristiques de l'AS.

3.8.2. Les problèmes de la TOP des textes en Arabe Standard

Une langue naturelle comporte souvent des ambiguïtés. En ce qui concerne la langue Arabe, peu de recherches ont porté sur la Transcription

Orthographique Phonétique et cela à cause de la complexité liée à la langue elle même ainsi qu'aux problèmes suivants :

- Des graphèmes qui ont plusieurs réalisations phonétiques, le /و / [w] de «بوم » [buum] et «موز » [mawz] correspondent à deux sons différents ayant la même graphie [28].

- Plusieurs graphèmes qui ont la même réalisation phonétique, c'est la cas de /ة / et /ت / qui ont le même phonème [t] par exemple «حياة » [hayaat] et « جنات» [zannaat].
  On peut citer aussi Le /ن / [n] dans «يَصلون » [yu$alluuna] et «انزلق » [EinZalaqa]

- Des graphèmes qui ne sont pas pris en compte. Le /ا / [alif] dans « قاموا » ne correspond pas à un son (c'est un silence) [28]. Même le /ا / [alif] dans « غدًا » correspond aussi à un silence.

- Une absence totale de correspondance graphème - phonème [28] :
  - « هذا » devrait être → «هاذا ».
  - « هؤلاء » → « هاؤلاء ».
  - « ذلك » → « ذالك ».

Il faut savoir aussi que mêmes les caractéristiques intrinsèques de l'Arabe imposent aussi des difficultés lors de l'étape de transcription, c'est le cas de la gémination [a§§addatu] par exemple ou chaque graphème géminé est transcrit avec un dédoublement de son qui lui correspond.

3.8.3. Un échantillon de règles de transcription graphème/phonème

Voici un exemple concret de règles de transcription [28] :

# : signe de début de phrase,	C : **C**onsonne,
$: signe de fin de mot ou de phrase,	V : **V**oyelle,
CS : **C**onsonne **S**olaire	CL : **C**onsonne **L**unaire

- [uu] = {CS}+{Damma}+{w} [uu] = $ +{Damma}+{w}
- [uu] = {CL}+{Damma}+{w}
  - lorsque le **/** و / [w] est précédé par la voyelle / ُ / [Damma] et suivi par une consonne, on obtient la voyelle longue [uu].
  - lorsque le /و / est précédé par la voyelle / ُ / [Damma] en position finale, on obtient la voyelle longue [uu].

  Exemple : دون [duuna] ; لبسوا [labisuuE]

- [aa] = {fatha}+{E}
  - lorsque le / ا **/** [alif] est précédé par la voyelle / َ **/** [fatha] on obtient la voyelle longue [aa] quel que soit ce qui suit. Exemple: لما [lammaa]

- [ii] = {CS}+ {kasra}+{y} [ii] = $+ {kasra}+{y}
  [ii] = {CL}+ {kasra}+{y}
  - lorsque le / ي / [y] est précédé par la voyelle / ِ **/** [kasra] et suivi par une consonne, on obtient le phonème de la voyelle longue [ii] ;
  - lorsque le / ي / est précédé par la voyelle / ِ / [kasra] en position finale de mot, on obtient la voyelle longue [ii]. Exemple : قليلا [qaliilanE] ; لمسني [lamasanii] [28].

- [CC] = {C} + {§adda }
  - Lorsque une consonne est suivie par la voyelle / ّ / [§adda], elle est doublée, on obtient alors les phonèmes [CC]. Exemple : ودّ [wadda] ;

- [C] = {C}+{ْ}

Lorsqu'une consonne C est suivie par / ْ / [sukuun], elle reste inchangée, on obtient alors le phonème [C]. Exemple يرغب [yargabu] ;

- [CS] + [E] = # + { al } + {CS}
  [CL] + [l] + [CL] = {CL} + { al } + {CL}

  - Lorsque le / أل / [al] est en début de phrase et suivi par une consonne solaire, il est équivalent à la non présence du /ل/ [l] ;
  - Lorsque le / أل / est entre deux consonnes lunaires, il est équivalent à l'absence du / أ /.

  Exemple : السميع [assamii£] ; منع الاكل [mana£a alakla] ;

- Si le premier caractère de la phrase est un / ا / « الالف الساكنة » il sera remplacé par /ء/ car s'il est à l'intérieur de la phrase, il n'est pas prononcé [29] ;

- S'il n'y a pas de voyelles à la fin de la phrase, celle-ci sera terminée par / ْ / /السكون/ [assukuun] [29] ;

- Les règles de caractère / مْ / [m] « أحكام الميم الساكنة »
  il n' y a qu'une séquence à détecter : / ب ْ م / elle est remplacée par / ب ْ مّ / tel que / مّ / est la consonne / م / affectée d'une nasalisation [29] ;

  Exemple : جمبري [zambarii] ;

- Les règles de caractère / ر / [r] « احكام الراء »
  ce caractère a deux prononciations différentes, ce qui entraîne deux phonèmes différents le / ر / emphatisé /مفخم/ [mufaxxam] est représenté par / R / [29] ;

  Exemple : الرَّحمَان [arrahmaan] ;

- La règle de renversement « القلب » ou l'assimilation de la consonne / ْن /. La séquence / **ب** ْ **ن** / est remplacée par / **ب** ْ **مْ** / [29] ;

  Exemple : المنبع [alma**n**ba£] ⟶ المبع [alma**m**ba£] ;

- Les règles de collage « ادغام » de la consonne / ْ**ن** / :

  [ر] + [ر] = {ر} + {ْ} + {ن}
  [ل] + [ل] = {ل} + {ْ} + {ن}
  [م] + [مْ] = {م} + {ْ} + {ن}
  [و] + [وْ] = {و} + {ْ} + {ن}
  [ى] + [ىْ] = {ى} + {ْ} + {ن}

- La règle de masquage « إخفاء » de la consonne / ْ **ن** / :

  { C }+{ْ} **+** { **نْ** } = {C}+ {ْ}+ {**ن**}
  avec C € { ص د ث ئ ج ش ق س د ط ز ف ت ض ظ ن } ;

  Exemple من كان [man kaana] ⟶ مكان [makaana] ;

- La règle de [attanwin] « التنوين »

  [ن] + [a] + [C] = $ +{ً}+ { C } [ن] + [a] + [C] = § + {ً}+ { C }
  [ن] + [u] + [C] = $+{ٌ}+ { C } [ن] + [u] + [C] = § + {ٌ}+{ C }
  [ن] + [i] + [C] = $+{ٍ}+{ C } [ن] + [i] + [C] = §+ { ٍ }+{ C }.

3.9. Quelques travaux antérieurs en TOP de l'Arabe Standard

A l'heure actuelle peu de travaux ont été développés dans la branche de Transcripteur Orthographique Phonétique des textes arabes. Et si ces quelques travaux existent, ils se basent sur un principe identique aux transcripteurs des autres langues (Français, Anglais, etc.), par exemple :

- Le système de synthèse à partir du texte MBROLA qui utilise le code SAMPA durant la phase de transcription, dans ce cas l'utilisateur doit respecter la forme de code SAMPA qui n'est pas un code international [2] ;

- Le travail de S. BALOUL qui représente un exemple concret de transcription des mots; en se basant sur l'analyse morphologique, et sur les études des pauses pour générer la prononciation des textes [3].

De plus trois générations d'outils existent déjà dans l'axe de transcription en Arabe Standard, nous pouvons les résumer par :

- SYAMSA (SYstème d'Analyse Morphosyntaxique de l'Arabe), qui a été réalisé par SAROH. Selon lui, « la phonétisation de l'Arabe repose en particulier sur l'emploi de lexiques et d'un analyseur morphologique pour la génération des différentes formes d'un mot. Par ailleurs, ce sont les phénomènes d'interaction entre les mots (liaison, élision, etc.) et les phénomènes d'assimilation qui suggèrent l'utilisation de règles phonologiques » [3]. Cet outil assure pour chaque mot en entrée, la racine qui lui correspond, ainsi que les représentations morphologiques et phonétiques de cette dernière.

  Par conséquent, l'opération de transcription dans ce logiciel n'est d'autre qu'une comparaison de mots en entrée avec ceux issus de l'analyse morphologique. S'il y a une correspondance entre le mot en entrée et un des mots résultants de l'analyse morphologique le système fournit directement la représentation phonétique de ce dernier ;

- Le projet de GAZALI

  Le travail de S. GAZALI a été effectué à l'IRSIT (Institut Régional des Sciences Informatiques et des Télécommunications) de Tunis, c'est un travail de transcription qui s'insère dans le cadre d'une réalisation d'un système de SAT (Synthèse A partir du Texte) la particularité de ce système réside dans l'application d'un ensemble de *règles de propagation de l'emphase* ;

- SYNTHAR+

  C'est un outil étudié par Z. ZEMIRLI à l'INI (l'Institut National d'Informatique d'Alger), il assure l'étape de transcription pour un système de Synthèse à partir du texte (SAT) de telle sorte qu'il

transmette la représentation phonétique du texte au synthétiseur MULTIVOX. Il faut savoir que SYNTHAR+ se base sur une analyse morphologique avant de réaliser la transcription.

L’utilisation de la notion de code et l’introduction des niveaux élevés d’analyse (morphologiques, syntaxiques, et pragmatiques..) rend la tâche de transcription complexe, et nécessite des études approfondies dans la langue elle-même. La particularité de notre travail par rapport à tout ce qui existe réside dans la transcription à base de graphèmes, c'est-à-dire l’utilisation des caractères arabes comme des unités de base pour transcrire directement le texte. Ce qui facilite la tâche de transcription et donne des résultats très acceptables.

3.10. Conclusion

Dans ce chapitre nous avons vu en détail la phase de Transcription Orthographique Phonétique, qui représente le maillon de base de toute système de synthèse de la parole à partir du texte, et puisque notre travail est appliqué sur l'Arabe Standard nous avons mis l'accent sur les particularités phonologiques et phonétiques de cette dernière.

Nous essayerons de détailler toutes les étapes nécessaires pour réaliser un tel type de système dans le prochain chapitre qui représente la mise en œuvre d'un système de Transcription Orthographique Phonétique du texte Arabe en vue de réaliser une lecture automatique de ce texte, avec tous les problèmes et les cas particuliers rencontrés.

CHAPITRE 4
CONCEPTION ET IMPLEMENTATION DE TOP-AS

4.1. Introduction

Après passage en revue des notions de bases nécessaires, et après avoir situé notre recherche dans un cadre théorique, nous verrons la méthodologie utilisée et les résultats obtenus de notre travail qui est réalisé en deux grandes étapes : la création de la base de données sonores, et la transformation d'un texte écrit à un texte lu. Cette transformation est réalisée par :

- Une Transcription Orthographique Phonétique de n'importe quel texte écrit en Arabe Standard et cela pour le transformer en une chaîne phonétique ;
- Une génération de signal de la parole qui correspond à cette chaîne.

