République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE POLYTECHNIQUE

DÉPARTEMENT D'ÉLECTRONIQUE

Projet de fin d'études

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

Thème :

Elaboration d'un Système de Calculatrice Parlante pour Les Non-Voyants Arabophones

Encadré par :

Mme GUERTI Mhania

Réalisé par :

Mlle KADRI Meriem

Promotion : Juin 2015

ملخص:

في مشروعنا هذا قمنا بتطوير آلة حاسبة للمكفوفين الناطقين بالعربية (SCPNVA)، هذه الأخيرة عبارة عن نظام لإصطناع الكلام يعتمد على تحليل وتركيب كلمات مع بعضها البعض. لأجل هذا قمنا بتسجيل قاعدة بيانات خاصة، وتحليلها صوتيا ثم تجزئتها الى الوحدة المطلوبة. لتقييم واختبار هذا النظام قمنا بمحاكاته عن طريق واجهة تم انشاءها في برنامج MATLAB وقد شارك 10أشخاص في هذا التقييم، بالإضافة إلى دراسة تقارنية بين إشارة أصلية وأخرى مركبة لدراسة مدى طبيعية ووضوح الصوت المنتج.

الكلمات المفتاحية :إصطناع الكلام، تركيب الكلمات، قاعدة بيانات، اللغة العربية الفصحى، تحليل صوتي، آلة حاسبة للمكفوفين الناطقين بالعربية،MATLAB

Résumé :

Dans ce projet de fin d'étude nous avons élaboré un système de **C**alculatrice **P**arlante pour les **N**on-**V**oyants **A**rabophones (**SCPNVA**).Ce dernier est un système de synthèse de parole basé sur la concaténation des mots combinés, pour cela une étape d'enregistrement de corpus suivie d'une analyse acoustique et segmentation ont été faites. Pour tester et évaluer notre système une interface de simulation sous le logiciel MATLAB a été réalisée. 10 personnes ont contribuées dans cette évaluation en plus d'une étude comparative entre un signal réel et synthétisé pour tester l'intelligibilité et la naturalité de la voix résultante.

Mots clés : Synthèse de la Parole, Concaténation des mots, corpus, Arabe Standard, analyse acoustique, **C**alculatrice **P**arlante pour les **N**on-**V**oyants

Abstract :

In this final study project, we developed a Talking Calculator System for blinds Arabic speakers (SCPNVA). This latter is a speech synthesis system based on the concatenation of combined words, for this a step of corpus recording followed by an acoustic analysis and segmentation have been done. To test and evaluate our system a simulation interface in MATLAB software was realized. 10 peoples have been contribute in this assessment in addition to a comparative study between a real signal and synthesized one, to test the intelligibility and naturalness of the resulting voice.

Keywords: Speech Synthesis, word concatenation, corpus, Standard Arabic, acoustic analysis, talking calculator for the blinds Arabic speakers, MATLAB.

إهـــــــــداء

أهدي هذا العمل إلى

رسول الأنام عليه الصلاة والسلام ...

مصدر الحب الذي استمر في تشجيعي والدعاء لي... حبيبتي دومًا أمي "زهرة"

من تعلمت منه معنى الكفاح والصبر... أبي الغالي " أحمد"

زهرات بيتنا أخواتي "هناء" ، "جهينة" ، " أسماء" ، "فاطمة الزهراء" ، "راضية"

أخوالي "حيدر" و "باباعيد" وجميع أفراد عائلة "قادري".

من شاركني ساعات الأمل و أمضيت معهم أحلى الأيام صديقاتي كل بإسمها

خاصة "رشا" ، "سمراء" ، "نجلاء" ، " أمينة" ، "وفاء" ، "صوفيا"

كل زملائي بالمدرسة الوطنية المتعددة التقنيات خاصة دفعة ا لإلكترونيك 2015 و إلى كل من

مر بحياتي و ترك فيها ذكرى و أثرا طيبًا

إلى كل هؤلاء أهدي ثمرة جهدي

مريم

Remerciements

Que Dieu soit loué pour nous avoir permis d'arriver au terme de ce travail.

Je remercie et exprime ma reconnaissance à quiconque ayant allumé une bougie dans le chemin de la science et ayant occupé les tribunes du savoir pour m'éclairer.

Je tiens aussi à exprimer mes remerciements et mon gratitude à :

- Mon promotrice, Pr M. GUERTI pour avoir dirigé ce travail jusqu'à son terme, pour le temps qu'elle m'a consacré et pour ses précieux conseils ; Les Membres du jury, Mr R. ZERGUI et Mr M. MAMRI, pour l'enrichissement de cette recherche à travers leurs bénéfiques remarques et orientations conséquentes ;
- Mlle BETTAYEB pour ses précieuses aides et encouragements ;
- Mr TIDJANI pour l'aide et les conseils qu'il m'avait apporté ;
- Je remercie également tous les enseignants de l'Ecole Nationale Polytechnique, et spécialement ceux des départements des Sciences Fondamentales et de Génie Electrique, pour leur apport en savoir ;
- Je remercie enfin tous mes collègues et amis qui m'ont aidé, de près ou de loin, ne serait-ce qu'à travers leurs encouragements.

Liste des abréviations

API : **A**lphabet **P**honétique **I**nternationale.

ARMA : **A**uto **R**égressif à **M**oyenne **A**justée.

AS : **A**rabe **S**tandard.

BD : **B**ase de **D**onnées.

CGP : **C**onversion **G**raphème-**P**honème

CVC : **C**onsonne-**V**oyelle-**C**onsonne.

EJA : **E**coles pour **J**eunes **A**veugles.

FFT : **F**ast **F**ourier **T**ransform.

ISMAS : l'**I**nstitut **S**upérieur des **M**étiers des **A**rts du **S**pectacle et de l'**A**udiovisuel.

LPC : **L**inear **P**redictive **C**oding

MIC : **M**odulation par **I**mpulsions **C**odées.

NV : **M**al-**V**oyants

OE : **O**reille **E**xterne.

OI : **O**reille **I**nterne.

OM : **O**reille **M**oyenne.

OT : **O**ptimality **T**heory

RAP : **R**econnaissance **A**utomatique de la **P**arole.

SAT : **S**ynthèse **A** partir de **T**exte.

SCP : **S**ystème de **C**alculatrice **P**arlante.

SPC : **S**ynthèse **P**ar **C**oncaténation.

TAP : **T**raitement **A**utomatique de la **P**arole.

TF : **T**ransformée de **F**ourier

TFD : **T**ransformée de **F**ourier **D**iscrète

TFR : **T**ransformée de **F**ourier **R**apide

TPZ : **T**aux de **P**assage par **Z**éro

VODer : **Voice O**perating **D**emonstrator.

VOT : **V**oice **O**n **T**ime

Liste des figures

Liste des tableaux

Table des matières

Chapitre 2 : Analyse et Synthèse de la Parole

Chapitre 3 : Elaboration d'un système de Calculatrice Parlante

Chapitre 4 : Evaluations et Discussions

Introduction Générale

Introduction Générale

La parole est un atout que seul nous, êtres humains, possédons dans tout le monde animal. La générer naturellement résulte d'une combinaison complexe de phénomènes physiques et d'interprétations psychoacoustiques. Cependant, de tous temps, l'homme a toujours voulu comprendre, expliquer ses phénomènes, mais également les générer artificiellement. Cela rend le **T**raitement **A**utomatique de la **P**arole (**TAP**) un composant fondamental des sciences de l'ingénieur et d'un domaine de recherche actif. Depuis les années 60, le TAP bénéficie d'efforts de recherches très importants, liés au développement des moyens techniques, de télécommunications et du traitement numérique de l'information. Ces efforts se sont concrétisés grâce à plusieurs applications du TAP, telles que le codage, les perceptions auditive et visuelle, la **R**econnaissance **A**utomatique de la **P**arole (**RAP**) et la **S**ynthèse de la **P**arole (**SP**). Un thème important de la recherche actuelle dans le domaine du TAP, est la réalisation de véritables systèmes de dialogue oral entre l'Homme et la Machine

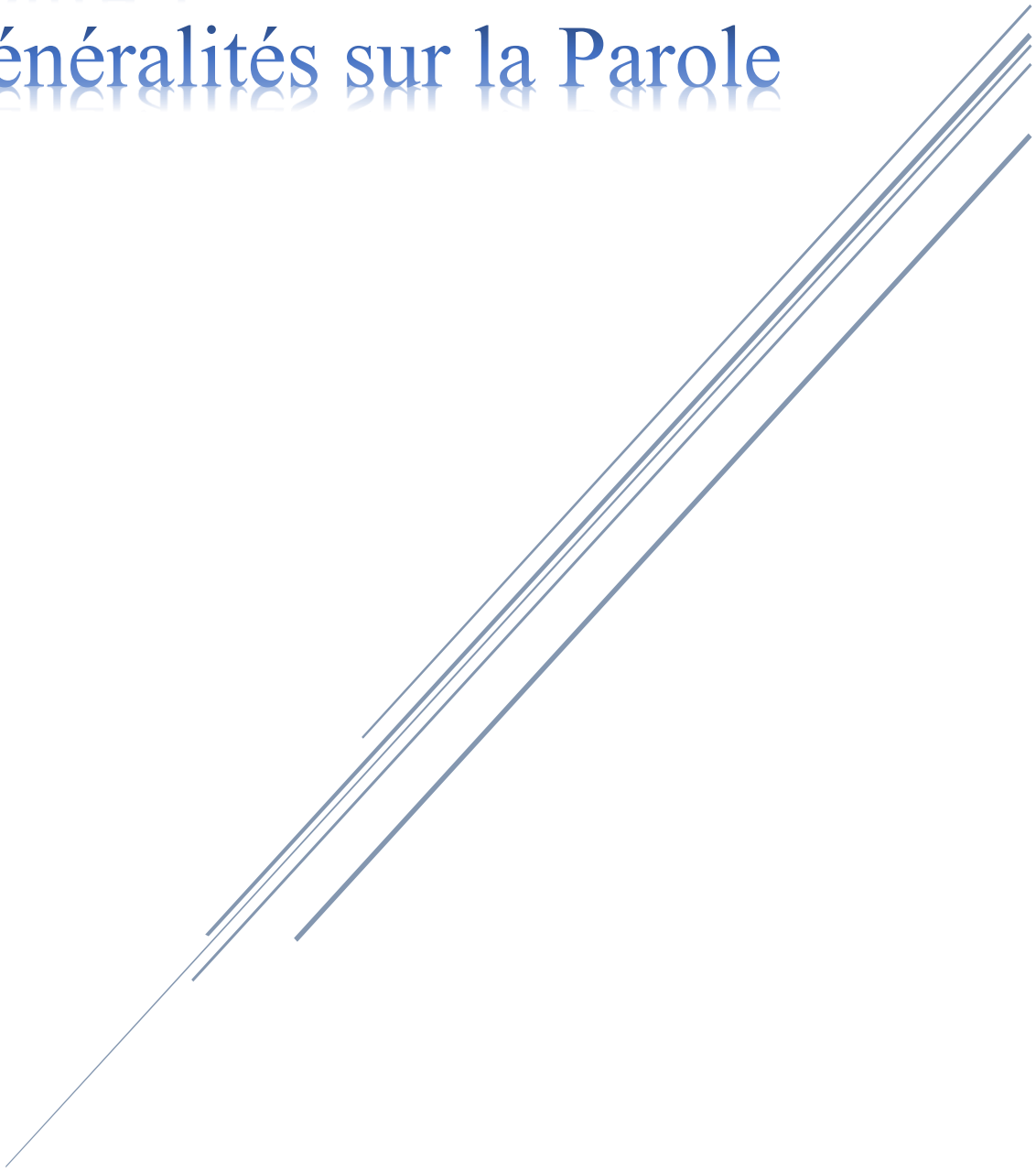
Le but de notre travail est d'élaborer un outil d'aide en apprentissage des sciences mathématique, destiné aux élèves non-voyants arabophones du niveau primaire. Cet outil est une **C**alculatrice **P**arlante (**CP**) basée sur la synthèse de la parole par concaténation de mots combinés. Le système **SCPNV** (Système de Calculatrice Parlante pour les Non-Voyants Arabophones) joue le rôle de lecteur de texte en Arabe Standard de l'opération de calcul et le résultat obtenu, via une base de données préconçue. Une présentation en code Braille a été choisie pour la conception matérielle du système.

Tout système de synthèse de parole à partir du texte est représenté par trois problèmes de natures différentes. Il s'agit dans un premier temps d'analyser et de structurer le texte afin de déterminer un mode de prononciation cohérent ; par la suite, le texte analysé doit être transformé en une suite de sons de parole accompagnée d'indications concernant leur agencement ; enfin, il faut générer un signal acoustique qui « retranscrit » cette suite de sons tout en possédant les caractéristiques apparentes de la parole naturelle.

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

- dans le premier, nous allons décrire d'une manière générale des notions sur la parole ainsi que sa production, les appareils phonatoire et auditif de l'être humain, des spécifications du signal vocal et des notions fondamentales sur l'Arabe Standard ;
- le deuxième, nous donne une brève définition d'analyse et de synthèse de la parole, et du traitement linguistique du texte, En outre, nous étudions les différentes techniques d'analyse du signal vocal. Puis nous expliquons les méthodes de la synthèse de la parole ainsi que ses différentes applications.
- dans le troisième, nous présentons le système de la « SCPNVA », avec une brève explication du processus de construction de notre BD, à l'aide de l'outil d'analyse Praat, et la procédure de concaténation des unités sonores.
- dans le dernier chapitre, nous présentons le programme de la simulation que nous avons développée en vue d'obtenir une CP, nous allons faire une étude comparative entre les phrases synthétisé et naturelles, et de tester l'intelligibilité. En dernier lieu nous finissons par des conclusions générales et perspectives.

Chapitre 1 :
Généralités sur la Parole



1.1 INTRODUCTION

Le but de ce chapitre est de présenter le mécanisme de la perception et la production de la parole chez l'être humain, puis de définir les notions fondamentales utilisées dans le domaine du traitement de la parole. Afin de comprendre les différents niveaux de complexité de problème, nous allons expliquer les principales caractéristiques du signal vocal.

1.2 DEFINITION DE LA PAROLE

La parole est définie comme étant l'expression concrète de la langue. C'est un mode propre à l'Homme. Elle est réalisée en contexte d'énonciation, est plus ou moins dépendante de ce contexte. Elle représente l'image auditive qui vient s'associer à un concept cognitif [1].

Sur le plan physique, la parole est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur. Il s'agit d'une onde sonore créée par le passage de l'air expulsé des poumons dans l'appareil phonatoire et articulatoire du locuteur, ce qui provoque une modification de cette onde puis elle se propage dans l'air. La production de la parole est rapide : 150-300 mots/min. 3-5 syllabes/sec. 10-15 phonèmes/sec. La parole est basée sur l'utilisation des sons d'une langue, c'est-à-dire des phonèmes répertoriés et significatifs pour une langue donnée. L'une des difficultés en traitement automatique de la parole est l'absence de marqueurs entre les mots comme on a les espaces pour l'écrit ainsi que l'extrême variabilité des productions de parole possibles. Le schéma suivant présente le fonctionnement de la communication entre deux locuteurs donnés [2] (figure 1.1).

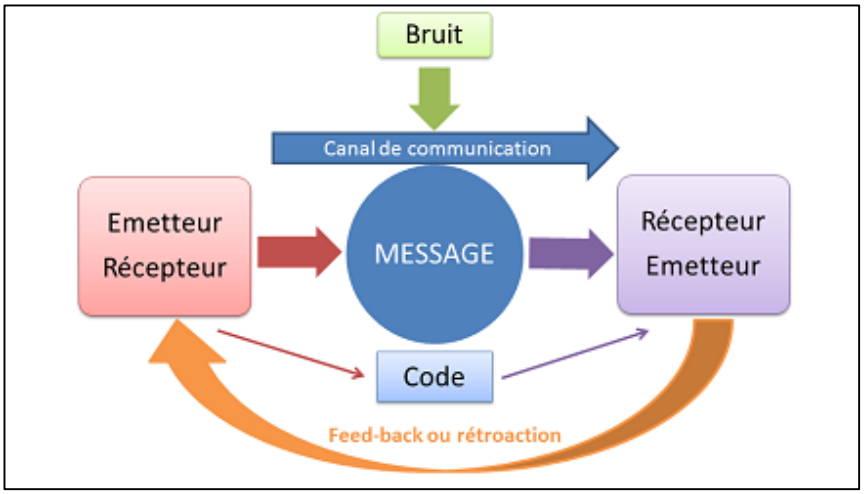


Figure 1.1 : *Schéma du fonctionnement de la communication* [2]

1.3 PRODUCTION DE LA PAROLE

Décrire le processus de production de la parole en vue de spécifier le signal ainsi produit, nécessite l'acquisition de certains nombres de connaissances liées à la complexité du processus de génération et de ses difficultés de mesure.

1.3.1 Appareil phonatoire humain

L'appareil phonatoire humain peut se présenter comme un système source/filtre avec nos poumons comme un réservoir énergétique.

1.3.1.1 Physiologie des organes de phonation

Les fonctions essentielles dans l'acte de parole, ou phonation sont réalisées par trois groupes d'organes :

- **l'appareil respiratoire** : (diaphragme, poumons, trachée), soufflerie qui fournit l'énergie et la quantité d'air nécessaire à la production de sons en poussant de l'air à travers la trachée-artère ;
- **le larynx « l'organe vibrant »** : se trouve au sommet supérieure de trachée-artère, où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles. **Les cordes vocales** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée **glotte**. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer (figure 1.2) ;

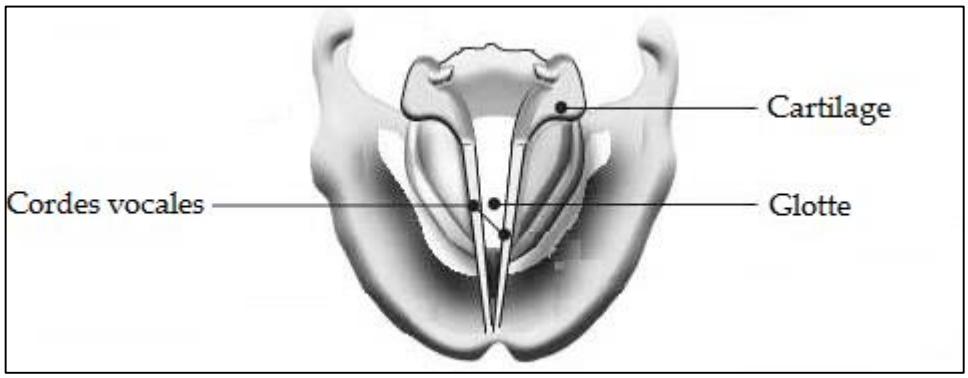


Figure 1.2 : *Le larynx*

• **Le conduit vocal**, formé des cavités résonantes supra-laryngées (pharynx, bouche, nez) où s'effectue l'articulation proprement dite par les changements de forme du conduit vocal. Ces changements résultent surtout des mouvements des lèvres, de la langue, du voile du palais (dont l'abaissement fait intervenir une cavité supplémentaire, les fosses nasales) et de la mâchoire inférieure [3] (figure 1.3).

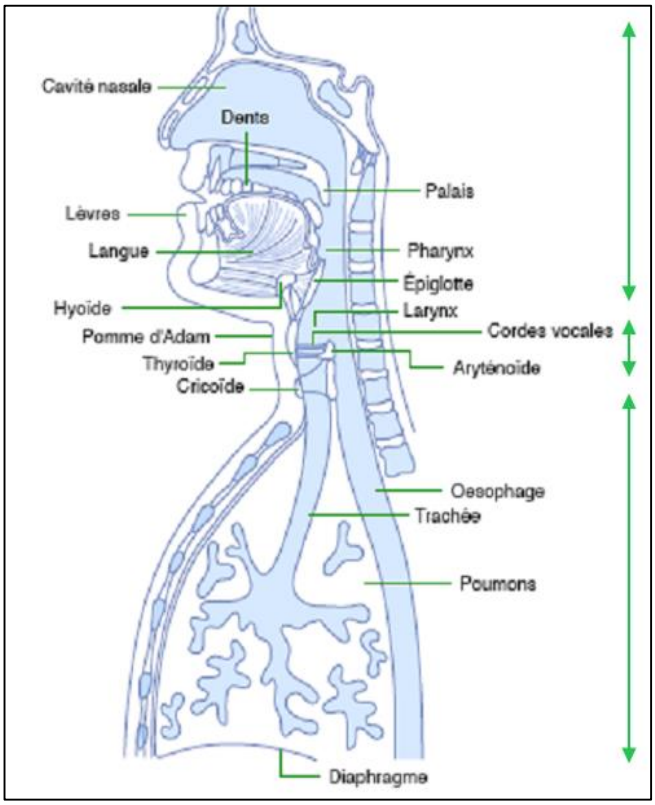


Figure 1.3 : *Appareil phonatoire humain* [4]

1.3.1.2 Processus de la phonation

Le fonctionnement de cet appareil est déclenché et contrôlé par le système nerveux central du locuteur, après avoir pris la décision de parler, les muscles du diaphragme aident à gonfler et dégonfler les poumons, ces derniers envoient l'air vers le conduit vocal, les cordes vocales font vibrer l'air en provenance des poumons, le nez et la bouche font modifier l'onde sonore pour former le mot à prononcer (figure 1.4).

