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

Université Saâd Dahlab - Blida
USDB

Faculté des sciences
Département d'Informatique

Mémoire pour l'obtention
d'un diplôme d'Ingénieur d'Etat en Informatique.
Option : Systèmes d'Informations

Sujet :

Elaboration d'une Base de Données des Sons de l'Arabe Standard

Présenté par : ZHOR BENDRAOUA
FATIMA BANDOU

Promotrice : Dr M.GUERTI

Organisme d'accueil : ENP ALGER

Soutenu le : 15 Octobre 2006, devant le jury composé de :

M^elle	N. BENBLIDIA	CC	USDB BLIDA	**Présidente**
Mme	M. GUERTI	MC	ENP ALGER	**Promotrice**
M^elle	F.Z REGUIEG	CC	USDB BLIDA	**Examinatrice**
M	Z. BENSLAMA	CC	USDB BLIDA	**Examinateur**

- N° / Promotion 2006

Dédicaces

Je dédie ce travail en signe de reconnaissance à mes très chers parents qui ont tout fait pour me soutenir et encourager dans mes études.

A mes sœurs, et mes frères.

A tout membre de ma famille.

A ma promotrice M. Guerti.

A mon binôme F. Bandou et sa famille.

A tous mes amis.

BENDRAOUA ZHOR

Dédicaces

Ce mémoire est dédié à mes parents, qui m'ont toujours poussé et motivé dans mes études.

A mes frères abdelrahim, abdelmadjid et Nadjib.

A mes sœurs Safia et Amel.

A mon binôme Z. Bendraoua et sa famille.

A ma promotrice M.Guerti.

A tous mes amis.

BANDOU FATIMA

Résumé :

Notre travail porte sur l'élaboration d'une base de données des sons de l'Arabe Standard utile pour le traitement automatique de la langue arabe. L'élaboration d'une telle base de données nécessite l'enregistrement d'un corpus, ce dernier est constitué de 76 phrases et 15 mots en Arabe Standard. Une étape de segmentation en unités acoustiques basée sur une analyse sonagraphique a l'aide de l'outil d'analyse PRAAT a été appliquée afin d'extraire les paramètres pertinents du signal vocal et les sauvegarder dans cette base de données.

Mots Clés : segmentation, analyse sonagraphique, base de données, Arabe Standard.

Abstract:

Our research is the elaboration of database for sounds of the standard Arabic, used for automatic speech processing. The elaboration of this database needs developpement of corpus, this one is constitude by 76 sentences and 15 isolated words in standard Arabic, a step of segmentation to acoustic unites based on the sonagraphic analysis with the programm PRAAT is necessary in order to extract pertinent parameters that constitute the vocal signal.

Key words :segmentation, sonagraphic analysis, database, standard arabic.

ملخص:

يدور عملنا حول انجاز قاعدة معلومات لأصوات اللغة العربية الفصحىالمستعملة فى المعالجة الالية للغة العربية.لانجاز هذه القاعدة يستلزم تسجيل مدونة .هذه الأخيرة مكونة من 76 جملة و 15 كلمة باللغة العربية الفصحى.

مرحلة التقسيم الى وحدات صوتية والتى تعتمد على التحليل المطيافي وهذا باستعمال برنامج PRAAT كانت مطبقة و هذا لاستخراج المميزات المحتملة للاشارة الصوتية و من ثم تسجيلها داخل قاعدة المعلومات.

كلمات المفاتيح: تقسيم، التحليل المطيافى قاعدة معلومات، اللغة العربية الفصحى.

Remerciements

Nous remercions avant tout ALLAH, le généreux, l'exalté de nous avoir guidées dans la réalisation de ce PFE.

Nous remercions tout particulièrement M. GUERTI Maître de conférence à l'ENP Alger pour avoir accepté et dirigé ce PFE et pour ses conseils et ses recommandations tout au long de ce projet.

Nous remercions aussi Mme la présidente ainsi que les membre du jury qui ont bien voulu nous faire l'honneur d'examiner ce travail.

Nous remercions Mr I. Mezaourou pour son aide précieuse.

Nos remerciements vont également à nos enseignants, qui nous ont enseigné durant toutes nos études.

Liste des figures

Liste des tableaux

Liste des abréviations

TAP : Le Traitement Automatique de la Parole

TOP : Transcription Orthographique et Phonétique

AS : Arabe Standard

Fo : Fréquence Fondamentale

SQL : Structured Querry Language

RAP : La Reconnaissance Automatique de la Parole

DAP : Décodage Acoustico- Phonétique

HMM : Hidden Markov Model

SGBD : un Système de Gestion de Base de Données

DBMS : Data Base Management System

BDOO : Base de Données Orientée Objet

BD : Base de Données

MCD : Modèle Conceptuel de Données

VB : Visual Basic

Sommaire

Sommaire

Chapitre 2 : Segmentation et Etiquetage de la parole

Chapitre 3 : Généralités sur les Bases de Données

Chapitre 4 : Elaboration de BDSONARABE

Introduction Générale

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

L'importance particulière du TAP dans le cadre le plus général s'explique par la position privilégiée de la parole comme vecteur d'informations dans la société humaine. La production de la parole comme moyen de communication nécessite deux mécanismes l'audition et la perception, à partir de ces deux mécanismes nous pouvons extraire une représentation sonagraphique et ainsi les paramètres pertinents du signal de la parole par l'intervention d'un outil d'analyse.

La nécessité de disposer des bases de données sonores a émergé il y a une dizaine d'années, sous la poussée des méthodes de Reconnaissance Automatique de la Parole (RAP) fondées sur l'apprentissage. Dans le milieu des années 80, des travaux ont été coordonnés en France autour de BDSONS, et prolongés dans des actions européennes (projets SAM I et SAM II). A l'heure actuelle, le besoin de données sonores reste encore une préoccupation essentielle.

Notre travail s'inscrit dans le cadre de l'élaboration d'une base de données des sons de l'Arabe Standard (BDSONSARABE), utile au Traitement Automatique de la langue Arabe, elle sera utilisée pour l'analyse des phénomènes acoustico-phonétiques, l'apprentissage, et l'évaluation des systèmes de RAP. La conception de ce type de système nécessite des conditions à respecter :
l'enregistrement d'un corpus de parole en Arabe Standard ;
afin d'extraire les paramètres pertinents nécessaires pour l'élaboration de notre BD, nous avons opté à une étape de segmentation et d'étiquetage manuel à l'aide du logiciel d'analyse PRAAT. La segmentation est une étape très délicate à cause de problèmes de variabilité du signal, et de la coarticulation .

Dans notre étude nous avons utilisé un corpus contenant 76 phrases et 15 mots en Arabe Standard prononcé par un locuteur jordanien, dans un milieu calme,

et comportant les sons dans les différentes positions (Initiale, Médiane, Finale) et les différentes voyelles (fatha, damma, kasra).

Nous pouvons présenter notre PFE dans quatre chapitres :

- dans le premier Chapitre nous commençons par définir les deux mécanismes phonatoire et auditif de l'être humain, ensuite nous présentons les notions fondamentales et les classes des sons de l'AS ;
- dans le second nous allons parler des systèmes de RAP, nous présentons quelques outils d'analyses, nous illustrons la notion des unités acoustiques et de segmentation et étiquetage, ainsi que les modes de segmentations et d'étiquetage connus ;
- le Troisième expose les BD en générale sans oublier de donner des notions sur les BD sonores comme TIMIT et BDSONS ;
- le dernier concerne l'application de notre travail ;

Enfin des conclusions générales et des perspectives finissent ce travail.

Chapitre 1 : Notions fondamentales sur la parole et l'AS

1.1. Introduction

Le Traitement Automatique de la Parole (TAP) est une discipline qui associe étroitement linguistes et informaticiens .Il repose sur la linguistique, les formalismes (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'informatique.

Ce chapitre a pour but de présenter les notions élémentaires de la parole et de l'Arabe Standard (AS). Nous y exposerons tout d'abord le fonctionnement des appareils phonatoire et auditif de l'être humain. Nous présenterons ensuite les propriétés spécifiques du signal vocal, la phonétique et la phonologie de l'AS, et on termine par la Transcription Orthographique et Phonétique (TOP) des consonnes de l'AS.

1.2. Production de la parole

L'appareil phonatoire nous permet de produire des sons très variés dans un espace fréquentiel et énergétique pourtant limité. L'appareil phonatoire humain a été la base de recherches visant à simuler mécaniquement ses capacités, recherches ayant permis, en retour, de mieux comprendre son fonctionnement.

1.2.1. Description de l'appareil phonatoire

L'appareil vocal humain peut être comparé à la fois à un instrument de musique à vent et à cordes. Il comprend une source de vent, les poumons; une structure qui vibre, les cordes vocales dans le larynx; et une série de caisses de résonance que forme le pharynx, la bouche et les fosses nasales (Fig.1.1).

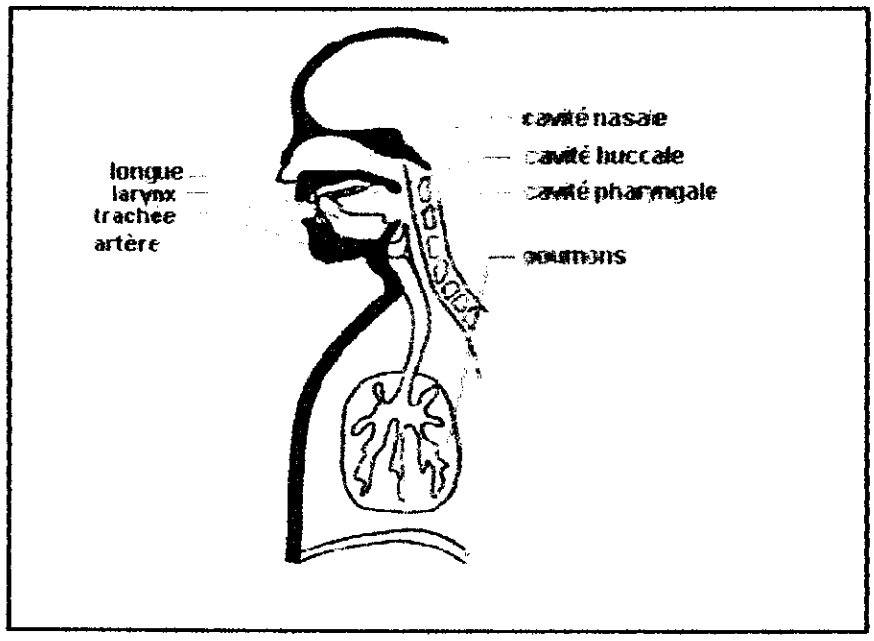


Figure 1.1 : L'appareil phonatoire humain [1]

1.2.2. Fonctionnement de l'appareil phonatoire

L'air pulmonaire, que l'on pourrait définir comme le " générateur ". Quand on parle, les phases d'inspiration de notre respiration deviennent plus rapides et plus courtes. On respire également davantage par la bouche, alors que l'inspiration est normalement exclusivement nasale. Du côté de l'expiration, le volume et la pression de l'air expiré sont augmentés pour pouvoir faire vibrer les cordes vocales situées dans le larynx. Le larynx se compose d'une série de muscles et de cartilages plus ou moins mobiles, il peut être relevé ou abaissé. La Cavité nasale, lors de la déglutition, le larynx s'élève tandis que l'épiglotte, cette lame cartilagineuse située à l'entrée du larynx, se rabat vers l'arrière.

Lorsqu'on parle, l'air expulsé des poumons emprunte la trachée avant d'arriver dans le larynx où il va rencontrer les cordes vocales. Celles-ci sont en fait une paire de muscles et de ligaments de 20 à 25 millimètres de long et recouverts d'une muqueuse. C'est la seconde composante de l'appareil phonatoire : le " vibrateur ".

Sous la pression de l'air expiré, les cordes vocales s'écartent, puis se referment aussitôt, entraînant à nouveau une hausse de la pression sous la glotte. En ouvrant et fermant la glotte lors de la phonation, les cordes vocales libèrent de façon saccadée l'air emmagasiné dans les poumons.

Au cours d'une phrase, le locuteur modifie ainsi plusieurs fois la fréquence de vibrations des cordes vocales pour produire les vibrations acoustiques correspondant à différents sons. Mais ces sons ne constituent pas encore des mots, ils doivent être sculptés par le reste de l'appareil phonatoire. La première transformation du son se fait dans la cavité du pharynx, le carrefour où se croisent les voies respiratoires et digestives. Le pharynx et les différentes cavités avec lesquelles il communique (les cavités pharyngales, nasales, buccales, et labiales) jouent le rôle de "résonateurs ou d'amplificateurs" qui modulent les sons émis au niveau des cordes vocales. Certaines fréquences seront amplifiées, d'autres atténuées.

1.3. Audition et Perception

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur. Nous allons maintenant présenter l'anatomie du système auditif humain, qui est récepteur de l'information sonore, et les capacités de perception qui le caractérise lorsqu'il est en parfait état.

1.3.1. Anatomie du système auditif

On distingue trois parties dans l'appareil auditif humain (figure1.2).

- l'oreille externe inclut la partie visible et le conduit auditif. Elle collecte les sons, les dirige vers le tympan, et ensuite vers les parties moyennes et internes de l'oreille ;
- l'oreille moyenne est un espace rempli d'air, constitué de trois osselets : le marteau, et l'étirer, qui permettent une adaptation de la transmission du signal de parole entre l'air ambiant et le liquide de l'oreille interne ;
- l'oreille interne abrite une structure en forme de coquille d'escargot, appelée la cochlée, qui contient l'organe de l'audition à proprement parler. Bien que la cochlée soit approximativement de la taille d'un petit pois, sa structure est très complexe. Elle est protégée par l'os temporal, l'os le plus dur du crâne.

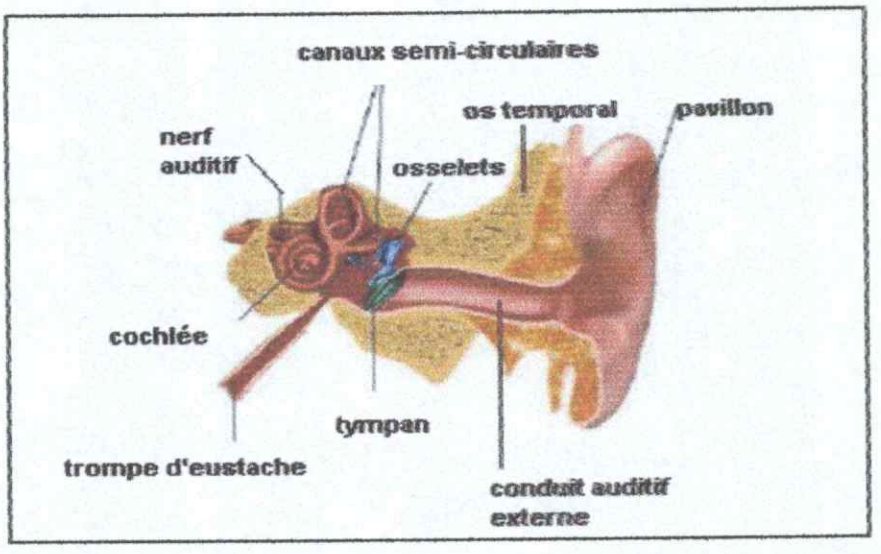


Figure 1.2 : L'appareil auditif humain [2].

1.3.2. Fonctionnement de l'oreille humaine

Le son, vibration de l'air, frappe le pavillon de l'oreille qui l'amplifie et le concentre vers l'orifice du conduit auditif externe. Le son chemine dans le conduit et vient frapper le tympan qui se met à vibrer à la façon d'une peau de tambour.

Cette vibration traverse l'oreille moyenne. Elle est transmise par la chaîne des osselets sur la membrane de la fenêtre ovale qui donne sur l'oreille interne. La membrane de la fenêtre ovale étant bien plus petite que celle du tympan, sa vibration sera plus intense.

La répercussion atteint le liquide de la cochlée qui, dans ses mouvements, va mobiliser les cils des cellules auditives, réceptrices des vibrations. Excitées, celles-ci vont émettre des signaux codés, sous forme de petits courants électriques acheminés vers le cerveau par le nerf auditif. Le cerveau va intégrer les messages arrivant des deux oreilles, procéder au décodage et analyser l'ensemble. L'individu entend, localise, comprend le son. Savoir d'où vient un bruit, séparer des bruits mélangés (comme comprendre un interlocuteur au milieu d'autres conversations) nécessitent le fonctionnement simultané des deux oreilles en bon état.

1.4. Propriétés spécifiques du signal vocal

La Reconnaissance Automatique de la Parole (RAP) est un domaine de la science qui traite la communication entre l'homme et la machine et ayant toujours eu un grand attrait auprès des chercheurs comme auprès du grand public.

La grande difficulté de la reconnaissance automatique de la parole provient du caractère du processus de la communication parlée et des caractéristiques du signal vocal [3].

1.4.1 Continuité

Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont l'occurrence est aléatoire. Il peut très bien y avoir des intervalles de silences au milieu d'un mot et aucun intervalle entre deux mots successifs. Par conséquent, il est très difficile de déterminer le début et la fin des mots composant la phrase [4].

1.4.2. Variabilité

Le terme de variabilité, qui est assez générique, peut englober plusieurs problèmes totalement indépendants du point de vue des techniques actuellement utilisées pour les résoudre. Il est ainsi possible d'isoler une variabilité du signal de parole relativement aux classes phonétiques .Il est aussi possible d'isoler la variabilité de l'environnement sonore d'un système de reconnaissance. À un niveau beaucoup plus abstrait, celui de la sémantique, il est également possible de parler de variabilité, certaines phrases ne pouvant pas être comprises lorsqu'elles sont considérées hors contexte, imposant ainsi de définir des mécanismes de gestion de l'historique du dialogue.

Nous allons maintenant voir les problèmes directement liés à la parole. Ceux-ci sont relatifs à la différence innée de prononciation vis-à-vis d'un ou plusieurs locuteurs [5].

