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 :

Annonces Vocales Automatiques des Stations d'Arrêt du Tramway d'Alger

Réalisé par :

Mr **BOUMAZA SOUFYANE**

Proposé et Dirigé par :

M. GUERTI Professeur ENP

Promotion : Juin 2013

Dédicaces

Je dédie ce travail à :

- *Mes parents qui m'ont soutenu, orienté, aidé et encouragé le long de ma vie.*
- *mes frères et sœurs avec qui j'ai passé les plus beaux moments de ma vie.*
- *tous mes amis : Amine, Sid Ali, Mohamed El Amine, Moussaab, Raouf, Houari, Abdelkader, Ismail*
- *A tous ceux pour qui je compte et à tous ceux qui comptent pour moi.*

Soufyane

Remerciements

Tout d'abord je remercie Dieu de m'avoir donné la force et le courage d'accomplir ce travail.

Je remercie vivement ma promotrice Professeur GUERTI Mhania pour m'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. Je la remercie pour ses compétences, son ouverture d'esprit et sa grande disponibilité.

J'exprimer ma profonde gratitude à Monsieur Boualem BOUSSEKSOU, chargé de cours à l'ENP, pour l'honneur qu'il me fait en acceptant de présider le jury de ce PFE.

Je remercie vivement Monsieur Llies SAADAOUI, Chargé de Cours à l'ENP, pour m'avoir fait le plaisir d'accepter de participer à mon jury en tant qu'examinateur.

Mes remerciements vont également à tous mes enseignants à l'Ecole Nationale Polytechnique qui ont contribué à ma formation.

Je remercie tous ceux, qui de près ou de loin, m'ont apporté leur contribution pour la réalisation de ce travail.

ملخص

الهدف من عملنا هذا هو إ نشاء نظام الإعلانات الصوتية تلقائية لمحطات التوقف تراموای الجزائر بالعربية الفصحى والفرنسية. ولهذا قمنا بدراسة أسلوب تركيب الكلام بالتسلسل الموجي، واهتممنا في دراستنا هذه بنوعية الكلام المركب من حيث الوضوح وطبيعة الكلام.من أجل ذلك قمنا بتسجيل مدونة صوتية بتقنيات عالية وفي شروط ملائمة ثم أجرينا عليها مجموعة من التحاليل الصوتية الأكوستيكية للتأكد من نوعية الكلام المركب، وقمنا أيضا باختبار ذاتي لعشرين شخصا للتأكد من وضوح وطبيعية الكلام المصطنع.وفي الأخير قمنا بمحاكاة مشروعنا هذا ببرنامج visuel studio وفق لغة البرمجة C#

كلمات المفاتيح: تركيب الكلام، تقنية التسلسل الموجي، اللغة العربية الفصحى، اللغة الفرنسية، التحليل الصوتي، الاختبار الذاتي.

Résumé

Le but de notre travail est d'élaborer un **S**ystème des **A**nnonces **V**ocales Automatiques des **S**tations d'arrêt du **T**ramway d'**A**lger (**SAVSTA**), en Arabe Standard et en Français. Pour cela, nous avons étudié les techniques de synthèse de la parole par concaténation en forme d'ondes. Nous nous sommes intéressés à l'étude de la qualité de la parole en vue de l'intelligibilité et de son aspect naturel. Pour ce faire, nous avons élaboré un corpus avec des bonnes conditions d'enregistrements, et nous avons fait des analyses acoustiques comparatives pour confirmer la bonne qualité de la parole synthétique. Le travail a été évalué à l'aide des tests subjectifs par 20 personnes. Nous avons fait une simulation de SAVSTA avec une interface graphique. Cette simulation a été réalisé par l'environnement visuel studio avec un langage de programmation C#.

Mots clés : Synthèse de la Parole, Méthode de concaténation, Arabe Standard, Français, Corpus, Analyse de la parole, Tests d'évaluations.

Abstract

The objective of our work is to develop ads **V**oice **S**ystem of **A**rrest **S**tations **T**ram **A**lgiers (**VSASTA**, in Standard Arabic and French languages. For this, we studied speech synthesis technique's by concatenation waveform. We are interested in the intelligibility and naturalness of speech quality. So, we have developed an VSASTA_corpus with good condition of records, and we made comparative acoustic analysis to confirm the quality of synthetic speech. The work was evaluated using subjective tests by 20 people. We made a simulation VSASTA with a graphical interface. Witch was programmed by C# language carried out by the visual studio environment.

Keywords : Synthesis of the Word Method concatenation, Standard Arabic, French, Corpus, Speech analysis, evaluation tests.

Liste des Abréviations

SP	**S**ynthèse de la **P**arole
TAP	**T**raitement **A**utomatique de la **P**arole
TOP	**T**ranscription **O**rthographique de la **P**arole
LP	**L**inear **p**rediction
LPC	**L**inear **P**redictive **C**oding
TTS	**T**ext-**T**o-**S**peech
CT	**C**ourt-**T**erme
F0	**F**réquence Fondamentale
Fe	**F**réquence d'**é**chantillonnage
FT	**F**onction de **T**ransfert
API	**A**lphabet **P**honétique **I**nternational
PSOLA	**P**itch **S**ynchronous **O**ver**l**ap and **A**dd
TFD	**T**ransformée de **F**ourier **D**iscrète
TF	**T**ransformée de **F**ourier
FFT	**F**aste **F**ourier **T**ransform
TFI	**T**ransformée de **F**ourier **I**nverse
TPZ	**T**aux de **P**assage par **Z**éro
TD-PSOLA	**T**ime **D**omain **PSOLA**
AR	**A**uto**R**égressif
ARMA	**A**uto **R**égressif à **M**oyenne **A**justée
AS	**A**rabe **S**tandard
MA	**M**oyenne **A**justée
MFCCs	**M**el scaled **F**requency **C**epstral **C**oefficients
TFCT	**T**ransformée de **F**ourier à **C**ourt **T**erme
AVAST	**A**nnonces **V**ocales **A**utomatiques des **S**tations d'Arrêt du **T**ramway d'Alger
SAVSTA	**S**ystème d'**A**nnonces **V**ocales Automatiques des **S**tations d'Arrêt du **T**ramway d'**A**lger
RAP	**R**econnaissance **A**utomatique de la **P**arole

Liste des tableaux

Table des Figures

Sommaire

CHAPITRE 2 : ANALYSE ET SYNTHESE DE LA PAROLE

CHAPITRE 3 : ANALYSE ACOUSTIQUE DU CORPUS

CHAPITRE4 : ANNONCES VOCALES DES STATIONS D'ARRET DU TRAMWAY D'ALGER

INTRODUCTION GENERALE

Le traitement de la parole est un vaste domaine de recherche qui demande l'intervention des experts de plusieurs spécialités. Malgré le développement remarquable des outils et les programmes informatiques, les systèmes à commandes vocales n'ont eu du succès que ces dernières années.

Le **T**raitement **A**utomatique de la **P**arole (**TAP**) est une composante fondamentale des sciences de l'ingénieur et un domaine de recherche actif, au croisement du traitement du signal numérique et du traitement symbolique du langage. Depuis les années 60, le TAP bénéficie d'efforts de recherche très importants, liés au développement des moyens et 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, 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.

De nos jours, la synthèse vocale n'est plus un concept avant-gardiste mais 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 et les annonces vocales. 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 la parole tout à fait intelligible, les intonations et l'expressivité ne sont pas encore au point.

L'objectif de notre travail est de réaliser un **S**ystème parlant qui fait les **A**nnonces **V**ocales Automatiques des **S**tations d'Arrêt du **T**ramway d'**A**lger (**SAVSTA**). Ces annonces seront faites pour indiquer les stations prochaines avant d'y arriver. Notre système va se déclencher automatiquement par un lancement de signal vocal qui prononce le nom de cette station, et en parallèle afficher le nom de la station concernée en Arabe Standard et en Français.

Nous avons structuré notre travail en quatre chapitres :

- dans le premier, nous avons décrit de manière générale des notions sur le traitement de la parole, des spécifications du signal vocal et des notions fondamentales sur les langues Française et Arabe Standard ;
- le deuxième présente les principes de la synthèse vocale, suivi d'une description bien détaillée des différentes techniques et méthodes utilisées ;
- dans le troisième chapitre, nous avons fait l'analyse acoustique des sons pour étudier les caractéristiques (formants, fréquence fondamentale, intensité) de notre corpus que nous avons nommé : **A**nnones **V**ocales **A**utomatiques des **S**tations d'Arrêt du **T**ramway d'Alger (**AVAST).** Ensuite, nous exposons les grandes lignes introduites dans les étapes de l'élaboration de notre outil d'analyse : le logiciel Praat ;
- dans le dernier chapitre, nous avons présenté une simulation du **S**ystème d'**A**nnonces **V**ocales Automatiques des **S**tations d'arrêt du **T**ramway d'**A**lger (**SAVSTA**), avec une interface graphique dans l'environnement visuel studio, et finissons par des tests 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 terminons notre travail par des conclusions générales et perspectives.

Chapitre I :

Généralités sur la parole

1.1 INTRODUCTION

Dans ce chapitre nous allons décrire de manière générale des notions sur le traitement de la parole, des spécifications du signal vocal et des notions fondamentales sur le la langue Française et l'Arabe Standard.

1.2 GENERALITES SUR LA PAROLE

L'importance particulière du traitement de la parole s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine. L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau Humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

La parole est une faculté, propre à l'Homme, de communication par des sons articulés. Elle met en jeu des phénomènes de natures très différentes et peut être analysée de bien des façons. On distingue généralement plusieurs niveaux de description non exclusifs : physiologique, phonologique, phonétique, acoustique, morphologique, syntaxique, sémantique, et pragmatique **[1]**.

1.3 L'APPAREIL PHONATOIRE HUMAIN

La production de la parole est assurée, chez l'Homme, par plusieurs organes successifs. Les poumons sont indispensables dans ce processus puisqu'ils assurent la génération d'un composant incontournable : de l'air sous pression. Cet air, expulsé, traverse alors les cordes vocales qui entrent ou non en action pour produire un voisement. Ce voisement correspond à la fréquence fondamentale qui est le timbre de la voix. Cette fréquence fondamentale étant produite, elle est propagée dans l'ensemble du conduit vocal. Ce conduit est de forme et de volume variable. Plusieurs organes concourent à ces possibles modifications qui permettent de produire des sons différents. Parmi ces organes se trouve la langue, acteur principal des modifications qui peut agir par constriction ou occlusion du conduit vocal [2].

Les dents et les lèvres agissent également par occlusion ou constriction, à des degrés cependant moindres. Le conduit vocal est, la plupart du temps, constitué du seul conduit buccal. La luette et son prolongement vers le palais, le vélum, assurent normalement la

fermeture du conduit nasal pendant la production de parole. Le conduit nasal peut, dans certains cas, être connecté au conduit vocal. Cette connexion permet de générer des sons supplémentaires en modifiant le volume de la caisse de résonance normalement constituée par le seul conduit buccal. Une coupe de l'appareil phonatoire Humain est fourni en (figure 1.1).

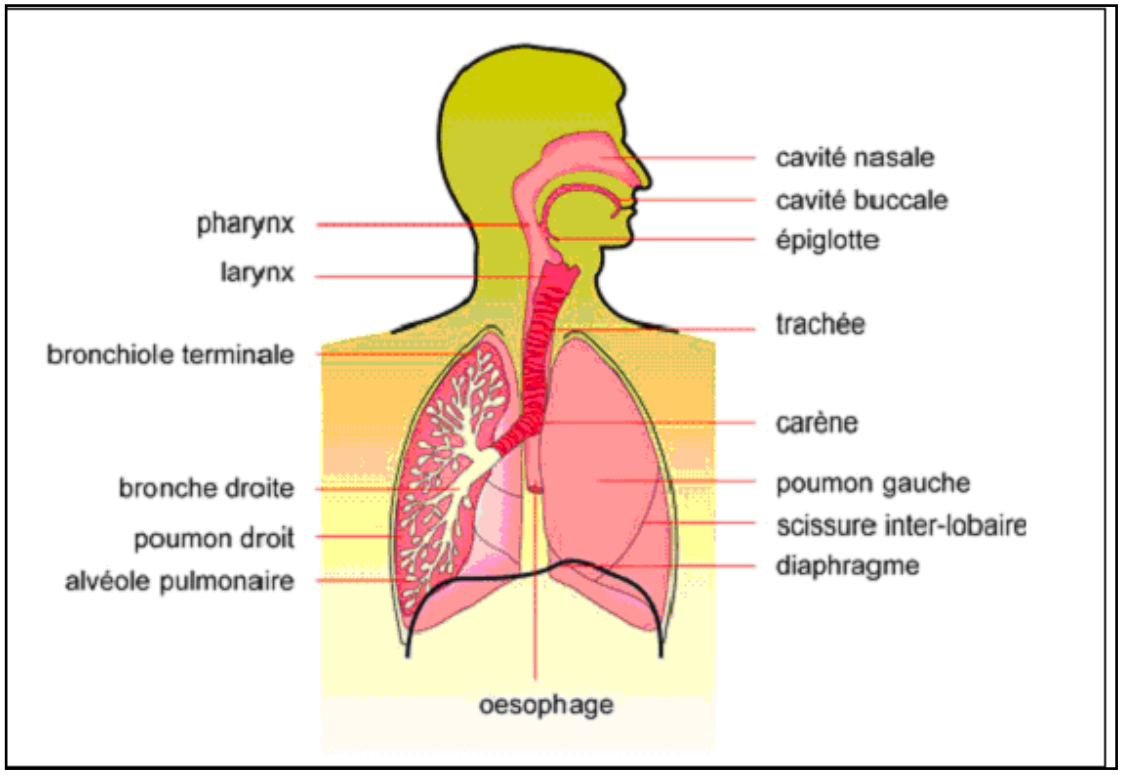


Figure 1.1 : Coupe sagittale de l'appareil phonatoire Humain [3].

1.3.1 Fonctionnement de l'appareil phonatoire Humain

Comprendre le mécanisme de production de la parole est un aspect d'une grande importance. En effet, c'est l'étude du système de phonation qui va nous permettre d'identifier et de caractériser les grandes classes de sons élémentaires et d'expliquer les variations de ces derniers dans les différents contextes. De plus, les algorithmes de para métrisation du signal vocal sont obtenus à partir de modèles du conduit vocal.

Les traits acoustiques du signal de parole sont évidemment liés à sa production. L'intensité du son dépend de la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que celle du rythme d'ouverture/fermeture des cordes vocales, induit par la tension de muscles qui les contrôlent. Son spectre résulte du filtrage du signal glottique

(impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses [4].

La parole est articulée en interrompant et en modulant le flux d'air à l'aide des lèvres, de la langue, des dents, de la mâchoire inférieure et du palais (Figure .1.2).

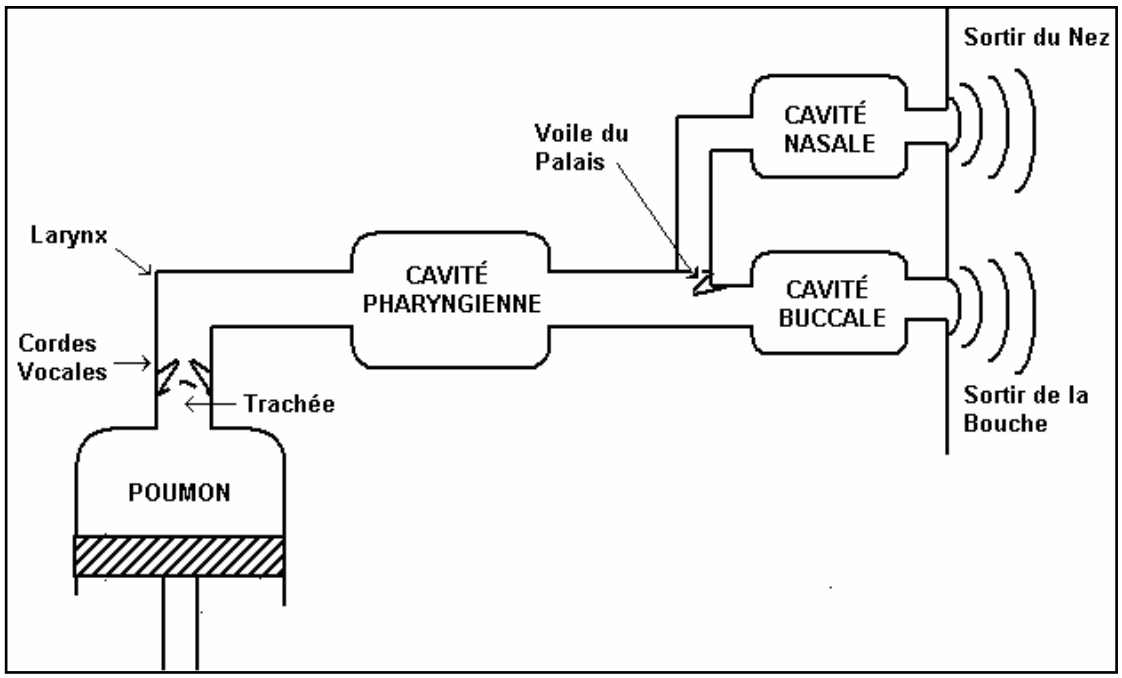


Figure 1.2 : Modèle mécanique de production de la parole [5].

1.3.2 Production de l'onde glottique

L'air produit par excès de pression dans les poumons rencontre un premier obstacle qui sont les cordes vocales (source d'excitation). Ces dernières accolées, sous l'effet de la pression sub-glottique se mettent à vibrer laissant passer l'air par impulsions. C'est ainsi que se forme l'onde glottique dont la fréquence d'oscillations notée F_0 (fréquence fondamentale ou pitch), est déterminée par la masse et la tension des cordes vocales ainsi que la pression sub-glottique. Quand elles vibrent, il y a émissions de sons dits sons voisés ou sonores par opposition aux sons non voisés ou sourds qui sont assimilables à un bruit blanc [4].

1.3.3 Fonction résonateur du conduit vocal

Le conduit vocal imprime au son émis les caractéristiques spécifiques permettant de distinguer les différents phonèmes et ceci selon deux fonctions, en tant que :

- résonateur de l'onde glottique pour la production des phonèmes sonores ;

- générateur de bruit pour la production des consonnes sourdes.

En effet, l'onde glottique est modifiée lors de son passage à travers le conduit vocal. Les positions de la mâchoire et de la langue déterminent les cavités qui jouent le rôle de caisses de résonance en renforçant certaines régions du spectre acoustique. Les maxima de la courbe de réponse en fréquences du conduit vocal sont appelés formants [4].

1.3.4 Fonction générateur de bruit du conduit vocal

Le flux d'air créé peut rencontrer soit, un obstacle partiel tel un rétrécissement du conduit vocal pour générer un bruit caractéristique des sons fricatifs ou constrictifs, soit un obstacle total produisant une augmentation de la pression en amont de l'obstacle (lieu d'articulation) suivi d'un relâchement brusque. Ce phénomène engendre la formation des sons occlusifs [4].

1.4 ARTICULATIONS COMPLEXES

Les vibrations des cordes vocales ne suffisent pas à produire un son intelligible, tout un système articulatoire en aval, assure la propagation de l'air vibrant ou non vibrant. Nous pouvons citer l'épiglotte, la luette, etc.

1.4.1 L'épiglotte

C'est une structure cartilagineuse reliée au larynx qui coulisse vers le haut quand les voies aériennes sont ouvertes, Elle aide à obstruer l'entrée de la trachée au moment de la déglutition. Elle descend légèrement vers le bas, afin d'entrer en contact avec le larynx qui s'élève, formant ainsi un verrou au-dessus du larynx. Il se peut que de temps à autre, lorsqu'on mange trop vite, des aliments liquides ou solides ingérés pénètrent dans le larynx avant que l'épiglotte n'ait pu se rabattre sur celui-ci. De tels cas peuvent s'avérer très dangereux du fait que les voies respiratoires peuvent se boucher et empêcher l'air de pénétrer dans les poumons.

1.4.2 La luette

La luette ou uvule est une saillie allongée mobile qui termine le voile du palais et qui contribue, lorsqu'elle se détache de la paroi pharyngale, à permettre à l'air provenant des poumons et du larynx de se diriger non seulement vers la bouche, mais également vers les fosses nasales. Lorsque la luette s'appuie sur la paroi pharyngale, elle empêche l'air de

pénétrer dans les fosses nasales et ne le laisse s'échapper que par la bouche (articulations orales) [3].

1.4.3 La langue

La langue est une masse musculaire divisée en trois parties :

- la pointe (apex) qui sert d'articulateur pour les articulations apicales, le dos pour les articulations pré médio ou post-dorsales, et la racine dans le cas des articulations radicales. Elle constitue l'articulateur principal des différents sons (figure 1.3).
- La langue permet le blocage d'air venant des poumons pour produire les consonnes occlusives, le resserrement de la cavité buccale inhérent à la production des consonnes constrictives, lorsqu'elle demeure suffisamment éloignée de la voûte du palais, elle permet la réalisation des différentes voyelles [3].

1.4.4 Les lèvres

Ce sont les parties charnues qui bordent extérieurement la bouche. Elles s'amincissent pour se joindre aux commissures. La lèvre supérieure est limitée par le nez, alors que la lèvre inférieure est limitée par le sillon mentonnier. Lorsqu'elles sont projetées et arrondies, les lèvres forment une cavité qui sert de résonateur lors de la réalisation des voyelles arrondies et des consonnes labialisées. En revanche, lorsque les lèvres sont rétractées, les voyelles sont non arrondies et les consonnes non labialisées. La lèvre supérieure peut également agir comme lieu d'articulation par exemple, alors que la lèvre inférieure peut agir comme articulateur pour les consonnes labialisées [3].

1.4.5 Les dents

Les dents bien que pas très coopératives à la phonation, leur absence rend le système phonatoire mécaniquement défectueux, en atrophiant les ouvertures des lèvres et la prononciation des labio-dentales. Chacun des ces organes est à la base de la production d'un son élémentaire appelé phonème. Ce dernier est la contribution distribuée du système phonatoire. La participation de chaque intervenant dépend de la langue prononcée. Les nasalisations, les roulements des [r]…. lorsqu'elles sont exagérées sont à la base des défauts langagiers et sont considérées comme pathologies nécessitant un traitement de réapprentissage de la prononciation. La figure 1.1 illustre l'appareil phonatoire humain.

1.5 SONS VOISES ET NON VOISEES

Quand le conduit vocal est excité par des impulsions périodiques de pression, résultantes des oscillations des cordes vocales, la pression accumulée puis libérée inopinément par l'ouverture de la glotte, crée des sons appelés voisé, ce sont des sons qui constituent entre autres les voyelles. Le spectre d'un tel son est esquissé à la figure 1.3, il est caractérisé par des pics épars, le premier correspond à la fréquence fondamentale, les autres à des fréquences appelées formants. Les trois premiers formants sont nécessaires pour décrire un spectre vocal, les formants d'ordres supérieurs ont des applications diverses telles que la reconnaissance de voix.

Le resserrement du conduit vocal entraine des sons semblables à des consonnes, en plus de ce resserrement, les cordes vocales n'entrent pas en vibrations, elles restent écartées, c'est la raison pour laquelle ces sons sont apériodiques, ils sont généralement assimilés à un bruit blanc à la sortie d'un filtre constitué par la partie du conduit vocal sise entre la constriction et les lèvres.