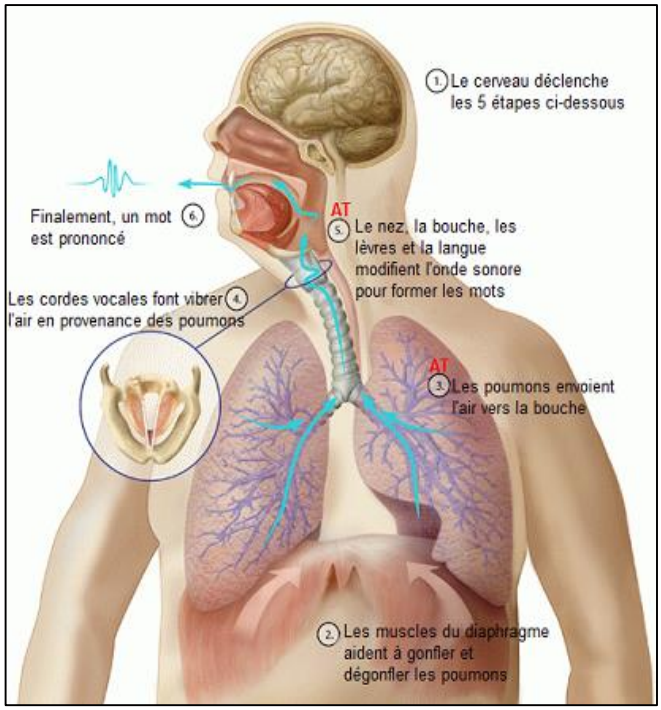


Figure 1.4 : *Mécanisme de la phonation humaine* [5]

1.3.2 Modélisation Source/Filtre

Le modèle source/filtre, considère le signal de parole s(n) comme le résultat de la convolution du signal glottique e(n) (la source) par un filtre h(n) qui représente le comportement fréquentiel du conduit vocal soit (figure 1.5) :

$$s(n) = e(n) * h(n) \tag{1.1}$$

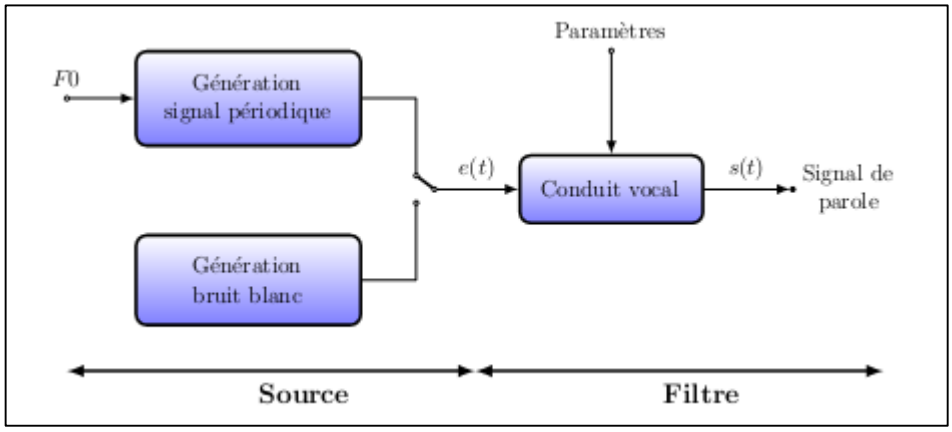


Figure1.5 : *Modélisation Source /Filtre de la parole* [6]

Dans sa représentation la plus simple, ce modèle repose sur deux contraintes fortes [6] :

- le filtre est supposé comme un système linéaire ;
- le filtre et la source sont indépendants.

Lors de la phase d'analyse du signal de parole, seul s(n) est connu. Représenter un signal selon une modélisation source/filtre consiste donc à résoudre une équation à deux inconnues ce qui implique de faire des hypothèses simplificatrices sur la source ou sur le filtre. Généralement, ces hypothèses sont appliquées au modèle de la source qui est réduit à deux états possibles : la source correspond à un signal périodique si le son est voisé, ou elle correspond à un signal bruité si le son est non-voisé. Le filtre correspond ici à un spectre qui, par convolution, va permettre d'amplifier certaines fréquences du signal issues de la source.

Afin de simplifier l'opération de déconvolution et ainsi de déterminer plus aisément la contribution de la source et du filtre, cette opération est effectuée dans l'espace dit cepstral. Cet espace, appelé domaine quéfrentiel, est obtenu en effectuant une Transformée de Fourier Inverse sur le logarithme du spectre, ce qui permet de substituer l'opérateur de convolution à un opérateur d'addition. La relation suivante est alors obtenue :

$$\tilde{c}(n) = \tilde{e}(n) + \tilde{h}(n) \tag{1.2}$$

Où :

- $\tilde{c}(n)$ correspond aux coefficients dits cepstraux ;
- $\tilde{e}(n)$ correspond à e(n) ;
- $\tilde{h}(n)$ donne h(n), dans le domaine quéfrentiel [7].

1.4 SYSTEME AUDITIF ET PERCEPTION DE LA PAROLE

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille humaine est aussi importante qu'une maîtrise des mécanismes de production.

1.4.1 Système auditif humain

L'appareil auditif comprend l'**O**reille **E**xterne (**OE**), l'**O**reille **M**oyenne (**OM**), et l'**O**reille **I**nterne (**OI**). Le conduit auditif relie le pavillon au tympan : c'est un tube acoustique de section uniforme, fermé à une extrémité, son premier mode de résonance est situé vers 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'OI

(marteau, étrier, enclume) permet une adaptation d'impédance entre l'air et le milieu liquide de l'OI. Les vibrations de l'étrier sont transmises au liquide de la cochlée. Celle-ci contient la membrane basilaire qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée ; elle est le support de l'organe de Corti qui est constitué par environ 25000 cellules ciliées raccordées au nerf auditif (Figure 1.6).

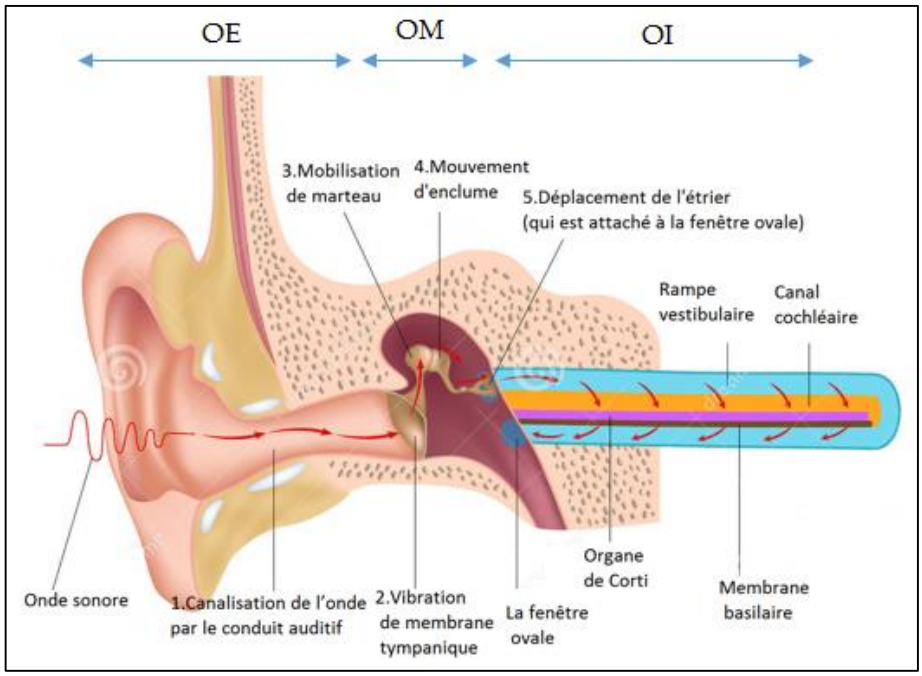


Figure 1.6 : *Processus de la perception humaine*

1.4.2 Le champ auditif

L'oreille humaine ne répond pas également à toutes les fréquences, son champ auditif est délimité par un seuil d'audition et un seuil de douleur. Sa limite supérieure en fréquence (≈ 16000 Hz), variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (≈ 32000 Hz) [8].

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La figure (1.7) a fait apparaître les courbes d'égale impression de puissance auditive (aussi appelée sonie, exprimée en sones) en fonction de la fréquence. Elles révèlent un maximum de sensibilité dans la plage (500 Hz, 10 kHz), en dehors de laquelle les sons doivent être plus intenses pour être perçus.

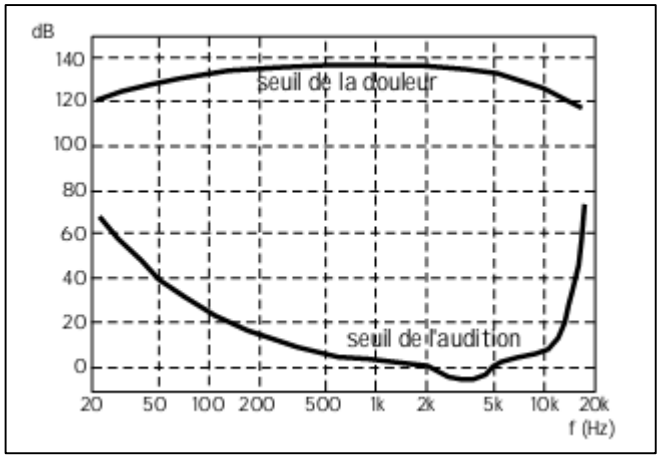


Figure 1.7 : *Le champ auditif humain* [8]

1.5 CARACTERISTIQUES DU SIGNAL DE LA PAROLE

L'information portée par le signal de la parole peut être décrite par plusieurs niveaux de description : phonétique, phonologique, acoustique, …

1.5.1 Niveau phonétique

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

1.5.1.1 Classification des sons du langage

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode et lieu d'articulation. Dans la cavité buccale, le point d'articulation est l'endroit où se trouve un obstacle au passage de l'air. D'une manière générale, le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air (figure 1.8).

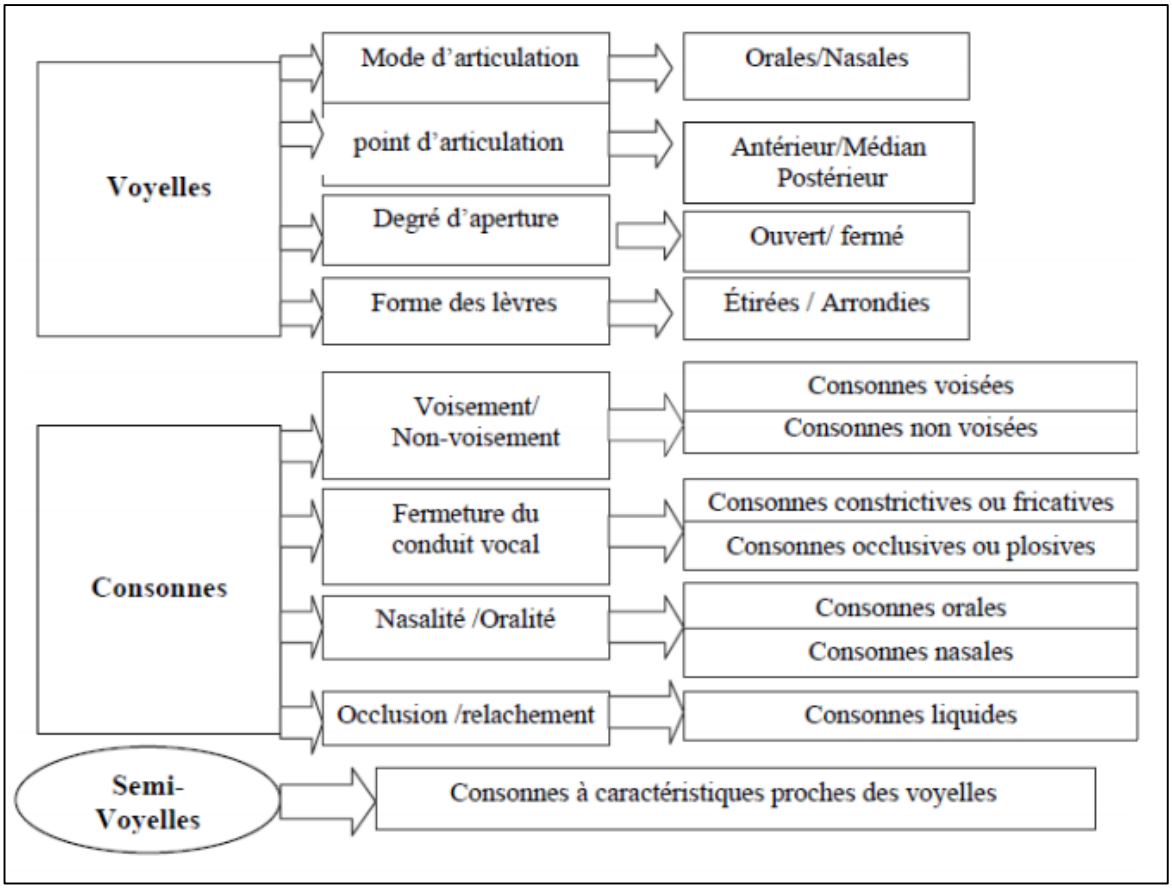


Figure 1.8 : *Classification des sons de langage* [9]

1.5.1.1.1 Sons voisés

Les sons voisés, tels que les voyelles, semi-voyelles et les consonnes nasales, sont produits par le passage de l'air des poumons à travers la trachée qui met en vibration les cordes vocales. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité et une énergie élevée (figure 1.9).

1.5.1.1.2 Sons non voisés

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

Les consonnes sont un exemple de son non voisé, apériodique. Ces sons sont considérés comme ayant les mêmes caractéristiques que le bruit (figure 1.9).

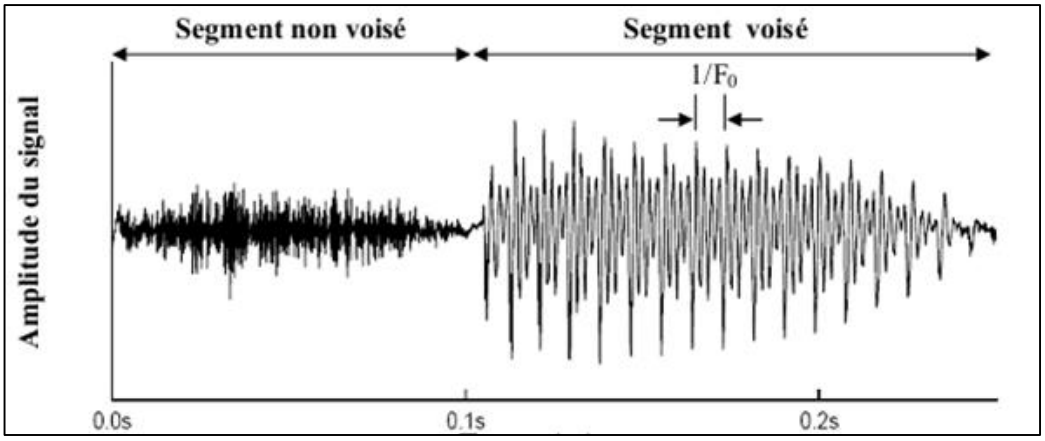


Figure 1.9 : *Représentation temporelle des segments de sons voisés et non voisés* [10]

1.5.1.1.3 Voyelles

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Quand ce dernier est suffisamment ouvert pour que l'air expiré par les poumons, le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la cavité buccale se réduit alors à une modification du timbre vocalique.

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

Les voyelles orales sont dues à une élévation du palais qui détermine la fermeture des fosses nasales ainsi qu'à l'écoulement de l'air expiré à travers la cavité buccale. Par contre, les voyelles nasales sont caractérisées par l'écoulement d'une partie de l'air à travers la cavité nasale [11].

1.5.1.1.4 Consonnes

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

- intervention ou mise en vibrations des cordes vocales : articulation sonore ;
- fermeture momentanée du passage de l'air suivie d'une ouverture brusque (explosion) : articulation occlusive ;
- rétrécissement du passage de l'air qui produit un bruit de friction : articulation fricative ;

- position abaissée du voile du palais : articulation nasale ;
- contact de la langue au milieu du canal buccal ; l'air sort des deux côtés ;
- une série d'occlusions brèves ; séparées de la luette : articulation vibrante.

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

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

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

1.5.1.1.5 Les semi-voyelles

Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, elles sont assez difficiles à classer [11].

1.5.1.2 Langue Arabe Standard

L'Arabe est la langue officielle dans plus de 22 pays, elle est utilisée par 1.62 milliard de musulmans dans le monde. Nous distinguons deux types de langue Arabe : les différents dialectes Arabes utilisés dans chaque pays Arabe et l'Arabe Standard (AS) qui est la langue utilisée dans les cadres officiels, enseignée dans l'école et la langue du Saint Coran.

Le système phonétique arabe contient 40 phonèmes : 26 consonnes, trois voyelles courtes, trois voyelles longues, 2 semi-voyelles et six variantes vocaliques en contexte emphatique.

Généralement ces trois voyelles courtes ne sont pas présentées dans l'écriture arabe, ils sont ajoutés avec d'autre signes diacritiques comme : la [*ʃadda*] « gémination ou dédoublement d'une consonne) » ; le [*suku:n*] « qui désigne que la consonne n'est pas suivie d'une voyelle », pour faciliter la compréhension du contexte Arabe.

L'AS ne possède pas de voyelles nasales. Elles sont représentées sur un plan dont les axes sont les formants F_1 et F_2. Elles tracent alors un triangle dont les extrémités sont occupées par les voyelles [i, u, a]. Ce triangle représente également les positions de la langue dans la cavité buccale selon deux axes : antérieur à postérieur (avant et arrière) et de fermé à ouvert, selon que la langue est massée en avant et vers la zone dentale pour [i], basse et étalée loin du palais pour [a] (ouvert), ou massée postérieurement vers le voile pour [u] dont laquelle les voyelles soulignées sont labialisées (arrondies) [9] (figure 1.10).

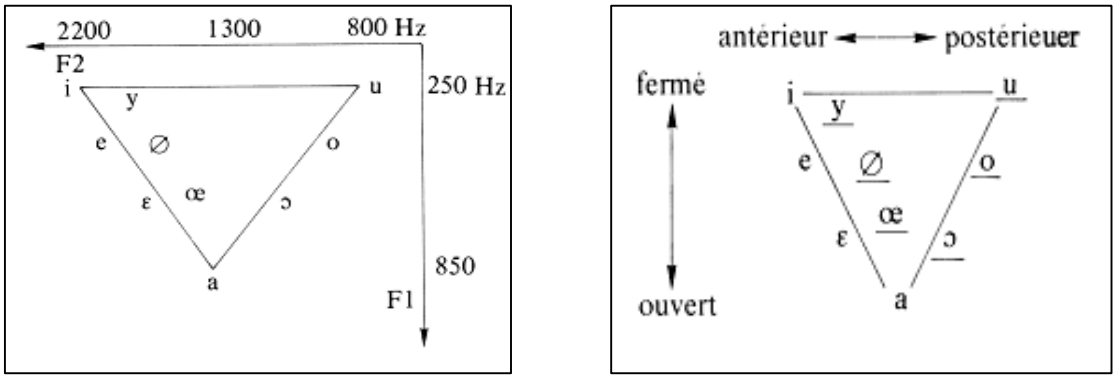


Figure 1.10 : *Triangle vocalique des voyelles « Caractéristiques acoustiques et articulatoires »* [7]

Le tableau suivant (Tableau 1.1) montre les modes et les lieux d'articulation des différentes consonnes et semi-voyelle de l'AS.

Tableau 1.1 : *Classification des consonnes et semi-voyelles de l'Arabe Standard*

Mode	Type de phonème		Phonèmes Arabes	Lieux d'articulation
Occlusives	Voisées		ب	Bilabiale
			د	Alvéodentale
	Non- Voisées		ق	Uvulaire
			ت	Alvéodentale
			ك	Postpalatale
			ء	Glottale
	Voisée	Emphati-ques	ظ	Alvéolaire
	Non- Voisée		ط	Alvéodentale
Fricatives	Voisées		ز	Sifflante
			ذ	Dorsoalvéolaire
			غ	Interdentale
			ع	Uvulaire
	Non-Voisées		س	Sifflante dentale
			ث	Interdentale
			ف	Labiodentale
			ش	Chuinchante palatale
			خ	Vélaire
			ه	Glottale
			ح	Pharyngale
	Voisées	Emphati-ques	ص	Dorsealvéodentale sifflante
	Non-Voisées		ض	Interdentale
Nasales	Voisées		م	Bilabiale
			ن	Alvéodentale
Liquide	Voisées		ل	Dentale
Affriquée	Voisées		ج	Alvéopalatale
Vibrante	Voisées		ر	Apico-alvéolaire
Semi-voyelles	Non-Voisées		و	Bilabiale
			ي	Palatale

1.5.1.3 Alphabet Phonétique International

L'**A**lphabet **P**honétique **I**nternational (**API**) associe des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations, la figure suivante présente tous les symboles de l'alphabet de l'AS (figure 1.11).