- La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humour ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation. L'autre type de variabilité intra-locuteur lié à la phase de production de parole ou de préparation à la production de parole. Cette variation est due aux phénomènes de coarticulation ;

- La variabilité inter-locuteur est un phénomène majeur en reconnaissance de la parole, un locuteur reste identifiable par le timbre de sa voix malgré une variabilité qui peut parfois être importante. La contrepartie de cette possibilité d'identification à la voix d'un individu est l'obligation de donner aux différents sons de la parole une définition assez souple pour établir une classification phonétique commune à plusieurs personnes. La cause principale des différences inter-locuteur est de nature physiologique. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une

fréquence de base, le fondamental. Cette fréquence de base sera différente d'un individu à l'autre, une voix d'homme étant plus grave qu'une voix de femme ;

- La variabilité due à l'environnement peut, parfois, être considérée comme une variabilité intra-locuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution.

1.5. Les ressources de connaissance dans les systèmes de RAP

Le système de la reconnaissance automatique de la parole (RAP) dispose des ressources de connaissance suivantes : la phonétique, la phonologie, la prosodie, le lexique, la syntaxe, la sémantique et la pragmatique.

L'étude phonétique d'une langue peut se faire sans le sens ou la signification (sémantique). A la limite on pourrait étudier les caractéristiques phonétiques d'une langue qu'on ne comprenait même pas, Par contre, la phonologie s'occupe de la fonction des sons dans la transmission d'un message [2].

1.5.1. La phonétique

La phonétique est la science des sons langagiers tels qu'ils existent dans la réalité. Cette science peut être abordée sous trois aspects différents: la production du son (phonétique articulatoire), la transmission des sons par les airs (phonétique acoustique) et la réception de ces sons par l'oreille de l'interlocuteur (phonétique auditive). La phonétique auditive est rarement étudiée sauf pour l'élaboration de traitements orthophoniques. La phonétique acoustique permet une description précise des sons, mais la variation qui survient est telle qu'il est plus aisé de décrire les sons articulatoirement puis d'en vérifier la structure acoustique.

1.5.2. La phonologie

La phonologie a comme objectif d'étudier les variantes phonétiques contextuelles. En reconnaissance de la parole, la phonologie regroupe l'ensemble des modules de traitement des altérations possibles d'un phonème (unité minimale de son) ou d'un mot dans un contexte donné [3].

La phonologie regroupe trois types d'informations :

- les altérations phonologiques dans le mot, (variantes de prononciation) ;
- les altérations phonologiques dues aux flexions en fin de mot (conjugaisons

des verbes, pluriels des noms et des adjectifs) ;

- les altérations qui apparaissent à la jonction de deux mots (liaison) ;

1.5.3. La prosodie

La prosodie est une branche de la linguistique de l'expression (phonétique et phonologie) qui s'attache à la description et à la représentation physique et formelle des éléments phoniques systématiques du langage différents des phonèmes tels que l'accent, l'intonation, dont la manifestation concrète est associée à des variations de la F_0, de la durée et de l'intensité (paramètres prosodiques physiques) qui sont perçues comme des changements de hauteur ou de mélodie, de longueur et de volume sonore (paramètres prosodiques subjectifs) [6].

1.6. L'étude de la prosodie

Le terme prosodie recèle des notions différentes selon le point de vue adopté pour son étude. Du point de vue acoustique, la prosodie se définit au moyen des paramètres de la fréquence du fondamental (estimation du son laryngien à un instant donné sur le signal), de la durée et de l'intensité. Du point de vue de la perception de la parole, elle concerne l'étude des phénomènes de l'accentuation et de l'intonation (variation de hauteur, de rythme et d'intensité) permettant de véhiculer de l'information liée au sens de la phrase [4]. Toutefois, il est difficile d'établir une correspondance directe entre les paramètres physiques et les corrélats perceptifs.

1.6.1. Fréquence Fondamentale (F_0)

La fréquence fondamentale ou fréquence laryngienne (notée F_0) représente la fréquence de vibrations des cordes vocales. Son estimation est liée à la localisation de portions voisées sur le signal de parole, les sons non voisés ayant une fréquence nulle. Les algorithmes d'extraction de F_0 peuvent être de type temporel ou fréquentiel. Les premiers se basent directement sur la description temporelle du signal pour le calcul de F_0 (F_0=1/Période), alors que les seconds s'appuient sur les fréquences des harmoniques (fréquence de résonance) qui peuvent être représentés graphiquement sur un spectrogramme.

La figure (Fig. 1.4) représente une description temporelle (fenêtre du haut) et spectrale (fenêtre du bas) de la phrase الدرس العاشر / addarsul AOir/ (le dixième cours). Les trames blanches correspondant aux harmoniques sur le spectrogramme désignent les portions voisées (voyelles /[a]/ /[u]/ /[i]/) du signal acoustique.

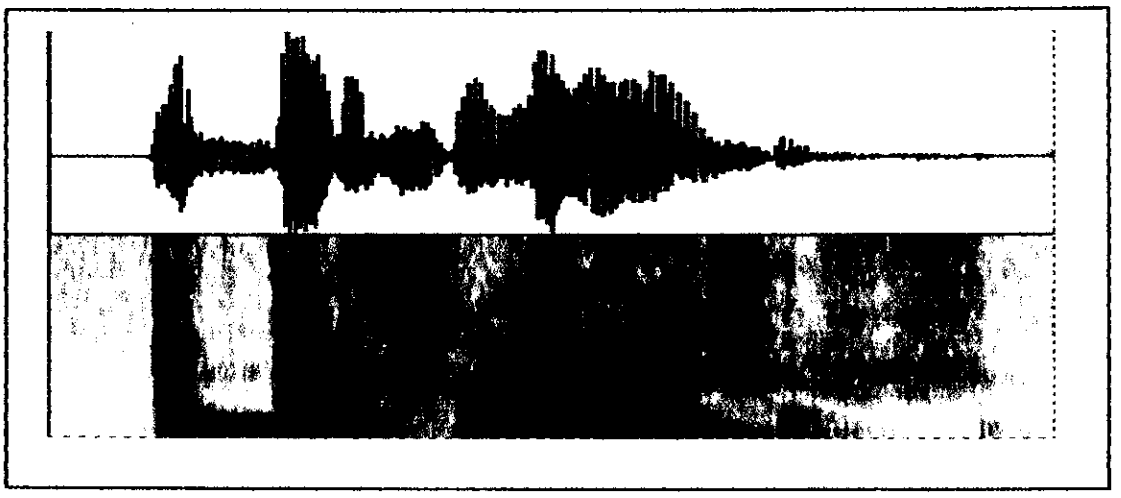


Figure 1.3 : Représentations temporelle et fréquentielle de la phrase /addarsul AOir/.

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

$$F = 6 \times ech \times \frac{\ln(\frac{F2}{F1})}{\ln 2} \qquad (1.1)$$

Dans cette formule, si *ech*=2 alors F est exprimé en demi-ton, si *ech*=4, F est exprimé en quart de ton, etc.

La fréquence fondamentale s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants.

1.6.2. L'intensité

Résultant de la pression sous-glottique, l'intensité est le paramètre prosodique le plus simple à calculer. Elle est mesurée sur des portions de signal allant de 5 à 10 ms (énergie à court terme) et exprimée en décibels (db) pour respecter l'échelle perceptive. Sa formule est la suivante :

$$E_{db} = 10 \times \log_{10}\left(\sum_{i=A}^{I} S_i^2\right) \qquad (1.2)$$

Ce paramètre est le plus souvent négligé en génération de la prosodie. Bien que considéré comme dépendant de F_0, ces deux paramètres peuvent varier indépendamment dans l'interrogation.

1.6.3. La durée

Pour beaucoup d'auteurs, le paramètre de la durée est difficile à calculer car il ne dépend d'aucun corrélat biologique, contrairement à F0 et à l'intensité (qui dépendent respectivement de la tension des cordes vocales et de la pression sous-glottique). Pour calculer la durée d'un phénomène, il faudrait se fixer deux événements qui délimitent ses repères initial et final. Sur un signal de parole.

La durée peut être corrélée à une multitude de facteurs complexes de nature linguistique (accent, position des mots dans la phrase, catégorie grammaticale, etc...) et extra-linguistique (débit de parole, expressivité, etc.). Certains d'entre eux peuvent être privilégiés par rapport à d'autres selon le type de corpus d'analyse et le style de lecture employés, il existe l'analyse des :

- logatomes dans des phrases porteuses, dont le but est d'uniformiser l'environnement linguistique (syntaxique, sémantique...), ne rend compte que des phénomènes intra-mots (phonétique, phonologique, etc.) ;
- phrases rend compte de phénomènes d'interaction entre les mots (syntaxique, sémantique, etc.). La difficulté ici est de faire la part des choses entre ce qui relève du contexte phonétique, syntaxique, sémantique, etc. De plus, les proportions d'influence des facteurs en question peuvent être de degrés divers ;
- corpus de lecture spontanée rend compte de phénomènes liés à l'hésitation, etc.
- Enfin, plusieurs lectures de locuteurs différents rendent compte des variabilités individuelles (physiologique, régionale, etc.) par rapport aux autres variabilités [4].

1.6.4. Les formants

Les cordes vocales produisent l'énergie sonore qui est résonnée et filtrée par des cavités de résonance. Lors de la phonation, certaines fréquences du son produit par les cordes vocales sont amplifiées par les cavités de résonance en fonction de la fréquence de résonance propre à chaque cavité. Les fréquences qui sont amplifiées sont celles qui sont voisines de la fréquence caractéristique de la cavité, les autres fréquences étant affaiblies. Ce sont les fréquences renforcées que l'on nomme « formants ».

Les formants permettent de discriminer des voyelles ayant la même F_0, la même amplitude et la même durée. Les formants (abréviation F1, F2, F3, F4) permettent en fait

d'identifier le timbre des sons. Les valeurs en Hertz des quatre premiers formants des sons humains sont particulièrement significatives.

1.7. Le système phonétique de l'Arabe Standard (AS)

L'Arabe Moderne ou l'Arabe Standard est, la langue de communication commune à l'ensemble du monde arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel. La langue arabe appartient à la famille des langues sémitiques. L'étude de la grammaire arabe a commencé très tôt au milieu du 11^{éme} siècle de l'hégire et a donné lieu à d'énormes productions, avant de connaître une période de stagnation qui a duré plusieurs siècles. Ces dernières années, elle connaît un regain d'intérêt, entre autres dans le domaine du traitement automatique.

1.7.1. Phonétique et Phonologie de la langue arabe

Nous présentons ci-dessous certaines caractéristiques phonétiques de l'AS.

- **Le système vocalique**

Le système vocalique comprend trois voyelles brèves /[a]/[u]/[i]/, et trois voyelles longues /[A]/[U]/[I]/ qui s'opposent aux précédentes par une durée plus importante sur le plan temporel. L'ensemble des voyelles brèves et longues est dit oral car elles sont émises sans l'intervention de la cavité nasale. Elles sont généralement classées selon le degré d'ouverture du conduit vocal (ouvert /[a]/, fermé /[u]/, /[i]/) et sa position de constriction (/[i]/ antérieure, /[u]/ postérieure) [6].

Ces voyelles peuvent avoir des timbres différents selon leur contexte d'apparition :

- dans un contexte emphatique (au contact des consonnes ص/[S]/, ض/[D]/, ط/[T]/, ظ/[Z]/), le point d'articulation des voyelles est reporté à l'arrière ;
- après les consonnes labiales م/[m]/ et ب/[b]/, les voyelles sont plus arrondies et se rapprochent du phonème /[u]/.

- **Le système consonantique**

L'arabe standard contient 28 consonnes qui correspondent chacune à un phonème, les consonnes de l'arabe sont classées selon leur mode d'articulation (occlusif, fricatif, nasal, glissant ou liquide), leur lieu d'articulation (labial, dental ou vélo-palatal) et leur voisement (sonore ou sourd). Nous proposons de les grouper en fonction de leurs équivalences dans les autres langues :

• Les phonèmes spécifiques à l'arabe qui n'ont pas d'équivalent dans les langues européennes.ظ /[Z]/, ط/[T]/, ض/[D]/, ص/[S]/, ح/[H]/, ق/[q]/, ع/[ε]/.

• Les phonèmes qui ont des équivalents dans la langue française : ت/[t]/, ز/[z]/, د/[d]/, س/[s]/, ش/[O]/, غ/[G]/, ك/[k]/, ج/[j]/, ف/[f]/, ب/[b]/, ل/[l]/, م/[m]/, ن/[n]/, و/[w]/, ي/[y]/.

• Les phonèmes qui ont des équivalents dans plusieurs langues telles que l'espagnol, l'allemand ou l'anglais :ر/[r]/, ذ/[v]/, ه/[h]/.

1.7.2. particularités phonologiques

Les caractéristiques phonologiques de l'arabe sont l'emphase, la gémination et le madd :

- l'emphase est habituellement utilisée pour rendre compte des manifestations prosodiques liées à l'accentuation volontaire d'une syllabe, les consonnes ظ /[Z]/, ط/[T]/, ض/[D]/, ص/[S]/) sont dites emphatiques, Certaines des études affirment que le phénomène de l'emphase dépasse le cadre de la voyelle (ou des voyelles) adjacente(s) et se propage aux phonèmes voisins comme dans le mot $[C_1V_1C_2V_2...]$ ([C]=consonne,[V]=voyelle), si $[C_1]$ est emphatique, alors la synthèse est plus naturelle quand la propagation de l'emphase arrive jusqu'à$[C_2]$. En revanche, il existe des divergences sur la portée de cette propagation, en d'autres termes, sur la taille du segment sonore affecté par la consonne emphatique ;

- la gémination est symbolisée par le signe de la chadda qui signifie le dédoublement de la consonne. Une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position (les lèvres ne se referment pas après le premier /b/ dans /kabbara/), d'où la transcription /kab:ara/ qui est plus appropriée Dans beaucoup de langues, ce phénomène permet de mettre en relief un mot dans son contexte, alors qu'il s'avère être un élément distinctif sur les plans morpho-sémantiques en langue arabe :

حضر/haDara/ (il a assisté) est différente de حضّر /haDDara/ (il a préparé) – la deuxième consonne est géminée.

- le madd concerne l'allongement des voyelles. Il est provoqué par la présence d'une voyelle longue (و /[U]/, ا/[A]/ ou ي/[I]/) La lecture de textes arabes est régie par des règles phonologiques qui ont trait à la contraction des sons, leur élision et à l'assimilation homo-organique des nasales. Certaines de ces règles sont obligatoires, d'autres facultatives ou réservées à certains types de textes, comme le Coran. Exemple le mot نام/ nAma, la voyelle [A] représente le madd dans le mot.

1.8. Classification des sons

La taxonomie des sons est définie de deux manières, grâce à la phonétique et à la phonologie. Alors que la phonétique peut être considérée comme véritablement

descriptive, associant chaque son de la langue à un symbole et à une classe, la phonologie s'intéresse, elle, à la description des interdépendances entre sons et au codage effectif des mots du langage lors du processus d'oralisation. La phonologie essaie donc plus particulièrement d'expliquer les différences qui peuvent exister entre la transcription phonétique d'un mot du langage et la transcription phonétique exacte du mot qui est effectivement prononcé.

- **Description des voyelles**

Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle [4].
Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais se sont principalement les formants en basses fréquences qui caractérisent les voyelles.

- **Description des consonnes**

Les consonnes sont caractérisées par la présence de bruits sans définition périodique précise [5].
On distingue deux types de consonnes :

- Les occlusives

Les phonèmes de cette classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont donc constituées de deux parties successives, une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion, au moment du relâchement.
Les occlusives peuvent être voisées, à la manière des voyelles, ou sourdes, c'est-à-dire non voisées. Les occlusives voisées peuvent également être appelées occlusives sonores. Il existe deux types d'occlusives : nasales et orales.

-Les occlusives nasales sont produites avec la participation de la cavité nasale ;

-Les orales sont produites avec le velum en position relevée, c'est-à-dire que l'air ne passe pas dans les fosses nasales ;

- Les fricatives

Dans cette classe les sons produits sont regroupés par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences, et peut être voisée ou sourde.

Les semi consonnes ou semi voyelles combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes [4].

1.8.1. Modes et lieux d'articulation

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- libre passage, avec mise en vibration, de l'air au niveau de la glotte (sonore ou sourde) ;
- libre passage, ou non, en un point quelconque (le lieu d'articulation) des cavités supra-glottique (voyelles ou consonnes) ;
- passage par une voie unique ou deux voies différentes (orale ou nasale) ;
- passage, dans le conduit buccal, par une voie médiane ou latérale (la plupart des articulations opposées aux latérales).

Le lieu d'articulation est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. Il peut se situé aux endroits suivants :

- les lèvres (articulation labiale ou bilabiale) ;
- les dents (articulation dentales) ;
- les lèvres et les dents (articulation labio-dentale) ;
- les alvéoles (c'est-à-dire les gencives internes des incisives supérieures, articulations alvéolaires) ;
- le palais (vue sa grande surface, on peut distinguer des articulations pré-palatales, médio-palatales et post-palatales) ;
- le voile du palais (palais mou, articulation vélaires) ;
- la luette (articulations dites uvulaires) ;
- le pharynx (articulations pharyngales) ;
- la glotte (articulations glottales).

1.8.2. Transcription Orthographique Phonétique (TOP)

La TOP permet de représenter le texte tel qu'il sera prononcé par le système. La complexité de cette tâche varie selon la langue traitée. Ainsi, la transcription de l'arabe ou de l'espagnol est relativement directe par rapport à celle de la langue française qui présente de nombreuses ambiguïtés de prononciation que seul le contexte syntaxique permet de lever. L'approche utilisée pour la transcription des différentes langues est de

type système expert, hormis l'anglais pour laquelle une approche par analogie est appliquée. Le tableau 1.1 donne la TOP des différentes consonnes de l'AS.

Tab 1.1 : Transcription orthographique et phonétique Des consonnes de l'Arabe Standard [9]

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

1.9. Conclusion

Dans ce chapitre, nous avons passé en revue les principales caractéristiques acoustiques et phonétiques du signal de parole que nous avons essayé de corréler avec les processus de production et l'audition de ce signal.

Nous avons aussi présenté quelques notions de base sur la phonétique et la phonologie de l'AS, et on a fini par la classification des sons de l'AS.

Nous allons parler dans le chapitre suivant du TAP et ses domaines, de l'analyse la segmentation et l'étiquetage du signal de parole.

Chapitre 2 : Analyse, Segmentation et Etiquetage du signal vocal

2.1. Introduction

L'étiquetage et la segmentation de corpus de parole sont deux tâches fondamentales qui sont nécessaires au TAP.