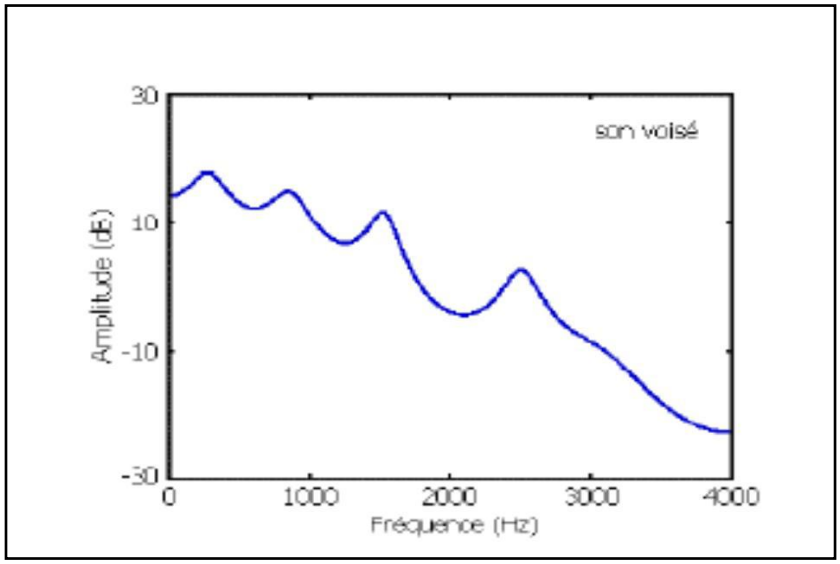


Figure 1.3 : Le spectre d'un son voisé [i] **[6]**

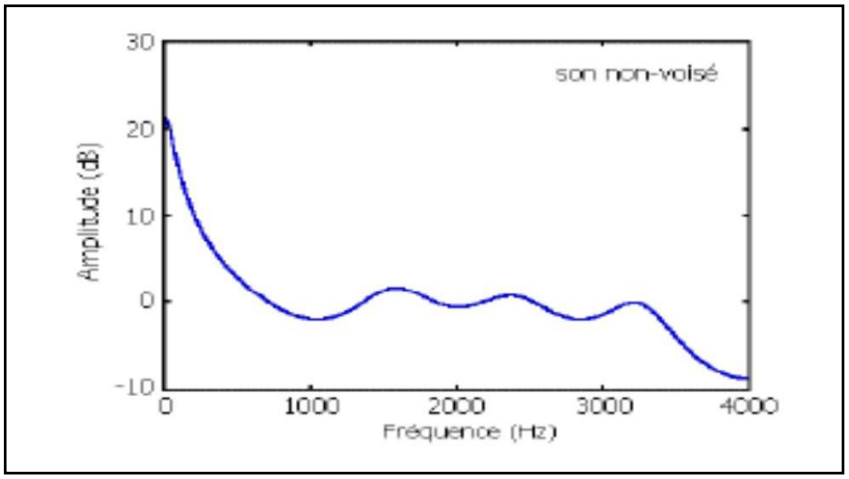


Figure 1.4 : Le spectre d'un son non voisé [p] **[6]**

Le spectre d'un son non voisé ne possède pas de pitch à la différence du voisé, mais il garde quelques attributs de ce dernier, en ce qui concerne sa structure formantique. Le spectre d'un tel son est donné par la figure 1.4 [6].

1.6 Caractéristiques phonétiques

La production des sons ou d'un mot réside dans la production en série de tous les phonèmes constituant ce mot .Un phonème est la plus petite unité présente dans la parole. Le nombre de phonèmes est toujours très limité (normalement inférieur à cinquante) et ça dépend de chaque langue. Les phonèmes peuvent être classés en fonction de trois variables essentielles : le voisement (activité des cordes vocales), le mode d'articulation (type de mécanisme de production) et le lieu d'articulation (endroit de resserrement maximal du conduit vocal) [5].

Table 1.1 : Les phonèmes du Français [7]

TABLEAU I. — *Les phonèmes du français*

Consonnes			
	[p] paie	[t] taie	[k] quai
	[b] baie	[d] dais	[g] gai
	[m] mais	[n] nez	[ɲ] gagner
	[f] fait	[s] sait	[ʃ] chez
	[v] vais	[z] zéro	[ʒ] geai
	[w] ouais	[ɥ] huer	[j] yéyé
		[l] lait	[R] raie
Voyelles			
	[i] lit	[y] lu	[u] loup
	[e] les	[ø] leu	[o] lot
	[ɛ] lait	[œ] leur	[ɔ] lotte
	[a] là	[ə] le	
	[ɛ̃] lin	[ɑ̃] lent	[ɔ̃] long

Note : Les distinctions vocaliques [e]-[ɛ], [ø]-[œ] et [o]-[ɔ] ne sont pas faites dans tous les contextes et par tous les locuteurs du français. Par contre, certains locuteurs font aussi des distinctions entre patte et pâte, ([a]-[ɑ]) ainsi qu'entre brin et brun ([ɛ̃]-[œ̃]).

Ces phonèmes forment les unités phonétiques qui sont classées en voyelles, consonnes et semi-voyelles, etc.

1.6.1 Voyelles

Les voyelles sont des sons voisés qui résultent de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales. Chacune des voyelles correspond à une configuration particulière du conduit vocal. Les voyelles se différencient principalement les unes des autres par leur lieu d'articulation, leur aperture, et leur nasalisation. On distingue ainsi les voyelles antérieures, moyennes et postérieures, selon la position de la langue, et les voyelles ouvertes et fermées, selon le degré d'ouverture du conduit vocal. Il y a deux types de voyelle :

- les voyelles orales qui sont émises sans intervention de la cavité nasale ;
- les voyelles nasales qui font intervenir la cavité nasale.

La langue française comprend douze voyelles orales émises seulement par la bouche, ainsi que quatre voyelles nasales correspondant à la mise en parallèle des cavités nasales sur la cavité buccale par abaissement du voile du palais. Chaque voyelle se caractérise par les résonances du conduit vocal qu'on appelle "les formants".

En général, les trois premiers formants sont suffisants pour caractériser toutes les voyelles. Il est commode de représenter une voyelle sur un plan F1, F2 pour voir le “triangle articulatoire ” ou “triangle vocalique ” de la phonétique. Ce triangle représente la position de la langue dans la cavité buccale selon les 2 axes F1 “antérieur-postérieur ” et F2 “ouvert-fermé ”, selon que la langue est massée en avant vers la zone dentale (i), basse et étalée loin du palais (a), ou massée postérieurement vers le voile (u). F1 représente la position de la langue. F2 dépend de l'ouverture de la cavité buccale. Les autres formants représentent d'autres facteurs comme l'arrondissement des lèvres [5].

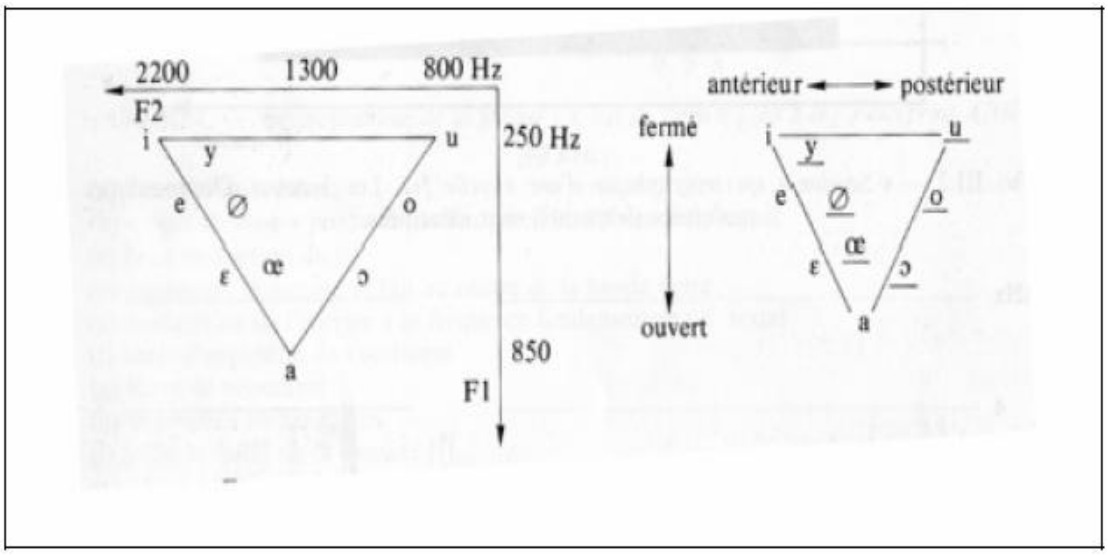


Figure 1.5 : Triangle vocalique pour le français [7]

1.6.2 Consonnes

Les consonnes sont des sons qui sont produits par une turbulence créée par le passage de l’air dans une constriction du conduit (les consonnes non voisées) ou une source périodique liée à la vibration des cordes vocales s’ajoute à la source de bruit (les consonnes voisées). Il y a trois types de consonnes : les fricatives (ou constrictives), les occlusives et les nasales.

- Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d’articulation (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 turbulente : les cordes vocales s’ouvrent et se ferment périodiquement, mais la fermeture n’est jamais complète.

- Les occlusives correspondent quant à elles à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal (les palais [k, g], les dentales [t, d], ou les lèvres [p, b]), puis relâché brusquement. La période d'occlusion est appelée la phase de tenue. Pour les occlusives voisées [b, d, g] un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées [p, t, k], la tenue est un silence.

- Les consonnes nasales [m, n, ï] font intervenir les cavités nasales par abaissement du voile du palais. Les consonnes sont caractérisées par la fréquence de spectre, la durée d'existence et la transition du son [5].

Tableau 1.2 : La classification des phonèmes du français [7]

CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales	← Lieu d'articulation
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[z]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[ɥ]	[j]	
Liquides		[l]	[R]	
VOYELLES				
Orales	Antérieures		Postérieures	
	Non arrondies		Arrondies	
Fermées	[i]	[y]	[u]	
	[e]	[ø]	[o]	
	[ɛ]	[œ]	[ɔ]	
Ouvertes	[a]			
Nasales	Antérieures		Postérieures	
Fermées	[ɛ̃]		[õ]	
Ouvertes		[ã]		

1.7 LES PARAMETRES PROSODIQUES D'UN SIGNAL VOCAL

La prosodie est une science de la linguistique qui étudie les éléments phoniques (l'accent, l'intonation, etc.) de n'importe quelle langue, et puisque la parole est un signal réel d'énergie finie, continu, 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 [8].

1.7.1 La Fréquence Fondamentale

La Fréquence Fondamentale 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 (figure1.6). Elle permet de diviser l'ensemble des sons de parole en trois grandes macros classes :

- 70 -250 Hz pour les Hommes ;
- 150 - 400 Hz pour les Femmes ;
- 200 - 600 Hz pour les Enfants [2].

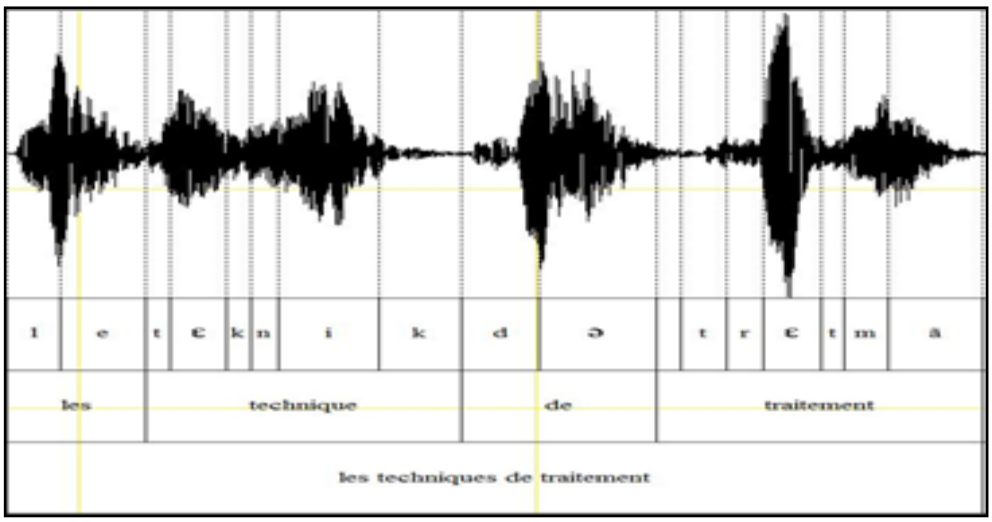


Figure 1.6 : Evolution de la fréquence de vibrations des cordes vocales de la phrase : "les techniques de traitement numérique de la parole"

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 [8].

1.7.2 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 [8].

1.7.3 L'Intensité ou l'énergie

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

$$\mathrm{E}=\frac{1}{T}\sum_{n=1}^{T} X_n^2 \qquad (1.1)$$

$$E_{dB} = 10 \times \log_{10}\left(\frac{1}{T}\sum_{n=1}^{T} x_n^2\right) \qquad (1.2)$$

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 1.7 et 1.8 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 [8].

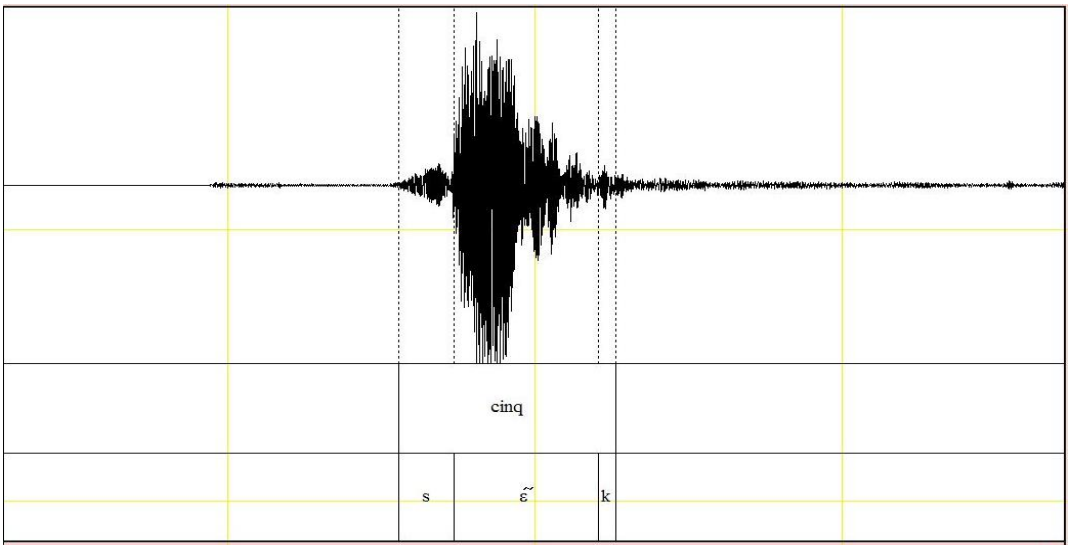


Figure 1.7 : Audiogramme du signal de parole du mot [$s\tilde{\varepsilon}k$]

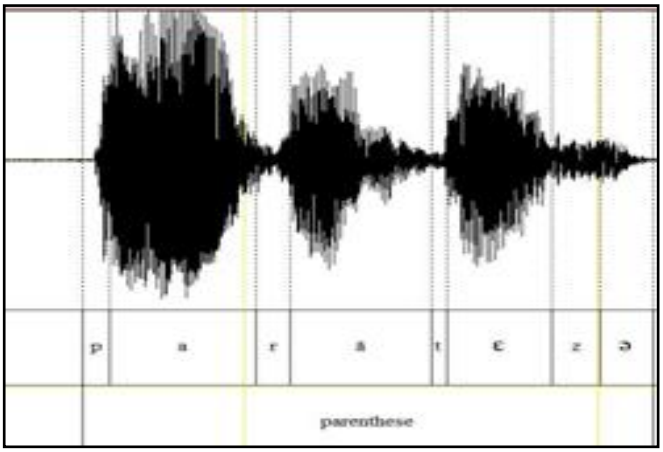


Figure 1.8 : signale audio du mot « Parenthèse » [par$\tilde{a}$tεz∂]

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].

1.8 COMPLEXITE DU SIGNAL VOCAL

La parole est un signal continu d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps ; tantôt périodique ou plus exactement pseudo périodique pour les sons voisés, tantôt aléatoire pour les sons fricatifs et tantôt impulsionnelle pour les sons occlusifs [4].

1.8.1 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 [4].

1.8.2 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 [4].

1.8.2.1 Variabilité intra-locuteur

La variabilité intra-locuteur concerne les différences de production du signal 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 [4].

1.8.2.2 Variabilité inter-locuteur

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 [4].

1.8.2.3 Variabilité contextuelle

Les mouvements articulatoires peuvent en effet 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 même dans le parlé 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 [4].

1.8.3 Coarticulation

La coarticulation est l'effet contextuel d'un son sur son voisin. Les contraintes introduites par les mécanismes de production créent ce genre de phénomènes. La production d'un son est fortement influencée par les sons qui le précèdent mais aussi par ceux qui le suivent en raison de l'inertie du geste articulatoire.

1.8.4 Redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, en effet, 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 [4].

1.9 NOTIONS FONDAMENTALE SUR L'ARABE STANDARD

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'Arabe Standard** (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 [8].

1.9.1 Système phonétique de l'Arabe Standard

L'Arabe Standard (AS) compte 34 phonèmes: 6 voyelles et 28 consonnes. 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.

Tableau 1.3: Transcription Orthographique Phonétique de l'AS [1]

Consone arabe	Transcription Des arabisants	Mode et lieu d'articulation
ا	[ʔ]	Laryngale occlusive
ب	[b]	Labiale occlusive sonore
ت	[t̪]	Dentale occlusive sourde
ث	[θ]	Interdentale émise en insérant le bout de la langue entre les dents ; fricative sourde
ج	[ʤ]	Affriquée palatale sonore
ح	[ħ]	Fricative laryngale sourde
خ	[x]	Vélaire fricative sourde
د	[d̪]	Dentale occlusive sonore

ذ	[ð]	Interdentale fricative sonore émise en insérant le bout de la langue entre les dents
ر	[r]	Vibrante linguale sonore
ز	[z]	Dentale fricative sonore
س	[s]	Dentale fricative sourde
ش	[ʃ]	Palatale fricative sourde
ص	[ʂ]	Emphatique ; dentale fricative sonore vélarisée
ض	[ð̴]	Emphatique ; Interdentale occlusive sonore vélarisée
ط	[t̴]	Emphatique ; dentale occlusive sourde vélarisée
ظ	[z̴]	Emphatique ; Interdentale fricative sonore vélarisée
ع	[ʕ]	Laryngale fricative sonore
غ	[ɣ]	Vélaire fricative sonore
ف	[f]	Labiodentale fricative sourde
ق	[q]	Occlusive arrière-vélaire sourde accompagnée d'une explosion glottale
ك	[k]	Palatale occlusive sourde
ل	[l]	Linguale ; sonore souvent appelée « liquide »
م	[m]	Labiale nasale sonore
ن	[n]	Dentale nasale sonore
هـ	[h]	Fricative glottale sonore
و	[w]	Semi-voyelle vélaire labiale sonore
ي	[j]	Semi-voyelle palato-alvéolaire sonore

Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscules/majuscules n'existe pas. 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 (Tableau 1.1) [8].

1.9.1.1 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 : [a], [u], [i], ces voyelles sont représentées dans un texte voyellé au dessus ou au dessous de la consonne, (ِ , ُ , َ) , exemple : تُرِكَ (turika) ;
- les voyelles longues : [aa], [uu], [ii], ces voyelles sont écrites sous forme de caractères consonantiques (ا,و,ي) et sont obligatoirement représentées dans un texte écrit (sauf dans certains cas particuliers), exemple : مُسَافِرُونَ (musaafiruuna).

1.9.1.2 Les consonnes

Les consonnes de l'arabe peuvent être classées suivant plusieurs critères comme suit (Tableau 1.1) :

• vibration des cordes vocales: les consonnes articulées avec une vibration des cordes vocales sont dites sonores (ou voisées), sinon elles sont dites sourdes (non voisées) ;

• le franchissement de l'air à travers le conduit vocal :

- Les fricatives qui sont caractérisées par un frottement sur les parois du conduit vocal. On distingue les fricatives non voisées comme س[s] et les fricatives voisées comme ز [z].
- Les occlusives qui sont caractérisées par un passage de l'air momentanément arrêté en un point quelconque de l'articulation, l'échappement de l'air s'effectue avec une petite explosion. On rencontre des dentales, des labiales et des glottales qui peuvent être aussi voisées et non voisées comme ب [b] et د [d] ;

- Un liquide caractérisé par un passage de l'air sur les côtés de la langue : ل [l] ;
- Deux nasales caractérisées par un échappement de l'air en même temps par la bouche et par le nez: م [m], ن [n] ;
- Une vibrante caractérisée par la vibration de la langue au passage de l'air: ر [r] ;
- Deux semi-consonnes (ou semi-voyelles) caractérisées par un passage rapide de l'air à travers la bouche accompagné de frottement consonantiques: ي [j], و [w].

• le mode d'articulation : suivant le mode d'articulation, on distingue les consonnes géminées et les consonnes emphatiques. Toute consonne géminée est formée par l'assemblage de deux consonnes identiques fortement articulées. La gémination est indiquée par un signe graphique spécifique appelé chadda (ّ). Les consonnes emphatiques ط [t̪], ض [ð̪], ص [s̪] , ظ [z̪] sont caractérisées par une forte tension des différents organes du conduit vocal.

Les voyelles brèves sont figurées par des symboles appelés signes diacritiques. Ces symboles sont absents à l'écrit dans la majorité des textes arabes ce qui peut engendrer des ambiguïtés de prononciation dans un système de TTS. Au nombre de trois, ces symboles sont transcrits de la manière suivante :

- la fetha [a] est symbolisée par un petit trait sur la consonne (بَ [ba]) ;
- la damma [u] est symbolisée par un crochet au-dessus de la consonne (بُ (bu]) ;
- la kasra [i] est symbolisée par un petit trait au-dessous de la consonne (بِ [bi]) ;
- un petit rond ° symbolisant la soukoun (سكون) est apposé sur une consonne lorsque celle- ci n'est liée à aucune voyelle (بَعْدَ [baʕda]) [1].

1.9.1.3 Le tanwiin

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

1.7.1.4 La chadda

Le signe de la chadda peut être placé au-dessus de toutes les consonnes en position non initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiques : Signe (كَلَّمَ [kallama] ″il a parlé à″) [1].

