REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

Ecole Nationale Supérieure Polytechnique

Département d'Electronique

Thèse de Doctorat en Electronique

Présentée par

CHENTIR Amina
Chargée de Cours au Département d'Electronique – USD-Blida

THEME

Etude de la Microprosodie en vue de la Synthèse de la parole en Arabe Standard

Soutenue le : 01 Octobre 2009

Devant le jury composé de :

M^{me} Latifa HAMAMI	Professeur à l'ENSP	Présidente
M^{me} Mhania GUERTI	Professeur à l'ENSP	Rapporteur
M Daniel J. HIRST	Directeur de recherche (LPL, France)	Co-Rapporteur
M^{me} Malika TALHA	Maître de Conférence à l'USTHB	Examinateurs
M Halim SAYOUD	Maître de Conférence à l'USTHB	
M Hocine TEFFAHI	Maître de Conférence à l'USTHB	

Je dédie cet événement marquant de ma vie à la mémoire de mon père disparu trop tôt pour moi. J'espère que, du monde qui est le sien maintenant, il apprécie cet humble geste comme preuve de reconnaissance de la part de sa fille qui a toujours prié pour le salut de son âme. Puisse Dieu, le tout puissant, l'avoir en sa sainte miséricorde !

A ma mère, à celle qui est toujours présente et continue de l'être pour faire mon bonheur. Merci pour t'être sacrifiée pour que tes enfants grandissent et prospèrent. Enfin ! Merci tout simplement d'être… ma mère.

ملخص

يقوم عملنا علي تطبيق البروزوديا من أجل تحسين جودة الكلام الإصطناعي من نص الخطاب في اللغة العربية الفصحى. لذلك، قمنا بتطوير 36 جملة مفيدة، متكونة من جميع الأصوات الأساسية للغة العربية الفصحى. تمّ التسجيل داخل غرفة صماء بأسلوب متعدد المتحدثين (رجلين و ثلاثة نساء). بعد هذه الخطوة، قمنا بإستعمال برنامج للتحليل اللفظي و النسخي. هذا الأخير، سمح لنا بالقيام بعملية تحليل طيفي لمجمل الأصوات المزمع دراستها، لنسخها بغرض تقسيمها و موائمتها بطريقة شبه آلية من أجل توليد البروزوديا تلقائيا، قمنا بإقتراح طريقتين :

- الأولى، علي أساس تصنيف بطريقة التحليل التمييزي لعنصر البروزوديا الطاقة. بإستعمال كلمات ثلاثية، قمنا بإستخراج النبرات المختلفة (الأولية، الثانوية و الثلاثية) حيث تمّ تعيين النبرة الأولية بنسبة 78% . هذه النتيجة تدل علي فعالية هذه الطريقة اللتي يمكن أن تعزز الأساليب اللتي تستند فقط علي معيار التردد الأساسي . في الأخير، تمّ إقتراح نسختين محسّنتين ، اللتين أفرجتا عن نسبة تعيين النبرة الأولية بقيمة 85 %؛
- الثانية تعتبر طريقة جديدة، تسمح بإستخراج تلقائي للمعلومة الميكروبروزودية للإشارة الكلامية بإستخدام المنحنى الأصلي للتردد الأساسي و المنحنى المتحصل عليه عن طريق برمجية مومال. النتائج المحصل عليها، تعزز فكرة أنّ التأثير الميكروبروزودي موجود حقا و أنّ إنخفاض نسبي لمنحنى مومال، لكل حرف، يكون كافيا لتوفير تحسين طبيعية الكلام الإصطناعي.

الكلمات الرئيسية: البروزوديا، توليف اللغة، العربية الفصحى، النبرة، التحليل التمييزي، ميكروبروزوديا .

RÉSUMÉ

Notre travail consiste en l'application de la prosodie en vue d'améliorer la qualité de la parole synthétique à partir du texte (Text-To Speech : TTS) de l'Arabe Standard (AS).

Pour cela, nous avons élaboré un corpus de 36 phrases affirmatives, comprenant tous les phonèmes de l'AS. Un enregistrement dans une chambre sourde, a été fait en mode multilocuteur (2 hommes et 3 femmes). Après cette étape, nous avons utilisé un logiciel d'analyse et de transcription phonétique, PRAAT. Ce dernier nous a permis de faire une analyse sonagraphique de tous les phonèmes à étudier, de les transcrire afin de les segmenter et de les aligner d'une manière semi-automatique.