Cette tâche nécessite une grande base de données de parole qui regroupe l'ensemble des unités acoustiques. Pour aboutir à cette fin, il faut utiliser des outils d'analyse du signal vocal.

Dans ce chapitre nous essayons d'illustrer le TAP et ses activités, l'analyse et les outils d'analyse de la parole, la segmentation et l'étiquetage du signal vocal.

2.2. Le Traitement Automatique de la Parole (TAP)

Le Traitement Automatique de la Parole est un domaine de recherche pour lequel un effort important a été consenti au cours des trois dernières décennies. Les problèmes à résoudre sont considérables et de nature fondamentale. De plus, ils sont pluridisciplinaires : traitement du signal, reconnaissance des formes, intelligence artificielle, informatique, phonétique, linguistique, ergonomie, neurosciences interviennent à des degrés divers dans les solutions apportées.

Le TAP recouvre les activités liées à l'analyse de la parole, à son codage à débit variable afin de la stocker ou la transmettre, à sa synthèse, en particulier à partir du texte, à sa reconnaissance et à sa compréhension, soit pour une transcription, suivie éventuellement d'une indexation, soit dans le cadre d'un dialogue personne-système ou entre humains assisté par une machine. Il comprend également la reconnaissance du locuteur et la reconnaissance de la langue parlée. Ces différents traitements peuvent se faire dans un contexte bruyant, ce qui rend le problème encore plus difficile.

2.2.1. L'analyse de la parole

L'étude de l'évolution temporelle et fréquentielle d'un signal de parole permet de mettre en évidence les caractéristiques de ce signal. Cet objectif est atteint grâce aux méthodes modernes de traitement du signal qui permettent de calculer par exemple, la Fréquence fondamentale (F_0), la durée, l'intensité, les formants d'un signal de parole, et le spectrogramme qui représente l'évolution temporelle de ce spectre. L'analyse

acoustique du signal de parole est souvent utilisée dans l'évaluation et la discrimination des voix pathologiques.

2.2.2. La synthèse de la parole

La synthèse de la parole s'occupe de construire des automates capables de transformer des textes ou des concepts en messages vocaux. Les techniques d'analyse utilisées en synthèse proviennent largement de celles mises au point en codage de la parole. La synthèse par concaténation de phonèmes étant impossible puisque les transitions articulatoires transportent souvent seules l'information pertinente, deux méthodes d'assemblage ont vu le jour, la synthèse par règles et la synthèse par diphones. La première modélise les transitions entre phonèmes sous forme de règles, à partir de représentations formantiques. Les paramètres de commande du synthétiseur sont alors définis sur la base de valeurs cibles (peu nombreuses) et de règles (plusieurs centaines, pour une langue) qui tiennent compte des transitions des éléments à synthétiser. La synthèse par diphones (chaque son est analysé en deux phases, exemple: "bal" = [#b] + [ba] + [al] + [l#]) procède au stockage des transitions plutôt qu'à leur modélisation.

2.2.3. Le codage de la parole

Le codage consiste à prendre en compte le signal analogique à des instants discrets du temps que l'on représente par des nombres. Le débit du signal numérisé est fonction de la fréquence d'échantillonnage et du nombre d'éléments binaires nécessaires à la représentation des valeurs discrètes du signal. Au contraire, le décodage consiste à transformer le signal numérisé en signal analogique. Le signal numérisé comporte plusieurs avantages: immunité au bruit, universalité par rapport aux canaux de transmission. Le codage peut être temporel, obtenu par analyse et synthèse, ou mixte.

2.3. La Reconnaissance Automatique de la Parole (RAP)

Les premières applications de reconnaissance vocale ne fonctionnaient que pour reconnaître quelques mots bien définis, on trouve encore couramment ce genre de

logiciel principalement pour le fonctionnement de machines guidées par la voix ou pour les système tels que les serveurs téléphoniques. Grâce à l'avancé de la recherche et aux nouveaux processeurs de plus en plus puissants il a été possible de créer des systèmes fonctionnant avec un vocabulaire très vaste pouvant contenir des centaines de milliers de mots. Actuellement il existe des modes de reconnaissance que l'on utilise en fonction de l'architecture matérielle du système et de son objectif de fonctionnement.

2.3.1. La reconnaissance monolocuteur

Ce mode de fonctionnement est principalement utilisé pour de petits vocabulaires car il est demandé à l'utilisateur de prononcer tous les mots du vocabulaire une ou plusieurs fois afin de les apprendre et de s'adapter à sa prononciation. Cette méthode bien que très contraignante présente de meilleures performances. Il faut cependant noter qu'il est souvent demandé de prononcer les mots plusieurs fois car il est impossible de prononcer un mot deux fois exactement de la même façon, il existe toujours des variations dans le débit d'élocution, l'accentuation ou l'intonation du mot prononcé. Cette phase d'apprentissage permet de créer une image du mot qui nous servira de modèle pour la phase de reconnaissance.

2.3.2. La reconnaissance multilocuteurs

Les méthodes les plus couramment utilisées lors de reconnaissance vocale indépendante du locuteur sont basées sur des classificateurs neuronaux, ce qui explique pourquoi les chercheurs en reconnaissance vocale sont toujours à l'affût d'une avance dans le secteur des réseaux de neurones. La phase d'apprentissage nécessite un grand nombre de locuteurs servant à l'entraînement du système.

Le choix de ces locuteurs doit être fait de manière très précise car ils doivent être représentatifs de la population utilisant le système (rythme, intonation, accent,....) ils doivent alors prononcer tous les mots du vocabulaire à plusieurs reprises chacun. Le type d'algorithme utilisé est bien évidement un classificateur dans la plupart des cas on retrouve cependant quelque fois des mécanismes d'apprentissage statistique.

Ces méthodes permettent des taux de reconnaissance proche de 90%, il est cependant possible d'améliorer encore les performances du système par l'adjonction d'un système d'adaptation au locuteur. Ces systèmes déterminent les modifications nécessaires aux classes existantes pour se rapprocher du locuteur utilisant le logiciel de la RAP.

2.3.3. La reconnaissance des mots enchaînés (dictée continue)

Le domaine de la recherche permettant la dictée continue a donc naturellement commencé très tôt, mais les difficultés sont tout de même importantes, car en plus des problèmes liés à la reconnaissance des mots isolés, d'autres difficultés viennent s'y ajouter comme par exemple le phénomène de la coarticulation (qui fait que certains sons influencent les sons voisins), les liaisons, la recherche de séparation des mots.

2.3.4. Les méthodes utilisées dans la RAP

On distingue usuellement en reconnaissance de la parole l'approche analytique et l'approche globale. La première approche cherche à traiter la parole continue en décomposant le problème, le plus souvent en procédant à un Décodage Acoustico-Phonétique (DAP) exploité par des modules de niveau linguistique. La seconde consiste à identifier globalement un mot ou une phrase en les comparant avec des références enregistrées. La distinction entre globale et analytique a perdu de sa pertinence avec l'introduction des méthodes statistiques à base de modèles de Markov cachés pour la reconnaissance de la parole continue et le traitement de grands vocabulaires;

- **La méthode analytique**

L'approche analytique cherche à résoudre le problème de la parole continue en isolant des unités acoustiques courtes comme les phonèmes, ou les syllabes. Un exemple classique de cette approche est l'analyse par traits, des indices acoustiques sont calculés à partir du signal de parole, ils permettent de faire des hypothèses locales sur certains traits phonétiques, comme le voisement, la nasalisation, le lieu d'articulation ou le degré d'ouverture du conduit vocal. En fonction de ces traits, le signal acoustique est segmenté et une identification phonétique des segments est réalisée.

Le DAP ainsi obtenu est exploité par des modules d'ordre linguistique. Les niveaux lexical, syntaxique ou sémantique utilisent des sources de connaissances spécialisées et sont organisés avec le module acoustique.

Les systèmes analytiques, conçus avec des objectifs ambitieux, sont restés au stade expérimental. Leur faiblesse provient d'un processus de décision trop précoce, à savoir une segmentation préalable à l'identification ou une identification phonétique sans prise en compte des niveaux linguistiques. Les méthodes globales, développées pour la reconnaissance de mots isolés, ne font pas d'hypothèse sur la structure phonétique des mots, ce qui évite une erreur pénalisante au début du traitement.

- **La méthode globale**

Les méthodes globales identifient un mot ou une phrase en les considérant comme des entités élémentaires et en les comparant avec des références enregistrées. Leur essor en reconnaissance de parole est dû à l'exploitation de critères de comparaison performants, comme l'alignement temporel dynamique des formes acoustiques, et à leur application à des représentations adaptées du signal, qu'il s'agisse de l'analyse spectrale ou de la prédiction linéaire.

2.4. Les outils d'analyse

Afin de réaliser la phase de segmentation semi-automatique nous avons besoin d'un outil d'analyse qui facilite cette phase.

Nous pouvons citer quelques outils logiciel qui permettent de visualiser la forme d'onde et le spectrogramme d'un signal ou de paroles, d'éditer et d'aligner des transcriptions orthographiques et phonétiques sur ce signal, tels que PRAAT, CLAN, speech analysis, Goldwave, Cool Edit , etc....

2.4.1. Le logiciel CLAN (Computerized Language ANaalysis)

CLAN dont la traduction serait Analyse du langage par ordinateur. Chaque ligne/paragraphe correspond par exemple à une prise qui peut être alignées avec le signal audio ou audiovisuel. CLAN facilite aussi l'alignement temporel entre transcription et signaux audio ou vidéo. Il permet de communiquer des segments sonores au logiciel

d'analyse. De plus, CLAN permet l'insertion et la représentation de catégories syntaxiques, morphologiques et phonétiques sous formes d'annotations interlinéaires. Le logiciel offre un nombre important de routines d'analyse linguistique, de recherche et de calcul statistique. CLAN offre également de nombreuses possibilités d'analyses automatiques sur les données transcrites telles que le calcul de fréquence, la recherche de mots, les analyses interactionnelle et morphosyntaxique. Il permet de communiquer des segments sonores au logiciel d'analyse et d'annotation phonétique.

. IL est gratuit et accessible sur plusieurs environnements informatiques : Mac Classic, Mac Carbon (OSX), Windows, Unix [10].

Enfin, aucune possibilité d'import n'est envisageable alors qu'il est possible d'exporter un segment sonore vers PRAAT mais aucune transcrpition ou annotation n'est fournie dans la foulée : seul le son est exporté.

2.4.2. Le logiciel PRAAT

Le logiciel PRAAT a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam.

PRAAT est un logiciel d'analyse et de transcription phonétique (spectre, intonation, intensité etc.). Le logiciel comporte aussi des fonctionnalités importantes pour l'enregistrement, pour la manipulation et pour la synthèse de sons, pour la création d'algorithmes d'apprentissage, pour l'analyse statistique, ainsi que pour diverses expériences auditives. Praat est hautement portable, configurable et programmable [11]. En linguistique interactionnelle, le logiciel est utilisé pour divers types de transcription alignée de données sonores (éventuellement extraits d'une vidéo), pour aligner des transcriptions déjà réalisées en texte brut, mais aussi pour l'analyse et la transcription prosodiques. Avec ce logiciel, il est possible :

- d'enregistrer des fichiers audio qui pourront ensuite être analysés ;
- de transcrire, d'étiqueter et de segmenter des données audio (que les enregistrements aient été effectués sous Praat ou qu'ils proviennent d'autres fichiers, au format WAV, par exemple) ;

- d'effectuer des analyses phonétiques et acoustiques au niveau segmental (spectrogramme, analyse de formants, sonagrammes, etc.) et au niveau suprasegmental (pitch ou Fo, intensité et durée) ;
- de manipuler et modifier le signal de parole (utilisation de filtres ; modification des contours intonatifs et de la durée, etc.) ;
- de faire de la synthèse de la parole (créer des stimuli audio, synthèse articulatoire, analyse -synthèse de données modifiées, etc.) ;
- de faire des analyses statistiques à partir des études phonétiques (analyses de covariances, etc.). Nous pouvons résumer les fonctionnalités de ce logiciel dans la figure suivante (Figure 2.1).

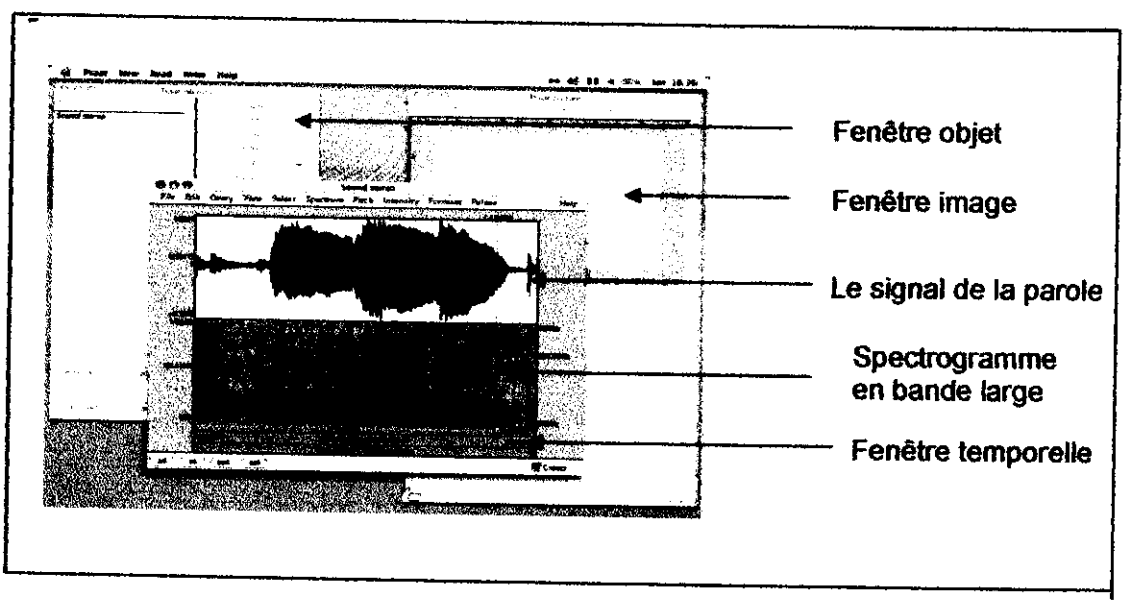


Figure 2.1 : Présentation du logiciel PRAAT

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

- La fréquence fondamentale : cochez « Show pitch » dans le menu « Pitch », et elle apparaît (c'est une courbe de couleur cyan). Sa valeur moyenne (en Hz) s'affiche à droite. ;

- Les formants : cochez « Show formants » dans le menu « Formant », et ils apparaissent en pointillés rouges. Pour les afficher sur toute la longueur de la fenêtre, affichez la fenêtre « Formant Settings » du menu « Formant », et dans le champ « Maximum Duration », entrez la durée de la fenêtre, en secondes, à la place de la valeur initiale.

- Les périodes du signal sonore : cochez « Show pulses » dans le menu « Pulses ». Chaque période est représentée, sur l'enveloppe, par un trait bleu vertical.

2.5. Les unités acoustiques

Les unités acoustiques sont des unités utilisées lors de la segmentation, ils sont divisées en plusieurs types : phonèmes, diphones, syllabes, triphones, demi-syllabe,...etc.

Nous présentons ci- dessous quelques unités acoustiques.

2.5.1. Phonème

Un phonème est la plus petite unité discrète ou distinctive (c'est-à-dire permettant de distinguer des mots les uns des autres) que l'on puisse isoler par segmentation dans la chaîne parlée. Un phonème est en réalité une entité abstraite, qui peut correspondre à plusieurs sons [8].

Il est en effet susceptible d'être prononcé de façon différente selon les locuteurs ou selon sa position et son environnement au sein du mot. On transcrit traditionnellement les phonèmes par des lettres placées entre des crochets: [a], [t], [r], etc...

L'identification des phonèmes d'une langue se fait en construisant des paires minimales, c'est-à-dire des paires de mots de sens différents et qui ne diffèrent dans leur forme sonore que par un seul son (ce son peut alors être considéré comme un phonème).Exemples : pas [pa] et bas [ba] sont deux mots différents de la langue

française, et il n'y a qu'un seul son différent (le premier). Donc, on peut conclure que le [p] et le [b] sont des phonèmes différents.

- **Les problèmes avec les phonèmes**

Le phonème est défini en termes articulatoires, alors que les sons de parole arrivent sous forme acoustique. Les problèmes principaux sont :

- la variabilité et la coarticulation ;
- la segmentation : les frontières entre les phonèmes ne sont pas toujours évidentes dans le signal acoustique ;
- Certains sons appartiennent à plus d'un phonème.

2.5.2. Qu'est ce qu'un diphone ?

Les premières études à avoir utilisé le diphone arabe comme unité de base ont été menées respectivement à L'ENPA en collaboration avec le CNET et à la faculté des sciences de RABAT. La technique de LPC a été d'abord utilisée pour la concaténation des unités acoustiques. Elle a été supplantée par la technique PSOLA dans d'autres systèmes de SAT (système de synthèse à partir du texte arabe) et plus récemment dans des systèmes intégrant le moteur de synthèse d'Elan Speech [6].

Chaque son est le résultat d'une position bien précise des lèvres, de la langue, du voile du palais, de la glotte... Or, les muscles articulatoires qui positionnent ces organes travaillent, pour ainsi dire, à l'économie: ils évitent les gymnastiques trop périlleuses dans le passage d'un phonème à un autre. Ils préfèrent enchaîner les mouvements souplement, en un mécanisme appelé coarticulation.

C'est pourquoi, en synthèse vocale, on ne travaille pas seulement à partir des phonèmes, mais aussi à partir de ce qu'on appelle les diphones, qui sont, pour ainsi dire, les "jointures" entre phonèmes. Dans l'exemple cité, [na], [ma] et [sa] sont quatre diphones différents, le diphone est un segment partant d'une zone stable du premier phonème et allant à la zone stable du deuxième phonème et qui contient en son centre toute la zone transition, Exemple : le mot مدرسة/ madrasatun (école) sera restitué à partir de la séquence de diphones [#m], [ma], [ad], [dr], [rs], [sa], [at], [tu], [un], [n#], les diphones [#m] et [n#] représentent respectivement un segment de silence suivi de la

première partie du phonème [m] et la deuxième partie du phonème [n] suivie d'un segment de silence.

