UNIVERSITE SAÂD DAHLEB DE BLIDA

Faculté des Sciences
Département d'Informatique

MEMOIRE DE MAGISTER

En Informatique
Spécialité : Ingénierie des Systèmes et de la Connaissance

SYNTHESE POLYPHONIQUE DE L'ARABE STANDARD

Par

TOUAHRI Dalila

Devant le jury composé de :

A. GUESSOUM	Professeur	USD Blida	Président
K. KARA	Maître de conférence	USD Blida	Examinateur
F.Z. REGUIEG	Chargée de Cours	USD Blida	Examinatrice
M. GUERTI	Professeur	ENP Alger	Rapporteur

Blida, 2008

RÉSUMÉ

Le travail de ce mémoire est une contribution à l'étude et au développement d'un système de synthèse de la parole à partir du texte ou TTS (Text-To-Speech) en Arabe Standard basé sur la concaténation de polyphones. Cette étude intervient à différents niveaux de ce système : l'élaboration de la base de données acoustiques polyphoniques, la segmentation du texte lu en polyphones, le traitement du problème de perturbations du signal de la parole synthétique dans les points de concaténations, l'évaluation des systèmes TTS. Pour la génération artificielle du signal vocal ; nous avons développée deux méthodes de sélection d'unités acoustiques de tailles variables. La première méthode utilise des polyphones. La seconde est basée sur des polyphones multi-représentés ou SPC (Synthèse Par sélection dynamique dans un Corpus). La SPC représente le dernier état de l'art de la synthèse de la parole. Pour améliorer la qualité de la parole générée par notre système de synthèse, nous avons développée la méthode SPC. L'évaluation que nous avons effectuée est globale, par le test MOS (Most Opinions Score ou le score moyen des opinons) pour la comparaison entre la première méthode de sélection et la seconde. Nos résultats ont été évalués avec l'utilisation des phrases de tailles différentes et des interlocuteurs naïfs (méconnaissance des phrases préalablement). Les résultats obtenus d'après les tests effectués sont très satisfaisant ; puisque en aucun cas nous n'avons eu une dégradation de la qualité de la parole synthétique.

Mots clés : Synthèse de la parole à partir du texte (TTS), Arabe Standard (AS), polyphone, synthèse par concaténation, synthèse par sélection dynamique dans un corpus (SPC), évaluation globale des systèmes TTS.

ملخص

يتمثل عملنا في هذه المذكرة في المساهمة لدراسة وانجاز نظام نطق اصطناعي لنص مكتوب باللغة العربية الفصحى معتمد على تركيب وحدات صوتية تتمثل في بوليفونات . هذه المساهمة تتمثل في عدة مستويات من هذا النظام الآلي للنطق الاصطناعي: انجاز قاعدة معطيات صوتية بوليفونية ، تقسيم النص المقروء إلى وحدات بوليفونية ، علاجات الانقطاعات التي تظهر في نقط الاتصال ، تقيم نظامات النطق الاصطناعي الآلي لنص مكتوب . للحصول على الكلام الاصطناعي قمنا بإنجاز طريقتين للبحث في القاعدة الصوتية. الطريقة الأولى تعتمد على وحدات صوتية ذات طول متغير (بوليفونات) . الطريقة الثانية تتمثل في البحث عن وحدات صوتية ذات طول متغير ووجود متعدد. هذه الأخيرة تمثل أخر ما توصلت إليه حاليا الأنظمة الآلية للنطق الصناعي. بهدف تحسين نوعية الصوت الاصطناعي الناتج عن نظامنا قمنا بإنجاز الطريقة الثانية, التقييم الذي قمنا به هو تقيم إجمالي باستعمال اختبار MOSأو معدل عدد نقط الآراء للمقارنة ما ببين الطريقة الأولى والثانية نتائجنا قيمت باستعمال جمل ذات طول مختلف ومستمعين لا يعرفون الجمل مسبقا. النتائج التي تحصلنا عليها مرضية وهذا لعدم وجود أية حالة تشير إلى تدهور الكلام الاصطناعي المتحصل عليه من طرف نظامنا.

كلمات المفاتيح : النطق الآلي الاصطناعي ، اللغة العربية الفصحى ،تركيب وحدات صوتية ، بوليفونات ، تقيم نظام النطق الاصطناعي ، النطق الاصطناعي بالاستعمال وحدات الصوتية متعددة

ABSTRACT

The work of this memory is a contribution to study and development of Standard Arabic text-to-speech (TTS) system based on the polyphones. This study takes place at various levels of the system: the elaboration of the acoustical polyphonic database, the segmentation of the read text to polyphones, the processing of disturbances problems in the synthetic speech signal at the concatenate point, the evaluation of TTS systems. On behalf of generation an artificial speech signal; we have developed tow methods selection of acoustical variables-units. The first method uses the polyphones. The second uses the multi-represented polyphones or synthesis with dynamic selection in large database. The second method is the last state of art in the speech synthesis. The second method was developed to ameliorate the quality of our speech synthesis system. The evaluation of our TTS system is global, by MOS (Most Opinions Score) test, to compare between the firs selection method and the second. Our results have been tested with different sentences and naïf auditors. We have obtained a satisfactory results, because we dont have any degradation in the quality of the synthetic speech signal.

Keys words: Text To Speech synthesis system (TTS), Standard Arabic (SA), polyphone, concatenation synthesis, Synthesis with dynamic selection in large database, TTS Systems evaluation.

REMERCIEMENTS

Tous d'abord je tiens à remercier le bon Dieu qui m'a protégée et guidée.

J'exprime particulièrement mes profondes gratitudes et mes vifs remerciements à ma Directrice de mémoire, M^{me} M. GUERTI Professeur au Département d'Electronique à l'Ecole Nationale Polytechnique d'Alger, pour son aide, sa patience et ses précieux conseils tout au long de l'élaboration de ce travail.

Je remercie vivement Mr A. GUESSOUM, Professeur à l'Université Saâd Dahleb de Blida, de m'avoir fait l'honneur de présider mon jury de mémoire.

Mes sincères remerciements à Mr K. KARA, Maître de conférences à l'Université Saâd Dahleb de Blida, et M^{me} F.Z. REGUIEG Chargée de cours à l'Université Saâd Dahleb de Blida, pour l'honneur qu'ils me font en faisant partie de mon jury.

Je tiens à remercier chaleureusement tous mes enseignants en Graduation et en Poste Graduation au Département d'Informatique de l'Université Saâd Dahleb de Blida.

Je remercie également tous les étudiants de la Post Graduation en informatique pour l'ambiance qu'ils ont entretenue et leurs encouragements constants, particulièrement mes remerciements vont à : N.Toubaline et H. Tebbi.

Enfin, je tiens à remercier l'ensemble des membres de ma famille qui m'ont tous apporté, à un moment ou un autre, un soutien pour en arriver là aujourd'hui. Je remercie plus particulièrement mes chers parents qui m'ont guidée sur le chemin des études.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

La parole constitue sans aucun doute l'un des moyens les plus utilisés pour la communication entre les êtres humains. Ceux-ci ont très rapidement cherché à l'intégrer dans les interfaces Homme Machine. Cela a été rendu réalisable grâce aux efforts consentis de nombreuses équipes de recherche à travers le monde entier, en reconnaissance automatique et en synthèse de la parole. Malgré les avancées réalisées dans ces domaines ces dernières années, la plupart des interfaces courantes privilégient essentiellement l'écrit et le visuel, alors même que la parole constitue un élément primordial de la communication Homme Machine, de ce fait d'autres progrès demeurent nécessaires pour accroître le confort et la robustesse d'utilisation des systèmes actuels.

La synthèse de la parole est le domaine qui vise la production artificielle des sons de la parole humaine par des machines appelées : synthétiseurs. Cette synthèse s'accomplit à partir d'une représentation phonétique du message (texte lu), qui se présente sous la forme d'une chaîne de symboles phonétiques enrichis par des marques de prosodie (hauteur, intensité, etc.), que le synthétiseur se charge de restituer physiquement sous la forme d'ondes sonores. La synthèse de la parole est un champ de recherche pluridisciplinaire. Il implique plusieurs sciences : l'informatique, la linguistique, le traitement du signal, la phonétique, la phonologie, etc. Ceci permet à la synthèse de la parole d'être un domaine de recherche varié et innovant. Depuis quelques années, un nombre croissant de produits industriels utilise la technique de synthèse vocale. Il s'agit donc d'une technique d'avenir aux applications potentielles importantes. En particulier, dans le domaine des nouveaux services de Télécommunications ; qui est le domaine "porteur" actuel pour l'exploitation à grande échelle des technologies vocales.

L'objectif de notre travail est la création d'un moyen pour la génération artificielle de la parole humaine à partir d'un texte Arabe Standard, et cela pour assurer l'intégration de l'Arabe dans des applications embarquées utilisables dans notre vie quotidienne, et pour aider les personnes handicapées dans leur vie courante : par exemple, des machines à lire pour les non-voyants, des synthétiseurs à sortie vocale pour les non parlants.

Il existe différentes méthodes de synthèse destinées à créer de la parole artificielle : par règles, articulatoire, par concaténation. Les méthodes de synthèse à partir du texte représentent le moyen mis en œuvre pour passer de la représentation symbolique du texte vers le signal acoustique. Elles sont classées en deux catégories selon qu'elles modélisent ou non le fonctionnement de l'appareil vocal.

Les systèmes de synthèse par règles et articulatoire ont pour but de modéliser et d'étudier le phénomène de production de la parole ; mais la qualité de la parole synthétique générée est très insuffisante d'un point de vue perceptif, du fait que le modèle de production est très souvent issu d'une formulation qui ne tient pas compte de tous les phénomènes qui interviennent dans la production de la parole. Contrairement aux deux premières méthodes les systèmes de synthèse par concaténation ne s'intéressent pas à la modélisation du phénomène de production, en effet celui-ci va être capturé au sein des unités acoustiques à concaténer. Ce type de système demande très peu de connaissances sur le signal de parole, les principales difficultés vont intervenir au niveau du choix des unités acoustiques et au lissage des discontinuités ou le post-concaténation.

L'approche par concaténation est de loin la plus utilisée de nos jours, car celle-ci utilise des sons préenregistrés et stockés, qui sont assemblés pour constituer des mots, puis des phrases. On pourrait penser qu'il suffit d'enregistrer les sons élémentaires ou phonèmes d'une langue, puis de les juxtaposer pour former des mots. Des tests perceptifs ont démontré que la disposition des enregistrements des phonèmes seuls ne suffit, du fait que les phonèmes ne tiennent pas compte des zones de transitions possibles. Le diphone est considéré comme l'unité minimale qui se prête bien à générer de la parole artificielle. Il est définit comme la portion de signal acoustique comprise entre deux parties stables de deux phonèmes consécutifs et contenant en son centre toute la zone de transition. Cependant la qualité du signal de synthèse s'améliore avec la taille des unités, étant donnée que certaines classes de sons, leur problème de coarticulation (Influence de certains sons sur les sons voisins) dépassent le cadre du diphone ; de ce fait on a recours à l'utilisation des polyphones ou des unités de tailles variables. Nous avons choisi la méthode de concaténation de polyphones pour la réalisation de notre système de synthèse de la parole ; étant donnée ce type de synthèse est considéré comme une méthode hybride qui cerne le problème de la coarticulation.

Notre mémoire de Magister est organisé en quatre chapitres :

- le premier rassemble des généralités sur le signal de parole, ainsi qu'une vue générale sur le Traitement Automatique de la Parole et quelques notions de base sur la langue Arabe Standard ;

- le second est consacré aux principales méthodes de synthèse de la parole, les modules qui le composent, leurs domaines d'application ainsi les travaux antérieurs dans le champ qui concerne notre étude qu'est la synthèse à partir d'un texte Arabe Standard;

- Le troisième chapitre, présente l'architecture de notre système de synthèse à partir du texte Arabe Standard, que nous avons appelé TALKARABIC ;

- Le quatrième concerne l'étude des principales méthodes qui existent pour l'amélioration de la qualité du signal vocal synthétique. Par la suite nous présentons notre algorithme de sélection dynamique dans un corpus en vue de l'amélioration de la qualité de la parole synthétique de notre système TALKARABIC.

Enfin, le mémoire se termine par des conclusions générales et des perspectives pour ouvrir des voies de recherches en vue de la réalisation de travaux futurs dans le domaine de la synthèse de la parole.

CHAPITRE 1
GENERALITES SUR LE TRAITEMENT AUTOMATIQUE DE LA PAROLE

1.1. Introduction

Ce chapitre rassemble quelques notions sur le domaine du Traitement Automatique de la Parole (TAP). Nous allons cependant tout d'abord parler des notions qui se rattachent à l'étude du signal vocal. Nous allons présenter de manière succincte quelques notions sur la langue Arabe Standard (AS) ; après avoir donné une courte vue d'ensemble du domaine.

1.2. Traitement de la parole

La parole constitue sans aucun doute l'un des moyens les plus utilisés pour la communication entre les êtres humains. Elle apparaît physiquement comme une variation de la pression de l'air, causée et émise par le système phonatoire. C'est un acte qui évolue dans le temps (suite de sons).

La parole joue un rôle fondamental dans l'expression linguistique que l'on peut représenter sous forme écrite.

En effet le signal de parole véhicule, au-delà du message linguistique proprement dit, d'autres types d'informations, notamment les caractéristiques du locuteur comme : le genre du locuteur (homme ou femme), son âge, son attitude et ses intentions par rapport au discours, le type de discours, la santé physique et psychique du locuteur, etc., et de l'environnement, de l'enregistrement. Toutes ces informations contenues dans un même signal contribuent à sa variabilité et mènent à des directions de recherche variées, telles que :

- la Reconnaissance Automatique de la Parole (RAP): qui comporte fondamentalement deux types de reconnaissance de la parole :
  La première est la reconnaissance du message dont l'objectif est de reconnaître ce qui est dit (mots isolés ou enchaînés) ;

La deuxième est la reconnaissance du locuteur dont l'objectif est de reconnaître la personne qui parle;

- l'analyse de la parole : dont le but est de mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçue, par la recherche d'indices acoustiques, de constituants du signal de la parole, estimation du spectre, et ce grâce à divers modes de représentation du signal. Les analyseurs de la parole sont utilisés soit comme composantes de base de systèmes de synthèse à partir du texte (TTS : Text-To-Speech), codage ou de reconnaissance, soit pour des applications spécialisées, comme l'aide au diagnostic médical pour les pathologies du larynx, par l'analyse du signal de la parole ;

- le codage de la parole : dont le but est de permettre la transmission ou le stockage de parole avec un débit réduit. Cela est rendu possible par l'élimination de la redondance du signal de la parole en calculant les indices les plus pertinents dans le but de restituer ce signal à partir de ces indices ;

- la synthèse de la parole à partir du texte désigne l'ensemble des traitements permettant à une machine de transformer un texte écrit dans une langue donnée en un message oral correspondant à la langue dont laquelle est écrit le texte. Le but recherché par les synthétiseurs à partir du texte est la production d'une voix synthétique qui apparente au mieux la voix humaine, tant au niveau de l'intelligibilité des sons qu'au niveau du naturel.

Ces domaines de recherches font appel aux connaissances de plusieurs sciences tels que les signaux émis par la parole, la phonétique, la phonologie, la linguistique, l'informatique, le traitement du signal, l'intelligence artificielle, les statistiques, etc.

1.3. Étude Acoustique de la parole

L'onde de la parole couvre quasiment toute l'étendue du spectre audible. En pratique, on peut se limiter à la bande 50-5000 Hz. Le signal de la parole est un signal très riche en informations et très complexe ; pour cela nous abordons le signal de parole dans ce paragraphe d'un point de vue acoustique en évaluant ses paramètres à savoir la fréquence fondamentale**,** l'intensité, la durée, l'intonation, la résonance, etc.

1.3.1. Fréquence fondamentale

La fréquence fondamentale ou fréquence laryngienne notée F_0 représente la fréquence de vibrations des cordes vocales. Son estimation est liée à la localisation de portions voisées sur le signal de la parole, les sons non voisés ayant une F_0 nulle.

Les algorithmes d'extraction de F_0 peuvent être de type temporel (AMDF, LPC..) ou fréquentiel. Les premiers se basent directement sur la description temporelle du signal pour le calcul de F_0 ($F_0 = 1/T_0$), alors que les seconds s'appuient sur les fréquences des harmoniques (fréquence de résonance) qui peuvent être représentés graphiquement sur un spectrogramme.

Sur le plan de la perception, la valeur de F_0 correspond en première approximation à la sensation de hauteur que procure un son. Les phonéticiens utilisent le ton pour exprimer le rapport de hauteur entre une fréquence F_1 et une fréquence F_2, étant donné que notre oreille a une perception logarithmique de la hauteur et non pas linéaire.

La figure 1.1 donne l'évolution temporelle de la F_0 de la phrase *"les techniques de traitement numérique de la parole"*. On constate qu'à l'intérieur des zones voisées la F_0 évolue lentement dans le temps.

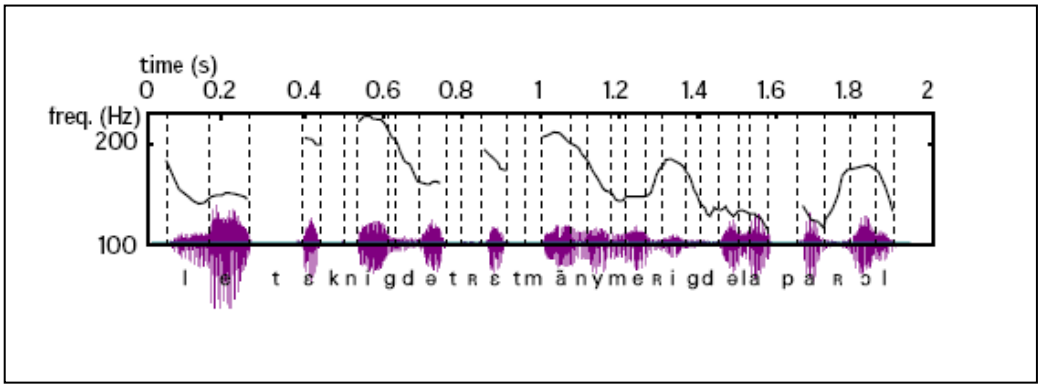


Figure 1.1 Evolution de la fréquence de vibration des cordes vocales dans la phrase : "*Les techniques de traitement numérique de la parole*"[10].

La fréquence fondamentale permet de diviser l'ensemble des sons de la parole humaine en trois grandes macro classes :

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

1.3.2. L'intensité

L'intensité est Résultante de la pression sous glottique. Elle est mesurée sur des portions de signal allant de 5 à 10 ms (énergie à court terme) et exprimée en décibels (dB) pour respecter l'échelle perceptive.

1.3.3. La durée

La durée est le paramètre acoustique le plus délicat à évaluer, car il ne dépend d'aucun corrélat biologique, contrairement à F_0 et à l'intensité (qui dépend respectivement de la tension des cordes vocales et de la pression sous glottique). Pour calculer la durée d'un phonème, il faudrait se fixer deux événements qui délimitent ses repères initial et final. La durée représente généralement le temps de la prononciation d'un phonème.

Pour mesurer une durée quelconque, il faudrait au préalable désigner, d'une part, les unités à mesurer et d'autre part, leurs repères (les frontières) dans le signal parole. Elles peuvent concerner les phonèmes, distance entre voyelles, les pauses, etc. 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.

1.3.4. L'intonation

Le terme de l'intonation a deux définitions possibles :

- au sens strict, ce mot désigne les changements relatifs à la hauteur de la voix, que certains chercheurs confondent avec le mot mélodie ;
- le sens le plus étendu de ce terme fait aussi référence aux changements de la durée et de l'intensité. Dans ce dernier cas, il s'apparente au mot prosodie [1].

1.3.5. La résonance

Un système vibratoire possède généralement une fréquence de vibrations dite propre, correspondant à son mode d'oscillation libre. En présence d'une excitation extérieure, ce système entre en vibrations à la fréquence imposée par l'extérieur. Mais l'amplitude dépend fortement de la fréquence ; elle est maximale lorsque la fréquence imposée est égale à la fréquence propre du système.

1.4. Représentation sonagraphique du signal de la parole

Un spectrogramme est une représentation de l'évolution du spectre fréquentiel d'un son en fonction du temps. Elle est la représentation généralement utilisée pour l'étude des sons d'instruments, de voix de chanteurs, de signaux sonores de machines, etc. On y porte le temps en abscisse et la fréquence en ordonnée ; l'amplitude est alors codée soit par un niveau de gris, soit par un code de couleurs. C'est en quelque sorte une vue de dessus du graphique ci-contre (figure 1.4.), mais dont la lisibilité est meilleure et la lecture aisée. Différents logiciels permettent l'obtention de telles représentations.

Différentes informations peuvent être tirées d'un spectrogramme en vue de l'analyse du signal de la parole (figure 1.4.) :

- en repérant les instants de transition, on peut faire correspondre des symboles phonétiques à diverses phases du spectrogramme;
- on peut étudier l'évolution mélodique d'un signal ;
- on peut étudier le rythme du signal en observant l'évolution des durées des segments phonétiques qui le composent.

Il existe deux types de spectrogrammes :

- Spectrogramme à Bande Large (SBL) ;
- Spectrogramme à Bande Etroite (SBE).

Pour déterminer le type de sonagramme à utiliser pour l'analyse de notre signal acoustique il s'agit essentiellement d'effectuer un choix entre les paramètres qui nous intéressent :

- un spectrogramme à bandes larges (150 - 300Hz) offre une meilleure résolution temporelle et permet de dégager les formants vocaliques, mais apporte moins d'éléments pour l'étude du domaine fréquentiel. (figure 1.2) ;
- inversement, un spectrogramme à bandes étroites (10 - 45Hz) offre une bonne résolution au niveau fréquentiel, mais l'analyse temporelle est moins fine (figure 1.3).

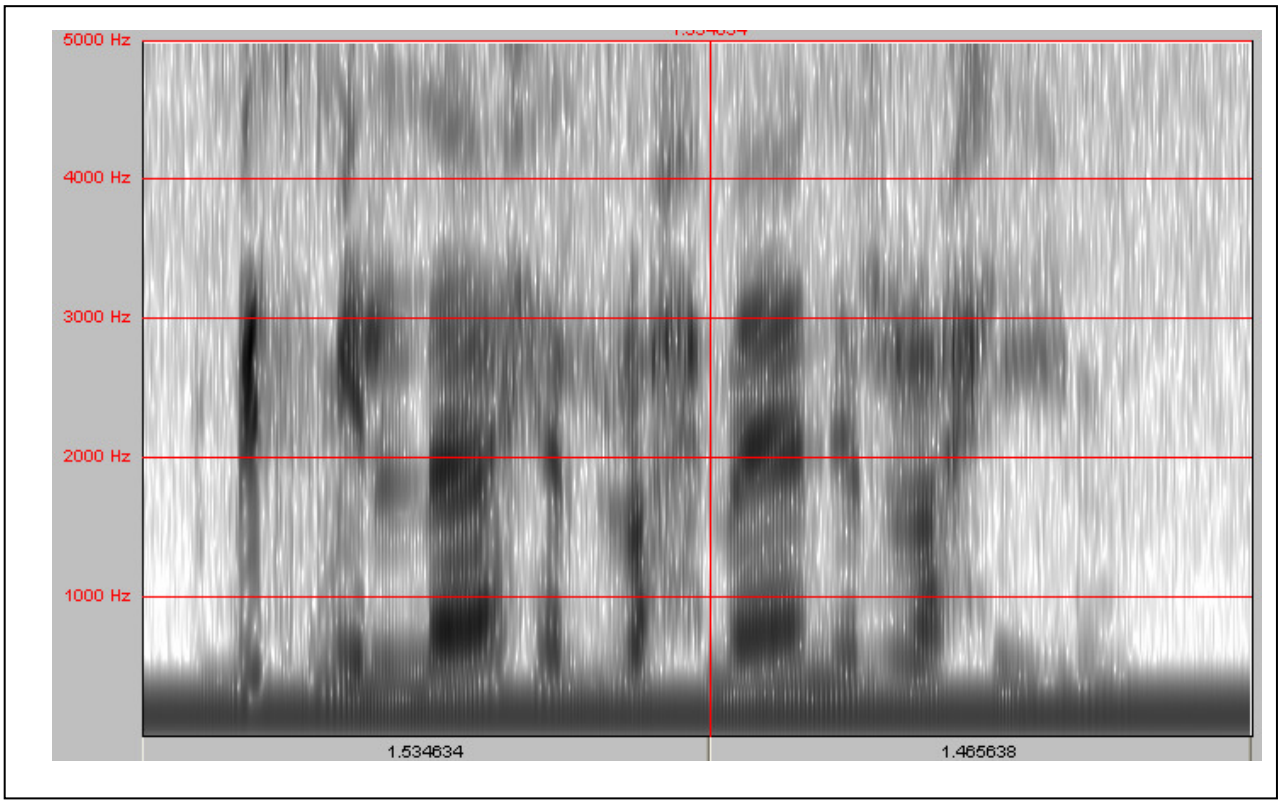


Figure 1.2 : spectrogramme à Bande Large de la phrase [bismi allaHi arrahmaan arahiim]

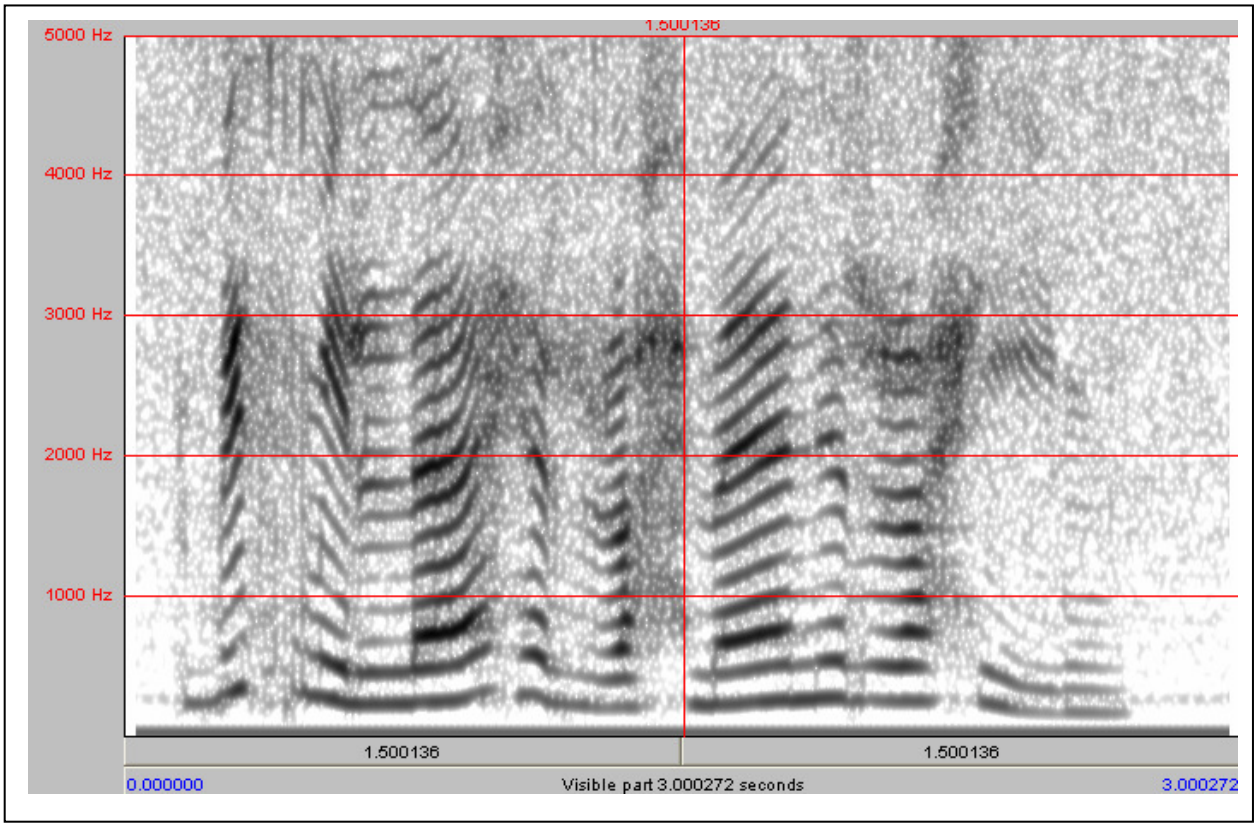


Figure 1.3 : spectrogramme à Bande Etroite de la phrase [bismi allaHi arrahmaan arahiim].

