N° d'ordre : 08/2013-D/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

FACULTE D'ELECTRONIQUE ET INFORMATIQUE

THESE

Présentée pour l'obtention du grade de **DOCTEUR EN SCIENCES**
En ELECTRONIQUE
Spécialité: Communication Parlée

Par **GHANIA DROUA-HAMDANI**

Sujet :

ENVIRONNEMENT POUR LES BASES DE DONNEES SONORES EN LANGUE ARABE : APPLICATION A LA VALIDATION D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE

Soutenue publiquement le 02/05/2013, devant le jury composé de :

SAYOUD Halim	Professeur	à l'USTHB	Président
BOUDRAA Malika	Professeur	à l'USTHB	Directeur de thèse
SELOUANI Sid-Ahmed	Professeur	à l'U. Moncton	Co-directeur de thèse
GUERTI Mhania	Professeur	à l'ENP	Examinateur
BELOUCHRANI Adel	Professeur	à l'ENP	Examinateur
GUESSOUM Ahmed	Professeur	à l'USTHB	Examinateur
BOUSBIA Salah-Hichem	Maître de Conférences/A	à l'ENP	Examinateur

Remerciements

Un travail de thèse se déroule sur plusieurs années qui sont à la fois une période de travail intense, de vie, de rencontres, d'échanges mais aussi de moments de doute. C'est pourquoi je tiens à exprimer mes remerciements et ma gratitude à toutes les personnes ayant apporté l'aide et l'assistance nécessaire à l'élaboration de ce travail.

Mes remerciements s'adressent en particulier au Pr. M. Boudraa, pour avoir accepté de m'accueillir au sein de son laboratoire et de son équipe. Je la remercie aussi pour son soutien et sa disponibilité tout au long de la réalisation de ce projet.

Mes plus vifs remerciements s'adressent aussi au Pr. Selouani pour son dynamisme et son infatigable énergie ainsi que pour le temps qu'il a bien voulu me consacrer durant ces longues et pénibles années.

Je souhaite également remercier les membres du jury pour avoir accepté de lire ce document, de l'évaluer et d'apporter les corrections nécessaires pour améliorer sa qualité.

Mes remerciements vont également au directeur, du Centre de Recherche Technique et Scientifique pour le Développement de Langue Arabe, Mr Rachid Benmalek, à la doyenne de la Faculté des Lettres Arabes d'Oran Mme Tahri et au doyen de la Faculté des Sciences Humaines et Sociales de Tlemcen Mr Saidi qui ont mis à ma disposition tous les moyens possibles pour que la collecte des enregistrements soit moins contraignante.

Mes remerciements vont aussi à mes collègues et amis du laboratoire de Traitement Automatique de la Parole (CRSTDLA) Mrs Khoudir Benbellil et Kamel Ferrat pour leurs encouragements permanents.

De même, je remercie Melle Lebal et Mr Bey Ahmed du laboratoire de Communication Parlée de l'Institut d'Electronique ainsi que tous les locuteurs, à travers le territoire national, qui ont prêté leurs voix et accepté sans hésitation de faire partie de cet ambitieux projet, en particulier le personnel de Nokia Siemens Network Algérie (NSN) d'Oran et de Constantine.

Je remercie, de tout mon cœur, ma famille adorée qui m'a toujours soutenue afin de ne jamais dévier de mon objectif final.

Merci Encore.

A la Mémoire de ma Très Chère Mère
et à mon Père Adoré.

"La plus grande gloire n'est pas de ne jamais
tomber, mais de se relever à chaque chute."
Confucius

ملخص

تتناول هذه الرسالة تصميم قاعدة بيانات صوتية باللغة العربية الفصحى المنطوقة في الجزائر (ALGASD). و قد نُطقت نصوص هذه القاعدة من قبل 300 متكلم جزائري اختيروا من 11 منطقة. إن أحد أهداف هذه المدونة الصوتية هوتمثيل اللهجات الإقليمية للغة العربية الفصحى الحديثة المتداولة في الجزائر. كما تحتوي على معلومات مفيدة عن المتكلمين مثل العمر والجنس ومستوى التعليم. وقد استعملت هذه المدونة في عدة دراسات منها الدراستين التجريبيتين التاليتين:

تتناول الدراسة الأولى اللإيقاع في الكلام حيث تمّ تمثيله حسب عوامل التغيرالتالية: العمر والجنس ومستوى التعليم ل 66 متكلم. وقد بينت مقارنة النتائج الصوتية و الإحصائية على مستوى اللغة المتداولة أي العربية الفصحى على اختلافات أساسية ما بين المتكلمين لم يشارإليها من قبل . كما بينت المقارنة على مستوى الدولي (اللغات الأجنبية و اللهجات العربية الشرقية والغربية) على أن إيقاع الكلام في النطق الجزائري للعربية الفصحى يختلف عن ما هو مورود في الدراسات النظرية. هذه النتائج هامة جدا يمكن استخدامها في ميادين شتى كالإستكشاف الآلي للغات و الإستكشاف الآلي للكلام ، إلخ.

أما الدراسة الثانية فإنها تتعرض إلى نتائج تطبيقية في الاستكشاف الآلي للكلام المستمر. فاستعرضنا أهم الخطوات المتبعة في تصميم نموذج أحادي الصوتيات لمعالجة التغير اللغوي على أساس (HMMs). و قد استخدمنا لذلك كل التسجيلات الصوتية الموافقة ل 6 مناطق من قاعدة البيانات الصوتية (ALGASD). نتائج اختبار جهاز الاستكشاف الآلي للكلام للنظام المرجعي مرضية جدا. ويمكن بذلك اتخاذه كأساس لأنظمة الاستكشاف الآلي للكلام المخصصة للغة العربية. وقد تطرقنا فيما بعد إلى تفصيل نتائج الاستكشاف لإظهار مدى تأثيرالإختلافات الإقليمية و الجنسية على أداءه. وقد أظهرت هذه التجربة تباين ما بين النتائج حسب المنطقة و جنس المتكلم. وأخيرا حللنا الأخطاء المنسوخة في النص النهائي الذي تم إنشاؤه بواسطة النظام.

Résumé

La thèse porte sur la conception et la réalisation d'une base de données sonores en Arabe Standard (ALGASD). Les textes de la base ont été lus par 300 locuteurs Algériens choisis parmi 11 régions présentant des variations régionales de prononciation. La ressource orale contient également des informations utiles sur les locuteurs comme l'âge, le genre et le niveau d'instruction. Le corpus sonore a été le fondement de plusieurs travaux expérimentaux comme ceux présentés ci-dessous :

La première étude réalisée traite la modélisation d'un paramètre prosodique : *le rythme de la parole*. Pour ce faire, nous avons considéré dans l'étude plusieurs facteurs de variabilité relevés sur 66 locuteurs. Les résultats acoustiques et statistiques montrent des différences significatives aux niveaux inter et intra langues. En effet, ils montrent que le rythme de l'Arabe Standard prononcé par des locuteurs Algériens diffère de ce qui est traditionnellement affiché dans les études théoriques. Ces résultats importants peuvent être utilisés dans plusieurs domaines comme dans : l'identification automatique des langues, la reconnaissance automatique de la parole, etc.

La deuxième étude concerne la réalisation d'un système de reconnaissance automatique de la parole continue et de sa validation avec le corpus ALGASD. L'expérience a été conduite avec les enregistrements de 6 régions en utilisant les HMMs. La performance du système est très satisfaisante. Cependant, les résultats détaillés des taux de reconnaissance par région et par genre montrent une variation selon ces deux sources de variabilité. De plus, une analyse des erreurs a été effectuée sur la transcription finale du système.

Abstract

The thesis focuses on the design of an Arabic speech database called (ALGASD). The texts are read by 300 Algerian speakers belonging to 11 regions that presenting pronunciation variation. The corpus also contains useful information such as: age, gender and level of education. ALGASD is the basis of several experimental studies:

The first study deals with the modeling of the speech rhythm (prosodic parameter). Several factors of variability are considered in the work. The acoustic and statistics results show significant differences at inter and intra languages levels. They show that the rhythm of Standard Arabic spoken by Algerian speakers differs from what is traditionally displayed in theoretical researches. Rhythm parameters can be used in: automatic identification of languages, automatic speech recognition, etc.

The second work concerns the development of an automatic speech recognition system of continuous speech using ALGASD corpus. The experiment is performed with recordings of six regions basing on HMMs models. The system performance is very satisfactory. However, the detailed results of recognition rates by region and by gender show variations. An analysis of transcription errors is performed.

Table des matières

Listes des Figures

Listes des Tableaux

Liste des Acronymes

ALGASD ALGerian Arabic Specch Database

ANOVA ANalysis Of VAriance

ANN Réseaux de Neurones Artificiels

AS Arabe Moderne Standard

C Consonne

Cc Corpus Commun

Ci Corpus individuel

Cr Corpus réservé

deltaV Ecart-types des intervalles vocaliques

deltaC Ecart-types des intervalles consonantiques

ELDA Evaluations and Langage Ressouces Distribution Agency

ELRA European Language Resources Association

GMM Gaussian Mixture Model

HMM Hidden Markov Models

HTK Hidden Markov Model Toolkit

IM Interval Measures

IPA International Phonetic Alphabet

LDC Linguistic Data Consortium

LPCC Linear Predictive Cepstral Coefficients

MFCC Mel Frequency Cesptral Coefficients

MIT Massachusetts Institute of Technology

OLA OverLap and Add

PCM Pulse Coded Modulation

PLP Perceptual Linear Predictive Analysis

PPE Phrases Phonétiquement Equilibrées

PVI Pairewise Variability Indices

RSB Rapport Signal-Bruit

SAMPA Speech Assessment Methods Phonetic Alphabet

SER Sentence Error Rate

SRAP Systèmes de Reconnaissance Automatique de la Parole

SRI Stanford Research Institute

TI Texas Instruments

TTS Text-To-Speech

percentV Proportion totale des intervalles vocaliques de la phrase

VarcoC Normalized consonantal Interval

VarcoV Normalized vocalic Interval

V Voyelle

WER Word Error Rate

INTRODUCTION GENERALE

Les corpora oraux constituent une assise de choix servant aussi bien à la supervision de la progression des recherches déjà établies qu'à l'émergence de nouvelles orientations et disciplines de recherches. De même, le besoin d'exploiter des corpora oraux dans les technologies de l'information (la synthèse automatique de la parole, la reconnaissance automatique de la parole, etc.) est une évidence marquée n'échappant pas à cette nécessité. Cependant, et sans aucune surprise, nous constatons que les ressources orales dédiées à la langue Arabe classique sont quasi inexistantes comparées à celles disponibles dans les autres langues. Et lorsqu'elles existent, elles se restreignent aux applications pour lesquelles elles ont été élaborées. Nous avons voulu, par la réalisation d'une base de données sonores *ALGerian Arabic Speech Database* (ALGASD), à partir d'enregistrements de locuteurs Algériens, apporter une modeste contribution, concrétisée dans l'uniformisation d'un corpus oral servant principalement à appuyer des travaux de recherches, dont la préoccupation principale est l'Arabe Standard, qui est rappelons-le la langue fédératrice de plusieurs millions d'Arabophones.

ALGASD se caractérise par de nombreux aspects comme : une bonne qualité des enregistrements, un grand nombre de locuteurs et des facteurs reflétant de nombreuses sources de variabilité comme : la région d'appartenance, l'âge, le genre et les niveaux d'instruction. Cette richesse, matérialisée par le nombre important d'enregistrements et la diversité des profils de ses locuteurs, lui confère la position de substrat idéal pour entreprendre différents axes de recherches tels la prosodie de la langue et la reconnaissance automatique de la parole. En effet, elle a été le fondement de deux études qui traitent respectivement la classification du rythme de l'Arabe Standard parlé par les Algériens et la réalisation d'un système de reconnaissance de la parole continue mutlilocuteurs pour la langue arabe.

Nous nous sommes intéressés à l'organisation temporelle des énoncés car elle connaît, depuis quelques années, un regain d'intérêt à tous les niveaux (articulatoire, acoustique, etc.). En raison de son impact positif sur les performances des systèmes automatiques ayant la langue comme substrat de communication comme : la synthèse automatique de la parole, l'identification automatique des langues, la reconnaissance automatique de la parole. Les études expérimentales se rapportant à

cet aspect du signal parole sont nombreuses, notamment celles consacrées à l'analyse de la durée segmentale des unités sonores (et ce pour toutes les langues y compris pour l'Arabe). De même, les investigations portant sur le rythme des langues ont intéressé bon nombre de chercheurs. Cependant, les études expérimentales abordant le rythme de l'Arabe sont très peu fréquentes. Le corpus sonore utilisé dans cette analyse, ALGASD, offre un terrain expérimental idéal compte tenu du nombre important de locuteurs ainsi que des divers facteurs de variabilités engagés. L'étude du rythme de la parole consiste donc en l'application des corrélats de mesures pour analyser l'organisation temporelle (rythme) dans la prononciation de l'Arabe Standard (AS) par des locuteurs Algériens.

La dernière partie de la thèse concerne la validation d'un système de reconnaissance automatique multilocuteurs de la parole continue avec l'utilisation de la ressource orale ALGASD. La langue arabe a un nombre limité d'études et d'outils dans le domaine de la reconnaissance vocale. La recherche documentaire a montré que les études sur l'Arabe ont été conduites principalement sur la reconnaissance automatique des digits, des voyelles et des mots isolés. Toutefois, les travaux concernant le développement d'un système de reconnaissance automatique de parole continue (SRAP) de l'Arabe ont été abordés, récemment, par un nombre restreint de chercheurs. Pour notre part, nous avons réalisé un système de reconnaissance automatique de la parole continue indépendant des locuteurs. Ce sytème est basé sur les Modèles de Markov Cachés (MMC) notés aussi par (*Hidden Markov Models* (HMM)).

Notre thèse de recherche est organisée en cinq chapitres:

Chapitre 1

Nous avons exposé, dans ce chapitre, les principaux critères nécessaires à la conception d'une base de données sonores tels que : les textes utilisés, les locuteurs et leurs profils, les conditions et les paramètres d'enregistrement, etc. La présentation de ces critères est suivie du classement de quelques ressources orales notemment celles dédiées à la langue arabe et à la reconnaissance automatique de la parole. La seconde partie du chapitre aborde les fondements théoriques sur lesquels se base la reconnaissance automatique de la parole : l'extraction des données acoustiques (la paramétrisation), les modèles acoustiques et le module de langage. Nous avons exposé à la fin du chapitre deux algorithmes d'optimisation de la performance de la reconnaissance automatique.

Chapitre 2

Le corpus ALGASD a pour vocation de refléter les principales variations de prononciation de l'Arabe Standard des locuteurs algériens en tenant compte des

divers facteurs régionaux et sociaux de chacun. C'est ainsi que ce chapitre aborde dans le détail l'architecture suivie dans l'élaboration du corpus en commençant par citer les motivations et les objectifs visés.

Pratiquement, la réalisation d'une ressource orale nécessite un plan d'actions bien établi et une orchestration minutieuse tout au long de sa réalisation : choix des locuteurs (leurs profils, leurs répartitions, etc.), choix des régions selon la prononciation émergente, choix des corpora textuels et leurs distributions sur les régions et les locuteurs, etc.

Chapitre 3

Comme nous nous intéressons au rythme de la parole, nous avons voulu, par ce chapitre, introduire ce paramètre suprasegmental de la prosodie, en accentuant toutefois sur les modélisations utilisées pour sa mesure. Nous avons conclu le chapitre par des notions de statistique descriptive indispensables à ce type d'études.

Chapitre 4

Ce quatrième chapitre traite le rythme de la parole à travers plusieurs facteurs de variabilité (genre, age, etc.). Pour ce faire, une série d'analyses acoustiques et statistiques des paramètres rythmiques a été effectuée.

Après avoir étudié le rythme au niveau intra langue, c'est à dire, selon les variantes apportées par les locuteurs ; la seconde partie du chapitre traite le rythme au niveau inter langues permettant ainsi de situer l'Arabe parlé par les Algériens parmi les langues du monde.

Chapitre 5

Le chapitre 5 décrit la conception du système de reconnaissance automatique de parole continue de l'Arabe Standard. Ce système, fondé sur le principe des Modèles de Markov Cachés -modèles monophones-, satisfait plusieurs conditions : la parole continue, multilocuteurs, large vocabulaire et est surtout dédié à la reconnaissance de la langue arabe. Pour ce faire, les enregistrements de 6 régions d'ALGASD ont été utilisés. Les sections du chapitre aborderont les différentes étapes de développement. La performance du système est très satisfaisante.

Nous terminons la thèse par une conclusion générale et une ébauche des perspectives concernant les possibilités de recherches offertes par la base ALGASD.

Chapitre 1

RESSOURCES ORALES ET SRAP

1.1 INTRODUCTION

De nos jours, une grande majorité de recherches fondamentales ou appliquées se fondent sur l'exploitation de "corpora oraux". Les grands corpora servent principalement de substrat documentaire de la langue pratiquée. Lorsqu'ils sont échantillonnés, en tenant compte des régions et des données socioéconomiques et culturelles, ils permettent de guider les politiques linguistiques à grande échelle pour former une solide assise au développement et à la publication de nombreux ouvrages didactiques, servant à l'enseignement de la langue comme langue maternelle ou langue étrangère (exemple : le British National corpus et l'éditeur Collins pour l'Anglais). Ces corpora peuvent être aussi à la base de l'émergence de nouvelles disciplines comme l'analyse conversationnelle, ou bien, servir de base comparative concernant le langage pour : évaluer le langage des enfants à divers stades d'acquisition, diagnostiquer les pathologies du langage, évaluer le degré d'accomplissement dans l'acquisition des langues maternelles et étrangères, connaître l'effet des influences régionales, etc. Comme ils peuvent être aussi exploités dans les nouvelles technologies de l'information : dialogue homme machine, reconnaissance et synthèse de la parole, traduction automatique, etc. Le principal avantage de l'utilisation de corpora standards est de donner aux chercheurs la possibilité de comparer les performances des différentes techniques d'analyses, sur des données communes pour en déterminer les approches les plus prometteuses à poursuivre [Baude et al., 2006].

Nous présentons dans ce chapitre les caractéristiques et les principales ressources orales notamment celles dédiées à la langue arabe et à la reconnaissance automatique de la parole. Ensuite, nous aborderons le volet de la reconnaissance en donnant des notions fondamentales dans le domaine.

1.2 RESSOURCES ORALES

Cette première partie du chapitre vise à fournir quelques éléments nécessaires à l'élaboration ou l'enrichissement des corpora oraux pour anticiper ainsi certaines difficultés qui risquent d'entraver leurs réalisations et leurs exploitations ultérieures. En effet, des informations importantes à la constitution du corpus de données orales sont présentées, ainsi que quelques propositions, qui nous semblent intéressantes, concernant les aspects organisationnels et matériels relatifs aussi bien à la collecte du corpus qu'à la structuration et à la mise en forme des données. L'objectif étant de favoriser la création de corpora standards structurés utilisables et réutilisables.

1.2.1 Eléments d'une ressource orale

1.2.1.1 Corpus textuel

Les bases de données sonores peuvent être enregistrées à partir de textes lus ou de conversations spontanées dans des domaines contrôlés ou non. Le terme domaine désigne le contenu du discours (thème) ou les situations dans lesquelles une communication verbale prend place comme : en météorologie, domaine médical, etc.

1.2.1.2 Locuteurs

Le second paramètre indispensable à l'élaboration d'un corpus oral est la sélection des participants chargés de l'enregistrement de ce corpus, à savoir les locuteurs. La précision de leurs nombres, de leurs caractéristiques (profils) et de leurs distributions sont donc des éléments fondamentaux à considérer, notamment lorsqu'il s'agit d'un corpus dédié à une application dans le domaine du traitement automatique de la parole à savoir les systèmes de synthèse et de reconnaissance de la parole [Schiel et al., 2004].

En ce qui concerne les caractéristiques des locuteurs, elles se doivent d'être minutieusement documentées avant d'être récoltées. Bien que ces détails puissent sembler peu intéressants pour les locuteurs eux-mêmes, leur importance peut s'avérer fondamentale, notamment, dans des recherches sociologiques, linguistiques, etc. Ces caractéristiques peuvent porter principalement, dans l'ordre décroissant d'importance, sur :

- la distribution du genre et de l'âge des locuteurs,
- la langue maternelle (locuteurs natifs) et la langue du corpus ;
- la distribution des langues dialectales ;
- les niveaux d'instruction et la profession
- d'autres facteurs peuvent être aussi considérés comme les : pathologies du langage, accents étrangers, débits de parole, etc.

1.2.1.3 Conditions d'enregistrement

Les conditions d'enregistrement varient selon les utilisations escomptées du corpus. Ainsi le principe d'acquisition des signaux repose sur des critères imposés par l'enquêteur tels : la nature de la source à enregistrer (un/plusieurs émetteurs), le contexte, les perturbations sonores (bruit), la durée des entretiens, etc. De même, qu' il est chargé de choisir le matériel adéquat pour mener à bien cette opération.

Matériel Le matériel nécessaire pour enregistrer une ressource orale est principalement constitué de platines d'enregistrement et de microphones. Parmi les qualités requises de ce matériel, nous citons : la facilité d'utilisation, la robustesse, l'ergonomie, etc.

L'avènement technologique a permis diverses évolutions tant sur les caractéristiques techniques que sur les formes de ce matériel notamment sur les platines d'enregistrements. En effet, ces dernières peuvent être : analogiques, numériques, portables sur cartes mémoire, etc. En revanche, la technologie de base des microphones est restée inchangée depuis ses origines et leurs choix semblent, de ce fait, occuper une position non négligeable dans une prise professionnelle de sons. En effet, la spécification technique des microphones est nécessaire car ils sont choisis selon : la bande passante (l'intervalle des fréquences captées (50Hz-20kHz)), la dynamique, la directionnalité et la sensibilité (impédance).

Lieux de l'enregistrement De même que pour le matériel, l'enquêteur est chargé de trouver les lieux où vont se dérouler les séances d'enregistrements. Plusieurs possibilités s'offrent à lui : les enregistrements en intérieur (la chambre sourde ou autre) ou les enregistrements en extérieur. Dès que l'on n'est plus dans les conditions qu'offre une chambre sourde, l'enregistrement en intérieur rencontre deux difficultés principales auxquelles l'enquêteur doit remédier : la réverbération et les bruits parasites (climatiseur, grincement de chaise, manipulation de documents ou autres, etc.).

Conditions techniques d'enregistrement Les conditions techniques d'enregistrement concernent les paramètres techniques choisis pour la numérisation, à savoir :

- fréquence d'échantillonnage fe,
- choix du codage des données : le plus utilisé est le PCM (*Pulse Coded Modulation*),
- nombre de canaux d'enregistrement (monophonie ou stéréophonie),
- format de stockage : il définit les règles d'écriture et d'organisation des données encodées. Les formats des fichiers audio sont nombreux à savoir : (RIFF/wav, MP3, etc.),

– distance du microphone par rapport au locuteur : un bon positionnement du microphone permettra de maximiser le rapport signal-bruit (RSB).

1.2.2 Bases de données mondiales

De part le monde, il existe deux principaux organismes chargés de récolter ou de produire les ressources linguistiques orales : *Linguistic Data Consortium* (LDC) et *European Language Resources Association/Evaluations and Langage Ressouces Distribution Agency* (ELRA/ELDA) [LDC, ELRA]. L'importance et l'impact que manifestent ces ressources dans les nouvelles technologies de communication sont indéniables et ce, quelque soit le domaine applicatif visé : systèmes de synthèse automatique de la parole (Text-To-Speech), systèmes de reconnaissance automatique de la parole (RAP), systèmes de traduction (Speech-To-Speech). En effet, nous assistons ces deux dernières décennies à une recrudescence en matière de collecte et de réalisation de tout type de bases de données :

• *multilingue*s qui constituent d'importants projets englobant plusieurs langues officielles ou dialectales comme : SpeechDat, pour les langues de l'Europe de l'est [H. van den Heuvel and Tropf, 1998], GlobalPhone, réalisé avec 15 langues [Schultz and Waibel, 1998] , BABEL pour le Bulgare, l'Estonien, le Hongrois et le Polonais [Roach et al., 1996], SPEECON [Siemund et al., 2000], Langues Indiennes : le Tamil, le Telug et le Marathi [Anumanchipalli et al., 2008] ; ou bien des corpora *monolingues* se limitant à une unique langue comme : TIMIT pour l'Anglais Américain LDC [LDC], ANDOSL pour l'anglais Australien [Vonwiller et al., 1995], Bref120 réalisé pour le Français et élaboré par ELRA [ELRA], RACAD pour le Français Canadien [Cichocki et al., 2008].

• *restrictives* à des domaines précis comme la téléphonie, l'aéronautique [H. van den Heuvel and Tropf, 1998, Petrovska et al., 2000].

• *monolocuteur* ou *mutlilocuteurs* ; de parole *spontanée* continue ou de *textes lus*, etc.

Parmi les corpora les plus utilisés dans le domaine de la reconnaissance vocale, nous citons TIMIT. La conception et la réalisation de ce corpus, reconnu mondialement, a nécéssité la collaboration de trois prestigieux instituts : *Massachusetts Institute of Technology* (MIT), *Stanford Research Institute* (SRI) et *Texas Instruments* (TI). Sa distribution par contre est assurée par LDC. Constitué d'un total de 6300 phrases, le TIMIT a été enregistré par 630 locuteurs présentant 8 variétés dialectales.

Un comparatif détaillé sur les principales ressources orales mondiales, notemment celles dédiées à la reconnaissance automatique de la parole, est présenté en annexe A.

1.2.3 Bases de données arabes

Les corpora de parole dédiés à la langue arabe sont relativement moins nombreux par rapport aux principales ressources linguistiques destinées aux autres langues. Ces dernières années, LDC a publié le premier corpus en Arabe conçu pour les études de reconnaissance de la parole. Ce corpus, appelé West Point, contient des données qui ont été recueillies et traitées par les membres du Département des langues étrangères de l'Académie militaire de West Point aux États-Unis. L'objectif initial de ce corpus était de former des modèles acoustiques pour la reconnaissance automatique de la parole qui pourront être par la suite utilisés comme une aide à l'enseignement de l'Arabe.

Parallèlement, LDC fournit aussi deux autres ressources dont la première est basée sur des dialogues téléphoniques spontanés recueillis par des locuteurs Egyptiens, Syriens, Palestiniens, Libanais et Jordaniens ; la seconde ressource est le BBN/AUB, elle concerne l'Arabe Levantin (parler au Liban, Syrie et Palestine) [LDC].

Par ailleurs, le projet GlobalPhone, distribué par ELRA, offre aussi un corpus oral arabe réalisé à partir de journaux politiques et économiques. De plus, ce même organisme est chargé de distribuer, d'une part le projet NEMLAR qui est un corpus de parole enregistré à partir d'émissions de radio [Campbell Jr and Reynolds, 1999, Paulsson et al., 2004], et d'autre part un important ensemble de ressources pour l'Arabe Standard et dialectal appelé *OrienTel* enregistré dans de nombreux pays arabes : l'Arabie Saoudite, Émirats Arabes Unis, Maroc, Tunisie, Egypte et Palestine. Les bases de données OrienTel sont toutes enregistrées à partir de réseaux fixes et mobiles et des canaux multiples [ELRA]. Parmi les ressources arabes nous citons aussi le Saudi Accented Arabic Voice Bank (SAAVB) [Alghamdi et al., 2008]. Le SAAVB est une base de données téléphoniques (fixe et mobile) qui a été recueillie par l'université de King Abdul Aziz City pour la Science et la Technologie (KACST) entre 2002-2003. Un résumé des principaux corpora oraux arabes consacrés au développement des systèmes de reconnaissance automatique de la parole et du locuteur est donné en annexe B.

1.3 SYSTEME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Après avoir exposé les principales ressources linguistiques dédiées à la reconnaissance de la parole, nous allons présenter dans la partie qui suit, ce domaine avec plus de détails. En effet, ayant atteint une place de choix dans le domaine de la communication parlée, la Reconnaissance Automatique de la Parole (RAP) est considérée comme une charnière essentielle située aux frontières de l'oral et de l'écrit.

La RAP se matérialise par l'extraction automatique de l'information linguistique contenue dans un signal de parole. En effet, elle rassemble les différentes transformations que subit la parole injectée en entrée du système RAP pour en sortir sous forme de textes écrits. Cette description, plutôt simpliste, du procédé est loin d'être évidente pratiquement vu la complexité du problème.

Le signal vocal s'inscrit dans le cadre de la communication parlée comme un phénomène des plus complexes. Cette complexité est étroitement liée aux caractéristiques notoires du signal de parole, à savoir : une extrême redondance, une grande variabilité (vitesse d'élocution, l'âge, l'état physique et physiologique du locuteur, l'accent régional, etc.), des effets de co-articulation et enfin les lieux d'interférences (l'acoustique du lieu, qualité du microphone, les bruits, etc.).

1.3.1 Classification des systèmes RAP

Les Systèmes de Reconnaissance Automatique de la Parole (SRAP) sont classés selon plusieurs critères :

- Le premier d'entre eux correspond à la dépendance ou non au locuteur - systèmes mono/ multilocuteurs- [Calliope, 1989]. Les systèmes réalisés pour fonctionner avec un unique locuteur sont habituellement plus faciles à développer. Pour ce faire, il est nécessaire de prendre en compte la variabilité du signal aux niveaux inter et intra locuteur c'est à dire son état physique, le contexte et l'environnement dans lequel il se trouve, etc. Par ailleurs, si nous souhaitons étendre la reconnaissance à un nombre de locuteurs plus important, donc réaliser un système indépendant du locuteur, la gestion de tous les paramètres de variabilités devient moins flexible.
- Le second critère de classification concerne la taille du vocabulaire, où l'étendue de ce dernier est étroitement liée à la complexité du système (allant de simples commandes vocales -quelques mots- jusqu'aux applications complexes comme la dictée automatique -dictionnaire-). La taille du vocabulaire peut être constituée par :
  - des dizaines de mots (petit vocabulaire)
  - de centaines de mots (vocabulaire moyen)
  - de milliers de mots (large vocabulaire)
  - de dizaines de milliers de mots (très-grand vocabulaire)
- Le troisième critère concerne le type d'élocution à reconnaître (parole continue, spontanée, lue, ou bien une suite de mots isolés, etc.). Les systèmes de reconnaissance de mots isolés requièrent une pause entre deux mots prononcés. Ce qui revient à considérer que le mot est pris comme une entité entière délimitée dans le temps et qui ne se soumet pas au problème de coarticulation

existant dans la parole continue.

– Le dernier critère correspond à la robustesse des systèmes en milieux bruités. En effet, la performance d'un système peut être dégradé par une série de conditions défavorables comme : les bruits dus à l'environnement (bruit d'une voiture ou d'une usine) ; les distorsions acoustiques (échos, acoustique des salles) ; microphones différents (omnidirectionnel, téléphone) ; bande passante de fréquence limitée (en transmission téléphonique), etc.

1.3.2 Différentes applications des SRAP

Dans le domaine de la reconnaissance vocale, on distingue trois grands types d'applications : les systèmes à commandes vocales, les systèmes de compréhension et les machines à dicter [Calliope, 1989].

– Multiples et variés, les systèmes à commandes vocales occupent une large gamme de produits allant du simple jouet jusqu'aux outils sophistiqués (aide aux handicapés : rééducation ou remplacement d'un membre défaillant, commandes de voitures, etc.).

– Les systèmes de compréhension, quant à eux, sont des SRAP connectés à des modules d'interprétation, dont le rôle est d'interagir aux messages vocaux émis par l'utilisateur, pour ainsi répondre par une action mécanique, après prise de décision. Ces applications futuristes restent actuellement confinées dans des configurations plutôt 'simples' vu la complexité de leurs réalisations.

– Les systèmes de dictées automatiques ont pour but de transcrire intégralement un texte dicté par un locuteur devant un microphone aussi fidèlement que possible, sans faire intervenir pour autant la compréhension des phrases. En effet, le système est supposé avoir *appris* les règles d'usage et d'accords orthographiques spécifiques à la langue pour pouvoir les respecter au mieux. Toutefois, transcrire des dialogues ou des conversations improvisés représente un des challenges les plus compliqués à relever. En effet, les systèmes se trouvent, dans ce cas présent, confrontés à des éléments improbables concrétisés par : les phrases incomplètes, les hésitations, les répétitions, etc.

Une comparaison entre les systèmes de reconnaissances de la parole est exposée en annexe C.

1.3.3 Principe d'un SRAP

Du fait de la nature aléatoire du signal parole, à laquelle s'ajoute des interférences environnementales, une formulation déterministe du procédé de reconnaissance devient utopique. Alors, pour être étudiée, la reconnaissance vocale se trouve trivialement transposée dans un cadre probabiliste approchée par une théorie dont

les principaux fondements sont émis par [Jelinek, 1997, Rabiner and Juang, 1993].

Soit $O = o_1 o_2 ... o_m$ une suite de données acoustiques pour laquelle le système doit prendre une décision concernant les mots prononcés. Soit $W = w_1 w_2 ... w_n$ une suite linguistique constituée de n mots appartenant à un vocabulaire $\mathcal{V}$.

Si $P(W \mid O)$ dénote la probabilité que les mots W soient prononcés lorsque les données O sont observées, alors le SRAP doit satisfaire l'équation mathématique (1.1) afin de déterminer la séquence de mots la plus probable $\hat{W}$ (correspondant à la séquence de parole enregistrée) :

$$\hat{W} = arg\,max_W P(W \mid O) \tag{1.1}$$

En utilisant les règles de Bayes, (1.1) peut s'écrire sous la forme :

$$P(W/O) = \frac{P(O \mid W).P(W)}{P(O)} \tag{1.2}$$

où :

- $P(W)$ est la probabilité à priori de la séquence de mots W et représente le *Modèle de Langage*.
- $P(O \mid W)$ est la probabilité d'observer les données acoustiques O étant donnée la séquence W et correspond au *Modèle Acoustique*.
- $P(O)$ est la probabilité que O soit observée.

Trouver la séquence de mots $\hat{W}$ revient donc à la maximisation du produit $P(O \mid W).P(W)$ d'où l'équation :

$$\hat{W} = arg\,max_W P(O \mid W).P(W) \tag{1.3}$$

1.3.3.1 Modules d'un SRAP

En général, un SRAP basé sur une approche probabiliste nécessite trois niveaux de traitements : un module de paramétrisation, un module acoustique et enfin un module de langage. La figure suivante illustre globalement ces différents niveaux.

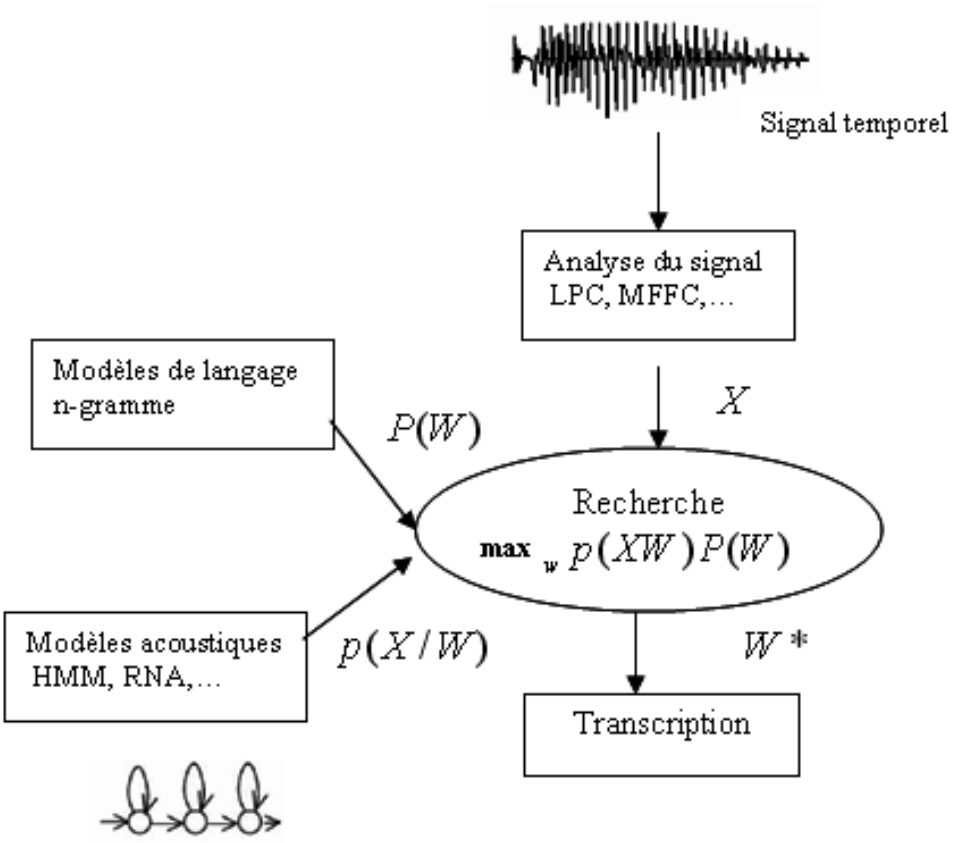


FIGURE 1.1: Différents niveaux d'un SRAP

Module d'analyse ou Paramétrisation

Pour concevoir un SRAP, il est important de préciser, au début du processus, le type de données acoustiques O à observer. Par conséquent, le signal parole capté par le microphone, est paramétré pour être transformé en une séquence de vecteurs observés o_i. Le but étant l'extraction de l'information linguistique utile contenue dans le signal vocal. Pour ce faire, les techniques de paramétrisation les plus souvent utilisées sont : les coefficients cepstraux MFCC (*Mel Frequency Cesptral Coefficients*), les LPCC (*Linear Predictive Cepstral Coefficients*) ou la PLP (*Perceptual Linear Predictive Analysis*) [Zheng et al., 2001, Huang et al., 2001]

Modèles du langage

La modélisation du langage a pour objectif de résumer les connaissances générales liées à un langage naturel. Deux types de modèles de langages se distinguent :

1. **Modèles à base de connaissances :** anciennement utilisée, cette modélisation s'inscrit dans le cadre des travaux en linguistique et se base principalement sur les règles de grammaire. Elle présente une contrainte majeure qui est la couverture de la langue. En effet, il est impossible de décrire la totalité des

règles grammaticales associées à un langage naturel donné. De plus, l'évolution permanente de la communication orale rend cette modélisation encore plus difficile à réaliser.

2. **Modèles probabilistes :** les limites que présentent les modèles précédents ont permis l'émergence d'une autre approche appelée les modèles probabilistes. Il s'agit actuellement du type le plus utilisé en reconnaissance de la parole. Ces modèles de langage se basent essentiellement sur l'apprentissage automatique à partir d'un corpus textuel sans se soucier des règles grammaticales existantes. Ils ont pour objet d'attribuer à une séquence donnée de mots une probabilité de calcul. De ce fait, l'équation (1.3) suppose que nous pouvons calculer, pour chaque mot W, la probabilité à priori $P(W)$. Cette probabilité peut être décomposée de plusieurs manières selon la formule de *Bayes*. Cependant, le système doit reconnaître W selon l'ordre de prononciation. Alors, la décomposition $P(W)$ se fait comme suit :

$$P(W) = \prod_{i=1}^{m} P(w_i \mid w_i, w_2, ..., w_{i-1}) \tag{1.4}$$

ou $P(w_i \mid w_1, w_2, ..., w_{i-1})$ représente la probabilité que w_i soit prononcé étant donnée la suite de mots $w_1, w_2, ..., w_{i-1}$ prononcée précédemment. Le choix de w_i dépend alors de la suite d'items qui le précède (l'historique). En effet, si $|\mathcal{V}|$ est la taille du vocabulaire, alors nous avons $|\mathcal{V}|^{\mathrm{i}-1}$ différentes suites de mots. En revanche, le nombre de probabilités à calculer se réduit drastiquement si nous considérons une suite limitée à quelques mots c'est à dire un historique de taille réduite. L'approximation la plus courante est les n-grammes où seuls les $n-1$ mots précédents sont pris en compte comme historique. Un n trop petit modélise mal les contraintes linguistiques tandis qu'un n trop grand va limiter la couverture du modèle. Généralement, l'utilisation de l'approche du *bigrammes* et *trigrammes* (respectivement n=2 et n=3) est la mieux adaptée pour les modèles de langage. Lorsque l'historique est basée uniquement sur le mot précédent (w_{i-1}), la probabilité est calculée comme suit :

$$P(W) = P(w_1)P(w_1 \mid w_2)...P(w_m \mid w_{m-1}) \tag{1.5}$$

Il est maintenant possible de calculer $P(w_i \mid w_{i-1})$ en évaluant le nombre de fois que la paire de mots (w_i, w_{i-1}) apparaît dans le corpus.

Les n-grammes peuvent être calculés sur n'importe quel type d'unités, que ce soit des mots, des phonèmes ou des syllabes, voire un mélange de ces unités. Pour la génération de tels modèles, il est nécessaire de posséder les transcriptions en unités désirées.

