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

Université d'Alger 2
Faculté des Lettres et des Langues
Département des Sciences du Langage

Mémoire de Magister

En Sciences du Langage et de la Communication Linguistique

Option : Traitement Automatique de la Parole

Présenté par :
Dalila BENSMAIL
Ingénieur d'État en Électronique
Option : Instrumentation USTHB-ALGER

Thème

SYNTHÈSE DE LA PAROLE EN ARABE STANDARD PAR PHRASES ET MOTS CONCATÉNÉS

Devant le Jury :

Mme. N. BENBLIDIA	MCA USD-Blida	Présidente
Mme. M. GUERTI	PROF ENP-Alger	Rapporteur
M. M^{ed}. AISSIOU	MC EPSTA-Alger	Examinateur

Octobre 2014

REMERCIEMENTS

Je remercie ALLAH le tout puissant qui m'a donné le courage et la patience d'effectuer et d'achever ce modeste travail.

Je remercie vivement mon encadreuse Professeur Guerti Mhania pour avoir bien voulu me confier ce travail, pour ses précieux conseils et ses orientations objectives tout au long de l'élaboration de ce Mémoire. Je la remercie aussi pour ses encouragements sa patience et sa grande disponibilité.

Je remercie également Madame Benblidia Nadjia d'avoir accepté de présider le jury ainsi que Monsieur Aissiou Mohamed de bien vouloir examiner ce présent travail.

Je tiens à remercier chaleureusement ma mère, et ma sœur qui m'ont soutenue et encouragé physiquement et moralement.

A tous ceux qui m'ont aidé de près ou de loin, qu'ils trouvent ici ma profonde gratitude.

Dalila BENSMAIL

ملخص

الهدف من عملنا هو تطوير نظام للتركيب التلقائي للكلام الاصطناعي باللغة العربية الفصحى بواسطة الجمل و الكلمات المتسلسلة ، من أجل إدراجها في حافلات نقل المسافرين لمدينة الجزائر. لإعداد نظامنا الذي سمي SYPHRAMO (التركيب بواسطة الجمل و الكلمات)، قمنا باستعمال مجموعة نصوص متكونة من جمل ثابتة و كلمات متغيرة ، ملفوظة من طرف متحدثة باللغة العربية الفصحى. ثم قمنا بتحليل و تجزئة مجموعة النصوص بواسطة برنامج Praat . ثم قمنا بتركيب الجمل و الكلمات بالاستعانة ببرنامج Delphi 7. أجرينا تقييما لنظام SYPHRAMO من أجل الحصول على كلام اصطناعي واضح و طبيعي قدر الإمكان.

كلمات المفاتيح : تركيب الكلام الاصطناعي، اللغة العربية الفصحى، تركيب الكلام الاصطناعي عن طريق تسلسل الجمل و الكلمات، برنامج Praat، Delphi 7، كلام واضح و طبيعي .

Résumé

Le but de notre travail est de réaliser un Système de Synthèse Automatique de la parole par concaténation de Phrases et Mots, en vue de son implémentation dans les bus destinés au transport des voyageurs de la ville d'Alger. Pour réaliser notre application nommée SYPHRAMO (SYnthèse par PHRases et MOts), nous avons élaboré un corpus constitué de phrases fixes et mots variables, en Arabe Standard, prononcés par une locutrice arabophone. Nous avons par la suite analysé et segmenté le corpus à l'aide du logiciel Praat. Puis nous avons procédé à la concaténation des phrases et mots à l'aide du logiciel Delphi 7. Nous avons fait une évaluation de SYPHRAMO afin d'obtenir une parole synthétique intelligible avec le plus naturel possible.

Mots clés : Synthèse de la parole, Langue Arabe Standard, Synthèse par concaténation de phrases et mots, logiciel Praat, Delphi 7, Intelligibilité et naturel de la parole.

Abstract

The purpose of our work is to develop an automatic speech synthesis system that concatenates sentences and words in Standard Arabic, in order to implement it in Algiers City buses. To realize our system named SYPHRAMO (Synthesis by sentences and words), we developed a corpus of fixed sentences and variable words in Standard Arabic, pronounced by an Arabic female speaker. After that, we analyzed and segmented the corpus by using Praat software. Then, we proceeded to the concatenation of the sentences and the words by utilizing Delphi 7 software. We conducted an evaluation of SYPHRAMO in order to get an intelligible synthetic speech with the naturalness as possible

Keywords: Speech synthesis, Standard Arabic Language, Concatenation speech synthesis by phrases and words, Praat software, Delphi 7, intelligibility and naturalness of speech.

Liste des Abréviations

API	Alphabet Phonétique International
ARPA	Advanced Research Project Agency
AS	Arabe Standard
ATR	Advanced Telecommunications Research institute international
DARPA	Defense Advanced Research Projects Agency
F_0	Fréquence fondamentale
F_e	Fréquence d'échantillonnage
GALE	Global Autonomous Language Exploitation
IBM	International Business Machines
LPC	Linear Predictive Coding
LP-PSOLA	Linear Prediction Pitch Synchronous OverLap and Add
MBROLA	Multi-Band Re-synthesis OverLap and Add
OLA	OverLap and Add
RAP	Reconnaissance Automatique de la Parole
SAMPA	Speech Assessment Methods Phonetic Alphabet
SAP	Synthèse Automatique de la Parole
SAPI	Speech Application Programming Interface
SPR	Synthèse de la Parole par Règles
SYPHRAMO	SYnthèse par PHRAses et MOts
TAP	Traitement Automatique de la Parole
TD–PSOLA	Time Domain PSOLA
TTS	Text-To-Speech
V/ NV	Voisé / Non Voisé
VODER	Voice Operation DEmonstratoR
VLSI	Very Large Scale Integration
WSOLA	Waveform Similarity OverLap and Add

Liste des Figures

Liste des Tableaux

Sommaire

Sommaire

Introduction Générale

Introduction Générale

La parole est le langage articulé qui exprime nos pensées. C'est aussi le moyen de communication propre aux êtres humains. Depuis très longtemps l'homme s'est intéressé à l'étude de la parole dans ses différents axes. L'axe qui nous intéresse le plus est le Traitement Automatique de la Parole (TAP) qui est un domaine pluridisciplinaire et qui se situe à l'intersection de l'acoustique, la phonétique, le traitement du signal et la programmation. Le TAP se subdivise en sous-domaines parmi lesquels : la Reconnaissance et la Synthèse Automatique de la Parole. La Reconnaissance permet d'analyser la parole captée puis la transforme en une commande. La Synthèse quant à elle permet d'obtenir une parole artificielle à partir d'un texte ou d'un concept. Notre travail s'insère dans ce dernier domaine qui utilise la méthode de Synthèse par concaténation de phrases et mots. Ce système devra prononcer à haute voix et le plus naturellement possible les messages concernant l'application choisie. Cette dernière nommée SYPHRAMO (SYnthèse par PHRAses et MOts) a pour rôle de signaler les arrêts d'un bus avant d'arriver à une station donnée.

Ce travail se veut également une contribution au développement d'un système d'information qui s'adresse aux passagers des bus et notamment les non-voyants, les personnes âgées, les illettrés et les étrangers à la ville ou au pays.

De part ses enjeux socio-économiques notre étude cherche à développer et améliorer une Synthèse Automatique de la Parole en Arabe Standard suffisamment mûre et adaptée pour être utilisable rapidement au niveau des transports publics (bus) des grandes villes algériennes.

Ce mémoire de magister se compose de quatre chapitres :

- dans le premier, nous donnerons des notions générales sur la parole, sa production et sa perception, les principaux traits pertinents de l'Arabe Standard, ainsi que les caractéristiques acoustiques du signal parole ;
- dans le deuxième, nous décrirons le TAP, les familles qui le composent ainsi que ses deux axes principaux : la reconnaissance et la synthèse automatique de la parole, leurs principales applications ainsi qu'un bref historique de chacune d'elle ;
- le troisième explique en détails la Synthèse Automatique de la Parole en donnant la structure détaillée d'un système de synthèse à partir du texte, les différentes méthodes de synthèse avec leurs avantages et leurs inconvénients, et quelques techniques utilisées en synthèse comme la technique LPC, la technique PSOLA et la technique MBROLA ;
- dans le dernier, nous illustrons, présentons et interprétons les résultats de notre application.

Enfin, nous présentons des conclusions générales et des perspectives concernant la thématique abordée.

Chapitre 1 :
Généralités sur la parole

1.1. Introduction

Étant réservée à l'être humain, la parole est un domaine pluridisciplinaire qui peut être traité suivant différents angles. Ce chapitre traite les généralités sur la parole. Nous commencerons par décrire de façon exhaustive la parole comme outil de communication. Nous présenterons ensuite le mécanisme de production de la parole ainsi que les organes impliqués dans le phénomène de phonation. Vu que nous utiliserons la langue arabe comme principal support d'information dans notre travail. Nous donnerons par la suite des notions et caractéristiques principales de l'Arabe Standard. Nous traiterons en fin les caractéristiques acoustiques qui définissent les sons de la parole.

1.2. Notions générales sur la parole

Depuis très longtemps, la parole s'est imposée comme le mode de communication entre individus par excellence. Mais, avant de pouvoir parler et communiquer librement avec ses semblables, l'être humain passe les premières années de sa vie à apprendre et comprendre comment les adultes parlent. Cette étape est indispensable à l'apprentissage et l'acquisition du langage. Ainsi, reconnaître et produire de la parole, n'est pas aussi facile que nous le pensons. A cet effet, le domaine de la parole aussi complexe qu'il soit a attiré l'attention d'une multitude de chercheurs de différentes disciplines au fil des siècles.

1.2.1. Définition de la parole

D'après la littérature, la parole est le langage articulé symbolique humain destiné à communiquer la pensée, et à distinguer des communications orales diverses, comme les cris, les alertes et les gémissements. « Articuler la parole » consiste à former des signes audibles, les syllabes, formant les mots qui constituent des symboles [1].

Du point de vue linguistique la parole est l'usage concret de la langue par les locuteurs. Celle-ci étant conçue comme un système abstrait [2].

1.2.2. Production de la parole

Les principaux organes impliqués dans la parole sont : le cerveau, la bouche, l'ensemble glottique et les poumons.

1.2.2.1. Le cerveau

Malgré que le cerveau humain ne représente que 2 % du poids du corps, il est le centre opérationnel de l'organisme. Les centaines de milliards de cellules (neurones) qui le composent sont le siège des fonctions intellectuelles, sensitives et motrices [3].

Au niveau physiologique, le cerveau est divisé en 2 hémisphères (droit et gauche) reliés au centre par un faisceau de fibres nerveuses par lesquelles transitent les échanges d'informations entre les 2 hémisphères. L'hémisphère gauche nous intéresse. Ce dernier commande pour les droitiers la partie droite du corps, contrôle l'écriture, le calcul et **la parole** [4].

De façon brève, pour prononcer un mot entendu, les informations perçues par les oreilles sont transmises vers l'aire auditive primaire puis à l'aire de Wernicke (zone de compréhension des mots parlés). Par la suite, il est transféré à l'aire de Broca (zone de production des mots parlés). Ensuite transmis au cortex moteur qui va actionner les muscles et les organes capables d'enclencher la phonation (Figure 1.1) [5].

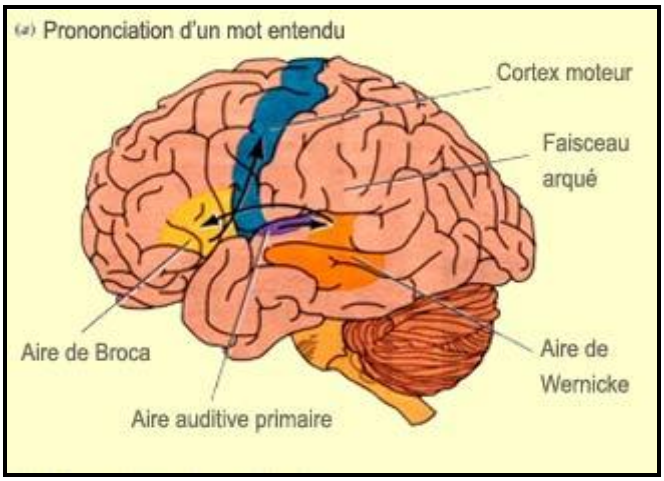


Figure 1.1 : Schéma représentant les aires du cerveau humain, impliquées dans la prononciation d'un mot entendu [5]

1.2.2.2. L'appareil phonatoire humain

Pour comprendre le fonctionnement de l'appareil phonatoire humain et ses possibilités articulatoires dans l'émission de la parole, il est important de connaître les différents organes qui le constituent et le rôle qu'ils jouent dans la phonation.

L’appareil phonatoire de l’être humain est constitué par :

- **les poumons, le diaphragme et la trachée artère :** assurent la respiration qui comporte deux phases l’inspiration et l’expiration. C'est l'air de l'expiration qui est utilisé pour la phonation ;

- **la bouche :** comporte les lèvres, les dents, la langue qui se divise en apex, dos et racine, le palais qui se subdivise en quatre zones : les alvéoles, le palais dur, le voile du palais ou palais mou et l’uvule ou luette.

Autour de la bouche se trouve la mâchoire qui est composée des maxillaires supérieure et inférieure.
La bouche englobe la cavité labiale et la cavité buccale (Figure 1.2).

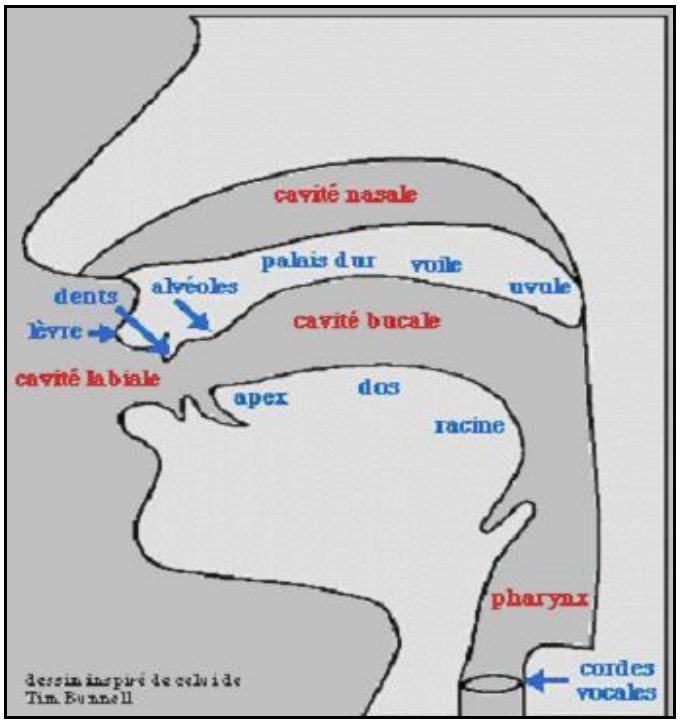


Figure 1.2 : Partie supérieure de l'appareil phonatoire [6]

- **le nez :** l’espace qui nous intéresse est l’espace compris entre les narines et le rhino-pharynx. Cet espace porte le nom de fosses nasales ou cavité nasale ;

- **le pharynx :** désigne la zone au fond du conduit buccal, il forme la cavité pharyngale;

- **le larynx :** se trouve en dessous du pharynx. C'est un conduit ostéo-cartilagineux rigide soutenu par : l'os hyoïde, les cartilages thyroïde, cricoïde, épiglottique et les aryténoïdes [7].

  Le larynx est aussi composé de plusieurs ligaments qui relient ses éléments squelettiques. Celui qui nous intéresse le plus est le ligament vocal (corde vocale). Il est tendu entre le cartilage aryténoïde et l'intérieur du cartilage thyroïde. Ainsi, une corde vocale est la superposition de deux muscles et d'un ligament [8] (Figure 1.3). Pendant la phonation, la vibration des cordes vocales produit des phonèmes voisés (ou sonores) et quand elles ne vibrent pas les phonèmes prononcés sont non-voisés (ou sourds). La glotte désigne l'ouverture entre les cordes vocales.

  Vue sa structure rigide le larynx joue un rôle important dans la respiration (inspiration - expiration), il permet aussi grâce à la fermeture automatique de l'épiglotte de protéger la trachée pendant la déglutition, comme il a pour rôle de protéger les cordes vocales et les muscles responsables de leurs mouvements.

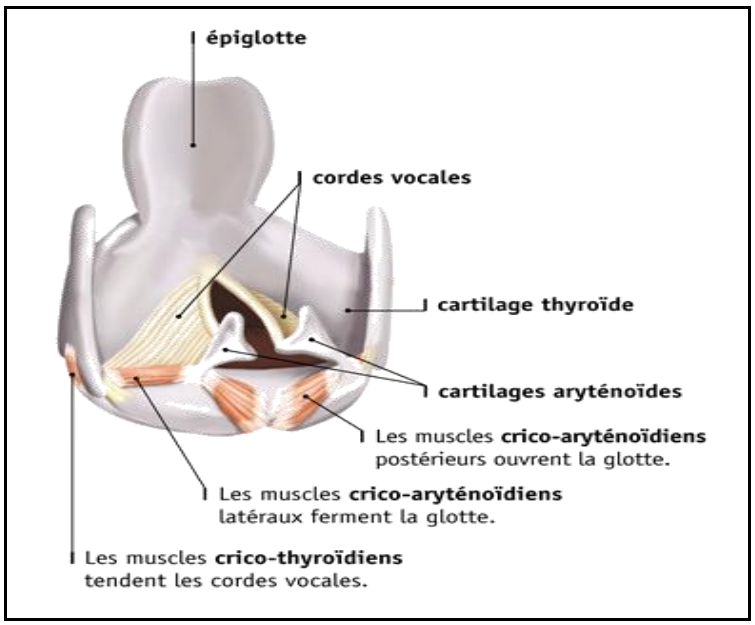


Figure 1.3 : Vue tridimensionnelle de l'emplacement des cordes vocales [9]

1.2.3. Notions sur la perception des sons

Avant de donner des notions sur la perception des sons, il est important d'avoir une idée sur l'anatomie de l'appareil auditif humain.

1.2.3.1. L'appareil auditif humain

L'oreille humaine est le capteur sensoriel qui collecte, analyse et transmet les sons au cerveau. Elle est aussi le siège de l'équilibre.

L'être humain, doté de deux oreilles, peut saisir l'effet stéréophonique et détecter la source du son par estimation de la différence d'intensité perçue par les deux oreilles.

Les sons perçus par l'oreille humaine varient entre 20 Hz et 20 kHz (Figure 1.4). Au dessous de 20 Hz, on trouve les infrasons. Ils sont très nocifs pour l'être humain au-dessous de 10 Hz, mortels au-dessous de 7 Hz. Malgré cela les éléphants les utilisent pour communiquer à une distance de plusieurs kilomètres. Au-delà de 20 kHz on trouve les ultrasons, seuls quelques animaux les interceptent tels que les chats les chiens, les chauves-souris et les dauphins.

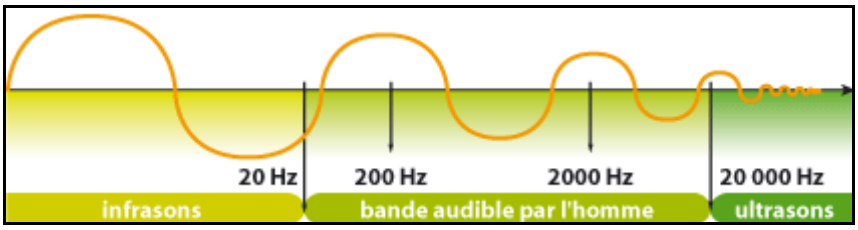


Figure 1.4 : Diagramme représentant les différentes bandes de fréquences sonores [10]

1.2.3.1.1. Anatomie de l'oreille humaine

L'oreille est un organe très sensible. Elle est composée de trois parties différentes mais complémentaires (Figure 1.5) :

- **l'oreille externe** est composée du pavillon et du conduit auditif. Son rôle est de protéger, capter, amplifier et transmettre les vibrations sonores à l'oreille moyenne ;

- **l'oreille moyenne** est composée du tympan, des trois osselets (marteau, enclume, étrier) et de la trompe d'Eustache. Son rôle est de protéger l'oreille interne, transformer les vibrations sonores en vibrations mécaniques (via les trois osselets) et limiter ainsi la perte d'énergie entre le milieu gazeux (oreille externe) et milieu liquide (oreille interne), contribuer à l'adaptation d'impédance entre ces deux milieux, amplifier l'énergie sonore et créer l'équilibre de pression atmosphérique sur les deux faces du tympan pour lui permettre de vibrer correctement via la trompe d'Eustache;

- **l'oreille interne** est constituée du vestibule, des canaux semi-circulaires et de la cochlée ou limaçon. Son rôle est d'établir l'équilibre du corps humain à travers le vestibule et les canaux semi-circulaires, transformer les vibrations mécaniques en signaux électriques au niveau de la cochlée puis les transmettre au cortex auditif au niveau du cerveau par le biais du nerf auditif.