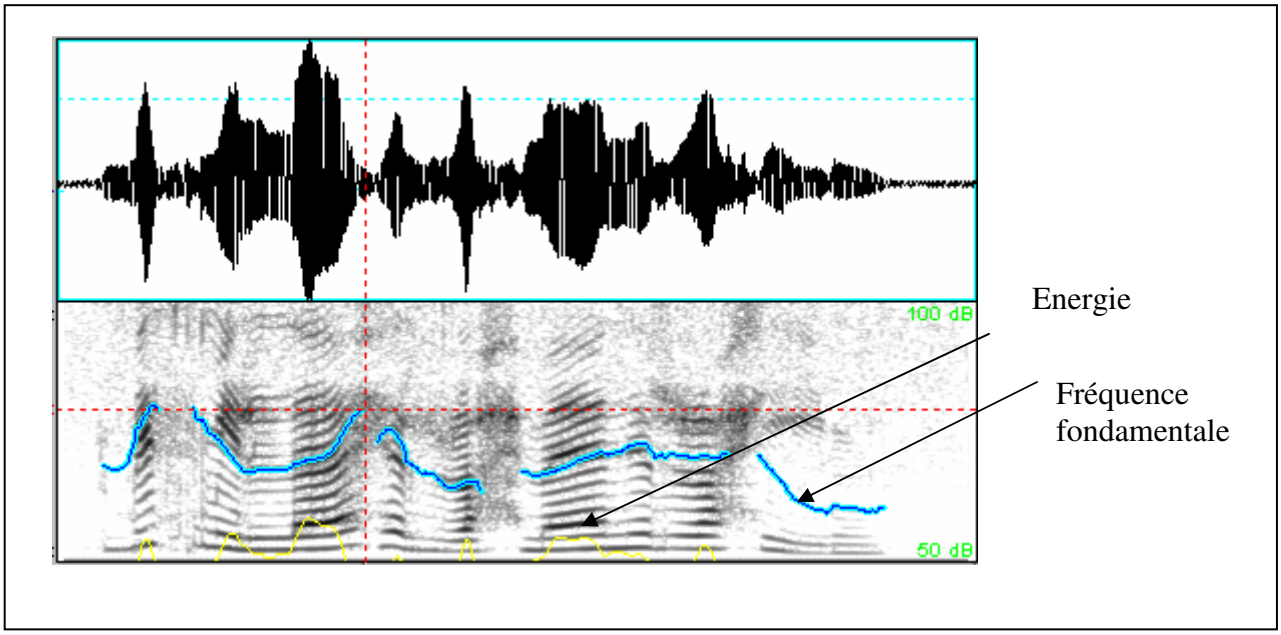


Figure1.4 : Sonagramme accompagné de quelques indices acoustiques

1.5. Alphabet Phonétique Internationale (API)

Les insuffisances des systèmes orthographiques traditionnels sont la conséquence directe qui a fait appel à la construction de l'API.

Quelques exemples : pour représenter le son [z], l'orthographe du Français utilise plusieurs graphèmes (« z », « s » (entre deux voyelles, liaison), « x » (liaison)). Inversement un graphème du Français peut renvoyer à plusieurs sons : le graphème « c » peut correspondre à [s], à [k], à [g] (ex. « second »), ou encore à « rien » (unité non prononcée, comme dans « tabac »).

Le besoin d'un système notationnel univoque : un son, un symbole. Ces symboles sont ceux qui sont proposés par l'Association Phonétique Internationale. Cette association a été créée dans le but d'uniformiser les diverses transcriptions phonétiques proposées à travers le monde. Cette association a donc proposé un API qui repose sur le principe qui veut qu'à chaque symbole corresponde un seul son et à chaque son correspond un seul symbole. Cela permet de:

- transcrire n'importe quelle langue avec le même jeu de symboles ;
- pouvoir lire une transcription phonétique d'une langue que l'on ne parle pas avec une précision relative.

L'API est prévu pour couvrir l'ensemble des langues du monde, il a été publié en 1888. Il fait l'objet des mises à jour régulières en fonction de l'avancement des recherches

dans le domaine de la Communication Parlée (CP). Sa dernière révision date de 2005 [27]. Cet alphabet phonétique représente un système partagé par la plupart des linguistes.

Quand on veut représenter les prononciations dans ce système on met la représentation entre crochets. Ainsi, pour écrire le son associé au mot *rat* on écrit [ra].

Le Tableau 1.1 représente un exemple de flexion orthographique phonétique des consonnes de l'Arabe Standard (AS) ; telle qu'il fournit pour chaque symbole arabe le code phonétique qui lui correspond.

Tableau 1.1 : Correspondances Graphèmes Phonèmes de l'AS suivant l'API, avec E : Emphatique et NE : Non Emphatique, V : Voisée et NV : Non Voisée.

Graphème en AS	**Phonème en API**	**E/NE**	**V/NV**	**Graphème en AS**	**Phonème en API**	**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 غ ج ث ف ش 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 ر	[l] [r]	E	V
Vibrante				**Semi-voyelles**			
ر	[r]	NE	V	و ي	[w] [y]	NE NE	V V

1.6. Transcription Orthographique Phonétique (TOP)

La Transcription Orthographique Phonétique est le passage d'un texte écrit vers un texte lu. Elle fournit la prononciation associée au texte qu'on veut entendre.

Nous illustrons la transcription à l'aide d'exemples : le signal et la représentation phonétique résultante de la transcription de mot Anglais [phonetician] (Figure 1.5.).

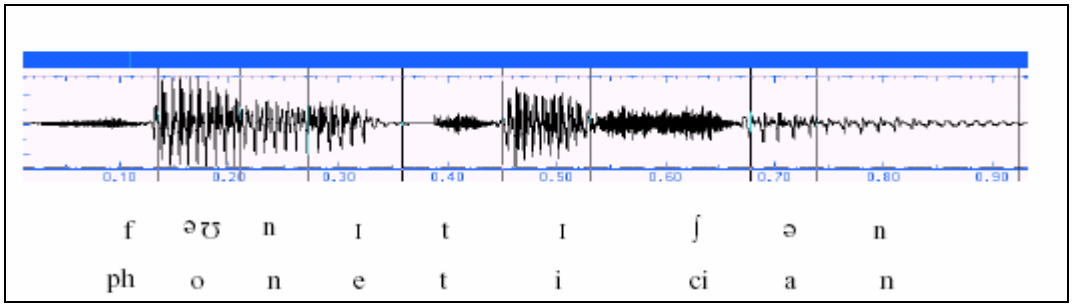


Figure 1.5. : Signal de parole et phonèmes (mot Anglais *phonetician*) [26]

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

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

- Notation
  [notasjõ]

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

- Exemple de transcription
  [ɛgzãmplə də trãskripsjõ]

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

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 ;
- lexical ;
- syntaxique, etc.

1.7. Phonétique

La phonétique est la discipline scientifique qui étudie les sons utilisés dans le langage humain, indépendamment du fonctionnement de ces sons au sein des langues du monde. La phonétique se divise en trois sous branches :

- la phonétique articulatoire: la plus ancienne des trois branches de la phonétique, elle étudie la manière dont les sons du langage humain sont produits. La description des articulations se fait à l'aide de trois variables : l'activité du larynx (voisement ou sonorisation), l'endroit où se situe le resserrement maximum au niveau du conduit vocal (point d'articulation), et la façon dont s'effectue l'écoulement de l'air à travers le canal phonatoire (mode d'articulation) ;
- la phonétique acoustique, étudie la transmission des sons dans l'air selon ses caractéristiques physiques (fréquence, intensité, durée, etc.) ;
- la phonétique auditive, étudie les processus d'audition du langage, la façon dont l'humain perçoit et reconnaît les sons.

1.8. 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 [28]. Cette science se base sur l'étude de l'écriture.

Le travail du phonologue est de déterminer le statut linguistique d'un son ou bien plutôt d'une propriété sonore au sein d'une langue. Parce que, certains sons ont une fonction distinctive, d'autres sont dûs à l'environnement.

Les sons distinctifs d'une langue forment un système sonore, doté d'une structure et d'une cohérence internes. Exemple : la différence fonctionnelle (dans l'écriture et dans le sens) entre les trois mots suivants :

نمل [na**m**lun] qui signifie fourmis

نحل [nahlun] qui signifie abeilles

نخل [naxlun] qui signifie palmiers.

1.9. Notions sur la langue Arabe Standard (AS)

La langue Arabe Standard est la langue dans laquelle est écrit le Saint Coran et que l'on trouve aussi enseignée dans les écoles. Elle est la langue officielle de nombreux pays, et elle est également la langue employée dans la plupart des écrits et, à l'oral, dans les situations officielles ou formelles (discours religieux, politiques, journaux télévisés, etc.).

L'étude de la grammaire Arabe a commencé très tôt au milieu du 11 siècle de l'Hégire et a donné lieu à des énormes productions, avant de connaître une période de stagnation qui a duré plusieurs siècles [8].

L'écriture arabe va de droite vers la gauche et lie les lettres de son alphabet. En revanche, la plupart des lettres s'attachent entre elles, et leur graphie diffère selon qu'elles sont précédées et/ou suivies d'autres lettres ou qu'elles sont isolées.

Exemple de la lettre /ت/ [t] elle correspond à 5 graphies voir le tableau (1.2). Il résulte 78 formes graphiques à partir des 28 lettres. Cependant, certaines lettres, ne s'attachent jamais à la lettre suivante, tels que elles sont présentées dans cet ensemble / ا, و, ر, ز, د, ذ / [E, w, r, Z, d, µ].

Tableau 1.2 : Exemples de variations de la lettre /ع /[£] et /ت/ [t] dans les différentes Positions : Initiale, Médiane, et Finale.

à la fin du mot (PF)		**au milieu du mot (PM)**		**au début du mot (PI)**		**Graphème**
مع وداع	ع ع	لعب	ـعـ	عيد	عـ	**[£]**
بيت كلية نجدة	ت ة ة	مكتب	ـتـ	تلميذ	تـ	**[t]**

L'Arabe Standard comporte 28 consonnes ou [huruuf] plus la hamza et 6 voyelles ou [harakaat] dont 3 voyelles courtes ; elles sont sujettes à l'allongement et donnent en conséquence 3 voyelles longues.

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

- 14 consonnes *solaires* /ن ل ظ ط ض ص ش س ز ر ذ د ث ت/ [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 **الشجرة** qui signifie un arbre, sera prononcé [a§azara] et pas [al§azara] ;

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

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 1.6) et leur voisement : sonore, sourd (Tableau 1.3).
Le mode d'articulation peut être (occlusif, fricatif, nasal, liquide) selon que les consonnes sont :

- 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 [29]. 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 ;

- nasale, dans ce cas le son est produit à travers un couplage entre les cavités pharyngo-buccale et nasales;

- liquides, leur articulation ressemble à celle d'une voyelle. La seule différence réside dans la fermeture partielle de conduit vocal (c'est le cas de / ل / [l]) ;

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

Le lieu d'articulation (la zone du conduit vocal qui participe à la formation du son) peut être (labial, dental ou vélo palatal) :

- labial : utilisation des Lèvres ;
- dentale : utilisation des dents ;
- palatale : utilisation du palais ; etc.

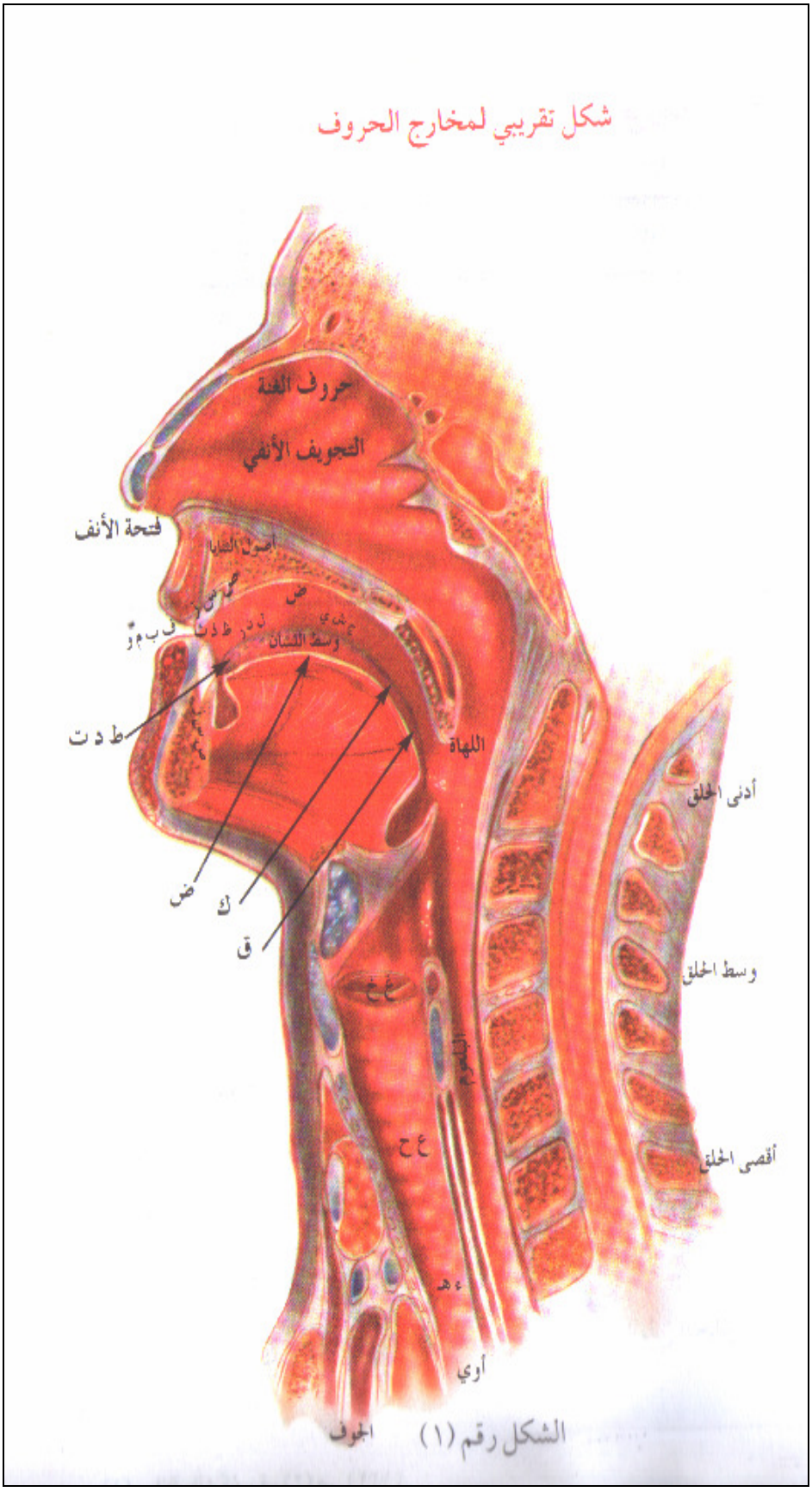


Figure 1.6. : Les lieux d'articulation des 28 consonnes de l'Arabe Standard [16].

Vibrantes	Affriquées	Liquides	Nasales	Fricatives		Occlusives		Semi-voyelles	
sonores	sonores	sonores	sonores	sonores	sourdes	sonores	sourdes	sonores	
			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	
	ج							Palatales	
					خ			Vélaires	
				غ			ق	Uvulaires	
				ع	ح			Pharyngales	
					ه	ء		Glottales	

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

En Arabe Standard chaque consonne (ou [harf]) est suivie par une voyelle [harakatun] pour qu'elle puisse être produite. Les voyelles courtes sont représentées par des symboles appelés signes diacritiques. Les quatre symboles sont transcrits de la manière suivante :

- la fetha [a] : est symbolisée par un petit trait sur la consonne ;
- la damma [u] : est symbolisée par un petit crochet au-dessus de la consonne ;
- la kasra [i] : est symbolisée par un petit trait au-dessous de la consonne ;
- le sukun : est apposé sur une consonne lorsque celle-ci n'est liée à aucune voyelle, il est symbolisé par un petit rond.

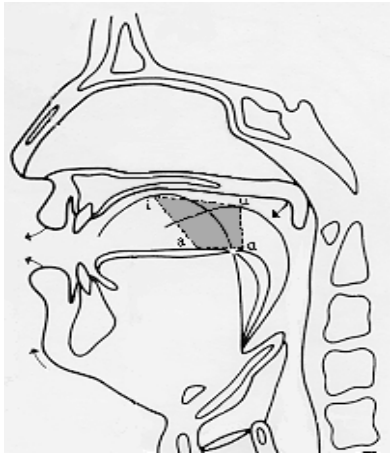
Ces symboles sont absents à l'écrit dans la majorité des textes arabes ce qui peut engendrer des ambiguïtés de prononciation.

Exemple :

كَتَبَ مُحَمَّدُ الدرسَ [kataba muhammadu addarsa] qui signifie Mohamed a écrit le cours

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

Selon le contexte la durée d'une voyelle longue est environ double de celle d'une voyelle courte. De plus les différentes voyelles courtes se diffèrent entre elles par leurs lieux d'articulation et le degré d'ouverture du conduit vocal (ouvert, fermé, semi fermé) (Figure 1.7.).



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

Figure 1.7. : Les lieux d'articulation des voyelles courtes de l'Arabe Standard [17].

La langue arabe comporte plusieurs caractéristiques phonétiques et phonologiques. On peut citer quelques caractéristiques :

- le tanwiin : le signe de tanwiin est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination et placé en début de mot. Les symboles de tanwiin au nombre de trois et sont constitués par dédoublement des signes diacritiques ci-dessus, ce qui ce traduit par l'ajout du phonème [n] au niveau phonétique, Exemples :
  - [an] : signe ً (بًا [ban])
  - [un] : signe ٌ (بٌ [bun])
  - [in] : signe ٍ (بٍ [bin])

- [almad], les voyelles longues sont caractérisées par une partie stable plus allongée que la partie stable des voyelles courtes et cela sur le plan acoustique (Figure 1.8.).

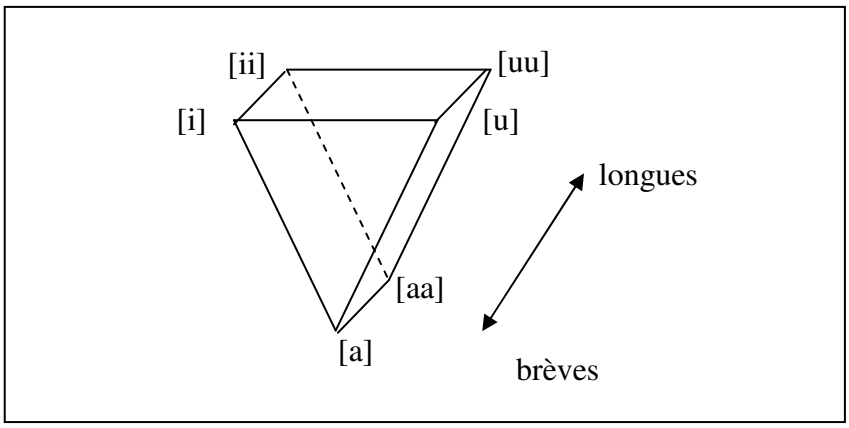


Figure 1.8. : Système vocalique de l'Arabe Standard [6].

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 le cas de /تفخيم/ [tafxiim] ou la présence des Consonnes Emphatiques qui reportent en arrière le point d'articulation des voyelles [a, u, i, aa, uu, ii] de sorte qu'elles deviennent [a, o, e, aa, oo, ee] [6].

On peut donner des exemples d'almad les deux mots :

- « جنات» [zan**aa**t] qui signifie les paradis ;
- « جَمَال » [zam**aa**l] qui signifie beauté.

- les phonèmes arrière : 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 [23] :

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

- la chadda ou la gémination: le signe de la chadda peut être placé au-dessus de toutes les consonnes en position non initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiques (dédoublement de consonnes identiques).

  En ce qui concerne la sémantique, la gémination peut changer totalement le sens des mots, exemple:

  حَضّرَ التلميذ [HaDDara] qui signifie l'élève a préparé;

  حَضَرَ التلميذ [HaDara] qui signifie l'élève est présent.

- l'emphase, On trouve dans le saint coran 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 [17].

  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 [24].

Il faut savoir que la notion d'emphase est applicable sur les consonnes ainsi que sur les voyelles telles que [17] :

- une consonne est emphatique si et seulement si elle appartient à l'ensemble /ظ, ص, ط, ض / et elle est emphatisée selon le contexte dans certains mots si elle appartient à /ل , ق /
- une voyelle est emphatisée si et seulement si elle est voisine proche d'une consonne emphatique, c'est le phénomène de la coarticulation. la voyelle est influencée par la consonne emphatique.

La figure 1.9 représente quelques caractéristiques de L'AS que nous avons évoquées.

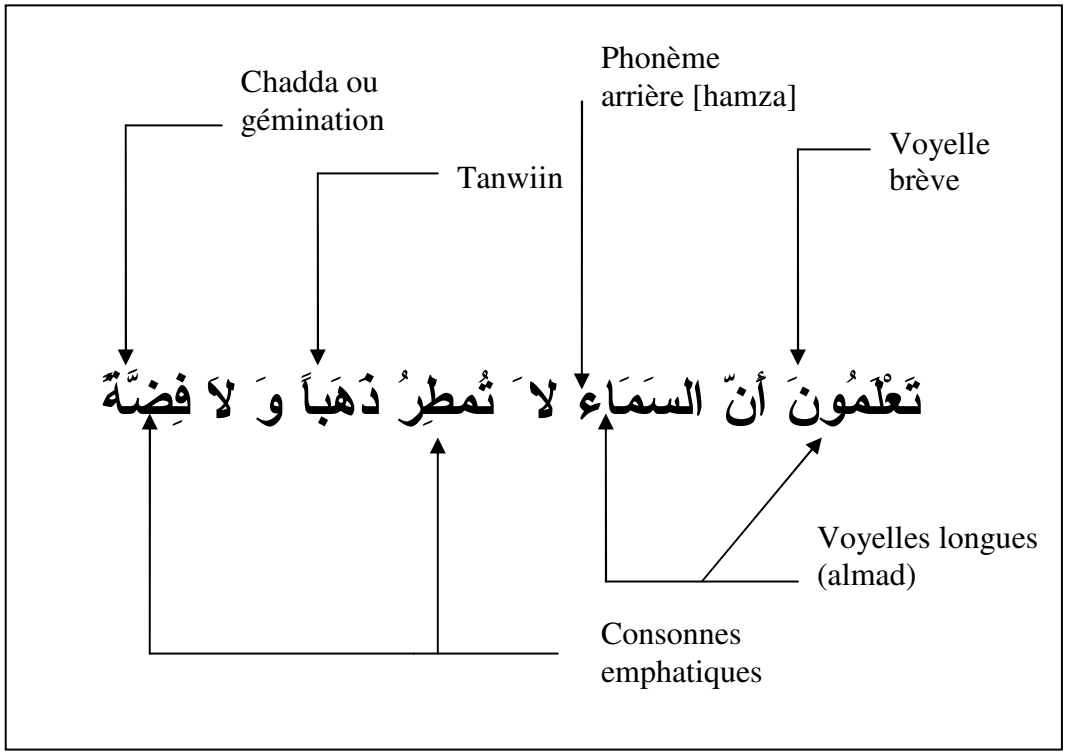


Figure1.9. : Exemples de quelques caractéristiques de l'AS.

1.10. Conclusion

Nous avons présenté dans ce chapitre quelques notions sur le signal de la parole ; qui vont nous servir dans le cadre de notre travail, nous avons mis l'accent sur les caractéristiques phonologiques et phonétiques de L'Arabe Standard. Par ses propriétés syntaxiques, phonétiques et phonologiques, on peut dire que la langue arabe est considérée comme faisant partie des langues difficiles à appréhender dans le domaine du traitement automatique du langage parlé et écrit.

Les principes cités dans ce présent chapitre sont fondamentaux pour appréhender les différentes parties qui suivent. Le but dans le chapitre suivant est de présenté les systèmes de synthèse à partir du texte et les différentes méthodes utilisées pour la génération artificielle de l'onde de la parole.