Modèles acoustiques

Les modèles acoustiques permettent de modéliser les unités acoustiques prononcées, en l'occurrence les phonèmes. Cette modélisation a connu, durant ces dernières décennies, une progression marquée allant de la reconnaissance analytique où l'extraction de tous les paramètres acoustiques sont nécessaires (pitch, formants, durée, etc.), à la reconnaissance globale (méthodes stochastiques) où l'analyse de la parole (MFCC, Cepstre, etc.) vise à trouver par le biais des probabilités une *image* acoustique de l'unité à reconnaître sans se soucier ni des traits phonétiques intrinsèques (voisement, friction, position des formants, etc.) ni de la longueur (phonème, syllabe, mot, etc.).

Les SRAP de parole continue actuels se basent sur l'approche stochastique (les Hidden Markov Models (HMM)[Huang et al., 2001, Rabiner and Juang, 1993] et les Réseaux de Neurones Artificiels (ANN) [Dreyfus, 2002]). Bien qu'ils s'appuient essentiellement sur des modèles et un formalisme mathématiques, ils nécessitent des connaissances acoustiques, phonétiques et linguistiques intégrées par le biais des dictionnaires de prononciation et la grammaire (les modèles de langage).

En se basant sur l'équation (1.3), les modèles acoustiques Markovien doivent pouvoir estimer la probabilité $P(O \mid W)$. Ces modèles sont entraînés dans une masse de données de parole contenant, plusieurs fois, les unités à modéliser en l'occurrence les phonèmes dans différents contextes.

1.3.3.2 Modèles acoustiques basés sur les HMMs

On appelle un modèle de Markov, tout automate stochastique M à nombre fini d'états ne dépendant, en un instant donné, que des états visités précédemment. Ce formalisme suppose que le signal est formé d'une séquence de segments stationnaires modélisés par les états HMM. Ces états sont caractérisés par des distributions de probabilité décrivant les probabilités d'observation des différents vecteurs acoustiques. La séquence d'états modélise, quant à elle, l'aspect séquentiel du signal, autrement dit, la structure temporelle de la parole comme une succession d'états stationnaires. Les transitions entre les états sont instantanées, caractérisées par une probabilité de transition. Les modèles sont dits cachés car la séquence d'états n'est pas observable contrairement à la séquence des observations (vecteurs acoustiques). La formulation mathématique des ces modèles est comme suit :

soit :

$\mathcal{S} = \{1, 2, ..., L\}$ l'ensemble des états de M,

s_t l'état de M visité à l'instant $t \in \mathbb{N}^*$,

Le modèle est alors paramétré sous forme d'un ensemble de probabilités de tran-

sition :

$$P(s_t = j \mid s_{t-1} = i, s_{t-2} = n, ...) \tag{1.6}$$

Si l'on suppose que M est de l'ordre 1 c'est à dire que la probabilité (1.4) de passer de l'état s_j à l'instant t ne dépendant que de sa valeur à l'instant $t-1$ lorsqu'il est à l'état s_i alors :

$$P(s_t = j \mid s_{t-1} = i, s_{t-2} = n, ...) = P(s_t = j \mid s_{t-1} = i) \tag{1.7}$$

Si l'on suppose que la chaîne de Markov est indépendante du temps, alors (1.4) s'écrit :

$$P(s_t = j \mid s_{t-1} = i) = p(j \mid i) = a_{ij} \quad \forall t \in \mathbb{N} \tag{1.8}$$

Les probabilités a_{ij} sont appelées *probabilités de transition* et elles doivent satisfaire les conditions suivantes : $a_{ij} \geq 0$, $\forall i, j$ et $\sum_{j=1}^{L} a_{ij} = 1$, $\forall i$

Un modèle de Markov discret est défini donc par :

1. le nombre d'états L,
2. la matrice de probabilités de transition A de dimension LxL avec : $A = \{a_{ij}\}$,
3. la distribution des probabilités des états initiaux : $\pi = (\pi_1, \pi_2, ..., \pi_L)$ où $\pi_i = P(s_1 = i)$ est la probabilité d'être à l'état i à l'instant $t = 1$.

Le HMM Λ est défini par un modèle discret M auquel on associe, à chacun des états une loi de probabilité. Les HMM visent à retrouver la séquence d'états pour laquelle une séquence d'observations est apparue. Le modèle Λ est alors décrit par $\Lambda = \{\pi, A, \mathcal{P}\}$:

1. M le modèle de base ;
2. $\mathcal{P} = \left\{P_\Lambda^1, P_\Lambda^2, ..., P_\Lambda^L\right\}$ l'ensemble des lois de probabilités associées aux différents états de M ;
3. s_t l'état de M à l'instant t ;
4. $O_t = (o_t^1, o_t^2 ..., o_t^N)$ vecteur d'observations à l'instant t de dimension N .

La figure suivante montre un modèle à 5 états (s_i) dont les transitions et les observations sont respectivement a_{ij} et o_i.

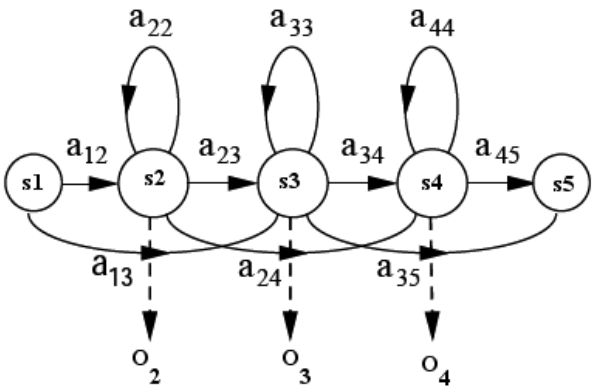


FIGURE 1.2: Représentation graphique d'un HMM à 5 états

Lorsque le système est à lois continues, les probabilités d'observation $b_j(o_t)$ sont représentées par des densités de *mélange gaussien* (*Gaussian Mixture Model* GMM).

$$b_j(o_t) = \sum_{m=1}^{M} c_{jsm}\mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}) \tag{1.9}$$

M est le nombre de composantes du mélange à l'état j pour le flux s, C_{jm} est le poids de chaque composante de l'état j et $\mathcal{N}(.; \mu, \Sigma)$ est une *gaussienne multivariée*. Sa formule est :

$$\mathcal{N}(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-1/2(o-\mu)'\Sigma^{-1}(o-\mu)} \tag{1.10}$$

avec μ et Σ sont respectivement le vecteur *moyenne* et la matrice de *covariance* exprimés par :

$$\mu = \frac{1}{T}\sum_{t=1}^{T} o_t$$

$$\Sigma = \frac{1}{T}\sum_{t=1}^{T}(o_t - \mu)(o_t - \mu)'$$

Le modèle général des HMMs est décrit alors par :

$$\Lambda_c = \{\pi, A, \mu_i, \Sigma_i \mid i = 1, ..., L\} \tag{1.11}$$

La figure 1.2 illustre un HMM à 3 états dont les observations sont des mono-gaussiennes.

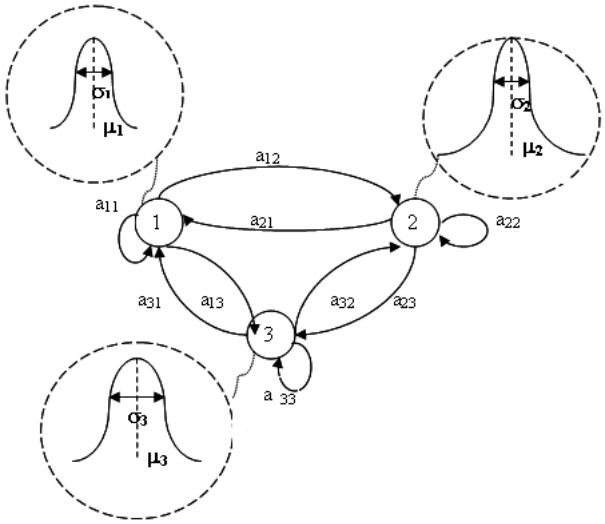


FIGURE 1.3: HMM à 3 états à monogaussienne chacun

1.3.3.3 Algorithmes d'optimisation

Pour optimiser la détermination de la suite de mots la mieux adaptée à une séquence de parole, les SRAP font intervenir 2 algorithmes robustes qui sont l'algorithme de Baum-Welch et l'algorithme de Viterbi [Huang et al., 2001, Rabiner and Juang, 1993].

Algorithme de Baum-Welch

Un algorithme particulièrement efficace dans l'estimation des paramètres d'une chaîne de Markov cachée est l'algorithme de Baum-Welch. Il permet de trouver la maximisation des probabilités d'observation étant donné un modèle $\Lambda = (A, B, \pi)$ où A est la matrice de transition, B vecteur d'observations et π la distribution des probabilités des états initiaux.

Le premier algorithme utilisé par Baum-Welch est l'agorithme *Forward* qui délivre deux informations :$P(O \mid \Lambda)$ et $\alpha_t(i)$.

La procédure *Forward*

Considérant $\alpha_t(i)$ la probabilité de la séquence d'observation partielle $(o_1, o_2, ...o_t)$ à l'instant t et à l'état i. $\alpha_t(i)$ est définie par :

$$\alpha_t(i) \quad = \quad P(o_1 o_2 ... o_t, s_t = i \mid \Lambda) \tag{1.12}$$

– *Initialisation*

L'algorithme Forward place dans un premier temps, α_1, la probabilité d'obtenir l'état caché i sachant que l'on a observé le symbole o_1, $\alpha_1(i)$ est alors calculé par:

$$\alpha_1(i) \quad = \quad \pi_i b_i(o_1) \quad 1 \leq i \leq N \tag{1.13}$$

avec π_i la probabilité d'avoir l'état i en premier, et $b_i(o_1)$ la probabilité d'observer o_1 lorsque l'état i est apparu.

– *Récurrence*

Cette étape représente la procédure pour atteindre l'état j en $t+1$ à partir des N états i possibles, $1 \leq i \leq N$, à l'instant t. $\alpha_t(i)$ est la probabilité que la suite $(o_1, o_2, ...o_t)$ soit observée à l'instant t et à l'état i. Le produit $\alpha_t(t)a_{ij}$ est la probabilité jointe que $(o_1, o_2, ..., o_t)$ soit observée lorsque l'état j est atteint à l'instant $t+1$ via l'état i à l'instant t. $b_j(o_{t+1})$ est l'observation obtenue à l'état j à l'instant $t+1$. α_{t+1} est alors calculé par :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\right] b_j(o_{t+1}) \quad 1 \leq t \leq T-1 \; et \; 1 \leq j \leq N \tag{1.14}$$

– *Fin*

La dernière étape correspond à la somme des termes de $\alpha_T(i)$ qui représente la probabilité recherchée $P(O \mid \Lambda)$, pour qui la chaîne $(o_1, o_2, ..., o_T)$ est entièrement observée.

$$P(O \mid \Lambda) \quad = \quad \sum_{i=1}^{N} \alpha_T(i) \tag{1.15}$$

En somme, l'algorithme stipule que pour calculer la probabilité d'être à un état j au temps $t+1$, il faut calculer tout d'abord la somme de toutes les probabilités jointes $P(O, s \mid \Lambda)$, à l'instant t pour tous les états i, puis les multiplier par les probabilités de transition correspondantes. L'expression obtenue représente alors $\sum_{i=1}^{N} \alpha_t(i)a_{ij}$ qui à son tour sera multipliée par la probabilité d'observation de l'état j au même instant c'est à dire à $t+1$ et qui est $b(o_{t+1})$.

La procédure *Backward*

Le deuxième algorithme utilisé par Baum-Welch est l'agorithme *Backward* qui délivre $\beta_t(i)$.

De la même manière que la procédure Forward, nous considérons donc la variable $\beta_t(i)$ définie par :

$$\beta_t(i) \quad = \quad P(o_{t+1}o_{t+2}...o_T \mid s_t = i, \Lambda) \tag{1.16}$$

$\beta_t(i)$ est la probabilité d'observation partielle de la séquence $t+1$ jusqu'à la fin du processus, étant donné l'état i à l'instant t et le modèle Λ. Le calcul de $\beta_t(i)$ se fait par induction comme suit :

– *Initialisation*

Cette étape consiste à mettre $\beta_T(i)$ égal à 1 pour tous les états i.

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \tag{1.17}$$

– *Itération*

Le point de repère dans ce cas est l'opposé de la procédure Forward, c'est à dire de la fin vers le début. Donc pour calculer la probabilité à l'état i à l'instant t, on calcule la somme des différentes probabilités de transitions vers i ainsi que les probabilités d'observation correspondantes à l'état j à l'instant $t+1$. Le résultat est par la suite multiplié par la probabilité jointe de l'état j.

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{1.18}$$

$$t = T-1, T-2, ..., 1 \quad 1 \leq i \leq N$$

Algorithme de VITERBI

L'algorithme de Viterbi est un algorithme dynamique qui va chercher à maximiser l'expression $P(O \mid W)P(W)$. Aussi, il parcourt, à tout instant, toutes les observations et construit au fur et à mesure les meilleurs chemins partiels pour ainsi trouver la séquence probable $s = (s_1 s_2 .... s_T)$. Pour cela on définit la quantité $\delta_t(i)$ qui représente le chemin des probabilités maximales en un temps t pour les t premières observations et qui fini à l'état i :

$$\delta_t(i) = \max_{s_1, s_2, ..., s_{t-1}} P\left[s_1 s_2 ... s_{t-1}, s_t = i, o_1 o_2 .... o_t \mid \Lambda\right] \tag{1.19}$$

Par induction, on a :

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] . b_j(o_{t+1}) \tag{1.20}$$

Pour trouver la séquence la plus probable, on doit récupérer dans $\psi_t(j)$ les traces des arguments qui maximisent l'équation (1.20) pour chaque t et j.

– *Initialisation*

$$\delta_1(i) = \pi_i b_i(o_1) \;\; 1 \leq i \leq N \tag{1.21}$$

$$\psi_t(i) = 0 \tag{1.22}$$

– *Récursion*

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \, . b_j(o_t) \;\; 1 \leq j \leq N \;\; et \;\; 2 \leq t \leq T \tag{1.23}$$

$$\psi_t(j) = arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \;\; 1 \leq j \leq N \;\; et \;\; 2 \leq t \leq T \tag{1.24}$$

– *Fin*

$$P* = \max_{1 \leq j \leq N} [\delta_T(i)] \tag{1.25}$$

$$s_T^* = arg \max_{1 \leq j \leq N} [\delta_T(i)] \tag{1.26}$$

Le chemin de la meilleure séquence en sens inverse est ainsi calculé par :

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \;\; t = T-1, T-2, ..., 1 \tag{1.27}$$

1.4 CONCLUSION

Le besoin d'exploiter des corpora oraux dans le but de faire évoluer les recherches fondées sur le langage en général ou la parole en particulier, notamment dans les nouvelles technologies de l'information, est d'une évidence marquée. De ce fait, nous avons voulu dans ce chapitre fournir les principaux constituants d'une base de données sonores: corpora textuels, les locuteurs, les conditions d'enregistrement, etc. Ces éléments ont été suivis par le classement de quelques ressources orales tout en s'attardant sur les corpora dédiés à la reconnaissance automatique de la langue arabe.

Nous avons présenté dans la seconde partie du chapitre, les fondements théoriques et les différents modules sur lesquels se basent les SRAP à savoir : l'extraction des données acoustiques, les modèles acoustiques basés sur les HMM, le module de langage et les algorithmes utilisés pour optimiser leurs performances.

Chapitre 2

ALGERIAN ARABIC SPEECH DATABASE (ALGASD)

2.1 INTRODUCTION

Les ressources orales dédiées à la langue arabe sont beaucoup moins nombreuses que celles disponibles pour les autres langues (voir chapitre 1). De plus, leur nombre semble insignifiant comparé au grand nombre d'arabophones dans le monde -environ 250 millions de personnes- sur une zone géographique allant du golf Arabo-Persique (Moyen-Orient) à l'Océan Atlantique. Nous avons voulu par la réalisation d'une base de données sonores, *ALGerian Arabic Specch Database* (ALGASD) à partir d'enregistrements de locuteurs algériens, apporter une modeste contribution pour essayer de réduire cette déficience en ressources orales.

Le corpus ALGASD a pour vocation de refléter les principales variations de prononciation de l'Arabe Standard entre Algériens en tenant compte de divers facteurs régionaux et sociaux. Les étapes de conception et de réalisation d'ALGASD sont présentées dans les sections suivantes.

2.2 NOTIONS SUR L'ARABE STANDARD

L'Arabe Moderne connaît deux variantes : l'Arabe Moderne Standard (représenté dans le document par AS), la langue du Saint Coran, et l'Arabe dialectal qui est un mélange de la première langue avec le parler des autochtones. Influencé par diverses traditions et cultures régionales, ce mélange offre, selon les régions, une marque linguistique particulière propre à chaque parler.

Le système phonétique de l'AS est principalement constitué de 34 phonèmes : 6 voyelles, 3 courtes ([a], [u] et [i]) auxquelles sont respectivement opposées 3 longues, et 28 consonnes [Watson, 2007]. La langue arabe est caractérisée par trois proprié-

tés fondamentales dont le rôle est la distinction entre deux mots phonétiquement proches. Ces propriétés sont : /ʔalmad/ (المد) correspond aux voyelles longues, la gémination /ʔataʃdid/ (التشديد) et l'emphase /ʔatafχim/ (التفخيم).

Appelée aussi redoublement, la gémination correspond au phénomène de renforcement d'une articulation consonantique. Cette dernière tend à prolonger la durée de la consonne tout en augmentant son intensité [Bonnot, 1979]. L'ensemble des consonnes arabes est concerné par cette propriété.

L'emphase se traduit par une double articulation. Elle est générée en deux temps synchronisés : le premier concerne le rétrécissement de la cavité pharyngale (pharyngalisation) par la racine de la langue et le second par l'incurvation ou l'aplatissement lingual pour former un creux. Par conséquent, deux lieux d'articulation se produisent simultanément. L'Arabe compte 4 phonèmes emphatiques [Jakobson, 1957]. L'annexe D illustre les phonèmes arabes transcris en *International Phonetic Alphabet* (IPA) et en *Speech Assessment Methods Phonetic Alphabet* (SAMPA) [IPA, 1999, SAMPA].

2.3 VARIETES LANGAGIERES ET PHONETIQUES EN ALGÉRIE

L'Algérie s'étend sur un vaste territoire de 2 380 000 km^2 partagé en 48 wilayas. Sa population s'élève actuellemnt à 38 millions d'habitants dont la majorité est concentrée dans le nord du pays.

La langue officielle des Algériens est l'AS. Il est enseigné à l'école du niveau primaire jusqu'au secondaire (de six à seize ans). L'AS est utilisé dans toutes les procédures administratives (gouvernement, médias, etc). Toutefois, son écriture diffère considérablement des parlers algériens (langues maternelles). En effet, environ 72% de la population parlent quotidiennement la "*Darija*" qui est l'Arabe dialectal Algérien et 28% d'entre eux pratiquent une seconde langue appelée le "*Tamazight*" qui est une variante de la langue berbère.

Les dialectes qui constituent la "Darija" découlent des différentes influences ethniques, géographiques et coloniales qu'à connu le pays à travers des âges [Taleb Ibrahimi, 1995]. Tandis que la Darija d'Alger communément appelée l'Algérois a été influencé par le berbère et le Turc, le Constantinois est influencé par l'Italien, l'Oranais par l'Espagnol, le Tlemcenien par l'Arabe de l'Andalousie, etc. En conséquence, au sein même de l'Arabe algérien, il existe, de ville en ville, d'importantes variations locales observées sur les plans phonétique, grammatical et lexical.

Les principales spécificités phonétiques sont relevées dans les régions suivantes :

- Le Jijelien remarquable pour la profusion de termes empruntés au Berbère

ainsi qu'à la prononciation particulière de certains phonèmes comme : le phonème [q] (ق) qui est remplacé par [k] (ك) [Marçais, 1956] ;

• Le Tlemcenien où le phonème [q] (ق) est remplacé par la glottale [ʔ] (ء) [Marçais, 1956] ;

• L'Arabe de l'Est et l'Ouest Algérien respectivement les régions d'Annaba et Oran dont les parlers sont proches des dialectes Tunisiens et Marocains. En ce qui concerne Oran le [q] (ق) est remplacé par [g] [Caubet, 2000] ;

• L'Arabe des Hâssanya : parlé principalement par les habitants du sud-ouest Algérien, le Maroc et la Mauritanie. Son système phonologique est à la fois novateur et conservateur. En effet, de nouveaux phonèmes sont introduits aux côtés de tous les phonèmes basiques présents dans la langue arabe [Caubet, 2000].

Essentiellement oral, le *Tamazight* est la langue des autochtones avant l'avènement de l'Islam. Il a été parlé dans toute la région qui s'étend de l'oasis de Siwa dans l'ouest de l'Egypte jusqu'aux îles Canaries, en passant par la Libye, la Tunisie, l'Algérie et le Maroc sans oublier le Mali et le Niger. À l'heure actuelle, les zones où le Tamazight est répandu sont les déserts ou certaines régions montagneuses. L'absence de contact, dans le passé, entre ces régions a conduit à un processus de dialectalisation important. Cependant, la nature de la variation dialectale est plus phonologique et lexicale que syntaxique (grammaticale) [Achab, 2001].

Le Tamazight a plusieurs variétés dont la première, la plus répandue, est le Kabyle. Cette variante est parlée principalement dans les wilayas de Tizi-Ouzou et de Béjaia. La seconde variante est pratiquée aux frontières Algéro-Marocaines notamment à Béchar. La troisième, courante à Ghardaia, est appelée le Mzab. La quatrième variante, le Chaoui, est parlée dans les régions montagneuses de l'Est (Aurès). Et enfin le Targui parlé par les nomades (les touaregs) du Sahara.

Ces deux langues maternelles en l'occurrence la Darija et le Tamazight constituent le principal substrat de la communication orale entre les Algériens. Toutefois, une catégorie d'Algériens utilise une autre langue complètement différente qui est la langue Française. Bien que cette dernière n'ait pas de statut officiel, le Français est largement utilisé par le gouvernement, les médias et les universités, etc. [Cheriguen, 1997, Benrabah, 2007].

Par conséquent, apprendre l'AS à une société multilingue, telle la société Algérienne, revient à lui faire enseigner des aspects lexicaux, grammaticaux et phonologiques souvent différents de ceux utilisés quotidiennement. Par conséquent, la maîtrise de l'AS rime nécessairement avec l'acquisition progressive de tous ces nouveaux fondements en commençant notamment par le plan phonologique (apprentissage de nouveaux phonèmes). Ceci relègue presque intuitivement l'AS au rang de *seconde langue* par rapport au vécu Algérien.

2.4 OBJECTIFS PAR ALGASD

La réalisation d'une base de données nécessite souvent une prise en charge sérieuse, de même que des études préliminaires rigoureuses et ce, avant d'entamer l'enregistrement. Cette phase de préparation consiste à apporter des éléments de réponses sur un ensemble de questions essentielles posées telles que : les objectifs visés par la base, les conditions d'élaboration, les pratiques de terrain choisies, etc.

Nous avons voulu, par le biais d'ALGASD, révéler la variété langagière et phonétique du peuple Algérien à travers des enregistrements sonores de phrases lues en AS, et ce, principalement, dans des zones géographiques caractérisées par des traits phonétiques spécifiques émergents (voir la section précédente). En plus du facteur régional, la base inclut, pour chaque locuteur, des renseignements personnels (genre, âge, niveau d'instruction et maîtrise de la langue). Ces données réflectives de la société algérienne constituent un substrat essentiel pour des études comparatives (sociolinguistiques) qui visent à illustrer de possibles différences et similitudes dans la production des phonèmes de l'AS. Comme elles peuvent être utilisées aussi dans le développement et l'évaluation d'un SRAP pour la langue arabe en général ou bien pour l'accent algérien en particulier.

2.5 ARCHITECTURE

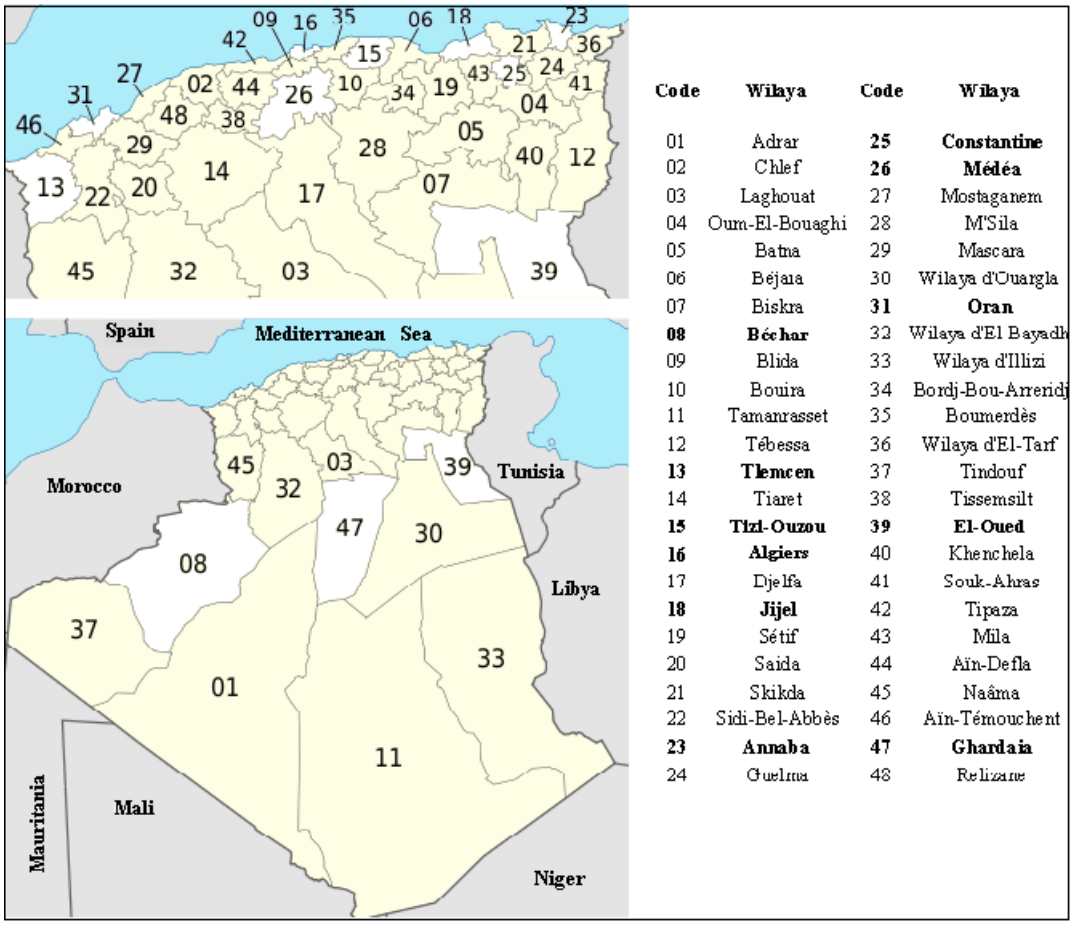
L'architecture de la base, ainsi que les différentes étapes de sa réalisation (les choix des régions et des locuteurs ainsi que les corpora textuels à lire) sont détaillés dans les sections suivantes.

2.5.1 Sélection des régions

ALGASD vise à être une base de données réaliste qui reflète les principales caractéristiques de prononciation liées aux différentes variations régionales et sociales. Ainsi, à la lumière des recherches linguistiques présentées dans la section 2.3, la couverture régionale, dans cette base, respecte les grands groupes dialectaux recensés. Nous avons proposé donc l'étude de 11 variations pouvant influencer éventuellement la prononciation de l'AS. Aussi, de l'Est à l'Ouest et du Nord au Sud, les variations de prononciations les plus remarquables sont observées dans : 8 régions du nord du pays - où la population est dense - et 3 autres du sud.
Pour le nord, nous avons :

1. 3 régions du Centre (Alger, Tizi Ouzou et Médea)
2. 3 régions de l'Est (Constantine, Annaba et Jijel)
3. 2 régions de l'Ouest (Oran et Tlemcen)

Pour le sud, nous avons : Bechar (sud-ouest), El Oued (sud-est) et Ghardaïa (centre). La figure 2.1 illustre la situation géographique des 11 zones soumises à l'étude. Les noms des régions sont transcrits en caractères gras sur la légende.



Code	Wilaya	Code	Wilaya
01	Adrar	**25**	**Constantine**
02	Chlef	**26**	**Médéa**
03	Laghouat	27	Mostaganem
04	Oum-El-Bouaghi	28	M'Sila
05	Batna	29	Mascara
06	Béjaia	30	Wilaya d'Ouargla
07	Biskra	**31**	**Oran**
08	**Béchar**	32	Wilaya d'El Bayadh
09	Blida	33	Wilaya d'Illizi
10	Bouira	34	Bordj-Bou-Arreridj
11	Tamanrasset	35	Boumerdès
12	Tébessa	36	Wilaya d'El-Tarf
13	**Tlemcen**	37	Tindouf
14	Tiaret	38	Tissemsilt
15	**Tizi-Ouzou**	**39**	**El-Oued**
16	**Algiers**	40	Khenchela
17	Djelfa	41	Souk-Ahras
18	**Jijel**	42	Tipaza
19	Sétif	43	Mila
20	Saida	44	Aïn-Defla
21	Skikda	45	Naâma
22	Sidi-Bel-Abbès	46	Aïn-Témouchent
23	**Annaba**	**47**	**Ghardaia**
24	Guelma	48	Relizane

FIGURE 2.1: Situation géographique des 11 régions sélectionnées dans ALGASD [Geo]

2.5.2 Sélection des locuteurs

En se basant sur les statistiques de l'Office National des Statistiques ONS (recensement de 1998) [ONS] [1], nous avons relevé le nombre d'habitants des 11 régions citées ainsi que leurs distributions selon le genre (Homme/Femme) (tableau 2.1). Nous notons à partir des valeurs que la distribution selon le genre est presque équivalente pour les deux sexes ($\backsimeq$ 50% d'hommes vs $\backsimeq$ 50% de femmes pour chacune des 11 régions).

1. Les travaux de cette thèse se basent sur les données démographiques disponibles entre 2006-2009 sur le site de l'Office National des Statistiques à savoir le recensement de la population de 1998 ($\approx$30 millions d'habitants). L'architecture et l'enregistrement du corpus oral ont été réalisés durant cette période.

Désignation	Régions	Femme	%	Homme	%	Total
R1	Alger	1287077	50,23	1275351	49,77	2 562 428
R2	Tizi Ouzou	553248	49,90	555460	50,10	1 108 708
R3	Médéa	408675	50,95	393403	49,05	802 078
R4	Constantine	406288	50,10	404625	49,90	810 914
R5	Jijel	287581	50,17	285627	49,83	573 208
R6	Annaba	280019	50,20	277799	49,80	557 818
R7	Oran	608832	50,16	605007	49,84	1 213 839
R8	Tlemcen	424140	50,37	417914	49,63	842 053
R9	Bechar	113764	50,44	111782	49,56	225 546
R10	El Oued	256656	50,88	247745	49,12	504 401
R11	Ghardaïa	226900	50,92	218719	49,08	445 619

TABLE 2.1: Distribution réelle de la population selon l'ONS (recensement de 1998)

2.5.2.1 Nombre de locuteurs

A partir du table 2.1, l'évaluation d'un échantillon représentatif réel de la population a été effectué et ce afin d'en déduire le nombre exact de locuteurs pour chaque région selon le genre. Les résultats obtenus montrent que pour un nombre global de 300 locuteurs, la distribution par région s'effectue de la manière suivante (tableau 2.2 et figure 2.2). Les régions 1 et 9, en l'occurrence Alger et Béchar, sont dotées respectivement du plus grand et du plus faible nombre de locuteurs (proportionnellement aux densités effectives en habitants).

Régions	Locuteurs Féminins	Locuteurs Masculins	Total Locuteurs/Régions
R1	40 (50%)	40 (50%)	80 (27%)
R2	17 (50%)	17 (50%)	34 (11%)
R3	13 (52%)	12 (48%)	25 (8%)
R4	13 (52%)	12 (48%)	25 (8%)
R5	09 (50%)	09 (50%)	18 (6%)
R6	09 (52%)	08 (48%)	17 (6%)
R7	19 (50%)	19 (50%)	38 (13%)
R8	13 (50%)	13 (50%)	26 (9%)
R9	04 (52%)	03 (48%)	07 (2%)
R10	08 (50%)	08 (50%)	16 (5%)
R11	07 (50%)	07 (50%)	14 (5%)
TOTAL	152 (51%)	148 (49%)	300 (100%)

TABLE 2.2: La distribution des locuteurs dans les differentes régions d'ALGASD

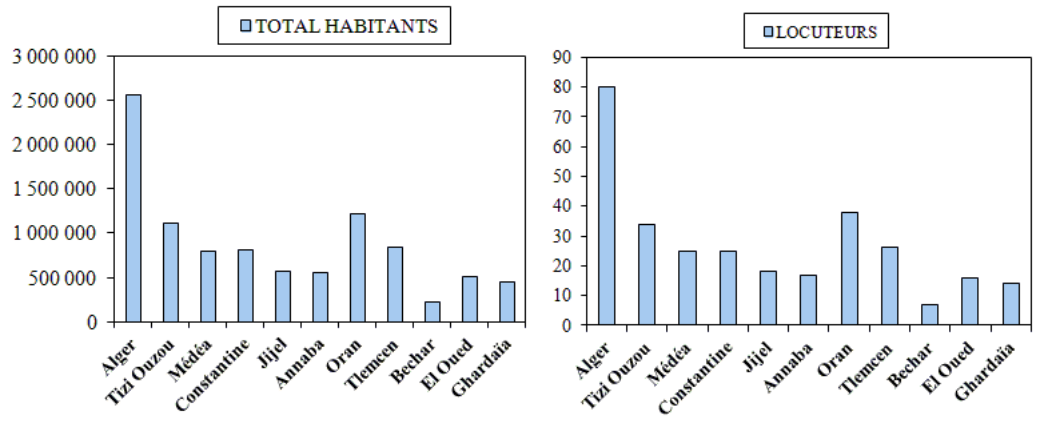


FIGURE 2.2: Nombre de locuteurs par région : représentation réelle (à gauche) ; représentation dans ALGASD (à droite)

2.5.2.2 Profils des locuteurs

Les locuteurs d'ALGASD sont tous natifs et vivent dans leurs localités respectives. Les participants appartiennent à des milieux socio-économiques différents (médecins, enseignants, étudiants, chômeurs, femmes aux foyers, etc.). En plus de cette variante sociale, la base a été conçue pour recueillir d'autres renseignements personnels comme le : genre, âge et niveaux d'instruction de chacun.
Aussi, nous avons pour chaque catégorie citée les sous-classes suivantes :

- Genre : homme/femme
- Âge:
  - jeune (18-30 ans)
  - moyen (30-45 ans)
  - âgé allant de 45 ans à plus
- Niveaux académiques :
  - ayant bénéficié d'un ou des paliers d'enseignement primaire, moyen et secondaire
  - ayant obtenu des diplômes universitaires (licence, ingéniorat, etc.)
  - ayant terminé où poursuivent toujours des études de post-graduation

2.5.3 Corpus

Les bases de données sonores peuvent être enregistrées à partir de textes lus ou à partir de conversations spontanées. L'élaboration d'ALGASD est basée sur la lecture d'un *corpus de référence* constitué de phrases écrites en AS. La répartition de ce corpus en trois sous-ensembles est inspirée de celle utilisée dans la réalisation du corpus TIMIT [Zue et al., 1990].

Nous exposons, dans ce qui suit, les propriétés du corpus de référence ainsi que sa répartition en sous-ensembles.

2.5.3.1 Corpus de référence

Pour construire les textes à lire d'ALGASD, nous avons utilisé les 200 Phrases Phonétiquement Equilibrées (PPE) conçues par [Boudraa et al., 2000]. Par définition, un corpus de PPE correspond à un ensemble de phrases répondant à certaines exigences comme : contenir tous les phonèmes de la langue, avec leurs véritables proportions d'apparition (nombres d'occurrence), respect des différents contextes et spécificités de la langue, des redondances, etc. En somme, construire un corpus de Phrases Phonétiquement Equilibrées revient à reproduire la langue à une échelle réduite en utilisant des techniques scientifiques adéquates.

Pour réaliser ces PPE, les auteurs se sont basés sur des résultats de travaux établis par [Moussa, 1973] sur les fréquences d'apparitions des voyelles (V) et des consonnes (C) dans les racines des mots arabes. La procédure de vérification expérimentale de ces occurrences a été inspirée de travaux effectués sur le Français. Ainsi, les auteurs ont procédé à des tests statistiques notamment le test du χ^2, pour en ressortir la distribution réelle des syllabes [CV] dans l'Arabe. L'analyse expérimentale leur a permis, par la suite, d'assembler ces unités pour en construire des phrases équilibrées réparties en 10 listes de 20 phrases. Chaque liste est constituée de 208 diphones [CV]. Les mots utilisés dans la construction de ces phrases sont communs et ne font pas partie d'un lexique spécialisé. Ces phrases ont été soumises à des évaluations sur les plans sémantique et syntaxique par des linguistes pour être par la suite validées.

La réduction du nombre de phrases allouées pour l'enregistrement (200 contre les 2432 de TIMIT) est "relativement compensée" par l'utilisation d'un corpus entièrement composé de phrases phonétiquement équilibrées. Aussi, en exploitant le corpus PPE, nous nous assurons d'une part, de l'utilisation de tous les phonèmes de la langue, d'autre part, nous éliminons d'emblée une tâche fastidieuse qu'est la procédure de constitution des phrases ainsi que toutes les étapes connexes telles que : l'exhaustivité des phonèmes dans le corpus, le calcul du nombre d'occurrences de chaque phonèmes, etc. Les 200 PPE transcrites sous la forme orthographique sont données en annexe E.

2.5.3.2 Corpus d'ALGASD

Les PPE sont originellement utilisées dans des tests effectués sur le codage de la parole et non pas conçues pour mettre en exergue les caractéristiques phonétiques de la langue. Pour les adapter à notre objectif, qui est l'enregistrement d'une base de données sonores, nous avons été amenés à apporter quelques réajustements mineurs

à l'organisation initiale des listes. Ainsi, nous avons déplacé certaines phrases (d'une liste à une autre) pour enrichir les textes à lire par des phonèmes peu fréquents. De ces nouvelles listes, trois sous-corpus ont été conçus. L'objectif de chaque sous-ensemble vise à mettre en exergue des aspects phonétiques bien déterminés :

Corpus commun (Cc)

Dans le but de montrer la variation dialectale notée parmi les locuteurs algériens, nous avons sélectionné 2 phrases parmi les 200. Ces deux phrases ont servi à l'élaboration du *Corpus Commun* (Cc). Le choix de ces phrases a été orienté de telle sorte qu'elles contiennent le plus grand nombre de consonnes non redondantes de la langue et qu'elles incluent le maximum de phonèmes susceptibles d'avoir des prononciations différentes relatives aux accents régionaux recensés comme le [q] (voir section 2.3). Aussi, les deux phrases "*dialectales*" sont constituées de 75% de consonnes soit : 12 consonnes différentes dans la première contre 9 dans la seconde phrase.

Corpus réservé (Cr)

Le Corpus réservé (Cr) a pour rôle de couvrir tout le système phonétique de l'Arabe (28 consonnes et 6 voyelles). En constatant que certains phonèmes sont peu fréquents comme dans le cas du phonème [d^{ς}] (ض), nous avons été contraints de casser l'équilibre phonétique préétabli, dans certaines listes, pour accroître le nombre d'occurrences de ces phonèmes et ce par le remplacement de certaines listes de phrases par d'autres.

Le corpus Cr est doté de 62 phrases dont 30 sont utilisées pour la formation de 10 textes ayant 3 phrases chacun. A chaque texte, nous avons rajouté une seule phrase -différente à chaque fois- des 32 sur les 62 restantes pour en former un nouveau texte. Par conséquent, le nombre total atteint est de 32 textes de 4 phrases. Leur répartition sur les locuteurs et leurs régions respectives sera détaillée dans la section d'enregistrement.

Corpus individuel (Ci)

L'ensemble restant de phrases soit 136 sur 200 a constitué le Corpus individuel (Ci). L'objectif de ce corpus est de collecter les différences individuelles de prononciation.

2.5.4 Choix du Protocole

Une base de données inclut, en plus des fichiers sonores, une multitude d'informations appelées le *metadata*. Ce dernier est une source de renseignements pour

diverses recherches linguistiques, phonétiques, sociolinguistiques, etc.

Le protocole d'ALGASD se rapporte à la manière avec laquelle la multitude de renseignements connexes du metadata est inscrite (textes, répartitions des textes et des locuteurs, informations concernant les locuteurs, etc.). Un code a été adopté pour normaliser tous les fichiers de la base. Optimal et unique, ce dernier rassemble en une seule entête le maximum d'informations (nom et prénom du locuteur, le genre, etc.).

