République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
ECOLE NATIONALE POLYTECHNIQUE

DÉPARTEMENT D'ÉLECTRONIQUE

Mémoire de fin d'études

En vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

Thème :

Synthèse de la Parole par Unités Variables en Vue d'un Guide Touristique en Algérie

Encadré par :

Mme M. GUERTI

Réalisé par :

Mr BOUALAM Mohamed El Amine

Mr BOUNABI Moussaab

Promotion : Juin 2013


بسم الله
الرحمن الرحيم

ملخص :

الهدف من عملنا هذا هو إعداد نظام لتركيب الكلام لغرض تطوير دليل سياحي جزائري متكلم، من أجل هذا قمنا بدراسة مختلف تقنيات تركيب الكلام واعتمدنا في مشروعنا على تقنية الجمع بين الجمل و المفردات المتصلة، وقمنا بإعداد مدونة من تعابير مستمرة ،تتكون من 32 كلمة منها جمل ثابتة وجمل متغيرة بالعربية الفصحى ،الدارجة الجزائرية والامازيغية.
تم تسجيل المدونة في المعهد العالي للأعمال الفنية والمسرحية و السمعي البصري -الجزائر- من طرف متكلمين (رجل وامرأة عربيا اللسان) في شروط ملائمة للتسجيل كما أجرينا عليها بعض التحاليل الأكوستيكية والعروضية المقارنة بين الكلام الأصلي والمركب، كما قمنا باختبار ذاتي سماعي لهذه المدونة للتأكد من جودة الكلام المركب، في النهاية قمنا بتطوير برنامج محاكاة الدليل السياحي بواسطة برنامج visuel studio المعتمد على لغة البرمجة C# كما قمنا بصنع دارة كهربائية منمذجة لمشروع الدليل السياحي.
كلمات المفاتيح : تركيب الكلام ، الجمع بين الجمل و المفردات، مدونة ،العربية الفصحى، التحاليل الأكوستيكية والعروضية دليل سياحي جزائري متكلم.

Résumé :

Le but de notre travail est d'élaborer un système de synthèse de la parole en vue d'un **G**uide **T**ouristique **P**arlant en **A**lgérie (**GTPA**). Pour cela, nous avons étudié les méthodes et les techniques de synthèse de la parole, en se basant sur la synthèse par concaténation des phrases et mots combinés. Après nous avons élaboré un corpus de parole continue, enregistré par 2 locuteurs, en Arabe Standard, dialecte et Amazigh, à l'ISMAS (Institut Supérieur des Métiers des Arts du Spectacle et de l'Audiovisuel) -Alger, suivie d'une analyse acoustique et prosodique. Puis nous avons développé un programme de simulation du GTPA sous l'environnement Visual studio 2010 avec le langage C #. Pour évaluer la qualité de la parole synthétique, une comparaison a été faite entre le signal original et le synthétique avec des tests de perception avec 10 personnes en vue de l'intelligibilité et l'aspect naturel. Suite à ce travail une réalisation pratique a été faite.
Mots clés: Synthèse de la Parole, Concaténation de phrases et mots, corpus, Arabe Standard, analyse acoustique et prosodique, Guide Touristique Parlant.

Abstract:

The objective of our work is to develop a system of speech synthesis for a Speaking Tourist Guide in Algeria (STGA). For this, we studied the methods and techniques of speech synthesis. It is based on the synthesis by concatenating of phrases and words; on those we have developed a corpus of continuous speech recorded by two speakers in Standard Arabic, dialect and Amazigh in High Institute of Arts and Theater Profession and Audiovisual-Algeria-, followed by an acoustic and prosodic analysis. Then we developed a simulation program STGA under Visual Studio environment 2010 with C # language. To evaluate the quality of synthetic speech, a comparison was made between the original signal and the synthetic with perception tests on 10 peoples for the intelligibility and naturalness.

Following this work a practical implementation has been made.
Keywords: Speech synthesis, Concatenation of sentences and words, corpus, Standard Arabic, acoustic and prosodic analysis, Tourist Guide Speaking.

DEDICACES

Je dédie ce modeste travail à celle qui m'a donné la vie, le symbole de tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite, à ma mère.

A mon père, école de mon enfance, qui a été mon ombre durant toutes les années des études, et qui a veillé tout au long de ma vie à m'encourager, à me donner l'aide et à me protéger.

Que dieu les gardes et les protège.

A mes sœurs Fatiha, Ibtissem et Zakia et a toutes ma famille.

A mon binôme Moussaab.

A tous mes amies.

A tous ceux qui m'aiment.

A tous ceux que j'aime.

Je dédie ce travail

Mohamed El Amine

Dédicaces

Je dédie cette humble travaille à ma très chère et tendre mère qui ces occupé de moi avec une grande affection et amour. A mon père qui a été et sera toujours derrière moi en m'encourageant a perfectionné mon travail.

Je le dédie aussi à :

- Mon frère Aymen et a mes deux sœurs Zineb et Fatima el Zahraa qui me soutiennent toujours dans le meilleure et dans le pire
- mon binôme Mohamed El-Amine qui a sue m'épaulé tous le long de ce travail
- Mes amis qui ont été toujours à mes coté dans les moments les plus rigoureux : Aymen, Billel, Yazid, Minou, Zinou, El bouz, amine, Khali, les « Samir », Oussama, Salah, djawade, Saad, Houdayfa, Ismail, Ben Taleb et surtout à celui que j'admire et que je respecte le plus qui est l'honorable « Cheheb Lotfi »
- et je n'oublie pas le reste de ma famille et tous ce qui m'aimes de m'avoir aidé commensurablement et indirectement à devenir l'être que je suis
- et je n'oublie pas aussi l'association scientifique « ELMAARIFA » et le GROUP « MAZAL WAKFIN » qui ont pris soin de moi.

MOUSSAAB

Remerciements

Tout d'abord nous remercions Dieu de nous avoir donné la force et le courage d'accomplir ce travail.

Nous remercions vivement notre promotrice Professeur ***GUERTI Mhania*** *pour nous avoir confié ce travail d'abord et pour son soutien constant, son rôle majeur et sa grande patience ainsi que ses encouragements durant toute la période de ce travail. Nous la remercions pour ses compétences, son ouverture d'esprit et sa grande disponibilité.*

Nous remercions les membres du jury, qui nous ont fait l'honneur de participer au jugement de ce travail.

Nous exprimons notre reconnaissance à Monsieur ***LARBES chérif****, Professeur à l'Ecole Nationale Polytechnique, d'avoir accepté de présider le jury de notre PFE.*

Nous remercions également Monsieur ***MAMMERI Mohamed****, Maître de conférences à l'Ecole Nationale Polytechnique, d'avoir accepté de faire partie de notre jury.*

Nous Exprimons notre profonde gratitude à Monsieur ***KABACHE Mahrez*** *de l'institut Supérieur des Métiers des Arts du Spectacle et de l'Audiovisuel (ISMAS), pour sa gentillesse en acceptant d'enregistrer notre corpus à l'ISMAS*

Nous tenons à remercier également l'ensemble des enseignants qui ont contribué à notre formation.

Nous remercions tous ceux, qui de près ou de loin, nous ont apportés leur contribution pour la réalisation de ce travail.

LISTE DES ABREVIATIONS

TAP : **T**raitement **A**utomatique de la **P**arole
PSG : **P**ression **S**ous **G**lottique
PSI : **P**ound **S**quare **I**nch
PIO : **P**ression **I**ntra **O**rale
RAP : **R**econnaissance **A**utomatique de la **Pa**role
API : **A**lphabet **P**honétique **I**nternational
AS : **A**rabe **S**tandard
F_0 : **F**réquence **f**ondamentale
F_1, …, F_5 : Formants
OCDE : **O**rganisation de **C**oopération et de **D**éveloppement **E**conomiques
TTS : **T**ext-**T**o-**S**peech (Un Système de Synthèse à Partir du Texte)
OCR : **O**ptical **C**haracter **R**ecognition (un système de reconnaissance optique des caractères)
TOP : **T**ranscription **O**rthographique-**P**honétique
TPZ : **T**aux de **P**assage par **Z**éro
LPC : **L**inear **P**rédictive **C**oding (Codage Prédictif Linéaire)
LPCC : **L**inear **P**rédictive **C**epstral **C**oefficients (Coefficients Cepstraux Prédictifs Linéaires)
MFCCS : **M**el **S**caled **F**requency **C**epstral **C**oefficients
TF : **T**ransformée de **F**ourier
TFD : **T**ransformée de **F**ourier **D**iscrète
TFR : **T**ransformée de **F**ourier **R**apide (**FFT**)
AR : **A**uto **R**égressif
ARMA : **A**uto **R**égressif à **M**oyenne **A**justée
MA : **M**oyenne **A**justée
SPR : **S**ynthèse **P**ar **R**ègles
TFI : **T**ransformée de **F**ourier **I**nverse
V-C-V : **V**oyelle-**C**onsonne-**V**oyelle
PSOLA : **P**itch **S**ynchronous **O**ver **L**ap and **A**dd
CT : **C**ourt **T**erme
OLA : **O**ver **L**app and **A**dd
SI : **S**ystème **I**nternational
TD-PSOLA : **T**ime **D**omain-Pitch Synchronous **O**ver **L**ap and **A**dd
VO : **V**ersion **O**riginale
GTPA : **G**uide **T**ouristique **P**arlant en **A**lgérie
VOT : **V**oice **O**n **T**ime
OT : **O**ptimality **T**heory
ISMAS : **I**nstitut **S**upérieur des Métiers des Arts du Spectacle et de l'Audiovisuel

Liste des Tableaux

Liste des Figures

page

Table des Matières

Introduction Générale

Introduction Générale

La parole étant le moyen de communication le plus naturel chez l'Homme, celui-ci a très vite cherché à l'intégrer dans les interfaces Homme-Machine. Cela a été rendu possible grâce aux efforts consentis en reconnaissance et en synthèse de la parole, cependant la première vise à reconnaitre les messages de l'utilisateur pour les traduire en action, alors que la seconde a pour objectif de doter l'ordinateur d'une capacité à lire des textes à haute voix. Cela rend le **T**raitement **A**utomatique de la **P**arole (**TAP**) un composant fondamental des sciences de l'ingénieur et d'un domaine de recherche actif. Depuis les années 60, le TAP bénéficie d'efforts de recherches très importants, liés au développement des moyens techniques, de télécommunications et du traitement numérique de l'information. Ces efforts se sont concrétisés grâce à plusieurs applications du TAP, telles que le codage, les perceptions auditive et visuelle, la **R**econnaissance **A**utomatique de la **P**arole (RAP) et la **S**ynthèse de la **P**arole (**SP)**. Un thème important de la recherche actuelle dans le domaine du TAP, est la réalisation de véritables systèmes de dialogue oral entre l'Homme et la Machine

Le but de notre travail est de réaliser un Guide Touristique Parlant en Algérie (GTPA) en se basant sur la méthode de synthèse de la parole. Le choix de la langue native qui est le Français vers la langue cible qui sera l'**A**rabe **S**tandard (**AS**), le dialecte du centre d'Algérie et l'Amazigh (chawia).

Des étrangers, peuvent perdre facilement leurs repères en territoire inconnu, à cause de la barrière de la langue, mais en grande partie à cause de la culture, des traditions et des coutumes. Ils préparent alors son séjour, en se munissant des ouvrages touristiques les plus réputés, pour retenir les sites à visiter en fonction de leurs centres d'intérêts, et choisir ainsi les itinéraires adéquat. Malgré cela, ils demandent souvent de l'aide sur place, Ce n'est qu'à la fin du séjour qu'ils prennent conscience de tout ce qu'ils auraient pu découvrir et apprendre. Il est aussi vrai que notre soif de découverte, de culture et de savoir n'a pas de limite.

Tout système de synthèse de parole à partir du texte, est amené à répondre de manière plus ou moins précise et développée selon la qualité et la finalité du système, représentés par trois problèmes de natures différentes : il s'agit dans un premier temps d'analyser et de structurer le texte afin de déterminer un mode de prononciation cohérent ; par la suite, le texte analysé doit

être transformé en une suite de sons de parole accompagnée d'indications concernant leur agencement ; enfin, il faut générer un signal acoustique qui « retranscrit » cette suite de sons tout en possédant les caractéristiques apparentes de la parole naturelle.

Pour atteindre notre objectif, nous avons structuré notre travail en quatre chapitres :

- dans le premier, nous allons décrire d'une manière générale des notions sur le traitement de la parole ainsi que sa production, les appareils phonatoires et auditifs de l'être humain, des spécifications du signal vocal et des notions fondamentales sur l'Arabe Standard ;
- le deuxième, nous donne une brève définition de la synthèse de la parole, et du traitement linguistique du texte, En outre, nous étudions les différentes techniques d'analyse du signal vocal. Puis nous expliquons les méthodes de la synthèse de la parole ainsi que ses différentes applications.
- dans le troisième, nous nous intéressons à faire une analyse acoustique de notre corpus en étudiant les caractéristiques et les paramètres pertinents de ce signal vocal (fréquence fondamentale, formants et intensité). Nous introduisons les étapes de l'élaboration de notre corpus et son traitement. Nous expliquons le logiciel Praat et finissons par une étude comparative pour quelques signaux vocaux avant et après la concaténation des unités sonores.
- dans le dernier chapitre, nous présentons le programme de la simulation que nous avons développée en vue d'obtenir un Guide Touristique Parlant en Algérie, après une brève introduction au langage de programmation C# sous l'environnement Visual Studio. Nous allons faire aussi un test de perception subjective afin de pouvoir évaluer les résultats obtenus (signal vocal de sortie de l'interface), qui concerne l'intelligibilité et l'aspect naturel. En dernier lieu nous finissons par une réalisation pratique du circuit (guide parlant), et on parlera des conclusions générales et perspectives.

Chapitre 1 :

Notions sur la Parole et l'Arabe Standard

1.1 INTRODUCTION

La parole est le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés. Dans ce chapitre nous allons décrire de manière générale des notions sur le traitement automatique de la parole et de sa production, ensuite nous présentons les appareils phonatoires et auditifs humains, les spécifications du signal vocal et des notions fondamentales sur l'Arabe Standard.

1.2 QU'EST-CE-QUE LE TRAITEMENT AUTOMATIQUE DE LA PAROLE (TAP) ?

Le **T**raitement **A**utomatique de la **P**arole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques). Cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

Les techniques de TAP tendent cependant à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions

- **les analyseurs** de parole cherchent à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (on parle alors d'analyseur perceptuel), mais jamais tel qu'il est compris, ce rôle étant réservé aux Reconnaisseurs. Les analyseurs sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse, soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical ou l'étude des langues ;
- **les reconnaisseurs** ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la reconnaissance du locuteur, dont l'objectif est de reconnaître la personne qui parle, et la reconnaissance de la parole, où l'on s'attache plutôt à reconnaître ce qui est dit ;
- **les synthétiseurs** ont quant à eux la fonction inverse de celle des analyseurs et des reconnaisseurs de parole : ils produisent de la parole artificielle. On distingue deux types de synthétiseurs : les synthétiseurs de parole à partir d'une représentation numérique, inverses des analyseurs, dont la mission est de produire de la parole à partir des caractéristiques

numériques d'un signal vocal telles qu'elles sont obtenues par analyse, et les synthétiseurs de parole à partir d'une représentation symbolique, inverse des reconnaisseurs de parole et capables en principe de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, on classe également les synthétiseurs en fonction de leur mode opératoire :

- les synthétiseurs à partir du texte reçoivent en entrée un texte orthographique et doivent en donner lecture ;
- les synthétiseurs à partir de concepts, appelés à être insérés dans des systèmes de dialogue Homme-Machine, reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue.

- enfin, le rôle des **codeur**s est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole [1].

1.3 ANATOMIE DE L'APPAREIL PHONATOIRE

L'appareil phonatoire est l'ensemble des organes qui permettent de produire les sons constituant la voix. Chez l'Homme, cet appareil est confondu avec l'appareil respiratoire et ses organes sont répartis entre le thorax, le cou et la tête. D'une manière générale, l'appareil phonatoire est décomposé en trois parties correspondant à trois entités fonctionnelles différentes : les voies aériennes inférieures composées des poumons et de la trachée artère, le larynx, et le conduit vocal (Fig.1.1).

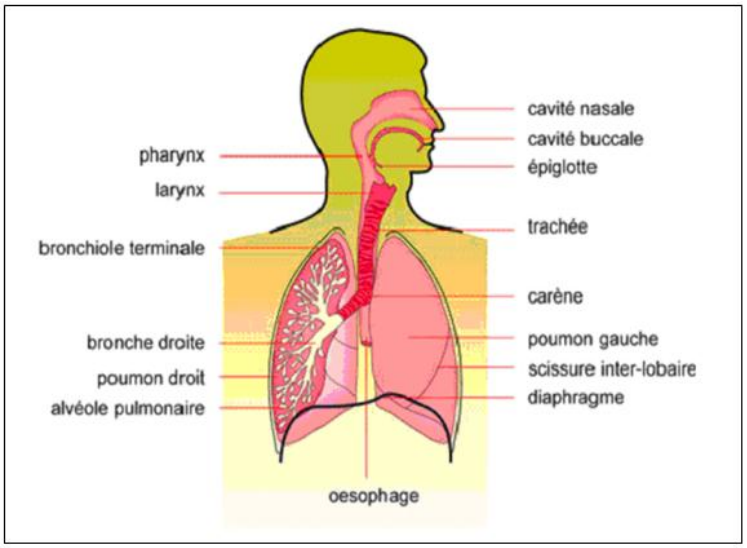


Figure 1.1 : Modèle simplifié de l'appareil phonatoire [2]

1.3.1 Les voies aériennes inférieures

Les voies aériennes inférieures correspondent à la partie de l'appareil phonatoire située dans le thorax et sont composées de deux poumons reliés à la trachée artère qui elle-même remonte jusqu'aux voies aériennes supérieures. La fonction première des poumons est d'assurer la fonction de respiration en permettant l'échange d'oxygène et de dioxyde de carbone entre le sang et l'air extérieur. Lors de la phonation, les poumons jouent le rôle de réservoir de pression et permettent de générer l'écoulement d'air à l'origine de la production de sons et notamment des vibrations des cordes vocales. La circulation de l'air entre les poumons et l'extérieur est réalisée grâce aux mouvements du diaphragme et aux contractions et relâchements des muscles de la cage thoracique. Cette ventilation se fait ainsi dans un mouvement de va-et-vient correspondant alternativement à l'inspiration et à l'expiration. Sauf pour des cas atypiques, la phonation intervient durant la phase d'expiration.

1.3.2 Le larynx

Le larynx est l'organe qui fait la jonction entre la trachée et le pharynx. Il se situe dans la gorge et remplit trois fonctions. La première est respiratoire puisque qu'il est le premier organe des voies aériennes supérieures. Ensuite, il joue un rôle dans la déglutition en bloquant l'accès aux poumons pour les aliments en cas de « fausse route», c'est- à -dire un dysfonctionnement de l'épiglotte sensée diriger ces aliments vers le tube digestif. Enfin, il abrite les cordes vocales et est donc le siège de la production des sons voisés, qui implique la vibration des cordes vocales. Le larynx est composé de cartilages, de muscles et de muqueuses. Une coupe coronale d'un pli vocal montre que celui-ci est constitué de plusieurs couches de tissus musculaires, ligamentaires et muqueux, dont les propriétés sont différentes (Fig.1.2).

La structure musculo-cartilagineuse autour des cordes vocales permet de faire varier leur tension, leur position et l'espace compris entre ceux-ci, appelé la glotte. Lors de la production des sons voisés de la parole (comme les voyelles par exemple), c'est la vibration des cordes vocales qui constitue la source des ondes acoustiques.

1.3.3 Le conduit vocal

Le conduit vocal est la partie des voies aériennes supérieures située au-dessus du larynx, il est localisé dans la tête. Constitué du pharynx et de deux cavités résonnantes séparées par le palais : la cavité orale (ou buccale) et la cavité nasale. Lorsque les cordes vocales vibrent, les

ondes acoustiques générées se propagent dans ces cavités qui agissent comme un résonateur acoustique (Fig.1.2).

La géométrie du conduit vocal influe ainsi sur les fréquences de résonance et sur le signal acoustique de parole émis. Pour un individu, la forme de la cavité nasale est invariable. En revanche la forme de la cavité orale peut être modifiée grâce à plusieurs articulateurs comme la langue, les lèvres ou la mâchoire. Le voile du palais (ou velum) situé à l'extrémité du palais dur est constitué d'une membrane et d'un muscle et joue le rôle de clapet entre ces deux cavités. Selon sa position, il permet soit d'obturer la cavité nasale, soit d'obturer la cavité orale ou soit de relier les deux cavités. Certains sons de la parole (comme les consonnes fricatives, par exemple) sont produits dans le conduit vocal et n'impliquent pas les vibrations des cordes vocales.

L'appareil phonatoire humain permet de générer une grande variété de sons. En contrôlant simultanément et de façon dynamique la pression pulmonaire, les propriétés des cordes vocales et la configuration géométrique du conduit vocal, un locuteur peut produire l'ensemble des voyelles et des consonnes qui constituent un langage [3].

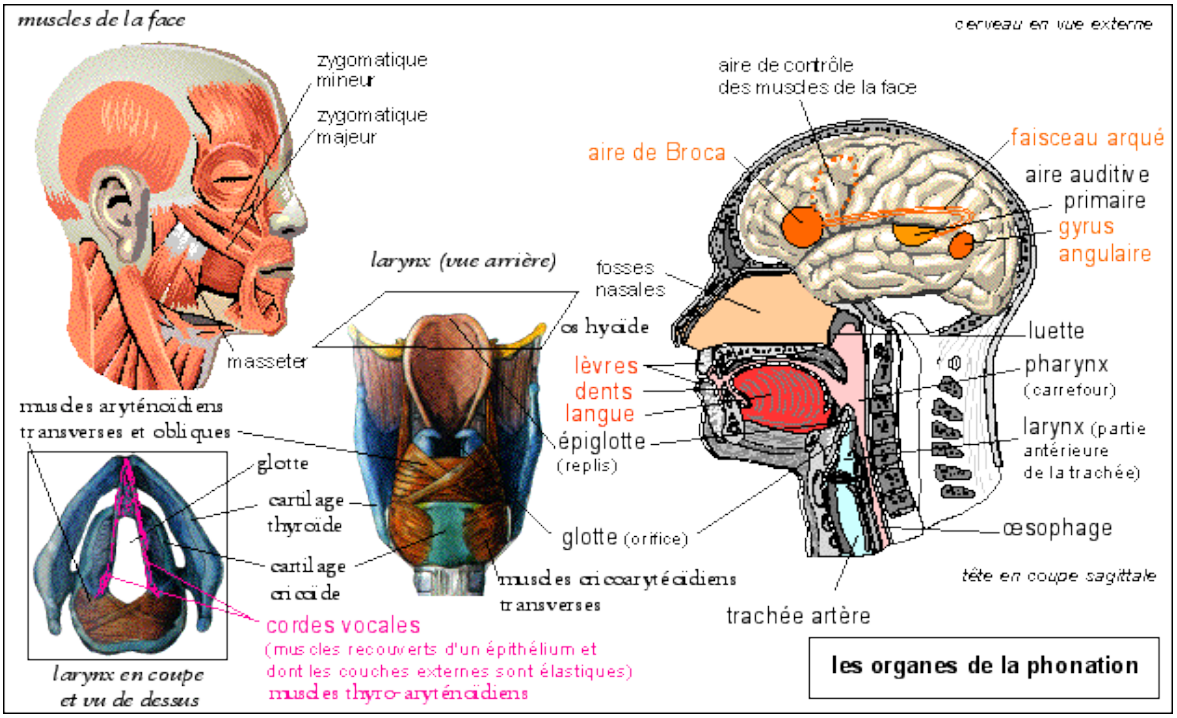


Figure 1.2 : les organes de la phonation [4]

1.4 LA PRODUCTION DE PAROLE

La production de la parole est l'opération la plus complexe de l'activité biologique humaine, et du monde vivant connu. Elle met en jeu un très grand nombre de muscles aux mouvements particulièrement précis, caractérisés par de très nombreuses unités motrices, dont la synchronisation doit être parfaitement contrôlée pour créer l'objet sonore porteur de sens.

La production de la parole est un système dynamique, dont le comportement à un moment donné dépend de ses états antérieurs. Le système est donc dépendant d'une variable paramétrable en fonction du temps qui dans ce cas est un geste articulatoire.

La phonologie articulatoire est basée sur la définition des phonèmes en termes de gestes, qui sont les unités d'action et les bases de contraste des items linguistiques, les atomes de la description phonologique. Les phonèmes sont donc définis par des groupes de gestes.

Ainsi, la phonologie peut être considérée comme un ensemble de relations parmi les gestes, et les événements physiques réels, qui caractérisent les systèmes de production de la parole.

Il existe deux classes de gestes dans la production de la parole.

- d'une part les mouvements des organes articulateurs constituant le conduit vocal, qui permettent de produire la parole.
- d'autre part, les mouvements qui, pendant le discours, apportent dans des proportions variables un complément d'information au message parlé et ne participent pas directement aux mécanismes de sa création. Nous les appelons, par simplification, mouvements d'accompagnement de la parole. Ils sont généralement des mouvements de mimiques de la face, des mouvements de la tête et des segments des membres supérieurs. Ils suivent surtout les variations prosodiques du discours et sont caractérisés par une vitesse de réalisation relativement faible. La technique de choix pour l'étude de ces mouvements est la vidéo cinématographique qui permet de les décomposer image par image et ainsi « d'arrêter le temps » pour les étudier.

1.4.1 Les gestes articulatoires

Dans la production de la parole, les gestes sont définis en termes de constriction dans le conduit vocal et sont, pour la dynamique des tâches, des mouvements ayant une trajectoire vers une cible spatiale. Pour générer ces trajectoires, il faut connaître pour chaque variable pertinente l'état instantané du système, la position de la nouvelle cible et les valeurs d'amortissement du système.

1.4.1.1 La phonologie articulatoire

La phonologie articulatoire est dérivée du modèle de la dynamique des tâches dont le but principal est l'étude des mouvements, de leurs compensations et de la chronologie de ces compensations.

La phonologie articulatoire proposée par Browman et Goldstein est le seul modèle qui introduit explicitement par sa nature spatiotemporelle, un aspect dynamique en phonologie qui peut être alors considérée comme un ensemble de relations parmi les gestes, événements physiques réels, qui caractérisent les systèmes et les modèles de production. Dans cette perspective, il est donc essentiel de pouvoir observer, enregistrer et mesurer les gestes articulatoires [5].

1.4.1.2 Les organes articulateurs et leur contrôle

Les organes articulateurs mis en jeux dans la parole ne sont pas spécifiques à sa production. Pour parler, l'homme utilise les deux grandes fonctions physiologiques que sont la respiration et la digestion. Si cette dernière n'est utilisée que partiellement pour la production de la parole dans ses voies supérieures (mastication et déglutition), en revanche la respiration est utilisée dans sa totalité.

Les organes actifs dans la production de la parole sont dans l'ordre anatomique du conduit vocal (de bas en haut) : la musculature respiratoire, le larynx, le pharynx, le voile du palais, la langue, la mandibule et les lèvres. Les sons de la parole ont pour origine des phénomènes aérodynamiques et acoustiques, à partir de l'air mis en pression dans les poumons et modulé par les différentes constrictions et les variations de longueur du conduit vocal ; la constriction laryngienne, source de la voix, les constrictions vélo-pharyngale, linguo - palatine et labiale qui différencient les segments phonologiques des langues, consonnes et voyelles (Fig.1.3).

Avec un tel instrument à sa disposition, l'homme est donc capable de produire un grand nombre de sons différentiés, de l'ordre de 150 [5], dans lesquels il pioche pour établir des codes phonologiques très variés lui permettant de communiquer par la parole. Ceci explique la variété infinie des langues du monde.

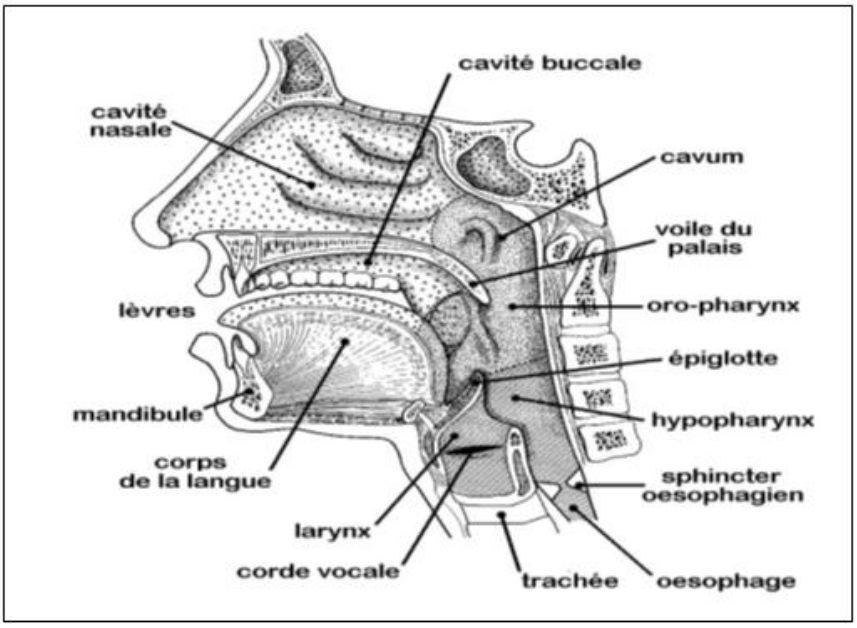


Figure 1.3 : Coupe du conduit vocal et des principaux organes de la production de la parole [5]

1.4.2 Les méthodes aérodynamiques

Le domaine de l'aérodynamique dans la production de la parole est fondamental par le fait qu'il est à l'origine de la création de tous les sons porteurs du code linguistique d'une langue. Ce sont toujours des phénomènes aérodynamiques qui sont à l'origine des bruits des consonnes plosives (explosion) et fricatives (turbulences d'écoulement de l'air) au niveau des constrictions du conduit vocal (glottale, linguo-palatale, labiale).