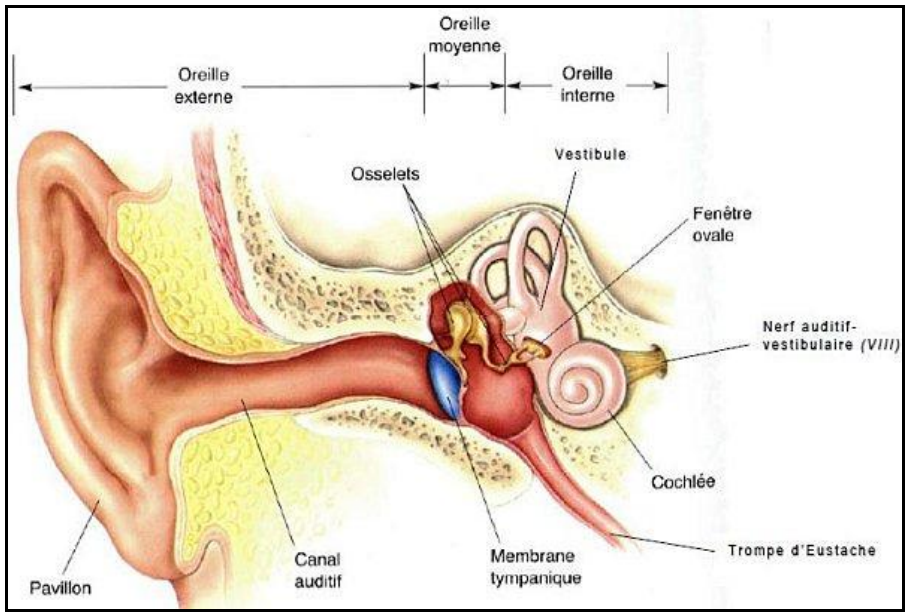


Figure 1.5 : Anatomie de l'oreille humaine [11]

1.2.3.2. Mécanismes de la perception des sons

La cochlée est un organe creux à la forme d'un escargot. Elle comporte l'organe sensible de l'ouïe (l'organe de Corti) qui est composé d'environ 14000 cellules ciliées, baignant dans l'endolymphe, elles font la transduction mécano électrique qui se produit comme suit : les mouvements de l'étrier font vibrer les liquides de la cochlée qui vont à leur tour faire bouger un groupe de cellules ciliés bien déterminées. Le mouvement de ces cellules va se transformer en un signal électrique véhiculé au cerveau à l'aide du nerf auditif.

En effet, tout au long de la cochlée chaque cellule répond préférentiellement à une certaine fréquence, pour permettre au cerveau de différencier la hauteur des sons. Ainsi, les cellules ciliées les plus proches de la base de la cochlée (près de l'oreille moyenne) répondent aux hautes fréquences (sons aigus). Celles situées en son apex (dernier tour de la cochlée) répondent au contraire aux basses fréquences ou sons graves (Figure 1.6).

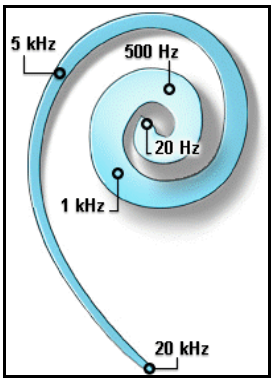


Figure 1.6 : Distribution des fréquences le long de la cochlée [12]

Dans le cadre de notre étude, nous avons choisi pour notre application comme support d'information la langue arabe pour cette raison nous allons donner quelques notions sur cette langue ainsi que ses principales caractéristiques.

1.3. Quelques notions et caractéristiques de l'Arabe Standard (AS)

La langue Arabe appartient à la famille des langues sémitiques, tels que l'Hébreu, le Phénicien, l'Araméen et le Syriaque. Elle s'écrit de la droite vers la gauche.

C'est grâce à l'Islam que cette langue s'est répandue à travers le globe, elle est parlée par environ 240 millions de personnes. Ce qui nous intéresse dans notre étude est l'Arabe Standard, la langue qui est utilisée dans les écoles, les universités et les administrations.

L'AS est constitué de 34 phonèmes, 28 consonnes (Tableau 1.1) et 6 voyelles dont 3 courtes et 3 longues.

Toutes les voyelles sont des sons voisées, orales (l'échappement de l'air se fait par la cavité buccale uniquement - luette relevée -), leur réalisation dépend de la position de la langue (antérieure - postérieure), du degré d'aperture (d'ouverture) et de la position des lèvres (arrondie dans le cas de [u], écartée dans le cas de [i]). Elles sont reparties comme suit :

- les voyelles courtes : [a], [u], [i] (ـَ , ـُ , ـِ) sont représentées au-dessus ou au-dessous de chaque consonne dans un texte voyellé, exemple : كُتِبَ [kutiba] ;
- les voyelles longues : [a:], [u:], [i:] sont représentées respectivement par les consonnes ا , و , ي , exemple : كَاتِبُونَ [ka:tibu:na], اَلبَدِيلُ [al-badi:lu].

Modes d'articulation \ Lieux d'articulation		Bilabiale	Labiodentale	Interdentale	Dentale	Alvéolaire	Palatale	Vélaire	Uvulo-vélaire	Uvulaire	pharyngale	Glottale
Fricatives	Emphatique V			ظ [ẓ]								
	Emphatique NV					ص [ṣ]						
	Non Emphatique V			ذ [d]		ز [z]				غ [ġ]	ع [ɛ]	
	Non Emphatique NV		ف [f]	ث [t]		س [s]	ش [š]			خ [ḫ]	ح [ḥ]	ه [h]
Occlusives	Emphatique V				ض [d]							
	Emphatique NV				ط [t]							
	Non Emphatique V	ب [b]			د [d]							
	Non Emphatique NV				ت [t]			ك [k]	ق [q]			ء [']
Nasales		م [m]			ن [n]							
Liquide					ل [l]							
Affriquée							ج [ǧ]					
Vibrante						ر [r]						
Semi-voyelles		و [w]					ي [y]					

Tableau 1.1 : Modes et lieux d'articulation des phonèmes de l'Arabe Standard

Les consonnes se divisent selon leur mode d'articulation en macro classes qui sont :

- **les occlusives** sont produites par l'obstruction du conduit vocal dans le lieu d'articulation puis sa réouverture brusque. Elles se subdivisent en: occlusives sonores ([b], [d]) et occlusives sourdes ([t], [k]). Comme il est à distinguer les occlusives orales (luette relevée) des occlusives nasales. Les occlusives nasales sont voisées et se réalisent par l'échappement de l'air par les deux cavités buccale et nasale grâce à l'abaissement de la luette, et elles sont: [m], [n].
- **les fricatives** se produisent par le rétrécissement du conduit vocal (sans qu'il ait une fermeture totale comme c'est le cas pour les occlusives) au point d'articulation laissant ainsi passer l'air suivi de bruit tels que sifflement ou chuintement . Il existe des fricatives voisées ([ḍ], [z]) et des fricatives non voisées ([f], [s]).
- **les emphatiques et non emphatiques** le phénomène d'emphase est propre à la langue Arabe. Les consonnes emphatiques prennent naissance aux mêmes lieux d'articulation que leurs correspondantes non emphatique à la différence qu'il ya recule de la langue au fond de la bouche agrandissant ainsi le volume de la cavité buccale. Elles sont au nombre de quatre (Tableau 1.2).

Les Emphatiques	ظ [ḍ]	ص [ṣ]	ط [ṭ]	ض [ḍ]
Les Non Emphatiques	ذ [d]	س [s]	ت [t]	Pas de consonne correspondante

Tableau 1.2 : Consonnes emphatiques de l'AS et leurs correspondantes non emphatiques

- **les liquides** sont voisées et comprennent **le phonème latéral [l]** qui se réalise quand l'apex se met contre les incisives supérieures favorisant le passage de l'air par les deux côtés latéraux de la langue, et **la vibrante [r]** qui se produit lorsque la langue se met contre les dents supérieures, mais produit un battement qui laisse passer de l'air.

- **les affriquées** se constituent de deux phases, la première est l'occlusion suivi de la seconde qui est la friction (par exemple : le phonème **[tʃ]** dans le mot « tea**ch »** en Anglais). Il existe une seule consonne affriquée en Arabe Standard qui est : [ǧ].
- **les semi-voyelles** ou semi-consonnes sont voisées et possèdent quelques caractéristiques des consonnes. L'AS compte deux semi-voyelles : [w] et [y].

1.4. Les caractéristiques acoustiques de la parole

Avant d'aborder les caractéristiques acoustiques de la parole il est nécessaire de définir le son du point de vue acoustique.

1.4.1. Définition du son

Le son est une vibration acoustique qui transmet depuis **une source**, jusqu'à **un récepteur** (l'oreille ou un micro) engendrant une sensation auditive ou un signal sonore électrique [13].

Le son se propage dans tous les milieux élastiques (solides, liquides ou gazeux) sauf le vide. La vitesse de propagation du son varie d'un milieu à un autre. Elle varie aussi dans le même milieu à température et altitude différentes. Par exemple dans l'air la vitesse de propagation du son à -10°C sous une pression atmosphérique égale à 1,341 kg/m³ est de **325,4 m/s**, et à +15°C sous une pression atmosphérique égale à 1,225 kg/m³ elle est à **340,5 m/s** [14].

1.4.2. Les caractéristiques acoustiques du signal de parole

Le signal de parole est caractérisé par :

- il est quasi-stationnaire ;
- le signal de parole est 70% du temps pseudo périodique, bruit ou silence, le reste du temps ;
- la parole est un signal à bande limitée (0-8000 Hz essentiellement) [15] ;
- l'appareil phonatoire humain possède deux sources, une source de bruit qui est **les poumons** et la source du son simple qui est **glotte-cordes vocales**. Comme il possède plusieurs résonateurs qui sont : la cavité buccale, la cavité labiale, la cavité pharyngale et la cavité nasale ;
- le signal de parole possède une fréquence fondamentale et les harmoniques de cette même fréquence ;

- le signal de parole est complexe parce qu'il fait intervenir plusieurs paramètres qui sont : les phénomènes de coarticulation, la prosodie, la variation inter et intra locuteur, les variations dues à l'environnement (bruits), la durée, l'âge, le sexe et l'état de santé physiologique et psychologique de l'interlocuteur.

Ce spectrogramme présente un enregistrement du mot [al-qubba], prononcée par un locuteur arabophone et effectué en mono son à l'aide du logiciel Praat. Le schéma en haut représente l'intensité (dB) en fonction du temps (ms) et celui du bas représente l'évolution de la fréquence (Hz) en fonction du temps (ms). La courbe en bleu représente la fréquence fondamentale, la courbe en jaune désigne l'intensité ou énergie du signal et les points en rouge représentent les formants (Figure 1.7).

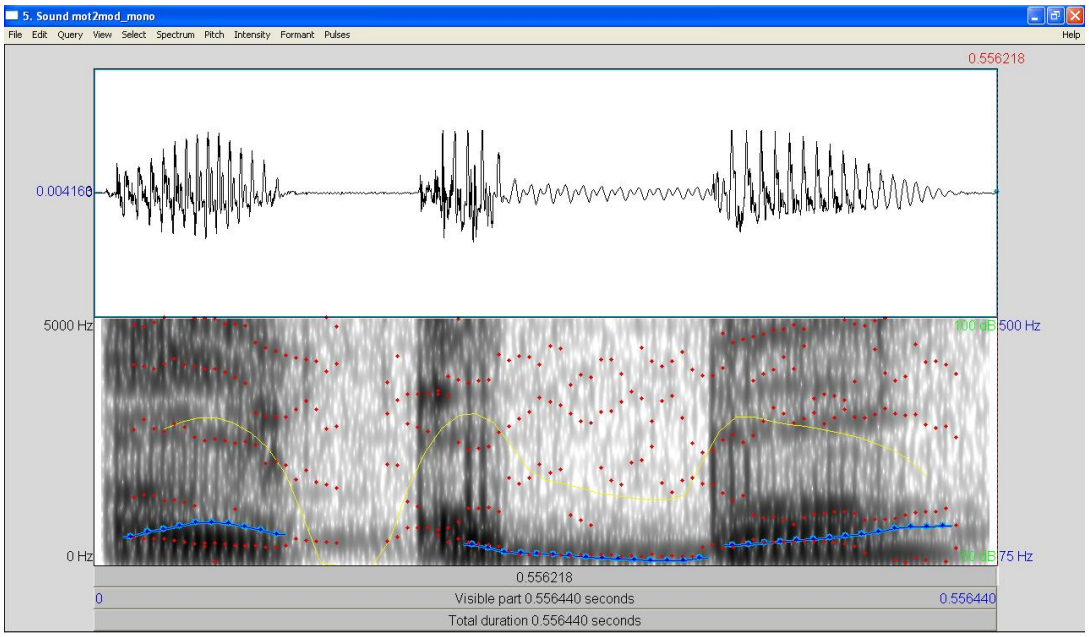


Figure 1.7 : Spectrogramme du mot / القبة / [al-qubba]

1.4.2.1. Définition de la fréquence fondamentale

La fréquence fondamentale ou pitch est la vitesse de vibration des cordes vocales elle est notée $\mathbf{F_0}$ et s'exprime en Hertz. Elle est liée à la hauteur qui spécifie si le son perçu est aigu ou grave.

Plus la fréquence augmente, le son est aigu et plus la fréquence diminue le son est grave. Comme elle détermine si le phonème prononcé est voisé (sonore) ou non voisé (sourd).

- Chez les hommes elle varie entre : 70 - 200 Hz (grave);
- Chez les femmes elle varie entre : 150 - 400 Hz (aigu) ;
- Chez les enfants elle varie entre : 200 - 600 Hz (aigu) [16].

Le son grave ou aigu dépend de la longueur des cordes vocales. Plus les cordes vocales sont longues plus le son est grave. C'est pour cette raison que d'un point de vue général, les voix masculines sont graves et les voix féminines et celles des enfants sont aiguës. La longueur des cordes vocales des hommes varie entre 17 et 24 mm et celle des femmes varie entre 14 et 18 mm [8].

1.4.2.2. Définition de l'Intensité

L'intensité détermine l'amplitude du signal, elle correspond à la distance entre le point de repos et le point le plus élevé de la courbe. L'amplitude diminue avec la distance parcourue et si le son rencontre un obstacle. C'est une grandeur qui détermine si le son est fort ou faible. Elle est mesurée en décibel (dB) et s'exprime pour un signal discret x_n de la manière suivante :

$$E = \frac{1}{T}\sum_{n=1}^{T} x_n^2 \qquad 1.1$$

$$E_{dB} = 10\ log_{10}(\frac{1}{T}\sum_{n=1}^{T} x_n^2) \qquad 2.2$$

La zone audible par les êtres humains possède deux limites (Figure 1.8) :

- 0 dB : c'est le seuil de la perception (silence absolu) ;
- 120 dB : c'est le seuil de la douleur.

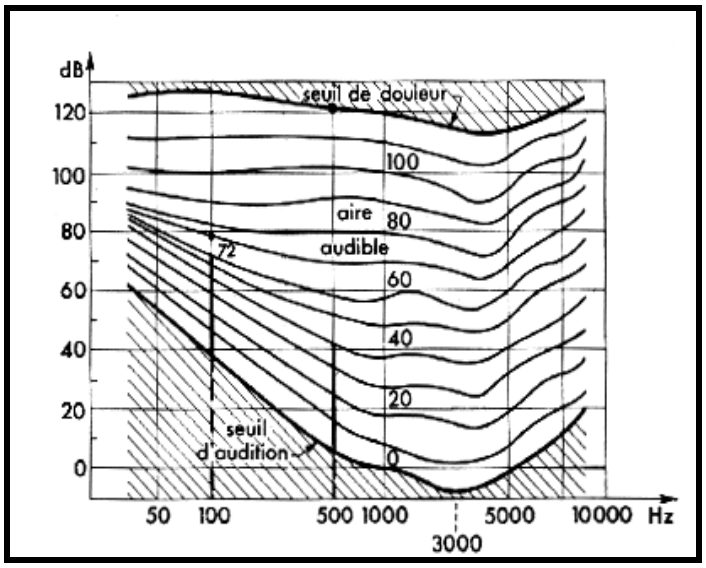


Figure 1.8 : Diagramme représentant les seuils de perception et de douleur chez l'être humain [17]

1.4.2.3. Définition des formants

L'appareil phonatoire humain est composé de plusieurs cavités qui jouent le rôle de résonateurs. Chaque cavité possède une fréquence de résonance propre à elle qui dépend de son volume (plus le volume est grand, plus la fréquence de résonance est petite). Le signal de parole émie par la source (glotte-cordes vocales) se compose de la fréquence fondamentale F_0 et de ses harmoniques 2 F_0, 3 F_0, 4 F_0, etc. Ce signal va être modifié quand il va traverser les cavités, une à une, et qui vont le filtrer de manière que les fréquences du son, proches de la fréquence de résonance de la cavité sont amplifiées et les autres sont atténuées.

Les **formants** représentent les maximas du spectre du signal à la sortie des cavités de résonance. Chaque formant désigne une cavité de l'appareil phonatoire, le formant correspondant à la cavité pharyngale est noté $\mathbf{F_1}$, le formant correspondant à la cavité buccale est noté $\mathbf{F_2}$ et le formant correspondant à la cavité labiale est noté $\mathbf{F_3}$.

Pour illustrer ce qui précède, la voyelle mi-fermée, centrale et non arrondie du Français [ə] est prise pour exemple, les signaux de sa prononciation sont représentés sur les 3 figures suivantes.

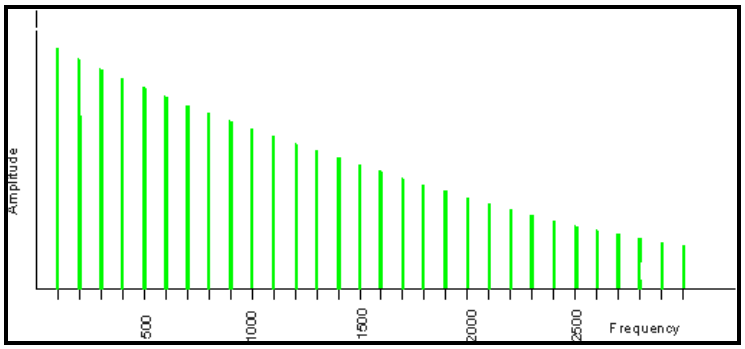


Figure 1.9 : Spectre de l'onde glottale [18]

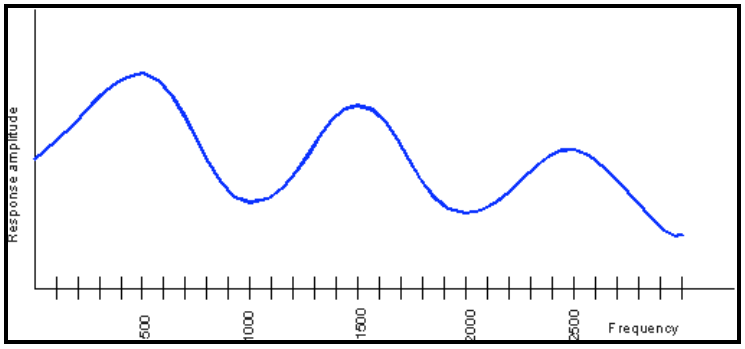


Figure 1.10 : Courbe de réponse du conduit vocal (Filtre) [18]

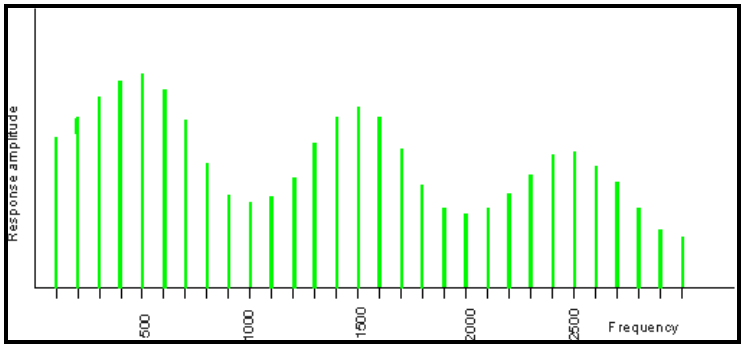


Figure 1.11 : Spectre du son résultant [18]

Dans cet exemple, et comme l'indique la figure 1.11, les valeurs des 3 premiers formants sont: $\mathbf{F_1}$ = 500 Hz, $\mathbf{F_2}$ = 1500 Hz et $\mathbf{F_3}$ = 2500 Hz [18].