Nous exposons dans ce qui suit le protocole d'ALGASD ainsi que les formulaires élaborés pour faciliter la manipulation des fichiers du metada. Ces documents sont appelés *fiches de renseignements*.

2.5.4.1 Fiche de Renseignement 1

Le but de ce document est de lister tous les noms et prénoms des locuteurs d'ALGASD ainsi que leurs renseignements respectifs.

Deux feuilles distinctes constituent cette première fiche : la première, liste les locuteurs selon la région d'appartenance et la seconde, constitue une synthèse de tous les éléments recueillis dans les fiches de renseignements 2 (voir ci-dessous).

2.5.4.2 Fiche de Renseignement 2

La fiche de renseignement 2 rassemble toutes les informations relatives aux locuteurs et aux enregistrements réalisés : la région, les nom et prénom, le genre, l'âge, le niveau d'instruction et le corpus à lire (écrit en Arabe et transcrit en phonétique SAMPA).

Les modèles retenus pour la conception des fiches de renseignements 1 et 2 sont donnés dans l'annexe F. Le tableau suivant expose les codes adoptés pour ces différents renseignements.

Désignation	Signification	Objectifs
ID	Nom du locuteur (3 initiales) + chiffre (0-9)	Distinction entre locuteurs
m/f	Masculin /Féminin	Genre du locuteur
DatNais	Date de Naissance du locuteur	Âge du locuteur
NIns	Niveau d'Instruction	Post Graduation/Universitaire ou Moyen
Rn	Région étudiée	Accent régional
DatEnregis	Date d'Enregistrement des phrases	
C	Type de phrases enregistrées	Cc / Cr ou Ci

TABLE 2.3: Renseignements collectés

– ID : Pour identifier le locuteur et afin d'assurer son anonymat, nous avons opté pour l'utilisation d'un code d'identification remplaçant son vrai nom. Ce

code est constitué des initiales du nom et prénom (au nombre de 3). Dans le cas où la troisième initiale vient à manquer, elle est remplacée par la lettre X. Cette dernière ne sera probablement pas confondue avec les initiales réelles du locuteur car le graphème X n'existe pas dans la langue arabe. Le nombre d'initiales a été fixé à 3 car, en Algérie, plusieurs cas de figures concernant les noms existent. En effet, en plus des noms que nous appelons simples (constitués d'un seul nom de famille et d'un seul prénom), nous pouvons trouver des noms et/ou prénoms composés. Pour plus de précisions voir le tableau suivant.

Nom	Prénom		Code
type Simple			
Droua	Ghania		DGX
type Composé			
Sellami	Sid Ahmed		SSA
Hamdani	Mohamed	Saïd	HMS
Châaban Chaouch	Hassina		CCH

TABLE 2.4: L'identification du locuteur ID

- Le n de l'ID sert à distinguer entre les locuteurs qui partagent les mêmes initiales.
  Exemples :
  - Lahcen Siham —>LSX0
  - Louari Salma —> LSX1
- M/F : Cette désignation permet de séparer les locuteurs selon le genre, dans le cas où nous avons deux locuteurs ayant les mêmes initiales mais de sexes opposés.
  - Exemples :
    −Lahcen Siham —> (f) LSX0
    −Larbi Sami —> (m) LSX0
- NIns : Le NIns nous renseigne sur le niveau d'instruction de chaque individu. L'objectif est de récolter un large éventail de données incluant toutes les couches de la population, locuteurs instruits ou non, (voir table 2.5).

Désignation	Signification
NM	niveau d'instruction moyen (Primaire, Moyen et Lycée)
NU	niveau d'instruction universitaire (Licence, Ingéniorat, etc.)
NH	très haut niveau d'instruction (Magister, Doctorat, etc.)

TABLE 2.5: Codification des niveaux d'instruction

A partir de toutes ces informations, chaque fichier d'ALGASD est codé comme suit :

ALGASD- R1-mrsa0-Cr1 :

- IDbase : ALGASD
- Région (1) : Alger
- Genre locuteur (m) : masculin
- ID locuteur (rsa0) : nom+prénom + 0
- Corpus (Cr) : Corpus réservé
- N° de la phrase (1) : première phrase de Cr

2.5.5 Conditions d'enregistrement

La charpente d'ALGASD terminée, la seconde étape de la réalisation se rapporte à la procédure d'enregistrement. Pour que l'acquisition des signaux de parole soit de haute qualité, une recherche documentaire concernant les conditions d'enregistrements et le matériel adéquat à utiliser a été nécessaire [Baude et al., 2006]. A l'issue de cette prospection, nous avons convenu que les enregistrements se fassent en intérieur, dans des milieux calmes, respectant les mêmes conditions et précautions de prise de son, c'est à dire, utilisation d'un matériel d'enregistrement similaire pour toutes les régions. Cette méthode contrôlée vise à maîtriser le maximum de paramètres de variations extra-locuteurs, ce qui permet un suivi permanent de tout le processus d'acquisition du signal parole. Le matériel utilisé, les modalités ainsi que les conditions techniques d'enregistrement sont comme suit :

2.5.5.1 Matériel et lieu d'enregistrement

Pour une meilleure acquisition des signaux de parole, un matériel de bonne qualité est recommandé qu'il s'agisse des microphones où bien des plateformes d'enregistrement employées. En ce qui concerne les microphones, nous avons utilisé 2 types de microphones professionnels à savoir : Shure MS 58 et PHILIPS MD-109 (plus de 80% des enregistrements ont été fait avec le Shure MS 58). Pour ce qui est du support d'enregistrements, plusieurs modèles existent, mais nous avons pour notre part, opté pour des micro-ordinateurs - portable et de bureau - de marque IBM et HP dotés de cartes son 16 bits. Les caractéristiques techniques du matériel sont détaillées en annexe F.

Le choix des lieux d'enregistrements était orienté vers l'utilisation d'endroits de dimensions similaires, calmes, fermés, meublés et connus par les locuteurs. En ce qui concerne les bruits ambiants (rapport signal/ bruit) et l'*effet de masque* [Boite, 2000], ils sont relativement très faibles. Toutes ces précautions ont été respectées pour avoir une acoustique relativement identique durant toutes les séances qui ont été nécessaires pour l'acquisition des signaux sonores du corpus.

2.5.5.2 Spécifications techniques

Les spécifications techniques sont :

- Enregistrement et numérisation avec l'un des deux logiciels libres *Wavesurfer* [Sjolander and Beskow, 2000] ou *Praat* [Boersma and Weenink, 2009].
- Respect des paramètres techniques suivants :
  - nombre de canaux: mono
  - fréquence d'échantillonnage *fe* : 16 kHz
  - codage en nombre de bits: 16 bits PCM signé
  - format d'enregistrement: wave (.wav)
- Eviter la saturation du signal et l'effet *Larsen* [Fillon, 2004].

2.5.5.3 Spécifications de lecture

Avant d'entamer l'enregistrement, il est important de mettre en confiance les locuteurs. En effet, en plus de l'utilisation de lieux connus par les participants, il est important de leur expliquer le but du corpus ALGASD, ce qu'ils devront faire et surtout, leur laisser le temps de se préparer et de ne lancer l'enregistrement que lorsqu' ils sont prédisposés à le faire. Les recommandations de lecture sont :

- Lecture avec un débit d'élocution normal (3-4 syllabes/seconde) ;
- Délimitation de l'enregistrement par deux silences ;
- Réduction au maximum de l'effet *Lombard* (modification de la prononciation (fréquence fondamentale, intensité, articulation, allongement des voyelles) pour compenser la présence de bruits environnants) ;
- Vérification perceptive des enregistrements et inspection visuelle des spectrogrammes ;
- Sélection de la meilleure lecture parmi plusieurs ;

- Suppression des phrases contenant des hésitations ;
- Réenregistrement des phrases mal exprimées (erreur de prononciation, signal trop faible, etc.).

2.6 PHASE D'ENREGISTREMENT

La procédure d'enregistrement des trois corpora d'ALGASD s'est effectuée de la manière suivante :

2.6.1 Enregistrement du corpus Cc

Le texte du Corpus commun (Cc), constitué de deux phrases dialectales (Cc1 et Cc2), a été lu par l'ensemble des locuteurs d'ALGASD à savoir les 300 participants. De ce fait, le nombre d'enregistrements atteint 600 phrases. Le tableau 2.6 illustre en détail le nombre d'enregistrements réalisés pour chacune des deux phrases selon le genre (hommes (H) et femmes (F)) ainsi que leurs régions d'appartenance.

Régions	Cc1		Cc2		Enregistrements
	F	H	F	H	
R1	40	40	40	40	160
R2	17	17	17	17	68
R3	13	12	13	12	50
R4	13	12	13	12	50
R5	9	9	9	9	36
R6	9	8	9	8	34
R7	19	19	19	19	76
R8	13	13	13	13	52
R9	4	3	4	3	14
R10	8	8	8	8	32
R11	7	7	7	7	28
Total					600

TABLE 2.6: Répartition du corpus (Cc) par région et par genre (F : femme, H : homme)

2.6.2 Enregistrement du corpus Cr

L'enregistrement de (Cr) s'est effectué en deux étapes : une phase de répartition simple suivie d'une optimisation dans les enregistrements.

Première étape A la base, le Corpus réservé (Cr) était doté de 30 phrases utilisées dans la formation de 10 textes de 3 phrases chacun. La première étape consistait en une distribution périodique de ces 10 textes sur les 11 régions d'ALGASD pour être lus par 3 locuteurs (2 hommes et 1 femme) sauf pour R9 où nous avons seulement 2 locuteurs (1 homme et 1 femme). Le nombre total de textes lus est alors de 32.

Optimisation des enregistrements Afin d'augmenter le nombre d'enregistrements, trois suggestions sont possibles : augmenter les textes à lire ou bien le nombre de locuteurs sinon les deux à la fois. Pour optimiser les enregistrements de Cr nous avons opté pour la dernière proposition.

L'accroissement des textes à lire s'est effectué par le rajout d'une phrase différente à chaque texte pour en former au total 32 nouveaux textes $(t_1, ..., t_{32})$ de 4 phrases

chacun. Pour ce faire, 32 nouvelles phrases ont été utilisées. Par conséquent, le nombre total de phrases de Cr passe de 30 à 62 phrases. Le nombre de locuteurs à ce niveau de l'enregistrement est de 32 participants. Le tableau suivant expose la distribution optimisée provisoire des textes par région et par locuteur.

Régions	Textes des locuteurs masculins	Textes des locuteurs féminins
R1	$(t_1 + t_2) \rightarrow$ 2H	$t_3 \rightarrow$ 1F
R2	$(t_4 + t_5) \rightarrow$ 2H	$t_6 \rightarrow$ 1F
R3	$(t_7 + t_8) \rightarrow$ 2H	$t_9 \rightarrow$ 1F
R4	$(t_{10} + t_{11}) \rightarrow$ 2H	$t_{12} \rightarrow$ 1F
R5	$(t_{13} + t_{14}) \rightarrow$ 2H	$t_{15} \rightarrow$ 1F
R6	$(t_{16} + t_{17}) \rightarrow$ 2H	$t_{18} \rightarrow$ 1F
R7	$(t_{19} + t_{20}) \rightarrow$ 2H	$t_{21} \rightarrow$ 1F
R8	$(t_{22} + t_{23}) \rightarrow$ 2H	$t_{24} \rightarrow$ 1F
R9	$(t_{25}) \rightarrow$ 1H	$t_{26} \rightarrow$ 1F
R10	$(t_{27} + t_{28}) \rightarrow$ 2H	$t_{29} \rightarrow$ 1F
R11	$(t_{30} + t_{31}) \rightarrow$ 2H	$t_{32} \rightarrow$ 1F
Total	32 textes	

TABLE 2.7: Répartition provisoire des textes et des locuteurs

L'augmentation du nombre total de phrases enregistrées par l'optimisation du nombre de locuteurs consiste, dans sa plus simple configuration, à faire participer pour l'enregistrement le même nombre de locuteurs pour chaque région comme dans TIMIT. Cependant, rajouter un nombre identique de lectures pour chaque texte est impossible dans notre cas, car nous sommes tenus de respecter d'une part, un total fixe de locuteurs pour toute la base à savoir 300 locuteurs contrairement à TIMIT qui est libre de cette contrainte, et d'autre part, une répartition irrégulière du nombre de locuteurs par région qui est proportionnelle à la population réelle dans celle-ci (voir tableau 2.2). Par conséquent, le nouveau nombre de locuteurs a été calculé statistiquement en respectant le genre et la région des locuteurs. Les détails de cette nouvelle répartition sont exposés dans le tableau ci-dessous. Nous remarquons à partir des résultats que R1 et R9 sont respectivement dotées du plus haut et plus bas nombre de lectures conformément aux statistiques calculées.

Régions	Hommes		Femmes	Textes/Régions
R1	t_1x6	t_2x6	t_3x11	23
R2	t_4x3	t_5x2	t_6x5	10
R3	t_7x2	t_8x2	t_9x3	7
R4	t_{10}x2	t_{11}x2	t_{12}x3	7
R5	t_{13}x2	t_{14}x1	t_{15}x2	5
R6	t_{16}x2	t_{17}x1	t_{18}x2	5
R7	t_{19}x3	t_{20}x3	t_{21}x5	11
R8	t_{22}x2	t_{23}x2	t_{24}x3	7
R9	t_{25}x1	_	t_{26}x1	2
R10	t_{27}x2	t_{28}x1	t_{29}x2	5
R11	t_{30}x1	t_{31}x1	t_{32}x2	4

TABLE 2.8: Répartition finale des textes du corpus Cr par région

Le nombre d'enregistrements de Cr atteint alors 344 phrases lues par un total de 86 locuteurs. Le tableau suivant montre le nombre de locuteurs et d'enregistrements par région et par genre.

Régions	Hommes	Femmes	Enregistrements
R1	12	11	92
R2	5	5	40
R3	4	3	28
R4	4	3	28
R5	3	2	20
R6	3	2	20
R7	6	5	44
R8	4	3	28
R9	1	1	8
R10	3	2	20
R11	2	2	16
	47	39	
Total	86		344

TABLE 2.9: Enregistrements du corpus Cr par locuteur et région

2.6.3 Enregistrement du corpus Ci

Les 136 phrases constituant le corpus individuel (Ci) ont été distribuées aléatoirement sur les 11 régions d'ALGASD et lues par 136 locuteurs (1 phrase par locuteur) en respectant toujours les rapports indiqués dans le tableau 2.2. Cependant, lors de la répartition de Ci, la région R9 devait être dotée uniquement de 3 phrases. Mais lorsque la distribution sur toutes les régions est terminée, 2 phrases n'ont pas été attribuées (restées). Ces deux phrases ont été rajoutées à R9 afin d'augmenter

le nombre de ses enregistrements. Le tableau 2.10 expose la répartition de Ci sur l'ensemble des régions selon le genre.

2.6.4 ALGASD dans sa globalité sonore

A l'issue de l'enregistrement des 3 corpora, chaque locuteur a lu un ensemble de textes constitués de 2, 3 ou 6 phrases :

- 2 phrases dialectales communes Cc
- 1 phrase individuelle de Ci avec les 2 phrases dialectales communes Cc
- 4 phrases réservées de Cr avec les 2 phrases dialectales communes Cc

Régions	Hommes	Femmes	Enregistrements
R1	18	18	36
R2	7	8	15
R3	6	5	11
R4	8	3	11
R5	2	6	8
R6	4	4	8
R7	8	9	17
R8	5	7	12
R9	3	2	5
R10	3	4	7
R11	2	4	6
Total	66	70	136

TABLE 2.10: Répartition de Ci par genre et par région

Le tableau 2.11 montre le nombre de locuteurs par régions pour lesquels sont attribués les trois types de textes cités ci-dessus.

	Régions	Nombres de Phrases		
		6	3	2
Nombre de Locuteurs	R1	23	36	21
	R2	10	15	9
	R3	7	11	7
	R4	7	11	7
	R5	5	8	5
	R6	5	8	4
	R7	11	17	10
	R8	7	12	7
	R9	2	5	0
	R10	5	7	4
	R11	4	6	4
	TOTAL	86	136	78

TABLE 2.11: Nombre de phrases par locuteurs

Les enregistrements d'ALGASD, tous les corpora confondus, totalisent 1080 phrases lues : 600 enregistrements pour Cc, 344 pour Cr et 136 pour Ci (tableau 2.12).

Corpus	Phrases	Locuteurs	Enregistrements
Corpus Commun (Cc)	2	300	600 (55.5%)
Corpus Réservé (Cr)	62	86	344 (31.8%)
Corpus Individuel (Ci)	136	136	136 (12.5%)
TOTAL	200	300	1080 (100%)

TABLE 2.12: Principaux résultats d'ALGASD

2.7 FICHIERS D'ALGASD

En plus des enregistrements sonores, ALGASD est doté de fichiers textuels contenant les alignements des unités de parole avec les transcriptions phonétiques correspondantes. Ces alignements sont le résultat de deux tâches coûteuses en temps et en efforts à savoir: la segmentation et l'étiquetage.

2.7.1 Segmentation

La segmentation de la parole continue en unités plus petites constitue un aspect crucial du traitement de la parole car elle vise à extraire du continuum acoustique des unités correctement délimitées.

Deux méthodes de segmentation existent: la manuelle et l'automatique. Tandis que la première est fastidieuse, la seconde méthode est moins contraignante mais présente en revanche plus d'erreurs que son homologue manuelle.

Les fichiers sons d'ALGASD ont tous été analysés avec les mêmes logiciels que ceux utilisés pour l'enregistrement à savoir Wavesurfer et Praat. En effet, chaque fichier a été segmenté manuellement en phones par la perception ainsi que par l'inspection visuelle de leurs ondes temporelles et leurs spectrogrammes. L'objectif étant de déterminer les zones de coarticulations (transitions formantiques) afin de délimiter les frontières de ces phonèmes. Pour ce faire, les critères standards de segmentation ont été suivis [Turk et al., 2006]. Les deux logiciels d'analyses cités ci-dessus offrent trois niveaux de segmentation : phrase, mot ou bien phonème (la plus petite unité susceptible, par sa présence, de changer la signification d'un mot).

2.7.2 Etiquetage

La seconde étape de l'alignement est l'étiquetage. Elle consiste à décoder le segment pour lui attribuer une appellation où une transcription adéquate. Cette

opération peut se faire simultanément avec la segmentation. La transcription des unités sonores est basée sur les symboles de la transcription SAMPA à laquelle est rajoutée le symbole (sil), indiquant le silence qui précède le signal parole ou lui succède. La figure suivante illustre un exemple de segmentation et d'étiquetage réalisé avec le logiciel Wavesurfer sur un signal de la base.

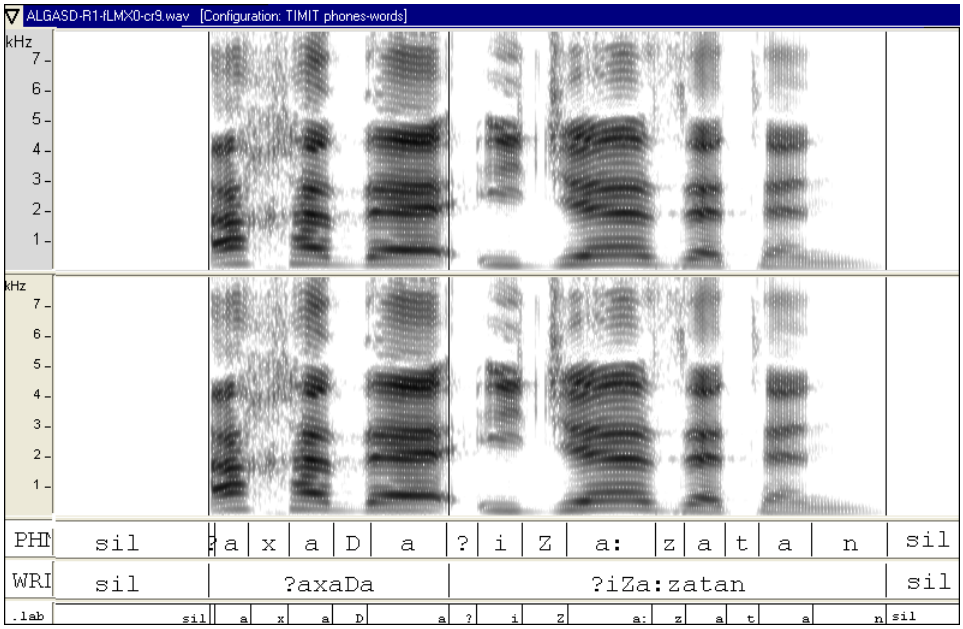


FIGURE 2.3: Exemple de segmentation et d'étiquetage

2.8 Fichiers Associés

A partir de cet alignement, nous avons associé à chaque fichier sonore (.wav) :

- la phrase textuelle orthographique correspondante (.txt) ;
- la phrase transcrite phonétiquement avec SAMPA (.txt) ;
- les mots de la phrase balisés dans le temps (.wrd) ;
- les phonèmes délimités dans le temps pour chaque mot de la phrase (.phn) et (.lab).

Le tableau suivant montre un exemple de fichiers associés.

ALGASD-R1-fAAX0.ccl	
qa:dana: wa-lam yad`t`ahidkum	**ALGASD-R1-fAAX0-ccl.txt**
0 4776 silence 4776 16889 qa:dana: 16889 22864 wa-lam 22864 38233 lam yad`t`ahidkum 38233 43206 silence	**ALGASD-R1-fAAX0-ccl.wrd**
0.0000000 0.2985200 sil 0.2985200 0.3626418 q 0.3626418 0.5823855 a: 0.5823855 0.6713191 d 0.6713191 0.7529135 a 0.7529135 0.8306225 n 0.8306225 1.0555468 a: . . 2.3895507 2.7003866 sil	**ALGASD-R1-fAAX0-ccl.phn (.lab)**

TABLE 2.13: Exemples de fichiers associés

2.8.1 Caractéristiques d'ALGASD

En résumé, les caractéristiques du corpus ALGASD sont donc :

1. Locuteurs : 300
2. Variabilité Régionale : 11
3. Corpus Textuel : 200 phrases phonétiquement équilibrées
4. Texte/ locuteur : 2 à 6 phrases.
5. Total fichiers/locuteur : 8-24
6. Total fichiers : 4632
7. Total Fiches Techniques : 336 = (300 (locuteurs) + 2 fiches /régions + 1 texte/région).
8. Taille de la base : 117 MB
9. Informations Fournies :

   (1) Répartition des locuteurs selon la distribution officielle de la population

   (2) Genre du locuteur (152 femmes/148 hommes)

   (3) Niveaux d'instruction des locuteurs

   (4) Âges des locuteurs

   (5) Textes avec les signes diacritiques

   (6) Segmentation manuelle des signaux de parole

   (7) Etiquetage (transcription) manuel des signaux de parole avec l'alphabet universel SAMPA sur deux niveaux : mots et phonèmes.

2.8.2 Organisation

Une base de données est un lot d'informations stockées d'une manière organisée et structurée. Dans la mesure du possible, ces données doivent pouvoir être à la disposition des utilisateurs pour une consultation, une saisie ou bien une mise à jour. Ces différentes manipulations requièrent donc un système de gestion que l'on appelle *Système de Gestion de Bases de Données* (SGBD) [Adiba and Collet, 1993]. Les données d'ALGASD ont été organisées de deux manières différentes. Tandis que la première est classique (système d'arborescence), la seconde organisation est sous forme d'application programmée spécialement conçue pour la gestion de ces données.

2.8.2.1 Organisation classique

Les données sont stockées et organisées sous forme d'arborescence. Chaque dossier, représentant l'une des 11 régions étudiées, contient plusieurs sous-répertoires portant les codes des locuteurs. Le nombre de répertoires dépend du nombre de locuteurs de chaque région. Le sous répertoire contient à son tour tous les fichiers associés (.txt, .lab, .wrd, etc.).

2.8.2.2 Application programmée

Bien qu'il existe de nombreux systèmes de gestion de bases de données comme : Borland Paradox, IBM DB2, Interbase, etc., nous avons programmé une application personnalisée, adaptée à notre base et qui se charge de la gestion des différentes données hétérogènes proposées par ALGASD (fichiers waves, textes, etc.) (figure 2.4). Cette application caractérisée par une architecture simple est programmée grâce à l'utilisation du langage de programmation BORLAND-DELPHI.7.

L'application est basée sur l'utilisation d'une interface principale présentant une panoplie de boutons permettant de réaliser plusieurs requêtes comme :

- la *recherche globale* sur les données de la base (les textes, les fichiers sons, les locuteurs, les fiches de renseignements, etc.) avec une possibilité d'affinage de cette recherche selon les facteurs sociaux des locuteurs tel le genre.
- la *recherche avancée* qui permet de visualiser des informations ciblées relatives à une région bien définie d'ALGASD avec toujours l'option d'affinage de la recherche.
- de détailler les 3 corpora textuels d'ALGASD (Cc; Cr et Ci);
- d'accéder à des informations générales sur le système phonétique arabe, sur l'Algérie etc.

2.9 CONCLUSION

La base de données ALGASD se caractérise par de nombreux aspects comme : une haute qualité des enregistrements, un grand nombre de locuteurs, des facteurs sociaux reflétant de nombreuses différences entre les locuteurs comme : l'âge, le genre, les niveaux d'instruction et les variétés dialectales. Toutes ces caractéristiques offrent une intéressante ouverture sur de nouvelles perspectives de recherches dans le domaine des sciences du langage et de la communication parlée comme : la reconnaissance automatique de la parole, l'analyse phonétique et acoustique de la langue arabe, les études sur la perception des signaux de parole, la classification des différentes variétés régionales parlées en Algérie, les études prosodiques, comparaison de l'AS des Algériens avec celui des pays du Maghreb ou autres, etc.

Figure 2.4: L'application d'ALGASD

La base de données ALGASD a été utilisée jusqu'à présent dans différents domaines [Droua-Hamdani et al., 2010b] : Etudes acoustique et statistique des variations qualitatives et quantitatives des voyelles courtes et longues de l'Arabe en fonction des niveaux d'instruction des locuteurs [Droua-Hamdani et al., 2009]. Etude prosodique sur la classification du rythme de la parole [Droua-Hamdani et al., 2010a].

En ce qui concerne son utilisation dans la reconnaissance automatique de la parole continue, nous avons plusieurs travaux : reconnaissance vocale basée sur des modèles monophones [Droua-Hamdani et al., 2010c], reconnaissance vocale basée sur des modèles triphones [Droua-Hamdani et al., 2012a]. De même que ce corpus a servi dans l'étude de l'effet de la variabilité régionale et le genre du locuteur sur les performances d'un SRAP [Droua-Hamdani et al., 2012b, 2013].

Chapitre 3

RYTHME & NOTIONS DE STATISTIQUE DESCRIPTIVE

3.1 INTRODUCTION

L'organisation temporelle des énoncés connaît un regain d'intérêt à tous les niveaux de recherches liés à la communication parlée : acoustique, articulatoire et perceptif. Elle s'est octroyée une place de choix depuis la quantification, plus que positive, de son impact sur les performances des systèmes automatiques qui ont la langue comme substrat de communication notamment la synthèse automatique de la parole [Droua-Hamdani and Guerti, 2007]. En effet, pour générer une parole artificielle approchant le plus possible la parole humaine, des réajustements sur les durées des unités sonores sont souvent nécessaires. Ceci se traduit au niveau de la phrase par des modifications du rythme de la parole. Ce même intérêt est observé dans le domaine de l'identification automatique des langues où l'on fait souvent appel au rythme pour pouvoir optimiser le taux de reconnaissance d'une langue donnée parmi un ensemble de langues. Aussi, plusieurs études ont vu dans l'intégration de ce paramètre prosodique, dans les systèmes de reconnaissance automatique de la parole ou dans l'identification du locuteur, un moyen important pour accroître leurs performances.

Les études expérimentales se rapportant à l'aspect temporel de la parole sont nombreuses, notamment celles qui ont trait à l'analyse de la durée segmentale, et ce pour toutes les langues y compris pour la langue arabe [Klatt, 1976, De Jong and Zawaydeh, 1999, Van Santen, 1992]. De même, les travaux théoriques portant sur le rythme des langues ont intéressé beaucoup de chercheurs depuis plusieurs décennies. En revanche, les études expérimentales qui abordent son étude sont assez récentes [Jang, 2009, O Rourke, 2008, Ling et al., 2000].

Les investigations dédiées à l'étude du rythme de la langue arabe sont très peu

fréquentes comparées aux autres langues [Hamdi et al., 2004].

3.2 PROSODIE

La prosodie est une source supplémentaire de connaissances qui est inhérente et exclusive à la langue parlée. La linguistique lui accorde le traitement de tous les phénomènes dits suprasegmentaux qui échappent au découpage de la chaîne parlée. Elle concerne, entre autre, l'étude de la perception auditive véhiculée au niveau de la phrase par l'accentuation, l'intonation et le rythme. Ces derniers paramètres se traduisent sur le plan acoustique par la variation de : l'intensité, la fréquence fondamentale (pitch) et la durée segmentale. La figure 3.1 illustre la variation inter locuteur. Elle montre l'analyse acoustique de la prosodie (mélodie, intensité et rythme) de l'une des phrases du corpus ALGASD, prononcée par deux locuteurs de la base. Les courbes de l'énergie et de la mélodie (intonation) sont représentées sur les spectrogrammes. La durée des phonèmes varie aussi dans les deux spectrogrammes entraînant ainsi la variation du rythme de la parole.

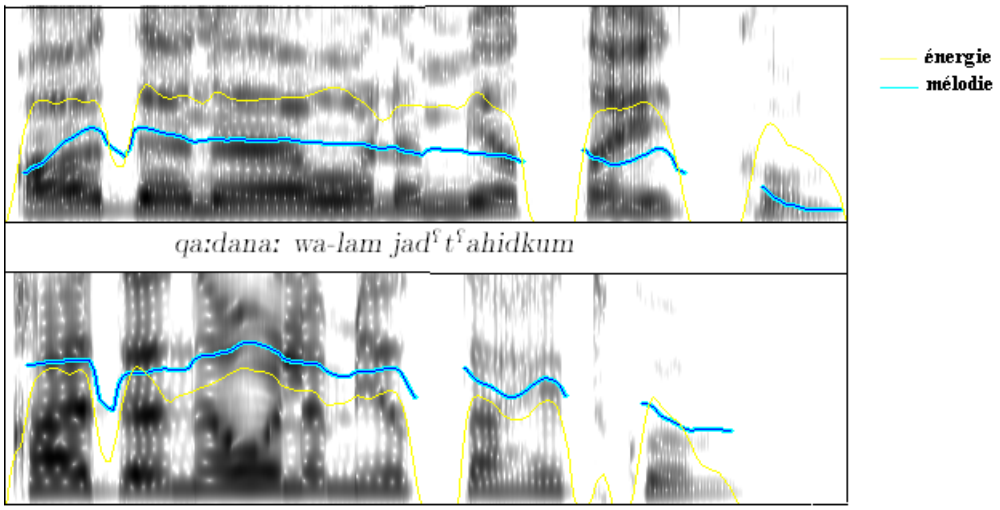


FIGURE 3.1: Spectrogrammes d'une même phrase prononcée par deux locuteurs

La prosodie a longtemps été étudiée comme une source importante de compréhension de la parole. Ces dernières années, beaucoup de travaux destinés à sa modélisation ont été menés pour différentes applications comme dans la synthèse et la reconnaissance automatiques de la parole et du locuteur.

– La modélisation de la prosodie est utilisée pour le lissage de la voix articficielle générée par les systèmes de synthèse automatique de la parole (Text-To-Speech TTS) (figure 3.2). L'objectif du lissage est la réduction des discontinuités observées aux lieux de jonction des segments de parole afin d'approcher le plus possible la voix naturelle (humaine). La figure montre les spectrogrammes

d'une phrase du corpus ALGASD prononcée par un des locuteurs de la base vs la même phrase lue par un système de synthèse automatqiue de la parole.

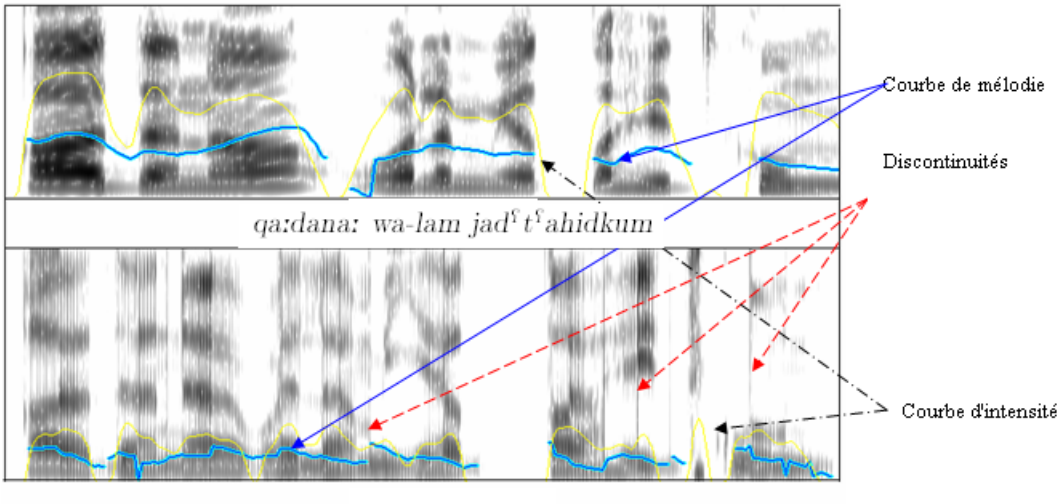


FIGURE 3.2: Exemple d'une phrase générée : naturellement par un locuteur d'ALGASD (haut) et artificiellement par un système de synthèse automatique (bas)

Au niveau du signal, le lissage est opéré par l'application de techniques de synthèse sur les différents paramètres acoustiques. La figure illustre des modifications apportées sur la durée d'une portion de signal parole. En effet, elle montre une réduction vocalique de la voyelle longue [a] par une méthode de synthèse appelée OLA (OverLap and Add) [Huang et al., 2001].

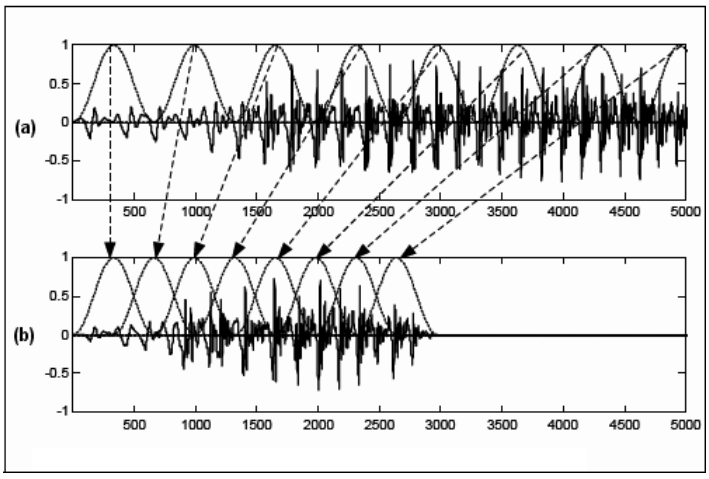


FIGURE 3.3: Modification de la prosodie d'un signal : réduction de la durée segmentale par la méthode de synthèse OverLap and Add (OLA) : (a) signal temporel (b) signal synthétisé

– L'objectif général par la modélisation des aspects prosodiques de la parole en reconnaissance vocale est l'amélioration de la performance des SRAP. L'utilisation de la prosodie dans les applications de compréhension de la parole est

importante, et il y existe un nombre croissant d'applications qui sont à l'étude. Les exemples incluent la résolution des ambiguïtés, la détection des limites de phrases, des erreurs de prononciation, etc. De même que la prosodie, par le biais de ses paramètres acoustiques (fréquences fondamentale, durée et intensité), est également utilisée dans les travaux sur la reconnaissance du locuteur [Shriberg and Stolcke, 2004, Wang, 2001, Waibel and Weibel, 1988].

3.3 ETUDE DU RYTHME DANS LA PAROLE

Parmi les corrélats acoustiques de la prosodie, nous nous intéressons particulièrement à l'organisation temporelle des unités de parole autrement dit à l'évolution du rythme dans la phrase (la durée des phonèmes, pauses, etc.).

3.3.1 Définition du rythme

La notion du rythme recouvre une multitude d'interprétations selon le domaine [Astésano, 2001]. Toutefois, les auteurs s'accordent à penser qu'il appartient essentiellement au domaine de la perception où il se révèle dans la répétition ou l'alternance périodique d'éléments saillants selon un schéma ou un modèle donné [Di Cristo and Hirst, 1993]. Au niveau acoustique, le rythme désigne la structure qui régit les durées des sons successifs (rallongements ou raccourcissements) et son estimation se traduit par la mesure de suites de durées : des phonèmes, des pauses du discours, des segments [CV], [CVC], etc. ce qui le relie au débit d'élocution ainsi que le nombre de pauses générées dans la phrase [Astésano, 2001].

3.3.2 Durées segmentales

Chaque phonème se caractérise selon son contexte par deux type de durées intrinsèques et co-intrinsèques. Les durées segmentales ont fait l'objet de plusieurs travaux et expérimentations surtout dans le domaine de la synthèse automatique de la parole [Klatt, 1976].

3.3.3 Débit

La notion de débit est souvent associée à la vitesse d'élocution. Considéré comme le principal organisateur de la dynamique verbale, son rôle est d'établir une gestion et un conditionnement des séquences de parole (allongement ou compression). Ainsi, selon qu'un locuteur parle lentement ou rapidement, les durées segmentales des phonèmes varient en conséquence. En effet, à un débit de parole donné, des

changements directs dans la structure rythmique sont perçus. Chaque niveau (segmental, lexical, syllabique, etc.) doit se soumettre à ces variations pour produire des durées qui respectent l'harmonie générale du message (conservation des contrastes phonémiques). Les différents facteurs qui peuvent altérer le débit sont nombreux, parmi eux nous citons: la variation du nombre de pauses générées, leurs durées, la fusion vocalique, les variations de durées segmentales.

3.3.4 Pauses

Les pauses générées suite à des phénomènes linguistiques ou extralinguistiques ont pour objectif la perception et la compréhension du message. Elles apparaissent, quant elles ne sont pas uniquement dues à une gêne respiratoire ou émotionnelle, pour supporter des fonctions particulières dans le message telles que les fonctions grammaticales et sémantiques. Les pauses silencieuses ou remplies se matérialisent dans l'incertitude de l'énoncé, le temps nécessaire pour accéder aux informations lexicales, la mise en relief d'une certaine partie du texte, etc.

3.4 RYTHMES DES LANGUES : VARIATIONS & TYPOLOGIE

La classification des langues et des dialectes selon leurs rythmes a inspiré, depuis de nombreuses décennies, plusieurs travaux de recherches théoriques [Pike, 1979, Abercrombie, 1967, Ladefoged and Johnson, 2010, Bloch, 1950]. La typologie rythmique s'est accordée dès lors à classer les langues en trois grandes catégories :

- Les langues syllabiques (syllable-timed) : La syllabe est décrite sur le plan acoustique comme une unité phonétique qui se prononce d'une seule émission de voix quelque soit sa constitution : CV, CVC, CVCC, etc. Les langues syllabiques sont caractérisées par des syllabes (groupes de sons) dont les intervalles de temps de prononciation sont équivalents, indépendamment de l'accent (intensité) mis sur certaines d'entre elles. Parmi les langues syllabiques nous citons : le Français, l'Espagnol, le Turc, etc. [Pamies Bertrán, 1999].
- Les langues accentuelles (stress-timed) se caractérisent, quant à elles, par la prononciation de syllabes accentuées à des temps approximativement constants. Ceci, indépendamment du nombre et de la durée des syllabes inaccentuées se trouvant entre deux accents. Nous citons pour exemple : l'Anglais, l'Allemand, le Néerlandais, l'Arabe, etc.
- Les langues moraiques (mora-timed) qui sont des langues basées sur la more (une unité plus fine que la syllabe) comme le japonais [Labrune, 2001].

Pour confirmer ou infirmer les catégorisations rythmiques interlangues diffusées dans la littérature, des travaux expérimentaux ont été entrepris ces dernières années afin d'exploiter de nouvelles pistes d'analyse visant la formalisation de toutes ces suggestions de classifications. Ces études expérimentales soulignent le début de l'ère de la modélisation du rythme de la parole.