1.4.2.1 L'air moteur de la parole

La parole est un objet sonore complexe, ces sons et bruits sont le résultat d'une utilisation rigoureuse et précise du courant d'air généré par les poumons. Le rôle essentiel du contrôle de la respiration lors de la production de la parole a été mis en évidence dans un grand nombre de travaux. Alors que la respiration dite normale est un phénomène "automatique", la respiration dans l'acte de parole est contrôlée et organisée de façon très fine. L'expiration doit permettre de fournir et de maintenir une pression sous glottique stable pendant toute la durée de la phrase. Le contrôle de l'expiration s'effectue par le jeu des muscles de la respiration essentiellement les intercostaux et le diaphragme dont le rôle n'est pas limité qu'à la phase inspiratoire.

Le groupement fonctionnel des muscles, contrôlés par un tel système, est connu comme "structure coordinative". Dans le cas de la phonation, il semble que la structure coordinative respiratoire initiée en début de la phrase (ligne de base correspondant aux groupes de souffle) réagisse localement aux résistances apportées à l'écoulement de l'air au niveau de la glotte et au niveau de la cavité buccale (distinction de voisement, et division voyelle et consonne).

1.4.2.2 Les paramètres aérodynamiques

Les paramètres aérodynamiques sont au nombre de quatre. Le principal paramètre physiologique de la production de la parole est la pression de l'air contenu dans les poumons ou **Pression Sous Glottique**(**PSG**). Elle s'exprime en hectoPascal (hP) relativement à la pression atmosphérique, le Pascal étant l'unité internationale de pression. On trouve dans les études antérieures aux années 80 des unités telles que le millibar (mB) ou le centimètre d'eau (cm H_20) qui ont des valeurs très proches de l'hectoPascal. On peut également rencontrer l'unité de pression anglo-saxonne « **P**ound **S**quare **I**nch » ou **PSI** qui vaut approximativement 75 hP.

La **Pression Intra Orale** (**PIO**), ou pression supra glottique, est la pression qui règne dans la cavité oro - pharyngale. Elle s'exprime également en hP. La différence entre la PSG et la PIO est le moteur de la source vocale. La pression intra orale est un indicateur de l'ensemble des états des constrictions du conduit vocal et à ce titre joue un rôle fondamental dans leur contrôle.

Un débit étant un déplacement d'air d'une zone de haute pression vers une zone de basse pression, la pression intra orale génère à travers les constrictions labiales et lingo -palatine le **débit d'air oral** émis à la bouche. Elle génère également à travers la constriction vélo pharyngale le **débit d'air nasal** émis aux narines. Ces débits s'expriment en litre ou décimètre cube (dm^3) par seconde ou encore leur sous multiple le centimètre cube (cm^3). L'amplitude de ces débits dépend de la pression différentielle aux bornes de la constriction et de la résistance qu'elle oppose à l'écoulement de l'air qui la traverse.

1.4.2.3 L'aérophonométrie

L'aérophonométrie est l'ensemble des techniques de mesures des paramètres aérodynamiques dans la production de la parole.

La connaissance des variations de ces paramètres en fonction des segments phonémiques prononcés, donne de bonnes informations sur les mouvements des organes articulateurs du conduit vocal. On doit là encore à Rousselot les premières codifications des principaux

paramètres aérodynamiques et la description de la majorité des mécanismes articulatoires des langues. Cependant il est illusoire d'attendre une relation linéaire entre les variations de débits, de pression et des gestes articulatoires. Cependant, ils permettent une remarquable description chronologique et sont indispensables dans l'étude des phénomènes consonantiques complexes et de la coarticulation, la bonne pratique des méthodes aérodynamiques nécessite des précautions rigoureuses, mais n'est pas très compliquée (Fig.1.4.) [5].

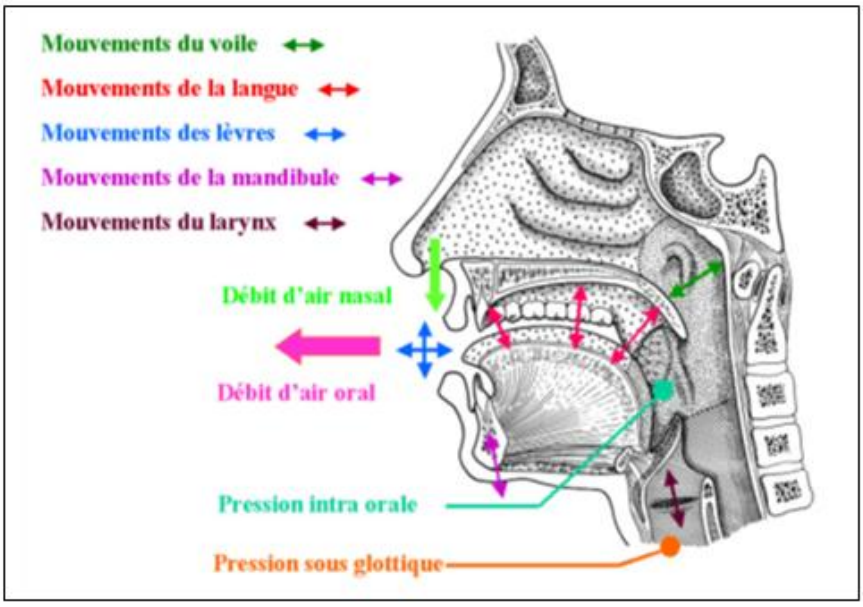


Figure 1.4 : Dynamique et lieux de constriction des principaux organes articulateurs du conduit vocal et situation anatomique des paramètres aérodynamiques [5]

1.5 SYSTEME DE PERCEPTION AUDITIVE

Le système auditif comporte trois parties qui jouent chacune un rôle spécifique dans la perception du son : l'oreille externe, l'oreille moyenne et l'oreille interne

1.5.1 L'appareil auditif

Cette division se fait en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan.

Une fois générées, ces vibrations sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en un flux nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement. Une description détaillée de l'oreille permettra au lecteur de mieux appréhender les différents organes qui la constitue et de mieux visualiser leur répartition. Il faut noter que la présence des deux oreilles permet d'effectuer, au niveau du cerveau, des traitements plus complexes que le simple décodage d'une scène auditive (Fig.1.5).

Le positionnement des oreilles de chaque côté du crâne permet en effet de profiter des capacités de la binauralité. Cette faculté permet de calculer la provenance d'un son en fonction du retard d'arrivée de ce son dans une oreille par rapport à l'autre. Il est à noter que cette binauralité permet à l'homme de discerner la position horizontale de l'émetteur d'un son mais pas sa position verticale [6].

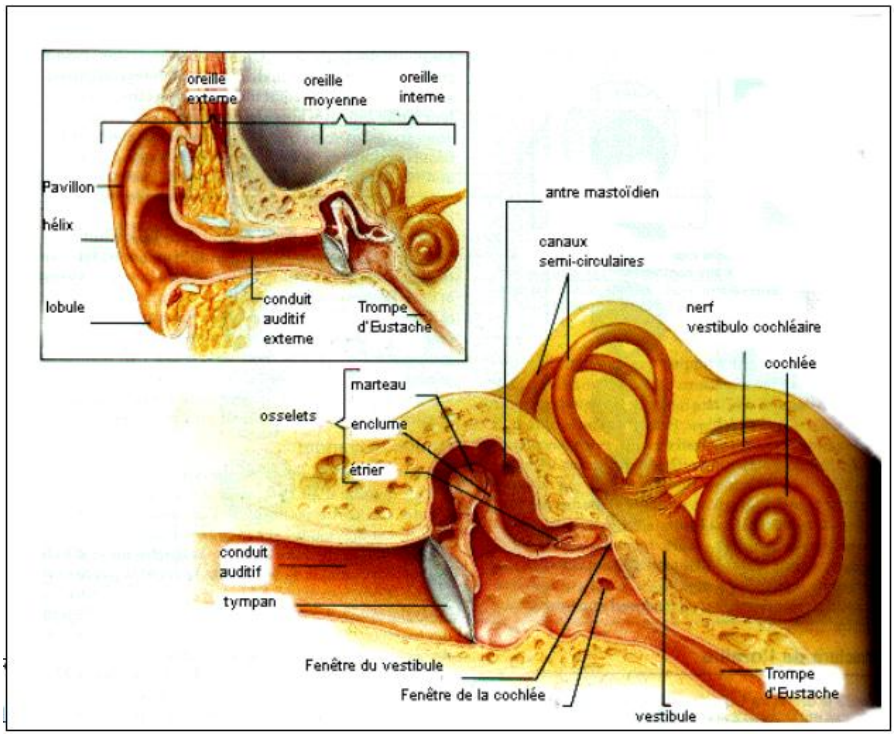


Figure 1.5 : Coupe de l'appareil auditif humain [6]

1.5.2 Perception des voyelles et des consonnes

Il est impossible d'identifier clairement les voyelles dans les notes aiguës chantées par les chanteurs d'opéra, lorsque la hauteur de la voix dépasse celle du second formant.

On ne peut imaginer une langue qui n'aurait que des voyelles, car combinées, elles formeraient des unités mal structurées et mal différentiables.

Même chose pour les consonnes. Les brèves [p, t, k] et les aiguës [f, s] sont trop difficiles à percevoir sans le secours d'un appui vocalique. Les consonnes fricatives ont des bruits de friction beaucoup plus élevés que les harmoniques des voyelles. [z] et [s] vont jusqu'à 8000 Hz alors que les harmoniques les plus hautes des voyelles ne dépassent guère 4000 Hz.

Les phonèmes les plus faciles à entendre sont les voyelles, parce qu'elles se situent dans les fréquences graves (qui sont les plus aisées à percevoir), et parce que leur sonorité intrinsèque est beaucoup plus forte que celle des consonnes.

Dans la chaîne sonore ce sont les consonnes qui assurent la compréhension du contenu de l'énoncé, par la structuration sonore qu'elles engendrent, alors que les voyelles assurent un certain niveau sonore d'audibilité. On parlera alors de « structurabilité consonantique » assurant le sens de l'énoncé et d' «audibilité vocalique». On explique ce phénomène par le fait que les voyelles ont moins de traits distinctifs entre elles que les consonnes. La structuration plus complexe des consonnes en termes de point d'articulation permet à l'auditeur de les localiser plus facilement et plus spécifiquement. L'articulation plus ouverte des voyelles en revanche permet de les rendre plus sonores.

Notre oreille à ses limites et le champ de la perception auditive se situe entre deux seuils :

- le seuil d'audibilité
- le seuil de douleur;

Elle ne perçoit pas les sons trop graves (infrasons), et les sons trop aigus (ultrasons). La hauteur et l'intensité vont se combiner pour déterminer les seuils de perception (Fig.1.6).

L'oreille ne supporte pas les sons trop intenses. Le seuil de douleur (ou de traumatisme) se situe au-dessus de 140 dB. C'est sur une étendue d'environ 10 octaves, entre 16 et 16 000 Hz que se situent les sons théoriquement audibles [7].

La zone de perception va varier pour les différents auditeurs car elle est liée à des facteurs comme l'âge ou l'état de santé.

Notre oreille est aussi sélective. Elle ne perçoit que ce qu'elle a appris à percevoir. L'acte perceptif est un phénomène très complexe [7].

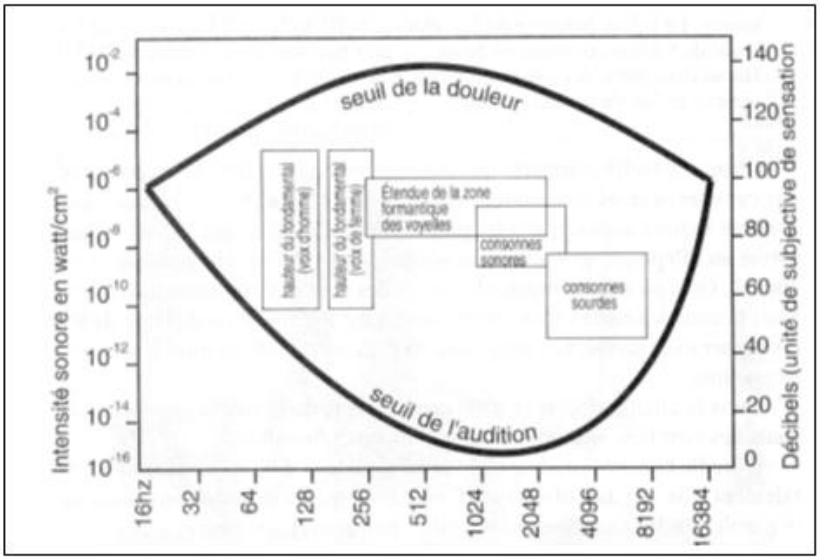


Figure 1.6 : Aire des sons de la parole [7]

1.6 LES PARAMETRES PROSODIQUES ET ACOUSTIQUES D'UN SIGNAL VOCAL

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

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

1.6.1 La Fréquence Fondamentale

La Fréquence Fondamentale ou $\mathbf{F_0}$ est la fréquence de vibrations des cordes vocales, elle varie d'une personne à une autre en fonction de la longueur et de la masse des cordes vocales de chaque personne (Fig.1.7).

Elle permet de diviser l'ensemble des sons de parole en trois grandes macros classes [8]:

- 70 - 250 Hz pour les hommes ;

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

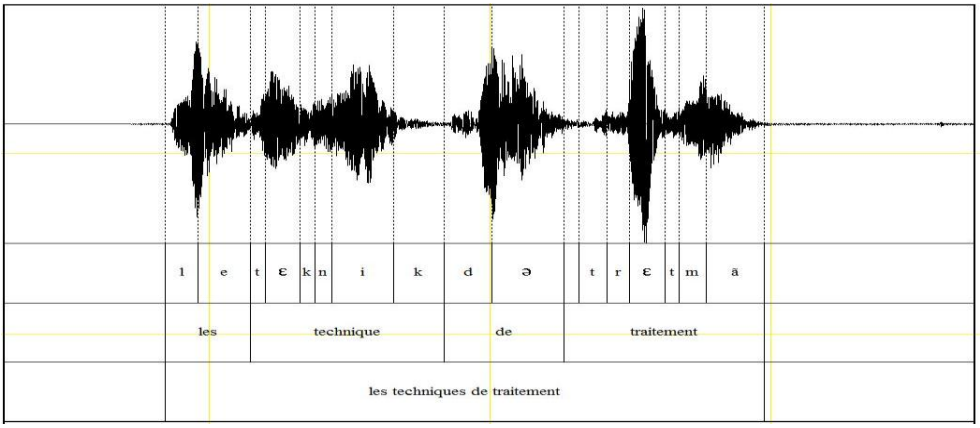


Figure 1.7 : Evolution de la fréquence de vibrations des cordes vocales de la phrase : "Les techniques de traitement "

Les variations de la fréquence au cours de la parole constituent ce qu'on appelle la mélodie ou l'intonation. Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la F_0.

1.6.2 La durée

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

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

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

1.6.3 L'intensité ou l'énergie

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

$$E=\frac{1}{T}\sum_{n=1}^{T} x_n{}^2 \quad (1.1)$$

$$E_{db}=10*\log_{10}\left(\frac{1}{T}\sum_{n=1}^{T} x_n{}^2\right) \quad (1.2)$$

Le rythme d'élocution correspond à la vitesse du débit de parole. On peut faire varier ce paramètre de manière à ce qu'une phrase prononcée trop rapidement puisse être « ralentie » pour la rendre plus compréhensible lors de l'apprentissage d'une langue étrangère par exemple. L'intensité du son émis est liée à la pression de l'air en amont du larynx. Les figures 1.8 et 1.9 représentent l'évolution temporelle du signal vocal pour les mots cinq et parenthèse, elles donnent un exemple des parties voisées et non voisées du signal vocal.

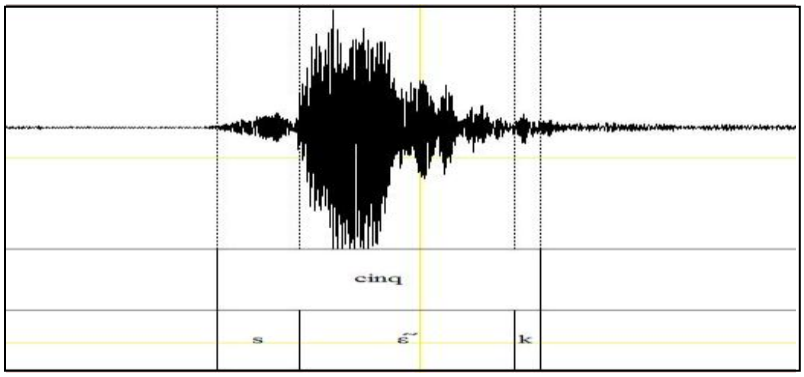


Figure 1.8 : Audiogramme du signal de parole du mot « Cinq » [$s\tilde{\varepsilon}k$]

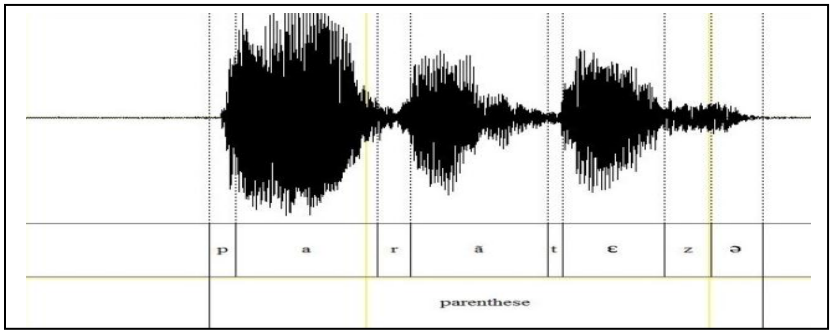


Figure 1.9 : Audiogramme du signal de parole du mot « Parenthèse » [par$\tilde{a}$tɛzə]

Tout l'enjeu du traitement de la parole est de modéliser l'appareil phonatoire humain de façon à créer un signal de parole synthétique aussi réaliste que l'original. Il existe plusieurs manières de le faire, notamment en utilisant un vocodeur à prédiction linéaire qui, dans un premier temps, code le signal vocal de manière à réduire le débit d'informations puis le restitue à l'aide de paramètres qui caractérisent la fonction de transfert du conduit vocal. Ces paramètres étant réactualisés toutes les 20 ms environ. En fait, on part de l'hypothèse qu'un échantillon de parole peut être prédit à partir d'une pondération linéaire d'un nombre fini d'échantillons précédents.

Cette hypothèse se justifie par le fait que la forme du conduit vocal n'évolue pas rapidement. En général, on considère que l'appareil vocal est quasi stationnaire sur un intervalle de temps de l'ordre d'une vingtaine de millisecondes. On parle donc ici de statistique du signal à court terme. Cette méthode a l'avantage de donner de bons résultats au niveau du signal synthétique mais demande des capacités de calcul important pour la réalisation en temps réel [8].

1.6.4 Les Formants

Lorsqu'un excitateur entre en vibrations et fournit un signal, ce dernier passe à travers une cavité de résonnance (le résonateur) qui a amplifié certains composants. On obtient alors ce qu'on appelle les formants qui sont des facteurs essentiels dans la caractérisation du timbre.

L'appareil phonatoire étant constitué de différentes cavités. Lors du passage de l'air à travers ces cavités, il est amplifié et subit différentes transformations dues aux degrés d'ouverture et de fermeture au niveau de chaque cavité, à la position de la langue, des lèvres, etc. Ces cavités possèdent des fréquences de résonance qui renforcent certaines régions du spectre de sources excitatrices. Les maxima de la courbe de réponse en fréquences du conduit vocal sont appelés Formants. Chaque son à ses formants caractéristiques. Sur un spectrogramme, les formants sont représentés par des bandes noires (le degré de noirceur correspondant à l'énergie) (Fig.1.10).

La fréquence fondamentale (fréquence de vibrations des cordes vocales) est responsable de la hauteur perçue d'un son. Les fréquences d'harmonique renforcées, responsables du timbre d'un son, sont elles aussi numérotées. Le premier Formant F_1 correspond à la première zone des harmoniques renforcées, F_2 à la seconde et ainsi de suite jusqu'à F_5.

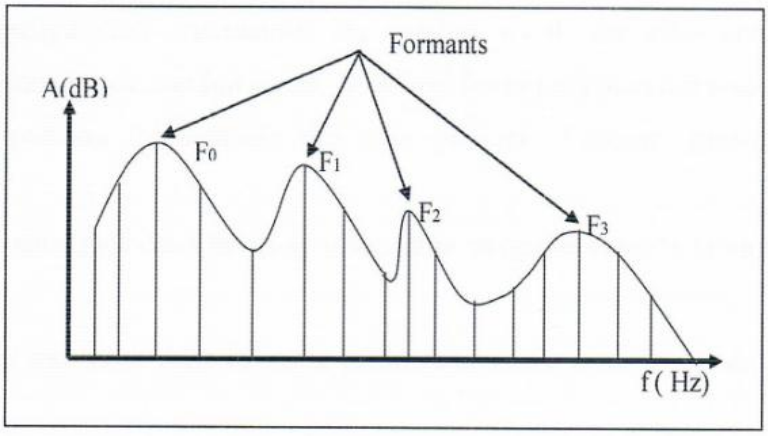


Figure 1.10 : Représentation des Formants d'un son voisé [17]

Les valeurs des formants sont très influencées par le lieu d'articulation des phonèmes. Elles donnent une image de la configuration articulatoire du conduit vocal, car elles correspondent aux fréquences de résonances du conduit vocal.

Le nombre des formants, selon les caractéristiques du résonateur (volume, forme et ouverture), est variable : d'un seul (théoriquement) à une infinité. Néanmoins, du point de vue perceptif, seuls quelques-un d'entre eux jouent un rôle central au niveau de la parole. Par exemple, on peut caractériser toute voyelle en ne prenant compte que ses trois premiers formants F_1, F_2 et F_3. (Pour une réalisation de la voyelle [i] par exemple, les trois premiers formants pourraient se situer respectivement à 300, 2200 et 3000 Hz).

Généralement, nous pouvons aller jusqu'à cinq ou six formants pour produire une parole de très haute qualité. Les formants nous permettent de décrire aussi les cibles vocaliques correspondant aux zones stables ainsi que les zones de transitions (passage entre deux sons consécutifs) ce qui montre leur très grande importance pour l'analyse acoustique en phonétique au moins trois formants sont exigés pour produire les différentes voyelles généralement, on peut aller jusqu'à cinq formants pour produire une parole de haute qualité.

Les valeurs des formants sont très influencés par le lieu d'articulation, ils donnent une image de la configuration articulatoire du conduit vocal, car elles correspondent aux fréquences de résonance du conduit vocale, de même des expériences qui restent à vérifier ont montré que la position fréquentielle des trois premiers formants caractérise le timbre vocalique [17] :

- F_1 prend naissance dans la cavité résonante comprise entre le larynx et le dos de la langue ;
- F_2 prend naissance dans la cavité résonante située entre le dos de la langue et les lèvres ;
- F_3 dépond de l'arrondissement des lèvres.

A cela on ajoute le fait que plus la cavité de résonance est large, plus la fréquence correspondante est basse. Réciproquement plus cette cavité est petite plus la fréquence et haute.

1.7 LA COMPLEXITE DE SIGNAL VOCAL

La grande difficulté du TAP et en particulier celui de la **R**econnaissance **A**utomatique de la **P**arole **(RAP)** provient du caractère du processus de la Communication Parlée et des caractéristiques intrinsèques du signal vocal. La parole est un signal continu d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps ; périodique ou plus exactement pseudo périodique pour les sons voisés, aléatoire pour les sons fricatifs et impulsionnels pour les sons occlusifs

1.7.1 Continuité

Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont l'occurrence est aléatoire. Il peut très bien y avoir des intervalles de silence au milieu d'un mot et aucun intervalle entre deux mots successifs. Par conséquent, il est très difficile de déterminer le début et la fin des mots composant la phrase.

1.7.2 Variabilités

La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que ce soit pour un même ou plusieurs locuteurs. Parmi ces facteurs, les perturbations apportées par le microphone (selon le type, la distance et l'orientation) et l'environnement (bruit et réverbération). De telles variations ne donnent pas naissance à de nouveaux phonèmes, puisqu'elles ne portent aucune information sémantique. Ainsi, les phonèmes apparaissent sous une multitude de formes articulatoires, appelées allophones ou variantes.

1.7.2.1 Variabilité intra-locuteur

La variabilité intra-locuteur concerne les différences de production du signal de parole chez un même locuteur. Plusieurs critères peuvent être responsables de ces différences :

- la fatigue ;
- l'état émotionnel du sujet qui affecte le timbre et le rythme de la voix ;
- les maladies affectant les organes de la voix.

1.7.2.2 Variabilité interlocuteur

Des différences acoustiques apparaissent dans un mot prononcé par plusieurs locuteurs. En effet, des contrastes considérables peuvent se manifester suivant l'âge, le sexe, l'origine géographique et le milieu social.

1.7.2.3 Variabilité contextuelle

En effet, les mouvements articulatoires peuvent être modifiés de façon à minimiser l'effort à produire pour les réaliser à partir d'une position articulatoire donnée, ou pour anticiper une position à venir. Ces effets sont connus sous le nom de *réduction, d'assimilation* et de *coarticulation.*

Les phénomènes coarticulatoires sont dûs au fait que chaque articulateur évolue de façon continue entre les positions articulatoires. Ils apparaissent mêmes dans le parler le plus soigné.
Au contraire, la réduction et l'assimilation prennent leur origine dans des contraintes physiologiques et sont sensibles au débit de la parole. L'assimilation est causée par le recouvrement de mouvements articulatoires et peut aller jusqu'à modifier un des traits phonétiques du phonème prononcé. La réduction est plutôt due au fait que les cibles articulatoires sont moins atteintes dans le parler rapide.
Ces phénomènes sont en grande partie responsables de la complexité des traitements réalisés sur les signaux de parole [9].

1.7.3 Coarticulation

Le signal de parole est constitué d'une succession d'unités différentes. Cependant, contrairement à ce qu'on pourrait croire, ces unités ne sont pas indépendantes les unes des autres mais s'influencent mutuellement : c'est le phénomène de coarticulation. En effet, quand on produit de la parole, on ne produit pas des segments individuels les uns après les autres : la parole n'est pas de l'épellation. Au contraire, la parole est produite par les gestes des différents articulateurs du conduit vocal (larynx, langue, lèvres, mâchoire, velum) qui se chevauchent en partie au cours du temps car ils subissent des influences diverses [10].

1.7.4 Redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter les durées du traitement, lequel doit se faire en temps réel. A l'inverse, le débit ne doit pas être trop faible pour conserver un bon rapport signal/bruit. En effet, Il existe une grande disproportion entre le débit du signal enregistré et la quantité utile pour une tâche de reconnaissance [9].

1.8 CLASSIFICATION DES SONS

D'un point de vue linguistique, la production des sons ou d'un mot réside dans la production en série de tous les phonèmes constituant ce mot. Ces phonèmes forment les unités phonétiques qui sont classées en voyelles, consonnes et semi-voyelles.

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode et lieu d'articulation. Dans la cavité buccale, le point d'articulation est l'endroit où se trouve un obstacle au passage de l'air. D'une manière générale, le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air (Fig.1.11).

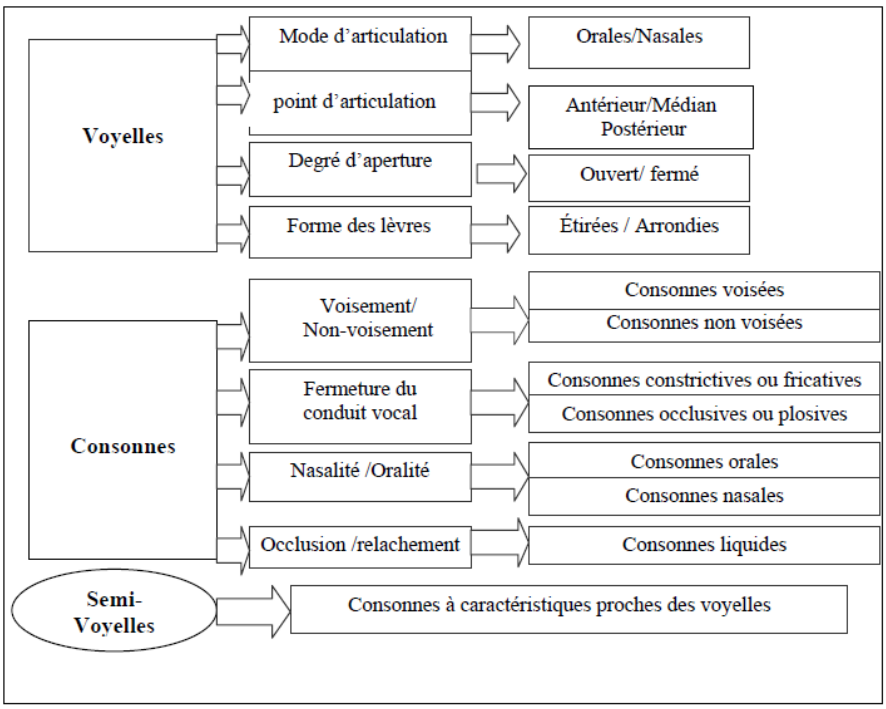


Figure 1.11 : Classification des sons du langage [9]

1.8.1 Les sons voisés

Les vibrations des cordes vocales produisent les sons voisés (voyelles, semi-voyelles, consonnes nasales, etc.). Les cordes vocales sont des replis musculaires recouverts d'une muqueuse, attachés aux trois cartilages (thyroïde, aryténoïdes) situés à l'extrémité de la trachée artère et constituant le larynx. Leur vibration est en fait leur accolement, puis leur séparation sous l'effet de la pression de l'air provenant des poumons, et de nouveau leur accolement sous l'effet des forces de Bernoulli produites par le passage de l'air. Les cartilages sur lesquels s'accrochent les cordes vocales régularisent la tension des cordes vocales, donc la fréquence des vibrations, au moyen des muscles du larynx s'appelle la fréquence fondamentale ou $\mathbf{F_0}$.