Pour une bonne représentation des sons en vue de leurs traitements, en reconnaissance ou en synthèse, le plus grand nombre de formants est requis (six formants).

L’exemple illustratif de la figure 1.12 explique la relation entre la variation des volumes des cavités pendant la prononciation des trois voyelles de l’AS [i], [u], [a] avec la variation des valeurs des formants **F_1**, **F_2**, **F_3** et **F_4** .

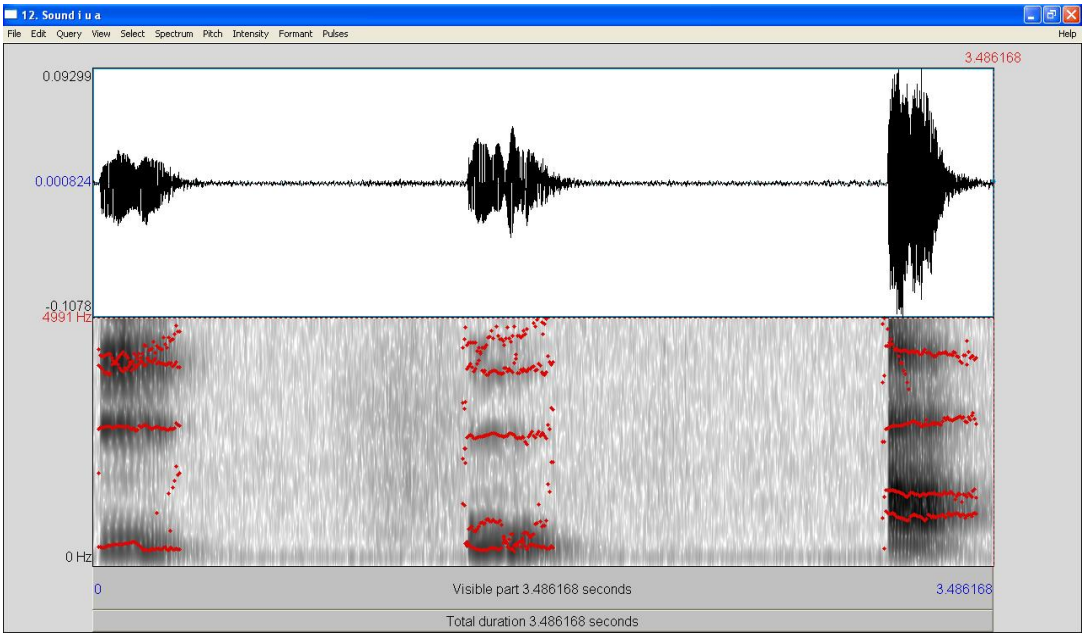


Figure 1.12 : Spectrogramme des voyelles [i], [u] et [a] de l’AS avec leurs formants **F_1**, **F_2**, **F_3** et **F_4** correspondants

1.4.2.4. Définition du timbre

C’est la qualité qui nous permet de différencier entre deux sons qui ont la même intensité et la même hauteur. Par exemple en musique, les notes musicales (Do et Ré) sont perçues différemment si elles sont jouées sur une guitare ou sur un piano. Pour la parole, le timbre nous permet de différencier entre une voix familière et une autre étrangère. En réalité, le timbre c’est l’empreinte vocale de chaque individu. Il est de ce fait lié aux formants qui dépendent de la variation des dimensions des cavités au cours de la phonation. Les dimensions des cavités sont corrélées à la position de la langue, des dents et des lèvres pour un interlocuteur donné.

1.4.2.5. Définition des antiformants

Les antiformants sont une caractéristique des consonnes occlusives nasales [m] et [n]. Leur prononciation implique l’abaissement du velum, favorisant ainsi le passage de l’air dans la cavité nasale. Puis échappement de celui-ci par les deux cavités buccale et nasale.

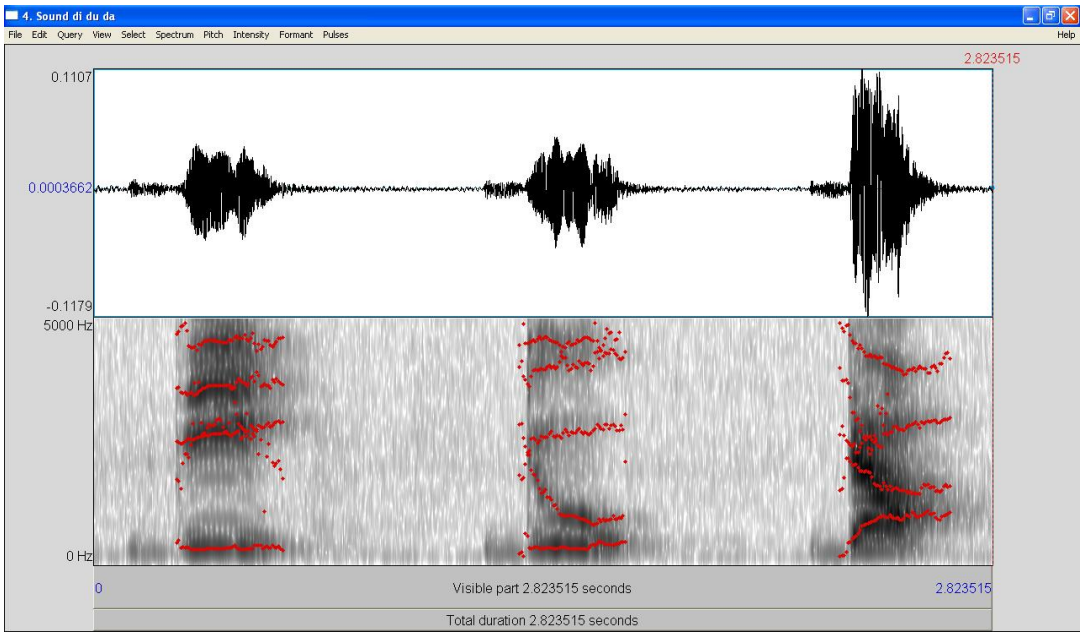
L’articulation des nasales crée des antirésonances dans le conduit vocal. Ces antirésonances ou antiformants sont des régions de fréquence dont les amplitudes du signal source sont atténuées car la cavité nasale absorbe l’énergie de l’onde sonore [19].

1.4.2.6. Définition des transitions formantiques

Les transitions sont les variations fréquentielles (sous forme de pente) des formants de la voyelle au contact de la consonne. Le type de pente (montant, descendant ou plat) va dépendre du lieu d’articulation de l’occlusive ainsi que du type de voyelle produit.

Les transitions formantiques résultent des phénomènes dus à la coarticulation qui peut être décrite comme le chevauchement et l’interaction des différents articulateurs au cours de la production de segments phonétiques successifs. Elles sont très importantes en Traitement Automatique de la Parole (TAP) [20].

P. Delattre est l’un des pionniers à faire des recherches sur les transitions formantiques. Il explique dans ses travaux que la perception des consonnes (précisément les occlusives) n’est réellement possibles que si il y a une transition Consonne-Voyelle ou Voyelle-Consonne : « Les transitions sont la clef de la perception de la consonne » [21] (Fig.1.13).



[di] **[du]** **[da]**

Figure 1.13 : Transitions formantiques du phonème [d] de l’AS

1.4.2.7. Définition de la prosodie

La prosodie est une branche de la linguistique qui se consacre à l'étude des phénomènes suprasegmentaux tels que l'accent, l'intonation, le rythme, etc. Elle est associée à des paramètres physiques comme la fréquence fondamentale, la durée et l'intensité qui représentent le côté objectif de la prosodie, et des paramètres perceptifs tels que la hauteur, le timbre, le rythme qui quant à eux représentent les paramètres subjectifs de la prosodie.

1.4.2.8. Définition de la durée

La durée est une mesure très variable. Elle représente le temps de la prononciation d'un phonème [22]. Sa variabilité est essentiellement due à la complexité et la variabilité du signal de parole qui n'est pas stationnaire et variable des côtés interlocuteur et intra locuteur.

1.5. Conclusion

Dans ce chapitre nous avons abordé les notions théoriques de base sur le domaine de la parole. En effet, nous avons commencé par des généralités sur la parole, les organes impliqués dans la parole et leurs rôles dans le mécanisme de phonation, les caractéristiques essentielles de l'Arabe Standard et les caractéristiques acoustiques du signal de parole. Cette étude théorique sera compléter dans le chapitre suivant par une étude approfondie sur le Traitement Automatique de la Parole.

Chapitre 2 :
Notions sur le Traitement Automatique de la Parole

2.1 Introduction

Ce chapitre explique en détails le domaine de Traitement Automatique de la Parole. Il décrit les grandes familles qui le composent. Deux d'entre elles seront exposées dans ce chapitre : la Reconnaissance Automatique de la Parole et notre sujet d'étude qui est la Synthèse Automatique de la Parole. Pour chacune d'elles nous avons donné une définition, expliqué ses principales applications et donné un bref historique sur son évolution.

2.2. Le Traitement Automatique de la Parole (TAP)

Le TAP est une discipline scientifique située au croisement du traitement du signal numérique et du traitement du langage. Dés les années 60, son développement considérable lié au développement des domaines électronique et informatique, lui a permis de s'imposer comme domaine leader dans différents secteurs.

Il existe quatre grandes familles de modules vocaux :

- **les analyseurs de parole :** ont pour objectif de mettre en évidence les caractéristiques du signal vocal du point de vue de sa production, ou parfois du point de vue de sa perception, mais jamais du point de vue de sa compréhension, ce rôle étant réservé aux reconnaisseurs. Les analyseurs peuvent être utilisés en tant que composant de base des systèmes de codage, de reconnaissance ou de synthèse, ou bien en tant qu'analyseurs 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.
- **les reconnaisseurs :** ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. Il existe deux types de reconnaissance selon l'information à extraire du signal vocal :
  - la reconnaissance du locuteur : qui a pour objectif de reconnaître la personne qui parle ;
  - la reconnaissance de la parole : qui a pour objectif de reconnaître ce qui est dit.
- **les synthétiseurs :** ont la fonction inverse de celle des analyseurs et des reconnaisseurs de parole, ils produisent de la parole synthétique. Ils sont classés en 2 types, à partir :

- d’une représentation numérique (inverse des analyseurs) ont pour mission de produire de la parole à partir des caractéristiques numériques d’un signal vocal telles qu’obtenues par analyse ;
- d’une représentation symbolique (inverse des reconnaisseurs) sont en principe capables de prononcer n’importe quelle phrase sans qu’il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, les synthétiseurs sont également classés en fonction de leur mode opératoire :
  - les synthétiseurs à partir du texte reçoivent en entrée un texte orthographique et doivent en donner la lecture ;
  - les synthétiseurs à partir de concepts sont appelés à être insérés dans des systèmes de dialogue Homme-Machine. Ils reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue (exemple : traduction parole-parole).

- **les codeurs :** leur rôle est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole [23].

2.3. La Reconnaissance Automatique de la Parole (RAP)

La reconnaissance de la parole est l’habilité d’une machine ou d’un programme à identifier les mots et les phrases d’une langue parlée et le convertir en un format décodable par une machine. Les systèmes rudimentaires de reconnaissance vocale possédaient un vocabulaire limité de mots et phrases qui pouvaient être identifié s’ils étaient très clairement prononcés. Le terme reconnaissance de la parole et reconnaissance du locuteur sont des termes qui ont des sens différents. Alors que le premier est utilisé pour l’identification des mots de la parole dans une langue donnée, le deuxième est une technologie biométrique qui vise à identifier la voix particulière d’un individu.

2.3.1. Les sources de variabilité de la RAP

La Reconnaissance Automatique de la Parole est confrontée aux sources de variabilité suivantes :

- les facteurs intra-locuteurs : qui comportent les phénomènes de coarticulation, la variation dans la prononciation.

- les facteurs interlocuteurs : elles englobent la physiologie, l'âge, le sexe, la psychologie, la familiarité avec l'application, etc.

- l'environnement : qui sont le bruit, le microphone, le canal de transmission, la présence d'autres locuteurs, etc. [15].

2.3.2. Applications de la RAP

Les principales applications de la RAP sont :

- **télématique et services vocaux :** composeur vocal, serveurs vocaux interactifs, consultation de messagerie vocale ;
- **bornes interactives :** renseignements sur les horaires (train, avion, bâteau) et prise de réservations ;
- **bureautique :** services télématiques vocaux et commandes vocales d'éditeur ;
- **contrôle de qualité et saisie de données :** l'interface vocale libère la vue et les mouvements. L'utilisateur peut donc se déplacer librement pour manipuler des objets ou entrer des données ;
- **aide à la conception graphique :** système d'interaction multimodale, incluant parole, geste et vision ;
- **avionique :** permet aux pilotes une meilleure attention visuelle ;
- **identification/vérification du locuteur :** pour assurer une meilleure sécurité pour l'accès en direct à des bases de données confidentielles ;
- **aide à la navigation en voiture :** permet le positionnement du véhicule, la planification de l'itinéraire et notamment le guidage du conducteur par des messages vocaux ;
- **aide à la formation :** apprentissage des langues, de la lecture, formation des contrôleurs aériens (meilleure connaissance de la phraséologie spécialisée du domaine) ;
- **aide au handicap :** aide à la rééducation de la voix, contrôle d'objets de l'environnement pour les tétraplégiques, consultation de documents pour les non voyants (tâches d'édition et de consultation) ;
- **dictée automatique ou entrée vocale de textes :** contrôle d'un microscope, interrogation vocale d'une base de données, constitution automatique de rapports médicaux par dictée vocale ;

- **traduction automatique :** de conversations téléphoniques avec un interlocuteur de langue étrangère [24].

2.3.3. Historique de la reconnaissance vocale

Les travaux sur la reconnaissance de la parole datent du début du XX^{ème} siècle. Le premier système pouvant être considéré comme faisant de la reconnaissance de la parole date de 1952. Ce système électronique développé par Davis, Biddulph et Balashek aux laboratoires Bell (USA) était essentiellement composé de relais et ses performances se limitaient à reconnaître des chiffres isolés. La recherche s'est ensuite considérablement accrue durant les années 1970 avec les travaux de Jelinek chez IBM (1972-1993). La société Threshold Technologies fut la première à commercialiser en 1972 un système de reconnaissance d'une capacité de 32 mots (**VIP100)**. Aujourd'hui, la reconnaissance de la parole est un domaine à forte croissance grâce à la déferlante des systèmes embarqués [25].

Le tableau 2.1 présente quelques dates qui ont été des faits marquants pour l'évolution de la RAP :

Dates	Système de reconnaissance
1952	Reconnaissance des 10 chiffres, pour un monolocuteur, par un dispositif électronique câblé
1960	Utilisation des méthodes numériques
1965	Reconnaissance de phonèmes en parole continue
1968	Reconnaissance de mots isolés par des systèmes implantés sur de gros ordinateurs (jusqu'à 500 mots)
1969	Utilisation des informations linguistiques
1971	Lancement du projet ARPA aux USA pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
1972	Premier appareil commercialisé de reconnaissance de mots
1976	Fin du projet ARPA, les systèmes opérationnels sont HARPY, HEARSAY I et II et HWIM
1978	Commercialisation d'un système de reconnaissance à base de microprocesseurs sur une carte de circuits imprimés
1981	Utilisation de circuits intégrés VLSI spécifiques du traitement de la parole
1981	Système de reconnaissance de mots sur un circuit VLSI
1983	Première mondiale de commande vocale à bord d'un avion de chasse en France
1985	Commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
1986	Lancement du projet japonais ATR de traduction automatique en temps réel
1988	Apparition des premières machines à dicter par mots isolés
1989	Recrudescence des modèles connexionnistes neuromimétiques
1990	Premières véritables applications de dialogue oral Homme-Machine
1994	IBM lance son premier système de reconnaissance vocale sur PC
1997	Lancement de la dictée vocale en continu par IBM (Via Voice)
1998	Phillips commercialise un logiciel de reconnaissance vocale nommé Freespeech et Lernout & Hauspie lance ses premiers produits VoiceXpress, VoiceXpress+ et VoiceXpress Pro
2000	Microsoft développe SAPI 5 (Speech Application Programming Interface)
2006	DARPA (Defense Advanced Research Projects Agency) prépare un projet de traduction nommé GALE (Global Autonomous Language Exploitation)
2011	Google annonce le lancement de la recherche vocale sur son moteur de recherche (disponible sur Google Chrome et en Anglais seulement) et Apple lance Siri, l'application d'assistant personnel pour système d'exploitation de l'iPhone 4S (disponible en anglais, allemand et français)
2012	CES (Consumer Electronics Show) présente en Janvier, la smart TV de Samsung avec reconnaissance vocale et faciale

Tableau 2.1 : Faits marquants de l'évolution de la RAP [26]

2.4. La Synthèse Automatique de la Parole (SAP)

2.4.1. Définition

La synthèse vocale est une technique informatique de synthèse sonore qui permet de créer de la parole artificielle à partir de n'importe quel texte orthographique ou concept (Figure 2.1).

D'où la définition suivante :

« La synthèse de la parole permet de produire des sons de la parole à partir d'une représentation phonétique du message » [22].

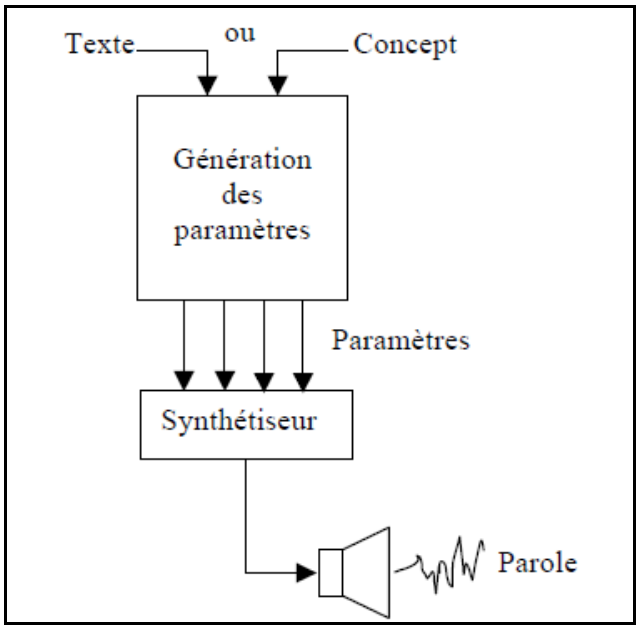


Figure 2.14 : Système de synthèse de la parole [22]

La parole synthétique résultante est supportée par deux types d'informations, au niveau :

- segmental, avec les unités phonémiques ;
- suprasegmental, avec la génération de la prosodie, nécessaire à l'intelligibilité et le naturel de la parole.

2.4.2. Historique de la SAP

- Les résultats fructueux des recherches sur la production d'une parole synthétique n'ont vu le jour qu'au 18ème siècle. En effet, en 1780 le lauréat de la compétition scientifique de l'Académie des Sciences de Saint-Pétersbourg Christian Gottlieb réalise 5 résonateurs acoustiques qui modélisent l'appareil phonatoire humain (Figure 2.2), pour démontrer les différences physiologiques entre les voyelles [a], [e], [i], [o], [u] (sujet de la compétition). Le son de chaque voyelle est produit par la vibration de la lame (représentant les cordes vocales) actionnée par l'air soufflé à l'extrémité inférieure de chaque résonateur [27], [28].