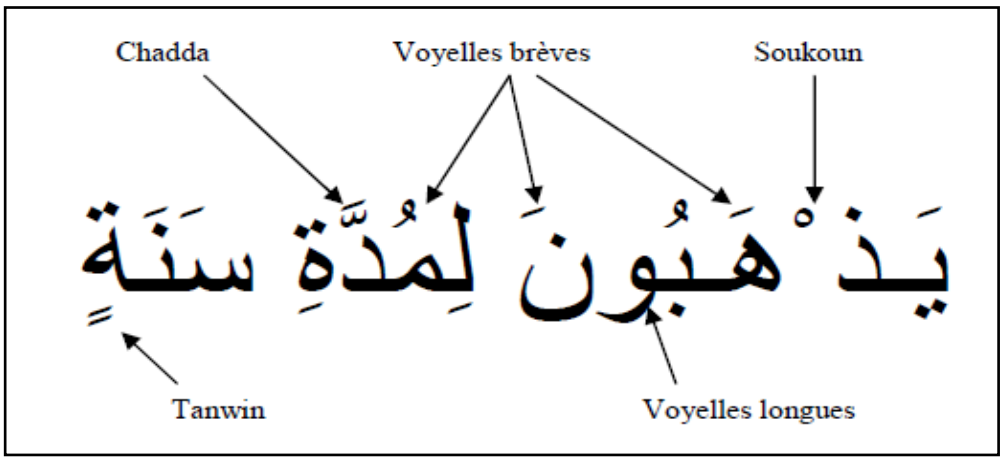


Figure 1.10 : Exemple d'une phrase voyellée /YavhabUna limuddati sanatin/ [3]

1.9.2 Système vocalique de l'Arabe Standard (AS)

La réalisation des voyelles est la classe de phonèmes issus des vibrations continues des cordes vocales, sans obstruction. Nous distinguons 3 voyelles [a], [i], [u], appelées dans l'ordre :الفتحة / الضمة / الكسرة et [A], [I], [U], représentant les voyelles longues.

La réalisation phonétique des voyelles est très variable et dépend à la fois de :

- ✓ l'origine géographique des locuteurs ;
- ✓ l'environnement consonantique et de la place de la voyelle dans le mot ;
- ✓ la place de l'accent du mot (tendance à abréger les longues non accentuées chez beaucoup d'arabophones).

Afin de situer les voyelles en termes de degré de durée de voisement, celles-ci sont mentionnées dans un triangle vocalique contenant :

- deux plans représentant la durée de voisement, de la plus brève à la plus longue ;
- le bas des triangles représente le degré d'aperture maximale, le haut des triangles représente le degré d'aperture minimale (figure 1.10).

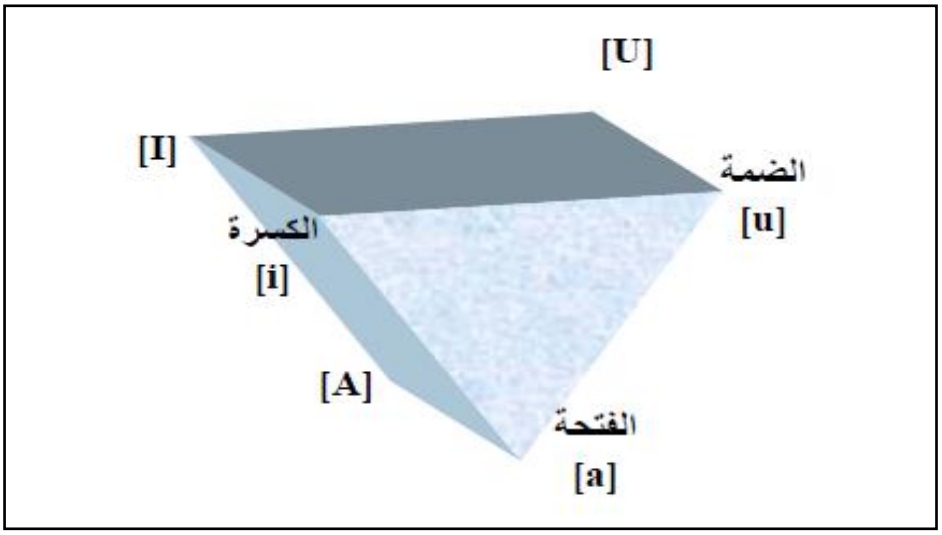


Figure 1.11 : Triangle vocalique de la langue Arabe [3].

Toutes les consonnes peuvent être dédoublées. Ou bien géminées. On prolonge et on renforce l'articulation de cette consonne. La gémination est indiquée par un signe graphique spécifique appelé "chadda".

La gémination joue un rôle très important en morphologie. Il est donc essentiel de bien l'entendre et de bien la réaliser.

Exemples :

درس : [darasa] "il a étudié"
درّس : [darrasa] "il a enseigné"

Les consonnes emphatiques [ص] ,[ظ],[ط],[ض] sont caractérisées par un trait articulatoire spécifique de la "pharyngalisation" : le son produit est plus grave que pour le son non emphatique correspondant. On l'obtient en modifiant la forme du résonateur buccal dans sa partie arrière par rétraction et exhaussement de la racine de la langue.
Il est très important de bien distinguer le son emphatique du non emphatique correspondant.
Exemples :

سيف [**s**ayf] "épée"
صيف [**S**ayf] "été"

Souvent la présence d'une consonne emphatique dans un mot "influence" l'environnement, consonantique et vocalique, et c'est toute la syllabe qui est emphatisée [3].

1.9.3 Particularités phonologiques de L'AS

Les caractéristiques phonologiques de l'AS sont l'emphase, la gémination et le madd.

1.9.3.1 L'emphase

Le mot emphase est habituellement utilisé pour rendre compte de manifestations prosodiques liées à l'accentuation volontaire d'une syllabe. Chez les linguistes arabes, il désigne certaines qualités que possèdent les consonnes :

- l'itbaq : les consonnes qui ont cette qualité sont ص [ʂ], ض [ð̪], ط [t̪], ظ [z̪]. 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 meut vers la partie postérieure de la cavité buccale, avec ou sans tafkhiim.

Seules les consonnes ص [ʂ], ض [ð̪], ط [t̪], ظ [z̪] 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'intérêt 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 oeuvre est directement liée à la technique de synthèse utilisée. Dans une approche par diphones, ont défini des unités acoustiques incluant les variantes emphatisées des voyelles avec l'ensemble des autres phonèmes. Les trajectoires des formants se trouvent ainsi préservées et fidèlement restituées [1].

1.9.3.2 La gémination

Au niveau graphique, elle est symbolisée par le signe de la chadda qui signifie le doublement 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 [حضَرَ : [hazðạra] ″ il a assisté″ est différente de حضَّرَ [haððạra] ″il a préparé″ où la deuxième consonne est géminée [1].

1.9.3.3 Le madd

Ce phénomène concerne l'allongement des voyelles. Il est provoqué par la présence d'une voyelle longue (ا [aa], و [uu], ي [ii]).

La lecture de textes arabes est régie par des règles phonologiques qui ont trait à la contraction des sons, leur élision et à l'assimilation homo-organique des nasales. Certaines de ces règles sont obligatoires, d'autres facultatives ou réservées à certains types de textes, comme le Coran. Nous présentons ci-dessous des définitions brèves de ces phénomènes :

- la contraction : elle est utilisée à cause de la lourdeur de la liaison de deux phonèmes identiques. Elle peut être obligatoire (لَهُ قلْ [qul] [lahu] = قلَّهُ [qullahu]), interdite (dans مَلَلْتُ [malaltu], le premier [l] ne doit pas être contracté avec le second [l]) ou permise (سَرَرَ [sarara] = [sarra]) ;
- l'élision : c'est le changement qui se produit dans la prononciation du phonème [n] qui porte une soukoun devant certaines consonnes ;
- l'assimilation homo-organique des nasales : elle concerne la substitution d'une consonne nasale par une autre consonne. Elle peut se produire à l'intérieur du mot (أَنْبَتَتْ [ʔa**n**batat] = أَمْبَتَتْ [ʔa**m**batat]) ou à la frontière de deux mots successifs (بَعْدْ منْ [min baʕd] = ممْبَعْدْ [mimbaʕd]).

1.9.4 Problème de la langue arabe en traitement automatique

La langue arabe rencontre deux principaux problèmes en traitement automatique : le premier, général, concerne l'agglutination des mots ; le second, spécifique, a trait à l'absence de voyelles à l'écrit.

1.9.4.1 Agglutination des mots

La plupart des mots en AS sont composés par agglutinations d'éléments lexicaux de base (proclitique + base + enclitique). Par exemple, la détermination peut s'exprimer par agglutination de l'article ال [ʔal] avant le mot (الولد , [alwaladu] , ″l'enfant″) ou par agglutination d'un pronom personnel après celui-ci (وَلَدُهُ , [waladuhu] , ″son enfant″). De même, les pronoms personnels peuvent se rattacher aux verbes (ضرَبَهُ, [ðaṛabahu], ″il l'a frappé″), les particules régissant le cas indirect aux noms (أَدَارِهِ , [kadaarihi], ″comme sa maison″) et les conjonctions de coordination aux verbes (فَذَهَبَ , [faðahaba], ″et il est parti″), etc. [1].

Dans toute perspective de traitement automatique, le problème est donc de décomposer le mot en ces différentes parties. Cette décomposition nécessite des connaissances de niveau supérieur en cas d'ambiguïtés.

1.9.4.2 Voyellation

Les textes en AS sont ordinairement dépourvus de diacritiques. Pour les lire, tout un processus mental est nécessaire : identifier le mot comme appartenant au lexique puis lui attribuer ses voyelles dans son contexte, ce qui nécessite la compréhension du texte.

Pour le TA d'un texte, il est indispensable d'introduire les voyelles avant le traitement dans le cas d'une synthèse TTS ou après dans le cas d'une Reconnaissance Optique (RO). Cette opération, appelée voyellation ou vocalisation automatique, est effectuée par la machine et se déroule *généralement* en deux étapes, sous forme d'une analyse :

- morphologique qui va assigner à chaque mot non-voyellé l'ensemble des mots voyellés correspondants. Ce qui nécessite la présence d'un lexique total avec toutes les formes canoniques et fléchies des mots ;
- syntaxique pour réduire l'ambiguïté au vu du contexte grammatical ;
- sémantique qui est nécessaire pour réduire l'ambiguïté au vu du sens de la phrase [1].

1.10 CONCLUSION

Dans ce chapitre nous avons exposé des notions de base sur les généralités 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 II :

Analyse et Synthèse de la parole

2.1 INTRODUCTION

La qualité d'un système de synthèse de la parole dépend de l'intelligibilité, du naturel, de la parole générée et des caractéristiques propres à la voix produite. Ces caractéristiques dépendent des techniques et des méthodes de synthèse, mais également du soin apporté à la modélisation linguistique et prosodique. Plusieurs travaux soulignent le fait que des structures linguistiques entretiennent des liens étroits avec les réalisations prosodiques. Dans ce chapitre, nous allons introduire le cadre technique de notre étude : la synthèse de la parole. Le chapitre s'articule autour des principes de la synthèse vocale, suivi d'une description bien détaillée des différentes techniques et méthodes utilisées.

2.2 HISTORIQUE DE LA SYNTHESE DE LA PAROLE

À plusieurs reprises au cours de l'histoire, on a tenté de reproduire la voix humaine. Au XVIII^{ème} siècle, on met au point à cet effet des dispositifs mécaniques équipés de soufflets et d'anches vibrantes. Au XX^{ème} siècle, l'apparition de l'électricité et de l'électronique autorisent des tentatives plus ambitieuses : en 1922, J.C. Stewart fabrique une machine capable de reproduire des voyelles, des diphtongues et quelques mots simples ; plusieurs années plus tard en 1939, H. Dudley présente, à l'occasion de l'exposition universelle de New York, le VODer (Voice Operation Demonstrator), appareil mis au point par les laboratoires Bell.

Mais ce n'est que dans les années cinquante que les premiers véritables synthétiseurs de la parole font leur apparition, avec, par exemple, le Pattern Playback, système mis au point par les laboratoires Haskins aux USA, qui se présente comme un lecteur de sonagraphe (un faisceau de lumière produit, après amplification, des sons à partir de la représentation de leur durée, de leur fréquence et de leur intensité).

Depuis les années soixante-dix, des progrès considérables ont été accomplis, avec notamment le développement de l'utilisation des calculateurs numériques. Aujourd'hui encore, ces progrès se poursuivent, dans plusieurs directions (perfectionnement des synthétiseurs à formants, des synthétiseurs à prédiction linéaire, etc.) [8].

2.3 PRINCIPE DE SYNTHESE DE LA PAROLE

Qu'est-ce que la synthèse de la parole ?

Une simple réponse à cette question pourrait être : « la production de la parole par une machine ». Mais chacun sait qu'un magnétophone peut produire de la parole sans que l'on n'ait jamais songé à l'appeler « synthétiseur ».
Une meilleure définition serait alors : « la production par une machine de sons ou de mots qui n'ont jamais été prononcés auparavant par un être Humain ». Mais cette définition est trop restrictive car elle ne tient pas compte des techniques de synthèse par assemblage d'éléments préenregistrés.

Si l'on peut simplement définir cette technique en fonction de la sortie, considérons alors le type d'entrée qui va engendrer une parole de synthèse. Deux cas peuvent se présenter : Soit l'entrée est une succession de concepts, ou bien c'est une chaîne de caractères orthographiques. Dans un cas comme dans l'autre, l'émission de la parole sera déterminée par une représentation phonétique de ce qui doit être dit. Nous adoptons donc la définition suivante : «La synthèse de la parole permet de produire des sons de la parole à partir d'une représentation phonétique du message».

Le message vocal est un continuum acoustique dans lequel il n'y pas de frontière marquée entre les mots ni entre les sons élémentaires (ou phonèmes) du langage. En synthèse, la reproduction de ce message résulte de l'encodage d'information au niveau :

- Segmental par le choix des unités phonétiques et de leurs enchaînements ;
- Suprasegmental par la génération automatique de la prosodie donnant à ces unités une importance de nature linguistique et expressive.

A cette étape, il est important de bien distinguer la différence qui existe entre synthèse de la parole (on l'appelle parfois synthèse de la parole à partir du texte) et un synthétiseur de parole, ainsi nous nommons :

- un système de synthèse de la parole comme étant capable de reproduire des sons « parlés » à partir d'un texte ou d'une entrée conceptuelle (Figure 2.1) ;
- un synthétiseur de parole comme étant la dernière étape de la transformation d'un certain nombre de paramètres de contrôle en parole [8].

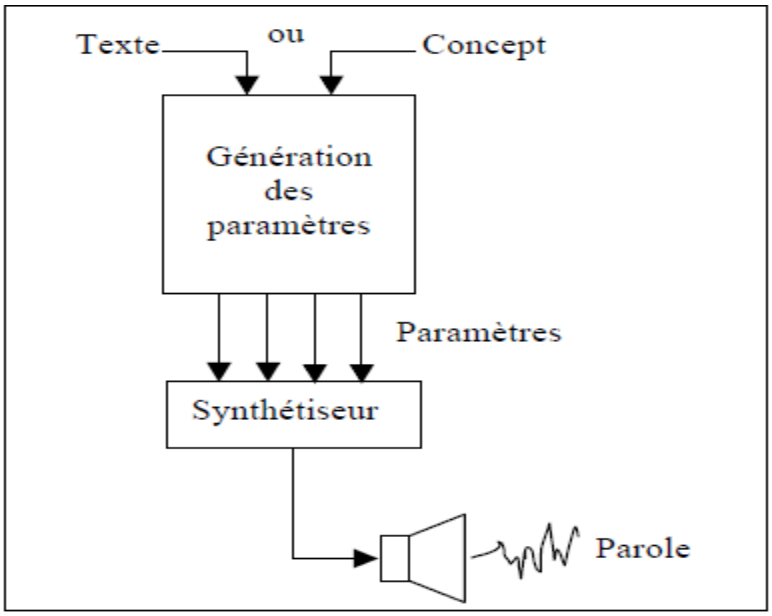


Figure 2.1 : Système de synthèse de la parole [8]

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 fondamentalement 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'obtenues par analyse ;

- 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 dedialogue Homme-Machine, reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue [8].

2.4 ARCHITECTURE D'UN SYSTEME DE SYNTHESE DE LA PAROLE

Tout système **TTS** (**T**ext- **T**o -**S**peech) est généralement constitué de deux blocs de traitements principaux : un bloc de traitements linguistiques et un bloc de traitements acoustiques. Le premier bloc vise à analyser et à structurer le texte afin de déterminer un mode de prononciation cohérent, puis à transformer le texte analysé en une séquence de descripteurs symboliques décrivant les unités cible. Le deuxième bloc consiste à générer un signal acoustique adapté à cette séquence symbolique.

La **Figure 2.2** présente l'architecture générale d'un système de synthèse de la parole à partir du texte. Les deux premières parties qui concernent les traitements de *haut niveau* permettent le passage de la représentation orthographique du texte en entrée à une représentation phonétique munie d'une description prosodique. La dernière partie englobe les traitements de bas niveau du synthétiseur qui permettent la génération proprement dite du signal acoustique [8].

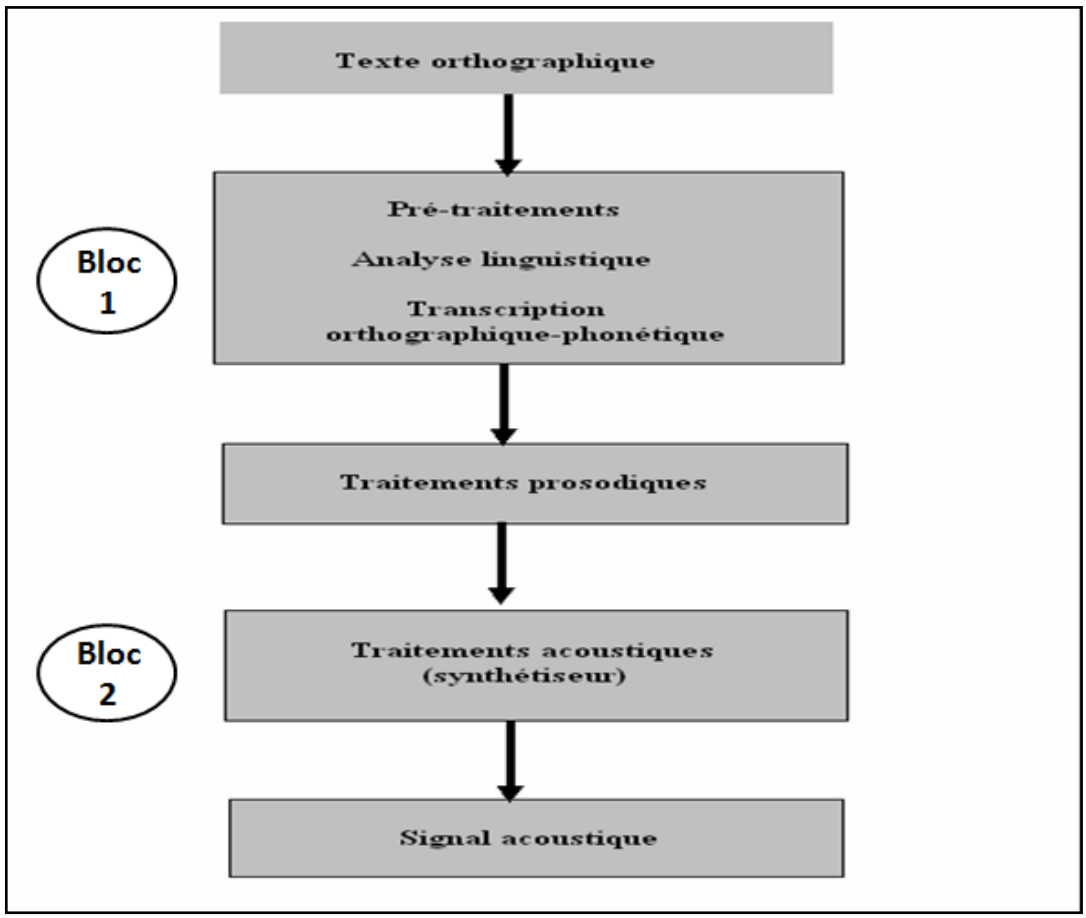


Figure 2.2 : Architecture générale d'un système de synthèse de la parole à partir du texte

2.5 APPLICATIONS DE SYNTHESE DE LA PAROLE

Les applications actuelles de synthèse de la parole à partir du texte peuvent être regroupées en cinq grands domaines :

- Aide aux personnes handicapées :

  - lecture d'écrans ou de documents écrits pour non-voyants ;
  - Aides à la communication vocale pour personnes muettes, laryngectomisées ou à infirmité motrice cérébrale ;
  - journaux vocaux, etc.

- Outils d'Enseignement Assisté par Ordinateur (OEAO) :

  - système de dictées automatiques ;
  - système d'apprentissage des langues.

- Applications industrielles :

  - Serveurs d'alerte, de surveillance de sites et de supervision de réseaux ;
  - télémaintenance ;
  - fonctions d'aide dans les postes de pilotage ;
  - Fonction de vérification vocale dans les postes d'édition (correction des épreuves) ou de saisie d'informations écrites (bases de données), etc.

- Applications grand public non téléphoniques :

  - Domotique (alarmes, appareils domestiques parlants, etc.) ;
  - Micro-informatique (jeux et CD parlants, bureautique, etc.).

- Télématique vocale :

  - Serveurs vocaux d'informations (la synthèse remplaçant la parole naturelle enregistrée pour des informations rapidement évolutives et disponibles sous forme textuelle) ;
  - Serveurs de lecture vocale de FAX ou de messages électroniques (e-mails);
  - Automatisation de services de renseignements (Annuaires, standards d'entreprises, etc.) [1].

2.6 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, 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 qui est souvent non récursif du 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.

Le signal vocal peut être analysé soit, en tenant compte des mécanismes de production en utilisant les méthodes paramétriques, soit en les ignorant 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 de l'autocorrélation font partie des méthodes temporelles. Les coefficients les plus utilisés en RAP sont certainement les cepstres. Ils peuvent être extraits de deux façons : soit par l'analyse paramétrique, à partir du Codage Linéaire Prédictive ou **L**inear **P**redictive **C**oding (**LPC**) en Anglais, soit par l'analyse cepstrale [4].

2.6.1 Analyse par FFT

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**) en Anglais. 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 [4].

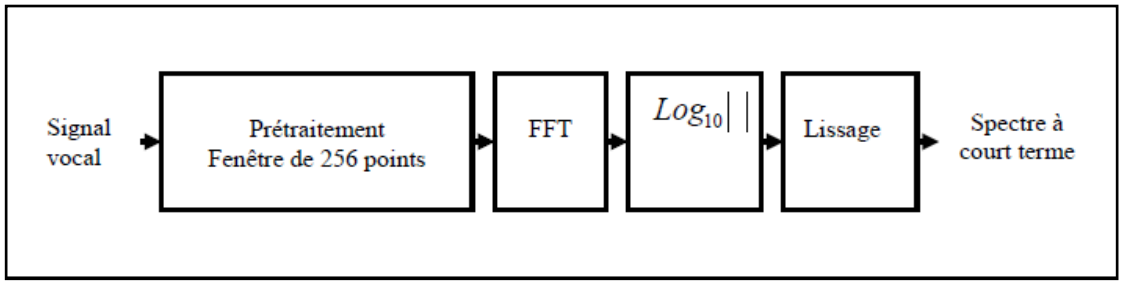


Figure 2.3 : Analyse numérique du signal parole par FFT

2.6.2 Codage Prédictif Linéaire (LPC)

La prédiction linéaire est une technique importante pour la compression d'un signal de parole en modélisant le conduit vocal par un filtre dont les coefficients sont déterminés par le biais de l'analyse de la redondance du signal, cela est réalisé par la prédiction linéaire qui exploite cette redondance pour prédire un échantillon par une combinaison linéaire d'échantillons antérieurs, c'est l'idée de base du codage par prédiction linéaire.