CHAPITRE 2
METHODES ET TECHNIQUES DE LA SYNTHESE VOCALE

2.1. Introduction

Notre objectif dans ce chapitre est de faire une étude approfondie des systèmes de synthèse de la parole ; pour cela nous allons tout d'abord présenter l'historique de la synthèse de la parole ; avec un rappel sur les principales périodes d'inventions des synthétiseurs de la parole. Nous exposons ensuite le fonctionnement général des systèmes TTS et les méthodes employées pour la génération du signal acoustique synthétique, ainsi sans oublier d'octroyer quelques domaines d'applications pour la synthèse de la parole et quelques systèmes TTS qui existent à l'heur actuelle.

2.2. La synthèse vocale

La synthèse vocale est la génération automatique, par des dispositifs matériels et/ou des algorithmes, de parole artificielle. Plusieurs méthodes de synthèse vocale existent, nous pouvons citer à titre d'exemple la synthèse de la parole à partir de texte (Text –To- Speech) dont le but est de produire de la parole à partir d'un texte a priori inconnu (figure2.1).

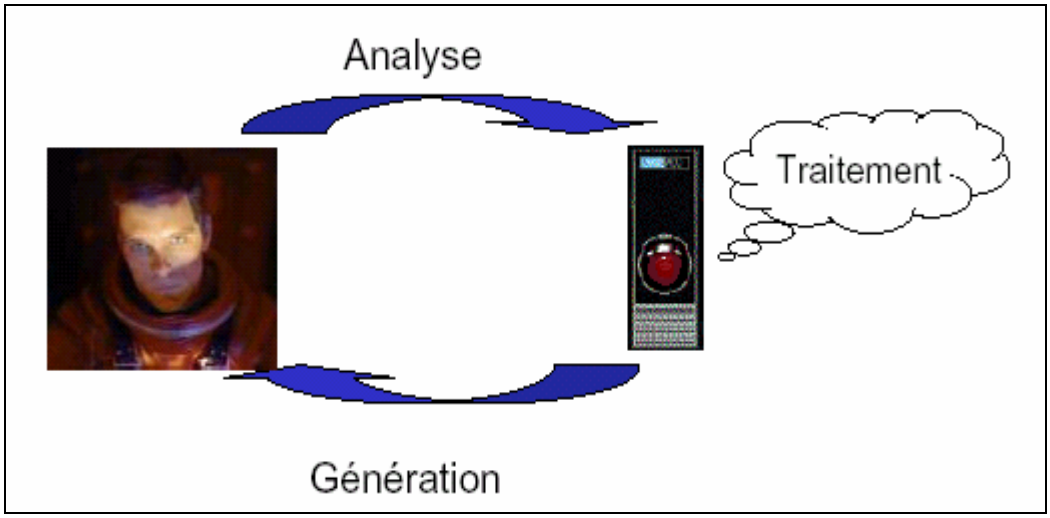


Figure 2.1. : Analyse de la parole humaine en vue de sa reproduction artificielle

2.3. Historique de la synthèse de la parole

Historiquement la synthèse de la parole à commencer à partir d'une vieille idée « *faire parler le non-vivant* ». Cette dernière à pousser l'être humain à faire plusieurs tentatives pour concevoir des machines parlantes. Les premières tentatives pour la reproduction de la voix humaine sont des synthétiseurs mécaniques. C. Kratzenstein professeur de physiologie à Copenhague réussit à synthétiser quelques voyelles en utilisant des tubes résonants connectés à des pipes (des dispositifs mécaniques équipés de soufflets et d'anches vibrantes). La machine construite en 1885 par J. Faber présentait un modèle de la langue et une cavité pharyngale de forme contrôlable et peut ainsi être considérée comme un progrès dans la synthèse. Aux environs de 1937, R. Riesz (USA) construit une machine similaire aux deux précédentes avec un conduit vocal de forme proche du naturel. La figure 2.2. montre le Perfectionnement de la synthèse mécanique de Faber à Riesz

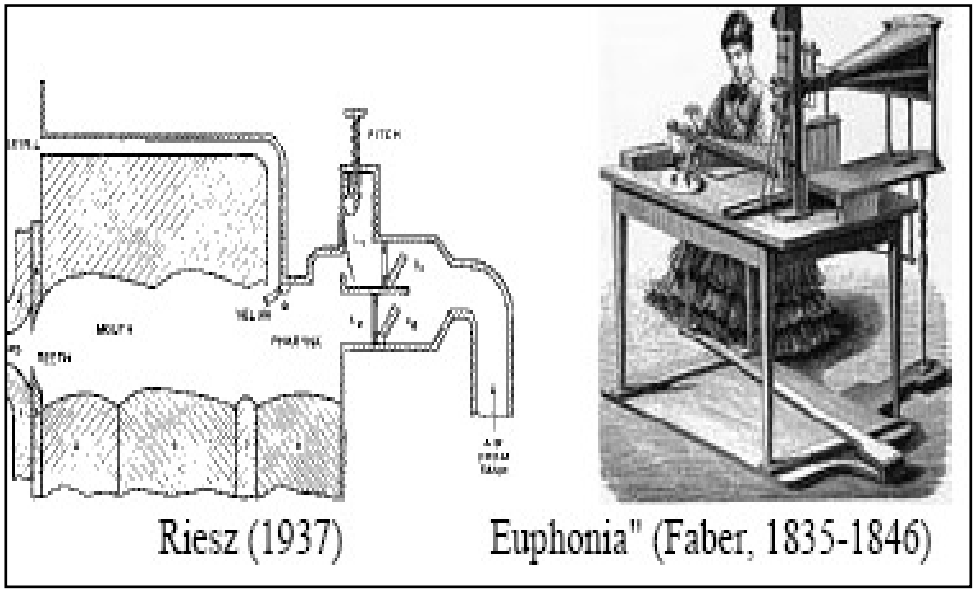


Figure 2.2. : La synthèse mécanique de Faber à Riesz [14]

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, 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 dans les laboratoires Bell (figure 2.3.). 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 F. Cooper dans les laboratoires Haskins aux USA. Cette machine servit

principalement aux recherches dans la perception de la parole et fonctionne à l'inverse du spectrographe, c'est-à-dire que les contours d'énergie du sonagramme (naturel ou artificiel) sont convertis en ondes de pression audibles [30]

Récemment, avec la diffusion des potentialités de l'ordinateur, la popularisation d'Internet et l'émergence de la Société de l'Information, des progrès considérables ont été accomplis en matière de synthèse de la parole, ce qui se confirme par le nombre important d'applications de la synthèse vocale disponible sur le marcher du logiciel. Malgré ce grand nombre d'application vocale ; il serait faux de croire que la synthèse de la parole constitue une technique entièrement maîtrisée. Cependant, les travaux de recherche menés depuis des années ont permis d'atteindre une qualité qui se rapproche de plus en plus de la voix naturelle.

Nous présentons dans le paragraphe suivant les principales modules qui composent les système de synthèse vocale actuel.

Figure 2.3. : Le système VODER [31].

2.4. Architecture générale d'un système de synthèse à partir du texte

La Figure 2.4 donne le diagramme fonctionnel d'un synthétiseur TTS. On y retrouve le module de Traitement du Langage Naturel, capable de produire la Transcription Orthographique Phonétique du texte à lire, et d'y associer une prosodie aussi naturelle que possible. Ce module génère des informations symboliques. Ces dernières sont transmises au module de traitement du signal, homologue fonctionnel de l'appareil phonatoire, qui transforme cette information symbolique en un signal acoustique de parole.

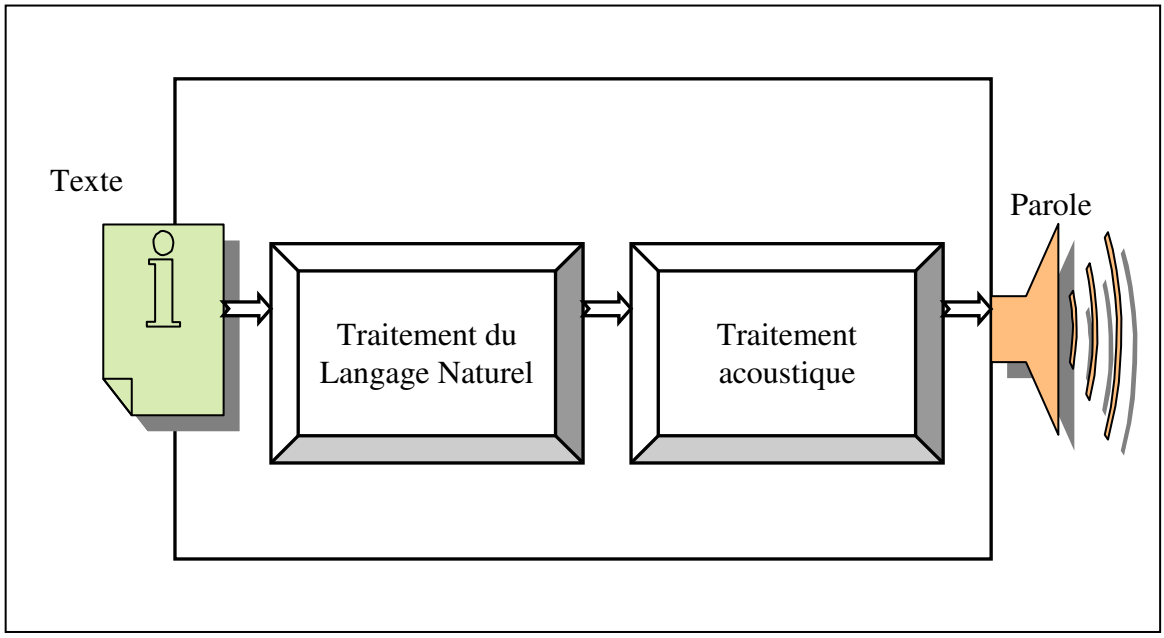


Figure 2.4. : Diagramme fonctionnel d'un synthétiseur TTS

Les deux modules cités précédemment qui composent tout système de synthèse à partir du texte sont eux-mêmes composés de plusieurs sous-modules qui sont indispensables et complémentaires entre eux. Nous pouvons décrire le fonctionnement général de chaque module ci-dessous :

- avant de faire la Transcription Orthographique Phonétique d'un texte on est amené de passer par plusieurs étapes : la première étape est de repérer et de traiter les « anomalies » ou « noms, chiffres, unités » du texte, La deuxième étape est de générer la prononciation du texte. Il y a une différence importante entre la façon dont le texte est écrit sous forme de lettres et la façon dont il va être prononcé. En élaborant une série de règles de prononciation. C'est ce qu'on appelle la phonétisation, ou transcription du texte orthographique sous forme de

texte phonétique. Quand on sait comment prononcer le texte, on s'intéresse ensuite à la structure du texte, à la prosodie qu'il faut mettre sur le texte : quand est-ce que cela commence, ou s'arrête ? Quelles sont les nuances ? Quels sont les types de mélodie ou de rythmes ? C'est ce qu'on appelle la prosodie du texte ;

- le deuxième module, concerne les étapes de l'élaboration de la parole de synthèse proprement dite. Il existe plusieurs méthodes de synthèse (méthodes de génération du signal acoustique) certains parmi elles possèdent plusieurs techniques que nous voyons dans les prochains paragraphes. Le choix de la méthode et de la technique est très important puisque ces dernières influentes directement la qualité de la parole synthétique générée. Chaque méthode possède des sous-modules spécifiques. En conséquence, le concepteur de ce module doit avoir des connaissances des avantages et des inconvénients de chaque méthode; pour qu'il puisse résoudre convenablement la problématique de la synthèse de la parole.

La figure 2.5 montre l'enchaînement des étapes décrites ci-dessus et qui modélise le schéma général des systèmes de synthèse à partir du texte actuel.

Nous présentons dans la partie suivante les principales classes de méthodes pour la génération de la parole synthétique.

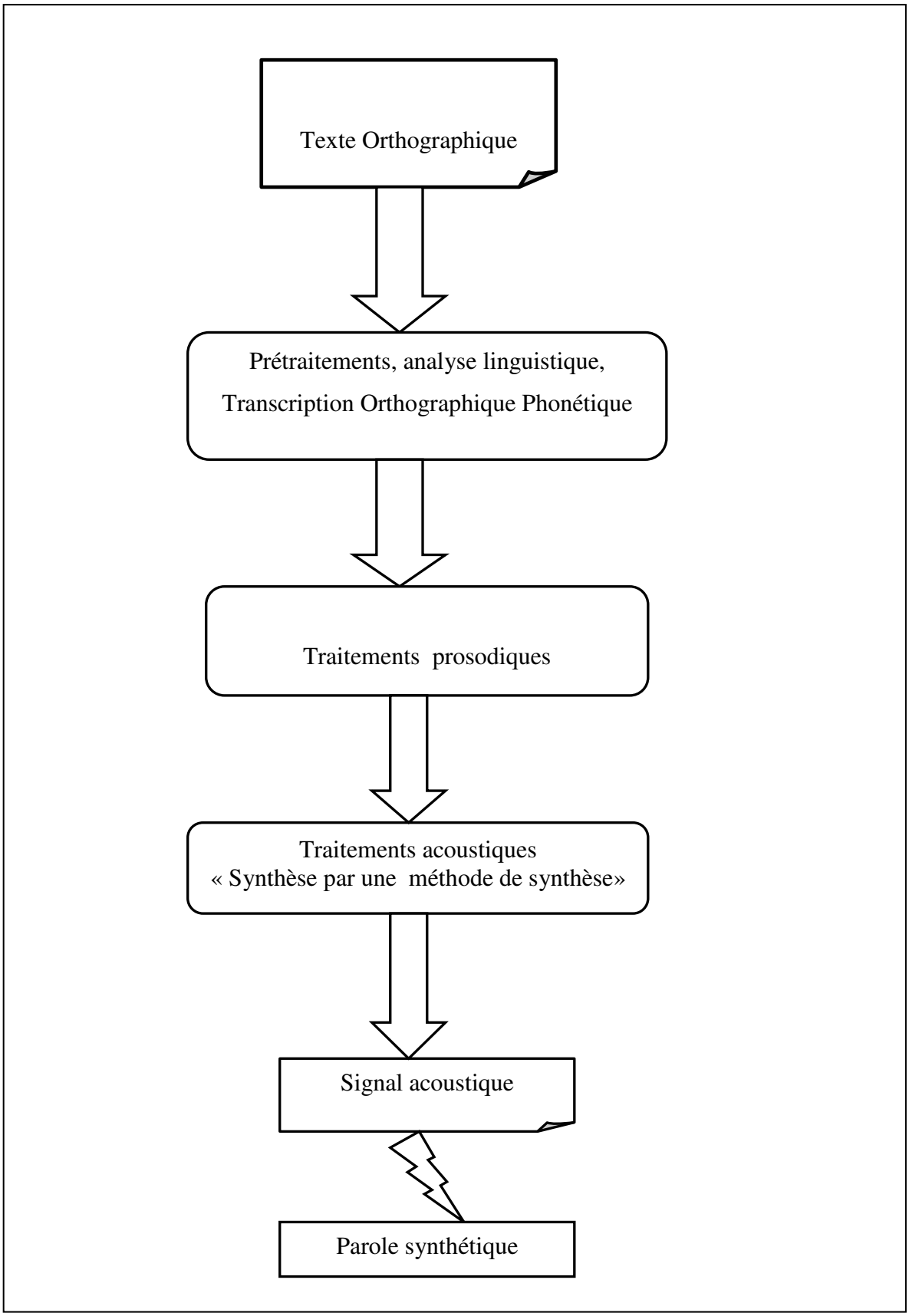


Figure 2.5. : Schéma général d'un système de synthèse à partir du texte

2..5. Les classes de la synthèse de la parole à partir du texte

Il existe trois grandes classes de méthodes pour réaliser la synthèse sonore à partir des informations phonétiques : la synthèse articulatoire, la synthèse par règles et la synthèse par concaténation d'unités acoustiques.

2.5.1. La synthèse par règles

La synthèse par règles est basée sur la modélisation de la parole à partir d'un spectre sonore. Des règles peuvent être écrites pour générer un spectre sonore artificiel. Les synthétiseurs par règles sont basés sur l'idée que, si un phonéticien expérimenté est capable de «lire» un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel (figure 2.6) pour une suite donnée de phonèmes. Les paramètres acoustiques utilisés peuvent être des fréquences formantiques, c'est-à-dire les fréquences de résonances du conduit vocal. On entend par règles, les règles d'évolution dans le temps de ces paramètres pour une unité donnée (un phonème par exemple) en fonction de son contexte linguistique. Ces paramètres doivent retracer le geste articulatoire effectué lors de la phonation avec précision pour une parole au moins intelligible (Figure 2.7). L'élaboration de ces règles est difficile et nécessite beaucoup de connaissances phonétiques.

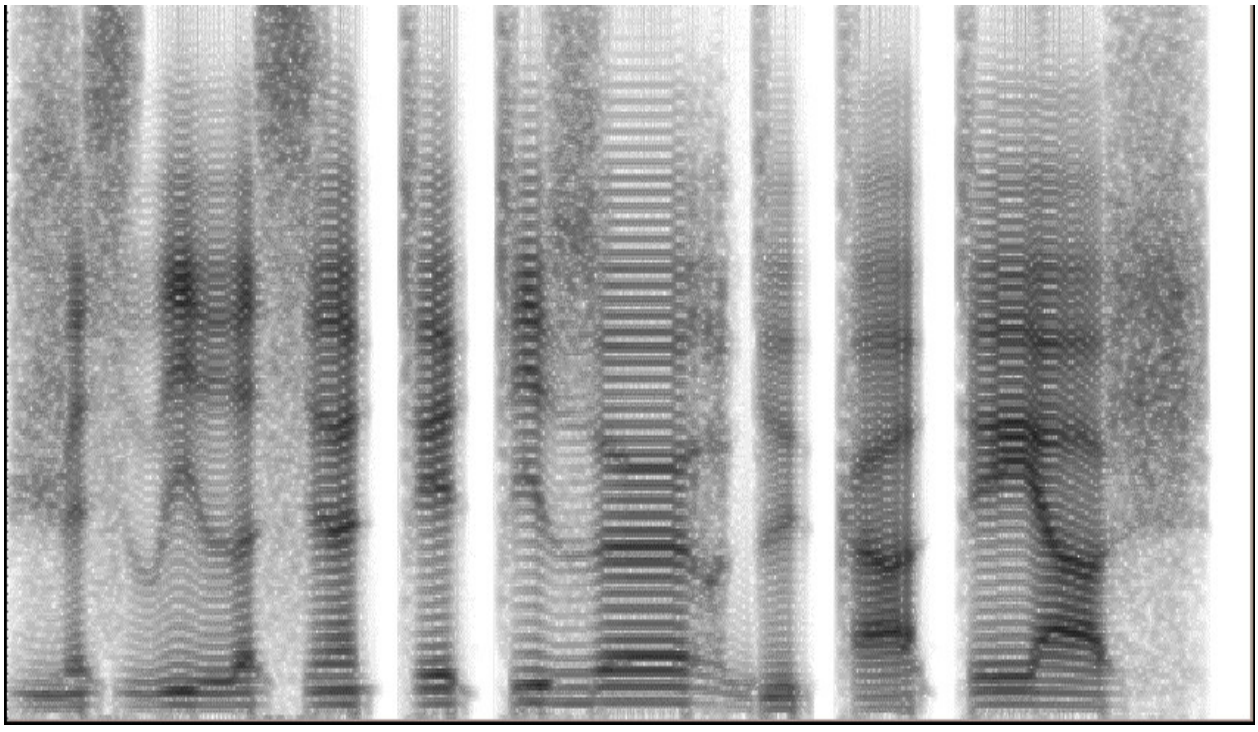


Figure 2.6. Spectrogramme d'une phrase synthétisée par règles [25].

Symbol	C/V	Min	Max	Name
DU	C	30	5000	Duration of the utterance(ms)
NWS	C	1	20	Update interval for parameter reset(ms)
SR	C	5000	2000	Output sampling rate(Hz)
NF	C	1	6	Number of formants in cascade branch
SW	C	0	1	0= Cascade, 1=Parallel tract excitation by AV
GO	C	0	80	Overall gain scale factor (dB)
F0	V	0	500	Fundamental frequency (Hz)
AV	V	0	80	Amplitude of voicing (dB)
AVS	V	0	80	Amplitude of quasi-sinusoidal voicing (dB)
FGP	V	0	600	Frequency of glottal resonator “RGP”
BGP	V	50	2000	Bandwidth of glottal resonator “RGP”
FGZ	V	0	5000	Frequency of glottal anti-resonator ”RGZ”
BGZ	V	100	9000	Bandwidth of glottal resonator “RGZ”
BGS	V	100	1000	Bandwidth of glottal resonator “BGS”
AH	V	0	80	Amplitude of aspiration (dB)
AF	V	0	80	Amplitude of frication (dB)
F1	V	180	1300	Frequency of 1st formant (Hz)
B1	V	30	1000	Bandwidth of 1st formant (Hz)
F2	V	550	3000	Frequency of 2nd formant (Hz)
B2	V	40	1000	Bandwidth of 2nd formant (Hz)
F3	V	1200	4800	Frequency of 3rd formant (Hz)
B3	V	60	1000	Bandwidth of 3rd formant (Hz)
F4	V	2400	4990	Frequency of 4th formant (Hz)
B4	V	100	1000	Bandwidth of 4th formant (Hz)
F5	V	3000	6000	Frequency of 5th formant (Hz)
B5	V	100	1500	Bandwidth of 5th formant (Hz)
F6	V	4000	6500	Frequency of 5th formant (Hz)
B6	V	100	4000	Bandwidth of 5th formant (Hz)
FNP	V	180	700	Frequency of nasal pole (Hz)
BNP	V	40	1000	Bandwidth of nasal zero (Hz)
FNZ	V	180	800	Frequency of nasal zero (Hz)
BNZ	V	40	1000	Bandwidth of nasal zero (Hz)
AN	V	0	80	Amplitude of nasal formant (dB)
A1	V	0	80	Amplitude of 1st formant (dB)
A2	V	0	80	Amplitude of 2nd formant (dB)
A3	V	0	80	Amplitude of 3rd formant (dB)
A4	V	0	80	Amplitude of 4th formant (dB)
A5	V	0	80	Amplitude of 5th formant (dB)
A6	V	0	80	Amplitude of 6th formant (dB)
AB	V	0	80	Amplitude of bypass path (dB)

Figure 2.7 : paramètres utilisés dans la synthèse par formants [24]

Les synthétiseurs par règles sont organisés comme à la figure 2.8. :

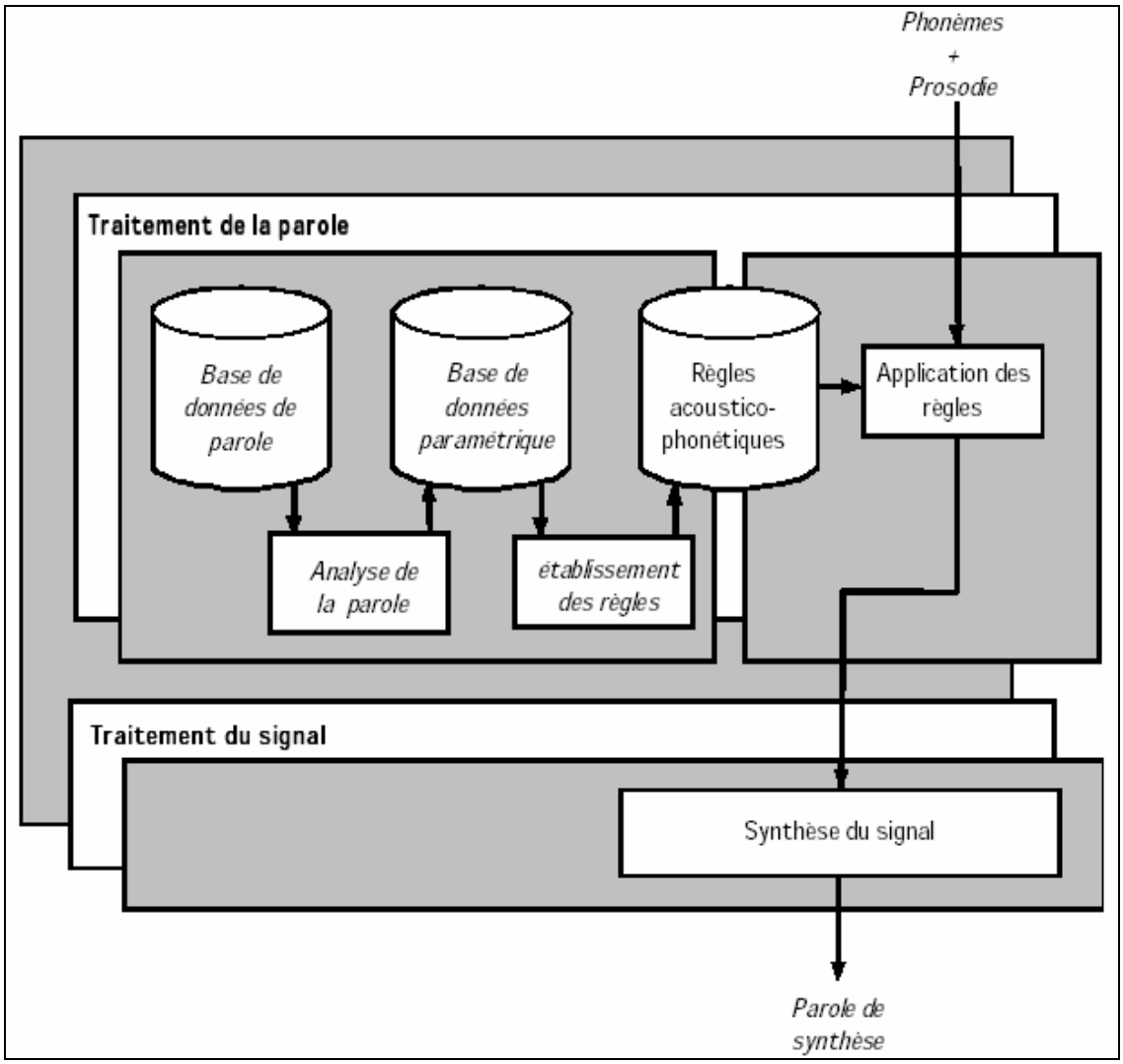


Figure 2.8. : Schéma de conception et fonctionnement typique d'un système de synthèse par règles [10].