3.5 MODELISATION DU RYTHME

Les chercheurs ont développé un certain nombre de paramètres qui permettent de quantifier le rythme dans les langues [Ramus, 2002, Grabe, 2002, Dellwo, 2006, Arvaniti, 2009]. Ces paramètres sont basés sur la mesure acoustique de la durée des intervalles vocaliques et consonantiques dans la parole continue. Ils peuvent être calculés dans les deux formes simple et normalisée. Bien que ces mesures ne donnent qu'une image partielle de la notion générale du rythme dans la langue, comme l'ont souligné [Arvaniti, 2009, Asu and Nolan, 2005], ces paramètres sont un moyen formel, qui vient conforter ou non, la classification des langues. De ce fait, trois types de mesures sont exploitées : les mesures d'intervalles (IM) proposées par [Ramus et al., 1999], les mesures normalisées établies par [Dellwo et al., 2004] et enfin les Pairewise Variability Indices (PVI) développés par [Grabe, 2002].

3.5.1 Mesures d'intervalles (IM)

L'approche suggérée par [Ramus, 2002] consiste à calculer trois variables considérées comme étant les corrélats acoustiques représentatifs des classes de rythme. Ces trois paramètres ou intervalles de mesure sont issus de la segmentation du signal de parole en ses unités de base, à savoir les voyelles et les consonnes. Aussi, la méthode propose, dans sa globalité, deux types d'intervalle, vocalique et consonantique, dans lesquels sont inclus, respectivement toutes les durées des voyelles et les séquence consécutives de voyelles, ainsi que toutes les durées de consonnes avec les durées des séquences successives de consonnes.

Les 3 variables IM sont :

- $\%V$: est la proportion totale des intervalles vocaliques inclus dans la phrase ;
- $\triangle V$: les écart types des intervalles vocaliques ;
- $\triangle C$: les écart types des intervalles consonantiques.

L'analyse a été effectuée sur un corpus de phrases, de durées moyennes ($\approx$ 3 secondes), enregistrées en huit langues : Anglais, Néerlandais, Polonais, Français, Espagnol, Catalan, Italien et Japonais.

Les résultats expérimentaux de l'analyse rythmique ont montré une étroite corrélation (r) entre les variables ΔC et $\%V$ ($r = 0,93, p < 0.01$). Les scores calculés

ont permis de classer les langues étudiées en trois groupes correspondant aux classes rythmiques prédéfinies dans la typologie des langues. En conclusion, la projection plane de la combinaison des corrélats rythmiques (%V, $\triangle C$) fournit une catégorisation expérimentale des langues.

3.5.2 Mesures normalisées

Les expériences pratiques ont montré que $\triangle C$ et $\triangle V$ sont inversement proportionnels à la vitesse d'élocution, dès lors [Dellwo, 2006, White and Mattys, 2007] proposent l'utilisation d'une version normalisée des variables IM à savoir les *VarcoV* et *VarcoC* respectivement pour les calculs des voyelles et des consonnes. Ces nouvelles variables sont données par les formules suivantes :

$$VarcoV = \frac{\triangle V}{meanV} \tag{3.1}$$

$$VarcoC = \frac{\triangle C}{meanC} \tag{3.2}$$

avec $meanV$ et $meanC$ sont les moyennes respectives des durées vocaliques et consonantiques.

3.5.3 The Pairewise Variability Indices

Les Pairewise Variability Indices (PVI) sont des mesures quantitatives qui déterminent le niveau de variabilité entre les mesures qui se suivent. Ils calculent la différence de durées entre deux intervalles vocaliques et intervocaliques subséquents [Grabe, 2002].

Le raw Pairewise Variability Index ($rPVI$), réservé pour le calcul des durées consonantiques, est utilisé dans le débit de parole non normalisé. Sa formule est :

$$rPVI = \frac{\left(\sum_{k=1}^{m-1} \left|_{d_k} - _{d_{k+1}}\right|\right)}{m-1} \tag{3.3}$$

où m et d sont respectivement le nombre et la durée des intervalles.

La version normalisée du PVI est calculée à partir de la moyenne des différences des durées entre les intervalles successifs divisée par la somme de ces intervalles. Le résultat est par la suite multiplié par 100. La formule du $nPVI$, dédiée aux calculs des durées vocaliques, sert à corriger les fluctuations engendrées au niveau des voyelles. Le $nPVI$ est calculé avec :

$$nPVI = 100 \frac{\left(\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{|d_k + d_{k+1}|}\right)}{m-1} \tag{3.4}$$

où m et d sont respectivement le nombre et la durée des intervalles.

Nous notons pour nos expériences futures $rPVI$ et $nPVI$ respectivement par $rPVI - C$ et $nPVI - V$.

Les PVI ont été calculés à partir de l'enregistrement du texte " *la bise et le soleil* " disponible dans le manuel de [IPA, 1999] dans 18 langues et dialectes. Le corpus a été lu par un seul locuteur pour chaque langue sauf pour le Français et l'Espagnol où l'enregistrement a été effectué par 7 locuteurs.

3.6 ETUDES TRAITANT LE RYTHME DE LA LANGUE ARABE

La classification des études théoriques du rythme octroie à l'AS ainsi qu'à tous ses dialectes une place dans la catégorie accentuelle (stress-timed) [Abercrombie, 1967, Pike, 1979]. Par opposition à cette profusion de la littérature, les études rythmiques expérimentales dédiées à la langue arabe, sont quasi inexistantes hormis celle de [Hamdi et al., 2004] dont l'objet principal est la présentation d'une analyse acoustique des différentes structures rythmiques des dialectes arabes, afin d'examiner la possibilité d'en obtenir une typologie de sous-classes rythmiques liées aux variations prosodiques inter-dialectales. Pour ce faire, l'auteur a procédé à la description et à l'analyse des différents parlers arabes représentatifs du Maghreb et du Moyen-Orient en se basant sur les deux principales approches rythmiques citées ci-dessus, à savoir les IM mesures et les PVI (tableau 3.1).

Arabe Maghrébin	Arabe du Moyen-orient
Marocain (Rabat - Casablanca)	Libanais (Beyrouth)
Algérien (Alger - Jijel)	Jordanien (Irbid)
Tunisien (Tunis)	Égyptien (Caire)

TABLE 3.1: Les variétés dialectales arabes étudiées

Le corpus utilisé dans l'étude se compose de la traduction spontanée du texte de la *'La bise et le soleil'* . Le nombre de locuteurs par variété dialectale est de 10.

3.7 NOTIONS DE STATISTIQUE DESCRIPTIVE

Pour faciliter la compréhension de la démarche adoptée dans le chapitre suivant et ce lors de l'analyse du rythme de l'AS, nous avons jugé utile de donner dans la section suivante quelques notions essentielles de statistique comme l'analyse de la variance, considérée comme un test incontournable pour vérifier l'effet d'un facteur donné sur des échantillons de variables indépendantes. Ce test est basé sur la comparaison des moyennes des groupes existants dans l'ensemble des échantillons. Pour cela, nous rappelons tout d'abord les formules de calcul de la moyenne et de la variance, puis nous aborderons en détails l'analyse de cette dernière [Morgenthaler, 2007].

3.7.1 Moyenne et variance

Soient $y_1, y_2, ..., y_n$ des variables aléatoires indépendantes de même loi de répartition, d'espérance μ et de variance σ^2. La moyenne et la variance sont calculées comme suit:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\sigma^2 = \frac{1}{n-1}\sum(yi - \bar{Y})$$

avec n le nombre d'échantillons.

3.7.2 Principe de l'Analyse de la Variance (ANOVA)

L'Analyse de la Variance (ANalysis Of VAriance) ou l'analyse factorielle connue sous l'acronyme de ANOVA est l'un des tests fondamentaux des statistiques. Elle permet de vérifier si une ou plusieurs variables *dépendantes* (endogènes) sont en relation avec une ou plusieurs variables *indépendantes* ou *explicatives* (exogènes). En d'autres termes, elle est utilisée dans la comparaison des moyennes de k échantillons appartenant à des modalités différentes issues d'un ou de plusieurs critères (facteurs).

Les analyses de la variance se distinguent selon le nombre de facteurs intervenants. On parle d'analyse à un facteur, lorsque l'analyse porte sur un modèle décrit par un facteur de variabilité (One way ANOVA), d'analyse à deux facteurs (Two way ANOVA) ou d'analyse multifactorielle lorsque le nombre de facteurs est supérieur à 2 (MANOVA). Le calcul de la variance dépend du nombre de facteurs, de leurs critères (quantitatif ou qualitatif) ainsi que des différentes modalités qui leurs sont associées. Nous développons dans ce chapitre la démarche adoptée pour une One-way ANOVA, donc à un facteur, puis nous donnerons le modèle général qui est

le modèle multifactoriel.

3.7.2.1 Conditions d'application

L'analyse de la variance repose sur le test de Fisher qui nécessite, pour procéder au calcul, la vérification de deux conditions d'application : la première concerne la normalité de distribution des échantillons indépendants (variables dépendantes qui doivent suivre une *loi normale*) dont la moyenne dépend éventuellement de la valeur des facteurs. La seconde condition est l'*homoscédascticité* ou homogénéité des variances.

3.7.2.2 Hypothèse à tester

Soit le facteur contrôlé A à p modalités ($1 \leq i \leq p$). Soit j le nombre de répétitions pour une modalité donnée i noté n_i (le nombre de répétitions peut varier d'une modalité à une autre). La valeur de la variable aléatoire Y pour la modalité i du facteur A à la répétition j est notée y_{ij} et la valeur moyenne pour chaque modalité est nommée $\overline{y_i}$.

L'objectif de l'ANOVA est de savoir si une variable numérique a des valeurs significativement différentes ou non. Nous testons pour cela l'hypothèse nulle (H0) ou bien l'hypothèse alternative (H1) pour confirmer ou infirmer l'influence d'un facteur donné sur les variables mesurées.

Hypothèse nulle (H0) : Homogénéité des données

Sous l'hypothèse nulle (H0), l'effet du facteur A est inexistant car les p moyennes des différentes modalités sont égales à une même moyenne μ.

Donc sous H0 :

- $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_p = \mu$
- $y_{ij} = \mu + \varepsilon_{ij}$

ε_{ij} correspondent aux erreurs expérimentales pour chaque valeur de la variable y_{ij}. Elles suivent une même loi normale $\mathcal{N}(0, \sigma)$.

Hypothèse alternative H1 : Hétérogénéité des données

L'hypothèse alternative stipule qu'au moins deux moyennes sont significativement différentes. Ce qui revient à dire que l'effet du facteur A est non nul et agit sur les variables dépendantes. Cette influence est représentée par α_i.

Sous H1 les équations prennent la forme :

- $\mu_i \neq \mu_j$
- $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

avec :

- ε_{ij} : variables aléatoires indépendantes suivant une même loi normale $\mathcal{N}(0, \sigma)$.
- α_i : l'effet de la modalité i du facteur A sur la variable Y
- μ : moyenne des différents échantillons quelque soit la modalité.

Sous l'hypothèse H_1, certaines moyennes (ou toutes) sont différentes les unes des autres, suite à l'influence du facteur testé. Dés lors, nous attestons que ce facteur a un effet *significatif* sur la variable mesurée.

3.7.3 Modèle de l'ANOVA à un facteur

L'estimation des paramètres des modèles se fait :

Sous H0 :

$$y_{ij} = \mu + \varepsilon_{ij}$$
$$\overline{y} = \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij} \quad \text{avec} \quad N = \sum_{i=1}^{p} n_i$$

- p : nombre de modalités du facteur
- n : nombre d'éléments dans la modalité i

Sous H_0, chaque valeur y_{ij} peut être estimée à partir de la moyenne totale des y_{ij}, c'est à dire $\overline{y}$ à laquelle s'ajoute l'erreur de mesure (aléatoire) ε_{ij}.

Sous H1 :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{3.5}$$

soit : $\overline{y_i} = \mu + \alpha_i$ et $\overline{y_i}$ la moyenne des y_{ij} pour la modalité i avec $\overline{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ en remplaçant μ par $\overline{y}$, nous avons :

$$\alpha_i = \overline{y_i} - \overline{y} \tag{3.6}$$

et

$$\varepsilon_{ij} = y_{ij} - \overline{y} - \overline{y_i} + \overline{y} = y_{ij} - \overline{y_i}$$

ainsi :

$$\varepsilon_{ij} = y_{ij} - \overline{y_i} \tag{3.7}$$

3.7.3.1 Décomposition de la variance à un facteur

Le modèle général sous H_1 :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

en remplaçant les estimateurs (3.6) et (3.7) l'équation (3.5) devient :

$$y_{ij} = \overline{y} + (\overline{y_i} - \overline{y}) + (y_{ij} - \overline{y_i})$$

$$y_{ij} - \overline{y} \;\; = \;\; (\overline{y_i} - \overline{y}) + (y_{ij} - \overline{y_i})$$

avec l'écart quadratique :

$$(y_{ij} - \overline{y})^2 = (\overline{y_i} - \overline{y})^2 + (y_{ij} - \overline{y_i})^2 + 2(\overline{y_i} - \overline{y})(y_{ij} - \overline{y_i})$$

La somme sur tous les j nous donne :

$$\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = n_i(\overline{y_i} - \overline{y})^2 + \sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 + 2(\overline{y_i} - \overline{y})\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i}) \tag{3.8}$$

or $2(\overline{y_i} - \overline{y})\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i}) = 0$ car $\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i}) = 0$ (pas de variations de l'erreur entre les modalités ($E(\varepsilon_{ij}) = 0$)). L'équation (3.8) devient:

$$\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = n_i(\overline{y_i} - \overline{y})^2 + \sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 \tag{3.9}$$

En procédant à la somme de toutes les modalités du facteur, (3.9) prend la forme suivante:

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{p} n_i(\overline{y_i} - \overline{y})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 \tag{3.10}$$

$$SCT = SCE + SCR$$

où

- SCT correspond à la *variabilité totale* dans l'échantillon de données. Elle représente la somme des carrés totaux (indépendante des groupes). La dispersion totale est alors égale à: $\frac{SCT}{ddl_t}$ avec $ddl_t = N - 1$ et $N = \sum_{i=1}^{p} n_i$.
- SCE illustre la somme des carrés *interclasses*. Elle correspond à la variation entre groupes selon les modalités du facteur. La variance due au facteur A, appelée aussi CM_{inter}, est calculée avec: $\frac{SCE}{ddl_{inter}}$ avec $ddl_{inter} = p - 1$.
- SCR, appelée variabilité résiduelle, est la dispersion à l'intérieur des groupes. Elle correspond à la somme des carrés *intra-classes*. La dispersion résiduelle, appelée CM_{intra}, est calculée à partir de: $\frac{SCR}{ddl_{intra}}$ avec $ddl_{intra} = N - p$.

ddl_t, ddl_{inter} et ddl_{intra} correspondent aux degrés de liberté associés à chaque catégorie

3.7.3.2 Région critique

Après avoir calculé les sommes et les degrés de liberté, nous procédons au calcul d'un rapport F entre la variabilité expliquée et la variabilité résiduelle corrigée par les degrés de liberté:

$$F = \frac{CM_{inter}}{CM_{intra}} = \frac{\frac{SCE}{ddl_{inter}}}{\frac{SCR}{ddl_{intra}}} \tag{3.11}$$

Ce rapport est comparé par la suite à une valeur théorique $F_{1-\alpha}$ fournie par la table de *Fisher-Snedecor* pour un risque d'erreur α fixé et $(p-1, N-p)$ degrés de liberté. La région critique (R.C) du test au risque α s'écrit :

$$R.C. : \quad F \geq F_{1-\alpha}(p-1, N-p)$$

- si $F > F_{1-\alpha}$ l'hypothèse H_0 est rejetée au risque d'erreur α: le facteur contrôlé A a un effet significatif sur les valeurs de la variable étudiée.
- si $F \leq F_{1-\alpha}$ l'hypothèse H_0 est acceptée: le facteur contrôlé A n'exerce aucun effet significatif sur les valeurs de la variable étudiée.

La procédure générale de calcul de l'ANOVA pour deux facteurs ou plus est similaire à la démarche détaillée pour un facteur. Cependant, il faudra dans ce cas de figure, tester l'effet de chaque facteur individuellement puis les interactions entre facteurs.

3.7.4 Modèle multifactoriel

Pour pourvoir aborder l'analyse de la variance multifactorielle, nous souhaitons commencer par la démonstration de l'analyse à 2 facteurs pour passer ensuite à la généralisation au modèle multifactoriel.

Soient deux facteurs de variabilité pouvant prendre respectivement les niveaux $i = 1, ..., p$ et $j = 1, ..., q$, n_{ij} le nombre d'individus dans le niveau i du premier facteur et le niveau j du second facteur, n le nombre d'individus total et r le nombre d'individus dans chaque sous-groupe (pour un niveau i et un niveau j donnés). La variable à expliquer y_{ijk} s'écrit avec $i = 1, ..., p$, $j = 1, ..., n_i$ et $k = 1, ..., m_j$.

La variable à expliquer peut être modélisée par la relation :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{3.12}$$

avec:

- α_i l'effet du niveau i du premier facteur
- β_j l'effet du niveau j de second facteur
- γ_{ij} représente l'effet de l'intéraction entre les deux facteurs
- ε_{ijk} est l'erreur aléatoire

En remplaçant chaque terme ci-dessus, comme dans l'analyse à un facteur, nous obtenons:

$$SCT = SCE_{facteur1} + SCE_{facteur2} + SCE_{interaction} + SCR \tag{3.13}$$

où:

- $SCE_{facteur1} = rq\sum_{i=1}^{p}(\bar{y_i} - \bar{y})^2$
- $SCE_{facteur2} = rp\sum_{j=1}^{q}(\bar{y_j} - \bar{y})^2$
- $SCE_{interaction} = r\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y_{ij}} - \bar{y_i} - \bar{y_j} + \bar{y})^2$
- $SCR = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k-1}^{n_{ij}}(y_{ijk} - \bar{y_{ij}})^2$

Le modèle multifactoriel est donc la généralisation de 3.12:

$$Y_i = \mu + \sum_{j}\alpha_j + \sum_{j,k}\gamma_{ij} + \varepsilon_i \tag{3.14}$$

avec α_j l'effet du $j^{ème}$ facteur et γ_{ij} l'interaction entre le $j^{ème}$ et le $k^{ème}$ facteur.

3.8 CONCLUSION

Nous avons voulu aborder dans ce chapitre un *nouvel axe de recherche* presque méconnu des domaines de la communication parlée consacrée à la langue arabe. Ce domaine est le rythme de la parole. En effet, les études expérimentales traitant ce paramètre prosodique sont quasi inexistantes pour l'AS. Aussi, ce chapitre a constitué le préambule d'une étude proposée dans le chapitre suivant.

Nous avons donc tout d'abord décrit le rythme dans la parole, puis exposé les différentes classifications rythmiques théoriques de plusieurs langues du monde en citant la particularité de chaque classe rythmique. Nous avons présenté, par la suite, les plus importantes études expérimentales dans le domaine afin de donner au rythme une dimension formelle comme les modèles des intervalles de mesure (IM) et le modèle des Pairewise Variability Indices (PVI). Ces modèles sont essentiellement basés sur les durées consonantiques et vocaliques de la parole. Le chapitre a été clôturé par quelques notions fondamentales de statistiques descriptives nécessaires à l'étude du rythme.

Chapitre 4

RYTHME DANS L'ARABE STANDARD ALGERIEN

4.1 INTRODUCTION

Dans la typologie du rythme, l'AS est décrit comme appartenant à la famille stress-timed c'est à dire une langue accentuelle [Abercrombie, 1967]. Les études expérimentales sur l'AS sont beaucoup moins nombreuses que celles traitant des autres langues comme : l'Anglais, le Néerlandais, le Coréen, le Français, l'Espagnol, le Portugais, le Grec, etc. [Baltazani, 2007, Ramus et al., 1999, O Rourke, 2008]. Certaines de ces recherches ont montré que les paramètres rythmiques sont sensibles aux différences dialectales d'une langue, comme pour l'Anglais Américain [Thomas, 2007] et l'Italien [Giordano and DAnna, 2010]. En ce qui concerne l'Arabe, le peu de recherches trouvées dans la littérature s'intéressent beaucoup plus à la variation rythmique des dialectes arabes [Hamdi et al., 2004].

4.2 OBJECTIFS

Comme présentée dans le chapitre 3, l'utilisation de la prosodie dans le domaine de la communication parlée est très importante. Elle permet entre autre de lisser le signal de parole synthétique, d'accroître la performance des SRAP, etc. Cependant, pour savoir quantifier les taux de modifications à apporter aux systèmes, une analyse des paramètres prosodiques (pitch, durée et intensité) est nécessaire. Nous nous intéressons pour notre part à l'analyse de l'organisation temporelle des signaux de parole précisément au rythme de la parole. En étudiant cet aspect prosodique, nous poursuivons des travaux antérieurs dans le domaine [Droua-Hamdani and Guerti, 2007, Droua-Hamdani, 2007]. En revanche, les analyses effectuées cette fois-ci, sont à un niveau hiérarchique supérieur à celui des anciennes études (durée phonémique).

L'idée consiste donc en l'application des corrélats de mesures pour analyser la variabilité rythmique dans la prononciation de l'AS par des locuteurs Algériens tout en examinant si ces mesures sont sensibles aux différences apparentes entre ces locuteurs pour être considérées comme une source de variabilité discriminante. Notons que le manque de pratique quotidienne de l'AS par la majorité de la population Algérienne lui assigne *presque* la place d'une *seconde langue*.

Par ailleurs, pour obtenir une image plus complète des effets de variabilité possibles, notre étude comprend, outre le niveau d'instruction des locuteurs, deux autres facteurs de variabilité qui sont l'âge et le genre. Ces trois paramètres jouent un rôle prépondérant dans la structuration de la variation linguistique dans les sociétés Arabes autant que pour les autres sociétés [Bassiouney, 2009, Haeri, 2000]. Par ailleurs, [Wiget et al., 2010] constatent que les locuteurs, les corpora de parole ainsi que les appareils de mesure représentent aussi d'importantes sources de variabilité qui sollicitent une attention particulière des chercheurs, de même qu'ils exposent que les voyelles ont un pouvoir discriminant supérieur à celui des consonnes.

La seconde partie du chapitre concerne la comparaison du rythme de l'AS avec les langues du monde. Nous allons dans ce volet étudier la variation du rythme au niveau de la langue indépendamment de la variation intra locuteur. En effet, comme nous l'avons cité précédemment, la langue Arabe est traditionnellement située parmi les langues accentuelles. Mais qu'en est t-il de l'AS prononcé par des locuteurs algériens ?

4.3 METHODOLOGIE

La signal parole est variable tant par les contextes prononcés que par les locuteurs qui les produisent. Pour réduire ces variations et n'en garder que celles qui intéressent notre étude, nous avons contrôlé deux sources importantes de fluctuations : les erreurs de mesures relatives aux conditions d'enregistrements et celles issues des corpora (textes lus). Par l'utilisation d'ALGASD, ces deux paramètres semblent être relativement maîtrisés. Ce qui ramène notre principale source de variabilité essentiellement aux locuteurs (genre, âge et niveau d'instruction). Par ailleurs, pour considérer le maximum de cas, nous avons convenu d'utiliser un échantillon important de locuteurs contrairement aux autres travaux traitant le rythme.

4.3.1 Locuteurs

Afin d'étudier la corrélation entre la prononciation de l'AS et le niveau d'instruction des Algériens, nous avons prélevé à partir de la base complète d'ALGASD des enregistrements relatifs à la première région (R_1) qui correspond à Alger et ce pour éliminer la variante régionale. Le choix de la région d'Alger est justifié par le nombre

important d'habitants et par conséquent de locuteurs dans ALGASD. Bien que R_1 soit cosmopolite, rappelons que les locuteurs sont tous natifs et vivent dans la dite région. L'analyse statistique du rythme a été effectuée sur les enregistrements de 66 locuteurs. Ce nombre de participants dépasse de loin ceux habituellement utilisés dans des études similaires (maximum 10 locuteurs).

Rappelons que la base est conçue pour inclure différentes tranches de la sociétés (enseignants, docteurs, étudiants, journalistes, sans-emploi, etc.). Les enregistrements ont été partagés en groupes de locuteurs selon les trois facteurs suivants :

1. Genre : masculin (m) ou féminin (f) ;
2. Âge : jeunes locuteurs (18-30 ans) représentés par (a_1) ; locuteurs d'âge intermédiaire (30-45 ans) et enfin locuteurs âgés (+45 ans) correspondant respectivement à (a_2) et (a_3) ;
3. Instruction : 3 niveaux ont été proposés :
   - Le premier groupe, appelé catégorie moyenne (C_1), se compose des locuteurs ayant bénéficié d'un ou des palier(s) d'instruction suivant(s) : primaire, moyen et secondaire. Généralement, les 22 locuteurs de C_1 n'utilisent pas l'AS ni dans la vie quotidienne ni dans leurs occupations professionnelles.
   - Le second groupe noté par C_2 inclut 18 locuteurs ayant poursuivi leurs études afin d'obtenir des diplômes universitaires ou plus (la graduation et post-graduation). La langue utilisée dans le cursus universitaire de ces locuteurs est la langue française. De même que les professions exercées par la suite font appel plus au Français qu'à l'Arabe telles : les enseignants des filières techniques, les médecins, les journalistes de la presse française, etc.
   - La dernière catégorie (C_3) comprend 26 locuteurs. Ces derniers ont d'une part terminé leurs études universitaires ou de graduation en langue Arabe et d'autre part pratiquent l'AS dans leurs occupations professionnelles comme : les enseignants de la littérature arabe et des sciences sociales, les journalistes de presse arabe, les avocats, etc.

Le tableau suivant expose la répartition détaillée des 66 locuteurs selon les trois facteurs de variabilité cités.

	C_1		C_2		C_3		
Âge	f	h	f	h	f	h	Total
a_1	4	5	6	3	3	5	26
a_2	6	3	4	3	4	6	26
a_3	2	2	1	1	4	4	14
	12	10	11	7	11	15	
Total	22		18		26		66

TABLE 4.1: Distribution des locuteurs selon les facteurs acteurs de variabilité

4.3.2 Corpus de parole et mesures

Afin de limiter les variations dues au texte lu, nous avons effectué l'analyse sur les deux phrases du Corpus communs (Cc). Ces phrases, transcrites en IPA [IPA, 1999], sont présentées ci-dessous. Elles sont lues par les 66 locuteurs. 1254 voyelles et 1716 consonnes ont été analysées pour l'étude.

قادنا و لم يضطهدكم

qa:dana: wa-lam jadˤtˤahidkum

أخطأت فآثر صيدنا

ʔaχtˤaʔta faʔa:θara sˤajdana:

Les scores des sept paramètres du rythme ont été calculés pour chaque phrase de chaque locuteur. Les sept mesures rythmiques que nous examinons sont comme suit : 3 mesures d'intervalles (IM) à savoir $\%V$, $\triangle V$ et $\triangle C$; la forme normalisée des intervalles vocaliques et consonantiques $VarcoV$ et $VarcoC$ et enfin les Pairewise Variability Indices respectivement pour les voyelles et les consonnes $nPVI - V$ et $rPVI - C$.

4.4 CONTRASTE VOCALIQUE DE L'AS

La première étape de l'étude consiste en l'analyse acoustique des durées des voyelles (brèves v_b et longues v_L). Dans une recherche antérieure, nous avons montré que les durées des voyelles longues et courtes de l'AS produites par des locuteurs algériens étaient corrélées avec le niveau d'instruction de chacun [Droua-Hamdani et al., 2009]. Le tableau suivant présente des mesures acoustiques pertinentes basées sur la lecture des deux phrases (Cc) par les 66 locuteurs. Les données comprennent la durée moyenne des voyelles courtes et longues et les ratios des durées long/court des voyelles pour les trois groupes d'instruction cités ci-dessus.

Catégories	Durées v_b (ms)	Durées v_L (ms)	v_b/v_L
C_1	67.70	113.69	1.7
C_2	63.27	103.57	1.6
C_3	75.12	175.59	2.3

TABLE 4.2: Durées moyennes des voyelles par groupe d'instruction

La figure 4.1 fournit des informations complémentaires sur les moyennes et écarts-types des durées de voyelles longues/courtes pour les trois groupes de locuteurs selon le niveau d'instruction.

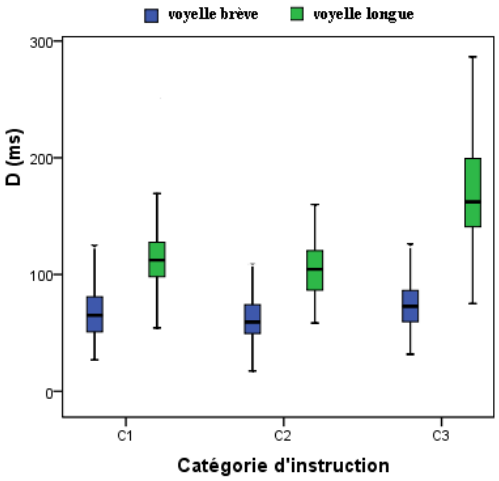


FIGURE 4.1: Boîtes à moustaches des voyelles courtes et longues selon le niveau d'instruction des locuteurs

Les résultats montrent que le contraste de durée des voyelles est plus important dans le groupe C_3, qui a le plus haut niveau d'instruction en AS. Le groupe C_2, avec un haut niveau d'instruction en Français et le groupe de C_1, avec une instruction moyenne en AS, ont des taux similaires. Ceci nous amène à dire qu'il existe une différence significative dans la façon dont les durées des voyelles longues et courtes sont réalisées en AS par les locuteurs algériens qui ont des niveaux d'instruction différents. En effet, les résultats montrent que les taux décroissent selon les *compétences* des locuteurs en AS.

4.5 VARIATIONS RYTHMIQUES DE L'AS CHEZ LES LOCUTEURS ALGERIENS

La présente étude concerne le calcul des différents paramètres rythmiques selon les trois facteurs de variabilité suivants : le niveau d'instruction, l'âge et le genre. Une analyse statistique a été appliquée pour chaque facteur.

4.5.1 Etude du rythme selon le niveau d'instruction

Cette étude consiste à calculer des paramètres afin d'analyser la variabilité rythmique dans la prononciation de l'AS tout en examinant si ces mesures sont sensibles aux différences apparentes dans les niveaux d'instruction des locuteurs.

4.5.1.1 Mesures des paramètres rythmiques

Pour infirmer ou confirmer l'influence du niveau d'instruction du locuteur sur les durées phonémiques, nous avons calculé pour l'ensemble des durées vocaliques

et consonantiques les 7 paramètres rythmiques suivants : %V, $\triangle$V, $\triangle$C, VarcoV, VarcoC, nPVI-V et le rPVI-C. La valeur moyenne de chacun d'eux est affichée dans le tableau ci-dessous.

	C_1	C_2	C_3
$\triangle V$	28.31	25.13	48.46
$\triangle C$	51.93	49.09	59.85
$VarcoV$	36.81	35.40	50.01
$VarcoC$	55.00	53.46	55.93
$\%V$	43.71	42.59	45.99
$nPVI-V$	16.95	15.85	20.60
$rPVI-C$	45.13	40.57	49.74

TABLE 4.3: Valeurs moyennes des paramètres rythmiques selon le niveau d'instruction

4.5.1.2 Analyse statistique

L'analyse statistique consiste en l'application du test de la variance ANOVA (*ANalysis Of VAriance*) à 1 facteur sur les valeurs du tableau 4.3 avec un indice de confidence $\alpha = 0.05$. Les résultats montrent que six des sept mesures rythmiques à savoir : les trois paramètres de mesure d'intervalles (IM) de temps et les deux paramètres PVI ainsi que la durée vocalique normalisée (VarcoV) - sont sensibles au facteur de l'instruction et montrent un effet significatif à ce facteur (tableau 4.4).

	F-test	P
$\triangle V$	$F(2, 129) = 54.29$	$p < 10^{-12}$
$\triangle C$	$F(2, 129) = 11.10$	$p < 10^{-4}$
$VarcoV$	$F(2, 129) = 30.44$	$p < 10^{-10}$
$VarcoC$	$F(2, 129) = 1.24$	$p = 0.29$
$\%V$	$F(2, 129) = 9.54$	$p = 0.0001$
$nPVI-V$	$F(2, 129) = 13.87$	$p < 10^{-5}$
$rPVI-C$	$F(2, 129) = 3.51$	$p = 0.032$

TABLE 4.4: Résultats du test ANOVA selon les niveaux d'instruction des locuteurs

4.5.1.3 Discussion

Nous remarquons du tableau 4.3 que les valeurs obtenues pour C_3 sont nettement plus grandes comparées à celles de C_1 et C_2 notamment en ce qui concerne les variations des durées vocaliques à savoir $\triangle V$ et $VarcoV$. Ces écarts sont clairement observés dans la figure 4.2.

L'analyse statistique conforte ces résultats car elle montre que les valeurs moyennes des quatre paramètres vocaliques significatifs sont : %V, $\triangle V$, $VarcoV$ et $nPVI-V$. Ceci confirme les mesures acoustiques du tableau 4.2.

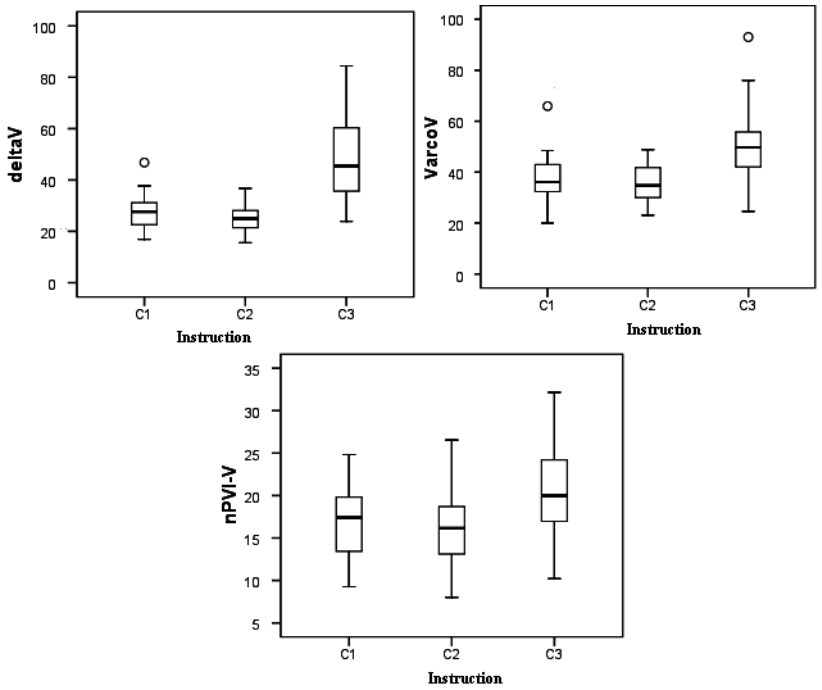


FIGURE 4.2: Représentation des différentes catégories d'instruction

Les trois paramètres affichent la même tendance : lorsque le rapport des durées vocaliques long/court augmente, les valeurs des mesures rythmiques aussi. $\triangle V$, $VarcoV$ et $nPVI-V$ sont plus sensibles aux différences entre les catégories d'instruction des locuteurs que $\%\ V$. Cette tendance suggère que l'examen des écarts-types des paramètres ΔV, $VarcoV$ et $nPVI-V$ sont de meilleurs détecteurs de contraste entre les durées des voyelles longues/courtes (entre locuteurs d'une même langue) que la proportion des intervalles vocaliques, mesurée par le paramètre $\%V$.

Les tests ANOVA montrent des différences significatives entre les groupes d'instruction pour ΔC et le $rPVI-C$ mais pas pour l'intervalle consonantiques normalisé $VarcoC$.

4.5.2 Etude du rythme selon l'âge

Nous nous intéressons, dans cette analyse à la vérification d'une éventuelle corrélation entre les scores rythmiques et un autre facteur de variabilité à savoir l'âge du locuteur.

4.5.2.1 Mesures des paramètres rythmiques

Nous avons calculé pour l'ensemble des durées vocaliques, les durées moyennes des voyelles courtes et longues et les ratios de durées long/court, puis évalué les 7 paramètres rythmiques habituels des voyelles et des consonnes. Les mesures acoustiques ainsi que les valeurs moyennes rythmiques obtenues pour les différentes classes

d'âges sont affichées dans les tableaux ci-dessous.

Catégories	Durées v_b (ms)	Durées v_L (ms)	v_b/v_L
a_1	66.65	118.19	1.77
a_2	66.38	135.27	2.03
a_3	81.78	175.70	2.14

TABLE 4.5: Durées moyennes des voyelles (courtes/longues) selon les 3 tranches d'âge

	$\triangle V$	$\triangle C$	$VarcoV$	$VarcoC$	$\%V$	$nPVI-V$	$rPVI-C$
a_1	29.77	49.97	38.30	53.49	43.91	17.18	39.73
a_2	36.18	55.05	43.66	56.11	43.94	18.21	47.22
a_3	46.99	62.26	45.46	55.70	45.99	20.00	55.41

TABLE 4.6: Valeurs moyennes des paramètres rythmiques selon l'âge

4.5.2.2 Analyse statistique

Les résultats de l'application du test ANOVA sur les 7 paramètres rythmiques en considérant que le paramètre *âge* est le facteur indépendant de l'analyse sont montrés dans le tableau 4.7. De même que pour l'instruction, les paramètres rythmiques montrent majoritairement une influence significative de la part de ce facteur.

	F-test	P
$\triangle V$	$F(2,129) = 12.11$	$p < 10^{-4}$
$\triangle C$	$F(2,129) = 10.26$	$p < 10^{-4}$
$\%V$	$F(2,129) = 2.82$	$p = 0.062$
$VarcoV$	$F(2,129) = 4.25$	$p = 0.016$
$VarcoC$	$F(2,129) = 1.90$	$p = 0.153$
$nPVI-V$	$F(2,129) = 2.90$	$p = 0.058$
$rPVI-C$	$F(2,129) = 9.31$	$p < 10^{-3}$

TABLE 4.7: Résultats du test ANOVA selon l'âge

4.5.2.3 Discussion

Les résultats montrent que les durées des v_b et v_L pour a_3 sont plus importantes que pour a_1 et a_2. Cependant le contraste de durée c'est à dire le rapport $\frac{v_L}{v_b}$ semble être gardé entre a_2 et a_3. La figure 4.3 confirme les résultats du tableau et offre des informations complémentaires sur les écarts de mesure pour chaque tranche d'âge.

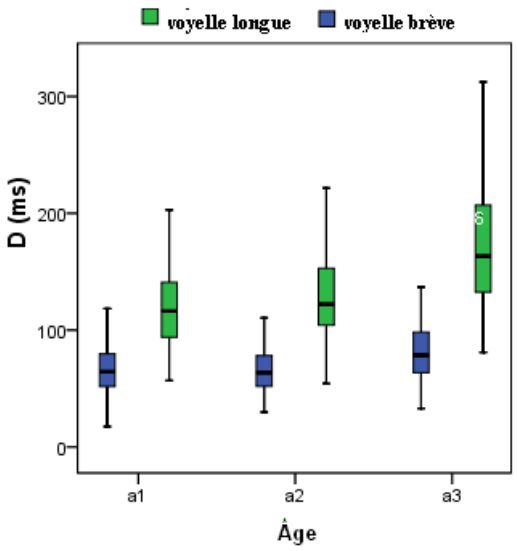


FIGURE 4.3: Comparaison des durées vocaliques selon les tranches d'âge

De même que $nPVI-V$ dépasse de peu le seuil de signification contrairement à $\%\ V$ qui est au dessus de l'indice de confidence α. Ce dernier résultat suggère que la proportion vocalique n'est pas respectée entre tranches d'âge car il existe des différences à l'intérieur de la classe elle même, ce qui est complètement plausible du moment que chaque tranche d'âge contient les trois niveaux d'instruction cités.

Les tests ANOVA montrent aussi des différences significatives pour ΔC et le $rPVI-C$ contrairement à $VarcoC$.

4.5.3 Etude du rythme selon le genre

Cette dernière analyse concerne le dernier facteur considéré, c'est à dire le genre du locuteur. L'expertise vise à montrer une éventuelle influence du genre sur les paramètres rythmiques.

4.5.3.1 Mesures des paramètres rythmiques

Nous avons dans cette expérience calculé pour l'ensemble des durées vocaliques les durées moyennes des voyelles courtes et longues et les ratios des durées long/court ainsi que les scores moyens des 7 paramètres rythmiques selon le genre du locuteur (f /h) et ce quelque soit le niveau d'instruction et l'âge du locuteur. Les moyennes obtenues sont montrées dans les tableaux suivants :

Catégories	Durées v_b(ms)	Durées v_L (ms)	v_b/v_L
f	70.86	131.46	1.85
h	68.10	141.33	2.07

TABLE 4.8: Durées moyennes selon le genre

	$\triangle V$	$\triangle C$	$VarcoV$	$VarcoC$	$\%V$	$nPVI-V$	$rPVI-C$
f	33.20	54.01	39.28	54.90	44.51	17.45	45.03
h	38.33	54.81	44.52	55.02	44.15	18.87	46.56

TABLE 4.9: Valeurs moyennes des paramètres rythmiques selon le genre

4.5.3.2 Analyse statistique

De même que pour les deux facteurs précédents, nous avons appliqué une ANOVA à un facteur sur le genre du locuteur pour déceler une éventuelle influence de ce facteur sur les paramètres rythmiques. Les résultats statistiques sont affichés dans le tableau ci-dessous.