Lors de l'analyse à court terme, la redondance proche entre les échantillons du signal de parole est supprimée par un filtre d'analyse LP, représentant le conduit vocal. Ce filtre permet d'extraire la structure des formants du signal d'entrée et d'obtenir un signal de sortie de faible énergie qui correspond à l'erreur de prédiction appelée signal résiduel ou

d'excitation. Le filtre inverse d'analyse est le filtre de synthèse LP, dont la fonction transfert décrit l'enveloppe spectrale du signal de la parole, il génère le signal de parole synthétisée.

Chaque trame de parole est donc modélisée en sortie du système linéaire par un signal d'excitation. Un meilleur codage de celui-ci pourrait être obtenu en utilisant un prédicteur à long terme qui prendra en compte la corrélation entre les échantillons distants du signal de parole. L'extraction de cette périodicité est obtenue par un estimateur du pitch. Cette analyse n'aura aucun effet sur les sons non voisés [6].

Dans cette présente étude nous avons utilisé la méthode par prédiction linéaire pour l'extraction des formants. Nous allons donc présenter cette méthode afin d'avoir une idée sur la façon avec laquelle nous avons calculé les formants. Le « Linear Predictive Coding »ou LPC repose sur un modèle simple décrivant le Comportement des organes vocaux lors de la synthèse d'un son. Ce modèle a été conçu à partir d'un modèle mathématique appelé « modèle autorégressif ». Nous allons donc commencer par présenter le modèle autorégressif afin de mieux comprendre le modèle LPC [10].

2.6.2.1 Le modèle autorégressif (AR)

Un processus AR peut être modélisé par la sortie d'un filtre linéaire et invariant dans le temps.
Son équation aux différences s'écrira :

$$x(n)+\sum_{i=1}^{M} a_i\, x(n-i)=e(n) \qquad (2.1)$$

La fonction de transfert H(z) s'écrit comme suit :

$$H(Z)=\frac{X(Z)}{e(Z)}=\frac{1}{\sum_{1}^{M} a_i Z^{-i}}=\frac{1}{A(Z)} \qquad (2.2)$$

Notons que H(z) ne contient alors que des pôles et c'est pour cette raison que ce modèle est aussi appelé modèle tous pôles. La condition de stationnarité du signal x(n) est équivalente à la condition de stabilité du filtre est que celle-ci n'est assurée que si les racines des polynômes H(z) (donc les coefficients a(i) du filtre), sont de modules inférieurs à 1 [10].

2.6.2.2 Modèle AR et modèle de prédiction linéaire

Si l'entrée d'un modèle autorégressif est inconnue alors nous pouvons estimer ce dernier par un modèle ayant les mêmes propriétés appelées modèle de prédiction linéaire Son équation aux différences $\hat{X}$(n) où il est l'estimé de *X(n)*, s'écrira alors :

$$\hat{X}(n) = -\sum_1^M \hat{a}_i X[n-i] \quad (2.3)$$

$\hat{a}_i$: les estimés des coefficients a(i) du filtre AR ;

P : ordre de prédiction.

On note *e(n)* l'erreur de prédiction définie comme suit :

$$e[n]=X[n]-\hat{X}[n]]=-\sum_1^M \hat{a}(i)\, X[n-i] \text{ ; } a(0) = 1. \quad (2.4)$$

2.6.2.3 Pourquoi utilise-t-on le modèle autorégressif

On peut assimiler le mécanisme phonatoire à un système de transmittance :

Avec :

$$H(Z) = \frac{G}{A(Z)} \quad (2.5)$$

$$A(Z) = \sum_{i=1}^{p} a_i\, z^{-i}$$

a(0)=1 et *A(z)* est un polynôme qui s'écrit comme suit:

$$x(n)+\sum_{i=1}^{M} a_i\, x(n-i)=G\ U(n) \quad (2.6)$$

G : le gain de ce système

Dans le domaine temporel :

$$X(z) = U(z) \cdot H(Z) \quad (2.7)$$

Ce modèle de production d'un signal est appelé AR (autorégressif) avec :

Si on suppose que notre système est excité par une excitation U (n) qui se présente comme :

- des **sons voisés** (ou sonores): l'excitation est un train périodique d'impulsions ;

- des **sons non voisés (sourds)** : l'excitation est un bruit blanc centré (de moyenne nulle et variance nulle) (Figure 2.4**).**

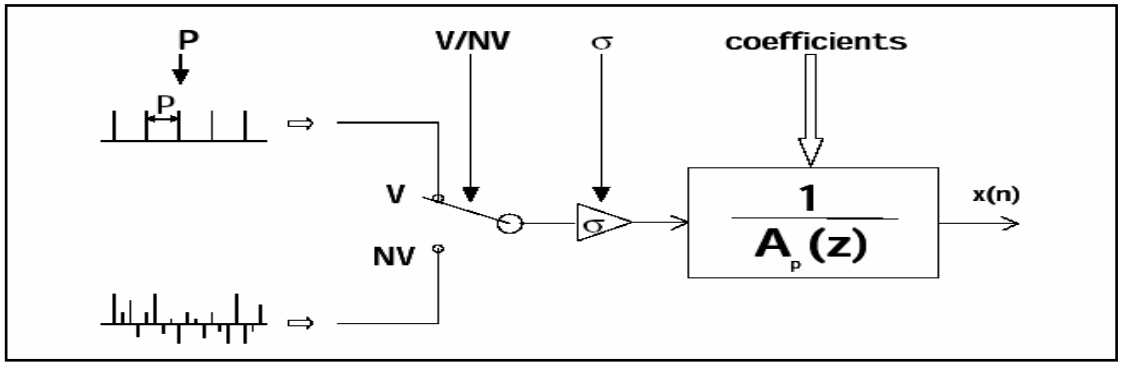


Figure 2.4 : Modèle de production de la parole [10]

P : l'ordre de prédiction : voisé, nv : non voisé,σ : $le\ gain$, $\frac{1}{A(z)_p}$: fonction de transfert

La transmittance *H (z)* est celle d'un filtre polynomial, On définit le filtre inverse dont la transmittance est définie par :

$A(Z) = \sum_{i=1}^{p} a_i\, z^{-i}$ *a(0)*=1 ; (2.8)

Ce filtre excité par le signal original, engendre en sortie l'erreur de prédiction [10].

2.6.2.4 Estimation des coefficients de prédiction linéaire

Le critère usuel pour l'estimation des coefficients du modèle de prédiction est la minimisation de l'erreur quadratique de ce dernier ou de la variance. La variance est définie sous la forme suivante :

$\sigma_e^2 = R_e(0) = \sum_{i,j=0}^{p} a(i)\ a(j)\ \overline{X(n-i)\ X(n-j)}$ (2.9)

$\sigma_e^2 = \sum_{i,j=0}^{p} a(i)\ a(j)\ R_x(i-j)$ (2.10)

La minimisation par rapport aux coefficients *a(i)*, nous mène à calculer la dérivée partielle suivante par rapport à *a(i)* :

$\frac{\partial \sigma_e^2}{\partial a\,(i)} = \sum_{j=0}^{p} R_x(i,j)a(j) = 0$ (2.11)

$$\sum_{j=0}^{p} R_x(i-j)a(j) = -R_x(\mathrm{i}) \tag{2.12}$$

L'autocorrélation est l'une des méthodes les plus utilisées des **LPC**, la variance de l'erreur de prédiction, sous forme quadratique :

$$\sigma_e^2 = [1 \quad a].\, R_{xx}^{p} \,.\begin{bmatrix} 1 \\ a \end{bmatrix} \tag{2.13}$$

$$a = [1, a(1), a(2), \quad , \dots , \quad a(p)] \tag{2.14}$$

La méthode d'autocorrélation assure la stabilité du modèle AR et conduit à un système de matrice de Toeplitz (symétrique et égalité des éléments diagonaux de la matrice) qui s'écrit :

$$R_{xx}\ (\mathrm{p}) = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(p) \\ R_{xx}(1) & R_{xx}(2) & \dots & R_{xx}(p-1) \\ \vdots & \vdots & \dots & \vdots \\ R_{xx}(p) & R_{xx}(p-1) & \dots & R_{xx}(0) \end{bmatrix} \tag{2.16}$$

$$R_x = [R_{xx}(1),\ R_{xx}(2),\ R_{xx}(3), \dots ,\ R_{xx}(P)] \tag{2.17}$$

On écrit dans ce cas :

$$R_{xx}(p) = \begin{bmatrix} R_{xx}(0) & R_x \\ R_x & R_{xx}(p-1) \end{bmatrix} \tag{2.18}$$

On aura donc d'après la forme quadratique de la variance de l'erreur [10] :

$$\sigma_e^2 = \sigma_e^2 + 2.a.R_x + a\, R_{xx}^{p-1} a \tag{2.19}$$

$$R_{xx}^{p-1} a = -R_x \tag{2.20}$$

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(p) \\ R_{xx}(1) & R_{xx}(2) & \dots & R_{xx}(p-1) \\ \vdots & \vdots & \dots & \vdots \\ R_{xx}(p) & R_{xx}(p-1) & \dots & R_{xx}(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ \vdots \\ R_{xx}(p) \end{bmatrix} \tag{2.21}$$

Le calcul des coefficients de prédiction a_i se résume donc à inverser la matrice d'autocorrélation R_{xx}.c'est une matrice de Toeplitz symétrique et le système découlant de l'équation matricielle (2. 20) est un système de Yule-Walker, l'inversion de la matrice R_{xx} est une opération qui nécessite un calcul fastidieux et long, un algorithme efficace peut être employé pour déterminer les coefficients LP. C'est l'algorithme de Levinson-Durbin [6].

2.6.3 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 MFCCs 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 est soit une séquence d'impulsions (périodiques, de période T0, 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.

$$Y(z) = B(z).U(z) \tag{2.22}$$

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

$$\log|Y(z)| = \log|U(z)| + \log|B(z)| \tag{2.23}$$

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 :

$$C(n) = FT^{-1}\{\log(FT\{y(n)\})\} \tag{2.23}$$

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 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, débarrassé des lobes, dû à la contribution de la source. Ces deux contributions peuvent être séparées par une simple fenêtre temporelle notée $\boldsymbol{F}$ (filtrage) 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 formantiques) est obtenue par une transformation supplémentaire (Figure 2.4). Le spectre lissé débarrassé théoriquement de la contribution de la source ne contient que des informations sur le conduit vocal et en particulier ses extrema (Formants) [4].

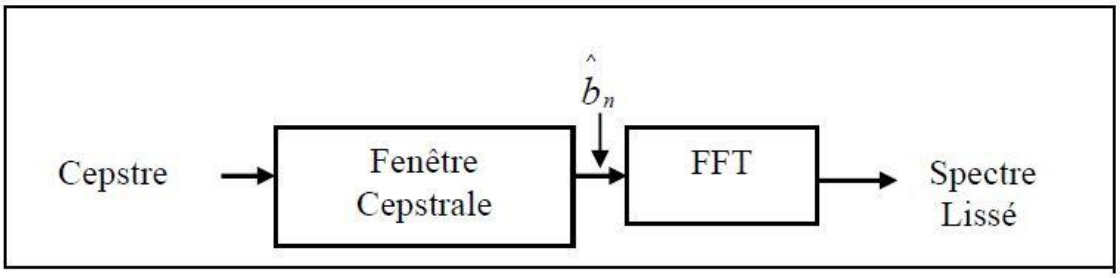


Figure 2.5 : transformation schématique pour l'obtention de la structure formantique à partir du cepstre [4]

2.7 REPRESENTATIONS SPECTRALES DU SIGNAL DE PAROLE

Il existe plusieurs représentations spectrales du signal de parole, parmi elles nous citerons :

2.7.1 Spectre obtenu par FFT

Tout son est la superposition de plusieurs ondes sinusoïdales. Grâce à la *FFT*, on peut isoler les différentes fréquences qui le composent. On obtient ainsi une répartition spectrale du signal (figure 2.6). Les valeurs des formants sont calculées automatiquement dans le signal de parole au moyen d'un lissage spectral.

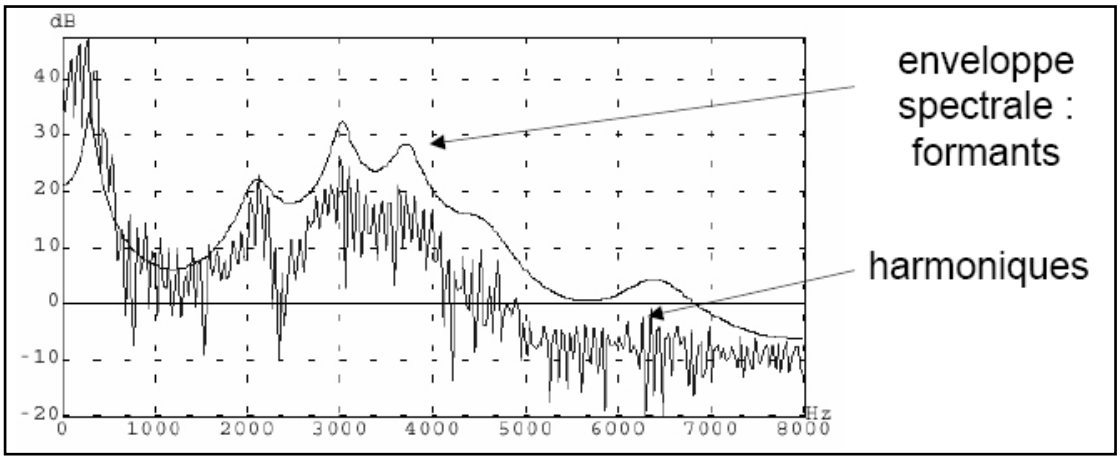


Figure .2.6 : Spectre obtenu par transformée rapide de Fourier (FFT) [10].

2.7.2 Spectre obtenu par prédiction linéaire (LPC)

Le spectre obtenu par LPC est plus lisse et permet ainsi de repérer plus facilement les Formants (figure 2.7).

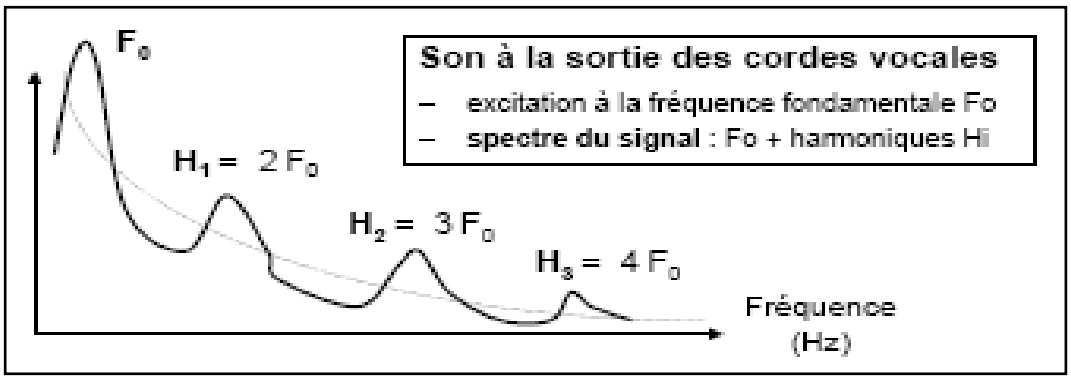


Figure 2.7 : Spectre lissé obtenu par prédiction linéaire (LPC) [10].

2.7.3 Le Spectrogramme

Le spectrogramme est un outil de visualisation utilisant la technique de la transformée de Fourier et donc du calcul de spectres. Il a commencé à être largement utilisé en 1947, à l'apparition du sonagraphe, et est devenu l'outil incontournable des études en phonétique pendant de nombreuses années.

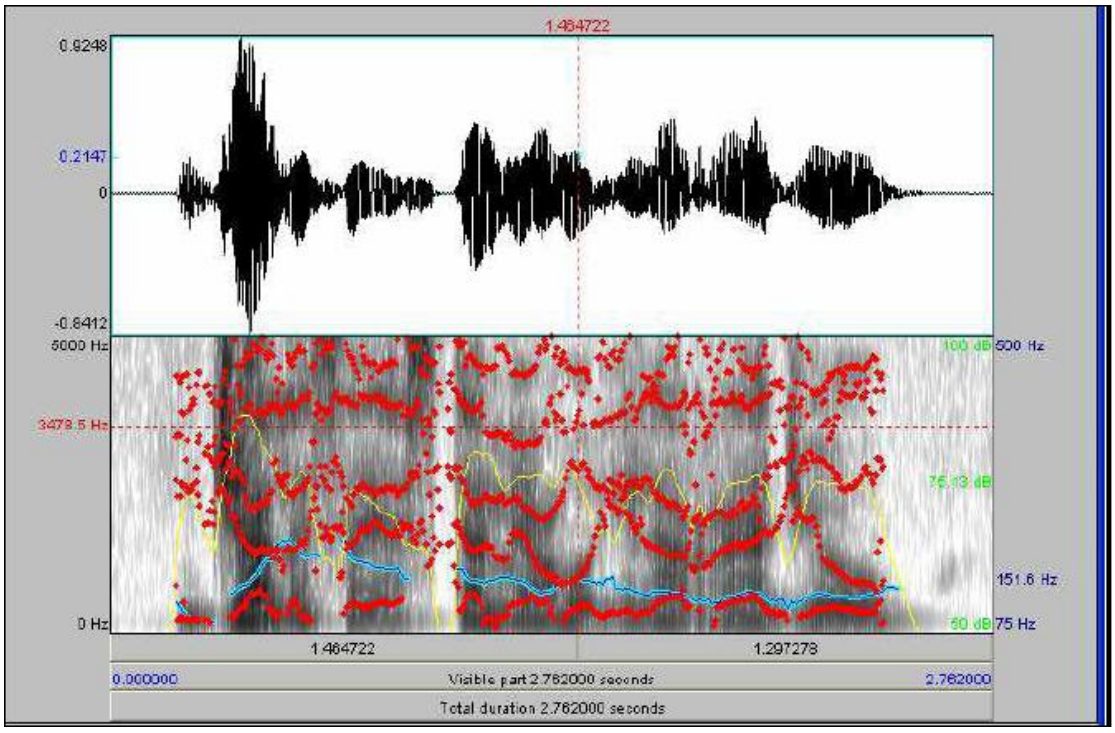


Figure 2.8 : Spectrogramme de la phrase / جلس يستمع إلى الراديو / [galasayastami£u ilaa arraadyuu] [4].

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 par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bandes étroites sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales. Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à tout instant (Figure 2.8) [10].

2.7.4 Intérêts de la représentation fréquentielle du signal de parole

La représentation fréquentielle de la parole est d'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 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 [10].

2.8 LES METHODES DE LA SYNTHESE DE PAROLE

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

2.8.1 Synthèse par règles

La synthèse par règles est une méthode qui a eu beaucoup de succès dans le contexte de la synthèse de la parole à partir du texte. Des règles sont utilisées pour estimer les paramètres nécessaires. Cette approche est fondée sur un modèle de production du signal vocal, modèle commandé par un nombre restreint de paramètres. La synthèse se décompose alors en deux étapes : une transformation des informations phonético- prosodiques, à l'aide de règles contextuelles, en commandes permettant de spécifier l'évolution temporelle des paramètres du modèle de synthèse; les paramètres ainsi déterminés sont utilisés pour synthétiser le signal acoustique.

Dans ce type de synthèse, les caractéristiques supra-glottiques sont modélisées à l'aide d'un filtre linéaire dont la fonction de transfert varie au cours du temps. Les paramètres utilisés pour le contrôle du filtre sont les paramètres formantiques, à savoir la fréquence centrale, la bande passante et l'amplitude des maxima significatifs de la fonction de transfert du conduit vocal. Pour obtenir une parole intelligible, il suffit de spécifier les paramètres des 3 à 4 formants les plus importants, d'où la dénomination de synthèse par formants couramment employée pour ce type de synthèse. Une telle approche ne permet pas de restituer un signal de parole apparaissant naturel. La qualité médiocre obtenue résulte d'une part de la difficulté à modéliser suffisamment finement les trajectoires acoustiques et d'autre part de la modélisation trop grossière du signal glottique.

Parmi les grands avantages de cette méthode, nous pouvons citer notamment la grande souplesse d'utilisation, la facilité d'extension, et surtout la grande portabilité de ces systèmes facilitant leur intégration dans une large gamme de produits.

Les synthétiseurs par règles sont organisés comme à la (figure 2.4) [8].

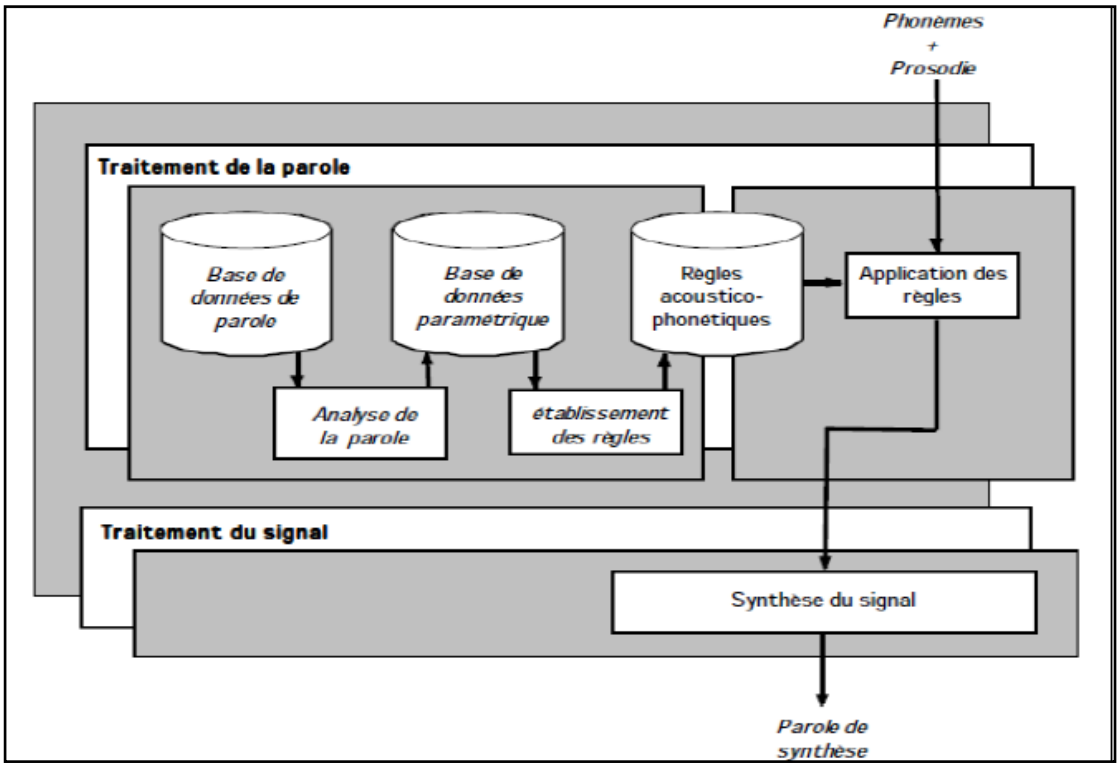


Fig. 2.9 : Schéma de conception et fonctionnement typique d'un système de la synthèse par règles [9].