Pour générer automatiquement la prosodie, nous avons proposé deux méthodes :

- la première est basée sur une Classification par Analyse Discriminante (CAD) du paramètre prosodique énergie. A partir de mots tri-syllabiques, nous avons extrait les différents Accents (Primaires, Secondaires et Tertiaires). Un pourcentage de détection de l'Accent Primaire (A_cP) de 78% a été obtenu. Ce résultat montre l'efficacité de cette approche qui pourra renforcer les méthodes basées uniquement sur le critère du fondamental. Deux versions améliorées sont ensuite proposées, donnant un pourcentage de détection de l'A_cP égal à 85 % ;
- la seconde est une nouvelle approche, permettant, le plus possible, une extraction automatique de l'information micromélodique du signal de parole à l'aide de la courbe issue de la fréquence fondamentale et de sa courbe macromélodique obtenue à l'aide de l'algorithme de MODelisation MELodique (MOMEL). Les résultats obtenus viennent renforcer l'idée que l'effet microprosodique existe bien et qu'au niveau micromélodique, un abaissement relatif à chaque consonne voisée de la courbe macromélodique s'avèrera suffisant pour apporter une amélioration au naturel de la parole synthétique.

Mots clés : Prosodie, Synthèse de la parole à partir du texte, Arabe Standard, Accent, Analyse Discriminante, Micromélodie.

ABSTRACT

Our work is about the application of prosody to improve the quality of synthetic speech from Standard Arabic (SA) text. (Text To Speech: TTS). To do this, we developed a corpus of 36 affirmative sentences, including all the phonemes of the SA. A record in a deaf room was made in a multi-speaker mode (2 men and 3 women). After this step, we used the software analysis and phonetic transcription, PRAAT. This allowed us to do an acoustical analyze for all phonemes to study, to transcribe them in order to segment and align them in a semi-automatic way.

To generate prosody automatically, we have proposed two methods :

- the first is based on a Classification by Discriminant Analysis (CDA) setting on the prosodic parameter energy. From tri-syllabic words, we extract different Accents (primary, secondary and tertiary). A percentage of detection of the Primary Accent of 78% was obtained. This result shows the efficiency of such an approach which could reinforce existing methods based exclusively on fundamental frequency. Two improved versions are then proposed, giving a percentage of detection of the Primary Accent equal to 85 %;
- the second is a new approach, allowing as much as possible, an automatic extraction of the micromelodic information of the speech signal using the original curved of the fundamental frequency and its macromelodic curve obtained using the algorithm of Modelling Melody (MOMEL). The results reinforce the idea that the microprosodic effect exists and seems to be possible to be included in a prosodic generating unit by a simple model to provide improved natural in speech synthesis.

Key words: Prosody, Text-To-Speech synthesis, Standard Arabic, Accent, Discriminant Analysis, Micromelody.

TABLE DES MATIERES

REMERCIEMENTS

Qu'il me soit permis de présenter ici mes remerciements à tout un petit monde de personnes qui ont rendu possible la présente étude et qui ont contribué à son élaboration sous quelque forme que ce soit.

Je tiens tout d'abord à dire ma reconnaissance envers Madame le Professeur Mhania GUERTI qui a accepté sans réserve, de diriger cette thèse. Ses conseils, sa patience, sa grande disponibilité et sa gentillesse ne sont que quelques-unes de ses nombreuses qualités. Je la remercie aussi pour le soin qu'elle a apporté à la lecture de mon manuscrit, pour ses remarques et conseils qui ont contribué à son amélioration. Je lui suis reconnaissante de m'avoir si bien dirigée tout au long de cette recherche et de m'avoir initiée à la prosodie. J'aimerais qu'elle sache que ma reconnaissance va bien au-delà de ces quelques lignes.

Ma gratitude va particulièrement à Monsieur Daniel HIRST, Directeur de recherche au CNRS qui, malgré les prérogatives qui sont les siennes, a accepté sans réserve, de co-diriger cette thèse. Qu'il accepte mes vifs remerciements pour toutes les facilités qu'il a mises à ma disposition tout au long de mon séjour au LPL. Je lui suis reconnaissante de m'avoir fait profiter pleinement des multiples facettes de ses compétences scientifiques. Je le remercie également pour sa disponibilité, sa simplicité, son humour et son souci incessant pour le bien et la satisfaction de ses étudiants. C'est un privilège et un honneur mais surtout un grand bonheur que de travailler avec lui.