2.5.3. Syllabe

La notion de syllabe est difficile à cerner, pour une bonne raison : elle varie selon la langue à analyser. Plusieurs approches sont possibles pour tenter de la définir. On peut, pour l'instant, se contenter de dire que c'est une unité phonétique plus grande que le phonème et plus petite que le mot et qu'un locuteur X est capable de découper un mot en syllabes dans sa langue, sans forcément savoir comment il procède. Un mot est donc composé de phonèmes, qui forment des syllabes.

2.6. Segmentation et Etiquetage phonétique de la parole

Selon le dictionnaire Larousse, le terme segmentation désigne la division d'un ensemble en portions bien délimitées. Autrement dit, c'est le processus de division d'une entité, généralement continue, en petites entités appelées segments. Chaque segment possède des propriétés propres qui permettent de le différencier des autres [9].

Les processus de segmentation et d'étiquetage de corpus de parole sont une forme d'annotation linguistique (Figure 2.2). Cette dernière désigne toute notation descriptive appliquée à des données audio, vidéo ou textuelles, et apportant des informations de nature interprétative à ces données brutes.

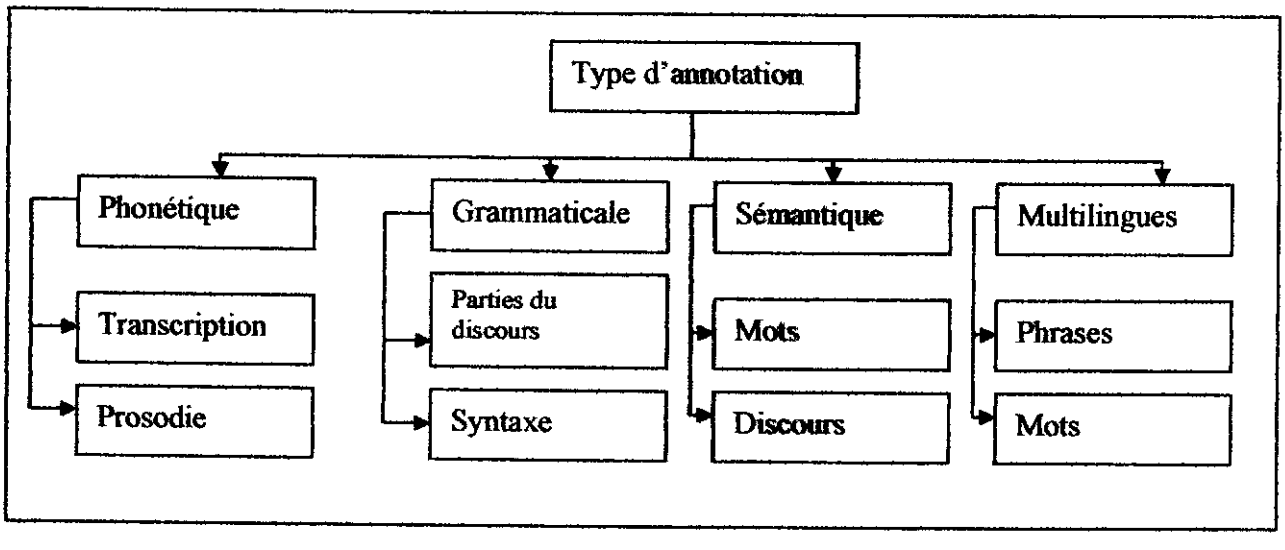


Figure 2.2 : différents types d'annotation linguistique.

La transcription phonétique signifiée dans la Figure 2.2, est celle généralement assorti d'un alignement temporel des phonèmes sur le signal de parole.
Cette transcription est généralement qualifiée d'étiquetage et segmentation phonétique de parole.

Nous pouvons distinguer, cependant, plusieurs autres types d'annotation (segmentation et étiquetage) de parole ; la segmentation :

- de parole pour l'indexation (l'étiquetage) des locuteurs ;
- bruit/silence/parole pour la détection de l'activité vocale ;
- bruit/silence/musique/parole, pour l'extraction de l'information des documents multimédia ;
- prosodique de parole.

La segmentation d'un signal de parole en phonèmes consiste à délimiter sur le continuum acoustique de ce signal une séquence de segments caractérisés par des étiquettes appartenant à un ensemble discret et fini d'éléments, qui est l'alphabet phonétique de la langue. [7]

D'un côté, nous constatons que l'élocution d'un énoncé se caractérise par un mouvement continu des organes de la parole et par l'absence d'un positionnement statique de ces organes. Le passage d'une cible articulatoire d'un phonème, à une autre cible articulatoire d'un autre phonème, se fait de manière continue, avec un chevauchement entre les deux configurations articulatoires, ce qui donne naissance au phénomène de coarticulation.

D'un autre côté, sur la base de notre perception de la parole, nous pouvons affirmer que ce signal se compose d'une série d'éléments sonores distincts. En effet, le spectrogramme d'un signal de parole permet de distinguer des zones spectralement homogènes (Figure 2.3).

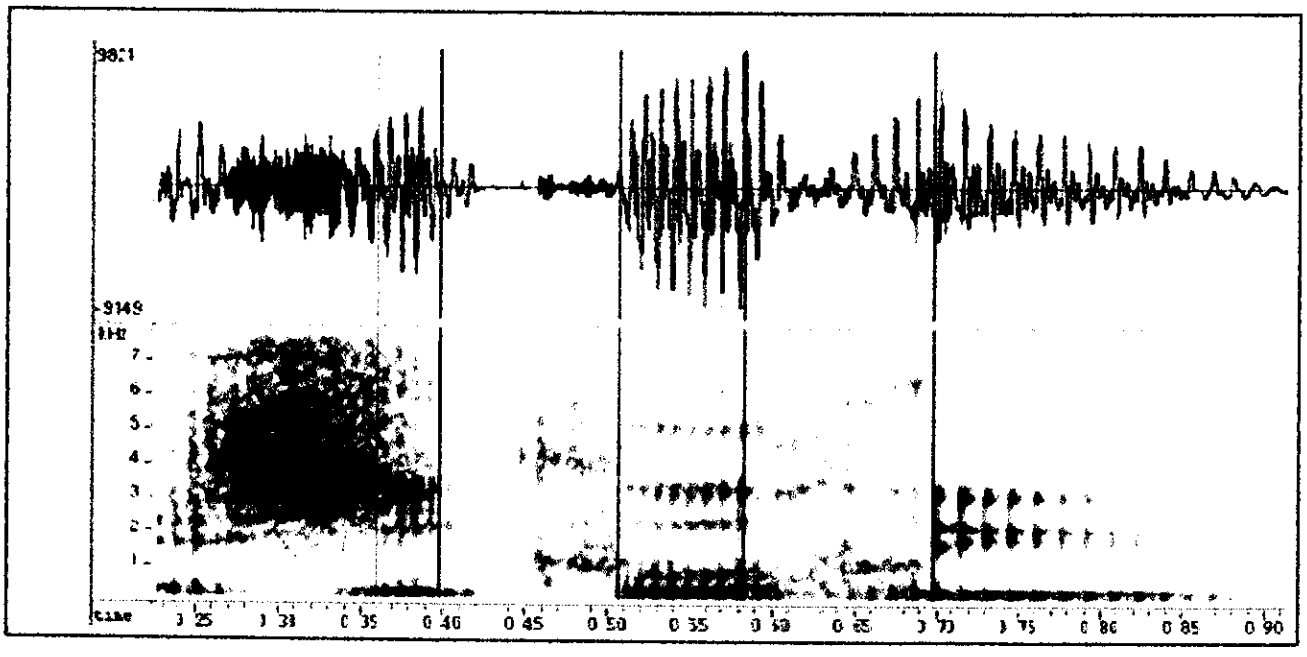


Figure 2.3 : forme d'onde et spectrogramme de la phrase " J'y couru"

2.6.1. Modes de segmentation

La segmentation phonétique de la parole est une tâche difficile car le signal de parole n'est pas clairement composé de segments discrets bien délimités.

2.6.1.1. Segmentation et Etiquetage manuels

L'étiquetage et la segmentation phonétique de parole, requièrent des experts phonéticiens de la langue, afin de faciliter cette intervention manuelle, il est nécessaire de fournir à ces experts une description des conditions d'enregistrement du corpus de parole (chambre anéchoïque ou environnement bruyant), des caractéristiques du locuteur enregistré (sexe, âge, taille, pays d'origine, qualité de la voix,pathologie, etc.)
A ce niveau, deux stratégies d'intervention manuelle peuvent être appliquées :

- la première stratégie s'appuie sur l'intervention manuelle dans tout le cycle de développement d'un corpus de parole segmentée et étiquetée ;
- la deuxième stratégie, semi-automatique, divise le cycle de développement en deux étapes, dont la première fait intervenir des méthodes de segmentation et d'étiquetage automatiques.

L'intervention manuelle dans la deuxième étape, consiste à effectuer une vérification et une correction de l'étiquetage et de la segmentation produits précédemment par les méthodes automatiques.

Afin de segmenter et d'étiqueter un énoncé de parole, l'expert humain, en s'aidant de l'écoute du signal de parole et de transcription orthographique de cet énoncé, recherche sur la forme d'onde et le spectrogramme de ce signal des événements qui permettent d'étiqueter et de délimiter les segments phonétiques constituant cet énoncé, on peut citer comme exemple le voisement ou le non voisement d'un son.

2.6.1.2. Caractérisation de la segmentation et de l'étiquetages manuelle

La segmentation manuelle de parole est effectuée par des opérateurs ayant une expertise en phonétique de la langue, et incontestablement à ce jour, la méthode de segmentation la plus précise pour les applications de la synthèse de parole par concaténation même constat peut être fait pour l'étiquetage phonétique manuel comparativement à l'étiquetage automatique.

Cependant, ce mode de préparation de corpus de parole pose premièrement un problème de disponibilité d'expert et deuxièmement un problème plus grave lié au temps prohibitif que nécessite l'intervention manuelle de ces derniers pour l'annotation de grands corpus de parole. L'intervention de plusieurs opérateurs humains pour segmenter et étiqueter un corpus de parole donné est possible mais elle pose des problèmes de consistance (la segmentation l'étiquetage d'un même énoncé de parole peut différer d'un opérateur à un autre).

Afin de segmenter et d'étiqueter un énoncé de parole, l'expert humain, en s'aidant de l'écoute du signal de parole et de la transcription orthographique de cet énoncé, recherche sur la forme d'onde et le spectrogramme de ce signal des évènements qui permettent d'étiqueter et de délimiter les segments phonétiques constituant cet énoncé.

2.6.1.3. Segmentation automatique de la parole

Il convient d'ores et déjà ce que nous entendons par segmentation automatique de parole. Actuellement, la segmentation complètement automatique de parole est une

tâche rarement possible. En effet, étant donnée la complexité des phénomènes acoustico-phonétiques à traiter, cette tâche nécessite très souvent une intervention manuelle, que ce soit pour la préparation des données (étiquetage phonétique) du traitement automatique. On peut voir l'automatisation de cette tâche comme un continuum de possibilités entre la segmentation et l'étiquetage purement manuel et l'automatisation complète. Globalement, les méthodes de segmentation acoustique de parole se divisent en deux grandes classes de méthodes :

- La première classe englobe toutes les méthodes qui permettent de segmenter un signal de parole sans connaissance a priori du contenu linguistique de ce signal. Ces méthodes produisent des segmentations d'un signal de parole en zones spectralement homogènes ;

- La deuxième classe englobe toutes les méthodes qui permettent de segmenter un signal de parole étant donnée une description linguistique de ce signal. Ces méthodes de segmentation sont dites contraintes linguistiquement.

Les méthodes de la première classe ont pour objectif de produire une segmentation acoustique du signal de parole sans lien a priori avec le contenu linguistique de ce signal. Autrement dit, ces méthodes ne fournissent pas un étiquetage linguistique des segments acoustiques qu'elles délimitent. Ces derniers reflètent cependant la réalité physique du signal car chacun de ces segments représentent une zone d'homogénéité (stabilité) spectrale du signal. C'est pourquoi, chacune des méthodes de segmentation de cette classe utilise une mesure de distance entre vecteurs acoustiques ou entre modèles statistiques, permettant de détecter et de limiter les segments acoustiques constituant le signal de parole.

Pour la deuxième classe, à partir d'un signal de parole et d'une description linguistique de ce signal, il s'agit de retrouver une association entre cette description linguistique et la séquence des trames acoustiques de cet énoncé. Cette description linguistique est constituée d'une séquence de symboles linguistiques, typiquement des phonèmes.

L'objectif des méthodes de segmentation qui utilisent ces types de données en entrée consiste à obtenir une séquence de segments acoustiques contigus et définis par des instants temporels de début et de fin. Le nombre de segments acoustiques

délimité doit être égal aux nombres d'étiquettes présentes dans la description. Chaque segment acoustique est caractérisé par son étiquette linguistique.

2.7. Etiquetage

L'étiquetage prosodique de grands corpus serait extrêmement utile pour l'étude de la parole en général, ainsi que pour des tâches d'ingénierie telles que la génération d'une prosodie de qualité en synthèse ou bien de l'utilisation de la prosodie en reconnaissance.

Il existe deux modes d'étiquetage :

- l'étiquetage manuel est considéré comme un caractère lent et difficile, donc il diminue la fiabilité des résultats, ou impose le recours à des contre-expertises qui augmentent les coûts ;
- l'étiquetage automatique, Il serait donc intéressant de disposer de systèmes d'étiquetage automatique permettant de s'affranchir de la phase d'intervention manuelle, ou de la réduire à une simple phase de vérification et de correction.

2.8. Décodage Acoustico-Phonétique (DAP)

Le DAP est une étape fondamentale de la reconnaissance de la parole continue. Le rôle du DAP est de transformer le signal acoustique, en une suite d'unités phonétiques.

Les méthodes actuellement les plus performantes dans ce domaine sont fondées sur les HMM (Hidden Markov Model). Des études sur les espaces de représentation ou sur le choix des modèles permettent une amélioration des performances. Cependant, le DAP est un processus au cours duquel la segmentation et l'identification sont étroitement liées, et les systèmes de reconnaissance à base de HMM ne permettent pas de localiser avec précision les frontières phonétiques. De plus, la variabilité inter-locuteurs rend le choix de l'ensemble d'apprentissage crucial pour une bonne estimation des densités de probabilité des modèles.

2.9. Conclusion

Afin de préparer une BD d'unités acoustiques qui sera utilisée dans le domaine de TAP, il est nécessaire de segmenter et d'étiqueter le plus souvent phonétiquement, le corpus de parole à partir duquel est extraite cette BD d'unités. Ces deux tâches sont en effet, fondamentales pour la mise en œuvre de telle BD.

Nous avons présenté dans ce chapitre les différentes méthodes de segmentation et d'étiquetage de la parole en unités acoustiques, nous avons aussi parler du DAP nous avons aussi exposé les outils d'analyse du signal de parole en citant l'outil que nous avons utilisé dans notre projet ;

Dans le chapitre suivant nous allons exposer tout d'abord un état de l'art des bases de données de façon générale et des BD spécialisées dans le TAP (BD sonores).

Chapitre 3 : Généralités sur les Bases de Données

3.1. Introduction

Les bases de données ont aujourd'hui pris une place essentielle dans les systèmes informatiques tant de point de vue pratique que théorique. On peut définir une BD comme un ensemble physique et cohérent de données facile à manipuler. La plupart des systèmes offrent aujourd'hui un Système de Gestion de Base de Données (SGBD). Un SGBD est une interface entre l'usager et les mémoires secondaires qui tendent à créer l'illusion que les données désirées par tout usager sont stockées sur mémoires secondaires, assemblées et codées comme souhaitées, comme si l'usager était seul à utiliser ces données.

Notre objectif dans ce chapitre est de présenter des généralités sur les BD, et notamment les BD sonores utilisées en TAP comme TIMIT et BDSONS.

3.2. Qu'est ce qu'une BD ?

Une BD est une entité dans laquelle il est possible de stocker des données de façon structurée et avec le moins de redondances possibles. Ces données doivent pouvoir être utilisées par des programmes, et des utilisateurs différents [12].

3.2.1. Utilité d'une BD

Une BD permet de mettre des données à la disposition des utilisateurs pour une consultation, une saisie ou bien une mise à jour, tout en s'assurant des droits accordés à ces derniers. Cela est d'autant plus utile que les données informatiques sont de plus en plus nombreuses.

Une BD peut être locale, c'est-à-dire utilisable sur une machine par un utilisateur, ou bien répartie, c'est-à-dire que les informations sont stockées sur des machines distantes et accessibles par réseau. L'avantage majeur de l'utilisation de BD est la possibilité de pouvoir être accédées par plusieurs utilisateurs simultanément.

3.2.2. Gestion des BD

Afin de pouvoir contrôler les données ainsi que les utilisateurs, le besoin d'un système de gestion s'est vite fait ressentir. La gestion de la BD se fait grâce à un système appelé SGBD ou en Anglais DBMS (DataBase Management System). Le SGBD est un ensemble de services (applications logicielles) permettant de gérer les BD, c'est-à-dire :

- de permettre l'accès aux données de façon simple ;
- d'autoriser un accès aux informations à de multiples utilisateurs ;
- de manipuler les données présentes dans la BD (insertion, suppression, modification).

Le SGBD peut se décomposer en trois sous-systèmes (figure3.1):

- le système de gestion de fichiers permet le stockage des informations sur un support physique ;
- le SGBD interne gère l'ordonnancement des informations ;
- le SGBD externe représente l'interface avec l'utilisateur.

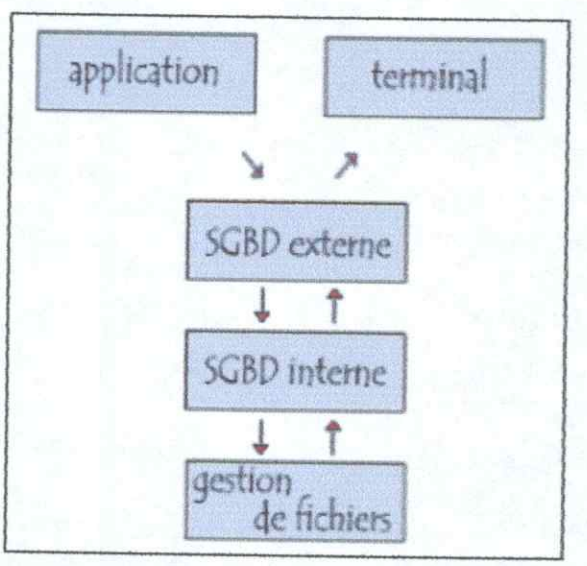


Figure 3.1 : Les trois couches d'un SGBD

3.3. Les types des BD

Il existe différents types des BD, les BD relationnelles, les BD orientées Objets, les BD multimédia, et les BD sonores qui sont spécifiées dans les domaines de TAP.