4.2. Spécification des besoins

Pour assurer un bon développement de notre démarche et pour obtenir une meilleure organisation de notre travail, nous avons défini les objectifs visés. Ces derniers peuvent se résumer par les points suivants :

- Notre principal objectif est de réaliser un système de Transcription Orthographique Phonétique en vue de la synthèse à partir d'une représentation textuelle en Arabe Standard ;
- Assurer une interactivité et une simplicité d'utilisation, puisque notre travail est destiné à des experts, et non experts surtout aux personnes mal voyantes qui doivent pouvoir comprendre et utiliser notre système ;

- L'utilisateur doit pouvoir comprendre les différentes phrases synthétisées qui doivent être claires et prononcées avec une qualité acceptable.

4.2.1. Les cas d'utilisation (use cases)

Dans cette phase on doit représenter le comportement et la réaction de notre outil face aux exigences et aux actions des différents acteurs (tout élément qui interagit avec notre système), dans notre cas ces derniers peuvent se résumer par *un acteur principal* qui est l'utilisateur de notre système.
Pour bien modéliser, et pour bien détailler les fonctionnalités de notre système, on a utilisé des schémas proches des notations d'UML (Unified Modeling Langage) [38].

4.2.1.1. Diagrammes de cas d'utilisation

Chaque acteur de système peut réagir d'une ou de plusieurs manières en réponse aux différentes interactions avec le système. On peut lui associer un ou plusieurs fonctionnalités (Tableau 4.1).

Acteur	**Buts utilisateur**
Utilisateur	Ouvrir le système Entrer (ouvrir, ou saisir) le texte arabe à lire Transcrire le texte Sauvegarder le résultat de la transcription Lire le texte Consulter l'API Générer acoustiquement les mots synthétiques Quitter l'application.

Tableau 4.1. Cas d'utilisations de notre système.

Il est clair que notre système de synthèse à partir texte en Arabe Standard (TOP-AS) doit comprendre trois modules :

- La création de la base des segments sonore ;
- La Transcription Orthographique Phonétique du texte ;
- La synthèse vocale de ce texte.

Cela donne naissance aux cas d'utilisation suivants :

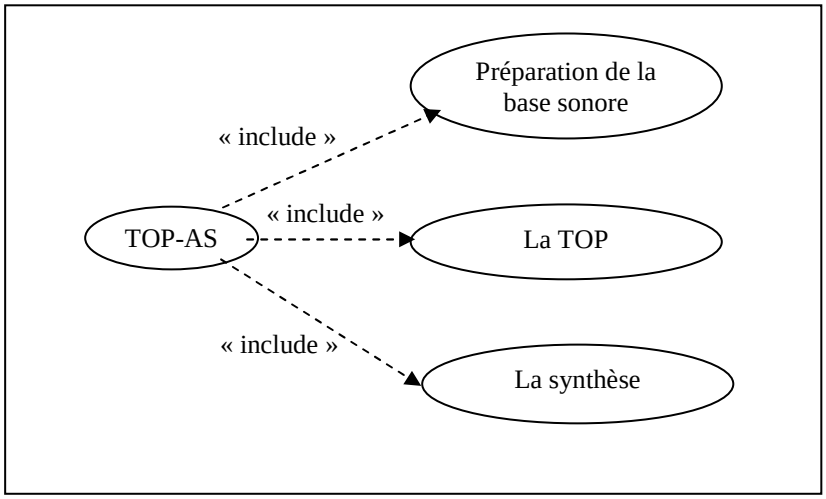


Figure 4.1 : Diagramme des cas d'utilisation principal.

- Cas d'utilisation « préparation de la base sonore »

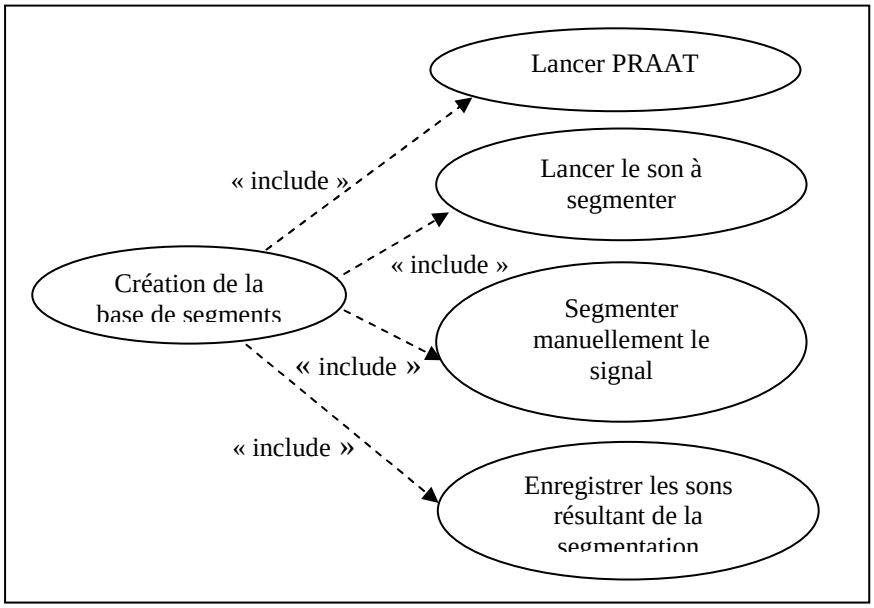


Figure 4.2 : Diagramme de use cases de cas « préparation de la base sonores »

o Cas d'utilisation « la Transcription Orthographique Phonétique »

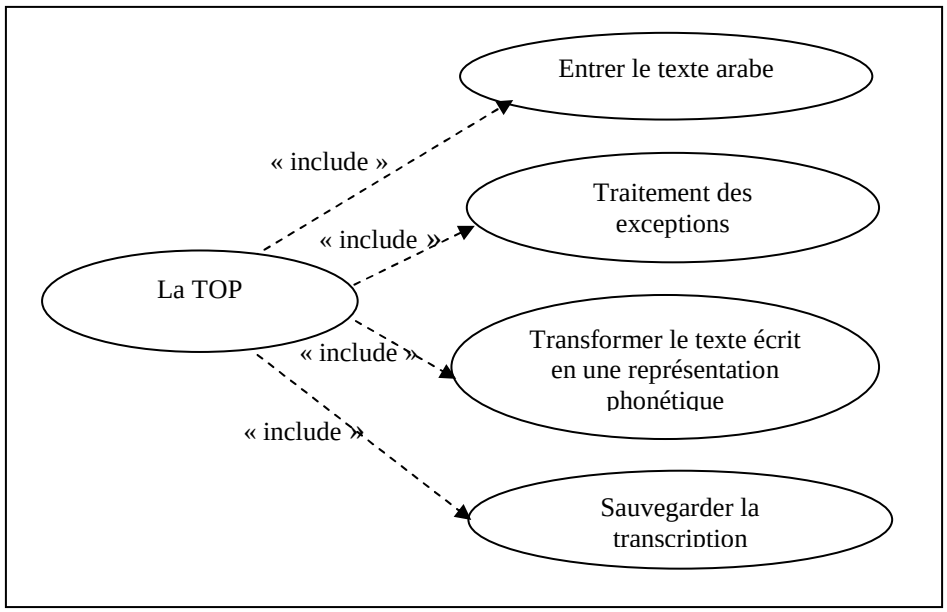


Figure 4.3 : Diagramme de use cases de cas « Transcription de texte »

o Cas d'utilisation « la synthèse»

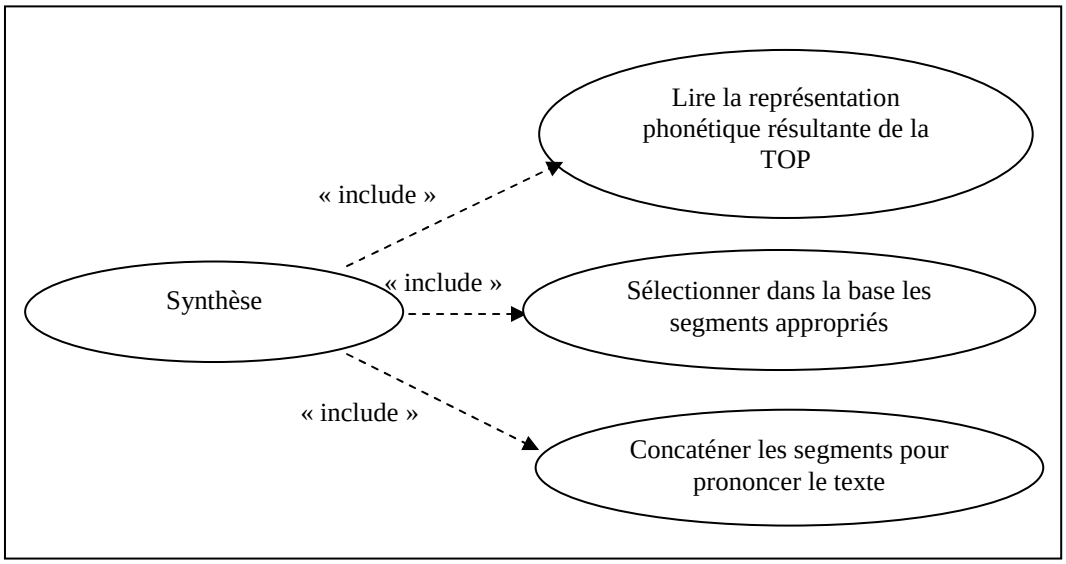


Figure 4.4 : Diagramme de use cases de cas « prononciation du texte »

4.3. Plan général de notre travail

Dans notre étude, nous avons mis l'accent sur un outil de génération de parole à partir d'un texte écrit en Arabe Standard. L'architecture générale de cette étude peut être résumée comme suit :