Je remercie le Laboratoire Parole et Langage (LPL) d'Aix En Provence, dans son ensemble et dans son unité, pour son accueil et son accompagnement, et individuellement, tous ses membres qui m'ont chacun à leur manière, aidée et soutenue à un moment ou un autre durant mon séjour.

Je tiens à remercier Madame Latifa HAMAMI, Professeur à l'ENSP, d'avoir accepté de présider le jury de cette thèse.

J'adresse aussi mes remerciements aux membres du jury, Madame Malika TALHA, Messieurs Halim SAYOUD et Hocine TEFFAHI, Maîtres de Conférences à l'Université des Sciences et de la Technologie Houari Boumediene (USTHB) d'Alger, pour avoir bien voulu examiner et juger ce travail.

Mes remerciements vont également aux enseignants du département d'Electronique de l'USD de Blida et à l'ensemble des personnes qui m'ont accompagnée, de près ou de loin, durant ces années et qui ont contribué, directement ou non, à l'aboutissement de ce travail. Un coucou particulier à ma meilleure amie N. Kahina.

Et j'en viens à ma famille. A ma sœur chérie, à l'unique sœur que j'ai au monde, Khadîdja, à mes trois frères, Mohamed Chérif, Abd El Hamid et Mahmoud. Merci d'être toujours à mes côtés, par votre présence, par votre aide, par votre amour, pour donner du goût et du sens à notre vie de famille. Merci aussi à mes adorables nièces et neveux. Merci de remplir ma vie de joie et de bonheur.

Enfin, j'espère du fond du coeur que tout ce petit monde, mon monde à moi, trouve ici un mot de reconnaissance, et que chacun se reconnaisse en ce qui le concerne. J'espère aussi que l'effort déployé dans le présent travail réponde aux attentes des uns et des autres.

LISTE DES ABREVIATIONS

Accent Primaire	A_cP
Accent Secondaire	A_cS
Accent Faible ou Tertiaire	A_cF ou A_cT
Analyse Discriminante	AD
Alphabet Phonétique International	API
Average Magnitude Difference Function	AMDF
Arabe Standard	AS
Codebook Excited Linear Prediction	CELP
Court Terme	CT
Classification par Analyse Discriminante	CAD
Durée	D
Énergie	E
Equivalent Rectangular Bandwidth	ERB
Extraction de l'Effet Microprosodique	EEM
Fréquence fondamentale	F_0
Frequency Domain Pitch Synchronous OverLap-Add	FD-PSOLA
Groupe-Inter-Perceptuel-Center	GIPC
Hidden Markov Model	HMM
Intensité	I
Linear Prediction	LP
Linear Prediction Coding	LPC
Linear Prediction Pitch Synchronous OverLap-Add	LP-PSOLA
Long-Term Average Spectrum	LTAS
MOdélisation MELodique	MOMEL
Multi-Band Excitation	MBE
Multi-Band Re-synthesis PSOLA	MBR-PSOLA
Optimality Theory	OT
Outils d'Enseignement Assisté par Ordinateur	OEAO
OverLapp-Add	OLA
Pitch Synchronous OverLap and Add	PSOLA
Reconnaissance Automatique de la Parole	RAP
Reconnaissance Optique	RO
Residual Excited Linear Prediction	RELP
Text-To-Speech	TTS
Traitement Automatique	TA
Time Domain Pitch Synchronous OvelLap-Add	TD-PSOLA
Traitement Automatique du Langage Ecrit	TALE
Traitement Automatique du Langage Naturel	TALN
Traitement Automatique du Langage Parlé	TALP
Traitement Automatique de la Parole	TAP
Transformée de Fourier Discrète	TFD
Transformée de Fourier Inverse	TFI
Transcription Orthographique Phonétique	TOP
Voice Operating DEmonstratoR	VODER

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GENERALE

Le Traitement Automatique de la Parole (TAP) est un domaine de recherche actif, au croisement du traitement du signal numérique et du traitement symbolique du langage. Depuis les années 60, il bénéficie d'efforts de recherche très importants, liés au développement des moyens et techniques de télécommunications et du traitement numérique de l'information. Ces efforts se sont concrétisés grâce à plusieurs applications du TAP, telles que le codage, la Reconnaissance Automatique et la synthèse de la parole. Malgré les avancées réalisées ces dernières années dans ces domaines, des progrès restent à faire pour accroître le confort d'utilisation des systèmes actuels. Ce confort se manifeste essentiellement dans le besoin de manipuler des systèmes de communication de plus en plus conviviaux.