3.3.1. Les bases de données relationnelles

Le modèle relationnel a été formalisé par CODD en 1970. Quelques exemples de réalisation en sont DB2 (IBM), INFORMIX, INGRES, ORACLE.

Dans ce modèle, les données sont stockées dans des tables, sans préjuger de la façon dont les informations sont stockées dans la machine. Un ensemble de données sera donc modélisé par un ensemble de tables [13]

Le succès du modèle relationnel auprès des chercheurs, concepteurs et utilisateurs est dû à la puissance et à la simplicité de ses concepts. il repose sur des bases théoriques solides, notamment la théorie des ensembles et la logique mathématique (théorie des prédicats d'ordre 1).

Les objectifs du modèle relationnel :

- proposer des schémas de données faciles à utiliser ;
- améliorer l'indépendance logique et physique ;
- mettre à la disposition des utilisateurs des langages de haut niveau pouvant éventuellement être utilisés par des non informaticiens ;
- optimiser les accès à la BD ;
- améliorer l'intégrité et la confidentialité.
- fournir une approche méthodologique dans la construction des schémas.

De façon informelle, on peut définir le modèle relationnel de la manière suivante :

- les données sont organisées sous forme de tables à deux dimensions encore appelées relations et chaque ligne n-uplet ou tuple ;
- les données sont manipulées par des opérateurs de l'algèbre relationnelle ;
- l'état cohérent de la base est défini par un ensemble de contraintes d'intégrité.

Au modèle relationnel est associée la théorie de la normalisation des relations qui permet de se débarrasser des incohérences au moment de la conception d'une BD. Nous pouvons donner des définitions brèves sur les termes utilisés dans les notions de BD relationnelle de la manière suivante :

- le domaine est un ensemble de valeurs ;
- la relation est un sous-ensemble du produit cartésien d'une liste de domaines caractérisés par un nom ;
- l'attribut est une Colonne d'une relation caractérisée par un nom ;
- le schéma de relation est le nom de la relation, suivi de la liste des attributs avec leurs domaines ;
- une BD relationnelle est une BD dont le schéma est un ensemble de schémas de relations et dont les occurrences sont les tuples de ces relations ;

- le SGBD relationnel est un logiciel supportant le modèle relationnel, et qui peut manipuler les données avec des opérateurs relationnels [13].

Les langages de manipulation de données relationnelles, dits assertionnels, sont basés sur la logique des prédicats d'ordre 1 et permettent de spécifier les données que l'on souhaite obtenir, sans dire comment y accéder. On doit y trouver des opérations permettant :

- la recherche pour retrouver des tuples vérifiant certains critères,
- l'insertion pour ajouter des tuples,
- la suppression pour enlever des tuples vérifiant certains critères,
- la modification pour modifier des tuples vérifiant certains critères.

Un langage de manipulation de données n'est pas utilisable à lui seul, il doit aussi pouvoir être incorporable dans un langage de programmation classique. On peut distinguer trois grandes classes de langages :

- algébriques basés sur l'algèbre relationnelle de CODD dans lesquels les requêtes sont exprimées comme l'application des opérateurs relationnels sur des relations. C'est dans cette catégorie que l'on trouve le langage SQL (Structured Query Language), standard pour l'interrogation de bases de données ;

- basés sur le calcul relationnel de tuples construits à partir de la logique des prédicats dans lesquels les variables manipulées sont des tuples ;

- basés sur le calcul relationnel de domaines, construit aussi à partir de la logique des prédicats mais en faisant varier les variables sur les domaines des relations.

Le langage SQL comprend à lui seul l'ensemble des instructions nécessaires à la spécification et à l'utilisation d'une BD relationnelle. C'est un langage de type déclaratif c'est-à-dire que l'on spécifie les propriétés des données que l'on recherche et pas comment les retrouver.

C'est à la fois un Langage :

- d'Interrogation de Données (LID) : select ;
- de Manipulation de Données (LMD) : update, INSERT, DELETE ;

- de Définition des Données (LDD) : ALTER, CREATE, DROP;
- de Contrôle des Données et des utilisateurs (LCD) : GRANT, REVOKE.

• Limites des SGBD relationnels

Les SGBD relationnels ont à leur avantage :

- le modèle de données est très simple et donc facile à comprendre pour les utilisateurs;
- le modèle repose sur une base formellement définie; ce qui a permis de définir des méthodes de conception de schémas (théorie de la normalisation) et des langages de manipulation de donnée (LMD) standardisés (SQL, QUEL...).

Mais, le développement de nouvelles applications, différentes des applications de gestion classiques, avec de nouveaux besoins, ont révélé les limites du modèle relationnel [14].

- le modèle de données est trop simple et ne permet pas de représenter facilement les entités du monde réel qui sont souvent plus complexes qu'une relation. Dans les schémas, les entités du monde réel sont éclatées en plusieurs relations, ce qui multiplie les jointures dans les requêtes des utilisateurs.
- de plus le modèle relationnel ne permet pas de représenter explicitement les différents types de liens sémantiques qui peuvent lier des entités : composition, généralisation / spécialisation, association...
- l'incompatibilité des LMD relationnels et des langages de programmation: les LMD sont déclaratifs et fournissent en résultat un ensemble de tuples, alors que les langages de programmation sont impératifs et travaillent sur un élément à la fois;
- les types de données manipulés par les langages de programmation sont plus complets et plus complexes que ceux des LMD relationnels.

3.3.2. Les bases de données orientées objet (BDOO)

Les BDOO sont nées de la convergence de deux domaines [15] :

· les BD ;

· les langages de programmation orientés objets, tels que Eiffel, Smalltalk, C++, Java...etc. L'objectif de ces langages est d'accroître la productivité des développeurs en permettant de créer des logiciels structurés, extensibles, réutilisables et de maintenance aisée. Leurs principes essentiels sont:

- les objets, qui comportent deux parties: leur valeur, et les opérations, appelées méthodes, qui permettent de les manipuler. La valeur est cachée. L'accès et la mise à jour des objets se fait par appel aux méthodes. Cela rend plus facile la maintenance des logiciels construits selon ce paradigme ;
- l'héritage, qui permet à une classe d'objets d'avoir les mêmes propriétés (structure de données et méthodes) qu'une autre classe sans avoir à les redéfinir. C'est l'héritage qui permet d'étendre et de réutiliser facilement des logiciels.

Les BDOO sont caractérisées par quatre points essentiels:

- un modèle de données qui permet de représenter des structures de données complexes;
- les données et les traitements ne sont plus séparés. La dynamique (la méthode) fait partie de la statique ;
- déclaration des objets;
- l'héritage.

Tout objet possède une identité qui le distingue de tout autre objet, même s'ils ont la même valeur.

3.3.3. Les BD multimédia

Une BD multimédia est un type de BD consacrée au stockage et à l'organisation de données multimédia : documents sonores, images, vidéos. Elles peuvent s'appuyer sur différentes architectures de bases de données, les types les plus utilisés étant le modèle relationnel et le modèle objet [16].

Les BD multimédia ne sont pas des applications encore très courantes, et posent encore des problèmes, notamment pour l'indexation de leur contenu et les recherches par contenu.

L'approche classique consiste à renseigner des mots-clés décrivant dans un vocabulaire restreint les caractéristiques principales et bien identifiables des documents stockés. Les limitations évidentes sont le manque de précision de la description et l'impossibilité de faire des recherches sur des informations existantes dans les documents mais non pertinentes du point de vue de l'indexation.

Une approche actuellement sujette à recherches consiste à disposer d'une indexation dynamique par l'application d'opérateurs de traitement d'image (pour les bases de données images et vidéo) ou de traitement du signal (pour des bases de

données sonores). L'objectif est de décomposer les documents en entités élémentaires structurées et reliées entre elles, de manière à ce que l'indexation permette de retrouver soit des formes (les structures) soit des objets (les entités) soit des combinaisons des deux. Le problème consiste alors à trouver les « bons » outils de traitement du signal (au sens large) pour faire le travail d'analyse des documents, et à construire le ou les index de manière à exploiter pleinement des informations extraites.

Le moteur de requêtes est lui aussi un sujet d'étude dans ces conditions, car les requêtes ne s'expriment donc plus en fonction de textes et de valeurs numériques « simples » mais en fonction de rapports spatiaux et temporels entre des entités qu'il faut par ailleurs décrire de manière structurée.

- **Problématiques des bases de données multimédia :**

Un des problèmes majeurs de l'évolution des BD multimédias est celui des performances lorsqu'il s'agit de traiter de gros corpus de textes, d'images, de sons et de vidéos. Les recherches et les développements dans ce domaine sont très actifs et plusieurs méthodes d'organisation et d'indexation des données multidimensionnelles ont été proposées et implantées. Aujourd'hui, certaines d'entre elles sont quelquefois proposées en option par les grands constructeurs de SGBD.

3.3.4. Bases de données sonores

Les sons digitalisés ont aussi été intégrés dans les systèmes informatiques et dans les BD. Le signal, par exemple d'une musique ou d'une phrase prononcée par un locuteur, ne pose pas de problème spécifique pour son stockage mais en pose pour sa recherche et son traitement. Là encore, extraire des informations pertinentes à partir d'un son digital fait partie de la recherche du domaine (par exemple, la reconnaissance de la parole). Il est possible cependant de stocker toutes sortes d'informations descriptives associées à un signal digitalisé et de constituer des banques de sons où les requêtes exploitent ces informations comme des attributs spécifiques. Comme pour les images, des interfaces spécifiques sont nécessaires pour poser les requêtes et entendre le résultat [17].

La nécessité de disposer des BD sonores a émergé il y a une dizaine d'années, sous la poussée des méthodes de reconnaissance de la parole fondées

sur l'apprentissage. Dans le milieu des années 80, des travaux ont été coordonnés en France autour de BDSONS, et la BD TIMIT vers 1990. A l'heure actuelle, le besoin de données sonores reste encore une préoccupation essentielle.

3.3.4.1. Description de la base TIMIT

La BD TIMIT est une BD acoustique et phonétique dédiée à la reconnaissance de la parole indépendamment du locuteur.

Elle contient les enregistrements de 630 locuteurs américains, répartis en 8 dialectes régionaux et prononçant chacun 10 phrases. Ces phrases proviennent de 3 corpus[17] :

• 2 phrases de calibration, prononcées par tous les locuteurs, servent à illustrer les variations régionales (identifiées "sa1" et "sa2");

• 5 phrases sont tirées au sort parmi 450 phrases phonétiquement équilibrées (Identifiées "sx3" à "sx452").

• 3 phrases sont choisies pour maximiser les contextes acoustiques; chaque phrase n'est prononcée qu'une seule fois, soit un total de 1890 phrases différentes pour les 630 locuteurs (identifiées "si453" à "si2342"); Le vocabulaire total de la base est de 6100 mots. Le texte est lu, et les conditions d'enregistrement sont bonnes. Les 630 locuteurs de la base (438 hommes et 192 femmes) sont répartis entre l'ensemble d'apprentissage (462 locuteurs dont 326 hommes et 136 femmes) et l'ensemble de test (168 locuteurs dont 112 hommes et 56 femmes). Chaque locuteur est identifié par une lettre indiquant son genre (m" pour les hommes et "f' pour les femmes), ses 3 initiales et un chiffre. L'Ensemble d'Apprentissage et l'Ensemble de Test sont respectivement notés EA et ET. Un sous-ensemble de l'ensemble de test, appelé noyau de test, ne contient que 24 locuteurs de test: deux hommes et une femme pour chacun des 8 "dialectes". Le noyau de test est noté ET192.

Pour chaque phrase, nous disposons du texte en anglais, du signal échantillonné à 16 kHz sur 16 bits, de la segmentation phonétique en 61 classes, et de la segmentation en mots. La segmentation phonétique est très fine; en particulier, l'occlusion précédant l'explosion des occlusives et des affriquées est étiquetée individuellement. Le tableau 3.1 donne la liste des étiquettes de la BD TIMIT, le phonème correspondant dans l'Alphabet Phonétique International (API) est une mise en contexte de ce phonème dans un mot anglais.

TIMIT	API	Exemple	TIMIT	API	Exemple	TIMIT	API	Exemple
Occlusives:			*Nasales*:			*Voyelles*:		
pcl p	p	pea	m	m	mom	iy	iʸ	beet
tcl t	t	tea	em	m̩	bottom	ih	i	bit
kcl k	k	key	n	n	noon	ix	I	debit
bcl b	b	bee	nx	ɾ̃	winner	eh	E	bet
dcl d	d	day	en	n̩	button	ae	æ	bat
gcl g	g	gay	ng	N	sing	aa	A	bott
dx	ɾ	muddy	eng	Ṇ	washington	ao	O	bought
q	ʔ	bat	*Liquides*:			uh	Ʊ	book
Affriquées:			l	l	lay	uw	u	boot
dcl jh	j&	joke	el	l̩	bottle	ux	ʉ	toot
tcl ch	c&	choke	r	r	ray	ax	ə	about
Fricatives:			*Semi-voyelles*:			ax-h	ə̥	suspect
f	f	fin	w	w	way	ah	Ʌ	but
th	Q	thin	y	y	yacht	er	ɝ	bird
s	s	sea	*Fricative glottale*:			axr	ɚ	butter
sh	s&	she	hh	h	hay	*Diphtongues*:		
v	v	van	hv	H	ahead	ey	eʸ	bait
dh	D	then	*Silences*:			ay	Aʸ	bite
z	z	zone	h#			oy	Oʸ	boy
zh	z&	azure	pau			aw	Aʷ	bout
			epi			ow	oʷ	boat

Tableau 3.1 Étiquetage de TIMIT, code API correspondant et exemple de mot anglais contenant le phonème

- **Regroupement d'allophones**

Un allophone est l'une des réalisations sonores possibles d'un phonème. Au sein d'une même langue. L'étiquetage d'origine en 61 classes est généralement jugé trop détaillé pour l'apprentissage de modèles phonétiques, et une réduction du nombre de classes phonétiques par regroupement d'allophones est réalisée. K.F. Lee et H.W. Hon ont proposé un regroupement en 39 classes phonétiques [Lee & Hon, 1989] qui a été ensuite réutilisé par d'autres chercheurs à des fins de comparaison [Robinson et al. 1990; Niles, 1991; Chigier & Leung, 1992]. Ce regroupement s'effectue en deux étapes:

• avant l'apprentissage des modèles, regroupement en 48 classes par fusion d'allophones ([m]/[em], [n]/[nx], [ng]/[eng],[ax]/[ax-h],[er]/[axr],[ux]/[uw],[hh]/[hv]), regroupement des silences et des occlusions (nouvelle étiquette sil pour regrouper les silences [h#]/[pau], [cl] pour les occlusions sourdes [pcl]/[tcl]/[kcl], [vcl] pour les occlusions voisées [bcl]/[dcl]/[gcl]) et suppression de l'étiquette q (qui ne correspond pas toujours à une occlusive);

• Des confusions sont autorisées entre certaines classes lors du calcul des taux de décodage ([sh]/[zh], [n]/[en], [l]/[el], [ih]/[ix], [aa]/[ao], [ax]/[ah], [sil]/[epi]/[cl]/[vcl]), conduisant finalement à des résultats sur 39 classes.

Nous présentons dans le tableau 3.2 des statistiques sur les 48 classes phonétiques d'apprentissage, en rappelant les regroupements réalisés, et en donnant pour chaque classe le nombre de représentants dans l'ensemble d'apprentissage ainsi que la durée moyenne des segments en millisecondes.

Etiquette(s)	Nombre	Durée (ms)	Etiquette(s)	Nombre	Durée (ms)
Occlusives:			*Semi-voyelles*:		
b	2181	17	w	2216	60
d	2432	24	y	995	54
g	1191	27	*Fricative glottale*:		
p	2588	44	hh,hv	1660	67
t	3948	49	*Voyelles*:		
k	3794	52	iy	4626	95
dx	1864	29	ih	4248	78
Affriquées:			ix	7370	51
jh	1013	61	eh	3277	93
ch	820	86	ae	2292	136
Fricatives:			aa	2256	123
f	2215	103	ao	1865	123
th	745	92	uh	500	76
s	6176	113	uw,ux	1952	100
sh	1317	118	ax,ax-h	3892	47
zh	149	81	ah	2266	89
v	1994	60	er,axr	4138	95
dh	2376	36	*Diphtongues*:		
z	3682	84	ey	2271	127
Nasales:			ay	1934	155
m,em	3566	65	oy	304	168
n,nx	6896	52	aw	728	161
en	630	78	ow	1653	128
ng,eng	1220	61	*Silences*:		
Liquides:			sil=(h#,pau)	8283	191
l	4425	61	cl=(pcl,tcl,kcl)	12518	58
el	951	90	vcl=(bcl,dcl,gcl)	7219	54
r	4681	56	epi	908	42

Tableau 3.2. Statistiques sur le nombre de représentants et la durée moyenne des 48 classes phonétiques (les confusions autorisées au décodage sont désignées par une accolade).

3.3.4.2. BDSONS

BDSONS est une BD de parole française, au format SAM, constituée de 32 voix (16 hommes, 16 femmes). La taille de cette base est approximativement de 3,5 Go. Ces données sont réparties sur 7 CDROM.

Les données sont divisées en deux groupes :

- "Evaluation": ce groupe comporte 32 locuteurs. Chaque locuteur a prononcé :
  - Un passage de 5 phrases et 54 logatomes dissyllabiques (syllabes : /[pa]/, /[si]/ [fu]/) ;
  - Des nombres et des chiffres : 400 chiffres isolés, 200 séries de 3 chiffres, 100 séries de 4 chiffres, et 100 séries de 5 chiffres
  - Des lettres et des noms : 432 lettres, 102 noms épelés.
- "Acoustique" : Ici, ne sont représentés que 12 locuteurs :
  - 600 mots, de type [CVCV] (Consonne-Voyelle-Consonne-Voyelle au sens phonétique) ;
  - 200 groupes consonantiques ;
  - 52 phrases phonétiquement équilibrées, 44 phrases nasales, 192 phrases incluant des mots réels en français avec 16 consonnes et 12 voyelles.