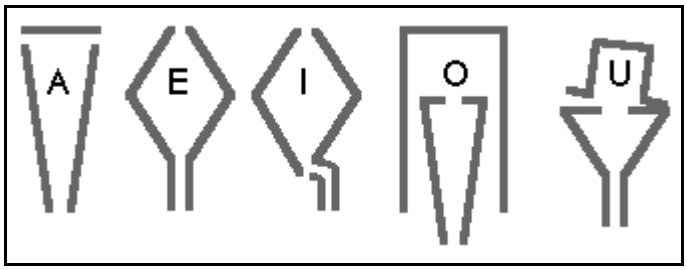


Figure 2.2 : Résonateurs de Kratztenstein pour la synthèse des voyelles [28]

- Quelques années plutôt, en 1769, le Baron Wolfgang Von Kemplen avait commencé ses travaux sur une machine qui va révolutionner le domaine de la parole synthétique. En 1791, parait son livre qui décrit avec précision sa machine parlante avec des schémas détaillés. Cette machine produisait, non seulement certaines voyelles mais surtout des mots entiers et des courtes phrases. Extérieurement, la machine était composée d'un caisson, d'un entonnoir en caoutchouc qui faisait office de bouche et d'un second, plus petit, divisé en deux, qui remplissait les fonctions d'un nez. Le mécanisme interne était un soufflet qui simulait les poumons, et un contrepoids pour simuler l'inhalation [29] (Figure 2.3).

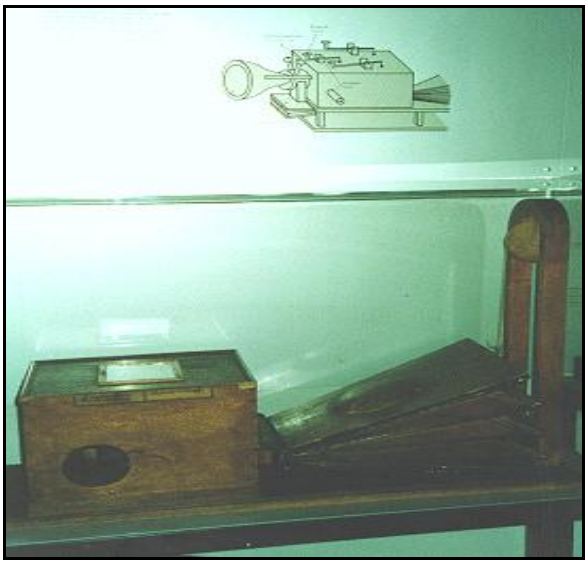


Figure 2.3 : Machine parlante de Von Kempelen [29]

- Une reconstitution de la machine de Von Kemplen démontré par Sir Charles Wheatstone en 1835 à Dublin, était différente de la version décrite par Von Kemplen dans son livre. En effet, cette machine a une cavité orale flexible et un contrôle actif de la voix mais ne possède pas un mécanisme de contrôle du pitch tel qu'a été inclus dans la version finale de Von Kemplen [29] (Figure 2.4).

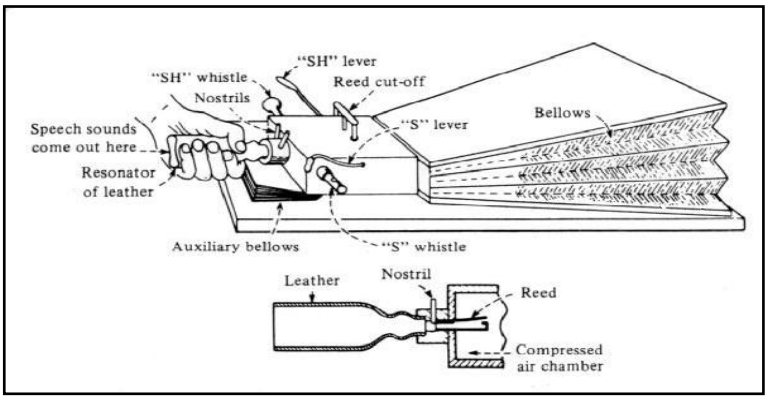


Figure 2.4 : Reconstitution par Wheatstone de la machine parlante de Von Kempelen [30]

- Au 19ème siècle, d'autres machines du même type (mécanique) ont été construites, mais elles n'ont pas vraiment apporté d'innovation fondamentale dans le domaine de la synthèse de la parole. Toutefois, la machine conçue par Joseph Faber en 1835 présentait quelques progrès du point de vue du mécanisme de production de la parole qui incluait un model de la langue et la cavité pharyngale dont les formes pouvaient être contrôlées [29].

Le dispositif complexe nommé « Euphonia » est contrôlé par dix-sept leviers, un soufflet, et une ligne télégraphique, cette machine est ornée de la réplique mobile d'un visage, qui était en mesure de reproduire fidèlement les sons de la parole. Elle possède aussi une pédale, un clavier, des leviers séparés qui contrôlent les mouvements de la langue, des lèvres, des joues, et des cordes vocales. En Effet, l'operateur de Euphonia pouvait lui faire parler n'importe qu'elle langue Européenne [31] (Figure 2.5).

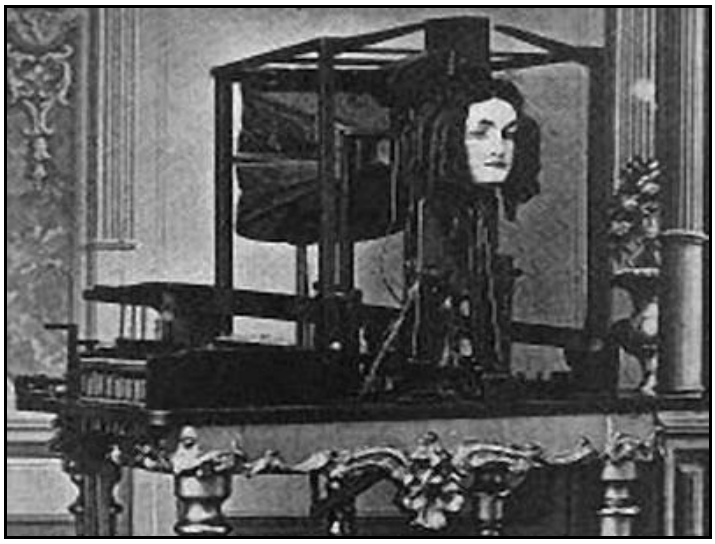


Figure 2.5 : Visage d'Euphonia, la machine parlante [31]

- « En 1937, R.R. Riesz a démontré son model mécanique parlant qui comme ses prédécesseurs, rappelait beaucoup plus l'instrument musical. L'appareil avait la forme de l'appareil phonatoire humain construit essentiellement avec du caoutchouc et du métal avec des touches comme ceux trouvé sur la trompette. La machine parlante mécanique produisait relativement une bonne qualité de parole synthétique avec un operateur entraîné… avec les 10 touches de contrôle (ou valves) utilisés simultanément avec les deux mains, le dispositif pouvait produire de la parole relativement articulée. Riesz a permis à travers l'utilisation des 10 touches, de contrôler le mouvement de presque toutes les parties mobiles de l'appareil phonatoire humain » (Figure 2.6) [32].

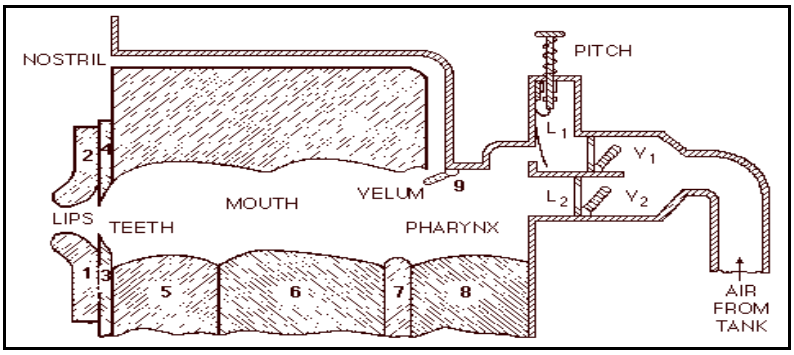


Figure 2.6 : Modèle mécanique de production de la parole conçue par Riesz [32]

- Au début du 20ème siècle, le progrès en génie électrique a rendu possible la synthèse de la parole à partir de moyens électriques. Le premier appareil de ce genre était le VODER (Voice Operating DEmonstratoR), développé par Homer Dudley, Riesz and Watkins aux laboratoires Bell et présenté à la Foire Internationale de New York en 1939. Malgré que le VODER ait attiré l'attention d'un très large public, le temps d'apprentissage nécessaire pour sa bonne utilisation était très long.

  Le VODER possède deux sources sonores une qui génère le souffle (générateur de bruit blanc) et l'autre génère un signal qui simule la vibration des cordes vocales (oscillateur à relaxation), cette dernière et lié à une pédale pour le contrôle du pitch. A la sortie des sources sonores, il existe un sélecteur de voisement ou non-voisement lié au poigné. Le VODER possède aussi 10 touches (5 touches pour chaque main) plus trois autres supplémentaires pour marquer le stop pour les plosives [t/d], [p/b] et [k/g]. Les signaux sont décomposés puis restitués grâce au bloc des 10 résonateurs filtres (sachant que la plage de fréquence choisie par Homer Dudley pour le VODER varie entre 300-3000 Hz alors les 10 filtres passe bandes auront une largeur de bande de 300 Hz chacun). Le signal synthétisé passe par un bloc amplificateur, relié un sélecteur pour augmenter ou diminuer le volume à la sortie du haut parleur. Le VODER permet de générer 23 sons de la parole avec des difficultés pour la prononciation de la consonne [l] [29], [33] (Figure 2.7).

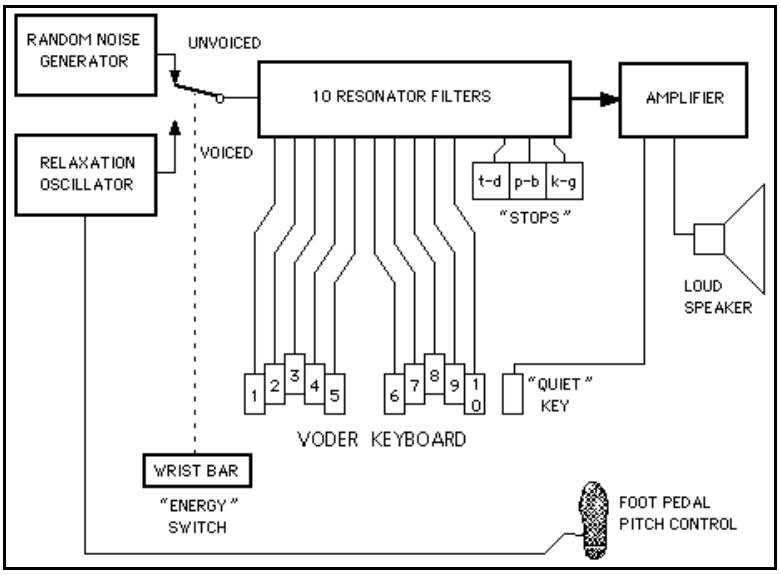


Figure 2.7 : Diagramme schématique du synthétiseur VODER [33]

- En 1950, Frank Cooper développe le Pattern Playback aux laboratoires Haskins (Figure 2.8). Le système de synthèse de la parole était différent de ses prédécesseurs. Il servait essentiellement à l'investigation dans la perception de la parole. Il effectuait le travail inverse du spectrographe.

  Une lampe produisait un rayon de lumière dirigée de façon radiale contre un disque tournant avec 50 pistes concentriques et dont la transparence varie de façon systématique pour produire 50 partiels avec une fréquence fondamentale de 120 Hz. La lumière est de plus en plus projetée contre le spectrogramme dont la réflexion ou la transparence (dans le mode d'opération alternative) correspond au niveau de la pression sonore de chaque partiels, et dirigé vers une cellule photovoltaïque par laquelle la variation en lumière est convertit finalement en variation de pression sonore. Le signal sonore résultant ressemble en quelque sorte à la parole originale [29].

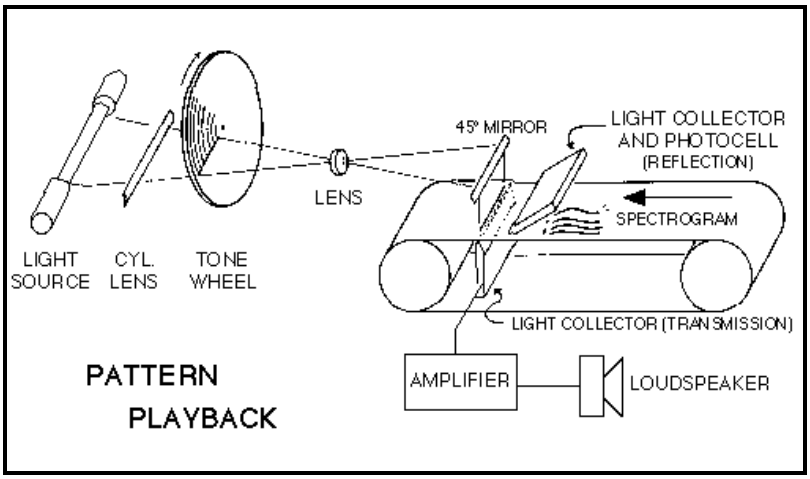


Figure 2.8: Schéma du Pattern Play-back [29]

- Dunn-Stevens (1953): premiers analogues électriques du conduit vocal. Ces machines simulaient le conduit vocal comme étant composé d'une succession de tubes élémentaires, le diamètre variable de chaque section représente la forme intérieure du conduit vocal.

- Premier synthétiseur à formants (1955)

2.4.3. Applications de la SAP

- **aides pour personnes handicapées :** lecture d'écran pour non-voyants, aides à la communication vocale pour personnes laryngectomisées, journaux vocaux ;
- **outils d'enseignement assisté par ordinateur :** dictée automatique, apprentissage des langues, rééducation, alphabétisation ;
- **applications industrielles :** serveurs d'alerte, surveillance de sites, supervision de réseaux, télémaintenance, aide dans les postes de pilotage, vérification vocale ;
- **applications grand - public non téléphoniques :** domotique (alarmes, appareils domestiques parlants), micro-informatique (jeux parlants, bureautique) ;
- **télématique vocale :** serveurs vocaux d'informations, serveurs de lecture vocale de FAX ou de messages électroniques, automatisation de services de prise de commande (vente par correspondance) ou de renseignements (annuaires, standards d'entreprises) [25] ;
- **clonage de voix d'utilisateurs et reconstruction de voix de malades :** à partir des pathologies de la parole (autisme, aphasie due à un accident vasculaire cérébrale) [34].

2.5. Conclusion

Nous pouvons conclure de ce qui précède que le TAP est un domaine multidisciplinaire qui peut être étudié suivant deux grands axes la RAP et la SAP. Les applications de la RAP et de la SAP s'étendent sur une multitude de secteurs. Le développement technologique au fil du temps a permis de concevoir des équipements de plus en plus performants dans les deux domaines.

Chapitre 3 :
Synthèse Automatique de la Parole

3.1. Introduction

Nous allons présenter dans ce chapitre les composantes essentielles d'un système de synthèse de la parole à partir du texte suivi d'une description du module de traitement du langage naturel et du module de traitement du signal numérique qui dépend principalement de la méthode de synthèse utilisée. L'une des trois méthodes de synthèse citée et expliquée dans ce chapitre sera utilisée dans notre travail, c'est la méthode de synthèse par concaténation d'unités acoustique. Par la suite, nous expliquerons quelques techniques d'analyse utilisées en SAP qui sont les techniques LPC, PSOLA et MBROLA.

3.2. Structure d'un système SAP

Le système de synthèse de la parole à partir du texte (en Anglais Text-To-Speech ou TTS) doit être conçu pour lire n'importe quel texte à haute voix, toute en ayant le naturel et l'intelligibilité de l'être humain. Cela est désormais possible, grâce au développement rapide dans les filières techniques, tel que l'informatique avec l'explosion d'internet et de la nanoélectronique qui permet désormais de concevoir des composants de plus en plus petits exécutants des tâches de plus en plus complexes. Ainsi le mythe dialogue Homme-Machine est devenue une réalité exploitable, d'où la conception de robots de plus en plus performants.

Un système TTS se compose essentiellement de 2 blocs importants comme il est décrit dans le schéma suivant [23] :

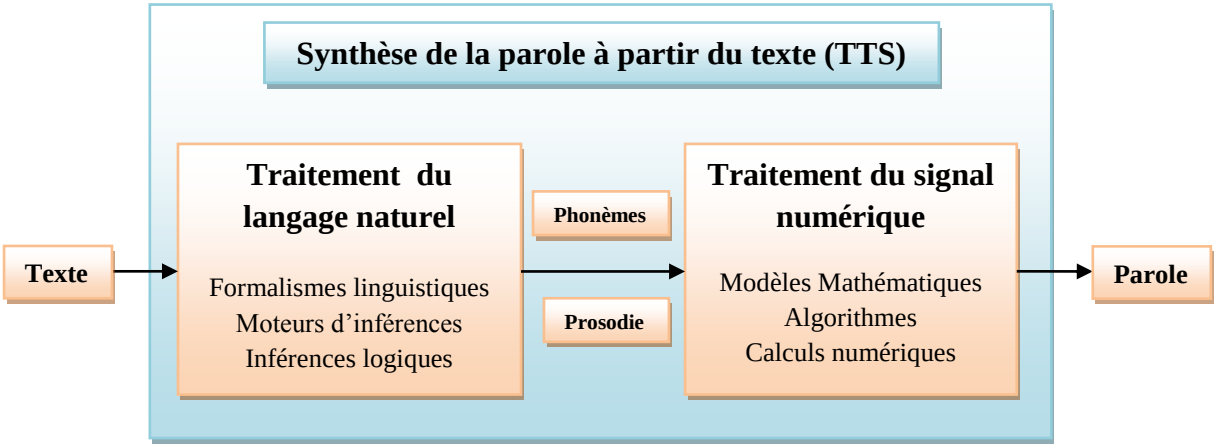
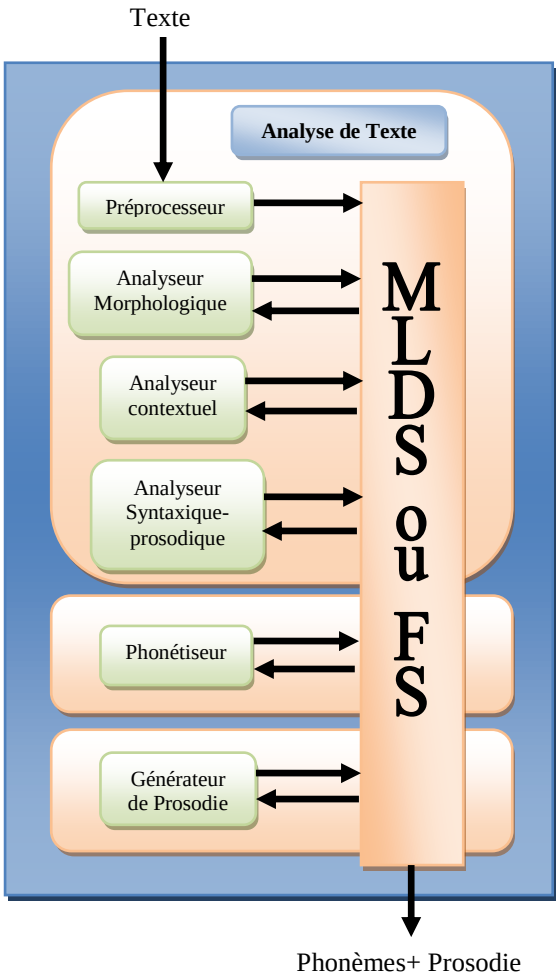


Figure 3.1 : Diagramme fonctionnel d'un système de synthèse de la parole à partir du texte [23]

3.3. Brève description du module de traitement du langage naturel

Le module de traitement du langage naturel (Figure 3.2) se compose :

- **d'un Préprocesseur :** joue le rôle d'interface entre le texte et la structure de données internes. Il permet d'identifier les caractères difficiles à prononcer tels que: nombres, abréviations, etc. et les transcrit en lettres ;
- **d'un Analyseur morphologique** : propose toutes les natures possibles pour chaque mot pris individuellement, en fonction de sa graphie ;
- **d'un Analyseur contextuel :** considère les mots dans leur contexte ;
- **d'un Analyseur syntaxique-prosodique :** examine l'espace restant et découpe le texte en groupes de mots pour y associer une prosodie ;
- **d'un phonétiseur :** génère la transcription phonétique des mots du texte ;
- **d'un générateur de prosodie :** attribue à chaque groupe de mots une courbe d'évolution du pitch et de la durée des phonèmes [23].