2.5.2. La synthèse articulatoire

Cette méthode de synthèse se distingue de la précédente par rapport à l'élément étudié. Alors que la première tente de générer un signal de la parole en reproduisant son spectre par exemple, la technique de la synthèse articulatoire s'appuie sur une simulation de l'appareil phonatoire (l'appareil anatomique responsable sur la production de la parole), en modélisant la source d'excitation, les cordes vocales et les différents articulateurs participant à la production. De par sa plus grande complexité, cette approche est essentiellement une voie de recherche et elle n'est pas sortie encore des tablettes des laboratoires de recherches.

2.5.3. La synthèse par concaténation d'unités acoustiques

Les méthodes décrites précédemment rendent compte de la mécanique de l'élocution mais ne permettent pas d'atteindre une qualité acceptable de la parole générée. Pour atteindre cette qualité quasi-naturelle, il faut abandonner l'idée de générer la parole au niveau 'phonémique' par le biais d'un modèle. Ce dernier est très souvent issu d'une formulation qui ne tient pas compte de tous les phénomènes qui interviennent dans l'objet que l'on veut modéliser et de ce fait, le résultat reste loin de la réalité. Pour cette raison, Une nouvelle voie de recherche est apparue avec le progrès technologique (notamment grâce aux capacités grandissantes de stockage en mémoire de données variées). La synthèse par concaténation d'unités vocales est née. La variabilité du signal de parole peut alors être contenue dans les différents segments de signal stockés, reproduits/retransmis dans le produit de synthèse concaténée ; ceci permet en outre d'améliorer nettement le naturel du son.

Ces dernières années, on assiste à la prédominance de la synthèse utilisant des segments de signal de parole préstockés. La synthèse par concaténation d'unités préstockées, repose sur la possibilité de concaténer des segments de signal de parole en nombre suffisant pour générer n'importe quel message écrit. Au contraire, des synthétiseurs par règles, les synthétiseurs par concaténation ont une connaissance très limitée du signal qu'ils mettent en forme. La plupart de ces connaissances se trouvent stockées dans les unités vocales mises en oeuvre par le synthétiseur. Ceci apparaît clairement dans la description générale d'un tel synthétiseur sur la figure 2.10.

Le choix de l'unité de synthèse est très important au niveau d'un système de synthèse de la parole par concaténation. Nous pouvons citer :

- Synthèse par phrases : il s'agit en fait d'un simple enregistrement des phrases à restituer à la demande. Ce type de synthèse vocale ne convient que pour un vocabulaire très limité et connu à l'avance. Comme exemple d'application, on peut citer les jouets pour enfants et certains messages vocaux. Certains systèmes de synthèse utilisent des phrases et des mots comme les horloges parlantes où le message vocal se compose d'une phrase fixe du type « *l'heure actuelle est* », et d'une partie variable indiquant l'heure courante ;

- Synthèse par mots : il s'agit également d'une synthèse à vocabulaire limité et connu car on ne peut pas stoker tous les mots du vocabulaire d'une langue donnée. La qualité de la parole est moins bonne que dans le cas des phrases à cause des pauses existantes entre les mots. Ce léger inconvénient est cependant compensé par une grande souplesse dans les messages à synthétiser ;

- synthèse par phonèmes, sa première utilisation, été dans le but de permettre la génération de n'importe quelles messages inconnus (vocabulaire illimité), mais dans la parole naturelle, la perception des phonèmes comme des sons distincts est une abstraction mentale. En 1953, Harris a enregistré des séquences Voyelle-Consonnes-Voyelle [VCV] et a extrait des segments de la taille du phonème qu'il a mis bout à bout pour constituer des mots. La parole qu'il a obtenue n'était pas intelligible, donc le phonème ne peut être utilisé pour la synthèse de la parole à partir du texte; car il ne permet pas d'obtenir la dynamique de la parole du processus de production de la parole ;

- Synthèse par diphones, grâce à l'utilisation du spectrographe pour l'analyse des sons et du Pattern Playback pour leur restitution, les chercheurs du laboratoire Haskins au USA montrent l'importance de la transition entre phonèmes dans l'intelligibilité de la parole [38], d'où la notion de diphones qui est définit comme le segment qui est compris entre deux parties stables de deux phonèmes consécutifs en prenant toute la transition. Malgré ceci, 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 polyphones ;

- Synthèse par polyphones, En 1986, le CENT introduit les polyphones [38]. Le polyphone est définit formellement comme une unité aux frontières de laquelle n'apparaissent pas de phonèmes spectralement instables ou sujets à des variabilités. La concaténation de polyphones est un phénomène qui est souvent dû 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 prendre complètement le phonème avec ses voisins pour construire une seule unité de synthèse (Figure 2.9.).

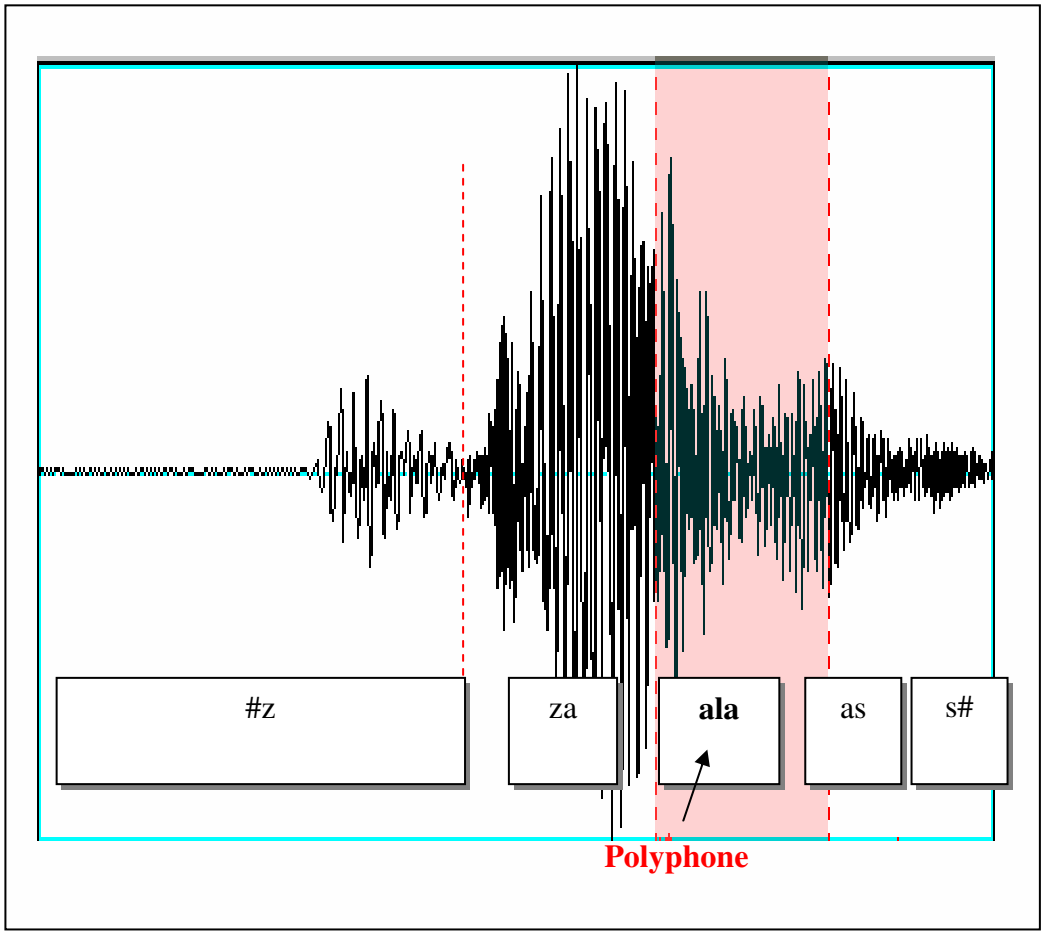


Figure 2.9 : Décomposition en polyphones du mot / جلس / [zalas]

La figure 2.10 montre le schéma général d'un système de synthèse par concaténation des segments de parole préenregistrés.

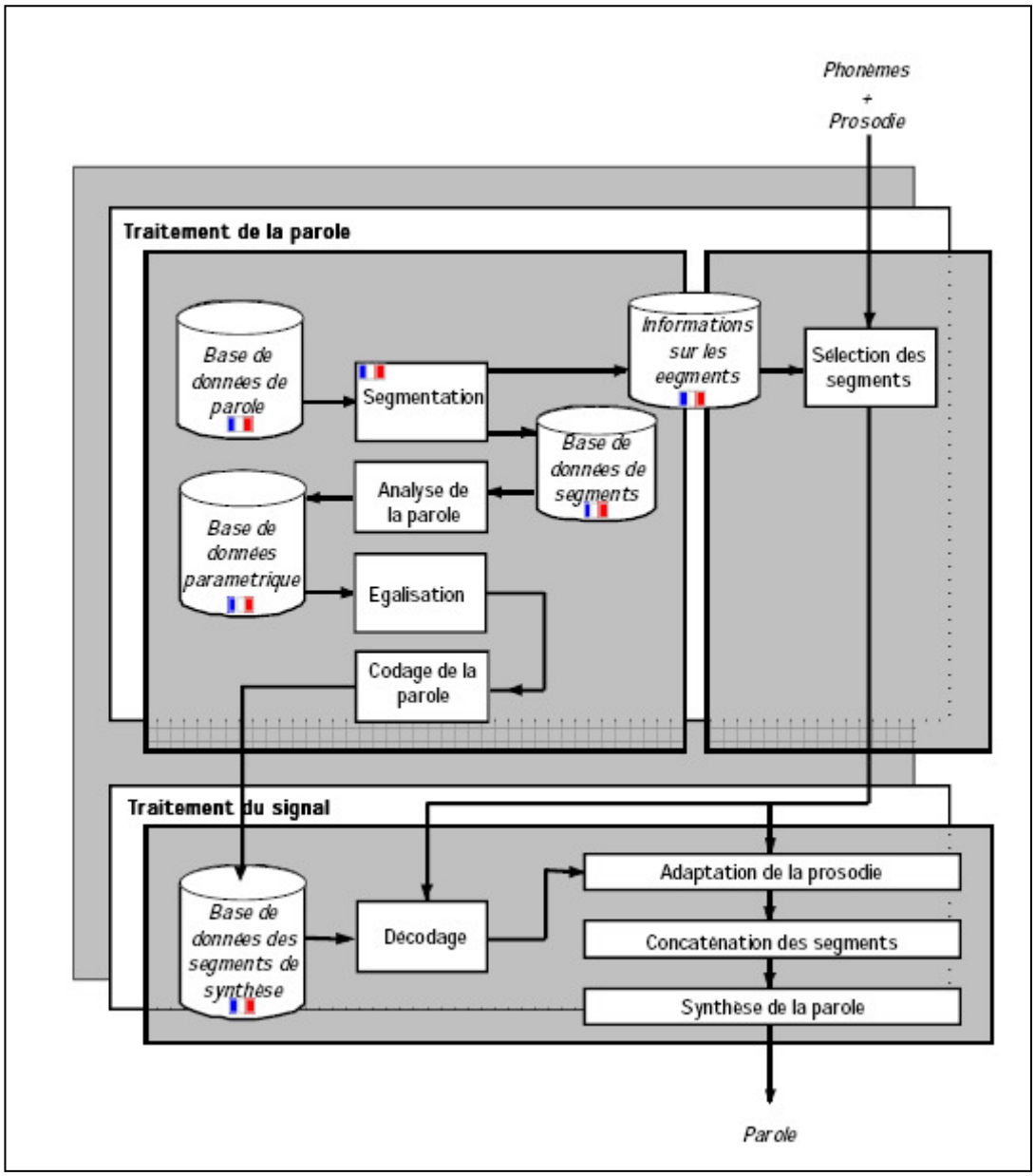


Figure 2.10 : Schéma général d'un synthétiseur par concaténation [10].

Pour la réalisation d'un système TTS utilisant la méthode de concaténation et qui fournit au moins une parole intelligible, la première des choses indispensables est la conception soigneuse de la base de données des unités acoustiques. Cette dernière est obtenue par segmentation du signal naturel. De ce fait, l'étape de la segmentation est une étape très importante pour la réussite d'un système de synthèse de la parole. Pour cette raison, la tâche de la segmentation fait l'objectif de notre étude dans le paragraphe suivant.

2.6. Segmentation

La segmentation de grands corpus est une tâche indispensable pour la mise en œuvre de nombreux systèmes de communication Homme-Machine comme les systèmes de synthèse de la parole et de reconnaissance vocale.

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

En Traitement Automatique de la Parole la segmentation consiste à couper les séquences audio enregistrées (une parole naturelle et continue) en unités acoustiques de tailles variables (phones, diphones, polyphones, mots, etc.), tel qu'on place des marqueurs temporels aux limites de ces unités phonétiques; et cela tout en mettant en correspondance le texte et l'audio (le signal de la parole) (Figure 2.11).

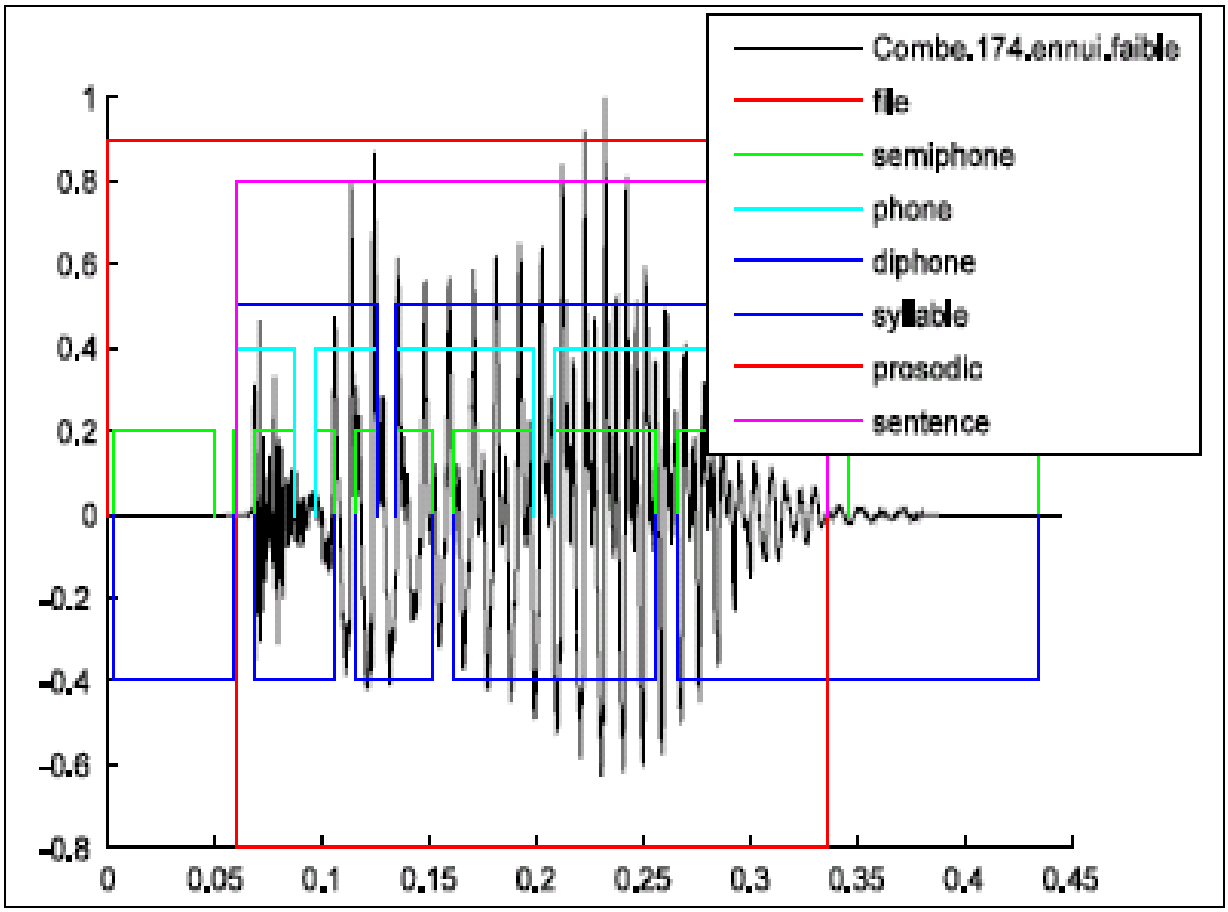


Figure 2.11. : Exemple de segmentation du mot « Comment ? »[34].

En ce qui concerne les modes de segmentation, nous distinguons deux catégories ; une manuelle et l'autre automatique :

- la première est assurée par des experts phonéticiens de la langue, une telle tâche nécessite un temps énorme et important durant l'annotation de grands corpus de parole. De plus son principale souci est dû grâce à la difficulté de bien préciser les frontières des unités segmentales, on utilise souvent des tests perceptifs répétitif pour les déterminées. Malgré le temps énorme que met la segmentation manuelle, cette dernière génère par conséquent des résultats très acceptables pour les systèmes TTS ;

- la seconde est une tâche qui est réalisée par plusieurs approches issues des laboratoires de recherche. Certaines techniques automatiques, comme l'approche basée sur les Modèles de Markov cachés (Hidden Markov Model), permettent d'acquérir une bonne précision acceptable dans certaines applications. Néanmoins, dans des applications comme la synthèse vocale, la précision de la segmentation issue de l'approche par HMM reste insuffisante et ne garantit pas une très bonne qualité de la voix de synthèse. Une autre approche générique et efficace développée par S. JARIFI & all [33] permettant de segmenter de grands corpus de parole. Cette approche est basée sur la fusion de plusieurs segmentations et permet de réduire de presque 60% les erreurs par rapport à la segmentation par HMM classique, les résultats obtenus par cette dernière approche sont acceptables mais ne garantissent pas une bonne qualité de la voix de synthèse. 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. 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.

Le principal objectif des équipes de recherche qui travaillent dans la branche de la segmentation automatique de la parole est d'avoir des résultats proches de la segmentation manuelle.

2.7. Les champs d'applications de la synthèse Vocale

Les champs d'applications des systèmes de synthèse de la parole sont nombreux. La synthèse existe là où la parole peut remplacer ou compléter une interface existante pour aider la machine à transmettre une information. Nous citons quelques unes à titres d'exemples :

- l'aide aux pour personnes handicapées : lecture d'écrans ou de documents écrits pour les non-voyants, aide à la communication vocale pour les personnes non- parlantes, laryngectomisées, la synthèse offre énormément de services et d'assistance pour eux, leur permettant d'avoir accès, sous forme vocale, aux informations écrites apparaissant sur leur écran de poste de travail. Elle peut garantir des systèmes pour commander des chaises roulantes vocalement, etc, C'est pour cela que la synthèse vocale est considérée comme un apport extrêmement important pour un public qui en a réellement besoin ;

- les applications grand public : tout appareil domestique parlant, tels que l'horloge parlante, l'appareil électroménager parlant, les 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 moyen de communication par excellence) ;

- la télématique vocale : les services de Télécommunications sont considérés comme le domaine "porteur" actuel pour l'exploitation à grande échelle des technologies vocales. serveurs vocaux d'informations (la synthèse remplaçant la parole naturelle enregistrée pour des informations rapidement évolutives et disponibles sous forme textuelle), par exemples les serveurs de lecture vocale, de FAX ou de messages électroniques, automatisation de services de prise de commande (prestation de service), automatisation de services de renseignements (Annuaires, standards d'entreprise, etc.), services de réponses pour des systèmes de vente. Telle société de vente par correspondance veut pouvoir donner par téléphone des informations sur son catalogue ;

- les outils d'enseignement : la synthèse vocale joue un rôle primordial pour l'apprentissage des langues, de rééducation, d'alphabétisation. Nous citons à titre d'exemples, 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 ;

- les applications industrielles : serveurs d'alerte, de surveillance de sites, de supervisions de réseaux, par exemple la surveillance dans un centre de contrôle industriel, dans ce cas 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. On peut retrouver aussi la synthèse vocale dans les environnements industriels comme des fonctions d'aide dans les postes de pilotage. De ce fait, la synthèse de la parole est couramment employée dans des situations où l'utilisateur d'un système informatique n'a pas le loisir de consulter un écran (cabine de pilotage d'un avion, systèmes industriels de fabrication, appareillage médical, etc). Dans ce type d'application le rôle de la synthèse de la parole consiste principalement à faire passer des informations brèves comme les messages d'aide au fonctionnement du système et les messages d'erreur de ce dernier ;

- Les systèmes embarqués représentent l'un des plus grands marchés à exploiter, où l'utilisation de la synthèse vocale s'impose naturellement. Ces combinent des parties matérielles et logiciels; le téléphone portable est un exemple type de ces systèmes ;

- 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 conséquents utiles pour la validation des théories relatives à la production, à la perception, ou à la compréhension de la parole ;

- les systèmes prototypes de dialogue vocal Homme-Machine : le couplage de la commande vocale avec la génération automatique d'énoncés (Figure 2.12)

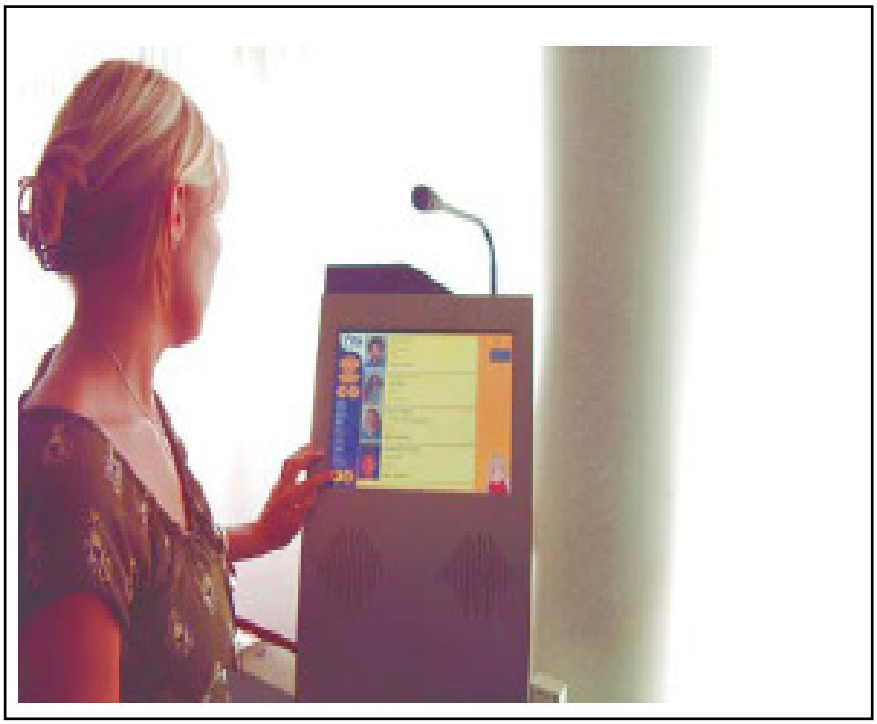


Figure 2.12 : Borne interactive *SpeechKiosk* à interface vocale [37]

2.8. Quelques systèmes de synthèse vocale

On assiste depuis quelques années, à l'émergence de nombreuses applications intégrant des systèmes de synthèse de la parole. Les premiers systèmes étaient pour la plupart câblés dont la mise en œuvre était chère et ardue. Avec l'apparition de calculateurs, la plupart des synthétiseurs actuels sont des logiciels, plus faciles à configurer et à mettre à jour et moins coûteux que leurs homologues câblés. Pour certaines applications spécifiques (serveurs vocaux ou applications embarquées), des implémentations matérielles sont encore souvent nécessaires.

La liste des systèmes de synthèse à partir du texte donnée ci-après rassemble quelques systèmes de synthèse les plus diffusés à travers le monde. Une liste plus exhaustive qui comprend notamment des logiciels en freeware (libre) peut être consultée grâce au réseau Internet sur le serveur du Center Spoken Language Understanding [32]. La plupart des systèmes de lecture automatique qui existe actuellement sont des systèmes multilingues, c'est-à-dire qu'ils sont capables de produire des voix de synthèse dans plusieurs langues différentes. Ces systèmes incluent tous, la synthèse de l'Anglais ; Cependant ne fournissant qu'un support très marginal pour les autres langues, ou dans la majorité des cas il n'y a aucun support pour une langue donnée. Nous allons citer ci-dessous quelques systèmes de synthèse de la parole.

2.8.1. Elan Speech

Elan speech est une société française spécialisée dans les technologies vocales, très connue sur le marché européen. Elan propose plusieurs produits à partir de deux technologies de synthèse multilingue. Elan Sayso est la toute dernière technologie de synthèse à partir du texte développée par Elan et est basée sur la sélection et la concaténation d'unités non uniformes. Cette technologie permet une voix de synthèse beaucoup plus naturelle. L'autre technologie Elan Tempo, est basée sur la concaténation de petites unités acoustiques (principalement des diphones) et permet d'avoir un système compact en termes de mémoire mais la qualité de la parole est inférieure par rapport Elan Sayso (au point de vue de l'intelligibilité et du naturel) . Elan Speech inclut la synthèse de l'Arabe Standard mais malheureusement cette dernière est développée par la technologie Elan Tempo.

2.8.2. AT & T Bell Labs

AT & T Bell labs est considéré comme l'un des laboratoires les plus ancien en synthèse de la parole, depuis sa démonstration du VODER en 1939. Actuellement AT & T Bell labs Propose un système de synthèse multilingue, « *AT & T Natural voice Text-To- Speech engine* » utilisant une approche par sélection d'unités non uniformes permettant ainsi une excellente qualité. Cinq langues sont actuellement disponibles (Anglais américain, Allemand, Français, Espagnol d'Amérique latine et l'Anglais britannique).

2.8.3. Festival

Un système de lecture à partir du texte développé par Black et Taylor au CSTR (Center for Speech Technologie Reaseach, le centre pour la recherche en technologie de parole) ; de l'université d'Edimbourg et en coopération avec CHATR au Japon. Festival est un outil qui est développé sous Unix et il supporte plus d'une langue naturelle. Il peut synthétiser l'Anglais et l'Espagnol. Le système utilise la technique LPC et PSOLA ainsi une base de données acoustiques de diphones du groupe MBROLA.

2.8.4. Loquendo

Loquendo est issu du groupe Telecom Italia et bénéficie ainsi de leurs nombreux travaux de recherche menés depuis des années 1970 en technologies vocales. Actuellement Loquendo est une grande société mondiale de technologie vocale, et a développé un ensemble de voix de synthèse de très bonne qualité à partir d'une approche par sélection d'unités non uniformes. Loquendo propose actuellement quatorze langues et vingt et une voix différentes. Les langues actuellement proposées comprennent l'Anglais britannique et américain, l'Italien, le Castillan, le Français, l'Allemand, le Portugais brésilien, le Portugais, le Mandarin, le Grec, l'Espagnol mexicain, l'Espagnol chilien, l'Espagnol argentinien, le Catalan ainsi que le Suédois, et d'autres langues s'ajouteront bientôt à cette liste.