L'avantage de BDSONS est son intégration dans un environnement informatique à l'aide du logiciel GERSONS2. Ce logiciel permet l'interrogation de la base pour en extraire un ensemble de signaux répondant à des critères particuliers.

3.4. Conclusion

Afin d'introduire la mise en œuvre du chapitre 4, nous avons exposé dans ce chapitre un état de l'art des BD en général et des BD sonores (TIMIT et BDSONS). Le prochain chapitre consiste à la modélisation et l'implémentation de notre BD nommée BDSONARABE.

Chapitre 4 :
Elaboration de BDSONARABE

4.1. Introduction

L'objectif de notre travail est de réaliser une interface d'acquisition et de stockage pour le développement de corpus de parole utiles dans les domaines de TAP suivants :

- apprentissage et évaluation des systèmes de reconnaissance et de synthèse de la parole ;
- analyse des caractéristiques des sons de la langue arabe (extraction des paramètres pertinents) ;

Il s'agit, soi d'utiliser les fichiers son qui sont déjà enregistrés ou bien d'élaborer des fichiers sons par le logiciel PRAAT, de les segmenter et de les stocker afin de les rendre accessibles aux traitements.

4.2. Le choix du corpus

Nous avons choisi, pour faciliter la segmentation, d'enregistrer un corpus de mots significatifs appartenant au vocabulaire arabe d'un locuteur masculin de nationalité jordanienne. Pour l'extraction de la totalité des phonèmes et diphones, et syllabe pris dans leurs trois positions (initiale, médiane et finale), et avec leurs trois voyelles (fatha, damma, et kasra) nous avons utilisé les enregistrements de près de 72 phrases et expressions utilisant le vocabulaire arabe usuel.

-أسعد زوجين.	-وحاولت.
-جلس يستمع إلى الراديو.	- إني أشعر هزة قوية.
-و يضاف إليه البصل.	-أرض المطار.
-في ما جاء في برنامج المرأة في ذلك.	-وكثير من الصناعات الخفيفة,وأنشئت.
-ظهور الإسلام. سيما.	-على تحسين وسائل النقل والمواصلات وإنشاء
-فإذا بصوت جميل.	المطارات.
-يوضع اللحم في القدر ثم يغطى.	-في صحبته.
-ولم يستطع أن يصبر فقام إلى أهله يقول لهم.	-ونظرة الحب تشع.
-لا بد لي من الزواج بهذه المرأة.	- على الغربة, كل هذا.

-أن تعرف طريقها إلى المعدة أولا.	-و تنظيم توزيعها,فتلجأ إلى بناء السدود.
- شوق.	-أهم رواد النهضة الحديثة.
-وانتظر ساعة ثم ساعة وساعة حتى تعب من الانتظار.	-نذكر في مقدمتهم رفاعة رافع الطهطاوي.
-من المطبخ والعرق يسيل من وجهها وقالت له آسفة	-بطرس البستاني ولد.
عملت لك بيضا مقليا خوفا.	-التي يسكنها الفلاحون.
-فوجده قد احترق و تحجر	-تسعين. تكوينه
-أسرعت الزوجة إلى الخروج كأنها على موعد.	-مجموع. رأينا.
-تقع الأقطار العربية عند ملتقى ثلاث قارات أوربا و آسيا وأفريقيا وتمتد سواحلها.	-عمر الولايات.
-والبحر الأبيض المتوسط وبحر العرب وعلى المحيطين الأطلسي والهندي.	-نصيب كبير, الدين.
-وتمتد جذور هذه الحضارة بعيدا في التاريخ كما نرى في آثار.	-تونس.
-العراق والفينيقيين في لبنان والرومان في سوريا والأردن.	-فإننا نلاحظ.
-فمع أن الأتراك والعرب قد عاشوا.	-أن تملك قلبه. فإذا ملكتها.
-لمذهب ديني.	-قل هو الله أحد,وظلت.
-والأرز في مصر والعراق والقمح في سوريا و ينتج التبغ	-وكانت الغاية.
في الجزائر والتمور.	-من سور القرآن تدخل في نفسها الطمأنينة فلم تتذكر.
-تشكل الزراعة موردا أساسيا في أكثر الأقطار العربية	-والعلوم المختلفة.
وتقوم على أساس.	-تشجيع.
-تستغل مياه النيل أحسن استغلال.	-مراكز.
-فراش الموت.	-في شبه.
	-والتاريخ كذلك لا يصلح وحده أن يكون الأساس الوحيد للقومية.
	-نهر الدانوق.
	-هذا الشعور.
	-خاصة للبنات.
	-الوزارات,غير. صاح,الرشيد
	-جامع الزيتونة تونس.
	--قطعة واحدة,من الأمم.

-ضعيف, أبي.	-شعور,وحدة المركز.
-المتفوقين للدراسة على حسابها.	-ومما يجب ذكره موضوعات صحيفة.
-الروح التي,أن يكون جوار.	-ومن أهم الصحف.
-نسبة المسيحيين.	-حكام العرب,الكاظمين. -مختلف العصور.
-حيث مي زيادة.	-الشعراء البارزون و المترجمون.
-خارج,روح التعاون. المشهور المسؤول.	-عاش أبناؤها محاطين.
-وأربعة عشر. -ولم يمضي,في ظلها.	-نشوء حزب البعث العربي الاشتراكي,ويضم.

4.3. Modélisation de la base BDSONARABE :

Nous présentons dans ce qui suit la modélisation de BDSONARABE.
Les fichiers son segmentés en utilisant le logiciel PRAAT, l'image du spectrogramme, ainsi que le fichier correspondant aux formants seront stockés dans des répertoires. Ces fichiers formeront un sous ensemble (BD1) de la base BDSONARABE.

Les informations relatives à chaque fichier son d'unité acoustique définie et stockée dans BD1, décrivant les caractéristiques des locuteurs, les informations concernant les conditions et l'environnement de l'enregistrement, les informations des sons (pitch, durée, intensité, spectrogramme) seront stockées dans l'autre sous ensemble BD2 de BDSONARABE.

L'interface utilisateur permet une saisie des informations relatives aux fichiers son après leur segmentation par PRAAT, une consultation permet la sélection ou choix du corpus suivant des spécificités déterminées par l'utilisateur, et permet la génération des fichiers segmentés, des images de spectrogrammes, des fichiers correspondant aux formants de chaque son .

La figure suivante illustre l'architecture de la BD BDSONARABE.

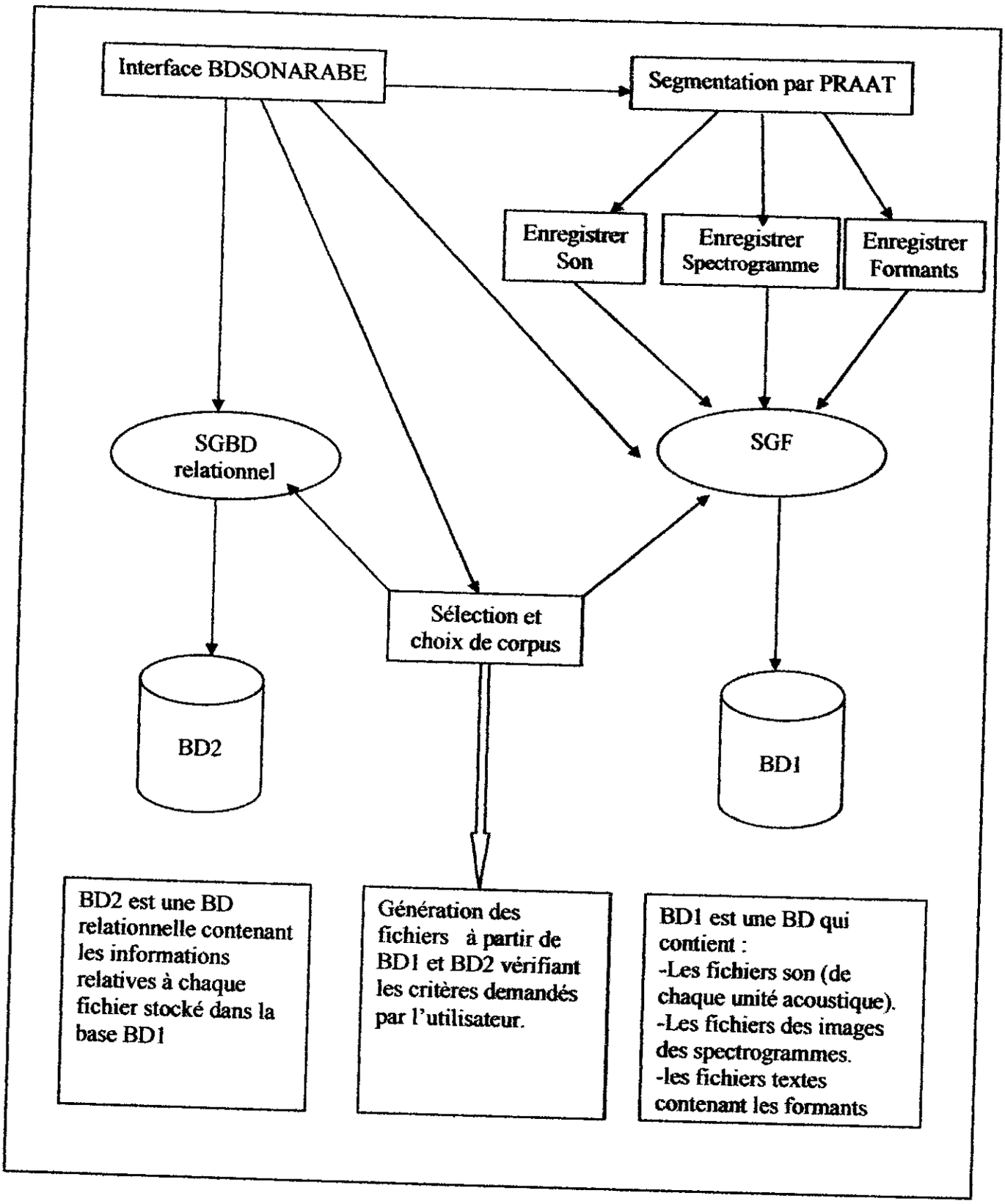


Fig 4.1 Schéma de modélisation de la BD BDSONARABE

Cette modélisation présente les avantages suivants :

- un accès rapide aux fichiers son stockés dans BD2, en profitant des avantages offerts par les SGBD relationnels. La consultation devient plus rapide vue que la taille de BD2 est réduite ;
- une optimisation de l'espace disque par le stockage unique des informations relatives aux corpus ;
- la possibilité de la répartition des données de BD1 sur plusieurs sites en gardant une vue générale sur toutes les informations relatives à ces données dans BD2 ;

Pour l'incorporation de la BDSONARABE dans un réseau d'ordinateur les données de BD1 peuvent être distribuées dans les différents postes du réseau, tandis que la BD2 sera implantée dans les postes pour que le volume de données dans BD2 soit moindre que celui de BD1.

La figure 4.2. illustre un schéma global explicatif .

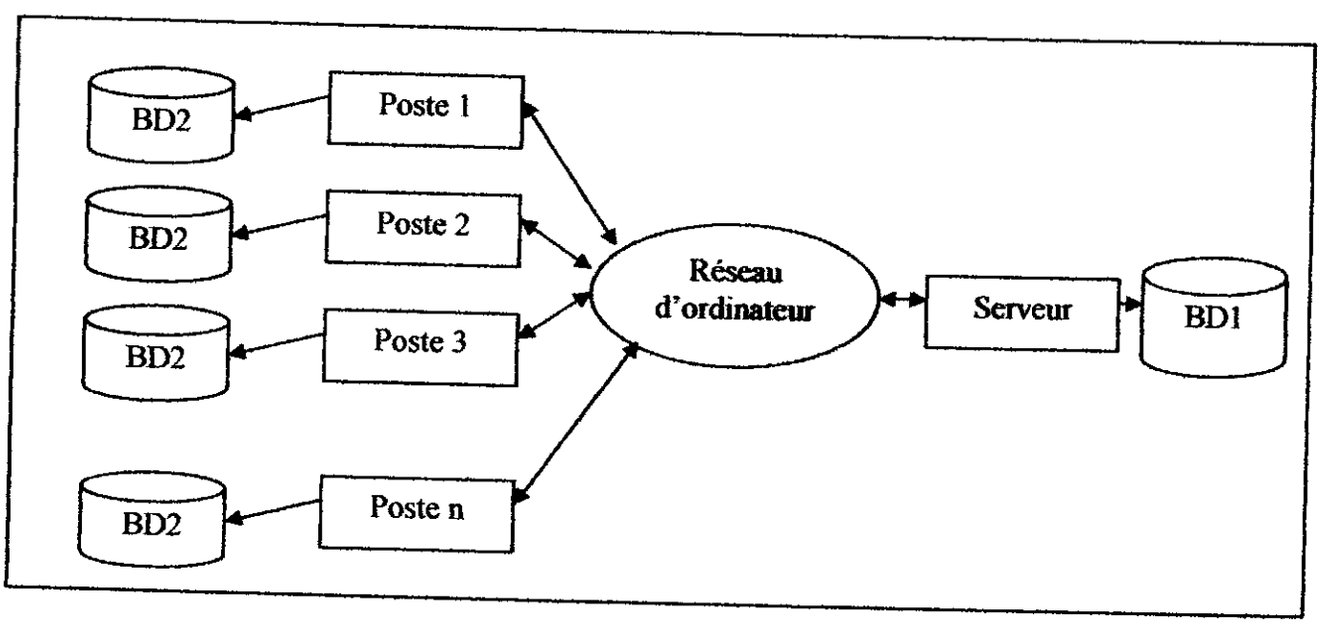


Fig 4.2 Mise sous un réseau de BDSONARABE

4.3.1. Description de BD1

Chaque fichier son, et les fichiers images des spectrogrammes, ainsi que les fichiers textes représentants les formants du signal vocal sont stockés dans un même répertoire. Un répertoire peut contenir par exemple, tous les fichiers d'une même catégorie de corpus (exemple : phonème, diphone, syllabe ...etc.).
Exemple :

E:\BDSONARABE\phonème\ba1\son.wav. // Fichier son du phonème [ba1].
E:\BDSONARABE\phonème\ba1\info.txt. // Les formants du phonème [ba1].
E:\BDSONARABE\phonème\ba1\praat.emf. // Fichier image du spectrogramme
E:\BDSONARABE\syllabe\ba3\son.wav. // Fichier son du syllabe [ba3].

4.3.2. Conception de BD2 :

Avant de décrire le schéma relationnel de BD2, nous avons procédé à la formalisation des données au niveau conceptuel en MCD (Modèle Conceptuel de Données) qui est la représentation schématique conforme à la sémantique des liens entre les données, et basé sur le formalisme individuel utilisant la notion d'entité et relation pour traduire les liens sémantiques de l'ensemble des données manipulées.

4.3.3. Dictionnaire des données conçu pour BD2 :

Code	Libellé	Type	Taille
mat_loc	Matricule du locuteur	Texte	8
Nom_loc	Nom du locuteur	Texte	20
Pren_loc	Prénom du locuteur	Texte	20
Sexe_loc	Sexe du locuteur	Texte	1
An_nais	Année de naissance du locuteur	Entier	4
Taille_loc	Taille du locuteur	Texte	4
Poid_loc	Poids du locuteur	Texte	5
Adr_loc	Adresse du locuteur	Texte	25
Tel_loc	Téléphone du locuteur	Texte	12
Mat_exp	Matricule de l'expert en segmentation	Entier	4
Nom_exp	Nom de l'expert en segmentation	Texte	20
Prén_exp	Prénom de l'expert en segmentation	Texte	20

Cod_lang	Code de la langue natale	Entier	2
Des_lang	Désignation de la langue natale	Texte	20
Cod_vil	Code de la ville	Entier	2
Lib_vil	Libellé de la ville	Texte	20
Cod_pay	Code du pays	Entier	2
Lib_pay	Libellé du pays	Texte	20
Cod_mil	Code du milieu d'enregistrement	Texte	2
Lib_mil	Libellé du milieu d'enregistrement	Texte	20
Cod_c_t	Code des caractéristiques temporaires	Texte	2
Lib_c_t	Libellé des caractéristiques temporaires	Texte	20
Cod_ph	Code du phonème	Texte	5
Dur_ph	durée du phonème	Réel simple	8
Pitch_ph	pitch du phonème	Réel simple	8
Int_ph	Intensité du phonème	Réel simple	6
Cod_syl	Code du syllabe	Réel simple	8
Dur_syl	durée du syllabe	Réel simple	8
Pitch_syl	pitch du syllabe	Réel simple	8
Int_syl	Intensité du syllabe	Réel simple	6
Cod_path	Code de la pathologie	Texte	2
Lib_path	Libellé de la pathologie	Texte	30