MLDS (Multi-Layers Data Structure): structure de données multi niveaux ;

FS (Feature Structure): structure d'attribut.

Figure 3.2 : Module de traitement du langage naturel d'un système de conversion texte-parole [23]

3.4. Les méthodes de synthèse da la parole

Le module de traitement du signal de parole dépendra de la méthode de synthèse utilisée. Il existe trois méthodes de synthèse de la parole.

3.4.1. Synthèse articulatoire

La synthèse articulatoire est basée sur une modélisation du conduit vocal. Les différents modèles de l'appareil phonatoire qui ont vue le jour, se basent tous sur la coupe sagittale de la tête.

Cette représentation généralement en 2D fournit des données de bonne qualité ainsi qu'une quantité d'information suffisante pour une bonne synthèse de la parole et cela est essentiellement due à la position de chaque articulateurs en temps réelle et à la prise en compte des phénomènes de coarticulations.

Pour construire un synthétiseur articulatoire, il faut élaborer les modèles suivants :

- un modèle articulatoire capable de produire des coupes sagittales à partir des paramètres articulatoires ;
- un modèle de passage de la coupe sagittale à une fonction d'aire représentant le conduit vocal par un ensemble de tubes acoustiques équivalents ;
- un modèle acoustique permettant de passer de la fonction d'aire au signal de parole [35].

3.4.1.1. Les classes de modèles articulatoires

Les principales classes de modèles articulatoires sont :

- **les modèles à fonction d'aire** sont très utilisées. C'est la plus simple des modélisations, il s'agit de calculer la fonction d'aire du conduit vocal qui est modélisé par quatre tubes de rayons variables où chaque tube représente une partie de l'appareil phonatoire. Chaque son correspond à une configuration de tubes aux rayons et longueurs différentes. Ce type de modèle ne cherche pas à représenter fidèlement le conduit vocal au sens anatomique, mais à simuler le comportement du passage de l'air dans le conduit pour en synthétiser le son ;

- **les modèles géométriques** représentent chacun des articulateurs du conduit vocal par une forme géométrique simple. Chaque articulateur est piloté par un nombre variable de paramètres qui agissent comme des commandes sur ces formes (rotations, translations, déformations…). Les inconvénients de ces modèles résident en l'emploi d'images réelles (rayons X), l'expertise humaine et l'intuition, la simplicité des formes géométriques utilisées ne rend pas compte de la complexité anatomique des formes du conduit vocal ;

- **les modèles statistiques** sont une alternative des modèles géométriques qui consistent en l'élaboration de modèles à l'aide d'une analyse factorielle basée sur des données articulatoires réelles. Ces modèles sont construits exclusivement à partir de données réelles, cet avantage est aussi leur inconvénient car la moindre fausse information va engendrer une faiblesse du modèle.

  L'un des modèles statistiques les plus connus est le modèle de S. Maeda qui décrit les formes de conduits vocaux à partir de contours dessinés manuellement sur des

images rayons X. La figure 3.3 décrit les sept paramètres du modèle de S. Maeda, où P1 représente la mâchoire, P5 est l'ouverture des lèvres, P6 est leurs protrusion, P2 est la position du corps de la langue, P3 est la forme de la langue, P4 est un terme qui contrôle la pointe de la langue, et enfin P7 qui détermine la hauteur du larynx [36] ;

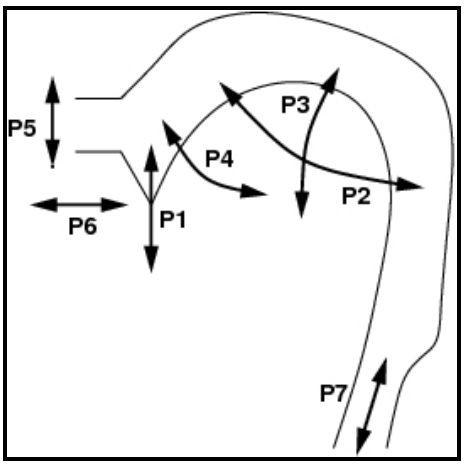


Figure 3.15 : Paramètres du modèle de S. Maeda [36]

- **les modèles biomécaniques** sont la forme la plus complexe, mais aussi la plus complète, de modèles de conduits vocaux. Elles reposent sur l'intégration d'un maximum de propriétés physiologiques des articulateurs et l'étude de leurs interactions avec des éléments externes (os, muscles). Le modèle biomécanique le plus répandu est le modèle masse-ressort. L'inconvénient majeur de ces modèles est que beaucoup de paramètres doivent être déterminés. Solution de modélisation la plus réaliste, elles demeurent encore aujourd'hui très coûteuse en ressources humaines et matérielles [36].

3.4.1.2. Les méthodes d'acquisition

En vue d'établir des modèles articulatoires réalistes et cohérents, de nombreuses méthodes ont été testées sur des locuteurs pour acquérir des données articulatoires du conduit vocal.

Parmi les méthodes d'acquisition nous pouvons citer :

- les méthodes d'acquisition physiologique telle que l'électromyographie qui mesure des courants électriques à partir d'électrodes collées sur le visage ;
- les méthodes d'acquisition aéroacoustiques permettent les mesures de flux d'air ;

- les méthodes qui permettent d'obtenir des informations anatomiques de position et de mouvement pour les confronter à des modèles articulatoires, telle que les techniques d'imagerie (cinéradiographies rayons X, microfaisceaux de rayons X, données électromagnétique, Echographie, IRM statique et dynamique) et les techniques qui permettent de récupérer la position de points (articulographie, palatographie) [36].

3.4.2. Synthèse de la Parole par Règles (SPR)

La SPR consiste en l'élaboration d'un ensemble de règles qui modélisent l'évolution des trajectoires décrites par les paramètres du signal acoustique cette méthode tient compte de l'ensemble des informations contenu dans les transitions entre phonèmes (transitions formantiques). Elle suppose donc une bonne maitrise de l'ensemble des paramètres caractérisant les différents phonèmes dans les différents contextes, il faut donc avoir des connaissances linguistiques profondes (Figure 3.4) [23].

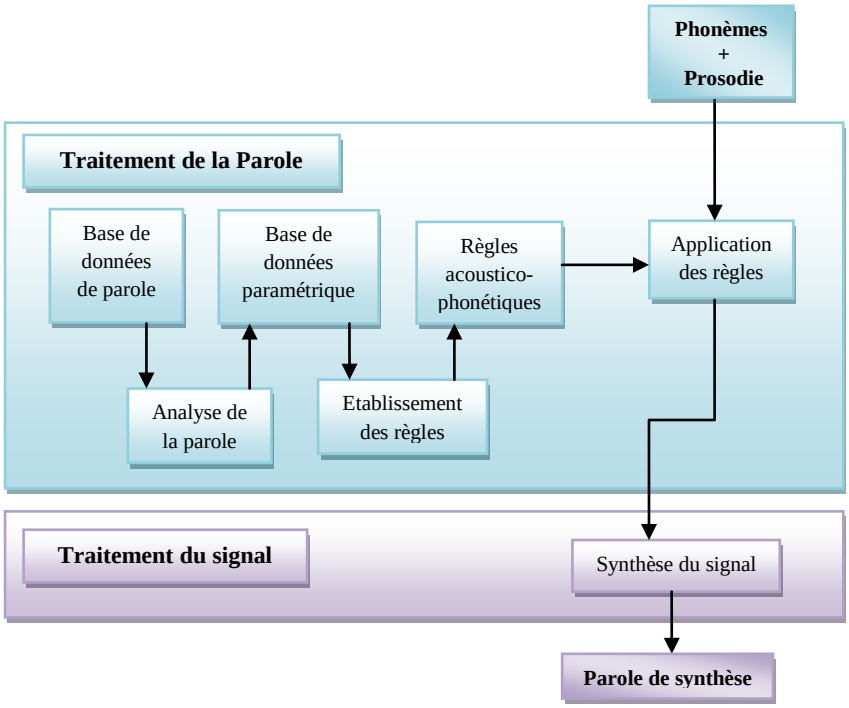


Figure 3.4 : Schéma de conception et de fonctionnement d'un système SPR [23]

Le principe est de constituer un dictionnaire de valeurs moyennes (table de valeurs paramétriques) et de formuler des règles qui permettent de décrire les transitions entre ces valeurs.

A la différence de la synthèse par concaténation où le dictionnaire contient des unités sonores préalablement enregistrées, celui utilisé par la SPR contient des paramètres

nécessaires à la restitution du segment de parole. De plus, à chaque combinaison de deux phonèmes est associés un ensemble de règles, définissant des trajectoires formantiques au cours de la transition. Les caractéristiques du modèle qui peut être modifié au cours du temps, sont définis en général par : la fréquence, la largeur de bande, l'amplitude de chaque formant, l'énergie, la durée, etc. Toutes les informations liées au voisement, à la nasalité, à la friction, turbulence. Soit environ 39 paramètres acoustiques selon un pas choisi (généralement un pas de 5 à 10 ms), pas au-delà de 20 ms.

Ce travail ne peut se faire manuellement, il est indispensable d'utiliser un compilateur de règles qui est un système expert qui permet de générer les règles et les valeurs intermédiaires (exemples de compilateurs de règles : COMPOST pour la langue française et RULSYS pour la langue Anglaise).

3.4.2.1. Les avantages de la SPR

- parole assez naturelle ;
- prise en compte des phénomènes de coarticulation (les suivre de façon précise) mais le lissage est indispensable ;
- vocabulaire illimité ;
- table des valeurs paramétriques limites (30 tables pour décrire l'Arabe Standard mémoire réduite) ;
- manipulation du pitch $\mathbf{F_0}$ (enfant, femme, homme) au niveau du générateur d'impulsions ;
- permet d'approfondir les connaissances linguistiques.

3.4.2.2. Les inconvénients de la SPR

- complexité dans la mise au point des règles ;
- le nombre de règles est assez important ;
- l'intelligibilité de la parole dépend du bon choix des règles ;
- le dictionnaire et les règles sont spécifiques pour chaque langue.

3.4.3. Synthèse par concaténation d'unités

Cette méthode consiste à synthétiser le signal de parole par concaténation d'unités acoustiques, c'est-à-dire de segments de parole préenregistrés. Cette technique, qui repose sur l'utilisation de segments de signaux extraits de la parole naturelle, est la seule qui permet à ce jour de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain [37] (Figure 3.5).

Les opérations qui dépendent de la langue sur la figure 3.5 sont indiqué par (**L**).

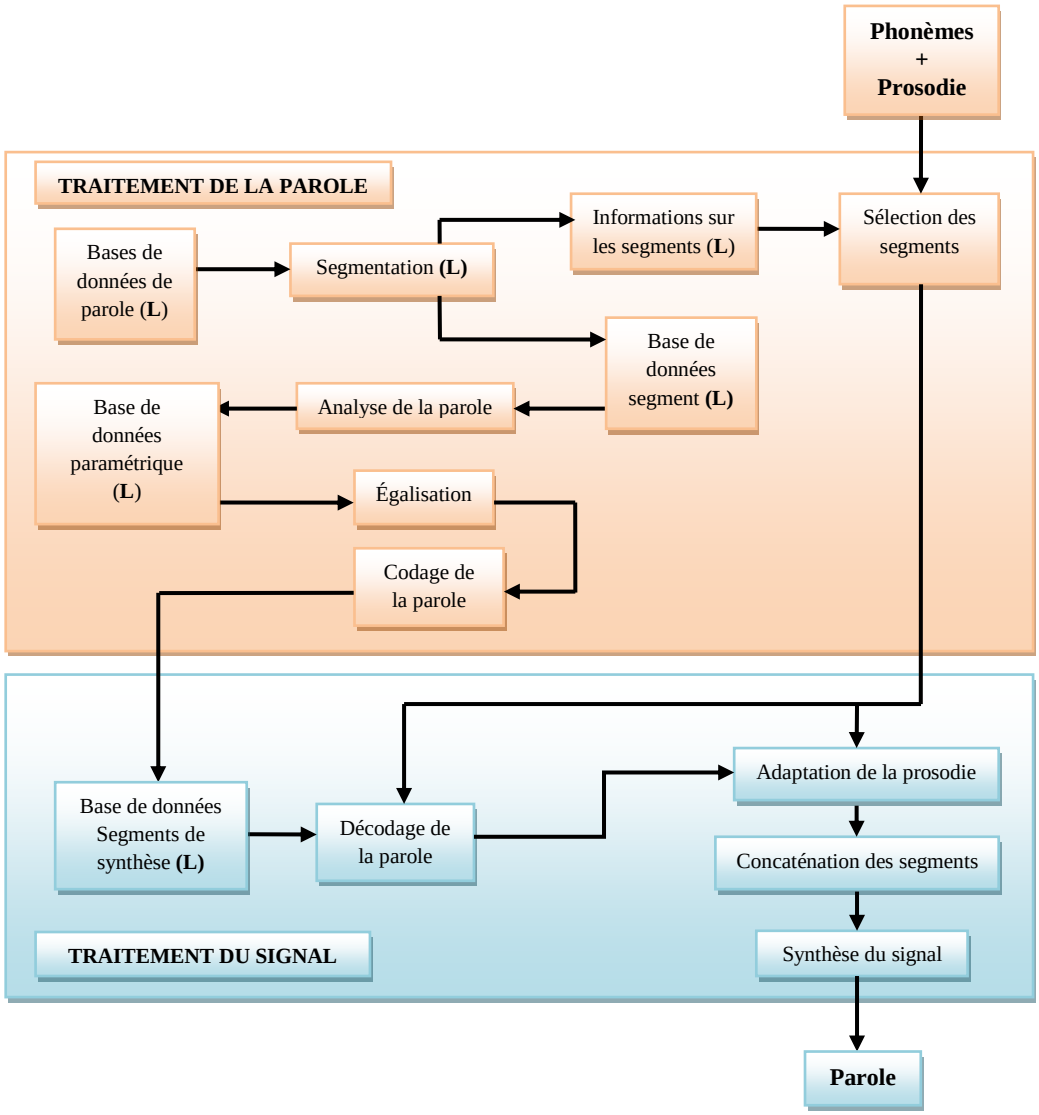


Figure 3.5 : Schéma général d'un synthétiseur par concaténation d'unités acoustiques [23]

3.4.4. Les méthodes de concaténation

Parmi les méthodes de synthèse par concaténation nous pouvons citer :

3.4.4.1. La concaténation de phonèmes

L'idée est de mettre bout à bout les unités sonores (phonèmes). Cette méthode présente l'inconvénient de produire une parole discontinue et cela est dû aux phénomènes de coarticulation.

3.4.4.2. La concaténation par diphones

Le diphone est l'unité acoustique qui débute de la partie stable d'un phonème et se termine par la partie stable du phonème adjacent et qui comporte dans son centre toute la transition (Figure 3.6).

La méthode par concaténation de diphones consiste à élaborer une base de données constituée de diphones enregistrés. C'est une alternative à la méthode précédente car elle prend en considération les transitions formantiques qui sont dues aux phénomènes de coarticulation et donnent de ce fait une qualité de synthèse satisfaisante. Par exemple le mot [kuba] sera décomposé en 5 diphones qui sont : [#k] [ku] [ub] [ba] [a#].

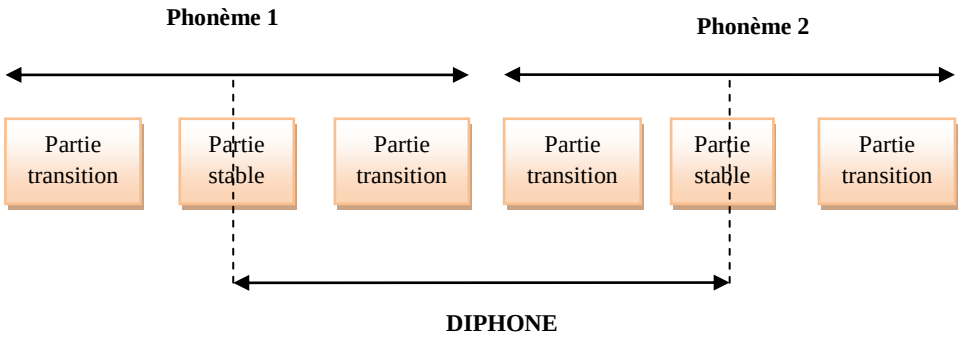


Figure 3.16 : Représentation du diphone dans une séquence sonore

3.4.4.3. La concaténation de mots

Il s'agit de mettre bout à bout des mots préenregistrés pour générer une phrase. Cette méthode produit des phrases de qualité inferieure à celle citée précédemment, et cela est dû aux problèmes de la coarticulation entre mots.

3.4.4.4. La concaténation de phrases

Cette méthode consiste à enregistrer des phrases en vue de leurs restitutions pour une application bien déterminée, tels que les opérateurs virtuels dans les trains, gares, horloge parlante, jouets parlants, etc. Cette méthode est à ce jour utilisée dans différents domaines, vue son faible coût et la facilité de son élaboration.

Les avantages et les inconvénients de quelques méthodes de synthèse par concaténation sont présentés dans le tableau suivant :

Méthode de synthèse	Avantages	Inconvénients
Phrases	Simples enregistrements de phrases, très bonnes qualité de restitution, excellente intelligibilité.	Domaine d'application limité, restitution de la parole fonction de la langue d'enregistrement.
Mots	Simple concaténation d'unités sonores, bonne qualité si les problèmes de coarticulation sont bien pris en compte, indépendance par rapport à la langue (sauf pour les règles de prosodie), règles d'assemblage assez simple.	Vocabulaire limité, domaine d'application limité, peu économique par stockage préalable des unités sonores sous forme codée, problème de l'intonation lors de la concaténation des mots.
Diphones	Possibilité de reconstituer n'importe quel texte d'une langue donnée, vocabulaire illimite, qualité satisfaisante, règles d'assemblage assez simple (simple concaténation)	Dictionnaire fonction de la langue et du locuteur, ne tient pas compte des effets de coarticulation pour les liquides et les semi-voyelles, nombre de diphones à enregistrer très élevé (1356 pour l'arabe Standard, 1156 pour le Français).

Tableau 3.1 : Tableau comparatif des avantages et inconvénients de quelques méthodes de synthèse par concaténation d'unités acoustiques

3.5. Les techniques utilisées en SAP

Avant d'aborder les principales techniques utilisées en SAP il est nécessaire de faire un tour d'horizon sur l'analyse et le traitement du signal de parole.

3.5.1. Modélisation du signal vocal

La modélisation la plus courante est le modèle source – filtre qui correspond respectivement en acoustique au flux d'air à travers les cordes vocales et aux cavités résonantes du conduit vocal (Figure 3.7) [38].

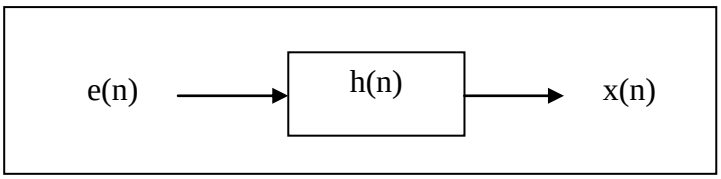


Figure 3.17 : Modèle de base de source-filtre pour le signal de parole [38]

Pour estimer le filtre plusieurs méthodes utilisées sont inspirées des modèles de production de la parole tels que le codage par prédiction et l'analyse cepstrale. Une fois le filtre estimé, la source peut être obtenue en faisant passer le signal de parole par le filtre inverse. La séparation entre source et filtre est l'un des défis les plus difficiles en traitement du signal de parole [38].