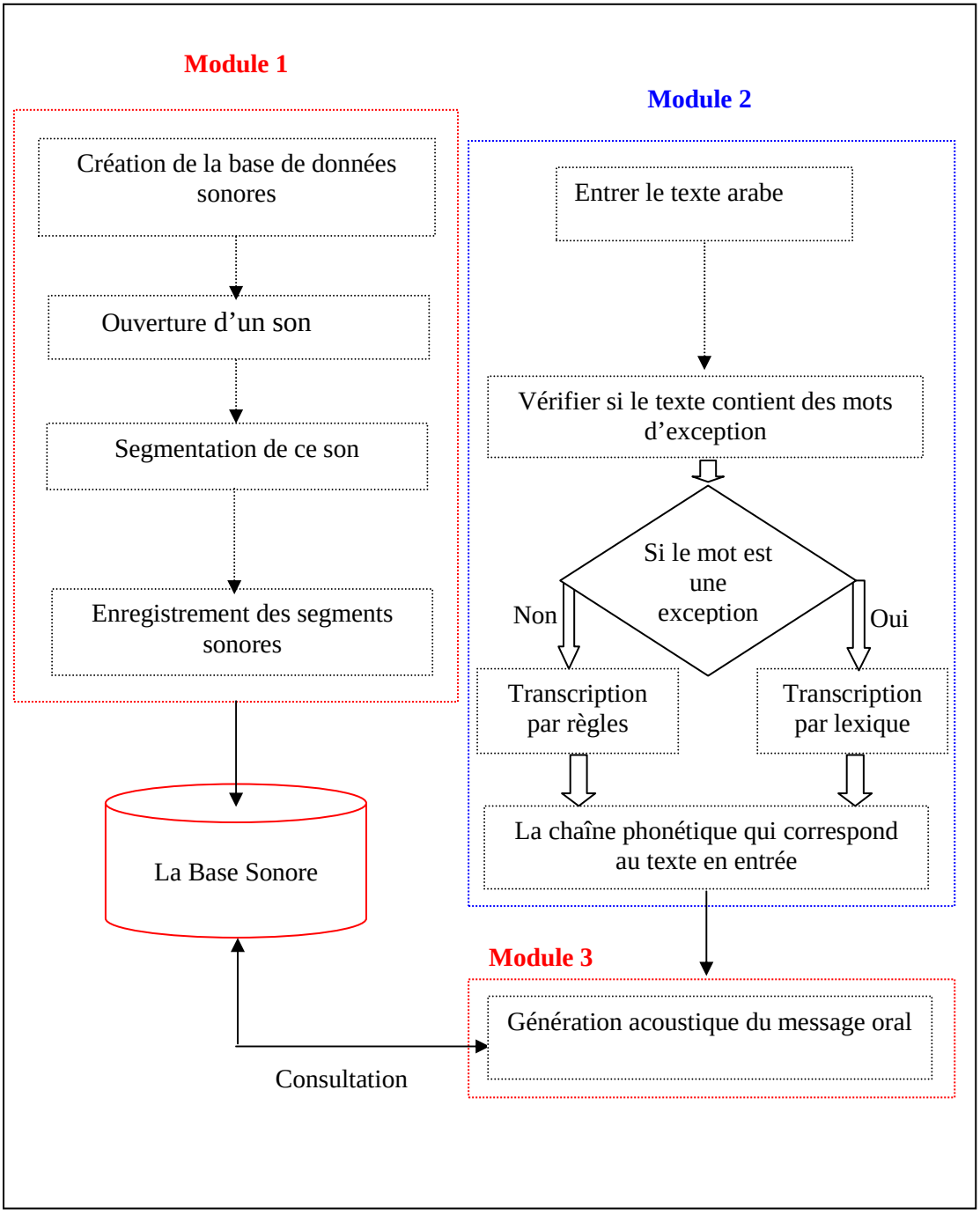


Figure 4.5 : Les trois blocs d'un système TOP-AS.

4.4. Création de Corpus

La plupart des travaux effectuer dans le domaine de la communication parlée nécessite souvent l'enregistrement, et la manipulation de corpus de parole continue, et cela pour mener à bien les études sur les effets contextuels, sur les indices acoustiques, et sur les variabilités intra et inter locuteurs.

Notre corpus à deux grandes sources sonores, la première contient des phrases en Arabe Standard qui sont bien prononcées par un locuteur jordanien avec un débit d'élocution normale ; la seconde source est une base des mots porteurs de polysons appelés aussi logatomes. Ces derniers sont prononcés par une personne de sexe féminin ayant une bonne élocution de l'Arabe. Ces deux corpus ont été enregistrés dans un milieu calme et avec un micro de bonne qualité.

Ces deux sources sont utilisées pour générer deux bases sonores : la première des phonèmes, et la seconde des diphones. Cela à travers l'application d'une procédure de segmentation.

4.4.1. Base des phonèmes

Le phonème est l'unité sonore atomique dans n'importe quelle langue, cette notion a été créée à la base de la construction des *paires minimales* c'est-à-dire la construction des paires de mots qui se diffèrent par un seul son (نحل [na**h**lun] / نمل [na**m**lun]), ce qui implique une différence au niveau du sens (sémantique).

Notre liste des phonèmes est constituée d'un ensemble des sons de base, et d'autres supplémentaires. L'ensemble des sons de base se compose des phonèmes qui correspondent aux 28 consonnes, et aux 6 voyelles (longues et courtes) de l'Arabe Standard, tandis que les sons supplémentaires sont constitués des trois sons de tanwin ([an], [un], [in]), et le silence qui est équivalent à une présence d'une ponctuation dans le texte à

lire. En plus des phonèmes précédents nous avons ajouté les sons qui correspondent aux différents mots d'exceptions.

4.4.2. Base des diphones

Pour améliorer la qualité des mots synthétisés par la méthode de concaténation de phonèmes, et pour réduire les effets de coarticulation, la solution consiste à enregistrer la transition qui existe entre phonèmes au lieu d'enregistrer le phonème, et cela parce que la transition est porteuse d'une quantité importante d'informations acoustiques par rapport au phonème lui-même. Chaque diphone varie de la partie stable d'un phonème jusqu'à la partie stable de phonème qui le suit en incluant la transition. Dans notre cas nous avons utilisé deux types de diphones :

- Le premier type se compose des diphones extraits à partir des phrases réelles et qui ont un sens, Exemple / جلس يستمع إلى الراديو / [zalasa yastami£u ilaa arraadyuu], qui peut être décomposée en diphones :

  { "debut_j.wav", "j_fatha.wav", "fatha_l.wav", "l_fatha.wav", "fatha_sine.wav", "sine_fatha.wav", "fatha_y.wav", "y_fatha.wav", "fatha_sine.wav", "sine_sekoune.wav", "soukoune_t.wav", "t_fatha.wav", "fatha_mime.wav", "mime_kasra.wav", "kasra_aine.wav", "aine_dama.wav", "dama_hamza.wav", "hamza_kasra.wav", "kasra_l.wav", "l_fatha.wav", "fatha_alif_lam_rae.wav", "rae_aa.wav", "aa_d.wav","\d_sukun.wav", "sukun_y.wav","y_oo.wav", "oo_fin.wav".**};**

  Voici un extrait de notre corpus de diphones :

بَابَا	[baabaa]	مَامَا	[maamaa]	جَلَسَ	[zalasa]
طَاولَة	[Taawilatun]	ألمَاء	[almaae]	تَمر	[tamr]
ألشَّمسُ	[a§§amsu]	ألقَمَرُ	[alqamaru]	أنِيسّ	[aniisun]
نَحلَة	[nahlatun]	أسعَدُ	[as£adu]	زَوجَين	[Zawzayni]
خَيرّ	[xayrun]	مُحَمَّدّ	[muhammadun]	مُسلِمّ	[muslimun]
ثَوَابّ	[&awaabun]	جَامِع	[zaami£]	لذِيذّ	[laɯiiɯun]

إلى [ilaa] صَوتٍ [$awtin] جَمِيلِ [zamiilin] يَستَمِعُ [yastami£u] ;

- Et le second type représente les diphones qui sont extraits à partir de l'ensemble des mots synthétiques et qui n'ont pas un sens, appelé aussi logatomes, ou mots porteurs. Chacun de ces derniers est utilisé pour extraire un seul diphone, c'est-à-dire chaque logatome contient un et un seul diphone et cela afin d'assurer une indépendance entre ce diphone et son contexte.

  Exemples des logatomes et des diphones :

  - c : représente une consonne
  - # : un silence de début ou de fin de mot
  - v : une voyelle

Logatomes	**Diphones**	**Exemples**
#katat**v#**	[v#]	#katat ـَ_**#** #katat ـُ_**#** #katat ـِ**#** #katat ا_**#** #katat **و**ُ #katat **ي**ـ**#**
# ta**cc**ata #	[cc]	# ta **b**_**b** ata #
#cata#	[#c]	**#**_**t** ata#
#kata**c#**	[c#]	#kata **f**_**#**
#a**cv**ta#	[cv]	#a **b**_ُta#
#at**vc**a#	[vc]	#at ـ**n** a#

Tableau 4.2 : Exemple des logatomes contenant des diphones [8].

4.4.3. Les étapes de la segmentation en phonèmes et en diphone

Une fois les enregistrements ont été faits, nous avons utilisé l'outil d'analyse du signal vocal *PRAAT*, en vue d'obtenir des spectrogrammes (ces

derniers assurent une représentation tridimensionnelle de signal de parole à travers trois axes : un vertical représente la fréquence du son en Hz, un autre horizontal représente l'évolution temporaire du son, et le degré de noircissement qui représente l'intensité ou l'énergie de son en dB). Afin de découper *manuellement* le signal de parole en une suite de segments associés chacun à un élément acoustique unitaire (phonème, ou diphone), puisque le processus de segmentation complètement automatique de corpus est jusqu'à présent peu concevable et peu fiable pour son utilisation dans les systèmes de synthèse par concaténation d'unités [35].

Segmenter le signal de parole, c'est effectuer une partition de ce signal en trames, telle que chacune d'entre elles possède au moins une caractéristique que les autres trames voisines n'ont pas [11]. Le but de la procédure de segmentation est d'isoler l'unité à étudier.