2.8.5. MBROLA

MBROLA est un système de synthèse vocale de dix langues différentes. Il se base sur la concaténation de phonèmes. Des outils et des bases de données pour le développement de systèmes TTS multilingues ont été récemment, et de façon indépendante, mis à disposition par quelques universités et centres de recherche européens. Parmi eux, la Faculté Polytechnique de Mons (FPMs) a contribué au développement de synthétiseurs multilingues «phonèmes vers parole» sous la forme du projet Internet MBROLA [35]. La FPMs a décidé d'étendre ce projet au développement d'un système TTS multilingue sous la forme du projet Euler. Il s'agit d'un projet de recherche et de développement qui vise à intégrer progressivement les résultats d'autres projets de recherche tant en synthèse de parole qu'en traitement du langage naturel. L'objectif principal de ce projet est de réunir, grâce à une collaboration internationale, un ensemble de ressources homogènes pour la réalisation d'un synthétiseur de parole multilingue libre de droits pour utilisation

non commerciale dans le plus grand nombre de langues et dialectes possibles, et ceci pour Windows, Linux, et Macintosh [36].

2.8.6. ScanSoft

ScanSoft est une grande entreprise américaine spécialisée en technologies vocales. Elle fait des partenariats avec plusieurs laboratoires et entreprises spécialisées dans les technologies vocales ce qui permet le partage des données et technologies avec ces entreprises. Ainsi ces partenariats font des progrès considérables pour ces produits et solutions vocales. ScanSoft propose maintenant toute une famille de produits de synthèse de parole multilingue ; et cela après l'achat des travaux Lernout & Hauspies, sachant que leur système de synthèse propose plusieurs langues dont l'Arabe.

2.8.7. Speech Dispatcher

Speech Dispatcher est une collection de synthétiseurs vocaux qui travaillent en collaboration, et cela pour exploiter les avantages de chaque synthétiseur.

2.8.8. Infovox

Infovox est un système embarqué qui contient une partie câblée 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).

2.8.9. 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.

2.8.10. HADIFIX

HAlbsilben, DIphone, SufFIXe, est un système de synthèse pour l'Allemand développé à Bonn. Il supporte deux types de voix (masculine et féminine) et permet le contrôle de paramètres tels que la durée, le pitch, le rythme, etc. Il utilise la concaténation de diphones, demi-syllabes et suffixes. 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.

2.9. Quelques Travaux antérieurs dans les systèmes TTS en Arabe Standard

Nous nous intéresserons exclusivement dans cette partie à l'étude des travaux réalisés dans le domaine de la synthèse de la parole en langue arabe utilisant la méthode de concaténation d'unités préenregistrées.

A l'heure actuelle peu de travaux ont été développés dans le domaine de la synthèse de la parole à partir des textes arabes si nous les comparons par rapport aux autres langues (Anglaise, Française, etc.). Les quelques travaux qui existent, se basent dans la majorité des cas sur le même principe « l'utilisation majoritaire du diphone comme unité de base », rappelons que le diphone est le segment qui est compris entre deux parties stables de deux phonèmes consécutifs en prenant toute la transition. Nous donnons ci-dessous quelques travaux réalisés :

- les premières études à avoir utilisé le diphone arabe comme unité de base on été mené respectivement à l'école Nationale Polytechnique d'Alger en collaboration avec le centre national d'études des télécommunications en France par M. GUERTI [6] ;
- les travaux effectuer dans le cadre d'une thèse par S. BALOUL [1] en collaboration avec la société française Elan speech se base aussi sur le diphone ;

Ces travaux antérieurs cités ci-dessus se basent sur la même unité acoustique de base « le diphone ». Mais le diphone n'est pas totalement adapté à la synthèse de la parole ; du fait que son utilisation peut engendrer des confusions dans la perception de certains sons ou groupes de sons persistent encore. Pour l'Arabe Standard, il apparaît que certains groupes consonantiques demeurent incorrectement perçus par des auditeurs naïfs. Ces défauts sont dûs à la grande variabilité de certaines consonnes comme les liquides [l] et les semi-voyelles [w, j].

- Les travaux de K. BENBLILI [7] utilise le polyson comme unité de base c'est-à-dire des unités de taille variables, les unités utilisées sont des diphones dans tous les types de consonnes à part les liquides et les semi-voyelles ou il utilise des triphones. Cette solution a permis de résoudre quelques cas qui posent des problèmes mais pas tous les cas, puisque il n'a pas prévu le cas de succession de deux consonnes transitoires (par exemples une semi-voyelle suivi d'une liquide) dans ce cas il faut utiliser une unité plus grande que le triphone ;

- Les travaux de T. SAIDANE & all [2] s'intègre dans le cadre du projet intitulé « *Oréodule* » : un système embarqué temps réel de reconnaissance, de traduction et de synthèse de la parole. l'objet de leur intérêt [2] dans le projet cité précédemment est la contribution à la réalisation d'un système de synthèse de la parole arabe et plus précisément du volet du traitement acoustique. Les unités acoustiques choisies sont de taille variable de trois types : phonème, diphone, triphones. Ces unités sont choisies d'une façon arbitraire tel que [CVV], [CV], [CC], [C], [VV], [V]. Le principal inconvénient dans ce travail est l'utilisation du phonème comme unité acoustique de concaténation, du fait que le phonème révèle le problème de coarticulation ; le deuxième inconvénient réside dans les triphones utilisés sont du type [CVV] (c'est une Consonne suivie d'une Voyelle Longue) ce qui ne permet pas de résoudre les cas qui pose des problèmes (les liquides et les semi-voyelles).

La Particularité de notre travail par rapport aux autres travaux existant préalablement à notre connaissance, réside dans le fait que nous utilisons des unités de taille variables qui couvrent pratiquement tous les cas de transition dans l'Arabe Standard. Ce qui facilite la tâche de synthèse et donne des résultats très acceptables. Et nous implémentons aussi un nouvel algorithme de sélection dynamique dans une large base de données qui contient des unités de tailles variables (polyphones) et muti représentées.

2.10. Conclusion

Dans ce chapitre nous avons passé en revue les principales méthodes de synthèse de la parole à partir de texte (méthodes utilisées, domaines d'application). L'état de l'art pour la synthèse de la parole à partir d'un texte en langue Arabe Standard a été présenté. Enfin, nous avons présenté un ensemble non exhaustif de système de synthèse existant à l'heure actuelle.

Nous avons vu que la synthèse de la parole à partir du texte était un mécanisme complexe faisant intervenir plusieurs modules ayant des tâches particulières, eux mêmes composés de sous modules. Ainsi à cause de la complexité des signaux de la parole, il n'existe pas de solution unique à la problématique de la synthèse de la parole à partir du texte. Dans le cadre de ces travaux de mémoire de magister nous allons utiliser un système de synthèse par concaténation de polyphones. Nous verrons plus en détails les choix adoptés pour notre système dans le chapitre suivant.

CHAPITRE 3
CONCEPTION ET IMPLIMENTATION DE TALKARABIC

3.1. Introduction

Dans ce chapitre, nous présentons en détail une conception d'un système de synthèse à partir du texte en langue Arabe Standard. Ce dernier fait l'objet de notre intérêt dans ce mémoire. Cette conception a été concrétisée par la mise au point du logiciel TALKARABIC. Nous abordons les différents modules qui le compose tels que, le module de transcription, le dictionnaire de polyphones (le corpus) correspondant. Nous détaillerons les étapes de constitution de ce dictionnaire et les difficultés rencontrées lors de son élaboration. Ainsi que le module de génération du signal acoustique. Ce module commence par l'élaboration d'un algorithme de décomposition du texte transcrit (chaîne phonétique) en polyphones (la syllabation en unités acoustiques de tailles variables) selon la méthode choisie.

3.2. Paramètres indispensables pour la mise en œuvre d'un synthétiseur vocal

La mise au point d'un système de synthèse de parole à partir du texte pour une langue donnée nécessite plusieurs paramètres de connaissances parmi ces derniers :

- l'existence de connaissances suffisamment élaborées à tous les niveaux de description de la langue visée (phonétique, phonologique, etc.) ;
- la mise en collaboration des diverses sources de connaissances (linguistique, algorithmique) par des interfaces et des structures de travail normalisées ;
- la définition de la stratégie d'utilisation de ces connaissances (contrôle du séquencement des opérations de transduction texte parole) ;

- Enfin, le manque d'harmonisation dans la conception des systèmes TTS rend leur comparaison qualitative, module par module, très difficile à réaliser, ce qui exige le choix d'une méthode de conception très adaptée à la problématique de la synthèse de la parole pour restreindre le temps de développement et pour une meilleure efficacité des systèmes TTS.

3.3. Les choix adoptés pour la mise en œuvre de notre outil de lecture automatique

Pour l'élaboration de notre système de lecture automatique de texte en AS nous avons adopté les procédés suivants :

- nous avons choisi la méthode de synthèse par concaténation d'unités préenregistrées. Ce choix s'est révélé fructueux car la qualité de la parole est très acceptable. Les unités acoustiques utilisées sont de tailles variables (polyphones), du fait que la synthèse polyphonique réduit le problème de coarticulation ;

- La méthode de conception des systèmes TTS doit être adaptable à la problématique de la synthèse de la parole à partir du texte, du fait qu'on peut trouver plusieurs modules et sous modules chacun a sa propre problématique. La solution proposée dans chacun des modules possède une influence directe sur la qualité de la parole synthétique (le résultat). De ce fait nous choisissons une méthode de conception par prototypage qui est une méthode du génie logiciel très appliqué pour ces catégories de problèmes, qui possède beaucoup de retours arrière dans chaque phase (nous testons s'il est réalisable nous continuons si non nous faisons un retour arrière pour modifier la politique employée ou la stratégie adoptée). L'approche de prototypage utilisée est celle du *développement incrémental* qui consiste à réaliser dès le début du cycle de vie un sous-ensemble du produit logiciel final. Ce sous-ensemble est alors raffiné incrémentalement jusqu'à obtenir le produit final ;

- pour assurer un bon développement de notre démarche et pour obtenir une meilleure organisation nous avons utilisé les schémas de la notation unifiée du langage UML (Unified Modeling Langage). Plus exactement pour décrire le cas d'utilisation principale de notre système TTS et ainsi les cas d'utilisation des différents modules de notre système.

La figure 3.1 donne le diagramme de cas d'utilisation principal de TALKARABIC

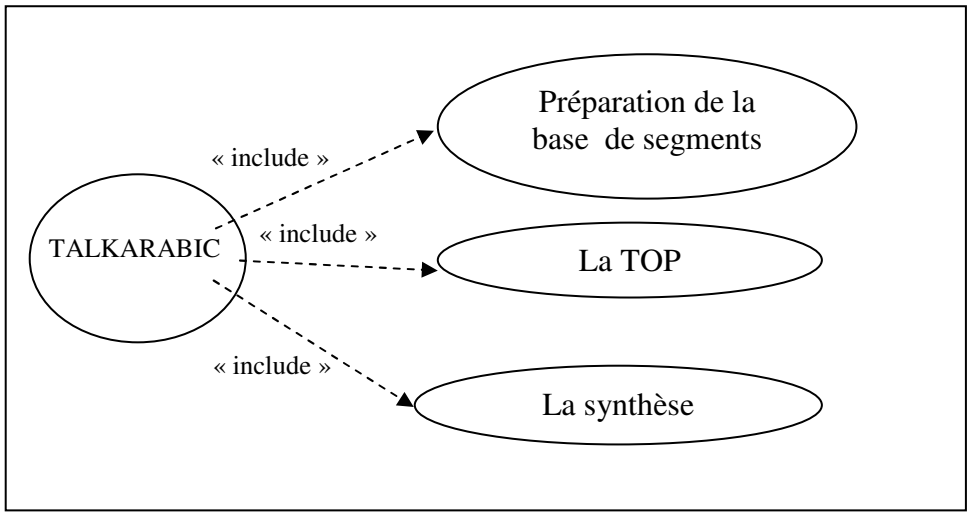


Figure 3.1 : Diagramme des cas d'utilisation principal.

Notre système de lecture automatique comprend donc trois modules essentiels qui sont indispensables pour sont fonctionnement (figure 3.2), à savoir :

- le module de Transcription Orthographique Phonétique de la langue Arabe Standard (TOP-AS) ;
- la base de données ou le dictionnaire des unités acoustiques (polyphones) ;
- le module de Traitement Acoustique (génération du signal vocal).

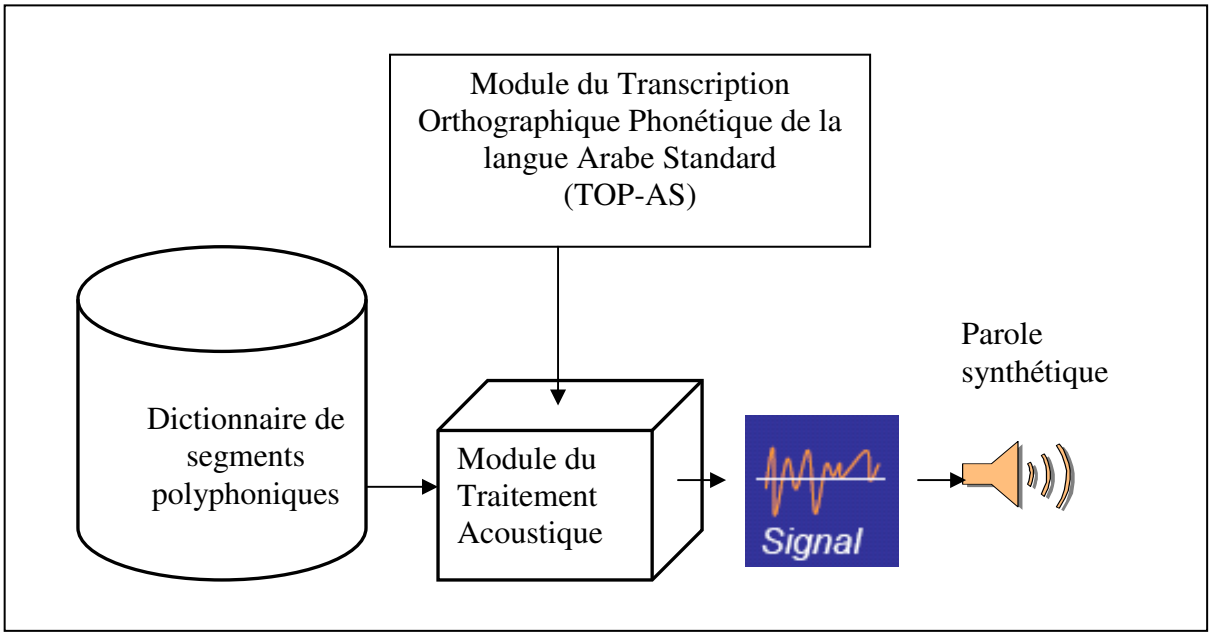


Figure 3.2 : Synoptique général du système TALKARABIC

3.4. Le module de TOP-AS

Ce module assure le passage du texte écrit en AS en texte lu. Un texte est vu comme un ensemble de chaînes de caractères que nous appelons de façon abusive mots. Ces derniers regroupent pour former des phrases et sont séparés par des caractères séparateurs qui sont :

- le caractère blanc qui sert à discerner un mot des mots adjacents ;
- la virgule ;
- le point virgule ;
- le point ;
- le point double ;
- les points de suspension ;
- le tiret délimitant une phrase intercalée ;
- les parenthèses et les crochets.

Les caractères inconnus sont les caractères qui ne figurent pas dans le texte de façon accidentelle ou délibérée.

Le diagramme de cas d'utilisation de la phase de transcription est schématisé dans la figure 3.3

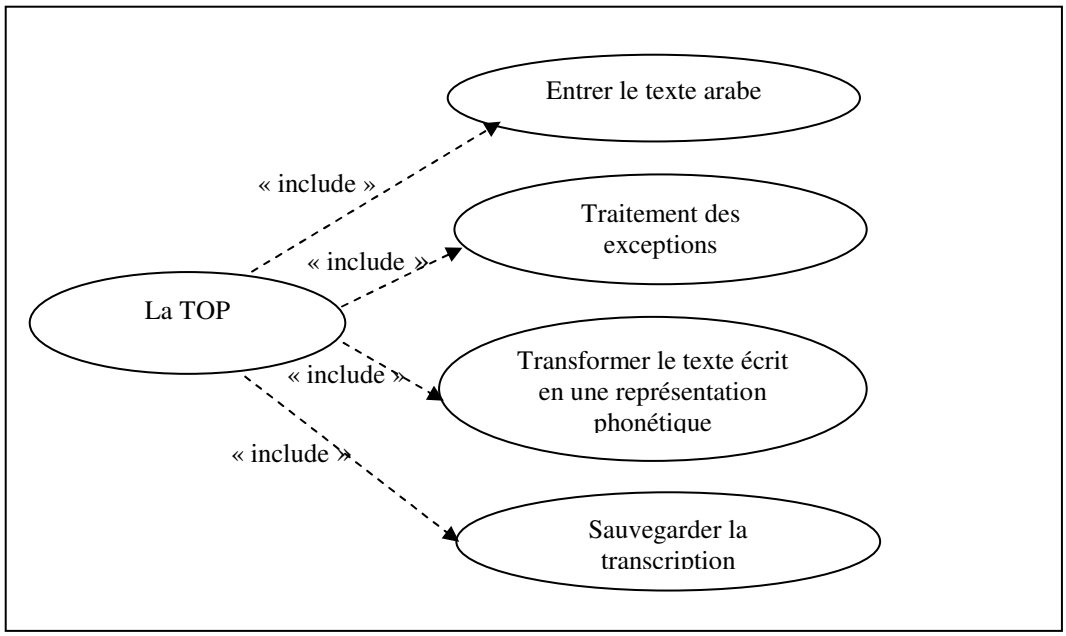


Figure 3.3. : Diagramme de use cases de cas « Transcription de texte »

Nous avons choisi un code (Tableau 3.1. et 3.2.) que nous avons utilisé pour l'identification des polyphones du dictionnaire.

Tableau 3.1 : Code proposé pour la transcription des consonnes qui composent les polyphones (Correspondance graphèmes phonèmes des consonnes de l'AS)

Harf en Arabe	Code adopté
Hamza	!
ب	b
ت	t
ث	c
ج	j
ح	h
خ	x
د	d
ذ	$
ر	r
ز	z
س	s
ش	g
ص	S
ض	D
ط	T
ظ	D
ع	v
غ	R
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
ه	H
و	w
ي	y

Tableau 3.2: Code proposé pour la transcription des voyelles qui composent les polyphones (Correspondance graphèmes phonèmes des voyelles de l'AS)

Voyelle en Arabe	Code adopté
Fatha	a
Dhama	u
Kasra	i
Fatha + elmad	A
Dhama + elmad	U
Kasra + elmad	I
Fatha + Tanwin	%
Dhama + Tanwin	§
Kasra + Tanwin	µ

Les consonnes de l'AS se prononcent toutes de la même manière, c'est-à-dire qu'à un graphème de l'alphabet correspond un et un seul phonème. Ce fait rend facile l'élaboration des règles de lecture pour la Transcription Orthographique Phonétique. La transcription par règles consiste à modéliser les connaissances linguistiques qui sont employées dans une langue par un groupe de règles de réécriture.

Les règles de lecture ne sont pas nombreuses en AS (Puisque l'AS ne comporte pas des ambiguïtés entre le texte lu et le texte orthographique).

Nous pouvons évoqué quelques règles du système TOP – AS à titre indicatif :

- « les règles de [tanwiin] » (figure 3.4);
- « les règles [almad] » (figure 3.5).

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";
   }
}
```

Figure 3.4. : Règles de [tanwiin]

Les règles [almad]

```
   If ( (grapheme[ig]== ' ٰ') && ((grapheme[ig+1]==' ا
')||(grapheme[ig+1]=='ى')) )
    {
       phoneme = phonem + "A";
       ig=ig+2;      //ig : indice de position dans la chaîne orthographique
    }

   If ( (grapheme[ig]== ' ُ') && (grapheme[ig+1]==' و') )
    {
       phoneme = phonem + "U";
       ig=ig+2;     ;  //ig : indice de position dans la chaîne orthographique
    }