3.5.2. Analyse de la parole

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage, ou de reconnaissance [23]. Toutefois, analyse et synthèse sont deux activités duales. En effet, l'analyse fournit une description du signal acoustique, que la synthèse utilise pour le reproduire.

L'Analyse acoustique est une partie importante dans le traitement du signal sonore dans le but de réaliser un système de haute qualité. Cette opération consiste à extraire à partir du signal vocal un ensemble de paramètres pertinents, discriminants et robustes susceptibles de le représenter [22].

3.5.2.1. Prétraitement

Le but de cette opération est de numériser le signal vocal (Figure 3.8). Une fois que la parole a été prononcée par le locuteur, elle sera captée par un microphone. Le signal résultant sera filtré par un filtre analogique passe bande dans le but de réduire la bande passante puis échantillonné à une fréquence $\mathbf{F_e}$ qui devra être supérieure ou égale à deux fois de la fréquence maximale contenue dans ce signal, selon le Théorème de Shannon [39].

Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la $\mathbf{F_e}$ peut varier de 6 à 16 kHz Pour la téléphonie elle est de 8 kHz [23].

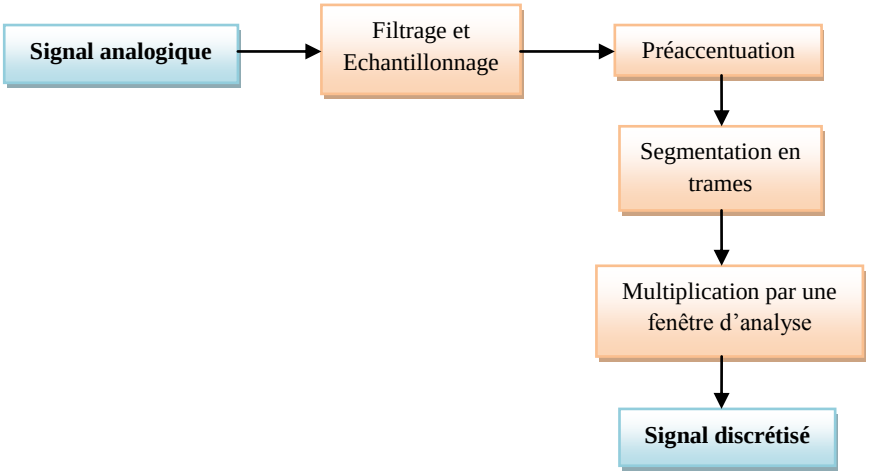


Figure 3.18 : Prétraitement du signal vocal [39]

3.5.2.2. Préaccentuation

Après avoir numérisé le signal de parole une préaccentuation est nécessaire pour élever les hautes fréquences qui sont moins énergétique que les basses. Pour cette raison un filtre de préaccentuation est utilisé, sa transmittance est :

$$H(z) = 1 - \mu\, z^{-1} \qquad 3.1$$

Où μ est le coefficient de préaccentuation inferieure et proche de 1.

3.5.2.3. Fenêtrage

Le signal résultant est décomposé en une succession de tranches élémentaires appelées fenêtres d'analyse ou trames. Chaque trame est constituée d'un nombre fixe d'échantillons de parole couvrant une durée de 20 à 30 ms. Le découpage en trames produit des

discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Ce phénomène s'appelle l'effet de bord. Pour compenser ces effets, chaque trame est multipliée par une fenêtre de Hamming qui a pour expression [39] :

$$W(t) = \begin{cases} 0.54 + 0.46 \cos 2\pi \frac{t}{T} & si\ t \in [0,T] \\ 0 & sinon \end{cases} \quad 3.2$$

Il existe différentes techniques d'analyse en SAP nous allons citer les plus connues.

3.5.3. Technique LPC

La méthode LPC ou Linear Predictive Coding est une méthode d'analyse très efficace basée sur le codage par prédiction linéaire, également connu par l'analyse par modélisation autorégressif (AR). Cette méthode est largement utilisé car elle est rapide et simple, mais aussi un moyen efficace d'estimer les principaux paramètres des signaux de la parole [38]. Ainsi, la fonction de transfert du filtre récursif tous pôles tel qu'il est présenté sur la figure 3.9 a pour équation :

$$H(z) = \frac{X(z)}{E(z)} = \frac{X(z)}{\sigma\, U(z)} = \frac{1}{A_p(z)} \quad 3.3$$

Avec

$$A_p(z) = 1 + \sum_{i=1}^{p} a_i\, z^{-i} \quad 3.4$$

Où p est l'ordre de prédiction, σ est le gain du système, et *U*(z) représente la transformée en Z du signal d'excitation u(n) qui est un train d'impulsion dans le cas de sons voisés et un bruit blanc pour les sons non voisés.

L'équation (3.3) va se simplifier :

$$X(z) = \sigma\ .\ \frac{U(z)}{A_p(z)} \quad 3.5$$

Qui va devenir :

$$X(z)\ .\ A_p(z) = \sigma\ .\ U(z) \quad 3.6$$

En appliquant la transformée en Z inverse à l'équation (3.6) on obtient dans le domaine temporel l'équation suivante :

$$x(n) + \sum_{i=1}^{p} a_i\, x(n-i) = \sigma.u(n) \qquad 3.7$$

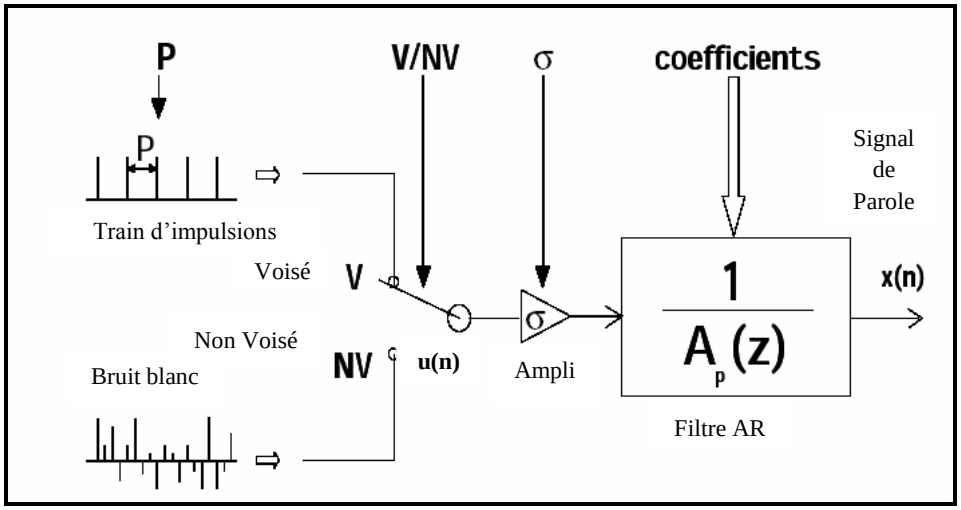


Figure 3.9 : Modèle autorégressif [40]

3.5.3.1. Estimation du modèle

Le codage prédictif linéaire tire son nom du fait qu'il prédit l'échantillon actuel comme une combinaison linéaire des p échantillons précédents [38] :

$$\tilde{x}(n) = \sum_{i=1}^{p} a_i\, x(n-i) \qquad 3.8$$

L'erreur de prédiction en utilisant cette approximation :

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{i=1}^{p} a_i\, x(n-i) \qquad 3.9$$

Pour estimer les coefficients de prédiction à partir d'un ensemble d'échantillons de parole, nous utilisons la technique de l'analyse à court terme. Nous allons définir $x_m(n)$ comme un segment de parole sélectionnée au voisinage de l'échantillon m :

$$x_m(n) = x(m+n) \qquad 3.10$$

Nous définissons l'erreur de prédiction à court terme pour ce segment comme :

$$E_m = \sum_n e_m^2(n) = \sum_n [x_m(n) - \tilde{x}_m(n)]^2 = \sum_n [x_m(n) - \sum_{j=1}^{p} a_j\, x_m(n-j)]^2 \qquad 3.11$$

Pour trouver les coefficients a_i qui minimisent E_m revient à annuler les dérivées partielles de E_m par rapport à ces coefficients.

L'équation (3.11) deviendra alors :

$$\sum_n x_m(n-i)\, x_m(n) = \sum_{j=1}^{p} a_j \sum_n x_m(n-i)\, x_m(n-j) \qquad i = 1,2,\dots,p \qquad 3.12$$

Pour plus de commodité, on peut définir les coefficients de corrélation comme :

$$\phi_m(i,j) = \sum_n x_m(n-i)\, x_m(n-j) \qquad 3.13$$

De sorte que les équations (3.12) et (3.13) peuvent être combinées pour obtenir ce qu'on appelle les équations de Yule-Walker :

$$\sum_{j=1}^{p} a_j\, \phi_m(i,j) = \phi_m(i,0) \qquad i = 1,2,\dots,p \qquad 3.14$$

La solution du groupe de *p* équations linéaires, résulte en *p* coefficients LPC qui minimisent l'erreur de prédiction. Avec a_i satisfaisant l'équation (3.14), l'erreur de prédiction de l'équation totale (3.11) prend la valeur suivante :

$$E_m = \sum_n x_m^2(n) - \sum_{j=1}^{p} a_j \sum_n x_m(n)\, x_m(n-j) = \phi(0,0) - \sum_{j=1}^{p} a_j \phi(0,j) \qquad 3.15$$

La solution des équations de Yule-Walker dans l'équation (3.14) peut être obtenue avec n'importe quel Algorithme standard d'inversion de matrice. En raison de la forme particulière de cette matrice, des solutions efficaces sont possibles et chaque solution offre un aperçu différent de l'autre.

Ces algorithmes sont : la méthode de covariance, la méthode d'autocorrélation, et la méthode treillis [38].

Deux méthodes seulement seront présentées ci-après :

3.5.3.2. La méthode de covariance

La méthode de covariance est calculée en définissant directement l'intervalle sur lequel la sommation dans l'équation (3.13) à lieu :

$$E_m = \sum_{n=0}^{N-1} e_m^2(n) \qquad 3.16$$

De telle manière que $\phi_m(i,j)$ devienne :

$$\phi_m(i,j) = \sum_{n=0}^{N-1} x_m(n-i)\, x_m(n-j) = \sum_{n=-i}^{N-1-j} x_m(n)\, x_m(n+i-j) = \phi_m(j,i) \qquad 3.17$$

Et l'équation (3.14) devienne :

$$\begin{pmatrix} \phi_m(1,1) & \cdots & \phi_m(1,p) \\ \vdots & \ddots & \vdots \\ \phi_m(p,1) & \cdots & \phi_m(p,p) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \phi_m(1,0) \\ \vdots \\ \phi_m(p,0) \end{pmatrix} \quad 3.18$$

Qui peut être exprime de la manière suivante :

$$\boldsymbol{\Phi a = \psi} \quad 3.19$$

Où la matrice **Φ** dans l'équation. (3.19) est symétrique et définie positive, pour laquelle des méthodes efficace sont disponibles, telle que la factorisation de **Cholesky**. Pour cette méthode, aussi appelée la méthode de la racine carrée, la matrice **Φ** est exprimé en :

$$\boldsymbol{\Phi = VDV^t} \quad 3.20$$

Où **V** est une matrice triangulaire inferieure (dont les éléments de la diagonale sont égaux à 1), et **D** est une matrice diagonale. Donc chaque élément de **Φ** s'exprime comme suit :

$$\phi(i,j) = \sum_{k=1}^{j} V_{ik}\, d_k V_{jk} \quad 1 \leq j < i \quad 3.21$$

Ou bien :

$$V_{ij} d_j = \phi(i,j) - \sum_{k=1}^{j-1} V_{ik} d_k V_{jk} \quad 1 \leq j < i \quad 3.22$$

Et pour les éléments diagonaux :

$$\phi(i,i) = \sum_{k=1}^{j} V_{ik} d_k V_{ik} \quad 3.23$$

Ou bien :

$$d_i = \phi(i,i) - \sum_{k=1}^{i-1} V_{ik}^2\, d_k \quad i \geq 2 \quad 3.24$$

Avec :

$$d_1 = \phi(1,1) \quad 3.25$$

La factorisation de Cholesky commence par l'équation (3.25) puis alterne entre les équations (3.22) et (3.24). Une fois que les matrices **V** et **D** ont été déterminés, les coefficients LPC sont résolus dans un processus en deux étapes. La combinaison des équations (3.19) et (3.20) peut être exprimées comme :

$$\mathbf{VY} = \boldsymbol{\psi} \tag{3.26}$$

Avec :

$$\mathbf{Y} = \mathbf{D}\mathbf{V}^{\mathbf{t}}\boldsymbol{a} \tag{3.27}$$

Ou bien :

$$\mathbf{V}^{\mathbf{t}}\boldsymbol{a} = \mathbf{D}^{-1}\mathbf{Y} \tag{3.28}$$

Par conséquent, la matrice **V** étant donnée et l'équation (3.26), **Y** peut être résolu par récursivement de la manière suivante :

$$Y_i = \psi_i - \sum_{j=1}^{i-1} V_{ij}\ Y_j \qquad 2 \leq i < p \tag{3.29}$$

Avec la condition initiale :

$$Y_1 = \psi_1 \tag{3.30}$$

Ayant déterminé **Y**, l'équation (3.28) peut être résolu récursivement de la même manière comme suit :

$$a_i = \frac{Y_i}{d_i} - \sum_{j=i+1}^{p} V_{ji}\, a_j \qquad 1 \leq i < p \tag{3.31}$$

Avec la condition initiale :

$$a_p = \frac{Y_p}{d_p} \tag{3.32}$$

Où l'indice *i* dans l'équation (3.31) se déroule à l'envers.

3.5.3.3. La méthode d'autocorrélation

La somme dans l'équation (3.13) n'a pas d'intervalle spécifique. Dans la méthode d'autocorrélation, nous supposons que $x_m(n)$ est 0 en dehors de l'intervalle $0 \leq n < N$ [38] :

$$x_m(n) = x(m+n)\, w(n) \tag{3.33}$$

Avec $w(n)$ une fenêtre (tel que la fenêtre de Hamming) qui prend la valeur 0 en dehors de l'intervalle $0 \leq n < N$. Avec cette hypothèse, l'erreur de prédiction $e_m(n)$ correspondante est

non-nulle sur l'intervalle 0 ≤ n < N + p, et par conséquent, l'erreur de prédiction totale prend la valeur :

$$E_m(n) = \sum_{n=0}^{N+p-1} e_m^2(n) \qquad 3.34$$

Avec cet intervalle, l'équation (3.13) peut être exprimé come suit :

$$\phi_m(i,j) = \sum_{n=0}^{N+p-1} x_m(n-i)\ x_m(n-j) = \sum_{n=0}^{N-1-(i-j)} x_m(n)\ x_m(n+i-j) \qquad 3.35$$

Ou alternativement :

$$\phi_m(i,j) = R_m(i-j) \qquad 3.36$$

Avec $R_m(k)$ la séquence d'autocorrélation de $x_m(n)$:

$$R_m(k) = \sum_{n=0}^{N-1-k} x_m(n)\ x_m(n+k) \qquad 3.37$$

En combinant l'équation (3.36) et l'équation (3.14), on obtient :

$$\sum_{j=1}^{p} a_j R_m(|i-j|) = R_m(i) \qquad 3.38$$

Qui correspond à l'équation suivante :

$$\begin{pmatrix} R_m(0) & \cdots & R_m(p-1) \\ \vdots & \ddots & \vdots \\ R_m(p-1) & \cdots & R_m(0) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_m(1) \\ \vdots \\ R_m(p) \end{pmatrix} \qquad 3.39$$

La matrice de l'équation (3.39) est symétrique et tous les éléments de sa diagonale sont identiques. De telles matrices sont appelées Toeplitz. La récurrence Durbin exploite ce fait entraînant un algorithme très efficace (pour plus de commodité, nous omettons l'indice m de la fonction d'autocorrélation) :

1- Initialisation

$$E^0 = R(0) \qquad 3.40$$

2- Itération pour i=1,..., p exécuter la récurions suivante :

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} a_j^{i-1} R(i-j)\right] / E^{i-1} \qquad 3.41$$

$$a_i^i = k_i \qquad 3.42$$

$$a_j^i = a_j^{i-1} - k_i\, a_{i-j}^{i-1} \qquad 1 \leq j < i \tag{3.43}$$

$$E^i = \left(1 - k_i^2\right) E^{i-1} \tag{3.44}$$

3- Solution finale :

$$a_j = a_j^p \qquad 1 \leq j < p \tag{3.45}$$

Où les coefficients k_i, nommés coefficient de réflexion, sont compris entre –1 et 1. Dans le processus de calcul des coefficients de prédiction d'ordre *p*, la récursion trouve la solution des coefficients de prédiction pour tous les ordres inférieurs à *p*.

En remplaçant *R(j)* avec les coefficients normalisés d'autocorrélation *r(j)* définis par :

$$r(j) = R(j)/R(0) \tag{3.46}$$

Résulte en coefficients LPC identiques, et la récursion est plus robuste aux problèmes de précision arithmétique. De même, l'erreur de prédiction normalisée à l'itération *i* est défini par: la division de l'équation (3.15) par R [0], ce qui, en utilisant l'équation (3.36), résulte en :

$$V^i = \frac{E^i}{R(0)} = 1 - \sum_{j=1}^{i} a_j\, r(j) \tag{3.47}$$

L'erreur de prédiction normalisée est, en utilisant l'équation (3.43) et l'équation (3.47) :

$$V^p = \prod_{i=1}^{p} (1 - k_i^2) \tag{3.48}$$

3.5.4. Technique PSOLA

PSOLA (Pitch Synchronous Overlap and Add) ou Superposition/Addition de fenêtres synchrones à la période fondamentale du signal ou plus communément « Recouvrement – Addition Synchrone au Pitch » est une technique de traitement du signal numérique qui est utilisée dans le domaine de synthèse de la parole. Cette technique ne fait pas la synthèse proprement dite mais permet de concaténer et lisser des segments de parole préenregistrés. Elle permet la modification de la durée et du pitch de ces segments. C'est une variante d'OLA (OverLap and Add) qui se ramifie en plusieurs techniques : SOLA (Synchronous OverLap and Add), TD–PSOLA (Time Domain PSOLA), FD–PSOLA (Frequency Domain PSOLA), LP–PSOLA (Linear Prediction PSOLA), WSOLA (Waveform

Similarity OverLap and Add) ainsi que MBROLA (Multi-Band Re-synthesis OverLap and Add) qui sera expliqué plus bas [41].

Pour la technique PSOLA, il s'agit de décomposer le signal échantillonné *s(n)* en des signaux dits à court terme $s_m(n)$ obtenus par une multiplication de *s(n)* par une suite de fenêtres d'analyse *h(n)* centrées sur les instants t_m marques de pitch ou marque de lecture.

$$s_m(n) = s(n)\, h(t_m - n) \qquad 3.49$$

Ces marques se succèdent à une cadence synchrone du pitch sur les segments voisés du signal vocal. Dans les parties non voisées les marques de pitch sont remplacées par un intervalle arbitraire fixe à 10 ms. La longueur des fenêtres est choisie de façon à ce que deux signaux élémentaires consécutifs présentent un recouvrement mutuel important variant typiquement entre 50% et 75%.

La modification des paramètres prosodiques (durée et pitch) consiste à produire du flux des signaux élémentaires d'analyse un flux de signaux élémentaires de synthèse *S(n)*, synchronisées sur une nouvelle suite d'instants t_s, appelés marques de synthèse. Ces modifications correspondent à la duplication ou l'élimination des fenêtres dont l'écartement peut être modifié.

La synthèse est la dernière étape qui consiste à calculer le signal de synthèse $\hat{S}(n)$ par simple superposition et addition des signaux élémentaires de synthèse qui présentent un taux de recouvrement important de façon analogue aux signaux d'analyse [39] (Fig. 3.10).