Arabic letter	Buckwalter	Amended SAMPA (original)	IPA
ا	A	aa	a:
ب	b	b	b
ت	t	t	t
ث	v	v	θ
ج	j	j (Z)	ʤ
ح	H	h (X)	ħ
خ	x	x	x
د	d	d	ð
ذ	*	D	X
ر	r	r	r
ز	z	z	z
س	s	s	s
ش	SH	s` (S)	ʃ
ص	S	S` (s.)	sˤ
ض	D	D` (d.)	dˤ
ط	T	T` (t.)	tˤ
ظ	Z	Z` (z.)	ðˤ
ع	E	E (H)	ʕ
غ	g	g (G)	ɣ
ف	f	f	f
ق	q	q	q
ك	k	k	k
ل	l	l	l
م	m	m	m
ن	n	n	n
ن	n	M (not provided)	ŋ
ن	n	c (not provided)	ɲ
ن	n	e (not provided)	ɲ
ه	h	h	h
و	w	w or uu	w or u:
ي	y	y or ii (j)	j or i:
أ	'	' (?)	ʔ
ـَ	a	a	a
ـُ	u	u	u
ـِ	i	i	i
ـً	F	an	an

Figure 1.11 : *Alphabet de l'AS en Buçkwalter, SAMPA et API* [12]

1.5.2 Niveau acoustique

Précédemment, On a défini le signal de la parole comme étant le résultat d'une variation de la pression produite par l'émission d'un son par un système articulatoire.

La phonétique acoustique étudie ce signal en le transformant dans un premier temps en un signal électrique. De nos jours, ce signal électrique résultant, est le plus souvent numérisé, l'opération de numérisation, requiert successivement (figure1.12) :

- **une transduction ;**
- **une préamplification ;**
- **un filtrage de garde** à une fréquence de coupure f_c **;**
- **un échantillonnage** à une fréquence f_e **;**
- **une quantification** avec un nombre de bits **b** et le pas de quantification **q.**

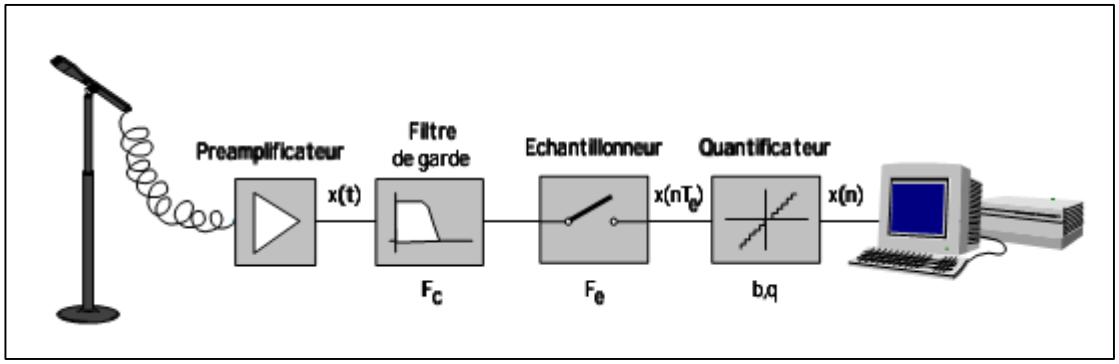


Figure 1.12 : *Enregistrement numérique d'un signal acoustique* [8]

Le signal numérisé peut être alors soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre.

1.5.2.1 Fréquence fondamentale

La fréquence fondamentale F_0 est le nombre de vibrations des cordes vocales par seconde au cours de la prononciation d'un son voisé. La gamme de variation moyenne de la fréquence fondamentale varie d'une personne à une autre en fonction de la longueur et de la masse des cordes vocales de chaque personne, donc elle dépend, essentiellement, de l'âge, de l'état et du sexe du locuteur. Elle peut varier de :

- 70 à 250 Hz chez l'homme ;
- 150 à 400 Hz chez la femme ;
- 200 à 600 Hz chez l'enfant [13].

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 $\boldsymbol{F_0}$.

1.5.2.2 Intensité sonore

L'intensité exprime le volume sonore d'un phonème et dans le cas d'un voisement elle représente l'amplitude des vibrations des cordes vocales, elle résulte de la pression sous glottique. Pour rendre compte de l'intensité d'un son, on utilise une unité de mesure relative, le décibel (dB).

Elle est exprimée pour un signal échantillonné s_n par :

$$\mathbf{E} = \frac{\mathbf{1}}{\mathbf{T}} \sum_{N=1}^{T} \boldsymbol{s_n}^{\mathbf{2}} \tag{1.3}$$

$$\boldsymbol{E_{dB}} = \mathbf{10} * \boldsymbol{log_{10}}\left(\frac{\mathbf{1}}{\mathbf{T}} \sum_{N=1}^{T} \boldsymbol{s_n}^{\mathbf{2}}\right) \tag{1.4}$$

1.5.2.3 Durée phonémique

La durée représente le temps de la prononciation d'un phonème. Elle est le paramètre acoustique le plus délicat à évaluer. La difficulté de mesure réside dans sa grande variabilité qui est due au contrôle quasi impossible du système phonatoire. Chaque phonème se caractérise par ses propres durées intrinsèques et extrinsèques.

1.5.2.4 Formants

Les formants sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants représentent les maxima de la courbe de réponse en fréquences du conduit vocal. Chaque son à ses formants caractéristiques (figure 1.13).

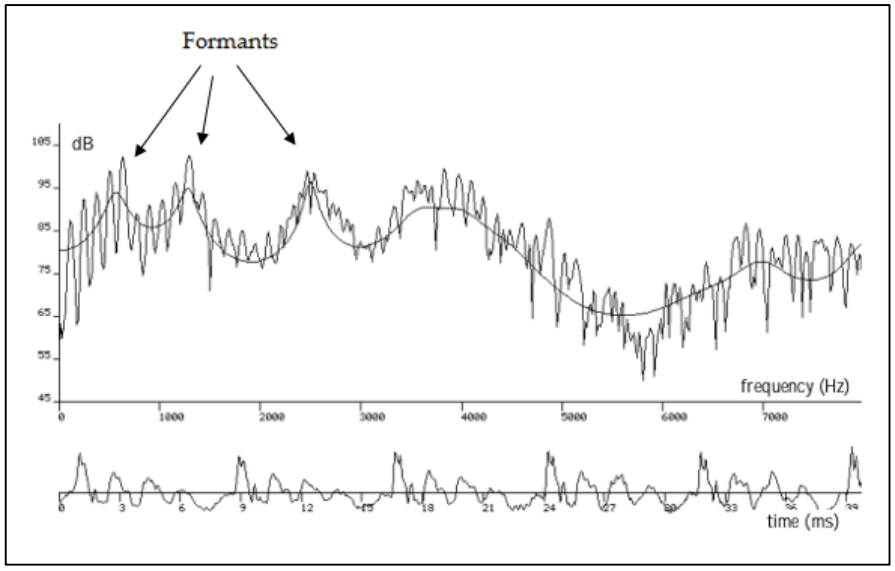


Figure 1.13 : *Représentation spectrale et temporelle d'un signal du son voisé*

1.6 COMPLEXITE DU SIGNAL DE LA PAROLE

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

1.6.1 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 [9].

1.6.2 Variabilité

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 [9].

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

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

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

- **Variabilité contextuelle :** en effet, les mouvements articulatoires peuvent être modifiés de façon à minimiser l'effort à produire pour les réaliser à partir d'une position articulatoire donnée, ou pour anticiper une position à venir. Ces effets sont connus sous le nom de réduction, d'assimilation et de coarticulation. Les phénomènes coarticulatoires sont dus au fait que chaque articulateur évolue de façon continue entre les positions articulatoires. Ils apparaissent mêmes dans le parler le plus soigné. Au contraire, la réduction et l'assimilation prennent leur origine dans des contraintes physiologiques et sont sensibles au débit de la parole. L'assimilation est causée par le recouvrement de mouvements articulatoires et peut aller jusqu'à modifier un des traits phonétiques du phonème prononcé. La réduction est plutôt due au fait que les cibles articulatoires sont moins atteintes dans le parler rapide.

Ces phénomènes sont en grande partie responsables de la complexité des traitements réalisés sur les signaux de parole.

1.6.3 Coarticulation

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

1.6.4 Redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter les

durées du traitement, lequel doit se faire en temps réel. A l'inverse, le débit ne doit pas être trop faible pour conserver un bon rapport signal/bruit. En effet, Il existe une grande disproportion entre le débit du signal enregistré et la quantité utile pour une tâche de reconnaissance [9].

1.7 TRAITEMENT ATOMATIQUE DE LA PAROLE (TAP)

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine. 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.

Les techniques modernes de traitement de la parole tendent cependant à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions :

1.7.1 L'analyse

Cherche à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu, mais jamais tel qu'il est compris, ce rôle étant réservé aux reconnaisseurs. Les analyseurs sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse, soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical (pour les pathologies du larynx, par analyse du signal vocal) ou l'étude des langues [8].

1.7.2 Reconnaissance Automatique de la Parole (RAP)

La RAP Sert à décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal.

1.7.2.1 Reconnaissance de locuteur

L'objectif de la reconnaissance de locuteur est de reconnaître la personne qui parle. On classe également les reconnaisseurs en fonction des hypothèses simplificatrices sous lesquelles ils sont appelés à fonctionner :

- **l'identification :** le problème est de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé ;
- **la vérification :** le problème est de vérifier que la voix analysée correspond bien à la personne qui est sensée la produire ;
- **dépendante de texte (avec texte dicté) :** la phrase à prononcer pour être reconnue est fixée dès la conception du système ;
- **indépendante du texte :** la phrase à prononcer pour être reconnue est fixée lors du test, donc la reconnaissance doit être assurée pour n'importe quelle phrase.

1.7.2.2 Reconnaissance de la parole

L'objectif de la reconnaissance de la parole est de reconnaître ce qui est dit, on classe également les reconnaisseurs de parole comme suit :

- **monolocuteur :** capable de reconnaître que la parole prononcée par la voix d'une seule personne ;
- **multilocuteur :** capable de reconnaître la parole prononcée par la voix d'un nombre fini de personnes ;
- **indépendante de locuteur :** capable de reconnaître la parole de n'importe qui ;
- **reconnaisseur de mots isolés :** le locuteur sépare chaque mot par un silence ;
- **reconnaisseur de mots connectés :** le locuteur prononce de façon continue une suite de mots prédéfinis ;
- **reconnaisseur de parole continue :** le locuteur prononce n'importe quelle suite de mots de façon continue.

1.7.3 Synthèse de la parole

La synthèse est la production de la parole artificielle. On distingue fondamentalement deux types de synthétiseurs 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 que celles obtenues par analyse ;

- **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, synthétiseurs à partir :

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

1.7.4 Codage

Le codage permet la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole.

1.8 CONCLUSION

Dans ce chapitre, nous avons présenté brièvement le phénomène de la parole « physiologique, phonétique et acoustique ». Nous avons pu ainsi voir qu'il s'agit d'un phénomène complexe qui repose sur de nombreux mécanismes physiologiques et cognitifs. En présentant le modèle source/filtre, nous avons pu introduire la description de signal de parole. Des généralités phonétiques, certaines propriétés spécifiques et une description simplifiée des différents sons de l'AS ont été présentées.

Chapitre 2 :
Analyse et Synthèse de la Parole

2.1 INTRODUCTION

La synthèse vocale à partir du texte permet de convertir un texte donné en un signal audio de parole. Ainsi peut-on imaginer de multiples usages de ces technologies, tous plus utiles les uns que les autres. Il ne s'agit pas ici de remplacer l'homme, mais de le décharger de tâches ingrates et contraignantes. Dès les années 1980, ses premiers utilisateurs furent les personnes malvoyantes. Les systèmes de synthèse de la parole leur permettent en effet d'avoir accès aux informations écrites, sous forme vocale.

Dans ce chapitre, nous allons introduire le cadre technique de notre étude : l'analyse et la synthèse de la parole. Nous allons donner une brève définition de la synthèse de la parole, et le traitement linguistique du texte, En outre, nous étudions les différentes techniques d'analyse du signal vocal ainsi que les méthodes et les différentes applications de la SP.

2.2 HISTORIQUE DE LA SYNTHESE DE LA PAROLE

Dans cette partie, nous allons présenter que les grandes lignes des évolutions de la synthèse vocale. Les premières machines parlantes voient le jour avec l'abbé MICAL et W.V. KEMPLEN (Figure 2.1) au 18ème siècle, premières grandes simulations mécaniques des phénomènes de production de la parole humaine associant source vocale et résonateurs supra-glottiques [7].

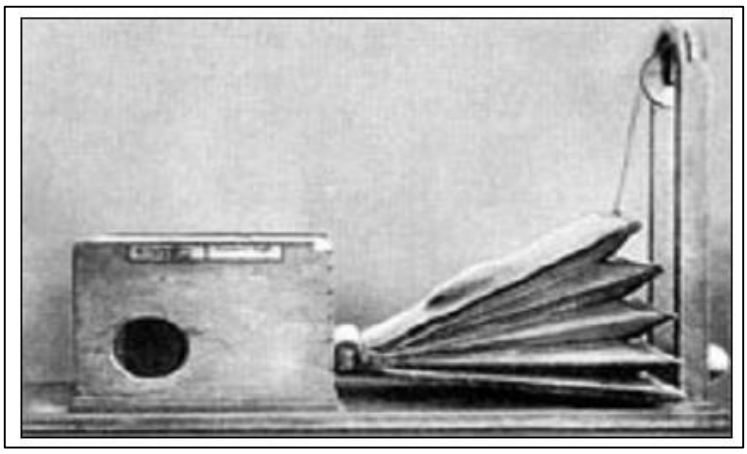


Figure 2.1 : *Machine à parler de Von Kempelen* [7]

Au 19ème siècle, la parole et la voix deviennent objets d'études scientifiques spécifiques. Inspirés par les travaux de leurs précurseurs, plusieurs chercheurs ont mis au point des machines simulant le conduit vocal. Parmi ceux-ci nous pouvons citer J. FABER avec son Euphonia (1830

- 40) ; C. WHEATSTONE avec le perfectionnement de la machine de KEMPLEN ; A. GRAHAM BELL pour une version simplifiée de la reconstitution de WHEATSTONE et R.R. RIESZ avec un appareil simulant les différentes sections du conduit vocal. Ce dernier permit une meilleure compréhension de la physiologie de l'appareil phonatoire humain, de la géométrie et du rôle de ses articulateurs.

Le début du 20ème siècle vit l'apparition de l'électricité et de l'électronique ce qui autorisa des tentatives plus ambitieuses : en 1922, J.C. STEWART fabrique une machine capable de reproduire des voyelles, des diphtongues (Voyelles complexes qui changent de timbre en cours d'émission) et quelques mots simples. Plusieurs années plus tard (1939), H. DUDLEY présente, à l'occasion de l'exposition universelle de New York, le **VODer** (**V**oice **O**perating **D**emonstrator), appareil mis au point par les laboratoires Bell, fondé sur le Vocodeur à Canaux. 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, qui se présente comme un sonagraphe fonctionnant à l'envers (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.) [14].

2.3 ANALYSE DU SIGNAL VOCAL

Le Traitement du signal vocal numérique passe par deux principales étapes : le prétraitement et l'analyse, ces étapes permettent d'extraire les caractéristiques du signal, cela peut aider pour augmenter les performances du signal et le rendre facile à traiter.

2.3.1 Prétraitement :

Un échantillonnage et une préaccentuation seront appliqués sur le signal vocal. Pour les techniques de reconnaissance, d'analyse ou de synthèse de la parole, la fréquence d'échantillonnage peut varier de 08 jusqu'à 16 kHz. Le filtre de préaccentuation de transmittance $H(z)$ est :

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

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

Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Pour compenser ces effets de bord, nous multiplions en général préalablement chaque tranche d'analyse par une fenêtre de pondération de type fenêtre de Hamming notée $W(n)$ [15] (figure 2.2).

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

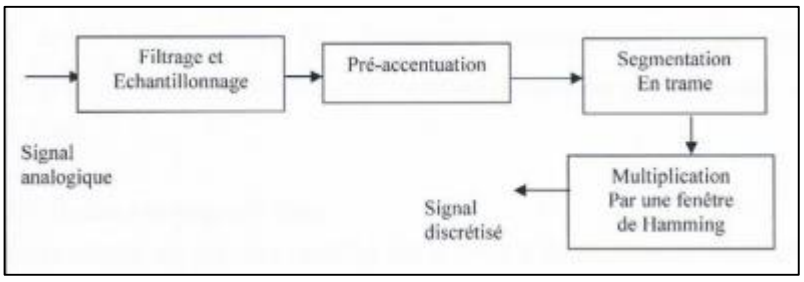


Figure 2.2 : *Prétraitement du signal vocal* [15]

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

Dans la plupart des méthodes d'analyse vocale, nous supposons que le signal de parole est localement stationnaire car les propriétés de ce signal varient très doucement en fonction du temps, d'où le recours aux méthodes d'analyse à court terme. Ainsi de courts segments de la parole sont analysés, on les appelle les trames d'analyse temporelle. Les mesures comme l'énergie, le **T**aux de **P**assage par **Z**éro (**TPZ**) et la fonction d'autocorrélation font partie des méthodes temporelles.

2.3.2 Méthodes non paramétriques

Le signal de parole peut être analysé dans le domaine temporel ou dans le domaine spectral par des méthodes non paramétriques, sans faire 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 il n'a pas d'intérêt que si elle s'applique à une période du signal vocal, donc sur une période assez courte. Actuellement, les spectres sont obtenus numériquement par la **T**ransformée de **F**ourier **D**iscrète (**TFD**), en particulier grâce à l'algorithme de la **T**ransformée de **F**ourier **R**apide (**TFR**) ou **F**ast **F**ourier **T**ransform (**FFT**). Cependant, le nombre de paramètres spectraux calculés sur une trame par FFT reste trop élevé pour un traitement automatique ultérieur. Pour une analyse très fine de la parole, la fenêtre de Hamming est déplacée à chaque fois de 128 points environ 10 ms (figure 2.3).

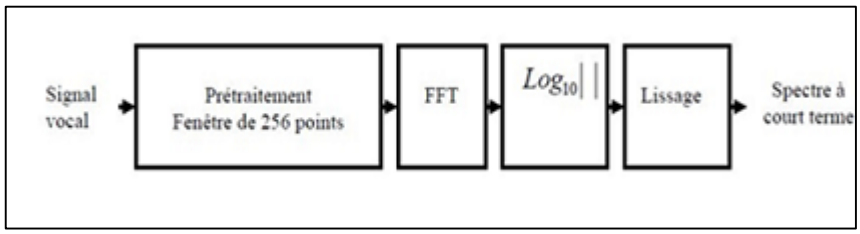


Figure 2.3 : *Analyse numérique du signal parole par FFT*

2.3.3 Méthodes paramétriques

Les méthodes paramétriques sont fondées sur une connaissance des mécanismes de production de la parole. Les plus utilisées sont celles basées sur l'analyse prédictive linéaire et l'analyse cepstrale. Hypothèse de base est que le conduit buccal est constitué d'un tube cylindrique de section variable. L'ajustement des paramètres de ce modèle permet de déterminer à tout instant sa fonction de transfert. Cette dernière fournit une approximation de l'enveloppe du spectre du signal à l'instant d'analyse. Ces méthodes consistent à ajuster un modèle aux données observées. Les paramètres du modèle, en nombre faible, caractérisent le signal, nous pouvons ainsi injecter des connaissances a priori sur le processus physique qui a engendré ce signal. Les avantages de cette approche sont la souplesse de l'analyse, l'introduction naturelle de l'information et les choix variés des espaces de représentations paramétriques.

2.3.3.1 Codage Prédictif Linéaire

Linear **P**redictive **C**oding (**LPC**) Cette méthode se fonde sur les connaissances de la production de la parole et suppose que le modèle de production de la parole est linéaire selon le schéma (figure 2.4).

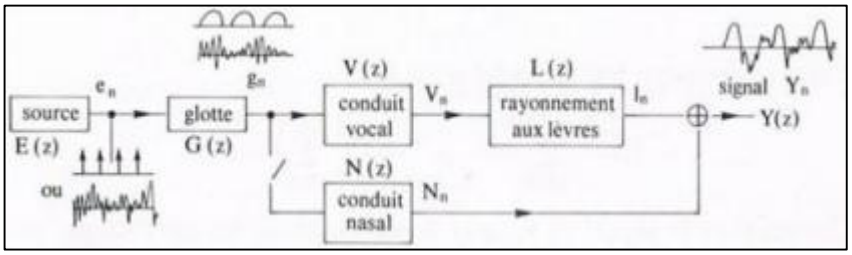


Figure 2.4 : *Modèle général de production de la parole* [7]

Globalement, ce modèle peut se décomposer en deux parties : la source active, le conduit passif de manière plus détaillée, il peut se décrire de la maniéré suivante : l'onde est modélisée comme la sortie d'un filtre passe bas à deux pôles de fréquence de coupure d'environ 100 Hz (glotte), l'entrée en de ce filtre est un train d'impulsions de période T0 pour les sons voisés ou un bruit blanc pour les sons non voisés (source). Le modèle du conduit vocal est un filtre tout pôle (**AR** : **A**uto - **R**égressif) d'ordre 2M décomposable en une cascade de résonateurs à 2 pôles en série (tuyaux résonants). Le modèle du conduit nasal est un filtre pôle zéro **ARMA** (**A**uto **R**égressif à **M**oyenne **A**justée) et le rayonnement aux lèvres peut se modéliser par un filtre tout zéro (**MA** : **M**oyenne **A**justée). L'ensemble des conduits se comporte donc comme un système linéaire ARMA [7].

Modèle glottal :

$$G(z) = \frac{1}{(1- e^{-2\pi f_g t}\, Z^{-1})^2} \qquad \text{Avec } f_g = 100 \text{ Hz} \qquad (2.3)$$

Modèle du conduit vocal :

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

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

Modèle du conduit nasal :

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

Avec F_n et $\acute{F}_n$ formant nasal ou anti formant nasal et respectivement, $B_n$et $\acute{B}_n$ leurs bandes passantes.