2.8.2 Synthèse par concaténation d'unités pré-stockées

La synthèse par concaténation d'unités pré-stockées est la génération des sons à partir de la juxtaposition d'un ensemble d'unités préenregistrées, ces dernières sont obtenues par une opération d'analyse du signal qu'on veut produire. Elle consiste à choisir dans une large base de données les unités sonores les plus appropriées pour construire, par concaténation la phrase à produire (Figure 2.5). En réalité dans cette approche on peut trouver plusieurs types d'unités (phonèmes, diphones, syllabes, polysons, mots, phrases). Parmi les méthodes de synthèse par concaténation on a [13] :

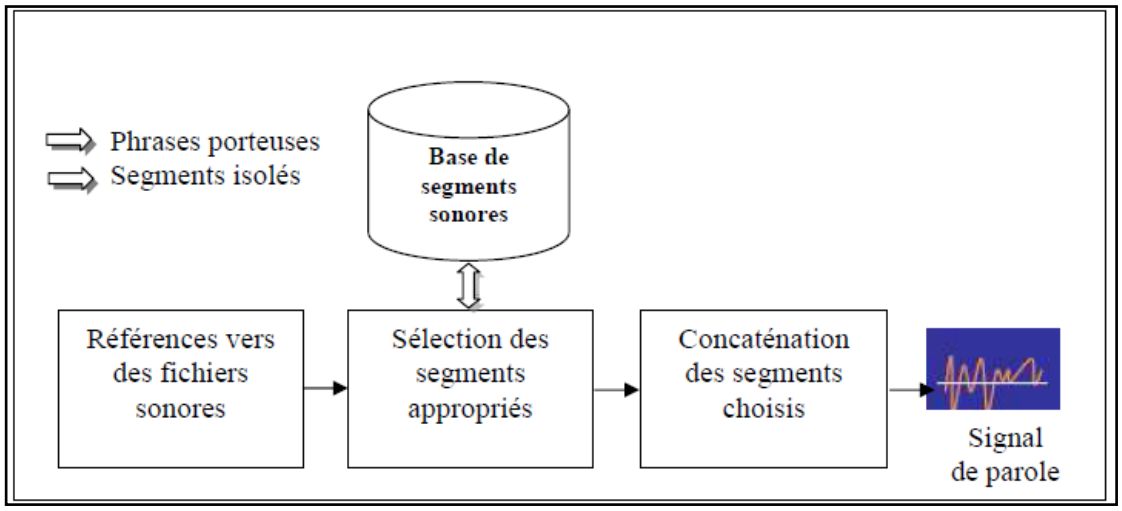


Figure 2.10 : Principe de base de la méthode de synthèse par concaténation [12].

2.8.2.1 La concaténation des phrases

Est une simple opération d'enregistrement et de restitution des phrases à synthétiser en vue d'une réalisation bien précise. Cette précision limite leur utilisation à un petit vocabulaire, ainsi qu'à des applications très restreintes telles que les jouets pour enfants, les répondeurs téléphoniques, l'horloge parlante, etc.

Cette dernière se compose de deux parties, une stable qui correspond à la phrase «il est : … heures ….minutes….secondes », et une autre variable où on trouve les nombres qui correspondent à la valeur actuelle de l'heure, des minutes, et des secondes ; Base de segments sonores Références vers des fichiers sonores Sélection des segments appropriés Concaténation des segments choisis Signal de parole.
Le principe de cette méthode est de stocker et de restituer de la parole continue ; en vue d'une application bien définie [13].

2.8.2.2 La concaténation des mots

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

2.8.2.3 La concaténation des phonèmes

Puisque les phonèmes représentent les éléments atomiques dans n'importe quelle langue, il suffit de les juxtaposer pour synthétiser un mot ou une phrase. Malgré la simplicité de cette méthode, elle présente l'inconvénient de discontinuité du signal généré et cela à cause du problème de la coarticulation qui est dû grâce à l'influence d'un son sur ses voisins. Pour résoudre ce problème la solution est de changer le phonème par une autre unité plus coûteuse en information qui est le diphone [13].

2.8.2.4 La concaténation par diphones

La synthèse par concaténation de diphones, ou plus simplement, synthèse par diphones, est née du double constat suivant :

- d'une part, les transitions entre phones ont un impact très important sur l'intelligibilité et le naturel perçus d'un signal de parole,
- d'autre part les modèles acoustiques existants ne permettent pas suffisamment de rendre compte de la complexité d'un signal de parole naturel, en particulier dans les zones transitoires.

De ce dernier point découle l'idée d'utiliser des unités acoustiques pré-enregistrées par un locuteur et stockées dans une mémoire informatique. La génération du signal consiste alors à récupérer les unités adéquates et à les juxtaposer (ou concaténer). Afin de préserver les zones transitoires entre phones, le choix de l'unité élémentaire s'est logiquement porté sur le diphone, c'est-à-dire l'unité acoustique qui s'étend du milieu d'un phone au milieu du phone suivant (figure 2.6). Ses frontières appartiennent donc à des zones acoustiquement stables, ce qui facilite les concaténations, à l'époque sous le nom de dyad .

Pour un alphabet (usuel en Français) de 35 phonèmes le nombre théorique de diphones est de $35 \times 35 = 1225$, mais dans la pratique une centaine de diphones inutilisés peuvent être exclus. Une occurrence de chaque diphone est enregistrée puis stockée dans le dictionnaire acoustique du système de synthèse. Pour pouvoir être enchaînés, ils doivent tous être enregistrés avec la même voix, à la même hauteur, à la même vitesse, etc. Un contexte phonétique neutralisant de type logatome 8 est généralement utilisé lors de la lecture par le locuteur [12].

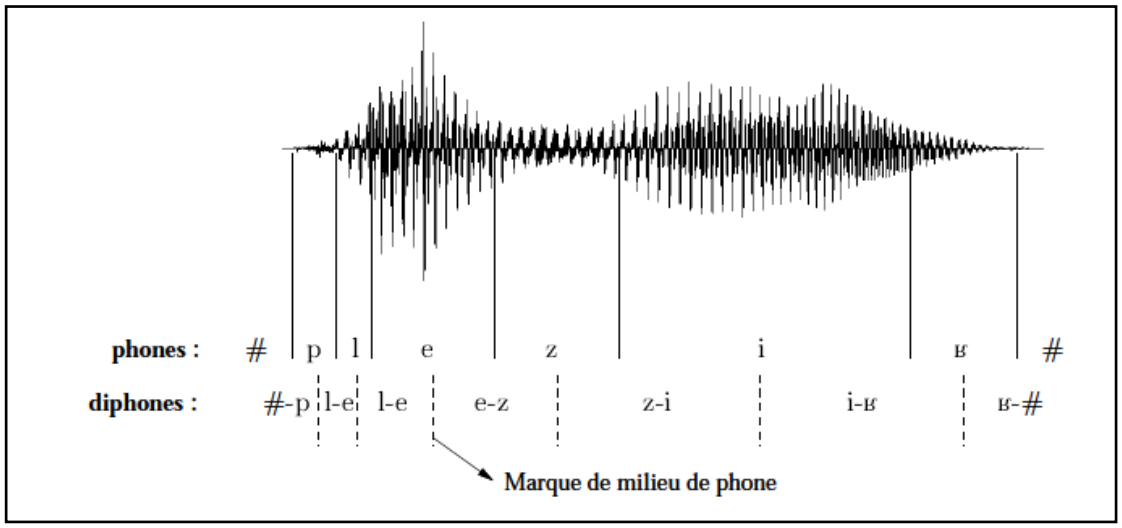


Figure 2.11 : Illustration d'un découpage en diphones, dans le mot plaisir [12]

- Le symbole # est une indication de pause.

Naturellement la parole obtenue par simple concaténation de ces diphones est dépourvue de prosodie : mélodie plate et rythme constant. Des algorithmes de traitement du signal sont alors utilisés pour plaquer une prosodie calculée par les hauts-niveaux. La prédiction de trajectoires prosodiques naturelles constitue alors une première difficulté des systèmes de synthèse par concaténation de diphones, Puis le plaquage des ces trajectoires sur le signal de parole représente un écueil supplémentaire. La méthode LPC était initialement la plus utilisée, bien qu'elle dégradât significativement le signal. La mise au point de la technique PSOLA, plus souple et performante, a changé la donnée par diphones.

2.8.2.5 Synthèse par polysons

La méthode de 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 a 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 a' 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 a' la synthèse par diphones.

La méthode de synthèse par concaténation fournit une parole synthétique de bonne qualité. Elle est utilisée dans un grand nombre de systèmes commerciaux et expérimentaux

tels que les systèmes british Telecomm's Laureate. Proverbe et Hadifix pour l'anglais .France Télécom au CNET a aussi développé des systèmes TTS pour le Français à base de diphones [13].

2.9 CONCLUSION

Nous avons exposé dans ce chapitre les principales méthodes et techniques utilisées dans la synthèse de la parole. La première génération des systèmes de synthèse de la parole avait pour objectif de minimiser le volume de la base de données pour réduire le coût de stockage et de rendre le système de synthèse flexible et facile à adapter pour une autre voix ou une autre langue. Cette flexibilité dépend de l'ensemble des règles qui doivent être élaborées soigneusement, ce qui induit à une complexité très élevée. Avec la génération actuelle, le problème de synthèse s'est réduit à un problème de base de données et d'optimisation de la sélection d'unités. L'objectif est donc de réduire au maximum la modification du signal des unités de synthèse afin de préserver l'aspect naturel de la parole.

Chapitre III :

Analyse acoustique du corpus

3.1 INTRODUCTION

L'objectif de notre travail est de réaliser un système de synthèse de parole par unités variables en vue des Annonces Vocales Automatiques des Stations d'Arrêt du Tramway d'Alger. Ce système fonctionne avec deux langues : Arabe Standard et Française. Ce système fait la production du signal vocal en temps réel en utilisant la méthode de concaténation par phrases et mots combinés.

Dans ce chapitre, nous nous intéressons à l'analyse acoustique des sons pour étudier des caractéristiques de notre corpus (fréquence fondamentale, formants, intensité) et puis nous exposons les grandes lignes introduites dans les étapes de l'élaboration de notre outil d'analyse.

3.2 SYNTHESE PAR CONCATENATION DES PHRASES ET MOTS COMBINES

La synthèse par concaténation d'unités pré-stockées est la génération des sons à partir de la juxtaposition d'un ensemble d'unités préenregistrées. Elle permet des applications concernant les annonces vocales des stations d'arrêt du tramway d'Alger. Le message est constitué d'une phrase porteuse fixe (المحطة التالية … La Prochaine station…) et de deux parties variables constituées par les noms (mots) des différentes stations (Fig. 3.1).

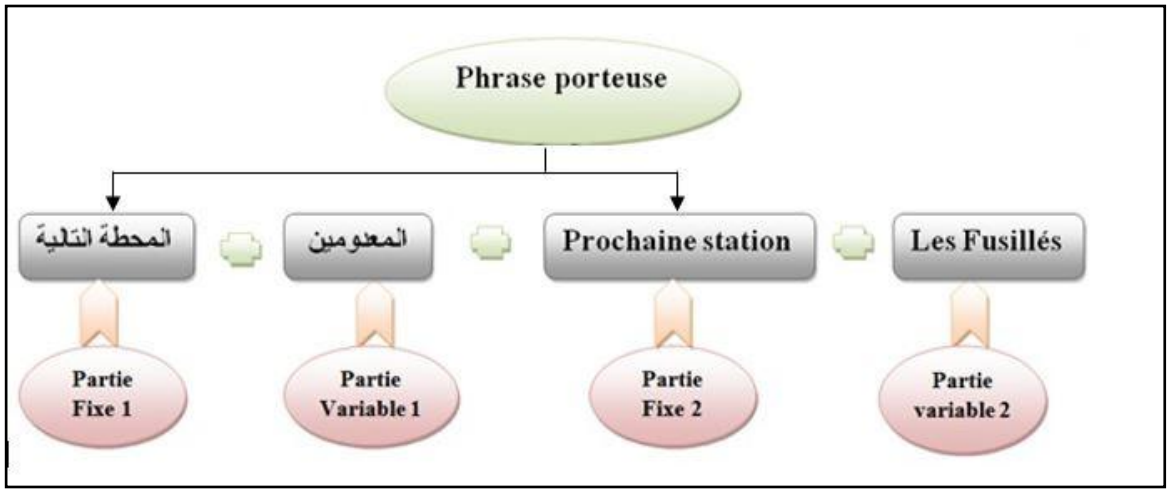


Figure 3.1 : Assemblage des parties fixes avec les parties variables

Dans ce cas, nous pouvons se contenter de faire enregistrer par une locutrice les phrases nécessaires pour former les stations requises (13 en Arabe et 13 en Français). L'enregistrement est numérique avec un codage de 24 bits correspondant à la bonne qualité. Il faudra analyser les mots et les phrases porteuses. Pour synthétiser une annonce,

il suffit de compléter les parties variables de la phrase porteuse avec les paramètres des mots que l'on souhaite émettre.

En effet, la mélodie issue d'un mot isolé n'est pas toujours identique avec ce qu'elle devrait une fois le mot mis en place dans la phrase porteuse. Ce problème peut être résolu soit en enregistrant plusieurs réalisations d'un même mot dans des environnements prosodiques différents, soit en modifiant les paramètres du synthétiseur (notamment la fréquence fondamentale) pour que la mélodie de la phrases reconstituée soit plus naturelle.

La qualité de la parole ainsi obtenue peut être tout à fait acceptable. Elle dépend essentiellement du type de synthétiseur choisi (codage à débit élevé, moyen ou faible). Cependant la qualité de mémoire requise croit linéairement avec la taille du vocabulaire de l'application. On est donc amené à trouver un compromis entre la taille de la mémoire de stockage et la qualité souhaitée.

3.3 L'OUTIL D'ANALYSE

Nous pouvons citer quelques outils qui permettent de visualiser la forme d'ondes et le spectrogramme d'un signal de parole, d'éditer et d'aligner des transcriptions orthographiques et phonétiques sur ce signal, tels que PRAAT, CLAN, Speech Analysis, Goldwave, Cool Edit , etc.[11]

3.3.1 Le logiciel Praat

Le logiciel Praat a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam. Il est un logiciel d'analyse et de transcription phonétique (spectre, intonation, intensité, etc.).

Le logiciel comporte aussi des fonctionnalités importantes pour l'enregistrement, pour la manipulation et pour la synthèse de sons, pour la création d'algorithmes d'apprentissage, pour l'analyse statistique, ainsi que pour diverses expériences auditives. Praat est hautement portable, configurable et programmable. En linguistique interactionnelle, le logiciel est utilisé pour divers types de transcription alignée de données sonores (éventuellement extraits d'une vidéo), pour aligner des transcriptions déjà réalisées en texte brut, mais aussi pour l'analyse et la Transcription Orthographique Phonétique. Avec ce logiciel, il est possible :

- d'enregistrer des fichiers audio qui pourront ensuite être analysés ;

- de transcrire, d'étiqueter et de segmenter des données audio (que les enregistrements aient été effectués sous Praat ou qu'ils proviennent d'autres fichiers, au format WAV, par exemple).
- d'effectuer des analyses phonétiques et acoustiques au niveau segmentai (spectrogramme, analyse de formants, sonagrammes, etc.) et au niveau suprasegmental (pitch ou Fo, intensité et durée).
- de manipuler et modifier le signal de parole (utilisation de filtres, modification des contours intonatifs et de la durée, etc.).
- de faire de la synthèse de la parole (créer des stimuli audio, synthèse articulatoire, analyse - synthèse de données modifiées, etc.).
- de faire des analyses statistiques à partir des études phonétiques (analyses de covariances, etc.). Nous pouvons résumer les fonctionnalités de ce logiciel dans la Figure 3.2 [11].

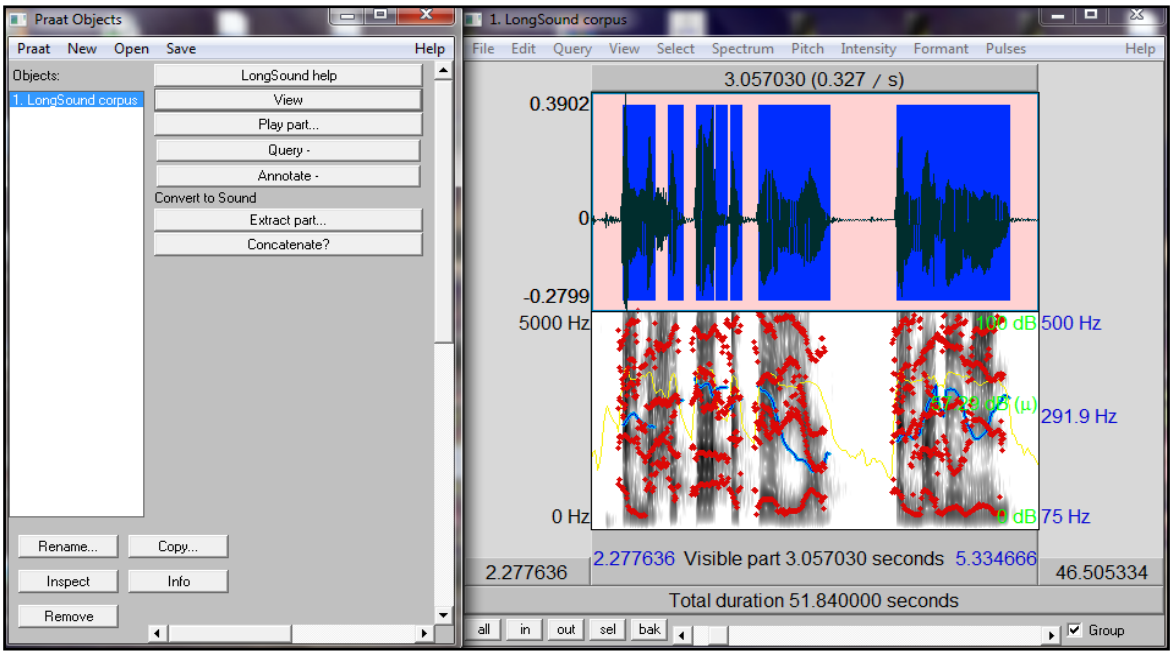


Figure 3.2 : Présentation du logiciel Praat

3.3.2 Visualisations par spectrogramme

On peut aussi visualiser différentes courbes en surimpression sur le spectrogramme :

- la fréquence fondamentale : cochez « Show pitch » dans le menu « Pitch », Sa valeur moyenne (Hz) s'affiche à droite ;

- les formants : cochez « Show formants» dans le menu « Formants », et ils apparaissent en pointillés rouges. Pour les afficher sur toute la longueur de la fenêtre, affichez la fenêtre « Formant Settings » du menu « Formants », et dans le champ « Maximum de la ration », entrez la durée de la fenêtre, en secondes, à la place de la valeur initiale,
- Les périodes du signal sonore : cochez « Show pulses » dans le menu « Pulses ». Chaque période est représentée, sur l'enveloppe, par un trait bleu vertical.

3.4 ELABORATION DU CORPUS

Dans le cadre de notre travail, nous avons utilisé un corpus en parole continue des phrases en Arabe Standard et en Français, prononcées par une locutrice arabophone, ce corpus contient des expressions utilisées pour des annonces vocales automatiques des stations d'arrêt du tramway d'Alger. Ce corpus contient 28 phrases (14 en Arabe Standard et 14 en Français) dont nous avons 2 phrases fixes et 26 phrases variables.
Ce corpus est surnommé **AVAST** : **A**nnonces **V**ocales **A**utomatiques des **S**tations d'arrêt du **T**ramway d'Alger.

Nous justifions le choix de ce type de AVAST (parole continue au lieu de l'utilisation de logatomes) par le fait qu'il est préférable d'étudier les segments dans un continuum vocal pour pouvoir prendre en considération les effets de coarticulation existants entre les phonèmes. Cette base de données prend 51.84 secondes et elle est stockée sur 7.12 Méga Octets (MO) de mémoire.

3.4.1 Enregistrement de corpus

L'enregistrement du corpus a été faite au sein de l'Institut Supérieur des Métiers des Arts du Spectacle et de l'Audiovisuel (ISMAS) d'Alger, avec les conditions d'enregistrement suivantes :

- La fréquence d'échantillonnages : Fe = 48 kHz, le codage : 24 bits ;
- le format : multiple mono (stéréo) ;
- logiciel utilisé Pro Tools version 8 ;
- la chambre d'enregistrement est sans bruit ;
- le type de parole : phrases continu en arabe ;

- les signaux acoustiques sont enregistrés en format (WAV).

3.4.2 Equipement utilisés en enregistrement

Le matériel utilisé est :

- Microphone professionnel unidirectionnel Electrodynamique [Beyer dynamic M 69 TG] (fig. 3.5) ;
- Station Pro Tools Version 8 et de Bonne qualité ;
- une cabine Speaker : c'est une chambre isolé et contient des Microphones et des casque, séparé à la cabine technique avec un verre transparent et isolant (fig. 3.5) ;
- une cabine technique : contient une table de mixage, une carte d'acquisition, un micro-ordinateur, des Haut-parleurs, des Microphones (fig. 3.5).

Figure 3.3: Microphone Beyer dynamic M 69 TG

Figure 3.4: Station Pro Tools

Figure 3.5: Cabine Speaker + cabine technique

Le corpus que nous avons enregistré est présenté sur Tableau 3.1.