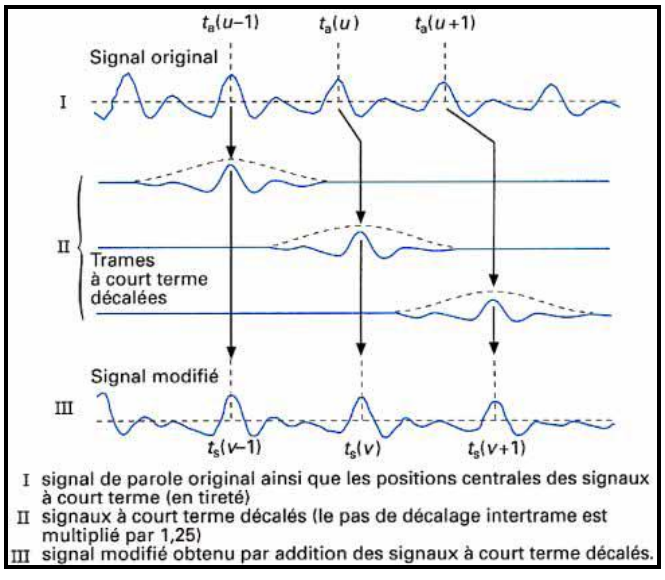


Figure 3.10 : Modification de la $\mathbf{F_0}$ par un facteur 0,8 avec la méthode PSOLA [37]

3.5.5. Technique MBROLA

MBROLA (Multi-Band Re-synthesis OverLap and Add), est une technique de synthèse qui permet de produire un énoncé naturel par concaténation de diphones préenregistrés, en spécifiant simplement leurs caractéristiques prosodiques.

Le projet MBROLA a été lancé en 1993 par le laboratoire TCTS de la faculté Polytechniques de Mons en Belgique et breveté internationalement depuis 1996, il a été développé par Thierry Dutoit.

La mise en disposition gratuite de cet outil auprès de la communauté internationale a encouragé de nombreux laboratoires dans plusieurs pays à constituer une base de données de diphones pour un but de développement de la synthèse vocale.

Après constitution de la base de données de diphones, le synthétiseur est mis à la disposition des utilisateurs potentiels pour toutes applications non commerciale et non militaire.

3.5.5.1. Caractéristiques principales de la méthode MBROLA

La synthèse par la méthode MBROLA utilise l'algorithme PSOLA pour la concaténation des diphones mais auparavant, des traitements spécifiques sont appliqués à la base de diphones originale composée de diphones extraits de signaux de parole naturelle. La nouvelle base de données est obtenue de la façon suivante:

La parole de départ est codée à l'aide d'un modèle d'excitation multi-bande: ceci permet de réduire considérablement la taille de la base de données. Des modifications ayant pour objectif de minimiser les problèmes de différence d'amplitude, de pitch et de phase lors de la concaténation des diphones sont ensuite appliquées aux unités de cette base; ainsi les diphones sont tous codés avec un pitch constant et l'amplitude est lissée en début et fin de diphone pour minimiser les différences d'amplitude lors du processus de concaténation.

3.5.5.2. Quels sont les étapes à suivre pour synthétiser un texte avec MBROLA?

Il faut tout d'abord installer le logiciel MBROLA avec toutes ses applications. Il s'agit en premier lieu d'écrire une phrase en symboles phonétiques **SAMPA** (Speech Assessment Methods Phonetic Alphabet) qui est un jeu de caractères phonétiques utilisable sur ordinateur utilisant les caractères ASCII 7-bits imprimables, basé sur l'Alphabet

Phonétique International (API) (Tableau 3.2) et de fournir les informations suivantes : pauses, durée de chaque phonème, mouvements de $\mathbf{F_0}$ (jusqu'à une quinzaine par phonème). Enregistrer le fichier en extension « .pho » si on l'ouvre avec MBROLA sinon en extension « .wav ».

SAMPA Consonnes Françaises	**Exemples**	**SAMPA Voyelles Françaises**	**Exemples**
p	**p**ont	**i**	s**i**
b	**b**on	**e**	s**e**s
t	**t**emps	**E**	s**ei**ze
d	**d**ans	**a**	p**a**tte
k	**c**oût, **qu**and, **k**oala	**A**	p**â**te
g	**g**ant	**O**	c**o**mme
f	**f**emme	**o**	gr**o**s
v	**v**ent	**u**	d**ou**x
s	**s**ans, de**ss**us, **c**erise	**y**	d**u**
z	**z**one, ro**s**e	**2**	d**eu**x
S	**ch**amp	**9**	n**eu**f
Z	**g**ens, **j**ouer	**@**	just**e**ment
j	**i**on [jo~]	**e~**	v**in**
m	**m**ont	**a~**	v**ent**
n	**n**om	**o~**	b**on**
J	oi**gn**on	**9~**	br**un**
N	campi**ng**		
l	**l**ong		
R	**r**ond		
w	q**u**oi [kwa]		
H	j**u**in [ZHe~]		

Tableau 3.2 : Symboles SAMPA du Français

Un exemple pour illustrer

Pour synthétiser le mot **bonjour**, on utilise les phonèmes SAMPA suivant :

« _b o~ Z u R _ »

; Synthèse du mot bonjour

_51

b 187 8 163 16 160

o~ 123 14 148 29 149 44 151 49 148

Z 88 11 139 25 137

u 95 8 122 15 122

R 163 8 109 16 108

_ 130

Explication de l'exemple

- le point virgule « ; » signifie que c'est un commentaire ;
- le signe « _ » permet d'insérer un silence, ici de 51ms ;
- o~ : « on » désigne le phonème que l'on veut prononcer ;
- 123 : est la durée en (ms) que l'on utilise pour prononcer ce phonème ;
- 14 148 : durant 14% de la durée, ce phonème est prononcé à 148 Hz ;
- 29 149 : entre 14% et 29% de la durée, ce phonème est prononcé à 149 Hz.

Vue les quantités importantes de phonèmes et de paramètres à implémenter vient la nécessité d'utiliser MBROLIGN (1997–2001) qui est un outil de MBROLA, fourni gratuitement dans les mêmes conditions que ce dernier. C'est un logiciel d'alignement automatique de phonème. Cet aligneur rapide basé sur le système TTS suit les étapes suivantes :

1- choisir la base de données à utiliser ;

2- ouvrir un fichier au format « .wav » ;

3- fournir une transcription phonétique ;

4- Lancer MBROLIGN ;

5- écouter le résultat synthétisé ;

6- enregistrer (en .wav ou en .pho).

Le résultat n'est vraiment bon que si le son ouvert a les caractéristiques suivantes :

- 16 kHz pour l'échantillonnage ;
- 16 bit pour le codage.

3.6. Conclusion

Dans ce chapitre nous avons décrit la SAP en donnant la structure d'un système de synthèse à partir du texte, les différentes méthodes de synthèse et en dernier nous avons expliqué quelques techniques utilisées en synthèse qui sont les techniques LPC, PSOLA et MBROLA.

Chapitre 4 : Implémentation des Algorithmes, Résultats et Évaluations

4.1. Introduction

Dans ce chapitre nous traitons la partie implémentation des algorithmes de notre application SYPHRAMO. Tout d'abord, nous décrivons le travail effectué puis nous présentons une brève description du logiciel d'analyse Praat. Par la suite, nous allons expliquer les éléments essentiels concernant la SYPHRAMO, et en dernier lieu nous expliquons les résultats obtenus et leurs évaluations.

4.2. Description du travail effectué

L'objectif de notre travail est la réalisation d'un système de synthèse vocale capable de produire une parole intelligible et de bonne qualité à partir d'un corpus constitué de phrases et mots préenregistrés, consacré à une application bien déterminé. Celle-ci sera nommée SYPHRAMO et consiste à annoncer les noms des stations de départ et d'arrivée (dans la ville d'Alger) ainsi que les noms des arrêts de bus des transports de voyageurs avant d'y arriver.

SYPHRAMO peut être utilisé dans n'importe quelle ville (quartier). Vue son aspect informationnelle. Ce système aura pour but d'annoncer les arrêts aux passagers pour leur permettre de se préparer à descendre. Il sera aussi plus bénéfique pour les non-voyants et les étrangers à la ville ou au pays, favorisant ainsi leur mobilité et leurs autonomies dans les transports publics.

Dans le cadre de ce mémoire, nous avons utilisé un corpus constitué de phrases affirmatives et mots en AS, prononcées par une locutrice arabophone.

Ces phrases ont été enregistrées et ont subi une analyse sonagraphique grâce au logiciel de transcription et d'analyse phonétique Praat.

Les moyens informatiques dont nous disposons sont constitués d'un micro-ordinateur portable de type SONY VAIO avec 2 Go de mémoire RAM et Windows XP comme système d'exploitation. Pour la programmation de SYPHRAMO, nous avons utilisé le logiciel Delphi 7, Microsoft Access 2007 et le langage Delphi.

- **Le logiciel Praat**

Praat a été développé par P. Boersma et D. Weeninck au cours des années 80 à l'Université d'Amsterdam. Praat est un logiciel de transcription, d'analyse et de traitement

de signal. Il est utilisé en phonétique, phonologie et dans d'autres domaines des sciences du langage, notamment en linguistique interactionnelle, ainsi que dans d'autres disciplines, tel que l'anthropologie, musicologie, et en médecine. Le logiciel est continuellement mis à jour et complété et peut être récupéré sur l'adresse suivante: www.praat.org [44].

Il est possible d'effectuer plusieurs tâches avec cet outil d'analyse (Figure 4.1), il permet entre autre de :

- enregistrer des fichiers audio qui pourront ensuite être analysés ;
- transcrire, étiqueter et segmenter des données audio (enregistrements effectués sous Praat ou provenant d'autres fichiers, au format wav, par exemple);
- effectuer des analyses phonétiques et acoustiques au niveau segmental (spectrogramme, analyse de formants, sonagramme) et au niveau suprasegmental (pitch, courbe de $\mathbf{F_0}$, intensité et durée) [23].

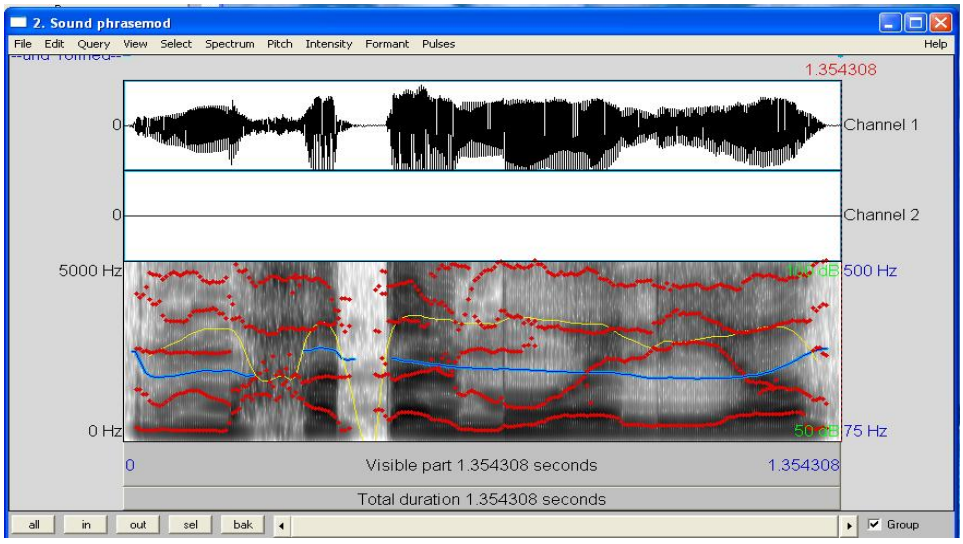


Figure 4.1 : Analyse à l'aide du logiciel Praat de la phrase /المحطة الموالية/ [al-maḥaṭa al-muwaliya]

4.3. Implémentation des Algorithmes de SYPHRAMO

Pour l'élaboration de SYPHRAMO nous avons utilisé à titre d'exemple un corpus constitué de 3 phrases et 10 mots. Ces phrases et mots sont utilisés pour signaler les arrêts ainsi que la station de départ et la station finale de la ligne (Tableau 4.1).

L'annonce automatique des arrêts par SYPHRAMO est possible grâce au signal que transmet un détecteur de proximité (D) installé à l'extérieur du bus et relié au système. Il

aura pour but de détecter l'arrêt à environ 20 mètres avant d'y arriver pour permettre aux passagers de se préparer à descendre.

Phrases et mots numérotés	Corpus utilisé
P 1	نعلم السادة المسافرين أن هذه الحافلة تبدأ رحلتها من محطة
M 1	بن عمر
P 2	المحطة الموالية
M 2	القبة
M 3	كالفير
M 4	الواحات
M 5	رويسو
GM 6	حمود بوعلام
GM 7	حديقة التجارب
M 8	الحامة
M 9	بلكور
P 3	نعلم السادة المسافرين أن المحطة النهائية هي
GM 10	ساحة أول ماي

Tableau 4.1 : Phrases et mots constituants le corpus utilisé

Avec : Pi, i= 1,…,3 et Mi, i=1,2,…, 10 (GM : groupe de mots)

4.3.1. Les étapes du programme de SYPHRAMO

L'organigramme de la figure 4.2 explique les principales étapes qui constituent le programme de SYPHRAMO :

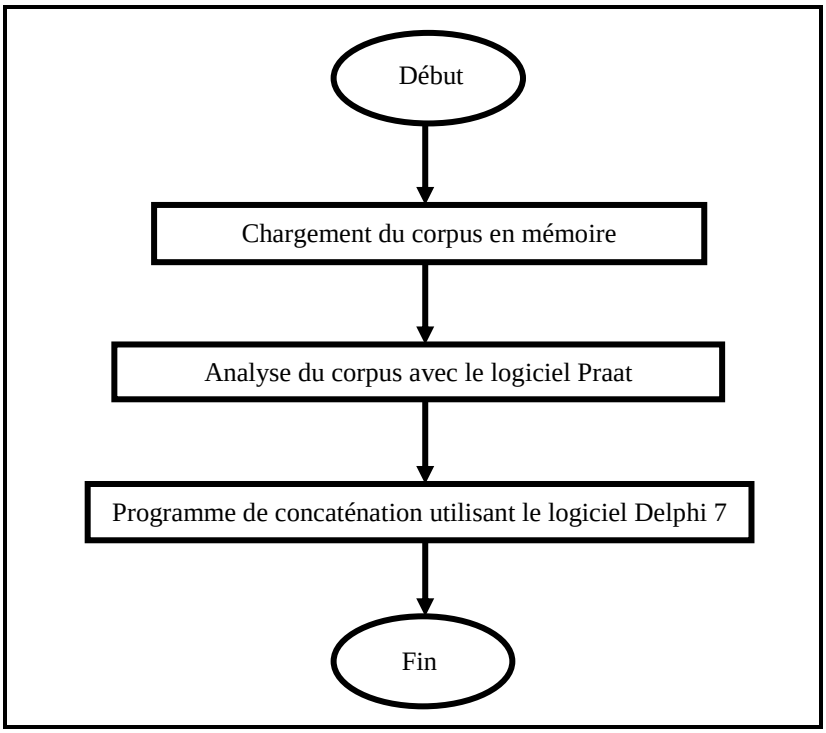


Figure 4.2: Principales étapes de la SYPHRAMO

4.3.1.1. Chargement du corpus en mémoire

Nous avons choisi d'enregistrer les fichiers sons par une locutrice arabophone, car les enregistrements effectués au préalable par un locuteur masculin ont donné des résultats pas très satisfaisants. Ces fichiers sont chargés en mémoire en vue de les utiliser par SYPHRAMO.

La gestion du enregistré est fait à l'aide de Microsoft Access 2007, qui permet de récupérer les fichiers sons un par un avec leur chemin d'accès en vue de leur appel par le programme principal.

4.3.1.2. Analyse de la base de données avec le logiciel Praat

L'analyse par le logiciel Praat consiste à éliminer les bruits et les silences indésirables avant et après chaque son de parole.

4.3.1.3. Programme de concaténation utilisant le langage Delphi 7

C'est l'étape qui décrit les éléments essentiels qui constituent le corps du programme de SYPHRAMO (Figure 4.3).

Dans la figure 4.3 : C désigne un compteur actionné par le détecteur de proximité D tel que : $1 \leq C < N+1$.

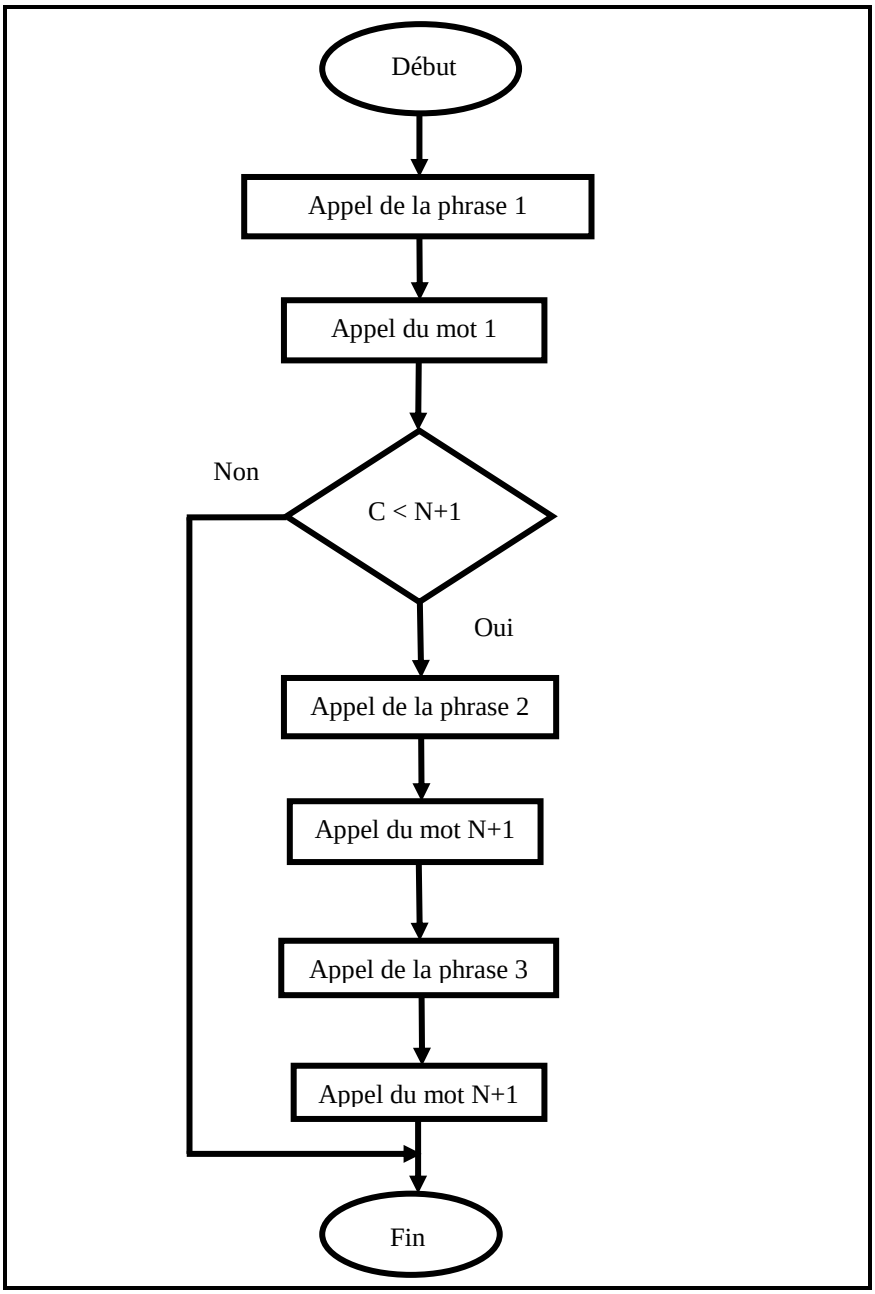


Figure 4.3 : Organigramme du programme de concaténation

Par exemple :

- Quand **C= 1** : le capteur D détecte à 20 m le **1er arrêt** qui correspond au son du **M 2** ;
- Quand **C= 2** : le capteur D détecte à 20 mètres le **2ème arrêt** qui correspond au son du **M 3** ;

Et ainsi de suite jusqu'au dernier Arrêt N+1 :

- Quand **C= N** : le capteur D détecte à environ 20 mètres le **N^{ème} arrêt** qui correspond au son du **M N+1** (dans notre cas N= 9).