	F-test	P
$\triangle V$	$F(1,130) = 3.499$	$p = 0.063$
$\triangle C$	$F(1,130) = 0.139$	$p = 0.709$
$\%V$	$F(1,130) = 0.263$	$p = 0.608$
$VarcoV$	$F(1,130) = 6.429$	$p = 0.012$
$VarcoC$	$F(1,130) = 0.008$	$p = 0.928$
$nPVI-V$	$F(1,130) = 0.263$	$p = 0.101$
$rPVI-C$	$F(1,130) = 0.282$	$p = 0.595$

TABLE 4.10: Résultats du test ANOVA selon le genre

4.5.3.3 Discussion

Une comparaison globale des durées des voyelles produites par des locuteurs féminins avec celles des locuteurs masculins montre, d'une manière générale, que les durées produites par les deux types de locuteurs sont voisines (figure 4.4). Le tableau 4.9 révèle que les scores rythmiques vocaliques et consonantiques sont très proches. Les grands écarts sont observés aux niveaux de $\triangle V$ et de $VarcoV$. L'analyse statistique confirme ces résultats. En effet, le seul paramètre sensible au facteur *genre* est le $VarcoV$.

4.5.4 Etude des interactions entre facteurs

Dans cette section, nous nous intéressons à l'analyse des effets combinés de plusieurs facteurs sur une même variable. En effet, nous allons commencer par des analyses à deux facteurs suivies par une autre à trois facteurs.

4.5.4.1 Interaction à deux facteurs

Dans cette expérience, nous avons procédé à l'analyse de deux séries d'interaction à 2 facteurs : *instruction* x *âge* et *instruction* x *genre*.

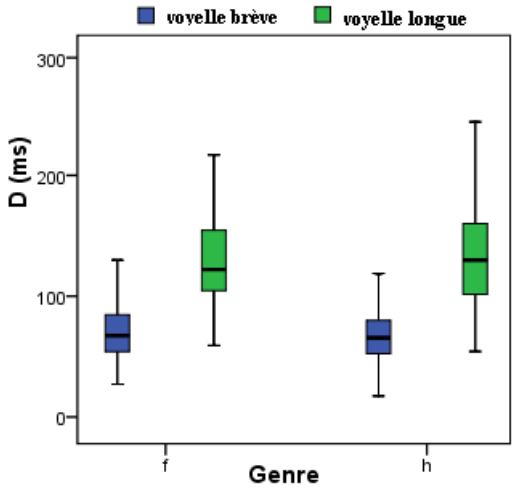


FIGURE 4.4: Comparaison des durées vocaliques (longues/courtes) selon le genre

Instruction et âge : La première combinaison étudiée concerne donc l'influence simultanée de l'instruction et de l'âge sur le rythme des locuteurs. Pour cela, nous avons procédé aux analyses rythmiques statistiques suivantes.

Paramètres rythmiques : Dans cette analyse, nous avons voulu voir s'il existait une influence simultanée de l'âge du locuteur et de son niveau d'instruction sur les paramètres rythmiques. Pour ce faire, nous avons calculé les moyennes pour chaque groupe de locuteurs en tenant compte de ces deux facteurs.

	C_1			C_2			C_3		
	a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3
$\triangle V$	28.73	27.01	30.98	23.29	27.37	25.51	37.29	50.60	58.37
$\triangle C$	48.50	53.13	58.64	47.00	47.52	64.03	54.43	62.06	63.17
$VarcoV$	38.32	35.44	36.34	33.57	39.16	30.44	43.00	54.21	52.64
$VarcoC$	53.23	56.52	55.75	53.32	53.82	52.82	53.92	57.35	56.40
$\%V$	43.75	43.70	43.62	42.84	42.83	40.63	45.15	44.95	48.23
$nPVI-V$	17.43	15.82	18.89	15.12	16.81	15.79	19.01	21.34	21.47
$rPVI-C$	37.14	47.77	61.13	37.70	40.81	52.62	44.35	51.20	53.96

TABLE 4.11: Valeurs moyennes des paramètres rythmiques selon l'âge et le niveau d'instruction

Analyse statistique des voyelles : Pour examiner les différences entre les locuteurs, nous avons réalisé des analyses de la variance à deux facteurs (two-way ANOVAs). Nous avons donc considéré les niveaux d'instruction et les tranches d'âge comme variables indépendantes. Les résultats de l'analyse statistique sont affichés dans le tableau suivant.

Paramètres		*Instruction*	*Âge*	*Instruction* x*Âge*
ΔV	*F-test*	$F(2,123) = 56.26$	$F(2,123) = 5.71$	$F(2,123) = 4.06$
	P	$p < 10^{-12}$	$p = 0.004$	$p = 0.0039$
VarcoV	*F-test*	$F(2,123) = 30.70$	$F(2,123) = 3.11$	$F(2,123) = 3.264$
	p	$p < 10^{-10}$	$p = 0.047$	$p = 0.0139$
%V	*F-test*	$F(2,123) = 10.650$	$F(2,123) = 0.054$	$F(2,123) = 1.67$
	P	$p < 10^{-4}$	$p = 0.94$	$p = 0.16$
nPVI-V	*F-test*	$F(2,123) = 10.60$	$F(2,123) = 0.91$	$F(2,123) = 1.12$
	P	$p < 10^{-4}$	$p = 0.40$	$p = 0.34$

TABLE 4.12: Résultats de l'ANOVA à 2 facteurs (âge et instruction) des paramètres rythmiques vocaliques

Analyse statistique des consonnes : De même que pour les voyelles, nous avons appliqué une ANOVA à 2 facteurs sur les paramètres rythmiques des consonnes. Les résultats sont affichés dans le tableau suivant.

Paramètres		*Instruction*	*Âge*	*Instruction* x*Âge*
ΔC	*F-test*	$F(2,123) = 5.306$	$F(2,123) = 8.72$	$F(4,123) = 1.255$
	P	$p = 0.0061$	$p < 10^{-3}$	$p = 0.291$
VarcoC	*F-test*	$F(2,123) = 1.01$	$F(2,123) = 1.42$	$F(4,123) = 0.24$
	p	$p = 0.36$	$p = 0.24$	$p = 0.91$
rPVI-C	*F-test*	$F(2,123) = 1.32$	$F(2,123) = 8.08$	$F(4,123) = 0.85$
	P	$p = 0.27$	$p < 10^{-3}$	$p = 0.49$

TABLE 4.13: Résultats de l'ANOVA à 2 facteurs (âge et instruction) des paramètres rythmiques consonantiques

Discussion Nous remarquons à partir des résultats que les scores rythmiques pour les voyelles ΔV et $VarcoV$ se distinguent nettement lorsqu'il s'agit des locuteurs de la catégorie C_3 notamment pour les locuteurs âgés de la dite catégorie. Ces résultats sont confirmés par la figure suivante.

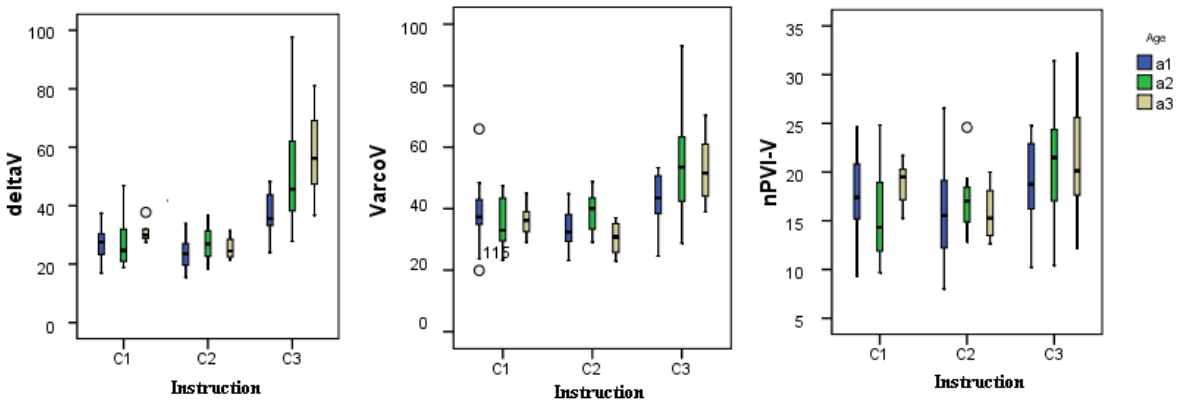


FIGURE 4.5: Interaction du nivau d'instruction avec l'âge

En ce qui concerne le $nPVI-V$, la même observation est valable mais avec un degré moindre. Par ailleurs, l'analyse statistique à 2 facteurs confirme ces résultats.

Pour ce qui concerne les paramètres rythmiques des consonnes, nous observons à partir du tableau 4.13 qu'il n'existe pas de corrélation entre l'âge et le niveau d'instruction du locuteur.

Instruction et genre : La seconde combinaison étudiée concerne l'influence simultanée du niveau d'instruction et du genre sur le rythme.

Paramètres rythmiques : Dans cette analyse, nous avons calculé les moyennes pour chaque groupe de locuteurs en tenant compte de ces deux facteurs cités (tableau 4.14).

	C_1		C_2		C_3	
	f	h	f	h	f	h
$\triangle V$	29.74	26.41	25.41	24.67	44.76	51.01
$\triangle C$	54.92	47.95	50.18	47.38	56.84	61.91
$VarcoV$	36.71	36.93	34.46	36.87	46.90	52.14
$VarcoC$	55.73	54.03	52.99	54.21	55.92	55.93
$\%V$	43.74	43.67	42.81	42.23	47.05	45.26
$nPVI-V$	17.28	16.50	14.91	17.33	20.18	20.89
$rPVI-C$	47.33	42.19	39.21	42.70	48.33	50.70

TABLE 4.14: Valeurs moyennes des paramètres rythmiques selon le genre et le niveau d'instruction

Analyse statistique des voyelles : Nous remarquons dans le tableau ci-dessous que les quatre mesures du rythme basées sur des durées vocaliques ne sont pas sensibles à l'interaction *instruction* x *genre* et que seule l'instruction est un facteur significatif.

Paramètres		*Instruction*	*Genre*	*Instruction* x*Âge*
$\triangle V$	*F-test*	$F(2,126) = 50.92$	$F(1,126) = 0.12$	$F(2,123) = 2.08$
	P	$p < 10^{-12}$	$p = 0.73$	$p = 0.12$
$VarcoV$	*F-test*	$F(2,126) = 26.98$	$F(1,126) = 2.11$	$F(2,123) = 0.73$
	p	$p < 10^{-9}$	$p = 0.15$	$p = 0.48$
$\%V$	*F-test*	$F(2,126) = 10.5$	$F(1,123) = 1.43$	$F(2,123) = 0.643$
	P	$p < 10^{-4}$	$p = 0.23$	$p = 0.52$
nPVI-V	*F-test*	$F(2,126) = 12.13$	$F(1,123) = 0.91$	$F(2,123) = 1.16$
	P	$p < 10^{-4}$	$p = 0.34$	$p = 0.31$

TABLE 4.15: Résultats de l'ANOVA à deux facteurs pour les paramètres rythmiques vocaliques

Analyse statistique des consonnes : Pour ce qui est de l'analyse statistique des consonnes, nous remarquons aussi que seule l'instruction est un facteur significatif pour $\triangle C$ et $rPVI-C$ et qu'il existe une interaction entre l'instruction et le genre pour $\triangle C$.

Paramètres		*Instruction*	*Genre*	*Instruction* x*Genre*
$\triangle C$	*F-test*	$F(2,126) = 10.82$	$F(1,126) = 0.60$	$F(2,126) = 3.49$
	P	$p < 10^{-4}$	$p = 0.43$	$p = 0.03$
VarcoC	*F-test*	$F(2,126) = 1.04$	$F(1,126) = 0.01$	$F(2,126) = 0.38$
	p	$p = 0.35$	$p = 0.90$	$p = 0.68$
rPVI-C	*F-test*	$F(2,126) = 2.96$	$F(1,126) = 0.006$	$F(2,126) = 0.84$
	P	$p = 0.05$	$p = 0.93$	$p = 0.43$

TABLE 4.16: Résultats de l'ANOVA à deux facteurs des paramètres rythmiques consonantiques

Discussion : Nous remarquons qu'il n'existe pas d'interaction entre le genre et le niveau d'instruction des locuteurs sauf pour $\triangle C$ et que seul l'instruction, comme facteur indépendant, est significatif (tableaux 4.15 et 4.16). La figure suivante expose les écarts obtenus par les paramètres vocaliques du rythme selon le genre des locuteurs.

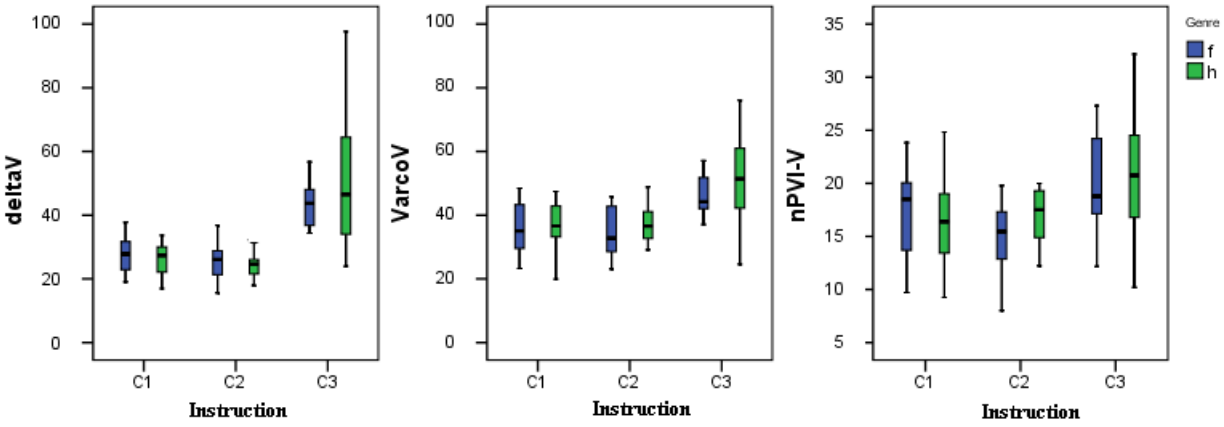


FIGURE 4.6: Interaction entre le niveau d'instruction et le genre

4.5.4.2 Interaction de trois facteurs :

Afin de déterminer si les textes lus introduisent des variations significatives sur les paramètres du rythme, nous avons réalisé une analyse ANOVA à trois facteurs avec mesures répétées sur les phrases. De même que pour les autres expériences, nous avons effectué deux séries d'analyses : l'une avec l'instruction et la tranche d'âge et l'autre avec l'instruction et le genre (tableaux 4.17 et 4.18).

Les scores des quatre mesures du rythme ont été examinés. Ils montrent qu'il existe un effet significatif de la phrase sur le $VarcoV$ et le $\%V$, ce qui est normal

car les deux phrases utilisées dans le corpus sont différentes d'où des quantités vocaliques différentes. Cependant, ce qui est intéressant à noter c'est le fait qu'il n'y ait pas d'interaction possible de la phrase avec les 2 autres facteurs à savoir l'âge et l'instruction. Ceci veut dire que les paramètres vocaliques du rythme sont toujours sensibles à l'âge, à l'instruction et à la phrase prononcée lorsqu'ils sont considérés indépendants. Ils sont aussi sensibles à l'interaction entre le niveau d'instruction et l'âge mais pas à l'influence simultanée de ces deux facteurs avec la phrase.

Facteurs	$\triangle V$	$VarcoV$	$nPVI-V$	$\%V$
E	$F(2,114)=53.64$	$F(2,114)=30.82$	$F(2,114)=10.26$	$F(2,114)=16,44$
	$p<10^{-12}$	$p<10^{-12}$	$p<10^{-6}$	$p<10^{-6}$
A	$F(2,114)=5.44$	$F(2,114)=3.12$	$F(2,114)=0.88$	$F(2,114)=0.08$
	$p=0.006$	$p=0.048$	$p=0.41$	$p=0.92$
P	$F(1,114)=0.41$	$F(1,114)=5.19$	$F(2,114)=0.97$	$F(2,114)=48.32$
	$p=0.523$	$p=0.025$	$p=0.32$	$p<10^{-6}$
ExA	$F(4,114)=3.87$	$F(4,114)=3.27$	$F(4,114)=1.09$	$F(2,114)=2.58$
	$p=0.005$	$p=0.014$	$p=0.36$	$p=0.04$
ExP	$F(2,114)=0.16$	$F(2,114)=0.09$	$F(2,114)=0.129$	$F(2,114)=0.21$
	$p=0.84$	$p=0.90$	$p=0.87$	$p=0.11$
AxP	$F(2,114)=0.42$	$F(2,114)=0.69$	$F(2,114)=0.677$	$F(2,114)=0.09$
	$p=0.65$	$p=0.50$	$p=0.51$	$p=0.91$
ExAxP	$F(4,114)=0.45$	$F(4,114)=0.37$	$F(4,114)=0.307$	$F(2,114)=0.45$
	$p=0.76$	$p=0.82$	$p=0.87$	$p=0.76$

E: instruction A: âge P: phrase

TABLE 4.17: Résultats du test MANOVA pour les paramètres rythmiques vocaliques : instruction, âge et phrase

Facteurs	$\triangle V$	$VarcoV$	$nPVI-V$	$\%V$
E	$F(2,120)=48.89$	$F(2,120)=27.01$	$F(2,120)=11,83$	$F(2,120)=16.12$
	$p<10^{-5}$	$p<10^{-6}$	$p<10^{-6}$	$p<10^{-6}$
G	$F(1,120)=0.11$	$F(1,120)=2.12$	$F(1,120)=0.88$	$F(2,120)=2.19$
	$p=0.73$	$p=0.14$	$p=0.348$	$p=0.14$
P	$F(1,120)=0.13$	$F(1,120)=5.35$	$F(2,120)=1.51$	$F(2,120)=56.38$
	$p=0.71$	$p=0.02$	$p=0.32$	$p<10^{-6}$
ExG	$F(2,120)=2.00$	$F(2,120)=0.73$	$F(2,120)=1.13$	$F(2,120)=0.98$
	$p=0.14$	$p=0.48$	$p=0.36$	$p=0.37$
ExP	$F(2,120)=0.27$	$F(2,120)=0.29$	$F(2,120)=0.34$	$F(2,120)=1.81$
	$p=0.759$	$p=0.74$	$p=0.71$	$p=0.16$
GxP	$F(1,120)=0.90$	$F(1,120)=0.11$	$F(1,120)=0.17$	$F(2,120)=1.77$
	$p=0.65$	$p=0.73$	$p=0.68$	$p=0.18$
ExGxP	$F(2,120)=0.82$	$F(2,120)=0.12$	$F(2,120)=0.45$	$F(2,120)=0.60$
	$p=0.76$	$p=0.88$	$p=0.63$	$p=0.54$

E: instruction G: genre P: phrase

TABLE 4.18: Résultats du test MANOVA pour les paramètres rythmiques vocaliques : instruction, genre et phrase

Les résultats du tableau 4.18 confirment ceux établis précédemment sur l'instruction et le genre en tant que facteurs indépendants (voir sections 4.7.1 et 4.7.3). De plus, les mesures montrent un effet significatif de la phrase sur le $VarcoV$ et le $\%V$.

Pour ce qui est des interactions à 3 facteurs:*instruction* x *genre* x *phrase*, il n' y a pas d'interaction. Les analyses statistiques sur les consonnes des deux séries : *instruction* x *âge* x *phrase* et *instruction* x *genre* x *phrase* sont données dans les tableaux suivants.

Facteurs	$\triangle C$	$VarcoC$	$rPVI-C$
E	$F(2,114)=5.38$	$F(2,114)=1.32$	$F(2,114)=1.85$
	$p=0.006$	$p=0.29$	$p=0.16$
A	$F(2,114)=8.86$	$F(2,114)=1.72$	$F(2,114)=11.30$
	$p<10^{-4}$	$p=0.18$	$p<10^{-5}$
P	$F(1,114)=5.73$	$F(1,114)=20.42$	$F(2,114)=43.99$
	$p=0.18$	$p<10^{-5}$	$p<10^{-5}$
ExA	$F(4,114)=1.27$	$F(4,114)=0.30$	$F(4,114)=1.19$
	$p=0.28$	$p=0.87$	$p=0.31$
ExP	$F(2,114)=1.11$	$F(2,114)=2.09$	$F(2,114)=0.23$
	$p=0.33$	$p=0.13$	$p=0.78$
AxP	$F(2,114)=0.11$	$F(2,114)=1.51$	$F(2,114)=0.47$
	$p=0.89$	$p=0.22$	$p=0.62$
ExAxP	$F(4,114)=019$	$F(4,114)=0.21$	$F(4,114)=0.38$
	$p=0.94$	$p=0.92$	$p=0.81$

E: instruction A: âge P: phrase

TABLE 4.19: Résultats du test MANOVA pour les paramètres rythmiques consonantiques : niveau d'instruction, âge et phrase

Facteurs	$\triangle C$	$VarcoC$	$rPVI-C$
E	$F(2,120)=11,35$	$F(2,120)=1.27$	$F(2,120)=4.00$
	$p<10^{-5}$	$p=0.28$	$p=0.021$
G	$F(1,120)=0.63$	$F(1,120)=0.01$	$F(1,120)=0.009$
	$p=0.42$	$p=0.89$	$p=0.92$
P	$F(1,120)=8.09$	$F(1,120)=23.09$	$F(1,120)=46.78$
	$p=0.05$	$p<10^{-5}$	$p<10^{-5}$
ExG	$F(2,120)=0.96$	$F(2,120)=0.46$	$F(2,120)=1.142$
	$p=0.02$	$p=0.63$	$p=0.32$
ExP	$F(2,120)=0.73$	$F(2,120)=2.89$	$F(2,120)=0.17$
	$p=0.48$	$p=0.05$	$p=0.83$
GxP	$F(1,120)=1.69$	$F(1,120)=0.83$	$F(1,120)=0.06$
	$p=0.19$	$p=0.36$	$p=0.79$
ExGxP	$F(2,120)=0.65$	$F(2,120)=0.96$	$F(2,120)=0.93$
	$p=0.52$	$p=0.38$	$p=0.39$

E:instruction G:genre P:phrase

TABLE 4.20: Résultats du test MANOVA pour les paramètres rythmiques consonantiques : niveau d'instruction, genre et phrase

Les résultats des deux tableaux ci-dessus montrent que les paramètres du rythme des consonnes ne sont pas sensibles aux interactions entre les facteurs de variabilité et les phrases.

4.6 RYTHME DES VOYELLES A TRAVERS LES ÂGES ET LES NIVEAUX D'INSTRUCTION DES LOCUTEURS

Pour illustrer l'importance de notre principale conclusion, nous discutons brièvement les résultats de ΔV (comme indiqué dans les tableaux 4.4 et 4.5). La figure 4.7 représente en moyenne ce paramètre pour les 3 catégories d'instruction à travers les groupes d'âge. Elle montre que les locuteurs dans le groupe de C_3 ont un plus grand score pour le ΔV contrairement aux locuteurs des groupes C_1 ou C_2. Cet effet est qualifié par une interaction significative entre les antécédents d'instruction et l'âge. La figure montre aussi que ΔV augmente dans tous les groupes d'âge parmi les locuteurs C_3 : les locuteurs plus âgés présentent un plus grand ΔV que leurs homologues moins âgés. Cependant, nous observons aussi une légère hausse de la valeur du paramètre parmi les groupes d'âge des locuteurs de C_1 et de C_2. Cet effet d'interaction est peut être dû justement à la différence d'âge des locuteurs. Ainsi, les locuteurs du groupe C_3 (les plus instruits et âgés) augmentent leur ΔV en augmentant le contraste des durées entre voyelles longues et courtes. Cet important constat peut être infirmé ou confirmé par des recherches futures qui fourniront des éléments de réponse supplémentaires à travers des explications socio-phonétiques.

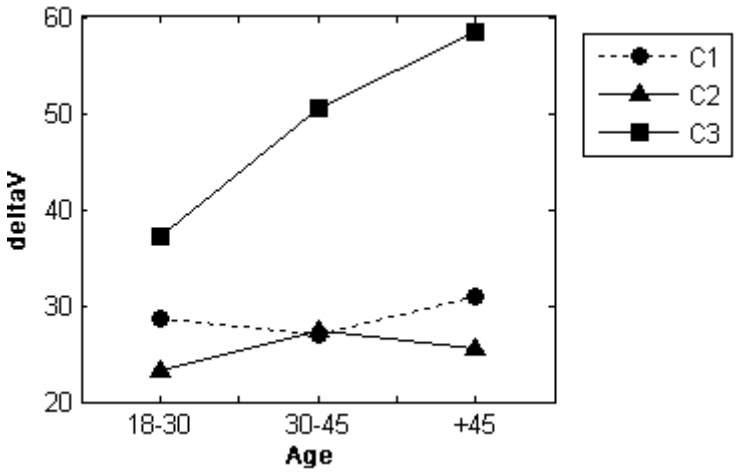


FIGURE 4.7: Evolution de $\triangle V$ selon les groupes d'instruction et d'âge

4.7 RYTHME DE L'AS ALGERIEN PARMI LES LANGUES ET LES DIALECTES ARABES

La deuxième partie du chapitre concerne l'étude des variations rythmiques indépendamment des textes. Autrement dit, nous comparons les corrélats rythmiques de l'AS algérien avec ceux obtenus par les études expérimentales pour les dialectes arabes et autres langues.

Les études sur le rythme, ne s'attardent généralement pas sur la qualité du phonème prononcé qu'elle qu'en soit sa nature : fricative, occlusive ou autre. Ceci n'est pas l'objet de l'étude. En effet, l'intérêt escompté est la comparaison de l'organisation temporelle dans la parole entre les langues/dialectes du monde au-delà de leurs systèmes phonétiques qui sont le plus souvent très différents [Ramus et al., 1999, Grabe, 2002, Arvaniti, 2009, Baltazani, 2007, Jacewicz et al., 2009]. Ce qui intéresse ce type d'analyse est de voir si les variétés considérées peuvent être discriminées sur la base de critères exclusivement rythmiques.

Les paramètres rythmiques ont été calculés à partir de la totalité des durées vocaliques et consonantiques extraites du corpus d'analyse présenté dans la méthodologie, et ce, pour toutes catégories confondues de locuteurs. Bien que la langue arabe soit traditionnellement située parmi les langues accentuelles, en entamant cette étude, plusieurs interrogations se sont posées. Mais qu'en est t-il de l'AS prononcé par des locuteurs algériens ? Les paramètres rythmiques lui confèrent t-il systématiquement une place parmi ces langues accentuelles ? Quelle est l'ampleur de l'influence de la variante régionale algérienne sur le patron rythmique ? En réponse à ces questions, des comparaisons entre les paramètres rythmiques ont été effectuées aux niveaux inter langues. Les investigations ont été poussées davantage pour déceler des différences ou similitudes notables entre l'AS algérien et les parlers arabes (dialectes étudiés par [Hamdi et al., 2004]).

4.7.1 Rythme de l'AS algérien parmi les langues du monde

La première étude concerne donc la localisation de l'AS des Algériens par rapport à la typologie rythmique des langues : stress-timed (Anglais, Allemand, Néerlandais, etc.) ; syllable-timed (Français, Italien, Espagnol, etc.) et mora-timed (Japonais). Les langues analysées dans cette section sont généralement celles qui ont été les plus soumises à des études expérimentales. Aussi, les scores rythmiques de l'AS (les mesures d'intervalles (IM), les PVI et IM normalisés) ont été tour à tour comparés aux différentes catégories de rythme.

4.7.1.1 Analyse des IM

Bien que les paramètres rythmiques IM soient au nombre de 3: $\triangle V$, $\triangle C$ et %V, nous n'allons dans notre cas étudier que $\triangle C$ et %V conformément aux résultats de [Ramus et al., 1999] (chapitre 3 section 3.5.1). L'analyse des paramètres IM de l'AS Algérien montre que la proportion notée des intervalles vocaliques dans la phrase est inférieure à 50% de la durée totale, comme il est généralement constaté dans le cas des langues accentuelles (tableau 4.21).

La comparaison des résultats montre que la proportion de %V de l'AS algérien apparaît plus importante que celle obtenue par [Ramus et al., 1999] pour les langues accentuelles notamment le Néerlandais et l'Anglais. De même, la valeur de ΔC obtenue pour les locuteurs algériens est aussi supérieure aux scores relevés dans la dite étude.

Langues	$\triangle C$	$\%V$
AS Algérien	54.39	44.33
Anglais	53.50	40.10
Polonais	51.40	41.00
Néerlandais	53.30	42.30
Français	43.90	43.60
Espagnol	47.40	43.80
Italien	48.10	45.20
Catalan	45.20	45.60
Japonais	35.60	53.10

TABLE 4.21: Comparaison de (ΔC, %V) de l'AS Algérien avec les langues du monde

La figure 4.8 illustre la projection plane de (ΔC , %V) de l'Arabe avec les différentes langues accentuelles ; syllabiques et moraiques [Ramus et al., 1999, White and Mattys, 2007]. Par ailleurs, en comparant le %V de l'AS algérien avec ceux des langues syllabiques calculés par les mêmes auteurs, nous constatons que le résultat est proche de celui obtenu pour le Français et ce dans les deux études.

4.7.1.2 Analyse des PVI

De même que pour l'analyse des IM, nous avons tout d'abord appliqué les formules des PVI sur les mesures de durées de l'AS algérien et comparé ensuite les résultats avec ceux de [Grabe, 2002, White and Mattys, 2007]. Le tableau 4.22 montre les valeurs obtenues pour l'Arabe algérien avec les résultats obtenus pour les autres langues.

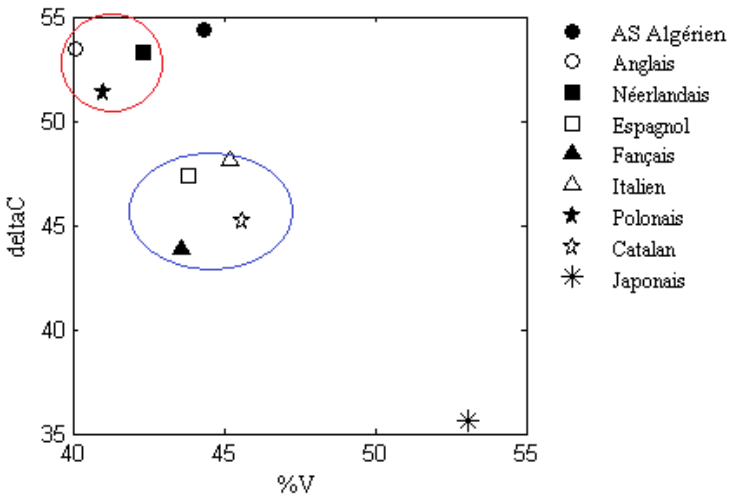


FIGURE 4.8: Comparaison des projections planes (ΔC , %V) de l'AS Algérien avec les langues du monde

Langues	nPVI	rPVI
AS Algérien	18.14	45.76
Anglais	57.20	64.10
Polonais	46.60	79.10
Néerlandais	57.40	65.50
Français	43.50	50.40
Espagnol	29.70	57.70
Thai	56.50	65.80
Catalan	44.60	67.80
Japonais	40.90	62.50

TABLE 4.22: Comparaison de (nPVI, rPVI) de l'AS algérien avec les langues du monde

La figure 4.9 illustre la position des PVI algériens par rapport à ceux trouvés par Grabe. Le graphe montre un détachement des locuteurs algériens de ceux des groupes des langues accentuelles notamment l'Anglais, le Néerlandais et le Thai.

4.7.1.3 Analyse des IM normalisés

La dernière expérience réalisée concerne les scores normalisés des durées consonantiques et vocaliques. Les valeurs rythmiques obtenues ont été comparées aux résultats de [White and Mattys, 2007] (tableau 4.23). Nous constatons, à partir de ces résultats, que le VarcoV algérien avoisine les mesures obtenues pour les langues syllabiques, notamment pour l'Espagnol. Ceci va à l'encontre de l'hypothèse émise concernant le fait que l'Arabe, en général, est une langue strictement accentuelle.

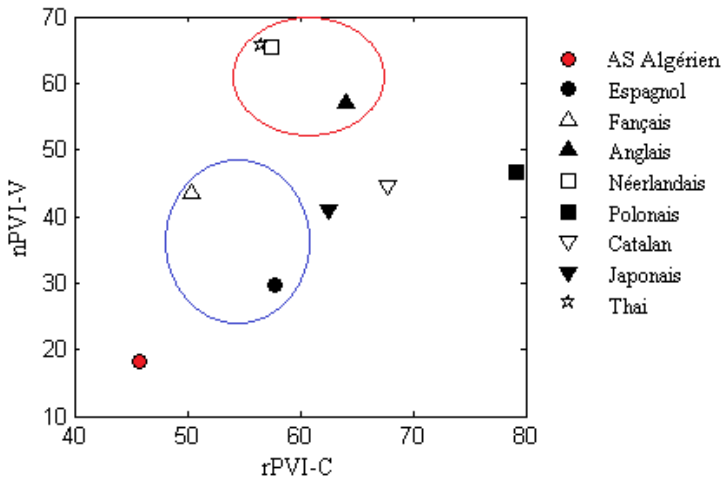


FIGURE 4.9: Projection plane des paramètres PVI des langues

Langues	VarcoV	VarcoC
AS Algérien	41.82	54.95
Espagnol	41	46
Français	50	44
Englais	64	47
Allemand	65	44

Table 4.23: Comparaison de (VarcoV, VarcoC) de l'AS Algérien avec les langues du monde

La figure suivante expose la projection plane des paramètres normalisés des durées VarcoV et VarcoC parmi les langues accentuelles et syllabiques.

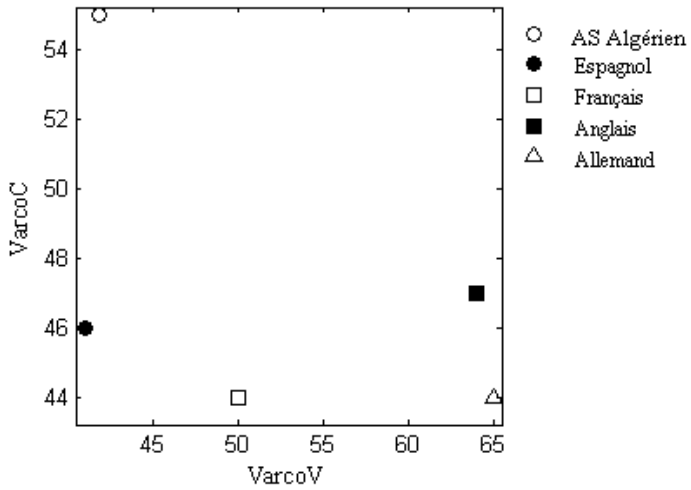


FIGURE 4.10: Projection plane des paramètres de mesures normalisée (VarcoV, VarcoC)

4.7.2 Situation de l'AS algérien parmi les dialectes arabes

Après avoir situé l'AS algérien parmi les langues du monde, notamment, les langues accentuelles, nous avons voulu par la présente expérience le situer parmi les dialectes arabes. Bien que les corpora d'analyse soient différents, nous pouvons, par cette étude, donner une vision globale du patron rythmique de l'AS et de son homologue dialectal. Pour ce faire, nous avons comparé et analysé les PVI et IM mesurés de l'AS algérien avec les mêmes paramètres obtenus pour les 6 dialectes arabes présentés dans l'étude de [Hamdi et al., 2004] : 3 maghrébins (Marocain, Algérien et Tunisien) et 3 du Moyen-orient (Egyptien, Libanais et Jordanien).

Le tableau suivant montre les paramètres rythmiques de l'AS ainsi que ceux de tous les dialectes considérés dans cette étude .

Langues	$\%V$	$\triangle V$	$\triangle C$	$nPVI$	$rPVI$
AS Algérien	44.33	35.68	54.39	18.14	45.76
D. Marocain	33.14	30.74	72.68	46.50	79.89
D. Algérien	33.10	32.41	68.10	46.08	78.73
D. Tunisien	35.42	28.64	56.85	44.41	63.74
D. Egyptien	37.41	31.53	53.67	45.53	57.37
D. Libanais	41.63	40.28	54.55	47.05	61.02
D. Jordanien	40.88	37.84	54.54	46.68	59.29

D.: dialecte

TABLE 4.24: Comparaison des paramètres rythmiques de l'AS algérien avec les dialectes arabes

Nous remarquons du tableau qu'il existe une réduction vocalique (%V) importante entre les dialectes maghrébins et l'AS algérien. Cette réduction est d'autant plus prononcée lorsqu'il s'agit du dialecte algérien. En revanche, cette réduction s'atténue dans les dialectes du Moyen-orient. A l'opposé, nous remarquons que $\triangle C$ est nettement plus important pour les dialectes maghrébins -notamment pour le Marocain et l'Algérien- comparé aux dialectes du Moyen-orient. Cette observation est valable aussi pour l'AS algérien qui se trouve plus proche des dialectes du moyen-orient que des dialectes maghrébins. La figure 4.11 illustre la projection des deux paramètres rythmiques $\triangle C$ et %V pour l'AS et l'ensemble des parlers: nous remarquons clairement, d'une part, cette réduction vocalique au niveau du dialecte algérien comparé à l'AS algérien et d'autre part, l'augmentation des proportions consonantiques.

Toujours dans le même cadre d'étude, nous avons cette fois-ci comparé les PVI de chaque dialecte avec ceux de l'AS algérien.

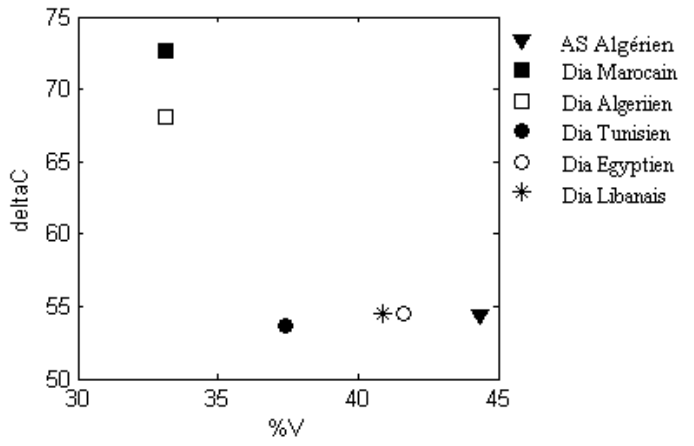


FIGURE 4.11: Comparaison de l'AS algérien avec les dialectes arabes : projection plane de (ΔC , %V) (Dia : dialecte)

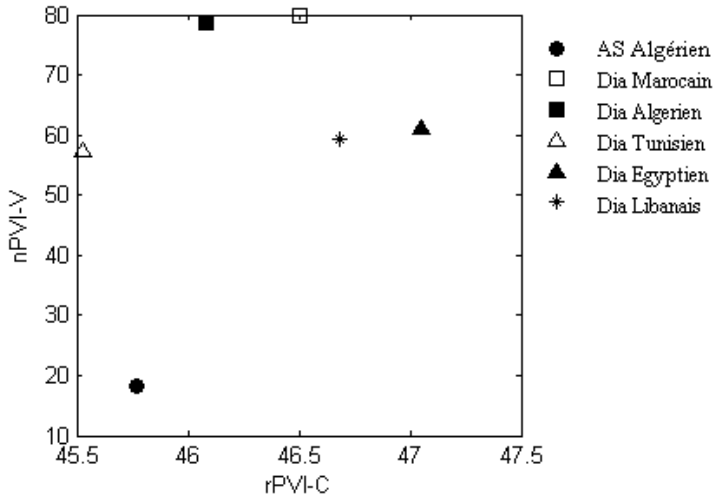


FIGURE 4.12: Comparaison de l'AS algérien avec les dialectes arabes : projection plane de (nPVI, rPVI) (Dia : dialecte)

4.8 DISCUSSION SUR LA LOCALISATION DE L'AS ALGERIEN

Bien que l'Arabe soit considéré comme une langue accentuelle, les analyses réalisées avec les différentes approches montrent que ceci n'est pas strictement établi lorsqu'il s'agît de l'AS prononcé avec l'accent algérien. En effet, les paramètres IM, c'est à dire ΔC et *%V,* affichent simultanément une tendance accentuelle en ce qui concerne le ΔC et un retrait net vers les langues syllabiques lorsqu'il s'agit du *%V*. Les mêmes résultats ont été obtenus lors de l'application des paramètres normalisés, c'est à dire *VarcoC* et *VarcoV*. Cependant, l'approche basée sur les PVI

présente des valeurs faibles à la fois pour les $rPVI-C$ et $nPVI-V$. Aussi, elle montre l'écartement de l'AS par rapport aux deux catégories rythmiques principales (accentuelles et syllabiques). De là, nous pouvons conclure que les locuteurs algériens, précisément les Algérois présentent, lorsqu'il s'agit du rythme de la parole, une particularité dans la prononciation qui peut être due aux différentes sources de variabilité. En effet, certains de ces locuteurs sont plus francophones qu'arabophones. Le manque de pratique quotidienne de l'AS et l'utilisation d'une seconde langue, en l'occurrence, le Français, pourraient expliquer cette hypothèse.