Exemple des différentes étapes de cette procédure :

- Visualiser le sonagramme associé au son porteur de l'unité à segmenter

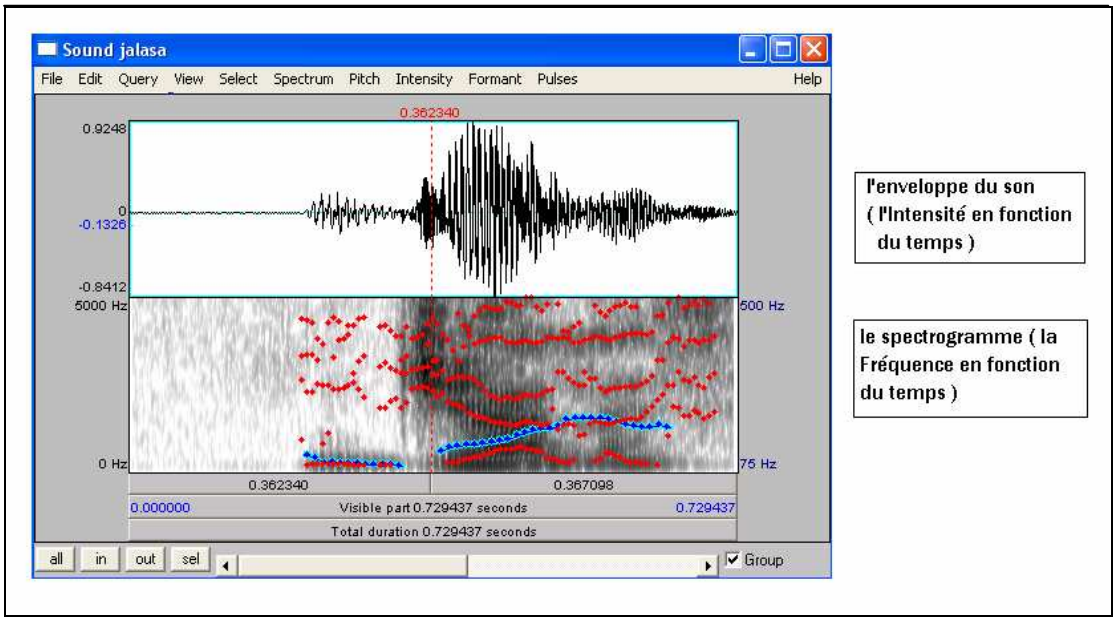


Figure 4.6 : Visualisation du son [zalasa] en entier par l'utilisation de la fenêtre *SoundEditor* de PRAAT

- Sélectionner par le curseur l'unité à extraire (phonème, diphone, ou polyson)

  Dans notre exemple nous voulons extraire le diphone [debut_z] à partir du son [#zalasa#] ;

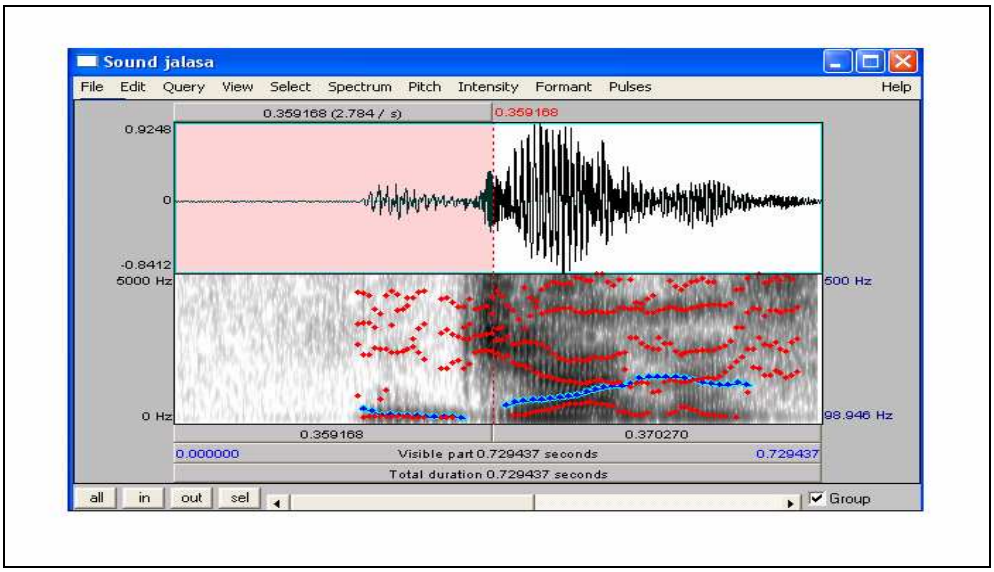


Figure 4.7. La sélection du diphone [debut_z]

- Enregistrer cette unité pour obtenir le nouveau son ; dans notre cas, c'est la création du nouveau son qui correspond au diphone [debut_z] ;

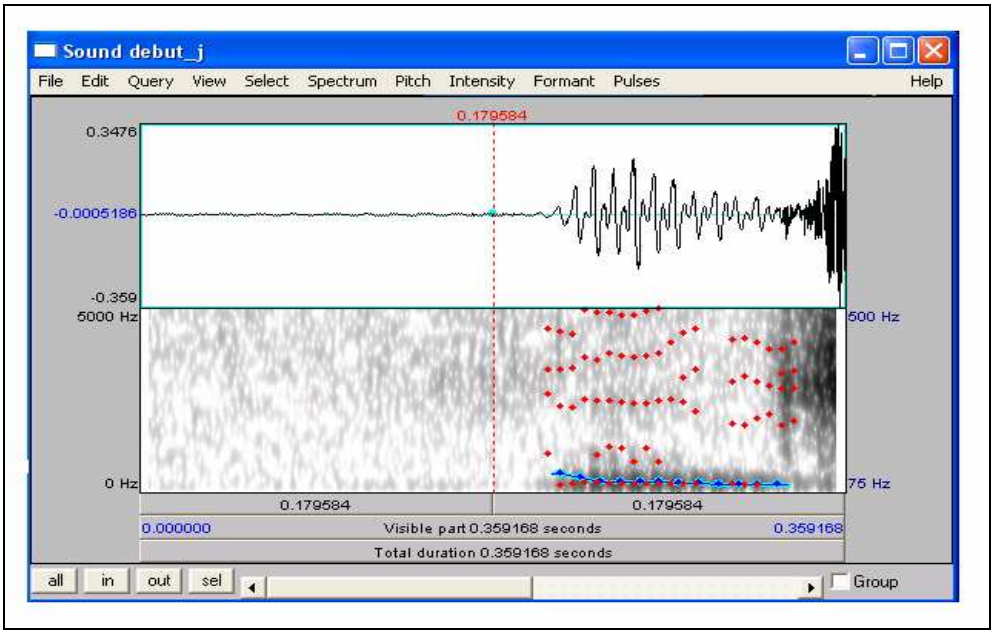


Figure 4.8. Visualisation du diphone [debut_z] qui est résultat de la segmentation par l'utilisation de la fenêtre *SoundEditor* de PRAAT.

4.5. Le passage d'un texte écrit en AS en un texte lu

Il a été montré que la majorité des erreurs de conversion graphème/phonème, pour les meilleurs systèmes opératoires provenaient des noms propres et les exceptions qu'ils posent [36].

Avant d'entamer la phase de transcription, il faut passer en premier par le module qui traite ces mots d'exceptions. Ce dernier fait appel à des connaissances linguistiques théoriques spécifiques à la langue traitée, dans notre cas c'est la langue arabe.

Les mots d'exceptions sont des mots qui ne se lisent pas suivant des règles d'écriture bien déterminées (Tableau 4.3).

Les mots d'exception	Prononciation correcte	Transcription en API
هذا	هاذا	[Haaµaa]
ذلك	ذالك	[µaalika]
كذلك	كذالك	[kaµaalika]
يأيها	ياأيها	[yaaayyuHaa]
يس	ياسين	[yaasiin]

Tableau 4.3. Quelques mots d'exceptions

Cette opération de transcription est appelée *transcription par lexique*, puisque pour chaque mot elle génère directement une entité lexicale qui représente la prononciation de mot.

Pour les mots qui n'appartiennent pas à l'ensemble des exceptions, nous leurs appliquons *une transcription par règles*. Cette dernière consiste à modéliser les connaissances linguistiques qui sont employées dans une langue par un groupe de règles de réécriture. Parmi ces règles nous pouvons citer par exemple :

- Les règles de [tanwin]

```
If ( grapheme[indice]= ='T' )
{
   if (API[position][0]= = ' ً')
     phoneme = phoneme +"an";
   else
   {
      if (API[position][0]= = ' ٍ')
        phoneme = phoneme + "in";
      else
        phoneme=phoneme+"un";
   }
}
```

- Les règles [almad]

```
If ( (grapheme[ig]== ' ◌َ') && ((grapheme[ig+1]==' ا')||(grapheme[ig+1]=='ى')) )
 {
    phoneme = phonem + "aa";
    ig=ig+2;                          //ig : indice dans la table API
 }

If ( (grapheme[ig]== ' ◌ُ') && (grapheme[ig+1]==' و') )
 {
    phoneme = phonem + "uu";
    ig=ig+2;                          //ig : indice  dans la table API
 }

If ( (grapheme[ig]== ' ◌ُ') && (grapheme[ig+1]=='ي ') )
 {
    phoneme = phonem + "ii";
    ig=ig+2;                          //ig : indice  dans la table API
 }
```

- Le traitement de « ال » [al]

  La base de ce traitement réside dans la nature de la lettre qui suit le déterminant « ال » ; c'est pour cela que dans notre travail nous avons rajouté une autre colonne dans la table d'API pour déterminer si la lettre est lunaire ou solaire.

  Exemple :

```
API[1][0]='ب ';   API[1][1]='b';   API[1][2]='L';   //L : Lunaire
API[2][0]='ت ';   API[2][1]='t';   API[2][2]='S';   //S :  Solaire
```

soit la procédure suivante qui résume le traitement de « ال » :

```
if ( (grapheme[ig]= =' أ') && (grapheme[ig+1]= =' ل') )
 {
  phoneme = phoneme+'a';
  ig=ig+2;
  position = chercher (grapheme[ig],API);

 if(API[position][2]=='L')          // si le caractère  qui suive  le « ال » est lunaire
 {
   phoneme = phoneme+ ' l '  +API[position][1];
   ig++;
 }
```

```
else                                    // si le caractère  qui suive  le « ال » est Solaire
    {
       if (API[position][2]=='S')
        {
           phoneme = phoneme+API[position][1];
           ig++;
        }
     }
}
```

Il existe d'autres règles telle que la règle de gémination, la règle de [wakf], etc.