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

$$H(z) = G(z).[(1 - \alpha)V(z)L(z) + \alpha N(z)] \qquad \text{Avec } 0 \leq \alpha \leq 1 \qquad (2.6)$$

Pour un son nasal $\alpha \cong 1$; pour un son non nasal $\alpha = 0$.

$H(z)$ Est en tout généralité un modèle ARMA d'ordre p

$$H(z) = \frac{B(z)}{A(z)} \qquad (2.7)$$

Dans le domaine temporel on aura :

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

Caractériser le signal $y_n$revient donc a estimé les coefficients $\{a_i;b_i\}$. Pour une source connue e_n (séquence d'impulsions ou bruit blanc). Souvent pour simplifier la résolution de ce problème, on suppose que b_i= 0, i ≥ 1 ce qui rend le modèle AR [7].

2.3.3.2 Analyse cepstrale

Le défaut majeur des méthodes d'analyse, comme la FFT, pour le calcul du spectre réside dans l'intermodulation source/conduit vocal qui rend difficile la mesure de la fréquence fondamentale 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 le spectre de la parole pour trouver les formants.

Pour cela, nous faisons 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 hypothèse qu'un signal excitateur est soit une séquence d'impulsions (périodiques, de période T_0, pour les sons voisés), soit un bruit blanc pour les sons non voisés, conformément au modèle

de production de la parole. Une transformation en Z permet de transformer la convolution en produit.

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

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

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

Par transformation inverse, nous obtenons le cepstre. L'expression du cepstre est donc :

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

Le cepstre qui ne fait appel à aucune information a priori sur le signal acoustique, est basé sur une connaissance du mécanisme de production de la parole. L'espace de représentation du cepstre ou espace quéferentiel est homogène par rapport au temps. Les premiers coefficients cepstraux contiennent l'information relative au conduit vocal. Cette contribution devient négligeable à partir d'un échantillon n_0 qui correspond à la fréquence fondamentale F_0 . Les pics périodiques visibles au-delà de n_0, reflètent les impulsions de la source.

Le spectre du cepstre pour les indices inférieurs à n_0 permet d'obtenir un spectre lissé, en éliminant les lobes secondaires dû à la contribution de la source. Ces deux contributions peuvent être séparées par une simple fenêtre temporelle notée F (liftrage) telle que le filtre rectangulaire.

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

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

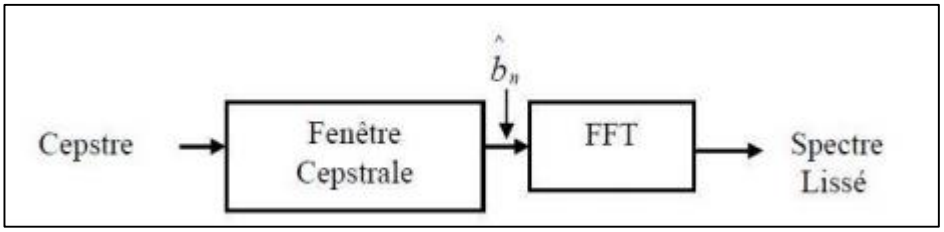


Figure 2.5 : *Obtention de la structure formantique à partir du cepstre* [9]

2.4 PRINCIPE DE LA 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 : ou bien 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 2.6) ;
- Un synthétiseur de parole comme étant la dernière étape de la transformation d'un certain nombre de paramètres de contrôle en parole [8].

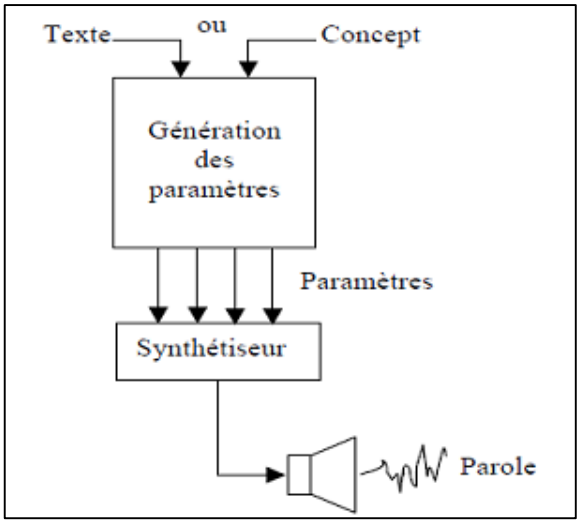


Figure 2.6 : *Système de synthèse de la parole* [8]

2.5 ARCHITECTURE D'UN SYSTEME DE SYNTHESE TTS

La synthèse de la parole à partir du texte (**T**ext-**T**o-**S**peech synthesis) a pour objectif de produire un signal de parole correspondant à un texte donné. L'architecture générale d'un système TTS de synthèse se compose ainsi de ces deux parties principales :

- les traitements linguistiques : cette première étape vise à analyser et à structurer le texte afin de déterminer un mode de prononciation cohérent, puis à transformer le texte analysé en une suite de sons de parole accompagnée d'indications concernant leur agencement ;
- la synthèse proprement dite : cette seconde étape consiste à générer un signal acoustique qui « retranscrit » cette suite de sons tout en possédant les caractéristiques apparentes de la parole naturelle (figure 2.7) [16].

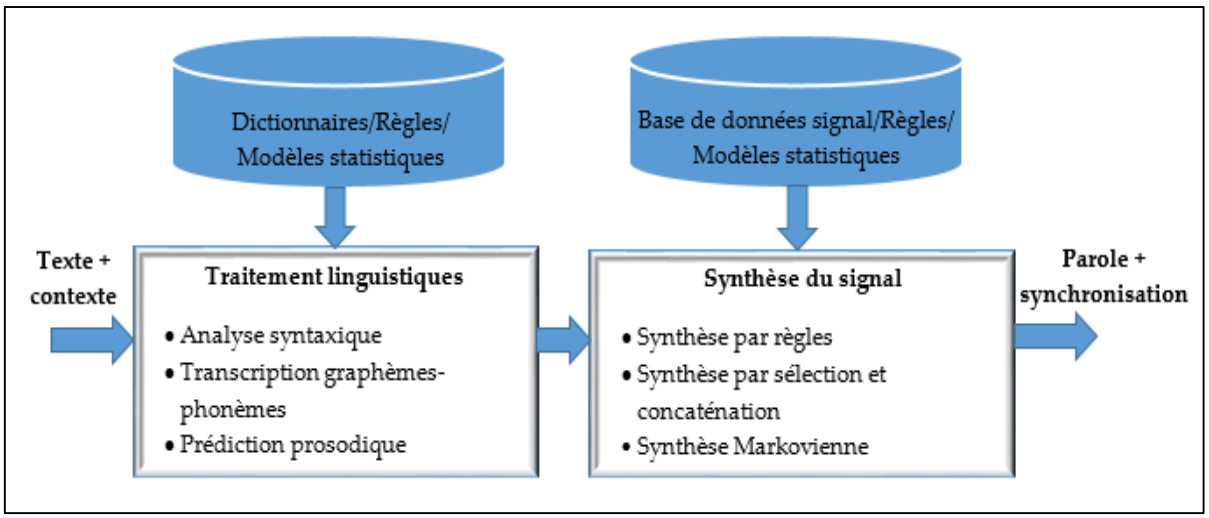


Figure 2.7 : *Diagramme fonctionnel d'un processus de synthèse d'un système TTS*

2.6 TRAITEMENT LINGUISTIQUE

Le bloc de traitements linguistiques (figure 2.7) regroupe les différents modules qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne de phonèmes éventuellement enrichis d'informations linguistiques et prosodiques caractérisant l'élocution. Ces différents modules sont : les prétraitements des éléments non lexicaux et l'analyse lexicale, l'analyse syntaxique, la transcription orthographique - phonémique et le traitement prosodique.

2.6.1 Prétraitement lexical et syntaxique

Cette étape de prétraitement permet de retranscrire en toutes lettres les chaînes non orthographiques. Il peut s'agir de chiffres, de dates (20/10/95, 19 Jan. 2008) ou plus généralement de sigles composés de caractères orthographiques et numériques [17].

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

2.6.1.1 Analyse lexicale

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

2.6.1.2 Analyse syntaxique

L'analyse syntaxique vise à déterminer la structure de la phrase. Elle est conduite par application de règles pouvant être de deux types. Dans certains cas, il peut s'agir heuristiques, résultant généralement de l'application de règles grammaticales standards (par exemple, on ne peut pas observer la succession de deux verbes conjugués). En complément ou à la place de ces heuristiques parfois très complexes, on utilise aussi fréquemment des règles probabilistes, exploitant des modèles de langage. Ces modèles sont fondés sur l'observation que toutes les séquences de catégories grammaticales dans une langue donnée ne sont pas équiprobables. La connaissance de la catégorie syntaxique exacte est également utile pour déterminer la prononciation correcte et notamment pour désambiguïser les homographes hétérophones [17].

2.6.2 Transcription Orthographique Phonétique (TOP)

Traditionnellement appelée **C**onversion **G**raphème-**P**honème(**CGP**), cette étape de Transcription constitue le noyau minimal, indispensable à tout système de synthèse de parole. Cette étape repose sur l'utilisation d'un automate paramétré appliquant un ensemble de règles de réécriture, qui permettent d'associer un phonème (ou un groupe de phonèmes) à un caractère (ou un groupe de caractères) orthographique en prenant en compte le contexte gauche et le contexte droit. Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales. Le nombre de règles nécessaires pour effectuer la TOP dépend de la langue que l'on considère [18].

2.6.3 Traitement prosodique

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

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

2.7 LES METHODES DE SYNTHESE DE LA PAROLE

Le choix d'une méthode de synthèse dépendra de l'application visée, de la qualité de la synthèse acceptée et du coût de revient du système. Il dépendra également de la taille du vocabulaire. Les systèmes sont dits à vocabulaire limité ou à vocabulaires illimité. Dans le premier cas, le degré de manipulation effectué sur le signal de parole pour obtenir une bonne qualité d'écoute est assez réduit : l'unité utilisée généralement est le mot. Dès que de nouvelles phrases se font nécessaires, l'explosion du vocabulaire oblige l'adoption de méthodes dont l'unité manipulée est inférieure au mot [18, 19].

Les méthodes de synthèse à vocabulaire illimité se divisent en synthèse par règles, synthèse par concaténation d'unités stockées et synthèse par systèmes dynamiques.

2.7.1 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 [18].

Les synthétiseurs par règles ont principalement la faveur des phonéticiens et des phonologistes. Ils permettent une approche cognitive, générative du mécanisme de la phonation [7] (figure 2.8).

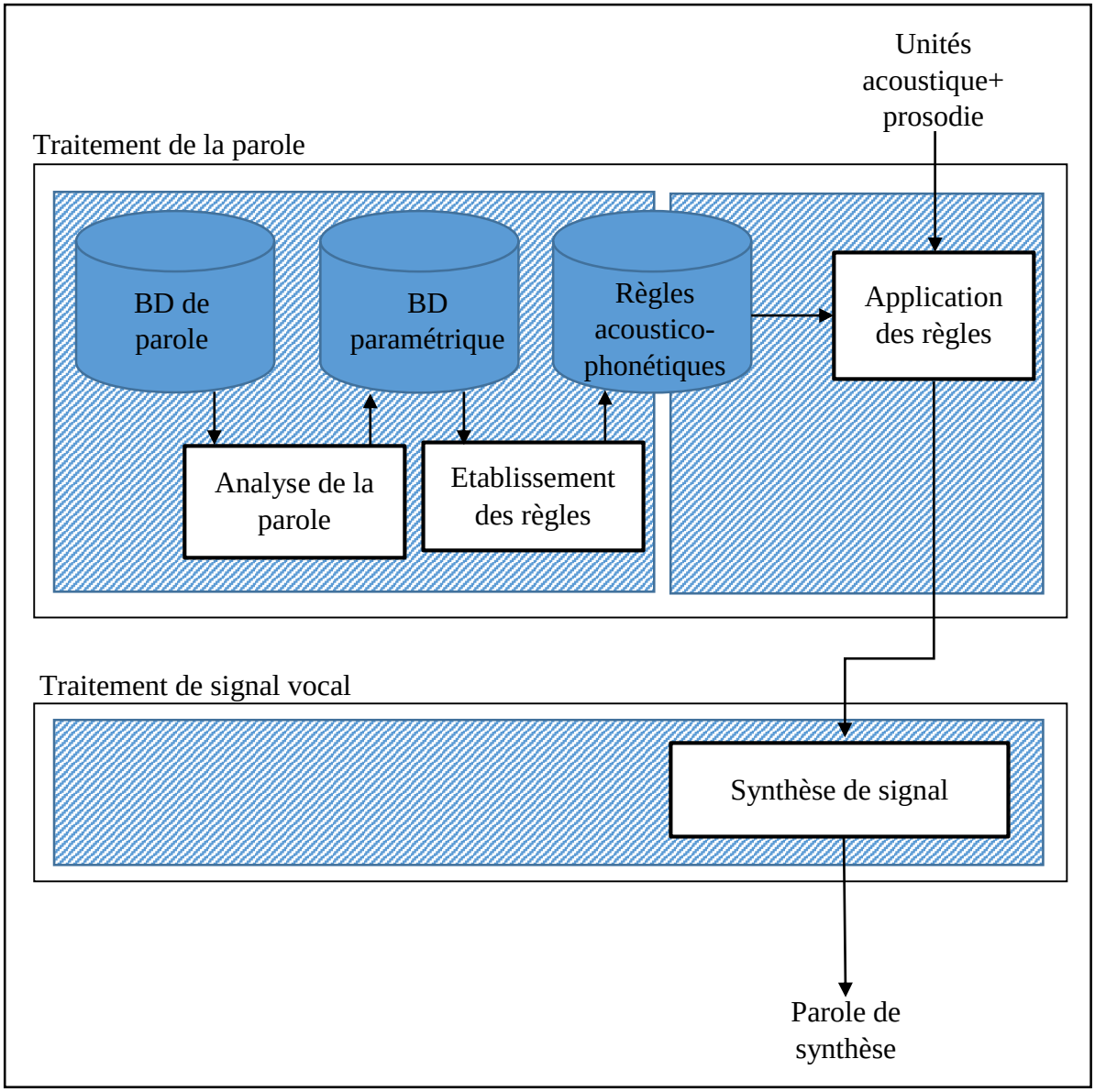


Figure 2.8 : *Schéma de conception et fonctionnement typique d'un système de synthèse par règles*

A partir du schéma précédent (figure 2.8) on peut résumer les étapes de la synthèse par règles comme suit :

- **enregistrement d'une BD de parole :** on fait lire par un locuteur professionnel d'un grand nombre de mots, généralement de type CVC (**C**onsonne-**V**oyelle-**C**onsonne) et on les enregistre sous forme numérique. Les mots sont choisis de façon à constituer un corpus représentatif des transitions phonétiques et des phénomènes de coarticulation dont on veut rendre compte ;
- **construction d'une BD paramétrique :** on modélise alors ces données numériques à l'aide d'un modèle paramétrique de parole, qui a pour rôle de séparer les contributions respectives de la source glottique et du conduit vocal et de présenter cette dernière sous

forme compacte, plus propice à l'établissement des règles. Celles-ci sont généralement proposées par des phonéticiens ;

- **établissement des règles :** on commence par inspecter globalement l'ensemble des données, de façon à établir la forme générale des règles à produire. On précise alors les valeurs numériques des paramètres intervenant dans ces règles (les fréquences des formants, ou les durées des transitions, par exemple) par un examen minutieux du corpus. Il est à remarquer que cette étape d'estimation est menée sur une seule voix : un moyennage interlocuteur aurait peu de signification dans ce contexte. De même, les règles provenant de synthétiseurs déjà existants ne peuvent resservir que dans la mesure où elles modélisent des caractéristiques articulatoires générales plutôt que des particularités du locuteur ayant enregistré le corpus [8, 18]. La mise au point du synthétiseur s'achève par un long processus d'essais-erreurs, afin d'optimiser la qualité de la synthèse.
- **synthèse du signal :** lorsqu'un nombre suffisant de règles ont été établies, la synthèse proprement dite peut commencer. Les entrées phonétiques du synthétiseur déclenchent l'application de règles, qui produisent elles-mêmes un flux de paramètres liés au modèle de parole, utilisé. Cette séquence temporelle de paramètres est alors transformée en parole par un synthétiseur, qui implémente les équations du modèle.

La synthèse par règles a connu un essor considérable dans les années 60-70. Elle n'est plus guère utilisée aujourd'hui que lorsque les contraintes de mémoire et de temps de calcul sont très importants. La qualité des voix disponibles n'est en effet pas aussi bonne qu'en synthèse pas concaténation, pour un coût de développement supérieur [14].

2.7.2 Synthèse Par Concaténation (SPC)

La Synthèse Par Concaténation produit de la parole par la mise bout à bout des segments acoustiques (phones, diphones, syllabes, mots, etc.) puisés dans une base de données de parole préenregistrée par un locuteur et segmentée, en imposant au signal de parole, synthétisé la prosodie prédite par le module de prédiction de prosodie.

La mise en œuvre d'une SPC nécessite des réponses à quelques questions relatives à la conception des bases de données de parole, à la sélection des unités dans ces bases, à leur modification prosodique et à leur concaténation. La figure 2.9 présente le schéma général de la chaîne de ces traitements. Le bloc hachuré représente le traitement relatif à l'étiquetage et à la segmentation [20].

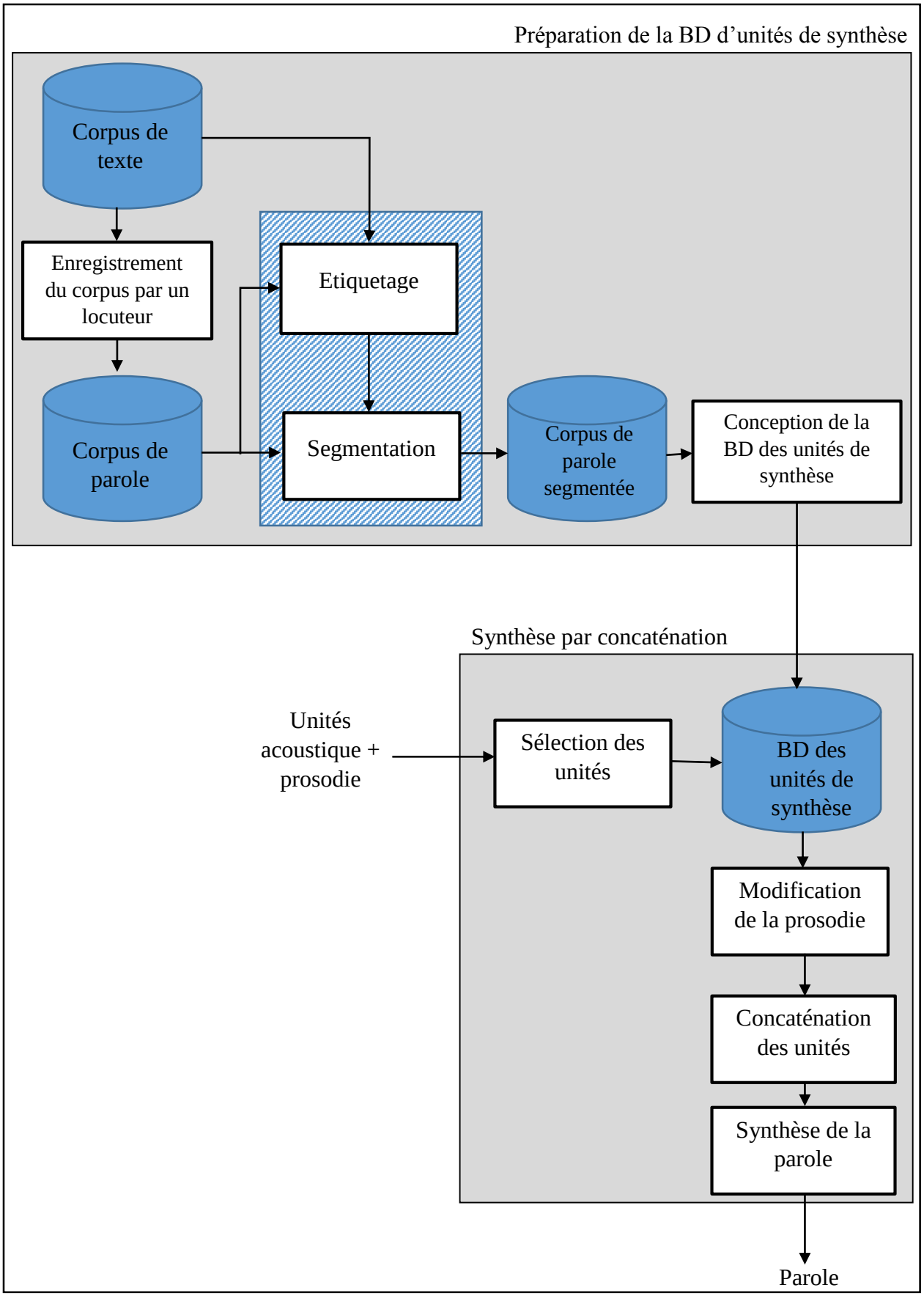


Figure 2.9 : *Processus de préparation d'une BD d'unités de synthèse et d'un SPC* [20]

2.7.3 Synthèse par système dynamique

Cette méthode se propose à modéliser l'organisation des gestes articulatoires ou des trajectoires acoustico-visuelles inspirées par le système biologique et les théories du contrôle moteur [18]. Un autre aspect prometteur de cette méthode est l'intégration possible entre signal visuel et acoustique dans un espace articulatoire commun. L'importance de la composante visuelle se fait sentir notamment dans les environnements bruités, où la récupération des cibles phonémiques est largement aidée par l'information fournie par les lèvres.