- **Liste des relations type**

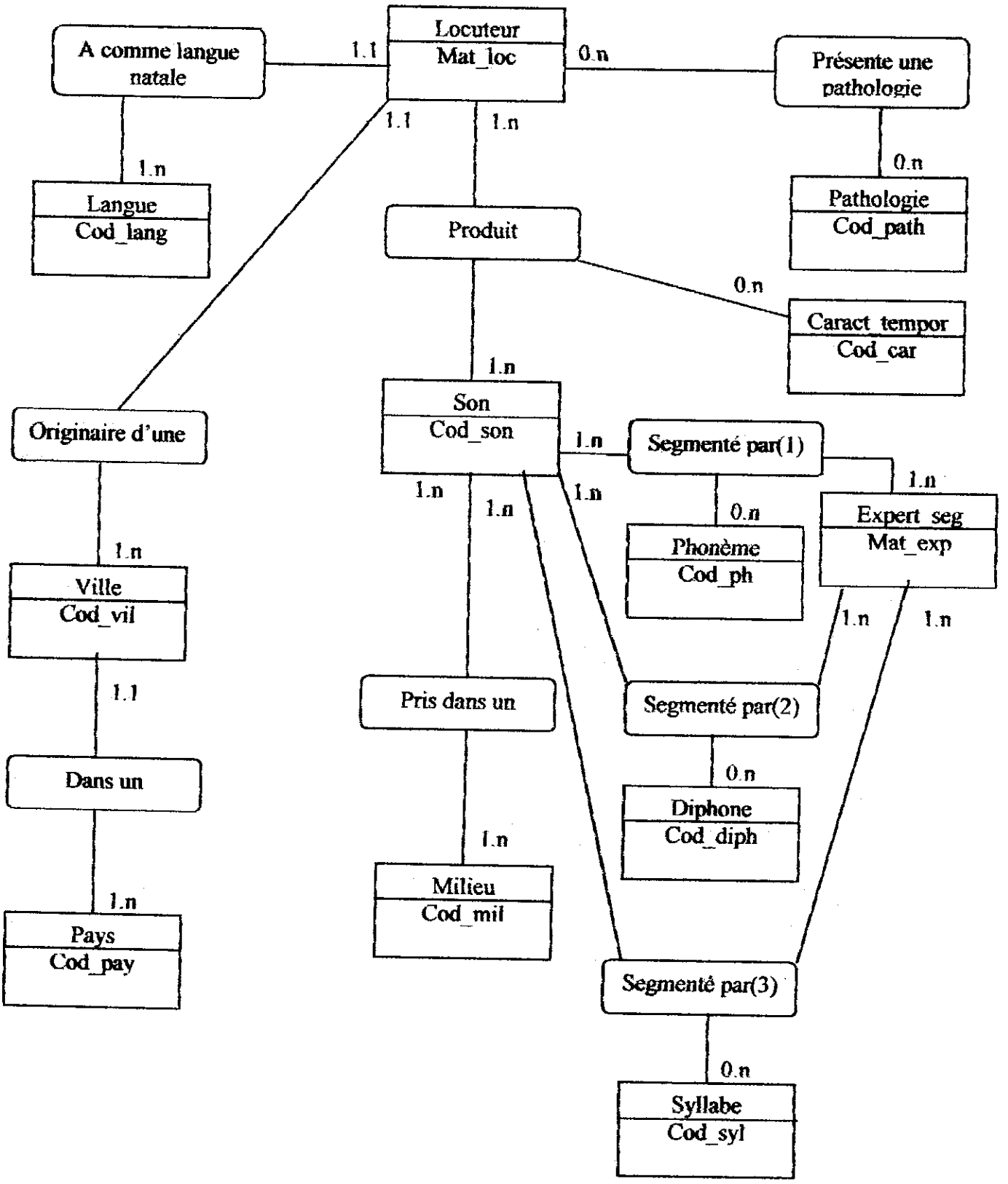
Relation	Identifiant	Propriétés
Produit_par	Mat_loc,cod_son, cod_car	-
Etiqueté par	Cod_son,mat_exp	-

- **Liste des individus types**

Individu	Identifiant	Propriétés
Locuteur	Mat_loc	Nom_loc, pren_loc, Sexe_loc An_nais,Taille_loc,Poid_loc,Adr_loc Taille_loc, Poid_loc, Adr_loc, Tel_loc

Expert	Mat_exp	Nom_exp, Prén_exp
Langue	Cod_lang	Lib_lang
Ville	Cod_vil	Lib_vil
Pays	Cod_pay	Lib_pay
Milieu	Cod_milieu	Lib_milieu
Pathologie	Cod_path	Lib_path
Son	Cod_son	
Caractéristique temporaire	Cod_car	Lib_car
Phonème	Cod_ph	Dur_ph, Pitch_ph, Int_ph
Syllabe	Cod_syl	Dur_syl, Pitch_syl, Int_syl
Diphone	Cod_diph	Dur_diph, Pitch_diph, Int_diph
Unité acoustique	Cod_unit	Lib_unit

4.4. Modèle conceptuel des données conçu pour BD2 :



4.5. Codification

Un code est une combinaison de symboles appartenant à une liste, il doit vérifier la non ambiguïté, l'adaptation à l'utilisateur, l'interprétation facile, l'extensibilité ...etc.
Nous avons utilisé les codifications suivantes :

- le locuteur est codifié par un code composé de l'année de naissance sur deux positions , son sexe sur une position, et d'un numéro séquentiel.
  Exemple : le locuteur 78M01.
- le son est codifié à partir du code de locuteur suivi du code de milieu d'enregistrement avec un numéro séquentiel.
  Exemple : le son 78M01BR01.
- le milieu d'enregistrement est représenté par les deux premières lettre du milieu. Exemple BR : milieu bruité, et CA : milieu calme.
- les entités langue, pays, ville, expert : sont codifiées par un numéro séquentiel.
  Exemple :
  01 : langue Arabe ; 03 : pays Algérie ; ... etc.
- la pathologie est codifié par un numéro séquentiel, elle peut être causée par :
  - un désordre articulatoire (distorsion, omission, ou substitution des sons de la parole) ;
  - un désordre de résonateurs (lésion des cavités orale, ou nasale ou pharyngale) ;
  - un désordre de la voix (infection des cordes vocales, par exemple le dévoisement d'une consonne voisé ... etc.) ;
  - une mauvaise maîtrise de la langue (exemple faire des pauses irrégulières, difficulté de trouver les mots de la langue, ...etc.),
  - un désordre dans le rythme (exemple omission au niveau des fins des énoncés, ...etc.)[18].
- L'ensemble d'unités acoustique est codifié de la manière suivante :

Le code de la lettre suivi de la voyelle et de la position.
Exemple :
ba1 :c'est le phonème [ب] avec la voyelle fatha et dans la position initiale.
to2 : c'est le phonème [ت] avec la voyelle damma et dans la position médiane.
si3 : c'est le phonème [س] avec la voyelle kasra et dans la position finale.

4.6. Choix du langage de programmation

Dans toute branche de l'ingénierie, les outils communément disponibles jouent un rôle considérable. Parmi ces outils, le langage de programmation qui occupe une place sans doute importante dans le domaine d'informatique. D'une importance capitale pendant les phases de réalisation et la maintenance du logiciel, son choix devient par ce fait très délicat.

Pour notre projet, le langage de programmation choisi est le *Visual Basic 6.0.*

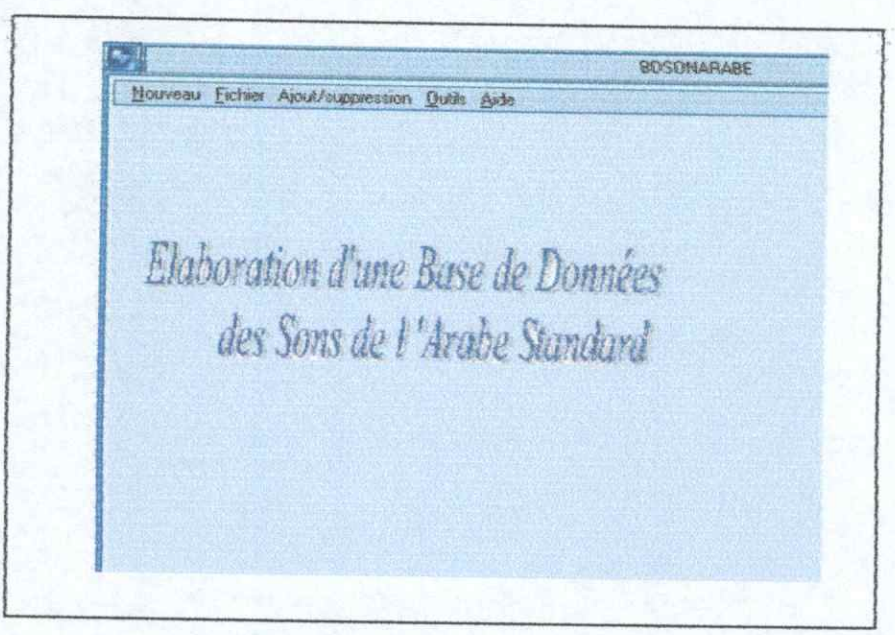
Présentation du langage de programmation

Visual Basic (VB) est un outil puissant pour le développement des applications windows, ses possibilités orienté objet et son approche basée sur les composants permettent d'améliorer la réutilisation du code VB.

Il associe la vitesse et la convivialité d'un environnement de développement visuel à la puissance et à la souplesse d'un langage objet au compilateur le plus rapide au monde et à une technologie de base de données de pointe.

4.7. Description du logiciel BDSONARABE

La fenêtre principale de notre application

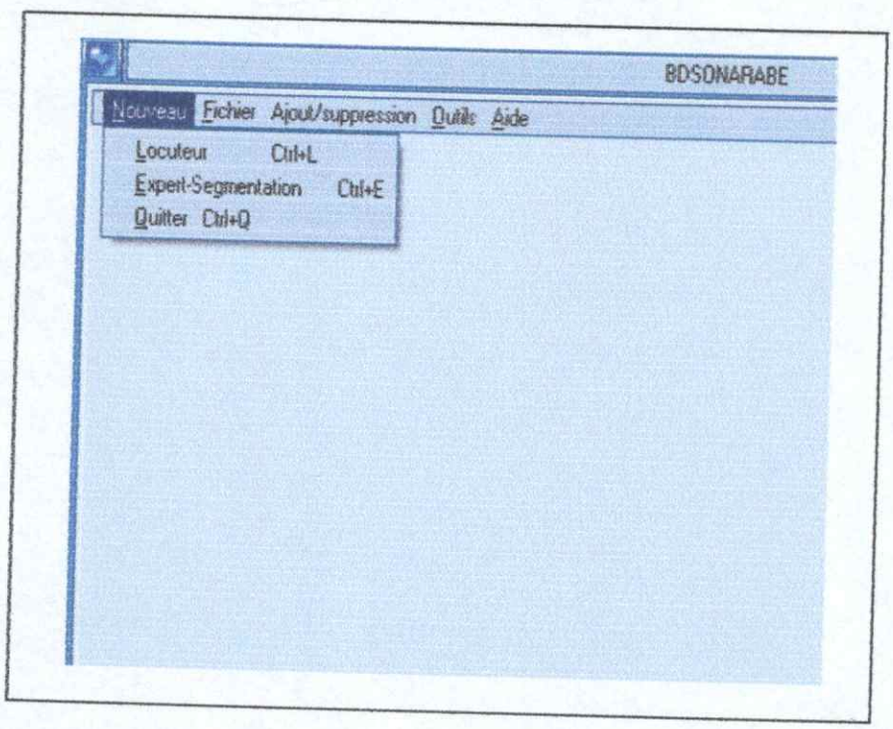


On commence par le menu Nouveau : il contient les fonctions suivantes :

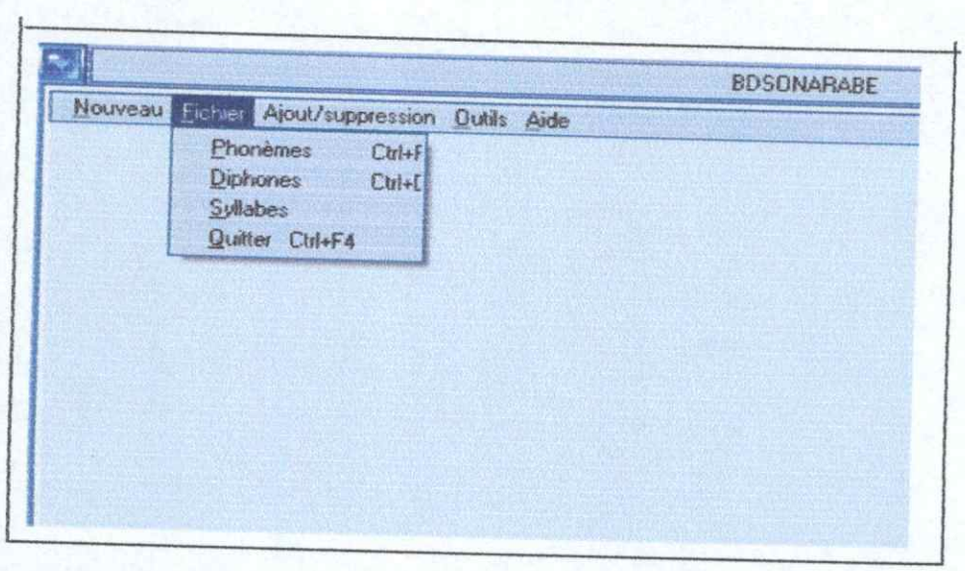
Accéder à la fenêtre locuteur

Accéder à la fenêtre expert en segmentation

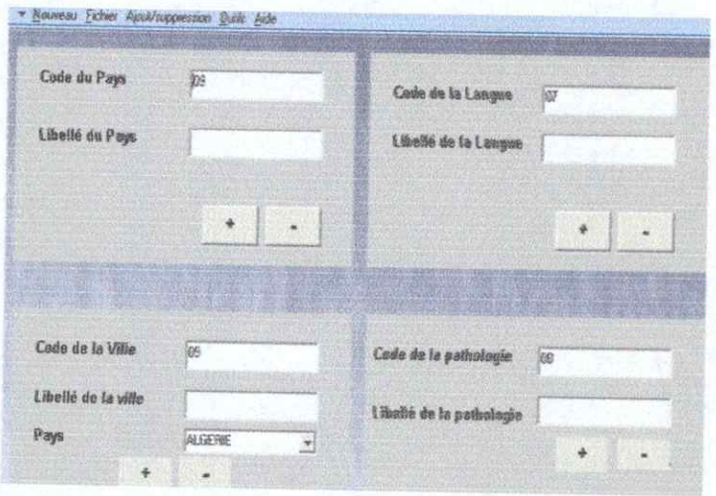
Quitter l'application



le menu fichier permet de consulter les fichiers correspondants aux unités acoustiques (phonèmes, syllabe, ...etc).

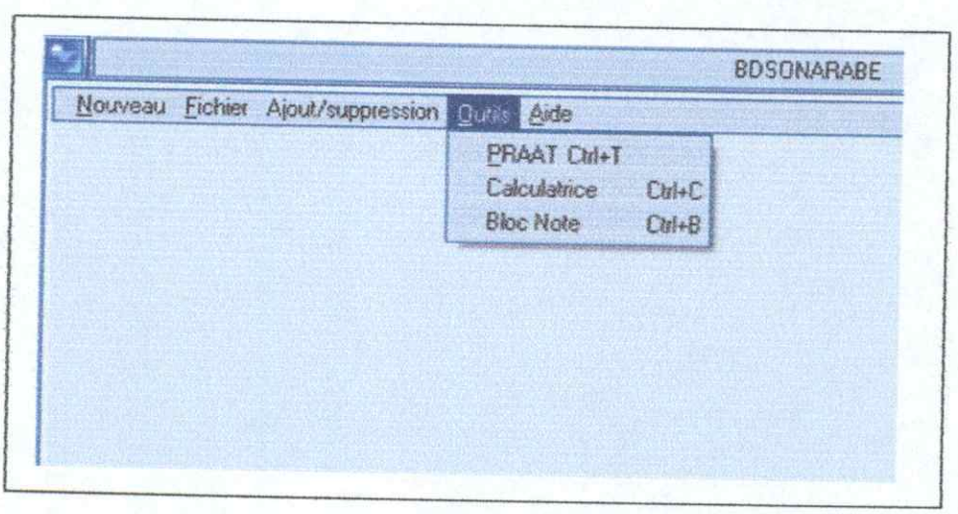


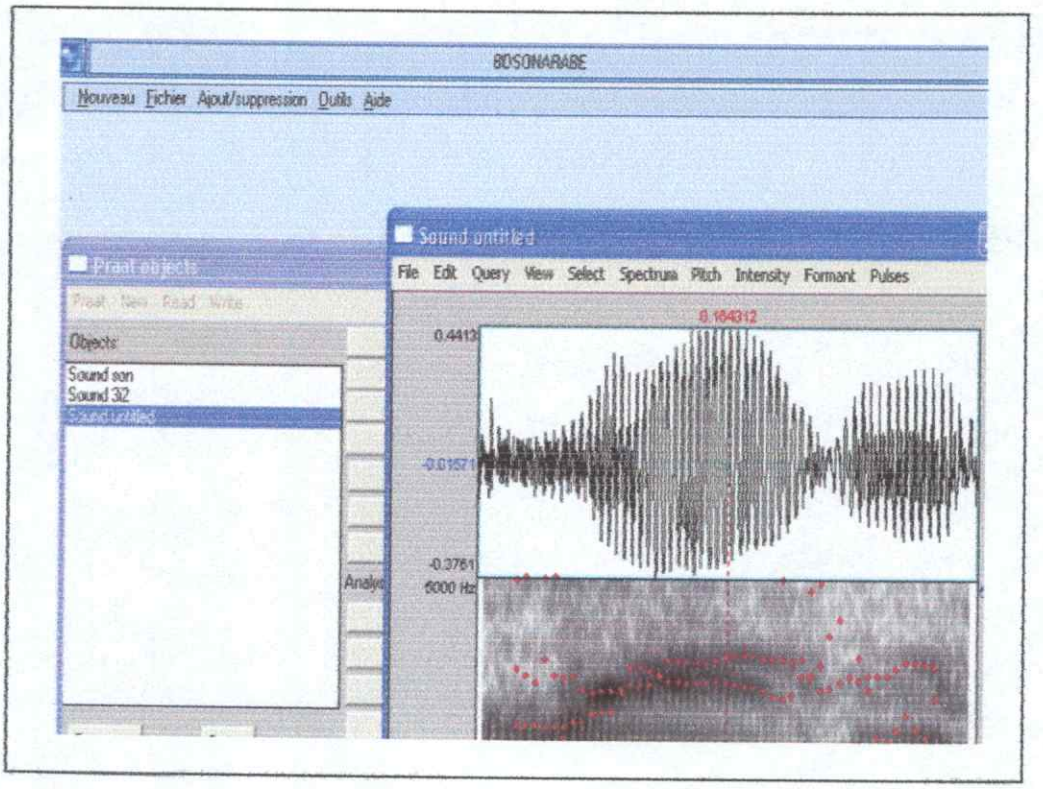
le menu Ajout/Suppression permet d'ajouter et de supprimer des pays, des villes ; des langues, et des pathologies.