1.8.2 Les sons non voisés

Le second mode d'excitation est obtenu par divers bruits produits par le passage de l'air en un point de resserrement du canal vocal ou par des bruits d'occlusion ou de plosion, provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais. Dans cette catégorie de sons, les cordes vocales ne vibrent pas.

1.8.3 Les voyelles

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Quand ce dernier est suffisamment ouvert pour que l'air expiré par les poumons, le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la cavité buccale se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. Une voyelle se caractérise par un passage libre de l'air dans le conduit vocal et par les vibrations des cordes vocales.

Elles se différencient principalement les unes des autres par leur lieu d'articulation (position de la langue), leur degré d'ouverture (espace compris entre la pointe de la langue et le palais), et leur nasalisation. Nous distinguons ainsi, selon la localisation de la masse de la langue, les voyelles antérieures ou avant, les moyennes, les voyelles postérieures (ou arrières), l'écartement entre l'organe, le lieu d'articulation, et selon les voyelles fermées et ouvertes.

Les voyelles orales sont dues à une élévation du palais qui détermine la fermeture des fosses nasales ainsi qu'à l'écoulement de l'air expirer à travers la cavité buccale. Par contre, les

voyelles nasales sont caractérisées par l'écoulement d'une partie de l'air à travers la cavité nasale. L'Arabe Standard (AS) ne possède pas de voyelles nasales. Elles sont représentées sur un plan dont les axes sont les formants F_1 et F_2. Elles tracent alors un triangle dont les extrémités sont occupées par les voyelles [i, u, a]. Ce triangle représente également les positions de la langue dans la cavité buccale selon deux axes : antérieur à postérieur (avant et arrière) et de fermé à ouvert, selon que la langue est massée en avant et vers la zone dentale pour [i], basse et étalée loin du palais pour [a] (ouvert), ou massée postérieurement vers le voile pour [u] dont laquelle les voyelles soulignées sont labialisées (arrondies) (Fig1.12).

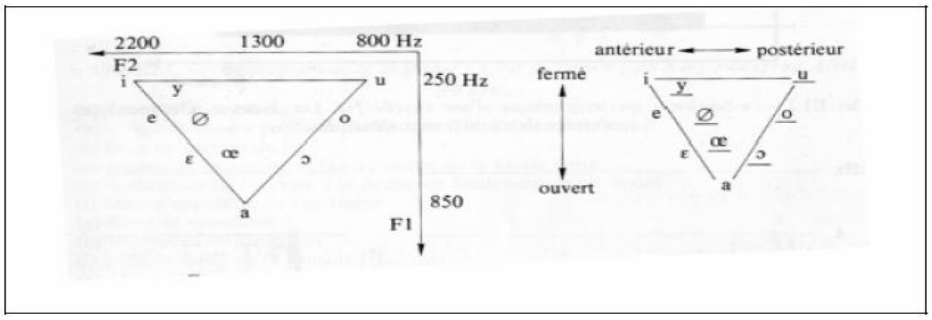


Figure 1.12 : Relation acoustico-articulatoire des voyelles orales du Français [11]

1.8.4 Les consonnes

Les consonnes se caractérisent par une fermeture partielle du conduit vocal ou constriction (constrictives ou fricatives) ou totale du conduit vocal (occlusion) : occlusives ou plosives. Nous classons principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- intervention ou mise en vibrations des cordes vocales : articulation sonore ;
- fermeture momentanée du passage de l'air suivie d'une ouverture brusque (explosion): articulation occlusive ;
- rétrécissement du passage de l'air qui produit un bruit de friction ou de frôlement : articulation fricative ;
- position abaissée du voile du palais: articulation nasale ;
- contact de la langue au milieu du canal buccal ; l'air sort des deux côtés ;
- une série d'occlusions brèves ; séparées de la luette : articulation vibrante.

La distinction du mode d'articulation conduit à deux classes : les fricatives ou constrictives et les occlusives ou plosives. Les consonnes fricatives appelées également spirantes sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais, les dents ou les lèvres. Les fricatives non voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et d'autres turbulentes : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète.

Les consonnes occlusives ou plosives sont reconnues grâce au silence provenant de la fermeture totale du conduit vocal ou occlusion. Cette dernière comporte trois phases :

- l'implosion ou fermeture ;
- l'occlusion proprement dite tenue de la fermeture ;
- l'explosion ou détente.

Les consonnes liquides combinent une occlusion et une ouverture simultanée du conduit vocal. Elles sont caractérisées par un degré de sonorité proche de celui des voyelles. Enfin, les consonnes nasales font intervenir la cavité nasale par abaissement du voile du palais. Elles sont produites par l'écoulement de l'air phonatoire dans le conduit nasal.

1.8.5 Les semi-voyelles

Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, elles sont assez difficiles à classer [9].

1.9 LE TIMBRE

La caractéristique première de la voix d'un locuteur est son timbre, qui n'est perceptible que sur les sons voisés, et surtout les voyelles et semi-voyelles. Le timbre est totalement exprimé par le spectre du signal. Or, le signal vocal est constitué de la convolution du signal glottique (excitation) et de la réponse impulsionnelle du canal vocal. Son spectre est le produit du celle de la source et de celui du canal [12].

1.10 LA MELODIE

La source (impulsion de glotte) est caractérisée non seulement par son spectre, mais aussi par la période de vibration. La fréquence de la source est la fréquence fondamentale, F_0, qui est le pitch. Cette fréquence n'est pas stable. Elle varie très rapidement en fonction du temps (mélodie), et porte une information sémantique au moyen des patrons intonatifs ou de la micro mélodie (évolution du fondamental d'un phonème à un autre, ou même au sein d'un même phonème). La mélodie porte également une information sur l'identité du locuteur qui apparaît dans la distribution statistique de la fréquence (pitch moyen, variance de pitch, …) et dans l'évolution temporelle de l'élocution, chaque locuteur ayant des patrons intonatifs favoris, mais cette dernière caractéristique est très sensible à l'imitation. Le pitch moyen permet de discriminer aisément la voix des hommes de celles des femmes et des enfants, dont la tessiture est en moyenne plus élevée d'une octave [12].

1.11 ALPHABET PHONETIQUE INTERNATIONAL (API)

L'**A**lphabet **P**honétique **I**nternational (**IPA**) associe des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations.

Tableau 1.1 : Symboles de l'Alphabet Phonétique International utilisés dans la transcription du Français

Les voyelles orales

symboles	*mot-clé*	*autres graphèmes*
[i]	**lit**	stylo, île, maïs, meeting
[y]	**lune**	sûr, j'ai eu, aigüe
[u]	**tout**	où, goûter, football, août
[ə]	**je**	
[e]	**télé**	parler, nez, pied, et
[ɛ]	**mère**	faire, secret
[ø]	**feu**	nœud, jeûne
[œ]	**fleur**	cœur, club
[o]	**vélo**	sauter, peau, nôtre
[ɔ]	**pomme**	album, alcool
[a]	**patte**	
[ɑ]		pâte

Les voyelles nasales

[ɑ̃]	**gant**	camp, cent, empereur, paon, Caen
[ɔ̃]	**bon**	ombre
[ɛ̃]	**lapin**	chien, pain, daim, imparfait, syndicat, sympa
[œ̃]	brun	parfum

Les consonnes

[s]	**se**	ce, poisson, citron, garçon, science, dix, démocratie
[z]	**zéro**	maison, dixième, blizzard, -s *en liaison* (plus◡actif)
[ʃ]	**chat**	short, schéma, fascisme
[ʒ]	**jeune**	âgé, mangeons
[f]	**fou**	affaire, pharmacie
[v]	**vin**	wagon
[R]	**rare**	beurre
[l]	**lait**	elle
[p]	**papa**	appartement
[b]	**bébé**	abbaye
[m]	**mais**	flamme
[n]	**non**	anniversaire
[t]	**table**	patte, sept, -d *en liaison* (un grand homme)
[d]	**dos**	addition
[k]	**car**	accord, qualité, képi, orchestre, ticket, coq
[g]	**gâteau**	bague, aggraver, second

Les semi-voyelles

[w]	**oui**	toit, loin, poêle, jaguar, aquarelle
[ɥ]	**puis**	continuer, linguistique
[j]	**pied**	travail, payer, grenouille

1.12 NOTIONS FONDAMENTALES SUR L'ARABE STANDARD (AS)

L'Arabe est une langue parlée par plus de 337 millions de personnes. Elle est la langue officielle d'au moins 22 pays. C'est aussi la langue de référence pour plus de 1,3 milliard de musulmans. Comme son nom l'indique, la langue Arabe est la langue parlée à l'origine par le peuple arabe. Dans le cadre de notre travail, nous parlerons de la langue Arabe en référence à ce qui est communément appelé ″l'**A**rabe **S**tandard″ (**AS**), c'est-à-dire, la langue de communication commune à l'ensemble du Monde Arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel [9].

1.12.1 Système phonétique de l'Arabe Standard

L'Arabe **S**tandard(**AS**) comprend 40 phonèmes, dont 3 voyelles courtes, 3 voyelles longues plus 6 variantes vocaliques en contexte emphatique et 28 consonnes (Tableau 1.2). Les phonèmes arabes se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques. La graphie des lettres est différente selon leur position dans le mot. Ainsi, la lettre **ب** [b] est transcrite بَيْتٌ [bajtun] (une maison) en début de mot, خُبْزٌ [xubzun] (du pain) en milieu de mot, كَلْبٌ [kalbun] en fin de mot et قُرْبَ [qurba] (à proximité de) isolé en fin de mot.

Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscules/majuscules n'existe pas [9].

Pour les besoins de la transcription les 28 consonnes arabes ont été divisées en deux groupes :

- 14 consonnes solaires qui assimilent le « ل » de l'article ;
- 14 consonnes lunaires qui n'assimilent pas le « ل » de l'article.

Les solaires se prononcent en double, comme par exemple avec le mot « soleil » شمس [chams], au lieu de prononcer الشمس, el-chams, on prononce ech-chams, car la lettre ش [chin], est une lettre solaire.

Les lettres lunaires, se prononcent normalement et simplement pour elles-mêmes, c'est-à-dire sans les doubler. Par exemple avec le mot « lune », قمر ([qamar] - lune), on prononce القمر, [el-qamar] tout à fait normalement, parce que la lettre ق [qaf] est une lettre lunaire.

1.12.1.1 Les voyelles

On distingue trois voyelles courtes opposées à trois voyelles longues. La durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales et sont réparties comme suit :

- les voyelles courtes: [i], [a], [u] sont représentées dans un texte voyellé au-dessus ou au-dessous de la consonne, (ِ . َ . ُ) , exemple : تُرِكَ [turika] ;
- les voyelles longues [huruuf al madd] : [ā] ou [aa], [ī] ou [ii] et [ū] ou [uu], sont écrites sous forme de caractères consonantiques (ي, و, ا), et sont obligatoirement, représentées dans un texte écrite exemple :

  مُسَافِرُونَ [musaafiruuna].

1.12.1.2 Les consonnes

Les consonnes de l'Arabe peuvent être classées suivant plusieurs critères

Tableau 1.2 : Transcription Orthographique Phonétique de l'Arabe Standard [13]

Mode	Type de phonème		Phonèmes Arabes	Transcription Arabisante	Lieux d'articulation
Occlusives	Voisées		ب	b	bilabiale
			د	d	alvéodentale
	Non-Voisées		ق	q	uvulaire
			ت	t	alvéodentale
			ك	k	postpalatale
			ء	'	glottale
	Voisée	Emphatiques	ظ	ḏ	alvéolaire
	Non-Voisée		ط	ṭ	alvéodentale
Fricatives	Voisées		ز	z	sifflante
			ذ	ḏ	dorsoalvéolaire
			غ	ġ	interdentale
			ع	ʿ	uvulaire
			ح	ḥ	pharyngale
	Non-Voisées		س	s	sifflante dentale
			ث	ṯ	interdentale
			ف	f	labiodentale
			ش	š	chuinchante palatale
			خ	h	vélaire
			ه	h	glottale
			ح	h	pharyngale
	Voisée	Emphatiques	ص	ṣ	dorsealvéodentale sifflante
	Non-Voisée		ض	ḍ	interdentale
Nasales	Voisées		م	m	bilabiale
			ن	n	alvéodentale
Liquide	Voisée		ل	l	dentale
Affriquée	Voisée		ج	ǧ	alvéopalatale
Vibrante	Voisée		ر	r	apicoalvéolaire
Semi-voyelles	Non-Voisées		و	w	bilabiale
			ي	y	palatale

1.12.2 Particularités de l'Arabe Standard

Le système phonétique de la langue arabe diffère de celui des autres langues par la présence de voyelles longues (huruuf al madd), de phénomènes d'emphase et de la gémination. Ces caractéristiques donnent une valeur particulière à cette langue.

1.12.2.1 Voyelles longues

En Arabe Standard les voyelles longues présentent une caractéristique très importante au niveau sémantique. Par exemple, les deux mots $\tilde{g}$amal (chameau) et $\tilde{g}$am$\tilde{a}$l (beauté) ne différent que par l'allongement de la voyelle finale.

Sur le plan articulatoire, il existe une similitude entre les voyelles [i] et [ī], [u] et [ū] cependant une différence existe entre les voyelles [a] et [ā] car la position de la langue est plus basse pour le [a] que pour le [ā].

1.12.2.2 Gémination

Au niveau graphique, la gémination est symbolisée par le signe de la chadda qui signifie le dédoublement de la consonne. Sur le plan phonétique, l'opposition simple/géminée peut se résumer de la manière suivante : pour une consonne non-occlusive, l'opposition se réduit essentiellement à l'opposition temporelle brève/longue ; pour une occlusive, elle réside au niveau de la durée du silence. Ce rallongement entraîne l'accentuation des propriétés de la consonne (augmentation du caractère emphatique). Une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position (les lèvres ne se referment pas après le premier [b] dans [ka**bb**ara]. Dans beaucoup de langues, ce phénomène permet de mettre en relief un mot dans son contexte, alors qu'il s'avère être un élément distinctif sur les plans morpho-sémantiques en langue Arabe حضر : [had ara] ″ il a assisté″ est différente de
حضَّرَ [had dara] ″il a préparé″ où la deuxième consonne est géminée [8].

1.12.2.3 Emphase

Le mot emphase est habituellement utilisé pour rendre compte d'une manifestation prosodique liée à l'accentuation volontaire d'une syllabe. Chez les linguistes arabes, ils désignent certaines qualités que possèdent les consonnes :

- l'itbaq : les consonnes qui ont cette qualité sont ص [ʂ], ض [ð̪], ط [t̪], ظ [d̪]. Celles-ci sont pressées et produites par la langue élevée vers le palais ;
- le tafkhiim : son contraire est le tarqiiq. Il traduit une expression acoustique grasse et épaisse de certaines consonnes ;
- l'istilaa : cette qualité décrit le mouvement articulatoire que fait la langue quand elle mue vers la partie postérieure de la cavité buccale, avec ou sans tafkhiim.

Seules les consonnes ص [ʂ], ض [ð̪], ط [t̪], ظ [d̪] possèdent ces trois qualités et sont appelées consonnes *emphatiques* (ou consonnes pharyngalisées). Si nous comparons le Français à l'Arabe, nous constatons que la différence entre *patte* et *pâte* par exemple est rarement faite en Français « Standard ». En revanche, cette postériorisation a suscité beaucoup d'intérêts en ce qui concerne l'Arabe. Du fait de sa pertinence au niveau perceptif, la modélisation de l'emphase est primordiale en synthèse de la parole à partir du texte de l'AS. Sa prise en compte passe par l'introduction de nouvelles variantes de voyelles dans les contextes emphatiques. Néanmoins, sa mise en œuvre est directement liée à la technique de synthèse utilisée [14].

1.13 CONCLUSION

Dans ce chapitre nous avons exposé des notions de base sur le traitement de la parole, des spécifications du signal vocal et quelques caractéristiques de la langue Arabe Standard.
Les objectifs de ce chapitre sont de définir les notions que nous utiliserons dans notre travail. Cette partie théorique sera complétée dans le chapitre suivant par une étude approfondie des systèmes de synthèse de la parole et ses variantes.

Chapitre 2 :

La Synthèse de la Parole

2.1 INTRODUCTION

Ce chapitre nous permet de présenter les principales techniques de la synthèse de la parole, en premier lieu, nous allons donner une brève définition de la synthèse de la parole, et le traitement linguistique du texte, En outre, nous étudions les différentes techniques d'analyse du signal vocal, puis nous expliquons les méthodes de la synthèse de la parole ainsi que ses différentes applications.

2.2 DEFINITION DE LA SYNTHESE DE LA PAROLE

De nos jours, la synthèse vocale n'est plus un concept avant-gardiste mais elle aboutit à des produits de bonne qualité. En effet, la synthèse vocale ressemble de moins en moins à une voix synthétique d'ordinateur et ouvre de nouvelles possibilités en alliant ainsi plusieurs technologies comme la reconnaissance vocale et la téléphonie.

La synthèse de parole présente plusieurs avantages, elle est d'une part plus naturelle pour le grand public, elle est plus rapide et efficiente qu'un message écrit court et le champ de vision reste libre pour effectuer une autre tâche de lecture.

Les deux principaux critères exigés par la synthèse de la voix sont l'intelligibilité et l'aspect naturel. Si de nos jours, le premier critère est atteint, le deuxième est encore au stade de développement. En effet, si les synthétiseurs reproduisent une voix tout à fait intelligible, les intonations et l'expressivité ne sont pas encore au point [10].

2.3 LE SYSTEME TEXT-TO-SPEECH (TTS) ET SON ARCHITECTURE

Un Système de Synthèse à Partir du Texte (**TTS** : **T**ext-**T**o-**S**peech) est une machine capable de lire a priori n'importe quel texte à voix haute, que ce texte ait été directement introduit par un opérateur sur un clavier alphanumérique, qu'il ait été scanné et reconnu par un système de reconnaissance optique des caractères (**OCR** : **O**ptical **C**haracter **R**ecognition), ou qu'il ait été produit automatiquement par un système de Dialogue Homme-Machine. Un tel système diffère fondamentalement d'autres machines parlantes en ceci qu'il est destiné à donner lecture de phrases qui n'ont en principe jamais été lues auparavant. Il est en effet possible de produire automatiquement de la parole en concaténant des mots ou des parties de phrases préalablement

enregistrées, mais il est clair dans ce cas que le vocabulaire utilisé doit rester très limité et que les phrases à produire doivent respecter une structure fixe, afin de maintenir dans des limites raisonnables la quantité de mémoire nécessaire à stocker les éléments vocaux de base. C'est le cas par exemple des annonceurs vocaux automatiques dans les gares. On définira donc plutôt la synthèse TTS comme la production automatique de phrases par calcul de leur transcription phonétique [1].

Le texte synthétisé peut provenir de plusieurs sources différentes comme par exemple, le clavier, la reconnaissance vocale, la reconnaissance d'écriture ou Internet. Le synthétiseur n'a alors aucune connaissance préalable du texte qu'il devra synthétiser, mais il est capable de reproduire un flux sonore correspondant au texte reçu comme si un humain le lisait. Le TTS représente une autre technologie qui exclut bien évidemment les systèmes concaténant plusieurs enregistrements sonores afin de créer une phrase [10].

Tout système de synthèse de parole à partir du texte est amené à répondre, de manière plus ou moins précise et développée, selon sa qualité et sa finalité, à deux types de problèmes de natures différentes :

- les traitements linguistiques : cette première étape vise à analyser et à structurer le texte afin de déterminer un mode de prononciation cohérent, puis à transformer le texte analysé en une suite de sons de parole accompagnée d'indications concernant leur agencement ;
- la synthèse proprement dite : cette seconde étape consiste à générer un signal acoustique qui « retranscrit » cette suite de sons tout en possédant les caractéristiques apparentes de la parole naturelle (Fig.2.1) [12].

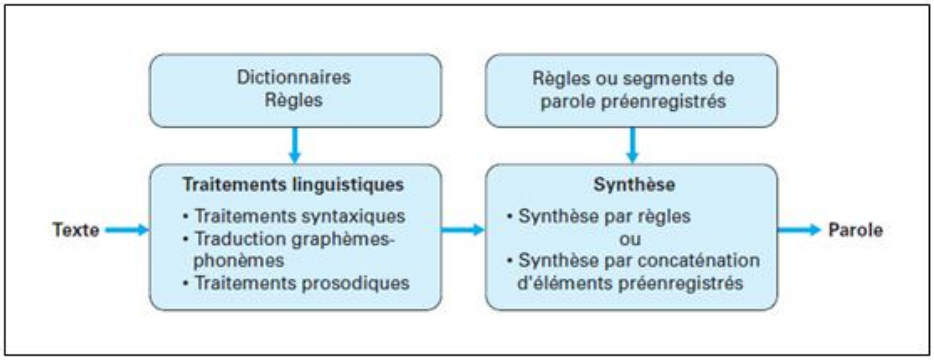


Figure 2.1 : Architecture classique d'un système de synthèse de la parole à partir du texte [16]

2.4 TRAITEMENTS LINGUISTIQUES

Le bloc de traitements linguistiques (Fig.2.1) regroupe les différents modules qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne de phonèmes éventuellement enrichis d'informations linguistiques et prosodiques caractérisant l'élocution. Ces différents modules sont : les prétraitements des éléments non lexicaux, l'analyse lexicale, l'analyse syntaxique, la transcription orthographique - phonémique et le traitement prosodique.

2.4.1 Prétraitement des éléments non lexicaux

Cette étape de prétraitement permet de retranscrire en toutes lettres les chaînes non orthographiques. Il peut s'agir de chiffres, de dates (20/10/95, 19 Jan. 2008) ou plus généralement de sigles composés de caractères orthographiques et numériques (vol AH2106, référence AM66).

En général, on fait appel à des règles de transcription pour le traitement des quantités numériques, des dates ou des sigles standard (SNCF, PTT, etc.). Si le système de synthèse est destiné à un domaine spécifique, le lexique propre à ce domaine sera appliqué.

2.4.1.1 Analyse lexicale

L'analyse lexicale consiste à déterminer dans un lexique les différents mots composant le texte orthographique à synthétiser. Cette analyse est réalisée en trois étapes : un découpage du texte en mots, une analyse morphologique et une analyse lexicale.

2.4.1.2 Analyse syntaxique

L'analyse syntaxique vise à déterminer la structure de la phrase. Elle est conduite par application de règles pouvant être de deux types. Dans certains cas, il peut s'agir d'heuristiques, résultant généralement de l'application de règles grammaticales standards (par exemple, on ne peut pas observer la succession de deux verbes conjugués). En complément ou à la place de ces heuristiques parfois très complexes, on utilise aussi fréquemment des règles probabilistes, exploitant des modèles de langage. Ces modèles sont fondés sur l'observation que toutes les séquences de catégories grammaticales dans une langue donnée ne sont pas équiprobables. La

connaissance de la catégorie syntaxique exacte est également utile pour déterminer la prononciation correcte et notamment pour désambiguïser les homographes hétérophones.

2.4.2 Transcription Orthographique-Phonétique (TOP)

Traditionnellement appelée Conversion Graphème-Phonème, l'étape de **T**ranscription

Orthographique-**P**honétique (**TOP**) constitue le noyau minimal, indispensable à tout système de synthèse de parole, aussi élémentaire soit-il. Cette étape repose sur l'utilisation d'un automate paramétré appliquant un ensemble de règles de réécriture, qui permettent d'associer un phonème (ou un groupe de phonèmes) à un caractère (ou un groupe de caractères) orthographique en prenant en compte le contexte gauche et le contexte droit. Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales. Le nombre de règles nécessaires pour effectuer la TOP dépend de la langue que l'on considère

2.4.3 Traitements prosodiques

La chaîne parlée est d'abord subdivisée en unités suprasegmentales qui facilitent le décodage du message par l'auditeur. La délimitation de ces unités est faite à l'aide de marqueurs dont la réalisation fait appel à des variations paramétriques, de durée, de fréquence et d'intensité.

Les traitements prosodiques sont complexes et s'articulent en différents modules (insertion des pauses, durées phonétiques et fréquence fondamentale). Cependant, l'apparition des techniques de synthèse par sélection dynamique d'unités non uniformes de segments de parole ont permis d'envisager des techniques nouvelles pour la génération de la prosodie. En effet, ces approches génèrent automatiquement la prosodie sans modèle a priori puisqu'elles utilisent une caractérisation symbolique fine des unités d'un corpus de grande taille, ce qui permet de conserver la prosodie originale des segments sélectionnés.

2.4.3.1 Insertion des pauses

Les pauses correspondent aux silences, de durées variables, qui s'insèrent à la fin de chacun des groupes de souffle. L'importance de la coupure syntaxique liée à un marqueur syntaxico-prosodique détermine la durée de la pause à insérer. Ce facteur est particulièrement important pour le naturel de l'élocution. La génération de pauses est absolument nécessaire à la synthèse de

la parole, et le réalisme de leur durée et de leur position est indispensable à la qualité de la synthèse résultante.

2.4.3.2 Durées phonétiques

Une bonne détermination des durées est cruciale pour assurer le naturel de l'élocution. Des durées erronées produisent une parole heurtée, chaotique et parfois difficilement intelligible. Deux approches existent, pour la modélisation de la durée :

- la première basée sur des règles et une bonne analyse statistique, en détermine la durée en prenant en compte différents facteurs, en particulier la durée intrinsèque des sons constituant le segment et le contexte. Parmi les facteurs influençant la durée phonétique, nous pouvons citer : le contexte phonétique (certains phonèmes ont tendance à allonger les phonèmes adjacents, d'autres auront tendance à les raccourcir), la position de la syllabe porteuse dans le groupe prosodique (en Français, par exemple, la syllabe finale de mots est généralement allongée, d'un facteur d'autant plus important que le groupe précède une frontière syntaxique majeure), la nature du groupe prosodique (sa fonction dans la phrase), la longueur du groupe prosodique, etc.
- la deuxième approche est basée sur des techniques d'apprentissage automatique. Celles-ci peuvent reposer sur l'utilisation de réseaux connexionnistes pour prédire la durée des syllabes et ainsi calculer les durées des phonèmes à partir de leur moyenne et de leur écart type.

2.4.3.3 Fréquence fondamentale

Le contrôle de la fréquence fondamentale, dont l'évolution dans le temps définit le contour mélodique, est le point essentiel pour la détermination de l'intonation. L'évolution de la fréquence fondamentale pour chaque phonème est spécifiée à l'aide d'un modèle prédictif complexe, prenant en compte deux types de phénomène globaux dits de macromélodie, et locaux dits de micromélodie [14].

2.5 TECHNIQUES D'ANALYSE DU SIGNAL VOCAL

Une fois que le son a été émis par le locuteur, il est capté par un microphone. Le signal vocal est ensuite numérisé à l'aide d'un Convertisseur Analogique/Numérique. Comme la voix humaine est constituée d'une multitude de sons, souvent répétitifs, le signal peut être compressé pour

réduire le temps de traitement et l'encombrement en mémoire. Ainsi comme prétraitement (Fig.2.2), nous échantillonnons et préaccentuons le signal vocal. Pour les techniques de reconnaissance, d'analyse ou de synthèse de la parole, la fréquence d'échantillonnage peut varier de 08 jusqu'à 16 kHz. Le filtre de préaccentuation de transmittance H(z) est :

$$H(z) = 1 - a.z^{-1} \qquad \text{avec : } a=0.95 \qquad (2.1)$$

Qui est souvent non récursif de premier ordre, permet d'égaliser les aigus toujours plus faibles que les graves. Aussi et vu qu'il est non stationnaire, nous réalisons un fenêtrage avec une fenêtre glissante ; chaque trame couvrant une durée de 20 à 30 ms sur laquelle le signal est supposé quasi-stationnaire. Le pas d'analyse entre deux trames successives est de l'ordre de quelques dizaines de ms.

Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Pour compenser ces effets de bord, nous multiplions en général préalablement chaque tranche d'analyse par une fenêtre de pondération de type fenêtre de Hamming notée W(n) [17].

$$W(n) = \begin{cases} 0.45+0.46.\ \mathrm{Cos}\ (\pi n/(n-1)) & n \in [0,\ldots, n-1] \\ 0 & \text{ailleurs} \end{cases} \qquad (2.2)$$

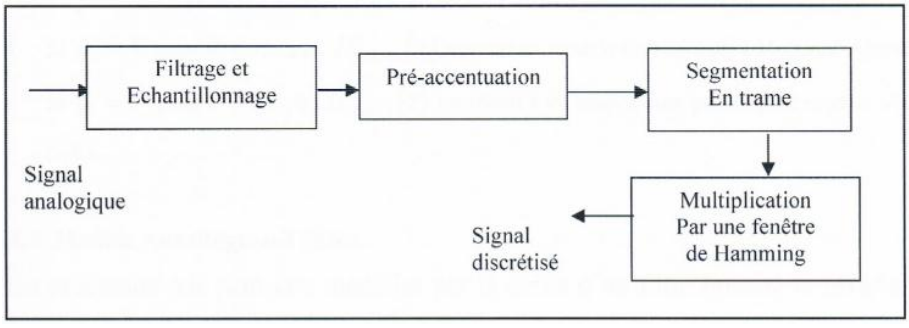


Figure 2.2 : Prétraitement du signal vocal

Le signal vocal peut être analysé soit, en tenant compte des mécanismes de production en utilisant les méthodes paramétriques, soit en utilisant les méthodes non paramétriques.

Dans la plupart des méthodes d'analyse vocale, nous supposons que le signal de parole est localement stationnaire car les propriétés de ce signal varient très doucement en fonction du temps, d'où le recours aux méthodes d'analyse à court terme. Ainsi de courts segments de la parole sont analysés, on les appelle les trames d'analyse temporelle.

Les mesures comme l'énergie, le **T**aux de **P**assage par **Z**éro **(TPZ)** et la fonction d'autocorrélation font partie des méthodes temporelles.

Les coefficients les plus utilisés en RAP sont les cepstres. Ils peuvent être extraits de deux façons : soit par l'analyse paramétrique, à partir du Codage Prédictif Linéaire ou **L**inear **P**rédictive **C**oding **(LPC)**, soit par l'analyse spectrale.