Tableau 3.1 : les phrases et les mots du corpus AVAST

المحطة التالية :	Prochaine station :
المعدومين	Les Fusillés
طرابلس	Tripoli
خروبة	Caroubier
لا قلاسيار	La Glacière
الديار الخمس	Cinq Maisons
الصنوبر	Les Pins
حي مختار زرهوني	Cité Mokhtar Zerhouni
حي رابية	Cité Rabia
جامعة باب الزوار	Université Bab Ezouar
باب الزوار الجسر	Bab Ezzouar Le Pont
برج الكيفان الثانوية	Bordj El Kiffan Lycée
موحوس	Mouhous
ميموني حمود	Mimouni Hamoud

A l'aide de l'outil Praat, nous obtenons les sonagrammes du signal vocal des AVAST (Fig. 3.6)

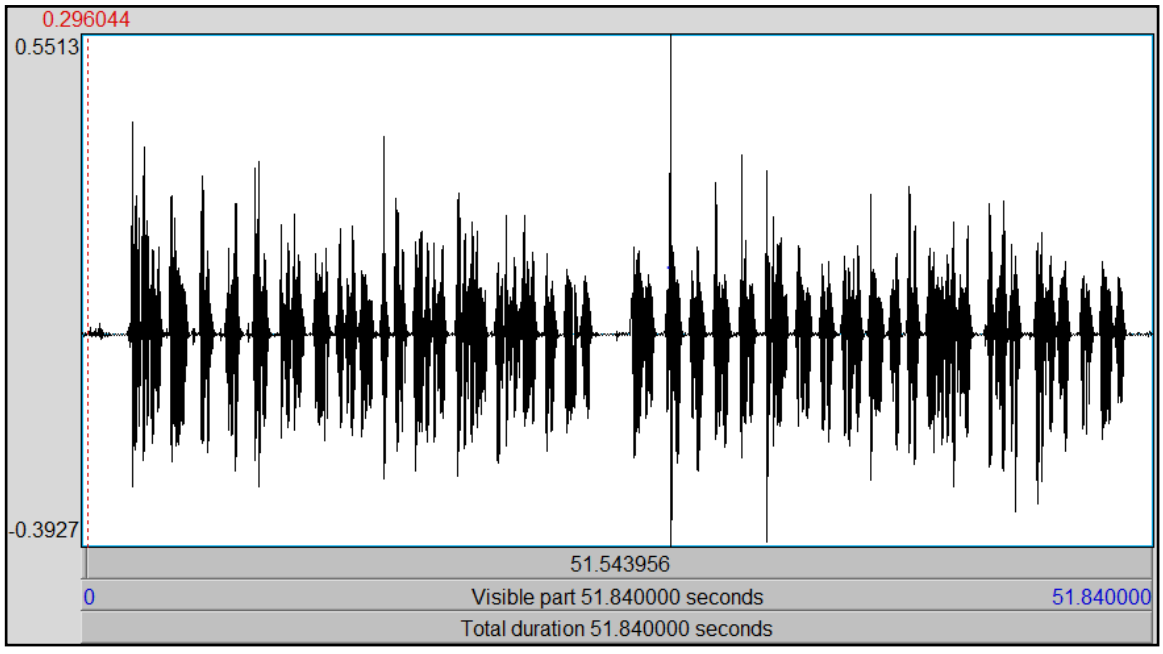


Figure 3.6: visualisation de signale audio des AVAST

3.4.3 Procédure de segmentation en phrases et mots

Les fichiers son segmentés en utilisant notre logiciel Praat, l'image du spectrogramme, ainsi que le fichier correspondant aux formant seront stockés dans des répertoires. La segmentation du corpus a été effectué manuellement ce qui justifie le temps, relativement long, alloué à cette opération. La procédure adoptée pour isoler les phonèmes à dégager l'unité à étudier de l'onde temporelle, et à effectuer des tests de perception pour s'assurer de la qualité de la segmentation (fig. 3.7).

La synthèse de parole par concaténation en forme d'ondes est une déclaration des segments sonores enregistrés, La combinaison entre ces segments impose de faire l'insertion de pauses silencieuses d'une durée appropriée entre les segments. Cette technique élimine les bruits et augmenter la qualité de parole synthétisée. L'opération d'insertions des pauses silencieuses a été faite par le logiciel Praat, lorsque nous faisons la segmentation des AVAST (fig. 3.8).

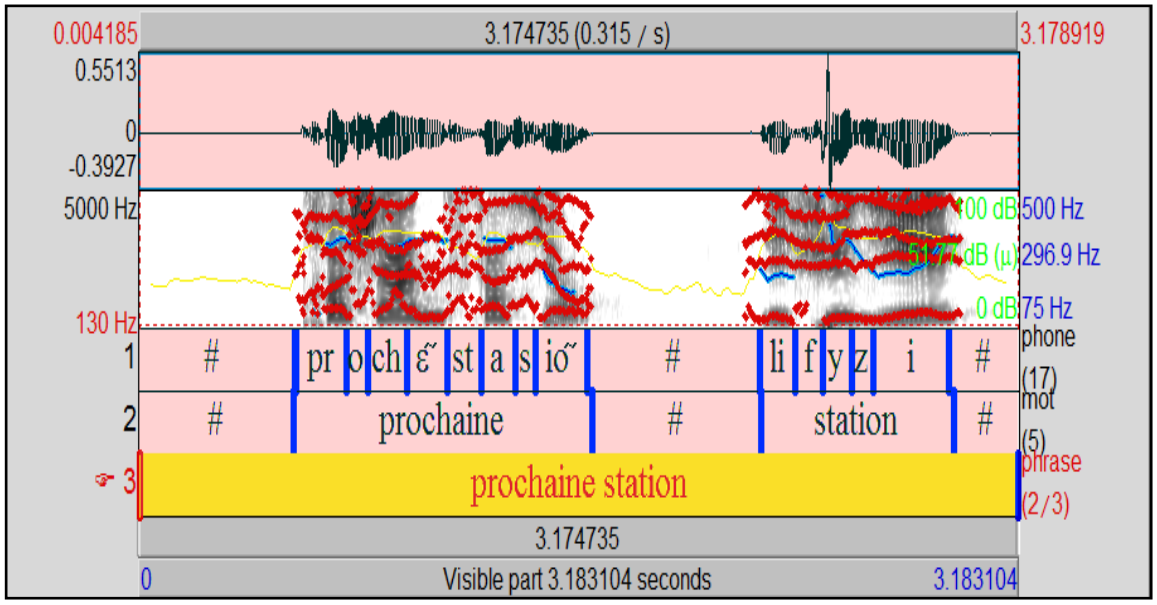


Figure 3.7: Segmentation de la phrase [prochaine station les fusillés] en phonèmes et mots

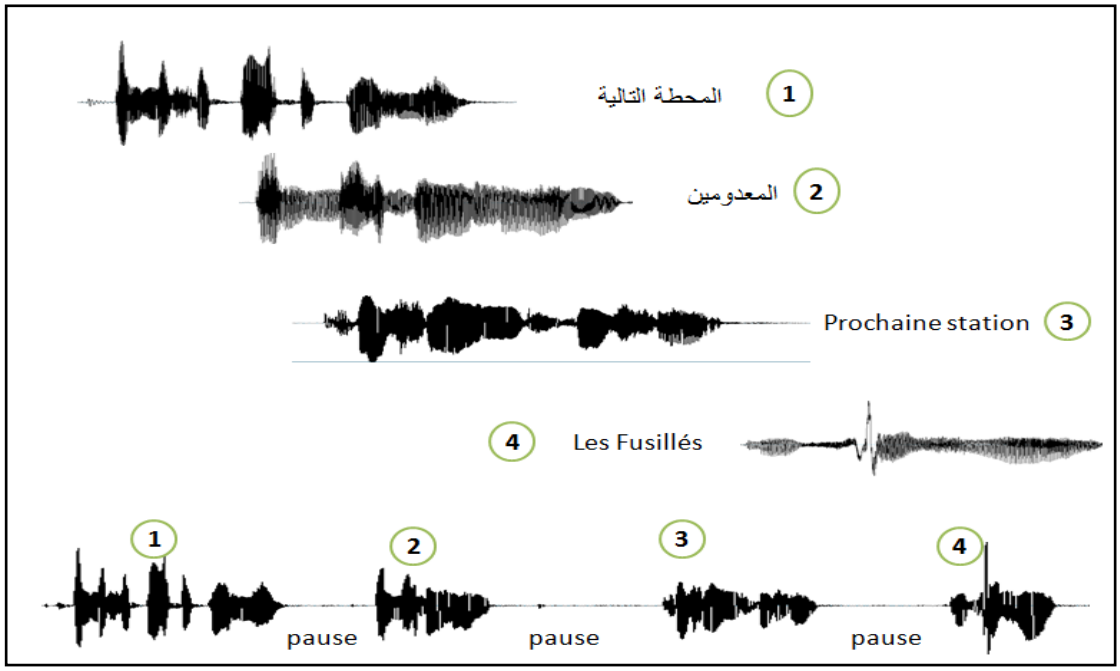


Figure 3.8 : Concaténation par forme d'ondes de la phrase Ph_7 avec Praat

Enfin, chaque tranche qu'on a déterminée nous vallons sélectionner et enregistrer sous format wav (Figure. 3.9).

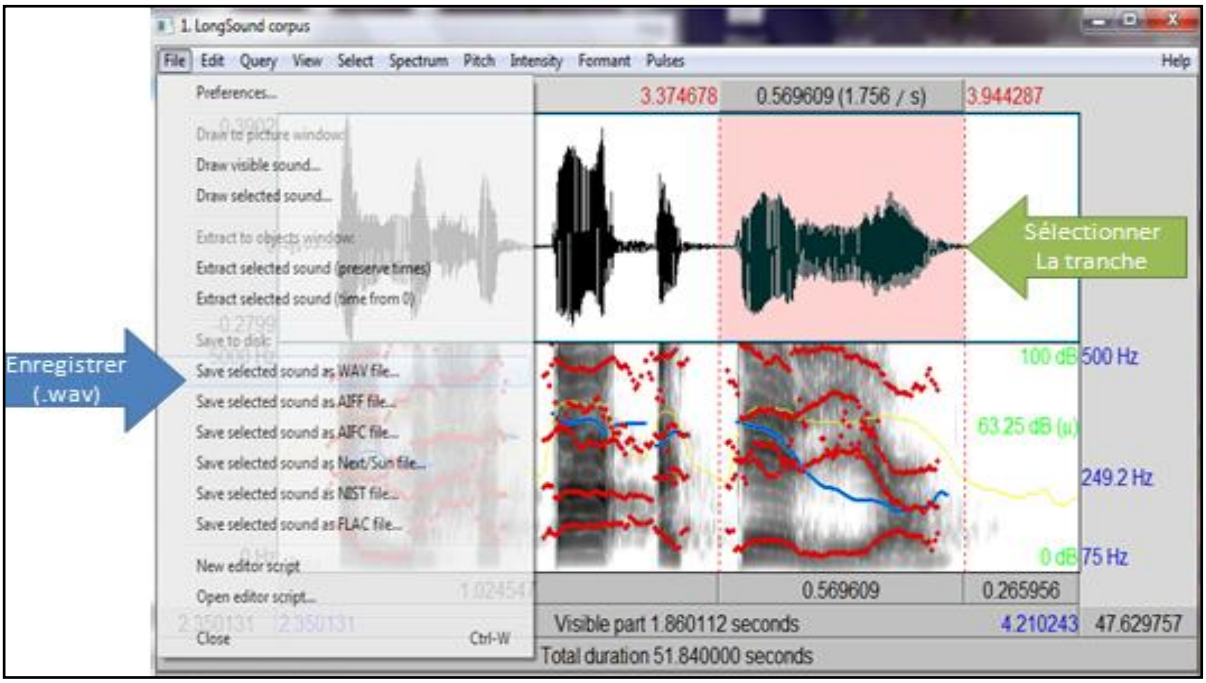


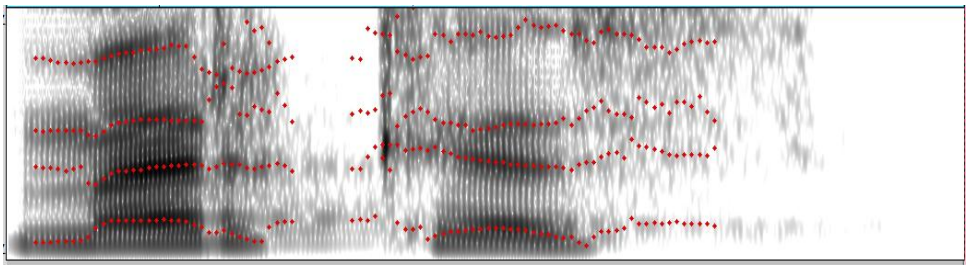
Figure 3.9 : Procédure de segmentation des AVAST

3.5 ANALYSE PAR SPECTROGRAMMES

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, et est devenu l'outil incontournable des études en phonétique pendant de nombreuses années [9].

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 bandes 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 y 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 y apparaissent sous la forme de bandes horizontales [9]. Avec un axe des abscisses de temps en millisecondes et des ordonnées des fréquences en Hz et l'intensité est donné par le degré de noirceur de la trace. (Fig.3.10).

a)



b)

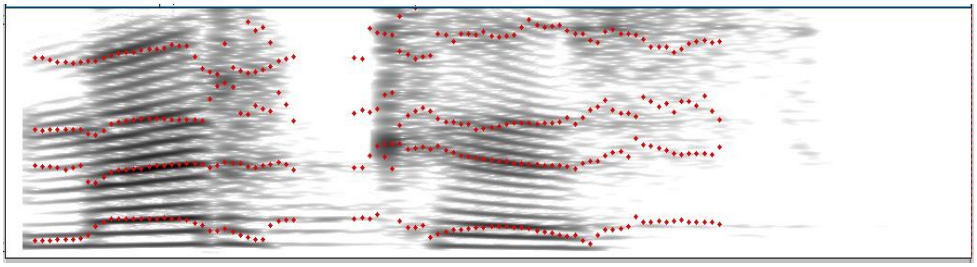


Figure 3.10 a, b: Spectrogramme de la phrase /prochaine station/ [prɔʃ$\tilde{\varepsilon}$ statiO$\tilde{}$] :

a) en bande large avec une fenêtre de Hamming de 5 ms ;
b) en bande étroite avec une fenêtre de Hamming de 30 ms.

3.5.1 Lecture de spectrogramme

La lecture de spectrogramme contient 4 étapes élémentaires :

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

Étape 2 : Savoir distinguer les consonnes et les voyelles :

- les consonnes sont des sons produits avec une constriction plus ou moins forte dans le conduit vocal. L'intensité du spectre est relativement faible et sur le spectrogramme sa noirceur n'est pas très forte ;
- alors que les voyelles sont des sons produits sans aucune constriction forte dans le conduit vocal, l'intensité du spectre est relativement élevée et sur le spectrogramme sa noirceur est relativement foncée.

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

- les occlusives sont produites par une occlusion complète dans le conduit vocal, donc pendant l'occlusion, l'air ne passe pas et sur le spectrogramme. Il correspond à un silence (sauf le voisement pour les sonores) ;
- les fricatives sont produites avec une forte constriction (mais pas complète) dans le conduit vocal. Il y a une turbulence de l'air dans le conduit vocal et sur le spectrogramme cette turbulence correspond au bruit de friction ;
- les sonantes [m, n, l, R] sont produites avec une constriction partielle dans le conduit nasal et vocal. L'air passe d'une façon relativement libre et sur le spectrogramme il y a des formants comme les voyelles, mais ces formants sont moins forts que ceux des voyelles ;
- il y a deux types pour les occlusives et les fricatives : sourdes et sonores. Pour les occlusives et les fricatives sonores, les cordes vocales vibrent alors sur le spectrogramme, ils présentent une barre de voisement. Tandis que, les cordes

vocales des occlusives et des fricatives sourdes ne vibrent pas, donc sur le spectrogramme il n'y a pas de barre de voisement.

Étape 4 : Savoir reconnaître les grandes classes de voyelles. Les voyelles se différencient les unes les autres par leurs formants. Un formant est la zone de fréquence où il y a une concentration (renforcement) d'énergie. Dans les voyelles orales, il y a en moyenne un formant par 1000 Hz (voix d'Homme). On utilise souvent le spectrogramme à bande large pour visualiser les formants et ces derniers y apparaissent sous les formes des bandes noires horizontales. Les voyelles orales sont divisées en des classes :

- les voyelles antérieures, la distance entre F_1 - F_2 est supérieure à la distance entre F_2 - F_3 ;
- les voyelles postérieures, la distance entre F_1 - F_2est inférieure à la distance entre F_2 - F_3 ;
- les voyelles centrales, les formants sont plus (ou moins) équidistants.

3.6 ETUDE DE LA PERFORMANCE DE LA CONCATENATION SUR AVAST

Cette étude représente une analyse comparative sur l'AVAST avant (signal vocal original) et après la concaténation (signal synthétique) ; en prenant des échantillons qui sont neuf phrases, pour les phrases avant concaténation nous avons fait un enregistrement en parole continue et pour les mêmes phrases après concaténation nous avons fait un réenregistrement du signal de sortie du notre interface graphique en temps réel avec les mêmes conditions d'enregistrement de l'AVAST. Cette étude représente un test objectif de l'AVAST du côté de l'intelligibilité et l'aspect naturel de la parole synthétique.

Cette étude concerne une analyse comparative des paramètres prosodiques de notre corpus. La comparaison sera faite avant et après la concaténation, pour cela nous allons étudier la performance de la concaténation des neuf phrases avec leur **T**ranscriptions **O**rthographiques **P**honétiques (**TOP**) (Tableau 3.2).

Tableau 3.2 : Transcriptions Orthographiques Phonétiques

	Les phrases a analysé	TOP
Ph_1	المحطة التالية المعدومين	[ʔlmaħaṭ ṭa ʔtaalija ʔlmaʕḏumiin]
Ph_2	المحطة التالية طرابلس	[ʔlmaħaṭ ṭa ʔtaalija ṭarablus]
Ph_3	المحطة التالية خروبة	[ʔlmaħaṭ ṭa ʔtaalija xarruuba]
Ph_4	Prochaine station Les Fusillé	[prɔʃ̃ɛ statiÕ li fyzi]
Ph_5	Prochaine station Tripoli	[prɔʃ̃ɛ statiÕ tripɔli]
Ph_6	Prochaine station Caroubier	[prɔʃ̃ɛ statiÕ carɔbj]
Ph_7	المحطة التالية المعدومين Prochaine station les Fusillés	[ʔlmaħaṭ ṭa ʔtaalija ʔlmaʕḏumiin prɔʃ̃ɛ statiÕ li fyzi]
Ph_8	المحطة التالية طرابلس Prochaine station Tripoli	[ʔlmaħaṭ ṭa ʔtaalija ṭarablus prɔʃ̃ɛ statiÕ tripɔli]
Ph_9	المحطة التالية خروبة Prochaine station Caroubier	[ʔlmaħaṭ ṭa ʔtaalija xarruuba prɔʃ̃ɛ statiÕ carɔbj]

Pour notre étude nous avons choisi un des échantillons parmi les 9 phrases en AS et en Français avant et après la concaténation. Le but de cette comparaison est l'étude de la qualité et la performance de parole obtenue par concaténation par rapport à celle d'origine. La comparaison sera basée sur l'analyse :

- générale : concernant la taille, durée, etc ;
- Formantique ;
- de l'intensité ;
- fréquentielle.

À l'aide de la distance euclidienne nous calculons la précision pour l'étude comparative concernant les neuf phrases telles que :

La distance euclidienne : d = $|(valeur\ avant) - (valeur\ après)|$

Précision (%) = $\frac{|(valeur\ \ avant) - (valeur\ \ après)|}{valeur\ \ avant}$ x 100

3.6.1 Analyse générale de l'AVAST

Cette analyse correspond aux différentes : tailles, durées, amplitudes, énergies et puissances moyennes (Tableau 3.4).

Tableau 3.3 (a, b) : Les paramètres généraux du nos phrases avant et après concaténation

a)

	Les phrases									
Type de comparaison	$\mathbf{Ph_{1av}}$	$\mathbf{Ph_{1ap}}$	$\mathbf{Ph_{2av}}$	$\mathbf{Ph_{2ap}}$	$\mathbf{Ph_{3av}}$	$\mathbf{Ph_{3ap}}$	$\mathbf{Ph_{4av}}$	$\mathbf{Ph_{4ap}}$	$\mathbf{Ph_{5av}}$	$\mathbf{Ph_{5ap}}$
Taille [Ko]	508	257	237	233	248	222	448	379	277	266
durée [secondes	3.60	2.73	2.52	2.51	2.63	2.62	3.18	2.94	1.45	1.40
amplitude Min [pascal]	-0.279	-0.279	-0.279	-0.278	-0.279	-0.270	-0.3926	-0.392	-0.225	-0.223
amplitude Max [pascal]	0.390	0.390	0.390	0.393	0.390	0.389	0.5513	0.5513	0.1600	0.1666
Energie [$\times \mathbf{10^{-5}}$J/$\boldsymbol{m^2}$]	1.542	1.542	1.332	1.337	1.388	1.379	1.373	1.733	0.9741	0.9678

b)

	Les phrases							
	Ph_{6av}	Ph_{6ap}	Ph_{7av}	Ph_{7ap}	Ph_{8av}	Ph_{8ap}	Ph_{9av}	Ph_{9ap}
La Taille [Ko]	275	267	533	499	509	487	522	500
La durée [seconde]	2.92	2.90	5.67	5.39	5.42	5.38	5.56	5.45
L'amplitude Min [pascal]	-0.2259	-0.2200	-0.3926	-0.3926	-0.2799	-0.2799	-0.2799	-0.2799
L'amplitude Max [pascal]	0.2787	0.2799	0.5513	0.5417	0.3909	0.3922	0.3902	0.3956
Energie [$\times \mathbf{10^{-5}}$J/$\boldsymbol{m^2}$]	0.9451	0.9367	2.9160	2.9189	2.3831	2.3830	2.3880	2.3831

La Figure 3.11 illustre la variation de la précision sur la taille, la durée, l'amplitude.

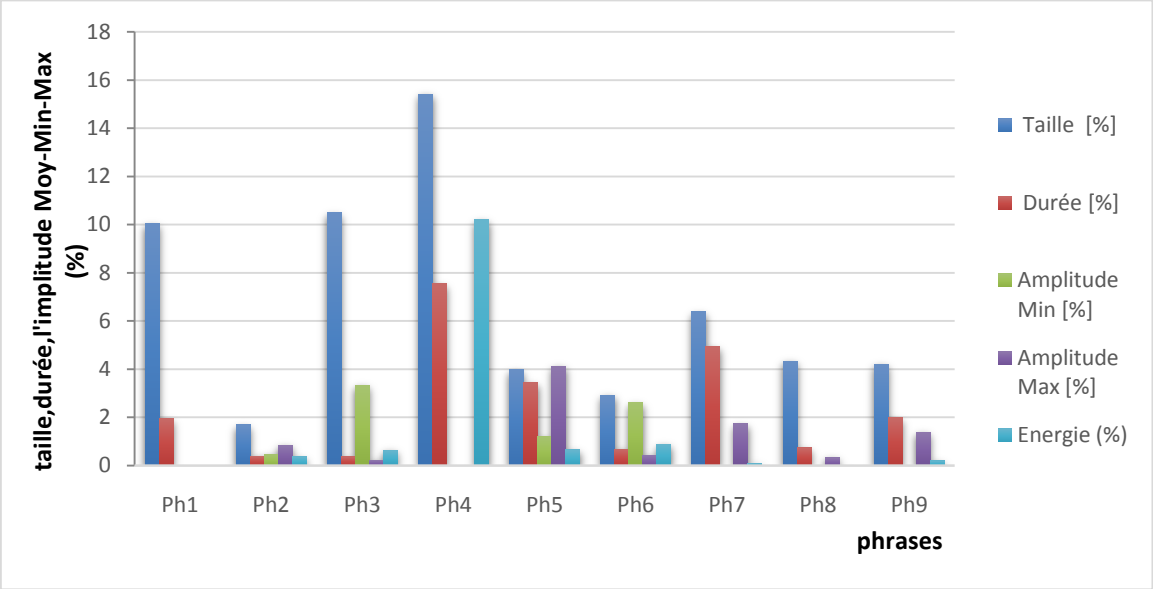


Figure 3.11 : Illustration de la précision de la taille, durée et l'énergie

3.6.2 Extraction des formants moyens

A l'aide de l'outil Praat, Nnous pouvons extraire les valeurs moyennes des formants ($F_1, F_2, \ldots, F_5$) Pour les neuf phrases avant et après la concaténation (Tab. 3.4).