Durant l'élaboration de cet ensemble de règles on a utilisé la représentation suivante qui joue le rôle de notre alphabet phonétique

Les consonnes

```
API[0][0]='ا';      API[0][1]='E';      API[0][2]='L'; // L : lunaire
API[1][0]='ب';      API[1][1]='b';      API[1][2]='L';
API[2][0]='ت';      API[2][1]='t';      API[2][2]='S'; // S : solaire
API[3][0]='ث';      API[3][1]='&';      API[3][2]='S';
API[4][0]='ج';      API[4][1]='z';      API[4][2]='L';
API[5][0]='ح';      API[5][1]='h';      API[5][2]='L';
API[6][0]='خ';      API[6][1]='x';      API[6][2]='L';
API[7][0]='د';      API[7][1]='d';      API[7][2]='S';
API[8][0]='ذ';      API[8][1]='µ';      API[8][2]='S';
API[9][0]='ر';      API[9][1]='r';      API[9][2]='S';
API[10][0]='ز';     API[10][1]='Z';     API[10][2]='S';
API[11][0]='س';     API[11][1]='s';     API[11][2]='S';
API[12][0]='ش';     API[12][1]='§';     API[12][2]='S';
API[13][0]='ص';     API[13][1]='$';     API[13][2]='S';
API[14][0]='ض';     API[14][1]='D';     API[14][2]='S';
API[15][0]='ط';     API[15][1]='T';     API[15][2]='S';
API[16][0]='ظ';     API[16][1]='^';     API[16][2]='S';
API[17][0]='ع';     API[17][1]='£';     API[17][2]='L';
API[18][0]='غ';     API[18][1]='g';     API[18][2]='L';
API[19][0]='ف';     API[19][1]='f';     API[19][2]='L';
```

```
API[20][0]='ق';      API[20][1]='q';      API[20][2]='L';
API[21][0]='ك';      API[21][1]='k';      API[21][2]='L';
API[22][0]='ل';      API[22][1]='l';      API[22][2]='S';
API[23][0]='م';      API[23][1]='m';      API[23][2]='L';
API[24][0]='ن';      API[24][1]='n';      API[24][2]='S';
API[25][0]='ه';      API[25][1]='H';      API[25][2]='L';
API[26][0]='و';      API[26][1]='w';      API[26][2]='L';
API[27][0]='ي';      API[27][1]='y';      API[27][2]='L';
API[28][0]='أ';      API[28][1]='a';      API[28][2]='L';
API[29][0]='ؤ';      API[29][1]='e';      API[29][2]='L';
API[30][0]='ى';      API[30][1]='e';      API[30][2]='L';
API[31][0]='ة';      API[31][1]='t';      API[31][2]='L';
```

Les voyelles

```
API[32][0]=' َ';      API[32][1]='a';      API[32][2]='V';  // V : voyelle
API[33][0]=' ُ';      API[33][1]='u';      API[33][2]='V';
API[34][0]=' ِ';      API[34][1]='i';      API[34][2]='V';
```

atanwin

```
API[35][0]=' ً';      API[35][1]='an';     API[35][2]='T';    // T : atanwin
API[36][0]=' ّ';      API[36][1]='un';     API[36][2]='T';
API[37][0]=' ٍ';      API[37][1]='in';     API[37][2]='T';
```

Silence

```
API[38][0]='º';      API[38][1]='º';      API[38][2]="si l";
```

Gémination

```
API[39][0]=' ّ';      API[39][1]='ñ';      API[39][2]='G';    // G : gémination
```

Les séparateurs ou ponctuations

```
API[40][0]=' ؟';      API[40][1]=' ? ';    API[40][2]='P';    // P : ponctuation
API[41][0]=',';       API[41][1]=',';      API[41][2]='P';
API[42][0]='.';       API[42][1]='.';      API[42][2]='P';
API[43][0]=':';       API[43][1]=':';      API[43][2]='P';
```

API[44][0]=' '; API[44][1]=' '; API[44][2]='P';

API[45][0]=';'; API[45][1]=';'; API[45][2]='P';

API[46][0]='ئ'; API[46][1]='e'; API[46][2]='L';

API[47][0]='ء'; API[47][1]='e'; API[47][2]='L';

Caractères inconnus

API [48][1]='*';

4.6. La génération acoustique du signal de vocal par le synthétiseur

La génération du signal vocal c'est la synthèse réelle de la parole. Cette opération consiste à transformer la chaîne phonétique (qui représente la prononciation du texte à lire) résultante de la transcription à sa substance c'est-à-dire à sa réalisation acoustique (Figure 4.9).

Dans cette étape nous essayons de choisir dans notre base sonore les unités (phonèmes, ou diphones) les plus appropriés pour construire, par modification et, concaténation la phrase à générer. Ce qui signifie que nous créons une fonction de lecture automatique, de telle sorte qu'à la fin l'utilisateur n'a qu'à écouter les phrases.

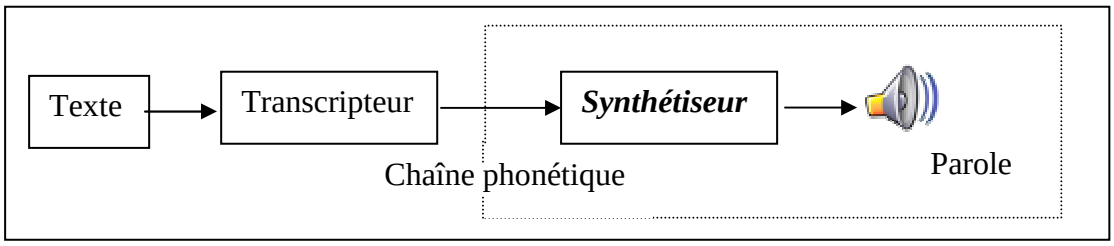


Figure 4.9 : Positionnement de la phase de génération acoustique de signal de la parole dans les systèmes TTS.

Pour assurer cette opération et pour générer toutes les phrases transcrites, nous commençons par la consultation de notre base sonore de phonèmes, qui contient les sons suivants :

{
"أ.wav"; "ب.wav"; "ت.wav"; ث.wav"; "ج.wav"; "ح.wav"; "خ.wav"; "د.wav"; "ذ.wav"; "ر.wav"; "ز.wav"; "س.wav"; "ش.wav"; "ص.wav"; "ض.wav"; "ط.wav"; "ظ.wav"; "ع.wav"; "غ.wav"; "ف.wav"; "ق..wav"; "ك.wav"; "ل.wav"; "م.wav"; "ن.wav"; "ه.wav"; "و.wav"; "ي.wav"; "fatha.wav"; "dama.wav"; "kasra.wav"; "an.wav"; "un.wav"; "in.wav"; "silence.wav";
}

Pour réduire les effets de coarticulation nous avons changé la base de phonèmes par une autre base de diphones, et nous avons décomposé la phrase transcrite avant de la générer oralement.
Par exemple :

- Nous avons le son qui correspond à la phrase /جلس يستمع الى ألراديو/ [zalasa yastami£u ilaa arraadyuu], c'est-à-dire nous avons les différents segments (diphones) de cette phrase ;
- Nous faisons la transcription de cette phrase ;
- Nous la décomposons en un ensemble de paires de caractères qui correspondent chacune à un diphone ;
- Enfin, nous régénérons cette phrase à travers la concaténation des sons qui représentent la prononciation de chacune de ces paires.

4.7. Configuration matérielle

Notre logiciel a été testé sous un enveronement Windows xp, et il a été compilé avec Builder C++, Ce dernier est un environnement de programmation visuel orienté objet qui assure le développement rapide de n'importe quelle application. Les fonctions du logiciel TOP-AS sont accessibles avec la sourie et le clavier. La principale caractéristique de TOP-AS est la possibilité de le réutiliser dans des systèmes de synthèse vocales (c'est un module réutilisable).

4.8. Présentation du notre logiciel TOP-AS

TOP-AS a été réalisé suivant une conception modulaire et parallèle, cette dernière accroît les performances de notre outil de Transcription Orthographique Phonétique d'un texte Arabe (TOP-AS) et cela tout en réduisant le temps nécessaire à son développement. Le schéma fonctionnel de TOP-AS est le suivant :

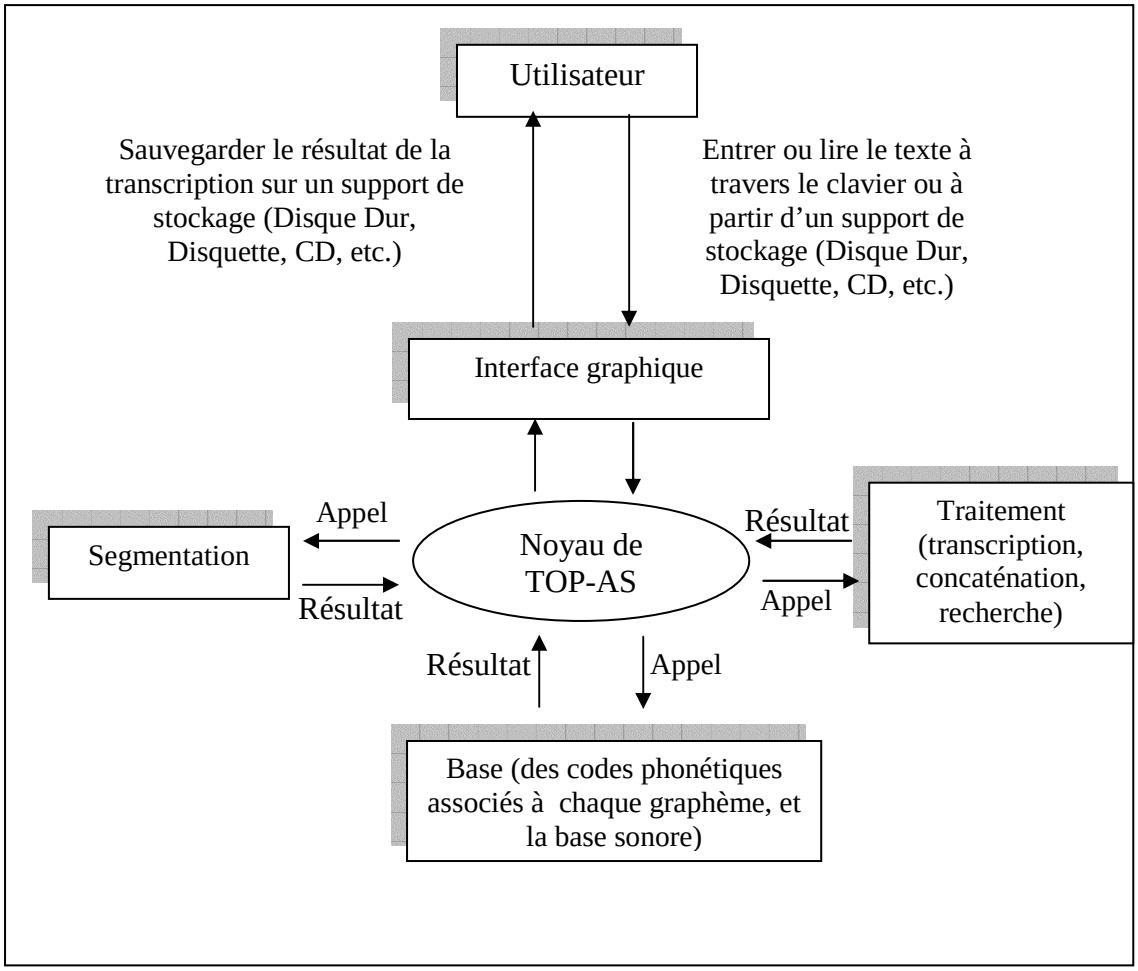
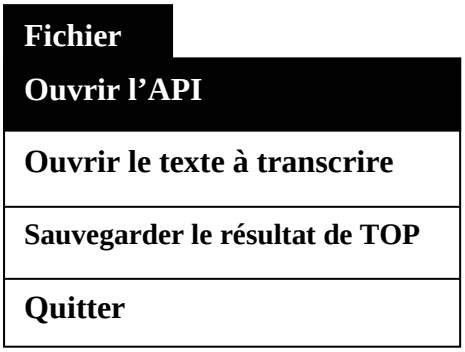