2.5.1 Méthodes non paramétriques

Le signal de parole peut être analysé dans le domaine temporel ou dans le domaine spectral par des méthodes non paramétriques, sans faire l'hypothèse d'un modèle pour rendre compte du signal observé. Les méthodes spectrales sont fondées sur la décomposition fréquentielle du signal sans connaissance a priori de sa structure fine. Une analyse spectrale du signal permet de mettre en évidence certaines caractéristiques de la production de la parole qui peuvent contribuer à l'identification phonétique. L'articulation des phonèmes a une influence directe sur la forme du conduit vocal et des cavités, et donc sur les résonances qui apparaissent dans l'enveloppe du spectre.

L'analyse fréquentielle de la parole se ramène aux opérations de la **T**ransformée de **F**ourier **(TF)** et n'a d'intérêt que si elle s'applique à une période du signal vocal, donc sur une période assez courte.

En RAP, il est important de connaître l'évolution de ce spectre dans le temps. Actuellement, les spectres sont obtenus numériquement par la **T**ransformée de **F**ourier **D**iscrète **(TFD),** en particulier grâce à l'algorithme de la **T**ransformée de **F**ourier **R**apide **(TFR)** ou **F**ast **F**ourier **T**ransform **(FFT)**. Cependant, le nombre de paramètres spectraux calculés sur une trame

par FFT reste trop élevé pour un traitement automatique ultérieur. Pour une analyse très fine de la parole, la fenêtre de Hamming est déplacée à chaque fois de 128 points environ 10 ms (Fig.2.3).

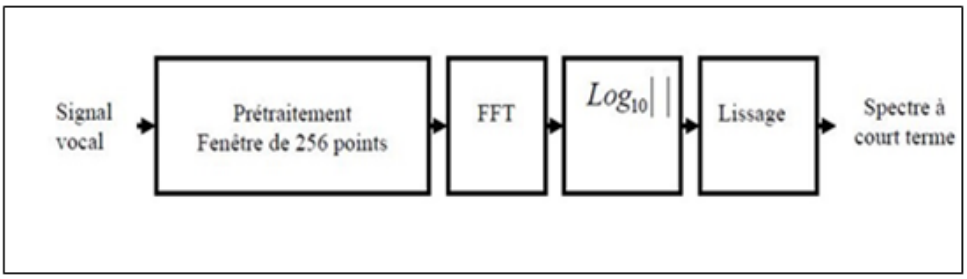


Figure 2.3 : Analyse numérique du signal parole par FFT

2.5.2 Méthodes paramétriques

Les méthodes paramétriques appelées aussi méthodes d'identification sont fondées sur une connaissance des mécanismes de production de la parole (Exemple : le conduit vocal). Les plus utilisées sont celles basées sur l'analyse prédictive linéaire et l'analyse cepstrale. L'hypothèse de base est que le conduit buccal est constitué d'un tube cylindrique de section variable. L'ajustement des paramètres de ce modèle permet de déterminer à tout instant sa fonction de transfert. Cette dernière fournit une approximation de l'enveloppe du spectre du signal à l'instant d'analyse. Ces méthodes consistent à ajuster un modèle aux données observées. Les paramètres du modèle, en nombre faible, caractérisent le signal, nous pouvons ainsi injecter des connaissances a priori sur le processus physique qui a engendré ce signal .

Les avantages de cette approche sont la souplesse de l'analyse, l'introduction naturelle de l'information et les choix variés des espaces de représentations paramétriques. Dans le cas de la modélisation du signal parole, nous n'avons accès qu'à une seule sortie du système alors que l'entrée n'est pas mesurée. Il en résulte un problème d'estimation non linéaire car nous ne disposons pas d'observation de l'onde glottique d'excitation. En conséquence, nous en sommes limités à faire quelques hypothèses relativement neutres sur l'entrée ; par exemple, bruit blanc à moyenne nulle et reporter tout l'effort de modélisation sur le système.

2.5.2.1 Codage Prédictif Linéaire (LPC)

Cette méthode connue de la production sous le sigle LPC (Linear Predictive Coding) se fonde sur les connaissances de la production de la parole et suppose que le modèle de production de la parole est linéaire selon le schéma (Fig.2.4).

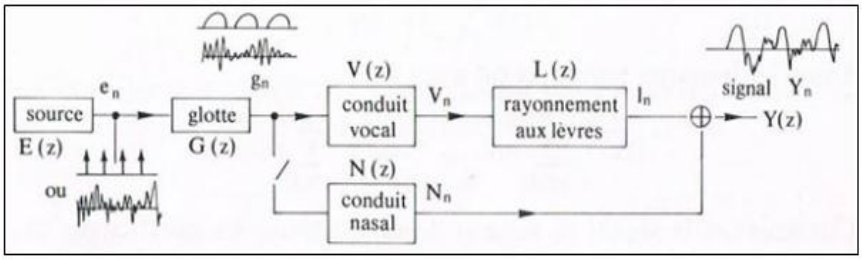


Figure 2.4 : Modèle général de production de la parole [11]

Globalement, ce modèle peut se décomposer en deux parties : la source active, le conduit passif de manière plus détaillée, il peut se décrire de la maniéré suivante : l'onde est modélisée comme la sortie d'un filtre passe bas à deux pôles de fréquence de coupure d'environ 100 Hz (glotte), l'entrée $\mathbf{e_n}$ de ce filtre est un train d'impulsions de période T_0 pour les sons voisés ou un bruit blanc pour les sons non voisés (source).

Le modèle du conduit vocal est un filtre tout pôle (AR : auto - Régressif) d'ordre 2M décomposable en une cascade de résonateurs à 2 pôles en série (tuyaux résonants). Le modèle du conduit nasal est un filtre pôle zéro ARMA (Auto Régressif à Moyenne Ajustée) et le rayonnement aux lèvres peut se modéliser par un filtre tout zéro (MA : Moyenne Ajustée).

L'ensemble des conduits se comporte donc comme un système linéaire ARMA.

Modèle glottale :

$$G(z) = \frac{1}{(1-e^{-2\pi f_g T z^{-1}})^2} \qquad f_g = 100Mz \tag{2.3}$$

Modèle du conduit vocal :

$$V(z) = \prod_{i=1}^{M} \left(\frac{1}{1-2e^{-2\pi B_i T}.\cos(2\pi F_i T)z^{-1}+e^{-4\pi B_i T} z^{-2}}\right) \tag{2.4}$$

F_i : Fréquence du formant n° i, B_i sa bande passante

Modèle du conduit nasal :

$$N(z) = \frac{1-2e^{-2\pi B'_N T}.\cos(2\pi F'_N T)z^{-1}+e^{-4\pi B_N T}z^{-2}}{1-2e^{-2\pi B_N T}.\cos(2\pi F_N T)z^{-1}+e^{-4\pi B_N T}z^{-2}} \quad (2.5)$$

Avec F_N et F'_N formant nasal ou anti formant nasal et respectivement, B_N et B'_N leurs bandes passantes.

Si on suppose qu'une partie α du signal g_n est dérivée vers le conduit nasal le modèle du conduit peut se mettre sous la forme :

$$H(z) = G(z).[1-\alpha).\mathrm{V(z)L(z)} + \alpha\mathrm{N(z)}] \quad (2.6)$$

Avec $0 \leq \alpha \leq 1$ pour un son nasal α=1 ; pour un son non nasal α=0.

$H(z)$ Est en tout généralité un modèle ARMA d'ordre p : $H(z) = \frac{B(z)}{A(z)}$ (2.7)

Dans le domaine temporel on aura :

$$y_n + \sum_{i=1}^{P} a_i y_{n-p} = e_n + \sum_{i=1}^{q} b_i e_{n-p} \quad (2.8)$$

Caractériser le signal y_n revient donc a estimé les coefficients $\{a_i ; b_i\}$.

Pour une source connue e_n (séquence d'impulsions ou bruit blanc). Souvent pour simplifier la résolution de ce problème, on suppose que $b_i = 0$, $i \geq 1$ ce qui rend le modèle AR [11].

2.5.2.2 Analyse cepstrale

Le défaut majeur des méthodes d'analyse, comme la FFT, pour le calcul du spectre réside dans l'intermodulation source/conduit vocal qui rend difficile la mesure du fondamental F_0 et des formants.

Le lissage cepstral est une méthode qui vise à séparer la contribution du conduit vocal de l'excitation glottique. Cette séparation est réalisée par un homomorphisme qui transforme la convolution des signaux dans le domaine temporel en une addition dans le domaine cepstral. En

outre, cette méthode permet de fournir un vecteur spectral des MFCC pour des fins de la RAP et de lisser le spectre de parole pour trouver les formants.

Pour cela, nous faisons l'hypothèse que le signal vocal y_n est produit par le signal excitateur $u_n$traversant un système linéaire de réponse impulsionnelle b_n.

Le but du cepstre est de séparer ces deux contributions par déconvolution. Il est fait l'hypothèse qu'un signal excitateur est soit une séquence d'impulsions (périodiques, de période T_0, pour les sons voisés), soit un bruit blanc pour les sons non voisés, conformément au modèle de production de la parole. Une transformation en Z permet de transformer la convolution en produit.

$$Y(z) = B(z).U(z) \tag{2.9}$$

Le logarithme du module uniquement (car nous ne s'intéressons pas à l'information de phase) transforme le produit en somme. Nous obtenons alors :

$$\log|Y(z)| = \log|U(z)| + \log|B(z)| \tag{2.10}$$

Par transformation inverse, nous obtenons le cepstre. Dans la pratique, la transformation en Z est remplacée par une TFR. L'expression du cepstre est donc :

$$C(n) = FT^{-1}\{\log(FT\{y(n)\})\} \tag{2.11}$$

Le cepstre qui ne fait appel à aucune information a priori sur le signal acoustique, est basé sur une connaissance du mécanisme de production de la parole. L'espace de représentation du cepstre ou espace quéferentiel est homogène par rapport au temps. Les premiers coefficients cepstraux contiennent l'information relative au conduit vocal. Cette contribution devient négligeable à partir d'un échantillon $\boldsymbol{n}_0$ qui correspond à la fréquence fondamentale $\boldsymbol{F}_0$ Les pics périodiques visibles au-delà de $\boldsymbol{n}_0$, reflètent les impulsions de la source.

Le spectre du cepstre pour les indices inférieurs à $\boldsymbol{n}_0$ permet d'obtenir un spectre lissé, en éliminant les lobes secondaires dû à la contribution de la source. Ces deux contributions peuvent être séparées par une simple fenêtre temporelle notée $\boldsymbol{F}$ (liftrage) telle que le filtre adouci ou le filtre rectangulaire.

La présence d'un pic important dans le cepstre renseigne d'une part sur le caractère voisé ou non du son et d'autre part constitue une bonne indication sur la fréquence fondamentale.
L'enveloppe spectrale du conduit vocal (structure formantique) est obtenue par une transformation supplémentaire (Fig.2.5).

Le spectre lissé débarrassé théoriquement de la contribution de la source ne contient que des informations sur le conduit vocal et en particulier sur ses extrema (Formants) [9].

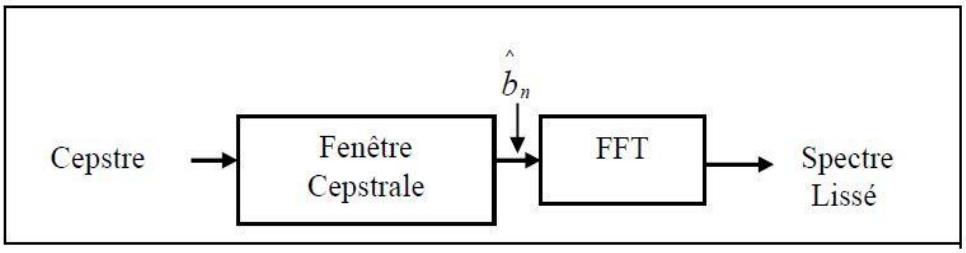


Figure 2.5 : Obtention de la structure formantique à partir du cepstre

2.6 LES METHODES DE SYNTHESE DE LA PAROLE

On peut établir une analogie fonctionnelle entre le rôle joué par le module de traitement du signal et celui du système phonatoire humain, qui contrôle en permanence l'activité de tous ses muscles (y compris de ceux qui règlent la fréquence de vibrations des cordes vocales) de façon à produire le signal voulu. Pour y arriver, il est clair que ce module doit, dans une certaine mesure, prendre en compte les contraintes articulatoires. En effet, on sait depuis longtemps que les transitions phonétiques contribuent plus à l'intelligibilité du signal vocal que les zones stables des phonèmes. On peut alors envisager de le faire de façons :

- explicite, sous la forme d'une série de règles décrivant formellement l'influence des phones les uns sur les autres ;
- implicite, en enregistrant des exemples de transitions entre phones dans une base de données de segments de parole, et en les utilisant tels quels comme unités de parole (en lieu et place des phones).

Cette alternative a donné lieu à deux grandes familles de synthétiseurs: la synthèse par règles et la synthèse par concaténation

2.6.1 Synthèse Par Règles (SPR)

La **S**ynthèse **P**ar **R**ègles (**SPR**) a connu un essor considérable dans les années 60-70. Elle n'est plus guère utilisée aujourd'hui que lorsque les contraintes de mémoire et de temps de calcul sont très importants. La qualité des voix disponibles n'est, en effet, pas aussi bonne qu'en synthèse pas concaténation, pour un coût de développement supérieur .

Les synthétiseurs par règles ont principalement la faveur des phonéticiens et des phonologistes. Ils permettent une approche cognitive, générative du mécanisme de la phonation. Ils sont basés sur l'idée que, si un phonéticien expérimenté est capable de «lire» un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel pour une suite de phonèmes donnés. Une fois le spectrogramme obtenu, il ne reste plus alors qu'à générer l'audiogramme correspondant (Fig.2.6).

Dans un premier temps, on fait lire par un locuteur professionnel un grand nombre de mots, généralement de type **C**onsonne-**V**oyelle-**C**onsonne (**CVC**) et on les enregistre sous forme numérique. Les mots sont choisis de façon à constituer un corpus. On modélise alors ces données numériques à l'aide d'un modèle paramétrique de parole, qui a pour rôle de séparer les contributions respectives de la source glottique et du conduit vocal et de présenter cette dernière sous forme compacte, plus propice à l'établissement des règles.

On commence par inspecter globalement l'ensemble des données, de façon à établir la forme générale des règles à produire. On précise alors les valeurs numériques des paramètres intervenant dans ces règles (les fréquences des formants, ou les durées des transitions, par exemples) par un examen minutieux du corpus. Il est à remarquer que cette étape d'estimation est menée sur une seule voix : un moyennage inter-locuteur aurait peu de signification dans ce contexte. De même, les règles provenant de synthétiseurs déjà existants ne peuvent resservir que dans la mesure où elles modélisent des caractéristiques articulatoires générales plutôt que des particularités du locuteur ayant enregistré le corpus (sauf bien entendu si l'on cherche à produire des règles caractérisant précisément le passage d'une voix à une autre). La mise au point du synthétiseur s'achève par un long processus d'essais-erreurs, afin d'optimiser la qualité de la synthèse.

Lorsqu'un nombre suffisant de règles ont été établies, la synthèse proprement dite peut commencer. Les entrées phonétiques du synthétiseur déclenchent l'application de règles, qui produisent elles-mêmes un flux de paramètres liés au modèle de parole, utilisé. Cette séquence temporelle de paramètres est alors transformée en parole par un synthétiseur, qui implémente les équations du modèle.

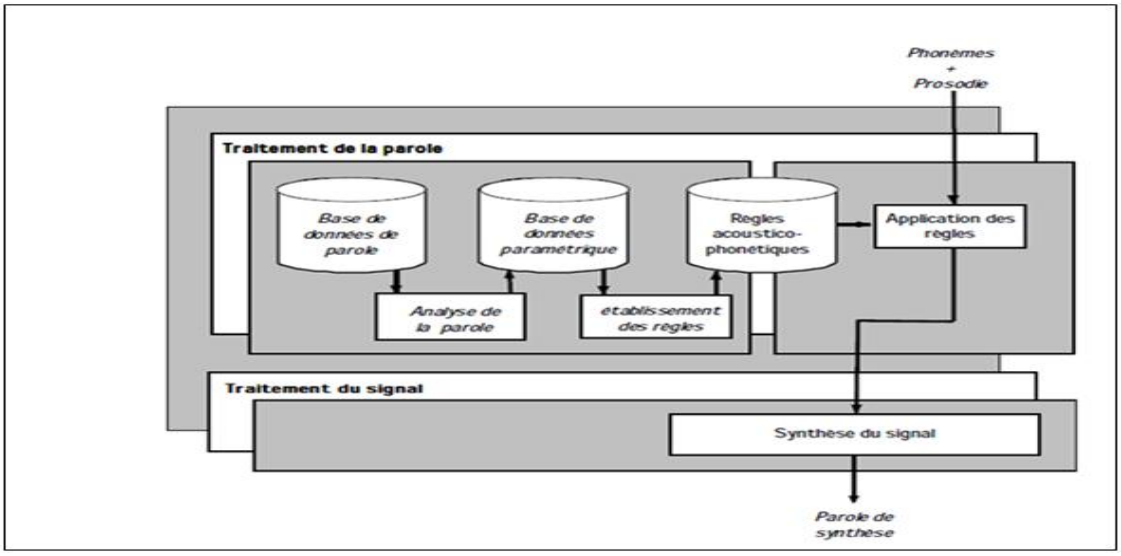


Figure 2.6 : Schéma de conception et fonctionnement typique d'un système de synthèse par règles

2.6.2 Synthèse par concaténation d'unités acoustiques

Cette seconde approche, qui ne fait pas explicitement référence à un modèle de production de la parole, consiste à synthétiser le signal par concaténation d'unités acoustiques, c'est-à-dire de segments de parole préenregistrés. Cette technique, qui repose sur l'utilisation de segments de signaux extraits de la parole naturelle, est la seule qui permet à ce jour de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain.

2.6.2.1 Choix des unités acoustiques

Le choix des unités joue un rôle primordial pour ce type de synthèse. La dialectique est simple : les unités acoustiques courtes sont économiques, mais elles n'en permettent pas d'obtenir une synthèse de bonne qualité. Les unités longues sont plus coûteuses mais permettent généralement d'obtenir une meilleure qualité de synthèse. Ce que l'on entend par « court » ou « long » dépend évidemment de l'application que l'on considère :

- S'agit-il réellement de synthèse de texte libre, ou le vocabulaire et la syntaxe sont-ils contraints ?
- certaines phrases ou construction de phrases sont-elles beaucoup plus fréquentes que d'autres ? de la qualité visée
- cherche-t-on à obtenir une synthèse simplement intelligible ou l'agrément de la voix joue-t-elle un rôle pour le service proposé ?

Et des moyens susceptibles d'être mis en œuvre pour atteindre l'objectif retenu (les impératifs ne sont pas les mêmes pour un système distribué sur micro-ordinateurs et un serveur de synthèse centralisé).

Ce que l'on peut dire, c'est que certaines unités sont définitivement trop courtes : les phonèmes par exemple sont inappropriés car ils ne permettent pas de capturer la dynamique du processus de production de parole : comme nous l'avons souligné précédemment, et contrairement à ce que pourrait laisser croire la théorie linguistique, la parole est essentiellement un processus continu et l'enchaînement des sons entre eux (qui n'est rien d'autre que la manifestation acoustique de l'articulation) est au moins aussi important, du point de vue de la perception, que les sons eux-mêmes.

L'unité minimale permettant d'obtenir une synthèse de qualité acceptable est le diphone, qui est défini comme la portion du signal de parole comprise entre les zones stables de deux phonèmes consécutifs. Le diphone, à l'inverse du phonème, capture la transition entre les différentes cibles articulatoires associées aux phonèmes, transition qui est cruciale pour la perception des différents sons. En théorie, le nombre de diphones est égal au carré du nombre de phonèmes c'est-à-dire environ à $36^2 = 1296$ diphones (en négligeant le fait que certaines transitions entre phonèmes sont impossibles en Français) (Fig.2.7) et (Fig.2.8).

Exemple de décomposition en diphones du mot : [samja]

[#s] [sa] [am] [mj] [ja] [a#], soit 6 diphones. Le symbole # est une indication de pause.

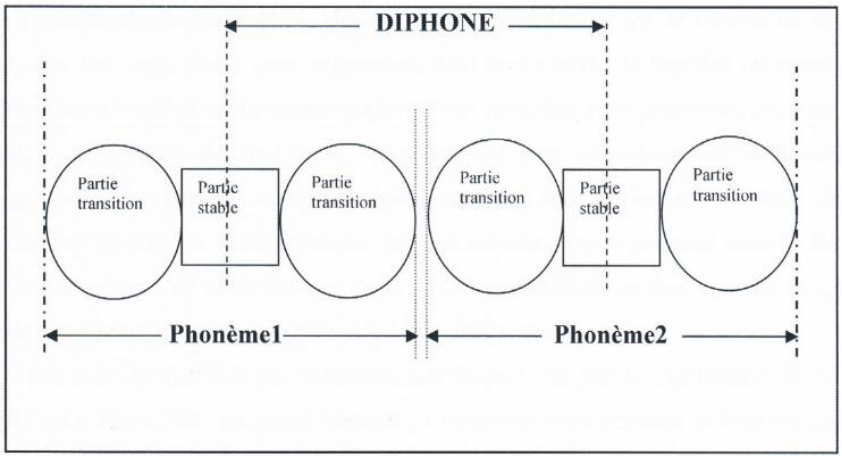


Figure 2.7 : Extraction de diphone dans une séquence sonore [17]

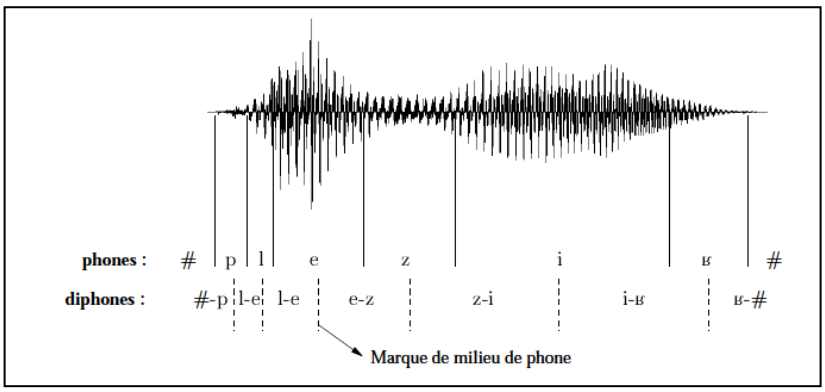


Figure 2.8 : Illustration d'une segmentation en diphones, dans le mot plaisir [18]

En pratique, pour la synthèse par diphones, le nombre d'unités utilisées est légèrement plus important (de l'ordre de 1 500 à 2 000) pour tenir compte des différentes variantes contextuelles des phonèmes composant le diphone (dans certains systèmes, plusieurs représentants de chaque diphone sont disponibles ; l'algorithme de concaténation choisit à chaque instant le « meilleur » représentant de façon à minimiser une fonction d'objectif). Le volume de stockage nécessaire est de l'ordre de 5 à 10 Mo (2 à 6 min de parole numérisée avec une fréquence d'échantillonnage de 16 kHz). Cette quantité de données (considérée il y a moins de dix ans comme une borne supérieure) semble aujourd'hui bien raisonnable au regard des possibilités de stockage offertes par les systèmes informatiques actuels [19].

Si l'utilisation de diphones améliore notablement la qualité de synthèse par rapport à la synthèse par unités phonétiques, des tests d'intelligibilité montrent que des confusions dans la perception

de certains sons ou groupes de sons persistent encore. Pour le Français, par exemple, il apparaît que certains groupes consonantiques demeurent incorrectement perçus par des auditeurs naïfs ; ces défauts sont dûs (tout au moins en partie), à la grande variabilité de certaines consonnes comme les liquides ([l], [r]) et les semi-voyelles.

Pour accroître la qualité, il est actuellement envisagé de considérer des unités plus longues que le diphone, aptes à prendre en compte des phénomènes de coarticulation à plus long terme (disons, pour simplifier, à l'échelle de la syllabe). Parmi celles-ci, les unités de la forme **V**oyelle-**C**onsonne-**V**oyelle **[V-C-V**] ou de façon plus générale du type [V-C-...-C-V] (deux voyelles séparées par un nombre quelconque de consonnes) apparaissent très prometteuses. Ces unités permettent de n'avoir à effectuer des concaténations que dans les zones réputées les plus stables du signal de parole, à savoir le centre des noyaux vocaliques. Elles capturent d'autre part, la coarticulation de voyelle à voyelle à travers la (ou les) consonne(s), la coarticulation qui joue vraisemblablement un rôle important à la fois pour l'intelligibilité et l'agrément de la voix de synthèse. Le problème est que le nombre d'unités ainsi obtenues est beaucoup plus important (de l'ordre de 10 000-15 000, en ne retenant que les unités apparaissant effectivement)... Un grand nombre de ces unités sont peu fréquentes et peuvent être éliminées pour satisfaire aux contraintes de taille [19].

2.6.2.2 Mise en œuvre

La synthèse proprement dite comprend trois étapes distinctes :

- **Sélection des unités acoustiques** : cette première étape consiste à choisir dans le répertoire d'unités acoustiques les unités qui seront effectivement utilisées pour synthétiser la succession de sons, désirée. Cette étape est à peu près évidente quand les unités sont régulières (à l'instar des phonèmes et des diphones) : seule la présence de plusieurs versions pour le même segment est à prendre en considération. Cette étape est en revanche plus délicate pour les systèmes d'unités de taille variable. Pour une suite donnée de sons, plusieurs choix d'unités sont en général possibles. Il faut alors arbitrer entre les différentes décompositions avec des critères composites.

- **Ajustement des paramètres prosodiques** : les unités acoustiques pré-enregistrées possèdent une prosodie intrinsèque (les sons qui la composent ont une certaine durée et la fréquence fondamentale décrit un certain contour). Bien sûr, cette prosodie intrinsèque n'a que très peu de

chances d'être conforme à la prosodie de synthèse, spécifiée par le module prosodique. Il va donc falloir utiliser une technique de traitement de signal pour ajuster aux valeurs cibles définies les paramètres prosodiques des unités de synthèse.

-Concaténation des unités : les unités acoustiques, quelles que soient les précautions prises lors de la sélection et de l'enregistrement des unités, ne possèdent pas exactement à leur frontière les mêmes caractéristiques acoustiques (en particulier énergétiques).

En l'absence de traitement, ces discontinuités vont engendrer des artefacts perceptibles et gênants. Il est donc important de lisser ces discontinuités en interpolant les trajectoires des différents paramètres caractéristiques de l'unité [19].

2.6.2.3 Synthèse fondée sur l'algorithme PSOLA

L'algorithme **PSOLA** (**P**itch **S**ynchronous **O**ver **L**ap and **A**dd) consiste à concaténer, à l'aide d'un lissage, des unités de parole pré-stockées en modifiant le pitch et la durée des segments. Cette technique est associée à la méthode de synthèse par concaténation. L'algorithme PSOLA permet la synthèse d'une parole de haute qualité.

Les différentes versions de PSOLA existantes fonctionnent selon le même principe. Le segment de signal de parole naturelle est subdivisé en un ensemble de signaux dits à **C**ourt-**T**erme (**CT**) en utilisant un fenêtrage synchronisé avec le pitch (trame voisée, Fig.2.9) et à intervalles fixes (trame non voisée). Le pitch est augmenté ou diminué en agissant sur la distance entre les signaux à CT durant le processus de synthèse. La durée est gérée par suppression ou duplication des signaux à CT.

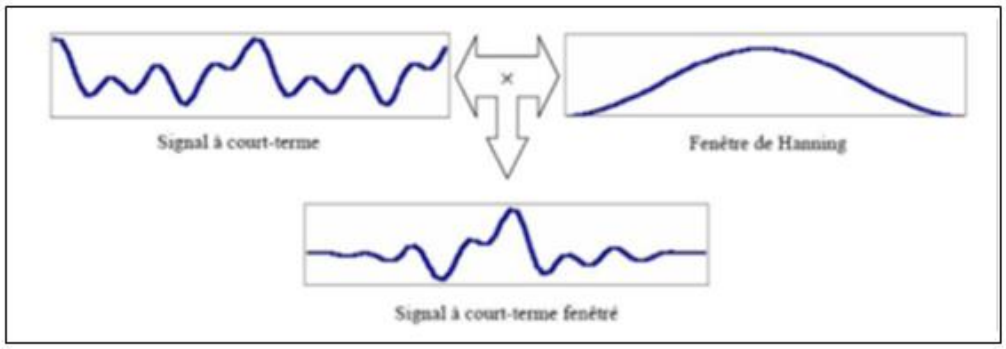


Figure 2.9 : Exemple de signal à Court-Terme

Les signaux à Court-Terme (CT) sont recombinés pour produire le signal de synthèse à l'aide d'une technique d'addition/recouvrement (OverLapp-Add : OLA) (Fig.2.10) et (Fig.2.11) [14].