2.8 LES TECHNIQUES DE LA SYNTHESE DE LA PAROLE

Il est possible de rassembler les différentes techniques offertes par les synthétiseurs de parole. En première approximation nous pouvons les regrouper comme suit, dans le domaine :

- **spectral :** synthétiseur à formants, et à prédiction linéaire ;
- **temporel :** synthétiseur par formes d'ondes, cette technique englobe l'ensemble des synthétiseurs telle que la compression de la parole numérisée MIC (Modulation par Impulsions Codées) et les autres basé sur la synthèse par concaténation.
- **articulatoire :** synthétiseur articulatoire, cette technique est potentiellement considérée comme la technique la plus performante car elle reflète théoriquement le processus physiologique, Elle est basée sur une modélisation géométrique du conduit vocal. Elle consiste à représenter le conduit vocal comme un tube de section variable, avec des embranchements et des sections parallèles, puis à y simuler le trajet des ondes produites au niveau de la glotte. Les modèles d'écoulement d'air (mécanique des fluides), de sources et de propagation acoustique (phénomènes physiques), en association avec des modèles articulatoires (mécaniques), permettent de constituer un synthétiseur articulatoire complet, contrôlé par deux jeux de paramètres : les paramètres supra-laryngés qui commandent le modèle articulatoire, et un jeu de paramètres qui pilotent les cordes vocales (pression sub-glottique, longueur des cordes vocales et hauteur de la glotte au repos).

2.8.1 Synthèse par formants

La technique la plus largement répandue de synthèse pendant de dernières décennies a été probablement la synthèse par formants qui est basée sur le modèle source/filtre de la parole décrit en chapitre 1. Il existe deux structures de base en général, en parallèle et en cascade, mais pour une meilleure exécution un certain type de combinaison de ces derniers est habituellement

employé. La synthèse par formants fournit également un nombre infini d'unités de sons, qui le rend plus flexible que par exemple les techniques basées sur la concaténation.

Les formants sont les fréquences propres du conduit vocal lors de la production d'un son voisé. Dans cette technique, le filtre du conduit vocal est composé d'un certain nombre de résonateurs similaires au nombre de formants de la parole naturelle. Les résonances formantiques naturelles du conduit vocal sont simulées par des filtres résonants du deuxième ordre caractérisés par une fréquence centrale et une largeur de bande spécifiques. Au moins trois formants sont généralement exigés pour produire le discours intelligible et jusqu'à cinq formants pour produire la parole de haute qualité.

La synthèse basée sur les règles formantiques est basée sur un ensemble de règles employées pour déterminer les paramètres nécessaires qui synthétisent une expression désirée à l'aide d'un synthétiseur à formants. Les paramètres d'entrée peuvent être par exemple les suivants, où le quotient d'ouverture de la glotte signifie le rapport du temps d'ouverture de la glotte durant toute la durée de la période (figure 2.10).

- Fréquence fondamentale de composante voisée (F_0)
- Quotient ouvert exprimé d'excitation (QO)
- Degré d'exprimer dans l'excitation (Vo)
- Fréquences de formants et amplitudes (F_1, …, F_3, et A_1,… , A_3)
- Fréquence d'un résonateur de basse fréquence additionnel (FN)
- Intensité de la basse et haute fréquence région (ALF, AHF)

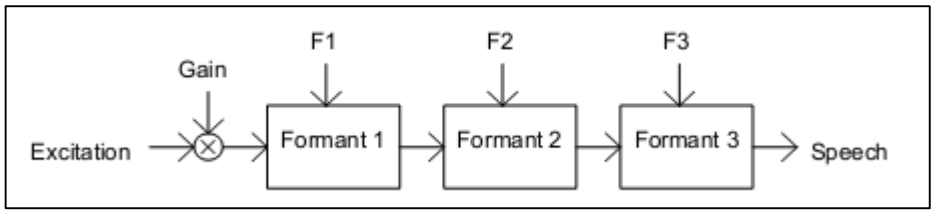


Figure 2.10 : *Structure basique d'un synthétiseur par formant en cascade* [21]

Une synthèse de qualité est obtenue par la simulation des quatre premiers formants. L'implantation de ces filtres peut se faire soit en cascade, soit en parallèle, soit de façon mixte (Figure 2.11). Pour les sons voisés, ce système est excité par une onde périodique dont la forme est aussi proche que possible de l'onde glottale. Pour les sons non voisés, l'excitation est un bruit blanc. Où :

- F_0 et A_0 sont respectivement la fréquence fondamentale et l'amplitude de la composante voisée ;
- Fn et Qn sont respectivement les fréquences de formants et leur bande passante ;
- VL et VH sont respectivement l'amplitude basse et haute de la composante voisée ;
- FL et FH sont respectivement la fréquence basse et haute de la composante non voisée ;
- QN est la valeur de la bande passante du formant nasal à 250 Hz.

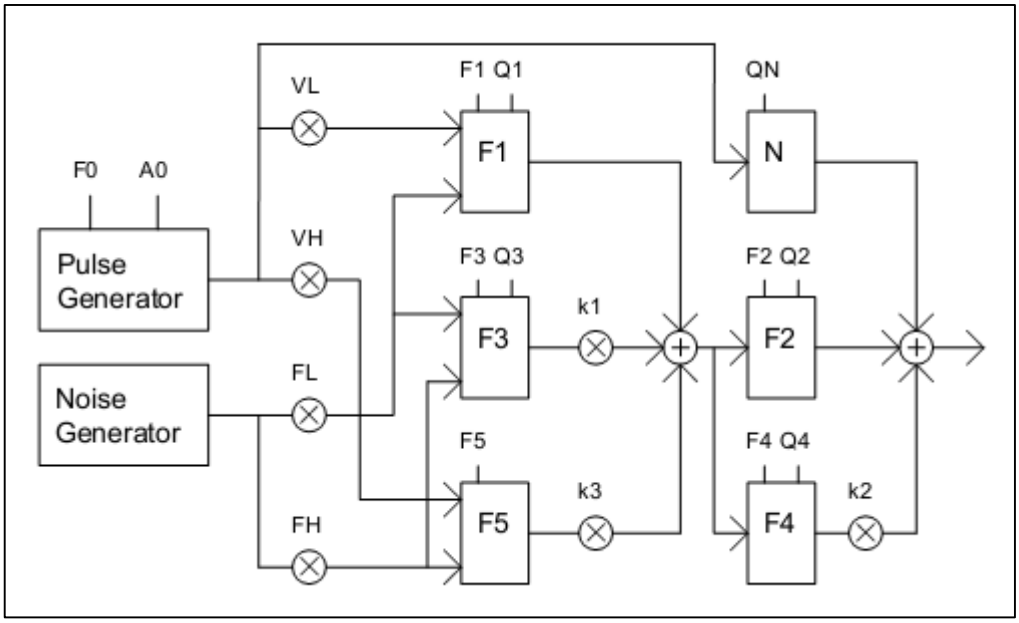


Figure 2.11 : *Synthèse par formants, Modèle de PARCAS* [21]

2.8.2 Synthèse par prédiction linéaire

La synthèse par prédiction linéaire est une méthode qui s'inscrit dans le cadre de la théorie source/filtre. Elle a largement été utilisée dans les systèmes par concaténation, car elle permet un codage rapide des unités à concaténer contrairement à la synthèse par formants. La parole de synthèse produite en utilisant la synthèse LP est loin d'être parfaite. La synthèse LP fondée sur la méthode d'autocorrélation ne reproduit pas correctement les fréquences et les bandes passantes des formants lors de la synthèse avec une fréquence fondamentale différente de la fréquence initiale [16]. La synthèse du signal de parole avec sa fréquence fondamentale d'origine conduit aussi à une dégradation du signal due à l'excitation utilisée. En effet, cette dernière est de nature très simplifiée par rapport au signal d'erreur réel. Plus particulièrement, dans le cas de sons voisés, d'autres informations ne sont pas prises en compte, conduisant à la dégradation du signal.

Pour pallier ce problème, une technique dite de prédiction linéaire par impulsions multiples est mise en œuvre. Elle consiste à construire une excitation composée de plusieurs impulsions pour chaque trame de parole analysée. La synthèse avec la combinaison de cette excitation et les coefficients LP produit un signal de parole très proche du signal naturel.

2.9 L'APPLICATION DE LA SYNTHESE DE LA PAROLE

Les applications de la synthèse de la parole sont très diverses et ciblent aussi bien un public spécifique comme les non-voyants que le grand public. Nous en exposons ci-dessous les plus courantes (tableau 2.1)

Tableau 2.1 : *Les application de la synthèse de la parole* [8,16]

Télécom	Portails vocaux, lecture de SMS ou d'emails, services de messagerie unifiée, CRM, annuaires et annuaires inverses, web parlant, standard automatique, serveurs vocaux, centres d'appels, etc.
Multimédia	Outils d'aide à la lecture et à l'apprentissage de langues, de relecture, d'assistance aux handicapés, assistant personnel, solution de lecture d'emails ou de fax, web parlant, assistance en ligne, productivité, agents en ligne, etc.
Automotive	Systèmes de navigation embarqué ou déporté, aide à la navigation, systèmes d'alerte et de diagnostic embarqués, info-trafic, lecture d'emails, réservation en ligne, accès Internet, etc.

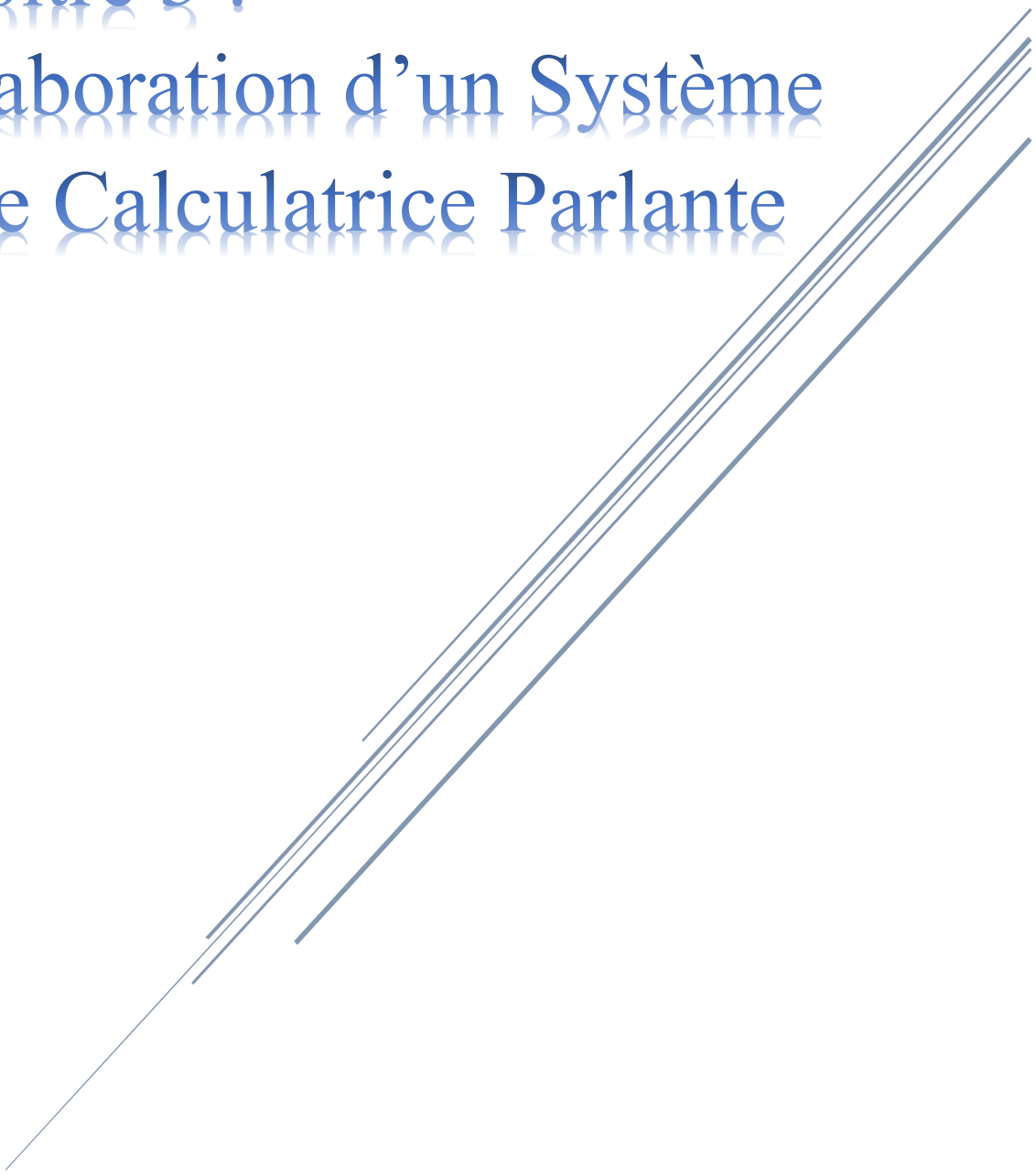
- **Outils d'aide aux personnes handicapées :** le système de SAT joue le rôle de lecteur de textes sur écran ou à distance via un serveur vocal, avec la possibilité pour l'utilisateur de configurer le système, en choisissant la voix de synthèse (masculine ou féminine), le débit de la parole et même la langue souhaitée si le document existe en plusieurs langues.
- **Dans les services de lecture vocale :** annuaire inverse où l'utilisateur obtient des informations sur une personne à partir de son numéro de téléphone ; consultation des télécopies par voie orale ; lecture vocale de journaux ; consultation de la messagerie écrite, etc.
- Dans le cadre des projets visant le dialogue homme-machine, des prototypes de système de dialogue existent en laboratoire où la synthèse vocale est couplée à un système de reconnaissance et exploite ainsi différentes ressources (informations syntaxiques, sémantiques…) utiles à la génération acoustique du signal. La SAT connaîtra sans doute de

nouvelles applications dans les années à venir, compte tenu de l'attrait grandissant pour les nouvelles technologies. Cela dépendra aussi de la robustesse et de la flexibilité de ces systèmes et de leur capacité à s'intégrer dans des applications manipulant la langue sous ses diverses formes (écrite ou or ale). Le principal enjeu des systèmes du futur est de pouvoir dialoguer dans la même langue que l'utilisateur en intégrant dans le signal synthétisé les caractéristiques propres de sa voix, comme le timbre ou encore le rythme.

2.10 CONCLUSION

Dans ce chapitre nous avons introduit les principales méthodes et techniques de synthèse de la parole courantes à ce jour, ainsi les différentes domaines des applications de la synthèse, ce qui nous aide pour introduire le chapitre suivant, qui est le sujet de notre étude « Calculatrice Parlante pour les Malvoyants ».

Chapitre 3 :
Elaboration d'un Système de Calculatrice Parlante



3.1 INTRODUCTION

Le but de notre travail est de réaliser un système de synthèse de la parole par unités variables, en vue d'élaborer un **S**ystème de **C**alculatrice **P**arlante pour les **N**on-**V**oyants **A**rabophones (**SCPNVA**), en se basant sur la méthode de concaténation de mots combinés.

Dans ce chapitre, nous allons présenter le système de la « SCPNVA », avec une brève explication du processus de construction de notre BD, à l'aide de l'outil d'analyse Praat. Nous finissons par la procédure de concaténation.

3.2 SITUATION DES NON VOYANTS EN ALGERIE

L'Algérie dispose d'un parc public composé de 271 établissements spécialisés à caractère administratif, doté de la personnalité morale et de l'autonomie financière. Ces établissements sont gérés par le secteur de la Solidarité Nationale, parmi ces derniers, on dénombre 19 **E**coles pour **J**eunes **A**veugles (**EJA**) réparties dans 19 Wilayas avec une capacité théorique de 3190 places pédagogiques, et un effectif réel de 2014 élèves, soit un taux d'occupation de 63,13%. Les élèves bénéficient, au sein de ces écoles, d'une scolarité adaptée à la nature, et au degré de leur cécité ainsi qu'une prise en charge psychologique (pour les enfants Mal Voyants ou Non-Voyants) ; leur permettant de suivre une scolarité normale [22].

3.2.1.1 Didactique des sciences pour les Non-Voyants Algériens

La didactique des sciences, et particulièrement celle des mathématiques, devient une nécessité pour l'enseignement aux élèves ayant des troubles d'apprentissage. La majorité des enseignants sont confrontés à une nouvelle gamme d'élèves et des programmes d'enseignement qui sortent du cadre classique avec lequel ont été formés auparavant.

La plupart des problèmes que rencontrent les élèves NV sont liés aux difficultés de prise d'information visuelle et de transmission par l'écrit de ses connaissances et ceci dans toutes les disciplines. Si tous les contenus doivent être abordés, des difficultés liées à la lecture et à l'écriture naissent souvent. Pour compenser ces aspects négatifs :

- l'élève doit apprendre à développer des capacités et des moyens de compensation qui lui permettront d'obtenir une efficience comparable à celle de ses camarades ;
- l'enseignant va mettre en place des adaptations pédagogiques et proposer des aides techniques indispensables.

Conséquences de le Cécité totale :

- toutes les informations collectives doivent être auditives ;
- la lecture et l'écriture se font par l'intermédiaire des outils de transcription en code Braille et des techniques informatiques [23].

3.2.1.2 Ecriture Braille

Un caractère braille se compose de (un à six) points disposés sur deux colonnes de trois points. On obtient ainsi 63 caractères écrits. On a l'habitude de numéroter ces points de haut en bas de un à trois pour la colonne de gauche et de quatre à six pour la colonne de droite [24] (figure 3.1).

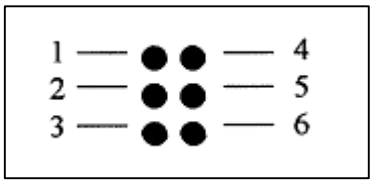


Figure 3.1 : *Cellule unitaire d'une écriture Braille* [24]

Les chiffres et le signe infini en Braille sont représentés dans le tableau 3.1.

Tableau 3.1 : *Ecriture Braille des chiffres et le signe infini* [24].

Configuration de points	Signification	Symbole
(Points 1 - 6)	Un	1
(Points 1 - 2 - 6)	Deux	2
(Points 1 - 4 - 6)	Trois	3
(Points 1 - 4 - 5 - 6)	Quatre	4
(Points 1 - 5 - 6)	Cinq	5
(Points 1 - 2 - 4 - 6)	Six	6
(Points 1 - 3 - 4 - 5 - 6)	Sept	7
(Points 1 - 2 - 5 - 6)	Huit	8
(Points 2 - 4 - 6)	Neuf	9
(Points 3 - 4 - 5 - 6)	Zéro	0
(Points 4 - 5, 1 - 4)	Infini	∞

Les signes des opérations mathématiques en Braille sont représentés dans le tableau 3.2.

Tableau 3.2 : *Ecriture Braille des signes d'opérations mathématiques*

Configuration de points	Signification	Symbole
(points 2 - 3 - 5)	Plus	+
(Points 3 - 6)	Moins	−
(Points 3 - 5)	Multiplié par	×
(Points 4 - 6, 3 - 4)	Divisé par	÷
(Points 2 - 3 - 5, 3 - 6)	Plus ou moins	±
(Points 3 - 5, 3 - 5)	Point multiplicatif, produit scalaire	·
(Points 4 - 5, 3 - 5)	Produit vectoriel, produit extérieure	^
(Points 2 - 3 - 5 - 6)	Egal	=
(Points 4 - 6, 2 - 3 - 5 - 6)	Défirent de	≠
(Points 4, 1 - 2 - 6)	Inferieur	<
(Points 4 - 5, 1 - 2 - 6)	Inférieur ou égal	≤
(Points 4, 3 - 4 - 5)	Supérieur	>
(Points 4 - 5, 3 - 4 - 5)	Supérieur ou égal	≥
(Points 4 - 5 - 6, 2 - 3 - 5)	Union	∪
(Points 4 - 5, 2 - 3 - 5)	Intersection	∩
(Points 4 - 5 - 6, 3 - 5)	Factoriel	!

3.3 PRESENTATION DE LA « CP » POUR LES NON VOYANTS

Avec le développement très important des moyens de calcul et de stockage, les interfaces Homme Machine sont devenues de plus en plus proches de l'interaction humaine naturelle. La CP est un outil d'aide aux élèves non-voyants du niveau primaire. Cette dernière est basée sur la synthèse de la parole par concaténation de mots combinés. Ce système joue le rôle de lecteur de texte de l'opération de calcul et le résultat obtenu, via une base de données préconçue.

Nous avons choisi les nombres en braille pour la conception matérielle, étant donnée que cette CP est destinée à l'utilisation des non-voyants.

3.4 METHODOLOGIE DU TRAVAIL

La synthèse par concaténation est basée sur une succession d'étapes, la figure 3.2 montre la méthodologie à suivre pour l'élaboration de notre « SCPNVA ».