   If ( (grapheme[ig]== ' ِ') && (grapheme[ig+1]==' ي') )
    {
       phoneme = phonem + "I";
       ig=ig+2;
    }
```

Figure 3.5. : Règles de [almad]

L'AS ne comporte pas des ambiguïtés entre le texte écrit et le texte lu (texte transcrit en API) sauf quelques mots d'exceptions. Un mot d'exception est un mot qui ne se lit pas conformément aux règles de lecture citées précédemment. Ces mots doivent donc être recensés et corrigés au début de la phase de traitement avant l'application des règles de réécriture. Nous rencontrons ces exceptions par exemple dans quelques pronoms et adjectifs démonstratifs (Tableau 3.3)

Tableau 3.3 : Quelques mots d'exceptions

Les mots d'exception	**Prononciation correcte**	**Transcription en API**
هذا	هاذا	[HAµA]
ذلك	ذالك	[µAlika]
كذلك	كذالك	[kaµAlika]
يأيها	ياأيها	[yA ?ayuHA]
يس	ياسين	[yAsIn]

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

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

La figure 3.6. Montre l'organigramme de la phase de Transcription Orthographique Phonétique de la langue Arabe Standard du module (TOP-AS).

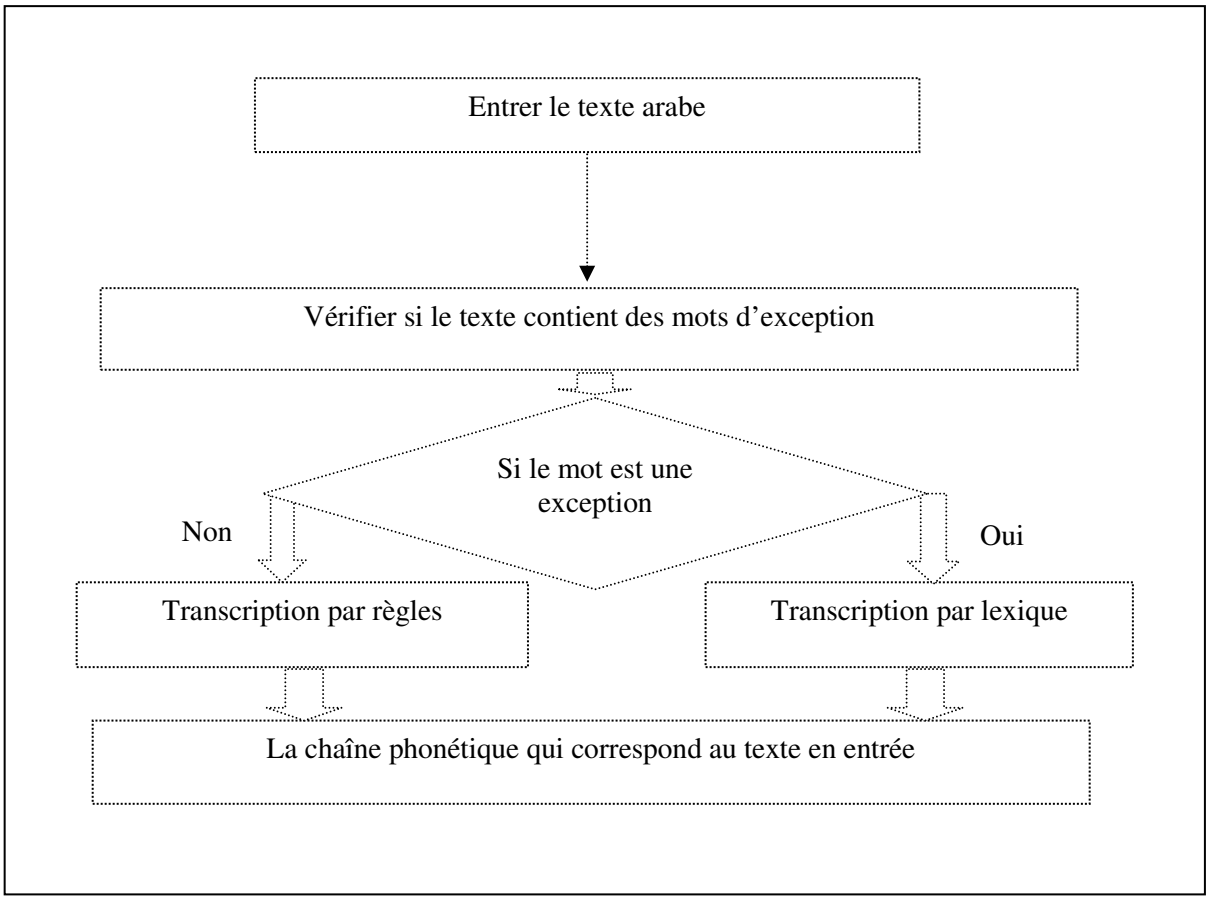


Figure 3.6 : Organigramme de la phase de Transcription Orthographique Phonétique (TOP-AS)

Une fois ces conventions adoptées, nous pouvons spéculer sur la forme littérale possible pour un mot en AS et détecter par conséquent, d'éventuelles erreurs dans le texte écrit. Un mot en AS s'écrit toujours sous la forme [CVX] où :

- [C] est une consonne car un mot en AS commence toujours par une consonne ;
- [X] est un groupement qui commence toujours par une consonne [C] et qui ne comporte jamais plus de deux consonnes qui se suivent (identique ou non). Dans ce groupement une voyelle est toujours suivie d'une consonne.

3.5. La base de données acoustiques polyphoniques

Le diagramme de cas d'utilisation de la création de la base de données acoustiques est représenté dans la figure 3.7. Nous allons présenter chaque étape de la conception d'une base de données servant de dictionnaire à notre système de lecture automatique dans les prochains paragraphes.

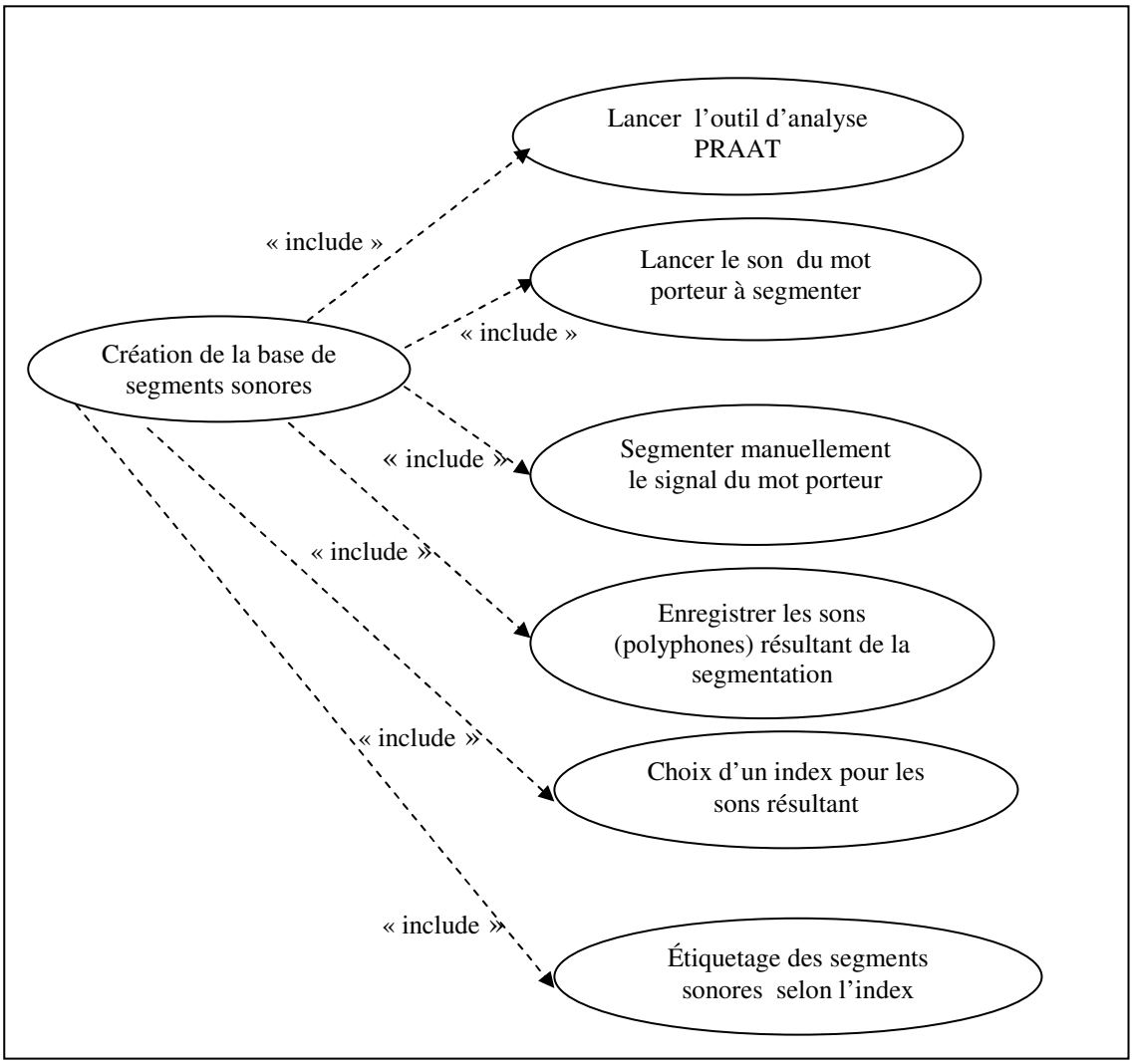


Figure 3.7. : Diagramme de use cases de cas « préparation de la base de segments sonores »

3.5.1. Corpus des mots porteurs de polyphones

Alors qu'on dispose de plusieurs centaines de millions de mots de textes écrits (et que le gigantesque réservoir qu'est le World Wide Web repousse chaque jour cette limite). On dispose de très peu de données sur l'oral. Les corpus de langue orale spontanée sont pourtant d'une importance fondamentale pour l'étude linguistique, comme pour la mise au point de nouvelles technologies vocales telle que la mise en œuvre des systèmes TTS. Des corpus oraux importants sont en cours de constitution pour diverses langues mais le développement de corpus oraux transcrits et annotés est extrêmement coûteux.

Au cours de la dernière décennie, les outils d'annotation et de traitement automatique des corpus écrits se sont fortement développés, mais l'on est bien loin de disposer d'outils équivalents pour l'oral.

Pour la réalisation de notre corpus des mots porteurs et afin de garantir une bonne qualité pour les unités constituant le dictionnaire, nous avons utilisé le logiciel PRAAT pour l'enregistrement du corpus sous un format numérique. Un mot porteur (un logatome) qui comprend le polyphone à extraire. Sa fonction assure une indépendance de l'unité par rapport à son contexte. Cette indépendance se traduit par une invariance relative dans le spectre temps-fréquence (contours formantiques pour les voyelles). Pour la dite unité. Le mot porteur est choisi de sorte que les phonèmes bordant l'unité minimisent l'effet de la coarticulation causé par les phonèmes adjacents sur cette unité. Nous avons suivi les travaux de M. Guerti tableau (3.4) pour l'élaboration de ce corpus de mots porteurs.

Exemples des logatomes et des diphones :

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

Tableau 3.4 : Exemple des logatomes contenant des diphones [15].

Logatomes	**Diphones**	**Exemples**
#katat**v#**	[v#]	#katat ـَ_# #katat ـُ_# #katat ـِ_# #katat ـَا_# #katat ـُو #katat ـِي_#
# 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#

La qualité du résultat final de la synthèse dépend directement de la qualité des enregistrements effectués lors de l'élaboration du dictionnaire d'unités acoustiques ; quelques précautions ont alors été prévues tels que :

- l'utilisation d'un seul locuteur par dictionnaire et la limitation des séances d'enregistrement ;

- le choix du locuteur pour l'enregistrement des mots porteurs (logatomes) est très important. Les logatomes de notre système sont prononcée par une locutrice algérienne ayant une bonne élocution en Arabe ; après d'avoir effectué un test lors d'enregistrement et segmentation de quelques logatomes si la voix correspond bien à une voix de synthèse ou non. Étant donné que l'étude de la qualité vocale nous permet de souligner la difficulté de déterminer a priori si une voix donnera de bon résultat en synthèse. La réussite d'une voix semble en effet ne pas dépendre des paramètres acoustiques utilisés pour caractériser les voix (F0, énergie, duré des segments) mais de grandeurs difficiles à détecter automatiquement telles les changements de registres ou plus généralement les variations du timbre de la voix. Cette dernière encourage fortement notre choix pour la méthode de conception par développement incrémentale. (Pas de temps perdu dans l'élaboration de la base de données sonores) ;

- les conditions d'enregistrements doivent répondre à des normes bien spécifiques (la non disposition du bruit dans le milieu des enregistrements, un microphone de bonne qualité, etc.).

3.5.2. Extraction de polyphones

Une fois le corpus des mots porteurs enregistrés, nous passons à la phase de segmentation. La méthode de segmentation que nous avons utilisée dans notre travail est manuelle et qui exige des connaissances phonétiques. Malgré le temps que met la segmentation manuelle, cette dernière génère par conséquent des résultats très acceptables pour les systèmes de synthèse. 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 ; du fait qu'ils ne garantissent pas une bonne qualité de la parole synthétique. Au cours de cette étape, l'identification des différentes

unités s'est faite à travers l'utilisation de l'outil d'analyse PRAAT pour visualiser la forme temporelle de l'onde acoustique correspondant à l'enregistrement, le critère de choix majeur pour la segmentation est l'audition .

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. Le but de la procédure de segmentation est d'isoler l'unité à étudier.

Exemple des différentes étapes de cette procédure avec le logiciel PRAAT :

- visualiser le sonagramme associé au son porteur de l'unité à segmenter ; (Figure 3.8.)

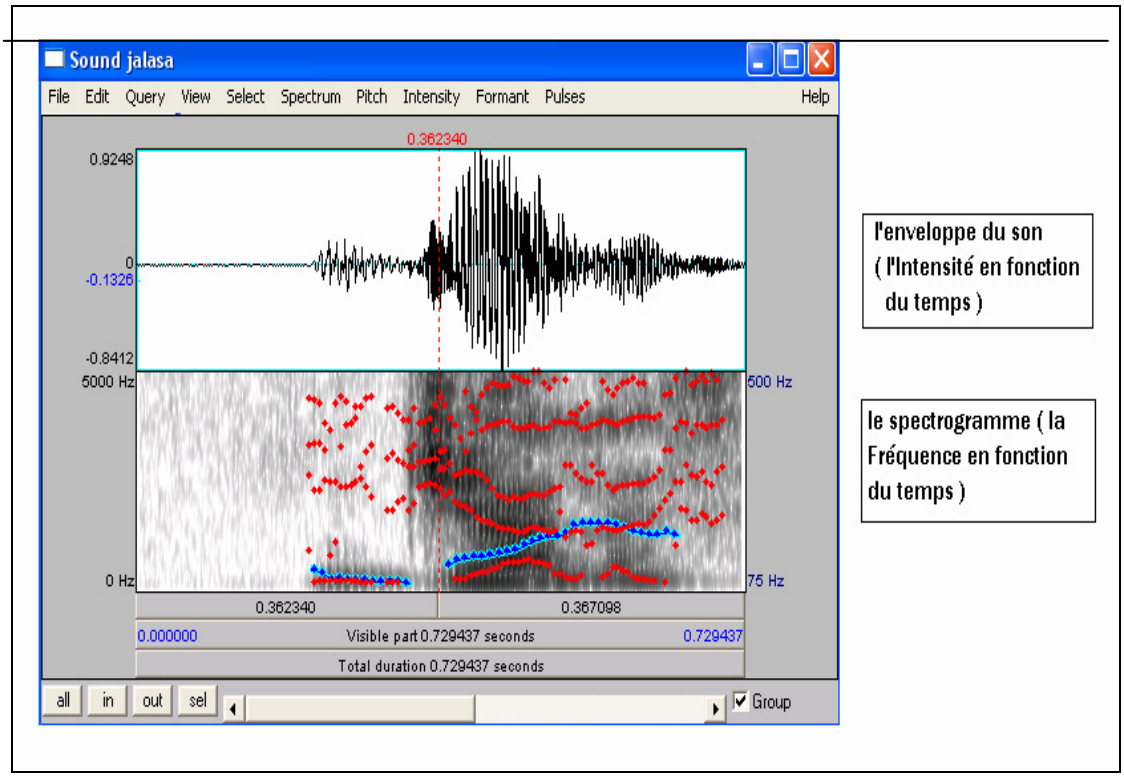


Figure 3.8.: Visualisation du son [zalasa] en entier par l'utilisation de la fenêtre *SoundEditor* de PRAAT

- 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#] ; (Figure3.9.)

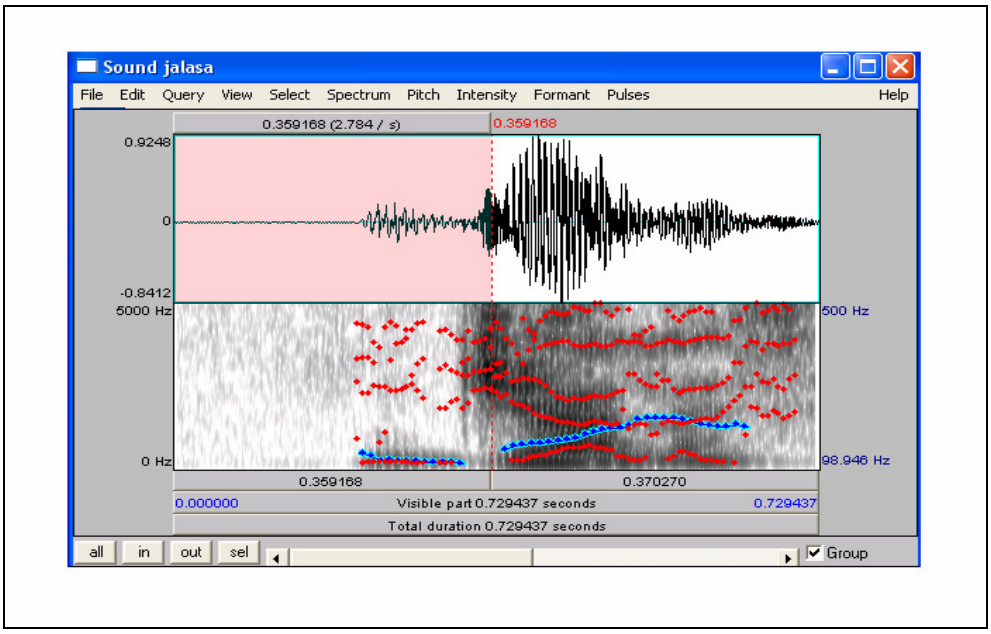


Figure 3.9. : La sélection du diphone [debut_z]

- 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 3.10.)

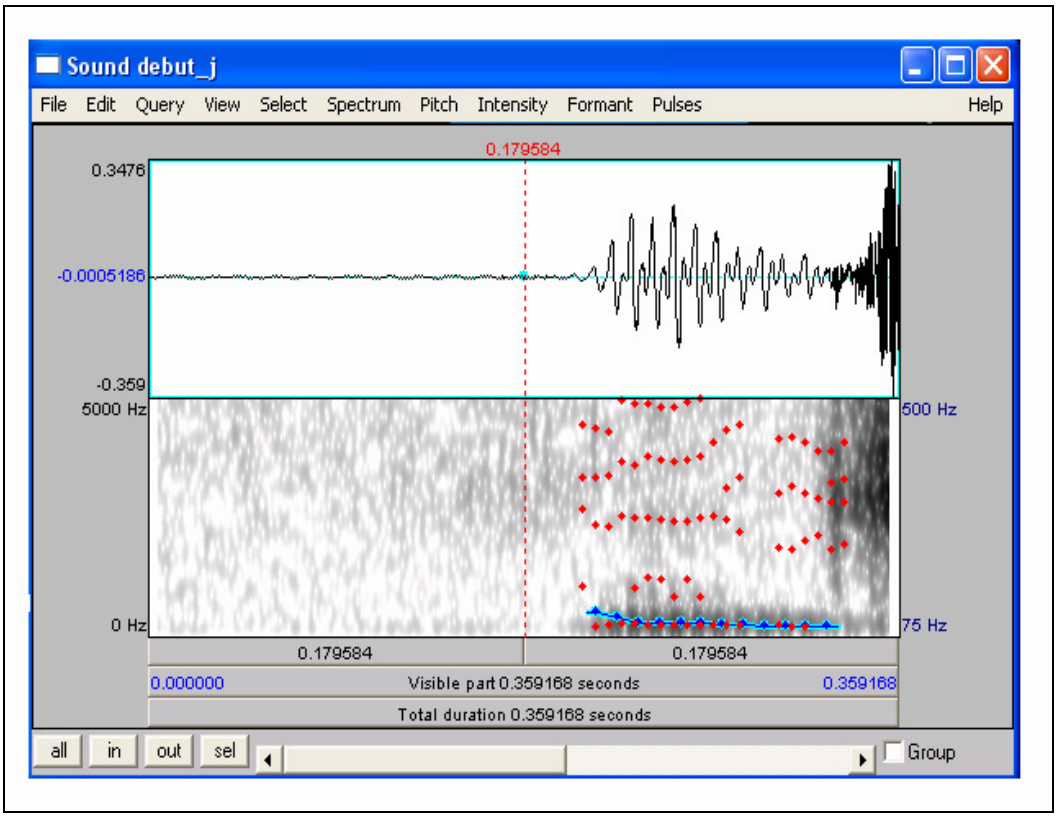


Figure 3.10. : Visualisation du diphone [debut_z] qui est résultat de la segmentation par l'utilisation de la fenêtre *SoundEditor* de PRAAT.

3.5.3. Dictionnaire des polyphones de notre système TALKARABIC

Le dictionnaire des polyphones est établi pour couvrir tous les cas de transition dans la langue Arabe Standard et ne comporte pas des combinaisons qui ne respectent pas les caractéristiques linguistique et phonétique de l'AS.

La table des segments polyphonique du système TALKARABIC est représentée dans le Tableau 3.5 dont le nombre total des unités acoustiques du dictionnaire polyphonique doit être 2620 unités pour la génération de tous les mots de L'AS.

Dans le tableau ci-dessous, [C] désigne une consonne, [V] désigne une voyelle et [T] une liquide ou une semi-voyelle, c'est dire un des phonèmes : [l], [r], [w] et [y].

Tableau 3.5.: les polyphones du dictionnaire du système TALKARABIC
(Les différentes combinaisons)

Types de polyphones	mots porteurs	nombre de réalisations
Consonne-consone CC	#ta[CC]ata#	22x22=484
Silence-Consonne #C	[#C]ata	22
Consonne-Silence C#	#Kata[C#]	22
Voyelle-Silence V#	#katat[V#]	6
Voyelle-Consonne VC	# !at[VC]a#	22x6=132
Consonne-Voyelle CV	# !a[CV]ta#	22x6=132
Voyelle-Transitoire-Voyelle VTV	# !at[VTV]ta#	6x4x6=144
Voyelle-Transitoire-Consonne VTC	# !at[VTC]a#	6x4x22=528
Silence-Transitoire-voyelle # TV	[#TV]ta#	4x6=24
Voyelle-Transitoire-Silence VT#	#Katat[VT#]	6x4=24
Voyelle-Transitoire-Transitoire-voyelle VTTV	# !atVTTVta#	6x4x4x6=576
Consonne-Transitoire-Voyelle CTV	# !a[CTV]ta	22x4x6=528

3.5.4. Index du dictionnaire et l'étiquetage des segments polyphoniques

L'indexation audio est un champ de recherche très active actuellement pour la sélection des enregistrements audio par le contenu. Le nombré élevé des segments exige en effet une organisation des données acoustiques de façon à optimiser la recherche et par conséquent la rapidité du processus de génération de l'onde acoustique.

Nous avons affecté à chaque phonème un nombre unique pour le calcul de l'index du dictionnaire polyphonique. Le calcul de l'index d'un segment acoustique se fait par la concaténation des différents codes affectées aux phonèmes qui composent la dite unité acoustique.

Exemple de calcul d'index de la chaîne phonétique «#bAb§# » (Tableau 3.6).

Tableau 3.6: Calcul index

Décomposition	#b	bA	b§	§#
Index segment1	0002			
Index segment 2		0201		
Index segment 3			0236	
Index segment 4				3600

L'avantage de notre indexation réside dans la souplesse du code phonétique mis en œuvre.

3.6. Génération du signal vocal

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.

Le diagramme de cas d'utilisation pour le module de la génération du signal vocal artificiel est représenté dans la figure 3.11.

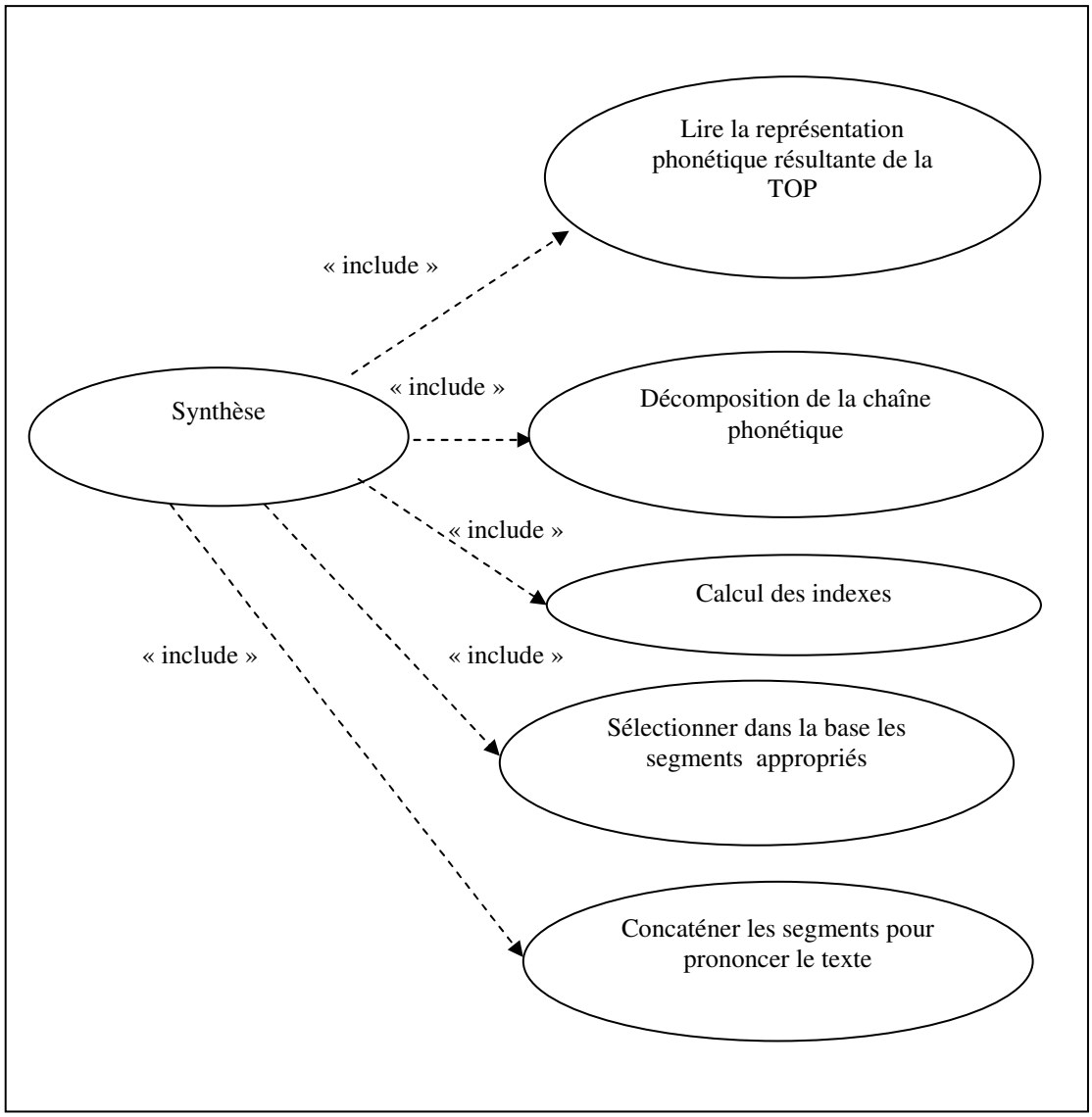


Figure 3.11. : Diagramme de use cases de cas « génération du signal vocal par polyphones»

La phase du génération du signal acoustique synthétique commence par la recherche des polyphones composant le texte à lire suivant un algorithme de décomposition. Le principe de la décomposition du texte transcrit en polyphones consiste dans la segmentation de ce texte en paire de caractères sauf dans le cas où une liquide ou bien une semi-voyelle est rencontrée, tel qu' il est représenté dans l'algorithme de décomposition en polyphones (figure 3.12).