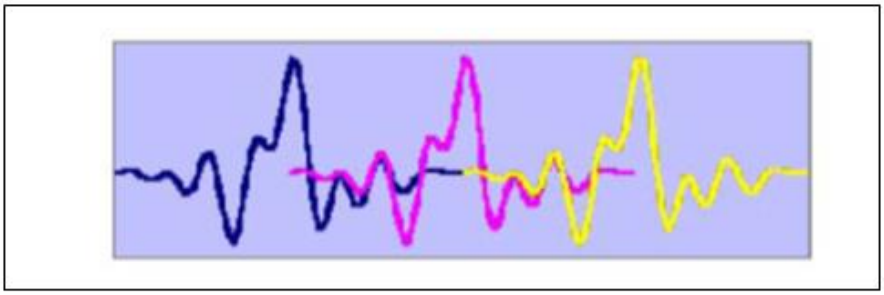


Figure 2.10 : Etape d'addition et recouvrement OLA

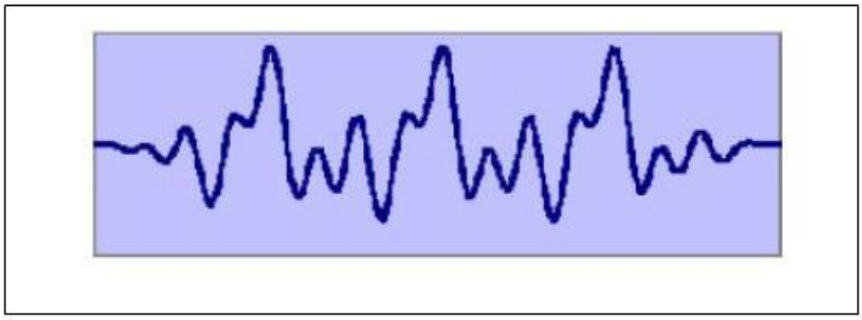


Figure 2.11 : Signal synthétisé avec PSOLA

2.6.2.4 Synthèse par polysons

La méthode par diphones a un inconvénient qui se résume en la difficulté de dégager une zone stable pour segmenter ; dans un contexte de liquides ou semi-voyelles très sensibles aux effets de la coarticulation. Pour remédier à ce problème ; on avait introduit la notion généralisée de polysons. Ces dernier sont constitués des diphones réalisés en segment les parties stables des phonèmes ; en incorporant à l'intérieur les phonèmes instables (liquides et semi-voyelles) ; lorsque ces derniers se trouvent dans le contexte. Cette méthode permet de tenir compte de la structure formantique des liquides et semi-voyelles. Elle vient pour apporter des correctifs à la synthèse par diphones [17].

2.7 LE FONCTIONNEMENT D'UN SYNTHETISEUR VOCAL

Pour créer un synthétiseur vocal, il est indispensable de passer par différentes étapes de traitement. Ce qui suit est le descriptif du fonctionnement du synthétiseur Kali de la société

Electrel (Fig.2.12). Certains synthétiseurs peuvent avoir un ordre de traitement légèrement différent mais les blocs de traitements sont sensiblement identiques.

2.7.1 Le prétraitement

Cette phase consiste dans la transformation d'un texte dans une suite de phrases, organisées en mots. Ce prétraitement a pour objectif de retranscrire en toutes lettres les chaînes non orthographique représentant la reconnaissance des unités de mesure du Système International (SI), les symboles non alphanumériques (ex. : antislash), les chiffres, les lettres et les motifs (ex. : extensions de fichiers). Il ne faut pas oublier les abréviations, les sigles comme « A+» que l'on trouve couramment dans les e-mails.. Là encore, il faut apprendre au système à reconnaître les abréviations les plus courantes.

2.7.2 Analyse syntaxique

L'analyse syntaxique découpe le texte en groupes de mots ou tronçons et établit leurs relations de dépendance. Elle permet une meilleure interprétation des mots pour la suite des opérations en découpant chaque mot en lexèmes et en déterminant leurs appartenances grammaticales.

2.7.3 Calcul prosodique

Le traitement prosodique sert à modéliser l'évolution temporelle de la fréquence fondamentale (vibrations des cordes vocales, prédire la durée des sons élémentaires et la durée des pauses). Si la prosodie d'un locuteur réel est recopiée sur la voix de synthèse, le résultat obtenu est sensiblement meilleur. En effet, l'impression de naturel et son intelligibilité s'améliorent. Le traitement prosodique est donc une composante tout à fait essentielle d'un système de synthèse de parole.

2.7.4 Transcription Graphème-Phonème

Le but de la transcription Graphème-Phonème est de passer du texte orthographique (plus ou moins traité par le module précédent) à une suite de symboles phonétiques. Plusieurs niveaux de connaissances sont pris en compte : phonétiques, phonologiques, lexicales, syntaxiques et même sémantiques. La prononciation du Français comporte ainsi plus de 1000 règles élémentaires, et plusieurs milliers de règles portant sur les noms propres et mots d'emprunt les plus courants.

Chaque langue est différente et emploie un certain nombre de règles spécifiques. Par exemple, l'espagnol ne nécessite que 50 règles pour obtenir une bonne synthèse.

2.7.5 Traitement acoustique

Le dernier traitement effectué par le synthétiseur est la conversion du texte phonétique en signal de parole. La voix synthétique s'obtient par l'extraction des diphones à partir de la voix d'un locuteur, un dictionnaire contenant entre 1000 à 2000 segments de signal seront ensuite concaténés par le synthétiseur pour former le signal de parole. Les paramètres acoustiques les plus utilisés pour représenter ces unités sont le codage par prédiction linéaire (LPC), la méthode TD-PSOLA ou encore celle de MBROLA.

2.7.6 Extraction des diphones

Une phrase articulée se compose d'une succession de «portions de signal sonores», les «diphones». Il s'agit de sortes d'unités phonétiques qui correspondent au son émis du milieu d'un phonème jusqu'au milieu du phonème suivant. Le Français comporte environ 1200 diphones.

Les diphones sont extraits, en laboratoire, lors d'enregistrements de parole d'un locuteur. Ces locuteurs nous permettent ainsi la création de plusieurs voix de synthèse. Il est évident que la création d'une nouvelle voix nécessite de longs enregistrements [15].

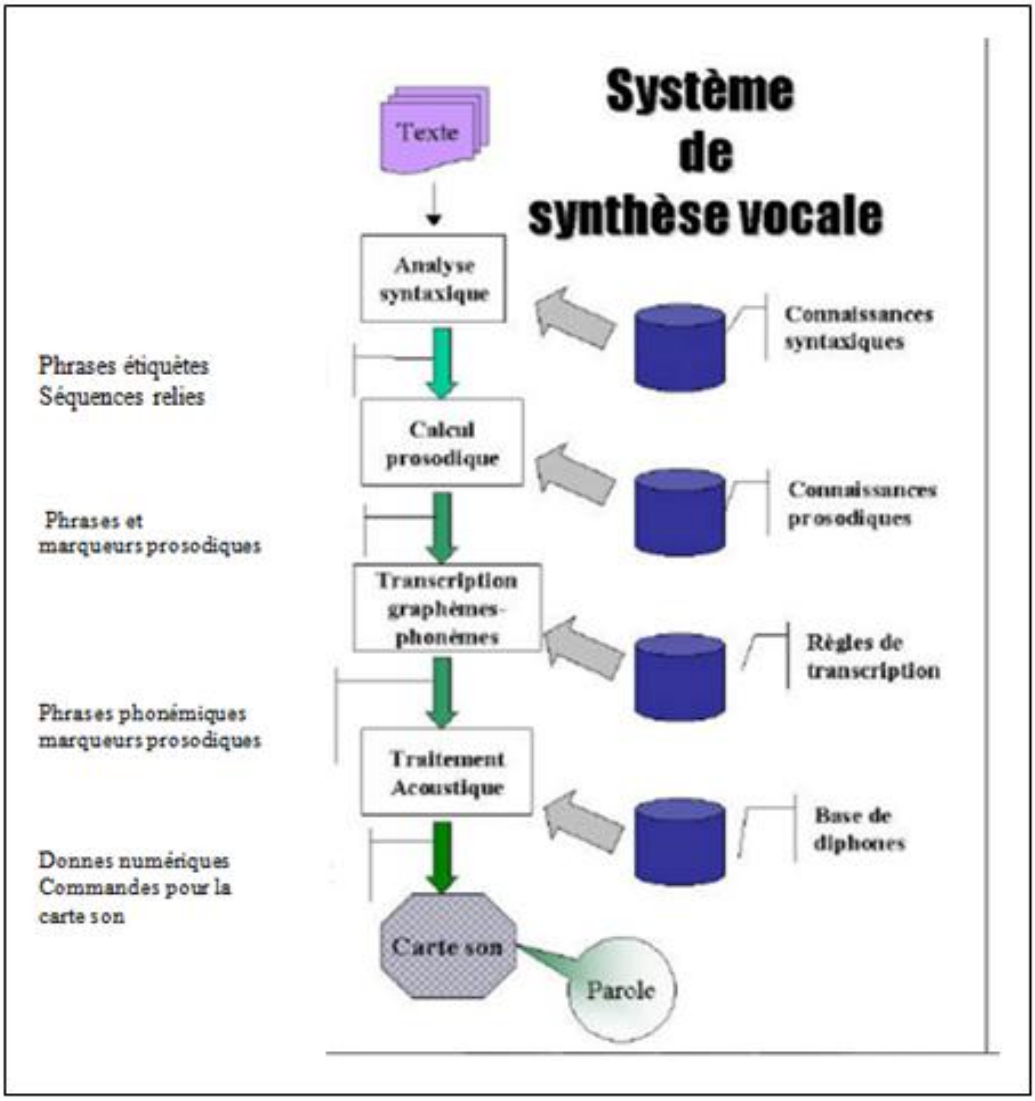


Figure 2.12 : Synthèse de la voix selon le synthétiseur Kali [15]

2.8 LES APPLICATIONS DE LA SYNTHESE DE PAROLE

D'après une étude du cabinet d'avocat américain Frost & Sullivan, le marché de la voix bénéficie d'une augmentation de 15% par an. Ce qui laisse à penser que d'autres sociétés vont se lancer dans les portails vocaux pour entreprise. Voici quelques exemples des différents secteurs où la synthèse pourra être utilisée :

- **Services de télécommunications**

La libéralisation du marché des télécommunications en Europe a récemment rendu les opérateurs de télécommunications plus sensibles au confort de leurs clients. En particulier, on cherche désormais à fournir un maximum de services, à moindre coût. Les synthétiseurs permettent précisément de rendre tout type d'information écrite disponible via le téléphone. On peut ainsi créer des serveurs vocaux diffusant les horaires des cinémas, des informations routières, l'état d'un compte en banque, ou encore des explications automatisées concernant la dernière facture de téléphone. Les requêtes se font soit par la voix (en combinant le synthétiseur avec un reconnaisseur), soit par le clavier du téléphone. AT&T a récemment testé certains services de ce type auprès de ses clients, et constaté un réel engouement, à condition que l'intelligibilité des voix de synthèse soit suffisante; il s'est avéré que le naturel n'est pas un facteur déterminant pour la plupart de ces services.

- **Apprentissage (ou perfectionnement) de langues étrangères**

Une synthèse de très bonne qualité couplée à un logiciel d'apprentissage constitue un outil très utile à l'apprentissage d'une nouvelle langue, en complément d'un cours avec un professeur. Si ce type de produit n'a pas encore percé sur le marché, c'est à cause de la mauvaise qualité des voix disponibles jusqu'à il y a peu. On voit par contre se multiplier les petits dictionnaires électroniques de poche, qui devraient rapidement être dotés de voix de synthèse. Il en va de même des traducteurs électroniques mot-à-mot qui sont apparus récemment. On pourra par exemple bientôt lire un ouvrage dans une langue étrangère et utiliser un stylo à lecture optique (intégrant un mini-scanner) pour obtenir instantanément la traduction d'un mot inconnu et sa prononciation.

- **Aide aux personnes handicapées**

Les handicaps liés à la parole sont soit d'origine mentale, soit d'origine motrice ou sensorielle. La machine peut être d'un grand secours dans le second cas. Avec l'aide d'un clavier spécialement adapté et/ou d'un logiciel d'assemblage rapide de phrases, un handicapé peut s'exprimer par la voix de son synthétiseur. Le célèbre astrophysicien Stephen Hawking donne tous ses cours à l'université de Cambridge de cette façon. La synthèse offre également des services aux personnes mal-voyantes, en leur donnant accès à l'information écrite "en noir" (Dans le

vocabulaire des aveugles, l'impression "en noir" s'oppose à l'impression en Braille), à condition de coupler le synthétiseur à un logiciel de reconnaissance des caractères.

- **Livre et jouets parlants**

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

- **Monitoring vocal**

Dans certains cas, l'information orale est plus efficace qu'un message écrit. L'utilisation d'une voix de synthèse dans un centre de contrôle de site industriel, par exemple, permet d'attirer l'attention du personnel de surveillance sur un problème urgent. De la même manière, l'intégration d'un synthétiseur dans la cabine de pilotage d'un avion permet d'éviter au pilote d'être dépassé par la quantité d'informations visuelles qu'il a à analyser. Et quand on voit à quoi ressemblera bientôt le tableau de bord de nos voitures, on comprend qu'elles ne tarderont pas à nous parler. La maison de demain, quant à elle, fera bien de se doter d'une voix de synthèse si elle veut avertir ses occupants d'une anomalie constatée sur un de ses circuits de surveillance.

- **Communication Homme-Machine, Multimédia**

A plus long terme, le développement de synthétiseurs de haute qualité (ainsi que la mise au point de reconnaisseurs fiables et robustes) permettra à l'homme de communiquer avec la machine de manière plus naturelle. L'explosion récente du marché du multimédia prouve bien l'intérêt du grand public en la matière.

- **Recherche fondamentale et appliquée**

Enfin, les synthétiseurs possèdent aux yeux des phonéticiens une qualité qui nous fait défaut : ils peuvent répéter deux fois exactement la même chose. Ils sont par conséquent utiles pour la validation de théories relatives à la production, à la perception, ou à la compréhension de la parole [1].

- **La synthèse vocale et la musique**

Un autre avantage de la synthèse vocale permettra de générer des choeurs à une vitesse exceptionnelle ou de faire revivre des chanteurs disparus. Ainsi l'utilisation de la synthèse fera gagner énormément d'argent puisque les répétitions destinées aux choeurs ne seront plus nécessaires et que le temps accordé à celles-ci sera utilisé pour la création du choeur mais cette fois à moindre coût puisque quelques personnes suffiront à le créer de manière synthétique.

- **La synthèse vocale et le 7ème Art**

La synthèse de la voix pourrait devenir un atout majeur dans le domaine du cinéma pour différentes raisons. En effet, les doublages sont souvent de moins bonne qualité que les films en Version Original (VO), car les voix qui sont associées aux personnages ne correspondent pas forcément à l'acteur jouant le personnage. La synthèse de la voix permettrait de reproduire le timbre de voix d'un acteur en le faisant s'exprimer dans une autre langue sans le moindre accent. A l'inverse, dans un film où l'acteur devrait s'exprimer dans une langue qu'il ne maîtrise pas, pourrait être doublé par un synthétiseur vocal et ainsi il donnerait l'impression de s'exprimer dans une langue étrangère sans le moindre accent [15].

2.9 PROJET NESPOLE

C'est un projet qui est en train d'être réalisé et qui permet de faire une traduction en temps réel. Ce dernier, co-financé par l'Union Européenne et la NSF (EU), adresse la problématique de la traduction automatique de parole et ses éventuelles applications dans le domaine du commerce électronique et des services. Les langues impliquées sont l'Italien, le Français, l'Allemand et l'Anglais. Les partenaires sont : ITC/IRST de Trento (Italie), ISL Labs. De UKA (Karlsruhe, Allemagne) et CMU (Pittsburgh, USA), Aethra (une société italienne spécialisée dans le domaine de la vidéoconférence), APT (une agence de tourisme dans la région du Trentin en Italie) et le laboratoire CLIPS (Grenoble, France).

Le scénario NESPOLE met en jeu un agent parlant italien, présent dans une agence de tourisme en Italie, et un client qui peut être n'importe où (parlant Anglais, Français ou Allemand) et utilisant un terminal de communication le plus simple possible (PC équipé d'une carte son et d'un logiciel de vidéoconférence). Ce choix correspond aux technologies disponibles aujourd'hui,

mais, dans un futur proche, les mobiles de troisième génération pourraient éventuellement être utilisés comme terminaux (Fig.2.13).

Le client veut organiser un voyage dans la région du Trentin en Italie, et navigue sur site Web de APT (l'agence de tourisme) pour obtenir des informations. Si le client veut en savoir plus, sur un sujet particulier, ou préfère avoir un contact plus direct, un service de traduction de parole en ligne lui permet de dialoguer, dans sa propre langue, avec un agent italien de APT.

Une connexion, via un logiciel de vidéoconférence, est alors ouverte entre le client et l'agent, et la conversation médiatisée (avec service de traduction de parole) entre les deux personnes peut alors démarrer. Dans le projet, l'accent est mis sur certains problèmes scientifiques en traduction automatique de parole : robustesse, extensibilité (extension de la couverture d'un domaine) et portabilité (passage d'un domaine à un autre) [20].

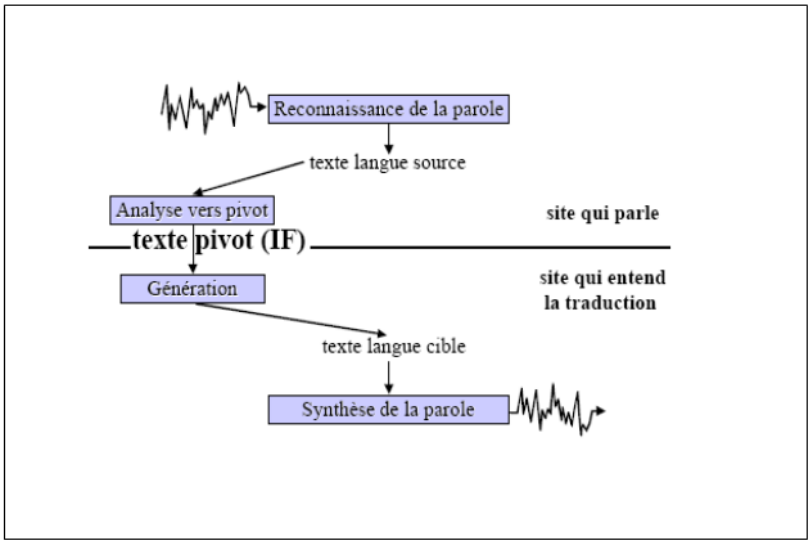


Figure 2.13 : Représentation du projet NESPOLE [19]

2.10 LES DEFAUTS ET LES LIMITES DE LA SYNTHESE VOCALE

Actuellement, la synthèse de la voix est limitée par ces différents aspects :

- les voix synthétiques manquent toujours d'expressivité et ne sont pas encore en mesure de simuler des attributs émotifs comme la joie, la colère ou la tristesse. En résumé, les voix artificielles ne disposent pas encore de la « palette vocale » étendue du locuteur humain ;

- les voix synthétiques sont très limitées. Dans le meilleur des cas, on dispose de quelques voix d'hommes et de femmes pour une langue donnée. Mais les voix d'enfants, d'adolescents ou de personnes plus âgées n'existent pas encore. Créer une nouvelle voix représente un effort majeur, même pour les grandes équipes dotées de financements importants.
- la synthèse des langues particulières (dialecte, variantes sociales, styles et types de parole) commence à peine à être abordée.

En dépit des améliorations indéniables apportées sur le plan de la qualité sonore, les capacités actuelles des synthétiseurs sont encore limitées. Dans le meilleur des cas, les synthétiseurs vocaux possèdent une bonne capacité à fournir un style de lecture à haute voix assez formel.

Mais aucun système n'est capable actuellement de produire une voix véritablement expressive. Les expressions de surprise, de tendresse, d'angoisse ou de déception sont très difficiles, voire impossibles à générer sur les systèmes actuels, compte tenu de la technologie utilisée. De plus, la plupart pour ne pas dire la totalité des synthétiseurs ne reproduisent pas tous les bruits comme par exemple la respiration. Il est important de préciser qu'un synthétiseur de voix est développé pour chaque langue car certaines caractéristiques sont spécifiques à certaines langues et nécessite un traitement particulier [15].

2.11. LA SYNTHESE VOCALE ET SES DANGERS

Si la synthèse de la voix devient comparable à la voix humaine, celle-ci pourrait poser quelques problèmes. En effet, dans le domaine de la biométrie, une méthode concerne l'identification d'une personne en fonction de sa voix. On s'aperçoit alors rapidement que si la synthèse vocale est utilisée de manière à reproduire la voix d'un individu, ce système de protection devient vite obsolète et non fiable.

L'autre danger consiste dans l'usurpation de l'identité d'un individu. En effet, grâce à cette technologie et en sachant que l'un de moyen de communication le plus répandu est la téléphonie, il serait aisé d'usurper l'identité de quelqu'un et de s'en servir de manière illégale ou incorrecte.

De plus, cette technologie pourrait bien accentuer l'effet de déshumanisation forçant ainsi les êtres humains à communiquer de moins en moins entre eux et plus avec des machines.

La synthèse de la voix vise à améliorer le quotidien, mais n'oublions pas que si elle atteint le niveau de conversation d'un humain, elle engendrerait aussi sa substitution dans certains domaines augmentant ainsi l'emprise de la machine sur l'homme [15].

2.12. CONCLUSION

Dans ce chapitre, nous avons abordé les principales méthodes et techniques de la synthèse de la parole, tout en commençant par le traitement linguistique du texte et le prétraitement d'un signal vocal afin de déterminer les paramètres acoustiques qui seront utilisés en synthèse de la parole.

Les deux principaux critères exigés par la synthèse de la voix sont l'intelligibilité et l'aspect naturel, d'où elle vise à améliorer le quotidien, mais n'oublions pas que si elle atteint le niveau de conversation d'un humain, elle engendrerait aussi sa substitution dans certains domaines augmentant ainsi l'emprise de la machine sur l'homme.

Chapitre 3 :

Analyse acoustique du corpus GTPA

3.1 INTRODUCTION

Le but de notre travail est de réaliser un système de synthèse de la parole par unités variables, en vue d'obtenir un **G**uide **P**arlant pour **T**ouristes en **A**lgérie (**GTPA**), en se basant sur la méthode de concaténation par mots et phrases combinés.

Dans ce chapitre, nous nous intéressons à faire une analyse acoustique de notre corpus en étudiant les caractéristiques et les paramètres pertinents de ce signal vocal (formants, fréquence fondamentale, intensité). Nous introduisons les étapes de l'élaboration de notre corpus et son traitement ainsi qu'une analyse prosodique. Nous expliquons le logiciel Praat et finissons par une étude comparative pour quelques signaux vocaux avant et après la concaténation.

3.2 REPRESENTATION SPECTRALE DU SIGNAL VOCAL

Il existe plusieurs représentations spectrales du signal de parole (FFT, LPC, spectrogramme, etc.).

3.2.1 Spectre obtenu par FFT

Un son quelconque est la superposition de plusieurs ondes sinusoïdales. Grâce au Transformée Rapide de Fourier (FFT*)*, nous pouvons isoler les différentes fréquences qui le composent. Ainsi nous obtenons une répartition spectrale du signal. Les valeurs des formants sont calculées automatiquement dans le signal de parole au moyen d'un lissage spectrale (Fig .3.1).

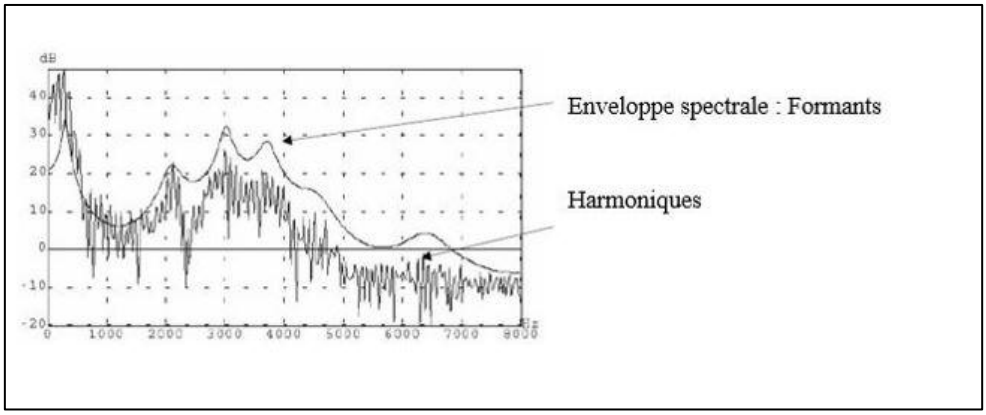


Figure 3.1 : Spectre obtenu par Transformée Rapide de Fourier (FFT)

3.2.2 Spectre obtenu par Codage Prédictif Linéaire (LPC)

Le spectre obtenu par LPC est plus lisse et permet ainsi de repérer plus facilement les formants (Fig. 3.2.).

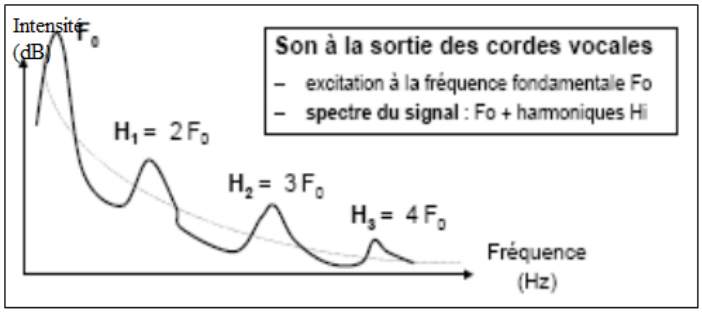


Figure 3.2. : Spectre lissé obtenu par prédiction linéaire

3.2.3 Spectrogramme

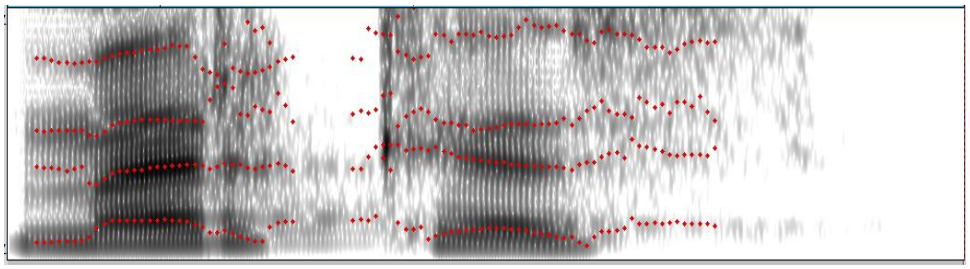
Le spectrogramme est un outil de visualisation utilisant la technique de la Transformée Rapide de Fourier et par conséquent du calcul de spectres. Il a commencé à être largement utilisé en 1947, à l'apparition du sonagraphe, puis il est devenu l'outil incontournable des études en phonétique pendant de nombreuses années [1].

L'apparition de l'informatique puis d'écrans graphiques de bonne qualité a permis d'abandonner tout matériel comme le sonagraphe, mais la technique du spectrogramme est encore aujourd'hui largement utilisée dans de nombreux domaines, du fait de sa simplicité de mise en œuvre et des résultats intéressants qu'elle procure. On parle de spectrogramme à larges bande**s** ou à bandes étroites selon la durée de la fenêtre de pondération. Les spectrogrammes à bandes larges sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms). Ils mettent en évidence l'enveloppe spectrale du signal et permettent de visualiser l'évolution temporelle des formants. Les périodes voisées apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bandes étroites sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées apparaissent sous la forme de bandes horizontales [1]. Le spectrogramme permet de mettre en évidence les différentes composants fréquentielles du signal à tout instant.

Avec un axe des abscisses de temps en millisecondes, des ordonnées de fréquences en

Hz et l'intensité est donné par le degré de noirceur de la trace (Fig.3.3).

a)



b)

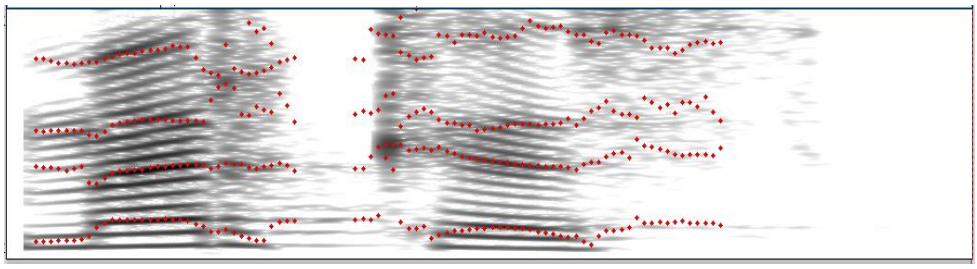


Figure 3.3 : Spectrogramme de la phrase[$markaz\ ti\tilde{g}aarii$]

a) en bande large avec une fenêtre de Hamming de 5 ms

b) en bande étroite avec une fenêtre de Hamming de 30 ms

3.2.4 Intérêts de la représentation fréquentielle du signal de parole

La représentation fréquentielle de la parole a une très grande importance dans le domaine de la Communication Parlée. Elle a permis l'extraction des paramètres pertinents du signal de parole comme la fréquence fondamentale, l'intensité et les formants. Ces paramètres sont d'une importance capitale dans de nombreux domaines comme :

- les différentes méthodes de synthèse ;
- la Reconnaissance Automatique de la Parole et du Locuteur ;
- l'Identification Automatique des Langues ;
- et bien d'autres domaines.

3.3 LECTURE DU SPECTROGRAMME

La lecture du spectrogramme contient 4 étapes élémentaires :

Étape 1 : Connaître les 3 dimensions du spectrogramme. L'énergie (l'intensité), le temps et la fréquence du spectre) ;

Étape 2 : Savoir distinguer les consonnes et les voyelles ;

Étape 3 : Savoir reconnaître les grandes classes de consonnes. Il y a 3 types de Consonnes, les occlusives, les fricatives et les sonantes ;

Étape 4 : Savoir reconnaître les grandes classes de voyelles. Les voyelles se différencient les unes des autres par leurs formants. Un formant est la zone de fréquence où il y a une concentration (renforcement) d'énergie. Nous utilisons souvent le spectrogramme à bande large pour visualiser les formants. Ces derniers apparaissent sous les formes des bandes noires horizontales.

3.4 L'OUTIL D'ANALYSE PRAAT

Praat est un logiciel libre pour l'analyse, la manipulation et l'annotation des sons. Ces fonctionnalités en font un outil complet, en particulier pour l'étude de la parole. Il permet également de tracer des graphiques, construire des grammaires basées sur la théorie de l'optimalité, de faire une synthèse articulatoire, de simuler des réseaux de neurones et de faire des analyses statistiques. Paul Boersma et David Weenink de l'Institute of Phonetic Sciences de l'Université d'Amsterdam ont créé Praat en 1996 et continuent activement de développer cet outil de manière très interactive avec la communauté des utilisateurs [21].