La prosodie concerne la musique de la parole, son effet sonore, les modulations de la voix. Elle est ainsi essentielle à la compréhension et au naturel de la parole, et par conséquent à la synthèse vocale. La prosodie est universelle mais est aussi spécifique à une langue. De plus, il n'existe pas qu'une seule prosodie.

L'organisation prosodique s'articule autour des structures syntaxique, sémantique (relative au sens de l'énoncé) et pragmatique (qui regroupe les informations relatives au contexte particulier de production des actes de parole). Elle repose sur un découpage en groupes prosodiques réalisé à partir d'une analyse syntaxique. La hiérarchie prosodique est généralement fonction de la relation de dépendance existant entre les groupes syntaxiques. Une analyse morphologique et la prise en compte de contraintes rythmiques est aussi nécessaire. Il existe enfin des contraintes extra-linguistiques qui influent sur les paramètres prosodiques : les contraintes phonotactiques d'une part qui sont relatives au nombre de syllabes par mot, par groupe syntaxique et par phrase; enfin, le contexte d'élocution, l'émotivité du sujet, la constitution physiologique des organes de production (liés principalement au sexe et à l'âge du locuteur), l'origine régionale et sociale du locuteur seront autant de paramètres qu'il faudra prendre en compte dans la structure prosodique de l'énoncé.

Les trois paramètres prosodiques classiquement extraits du signal acoustique sont l'énergie, la durée et la fréquence fondamentale (ou F_0). Dans la communication quotidienne, la prosodie joue un rôle irremplaçable pour la compréhension d'un message sonore. La réalisation différente du groupement rythmique, de l'accent, de la courbe mélodique et d'autres phénomènes, évoquerait un sens différent du message.

L'amélioration du naturel de la parole de synthèse repose sur le développement d'un modèle prosodique. Ce dernier doit être capable de reproduire, à partir d'un texte écrit, les phénomènes acoustiques impliqués dans la production de la parole naturelle. Ceci nécessite d'une part un formalisme linguistique qui interprète la prosodie d'un texte écrit et, d'autre part, d'un modèle phonétique qui permet la quantification de la prosodie pour qu'elle soit exploitable par le système de synthèse. Pour cela, la prosodie est essentielle à la compréhension et au naturel de la parole, et est donc indispensable pour un système de synthèse vocale.

Par ses propriétés morphologiques, syntaxiques, phonétiques et phonologiques, la langue Arabe est considérée comme faisant partie des langues difficiles à appréhender dans le domaine du Traitement Automatique du Langage Ecrit et Parlé (TALE et TALP). Dans le domaine du TALE de l'Arabe Standard (AS), les recherches ont débuté vers les années 1970, avant même que les problèmes d'édition de textes arabes ne soient complètement maîtrisés. Les premiers travaux concernaient notamment les lexiques et la morphologie. Depuis une dizaine d'années, l'internationalisation du Web et la prolifération des moyens de communication en langue arabe, ont révélé un grand nombre d'applications du Traitement Automatique du Langage Naturel (TALN) arabe. Les travaux de recherche ont ainsi commencé à aborder des problématiques plus variées comme la syntaxe, la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc.

Dans le cadre de cette thèse, nous avons utilisé un corpus de phrases affirmatives en AS, prononcées par 5 locuteurs arabophones, enregistrées dans une chambre sourde. Ces phrases ont alors subi une analyse sonagraphique grâce au logiciel de transcription et d'analyse phonétique PRAAT. Elles ont ensuite, été segmentées et alignées semi-automatiquement en phonèmes et à la fin, une Transcription Orthographique Phonétique (TOP) a été faite.

Dans le but de l'amélioration de l'intelligibilité et du naturel de systèmes de synthèse de l'AS, nous avons proposé deux méthodes :

- la première consiste en la détection de l'Accent Primaire en AS, en exploitant le paramètre acoustique énergie, grâce à une Classification par Analyse Discriminante (CAD) ;
- la deuxième est une nouvelle approche pour l'extraction de l'effet microprosodique en AS, basée sur l'exploitation de la courbe mélodique réelle et la courbe macromélodique obtenue grâce à l'algorithme de stylisation MOMEL.