```
Algorithme Decomp (String mot)
{
String c;
Int I, j, l;

L= length (mot);
I=1;
J=0;
C="";

While (i<=l-1)
{
    if not Transitoire (mot[i+1])
       {
        c= mot[i]+mot[i+1];
        j=j+1;
       res[j]=c ;
       i=i+1 ;
       c= ";
        }

else if Transitoire(mot [i+1]) and not Transitoire (mot [i+2])

       {
         c=mot[i]+mot[i+1]+mot[i+2] ;
          j=j+1 ;
          res[j]=c ;
          i=i+2 ;
          c= ";
        }

else if Transitoire (mot [i+1] )and Transitoire (mot [i+2])

       {
         c=mot[i]+mot[i+1]+mot[i+2] +mot[i+3] ;
          j=j+1 ;
          res[j]=c ;
          i=i+3;
          c= ";
        }

}
}
```

Figure 3.12. : Algorithme de décomposition d'un mot en polyphones

Une fois le dictionnaire de polyphones établi, ainsi que l'index de dictionnaire dont accomplis et la stratégie de décomposition de la chaîne transcrite est implémentée, il ne reste qu'à :

- Calculer les étiquettes de chaque segment phonétique obtenu, après la décomposition de la chaîne phonétique, selon l'index que nous avons adopté ;

- mettre les unités polyphoniques sélectionnées l'une à côté de l'autre.

3.7. Configuration matérielle et logicielle de TALKARABIC

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 TALKARABIC sont accessibles avec la souris et le clavier. La principale caractéristique de TALKARABIC est la possibilité de le réutiliser dans des systèmes de synthèse vocale (il contient des modules réutilisables tels que la base de données acoustiques et les algorithmes employés dans la partie traitement acoustique).

3.8. Présentation de notre logiciel TALKARABIC

TALKARABIC a été réalisé suivant une méthode de conception du génie logiciel appelé prototypage plus exactement c'est la méthode de *prototypage incrémentale.* Cette dernière accroît les performances de notre outil de Synthèse de la parole à partir d'un texte Arabe (TALKARABIC) et cela tout en réduisant le temps nécessaire à son développement et à la facilité de la maintenance puisque un logiciel doit pouvoir être maintenu (pour le corriger, l'améliorer, l'adapter aux changements de son environnement, ...).

L'interactivité de notre outil de synthèse vocale d'un texte écrit en Arabe Standard est assurée par le déclenchement d'une interface graphique principale (Figure 3.13.).

Figure 3.13. : Forme principale de notre système TALKARABIC

3.9. Test et résultats

La qualité de synthèse est un problème crucial. Il se manifeste essentiellement par le fait que la compréhension de la parole synthétique exige, de la part de l'auditeur, un effort plus important que pour la parole naturelle [9]. Cet effort supplémentaire est rendu nécessaire par les artefacts éventuels du traitement dû à la complexité et la richesse du signal de la parole, les erreurs de prononciation, la prosodie insuffisamment expressive (voix colère, voix joyeuse, etc.).

Pour le test nous avons choisi quelques phrases pour réaliser leur lecture automatiquement et évaluer la qualité de la parole synthétique produite par notre système TALKARABIC.

Sur l'ensemble des phrases testées, la parole générée par notre lecteur automatique est une parole synthétique intelligible dans la majorité des mots qui composent chaque phrase. Mais souffre de quelques artefacts au point de concaténation dûs à la différence des caractéristiques acoustiques (la fréquence fondamentale) qui nécessite un lissage par une méthode d'amélioration des systèmes TTS que nous aborderons dans le chapitre suivant.

La figure 3.14. Représente le spectrogramme de la phrase « المرور بمرحلة هامة » émis naturellement.

La figure 3.15. Représente le spectrogramme de la phrase «المرور بمرحلة هامة » produite par notre système TALKARABIC.

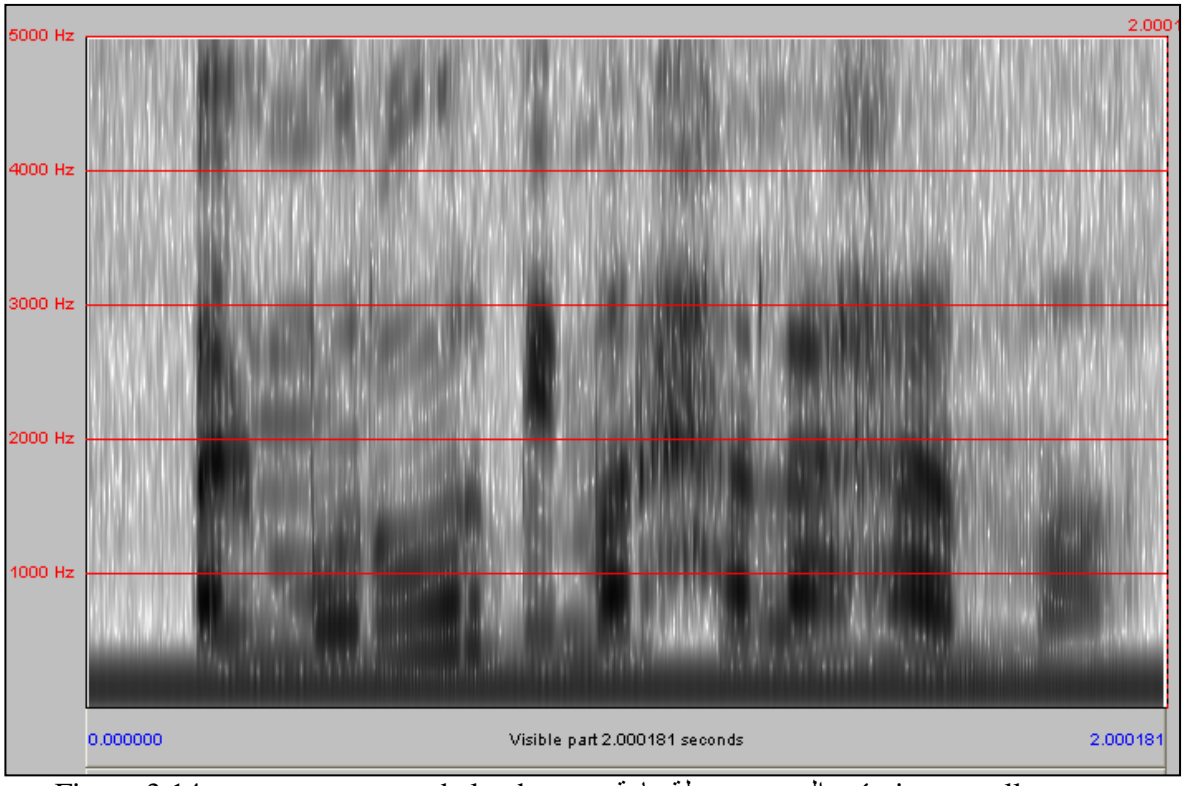


Figure 3.14.: spectrogramme de la phrase « المرور بمرحلة هامة » émis naturellement

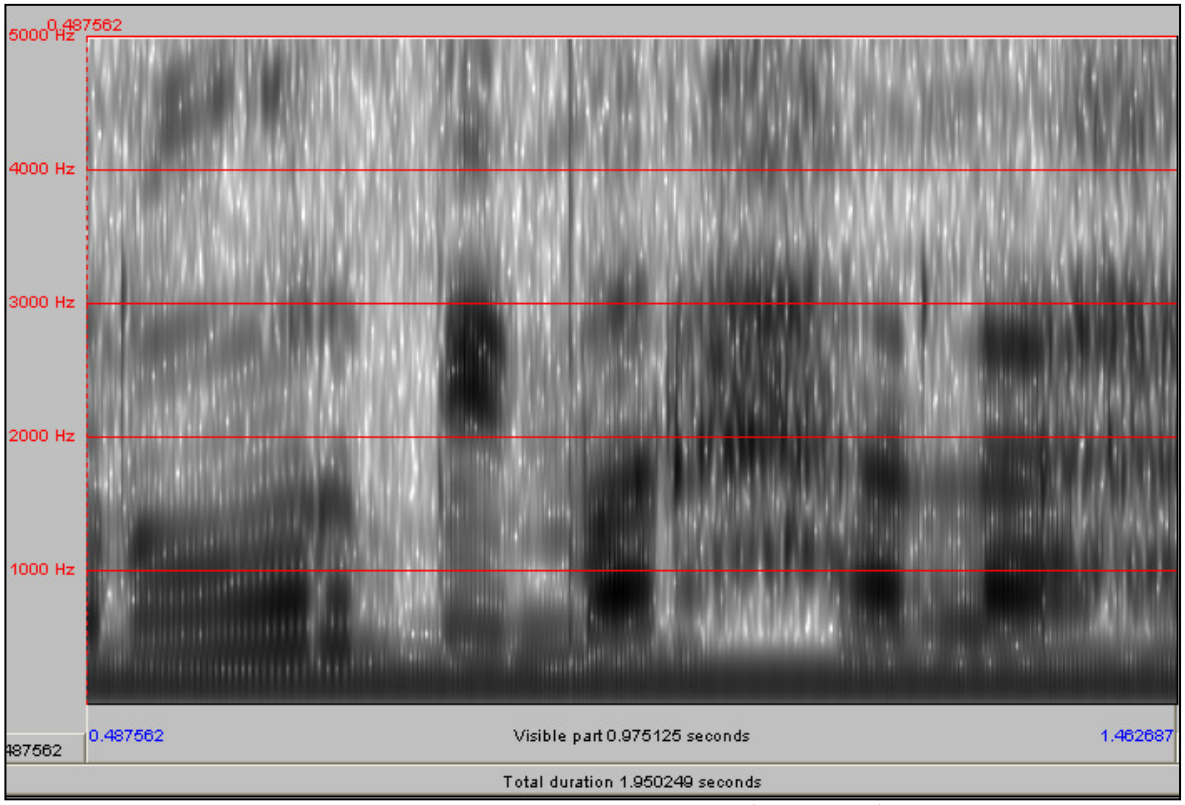


Figure 3.15. : Spectrogramme de la phrase « المرور بمرحلة هامة » produite par TALKARABIC

3.10. Conclusion

Dans ce chapitre, nous avons présenté la structure générale de notre outil de synthèse à partir de textes TALKARABIC. L'architecture modulaire qui le caractérise permet l'ajout de plusieurs fonctionnalités comme le traitement de textes spécialisés (changement du code de l'alphabet phonétique) ainsi que le choix de la voix et la langue avec lesquelles s'effectue la synthèse. Exception faite pour le dictionnaire qui exige des traitements précis.

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 vocaux qui existent. Cette dernière représente l'objectif de notre étude dans le chapitre prochain.

CHAPTIRE 4
AMELIORATION DE LA QUALITE DE LA PAROLE SYNTHETIQUE

4.1. Introduction

La concaténation d'unités pré-stockées donne un signal de parole intelligible. Cette dernière est obtenue grâce à la prise en compte de la coarticulation entre phonème composant le polyphone. Cette parole présente néanmoins des discontinuité et des irrégularités dans certains paramètres (la fréquences fondamentale, la phase, l'amplitude) localisé aux points de concaténation, ce qui rend la lecture inconfortable voire même incompréhensible (cas de la durée) comme dans le cas de la gémination ou d'elmad. Ces discontinuités et irrégularités sont le résultat des aléas qui caractérisent l'enregistrement et l'extraction des unités acoustiques à partir des mots porteurs, et il pratiquement impossible de les éviter à cause de la variabilité intra-locuteur du signal de parole. Dans ce chapitre nous décrivons le problème de perturbation du signal de la parole synthétique aux points de concaténation, aussi les méthodes permettant de le résoudre où de le réduire pour l'amélioration de la qualité de la parole synthétique. Ainsi nous proposons notre solution qui est un algorithme de sélection dynamique dans une grande base de données acoustiques pour l'amélioration des performances de notre système de synthèse à partir du texte écrit en Arabe Standard. Nous présentons également les paramètres et les méthodes d'évaluation des systèmes de synthèse de la parole en vue de la comparaison qualitative du signal synthétique générée par des méthodes différentes de synthèse vocale.

4.2. Modification de signal vocal synthétique

Ces méthodes modifient plusieurs paramètres acoustiques, on les appelle souvent *paramètres de lissage.*

Les paramètres concernés par un lissage qui sont la fréquence fondamentale, la phase et l'énergie ; ainsi les méthodes permettant de les calculer et de les modifier. Les trajets

formantiques peuvent aussi être sujets à des discontinuités et le choix d'unités comme le diphone ou le polyphone est fait justement dans le but de les supprimer sinon de les réduire. Des irrégularités peuvent apparaître dans la durée de phonèmes lors de la concaténation, notamment pour les voyelles longues. Le contrôle de ce paramètre est très important car il a une grande incidence sur le message véhiculé.

A l'heure actuelle, il existe un nombre important de méthodes pour le traitement de la discontinuité du signal de la parole synthétique aux points de concaténation. Nous présentons une méthode de modification du signal vocal ci –dessous :

◊ La méthode de lissage PSOLA (Pitch Synchronious Overlap and Add). Cette dernière se base sur la modification des paramètres prosodiques du son, ce qui se ramifie en plusieurs techniques tels que : TD PSOLA (Time Domain PSOLA) qui consiste à modifier le paramètre prosodique Temps (figure 4.1) et la seconde FD PSOLA (Frequency Domain PSOLA) qui modifie la fréquence, et LP PSOLA (Linear Prediction PSOLA)

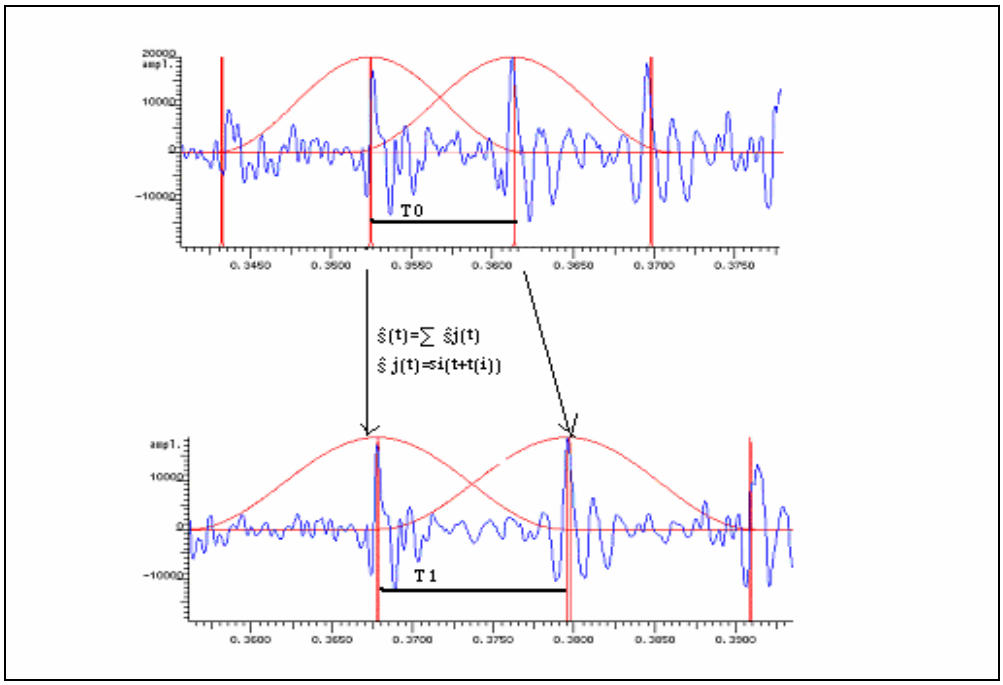


Figure 4.1 : Principe de fonctionnement de la technique TD PSOLA (superposer ou ajouter des segments dans le paramètre durée).

La génération de la prosodie par modification de signal synthétique est réalisée par la redéfinition de la courbe intonative par un modèle complexe spécifique à chaque langue. Ce dernier est élaboré par des linguistes spécialisés dans l'étude de l'intonation de la langue visée [5].

4.3. Synthèse par sélection dynamique dans un corpus

Tout système de synthèse de la parole basé sur la concaténation d'unité a besoin d'une base de données vocales contenant les différentes unités de parole à utiliser. La majorité de ces systèmes utilisent des bases de données vocales qui n'employaient qu'un seul exemplaire de chaque unité. Pour le traitement du problème de perturbations du signal et la génération de la prosodie, l'ensemble des chercheurs étaient obligés de régénérer la fréquence fondamentale, et la durée souhaitées. Malheureusement, ces modifications acoustiques apportées aux unités de manière à obtenir les caractéristiques prosodiques demandées entraînent une détérioration du naturel de la parole de synthèse et parfois même une détérioration dans l'intelligibilité.

Pour conférer à la parole de synthèse un caractère plus naturel, proche de celui de la parole humaine, les chercheurs [20, 21, 22] ont voulu mettre en oeuvre le principe de « *chose the best to modify the least* » [20] : la recherche de l'unité souhaitée est réalisée sur un corpus qui contient non plus un seul, mais plusieurs représentants de chaque unité, de sorte que les modifications acoustiques à apporter à l'unité sélectionnée soient réduites au strict minimum ou aucunes. L'approche de synthèse par sélection dynamique dans un corpus (SPC) repose sur la concaténation de segments de parole contenus dans une grande base de données enregistrée par un locuteur professionnel.

Le succès de la SPC ce tient au fait que, moyennant une couverture acoustico-prosodique suffisante, il devient possible de sélectionner une séquence d'unités acoustiques correspondant au contexte de synthèse. De ce fait, les modifications des unités de synthèse peuvent être diminuées ou évitées. Ce qui permet de préserver le naturel de la parole synthétique ainsi produite. Cependant, avec la SPC, la création de nouvelles voix de synthèse devient extrêmement coûteuse, car, outre l'enregistrement du corpus proprement dit, de nombreux traitements doivent être effectuées pour obtenir un dictionnaire acoustique utilisable par un système de synthèse. Parmi ceux-ci, les tâches de segmentations du corpus sont particulièrement critiques. En effet, même lorsque la chaîne

phonétique correspondant `a l'énoncé enregistré est connue, les méthodes de segmentation automatiques actuelles sont jugées trop peu précises pour pouvoir être utilisées telles quelles dans le processus de création de voix. Par conséquent, une étape de vérification manuelle de la segmentation demeure nécessaire. Cette étape, de loin la plus coûteuse, est un véritable frein à la diversification de voix dans le cadre de la SPC.

Actuellement, les systèmes automatiques de synthèse de la parole se dirigent vers l'utilisation de la SPC [12,13]. Du fait que cette dernière produise une parole très proche de la parole naturelle.

4.4. Proposition de notre solution d'amélioration de la qualité des systèmes TTS

Notre solution d'amélioration proposée appartient aux méthodes de synthèse à partir de bases de données élargies (les gigantesques dictionnaires acoustiques), plus exactement nous avons proposé un algorithme de sélection dynamique dans un corpus. Étant donné que les systèmes de synthèse par corpus ont amélioré significativement la qualité de la synthèse vocale. Leur succès est basé sur l'utilisation de grandes bases de données, associée à des algorithmes efficaces de sélection des unités. L'étape de sélection consiste à choisir la meilleure suite d'unités parmi toutes celles présentes dans le corpus.

Les fonctionnalités additives pour l'implémentation de la SPC dans TALKARABIC sont:

- l'analyse et la segmentation des enregistrements destinés à fournir le matériel source ;
- l'analyse en descripteurs sonores et la modélisation temporelle des unités sonores ;
- la gestion des fichiers de sons et de données dans la base de données;
- la recherche et la sélection d'unités de la base de données en fonction des paramètres cibles (algorithme proposé) ;

Le diagramme de cas d'utilisation pour le module de la génération du signal vocal artificiel par Sélection dynamique dans un corpus est représenté dans la figure 4.2.

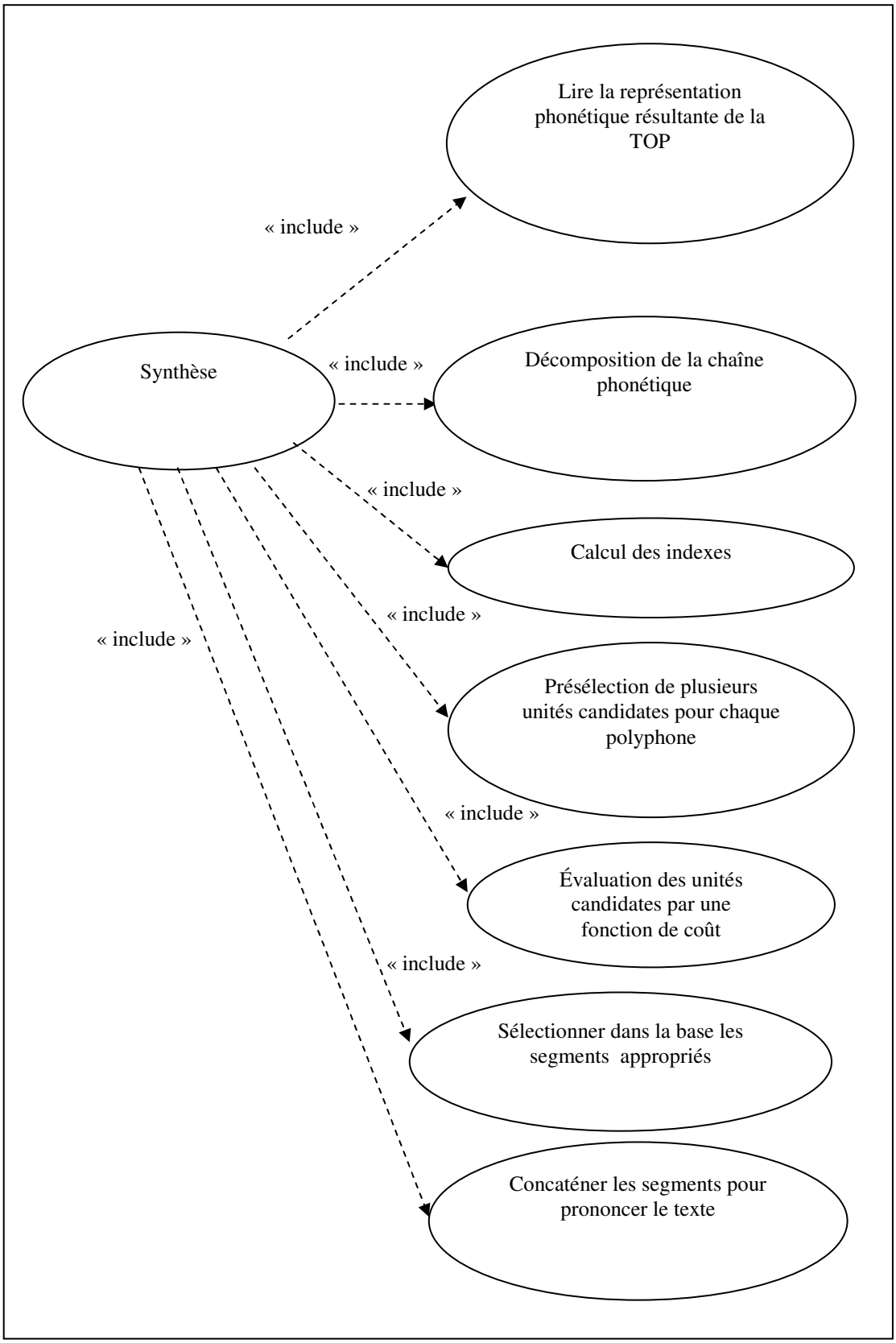


Figure 4.2. : Diagramme de use cases de cas « génération du signal vocal par la méthode SPC »

Pour la réalisation de la SPC, on doit d'abord minimiser une fonction de coût mesurant la fluidité du signal de parole synthétisé ainsi que son adéquation aux cibles issues des traitements linguistiques. Elle est généralement définie comme une somme pondérée de coûts-cibles et de coûts de concaténation [20], puis minimisée de manière optimale grâce à un algorithme d'optimisation combinatoire. Ces choix ont prouvé leur efficacité et donnent de bons résultats pour les systèmes TTS. La variation en terme de réalisation acoustique signifie qu'une même unité peut (et même doit) être présente plusieurs fois dans le corpus, chaque instance de l'unité se différenciant des autres au niveau acoustique. Les unités ne sont donc plus neutralisées ; elles conservent les variations obtenues au moment de l'élocution. Cependant la forme de la fonction de coût limite le type de contraintes qu'il est possible de prendre en considération : elles ne peuvent porter que sur des unités prises isolément (grâce au coût-cible), ou bien sur des couples d'unités consécutives (coût de concaténation).

L'entrée de notre algorithme de sélection dynamique dans un corpus est une chaîne phonétique, l'algorithme converge alors vers une suite optimale de représentants, visant à minimiser les discontinuités aux points de concaténation. La problématique de la sélection des unités a été formalisée via la minimisation d'une fonction coût.

Notre fonction de coût est définie comme la somme pondérée de coûts-cibles et de coûts de concaténation, l'équation (4.1).

$$C_L(s) = w_C \sum_{k=2}^{n} C_C(u_{k-1}, u_k) + w_T \sum_{k=1}^{n} C_T(u_k) \qquad \text{Equation(4.1)}$$

Où :

- S, désigne la séquence des unités $(u_1, u_2, \ldots, u_n)$,
- $C_L(S)$, le coût total associé à cette séquence,
- $C_C(u_{K-1}, u_K)$, le coût de concaténation entre les unités (u_{K-1}) et (u_K) ,
- $C_T(u_K)$, le coût cible associé à l'unité.

L'utilisation d'une telle fonction de coût est motivée par les contraintes suivantes : les unités doivent être choisies dans un contexte prosodique et linguistique adéquat (coût-

cible) et les transitions entre unités consécutives doivent être fluides (coût de concaténation). Une telle fonction de coût permet également une réduction de la complexité algorithmique. En effet, pour une séquence de *N* unités, chacune représentée par *M* occurrences, le nombre total de combinaisons est N^M ; mais la minimisation d'une fonction de cette forme peut être effectuée avec une complexité réduite par un algorithme d'optimisation combinatoire.

Notre algorithme de sélection dynamique dans un corpus est représenté dans la figure 4.3 :

1. créer un graphe d'états des unités candidates présélectionnés à l'aide de listes chainées.
2. initialiser les champs des entités du graphe d'états.
3. calcul du coût cible de toutes les unités candidates et le mettre dans les champs des entités correspondantes respectivement.
4. évaluation et sélection :

```
// N représente le  nombre d'unités de la chaîne phonétique.
// M  représente le  nombre de candidats d'une unité.
N=1 ; // la première unité
for (j=1 ; j++ ; j<= M)
{ Fct_coût(j)=  fct_cible(j) } ;
calcul du minimum de la fonction coût pour les candidats de la première unité
Marquer le candidat qui génère le minimum.

for (i=2 ; i++ ; i<= N)
{
   for (j=1 ; j++ ; j<= M)

     { Fct_coût(ij)=  fct_cible(ij) + fct_coût minimum des candidats précédents ;}

     calcul du minimum de la fonction coût pour les candidats de l'unité i
     Marquer le candidat qui génère le minimum.
 }
        //les unités marquées représentent la suite optimale de représentants