Il a été conçu à la fois pour les non-experts en traitement de la parole grâce à ses interfaces graphiques et menus simplifiés, cependant pour les utilisateurs avancés il ont de nombreuses possibilités de manipulations, d'analyses et de Scripting. Ce programme offre la possibilité d'effectuer de multiples tâches [23] :

- enregistrer des fichiers audio qui peuvent être ensuite analysés sous Praat. ils peuvent être aussi codés selon une multitude de formats audio ;
- segmenter, transcrire et annoter des fichiers audio dont la taille peut aller jusqu'à 2 Giga bytes, c'est-à-dire 3 heures d'enregistrement stéréo de qualité CD ou 16 heures

d'enregistrement mono à 22 kHz. ces enregistrements peuvent être effectués sous Praat ou provenir d'autres fichiers audio a divers format;

- effectuer des analyses phonétiques et acoustiques au niveau segmental. Il permet de calculer des paramètres prosodiques comme l'intensité, la fréquence fondamentale, le voisement, le timbre, …etc., et ceci selon plusieurs algorithmes. De mener des analyses spectrographiques et des mesures précises telles que la durée du VOT (Voice On Time) des plosives, les valeurs des différents formants d'une voyelle, …etc. ;
- étudier les paramètres prosodiques (F_0, durée et intensité), modifier par stylisation des courbes de fréquence fondamentale et d'intensité ;
- effectuer des manipulations et des modifications du signal de parole (utilisation de filtres, analyse-synthèse, …etc.) ;
- construire des outils d'apprentissage (réseau de neurones et élaboration de grammaires dans le cadre de la théorie de l'optimalité (OT : Optimality Theory) ;
- écrire des scripts pour effectuer plus rapidement certaines tâches d'analyse, d'extraction d'information ou d'édition, etc.

Le logiciel Praat propose une interface assez déroutante au premier abord, dans la mesure où elle est différente de celle fréquemment rencontrée.

Ainsi, au lancement du programme Praat, deux fenêtres s'ouvrent à l'écran (Fig 3.4.).

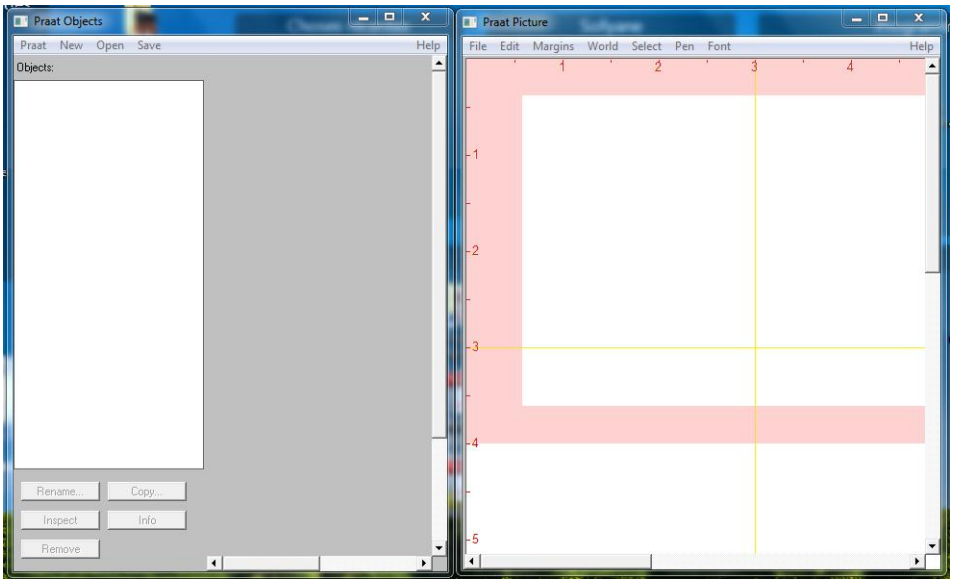


Figure 3.4. : Ecran à l'ouverture de Praat

La fenêtre de gauche est intitulée “Praat objects”, elle sert à “lister” les différents objets (fichiers sons, fichiers d’annotations, etc.) à partir desquels sont effectuées les analyses ou on trouve les résultats affiches. La fenêtre de droite, intitulée “Praat picture”, est utilisée pour reproduire des figures (sonagramme, courbe de F_0, etc.) qui peuvent être exportées vers d’autres logiciels (traitement de texte, …etc.) [22].

3.5 ELABORATION DU CORPUS CONCERNANT LE GTPA

Dans le cadre de notre travail, nous avons utilisé un corpus de parole continue composé de 32 mots. Qui se nomme « GTPA » (Tableau 3.1).

Tableau 3.1 : corpus GTPA

Phrase		Traduction en Français	TOP
P_1	أريد أقرب	Je veux le/la plus proche	[$uriidu\ 'qrab$]
P_2	شاطئ	Plage	[$\tilde{s}a\underline{t}i'$]
P_3	فندق	Hôtel	[$funduq$]
P_4	مستشفى	Hôpital	[$mustaš\tilde{f}a$]
P_5	مطعم	Restaurant	[$ma\underline{t}'am$]
P_6	مقهى	Cafétéria	[$maqha$]
P_7	مركز شرطة	Centrale de Police	[$markaz\ \tilde{s}ur\underline{t}a$]
P_8	مركز تجاري	Centre commercial	[$markaz\ ti\tilde{g}aarii$]
P_9	محطة حافلات	Station de bus	[$maha\underline{t}at\ haafilaat$]
P_{10}	محطة تاكسي	Station de taxi	[$maha\underline{t}at\ taaksii$]
P_{11}	محطة قطار	Station de train	[$maha\underline{t}at\ qi\underline{t}aar$]
P_{12}	أريد أن	Je veux :	[$uriidu\ 'n$]
P_{13}	أبلغ عن ضياع	Déclarer une perte	[$'ubaliga\ 'an\ \underline{d}aya'$]
P_{14}	أتصل بالإسعاف	Contacter les urgences	[$'atasila\ bil'is'af$]
P_{15}	أرسل طردا	Envoyer un courrier	[$ursila\ \underline{t}ardan$]
P_{16}	وين جاي سبيطار	où se trouve l’hôpital ?	[win $\tilde{g}ay\ sbiitaar$]
P_{17}	ماني يتيلي سبيطار	où se trouve l’hôpital ?	[$maanii\ ytiilii\ sbiitaar$]

Pour l'Arabe Standard ; concernant la première catégorie nous avons une phrase fixe (**P_1**) avec 10 parties variables (de **P_2** jusqu' a **P_{11}**). Pour la deuxième catégorie on a une phrase fixe (**P_{12}**) et 3 parties variables (**P_{13}**, **P_{14} et P_{15}**). Une phrase avec un dialecte du centre de l'Algérie (**P_{16}**) et en Amazigh (Chawia) (**P_{17}**).

Le GTPA a été enregistré après plusieurs essais par deux locuteurs adultes arabophones (Homme et Femme) en stéréo (gauche et droite). Ce corpus contient les informations qui peuvent être demandées par un touriste.

Nous justifions le choix de ce type du corpus (parole continue au lieu d'avoir utilisé les logatomes) par le fait qu'il est préférable d'étudier les segments dans un continuum vocal afin de pouvoir prendre en considération les effets de coarticulation existants entres les phonèmes.

Durant notre travail dans ce chapitre, nous allons suivre cet enchainement (Fig.3.5)

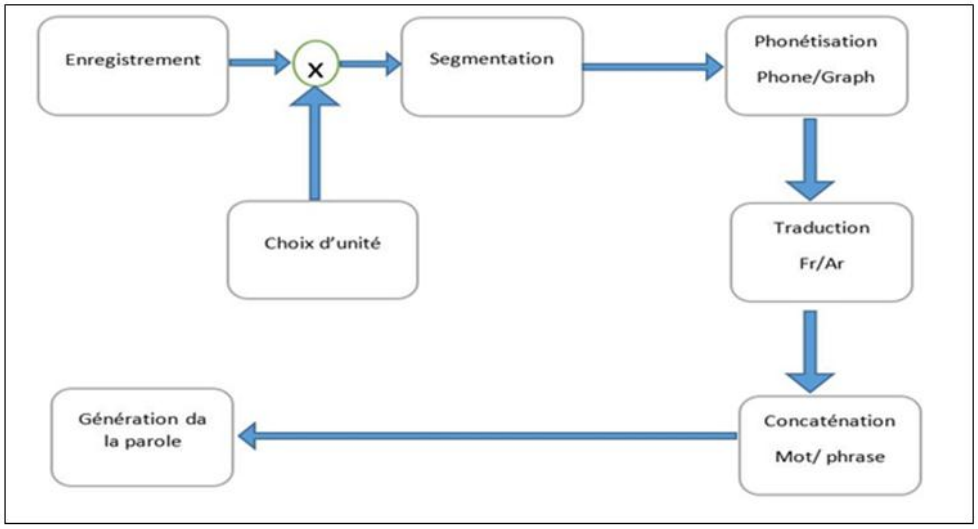


Figure 3.5 : Algorithme de la synthèse

Cette base de données choisie pour la suite du chapitre est celle de la locutrice et en format mono Left, de durée 21.07 secondes et de taille mémoire de stockage de 1.92 Méga Octet (Mo) (Fig.3.6).

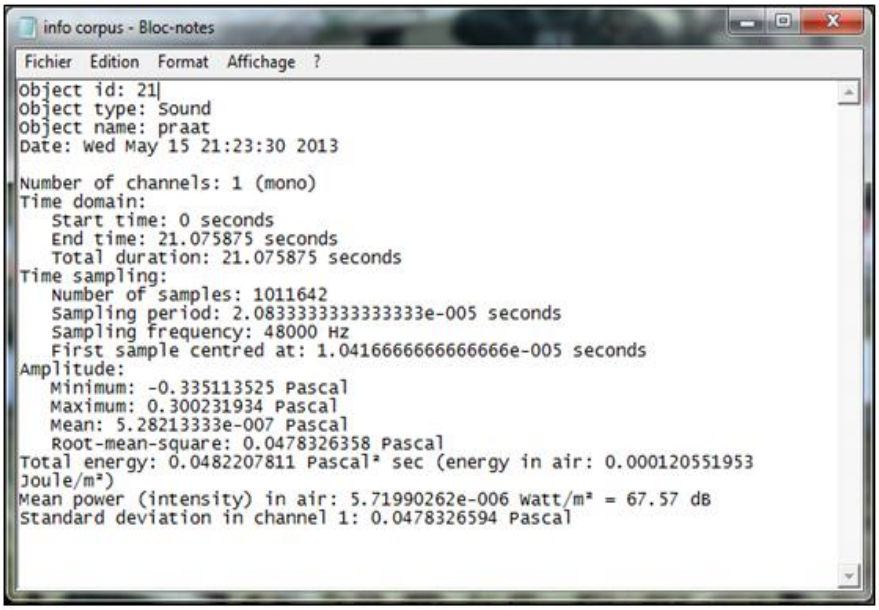


Figure 3.6 : Information sur le GTPA par Praat

- Une amplitude : - Maximum de 0.300 Pascal ;
  - Minimum de -0.335 Pascal;
  - Moyenne de 5.28 $x10^{-6}$ Pascal ;
- Une énergie Totale de 0.048 $Pascals^2$ sec.

3.6 ENREGISTREMENT DU CORPUS

L'enregistrement s'est fait dans une chambre sourde sans bruit ni réverbération au niveau de l'Institut Supérieur des Métiers des Arts du Spectacle et de l'Audiovisuel (ISMAS) – à Alger, avec de bonnes conditions d'enregistrement :

- les données sont échantillonnées avec une fréquence de 48 kHz codés sur 24 bits qui correspondent à une bonne qualité de la parole ;
- le format : multiple mono (stéréo) ;
- logiciel utilisé « Pro Tools speckles » version 8 ;
- le type de parole : phrases en parole continue ;
- les signaux acoustiques sont enregistrés en format (WAV).

3.7 EQUIPEMENT UTILISES EN ENREGISTREMENT

Le matériel que nous avons utilisé comprend :

- une cabine technique qui contient : une table de mixage, une carte d'acquisition, un micro-ordinateur, des Haut-Parleurs, des Microphones (Fig. 3.7);
- une cabine Speaker : c'est une chambre isolée qui contient des Microphones et des casques(les deux cabines sont séparées par un verre transparent et isolant) (Fig. 3.7) ;
- une Station Pro Tools, Version 8 et de Bonne qualité (Fig. 3.8) ;
- un Microphone professionnel unidirectionnel Electrodynamique [Beyer dynamic M 69 TG] (Fig. 3.9).

Figure 3.7 : Cabine Speaker + cabine technique

Figure 3.8 : Station Pro Tools

Figure 3.9 : Microphone Beyer dynamic M 69 TG

A l'aide de l'outil Praat nous obtenons les sonagrammes du signal vocal de GTPA (Fig. 3.10)

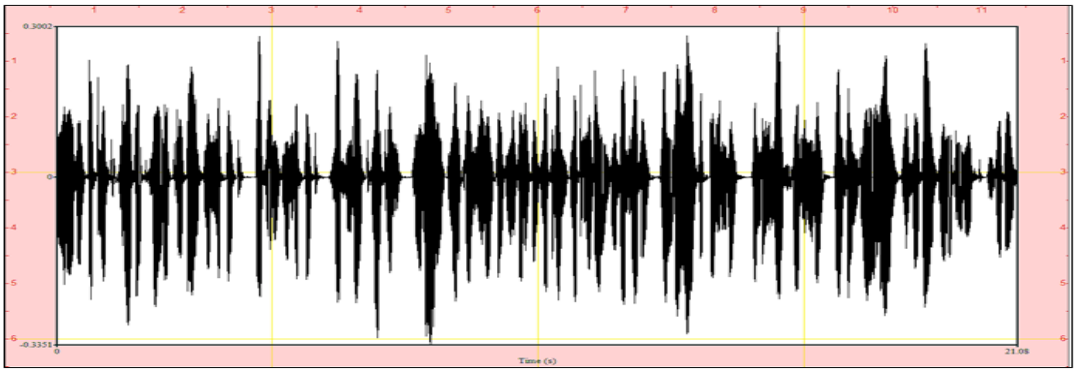


Figure 3.10 : Visualisation du signal audio du corpus par Praat Picture

3.8 PROCEDURE DE SEGMENTATION MANUELLE

Les fichiers ' son' segmentés en utilisant le logiciel Praat, l'image du spectrogramme, ainsi que le fichier correspondant aux formants seront stockés dans des répertoires. La segmentation du GTPA a été effectuée manuellement ce qui justifie le temps, relativement long, alloué à cette opération. La procédure adoptée pour isoler les phonèmes et dégager l'unité à étudier de l'onde temporelle, en fin effectué des tests de perception pour s'assurer de la qualité de la segmentation (Fig. 3.11).

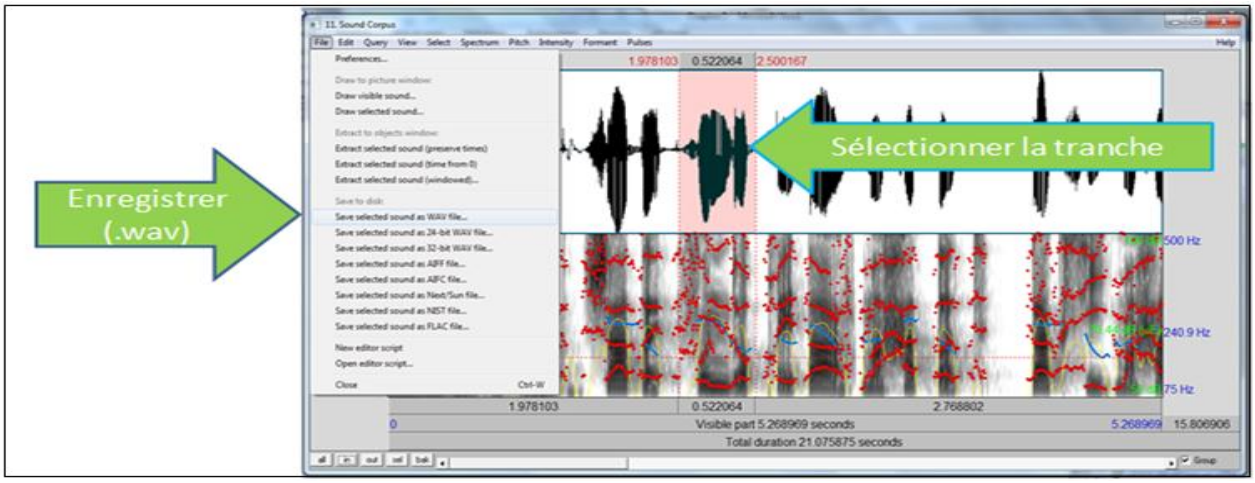


Figure 3.11: Procédure de segmentation du mot P_4

La première étape consiste à sélectionner l'objet son en cliquant sur son nom dans la liste.

Parfois il peut être pratique de segmenter un oscillogramme vocal et d'attacher des gloses à chaque segment pour une utilisation ultérieure :

- Sélectionner l'objet son original en cliquant sur son nom ;
- Sélectionnez en même temps l'objet 'Son' et l'objet Texte (ils ont le même nom) en utilisant la touche CTRL (cliquez sur l'objet 'Son', maintenez l'appui sur la touche CTRL et cliquez sur l'objet Texte).
- Aller sur 'Annotate' et sélectionner 'To TextGrid..'. Ceci fera apparaître la Fenêtre (Fig. 3.12)

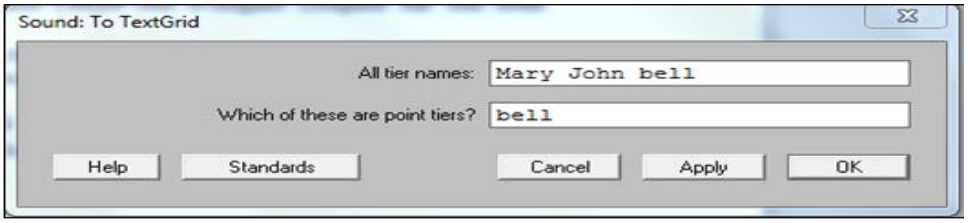


Figure 3.12 : segmentation par TextGrid

- Nous précisons le type de segmentation et nous le remplaçons dans la case ' Mary John bell ' : soit par phonème, diphone, mot, etc. ;
- Sur la droite de la fenêtre un nouveau menu devrait apparaître. Sélectionner 'Edit', ce qui fera apparaître une nouvelle image (évidemment le signal son sera différent avec votre échantillon) (Fig.3.13).

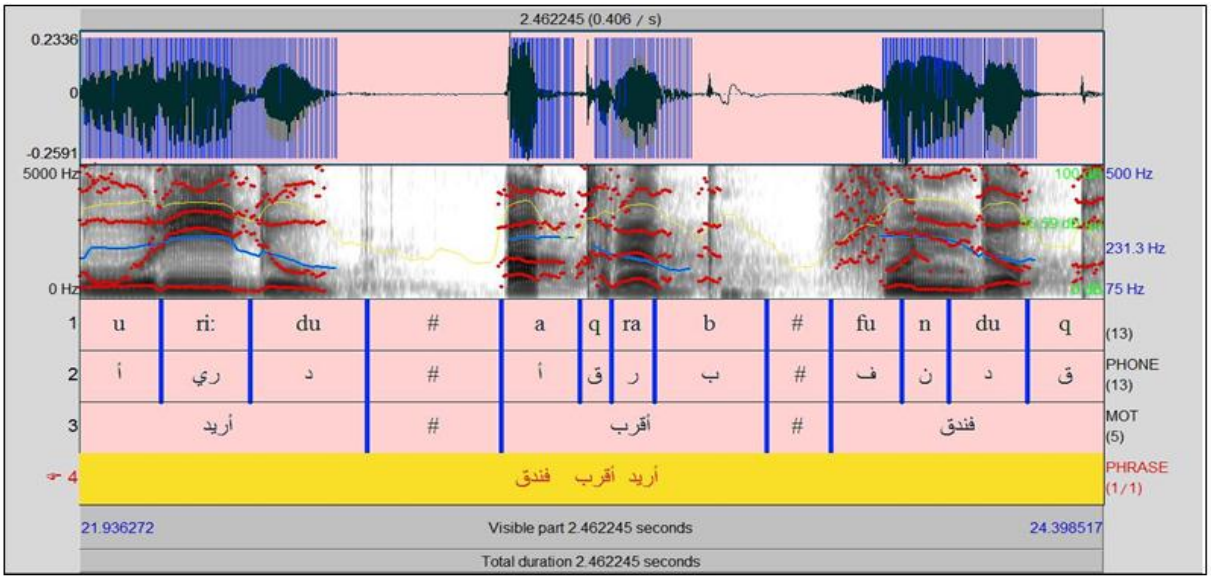
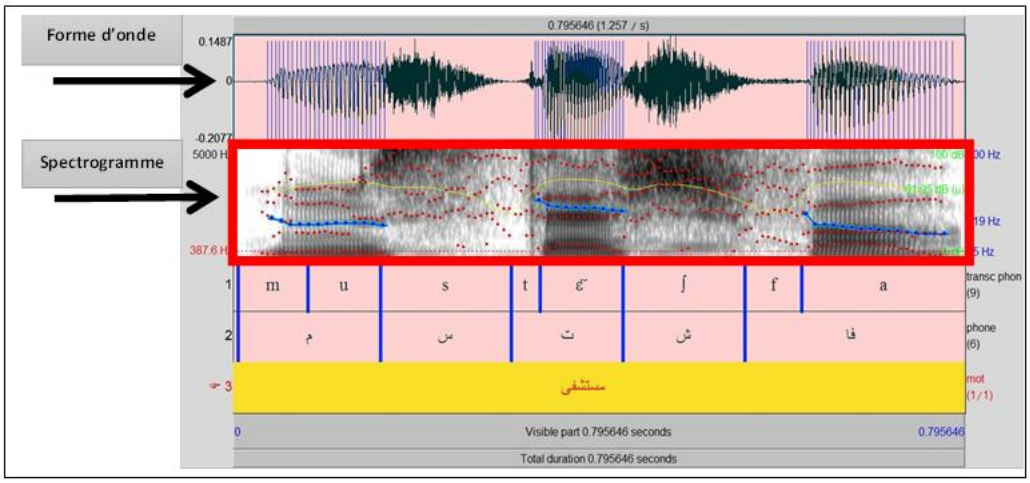


Figure 3.13 : Segmentation par phonème et mot de la phrase $P_1 + P_3$

Les principaux paramètres d'analyse : oscillogramme (forme onde ou audiogramme), intensité, spectrogramme, mélodie (pitch) et les formants ; Nous les obtenons suivant ces étapes :

- la fréquence fondamentale : cochez 'Show pitch ' dans le menu ' Pitch ', et elle apparaît (c'est une courbe de couleur cyan). Sa valeur moyenne (Hz) s'affiche à droite.
- les formants : cochez 'Show formants' dans le menu ' Formant ', et ils apparaissent en pointillés rouges. Pour les afficher sur toute la longueur de la fenêtre, affichez la fenêtre 'Formant Settings' du menu 'Formant', et dans le champ 'Maximum Duration ', entrez la durée de la fenêtre, en secondes, à la place de la valeur initiale.
- l'intensité : cochez ' Show Intensity ' dans le menu 'Intensity ', et elle apparaît (c'est une courbe de couleur Jaune). Sa valeur moyenne (dB) s'affiche à droite.
- les périodes du signal sonore: cochez 'Show pulses 'dans le menu 'Pulses' . Chaque période est représentée sur l'enveloppe par un trait bleu vertical (Fig.3.14).

a)



b)

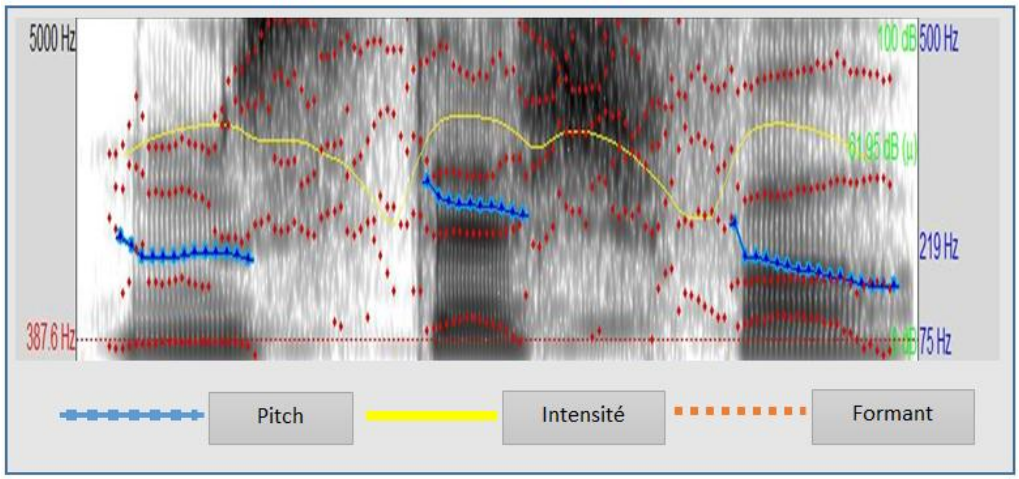


Figure 3.14 : a) visualisation du signal audio pour le mot P_4

b) visualisation des paramètres pertinents pour le mot P_4

La synthèse de parole par concaténation par forme d'ondes est une déclaration des segments sonores enregistrés, la combinaison entre ces segments imposent de faire l'insertion de pauses silencieuses d'une durée appropriée entre les segments. Cette technique élimine les bruits et augmente la qualité de la parole synthétique. L'opération d'insertions des pauses silencieuses a été faite par le logiciel Praat, lorsque nous faisons la segmentation du GTPA, cette méthode peut facilement créer une parole synthétisée de haute qualité presque naturelle.

Dans le cadre de notre travail, nous utilisons la concaténation par forme d'ondes pour 2 catégories (Fig.3.15) :

- la première : $P_1 + P_{2, 3,\ldots,10}$
- la deuxième : $P_{12} + P_{13, 14,15}$

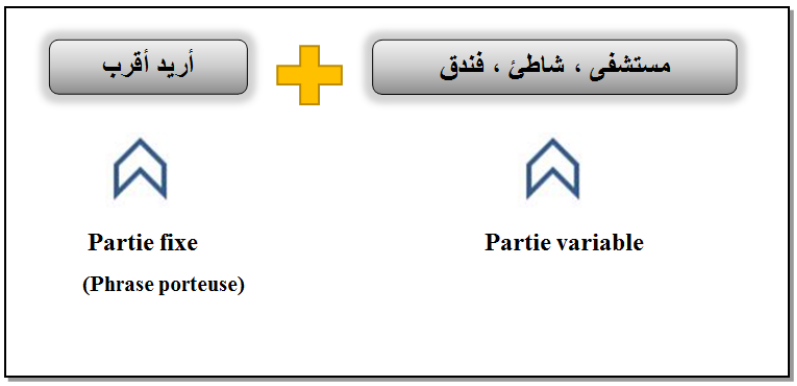


Figure 3.15 : Assemblage de la phrase fixe avec les mots variables

Dans la première catégorie nous avons 10 concaténations, cependant pour la deuxième catégorie nous avons 3 concaténations. (Ces deux catégories ont un seul point de concaténation à chaque phrase synthétisée) (Fig.3.16).

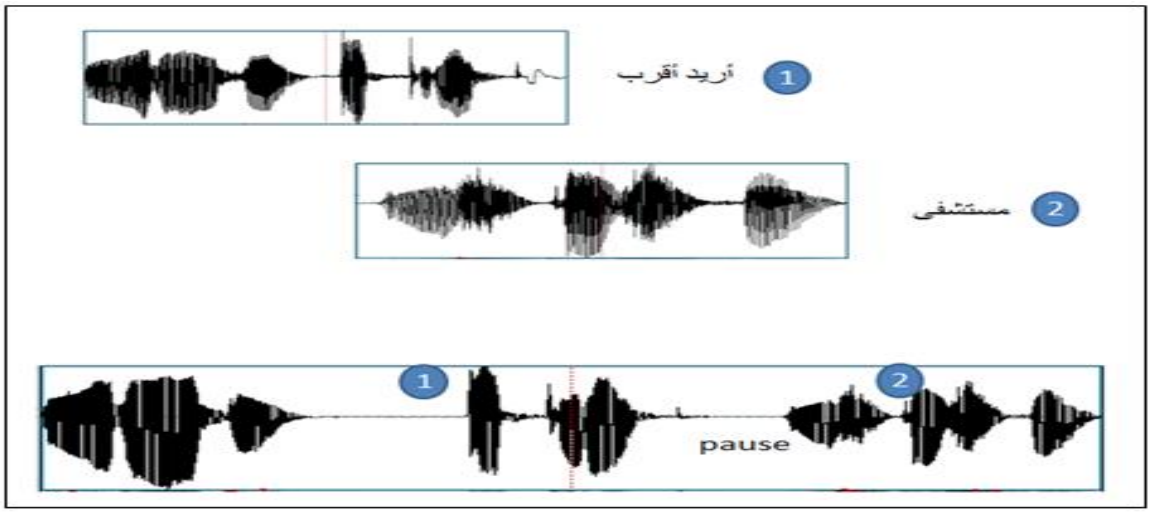
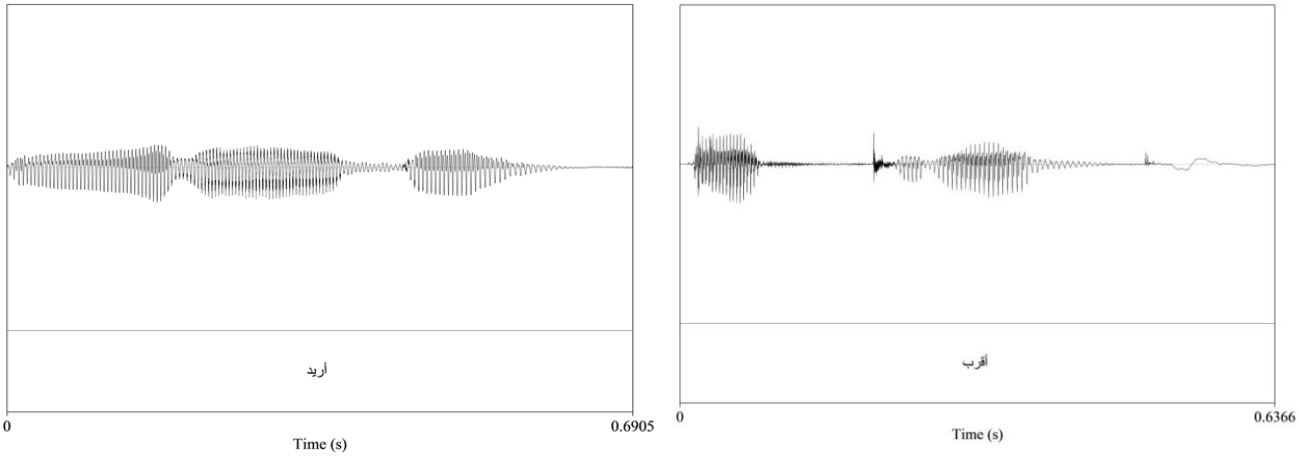


Figure 3.16 : Concaténation par forme d'ondes de la phrase $P_1 + P_4$

Les formes d'ondes de quelques signaux vocaux de notre GTPA (Fig.3.17)



a) Je veux b) le/la plus proche

Figure 3.17 : représentation spectrale de quelques signaux vocaux

3.9 TRADUCTION DU GTPA

Pour notre corpus, nous allons faire une traduction de la langue Française (langage du touriste) vers la langue cible qui est l'AS, dialecte du centre de l'Algérie et Amazigh (Tableau 3.1).