Tableau 3.4 (a, b): Comparaison formantique

a)

	Les phrases									
	Ph_{1av}	Ph_{1ap}	Ph_{2av}	Ph_{2ap}	Ph_{3av}	Ph_{3ap}	Ph_{4av}	Ph_{4ap}	Ph_{5av}	Ph_{5ap}
Formants [Hz]										
F_1	739.13	712.75	781.42	777.94	809.02	805.44	774.08	722.99	851.85	847.74
F_2	1870.02	1883.55	1850.09	1853.02	1769.00	1766.79	2112.08	2030.53	1984.11	1979.32
F_3	2860.67	2805.75	2794.07	2796.46	2781.92	2779.95	3112.94	3051.91	3067.23	3065.55
F_4	3921.80	3855.44	3875.54	3876.21	3887.61	3882.36	4203.67	4187.17	4142.41	4133.36
F_5	4648.77	4643.03	4689.36	4689.14	4645.30	4644.58	4677.77	4850.68	4788.62	4782.36

b)

Formants [Hz]	Les phrases							
	Ph_{6av}	Ph_{6ap}	Ph_{7av}	Ph_{7ap}	Ph_{8av}	Ph_{8ap}	Ph_{9av}	Ph_{9ap}
Formants [Hz]								
F_1	797.57	795.32	780.72	768.02	822.52	825.36	828.88	827.87
F_2	1932.13	1928.14	1996.42	1999.32	1931.07	1936.25	1869.94	1859.47
F_3	3022.69	3011.88	2965.61	2964.01	2948.10	2923.32	2913.94	2902.03
F_4	4105.29	4102.36	4030.32	4025.36	4018.07	4018.25	3996.80	3999.98
F_5	4786.27	4782.34	4700.42	4656.28	4728.88	4756.36	4703.92	4702.06

La Figure 3.12 présente la variation de la précision sur les différents formants.

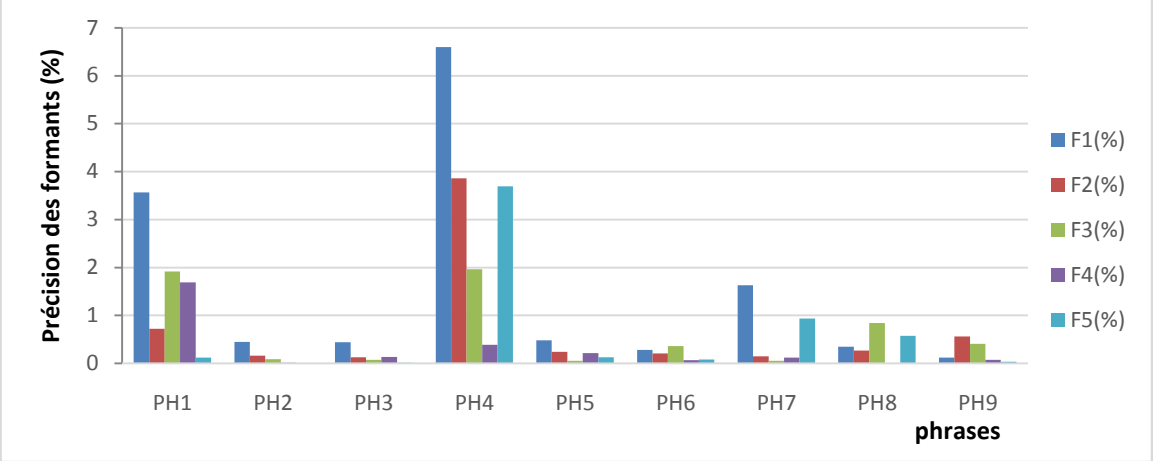
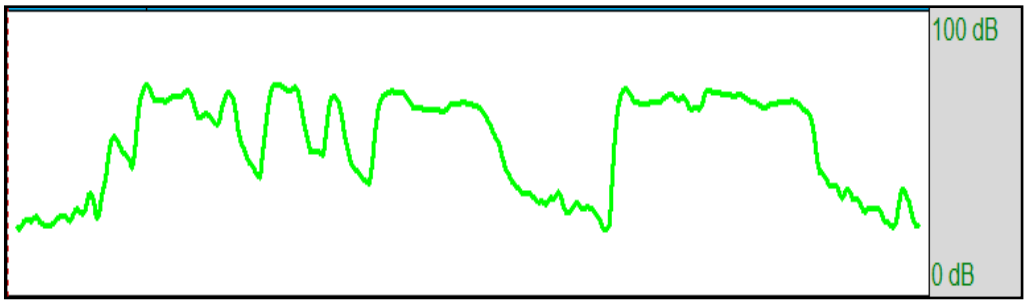


Figure 3.12 : Précision des formants

3.6.3 L'intensité et Le Gain

Nous analysons l'intensité, puis nous extractions les propriétés de chaque phrase et la comparaison sera basée sur la précision de : le Gain, l'intensité Min-Max-Moyenne. Nous pouvons aussi faire la comparaison avec la visualisation du diagramme de l'intensité, nous prenons, à titre exemple, la phrase Ph_1, nous trouvons une grande similarité entre les deux diagrammes de l'intensité (Fig. 3.13).

a)



b)

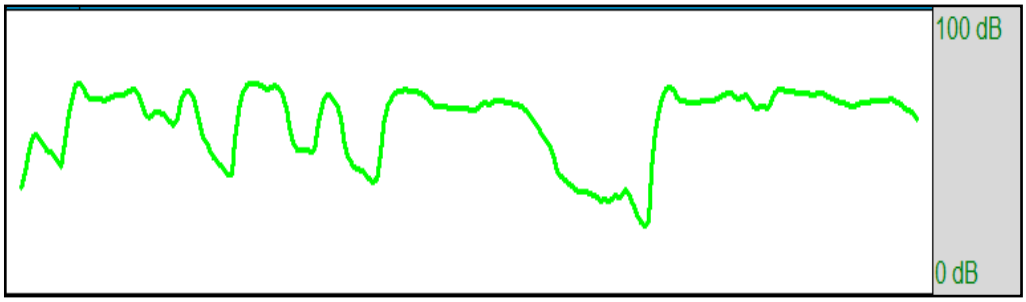


Figure 3.13 (a, b) : Diagramme de l'intensité de la phrase Ph_1 :

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.

Nous allons donner les différentes valeurs qui caractérisent le Gain et l'intensité dans le tableau 3.5.

Tableau 3.5 (a, b): comparaison par, le Gain, l'intensité Moyenne-minimum-maximum

a)

	Les phrases									
	Ph_{1av}	Ph_{1ap}	Ph_{2av}	Ph_{2ap}	Ph_{3av}	Ph_{3ap}	Ph_{4av}	Ph_{4ap}	Ph_{5av}	Ph_{5ap}
Type de comparaison										
Gain $[\times \mathbf{10^{-6}}]$	781	778	365.4	366	776.6	773	1066.5	1059.9	1551	1546
Intensité Moyenne (dB)	60.81	58.72	59.94	58.97	59.84	59.09	53.12	52.67	52.06	51.98
Intensité Min (dB)	23.31	23.65	31.88	31.65	29.77	28.95	23.62	24.86	24.87	24.56
Intensité Max (dB)	74.54	74.45	74.44	73.02	74.44	73.25	81.54	81.56	73.78	73.23

b)

	Les phrases							
	Ph_{6av}	Ph_{6ap}	Ph_{7av}	Ph_{7ap}	Ph_{8av}	Ph_{8ap}	Ph_{9av}	Ph_{9ap}
Type de comparaison								
Gain $[\times \mathbf{10^{-6}}]$	1590.2	1588.6	1614.67	1601.4	1081.25	998.93	2465.96	2397.86
Intensité Moyenne (dB)	52.57	52.04	56.34	55.86	55.19	54.7	55.71	54.31
Intensité Min (dB)	24.88	24.47	23.95	23.75	24.86	24.11	24.86	24.03
Intensité Max (dB)	73.78	73.14	81.56	81.21	74.44	73.99	74.45	74.01

La Figure 3.14 montre la variation de la précision sur le Gain et l'intensité des phrases Ph_1,…, Ph_9.

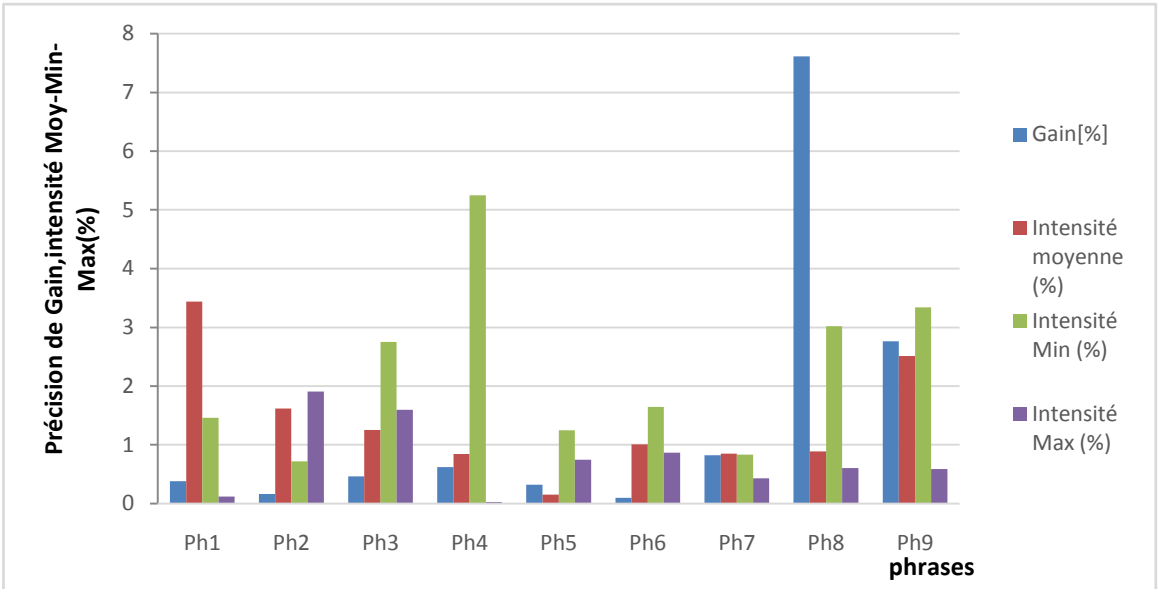
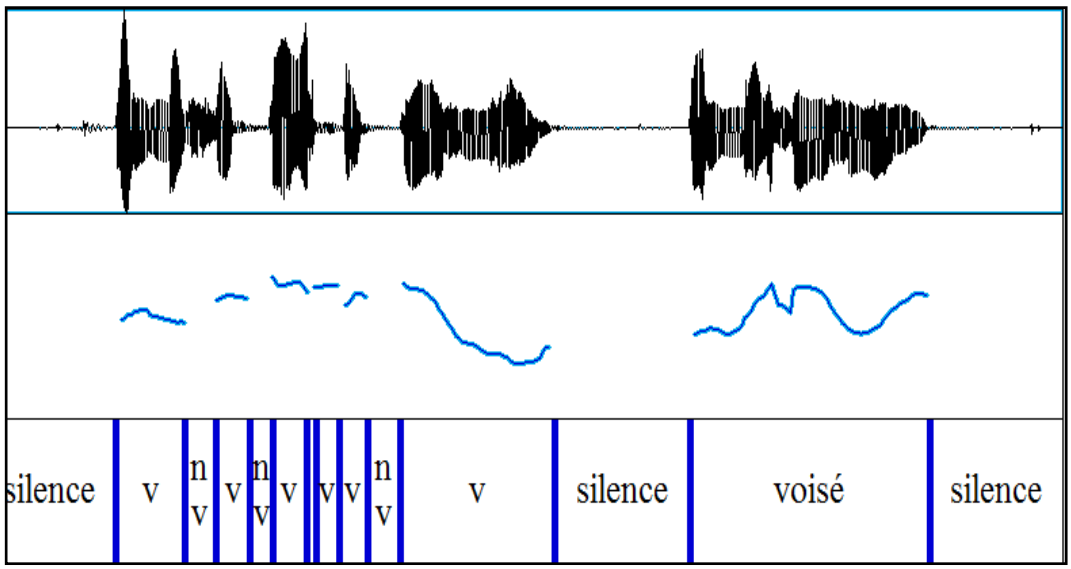


Figure 3.14 : Précision du Gain et des intensités moyenne-Min-Max

3.6.4 Analyses fréquentielles des AVAST (Pitch)

Nous avons fait l'analyse par visualisation de la variation de pitch (de mélodie) pour la phrase Ph_1, puis nous citons les zones des sons voisées et non voisées (Fig. 3.15).

a)



b)

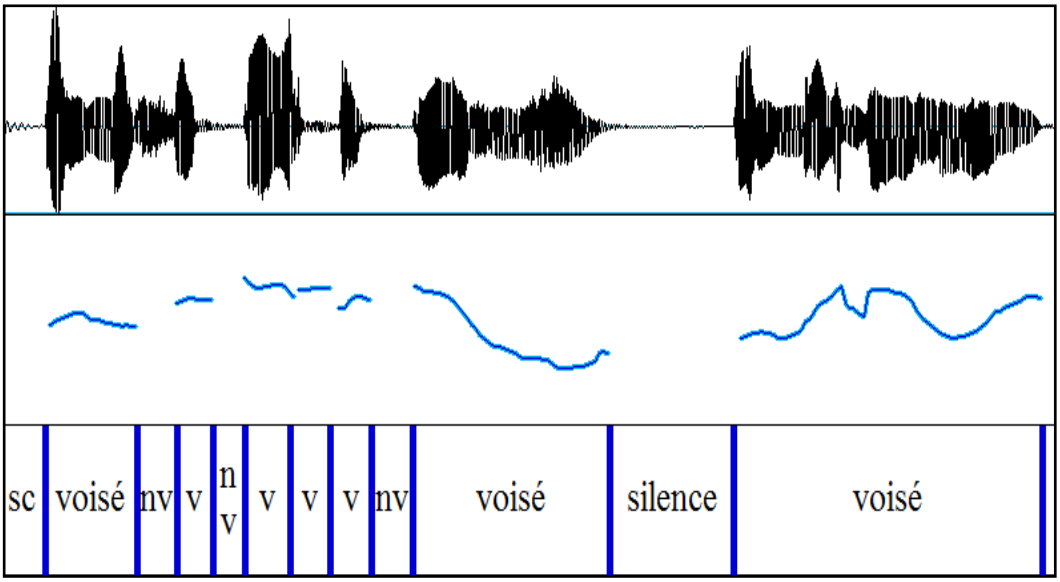


Figure 3.15 (a, b) : la mmélodie de la phrase Ph_1 :

a) Avant concaténation;
b) Après concaténation

Avec : **V** : Voisé, **NV** : Non Voisé, **SC** : Silence

Nous avons présenté les paramètres de la variation de la fréquence fondamentale des phrases Ph_1,…, Ph_9 dans le tableau 3.6 (a, b).

Tableau 3.6 (a, b) : les paramètres de la variation du pitch

a)

	Les phrases									
Type de comparaison	Ph_{1av}	Ph_{1ap}	Ph_{2av}	Ph_{2ap}	Ph_{3av}	Ph_{3ap}	Ph_{4av}	Ph_{4ap}	Ph_{5av}	Ph_{5ap}
F_0 MIN (Hz)	123.31	123.65	188.18	187.14	89.16	88.99	123.62	124.86	121.19	120.20
F_0 MAX (Hz)	374.54	374.45	366.90	365.64	367.69	366.95	481.54	481.56	450.59	448.47
F_0 moyenne (Hz)	291.75	292.26	291.13	289.93	284.54	282.91	295.95	293.98	302.18	300.86
Nombres des zones de son voisé	7	7	6	6	8	8	7	7	8	8
Nombre des zones de son non-voisé	8	8	6	6	6	6	8	8	6	6

b)

	Les phrases							
	Ph_{6av}	Ph_{6ap}	Ph_{7av}	Ph_{7ap}	Ph_{8av}	Ph_{8ap}	Ph_{9av}	Ph_{9ap}
Type de comparaison								
F_0 MIN (Hz)	119.68	118.87	119.68	118.11	121.19	119.95	88.84	87.77
F_0 MAX (HZ)	359.15	355.54	483.71	381.97	470.62	468.61	369.50	368.17
F_0 moyenne (Hz)	293.87	292.59	291.58	290.05	295.74	293.90	288.26	289.09
Nombres des zones de son voisé	07	07	14	14	15	15	14	14
Nombre des zones de son non-voisé	05	05	10	10	11	11	10	10

La Figure 3.16 illustre la variation de la précision sur la fréquence fondamentale et le nombre des zones de son voisé et non voisé au niveau des phrases Ph_1,…, Ph_9 .

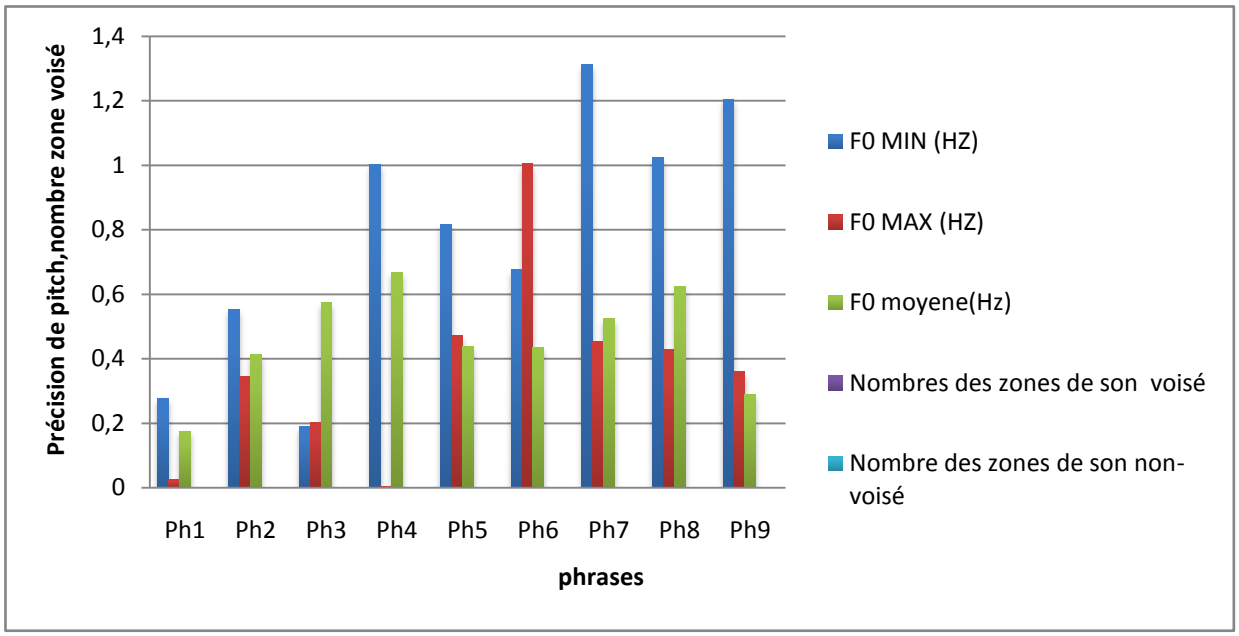


Figure 3.16 : Précision de pitch, nombre des zones voisées et non voisées

3.6.5 Interprétation générale

Nous remarquons pour chaque phrase des AVAST, que les variations des paramètres est acceptable, nous avons obtenu une bonne précision concernant la taille comprise entre 1% et 10 % sauf pour la phrase Ph_4, elle a une mauvaise précision de 15% .pour les autres paramètres nous avons aussi obtenu de bonnes précisions qui n'avaient pas dépassé 10% pour l'énergie et 7% pour la durée. Pour les formants. Nous remarquons que nous avons obtenu une meilleure précision qui ne dépasse pas 6.5%, Par conséquent, nous avons une grande similarité entre ces formants avant et après la concaténation.

Nous constatons qu'il y a une grande similarité pour le gain ; intensité moyenne et intensité maximum avant et après concaténation avec une variation pour l'intensité minimum ce qui justifie la bonne qualité de parole obtenue. Pour la fréquence fondamentale, nous avons une grande similarité pour les 5 paramètres de F_0, avant et après concaténation.

3.7 CONCLUSION

Le but de ce chapitre est de faire une analyse acoustique de notre corpus 'AVAST' en passant par plusieurs étapes d'analyse et visualisation. Nous avons fait des tests objectifs du signal original et le synthétisé. 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 préenregistrées est capable de produire des discours de haute qualité se rapprochant de naturel.

Chapitre VI :

Annonces vocales Automatiques des stations d'arrêt du tramway d'Alger

4.1 INTRODUCTION

Dans ce chapitre nous allons présenter une simulation d'un **S**ystème d'**A**nnonces **V**ocales Automatiques des **S**tations d'Arrêt du **T**ramway d'**A**lger (**SAVSTA**), avec une interface graphique qui est présentée dans l'environnement visuel studio, et finissons par des tests de perception subjective afin de pouvoir évaluer les résultats obtenus (signal vocal de sortie de l'interface) ce qu'il concerne l'intelligibilité et l'aspect naturel.

4.2 PRESENTATION DE LA SIMULATION D'ANNONCES VOCALES AUTOMATIQUES DES STATIONS D'ARRET DU TRAMWAY D'ALGER

Le Tramway est un moyen de transport d'une importance et d'une utilité primordiale pour les passagers, et grâce à lui le déplacement devient rapide et plus facile que les autres moyens de transport. Nous avons fait les annonces vocales pour attirer l'attention des voyageurs sur les stations d'arrêt prochaines.

L'objectif de notre travail est réalisé un **S**ystème parlant qui fait des **A**nnonces **V**ocales Automatiques des **S**tations d'arrêt du **T**ramway d'**A**lger (**SAVSTA)**, ces annonces seront faites pour des stations prochaines avant d'arriver, c'est-à-dire avant l'arrivée à la station désirée. Notre système va se déclencher automatiquement par un lancement de signal vocal qui annonce le nom de cette station, et en parallèle afficher les noms de ces stations en Arabe Standard et en Français.

4.3 REALISATION DU SAVSTA

La réalisation de ce projet a été faite sous l'environnement visuel studio en langage de programmation C#, et ce projet est basé sur plusieurs étapes que nous illustrons dans l'algorithme de ce projet.

4.3.1 ALGORITHME DE SIMULATION DU SAVSTA

La complexité de notre projet nécessite de faire un algorithme, ce dernier est basé sur :

- L'enregistrement du corpus ;
- Les segmentations sonores ;
- La base de données ;
- La sélection des segments appropriés ;
- La sélection des segments choisis ;

- Les annonces vocales (sortie orale).

L'enregistrement et segmentation sonores de corpus AVAST ont été expliqués dans le Chapitre 3. La segmentation du corpus implique d'ajouter des liens des fichiers sonores, chaque fichier de son doit être appelé par un lien (référence) pour être bien défini sur la base de données.

Les sélections des segments appropriés : les fichiers de sons doivent être classés en deux types de catégories, des parties une en Arabe Standard et l'autre en Français.

Chaque catégorie est divisée en deux :

- Phrase porteuse ;
- mots variables.