4.9 CONCLUSION

Nous avons procédé, dans ce chapitre, à l'étude d'un paramètre prosodique peu étudié pour l'AS et qui est le rythme. Les données utilisées sont extraites de la base de données ALGASD réalisées à partir d'enregistrements de 66 locuteurs. Le corpus est constitué des deux phrases communes de la base. Notre étude avait pour objectif de montrer l'influence des facteurs de variabilité sur la prononciation de l'AS à partir d'analyses acoustiques et statistiques sur les paramètres rythmiques.

Les résultats nous ont permis de noter des différences significatives entre des groupes de locuteurs, selon les tranches d'âge et le niveau d'instruction. Ce résultat est utile pour les travaux qui examinent les différences sociales dans la parole. Des études récentes, traitant les débits d'élocution, notent des différences entre les groupes sociaux en Anglais Américain [Jacewicz et al., 2009] et en Néerlandais [Verhoeven et al., 2004]. [Wiget et al., 2010] montrent que les métriques du rythme sont *sensibles* aux variations extérieures provenant de sources multiples.

Les résultats de notre étude démontrent que les paramètres rythmiques de l'AS, prononcé par les Algériens, sont dépendants des groupes de locuteurs choisis car ils constituent eux-mêmes d'importantes sources de variabilité. En effet, le rythme de l'AS varie d'un locuteur à un autre qu'il soit jeune ou moins jeune, qu'il soit instruit ou pas.

En plus de l'analyse du rythme au sein d'une même langue, nous avons étudié le rythme de l'AS au niveau inter langue en le comparant avec les différentes langues du monde, notamment, les accentuelles puis les dialectes arabes. Les résultats montrent que l'AS des Algériens présente quelques particularités. En effet, nous avons observé des réductions vocaliques voisines des mesures relevées pour les langues syllabiques et des proportions consonantiques approchant celles des langues accentuelles. Ceci nous amène à suggérer que l'AS parlé par les Algériens n'occupe pas systématiquement une place parmi les langues accentuelles. Il se localise dans une position intermédiaire entre les syllabiques et les accentuelles.

Nous pouvons donc conclure que ces études sur le rythme à deux niveaux inter

et intra langues, introduisant différents facteurs de varaibilité en l'occurrence les niveaux d'instruction des locuteurs, sont des travaux inédits pour les langues du monde et pour la langue Arabe.

Chapitre 5

UTILISATION D'ALGASD POUR LA VALIDATION D'UN SRAP

5.1 INTRODUCTION

Ce chapitre concerne la validation de la base de données sonores ALGASD par un système de reconnaissance de la parole. En comparaison avec d'autres langues, la langue arabe dispose d'un nombre limité d'études et d'outils dans le domaine de la reconnaissance vocale. Les recherches ont été menées principalement sur la reconnaissance des digits, des voyelles isolées et sur la reconnaissance des mots isolés [Elshafei et al., 2008, Satori et al., 2009, Alotaibi et al., 2010, Alotaibi, 2008]. Toutefois, les travaux concernant le développement d'un système de reconnaissance automatique de parole continue de l'Arabe ont été abordés récemment par un nombre restreint de chercheurs, que ce soit pour les systèmes dépendant ou indépendant du locuteur utilisant différentes plates-formes et diverses techniques: Sphinx; Julus; HTKtoolkit; Réseaux de Neurones; les HMM [Alkanhal et al., 2007, Alotaibi et al., 2008, 2010]. Dans cette section, nous décrivons la conception du système de reconnaissance automatique de parole continue de l'AS. Ce système, fondé sur le principe des HMM, est réalisé en utilisant le toolkit HTK.

5.2 HTK EN BREF

Développée par l'université de Cambridge pour les deux environnements Windows et Linux, la plate-forme HTK (Hidden Markov Model Toolkit) est un ensemble de bibliothèques et d'outils destinés à la construction des systèmes de reconnaissance automatique de la parole [Young et al., 2006]. Programmés en langage C, les outils de HTK sont basés sur les HMM. La topologie des modèles peut être configurée par l'utilisateur et servir à modéliser des mots ou des phonèmes. Les densités de pro-

babilités d'émission utilisées pour chaque état peuvent être discrètes ou continues (GMM). Tandis que l'initialisation des modèles se fait avec l'algorithme de Viterbi, la réestimation fait appel à l'algorithme de Baum-Welch. La reconnaissance, quant à elle, est assurée par l'application, une seconde fois, de l'algorithme de Viterbi en corrélation avec un réseau syntaxique précis [Rabiner and Juang, 1993]. HTK est structuré en 17 librairies utilisées par 18 outils de base. La communication entre ces outils se fait par la gestion de plusieurs types de fichiers : signaux, étiquettes, modèles acoustiques, modèles de langage, définition de réseau syntaxique. De par sa modularité, le toolkit offre la possibilité d'utiliser directement les outils disponibles dans sa bibliothèque ou d'intégrer les programmes de l'utilisateur. La figure 5.1 illustre les différentes étapes du traitement ainsi que les principales commandes du système.

HTK permet de réaliser deux types de configuration des HMM : des systèmes monophones qui dépendent de l'unité traitée à savoir le phonème, où bien des systèmes plus élaborés qui vont au delà d'une seule unité appelés les modèles triphones. Cela dit, cette dernière configuration est entièrement basée sur la première car les transcriptions monophoniques sont converties en transcriptions triphoniques et l'ensemble des modèles triphones sont créés par clonage à partir des modèles monophones avec des réestimations supplémentaires [Young et al., 2006]. En plus de HTK toolkit, il existe d'autres outils de développement utilisés pour la conception des SRAP. Des détails concernant ces toolkits sont donnés en annexe I.

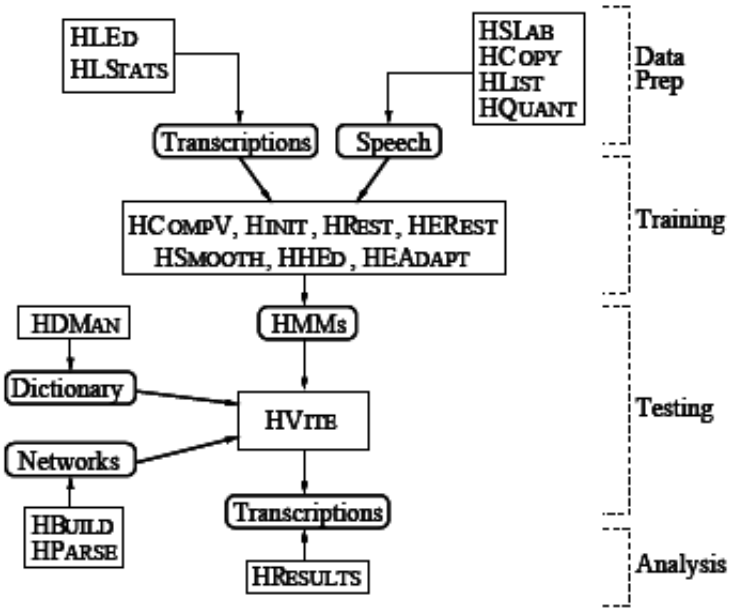


FIGURE 5.1: Description de HTK toolkit [Young et al., 2006]

5.3 DESCRIPTION DU SRAPC

Nous avons procédé à la réalisation d'un système de reconnaissance automatique de la parole continue basé sur les HMM. Pour concevoir ce dernier, nous avons installé la version HTK 3.3 Toolkit sous Windows XP. Notre choix s'est porté sur les modèles de Markov car ils ont prouvé leur efficacité en affichant des taux de reconnaissance élevés et ce dans des expériences similaires pour d'autres langues.

Le SRAP que nous avons souhaité mettre en place doit satisfaire certaines conditions qui sont :

- parole continue ;
- indépendance des locuteurs (multi locuteurs)
- grand vocabulaire
- dédié à la reconnaissance de la langue arabe

Chacun de ces objectifs constitue lui-même un challenge à relever. Par conséquent, réaliser un système incluant autant d'opérations complexes nécessite : l'écriture de plusieurs scripts, la manipulation et la gestion d'un nombre important de lignes de commandes ainsi qu'une multitude de fichiers sonores, de données textuelles et binaires.

5.3.1 Corpus Apprentissage et Test du SRAP

La construction d'un SRAP à grand vocabulaire exige à ce que tous les modèles acoustiques soient suffisamment entraînés durant le processus d'apprentissage. L'objectif est d'introduire dans cette phase les différentes variations inter et intra locuteurs qui peuvent se produire dans la parole ou lors de sa production telles : la coarticulation qui se traduit par l'influence du phonème adjacent, la variation du débit d'élocution, le genre, les différents accents régionaux des locuteurs, l'état émotionnel du locuteur, etc. La variabilité contextuelle nous contraint à considérer tous les phonèmes de la langue dans le maximum de contextes phonétiques possible. En pratique, cela signifie qu'en fonction de la qualité des échantillons et de la complexité des modèles acoustiques, des centaines de phrases sont nécessaires.

Les corpora utilisés dans notre cas sont tous extraits de la base ALGASD que ce soit pour les données textuelles ou sonores. Rappelons que cette dernière a été réalisée à partir de phrases phonétiquement équilibrées qui contiennent donc tous les phonèmes de l'Arabe. De plus, plusieurs paramètres de variabilité liés aux : genre, âge, niveau d'instruction, accent régional des locuteurs ont été considérés.

La plate-forme du SRAP a été conçue et testée sur 6 régions des 11 que compte ALGASD. Nous avons ciblé principalement par cette limitation la préparation d'un système multilocuteurs opérationnel. Ce système adaptable peut être par la suite développé pour contenir les régions restantes de la base. Par conséquent, les régions

étudiées sont : Alger (R_1), Tizi Ouzou (R_2), Jijel (R_5), Bechar (R_9), Ghardaia (R_{10}) et enfin El Oued (R_{11}).

Le nombre total de phrases utilisées pour le développement du SRAP s'élève à 592 enregistrements. Le corpus sonore alloué pour la phase d'apprentissage compte 434 enregistrements contre 158 pour le test soit $\approx 26\%$ du total ce qui est acceptable pour un corpus test. Cependant, il est important de souligner que les enregistrements utilisés dans le test n'apparaissent pas dans le corpus d'apprentissage. Le tableau suivant expose le nombre d'enregistrements utilisés dans le SRAP pour les deux phases apprentissage et test selon les régions.

Régions	Enregistrements	Apprentissage	Test
R1	276	198	78
R2	123	93	30
R5	57	45	12
R9	27	19	8
R10	59	45	14
R11	50	34	16
Total	592	434	158

Table 5.1: Nombre d'enregistrements utilisés dans le SRAP

5.3.2 Répartitions des locuteurs dans le SRAP

ALGASD compte un total de 300 locuteurs, cependant la limitation des régions étudiées ramène ce nombre à 167 locuteurs. Nous avons utilisé dans la réalisation du SRAP la totalité des locuteurs dont disposait ALGASD pour les 6 régions sauf pour Jijel (R5) où ce nombre est légèrement réduit (16/18). Tandis que l'apprentissage compte 119 locuteurs, le nombre réservé au test s'élève à 49 locuteurs qui n'ont pas participé à l'apprentissage. Le tableau 5.2 expose le nombre exact de participants pour chaque phase (apprentissage/ test) selon chaque région.

Régions	Apprentissage		Test		Locuteur/Région
	homme	femme	homme	femme	
R1	28	28	8	16	80
R2	11	13	5	5	34
R5	5	7	1	3	16
R9	3	2	-	2	7
R10	6	5	2	3	16
R11	6	4	1	3	14
Total	118		49		167

Table 5.2: Nombre de locuteurs du SRAP

5.4 DEVELOPPEMENT DU SRAP

Comme nous l'avons détaillé au chapitre 1, un système de reconnaissance de la parole continue nécessite dans sa globalité 3 phases de traitements à savoir : la préparation du système ou sa paramétrisation, l'apprentissage puis le test et l'analyse des résultats. Chacune de ces phases est elle-même partagée en plusieurs étapes de traitements. Les figures 5.2 et 5.3 résument les différentes étapes suivies dans l'élaboration des modèles monophones pour les deux phases d'apprentissage et de reconnaissance .

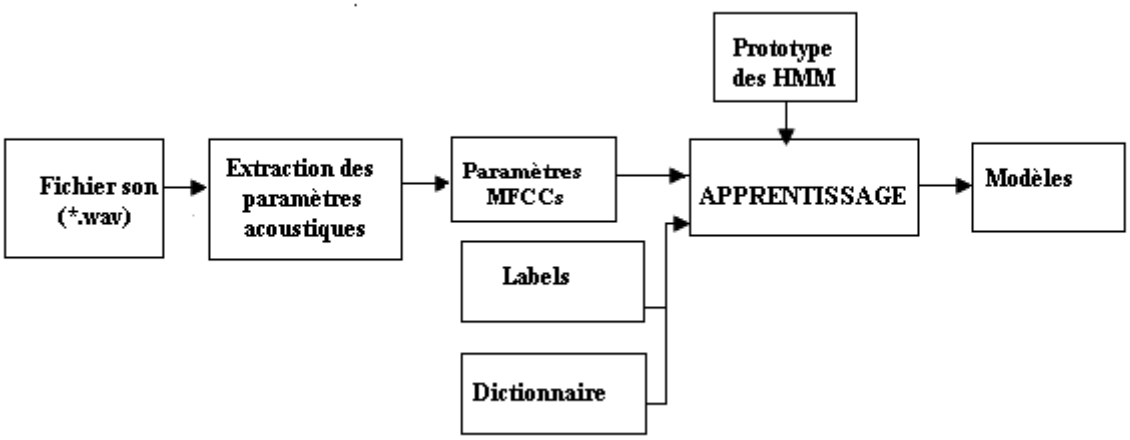


FIGURE 5.2: Phase d'Apprentissage

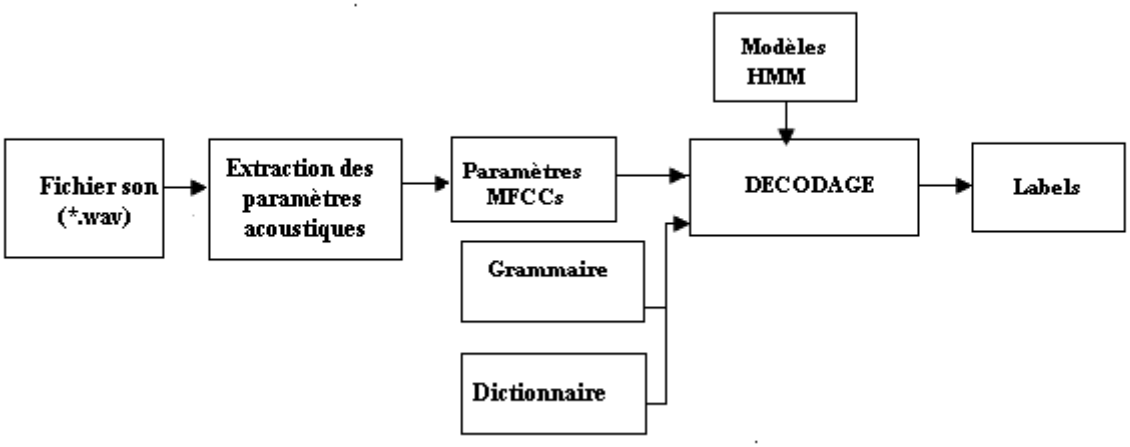


FIGURE 5.3: Phase de Reconnaissance

5.4.1 Retranscription des textes

Parallèlement aux traitements des signaux acoustiques qui seront réalisés dans les sections suivantes, un système de reconnaissance requiert un traitement parti-

culier appliqué sur les textes notamment sous le système d'exploitation Windows (Win XP). En effet, les phrases de ALGASD sont transcrites en alphabet SAMPA, mais lors de leur utilisation pour la réalisation du SRAP, nous avons rencontré certains problèmes notamment dans l'élaboration des modèles acoustiques. En effet, l'alphabet SAMPA est composé de 34 phonèmes dont certains sont désignés par la même lettre: une fois minuscule et une autre fois majuscule comme pour la plosive sourde (ت) représentée dans SAMPA par le symbole [*t*] et pour la fricative (ث) représentée à son tour par le symbole [*T*]. Or, lors de la compilation des modèles acoustiques sous Windows, ces deux phonèmes se trouvent être confondus. De même, nous avons constaté que des symboles non alphabétiques sont utilisés dans SAMPA pour désigner certains phonèmes tels la glottale [ʔ] (ء) (notée avec SAMPA par [?]), la pharyngale [ʕ] (ع) (notée en SAMPA par [?']), les voyelles longues, etc. Dans ce cas, la dénomination des modèles acoustiques est totalement rejetée par Windows. De ce fait, tous les symboles phonétiques non acceptés ont été remplacés par d'autres qui sont plus commodes à la manipulation de certaines phrases (tableau 4.3).

SAMPA	Symboles	SAMPA	Symboles
t'	T_	T	th
d'	D_	D	dh
s'	S_	S	sh
D'	DH_	Z	g
G	G_	a :	ae
X\	H_	u :	uh
?	Q_	i :	ih
?'(?\)	C		

Table 5.3: Liste de symboles SAMPA modifiés

5.4.2 Vocabulaire et dictionnaire de prononciation

Dans le cadre d'un système de reconnaissance automatique de la parole, la construction d'un lexique phonétisé nécessite tout d'abord la normalisation des textes du corpus d'apprentissage suivie de l'élaboration du vocabulaire à partir des mots du corpus. Ces étapes sont nécessaires pour le bon déroulement du processus de transcription finale (sortie textuelle) par le SRAP.

5.4.2.1 Vocabulaire

Nous entendons par normalisation des textes, l'opération qui consiste à : découper les phrases du corpus d'apprentissage en groupes de souffles (mots), à homogénéiser l'ensemble des mots c'est à dire les abréviations, etc., et à supprimer tous les

signes de ponctuation existants. Cette étape d'analyse tend à garder pour la conception du vocabulaire uniquement les entités lexicales. Dans notre corpus textuel, les phrases sont déjà séparées et ne contiennent que des groupes de souffles corrects (pas d'abréviations). Cependant, un traitement automatiquement, constitué de règles de transformation, a été élaboré pour éliminer les signes de ponctuations. Le vocabulaire de notre SRAP est construit en incluant tous les mots du corpus ALGASD. L'étendue de sa taille est de 5000 mots.

5.4.2.2 Dictionnaire de prononciation

Un réseau lexical décrit la séquence de mots qui doit être reconnue par le SRAP. Cependant, lorsque le système utilise des unités plus petites que le mot (phonème), le dictionnaire sert à décrire la séquence des HMM qui constituent chaque mot de ce réseau lexical. La construction du dictionnaire du SRAP revient donc à éditer une liste ordonnée alphabétiquement comportant tous les mots existants dans vocabulaire avec leurs transcriptions phonétiques correspondantes. Rappelons qu'un mot donné peut être prononcé de deux manières différentes. Le dictionnaire de prononciation se doit de contenir toutes les prononciations probables.

Le choix des entrées lexicales contribue largement au bon fonctionnement d'un SRAP. Tout mot absent du lexique ne peut pas être reconnu par le SRAP et peut engendrer des erreurs. Dans le domaine de la transcription graphèmes-phonèmes, deux approches sont principalement utilisées pour construire le lexique. La première repose sur l'utilisation d'un lexique phonétique de référence disponible (déjà existant) ; la seconde tend à le concevoir en utilisant soit un outil conçu à base de règles de phonétisation pour transcrire automatiquement les graphèmes en phonèmes comme le [Béchet, 2001], ou bien, le faire manuellement. Cette tâche est assez coûteuse en temps lorsque le lexique est conséquent. Lors de la création de notre dictionnaire, nous avons été confrontés à un certain nombre d'incommodités liées à la langue, au système d'exploitation et au dictionnaire de référence.

Dictionnaire de référence Le SRAP nécessite pour son élaboration un dictionnaire de prononciation de référence à partir duquel il puise les différentes prononciations probables. Ce type de document existe pour les autres langues tel le *BEEP* pour l'Anglais britannique [BEEP] alors qu'il est non disponible pour le grand public en Arabe. Dès lors, nous avons utilisé pour concevoir notre dictionnaire de phonétisation une combinaison des deux dernières approches citées dans le paragraphe précédent. En effet, nous avons tout d'abord réalisé notre propre outil de phonétisation automatique suivi de la vérification manuelle de toutes les transcriptions réalisées automatiquement.

Règles de conception du dictionnaire de phonétisation

1. Pour les langues autres que l'Arabe comme le Français, plusieurs mots peuvent être associés à une unique prononciation comme dans le pluriel (*mot/mots* → */mo/*) ou bien dans les homonymes (*maire/mer/mère* → */mɛr/*). En effet, un seul phonème peut être représenté par un seul ou plusieurs graphèmes comme pour : *f/ph* ; *c/s* ; *é/ei/et/ez/er*, etc. Aussi, le dictionnaire doit contenir toutes ces variantes graphologiques ainsi que leurs prononciations respectives. En ce qui concerne l'Arabe, la translittération entre le graphème (**حرف** */ħarf/*) et son homologue phonétique est bijective. Par conséquent, les deux constations précédentes ne sont pas rencontrées car chaque graphème a une unique prononciation mis à part le cas des semi voyelles */w/* et */y/* qui peuvent être soient des voyelles longues ou bien prononcées comme des consonnes. Par ailleurs, au niveau du mot, nous relevons certaines exceptions telles */ʔasm*a ːʔ *ʔalʔiʃ*a ːra/ (**أسماء الإشارة**) où l'écriture orthographique du mot correspond à une prononciation légèrement différente, due à un allongement de la durée vocalique (**مد**) */madd/* comme dans (**هذا**) */haða/*→/*h*a ːða/.
2. Le dictionnaire doit contenir une liste de tous les mots du corpus. Dans notre cas, les préfixes et suffixes de la langue arabe font parti du mot tels : la définition (*ʔataʕrif/* **التعريف**) ; la flexion des verbes (**صرف الأفعال**)/*ja/* et */ta/*, etc. ; et certaines particules (**حروف متصلة**) */ħur*u ːfs *ʔalmutasʕila/* commes/*bi/* et */fa/* etc.
3. Nous avons remarqué, durant la phase d'enregistrements des corpora, que certains locuteurs observaient un arrêt */waqf/* (**وقف**) lorsqu'ils lisaient une phrase se terminant par un */tanwin bi ʔalnasb/* (**تنوين بنصب**). En effet, ils remplaçaient cette propriété de la langue par un allongement de la voyelle finale */madd/* (**مد**). Par conséquent, nous avons doté le dictionnaire des deux prononciations possibles du même mot (exemple :/*musliman/* et */muslim*a ː/).

Listes des phonèmes La liste des symboles utilisés dans l'élaboration du dictionnaire et un extrait du lexique de phonétisation sont donnés respectivement dans les tableaux 5.4 et 5.5.

5.4.3 Modèles de langage

Les modèles de langage ont pour objectif de représenter les lois qui régissent le comportement de la langue. Ils sont essentiels dans un système de reconnaissance vocale car ils indiquent la probabilité à ce qu'une séquence de mots donnée soit prononcée parmi toutes celles qu'offre le corpus test. En effet, le modèle acoustique n'a pas d'*a priori* sur l'enchaînement entre les mots.

Transcription	Graphème	Transcription	Graphème
Q_	ء	r	ر
b	ب	z	ز
t	ت	s	س
th	ث	sh	ش
g	ج	S_	ص
H_	ح	D_	ض
x	خ	T_	ط
d	د	DH_	ظ
dh	ذ	C	ع
G_	غ	y	ي
f	ف	ih	الكسرة (ِ)
q	ق	ah	الفتحة (َ)
k	ك	uh	الضمة (ُ)
l	ل	iy	ـِي
m	م	ae	ـا
n	ن	uw	ـو
h	ه	♦▪♪╣	
w	و	in/un/an	ً / ٌ / ٍ

TABLE 5.4: Liste des phonèmes du dictionnaire

d'aX\aka :tiha :	D_ ah H_ ah k ae t ih h ae
d'a ?ula	D_ ah Q_ uh l ah
d'aru :ratun	D_ ah r uw r ah t uh n
Gadan	G_ ah d ah n
Gadrahum	G_ ah d r ah h uh m
Gafala	G_ ah f ah l ah
Gafat	G_ ah f ah t
Gala :	G_ ah l ae

TABLE 5.5: Extrait du dictionnaire de prononciation

Ceci revient donc à calculer la probabilité *a priori* d'une séquence de mots $W = w_1...w_n$ par la décomposition suivante :

$$P[w_1, w_2 \cdots w_N] \quad = \quad P[w_1] \prod_{i=2}^{N} P[w_i \mid w_1...w_{i-1}] \tag{5.1}$$

où la séquence $w_1, w_2 \cdots w_N$ est la séquence de mots. Ces probabilités conditionnelles sont calculées empiriquement par la technique du *maximum de vraisemblance* donnée par :

$$P[w_i \quad | \quad w_1 ... w_{i-1}] = \frac{C(w_1 ... w_i)}{\sum C(w_1 ... w_{i\text{-}1})} \tag{5.2}$$

où $C(w_1 ... w_i)$ est le nombre d'occurrences de la séquence $w_1 ... w_i$ dans le corpus d'apprentissage.

Le modèle de langage choisi pour la réalisation du réseau lexical de notre SRAP est fondé sur *les modèles bigrammes*. Basé donc sur l'hypothèse markovienne, ce modèle plutôt simple, suffit pour l'élaboration d'un réseau lexical efficace qui dépend d'un historique réduit du texte. En effet, le modèle bigramme suppose qu'un item donné, en l'occurrence un mot w_i dépend uniquement du mot qui le précède c'est-à-dire w_{i-1}.

Lissage des bigrammes Sur le plan pratique, la modélisation stochastique présente quelques insuffisances notamment lorsque les corpora d'apprentissage ne couvrent pas toutes les successions de mots possibles en particulier pour les tailles restreintes de vocabulaire. En effet, des successions de mots possibles peuvent ne pas être observées car les modèles bigrammes (n-grammes en général) tendent, dans le processus d'estimation des probabilités par le maximum de vraisemblance, à attribuer une probabilité nulle à tout n-gramme en l'occurrence bigrammes n'ayant jamais été rencontrés ou très peu fréquents dans le corpus d'apprentissage, même si cet n-gramme pourrait être parfaitement correct sur le plan linguistique. Ceci à pour conséquence directe l'absence quasi certaine de ces n-grammes (n'apparaîtront jamais) dans les transcriptions finales des systèmes de reconnaissance vocale et ce qui altérera indéniablement la performance de ces derniers. Pour éviter les probabilités nulles, et donc prévenir ces potentielles erreurs de reconnaissance, nous avons procédé à un *lissage* des probabilités afin d'uniformiser les probabilités d'apparition de tous les bigrammes. Plusieurs techniques de lissage existent [Huang et al., 2001], pour notre part, nous avons utilisé :

Techniques de Good-Turning et de repli La méthode de Good-Turing consiste à définir un facteur d'escompte (*discounting*) dont le rôle est de diminuer le nombre d'occurrences des bigrammes fréquents du corpus d'apprentissage, et par conséquent leurs probabilités d'apparition. Cette réduction en masse de probabilité est suivie par sa distribution sur les probabilités des n-grammes absents ou rares. Tous les mots du vocabulaire sont donc listés afin d'élaborer un fichier de statistiques contenant

toutes les probabilités d'apparition de chaque mot. Les probabilités de distribution sont établies en utilisant la formule suivante :

$$p(i,j) = \begin{cases} (N(i,j) - D)/N(i) & si\, N(i,j) > t \\ b(i)p(j) & ailleurs \end{cases} \tag{5.3}$$

où $N(i,j)$ est le nombre de fois que le mot j suit le mot i et $N(i)$ est le nombre de fois que le mot i apparaît. D est le facteur de réduction. t est le seuil de calcul pour inclure un bigramme donné. Si un quantificateur bigramme (dans le cas général n-grammes) a une valeur en dessous du seuil t, il est ramené à la probabilité unigramme (n-1 grammes) normalisée par un poids $\beta(i)$ afin d'assurer que toutes les probabilités bigramme pour une somme donnée soit égale à l'unité. Cette technique dite de *repli* (*back-off*) consiste donc à préférer en premier les modèles d'ordre supérieur c'est à dire les bigrammes, avant de traiter les modèles plus simples (unigramme). La probabilité unigramme est calculée en utilisant :

$$p(i,j) = \begin{cases} N(i)/N & if\ N(i) > u \\ u/N & ailleurs \end{cases} \tag{5.4}$$

où u est un seuil pour les unigram et N le nombre total de mots :

$$N = \sum_{i=1}^{L} max(N(i), u)) \tag{5.5}$$

Les poids sont estimés avec :

$$\beta(i) = \frac{1 - \sum_{j \in B} p(i,j)}{1 - \sum_{j \in B} p(i)} \tag{5.6}$$

B est l'ensemble de mots pour lesquels $p(i,j)$ ont un bigramme. $\beta(i)$ est appelé le *coefficient de repli*.

5.4.3.1 Évaluation des modèles de langage

La qualité d'un modèle de langage est jugée selon la capacité qu'à ce dernier à prédire les séquences de mots. Pour ce faire, la mesure la plus utilisée dans l'évaluation des modèles de langage est la *perplexité conditionnelle* [Huang et al., 2001]. Lorsque la perplexité est estimée sur le corpus d'apprentissage (*training-set perplexity*), elle définit si les modèles choisis modélisent correctement le corpus. Plus la valeur de la perplexité est petite, plus le modèle de langage possède des capacités de

prédiction importantes. En revanche, quant elle est calculée sur le corpus de test, elle tend à évaluer le degré de généralisation du modèle. Par ailleurs, elle n'est pas systématiquement corrélée avec la performance du SRAP. Pour des modèles n-grammes, la perplexité se calcule ainsi :

$$PP = 2^{-\frac{1}{n}} \sum_{t=1}^{n} log_2 P(w_t/h) \tag{5.7}$$

où $P(w_t/h)$ est la probabilité associée au n-gramme (w_t/h), h étant l'historique.

La perplexité du modèle de langage réalisé pour notre système, évalué à partir du corpus apprentissage, est assez faible. Elle est de l'ordre de 2.52. Ce résultat montre que la modélisation bigramme adoptée convient pour la prédiction des suites de mots improbables dans le corpus test.

5.4.4 Paramétrisation acoustique du système et codage de données

La paramétrisation du SRAP consiste en la conversion des fichiers de la base ALGASD en un format de paramètres acceptés par le système. Pour cela, nous avons tout d'abord configuré HTK de telle sorte à accepter le format *wave* qui est le format adopté dans ALGASD puis appliqué une paramétrisation acoustique (front-end) par le biais d'une technique d'analyse qui permet essentiellement d'extraire du signal de parole l'image acoustique *la plus significative possible*. Pour notre part, nous avons opté pour l'utilisation de la technique la plus communément répandue en reconnaissance vocale, vu les bons résultats qu'elle affiche à savoir celle basée sur les Coefficients Cepstraux MFCC (*Mel-Frequency Cepstral Coefficients*). Le calcul des MFCC se fait sur plusieurs étapes (figure 5.4).

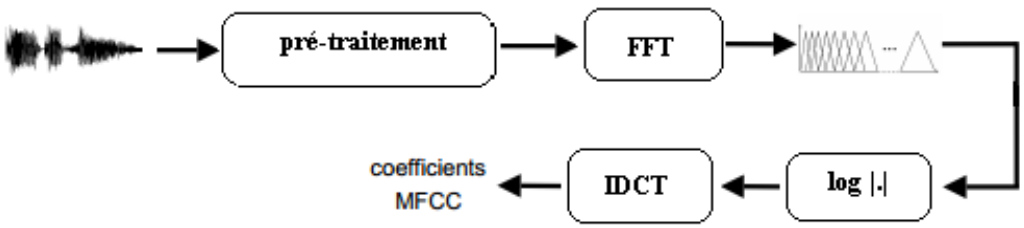


Figure 5.4: Etapes de calcul des MFCC

5.4.4.1 Prétraitement du signal

Pour calculer les MFCC, deux prétraitements du signal de parole sont nécessaires. Le premier étant le fenêtrage effectué avec l'application de la *fenêtre de Hamming* sur les trames du signal. L'objectif de cette opération est la réduction des discontinuités aux bords. La formule utilisée est donnée par:

$$w(n) = \begin{cases} 0.54 - 0.46cos(2\pi n/N - 1) & n = 0, 1, ...N - 1 \\ 0 & ailleurs \end{cases} \tag{5.8}$$

N représente la taille de la fenêtre.

Le second prétraitement, appelé *pré-emphase*, consiste dans le rehaussement des hautes fréquences du spectre grâce à l'application d'un filtre numérique du premier ordre calculé à partir de:

$$s(n) = s(n) - as(n-1) \quad 0.9 \leq a \leq 1 \tag{5.9}$$

Le résultat de la pré-emphase réalisé sur une portion du signal est illustré par la figure 5.5.

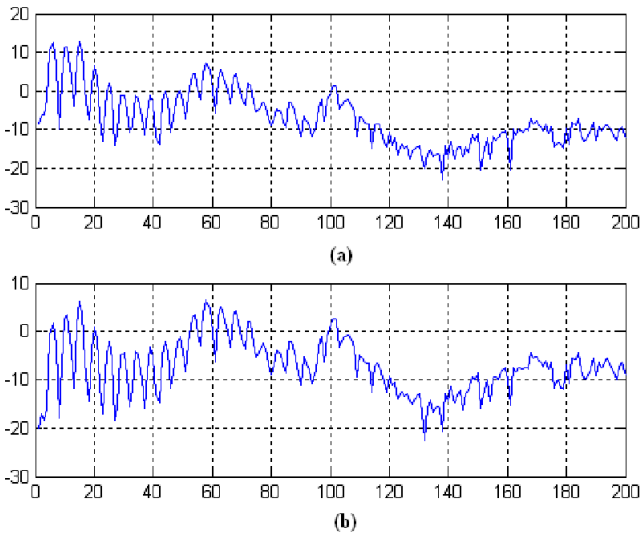


Figure 5.5: Pré-emphase du signal : (a) avant et (b) après

5.4.4.2 Calcul des MFCC(s)

Les coefficients MFCC ont été calculés à partir des énergies issues du banc de filtres m_j en utilisant la transformée en cosinus discrète dont la formule est la sui-

vante :

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j cos\left(\frac{\pi i}{N}(j-0.5)\right) \tag{5.10}$$

N est le nombre de canaux du banc de filtres. Ces filtres sont régulièrement espacés le long de l'échelle *mel* qui représente le mieux la perception auditive humaine [Huang et al., 2001] (figure 5.6).

La formule de l'échelle *mel* est donnée par:

$$f_{Mel} = 1127 log(1 + f_{Hz}/400) \tag{5.11}$$

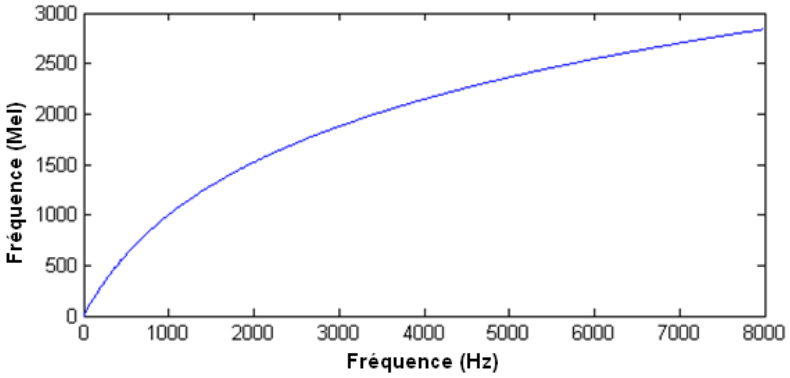


FIGURE 5.6: Echelle Mel

La forme des filtres est généralement triangulaire. La figure 5.7 illustre un banc de 24 filtres dont les fréquences centrales et les bornes sont données par:

$$T_m(k) = \begin{cases} \frac{k-k_m}{k_{1m}-k_m} & pour \;\; k_{1m} \leq k \leq k_m \\ \frac{k_m-k}{k_{2m}-k} & pour \;\; k_m \leq k \leq k_{2m} \end{cases} \tag{5.12}$$

avec T_m la fonction de transfert du filtre, k la variable de fréquence, k_m la fréquence centrale du filtre et la paire (k_{1m}, k_{2m}) représentent les bornes inférieure et supérieure du filtre.

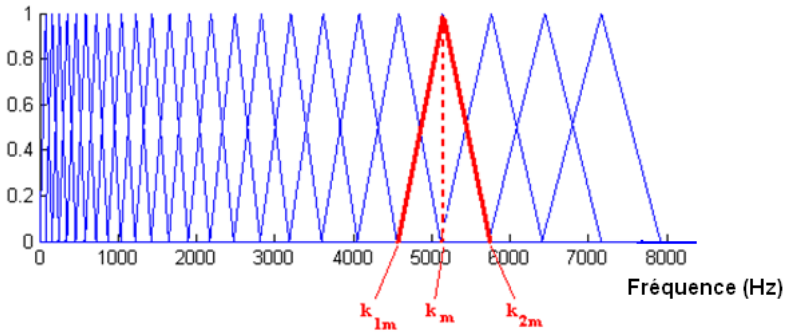


FIGURE 5.7: Banc de filtres triangulaires

Après avoir calculé les MFCC, nous appliquons à ces derniers un *liftrage cepstral* pour augmenter les magnitudes des coefficients d'ordre supérieur (figure 5.8). Il est effectué, dans notre cas, avec une fenêtre sinusoïdale dont l'équation est :

$$w(l) = 1 + \frac{T}{2} sin \frac{\pi l}{T} \quad l = 0, 2, ..., T-1 \tag{5.13}$$

l est la taille de la fenêtre de liftrage w.

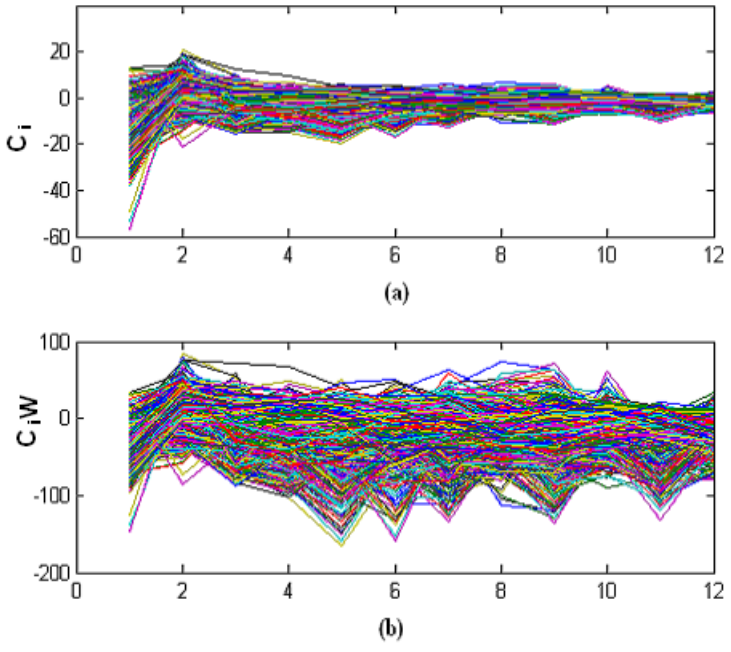


Figure 5.8: Valeurs des MFCC(s) : (a) avant et (b) après liftrage

Les différentes étapes de calcul des MFCC sont détaillées dans [Benbellil and Droua-Hamdani].

5.4.4.3 Dérivées △ et △△ des MFCC

La performance du système de reconnaissance vocale est considérablement améliorée avec l'ajout des dérivées premières et secondes des MFCC [Huang et al., 2001]. L'analyse de régression a été utilisée pour calculer les caractéristiques △ selon la formule suivante :

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{5.14}$$

d_t est le coefficient △ au temps t calculé en terme de coefficients statiques correspondant de $c_{t-\theta}$ à $c_{t+\theta}$. La même formule a été appliquée pour obtenir les coefficients d'accélération (△△).