3.10 ETUDE COMPARATIVE DU SIGNAL ORIGINAL AVEC LE SYNTHETISE

Cette étude représente une comparaison des phrases du GTPA avant (signal vocal original) et après concaténation. Prenant des échantillons de phrases qui sont nombre de cinq prononcé en AS par une locutrice, dont laquelle pour les 5 phrases avant concaténation nous les avons obtenus par un enregistrement en parole continue avec la même locutrice et pour celles après concaténation, nous avons fait un réenregistrement du signal de sortie de notre interface graphique en temps réel et dans les mêmes conditions d'enregistrement du corpus GTPA. Cette étude représente un test objectif du GTPA du côté de l'intelligibilité et l'aspect naturel de la parole synthétisée.

3.10.1 Transcription Orthographique Phonétique du GTPA

Avant de commencer l'analyse, et faire une comparaison, nous devrons passer par une TOP pour ces cinq phrases de GTPA (Tableau 3.1)

- P_1 : $P_1 + P_4$
- P_2 : $P_1 + P_3$
- P_3 : $P_1 + P_8$
- P_4 : $P_1 + P_5$
- P_5 : $P_{12} + P_{15}$

3.10.2 procédures d'analyse du p_i (i=1 ; 5)

Le but de cette comparaison est l'étude de la qualité et de la performance de la parole obtenue par concaténation par rapport à la parole naturelle qui a été enregistrée. La comparaison sera basée sur l'analyse :

- générale : concernant la taille, durée, etc. ;
- formantiques : pour les cinq premiers ;
- de l'intensité : le gain, l'intensité maximale, minimale et moyenne ;
- fréquentielle : F_0 maximale, minimale et moyenne ; le nombre de zone voisée et non voisée.

Dans notre analyse et a l'aide de la distance euclidienne d'ordre 1, nous écrivons [24] :

$$d_p = ||x - y||_p \quad (3.1)$$

Elle définit une distance qui correspond pour :

- p = 1 : à la valeur absolue moyenne ;
- p = 2 : à l'écart quadratique moyen ;
- p = ∞ : à l'écart maximum.

3.10.2.1 Analyse générale du GTPA du P_i (i=1 ; 5)

Cette analyse correspond aux : tailles, durée, amplitude, énergie et puissance moyenne (Tableau 3.2).

- P_{1av} : Phrase 1 avant concaténation ;
- P_{1ap} : Phrase 1 après concaténation.

Tableau 3.2 : Les paramètres généraux des phrases avant et après concaténation

	Les phrases									
	P_{1av}	P_{1ap}	P_{2av}	P_{2ap}	P_{3av}	P_{3ap}	P_{4av}	P_{4ap}	P_{5av}	P_{5ap}
Paramètres										
Taille en Ko	238	199	237	187	329	274	232	186	294	236
durée en s	2.54	2.12	2.52	1.98	3.49	2.91	2.47	1.97	3.12	2.51
Amplitude Min [Pascal]	-0.31	-0.24	-0.25	-0.25	-0.33	-0.33	-0.24	-0.24	-0.24	-0.24
Amplitude Max [Pascal]	0.21	0.233	0.233	0.233	0.271	0.244	0.22	0.23	0.21	0.21
Energie $[10^{-5}\mathrm{J}/m^2]$	1.36	1.097	1.4	1.4	2.22	2.046	1.14	1.14	1.73	1.69

Nous calculons la précision de la taille, la durée et l'énergie pour les cinq phrases telle que :

$$\text{Précision } (\%) = \frac{|valeur\ avant - valeur\ après\,|}{valeur\ avant} \times 100 \qquad (3.2)$$

Pour illustrer et comparer ces données nous présentons le graphe suivant (Fig.3.18)

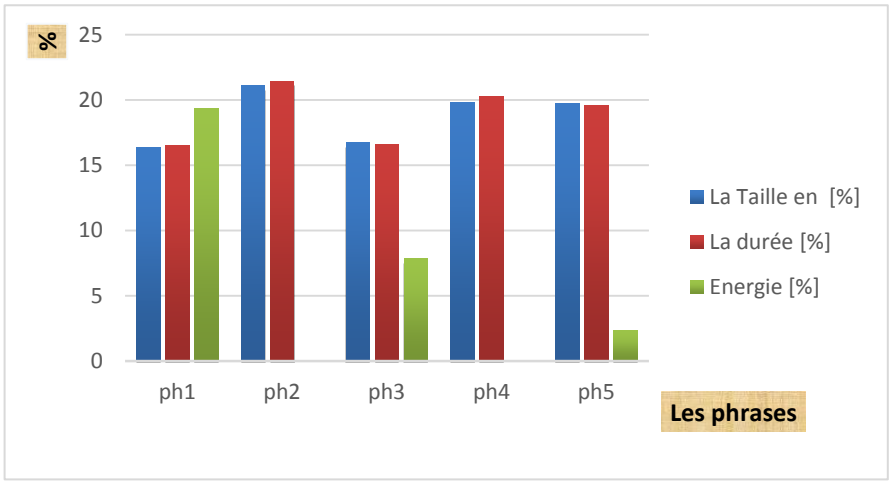


Figure 3.18 : précision de la taille, durée et l'énergie des phrases

3.10.2.2 Extraction des formants

Nous obtenons les valeurs moyennes des cinq premiers formants pour les cinq phrases avant et après concaténation (Tableau 3.3).

Tableau 3.3 : les valeurs des formants de 5 phrases

	Les Phrases									
	P_{1av}	P_{1ap}	P_{2av}	P_{2ap}	P_{3av}	P_{3ap}	P_{4av}	P_{4ap}	P_{5av}	P_{5ap}
Formants Hz										
F_1	687.4	711.9	601.5	601.4	597.9	601.7	605.8	602.1	505.3	498.7
F_2	1772.3	1719.1	1620.5	1590.6	1851.0	1842.0	1582.9	1532.1	1731.4	1697.2
F_3	2908.7	2912.1	2819.2	2782.4	2906.8	2880.5	2774.7	2715.5	2857.9	2839.6
F_4	4011.9	4065.4	4021.9	3999.2	4112.6	4113.3	3983.7	3913.6	4128.2	4146.2
F_5	4594.3	4720.7	4776.	4740.4	4721.0	4825.1	4862.1	4788.2	4888.7	4847.5

Nous illustrons les variations des trois premiers formants par le graphe (Fig.3.19).

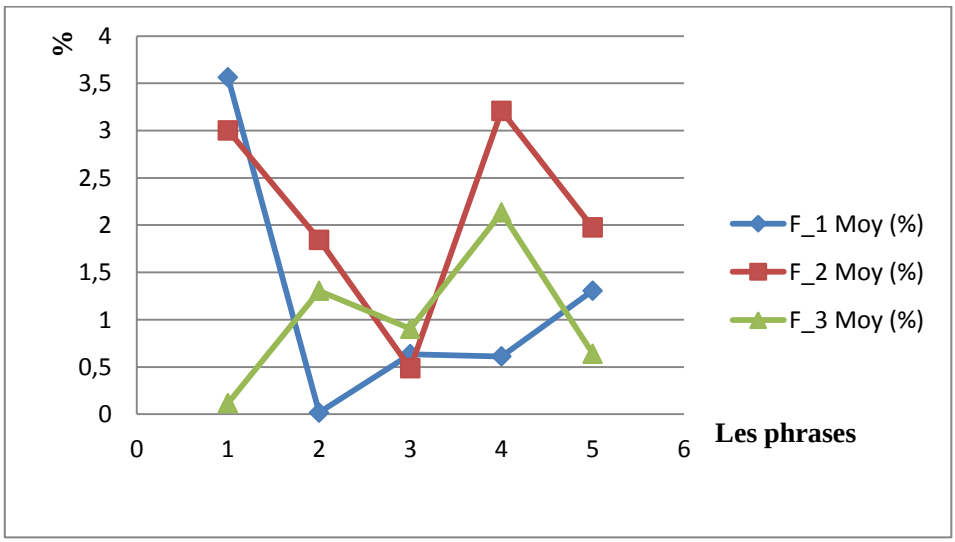


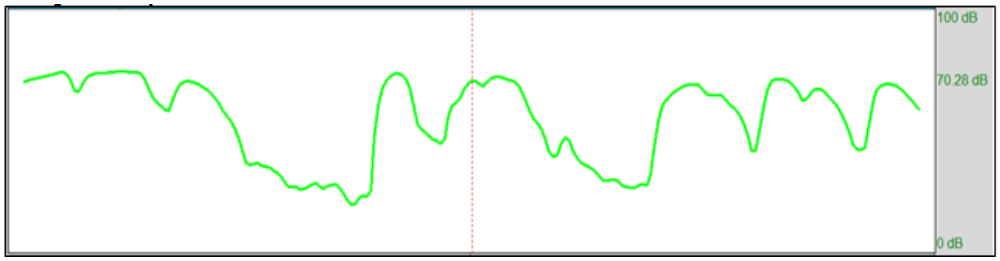
Figure 3.19 : Variations des formants

3.10.2.3 Analyse de l'intensité

Nous analysons l'intensité, puis nous extrayons les propriétés de chaque phrase en faisant une comparaison.

Pour le diagramme d'intensité de la phrase P_1, nous visualisons (Fig. 3.20)

a)



b)

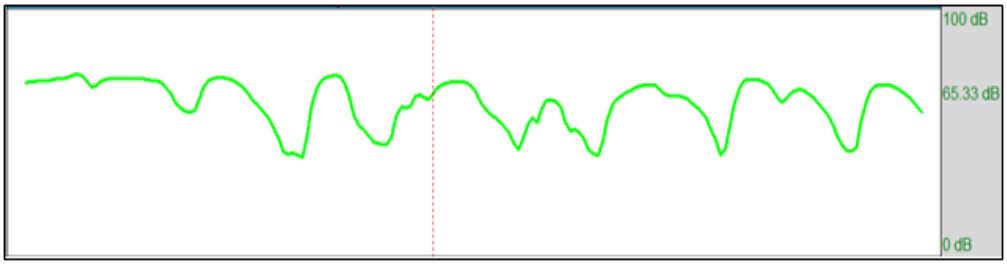


Figure 3.20 : diagramme de l'intensité de la phrase P_1

a) avant concaténation
b) après concaténation

Nous extrayons le gain de chaque phrase avec un ordre de prédiction M=12 ; plus M augmente plus l'enveloppe spectrale est semblable à celle du signal original. L'ordre est le résultat d'un compromis entre une bonne représentation de la structure formantique et la complexité de calcul. Pour satisfaire ce compromis, l'ordre est choisi généralement de 8 à 16.

Les différentes valeurs qui caractérisent l'intensité sont représentées sur le tableau 3.4

Tableau 3.4 : Paramètres caractéristiques de l'intensité

	Phrases									
	P_{1av}	P_{1ap}	P_{2av}	P_{2ap}	P_{3av}	P_{3ap}	P_{4av}	P_{4ap}	P_{5av}	P_{5ap}
Paramètres										
Gain $[10^{-5}]$	257.7	216.8	266	253.6	411	475.8	872.2	856.7	3001	3138.6
Intensité Moy (dB)	67.36	67.20	67.58	68.58	68.09	68.52	66.74	67.72	67.50	68.37
Intensité Min (dB)	19.54	39.02	23.21	37.84	19.72	21.67	22.44	40.30	23.89	27.13
Intensité Max (dB)	74.45	73.34	74.90	74.88	75.68	75.72	74.22	73.35	74.79	73.85

3.10.2.4 Analyses fréquentielles (Pitch)

L'analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou pitch.

Nous obtenons les caractéristiques de voisement ou non voisement suivant ces étapes :

a) sélectionner le fichier son ;

b) un clic sur 'Annotate' et après sur 'To TextGrid silence' (Fig.3.21)

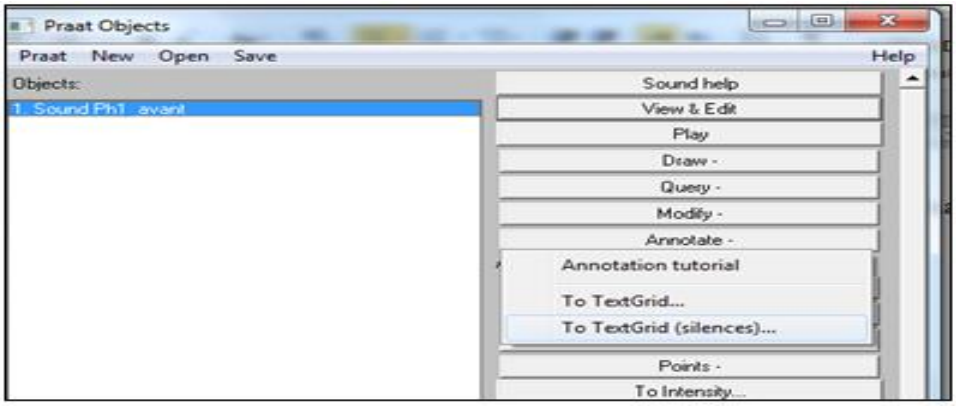
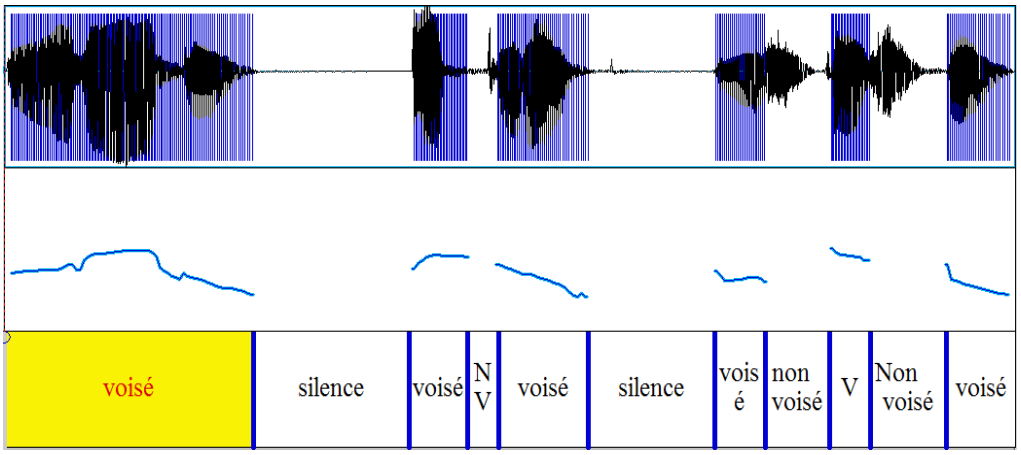


Figure 3.21 : Visualisation des zones sonores et silences

c) sélectionner les deux fichiers en cliquant sur ‘View & Edit’

Nous obtenons l’illustration suivante (Fig. 3.22)

a)



b)

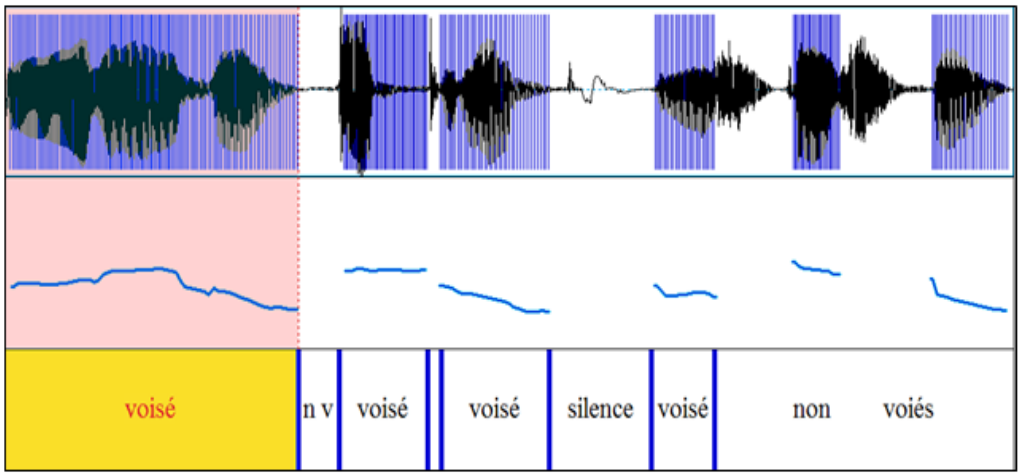


Figure 3.22 : Mélodie de la phrase P_1

a) avant concaténation
b) après concaténation

Par une analyse fréquentielle des cinq phrases nous obtenons les mesures qui sont définies dans le Tableau 3.5.

Tableau 3.5 : Quelques paramètres caractéristiques du pitch

	Phrases									
	P_{1av}	P_{1ap}	P_{2av}	P_{2ap}	P_{3av}	P_{3ap}	P_{4av}	P_{4ap}	P_{5av}	P_{5ap}
Paramètres										
F_0 MIN (Hz)	163.70	165.27	165.68	165.24	168.35	165.24	164.87	165.24	170.42	166.02
F_0MAX (Hz)	290.92	290.68	277.12	277.17	277.31	272.40	267.04	272.33	276.08	275.24
F_0MOY (Hz)	228.76	226.15	230.86	231.34	232.18	230.56	222.92	223.25	217.83	220.29
Nombres des zones de sons voisés	006	006	005	004	007	007	005	005	005	005
Nombres des zones de sons non-voisés	003	003	003	003	006	006	002	003	001	003

3.10.2.5 Interprétations générales sur l'étude comparative

Nous remarquons pour chaque phrase comparée du GTPA que les variations des paramètres sont acceptables, qui sont la taille mémoire, la durée et l'énergie. Pour les deux première on a une précision qui est comprise entre 17 % et 22 %, cependant pour le troisième paramètre nous avons une bonne précision qui est entre 0 et 7 % sauf pour la première phrase qui se présente comme le début du corpus. La variation de précision est due aux zones de silences que nous avons segmentées, dont laquelle ces zones ont une énergie, une durée et une taille mémoire propre a eu.

Nous constatons une grande similarité entre les trois premiers formants obtenus, elle se définit par une précision avant et après concaténation qui ne dépasse pas les 3.6 %. Même chose pour : le gain, l'intensité moyenne, l'intensité maximum, alors que l'intensité minimum avait une autre variation.

Pour les 5 paramètres de la fréquence fondamentale (pitch), nous avons obtenu de bons résultats avec des faibles variations. Ce qui justifie la bonne qualité de la parole obtenue.

3.11 ALGORITHME DU GTPA

1) Choix du corpus (GTPA) en langue Française ;
2) Enregistrement du corpus en parole continue ;
3) Choix du type de segmentation (mots et phrases combinés) ;
4) Phonétisation : Transcription Orthographique Phonétique du GTPA ;
5) Traduction du corpus, du Français vers l'Arabe Standard, dialecte du centre d'Algérie ou encore en Amazigh (chawia) ;
6) Concaténation des unités segmentées a 'laide du programme GTPA ;
7) Génération du signal de sortie (synthétisé).

3.12 CONCLUSION

Le but de ce chapitre est de faire une analyse acoustique de notre corpus 'GTPA' en passant par plusieurs étapes d'analyse et de visualisation. Terminé par un test objectif du signal original et du synthétique ; les résultats obtenus nous ont permis de confirmer le bon enregistrement du signal vocal sachant que la segmentation a été faite manuellement.

Malgré sa simplicité, la synthèse par concaténation en forme d'ondes préenregistrée est capable de produire des discours de haute qualité se rapprochant du naturel.

Chapitre 4 :

Application du GTPA

4.1 INTRODUCTION

Dans ce chapitre nous présentons le programme que nous avons développé en vue d'un Guide Touristique Parlant en Algérie, après une brève introduction au langage de programmation C# 2010 sous l'environnement Visual studio. Nous allons faire un test de perception subjective afin de pouvoir évaluer les résultats obtenus (signal vocal de sortie de l'interface), en ce qui concerne l'intelligibilité et l'aspect naturel. Nous finissons par la réalisation pratique (circuit du guide parlant).

4.2 PRESENTATION DU GUIDE TOURISTIQUE PARLANT EN ALGERIE

Le guide est destiné spécialement pour des touristes Français ou parlant la langue française, en faisant une traduction et une production d'un signal vocal en temps réel en utilisant la méthode de synthèse par concaténation de phrases et mots combinés. Dans cette méthode, les éléments du message, préenregistrés et mémorisés sous forme condensée. Ils constituent le dictionnaire de base de notre système. Nous procédons à une synthèse pour reconstituer des sorties vocales à partir d'une phrase fixe et des mots variables préalablement enregistrés. Cette synthèse se fait par la juxtaposition des mots extraits du dictionnaire, pour cela nous avons conçu une interface graphique qui illustre ce système de synthèse, à l'aide du langage C #.

Le guide est considéré comme un moyen qui facilite la communication entre le touriste et un citoyen algérien, afin de répondre aux besoins et aux questions posées par ce touriste. Il comprend deux locuteurs (un homme et une femme).

4.3 Microsoft Visual C # 2010

Microsoft Visual C# 2010 est un puissant langage orienté composant, créé par Microsoft. Il joue un rôle essentiel dans l'architecture de Microsoft .NET Framework. Certaines personnes ont comparé son rôle à celui joué par le langage C dans le développement d'UNIX. Si on connaît déjà un langage comme C, C++ ou Java, on trouvera que la syntaxe de C# se rapproche plus. Donc si on est habitué à programmer dans d'autres langages, on devra rapidement se familiariser avec la syntaxe de C# et on n'aura qu'à apprendre à placer les accolades et les points-virgules aux bons endroits [25].

4.4 INTERFACE GRAPHIQUE

Dans cette partie, nous présentons la simulation que nous avons développée dans le cadre de notre PFE. L'interface graphique se présente sous forme d'une fenêtre principale à l'aide du logiciel Auto Play, elle est dotée de deux boutons : un bouton de coupure de son, l'autre pour l'ouverture d'une petite fenêtre pour le choix du locuteur ou locutrice (Voix d'homme ou de femme) (Fig.4.1).

Figure 4.1 : interface graphique

Après un clic sur le bouton « démarrer », on devra choisir le type de locuteur (Fig.4.2)

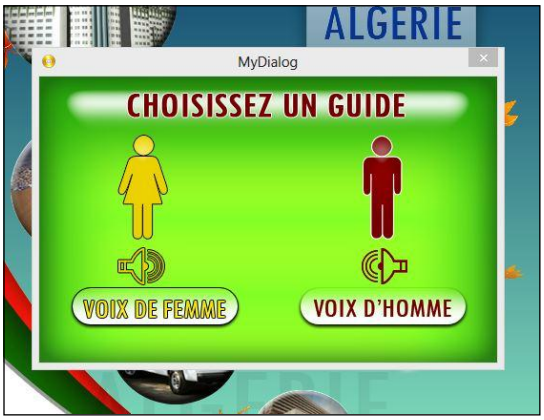


Figure 4.2 : Choix du locuteur

Nous obtenons une fenêtre qui regroupe toutes les phrases en langue Française, un clic sur la phrase, nous donne la dernière figure qui présente la phrase traduite en langage cible (Arabe Standard, Dialecte ou en Amazigh), avec une figure illustrative de la destination ou du besoin (Fig.4.3)

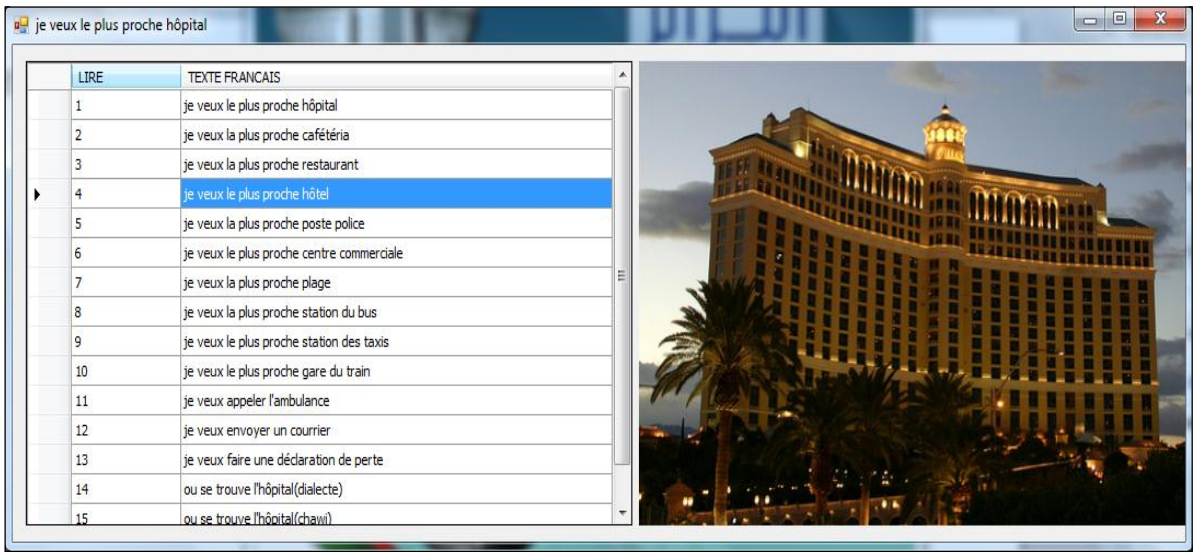


Figure 4.3 : Choix de phrase à prononcer

Après avoir choisi une phrase, nous obtenons sa traduction, et par un clic sur la commande « lire », nous obtenons le signal vocal synthétisé (Fig.4.4).

Figure 4.4 : Phrase traduite avec lancement du signal vocal

Le programme comporte un fichier :

- sound qui contient les segments sonores ;

- image qui contient les images affichées sur l'interface ;
- microsoft Access (base de données) dont laquelle nous pouvons l'implémenter ;
- exécutable (.exe) qui est notre interface.