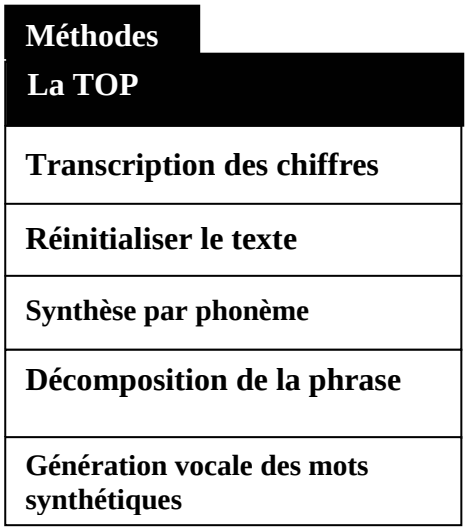
Figure 4.10 : Schéma fonctionnel du logiciel TOP-AS

Les menus qui constituent l'interface graphique de TOP-AS sont :

Le menu Fichier



Le menu Méthodes



Le menu Outils

Le menu Aide

L'interactivité de notre outil de Transcription Orthographique Phonétique en vue de la synthèse vocale d'un texte en Arabe Standard est assuré par le déclenchement d'une interface graphique principale (Figure 4.11).

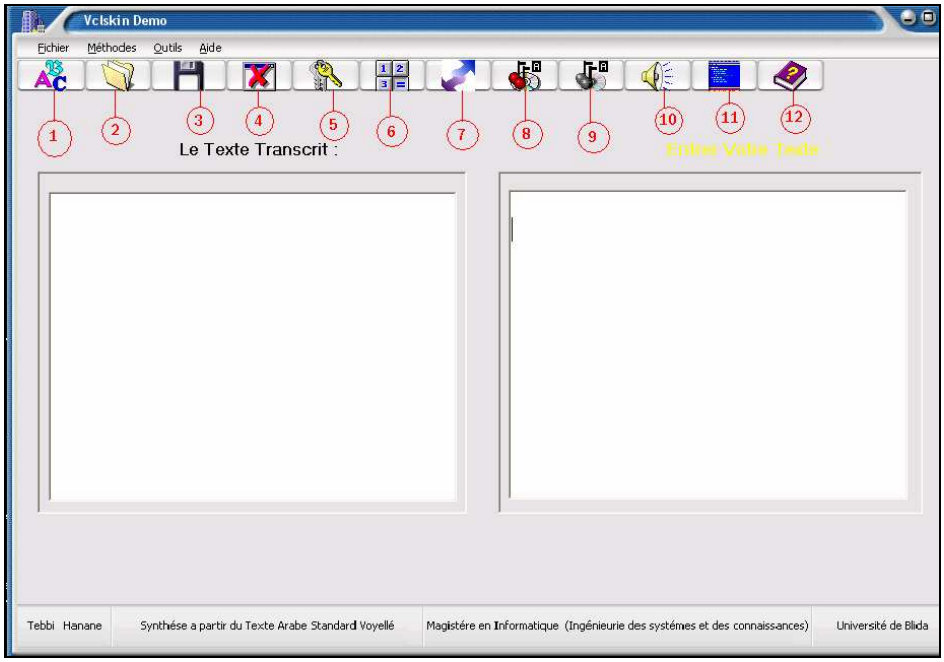


Figure 4.11 : Forme principale de notre système TOP-AS

1) Affichage du tableau de l'Alphabet Phonétique International (API)
2) Ouvrir le texte arabe à transcrire
3) Enregistrer le résultat de la transcription
4) Quitter l'application
5) Appliquer l'opération de transcription
6) Transcription des chiffres
7) Réinitialiser le texte à transcrire
8) Synthèse par phonèmes de notre texte
9) Synthèse par diphones de la phrase
10) Génération acoustique des mots synthétiques
11) Lancer l'outil d'analyse de la parole PRAAT
12) A propos.

4.9. Tests et résultats

Nous présentons dans cette partie le jeu de test utilisé pour la mise en œuvre de la transcription sur un ensemble de mots. Les résultats obtenus sont résumés dans les exemples : Transcription des phrases choisies aléatoirement, Transcription de سورة / الاخلاص / [suuratu alEixlaa$i],

Transcription des exceptions (Figure 4.12, 4.13, 4.14, et 4.15 respectivement) :

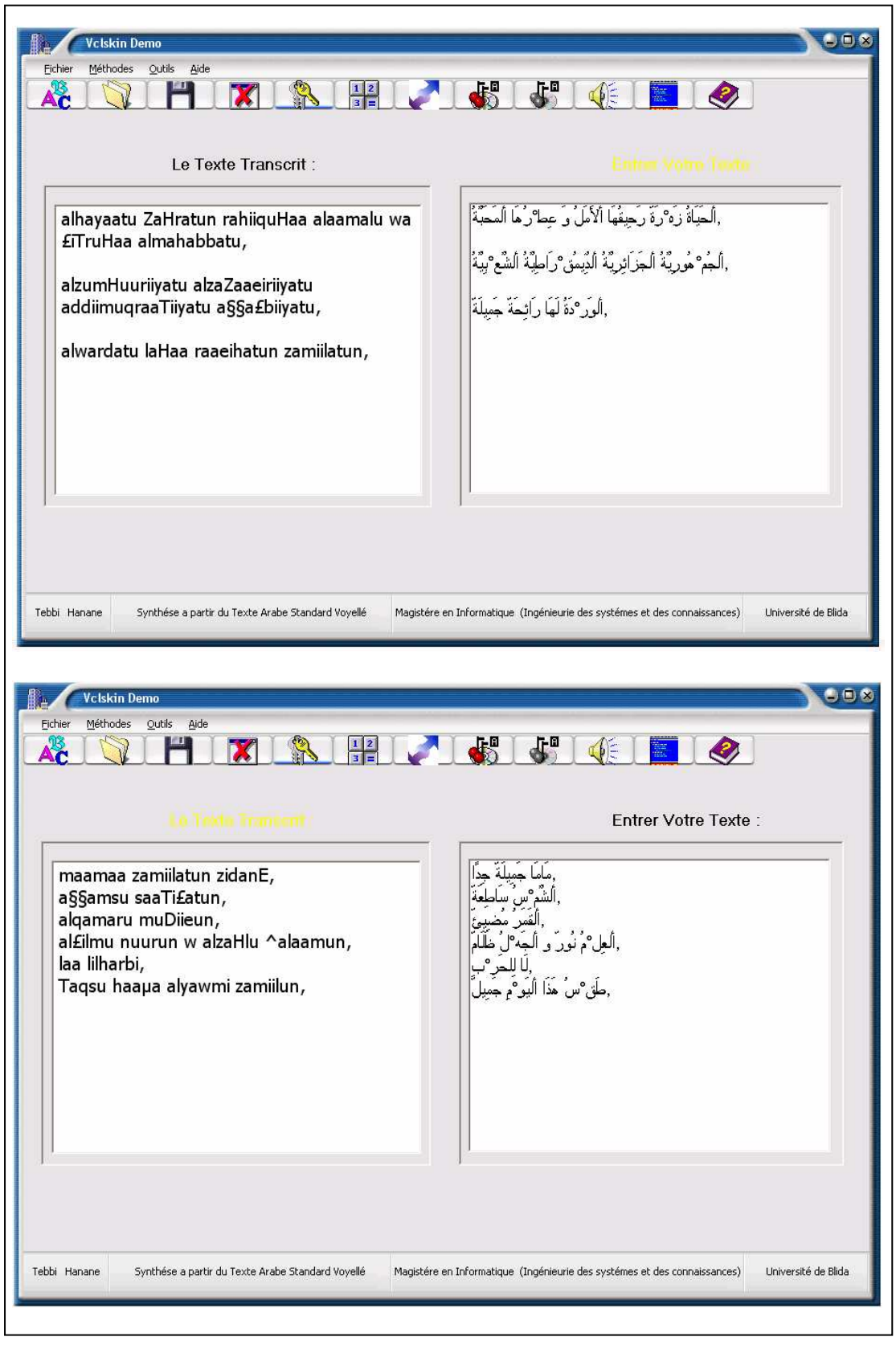


Figure 4.12 : Transcription des phrases aléatoires

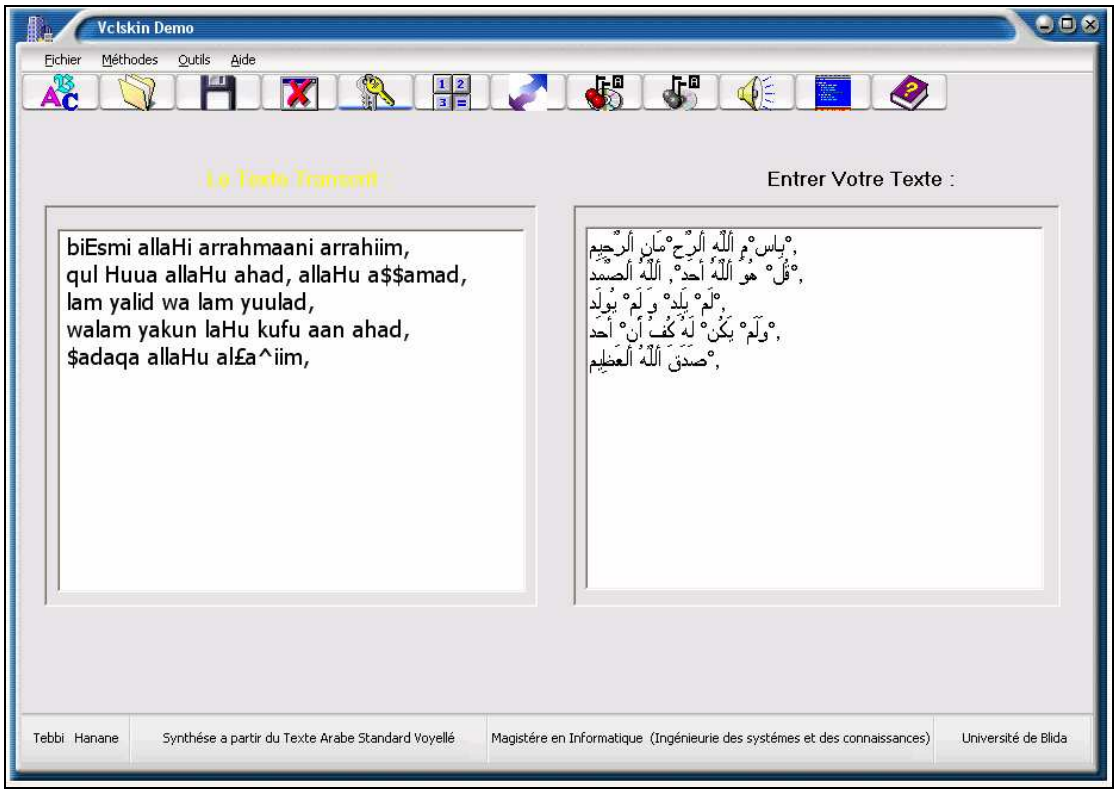


Figure 4.13 : Transcription de « سورة الاخـلاص ».