Cette thèse est organisée en cinq chapitres :

- le premier est consacré à la présentation des principes du Traitement Automatique de la Parole (TAP) et celui de l'Arabe Standard (AS). Une étude sur les principaux paramètres spécifiques à l'AS est exposée avec les problèmes rencontrés en Traitement Automatique ;
- le deuxième s'articule autour des principes de la synthèse vocale, des différents systèmes présents dans l'état de l'art, suivis d'une description détaillée des diverses techniques et méthodes utilisées ;
- le troisième aborde la thématique de la prosodie dans le contexte de la synthèse de la parole à partir du texte (Text-To-Speech : TTS) ;
- le quatrième introduit tout d'abord les notions de syllabe et d'accent en AS, suivies du principe de l'Analyse Discriminante. Une description détaillée de la méthode utilisée pour la détection de l'Accent Primaire à l'aide d'une Classification par Analyse Discriminante du paramètre énergie est développée. Deux versions améliorées sont proposées et les résultats obtenus sont présentés et commentés ;
- le cinquième est consacré à la présentation de la méthodologie suivie par l'extraction de l'effet microprosodique à partir de la courbe réelle de la fréquence fondamentale et de la courbe de modélisation mélodique obtenue grâce à l'application de l'algorithme MOMEL. Les résultats obtenus sont exposés et discutés à la fin du chapitre ;
- en dernier lieu, nous présentons des conclusions générales et exposons quelques perspectives pour la continuité et l'amélioration de ce travail de recherche.

CHAPITRE 1 :

GENERALITES SUR LA PAROLE

ET L'ARABE STANDARD (AS)

1.1. INTRODUCTION

Ce premier chapitre a pour but d'une part de présenter les connaissances essentielles qui décrivent les natures physiologiques et phonétiques de la parole, quelques méthodes efficaces d'analyse et de modélisation du signal de parole présentes dans l'état de l'art, et les spécificités de la langue Arabe Standard (AS).

1.2. GENERALITES SUR LA PAROLE

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

La parole est une faculté, propre à l'homme, de communication par des sons articulés.
Elle met en jeu des phénomènes de natures très différentes et peut être analysée de bien des façons. On distingue généralement plusieurs niveaux de description non exclusifs : physiologique, phonologique, phonétique, acoustique, morphologique, syntaxique, sémantique, et pragmatique. Nous survolons dans ce chapitre les quatre premiers niveaux qui sont les niveaux les plus concernés par notre étude.

1.2.1. Niveau physiologique

Les sons de la parole se produisent lors de la phase d'expiration au cours de laquelle un flux d'air contrôlé, en provenance des poumons passe à travers le larynx et le conduit vocal (conduit respiratoire). Ce flux d'air appelé *air pulmonaire* rencontre sur son passage plusieurs obstacles potentiels qui vont le modifier de manière plus ou moins importante. La figure 1.1 représente une vue globale de l'appareil phonatoire.

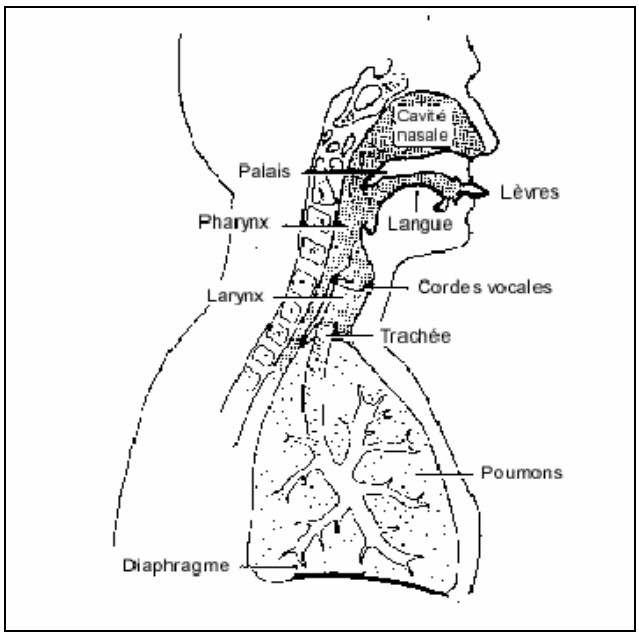


Figure 1.1 : Système vocal humain [1]

L'appareil vocal humain (figure 1.2) peut se présenter idéalement comme un système source - filtre avec notre poumon comme réservoir énergétique [2] :

- l'air sort des poumons et s'écoule dans le conduit vocal. Le son est produit lorsque le souffle passant au travers des cordes vocales les fait vibrer et est ainsi modulé par leur vibration ;
- le conduit vocal couvre le secteur de pharynx ainsi que les sinus nasal et buccal. Il représente un secteur de résonance, de sorte que le son rayonné au niveau des lèvres est le résultat d'un filtrage du signal généré au niveau de cordes vocales.