L'avantage de notre interface est la possibilité d'implémenter d'autres phrases (élargissement du corpus) par l'ajout des segments sonores des parties variables ou fixes au fichier sound puis l'image appropriée au fichier image, avec une définition des liens de ces segments dans le fichier Microsoft Access 'base de données' (Fig.4.5)

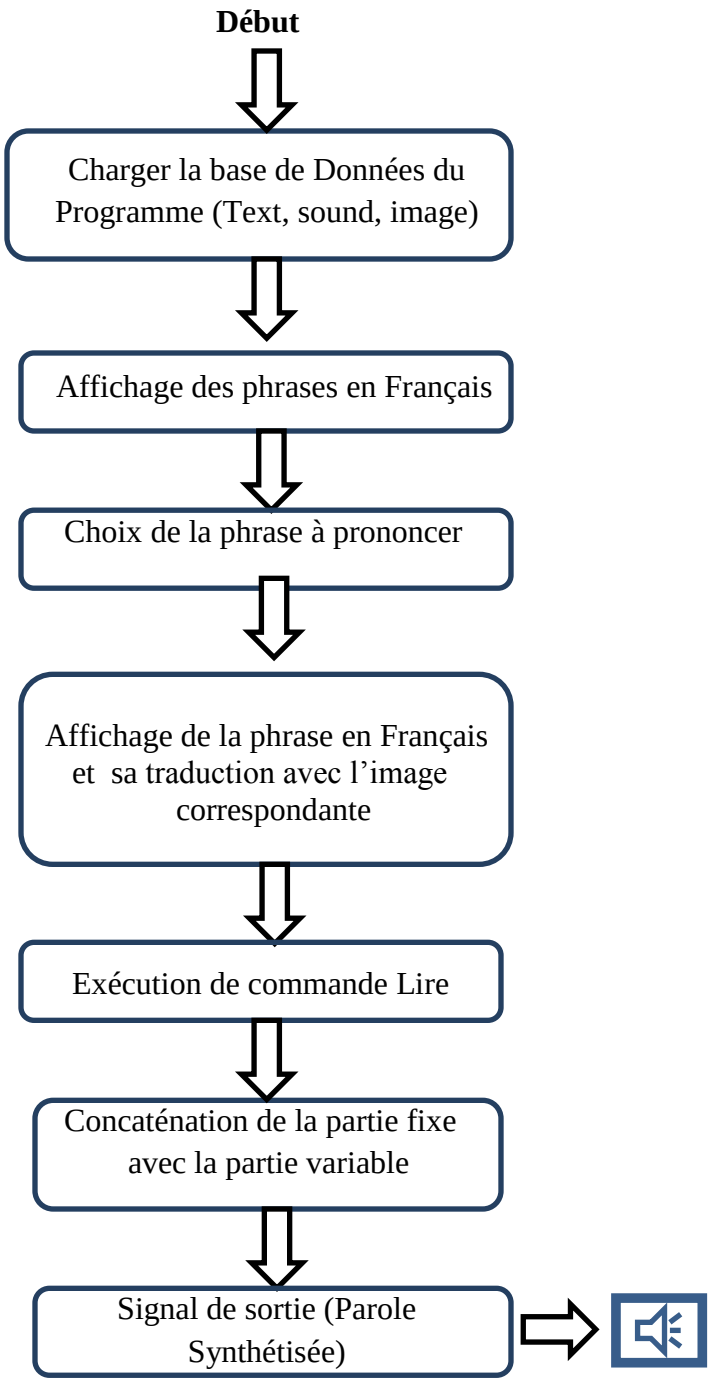
Table1

N°	TEXTE FRANCAIS	TEXTE ARABE	SOUND	IMAGE	SOUND2	SOUND3
1	je veux le plus proche Hôpital	أريد أقرب مستشفى	.\sound\أريد.wav	.\image\مستشفى.jpg	.\sound\أقرب.wav	.\sound\مستشفى.wav
2	je veux la plus proche cafétéria	أريد أقرب مقهى	.\sound\أريد.wav	.\image\مقهى.jpg	.\sound\أقرب.wav	.\sound\مقهى.wav
3	je veux la plus proche restaurant	أريد أقرب مطعم	.\sound\أريد.wav	.\image\مطعم.jpg	.\sound\أقرب.wav	.\sound\مطعم.wav
4	je veux le plus proche Hôtel	أريد أقرب فندق	.\sound\أريد.wav	.\image\2فندق.jpg	.\sound\أقرب.wav	.\sound\فندق.wav
5	je veux la plus proche centrale de police	أريد أقرب مركز شرطة	.\sound\أريد.wav	.\image\شرطة.jpg	.\sound\أقرب.wav	.\sound\مركز_شرطة.wav
6	je veux le plus proche centre commerciale	أريد أقرب مركز تجاري	.\sound\أريد.wav	.\image\مركز تجاري.jpg	.\sound\أقرب.wav	.\sound\مركز_تجاري.wav
7	je veux la plus proche plage	أريد أقرب شاطئ	.\sound\أريد.wav	.\image\شاطئ.jpg	.\sound\أقرب.wav	.\sound\شاطئ.wav
8	je veux la plus proche station du bus	أريد أقرب محطة حافلات	.\sound\أريد.wav	.\image\محطة حافلات1.jpg	.\sound\أقرب.wav	.\sound\محطة_حافلات.wav
9	je veux le plus proche station des taxis	أريد أقرب محطة تاكسي	.\sound\أريد.wav	.\image\محطة تاكسي 1.jpg	.\sound\أقرب.wav	.\sound\محطة_تاكسي.wav
10	je veux le plus proche gare du train	أريد أقرب محطة قطار	.\sound\أريد.wav	.\image\محطة قطارات.jpg	.\sound\أقرب.wav	.\sound\محطة_قطار.wav
11	je veux appeler l'ambulance	أريد أن أتصل بالإسعاف	.\sound\أريد.wav	.\image\سيارة الإسعاف.jpg	.\sound\أن.wav	.\sound\أتصل_بالإسعاف.wav
12	je veux envoyer un courrier	أريد أن أرسل طردا	.\sound\أريد.wav	.\image\طرد.jpg	.\sound\أن.wav	.\sound\أرسل_طردا_.wav
13	je veux déclarer une perte	أريد أن أبلغ عن ضياع	.\sound\أريد.wav	.\image\2أمن.jpg	.\sound\أن.wav	.\sound\أبلغ_عن_ضياع.wav
14	ou se trouve l'Hôpital (dialecte)	وين جاي سبيطار	.\sound\untitled.wav	.\image\مستشفى.jpg	.\sound\untitled.wav	.\sound\وين_جاي_سبيطار.wa
15	ou se trouve l'Hôpital (chawia)	ماتي يطي سبيطار	.\sound\untitled.wav	.\image\مستشفى.jpg	.\sound\untitled.wav	.\sound\chawi.wav
*	(Nouv.)					

Figure 4.5 : fichier Microsoft Access (base de données) de l'interface

4.5 ORGANIGRAMME DE LA SIMULATION

Le début de l'organigramme se présente par l'exécution de l'interface, et en suivant ces étapes jusqu'à l'obtention du signal de parole synthétisée.



4.6 TEST D'EVALUATION

Le test d'évaluation comprend 10 personnes qui ont participé à une session expérimentale d'écoute d'une phrase choisie aléatoirement et répété trois fois successivement. Nous avons considéré cinq niveaux de réponses (Mauvais, Médiocre, Passable, Bon et Excellent).

Les résultats obtenus sont définis dans le Tableau 4.1 et la figure 4.6

Tableau 4.1 : Résultats du test évaluatif

	Décision				
	Mauvais	Médiocre	Passable	Bon	Excellent
Décision /10	0	0	01	04	05
Pourcentage (%)	0	0	10	40	50

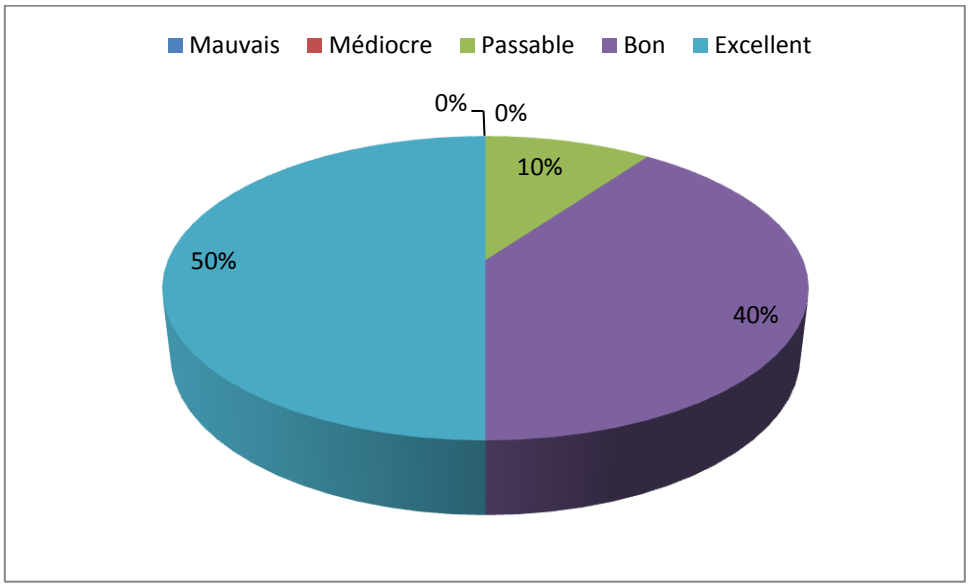


Figure 4.6 : Décision des 10 personnes sur la parole synthétisée

Interprétation

D'après le graphe que nous avons obtenu, les résultats sont de 50 % de décision « Excellent » avec 40 % « bon » et 10 % « passable », ceci montre que l'intelligibilité et le naturel des phrases ou des sons générés par notre système sont satisfaisants, dans la mesure où les auditeurs comprennent bien ce qui a été généré artificiellement.

Les résultats obtenus nous ont été satisfaisants, acceptables et avec une bonne intelligibilité et naturalisme des segments sonores concaténés. De plus, ils nous ont permis de confirmer les résultats du test subjectif qui a été fait dans le chapitre précédent.

4.7 CIRCUIT DU GUIDE TOURISTIQUE PARLANT EN ALGERIE

Nous avons élaboré et réalisé notre circuit au sein du Laboratoire Signal et Communications (LSC) à l'Ecole Nationale Polytechnique.

Le montage réalisé permet la traduction des messages écrits en Français sur un afficheur LCD, en paroles en Arabe Standard, dialectale et en Amazigh. Les messages en langue Française qui sont affichés sur l'écran LCD, sont stockés dans la mémoire programme du microcontrôleur, par contre les messages oraux en Arabe sont enregistrés dans la mémoire vocale.

La mémoire vocale peut stocker jusqu'à 60 secondes de parole, donc afin d'enregistrer plusieurs messages, la zone mémoire est partagée en 6 segments de 10 secondes. Nous avons la possibilité d'enregistrer 6 phrases de durée maximale de 10 secondes. L'adressage est effectué par le bus d'adresse de la mémoire vocale. Les voix et les signaux audio sont stockés directement dans la mémoire sous leur forme normale, fournissant la reproduction d'une parole de haute qualité.

Le pilotage de l'afficheur LCD, ainsi que la lecture des états des boutons de marche, arrêt et enregistrement est assuré par l'afficheur LCD (Fig.4.7).

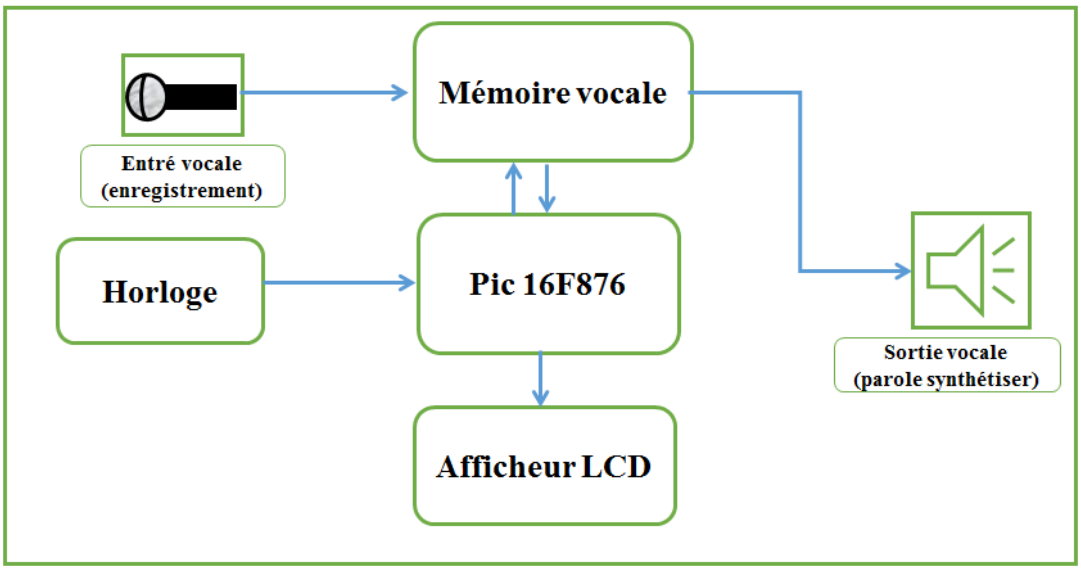


Figure 4.7 : Schéma bloc du circuit du Guide Touristique Parlant

Les composants que nous avons utilisés pour sa réalisation sont définis dans le Tableau 4.2

Tableau 4.2 : Les composants électroniques du Guide Touristique Parlant

Composant	Description
Afficheur LCD GDM1602A [26]	• Alimenter avec un +5V ; • 5*8 points avec le curseur ; • Contrôleur intégré (KS0066U ou équivalent) ; • BKL à conduire par pin1, 2, 15,16 ou A, K 6 ; • Contre-jour de LED
Mémoire vocale ISD2560 [27]	• D'une durée de 60 secondes • l'alimentation d'énergie de +5 volts ; • Entièrement accessible pour manipuler des messages multiples ; • 1 µA du courant de réserve (typique) ; • conservation du message pendant 100 ans;
Microcontrôleur PIC16F876 [28]	• Seulement 35 instructions de mots simples à apprendre • Modes d'adressage direct, indirect et relatif • Puissance faible, technologie à grande vitesse de CMOS FLASH/EEPROM • Températures ambiantes, industrielles et prolongées • Consommation de basse puissance : • < 0.6 mA @ 3V typique, 4 MHz • 20 µA @ 3V typique, 32 kHz • < 1 µA de courant • Possibilités d'interruption (jusqu'à 14 sources)
9 Résistances 4 capacités	
3 Boutons poussoirs	• Message (MSG) suivant • enregistré MSG • lire MSG
Alimentation	• 12 V
Microphone Haut-Parleur	
Horloge	• 4 MHz

Pour bien illustrer le schéma électrique du circuit complété du GTPA (Figure 4.8).

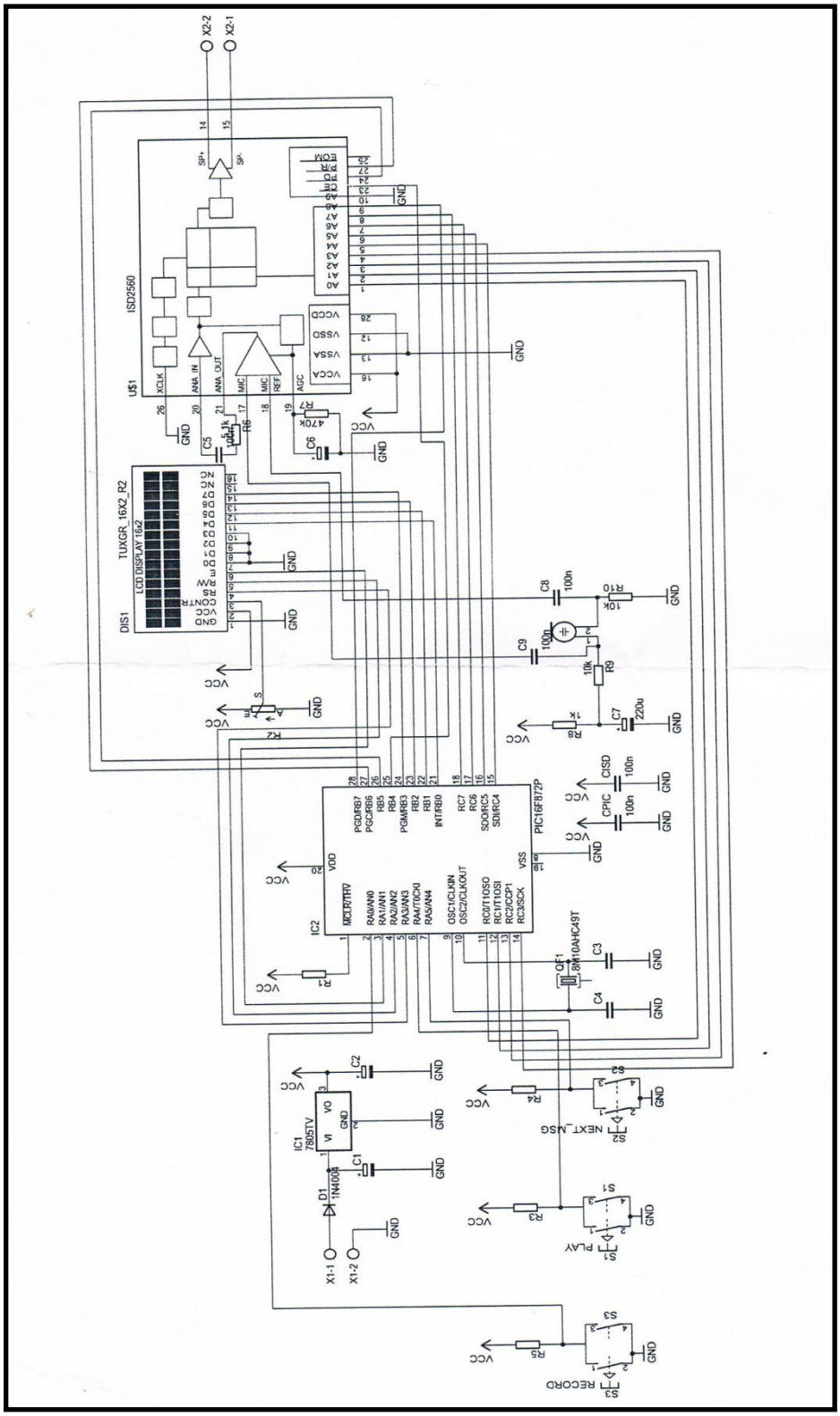


Figure 4.8 : Schéma électrique du Guide Touristique Parlant

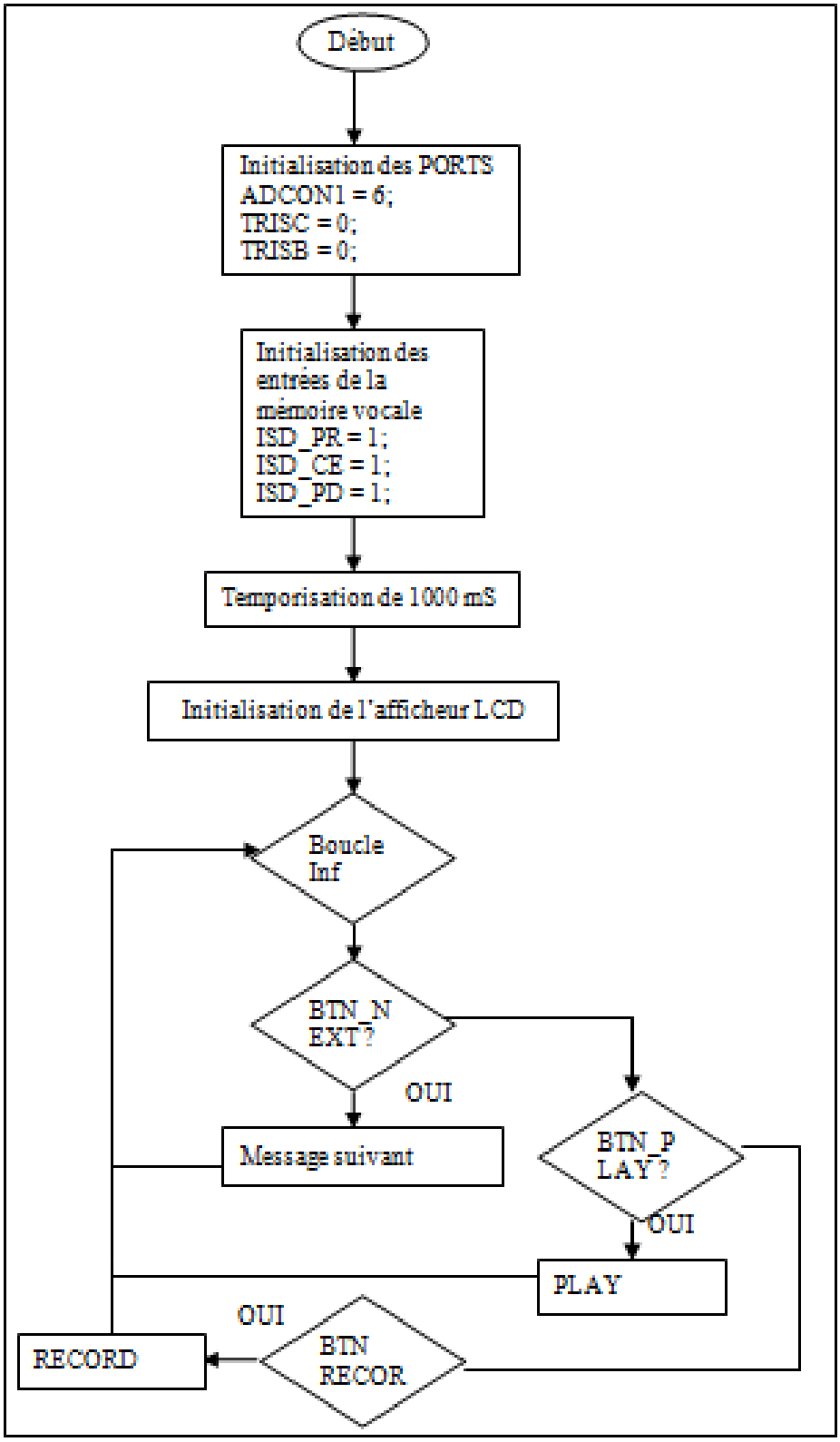


Figure 4.9 : organigramme du circuit électronique

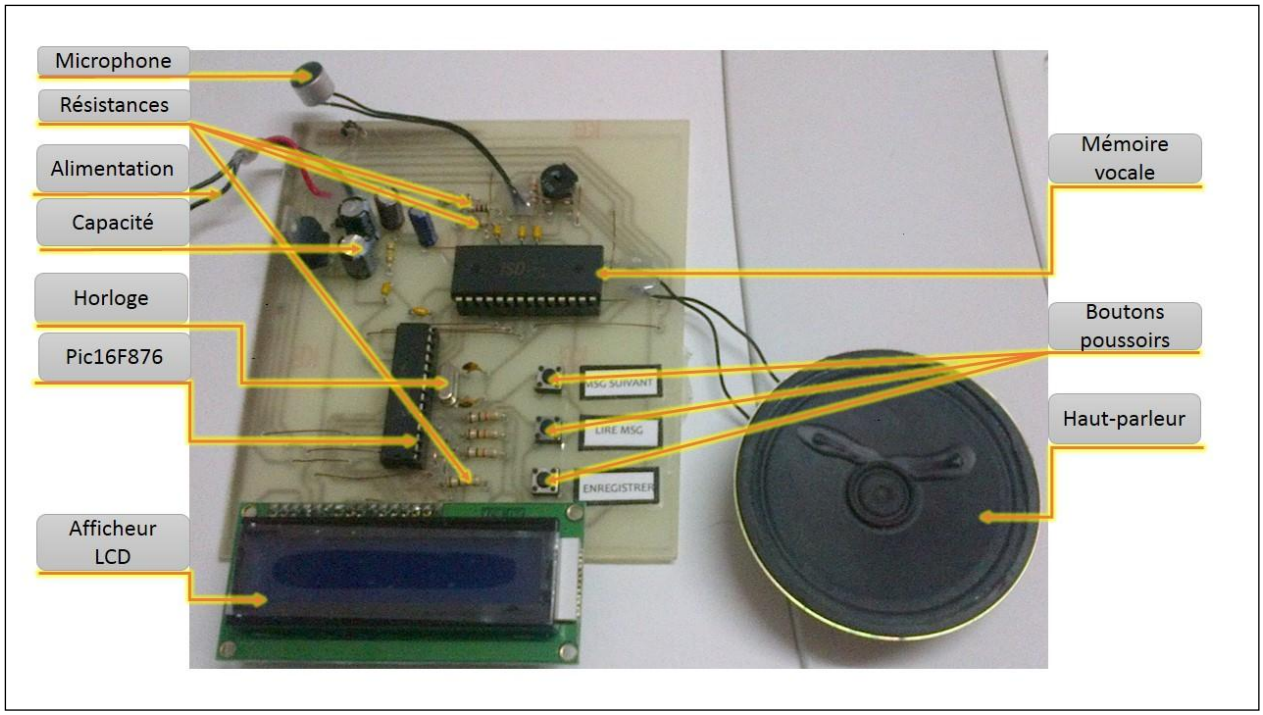


Figure 4.10 : Guide Touristique Parlant réalisé

4.8 CONCLUSION

Nous avons exposé dans ce chapitre un programme de simulation de la synthèse de parole par concaténation de phrases et mots ainsi que le langage de programmation C # sous l'environnement Visuel Studio, le circuit du Guide Touristique Parlant en Algérie et sa description électronique. Nos résultats ont été obtenus par un test de perception selon 10 personnes.

Les résultats (compréhension et l'intelligibilité des phrases prononcées) obtenus ont affirmé le bon enregistrement du corpus GTPA ainsi que la segmentation manuelle.

Conclusions générales et perspectives

Conclusions générales et perspectives

L'objectif de notre travail tout au long de ce **PFE** était la réalisation d'un système de synthèse de parole par unités variables (phrases et mots combinés), en vue d'obtenir un Guide Touristique Parlant en Algérie (GTPA). Ce travail est muni d'un programme de simulation et d'une application électronique du Guide.

Nous avons tout d'abord effectué des études générales sur la parole puis sur l'Arabe Standard, pour cela nous avons choisi la synthèse par concaténation des unités pré - stockés, puis un enregistrement du corpus GTPA qui a été fait par deux locuteurs arabophones (Femme et Homme), ce dernier représente la base de données de notre travail. L'étude du GTPA passe par plusieurs étapes d'analyse acoustique et de visualisations, qui nous ont permis une extraction des paramètres pertinents et acoustiques du signal vocal. Des tests de perception objective et subjective ont été effectués sur le signal original et le synthétisé.

En dernier lieu, les tests et les résultats correspondants étaient satisfaisants. Ils nous ont confirmé une bonne segmentation manuelle et un bon enregistrement du GTPA.

La synthèse par concaténation des formes d'ondes préenregistrées est capable de produire des annonces vocales de haute qualité se rapprochant du naturel.

La synthèse de parole présente plusieurs avantages, elle est d'une part plus naturelle pour le grand public, et d'autre part plus rapide et efficace qu'un court message écrit ainsi, que le champ de vision qui reste libre pour effectuer une autre tâche de lecture.

La qualité d'un système de synthèse vocale dépend : de son aspect naturel, de l'intelligibilité de la parole générée, des caractéristiques propres à la voix produite qui dépendent : des techniques, des méthodes de synthèse appliquées, et également du soin apporté à la modélisation linguistique et prosodique.

Comme perspective à ce travail, il sera très intéressant de faire une étude évaluative pour améliorer la qualité de la parole synthétisée, afin d'obtenir cette amélioration, nous proposons :

- un élargissement du vocabulaire du corpus GTPA ;
- l'utilisation d'autres langues (l'Anglais, par exemple) ;

- une intervention de la part des experts linguistes afin de mieux modéliser les classes des mots et phrases ;
- le choix des centres d'intérêts et des itinéraires doivent être définis par les professionnels du tourisme ;
- l'amélioration de la qualité de la voix synthétisée, avec une technique d'évaluation par ajustement des paramètres prosodiques du signal vocal, afin d'avoir une bonne qualité qui se rapproche du naturel ;
- l'amélioration de l'algorithme du programme (ajout des boutons de direction,… etc.) ;
- développement du programme de simulation pour des applications Smartphones et iPod.

Références Bibliographiques

REFERENCES BIBLIOGRAPHIQUES

[1] T. Dutoit, Introduction au traitement automatique de la parole notes de cours /DEC2, Faculté Polytechnique de Mons, LCTS Lab, France, 2000.

[2] http : //www.infovisual.info.

[3] J. Cisonni, Modélisation et inversion d'un système complexe de production de signaux acoustiques Application à la voix et aux pathologies, Thèse de Doctorat, Institut Polytechnique De Grenoble, France, Novembre 2008.

[4] http://users.skynet.be/illusionsauditives/images/page802.gif

[5] L'étude instrumentale des gestes dans la production de la parole ; importance de l'aérophonométrie ; Manuscrit auteur, publié dans "Les Dysarthries, P. Auzou (Ed.) (2007) 115-117

[6] D. Ducassou. Cours d'acoustique. Cours de 2ème année de médecine, Université de Nancy 1, France, 1991.

[7] J. Clarenc, les caractéristiques articulatoires et acoustiques des sons, analyse des structures phoniques (niveau segmental), Parcours fle (2006-2007).

[8] A. Ounnas, Synthèse de la parole en Arabe Standard, Mémoire de Magister, ENP, Alger, Algérie, 2011.

[9] M. Aissiou, Application des Algorithmes Génétiques au Décodage Acoustico- Phonétique de la parole en Arabe Standard, Thèse de Doctorat, ENP, Alger, 2008.

[10] V A. Dubesset, La Langue française Parlée Complétée (LPC) : Production et Perception, Thèse de Doctorat, Institut National Polytechnique De Grenoble, France, 2005.

[11] Calliope, La parole et son traitement automatique, Collection Techniques et Scientifiques des Télécommunications. Préface de G. Fant, CNET/ENST, Ed. Masson, 1989.

[12] M. Mouslem, Identification du locuteur indépendante du texte, Mémoire de Magister, ENP, Alger, Algérie.

[13] H. Lounis ; Analyse sonagraphique des consonnes fricatives [s] et [ŝ] et leurs opposées [z] et [ž] en vue de la RAP en Arabe Standard ; Projet de Fin d'Etudes ; Ecole Nationale Polytechnique -Alger ; 24 Juin 2007.

[14] A.Chentir ; Etude de la Microprosodie en vue de la Synthèse de la parole en Arabe Standard, Thèse de Doctorat, Ecole Nationale Polytechnique -Alger (Algérie), 01 Octobre 2009.

[15] C. Alessandro, La synthèse de la voix, ETR6 ,16-07-2002.

[16] G. Richard, O. Cappe, synthèse de la parole à partir du texte, techniques de l'ingénieur, Vol. H7 288, p.2.

[17] S. Djeghiour, Mémoire de Magister : Application des Réseaux de Neurones à la synthèse de la Parole En Arabe Standard, Ecole Nationale Supérieure Des Sciences Humaines, Bouzaréah-Alger (Algérie), 2011.

[18] D. Cadic ; Optimisation du Procédé de Création de Voix en Synthèse par Sélection ; Thèse de Doctorat ; Université de Paris Sud 11, Faculté des Sciences D'orsay, France, 10 Juin 2011.

[19] E. Moulines, O. Cappé, synthèse de la parole à partir du texte, Techniques de l'ingénieur, Vol. H1 960, p.6-7.

[20] S. Rossato, H. Blanchon & L. Besacier, Évaluation du premier démonstrateur de traduction de parole dans le cadre du projet NESPOLE, TALN 2002, Nancy, France, 24-27 juin 2002.

[21] J-P Goldman, Tutoriel Praat, Décembre 2006.

[22] E. Delais-Roussarie & G. Caelen-Haumont, Outils d'aide à l'annotation prosodique de corpus, Manuscrit auteur, publié dans "Bulletin PFC (Phonologie du Français Contemporain), N° 6 (2006) 7-26", hal-00256395, version 1 - 15 Février 2008.

[23] http://www.praat.org/.

[24] R. Boite & M. Kunt, Traitement de la parole, complément au traité d'électricité, Presses Polytechniques Romandes (Suisse), première édition, 1987

[25] http://www.programmez.com/livres/9-Sharp-Ch01.pdf.

[26] http://users.ece.gatech.edu/~hamblen/UP3/GDM1602A.pdf.

[27] Inbond Electronic's corp**,** ISD2560/75/90/120; Publication Release Date: May 2003; Revision 1.0.

[28] PIC16F87X, Data Sheet, 28/40-Pin 8-Bit CMOS FLASH, Microcontrollers, 2001 Microchip Technology Inc.