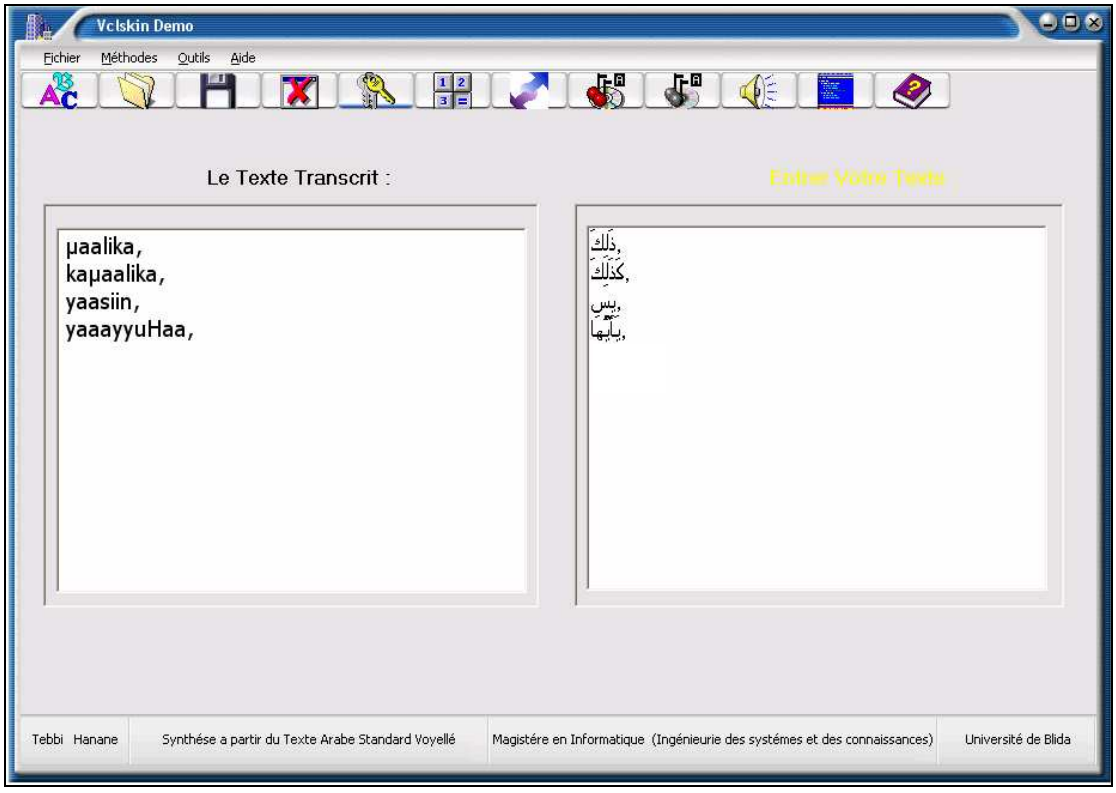


Figure 4.14 : Transcription des exceptions

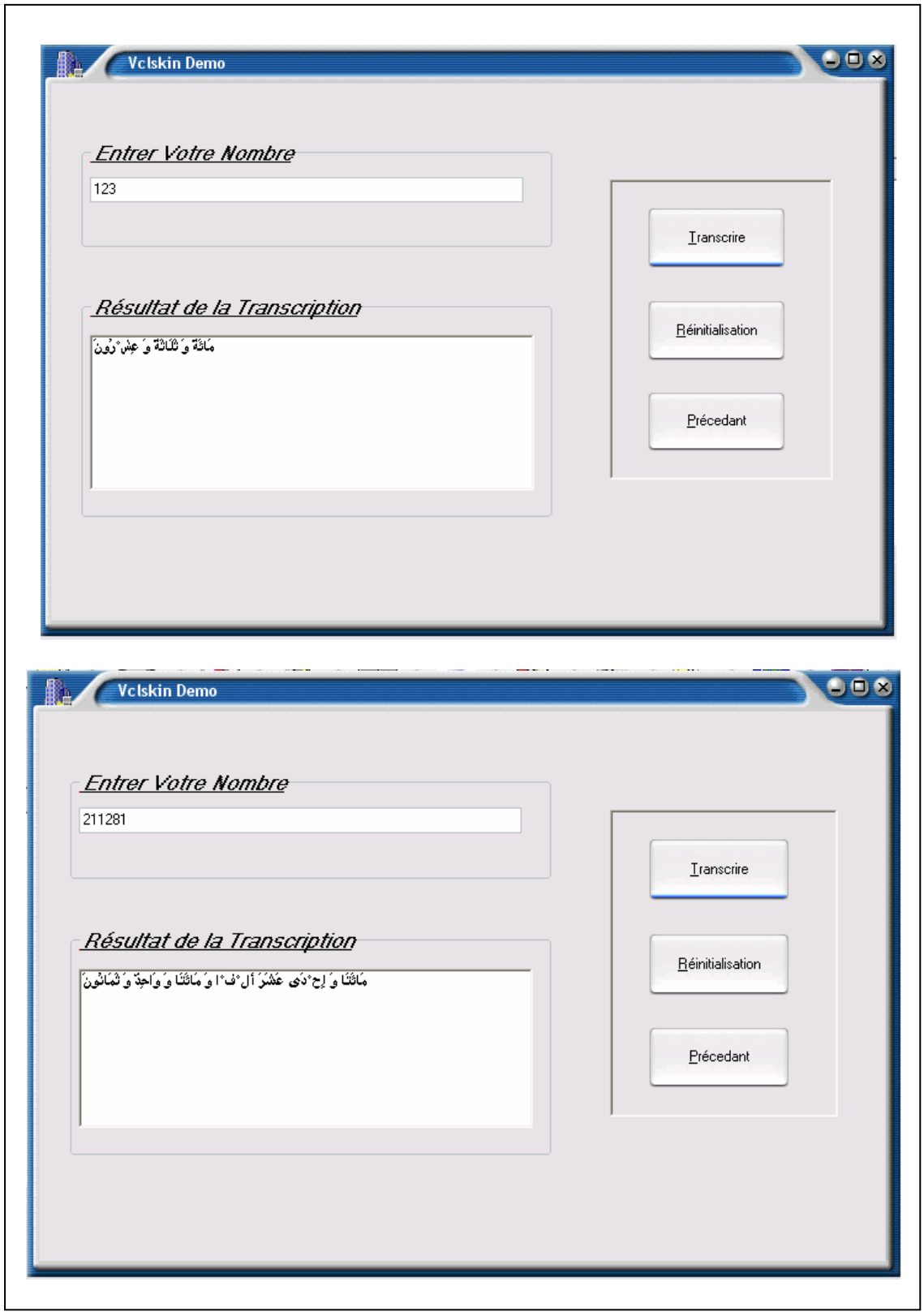


Figure 4.15 : Transcription des chiffres

Pour générer la sortie vocale qui correspond à cet ensemble d'exemples et n'importe quelle autre sortie vocale. L'utilisateur n'aura qu'à cliquer sur le menu « synthèse par phonèmes ». Dans le cas où il veut synthétiser les logatomes, il clique sur le menu « génération acoustique des mots synthétiques » qui fait appel à une autre forme où l'utilisateur doit saisir son logatome, puis il clique sur le button lire logatome pour l'entendre (Figure 4.16). Il faut savoir que Notre travail a été évalué par des expériences perceptives de la sortie vocale de notre logiciel.

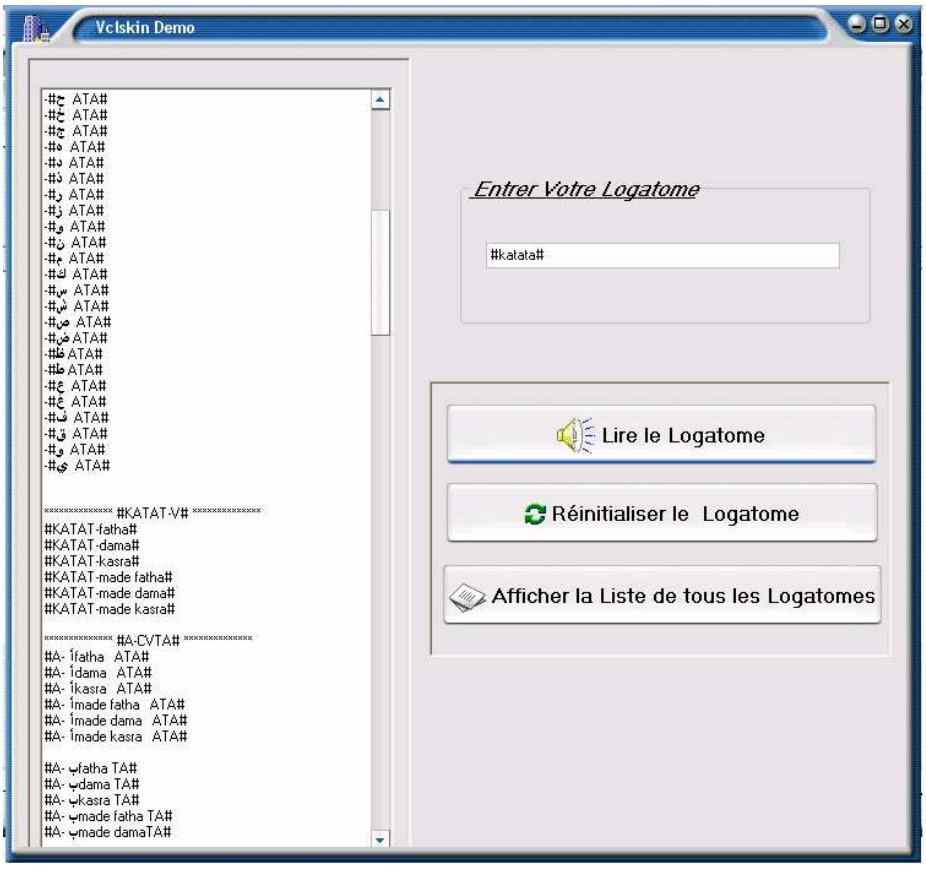


Figure 4.16 : Synthèse par diphones de quelques logatomes.