La configuration du SRAP inclut donc les spécifications suivantes :

1. Un vecteur MFCC de dimension standard constitué de 39 coefficients :12 MFCC (C_1 *à* C_{12}) + l'énergie normalisée (C_0) avec leurs coefficients △ et △△
2. Un coefficient de pré-emphase de 0.97
3. Une fréquence d'échantillonnage de $16\,kHz$
4. Un fenêtrage de Hamming avec une durée de 25 millisecondes et une taille de pas de 10 millisecondes
5. Un liftrage cepstral égal à 22
6. Un banc de 26 filtres
7. Une valeur de Θ fixée à 3 pour le calcul des coefficients △ et △△.

5.4.5 Modèles acoustiques

La parole est une concaténation d'unités acoustiques. Dans le cadre des SRAP Markoviens, ces unités acoustiques sont modélisées par des HMM. Il est possible de segmenter le signal de parole en succession de mots. Toutefois, l'utilisation des mots comme unités de base dans la conception d'un modèle acoustique est incommodant dès que l'on considère un vocabulaire conséquent. Cela supposerait, en effet, que l'on disposerait d'un corpus contenant suffisamment d'occurrences de chaque mot pour permettre un apprentissage efficace du modèle représentatif de chacun. Par conséquent, on utilise généralement une unité plus petite, les phonèmes. Leur nombre est effectivement limité — 34 pour l'AS — et ils conviennent parfaitement à la reconnaissance.

Les modèles acoustiques font intervenir trois niveaux de HMM. Ils cherchent dans un premier temps à reconnaître les types de son, autrement dit, à identifier les phonèmes parmi ceux possibles dans la langue. Pour ce faire, ils modélisent

généralement chaque phonème par un HMM à trois états représentant ses début, milieu (partie stable) et fin. Les états sont alors associés à un sous-phonème (figure 5.9). Pour introduire l'effet du contexte, c'est à dire l'effet de la co-articulation, il est possible de considérer pour la réalisation du SRAP des unités supérieures aux phonèmes comme les triphones (trois phonèmes successifs).

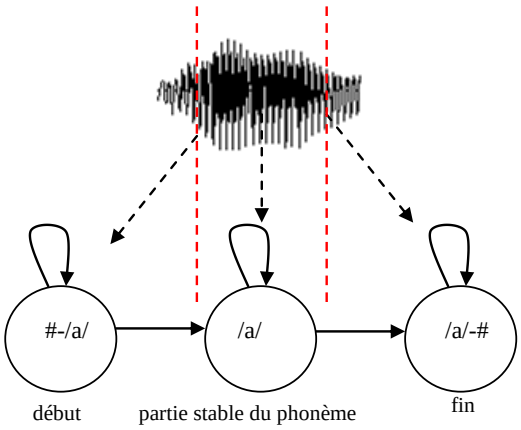


FIGURE 5.9: HMM du phonème /a/. (Le symbole # représente le silence)

Les mots sont ensuite construits en terme de phonèmes, les phrases en terme de mots. La concaténation de phonèmes en mots nécessite, cependant, l'appui du dictionnaire de prononciation pour déterminer la succession convenable de phonèmes à considérer. La construction de phrases quant à elle fait intervenir le modèle du langage (figure 5.10).

Nous avons donc considéré, pour notre part, les 34 phonèmes de la langue Arabe. Ces unités sont, pour l'instant, traitées hors contexte c'est à dire indépendamment des phonèmes suivants ou précédents. Sur le plan textuel les mots sont séparés les uns des autres par des pauses et les phrases sont précédées et succédées par des silences. Aussi, nous avons prévu pour notre modélisation acoustique 36 modèles :

- 28 modèles pour les 28 consonnes,
- 06 modèles pour les 6 voyelles (courtes et longues),
- 02 modèles pour le silence et la pause.

Toutes les unités sont construites sur une topologie de HMM dite de *Bakis* (gauche-droite) à 5 états dont 3 sont émetteurs. Les vecteurs de paramètres sont à densités de probabilités continues représentés par une *mono gaussienne* $g(\mu_i, \Sigma_i)$ (voir chapitre 1). La taille des vecteurs (μ, Σ) est de 39 composantes correspondant aux 12 coefficients MFCC, le log de l'énergie (C_0) et les dérivées premières et secondes $(\triangle, \triangle\triangle)$.

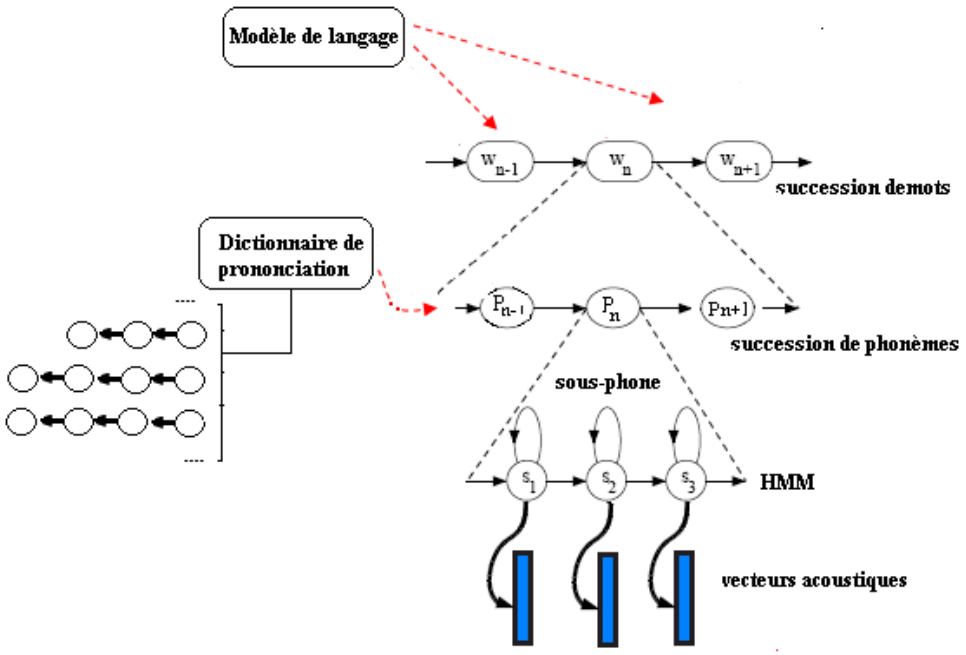


FIGURE 5.10: Niveaux d'analyse des modèles acoustiques

Nous avons calculé, par la suite, à partir de l'ensemble des fichiers de données MFCC, la moyenne globale et la variance de toutes les gaussiennes pour tous les modèles. Le nombre total de paramètres par modèle est de 117. Les nouveaux vecteurs de (μ, Σ) ont été par la suite soumis à une réestimation (un apprentissage) grâce à l'algorithme de *Baum-Welch* où les états d'occupation, les moyennes, les covariances de chaque modèle ont été cumulés progressivement.

Le modèle de la pause (sp) a été créé par l'ajout au modèle du silence (*sil*) des transitions supplémentaires dans sa partie centrale. Son intégration dans le calcul se fait au niveau suivant de l'apprentissage où il subit, par la suite, les différentes ré-estimations autant que les autres HMM.

5.4.6 Réalignement des données d'Apprentissage

Afin de considérer tous les mots listés dans le dictionnaire de phonétisation avec leurs prononciations respectives, principalement les mots qui contiennent de multiples prononciations, nous avons procédé à un réalignement forcé des données. Ce réalignement a pour but de faire correspondre les paramètres acoustiques de l'apprentissage aux différentes transcriptions disponibles dans le lexique. Cette opération permet donc d'accepter toutes les variantes de prononciations d'un mot donné en leurs assignant une seule transcription finale. Ceci est effectué grâce à l'utilisation de l'algorithme de Viterbi selon le processus indiqué dans la figure 5.11.

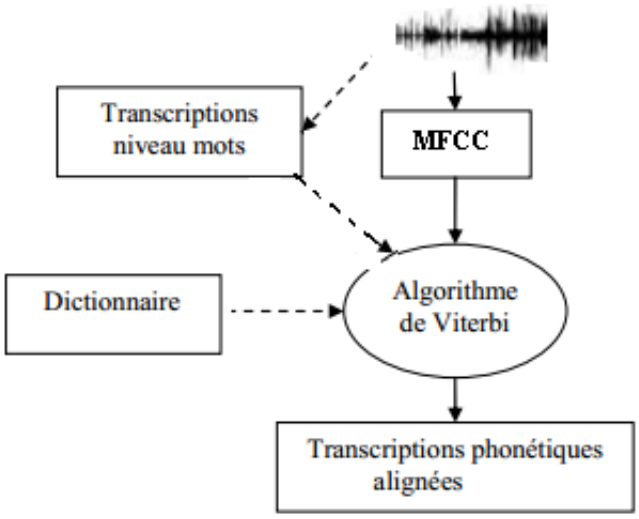


FIGURE 5.11: Alignement des transcriptions avec les vecteurs acoustiques

5.5 RECONNAISSANCE & EVALUATION

Le système, mis en place, est prêt pour l'évaluation de sa performance. Le décodeur cherche à construire un graphe des mots ayant pu être prononcés dans le signal à analyser, à partir des différentes informations transmises par le dictionnaire de prononciation, les modèles acoustiques $P(O \mid W)$ et le modèle du langage $P(W)$. La transcription textuelle issue de la reconnaissance permettra l'évaluation du SRAP.

5.5.1 Évaluation des transcriptions alignées

La qualité d'une transcription finale lors de la reconnaissance est caractérisée par son taux d'erreurs sur les mots (noté WER pour word error rate). Les SRAP sont donc généralement évalués en terme de taux de mots erronés. Ce taux est mesuré en alignant le texte hypothèse produit par le système (Hyp) avec la transcription de référence (Ref). L'alignement est réalisé en minimisant la distance d'édition [Levenshtein, 1966] entre Ref et Hyp.

Trois types d'erreur peuvent être calculés sur les mots reconnus :

- substitution (S) ou remplacement du mot correct par un autre mot ;
- suppression (D) ou omission d'un mot correct ;
- insertion (I) ou ajout d'un mot supplémentaire.

Le WER est calculé avec la formule suivante :

$$WER = 100\,\frac{I + D + S}{N}$$

avec N le nombre de mots dans la transcription de référence.

Par analogie, le SER (Sentence Error Rate) taux d'erreurs sur les phrases donne la proportion de groupes de mots ayant au moins une erreur de transcription. Le SER est calculé avec :

$$SER = 100\,\frac{nombre\,de\,groupes\,de\,mots\,avec\,au\,moins\,une\,erreur}{nombre\,de\,groupes\,de\,mots}$$

Un système performant correspond à un nombre minimal d'insertions, de substitutions et de suppressions de mots.

La figure suivante donne des exemples d'erreurs (insertion, suppression et substitution) entre une phrase de référence et de son hypothèse.

REF :	il AURA ALORS trois examens demain ————
HYP :	il ——- VALEUR trois examens demain SORTIR
	D S I

FIGURE 5.12: Alignement d'une transcription automatique (HYP) et une transcription de référence (REF)

Une autre possibilité d'évaluer la performance du SRAP est de la calculer en terme de mots reconnus par rapport aux mots de référence (word recognition rate noté WWR). C'est cette mesure qui a été utilisée pour exprimer nos résultats de reconnaissance. Les taux relatifs aux mots reconnus (correct%) et à la précision du système sont calculés respectivement par les formules :

$$Correct\% = 100\,\frac{N - D - S}{N}$$

$$Precision\% = 100\,\frac{N - D - S - I}{N} \tag{5.15}$$

N est le nombre total de mots dans les transcriptions de référence.

5.5.2 Performance du SRAP modèles monophones

Le SRAP a servi à l'analyse d'un corpus global contenant les enregistrements des 6 régions, soit un total de 158 phrases réservées exclusivement pour test et prononcées par un ensemble de 49 locuteurs n'ayant pas participé à l'apprentissage. A l'issue de cette expérience, nous avons détaillé les résultats de la reconnaissance pour montrer l'impact de la variabilité régionale et le genre du locuteur sur la performance du système. La paramétrisation (*front-end*) utilisée dans le test est identique à celle de l'apprentissage à savoir les 39 paramètres MFCC.

L'analyse globale des modèles monophones, c'est à dire à contexte indépendant, a révélé un résultat de reconnaissance très satisfaisant. En effet, la performance calculée est de 91.7 % avec une précision de 90.6 %. Par ailleurs, le taux de reconnaissance des phrases s'élève à 85.71 %. Le nombre de substitutions est de 32, le nombre de suppressions est de 6 et enfin le nombre d'insertions est de 5.

5.5.3 Performance du SRAP selon l'accent régional

Dans la deuxième expérience, nous avons détaillé le résultat global de reconnaissance pour observer la variation de la performance lorsque nous considérons l'accent régional. Le résultat précédent indique que la performance du SRAP est acceptable pour la reconnaissance automatique de la parole. Mais en tenant compte de l'accent régional, les résultats montent une variation considérable entre les localités. Les taux de reconnaissances pour les mots et les phrases respectivement (% Corr et % Correct), la précision dans la reconnaissance des mots (% Acc) ainsi que les taux d'erreurs (WER %) sont donnés dans le tableau 5.6 et la figure 5.13.
Les résultats montrent que les régions du nord R_1, R_2 et R_5 ont des taux de reconnaissance relativement élevés par rapport aux taux atteints par les régions du sud c'est à dire R_9, R_{10} et R_{11}. R_5 (Jijel) présente la meilleure performance du système, toutes les phrases produites par les locuteurs de cette région sont correctement reconnues par le SRAP. Rappelons que R_5 présente la prononciation particulière de plusieurs phonèmes comme il est cité dans le chapitre 2. En ce qui concerne R_2 (Tizi Ouzou), le système révèle également un taux de reconnaissance très élevé avec un WER de 2%. Les résultats de R_1, Alger, sont aussi très appréciables pour une reconnaissance de parole continue.
En ce qui concerne les régions situées dans le sud du pays, les résultats sont moins satisfaisants comme exposés dans le tableau 5.6 et figure 5.13. Les régions de Béchar (R_9) et El Oued (R_{10}) présentent des taux de reconnaissance relativement similaires avec des WER respectivement de (22% et 23%). Le plus mauvais score revient à la région centrale du sud à savoir Ghardaia (R_{11}) où la performance du système chute de près de 26% par rapport à la meilleure performance. Le WER obtenu pour cette région est de 28%.

Désignation	mots		phrases	WER (%)
	%Corr	%Acc	%Correct	
R1	94.34	93.87	88.89	6
R2	98.98	97.96	93.33	2
R5	100	100	100	0
R9	78.26	78.26	71.43	22
R10	80.85	76.60	71.43	23
R11	74.42	67.44	64.29	28

TABLE 5.6: Performance du SRAP par région

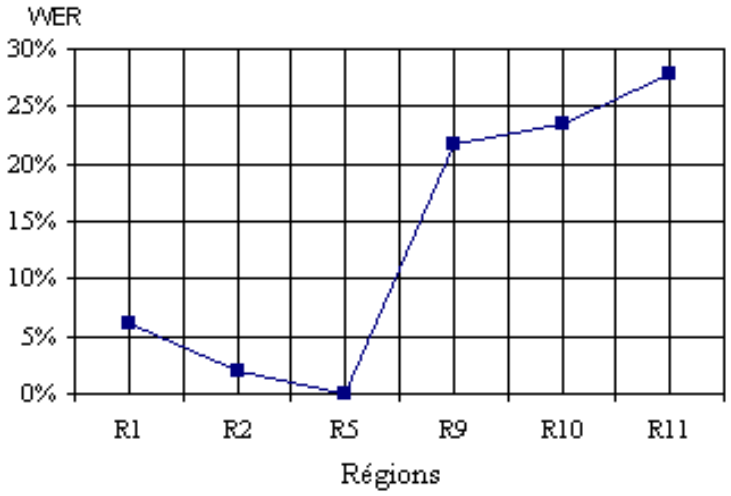


FIGURE 5.13: WER en pourcentage par région

La figure 5.14 montre les erreurs de suppressions, substitutions et insertions du système selon chaque région. Nous remarquons que les régions du sud présentent le nombre le plus élevé de substitutions, notamment R_{10} et R_{11}. Ce nombre est aussi important pour R_1. Le graphe montre également que la suppression de mots la plus importante est observée au niveau de R_1 suivie de R_{10}. Les autres régions ne sont pas touchées par ce type d'erreurs. Par ailleurs, les insertions de mots sont presque équivalentes pour (R_1, R_2) et (R_{10}, R_{11}), avec un léger dépassement pour R_{10}. En ce qui concerne les régions R_5 et R_9, elles ne présentent aucune erreur d'insertion.

5.5.4 Performance du SRAP selon le genre du locuteur

Nous nous intéressons dans cette partie aux locuteurs du corpus test. Les résultats montrent que le système a reconnu au moins une phrase par locuteur. Ceci nous amène à dire, que le système répond favorablement aux exigences d'une conception d'un système multilocuteurs. Cependant, le nombre de phrases correctement reconnues varie d'un participant à un autre et d'une région à une autre.

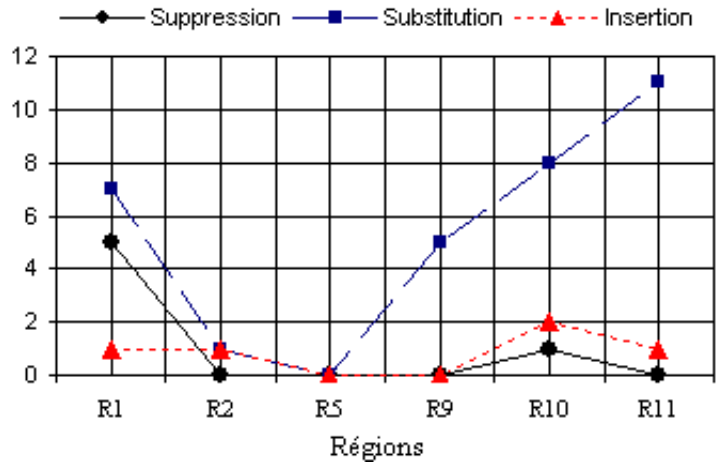


FIGURE 5.14: Comparaison des résultats au niveau des mots entre régions : suppressions, insertions et substitutions

La figure 5.15 montre par région et en pourcentage les locuteurs pour lesquels toutes les phrases prononcées sont reconnues par le SRAP. La figure montre aussi que par rapport aux localités du nord, les locuteurs du sud ont produit la majorité des groupes de mots non reconnus par le système soit ≈20% pour les régions nord contre ≈37% pour le sud. Nous pouvons aussi relevé que près de 50% de locuteurs de R_9 et R_{11}, ont prononcé des suites de mots qui présentent des erreurs de reconnaissance.

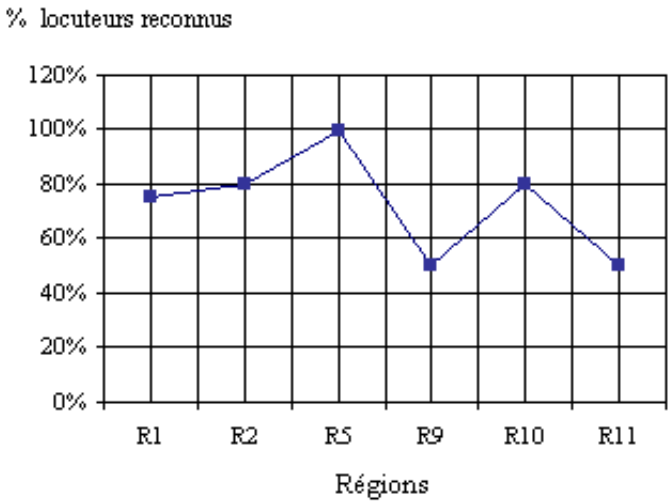


FIGURE 5.15: Nombre de locuteurs en pourcentage ayant produit des phrases sans erreurs de transcription

Afin de voir quel est le type de locuteur qui affecte le plus la performance du SRAP (féminin ou masculin), nous avons procédé à l'analyse des différentes erreurs relevées selon le genre. L'expérience a révélé que le rapport entre les locuteurs qui ont produit des phrases reconnues et ceux qui ont produit les non reconnues par

le système pour R_1 et R_2 est approximativement le même. Mais ce ratio augmente dans les localités du sud où les locuteurs masculins ont prononcé plus de groupes de souffles non reconnus par le système que les locutrices notamment pour R_{10} et R_{11}. Les figures 5.16 et 5.17 illustrent les comparaisons entre les nombres de locuteurs (féminins et masculins) ayant produit des phrases reconnues vs non reconnues par le SRAP.

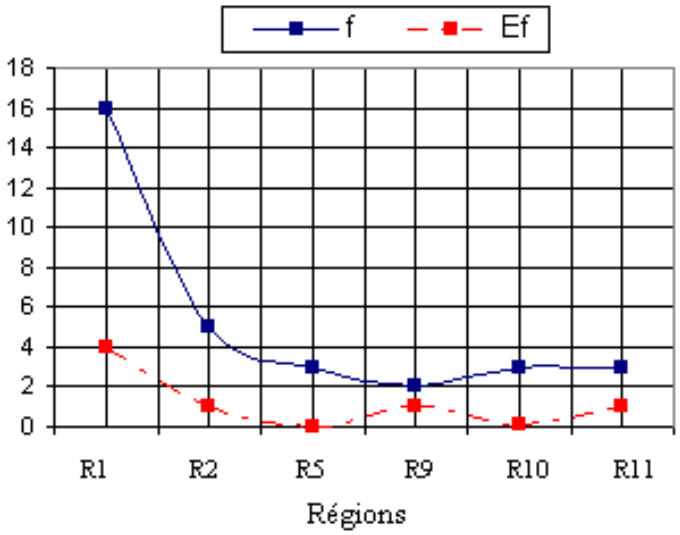


FIGURE 5.16: Comparaison entre le nombre de locutrices ayant produit : les phrases non reconnues (Ef) et reconnues (f)

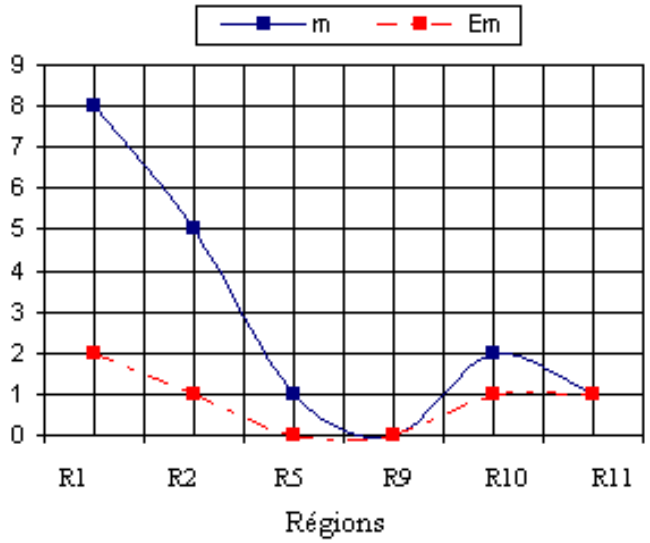


FIGURE 5.17: Comparaison entre le nombre de locuteurs ayant produit : les phrases non reconnues (Em) et reconnnues (m)

5.5.5 Résultats de la transcription

Comme cité précédemment, le rôle d'un SRAP est de produire une transcription du signal sonore. Le résultat est donc naturellement un texte. Le pourcentage de phrases non reconnues par notre système est de 13 % soit 21 sur 158 utilisées pour le test. Nous adopterons dans ce qui suit le terme de groupes de souffles, généralement utilisée en reconnaissance, pour désigner une succession de mots (phrases).

En étudiant les groupes de souffles non reconnus à savoir ceux présentant au moins l'une des trois erreurs : suppressions, insertions ou substitutions, nous avons remarqué que 81% des erreurs recensées se produisent aux extrémités de la phrase c'est à dire à son début (le premier mot) ou à sa fin (dernier mot). Les pourcentages calculés pour chaque type d'erreurs selon leurs positions dans la phrase sont exposés dans le tableau 5.7.

	début	milieu	fin
Sub (%)	59.3	15.6	25
Del (%)	100	-	-
Inser (%)	50	-	50

TABLE 5.7: Suppressions, substitutions et insertions en pourcentage selon la position dans la pharse

Parmi les erreurs de reconnaissance affichées dans la transcription finale en SAMPA, certaines sont récurrentes telles les erreurs sur des mots courts comme : le verbe *ka* : *na* (كان) et sa flexion à l'impératif *kun* (كن) qui constituent plus de 20% des erreurs de reconnaissance. Ces groupes de souffles sont remplacés dans de multiples phrases par l'une des deux propositions : *qa* : *dana* : (قادنا) ou l'absence de transcription. De même, les particules comme *ma* : (ما), *?iDa* : (إذا), *?'an* (عن) constituent près de 25% des erreurs relevées. Ces mots sont remplacés soit par d'autres particules ou mots sinon par rien. Par ailleurs, la grande majorité d'erreurs recensées sont commises sur les noms et verbes qui sont remplacés par d'autres groupes de souffles. Le tableau 5.8 donne quelques exemples d'erreurs de transcription.

Ref	Hyp
ka :na (كان)	qa :dana : (قادنا)
kana (كان)	Ø
kun (كن)	Ø
Ø	kun (كن)
?axaDa (أخذ)	qa :dana : (قادِنا)
laqad (لقد)	?axt'a ?ta (أخطأت)
d'a ?ula (ضؤل)	lam (لم)
saju ?Di :hima : (سيؤذيهما)	s'a :da (صاد)
saqatat (سقطت)	Ø
rafa ?'a (رفع)	?axaDa (أخذ)
ma : (ما)	?iDa : (إذا)
?'an (عن)	Ø
fama : (فما)	?iZa :zatan (إجازة)
Gafala (غفل)	qa :balna : (قابلنا)
qa :nu :nan (قانونا)	musliman (مسلما)

TABLE 5.8: Quelques erreurs de reconnaissance

5.5.6 Analyse des erreurs de transcription

Lors de la reconnaissance, il est possible de retenir plusieurs hypothèses pour un même groupe de souffles, certains mots mal reconnus dans la meilleure hypothèse proposée peuvent aussi apparaître correctement dans les suivantes. Le tableau 5.9 montre une référence (un groupe de mots) avec les différentes hypothèses affichées par le système.

REF	laqad ka :na musa :liman wa qutila (لقد كان مسالما وقتل)
HYP	?axt'a ?ta musa :liman wa qutila (أخطأت مسالما وقتل)
	?axt'a ?ta musa :liman jad't'ahidkum (أخطأت مسالما يضطهدكم)
	?axt'a ?ta wa lam jarX\al ?'awdatin (أخطأت و لم يرحل عودة)

TABLE 5.9: Différentes hypothèses pour une même référence

Ces erreurs de reconnaissance proviennent généralement soit d'une mauvaise réestimation des modèles acoustiques ou bien d'une insuffisance du modèle de langage. Lorsqu'elles sont issues du modèle de langage, elles peuvent être morphologiques, syntaxiques ou sémantiques.

La typologie des erreurs montre que les erreurs de transcription finale sont engendrées à la fois par des insuffisances liée aux deux modélisations acoustique et langage. Nous allons dans ce qui suit, décrire ces erreurs :

Modèles acoustiques

Les modèles acoustiques se basent essentiellement pour leurs estimations sur des indices acoustiques permettant ainsi la reconnaîssance des phones. Cependant, les contractions, sur le plan acoustique, de certains phonèmes sont particulièrement difficiles à modéliser puisque, le vocabulaire du SRAP a une taille limitée et ne les inclut généralement pas. Bien que la parole continue soit par définition constituée d'une suite de mots, dont les limites ne sont pas déterminées, la succession de mots contractés est prononcée d'une seule traite jusqu'à en paraître comme un seul mot. Ces contractions, lorsqu'elles existent, perturbent à la fois la modélisation acoustique des mots de même que le calcul des probabilités de séquences de mots par les modèles de langage. Nous citons pour exemple la transcription de la succession de mots *kun huna :* (**كن هنا**). La perception auditive de cette suite de mots montre une réduction du phonème fricatif [h] (**ه**) de *huna :*, d'où la concaténation des groupes de souffle à en former presque une seul mot *kunhuna :*. Cette entité lexical résultante est bien entendu inexistante dans le dictionnaire de prononciation d'où une probabilité faussée lors de son estimation par les modèles de langage et son remplacement par un mot de poids élevée notamment le mot *qa :dana :* . Nous avons aussi constaté notamment dans le groupe de souffles *s'a :da ?almawruTu mudliZan* (**صاد الموروث مدلجا**, que l'hypothèse du mot *s'a :da* est nulle (absence de transcription). En se référant au plan acoustique, nous constatons que la fricative initiale [s'] (**ص**) de *s'a :da* est presque imperceptible. Cette défaillance dans la production du phonème a faussé les calculs du HMM. Rappelons que ce mot est reconnu dans d'autres phrases de même type. Généralement, les contractions des phonèmes sont dues au débit rapide d'élocution.

Modèles du langage

Comme cité plus haut, les erreurs de transcription engendrées par le modèle de langage peuvent être morphologiques, syntaxiques ou sémantiques. La structuration des mots constitue une source d'information intéressante pour des langues morphologiquement très riches comme l'Arabe notamment, lorsque les hypothèses du SRAP affichent des erreurs de type : remplacement dans la transcription finale de *ka* : *na* (**كان**) par sa flexion à l'impératif *kun* (**كن**). La décomposition des mots en plusieurs constituants élémentaires grâce aux règles de flexions (dérivations), permet de réduire le nombre d'événements à envisager lors du calcul des probabilités et par conséquent réduit un potentiel d'erreurs dans la transcription. Les connaissances sur la morphologie sont intégrées au SRAP au moyen des modèles N-classes [Maltese and Mancini, 1992, Tahir et al., 2004].

Les hypothèses affichées par notre système montrent que les erreurs d'ordre mor-

phologique sont inexistantes. En ce qui concerne les erreurs sémantiques, elles sont aussi inexistantes. En revanche, sur le plan syntaxique, des erreurs ont été notées. Par définition, les mots (noms, verbes ou particules) ont des rôles syntaxiques bien précis selon les positions qu'ils occupent dans la suite de mots [Huet, 2007]. En ce qui concerne notre SRAP, la transcription finale montre que l'équilibre syntaxique de certains groupes de mots a été pertubé. Ce déséquilibre syntaxique, engendré par le remplacement de mots par d'autres n'occupant pas la même fonction grammaticale, peut être partiel (exemple : *ba :?'a* (**باع**) et *?ajna* (**أين**), ou bien total lorsque la structure de la succession de mots est complétement brisée comme dans l'hypothèse *qa :balna : laha : min* **(قابلنا لها من)**. Le tableau 5.10 montre un ensemble d'hypothèses syntaxiquement incorrectes donné par le système.

Ref	ka :na ?al ?aklu laDidan **(كان الأكل لذيذا)**
Hyp_1	wa lam ?al ?aklu laDidan (**و لم الأكل لذيذا**)
Hyp_2	?al ?aklu ladidan (**الأكل لذيذا**)

Ref	Gafala ?'an d'aX\aka :tiha : **(غفل عن ضحكاتها)**
Hyp	qa :balna : laha : min **(قابلنا لها من)**

Ref	ba :?'a t'ablan faxaraZa masru :ran **(باع طبلا فخرج مسرورا)**
Hyp	?ayna t'ablan faxaraZa masru :ran **(أين طبلا فخرج مسرورا)**

TABLE 5.10: Représentation de quelques erreurs syntaxiques dans la transcription finale

5.6 CONCLUSION

Nous avons réalisé, dans ce chapitre, la plate-forme d'un système de reconnaissance automatique de la parole basé sur les HMM. Le SRAP satisfait plusieurs conditions : l'indépendance aux locuteurs et à grand vocabulaire. La validation a été effectuée avec les enregistrements de 6 régions d'ALGASD (3 régions du nord du pays et 3 autres du sud).

Les différentes sections du chapitre ont abordé la stratégie adoptée, les étapes essentielles dans la construction du SRAP ainsi que les contraintes rencontrées. Le système, basé sur les modèles monophones, a servi à l'analyse globale de tous les enregistrements du corpus test.

La performance du SRAP est satisfaisante pour des modèles monophoniques. En effet, les nombres de locuteurs et de phrases alloués pour cette expérience sont suffisants pour tester la performance générale du système. Nous pouvons conclure que le SRAP reconnaît convenablement le système phonétique de l'AS et qu'il répond aux exigences recherchées dans la conception d'un système de reconnaissance de la

parole continue à grand vocabulaire indépendant du locuteur.

Les résultats détaillés de la reconnaissance selon les régions et les locuteurs (genre) ont été donnés. Ces derniers ont été suivis de la description des erreurs de transcription finale.

Certains travaux exposent qu'une réduction des erreurs de reconnaissance est possible si l'on apporte des améliorations aux deux modèlisations acoustique et langage. En effet, les modèles acoustiques, gagneraient davantage à exploiter des connaissances additionnelles sur le plan de la phonétique des sons [Huet, 2007, Huang et al., 2001]. Aussi, l'utilisation, des modèles acoustiques tenant compte des contextes droit et gauche du phonème, à savoir des modèles triphones, peuvent être envisagés dans notre cas à la place des monophones. D'autres études soulignent également, que la reconnaissance de traits phonétiques spéciques aux phonèmes pourraient être utilisée pour guider ensuite le décodage de la parole [Juneja, 2004].

En ce qui concerne le modèle de langage, le modèle bigrammes utilisé pour notre SRAP convient pour la prédiction des suites de mots improbables du corpus test. Cependant, il existe des techniques d'amélioration qui consistent à déterminer les événements impossibles dans le langage. [Langlois et al., 2003] montre qu' en affinant le calcul des probabilités des modèles N-grammes, nous pouvons exclure du modèle de langage toutes les suites de mots illicites comme dans notre hypothèse *lam ?al ?aklu laDidan* (**و لم الأكل لذيذا**).

CONCLUSIONS ET PERSPECTIVES

La nécessité d'exploiter des corpora oraux pour faire évoluer les recherches fondées sur le langage, notamment celles attelées aux nouvelles technologies de l'information, est d'une évidence marquée. Aussi, pour satisfaire cette nécessité une recrudescence importante en matière de collecte et de réalisation de tout type de bases de données standardisées, pour toutes les langues, a été observée ces deux dernières décennies. Le principal avantage de cette pratique est de donner aux chercheurs la possibilité de comparer les performances des différentes techniques d'analyse sur des données communes pour en déterminer les approches les plus prometteuses à poursuivre. Toutefois, les ressources orales dédiées à l'Arabe Standard sont peu fréquentes. Les corpora développés se restreignent généralement à des applications ciblées. C'est dans cette optique qu'est née l'initiative de concevoir la base de données *ALGerian Arabic Speech Database* (ALGASD) réalisée à partir d'enregistrements de textes arabes prononcés par des locuteurs algériens. Pour l'élaborer, une charpente dans laquelle sont détaillées toutes les étapes de construction depuis la définition des objectifs jusqu'à sa complète édification, a été échafaudée. Par ailleurs, ALGASD a nécessité en amont, une recherche documentaire approfondie traitant, d'une part, des données démographiques et statistiques de la population algérienne et d'autre part, des principales caractéristiques phonétiques reconnues comme discriminantes. C'est à la lumière de ce déblayage phonétique, que nous avons pu mettre en place une couverture régionale qui respecte à la fois les grands groupes de prononciations recensés et la population effective dans les régions étudiées. Aussi, nous avons suggéré comme échantillons de la variabilité régionale, 11 variations de prononciations, relevées à partir de 300 locuteurs répartis à travers tout le territoire national. Ces régions sont au nombre de 8 pour le nord (Centre : Alger, Tizi Ouzou et Médéa ; Est : Jijel, Constantine et Annaba ; Ouest : Oran et Tlemcen) et 3 pour le sud (Centre : Ghardaia ; Est : El Oued ; Ouest : Bechar). ALGASD compte un total de 1080 fichiers sons enregistrés à partir de 200 phrases phonétiquement équilibrées.

En plus du facteur régional, la base a inclus pour chaque locuteur, des renseignements : le genre, l'âge et le niveau d'instruction. Ces données descriptives de la

société algérienne constituent un substrat essentiel pour : des études comparatives (acoustiques, perceptives, etc.) visant à illustrer de possibles différences et similitudes dans la production des phonèmes de l'AS ; des études prosodiques concernant le rythme de la parole, la mélodie, etc. De même qu'elles peuvent être utilisées dans le développement et l'évaluation d'un système de reconnaissance automatique de la parole pour la langue arabe. En d'autres termes, le nombre important de participants, la diversité des profils retenus, la qualité des enregistrements, etc. offrent un large champ d'expérimentation à tous les niveaux de la langue.

Pour notre part, ALGASD a été la charnière de plusieurs travaux parmi lesquels ceux présentés dans ce document, à savoir l'étude du rythme de l'AS prononcé par les Algériens et l'implémentation d'un système de reconnaissance automatique de la parole continue multilocuteurs. Contrairement aux autres langues, les études expérimentales traitant le rythme dans la parole sont quasi inexistantes pour l'AS.

La typologie rythmique avancée dans la théorie, classe les langues du monde dans des catégories absolues : langues syllabiques, accentuelles et moraïques. L'Arabe et ses variantes régionales sont considérées comme appartenant aux langues accentuelles. Au niveau méthodologique, nous avons appliqué les sept corrélats acoustiques du rythme à savoir : les mesures d'intervalles et leurs normalisation ainsi que les Pairwises indices sur un large éventail de locuteurs.

Les résultats des analyses acoustiques et statistiques effectuées sur la région d'Alger, nous ont permis de noter des différences significatives au niveau inter et intra langue. En effet, nous avons observé que l'AS des Algériens, indépendamment du profil des locuteurs, n'occupe pas systématiquement une place parmi les langues accentuelles comme énoncé dans la littérature mais se situe à une position intermédiaire entre les langues syllabiques et les langues accentuelles.

Par ailleurs, l'étude du rythme au niveau intra langue a révélé qu'il existe des différences significatives dans la façon de réaliser les durées des voyelles longues et courtes de l'AS. Cette différence est dépendante du profil des groupes de locuteurs à savoir l'âge et le niveau d'instruction. En effet, les résultats montrent que les scores rythmiques décroissent selon les compétences des locuteurs en AS.

La dernière partie du document se rapporte à la validation d'un système de reconnaissance de la parole continue multilocuteurs grâce à l'utilisation de la base ALGASD. En effet, nous avons réalisé un système de base fondé sur les HMM. Pour ce faire, nous avons utilisé les enregistrements de 6 régions d'ALGASD (3 régions du nord du pays et 3 autres du sud). Le SRAP a révélé un résultat très satisfaisant en ce qui concerne sa performance 91.7%. Par ailleurs, l'analyse détaillée de la transcription finale a montré des taux de reconnaissance variant selon les régions et le genre des locuteurs. Une analyse des erreurs a été faite pour déterminer

leurs types : morphologiques, syntaxique, etc.

En perspective, nous souhaitons d'une part, élargir l'apprentissage du SRAP à la toute la base (11 régions) et d'introduire d'autre part, les modèles triphones ainsi que les paramètres prosodiques, notamment le rythme de la parole, dans la paramétrisation du SRAP afin d'améliorer les modèles acoustiques.

Bibliographie

D. Abercrombie. *Elements of general phonetics*, volume 203. Edinburgh University Press Edinburgh, 1967.

M.A.M. Abushariah, R.N. Ainon, R. Zainuddin, M. Elshafei, and O.O. Khalifa. Natural speaker-independent arabic speech recognition system based on hidden markov models using sphinx tools. pages 1–6, 2010.

K. Achab. The tamazight language profile. *Written for the Department of International languages. Ottawa-Carleton School Borad, Ontario Ministry of Education University of Ottawa.*, 2001.

M. Adiba and C. Collet. *Objets et bases de données : le SGBD O2*. Hermes, 1993.

M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, and A. Alenazi. Saudi accented arabic voice bank, 2008.

M. Alkanhal, M. Alghamdi, and Z. Muzaffar. Speaker verification based on saudi accented arabic database. In *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pages 1–4. IEEE, 2007.

Y.A. Alotaibi. Comparative study of ann and hmm to arabic digits recognition systems. *Engineering Sciences*, 19(1) :43–60, 2008.

Y.A. Alotaibi, S.A. Selouani, and D. O'Shaughnessy. Experiments on automatic recognition of nonnative arabic speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008 :1, 2008.

Y.A. Alotaibi, M. Alghamdi, and F. Alotaiby. Speech recognition system of arabic alphabet based on a telephony arabic corpus. In *Image and Signal Processing : 4th International Conference, ICISP 2010, Québec, Canada, June 30-July 2, 2010. Proceedings*, page 122. Springer-Verlag New York Inc, 2010.

B. AlQatab and R.N. Ainon. Arabic speech recognition using hidden markov model toolkit (htk). 2 :557–562, 2010.

G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S.P. Singh, RNV Sitaram, and SP Kishore. Development of indian language speech databases for large vocabulary speech recognition systems. Citeseer, 2008.