4.4. Présentation du langage Delphi 7

Delphi est un logiciel de développement visuel rapide sous Windows (RAD : Rapid Application Development) conçu pour créer des applications fenêtrées directement exécutables sous Windows. Sa simplicité d'emploi autorise une utilisation immédiate, car il suffit de cliquer-glisser des composants dans une fiche et de gérer quelques événements pour créer des applications simples (Figure 4.4).

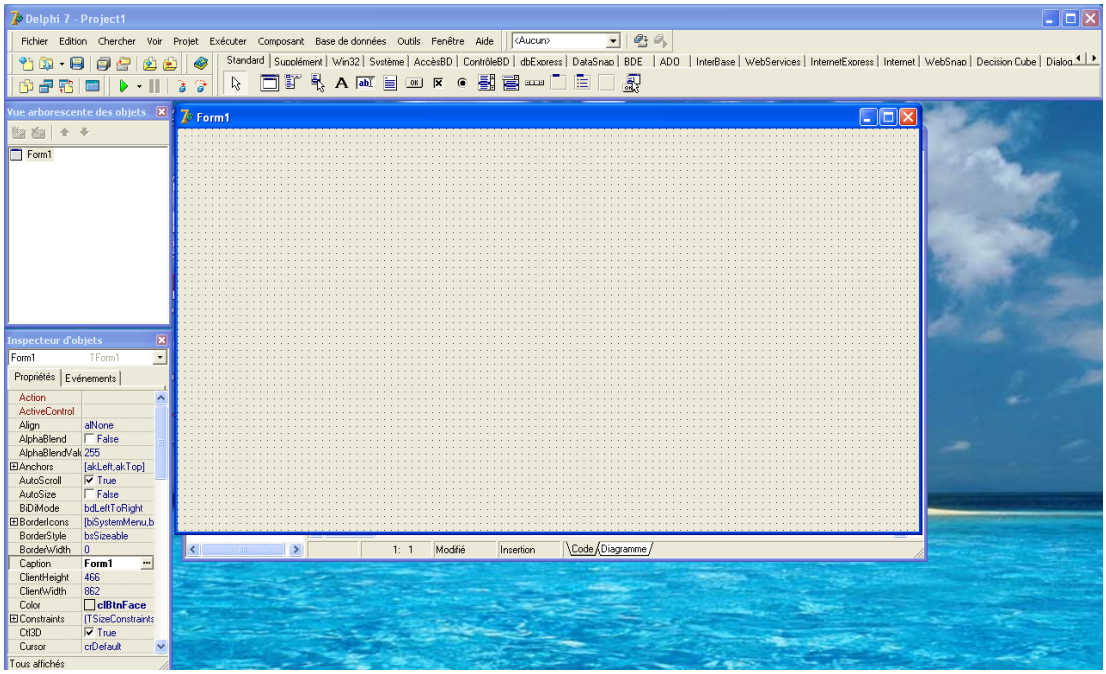


Figure 4.4 : Lancement de Delphi 7

4.5. Structure de SYPHRAMO

SYPHRAMO possède une interface graphique conçue avec le logiciel Delphi 7, qui permet de visualiser les stations de départ et d'arrivée et les différents arrêts en même temps qu'ils sont prononcés par l'opératrice virtuelle (Figure 4.5).

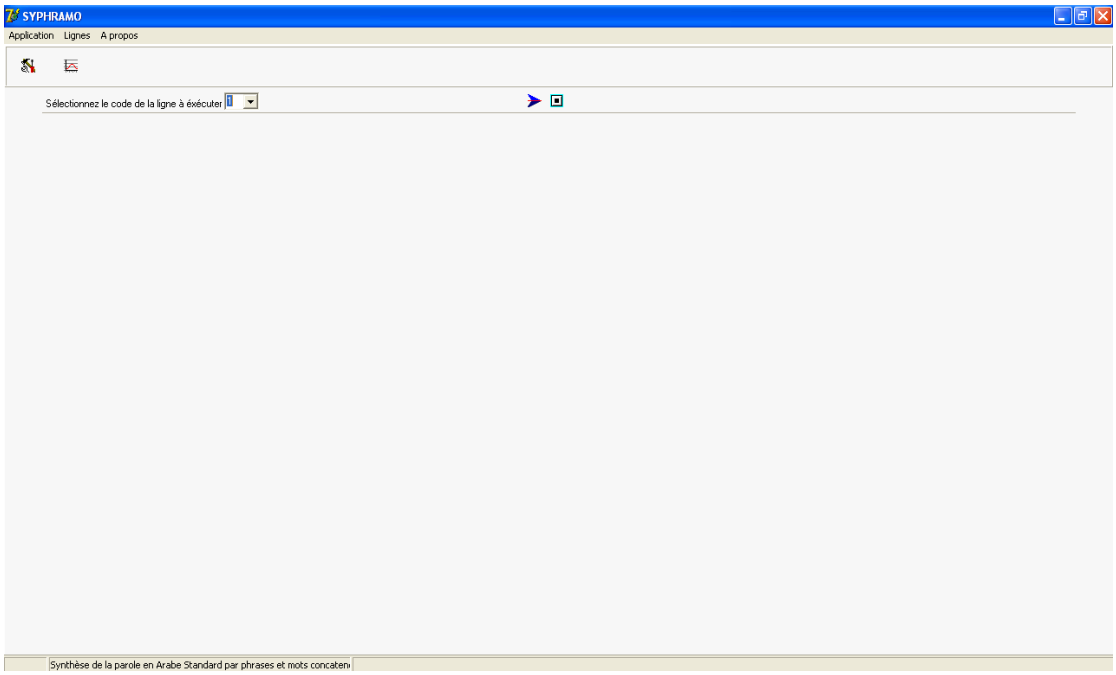


Figure 4.5 : SYPHRAMO

SYPHRAMO comporte les boutons suivants :

- marche : autorise le démarrage du système ;
- arrêt : permet l'arrêt de l'application;
- bouton (combobox) : permet de sélectionner le code de la ligne à exécuter ;
- gestion des lignes : permet de régler et de modifier les paramètres liés à la ligne (la direction, la distance parcourue, ajout de nouvelles lignes) ainsi que l'ajout ou la suppression d'arrêt avec leurs fichiers sons correspondants (Figure 4.6) ;
- paramètres : ce bouton permet de changer les fichiers sons des phrases fixes (phrase 1, phrase 2, phrase 3) à partir d'un nouveau corpus (Figure 4.7).

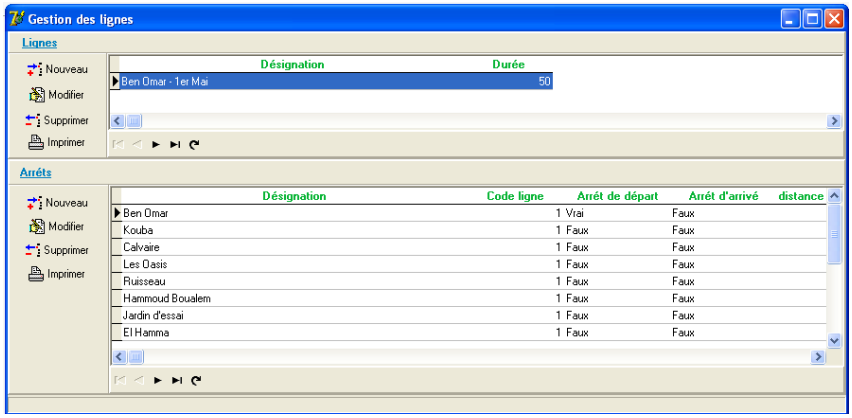


Figure 4.6 : Gestion des Lignes

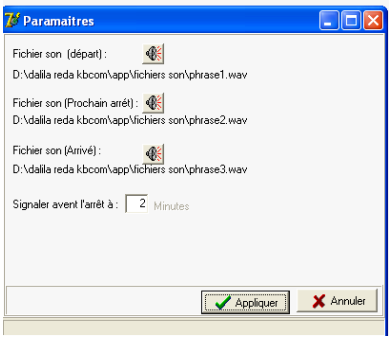


Figure 4.7 : Réglage des paramètres

4.6. Evaluation de SYPHRAMO

Pour effectuer une évaluation de SYPHRAMO nous avons pris pour exemple : la facturation en ligne de la société Algérie Telecom.

Le corpus utilisé dans ce cas est composé de 3 phrases et 3 mots. Les phrases et mots utilisés ont pour objectif d'informer le client d'Algérie Telecom du montant de sa facture bimestrielle et cela après avoir fait rentrer le code client (tableau 4.2).

Phrases et mots numérotés	Corpus utilisé
P 1	أهلا بكم في خدمة الزبائن لـ
P 2	أدخل رقم الزبون
P 3	قيمة الفاتورة تقدر بـ
GM 1	اتصالات الجزائر
M 2	ألفين
GM 3	دينار جزائري

Tableau 4.2 : Corpus utilisé

Pour plus de commodité nous avons changé l'interface graphique pour qu'elle simule l'opération de facturation (Figure 4.8).

Figure 4.8 : Interface graphique pour facturation en ligne

Après traitement des données nous avons constaté que SYPHRAMO a donné de bons résultats. Il pourra être utilisé de ce fait pour les systèmes comportant 3 phrases fixes et plus de 3 mots.

4.7. Test d'évaluation subjectif de la qualité de la parole

Afin d'évaluer la qualité de la parole synthétisée obtenue à l'aide de SYPHRAMO, nous avons fais recours aux tests d'évaluation subjectif. Il s'agit de faire entendre de la parole synthétisée à une dizaine de personnes de différents âges et sexes qui donneront leurs avis sur sa qualité du point de vue de son intelligibilité et son naturel (Tableau 4.3)

Personnes	**Age**	**Évaluation subjective**	
		Intelligibilité	**Naturel**
H1	65	excellent	très bon
F1	72	excellent	très bon
H2	33	excellent	bon
F2	23	excellent	bon
H3	57	excellent	très bon
F3	45	excellent	très bon
H4	47	excellent	bon
F4	34	excellent	très bon
H5	18	excellent	très bon
F5	20	excellent	bon
		Taux : 100%	**Taux : 60%**

Tableau 4.3 : Test d'évaluation subjectif

Nous pouvons conclure du tableau précédent que la parole synthétisée par SYPHRAMO est à 100% intelligible et 60% naturelle.

4.8. Conclusion

Dans ce chapitre nous avons expliqué les détails de notre conception. SYPHRAMO est un système de synthèse vocale à partir de phrases fixes et mots variables concaténés. Il est capable d'annoncer en AS avec une excellente intelligibilité et un très bon naturel, les noms des stations de départ et d'arrivée ainsi que les noms des arrêts de bus des villes Algériennes. Comme il peut être utilisé pour la facturation téléphonique en ligne.

Conclusions Générales et Perspectives

Conclusions et perspectives

L'objectif de notre travail a été une étude dans le domaine de la synthèse de la parole d'une part, et l'implémentation des résultats obtenus dans un système (SYPHRAMO) destiné à annoncer les arrêts de bus des transports de voyageurs dans la ville d'Alger, d'autre part.

Tout au long de ce mémoire nous avons essayé de décrire les systèmes de Traitement Automatique de la Parole en se focalisant sur la synthèse vocale. Notre travail se consacre à l'étude de la synthèse de la parole par concaténation d'unités acoustiques. Nous avons choisi les phrases fixes et mots variables comme unités acoustiques à cause de notre application qui nécessite une base de données réduite.

Nous pouvons conclure de ce que nous avons exposé tout au long de ce mémoire que nous avons donné des notions sur le TAP. Nous nous sommes familiarisées avec le logiciel Praat et le langage Delphi 7 qui nous ont permis d'acquérir des connaissances sur le traitement automatique de la parole en général et la synthèse vocale en particulier.

Notre système est flexible et facile à exploiter. La qualité de la parole est facilement intelligible et très naturelle mais en contre partie, elle donne des fichiers très volumineux en espace mémoire.

Les perspectives que nous suggérons sont :

- Il serait intéressant d'ajouter d'autres langues telles que le Français et l'Anglais,… etc. ;
- Ajout d'autres phrases :
  - message de bienvenue sur les lignes de la Société de Transport des Voyageurs ;
  - annoncer la possibilité de correspondance vers le métro, tramway ou bus d'une autre trajectoire ;
  - mise en garde de ne pas forcer les portes et de libérer le passage ;
  - annonce des arrêts sur écran LCD pour les mal-entendants.

Références Bibliographiques

Références Bibliographiques

[1] Parole, consulté durant 2012. http://fr.wikipedia.org/wiki/Parole

[2] Parole, consulté durant 2012. http://www.larousse.fr/dictionnaires/francais/parole/58286

[3] Les secrets du corps humain,consulté durant 2012. http://www.lecorpshumain.fr/corpshumain/motcle-cerveau.html

[4] Exploration du cerveau, consulté durant 2012. http://www.syti.net/Cerveau.html

[5] Le cerveau à tous les niveaux. Broca, Wernicke et les autres aires du langage, consulté durant 2012. http://lecerveau.mcgill.ca/flash/i/i_10/i_10_cr/i_10_cr_lan/i_10_cr_lan.html

[6] Phonétique. consulté durant 2012. http://www.ph-ludwigsburg.de/html/2b-frnz-s-01/overmann/baf3/phon/3k.htm

[7] Larynx, consulté durant 2012. http://fr.wikipedia.org/wiki/Larynx

[8] C. Gabriel, Cours de Claude Gabriel, Chapitre 9 : production de la parole et voix humaine, Haute-Ecole Libre de Bruxelles. consulté durant 2012. http://www.claudegabriel.be/Cine%20acoustique%209.pdf

[9] Dictionnaire visuel, La parole. consulté durant 2012. http://www.ikonet.com/fr/ledictionnairevisuel/static/qc/la_parole

[10] E-book de la sonorisation, Définition du son. consulté durant 2012. http://www.sonorisation-spectacle.org/definition-du-son.html

[11] Le Système Auditif Humain. consulté durant 2012. http://outilsrecherche.overblog.com/pages/Notes_111_Le_Systeme_Auditif_Humain-3080878.html

[12] Le son et l'audition: Fonctionnement de l'oreille. consulté durant 2012. http://anso.pagesperso-orange.fr/page_le_fonctionnement.htm

[13] C. Gabriel, Cours de Claude Gabriel, Chapitre 2 : notions d'acoustique physique. consulté durant 2012. http://www.claudegabriel.be/

[14] Vitesse du son, consulté durant 2012. http://fr.wikipedia.org/wiki/Vitesse_du_son

[15] I. Magrin – Chagnolleau, Le traitement automatique de la parole, Comment reproduire les processus physiologiques et cognitifs humains? Laboratoire Dynamique du Langage, CNRS, Lyon, France consulté durant 2012. http://perso.telecom-paristech.fr/~chollet/Biblio/Cours/Parole/IMC/ScCo.ppt

[16] P. Truillet, Interaction vocale, des modèles à l'interaction, 2012. consulté durant

2012. http://www.irit.fr/~Philippe.Truillet/

[17] Les sensations d'intensité, de hauteur, de timbre. ***consulté durant 2012.*** http://www.phaz.mc/edu/acoustique/sensations.html

[18] E. Marsico et B. Martinie, Cours de phonétique, 2ème année, Université Lumière Lyon 2, France, consulté durant 2012. http://lesla.univ-lyon2.fr/sites/lesla/IMG/pdf/doc-251.pdf

[19] Antiformant, consulté durant 2012. http://www.glottopedia.de/index.php/Antiformant

[20] C. Meunier, Phonétique acoustique. consulté durant 2012. http://aune.lpl.univ-aix.fr/~meunier/publi/neurophysio_ch13.pdf

[21] P. Delattre, Le jeu des transitions de formants et la perception des consonnes. http://www.haskins.yale.edu/Reprints/HL0040.pdf

[22] A. Ounnas, Synthèse de la parole en Arabe Standard, Mémoire de Magister, Ecole Nationale Polytechnique, Alger, 2012.

[23] T. Dutoit, Introduction au traitement automatique de la Parole, Notes de cours/ DEC2, Faculté Polytechnique de Mons, Belgique, 2000.

[24] M.J. Lebœuf, Reconnaissance et synthèse de parole principales applications, 2000, consulté durant 2012. http://www.esi.umontreal.ca/~leboeufm/blt6134/application.html

[25] La Reconnaissance Automatique de la Parole. consulté durant 2012. http://fr.wikipedia.org/wiki/Reconnaissance_vocale

[26] Historique de la reconnaissance vocale. consulté durant 2012. http://membres.multimania.fr/guillaumerey/reconnaissance_historique.htm

[27] J. Ohala, Christian Gottlieb Kratzenstein: Pioneer in speech synthesis, University of California, Berkeley, USA, 2011. consulté durant 2012. http://www.icphs2011.hk/resources/OnlineProceedings/SpecialSession/Session7/Ohala/Ohala.pdf

[28] Talking Heads: Simulacra Kratzenstein's resonators. consulté durant 2012. http://www.haskins.yale.edu/featured/heads/SIMULACRA/kratzenstein.html

[29] History of speech synthesis, 1770 – 1970: Wolfgang von Kempelen's speaking machine and its successors. consulté durant 2012. http://www2.ling.su.se/staff/hartmut/kemplne.htm

[30] Charles Wheatsone's refinements of Von Kempelen's talking machine, late 1800's. http://www.haskins.yale.edu/featured/heads/SIMULACRA/wheatstone.html

[31] Joseph Faber's Talking Euphonia, Irrational Geographic, 2009. consulté durant 2012. http://irrationalgeographic.wordpress.com/2009/06/24/joseph-fabers-talking-euphonia/

[32] R. R. Riesz's talking mechanism, 1937. consulté durant 2012. http://www.haskins.yale.edu/featured/heads/simulacra/riesz.html

[33] The Voder (1939). consulté durant 2012. http://www.haskins.yale.edu/featured/heads/SIMULACRA/voder.html

[34] J. Yamagishi, New and emerging applications of speech synthesis, Séminaire, The Centre for Speech Technology Research, University of Edinburgh, UK, 2011.

[35] S. Ouni, Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico – articulatoire, Thèse de Doctorat, Ecole doctorale IAEM Lorraine, Université Henri Poincaré - Nancy1, France, 2001.

[36] M. Aron, Acquisition et modélisation de données articulatoires dans un contexte Multimodal, Thèse de Doctorat, Ecole doctorale IAEM Lorraine, Université Henri Poincaré – Nancy 1, France, 2009.

[37] E. Moulines, O. Cappé, Synthèse de la parole à partir du texte, article, Techniques de l'ingénieur, Référence H1960, France, 1996.

[38] X. Huang, A. Acero, H. Hon, Spoken language processing: a guide to theory, algorithm, and system development, Ed. Prentice Hall, New Jersey, USA, 2001.

[39] S. Djeghiour, Application des réseaux de neurones à la synthèse de la parole en Arabe Standard, Mémoire de Magister, Ecole Normale Supérieure des Sciences Humaines, Alger, 2011.

[40] G. Richard, Traitement du Signal pour la parole, Ecole Nationale Supérieure des Télécommunications, Ecole d'été – Cargèse, France. consulté durant 2012. http://perso.telecom-paristech.fr/~chollet/Biblio/Cours/Parole/Cargese/Cours_VPL_Richard.pdf

[41] G. Peeters, Analyse et synthèse des sons musicaux par la méthode PSOLA, Article, Actes JIM98 (Journées d'Informatique Musicale), Agelonde, France, 1998.

[42] L. Balthasar & D. Valero, Transcription avec Praat – Mode d'emploi, ICAR, CNRS – Lyon 2 – ENS, France, 2005.

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي
جامعة الجزائر 2
كلية الآداب و اللغات
قسم علوم اللغة

تقديم الطالبة :
بن سماعيل دليلة
(مهندس دولة في الإلكترونيات)

مذكرة مقدمة لنيل شهادة الماجستير في علوم اللسان و التبليغ اللغوي

التخصص: العلاج الآلي للكلام

العنوان

تركيب الكلام الاصطناعي باللغة العربية بواسطة الجمل و الكلمات المتسلسلة

اللجنة المناقشة :
د. نجية بن بليدية (ئيسا)
د. مهنية قرتي (مقرارا)
د.محمد عيسيو (عضوا)

- أكتوبر2014 -