4.10. Conclusion

Dans ce chapitre nous avons présenté les différentes phases nécessaires (préparation de la base sonore, Transcription Orthographique Phonétique, génération acoustique) pour la mise en œuvre d'un outil de TOP en vue de réaliser un système de synthèse à partir d'un texte en Arabe Standard. Notre jeu de test commence par le test de l'étape de transcription puis celui de la génération vocale et cela par l'écoute de l'ensemble des phrases générées.

CONCLUSION

Dans ce travail nous avons abordé le problème de la mise en œuvre d'un logiciel de Transcription Orthographique Phonétique en vue de réaliser un système de lecture automatique d'un texte arabe. Notre approche est basé sur un formalisme divisé en trois grandes phases : une pour la préparation de la base sonore, une pour la TOP, et une autre pour la génération acoustique de signal vocal.

Nos études ont été plus approfondies sur la phase de transcription puisqu'elle représente l'étape clé des systèmes de synthèse de parole à partir du texte. Dans ce travail nous avons traité quelques ambiguïtés de la langue arabe tel que par exemple la présence de graphèmes qui ont plusieurs réalisations phonémiques (يَدًا [yadanE] / كَبِيرًا [kabiiranE]) et ceci par l'utilisation des deux techniques de transcription par règles et par lexique afin d'assurer une meilleure transformation du texte arabe en une chaîne phonétique.

Pour la synthèse nous avons utilisé la synthèse par concaténation d'unités sonores (phonèmes, diphone) et cela grâce à sa simplicité de mise en œuvre.

L'avantage principal de notre travail réside dans sa simplicité de plus, il assure les différentes étapes de n'importe quel outil de génération de parole à partir du texte, cet outil :

- est interactif ;
- il facilite la tâche de lecture aux mal-voyants ;
- il manipule des textes écrits en Arabe Standard ;
- il peut assurer la lecture des Emails ;

- et il peut être intégré dans des systèmes embarqués qui existent déjà tels que les téléphones portables, etc.

Dans notre cas nous avons abouti à de bons résultats pour l'étape de transcription graphèmes phonèmes tandis que la phase de synthèse souffre toujours de problème de coarticulation cela nous encourage de poursuivre la recherche pour améliorer la qualité des phrases générées afin d'aboutir à une parole naturelle.

Notre travail peut révéler un certain nombre de perspectives et cela pour poursuivre la recherche dans cette voie de synthèse à partir du texte. Ces dernières peuvent être résumées comme suit :

- améliorer la qualité de la parole en sortie et cela par l'utilisation des méthodes de modification des paramètres prosodiques ;

- utiliser d'autres techniques de synthèse telles que la synthèse par règles ou la synthèse par sélection d'unités préenregistrées ;

- automatiser si c'est possible l'opération de segmentation puisqu'elle représente le noyau des systèmes de synthèse vocale.

APPENDICE A

LISTE DES SYMBOLES ET DES ABREVIATIONS

AR : Auto Regressive

AS : Arabe Standard

API : Alphabet Phonétique International

C : Consonne

CS : Consonne Solaire

CL : Consonnes Lunaire

CP : Cavité Pharynx

CB : Cavité Buccale

CL : Cavité Labiale

CN : Cavité Nasale

dB : DéciBels

E : Energie (ou Intensité)

F_0 : Fréquence Fondamentale

IPA : International Phonetic Alphabet

Ou API : Alphabet Phonétique International

ICP : Institut de la Communication Parlée (Grenoble – France)

LPC : Linear Predictive Coding

LIMSI : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

LATL : Laboratoire d'Analyse et de Technologie du Langage

MIC : Modulation par Impulsions Codées

MBROLA	: MultiBand Overlap and Add
PSOLA	: Pitch Synchronous Overlap and Add
TAP	: Traitement Automatique de la Parole
TOP	: Transcription Orthographique Phonétique
TTS	: Text-To-Speech
TOP-AS	: Transcription Orthographique Phonétique d'un texte en Arabe Standard
UML	: Unified Modeling Langage
V	: Voyelle
Vocodeur	: Voice Coder
V/NV	: Voisés/Non Voisés

REFERENCES

1. Baloul. S, « Développement d'un système automatique de synthèse de la parole à partir du texte Arabe Standard voyellé ». Thèse de doctorat d'université, Le Mans, France, 27 Mai 2003.

2. Tassart, « Traitement du signal ». www.ircam.fr/ equipes/ analyse-synthése / tassart, 1998-1999.

3. www.rfv.insa-lyon.fr

4. Boite. R et Kunt. M, « Traitement de la parole ». Edition presses polytechniques romandes ,1987.

5. Didier. M, « Reconnaissance du locuteur en sciences forensiques ». L'apport d'une approche automatique Institut de Police Scientifique et de Criminologie, Université de Lausanne, 6 Décembre 1996.

6. Véronis. J, « Informatique et linguistique 1 ». Unité d'enseignement INF Z18, 1999-2001.

7. Marc Sato. M, « Representations verbales multistables en memoire de travail : vers une perception active des unites de parole » Docteur de l'INPG, specialite : sciences cognitives, Institut National Polytechnique de Grenoble, France, 30 Septembre 2004.

8. Guerti. M, « Contribution à la synthèse de la parole en Arabe Standard (Synthèse par diphones et techniques de Prédiction Linéaire), ILP-Alger, Algérie, 4 Mars 1984.

9. Benbellil. K, « Synthèse par polysons de l'Arabe Standard ». Mémoire de Magister, CRSTDLA - Université de Bouzaréah, Alger, Algérie, 17 Janvier 2005.

10. Song. J, « Relation entre le gain et le pith LPC ». Rapport RP/ LAA/ TSS/ RCP/ 294 Paris IX, 15 Avril au 15 Août 1983.

11. Ykhlef. F, « Modification de la fréquence fondamentale en vue de la synthèse de la parole a partir du texte de l'Arabe Standard ». Mémoire de Magister, département d'électronique, Université USD de Blida, Algérie, 2005.

12. http://www.ini.dz/ conference/ ecole2004/ Zemirli.htm.

13. Dutoit. T, « Introduction au traitement automatique de la parole notes de cours / DEC2 ». Collection électronique, Faculté Polytechnique de Mons, 2000.

14. http//post.queensu.ca/ ~lessardg/ cours.

15. Calliope, « La parole et son traitement automatique». Edition Masson , 1989.

16. http://www.lpl.univ-aix.fr/ lpl/ ressources/ ipa.html.

17. كيق تقرأ القران و فيه مقدمة في علومه للدكتور محمد ابو الفرج صادق
اليمامة للطباعة و النشر و التوزيع, دمشق, بيروت 2005

18. http://alis.isoc.org/glossaire/phonetique.htm.

19. Droua-Hamdani. G, « Prédiction de la durée des phonémes de l'Arabe Standard ». Mémoire de Magister, CRSTDLA - Université de Bouzaréah, Alger, Algérie, 18 Février 2004.

20. Douzidia. F. S, « Résumé automatique de texte arabe». Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en Informatique, Département d'informatique et de recherche opérationnelle, Faculté des arts et des sciences, Université de Montréal, Septembre 2004.

21. Guerti. M, « Contribution à la synthèse de parole en Arabe Standard », 16^{éme} JEP, SFA, pp.290-293, Hammamet, Tunisie, 5-9 octobre 1987.

22. Keller. E et Zellner. B, « Les théories de la parole dans l'éprouvette de la synthèse». Études des Lettres, vol 3: Les défis actuels en synthèse de la parole. pp. 9-27. Lausanne, Université de Lausanne, 16 Avril 2001.

23. Robiner. L. R et Schafer. R.W, « Digital Processing of Speech Signal ». Bell laboratories, 1978.

24. Belaid. A et Belaid. Y, « Reconnaissance des formes méthodes et applications ». Paris, Inter Editions 1992.

25. Bendraoua. Z et Bendou. F, « Elaboration d'une Base de Données des sons de l'Arabe Standard » Mémoire PFE, département d'Informatique, Université USD de Blida, Algérie, 2006.

26. Beller. Grégory et Marty. Aurélien, « TALKAPILLAR : outil d'analyse de corpus oraux », Actes des VIIèmes RJC ED268 'Langage et langues', Ircam, Institut de Recherche et de Coordination Acoustique/Musique1, place Igor Stravinsky Paris III, France, 15 Mai 2004.

27. http://www.irisa.fr/ra2000/cordial.

28. Saidane. T, Zrigui. M, et Ben Ahmed. M, « La transcription orthographique phonétique de la langue Arabe ». RECITAL 2004, Fès, 19-22 Avril 2004.

29. « Signaux systèmes et automatismes ». Proceedings du premier séminaire national, Blida, Algérie, un système de synthèse de la parole Arabe, 13-15, volume 2, Décembre 1992.

30. http://recherche.ircam.fr/ equipes/ analysesynthese/ Rapports Activités/ Rapport00.html, Rapport d'activité 2000.

31. http://www.bibliotheque.refer.org/ html/ parole/ sorin/ sorin.htm.

32. René-Lévesque. A, & Guyart. M, «Glossaire de la terminologie toponymique ». Traduite par la Commission de toponymie de l'Institut Géographique National de France et par la Commission de toponymie du Québec Paris et Québec, Décembre 1997.

33. http://aune.lpl.univ-aix.fr/lpl/presentation/equipes/pacomust.htm.

34. Kabache. M, « Application des réseaux de neurones a la reconnaissance automatique des phonèmes spécifiques en Arabe Standard ». Mémoire de Magistère, département d'électronique, Université USD de Blida, Algérie, 2004.

35. Saidane. T, Haddad. A, Zrigui. M, et Ben Ahmed. M, « Réalisation d'un système hybride de synthèse de la parole Arabe utilisant un dictionnaire de polyphones », JEP-TALN, Fès, Maroc, 20 avril 2004.

36. Boula de Mareüil. P, « Synthèse de la parole à partir de courriers et évaluation de la conversion graphème-phonème ». LIMSI-CNRS http://www.limsi.fr/ Individu/ mareuil/

37. http//:tcts.fpms.ac.be/ synthesis/ Mbrola.html.

38. www.uml.free.fr