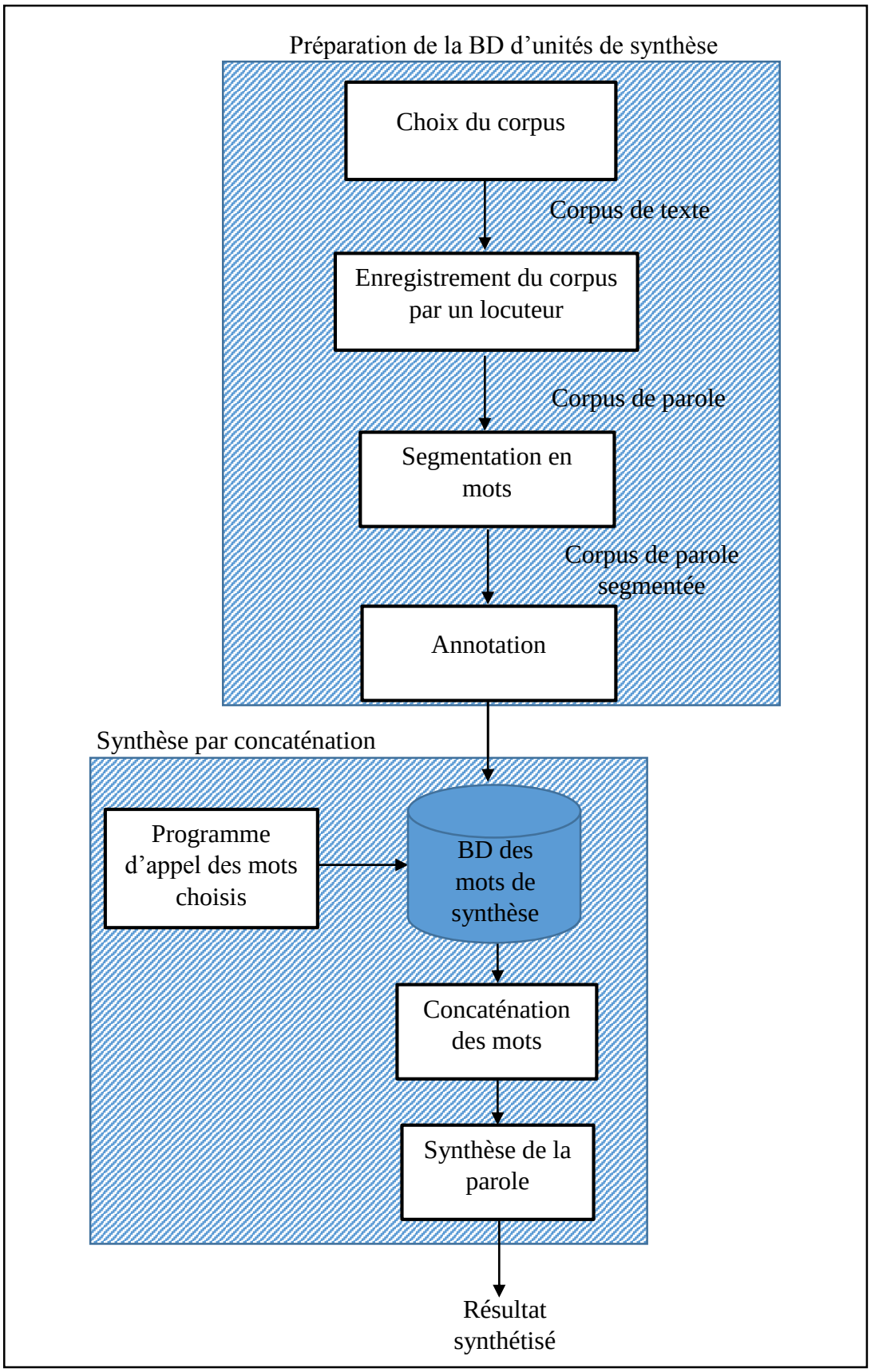


Figure 3.2 : *Schéma méthodologique du Système de la Calculatrice Parlante*

3.5 OUTIL D'ANALYSE PRAAT

Praat est un logiciel libre pour l'analyse, la manipulation et l'annotation des sons. Ces fonctionnalités en font un outil complet, en particulier pour l'étude de la parole. Il permet également de tracer des graphiques, construire des grammaires basées sur la théorie de l'optimalité, de faire une synthèse articulatoire, de simuler des réseaux de neurones et de faire des analyses statistiques. P. BOERSMA et D. WEENINK de l'Institute of Phonetic Sciences de l'Université d'Amsterdam ont créé Praat en 1996 et continuent activement de développer cet outil de manière très interactive avec la communauté des utilisateurs [25].

Il a été conçu à la fois pour les non-experts en traitement de la parole grâce à ses interfaces graphiques et menus simplifiés. Il propose une interface de deux fenêtres : Praat Objects, est l'espace de travail (figure 3.3) ; Praat Picture, est un afficheur de graphes (figure 3.4.).

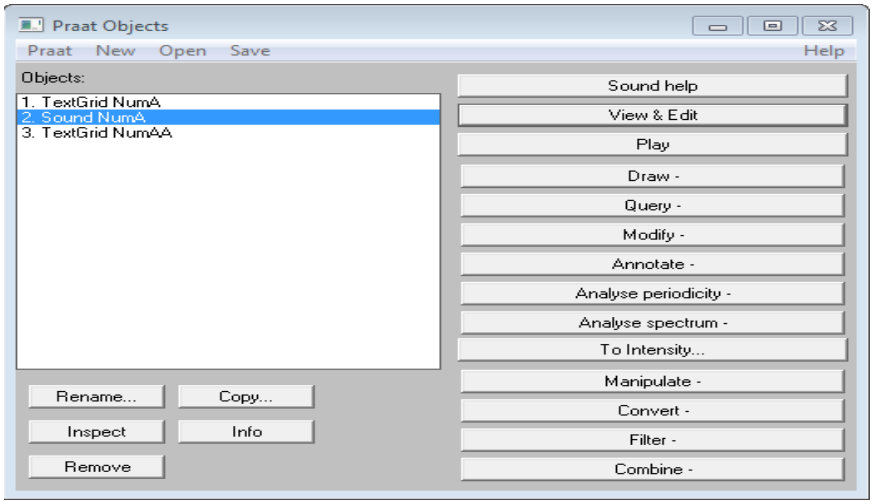


Figure 3.3 : *la fenêtre Praat Objects*

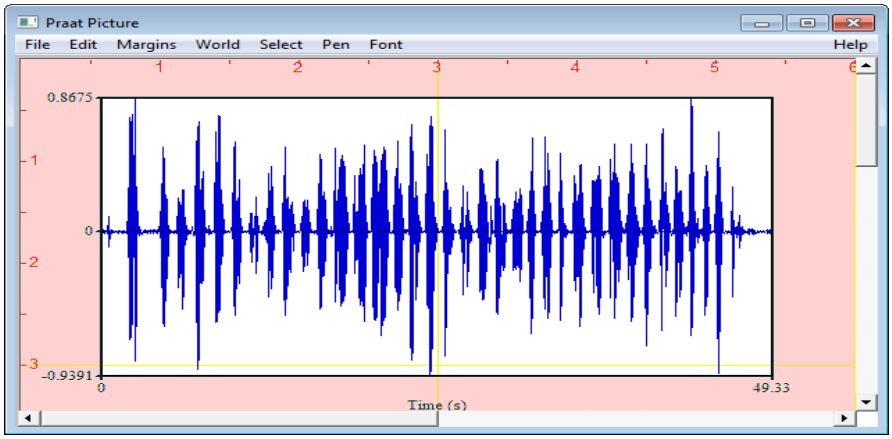


Figure 3.4 : *la fenêtre Praat Picture*

Pour les utilisateurs avancés ils ont de nombreuses possibilités de manipulations, d'analyses et de Scripting. Ce programme offre la possibilité d'effectuer de multiples tâches [25] :

- enregistrer des fichiers audio qui peuvent être ensuite analysés sous Praat. ils peuvent être aussi codés selon une multitude de formats audio ;
- segmenter, transcrire et annoter des fichiers audio dont la taille peut aller jusqu'à 2 Giga bytes, c'est-à-dire 3 heures d'enregistrement stéréo de qualité CD ou 16 heures d'enregistrement mono à 22 kHz. ces enregistrements peuvent être effectués sous Praat ou provenir d'autres fichiers audio a divers format ;
- effectuer des analyses phonétiques et acoustiques au niveau segmental. Il permet de calculer des paramètres prosodiques comme l'intensité, la fréquence fondamentale, le voisement, le timbre, etc., et ceci selon plusieurs algorithmes.
- De mener des analyses spectrographiques et des mesures précises telles que la durée du **VOT** (**V**oice **O**n **T**ime) des plosives, les valeurs des différents formants d'une voyelle, etc;
- étudier les paramètres prosodiques (F0, durée et intensité), modifier par stylisation des courbes de fréquence fondamentale et d'intensité (figure 3.5);
- effectuer des manipulations et des modifications du signal de parole (utilisation de filtres, analyse-synthèse, …etc.) ;
- construire des outils d'apprentissage (réseau de neurones et élaboration de grammaires dans le cadre de la théorie de l'optimalité (**OT** : **O**ptimalité **T**heory) ;
- écrire des scripts pour effectuer plus rapidement certaines tâches d'analyse, d'extraction d'information ou d'édition, etc.

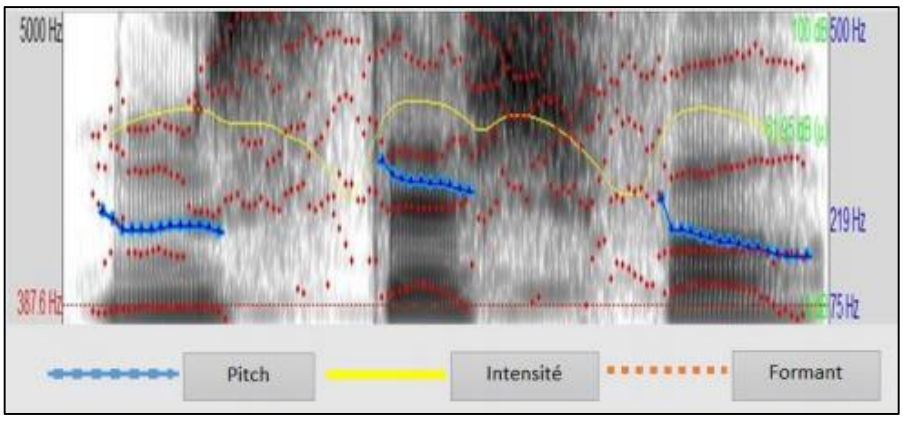


Figure 3.5 : *Spectrogramme et paramètres (F0, durée et intensité)*

3.6 CONSTRUCTION DE LA BASE DE DONNEES

3.6.1 Choix du corpus

Le choix du corpus est un élément clé pour la qualité d'un système de SPC. Définir le corpus revient à déterminer l'ensemble des unités à enregistrer de façon à obtenir un certain espace acoustico-prosodique meilleur.

Dans le cadre de notre travail, nous avons utilisé un corpus de parole qui englobe toutes les unités qui peuvent être utilisées dans une opération de calcul, qui se nomme « SCPNVA ». Ce corpus est continu et composé de 33 mots (Tableau 3.3).

Tableau 3.3 : *Corpus « SCPNVA »*

Mots en Arabe	Transcription en API	Mots en Arabe	Transcription en API
صفر	[*sˤifr*]	ستون	[*sittu:n*]
واحد	[*wa:ħid*]	سبعون	[*sabʕu:n*]
اثنان	[*ʔiθna:n*]	ثمانون	[*θama:nu:n*]
ثلاثة	[*θala:θa*]	تسعون	[*tisʕu:n*]
أربعة	[*ʔarbaʕa*]	مئة	[*miʔa*]
خمسة	[*xamsa*]	مئتان	[*miʔata:n*]
ستة	[*sitta*]	ألف	[*ʔalf*]
سبعة	[*sabʕa*]	ألفان	[*ʔalfa:n*]
ثمانية	[*θama:nija*]	آلاف	[*ʔa:laf*]
تسعة	[*tisʕa*]	زائد	[*za :ʔid*]
عشرة	[*ʕaʃara*]	ناقص	[*na:qisˤ*]
أحدا عشر	[*ʔaħada:ʕaʃar*]	في	[*fi:*]
اثنا عشر	[*ʔiθna:ʕaʃar*]	على	[*ʕala:*]
عشرون	[*ʕiʃru:n*]	فاصل	[*fa:sˤil*]
ثلاثون	[*θala:θu:n*]	تساوي	[*tusa:wi:*]
أربعون	[*ʔarbaʕu:n*]	وَ	[*wa*]
خمسون	[*xamsu:n*]		

3.6.2 Enregistrement du corpus

La génération d'une voix de synthèse est une étape importante dans le développement d'un système de synthèse par concaténation, car une voix de qualité médiocre dégrade d'autant la

qualité globale du système. En général, le choix d'une voix se fait selon un protocole bien défini. D'abord, un locuteur ou plus, présélectionné, enregistre sa voix dans des conditions de prise de son optimales (chambre sourde, avec un micro de bonne qualité). Ensuite, une évaluation permet de désigner la voix finale.

En outre, il faut bien vérifier que chaque unité a été bien prononcée selon la configuration désirée et le cas échéant de procéder à un nouvel enregistrement des unités mal prononcées. Signalons également que le fait de contraindre fortement le locuteur va prononcer un ensemble d'unités relativement courtes (des mots voire des expressions brèves) selon des schémas prosodiques prédéfinis, peut conduire à l'obtention d'une parole peu naturelle. Une façon plus réaliste et plus satisfaisante de procéder est de rechercher, parmi un vaste corpus textuel, un jeu minimal de phrases permettant une couverture symbolique acceptable.

Exemple : l'enregistrement de son **[***wa***]**, afin de pouvoir prendre en considération les effets de coarticulation existants entres les phonèmes, pour cela, nous avons enregistré 3 phrases, contenant ce son :

« **مئة وَ اثنا عشر**» [*miʔa**wa**ʔiθna:ʕaʃar*] ;

« **مئتان وَ واحد** » [*miʔata:n**wa**wa:ħid*] ;

« **ألف وَ سبعمئة** » [*ʔalf**wa**sabʕumiʔa*].

Après la phase de segmentation nous allons choisir le son le plus naturel.

L'enregistrement de notre corpus « SCPNVA », s'est fait dans une chambre sourde avec un minimum de bruit et de réverbération, et des conditions d'enregistrement moyennes :

- les données sont échantillonnées avec une fréquence de 44,1 kHz codés sur 32 bits qui correspondent à une bonne qualité de la parole ;
- **le format (nombre de canaux) :** mono sound ;
- **logiciel utilisé :** « Praat » version 5.4 ;
- **le type de parole :** mots en parole continue ;
- **les signaux acoustiques sont enregistrés en format :** (wav).

Figure 3.6 est une représentation temporelle du corpus « SCPNVA »

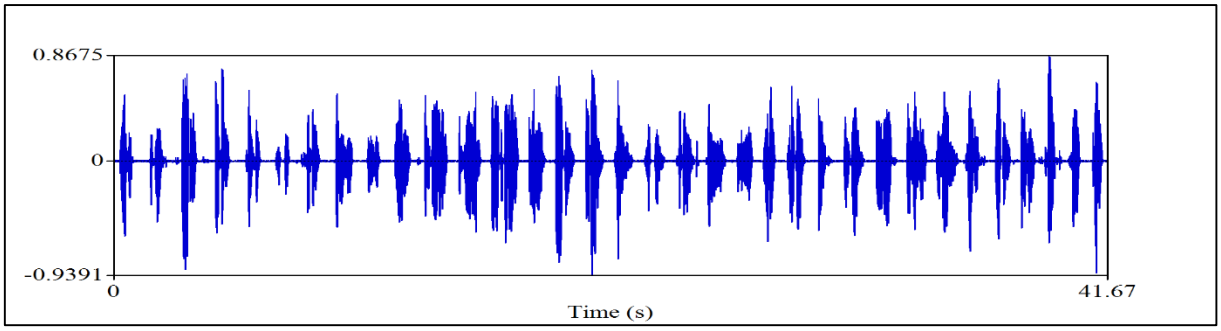


Figure 3.6 : *Visualisation du signal de corpus « SCPNVA » enregistré*

La figure (3.7) montre les principales informations acoustiques sur notre corpus avant segmentation.

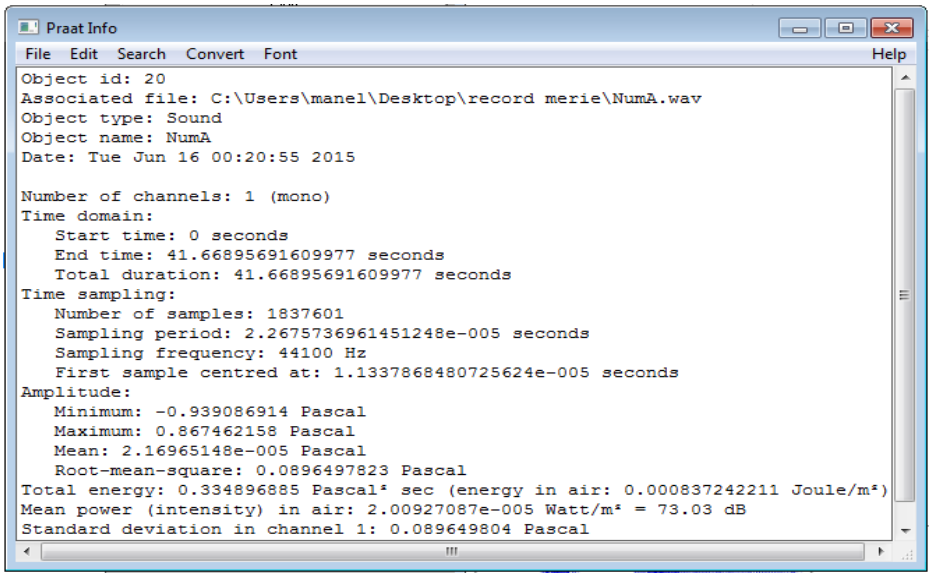


Figure 3.7 *: Information sur le corpus « SCPNVA »*

- durée totale : 41,67 s
- une amplitude : minimum : -0,939 pascal
  maximum : 0,867 pascal
  moyenne : $2{,}169.10^{-5}$ pascal
- une énergie totale de 0,335 pascal².sec ;
- puissance moyenne (intensité) en air : $2.009.\ 10^{-5}$ watt/m² = 73.03 db ;

3.6.3 Segmentation en mots

La segmentation consiste à extraire les mots du corpus enregistré. Cette segmentation a été effectuée manuellement à l'aide de l'outil **Praat.** La visualisation d'audiogramme et de

spectrogramme à la fois nous aide pour bien ajuster les limites de segmentation de chaque mot (figure 3.8).

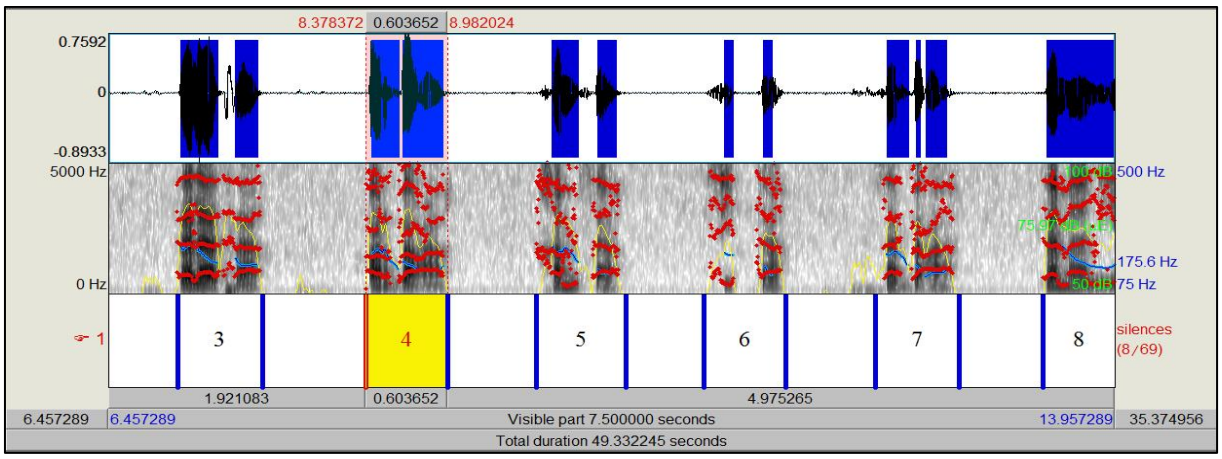


Figure 3.8 : *Segmentation en mots de corpus « SCPNVA »*

Les figures 3.9, 3.10 montre l'audiogramme de quelque segments (mot [*ɵala:ɵu:n*], et mot [*tisʕa*]).