A. Arvaniti. Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1–2) :46, 2009.

J. Ashraf, N. Iqbal, N.S. Khattak, and A.M. Zaidi. Speaker independent urdu speech recognition using hmm. pages 1–5, 2010.

C. Astésano. *Rythme et accentuation en français : invariance et variabilité stylistique*. L'Harmattan, 2001.

E.L. Asu and F. Nolan. Estonian rhythm and the pairwise variability index. In *FONETIK 2005*, page 29. Citeseer, 2005.

M. Baltazani. Prosodic rhythm and the status of vowel reduction in greek. In *Selected papers on theoretical and applied linguistics from the 17th international symposium on theoretical & applied linguistics*, volume 1, page 43, 2007.

R. Bassiouney. *Arabic sociolinguistics*. Edinburgh University Press, 2009.

O. Baude, C. Blanche-Benveniste, M.F. Calas, P. Cappeau, P. Cordereix, L. Goury, M. Jacobson, I. De Lamberterie, C. Marchello-Nizia, and L. Mondada. Corpus oraux, guide des bonnes pratiques 2006. 2006.

F. Béchet. Lia phon : un systeme complet de phonétisation de textes. *Traitement automatique des langues*, 42(1) :47–67, 2001.

The british english example pronunciation (beep) dictionary. Dernière consultation Octobre 2011 BEEP. http ://svr-www.eng.cam.ac.uk/comp. speech/section1/lexical/beep.html.

K. Benbellil and G. Droua-Hamdani. Implémentation des mfcc(s) et son application dans une commande vocale. *A paraitre dans AL-LISANIYYAT Revue algérienne des seciences et technologies du langage.*

M. Benrabah. Language-in-education planning in algeria : Historical development and current issues. *Language Policy*, 6(2) :225–252, 2007.

B. Bloch. Studies in colloquial japanese iv phonemics. *Language*, 26(1) :86–125, 1950.

P. Boersma and D. Weenink. Praat : doing phonetics by computer (version 5.1. 18) [computer program]. *(University of Amsterdam)*, 9 :2009, 2009.

R. Boite. *Traitement de la parole*. PPUR, 2000.

J.F. Bonnot. lÉtude expérimentale de certains aspects de la gémination et de lémphase en arabe, 1979.

M. Boudraa, B. Boudraa, and B. Guerin. Twenty lists of ten arabic sentences for assessment. volume 86, pages 870–882. S. Hirzel Verlag, 2000.

Calliope. *La Parole et son Traitement Automatique*. Masson, 1989.

J.P. Campbell Jr and D.A. Reynolds. Corpora for the evaluation of speaker recognition systems. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 829–832. IEEE, 1999.

D. Caubet. Questionnaire de dialectologie du maghreb (d'après les travaux de w. marçais, m. cohen, gs colin, j. cantineau, d. cohen, ph. marçais, s. lévy, etc.). Number 5, pages 73–90, 2000.

F. Cheriguen. Politiques linguistiques en algérie. volume 52, pages 62–73. ENS Editions, 1997.

W. Cichocki, SA Selouani, and L. Beaulieu. The racad speech corpus of new brunswick acadian french : design and applications. volume 36, pages 3–10, 2008.

K. De Jong and B.A. Zawaydeh. Stress, duration, and intonation in arabic word-level prosody. *Journal of Phonetics*, 27(1) :3–22, 1999.

V. Dellwo. Rhythm and speech rate : A variation coefficient for deltac. *Language and language-processing*, pages 231–241, 2006.

V. Dellwo, B. Aschenberner, P. Wagner, J. Dancovicova, and I. Steiner. Bonntempo-corpus and bonntempo-tools : A database for the study of speech rhythm and rate. In *Eighth International Conference on Spoken Language Processing*, 2004.

A. Di Cristo and D. Hirst. Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. syllabic rhythm, melodic rhythm and hierarchical representation of prosody in french. *Travaux de l'Institut de Phonétique d'Aix*, 15 :9–24, 1993.

G. Dreyfus. *Réseaux de neurones : Méthodologie et applications*. Eyrolles, 2002.

G. Droua-Hamdani. Durées des voyelles courtes/longues de l arabe standard en milieux emphatique et géminé. *Colloque international en Traductologie et TAL. Oran, Algeria*, April 09-11, 2007.

G. Droua-Hamdani and M. Guerti. Application of model to correct the phoneme ś duration produced by tts system -case of arabic voiced phonemes. *Colloque international sur le Traitement Automatique de la Langue Arabe (CITALA'07-IEEE), ISBN : 9954-412-12-3. Rabat, Maroc*, pages 63–70, June 18-19, 2007.

G. Droua-Hamdani, S.-A Selouani, M. Boudraa, and B. Boudraa. Algasd project : Statistical study of vocalic variations according to education levels of algiers speakers. In *In Intonational Variation in Arabic conference IVA09 . York, England*, September 28-29, 2009.

G. Droua-Hamdani, W. Cichocki, S.A. Selouani, and M. Boudraa. Algerian arabic rhythm classification. In Antonis Botinis, editor, *ISCA International Speech Communication Association, in Proceedings of the third ISCA Tutorial and Research Workshop Experimental Linguistics, ExLing 2010, ISBN : 978-960-466-067-7. Athens, Greece*, pages 37–41, August 25-27, 2010a.

G. Droua-Hamdani, S.A. Selouani, and M. Boudraa. Algerian arabic speech database (algasd) : Description and research applications. In *LREC 2010-Semetic Languages 2010. Valetta, Malta*, pages 37–41, May 17-19, 2010b.

G. Droua-Hamdani, S.A. Selouani, and M. Boudraa. Algerian arabic speech database (algasd) : Corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering, ISSN 1319-8025. 35(2C).*, 35(2C) :157–166, 2010c.

G. Droua-Hamdani, M. Boudra, and S-A. Selouani. Arabic speech continuous recognition system. *2012 International Conference on Mutimedia Computing ans Systems. In Proceedings, IEEE Catalog Number : CFP120050-1-4673-1519-7. Tangier, Morocco*, May 10-12, 2012a.

G. Droua-Hamdani, M. Boudraa, and S.A. Selouani. Speaker-independent asr for modern standard arabic : effect of regional accents. *International Journal of Speech Technology. ISSN 1381-2416. DOI : 10.1007/s10772-012-9146-4. Springer-Verlag eds*, pages 1–7, 2012b.

G. Droua-Hamdani, M. Boudra, and S-A. Selouani. Effect of characteristics of speakers on msa asr performance. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–5. IEEE, 2013.

Language Resources Association ELRA. http ://www.icp.grenet.fr/elra.

M. Elshafei, H. Al-Muhtaseb, and M. Al-Ghamdi. Speaker-independent natural arabic speech recognition system. In *The International Conference on Intelligent Systems ICIS 2008, Bahrain*, 2008.

T. Fillon. *Traitement numérique du signal acoustique pour une aide aux malentendants*. PhD thesis, Télécom ParisTech, 2004.

National Geographic Geo. http ://maps.nationalgeographic.com/maps.

R. Giordano and L. DAnna. A comparison of rhythm metrics in different speaking styles and in fifteen regional varieties of italian. *In Speech Prosody 2010*, 2010.

E. L. Grabe, E.; Low. Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, 7 :515–546, 2002.

V. Galounov H. van den Heuvel and H.S. Tropf. The speechdat (e) project : Creating speech databases for eastern european languages. In *Proceedings Workshop on Speech Database Development for Central and Eastern European Languages, Granada, Spain. 26 May*, 1998.

N. Haeri. Form and ideology : Arabic sociolinguistics and beyond. *Annual review of anthropology*, 29 :61–87, 2000.

R. Hamdi, M. Barkat-Defradas, E. Ferragne, and F. Pellegrino. Speech timing and rhythmic structure in arabic dialects : a comparison of two approaches. In *Proceedings of Interspeech 04 ; Jeju Island, Korea*, volume 4, 2004.

A. Harrag and T. Mohamadi. Qsdas : New quranic speech database for arabic speaker recognition. *Arabian Journal for Science and Engineering*, 35(2) :7, 2010.

X. Huang, A. Acero, and H.W. Hon. *Spoken Language Processing : A guide to theory, algorithm and system development*. Prentice Hall PTR, 2001.

S. Huet. *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole*. PhD thesis, Université Rennes 1, 2007.

International Phonetic Association IPA. *Handbook of the International Phonetic Association : a guide to the use of the international phonetic alphabet*. Cambridge Univ Pr, 1999.

E. Jacewicz, R.A. Fox, C. O'Neill, and J. Salmons. Articulation rate across dialect, age, and gender. *Language variation and change*, 21(02) :233–256, 2009.

R. Jakobson. *Mufaxxama : The emphatic phonemes in Arabic*. 1957.

T.Y. Jang. Automatic assessment of non-native prosody using rhythm metrics : Focusing on korean speakers' english pronunciation. In *Proc. of the 2nd International Conference on East Asian Linguistics*, 2009.

F. Jelinek. *Statistical methods for speech recognition.* the MIT Press, 1997.

A. Juneja. Speech recognition based on phonetic features and acoustic landmarks. 2004.

D.H. Klatt. Linguistic uses of segmental duration in english : Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59 :1208, 1976.

V.R.V. Krishnan, A. Jayakumar, and P.B. Anto. Speech recognition of isolated malayalam words using wavelet features and artificial neural network. pages 240–243, 2008.

L. Labrune. Structure de la syllabe japonaise. *Faits de langues*, (17) :111–122, 2001.

P. Ladefoged and K. Johnson. *A course in phonetics.* Wadsworth Pub Co, 2010.

D. Langlois, A. Brun, K. Smaïli, and J.P. Haton. Événements impossibles en modélisation stochastique du langage : Modélisation probabiliste du langage naturel. *TAL. Traitement automatique des langues*, 44(1) :33–61, 2003.

Linguistic Data Consortium LDC. http ://www.ldc.upenn.edu.

A. Lee, T. Kawahara, and K. Shikano. Julius—an open source real-time large vocabulary recognition engine. 2001.

V. I. Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.*, volume 10. Soviet Physics Doklady, 1966.

L.E. Ling, E. Grabe, and F. Nolan. Q uantitative characterizations of speech rhythm : Syllable-timing in singapore english. *Language and Speech*, 43(4) :377, 2000.

Z. Lishuang and H. Zhiyan. Speech recognition system based on integrating feature and hmm. 3 :449–452, 2010.

G. Maltese and F. Mancini. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. 1 :157–160, 1992.

P. Marçais. *Le parler arabe de Djidjelli, Nord constantinois, Algérie.* Librairie d'Amérique et d'Orient, 1956.

S. Morgenthaler. *Introduction à la statistique*. PPUR, 2007.

A. H. Moussa. Statistical study of arabic language roots in mo3jam al-sehah. 1973.

G. Muhammad, Y.A. Alotaibi, and M.N. Huda. Automatic speech recognition for bangla digits. pages 379–383, 2009.

E. O Rourke. Speech rhythm variation in dialects of spanish : Applying the pairwise variability index and variation coefficients to peruvian spanish. In *Proceedings of Speech Prosody 008 : Fourth Conference on Speech Prosody*, 2008.

Office National des Statistiques ONS. http ://www.ons.dz.

A. Pamies Bertrán. Prosodic typology : on the dychotomy between stress. timed and syllable-timed languages. *Language Design : Journal of Theoretical and Experimental Linguistics*, (2) :103–131, 1999.

N. Paulsson, K. Choukri, D. Mostefa, D. Dipersio, M. Glenn, and S. Strassel. A large arabic broadcast news speech data collection. Citeseer, 2004.

D. Petrovska, J. Hennebert, H. Melin, and D. Genoud. Polycost : a telephone-speech database for speaker recognition. volume 31, pages 265–270. Citeseer, 2000.

K.L. Pike. *The intonation of american english*. Greenwood Pub Group, 1979.

L. Rabiner and B.H. Juang. *Fundamentals of speech recognition*. Prentice hall, 1993.

F. Ramus. Acoustic correlates of linguistic rhythm : Perspectives. In *Speech prosody*, pages 115–120. Citeseer, 2002.

F. Ramus, M. Nesporb, and J. Mehlera. Correlates of linguistic rhythm in the speech signalq. *Cognition*, 73(265) :265–292, 1999.

P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel, et al. Babel : An eastern european multilanguage database. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1892–1893. IEEE, 1996.

The Speech Assessment Methods Phonetic Alphabet SAMPA. http ://www.phon.ucl.ac.uk/home/sampa/arabic.html. dernière consultation octobre 2011.

H. Satori, M. Harti, and N. Chenfour. Arabic speech recognition system using cmusphinx4. *arXiv preprint arXiv :0704.2201*, 2007.

H. Satori, H. Hiyassat, M. Harti, and N. Chenfour. Investigation arabic speech recognition using cmu sphinx system. *The International Arab Journal of Information Technology*, 6(2), 2009.

F. Schiel, C. Draxler, A. Baumann, T. Ellbogen, and A. Steffen. The production of speech corpora. *Bavarian Archive for Speech Signals : Munich. Accessed from www. phonetik. uni-muenchen. de/forschung/Bas/BasLiteratur. html on September*, 15 :2009, 2004.

T. Schultz and A. Waibel. Globalphone, das projekt globalphone : Multilinguale spracherkennung computers, linguistics, and phonetics between language and speech. In *Bernhard Schroder et al (Ed.) Springer, Berlin 1998, ISBN Proceedings of the 4th Conference on NLP - Konvens-98, Bonn, Germany, October*, 1998.

E. Shriberg and A. Stolcke. Prosody modeling for automatic speech recognition and understanding. *Mathematical Foundations of Speech and Language Processing*, pages 105–114, 2004.

R. Siemund, H. Hoge, S. Kunzmann, and K. Marasek. Speecon-speech data for consumer devices. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume 2, pages 883–886. Citeseer, 2000.

K. Sjolander and J. Beskow. Wavesurfer-an open source speech tool. In *Proceedings of ICSLP*, volume 4, pages 464–467. Citeseer, 2000.

AR Sukumar, AF Shah, and PB Anto. Isolated question words recognition from speech queries by using artificial neural networks. pages 1–4, 2010.

Y. Tahir, N. Chenfour, and M. Harti. Modélisation à objets d'une base de données morphologique pour la langue arabe. *JEP-TALN*, 2004.

K. Taleb Ibrahimi. *Les Algériens et leur (s) langue (s). Éléments pour une approche sociolinguistique de la société algérienne.* Alger, Dar El-Hikma, 1995.

E.R. Thomas. Phonological and phonetic characteristics of african american vernacular english. *Language and Linguistics Compass*, 1(5) :450–475, 2007.

A. Turk, S. Nakai, and M. Sugahara. Acoustic segment durations in prosodic research : A practical guide. *Methods in empirical prosody research (S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, J. Schliesser), (de Gruyter, Berlin)*, 3 :1–28, 2006.

J.P.H. Van Santen. Contextual effects on vowel duration. *Speech Communication*, 11(6) :513–546, 1992.

J. Verhoeven, G. De Pauw, and H. Kloots. Speech rate in a pluricentric language : A comparison between dutch in belgium and the netherlands. *Language and Speech*, 47(3) :297, 2004.

J. Vonwiller, I. Rogers, C. Cleirigh, and W. Lewis. Speaker and material selection for the australian national database of spoken language. volume 2, pages 177–211. Routledge, 1995.

A. Waibel and A. Weibel. *Prosody and speech recognition*. Pitman, 1988.

W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4 : A flexible open source framework for speech recognition. 2004.

C. Wang. *Prosodic modeling for improved speech recognition and understanding*. PhD thesis, Massachusetts Institute of Technology, 2001.

J.C.E. Watson. *The phonology and morphology of Arabic*. Oxford University Press, USA, 2007.

L. White and S.L. Mattys. Calibrating rhythm : First language and second language studies. *Journal of Phonetics*, 35(4) :501–522, 2007.

L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S.L. Mattys. How stable are acoustic metrics of contrastive speech rhythm ? *The Journal of the Acoustical Society of America*, 127 :1559, 2010.

SJ Young, G. Evermann, MJF Gales, D. Kershaw, G. Moore, JJ Odell, DG Ollason, D. Povey, V. Valtchev, and PC Woodland. The htk book version 3.4. 2006.

F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16(6) :582–589, 2001.

V. Zue, S. Seneff, and J. Glass. Speech database development at mit : Timit and beyond, 1990.

ANNEXES

Annexe A : Quelques ressources orales mondiales

SpeechDat

Le projet SpeechDat, a pour principale tâche l'aide à la création d'une infrastructure européenne pour la distribution et l'évaluation des ressources linguistiques qui ont attrait au développement et à la validation des systèmes à commande vocales [H. van den Heuvel and Tropf, 1998]. Ce projet est constitué d'un ensemble de bases de données multilingue construites à partir de communications par réseaux téléphoniques.

BREF120

Réalisé en langue française, le corpus BREF120 est constitué d'enregistrements d'articles tirés du quotidien "Le Monde" ELRA. Le but de sa réalisation est le développement et l'évaluation des systèmes RAP continue ainsi que la fourniture d'un large corpus pour l'acquisition des connaissances linguistiques. Les textes sélectionnés sont lus par 120 locuteurs (65 femmes et 55 hommes). La taille de son vocabulaire dépasse les 20 000 mots.

GlobalPhone

GlobalPhone est un corpus de base de données multilingues collectées par l'université allemande Karlsruhe. Englobant les 15 langues les plus répandues dans le monde Schultz and Waibel [1998], Globalphone fournit des données de parole transcrites pour le développement et l'évaluation des systèmes RAP continue à large vocabulaire. Chaque corpus comporte 100 phrases lues approximativement par 100 locuteurs adultes tous natifs du pays considéré. Les textes sont choisis à partir d'articles de journaux nationaux ou internationaux traitants les domaines politiques ou économiques. Le total de locuteurs du projet s'élève à 1500 pour un total de 300 heures d'enregistrements.

POLYCOST

Appartenant au projet européen COST 50, POLYCOST est dédié aux applications de reconnaissance du locuteur à travers les lignes téléphoniques. Utilisant l'anglais parlée par des locuteurs non natifs. Ce corpus est destiné pour couvrir 13 pays européens. Il est caractérisé par son important nombre de locuteurs (>1000), ses enregistrements comportent des chiffres lus et des discours libres avec environ 10 sessions d'appels par locuteur [Petrovska et al., 2000].

Dérivés du TIMIT

A partir de corpus principal, plusieurs autres ont été dérivésLDC. La majorité de ces sous corpus est conçue dans le but d'étudier la parole à travers une transmission téléphonique. La famille TIMIT englobe :

- **FMTIMIT :** comprend des enregistrements des sessions originales de TIMIT faites à partir d'un microphone secondaire de champ lointain.
- **NTIMIT :** créé par NYNEX Science and Technology Speech Communication Group, ce corpus correspond à la transmission des 6300 phrases du TIMIT original dans un réseau téléphonique.
- **CTIMIT :** élaboré par Lockheed-Martin Sanders, Inc. et produit par LDC, le CTMIT est doté de la même organisation que celle de TIMIT. En effet, il présente les même distributions : de locuteurs/corpus, textes, corpus d'apprentissage/test, les types de fichiers, etc. C'est un sous ensemble des données originales de TIMIT constitué par la transmission de 3367 des 6300 phrases à travers un téléphone cellulaire (la fréquence d'échantillonnage est de 8 kHz au lieu de 16 kHz pour TIMIT).
- **HTIMIT :** réalisé par MIT Lincoln Laboratory Speech Systems Technology Group et produit par LDC, est un re-enregistrement d'un sous corpus de TIMIT à travers différents combinés de téléphone. Le but est de créer un corpus pour l'étude de la transmission téléphonique et son effet (les canaux de téléphone et le bruit de fond variables). Il est constitué d'un ensemble de 10 phrases prononcées par 192 locuteurs masculins et par 192 locuteurs féminins par un haut-parleur stéréo.

Annexe B : Ressources orales arabes

SAAVB

Le Saudi Accented Arabic Voice Bank (SAAVB) collecté entre 2002 et 2003 à l'université des Sciences et de la Technologie du Roi Abdulaziz est conçue pour applications visant la reconnaissance vocale [Alghamdi et al., 2008]. Les locuteurs ont été enregistrés via les lignes téléphoniques dans différentes régions de l'Arabie saoudite :

- Nombre de locuteurs : 1033 natif (51% hommes/ 49% femmes).
- Age des locuteurs : 50% (16-30 ans), 35% (31-45 ans), et 15% (46-60 ans).
- Canal d'Enregistrement : Téléphonie (70% téléphones mobiles/ 30% téléphones fixes).
- Environnement d'Enregistrement : voiture, extérieur et intérieur.
- Texte : chiffres, mots, phrases, prononciation de l'alphabet arabe et anglais.
- Taille du vocabulaire : 34 961.
- Total parole : 2.59 GB.
- Taux d'échantillonnage : 8 kHz (discours de téléphonie).
- Codage : 16.
- Informations fournies : conditions d'enregistrement, genre, âge des locuteurs.
- Non accessibles au public.

BBN / AUB DARPA Babylon Levantine Arabic Speech Corpus (BBL)

Géré par DARPA (Defense Advanced Research Project Agency), la base de données BBN est développée en 2002, aux États-Unis d'Amérique (Boston) et au Liban (université américaine de Beyrouth), pour l'accent Libanais, Jordanien, Syrien [LDC].

- Nombre de locuteurs : 164 natif du Levant arabe (101 hommes /63 femmes)
- Répartition par âge des locuteurs : Non disponible
- Matériel d'Enregistrement : microphone antibruit casque, (l'électronique Andrea NC-65).

- Environnement d'Enregistrement : non mentionné.
- Texte : spontanée.
- Taille du vocabulaire : 15 ko.
- Total mots : 336 ko.
- Texte /locuteur : non mentionné.
- Total parole : 6,5 GB (1/3 de la durée de parole représente le silence).
- Taux d'échantillonnage : 16 kHz.
- Codage : 16 bits.
- Informations fournies : Aucune.
- Non disponible au grand public.

QSDAS

La base de données Quranic Speech Database for Arabic Speakers (QSDAS) a été récemment publié pour la reconnaissance des locuteurs [Harrag and Mohamadi, 2010].

- Nombre de locuteurs : 77 locuteurs masculins. Aucune mention sur l'origine des locuteurs.
- Age des locuteurs : 18-65 ans.
- Matériel d'Enregistrement : micro normal, un seul canal.
- Environnement d'Enregistrement : Non mentionné.
- Texte : chapitres du saint Coran (21 sourates).
- Taille du vocabulaire : 749.
- Texte/locuteur : 21 sourates.
- Total parole : 15,4 GB.
- Fréquence d'échantillonnage : 16 kHz.
- Codage : 16 bits.
- Informations fournies : pitch et formants.
- Non disponible au grand public.

West Point Arabic Speech Corpus

West Point Arabic Speech Corpus est collecté et traité par les membres du département des langues étrangères de l'Académie militaire américaine de West Point et le Center for Technology Enhanced Learning Langue (CTELL) [LDC]. Cette base de données contient de la parole arabe produite par des locuteurs arabophones et non arabophones. Ce corpus est adapté pour les applications de reconnaissance automatique de la parole.

- Nombre de locuteurs : Total 110. Natif 75 (41 hommes et 34 femmes) ; non natif 35 (25 hommes et 10 femmes).

- Age : non mentionné.
- Matériel d'Enregistrement : SHURE SM10A.
- Environnement d'Enregistrement : non mentionné.
- Texte : Phrases.
- Total données : 1,74 Go de données
- Fréquence d'Echantillonnage : 22,05 kHz.
- Codage : 16 bits.
- Informations fournies : Tous les scripts sont diacritisés.

GlobalPhone arabe

Le corpus GlobalPhone développé en collaboration avec l'Institut de technologie de Karlsruhe (KIT) a été conçu pour fournir des données de parole lue pour le développement et l'évaluation des grands systèmes de reconnaissance vocale pour les 18 langues les plus répandues dans le monde parmi elles la langue arabe Schultz and Waibel [1998].

- Nombre de locuteurs : 78 locuteurs (35 hommes/43 femmes). Les locuteurs sont Tunisiens, Palestiniens et Jordaniens.
- Age des locuteurs : 20 locuteurs (-19), 35 locuteurs (20-29), 13 locuteurs (30 -39), 6 locuteurs (40 -49), et 4 locuteurs (+ 50).
- Matériel d'Enregistrement : microphone de proximité (Sennheiser 440-6).
- Environnement d'Enregistrement : non mentionné.
- Texte : 100 Phrases du journal Assabah.
- Taille du vocabulaire : non mentionné.
- Total mots : non mentionné.
- Texte/locuteur : non mentionné.
- Total parole : non mentionné.
- Fréquence d'échantillonnage : 16 kHz.
- Codage : 16 bits.
- Informations fournies : âge, genre, profession des locuteurs.

OrienTel Maroc MCA

La base de données Maroc OrienTel MCA (arabe dialectal moderne) est élaboré dans le projet OrienTel et distribué par ELRA [ELRA]. Cette base de données est adaptée pour les applications de reconnaissance vocale automatique.

- Nombre de locuteurs : 772 locuteurs marocains (383 hommes et 389 femmes).
- Age des locuteurs : 381 (16-30), 262 (31-45), 129 (46 -60).
- Enregistrement : réseau téléphonique fixe et mobile.
- Environnement d' Enregistrement : non mentionné.

• Texte : chiffres, des mots, des phrases spontanées.

• Taille du vocabulaire : non mentionné.

• Total mots : non mentionné.

• Texte/locuteur : des chiffres isolés, deux séquences de 10 chiffres isolés, quatre mots phonétiquement riches, neuf phrases phonétiquement riches, etc.

• Total parole : non mentionné.

• Fréquence d'échantillonnage : 8 kHz.

• Codage : 8.

• Informations fournies : chaque fichier est accompagné d'un fichier d'étiquetage ASCII SAM.

OrienTel Tunisie MCA

La base de données Tunisie OrienTel MCA a été faite dans le projet OrienTel et distribué par ELRA [ELRA]. Cette base de données est adaptée pour les applications de reconnaissance vocale automatique.

• Nombre de locuteurs : 792 locuteurs tunisiens (426 hommes /366 femmes).

• Age des locuteurs : 516 locuteurs (16-30), 193 (31-45), 82 (46-60), 1 locuteur de 60 ans.

• Enregistrement : réseau téléphonique fixe et mobile.

• Environnement d'Enregistrement : non mentionné.

• Texte : chiffres, des mots, des phrases, spontanées.

• Taille du vocabulaire : non mentionné.

• Total mots : non mentionné.

• Texte/locuteur : des chiffres isolés, deux séquences de 10 chiffres isolés, quatre mots phonétiquement riches, neuf phrases phonétiquement riches, etc.

• Total parole : non mentionné.

• Taux d'échantillonnage : 8 kHz.

• Codage : 8 bits.

• Informations fournies : chaque fichier est accompagné d'un fichier d'étiquetage ASCII SAM.

OrienTel Egypte MCA

Conçue par [ELRA] pour les mêmes objectifs que OrienTel Tunisie MCA et OrienTel Maroc MCA, OrienTel Egypte MCA est caractérisé par :

• Nombre de locuteurs : 750 locuteurs égyptiens (398 hommes /femmes 352).

• Age des locuteurs : 379 locuteurs (16-30), 291 (31-45), 80 (46-60).

• Pour les autres caractéristiques voir la base OrienTel Tunisie MCA.

OrienTel Emirats Arabes Unis MCA

De même que OrienTel Egypte MCA, la base OrienTel Emirats Arabes Unis MCA est caractérisé par :

- Nombre de locuteurs : 880 locuteurs (432 hommes /448 femmes).
- Age des locuteurs : 488 locuteurs (16 -30), 309 (31-45), 83 (+ 46).
- Les autres caractéristiques sont identiques à OrienTel Tunisie MCA.

OrienTel Jordanie MCA

- Nombre de locuteurs : 757 locuteurs jordaniens (393 hommes /364 femmes).
- Age des locuteurs : 427 locuteurs (16-30), 230 (31-45), 100 locuteurs(46-60).
- Idem que OrienTel Tunisie MCA.

NEMLAR Broadcast News Speech Corpus

NEMLAR Broadcast News Speech Corpus a été développé entre 2002 et 2005 sous par NEMLAR (Network for Euro-Mediterranean Language Resources) et distribué par [ELRA]. Les émissions ont été enregistrées à partir de quatre stations de radio différentes : Medi1, Radio Orient, RMC - Radio Monte Carlo, RTM - Radio Télévision Maroc.

- Nombre de locuteurs : Un nombre inconnu d'arabophones.
- Age des locuteurs : Inconnu.
- Enregistrement : Informations (radio).
- Texte : Informations.
- Vocabulaire : 62 000 mots.
- Total mots : non mentionné.
- Texte/locuteur : non mentionné.
- Total parole : non mentionné.
- Fréquence d'échantillonnage : 16 kHz.
- Codage : 16 bits.
- Informations fournies : lexique de prononciation avec sa transcription phonétique en SAMPA.

Annexe C : Comaparaison de SRAP

[Muhammad et al., 2009]	Small vocabulary , Speaker independent, Isolated digit	MFCC	HMM	Bangia	>95% for digits (0-5) <90% for digits (6-9)
[Alotaibi et al., 2008]	Large vocabulary , Speaker independent, Phonetic/word	MFCC	HMM	Arabic	accuracy for non-native speakers
[Cichocki et al., 2008]	Large vocabulary , Speaker independent, Continuous Speech	MFCC	HMM	Acadian French	89.29%
[Krishnan et al., 2008]	independent, Phonetic/word independent, Isolated word	Discrete Wav-let Transform	ANN	Malayalam	89%
[Lishuang and Zhiyan, 2010]	Large vocabulary Speaker independent vowels	MFCC	Genetic Algo--rithem HMM	Chinese	effective and high accuracy
[AlQatab and Ainon, 2010]		MFCC	HMM	Arabic	97.99%
[Ashraf et al., 2010]	Small vocabulary , Speaker independent, Isolated word	MFCC	HMM	Urdu	Little variation WER WER for new speakers
[Sukumar et al., 2010]	Medium vocabulary, Speaker dependent, Isolated word	DWT	ANN	Malayalam	80%
[Satori et al., 2007]	Middle vocabulary, Speaker inde-pendent, Isolated word (digit)	MFCC	HMM	Arabic	(mean) 90,55%
[Abushariah et al., 2010]	Speaker independent, large Vocabulary, Continuous speech	MFCC	HMM	Arabic	For diffrent sentences and speakers 90.23%

Annexe D : Système phonétique arabe : IPA/SAMPA

SAMPA	IPA	Keyword	Orthography	SAMPA	IPA	Keyword	Orthography
b	**b**	ba:b	باب	G	**ɣ**	Garb	غرب
t	**t**	tis?`	تسع	X\	**ħ** [1]	X\ilm	حلم
d	**d**	da:r	دار	?` (?\)	**ʕ** [1]	?`alam	علم
t`	**tˤ**	t`a:bi?`	طابع	h	**h**	hawa:?	هواء
d`	**dˤ**	d`arab	ضرب	m	**m**	ma:l	مال
k	**k**	kabi:r	كبير	n	**n**	nu:r	نور
?	**ʔ**	?akl	أكل	r	**r**	rima:l	رمال
q	**q**	qalb	قلب	l	**l**	la:	لا
f	**f**	fi:l	فيل	l`	**lˤː**	?al`l`ah	الله
T	**θ**	Tala:T	ثلاث	w	**w**	wa:hid	واحد
D	**ð**	Dakar	ذكر	j	**j**	jawm	يوم
D`	**ðˤ**	D`ala:m	ظلام	i		D`il	ظل
s	**s**	sa?`i:d	سعيد	a		X\al	حل
z	**z**	zami:l	زميل	u		?`umr	عمر
s`	**sˤ**	s`aGi:r	صغير	i:		?`i:d	عيد
S	**ʃ**	Sams	شمس	a:		ma:l	مال
Z	**dʒ**	Zami:l	جميل	u:		fu:l	فول
x	**x**	xit`a:b	خطاب				

Annexe E : Les phrases phonétiquement équilibrées

أبصر ثعبانا ولم يظلمه
أورث لاقطا
ألفت العنان
كال وغبط الكبش
أمتعنا بنغم
لقد كان مسالما وقتل
كان اليمن في شقاء
استلزم العفن وعيكما
أرثته زوجا
غفل عن ضحكاتها
يكن مدركا
أين زوايانا و قانوننا ؟
ألى بالمال وظلمهم
صاد الموروث مدلجا
ناولها الأرض
أرغمتك ضرورة
ويل للجائر
كان في ظلمات و لم يرحل
ألم يستنصر بالقاتل؟
كل عيشك واستحضرلهما قانونا
هذى و رحل
لاح وبيل
سيؤذيهما إذا استبقيتكم
وانطفأ ضوؤهم
ألم يكن معروفا وشغل مذيعا؟
وقد آب الكلب
كانت ثكنتهم في سلام
لازم ظبيانا
بالوالدين إحسانا
ما خدع نمرها
و نجى ولم يستحسن ظلمكم
نهم بسائقنا
ابنك شروب
استقم كما أمرت
كان اللأكل لذيذا
لن يلامس وعلا
هل هار؟
نثرعليهم نهيموا

ضمنت فوزهما
جمع الموزو خلى
أدبر الآبق
أخذتنا في قارب لمال
ورما فلن يقاتلا
نطق صائم
هي هنا لقد آبت
غلا أكلهم
ولاؤها للقلب ولغبائهم
لو لم يشطبهم لقهرنا
أحفظ من الأرض
غفت فمات مختنقا
أين المسافرون؟
أين نذيرتكم؟
طلب رقصة من العروس
أقام العدل و آزر اسما
ألم يستمتع بثنمارها؟
يرثكم بال
سيؤذيهم زماننا
كافئ حذام
كنت قدوة لهم
قا دناف ولم يضطهدكم
ضؤل و لم يركع للواقف
عبد الإله
آزر صائما
إن هاسا
هل لذعته بقول؟
أخطأت فآثر صيدنا
عرف واليا و قائدا
خسرت ولائما
خلا بالنا منكما
نهم بسائقنا
حضر الوغل
طفح الكيل
لم يجاهد
لان و لم يكن شرسا
إبنكم فاروق ويظلم
ضمن ثروتهم
ثأرت منهما

أسعد الآل بالقذيف
أظفر بما أخذت
هنا لاط الهيمان
أتؤذيها بآلامهم؟
قاول في مئذنة
قابلنا وبرا
شربت من غروف
هل جاع أب؟
ذهب إلى تلمسان
لا ان يذيع الخبر
وضعت حملها
أكمل بالإسلام رسالتك
لا لن يذيقكم من أكلنا
سقطت إبرة
وقبل ظلمهم
من لم ينتفع؟
قابله يوما ما في مكانهما
و لماذا نشف مالهم؟
إستدركنا الجبان
نبه آباؤكم
باع طبلا فخرج مسرورا
قم و أظهره
نوى الأعزل
أصابكم بقذيفة
لسع وريا
لاث مدمن
ذهب صبانا
لن و لن ينالها
سعيكم مقرون بالعناء
ساح الماء لظمآن
أسقطت إنجازكم
واستبشع سوطكم
أبلغه بالأمر
رفع يديه للبارئ في منام
خرشت المرآة
هل كان يقابلكما؟
لاين ليثا

وقاك المذيم!
استغفر لذنبك
كان واهنا و لا حظ وألا
ها أنا ذي أنصت ليقل
سال مالهم
ما لبس ثوبا
فهل كانت حلب إمارة؟
كان منهم في ظلمات
رأيت قذيفا
أكرمه واعمل!
علا ولستعظم الإنسان
لا يا زينب شرم حبلا!
لماذا أخطأ آدم؟
علا و جار الإنسان
سيشكروننا في ندوة
وهل علت؟
قابلهم واصبر
سيسقط مؤامرتك
لا لم يبق زوجها
بعثت نذيرا
وأل يرث ملكنا
اذهب بأمان!
كن هنا!
صنع مدفأة ومعزفا
حفظت القرآن
لو لا أن مرضنا لخسروا
قد نضبطهم
لم يكتمه
سعيهم لأبنائهم
كان موتها صعبا بسبيلكم
زعمنا أن لن يشترك
ما لازم مغرورا و ما ورث
يقامرون بالمال
لم يستبح غدرهم
ناظر المذيع
ناهض المدمن وحشا
لن يستأمنها أبدا
علا صخب بالجامعة

هلك شوك
سار القائم
وي ! لنت فوافتنا !
لا تكن شرِسا
آثم الأبناءِ وزوجها
فسِدت ذات بينهم
كان صائما
وكع أبِوهما
ما قولك في ظلمهِم ؟
استمع للآذان
تدركهم غدا بيمامة
كن رذيلا
سيبعث الأِب آلهم
جالا وهلكا بالبأساء
لانت عناها
أين نام واقفا ؟
خزن القارورة وفرشا
نتن الأكل
جاء بضمانات لهروب منك
أخف والدها
لا ننس لوعظ والوصايا
خلق الإنسان من نطفة
وضع مصطلحا
ماذا يذيب؟
سمعنا قذيعة
أرسلهم لأوليائهم
لاينت من وفد إليكم
قتل وارث
وبماذا لاغزِكما؟
أصابتها بالكنف
قايظ آلهم
هاودتنا وكافأت العذير
زمن ولم ينتفع ببلسمهم
أبرم لنا أمورهم
وهتِ فآواها
لا تكابد ! لن يظلم سائقا
دخلوا المروج
كن صائغا

أسرونا بمنعطف
أخفق متآمرون
استقم كَما أُمرت
ضمنت شغفكم
لن يلامس علا
أخذ إجازة
ولى ولم نعقله
رفض الفدية فما لها من عودة

Annexe F : Corpus ALGASD

1. Fiches de renseignements

Fiche 1-1

Code de la Région		
N° de Loc.	Nom	Prénom
1	Nom 1	Prénom 1
2	Nom 2	Prénom 2
3	Nom 3	Prénom 3
.	.	.
n	Nom n	Prénom n

Fiche 1-2

ID	Sexe	DatNais	NIns	Corpus lu	DatEnregis
BFX0	M	05/01/1979	NH	Cc/Ci123	10/03/2008
GAX0	M	30/01/1989	NM	Cc/Ci13	10/03/2008
...	.	...	.	...	....

Fiche de Renseignement 2

Région:	2		
N° de la fiche:	28		
Code:	ALGASD/R2/CA/m BFX0		
Identification	m BFX0		
Sexe	M		
Date de Naissance	17/11/1989		
Niveau d'instruction	NM		
Corpus à lire	Cc	Cc1:	
		qAdanA wa-lam ya.d.tahidkum	
		Date enregistrement	11/03/2008
		Cc2:	
		a_h.ta'ta fa-'A_tara .saydanA	
		Date enregistrement	11/03/2008

2. Caractéristiques des microphones

Désignation	SHURE 58
Type	Dynamique
Modèle de Polarité	Cartoide (Unidirectionnel)
Bande passante	50 Hz à 15 000 Hz
Niveau de sortie (à 1000 Hz) :	-54.5 dB V/Pa (1.85 mV) 1 Pa = 94 dB SPL
Impédance	Impédance nominale 150 Ω (300 Ωeffective)
Connecteur	XLR mâle
Corps	Acier moulé avec grille sphérique en acier
Poids net	298 g
Dimensions	162 mm 51 mm

Désignation	PHILIPS DM-109
Type	Dynamique
Modèle de Polarité	Unidirectionnel
Bande passante	50 Hz à 15 000 Hz
Niveau de sortie (à 1000 Hz) :	-73 dB V/Pa (1.85 mV) 1 Pa = 94 dB SPL
Impédance de sortie	600 Ω ± 20 % à 1 kHz
Poids net	250 g
Dimensions	170 mm 53 mm

Annexe I : Outils de développement de SRAP

Sphinx

Le projet Sphinx [Walker et al., 2004] est développé conjointement par *Carnegie Mellon University*, SUN *Microsystems Laboratories*, *Cambridge Research Lab* de Hewlett-Packard et *Mitsubishi Electric Research Labs*. Il est lancé dans le but de concevoir un environnement pour la recherche dans le domaine de la reconnaissance automatique de la parole. CMU Sphinx 4 est une librairie de classes et d'outils disponibles en langage de programmation Java. Sphinx offre la possibilité de mettre en œuvre des systèmes de reconnaissance à grand vocabulaire, indépendants du locuteur, et traitant de la parole continue.

Julius / Julian

Julius 34 est un système open-source de reconnaissance de la parole continue dédié à la recherche et au développement [Lee et al., 2001]. Basé sur des N-grams de mots et des HMM dépendants du contexte, il permet des décodages quasi-temps réel sur les ordinateurs actuels pour des tâches de dictée vocale. Depuis la version 3.4, Julian, un parser pour une reconnaissance basée grammaire a été intégré au système Julius. Julian est une version modifiée de Julius qui utilisent des grammaires à états finis comme modèles de langage.