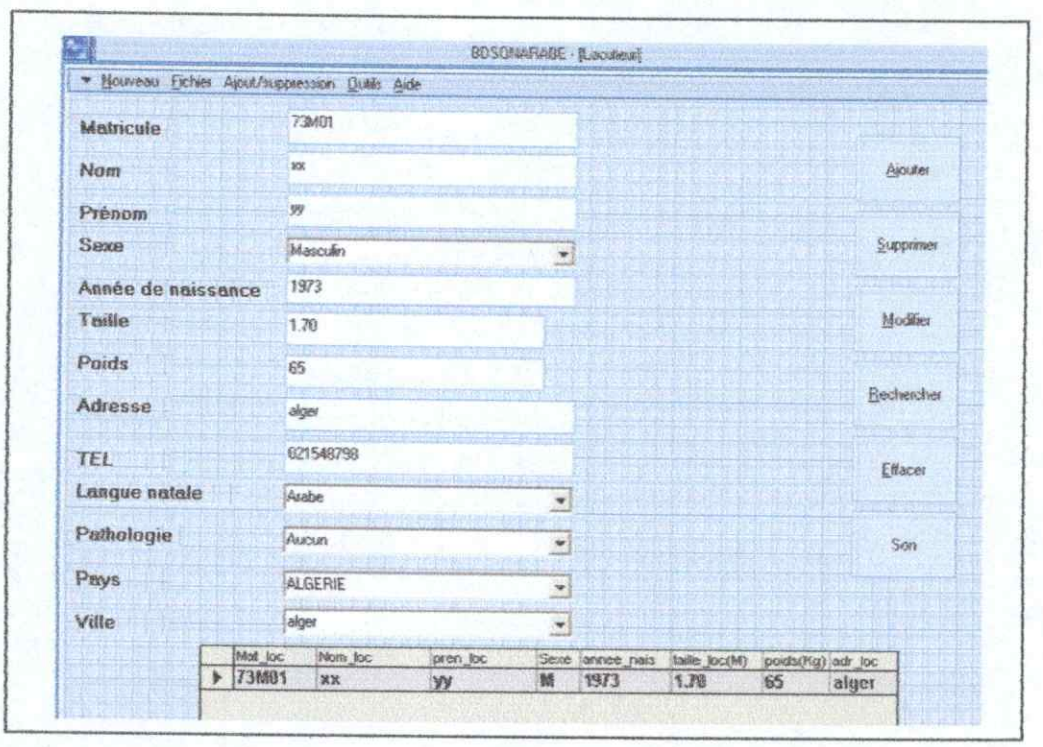
Le menu Outils :

- Praat : Fait appel au logiciel d'analyse Paat
- Calculatrice : Affiche la calculatrice de windows
- Bloc note affiche le bloc note



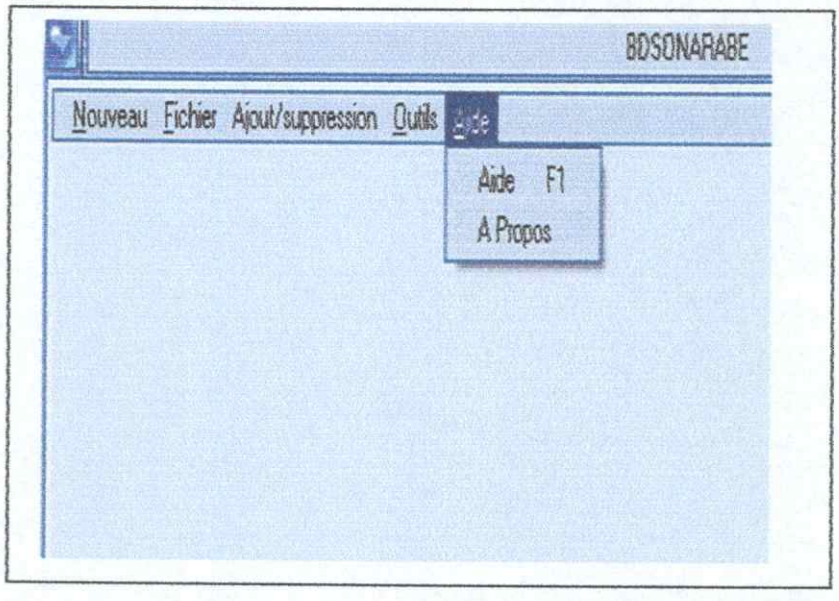


La fenêtre nouveau locuteur permet d'ajouter de supprimer, de modifier de rechercher, des locuteurs. et d'y accéder par la suite à la fenêtre d'enregistrement de son.



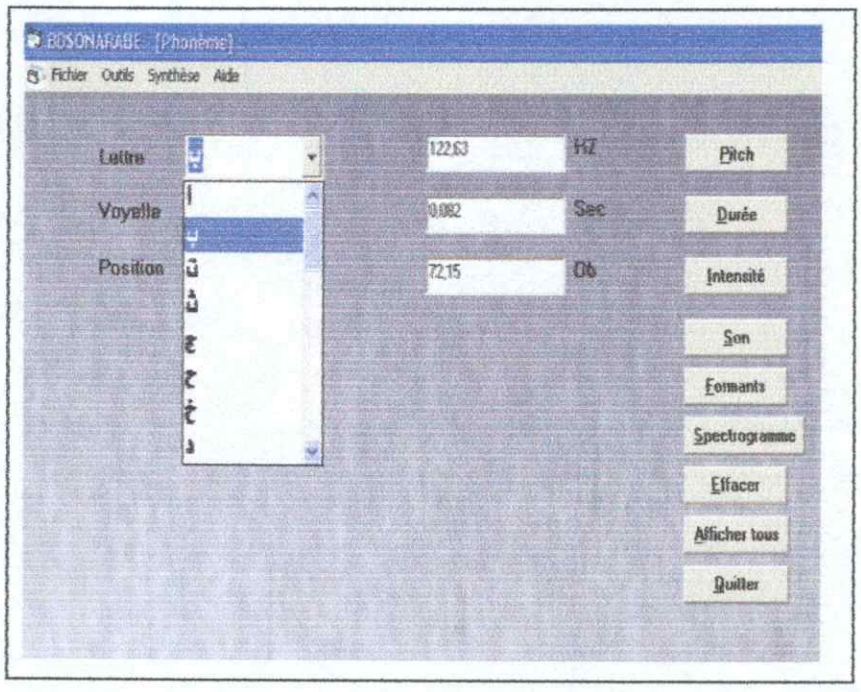
la fenêtre de son permet d'ajouter un son de le supprimer de rechercher l'ensemble des sons enregistrés par chaque locuteurs selon des caractéristique précises.

Le menu aide affiche l'aide du logiciel et l'à propos.



La fenêtre Phonème du menu Fichier contient les boutons suivants :

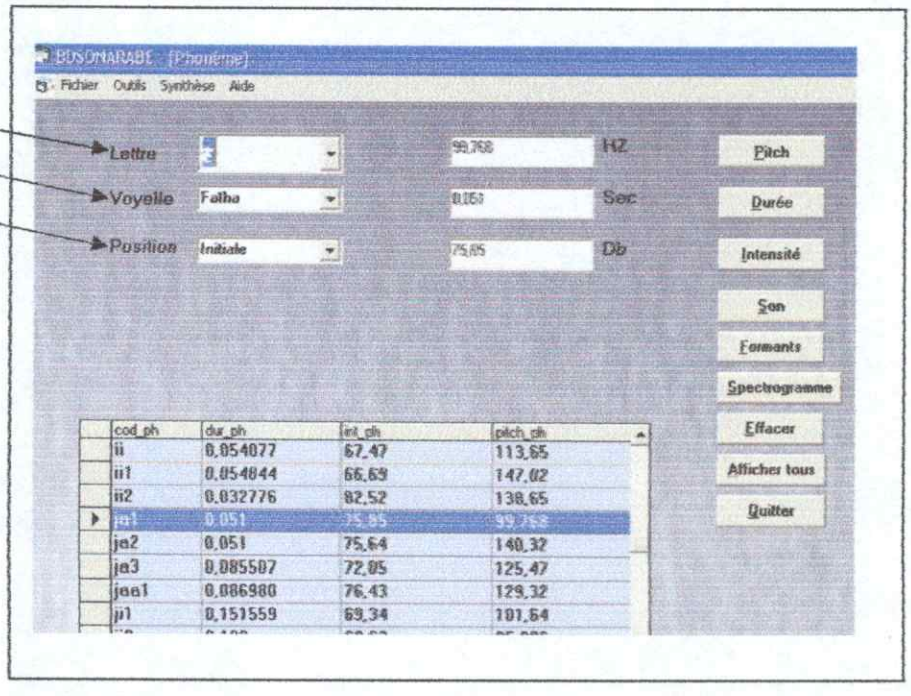
- Pitch : permet d'afficher le pitch du phonème sélectionné
- Durée : affiche la durée du phonème sélectionné.
- Intensité : affiche l'intensité
- Son : permet d'entendre le son du phonème en arrière plan
- Formants: affiche les formants du phonème
- Spectrogramme : affiche l'image du spectrogramme du phonème
- Effacer : permet d'effacer toutes les zones de texte
- Afficher tous : affiche les phonèmes dans un tableau pour lancer une recherche plus simple.
- Quitter : Fermer l'application.



Choisir la lettre

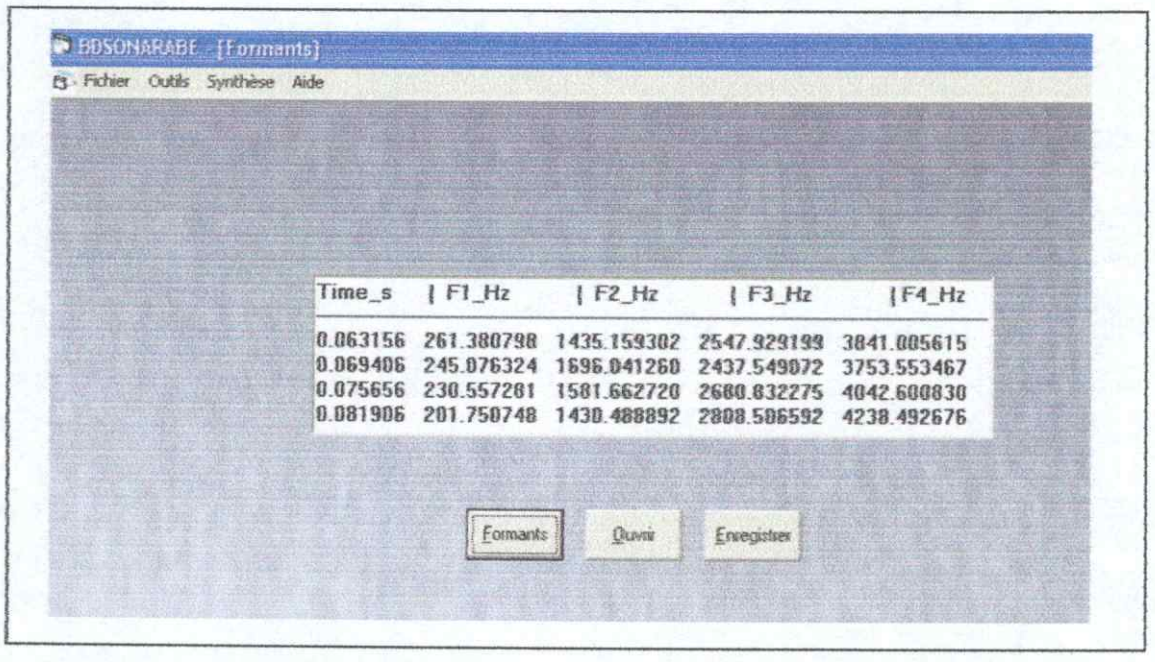
Choisir la voyelle

Choisir la position



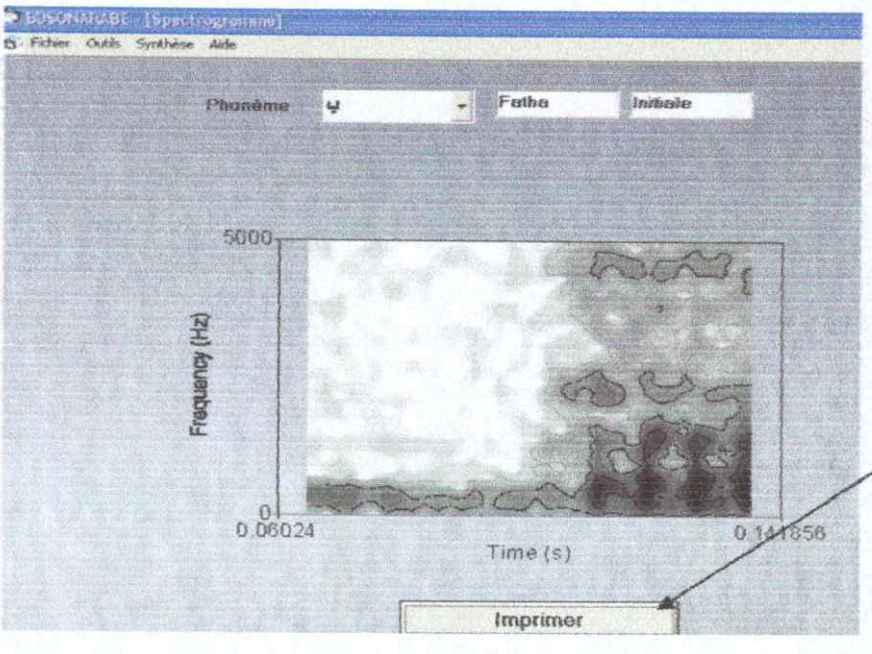
Remarque : Les fenêtres diphones, syllabes portent les mêmes fonctions que la fenêtre phonèmes.

En cliquant sur le bouton formant la fenêtre formants apparaîtra comme suit :



Dans cette fenêtre, on peut ouvrir d'autres formants pour d'autres phonèmes, comme on peut sauvegarder les formants affichés là ou on veut, avec une extension (.txt)

Pour le bouton spectrogramme sa fenêtre c'est la suivante :



Imprimer le Spectrogramme

4.8. Conclusion

L'application de BDSONARABE et toutes les fonctions qu'elle regroupe étaient des fonction de consultation et de MAJ c'est-à-dire que l'utilisateur peut ajouter une unité acoustique, la modifier la supprimer et la consulter, cette BD sera utilisée dans le domaine de traitement automatique de la parole de sorte que les gens qui font la reconnaissance ont besoin des informations tels que les formants, le pitch, la durée l'intensité sans refaire la segmentation. Même chose pour ceux qui travaillent sur la synthèse par concaténation d'unités acoustiques, ils ont besoins des mêmes informations dans leurs travaux.

Conclusions Générales et perspectives

Les BD de parole peuvent être intégrer dans les projets de traitement automatique de la langue, pour donner aux utilisateurs de la matière linguistique bien structurée, dont les conditions de production et d'acquisition sont clairement connue, ces BD peuvent être utilisée pour organiser la recherche fondamentale d'une manière simple et cohérente.

Nous avons dans notre travail élaboré un environnement d'enregistrement, d'analyse de stockage et de gestion de corpus. Nous avons utilisé un corpus constitué de 76 phrases et 15 mots isolés contenant les phonèmes, les diphones et les syllabes de l'AS dans différentes positions (Initiale Médiane et Finale) et différents contextes vocaliques (Fatha, Damma et kasra).

Une segmentation semi-automatique du corpus s'avère difficile à cause du problème de la variabilité de la parole (intra-locuteur, interlocuteur et contextuelle) cela nécessite de bonnes connaissances en phonétique et en phonologie.

A partir d'une analyse sonagraphique à l'aide du logiciel PRAAT nous avons extrait les paramètres pertinents du signal vocal la fréquence fondamentale(f_0), la durée, l'intensité, les formants, le son, ainsi que les images des spectrogrammes des unités acoustiques .Ces paramètres sont enregistrés dans une BD, cette dernière peut être manipulée de façon claire par l'utilisateur.

Nous suggérons des perspectives à notre travail telles que :

- L'élaboration du corpus : utilisation d'un corpus de grande taille en mode multi locuteur.
- Outil d'analyse : utiliser un autre outil d'analyse, et pouvoir de ce fait comparer entre les outils.
- Ajouter d'autres unités acoustiques, tel que les triphones, demi-syllabes, etc...

Références Bibliographiques

Références Bibliographiques

[01] www.lecerveau.mcgill.ca/flash/capsule/outil_bleu21.htm

[02] www.leger-transport.com/claudio/audition.html

[03] Calliope, La parole et son traitement automatique, Edition Masson, Paris 1989.

[04] R. Boite, M. Kunt, Traitement de la parole, presses polytechniques Romandes, 1987.

[05] L. Buniet, Traitement Automatique de la parole en milieu bruité : Etude de modèles connexionnistes statiques et dynamiques, Université Henry Poincaré Nancy France .Thèse de Doctorat 1997.

[06] S. Baloul,développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé , Université de Maine, France. Thèse de Docrtorat 2003

[07] www.fr.wikipedia.org

[08] S. Nefti, Segmentation Automatique de la parole en phones.Correction d'étiquetage par l'introduction de mesure de confiance, Université de Rennes1, France.thèse de Doctorat 2004

[9] www.isacolondecarvajale.perso.cegetel.net/maitrise_isacolon2004.pdf

[10] www.weblex.ens-lsh.fr/projets/xitools/logiciels/praat/praat.htm

[11] www.commentcamarche.net/bdd/bddintro.php

[12] www.lsi.supelec.fr/www/yb/poly_bd/poly9.html

[13] www.i3s.unices.fr/~crescen/publications/ofldb-essi3-rapport2002-04.pdf

[14] www.lbdwww.epfl.ch/e/teaching/slidesbda/support/oodb/bdoo.pdf

[15] www.fr.wikipedia.org/wiki/base-de-données-multimédia.html

[16] www.ebabylone.com/encyclopedie_Base_de_donn%E9es_multim%E9dia.html

[17] C. Barras, Reconnaissance de la parole continue adaptation au locuteur et contrôle temporel dans les modèle de Markov Caché, thèse de Doctorat. Université de Paris France 1996

[18] H. Khelifet, A. Brahmi, contribution à l'élaboration d'une Base de données de son de la langue Arabe BDSONs_A, PFE Informatique. Institut National de formation en Informatique(INI), 1997/1998.