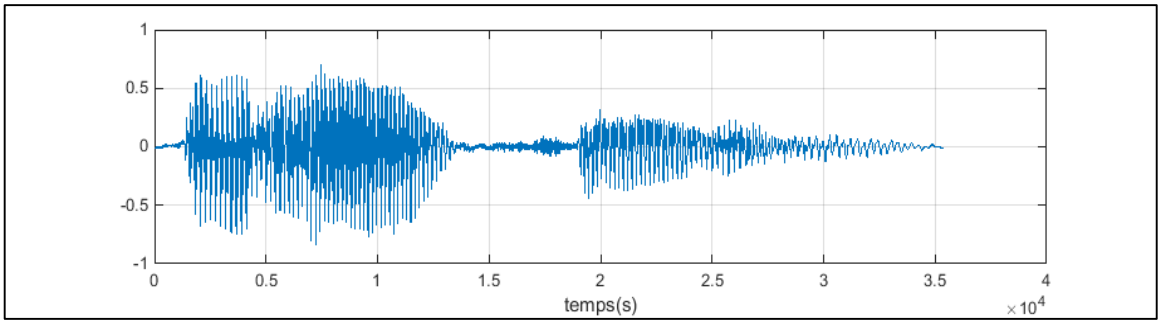


Figure3.9 : *Audiogramme de segment 30 [ɵala:ɵu:n]*

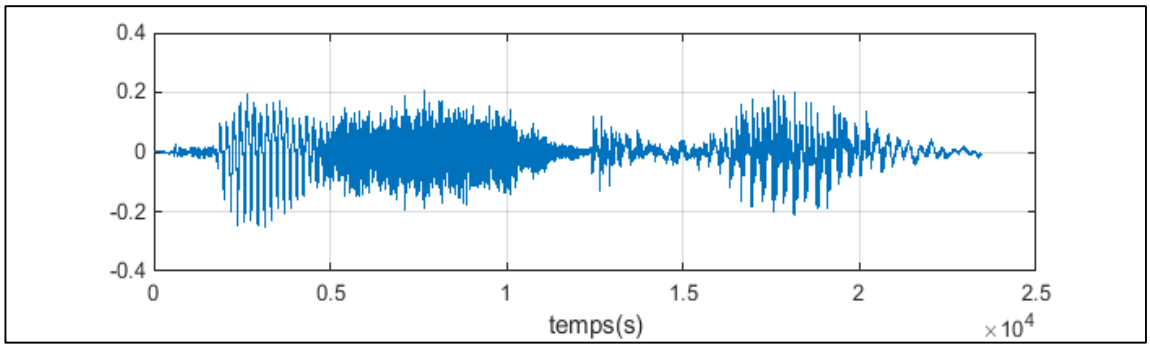


Figure 3.10 : *Audiogramme de segment 9 [tisʕa]*

3.6.4 Annotation

L'annotation est de faire décrire chaque mot segmenté par une note descriptive (symbole, chaine de caractères, etc.). Nous avons choisis la plus simple note qui peut décrire chaque mot, pour faciliter la manipulation des segments lors de la concaténation (tableau 3.4).

Tableau 3.4 : *Annotation des mots segmentés de corpus « SCPNVA »*

Mots	annotation	Mots	annotation
صفر	0	ستون	60
واحد	1	سبعون	70
اثنان	2	ثمانون	80
ثلاثة	3	تسعون	90
أربعة	4	مئة	100
خمسة	5	مئتان	200
ستة	6	ألف	1000
سبعة	7	ألفان	2000
ثمانية	8	آلاف	a1000
تسعة	9	زائد	+
عشرة	10	ناقص	-
أحداعشر	11	في	.
اثناعشر	12	على	..
عشرون	20	فاصل	,
ثلاثون	30	تساوي	=
أربعون	40	وَ	wa
خمسون	50		

3.7 CONCATENATION DES MOTS

Notre système « SCPNVA », se base sur la synthèse par concaténation des formes d'ondes d'une opération de calcul et son résultat, donc nous devons suivre cette procédure, la concaténation de :

- Premier terme (un nombre de plusieurs mots, $n \geq 1$), n= nombre de segments (mots) ;
- Signe d'opération (un seul mots, $n = 1$) ;
- Deuxième terme (un nombre de plusieurs mots, $n \geq 1$) ;
- Signe d'égalité (un seul mots, $n = 1$) ;
- Résultat (un nombre de plusieurs mots, $n \geq 1$).

Et comme il y a trois nombre à synthétisés, nous avons construit sous le logiciel MATLAB une fonction « NumA », spécialement pour la concaténation des nombres arabes. Pour cela un ordre d'exécution doit être respecté, la figure 3.11 montre cet ordre pour un nombre en AS de taille n=3.

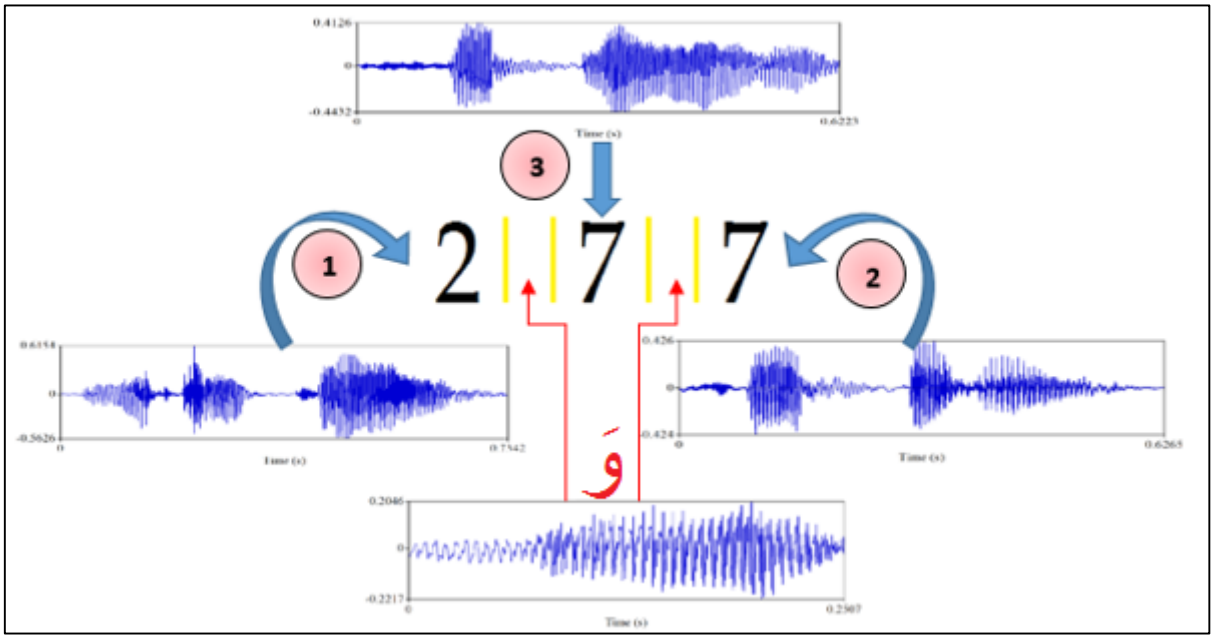


Figure 3.11 : *Ordre de concaténation d'un nombre en AS*

La figure 3.12 montre audiogramme de la sortie de fonction « NumA », avec le texte « 277 » en entré.

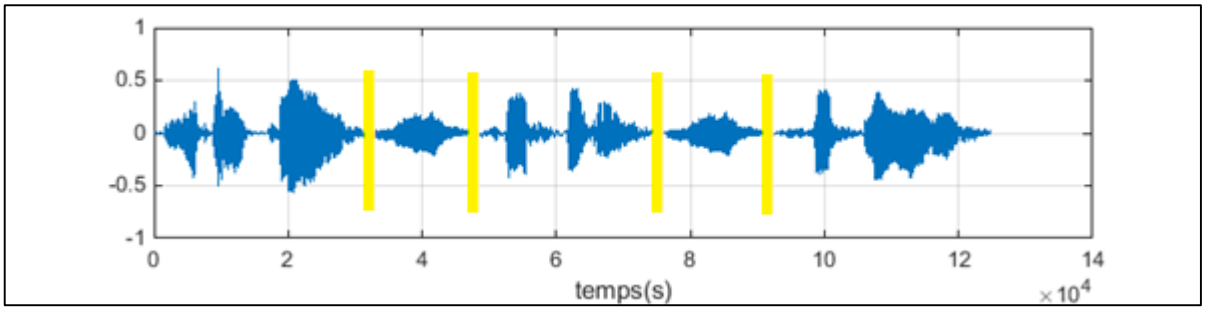


Figure3.12 : *Audiogramme de phrase concaténée « 277 »[miʔata:nwasabʕawasabʕu:n]*

3.8 CONCLUSION

Dans ce chapitre, nous avons présenté le système de la CP, avec une brève explication du processus de construction de notre BD, et aussi, la procédure de concaténation. Afin d'évaluer le système, dans le chapitre suivant des tests seront appliqués sur ce système.

Chapitre 4 :
Evaluations et Discussions

4.1 INTRODUCTION

Le but de ce chapitre est d'expliquer le fonctionnement de l'interface graphique « **SCPNVA** » qui est faite sous le logiciel MATLAB. Afin d'évaluer notre système, nous allons faire une étude comparative entre les phrases synthétisé et naturelles, et tester l'intelligibilité.

4.2 ARCHITECTURE DU PROGRAMME « SCPNVA »

Avant la conception matérielle d'un outil technique, une simulation de son système doit être élaborée, pour assurer le bon fonctionnement et améliorer les performances.

La simulation de notre « SCPNVA » est faite sous le logiciel MATLAB. Pour faciliter la tâche, nous avons décomposé le système en blocs fonctionnels (figure 4.1), chaque bloc représente une fonction principale de traitement.

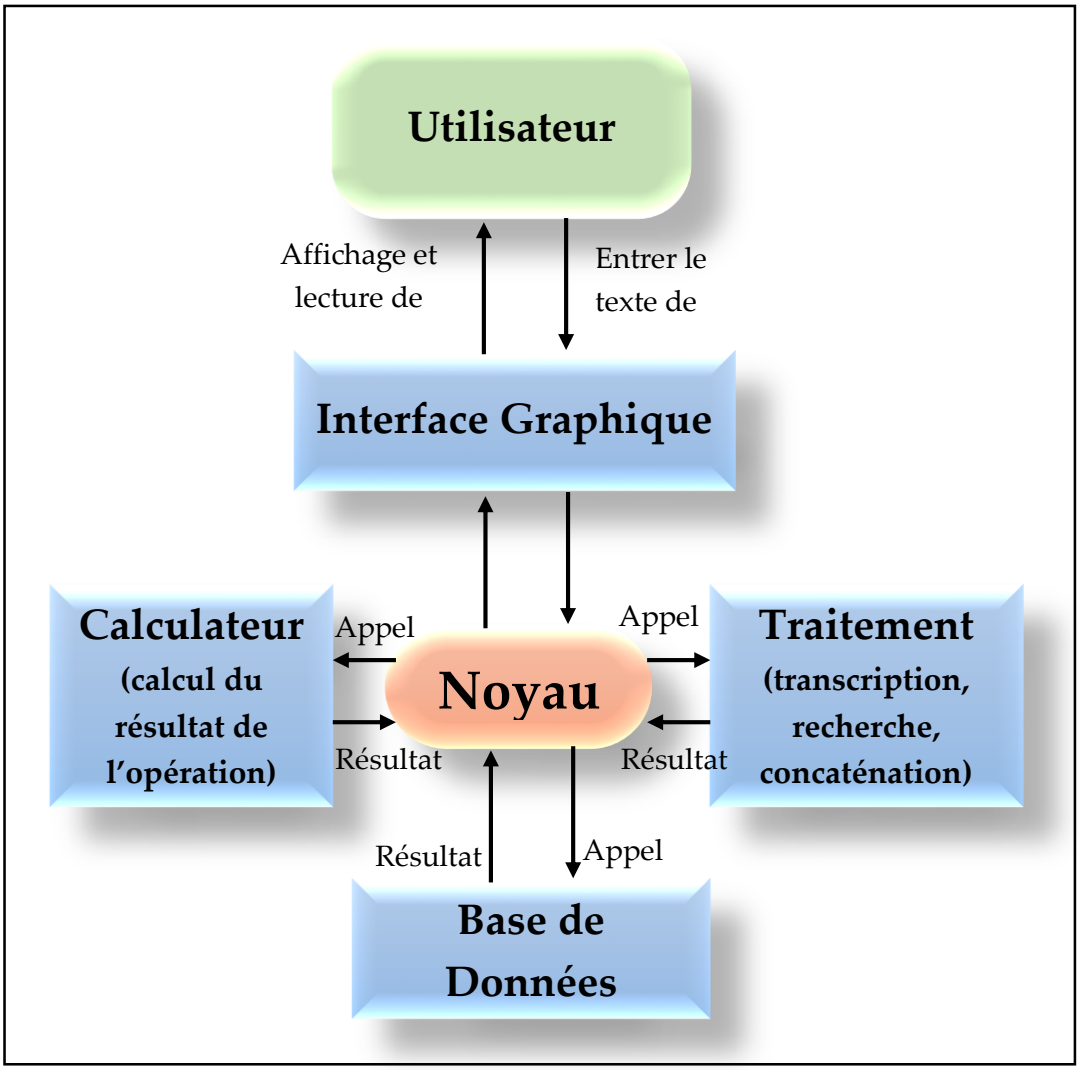


Figure 4.1 : *Schéma fonctionnel du « SCPNVA »*

4.2.1 Noyau

Le noyau est un programme administrateur, qui fait l'appel aux fonctions nécessaires dans chaque étape d'exécution, et gère l'affichage et le stockage de résultat.

4.2.2 Interface graphique

Une interface graphique a été conçue pour faciliter le test du système de la calculatrice parlante (figure 4.2), pour la conception matérielle une représentation en code Braille des chiffres et des signes doit être faite.

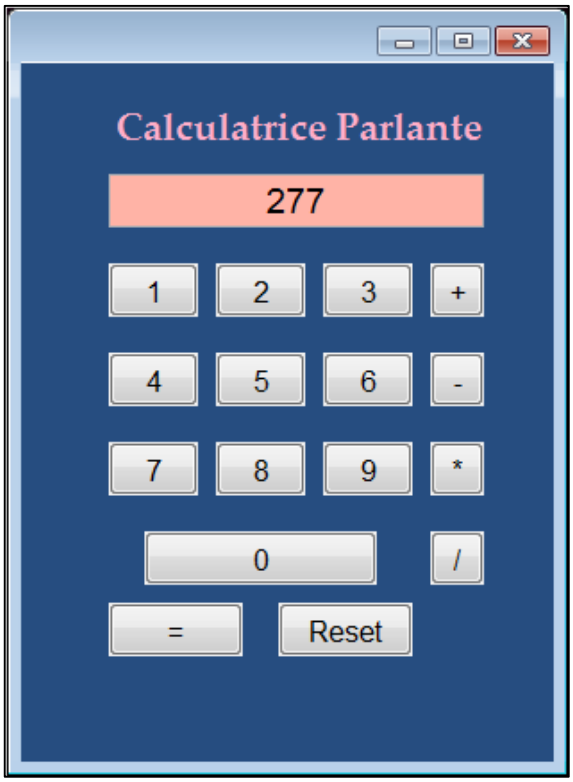


Figure 4.2 : *Interface graphique du « SCPNVA »*

4.2.3 Calculateur

Le programme calculateur est une fonction a comme entrées(le premier terme, le signe d'opération, le deuxième terme), et elle a comme sortie le résultat de calcul.

4.2.4 Fonction de traitement

La fonction de traitement est la fonction principale du système, elle a le rôle de :

- Transcription du chaque signe de texte en un son correspondant ;
- Concaténation des unités selon une règle prédéfinie.

4.3 EVALUATION

4.3.1 Etude comparative

Cette étude représente une comparaison des phrases « nombres concaténés» du « SCPNVA » avant (signal vocal original) et après concaténation. Prenant cinq échantillons, prononcés en AS par une locutrice. Pour les cinq phrases avant concaténation nous les avons obtenus par un enregistrement en parole continue avec la même locutrice et pour celles après concaténation, nous avons fait un réenregistrement du signal de sortie de notre interface graphique en temps réel, et dans les mêmes conditions d'enregistrement du corpus SCPNVA. Cette étude représente un test objectif de l'intelligibilité et le naturel de la parole synthétisée.

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

- générale : concernant la taille, la durée et l'énergie ;
- formantiques : pour les quatre premiers formants ;
- de l'intensité ;
- fréquentielle.

4.3.1.1 Analyse générale :

Cette analyse est faite pour les nombres : N_{77} ; N_{123} ; N_{1024} ; N_{1205} ; N_{30574}, avant et après concaténation (Tableau 4.1).

Tableau 4.1 *: Analyse général de quelques nombres avant et après concaténation*

	N_{77}		**N_{123}**		**N_{1024}**		**N_{1205}**		**N_{30574}**	
	avant	après	avant	après	avant	après	avant	après	avant	après
Taille [Ko]	194	159	302	371	276	312	272	317	376	382
Durée [s]	1.58	1.47	1.75	2.00	1.97	2.42	1.58	1.7	3.82	4.43
Amplitude Min [Pascal]	-0.41	-0.44	-0.64	-0.89	-0.65	-0.59	-0.47	-0.57	-0.75	-0.83
Amplitude Max [Pascal]	0.55	0.42	0.87	0.71	0.86	0.75	0.49	0.61	0.64	0.75
Energie [10^{-5} J/m^2]	2.41	2.74	6.61	7.60	7.98	6.71	4.90	5.68	8.12	9.07

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

$$\text{Précision (\%)} = \frac{|\text{Valeur avant} - \text{Valeur aprés}|}{\text{Valeur avant}} * 100 \qquad (4.1)$$

Pour illustrer et comparer ces données nous présentons le graphe suivant (figure 4.3)

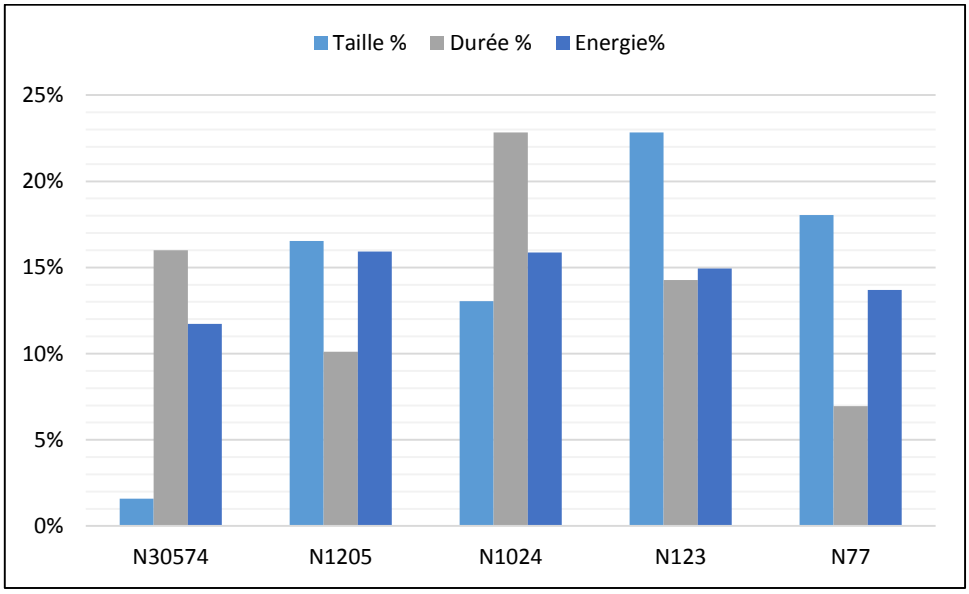


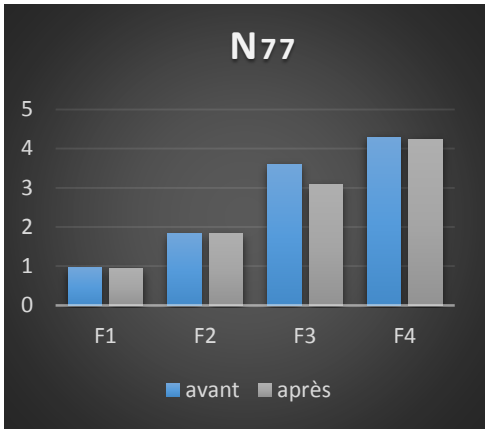
Figure 4.3 : *Précision de : Taille, Durée, Energie, des nombres concaténées*

4.3.1.2 Analyse formantique :

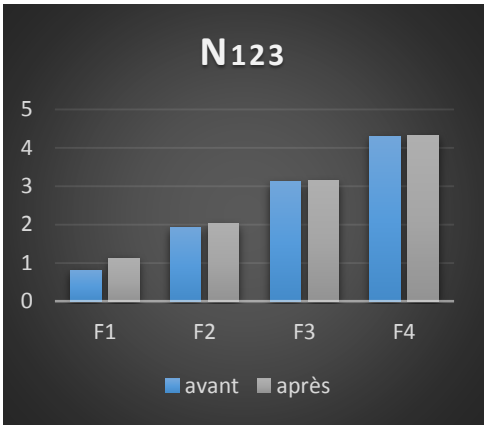
Nous obtenons les valeurs moyennes des quatre premiers formants pour les cinq nombres avant et après concaténation (Tableau 4.2).