Dans la Sélection des segments choisis : nous allons classer les fichiers sonores de telle manière que chaque étape (station) nous définissons la partie variable, la phrase porteuse reste fixe. Nous avons présenté l'algorithme du SAVSTA dans la Figure 4.1.

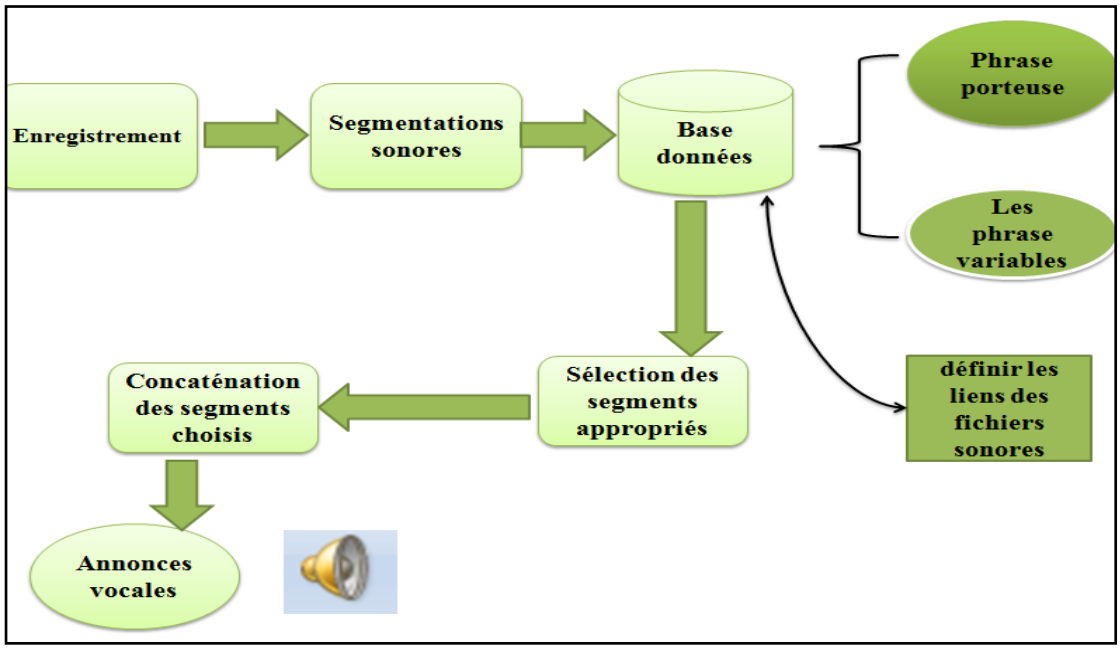


Figure 4.1 : Algorithme du SAVSTA

4.3.2 Les données du projet SAVSTA

Dans le tableau 4.1, nous avons présenté les stations du chemin de l'aller et du retour de tramway d'Alger. Nous avons 13 stations d'arrêt du tramway (en Arabe Standard et en Français), et les distances d_i entre deux stations successives.

Les points rouges que nous avons présentés dans la figure 4.2 sont les moments de lancement du signal vocal pour faire les annonces vocales des stations prochaines. Nous avons représenté l'emplacement des stations d'arrêt du tramway d'Alger dans la Figure 4.2.

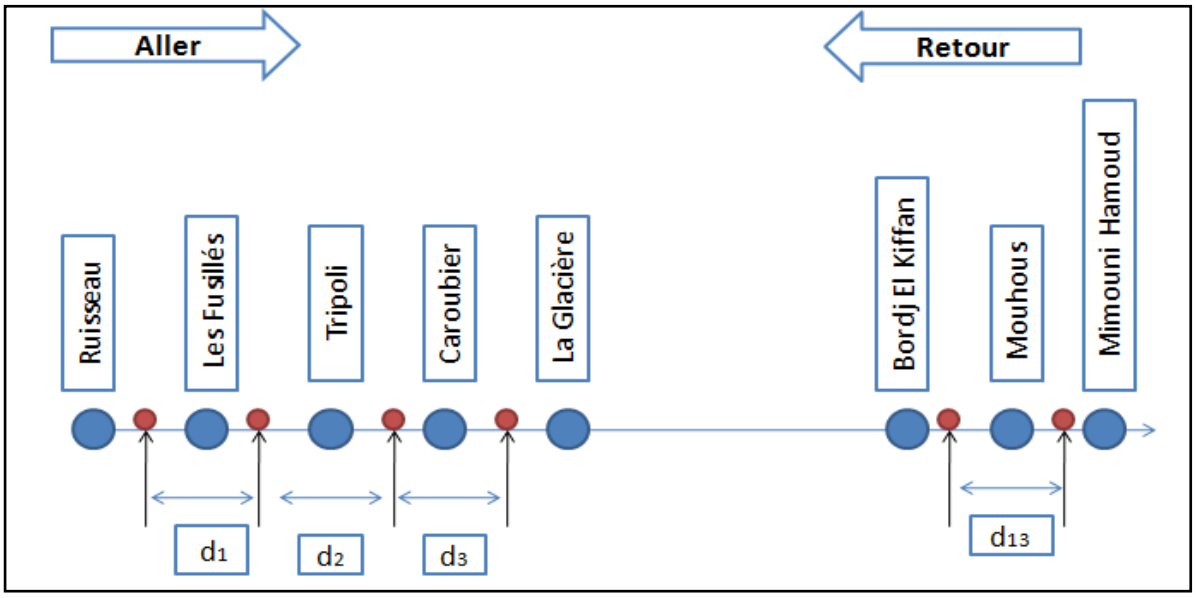


Figure 4.2 : Représentation des emplacements des stations sur le trajet du tramway d'Alger

S_i **:** sont des liens (références) des fichiers sonores des stations d'arrêt du tramway :

- S_1 **:** Un fichier sonore avec un lien '0', nous avons utilisé ce fichier pour attirer l'attention des voyageurs pour écouter l'annonce vocale.
- S_2 **:** Un fichier sonore avec un lien '1' qui contient la phrase :المحطة التالية .
- S_3 **:** L'ensemble des fichiers sonores (variables) avec des liens '2,…,14' qui contiennent les stations d'arrêt en Langue Arabe Standard.
- S_4 **:** Un fichier sonore avec un lien '15'qui contient la phrase : Prochaine station.
- S_5 **:** L'ensemble des fichiers sonores (variables) avec un lien '16,…,28' qui contiennent les stations d'arrêt en Langue Française.

Nous avons aussi des affichages des noms des stations d'arrêt en Arabe Standard et en Français au moment des annonces vocales. Nous allons présenter toutes les données concernant notre simulation de SAVSTA dans le Tableau 4.1.

Tableau 4.1 : les données de projet SAVSTA (distance, fichiers sonores, affichages)

	d_i(m)	s_1	s_2	s_3	s_4	s_5	Affichage
1	200	0	1	2	15	16	Prochaine station : Les fusillés المحطة التالية المعدومين
2	500	0	1	3	15	17	Prochaine station : tripoli المحطة التالية طرابلس
3	1000	0	1	4	15	18	Prochaine station :carobie المحطة التالية خروبة
4	200	0	1	5	15	19	Prochaine station : La glaciére لا قلاسيار المحطة التالية
5	2000	0	1	6	15	20	Prochaine station : Cinq maisons المحطة التالية لديار الخمس
6	4000	0	1	7	15	21	Prochaine station : Les pins المحطة التالية :الصنوبر
7	500	0	1	8	15	22	Prochaine station : Cité moktar zerhoni المحطة التالية : حي مختار زرهوني
8	400	0	1	9	15	23	Prochaine station : Cité rabia المحطة التالية حي رابية
9	500	0	1	10	15	24	Prochaine station :université-b-ezzouar المحطة التالية جامعة باب الزوار
10	200	0	1	11	15	25	Prochaine station : bab ezzouar le pont باب الزوار– الجسر المحطة التالية :
11	500	0	1	12	15	26	Prochaine station : Borj kifan برج الكيفان – الثانوية المحطة التالية :
12	2000	0	1	13	15	27	Prochaine station : mohos المحطة التالية موحوس
13	1000	0	1	14	15	28	Prochaine station : Mimouni hamoud المحطة التالية ميموني حمود

4.3.3 L'organigramme de la simulation (interface graphique)

Nous allons réaliser la simulation de SAVSTA dans l'environnement visuel studio avec un langage de programmation C#. L'algorithme de code source que nous avons utilisé sera basé sur l'organigramme suivant (Fig. 4.3) :

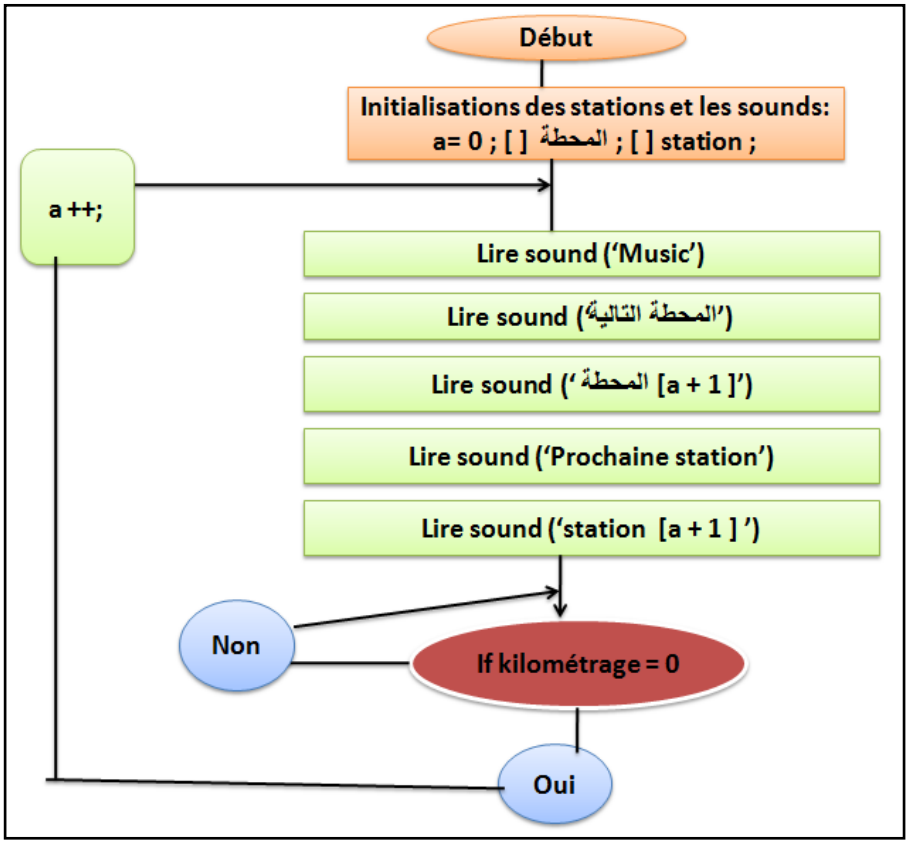


Figure 4.3 : Organigramme de la simulation du SAVSTA

4.3.4 Présentation de l'interface de simulation du SAVSTA

Dans cette partie nous allons présenter la simulation que nous avons réalisée dans le cadre de notre projet. Elle a été réalisée sous l'environnement visuel studio version 2010 avec le langage de programmation C#.

La simulation du SAVSTA contient un fichier :

- des segments sonores, surnommé : sound ;
- Microsoft Access qui contient la base de données (Fig. 4.4) ;
- L'exécutable de la simulation.

A l'aide du tableau de Microsoft Access, nous avons la possibilité d'implémenter d'autres stations d'arrêt supplémentaires (Fig. 4.4).

Table1

Station	StationAR	SoundFR	SoundAR	SoundFR1	SoundAR1	Music	Kilométrage
Les Fusillés	المحدومين	.\sound\15.wav	.\sound\1.wav	.\sound\16.wav	.\sound\2.wav	.\sound\0.wav	300
Tripoli	طرابلس	.\sound\15.wav	.\sound\1.wav	.\sound\17.wav	.\sound\3.wav	.\sound\0.wav	200
Caroubier	الخروبة	.\sound\15.wav	.\sound\1.wav	.\sound\18.wav	.\sound\4.wav	.\sound\0.wav	500
La Glacière	لا قلاسيار	.\sound\15.wav	.\sound\1.wav	.\sound\19.wav	.\sound\5.wav	.\sound\0.wav	1000
Cinq Maisons	الديار الخمس	.\sound\15.wav	.\sound\1.wav	.\sound\20.wav	.\sound\6.wav	.\sound\0.wav	200
Les Pins	الصنوبر	.\sound\15.wav	.\sound\1.wav	.\sound\21.wav	.\sound\7.wav	.\sound\0.wav	1500
Cité Mokhtar Zerhouni	حي مختار زرهوني	.\sound\15.wav	.\sound\1.wav	.\sound\22.wav	.\sound\8.wav	.\sound\0.wav	2000
Cité Rabia	حي رابية	.\sound\15.wav	.\sound\1.wav	.\sound\23.wav	.\sound\9.wav	.\sound\0.wav	500
Université de Bab Ezzou	جامعة باب الزوار	.\sound\15.wav	.\sound\1.wav	.\sound\24.wav	.\sound\10.wav	.\sound\0.wav	400
Bab Ezzouar - Le Pont	باب الزوار - الجس	.\sound\15.wav	.\sound\1.wav	.\sound\25.wav	.\sound\11.wav	.\sound\0.wav	500
Bordj El Kiffan - Lycée	برج الكيفان - الثانوية	.\sound\15.wav	.\sound\1.wav	.\sound\26.wav	.\sound\12.wav	.\sound\0.wav	200
Mouhous	موحوس	.\sound\15.wav	.\sound\1.wav	.\sound\27.wav	.\sound\13.wav	.\sound\0.wav	500
Mimouni Hamoud	ميموني حمود	.\sound\15.wav	.\sound\1.wav	.\sound\28.wav	.\sound\14.wav	.\sound\0.wav	2000

Figure 4.4 : Tableau Access de la simulation SAVSTA

4.3.5 Exécution du programme de SAVSTA

L'interface de la simulation SAVSTA se présente sous la forme d'une fenêtre principale dotée deux boutons (fig. 4.5) :

- afficher/Masquer les stations : ce bouton affiche la liste des stations d'arrêt du tramway d'Alger (fig. 4.6) ;
- démarrer : pour lancer l'application (Fig. 4.7).

Figure 4.5 : Présentation de l'interface de la simulation

Nous allons présenter la liste des différentes stations d'arrêt du tramway (Fig. 4.6).

Stations

N°	Station	StationAR
1	Ruiseau	العناصر
2	Les Fusillés	المعدومين
4	Tripoli	طرابلس
5	Caroubier	الخروبة
6	La Glacière	لا قلاسيار
7	Cinq Maisons	الديار الخمس
8	Les Pins	الصنوبر
9	Cité Mokhtar Zerhouni	حي مختار زرهوني
10	Cité Rabia	حي رابية
11	Université de Bab Ezzouar	جامعة باب الزوار
12	Bab Ezzouar - Le Pont	باب الزوار - الجس
13	Bordj El Kiffan - Lycée	برج الكيفان - الثانوية
14	Mouhous	موحوس

Figure 4.6 : Affichage de la liste des stations d'arrêt du tramway d'Alger

Pour le temps d'exécution de l'application ; nous avons :

- Une interface qui contient des affichages des stations prochaines en Arabe Standard et en Français, et qui affiche aussi la destination (Ruisseau ou Mimouni Hamoud) ;
- un compteur qui se décrémente avec une vitesse constante de celle du tramway et nous avons prise égale 50 m/s (en réalité elle est égale à 5 m/s) afin de visualiser et simuler l'application. Le kilométrage qui affiche chaque annonce a été calculé par un capteur. Ce capteur est placé dans le tramway, pour éviter tous les problèmes de retard ;
- des annonces vocales des stations d'arrêt de l'Aller puis du retour ; et l'application va continuer par basculement de l'aller vers le retour jusqu'au terminus du tramway.

La figure 4.7 présente l'interface d'exécution d'application SAVSTA en temps réel.

Figure 4.7 : L'exécution de l'application en temps réel

4.4. TESTS D'EVALUATION

Le test d'évaluation comprend 20 personnes que ont participé a une session expérimentale d'écoute d'une phrase choisie aléatoirement et répétée trois fois

successivement. Nous avons considéré cinq niveaux de réponses (Mauvais, Médiocre, Passable, Bon et Excellent).

Les résultats obtenus sont définis dans le Tableau 4.2.

Tableau 4.2 : Décision des 20 personnes sur la parole synthétique

	Mauvais	Médiocre	Passable	Bon	Excellent
Décision /20	0	0	04	10	06
Pourcentage (%)	0	0	20	50	30

Nous avons présenté le pourcentage de décision des 20 personnes dans la Figure 4.8.

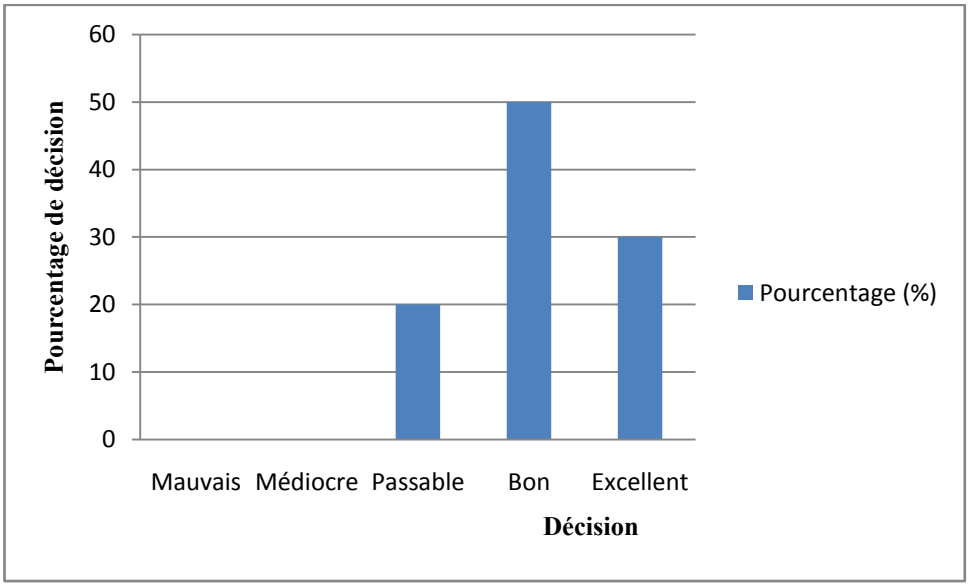


Figure 4.8 : Evaluation sur la parole synthétisée par 20 personnes

Interprétation

D'après le graphe que nous avons obtenu, les résultats sont de 30 % de décision Excellent avec 50 % bon et 20 % 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 ont bien compris la parole synthétique obtenue.

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

4.5 CONCLUSION

Nous avons exposé dans ce chapitre le programme de simulation de la synthèse de parole par concaténation de mots et phrases combinés ainsi que le langage de programmation C # sous l'environnement Visuel Studio. Nos résultats expérimentaux correspondant aux tests de perception subjective, subis par 20 personnes, ont montré le bon enregistrement du corpus AVAST ainsi que la bonne segmentation manuelle, ce qui implique qu'il y a eu compréhension des phrases prononcées avec une bonne qualité de la parole synthétique.

CONCLUSIONS GENERALES ET PERSPECTIVES

La simulation et les tests et les analyses acoustiques que nous avons réalisés tout au long de ce **PFE** nous ont permis d'apprendre des notions sur le traitement de la parole. Ils nous ont permis également de comprendre le fonctionnement d'un système de synthèse de la parole.

Le tramway est un moyen de transport d'une importance et d'une utilité primordiale pour les passagers, et grâce à lui le déplacement devient rapide et plus facile que les autres moyens de transport. Nous avons fait les annonces vocales pour attirer l'attention des voyageurs sur les stations d'arrêt prochaines.

Nous avons tout d'abord effectué des études générales sur la parole puis sur l'Arabe Standard et le Français, pour cela nous avons choisi la synthèse par concaténation des unités pré-stockés, puis un enregistrement du corpus AVAST par une locutrice, ce dernier représente la base de données de notre travail. L'étude de l'AVAST passe par plusieurs étapes d'analyse acoustique et de visualisation qui nous ont permis une extraction des paramètres pertinents du signal vocal. Des tests de perception objective et subjective ont été effectués sur le signal original et le signal synthétique. Les résultats obtenus nous ont permis de confirmer le bon enregistrement de corpus AVAST.

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

La performance d'un système de synthèse vocale dépend, de l'intelligibilité, du naturel de la parole générée et des caractéristiques propres à la voix produite. Ces caractéristiques dépendent des techniques et des méthodes de synthèse, mais également du soin apporté à la modélisation linguistique et prosodique.

Comme suite à ce travail, il serait très intéressant de faire une étude d'évaluation pour l'amélioration de la qualité de la parole synthétisée. Cette dernière sera faite par :

- Un élargissement du vocabulaire du corpus AVAST ;

- Une intervention de la part des experts linguistes afin de mieux modéliser les classes de mots et les phrases ;
- Une Amélioration de la qualité de la voix synthétisée avec une technique d'évaluation par ajustement des paramètres prosodiques de signal vocal pour avoir une bonne qualité se rapprochant du naturel ;
- Une Amélioration des performances de l'algorithme de simulation de SAVSTA.

REFERENCES BIBLIOGRAPHIQUES

[1] 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.

[2] L. Buniet, Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques, thèse de doctorat, Université Henri Poincaré - Nancy, France, 10 février 1997.

[3] Z. Benselama, pathologie du langage parlé Arabe cas des sigmatismes occlusifs et constrictifs , thèse de doctorat : Ecole Nationale Polytechnique-Alger (Algérie), 15/12/2007.

[4] M. Aissiou, Application des Algorithmes Génétiques au Décodage Acoustico-Phonétique de la parole en Arabe Standard, Ecole Nationale Polytechnique-Alger (Algérie), 30/06/2008.

[5] M. Tuan , Analyse acoustique des sons bien identifiés par un système de reconnaissance automatique de la parole, mémoire fin d'études , Institut de la Francophonie pour l'Informatique INRIA de Lorraine – LORIA (France), 30 septembre 2007.

[6] S. Bouguerra & H. Chougrani, Implémentation d'un codeur de parole CELP sur MATLAB, Projet de Fin d'Etudes, Ecole Nationale Polytechnique-Alger (Algérie), 2011.

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

[8] A. Ounnas, synthèse de la parole en Arabe Standard ; Mémoire de Magister ; Ecole Nationale Polytechnique-Alger (Algérie), Décembre 2011.

[9] T. Dutoit, introduction au traitement automatique de la parole, Notes de cours / DEC2 ; Faculté Polytechnique de Mons Belgique, Première édition , 2000.

[10] L. Hocine , 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 (Algérie), 24 Juin 2007.

[11] R. Gaël , synthèse de la parole à partir de texte, Ecole Nationale de Télécommunications, ENST-Paris (France) , documentation des archives techniques d'ingénieur , 20/10/2012.

[12] 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, 10 juin 2011.

[13] G. Djeghiour, thèse magister : application des réseaux de neurones a la synthèse de la parole en arabe standard, Ecole nationale supérieure des sciences humaines; 2011.

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

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