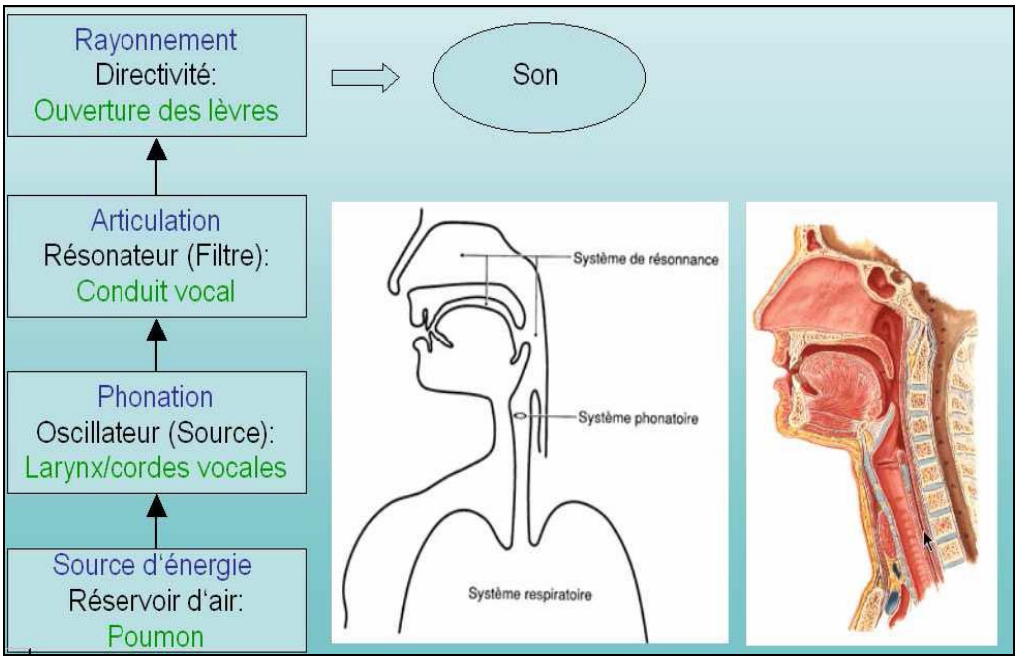


Figure 1.2 : Schématisation de l'appareil phonatoire [2]

A l'intérieur du larynx (figure 1.3) se situent les cordes vocales, organes vibratoires constitués de tissu musculaire et de tissu conjonctif résistant. Les cordes vocales sont reliées à l'avant au cartilage thyroïdien. Elles peuvent s'écarter ou s'accoler pour produire des ondes de pression. L'espace entre les cordes vocales est appelé glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non voisés (ou sourds[1]). Les sons voisés (ou sonores) résultent au contraire d'une vibration périodique des cordes vocales [3].

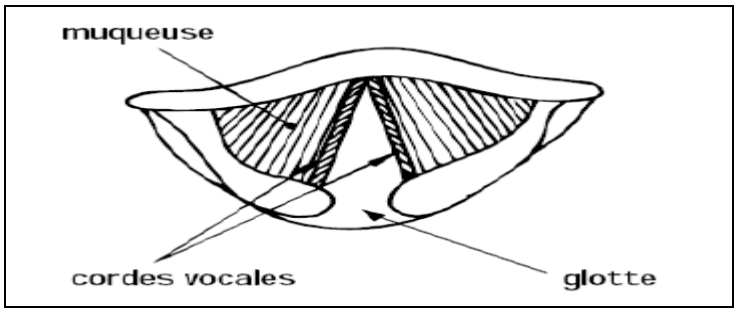


Figure 1.3 : Section du larynx [3]

1.2.2. Niveaux phonétique et phonologique

La phonétique et la phonologie sont deux branches de la linguistique qui interprètent le même matériau : la parole. La phonétique étudie les sons des langues du monde en tant que réalité physique (production, transmission et perception de ces sons), tandis que la phonologie recherche les principes qui régissent leur apparition et leur fonction de codage d'une langue particulière [3-4].