Tableau 4.2 *: Les valeurs moyennes des formants avant et après concaténation*

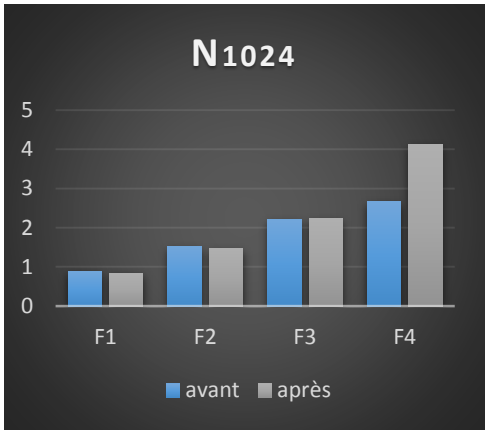
Formants [KHz]	**N_{77}**		**N_{123}**		**N_{1024}**		**N_{1205}**		**N_{30574}**	
	avant	après	avant	après	avant	après	avant	après	avant	après
F1	0,657	0,925	0,814	1,122	0,886	0,836	0,646	0,597	0,412	0,925
F2	1,836	1,839	1,917	2,021	1,518	1,464	2,032	2,082	2,635	1,839
F3	3,598	3,077	3,119	3,143	2,201	2,239	3,079	3,424	3,502	3,077
F4	4,300	4,240	4,299	4,329	2,661	4,122	4,456	4,471	4,403	4,240



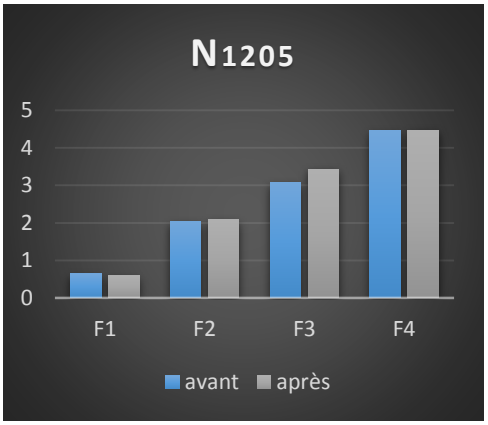
a)



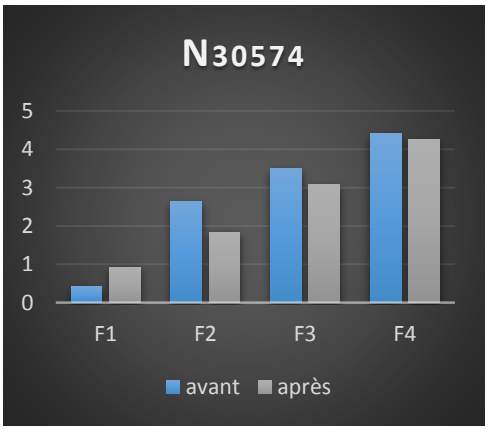
b)



c)



d)



e)

Figure 4.4 : *Formants avant et après concaténation de :*

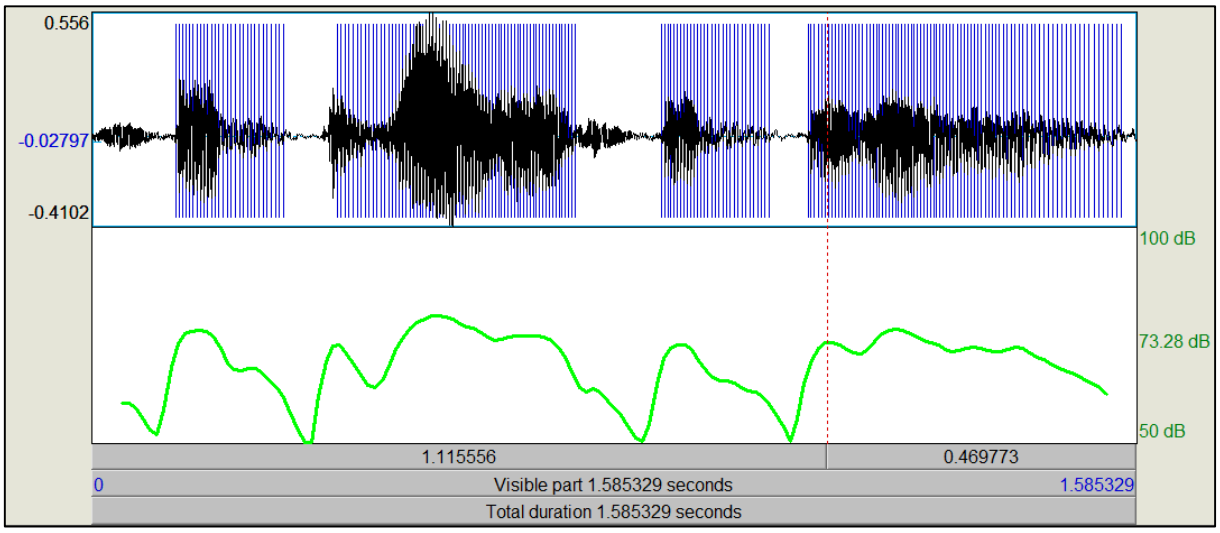
a) N_{77} ***b)*** N_{123} ***c)*** N_{1024} ***d)*** N_{1205}

e) N_{30574}

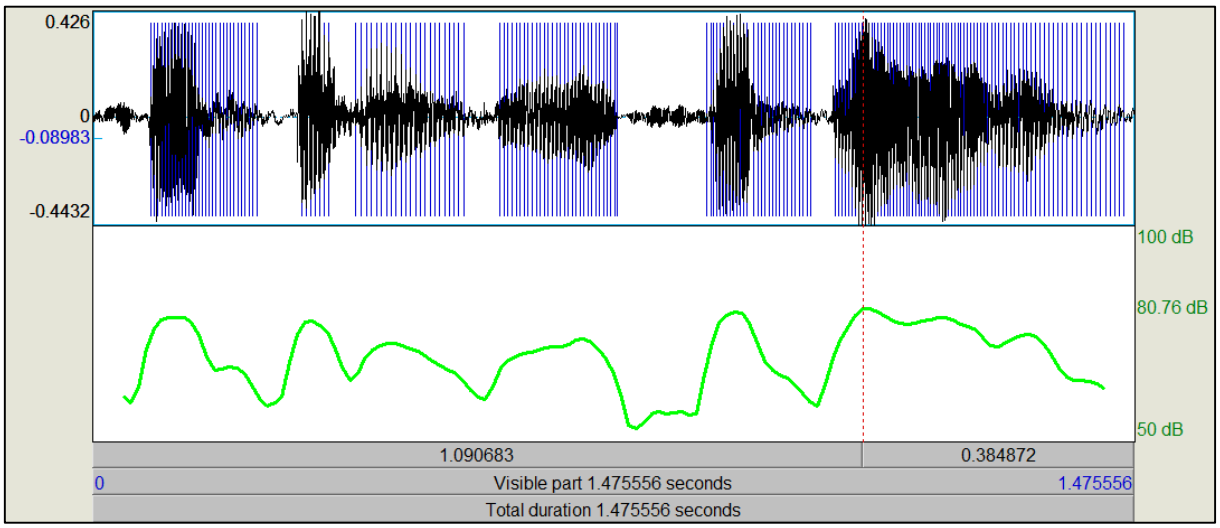
Interprétations : Remarquons ici (figure 4.4) une variation négligeable des formants, avant et après concaténation (e.g : formants de N_{1205}), ce qui est le résultat de la bonne segmentation et concaténation des unités. Exceptons deux cas (F4 de N_{1024} et F2 de N_{30574}) dont l'écart est un peu grand.

4.3.1.3 Analyse de l'intensité :

Pour une étude comparative, une comparaison d'intensités et nécessaire. A l'aide de l'outil Praat. Nous visualisons la courbe d'intensité de N_{77}, avant et après concaténation (figure 4.5).



a)



b)

Figure 4.5 : *la courbe d'intensité de N77*

a) *avant concaténation*

b) *après concaténation*

Intensité moyenne : ***a)*** 70.96 dB ; **b)** 58.25 dB

Interprétations : Remarquons que la variation d'intensité est similaire avant et après concaténation, avec une légère différence entre les intensités moyennes (12 ,71 dB), ce qui est dû à l'élimination des zones de silence lors de segmentation (les zones de silence ont une intensité moyenne faible).

4.3.1.4 Analyse fréquentielle :

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

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

1) Sélectionnons de la mélodie N_{77} (sound 77), puis choisissons (View & Edit) (figure 4.6) ;

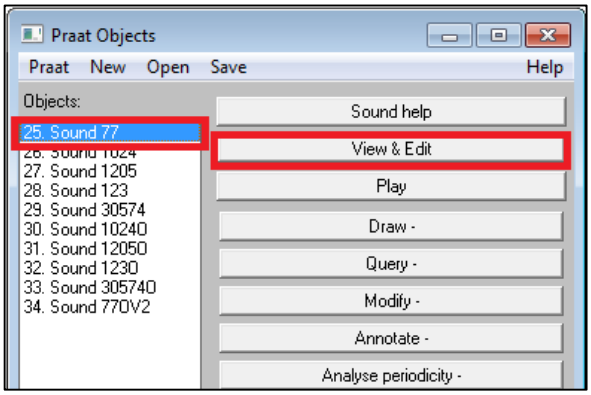
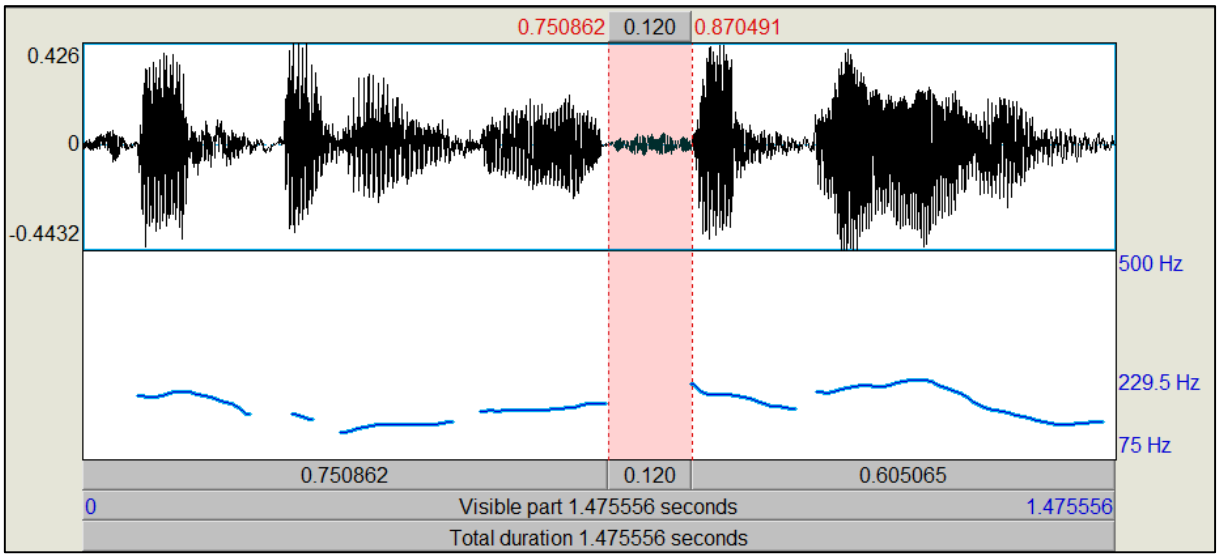
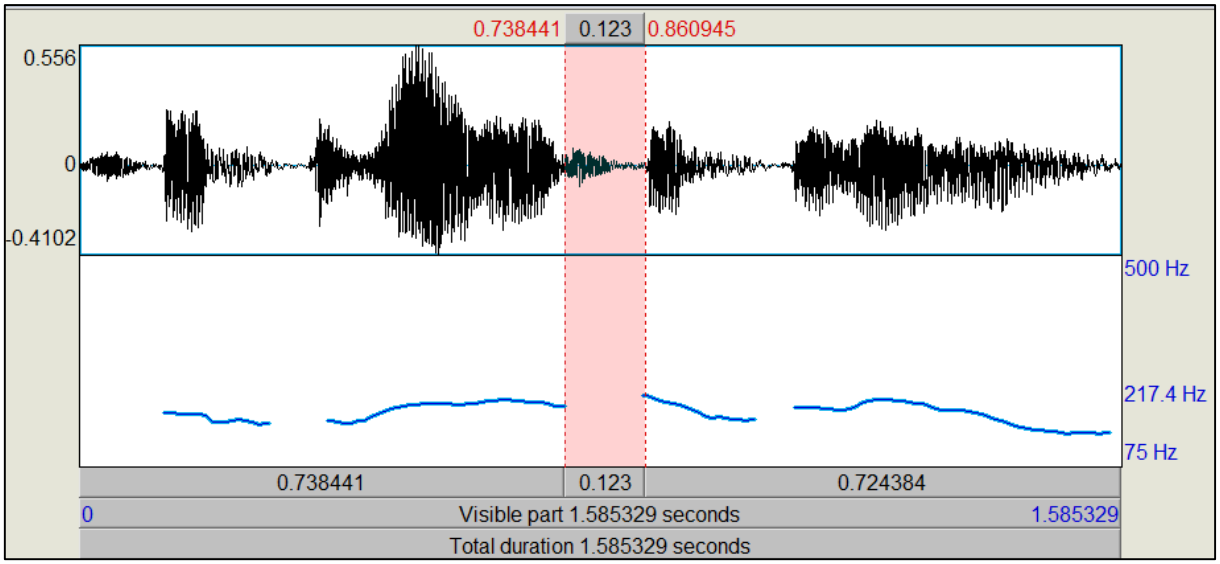


Figure 4.6 : *Visualisation des caractéristiques du son*

2) Une fenêtre s'ouvre, choisissons Pitch → show pitch (figure 4.7)



a)



b)

Figure 4.7 : *Visualisation de la fréquence fondamentale F_0 de N_{77}*

a) *avant concaténation ;*

b) *après concaténation.*

Remarquons que :

- la variation de F_0 de la phrase N_{77}, est presque la même dans les deux cas a et b ;
- a) 6 zones de son non voisé et 6 zones de son voisé ;
- b) 4 zones de son non voisé et 4 zones de son voisé.

Interprétation : l'élimination des zones de silence, justifie la diminution de nombres des zones des sons non voisés ou voisés.

4.3.1.5 Interprétation générale :

Nous remarquons pour chaque phrase (nombres concaténés) comparée du « SCPNVA » que les variations des paramètres sont acceptables, qui sont la taille mémoire, la durée et l'énergie. la variation de précision est due aux zones de silences que nous avons segmentées, dont laquelle ces zones ont une énergie, une durée et une taille mémoire propre a eu. nous constatons une grande similarité entre les trois premiers formants obtenus, elle se définit par une précision avant et après concaténation qui ne dépasse pas les 5 %. Même chose pour l'intensité moyenne et la fréquence fondamentale (pitch), nous avons obtenu de bons résultats avec des faibles variations. Ce qui justifie la bonne qualité de la parole obtenue.

4.1.1 Intelligibilité du système

Le test d'intelligibilité comprend 10 personnes de moyen âge (6 femmes, 4 hommes) qui ont participé à une session expérimentale d'exécution d'une opération de calcul, choisie aléatoirement, et d'écoute de résultat. Le test est répété trois fois successivement. Nous avons considère cinq niveaux de réponses (mauvais, médiocre, passable, bon et excellent). Les résultats obtenus sont définis dans le tableau 4.3 et la figure 4.8

Tableau 4.3 : *Résultat de test d'intelligibilité*

	Mauvais	**Médiocre**	**Passable**	**Bon**	**Excellent**
Décision	00	00	01	05	04
Pourcentage(%)	00	00	10	50	40

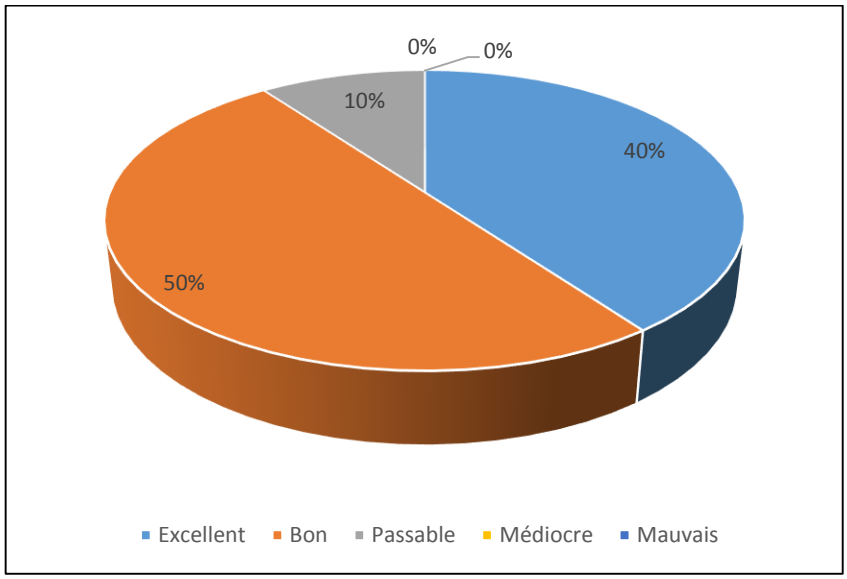


Figure 4.8 : *Décision sur la parole synthétisée*

Interprétation

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

4.2 CONCLUSION

Dans ce chapitre nous avons expliqué le fonctionnement du système de simulation « **SCPNVA** » qui est fait sous le logiciel MATLAB. Et afin d'évaluation notre système, nous avons fait une étude comparative entre les phrases synthétisé et naturelles, et un test d'intelligibilité comprenant 10 personnes de moyenne âge. Le système a répondu par des très bonnes caractéristiques acoustiques et des taux d'intelligibilité et naturel suffisants.

Conclusions Générales et Perspectives

Conclusions Générales et perspectives

L'objectif de notre travail tout au long de ce PFE était la réalisation d'un système de synthèse de parole par mots combinés, en vue d'obtenir « SCPNVA». Ce travail est muni d'un programme de simulation d'une application du CP. Nous avons tout d'abord effectué des études générales sur la parole puis sur l'Arabe Standard, pour cela nous avons choisi la synthèse par concaténation des unités pré-stockés, puis un enregistrement, qui a été fait par une locutrice arabophone, une segmentation et annotation du corpus « SCPNVA », ce dernier représente la base de données de notre travail. L'étude du « SCPNVA » passe par plusieurs étapes d'analyse acoustique et de visualisations, qui nous ont permis une extraction des paramètres pertinents et acoustiques du signal vocal. Des tests de perception objective et subjective ont été effectués sur le signal original et le synthétisé.

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

La synthèse par concaténation des formes d'ondes préenregistrées est capable de produire des annonces vocales de haute qualité se rapprochant du naturel. La synthèse de parole présente plusieurs avantages, elle est d'une part plus naturelle pour le grand public, et d'autre part plus rapide et efficace qu'un court message écrit ainsi, que le champ de vision qui reste libre pour effectuer une autre tâche de lecture. La qualité d'un système de synthèse vocale dépend : de son aspect naturel, de l'intelligibilité de la parole générée, des caractéristiques propres à la voix produite qui dépendent : des techniques, des méthodes de synthèse appliquées, et également du soin apporté à la modélisation linguistique et prosodique.

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

- un élargissement du vocabulaire du corpus « SCPNVA », pour une gamme plus grand de nombres ($\geq$ 1 million) ;
- l'amélioration de l'algorithme du programme, de façon qu'il peut calculer le maximum possible des opérations mathématiques, et les synthétisées ;
- l'utilisation d'autres langues (l'Anglais, le Français, par exemple) ;

- l’amélioration de la qualité de la voix synthétisée, avec une technique d’évaluation par ajustement des paramètres prosodiques du signal vocal, afin d’avoir une bonne qualité qui se rapproche du naturel ;
- l’utilisation de la RAP pour entrer l’opération à calculer ;
- La conception matérielle de la calculatrice, avec un guide d’utilisation parlant, et des boutons codés en Braille.

Références bibliographiques

[01] F. SAUSSURE, Cours de Linguistique Générale, 1975.

[02] J. LE GRAND. Parcours Traitement Automatique du Langage Naturel, Université Stendhal, Grenoble /France, 2012.

[03] http://www.claudegaberiel.be/Cineacoustique

[04] O. GODIN, Chapitre 5-Analyse de la parole IMN317, Université de Sherbrooke, Canada, Novembre 2011.

[05] A. ALMEIDA, T. LAVERGNE, B. MAILLOU, Éléments de production et de perception de la parole, Université de Maine, France, Novembre 2011

[06] S. LE MAGEUR, Thèse de doctorat : Evaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française, Université de Rennes, France Juillet 2013.

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

[08] T. DUTOIT, Introduction au Traitement Automatique de la Parole, Faculté Polytechnique de Mons, France ,2000.

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

[10] M. MEDJBER, Amélioration du standard G.729 -8Kb/s par la méthode de modification de l'échelle temporelle (WSOLA), Mémoire de magister, Ecole Nationale Polytechnique, Alger/Algérie, 2007.

[11] C. D'ALESSANDRO, G. RICHARD, Synthèse de la parole à partir du texte, Technique de l'ingénieur, Institut Mines-Télécom, France, 2013.

[12] A. RAMSAY, I. ALSHURHAN, H. AHMED, Generation of a phonetic transcription for modern standard Arabic, Université de Manchester, United Kingdomb, Université de Qatar, Qatar, Février 2013.

[13] S. OUAMOUR, Indexation automatique des documents audio en vue d'une classification par locuteurs Application à l'archivage des émissions TV et Radio, thèse de doctorat, Ecole Nationale Polytechnique, Alger/Algérie, 2009.

[14] https://www.wikipédia.com

[15] S. DJEGHIOUR, Mémoire de Magister : Application des Réseaux de Neurones à la synthèse de la Parole En Arabe Standard, Ecole Nationale Supérieure Des Sciences Humaines, Alger/Algérie, 2011.

[16] S. BALOUL, Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, Thèse de doctorat, Université du Maine, France, mai 2003.

[17] A. OUNNAS, Synthèse de la parole en Arabe Standard, Mémoire de magister, Ecole Nationale Polytechnique, Alger/Algérie, 2011.

[18] 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, 2009.

[19] P.A. BARBOSA, Caractérisation et génération automatique de la structuration rythmique du Français. Thèse de Doctorat en Signal, Image et Parole. Institut National Polytechnique de Grenoble, France, 1994.

[20] S. NEFTI, Segmentation automatique de la parole en phones, Thèse de doctorat, Université Rennes I, France, décembre 2004.

[21] S. LEMMETY, Review of Speech Synthesis Technology, Thèse de master, Université de technologie d'Helsinki/Finlande, 1999.

[22] L'Education pour l'Inclusion : La voie de l'avenir, Rapport de l'Algérie, 48^{éme} session de la conférence internationale de l'éducation, Alger/Algérie, 2010.

[23] http//:www.education.gouv.fr/handiscol/

[24] Notation Mathématique Braille, Commission pour l'Evolution du Braille Français, France, janvier 2007.

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