Sélectionner les unités Marquées.
```

Figure 4.3 Fonctionnement générale de notre algorithme de Sélection dynamique

Le changement à apporter dans l'interface Home Machine de TALKARABIC, consiste par l'ajout d'une fenêtre introductive qui devient la fenêtre principale de l'application (figure 4.4)

Figure 4.4 : Interface principale de TALKARABIC

À partir de la fenêtre principale nous pouvons lancer les deux interfaces respectivement, la synthèse polyphonique sans lissage, la synthèse par sélection dynamique dans un corpus (figure 4.5)

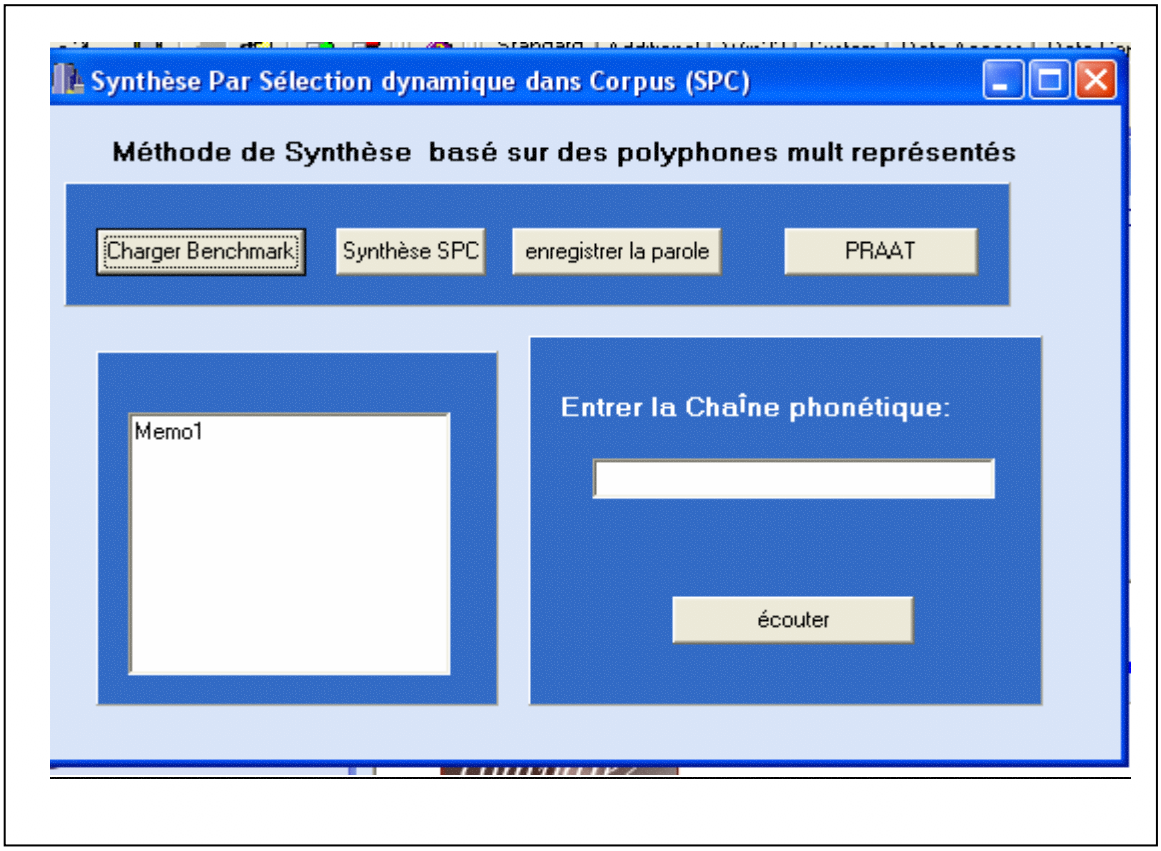


Figure 4.5 : Interface de la méthode de synthèse par sélection dynamique dans un corpus

4.5. Évaluation des systèmes de synthèse de la parole à partir du texte

L'évaluation des systèmes de synthèse de la parole n'est pas un problème résolu [3]. En conséquence 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 :

- 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;

- 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). L'interface d'une application de synthèse de la parole doit être approprié aux publics visés (les utilisateurs)

Les techniques de synthèse de la parole posent des problèmes d'évaluation nouveaux, puisque : l'architecture des systèmes de synthèse est différente, l'évaluation du « naturel » des voix devient importante et indépendamment de l'intelligibilité et de l'agrément (description de la voix comme agréable ou non (à écouter)). Ainsi, de nouveaux types de test d'agrément, plus fins, doivent être mis en oeuvre puisque la distance entre parole naturelle et parole synthétique diminue. Dans certaines applications de synthèse vocale la préservation du timbre est nécessaire ; ce qui génère les problèmes de cohérence de la voix. Cette nouvelle situation appelle de nouveaux types de tests.

On entend par évaluation globale l'évaluation de la sortie du système de synthèse sans se préoccuper de son fonctionnement interne et sans chercher la source des défauts éventuels. Les mesures suggérées pour l'évaluation globale sont l'intelligibilité, la compréhension, la charge cognitive et l'agrément (« Est-ce que la voix vous plait, est-ce qu'elle vous semble appropriée pour une application donnée »). La majeure partie des tests d'évaluation nécessite des expériences perceptives et fait donc appel à une expérimentation avec des sujets/auditeurs. Les deux méthodes utilisées sont :

- les tests MOS (Mean Opinion Score (test d'opinion moyen)) Les sujets (environ 10) écoutent des phrases et donne une note d'appréciation globale de la qualité de la parole synthétique;

- les tests SUS (Semantically Unpredictable Sentences (test d'intelligibilité)) Les sujets (environ 10) écoutent des phrases syntaxiquement correctes mais sémantiquement imprévisibles puis écrivent ce qu'ils ont entendu.

4.6. Évaluation comparative

Les premiers travaux de recherche dans l'évaluation comparative ont été fait par R. PRUDON & all [40].

Pour la comparaison qualitative entre la synthèse polyphonique et la synthèse par sélection dynamique dans un corpus; nous avons effectué une évaluation globale comparative, nous utilisons le test MOS avec des phrases de tailles différentes (courtes, moyennes, longues) et a l'aide des interlocuteurs naïfs (la méconnaissance de phrases préalablement) ayant une bonne connaissance en AS. Les sujets (auditeurs) utilisent le tableau (Tableau 4.1) pour la notation des phrases.

Tableau 4.1 : L'échelle de notation pour le test MOS

Note	Qualité
1	Mauvaise
2	Assez bien
3	Bien
4	Bonne
5	Excellente

Tableau 4.2 : La moyen des opinons

	Méthode classique (polyphones)	Méthode SPC (polyphones multi représentés)
Phrase 1	2.71	4
Phrase 2	2.42	3.71
Phrase 3	2.85	3.85
Phrase 4	3	4
Phrase 5	2.85	4
Phrase 6	3	4
Phrase 7	3	4

Les résultats de la moyenne des opinons obtenus (Tableau 4.2) montrent une nette amélioration de la qualité de la parole synthétique. Ceci confirme la robustesse de notre Algorithme de sélection.

D'après les différentes expériences et les résultas obtenus nous avons constaté que la taille du corpus influe sur la qualité de la parole synthétique, la manière de sélectionné dans le corpus influe directement aussi ; car si nous appliquons la politique de «sélectionner le meilleur pour modifier le moins » la parole générée est très acceptable d'un point de vue perceptif.

Pour la comparaison sonnagraphique des différentes paroles générer par plusieurs systèmes nous présentons leur représentation sonagraphique respectivement :

- spectrogramme de la phrase « حتمية التوسع في دراسات »de la parole naturelle (générer par le système phonatoire) figure 4.6 ;
- spectrogramme de la parole artificielle sans amélioration de la phrase « حتمية التوسع في دراسات » (générer par un notre système de lecteur automatique de texte) figure 4.7. ;
- spectrogramme de la phrase « حتمية التوسع في دراسات » de la parole artificielle avec amélioration (générer par un notre système de lecteur automatique de texte) figure 4.8.

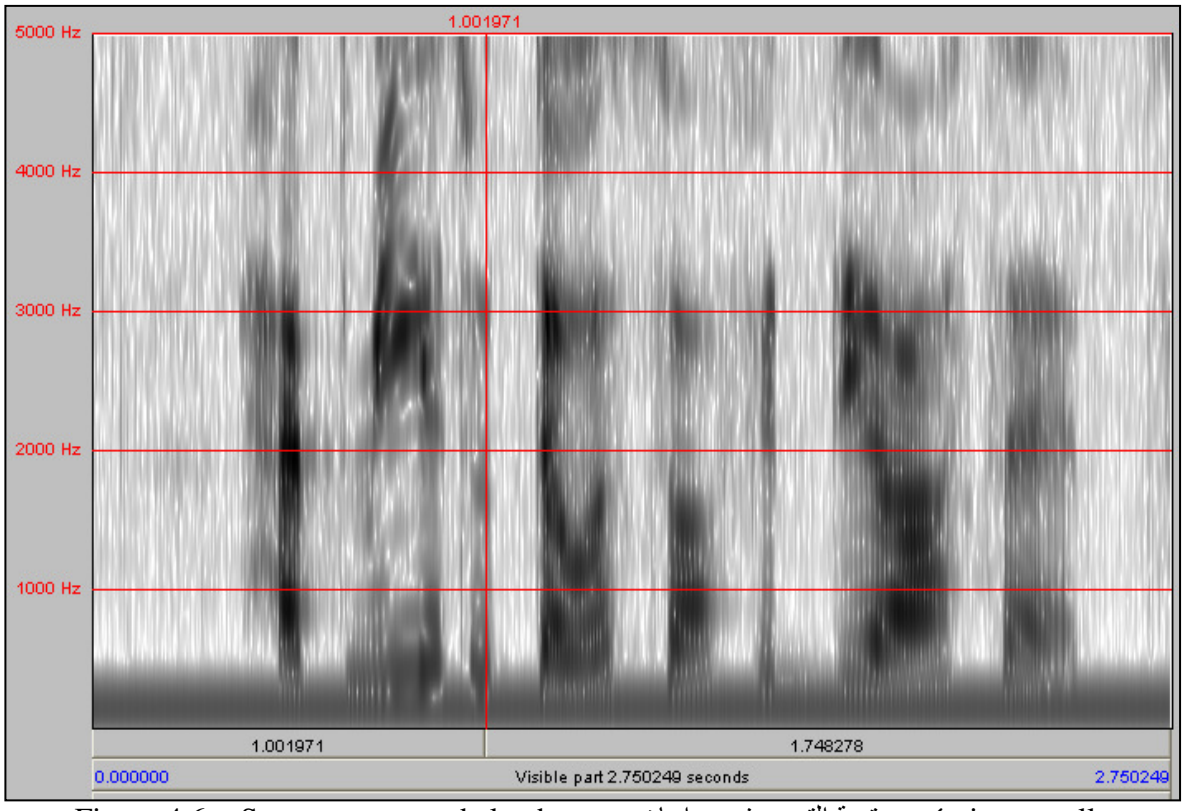


Figure 4.6. **:** Spectrogramme de la phrase « حتمية التوسع في دراسات » émis naturellement

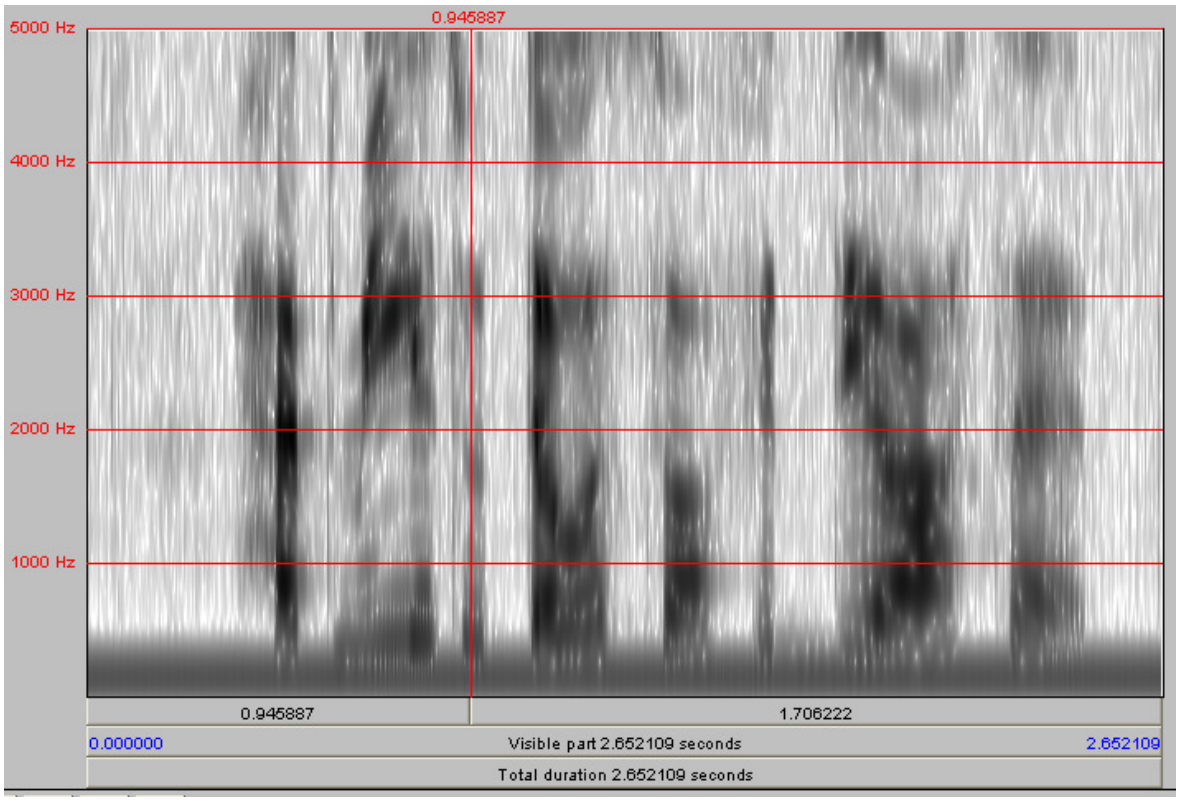


Figure 4.7. : Spectrogramme de la parole artificielle sans amélioration de la phrase « حتمية التوسع في دراسات »

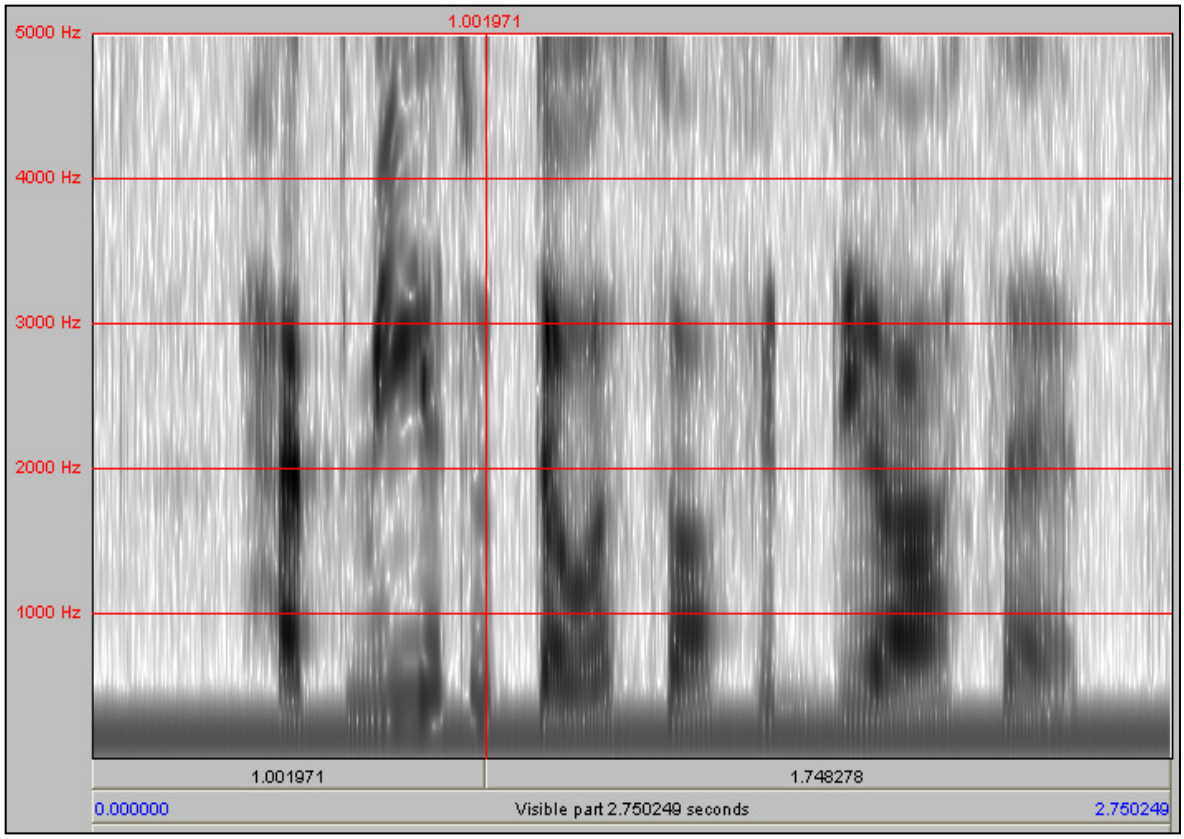


Figure 4.8. : Spectrogramme de la phrase « حتمية التوسع في دراسات » de la parole artificielle avec amélioration

4.7. Conclusion

Dans ce chapitre, nous avons traité le problème de l'apparition de perturbations du signal localisées aux points de concaténation dans la synthèse par concaténation. Ces perturbations concernent la fréquence fondamentale, l'énergie, la durée ainsi que la phase. Nous avons présenté un éventail de techniques de lissage dont l'intérêt est l'amélioration de la qualité de la parole générée et qui constitue de nos jours, un critère important pour le classement des systèmes de synthèse de la parole. En vue de l'application de ces techniques a notre système de synthèse à partir du texte écrit en AS que nous avons réalisé, nous avons exposé deux approches : la modification des segments composant le signal synthétique ainsi que la recherche du meilleur segment dans un dictionnaire élargi en minimisant une fonction coût. Nous avons également proposées notre solution, un algorithme permettant la sélection du meilleur segment parmi plusieurs segments candidats.

CONCLUSION

Dans ce travail nous avons abordé la problématique de la synthèse de la parole à partir du texte Arabe en Standard à différents niveaux tels que : la Transcription Orthographique Phonétique, l'élaboration de la base de données des segments acoustiques, et enfin la phase de génération de l'onde de la parole synthétique. Nos études ont été approfondies au niveau du module de génération du signal acoustique et la construction de la base de données sonores.

Nous avons mis en oeuvre un système de synthèse en langue Arabe Standard que nous avons nommé « TALKARABIC ». L'approche utilisée pour sa réalisation est basée sur la synthèse par concaténation d'unités de signal de parole. Ce choix s'est révélé fructueux car la qualité de la parole est très acceptable. L'unité de concaténation utilisée est le polyphone (c'est-à-dire des unités de tailles variables) ; pour éviter de segmenter les phonèmes instables tels que les liquides, les semi-voyelles ; en conséquence la synthèse polyphonique réduit considérablement le problème de la coarticulation. Cette méthode a l'avantage de produire une parole intelligible et plus proche du naturel mais nécessite toutefois quelques améliorations aux points de jonctions des unités acoustiques. Pour résoudre le problème de perturbations du signal acoustique synthétique dans les points de concaténations, nous avons présenté deux approches employées actuellement, la première est la méthode d'amélioration du signal synthétique par modification du signal acoustique (la fréquence fondamentale, la durée, etc.). Cette dernière entraîne dans la majorité des cas une dégradation de la qualité du signal acoustique « voix métallique », la deuxième méthode qui se base sur la concaténation sans modification, leur dictionnaire acoustique est très élargie, pour chaque unité on trouve plusieurs segments exemplaires sur lesquels un choix peut être effectué au cours du processus de concaténation. Pour améliorer notre système, nous avons proposé une solution sans modification du signal acoustique qui est l'algorithme de sélection dynamique dans un dictionnaire élargi.

D'après les différentes expériences effectuées, nous avons obtenu des résultas satisfaisants, du fait que la parole générée est très acceptable d'un point de vue perceptif. Ceci confirme la robustesse de notre Algorithme de sélections mis en œuvre pour l'amélioration de la qualité de la parole synthétique.

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 :

- l'enrichissement de la base de données acoustiques élargie (par l'ajout de plusieurs segments équivalents phonétiquement mais pas prosodiquement ou le contexte d'élocution) pour la synthèse par sélection dynamique dans un corpus ;

- l'établissement de nouvelles voix de synthèse pour avoir plus de diversités de locuteurs/locutrices virtuelles ;

- les unités acoustiques, quelles que soient les précautions prises lors de la sélection et de l'enregistrement des unités, ne possèdent pas exactement à leurs frontières les mêmes caractéristiques acoustiques. Il est alors préférable de procéder à un lissage des extrémités des unités acoustiques sans détérioration de la qualité de la parole synthétique (l'intelligibilité et le naturel de la parole).

Actuellement les travaux de recherche dans la synthèse de la parole s'attachent ainsi à :

- améliorer la variabilité de la voix de synthèse au cours du temps, à lui ajouter des possibilités d'expressivité accrue (voix joyeuse/triste, voix colère/calme, etc.) ;

- développer des méthodes de conversion de voix permettant de créer rapidement de nouvelles voix de synthèse;

- la segmentation automatique de la parole et l'indexation audio sont deux domaines de recherche en relation étroite à la synthèse de la parole.

APPENDICE A

LISTE DES SYMBOLES ET DES ABREVIATIONS

AS : Arabe Standard

AMDF : Average Magnitude Difference Function

API : Alphabet Phonétique International

C : Consonne

CS : Consonne Solaire

CL : Consonnes Lunaire

CP : Communication Parlée

dB : DéciBels

E : Energie (ou Intensité)

F_0 : Fréquence Fondamentale

FD PSOLA : Frequency Domain Pitch Synchronious OverLap and Add

FPMs : Faculté Polytechnique de Mons

IPA : International Phonetic Alphabet

Ou API : Alphabet Phonétique International

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

LPC : Linear Predictive Coding

LP PSOLA : Linear Prediction Pitch Synchronious OverLap and Add

LATL : Laboratoire d'Analyse et de Technologie du Langage

MBROLA : MultiBand Resynthesizer Overlap and Add

MOS : Most Opinions Score

PSOLA : Pitch Synchronous Overlap and Add

SBL : Spectrogramme à Bande Large

SBE : Spectrogramme à Bande Etroite

SUS : Semantically Unpredictable Sentences

SPC : Synthèse Par sélection dynamique dans un Corpus

TAP : Traitement Automatique de la Parole

TD PSOLA : Time Domain Pitch Synchronious OverLap and Add

TOP : Transcription Orthographique Phonétique

TTS : Text-To-Speech

TOP-AS : Transcription Orthographique Phonétique d'un texte en Arabe Standard

T : Transitoire (un phonème transitoire si il est une liquide ou une semi- voyelle, c'est dire un des phonèmes [l], [r], [w] et [y]).

T_0 : Période

UML : Unified Modeling Langage

V : Voyelle

REFERENCES

1. S. Baloul, « Développement d'un système automatique de synthèse de la parole à partir du texte arabe voyellé », Thèse de Doctorat, Université le Maine, France, 27 Mai 2003.

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

3. G. Gibert, « Conception et évaluation d'un système de synthèse 3D de la langue française parlée complétée à partir du texte », Thèse de Doctorat, INPG, Grenoble, France, 5 Avril 2006.

4. B.Grégory, « Synthèse concatenative de la parole par sélection d'unités », rapport de stage, IRCAM - PARIS VIII, 2004.

5. B.Bozkurt, T.Dutoit, V.Pagel, « Synthèse vocale par sélection d'unité: une méthode pour la redéfinition de la courbe intonative » les Journées d'Étude sur la Parole, Nancy, 24-27 juin 2002.

6. M.Guerti, « contribution a la synthèse de la parole en Arabe Standard » 16 journées d'étude sue la parole, Hammamet, Tunis, Octobre 1987.

7. K. Benbllil, « Synthèse par polysons de l'Arabe Standard », Thèse de Magister, Université Houari Boumediène, Alger, Algérie, 2004.

8. G.Bohas, « contribution à l'étude de la méthode des grammairiens arabe en morphologie et en phonologie d'après les grammairiens arabes tardifs », thèse de doctorat, université de Lille 3, Lille, France, 1979.

9. G. Richard, O.Caape, « Synthèse de la parole à partir du texte », Techniques de l'ingénieur, Volume H2, 7288, Novembre 2003.

10. T. Dutoit, « introduction au traitement automatique de la parole », Faculté polytechnique de Mons, Belgique.

11. P.Boula de Mareüil, « Étude linguistique appliquée à la synthèse de la parole à partir du texte », Thèse de doctorat de l'Université Paris XI, Orsay, France, 1997.

12. R. Beaufort, A. Ruelle, « eLite : système de synthèse de la parole à orientation linguistique », Actes des XXVI journées d'études sur la parole, Dinard, juin 2006.

13. V. Colotte, R. Beaufort, « Synthèse vocale par sélection linguistiquement orientée d'unités non uniformes : LiONS », Actes des JEP, Fès, Maroc, 2004.

14. G. Bailly, « Automates parlants », Institut de Communication Parlée, Grenoble, France.

15. M.Guerti, « Contribution à la synthèse de la parole en Arabe Standard (synthèse par diphones et techniques de prédiction linéaire) », ILP-Alger, Algérie, 4 Mars 1984.

16. محمدّ أبو الفرج، "كيف تقرأ القرآن"، اليمامة لطباعة و النشر و التوزيع، دمشق بيروت 2005 ..

17. H. Tebbi, « Transcription Orthographique Phonétique d'un texte écrit en Arabe Standard », Mémoire de Magister, Université Saad Dahleb, Blida, Algérie, Juin 2007.

18. B. Gosselin, «codage de l'information, représentation de l'information et quantification des signaux », notes de cours, Faculté polytechnique de Mons.

19. G. Droua-Hamadani, « Prédiction de la durée des phonèmes de l'Arabe Standard ». Mémoire de magister, CRSTDLA- Université de Bouzaréah, Alger, Algérie, 18 février 2004

20. A. Black, N. Campbell, « Optimising selection of units from speech databases for concatenative synthesis », In *Eurospeech'95*, volume I, pages 581–584, Madrid, Spain, 1995.

21. M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri, « Choose the best to modify the least: A new generation concatenative synthesis system », *Eurospeech'99*, pages 2291–2294, Budapest, Hungary, 1999.

22. B. Bozkurt, C. Alessandro, T. Dutoit, V. Pagel, R. Prudon, « Improving Quality of MBROLA Synthesis for Non-Uniform Units Synthesis », In *Proceedings of the IEEE TTS 2002 Workshop*, 2002

23. M. Kebache, « 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.

24. F. Ykhlef, « 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.

25. T. Dutoit, L. Couvreur, F. Malfrère, V. Pagel, C. Ris « Synthèse Vocale et Reconnaissance de la Parole : Droites Gauches et Mondes Parallèles »

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

27. http://www.arts.gla.ac.uk/IPA/ipa.html (le site officielle de l'Association de l'Alphabet Phonétique International)

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

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

30. Talking Heads Simulacra',http://www.haskins.yale.edu

31. T. Dutoit, « Traitement de la Parole à la FPMs (1983-2000) », TCTS Lab Faculté Polytechnique de Mons, Belgium.

32. le serveur du center spoken language understanding ou the center for spoken language understanding, Oregon, Health and sciences University.- http://cslu.cse.ogi.edu/tts

33. S. Jafri, D. Pastor, O. Rosec, « Segmentation automatique de corpus de parole continue dédiées à la synthèse vocale », Journée SC du Vendredi 17-11-2006, Département SC GET-ENST Bretagne, Brest.

34. B. Grégory, M. Aurélien, « TALKAPILLAR : outil d'analyse de corpus oraux », Actes des RJC ED268 'Langage et langues', Institut de Recherche et de Coordination Acoustique/Musique, place Igor Stravinsky Paris III, France, 15 Mai 2004.

35. T.Dutoit, V. Pagel, N. Pierret, F. Bataille, O. Van Der Vrecken, « The MBROLA project: towards a set of high quality speech synthetizers free of use for non commercial purpose », 1996.

36. le projet Euler « vers une synthèse de parole générique et multilingue »

37. R. Beaufort, « synthèse de la parole », MULTITEL – Département Synthèse de la Parole, Mons, Belgique.

38. P.Y. Le meur, « Synthèse de la parole par unités de taille variable », Thèse de Doctorat, 16 février 1996.

39. F. Emerad, « Les diphones et le traitement de la prosodie dans la synthèse de la parole », Bulletin de l'Institue de Phonétique de Grenoble, France, 1977.

40. R. Prudon, C. Alessandro, « A selection/concatenation TTS synthesis system: Databases developement, system design, comparative evaluation », In *ISCA/IEEE 4th Tutorial and Research Workshop on Speech Synthesis*, pages 201–206, Pitlochry, Schotland, August 29 – September 1 2001.