Autrement dit, la **phonétique** est l'étude scientifique des sons du langage humain. Elle exclut les autres sons produits par les êtres humains, même s'ils servent parfois à communiquer (les toux, les raclements de gorge). Elle exclut aussi les sons non humains. Elle se divise en trois domaines :

- la phonétique articulatoire s'occupe de l'activité des cordes vocales, de la bouche, etc. qui rendent possible la parole. Par exemple, nous savons que pour faire un [p] en Français, il faut mettre les deux lèvres ensemble, sortir un peu d'air des poumons, et ensuite ouvrir les lèvres ;
- la phonétique acoustique examine les caractéristiques sonores des sons du langage. Par exemple, nous savons que le son produit par la consonne [s] (exemple : sou [su]) en Français a une fréquence plus élevée que le son produit par une consonne comme [ʃ] (exemple : chou [ʃu]) ;
- la phonétique auditive examine les phénomènes de perception des sons du langage par les êtres humains. Par exemple, qu'est-ce qui nous permet de saisir une syllabe accentuée? Est-ce la durée, la force, la fréquence fondamentale ou une combinaison des trois?

[1] Les phonéticiens appellent sonore ou sourd ce que les ingénieurs qualifient de voisé ou non voisé.

Chaque langue retient pour son fonctionnement un ensemble de sons, parmi ceux que pourrait produire l'appareil vocal. Les plus petites unités sonores distinctives utilisées dans une langue donnée sont appelées phonèmes. Le **phonème** est la plus petite unité sonore qui, substituée à une autre, change le contenue linguistique d'un énoncé. Par exemple changer le premier son [p] de ″peau″ [po] en [b] aboutit à un mot différent : ″beau″ [bo]. On distingue donc les phonèmes [p] et [b].

L'ensemble de phonèmes généralement adopté pour une langue donnée sont regroupés par un système de transcription phonétique utilisé par les linguistes, représenté par l'Alphabet Phonétique International (API). Les phonéticiens regroupent les sons de parole en deux grandes classes phonétiques en fonction de leur mode articulatoire : les voyelles et les consonnes.

Les voyelles correspondent à une vibration périodique des cordes vocales et à une configuration stable du conduit vocal. Selon que la dérivation nasale est ouverte ou non (grâce à l'abaissement du voile du palais), les voyelles sont nasales ou sont orales. Les semi-voyelles sont produites lorsque l'excitation glottique périodique s'accompagne d'une évolution rapide du conduit vocal, entre deux positions vocaliques.

Contrairement aux voyelles, les consonnes sont produites lorsque le passage de l'air venant des poumons est partiellement ou totalement obstrué. Autrement dit, les consonnes correspondent à des mouvements rapides de constriction des organes articulateurs, donc souvent à des sons peu stables, qui évoluent dans le temps. Pour les fricatives, une constriction forte du conduit vocal provoque un bruit de friction. Les cordes vocales peuvent entrer en vibration en même temps que le bruit de friction, la fricative est alors voisée (ou sonore), ou laisser passer l'air sans émettre de son, la fricative est alors non voisée (ou sourde). Les plosives sont des occlusions complètes du conduit vocal, suivies d'un relâchement. Jointe à la vibration des cordes vocales, la plosive est voisée, sinon elle est sourde. Si la dérivation nasale est ouverte pendant la fermeture de la bouche, une nasale est produite. Les semi-voyelles sont des consonnes voisées, mouvements rapides qui passent par la position articulatoire d'une voyelle brève. Enfin, les liquides résultent d'une excitation voisée et de rapides mouvements articulatoires, principalement de la langue [5].

1.2.3. Niveau acoustique

La phonétique acoustique étudie le signal de parole en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur). De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut

alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques qui sont liés à sa production :

- la fréquence fondamentale (F_0) qui correspond à la fréquence du cycle d'ouverture/fermeture des cordes vocales ;
- L'énergie ou l'intensité (I) du son qui est liée à la pression de l'air en amont du larynx ;
- son spectre qui résulte du filtrage dynamique du signal en provenance du larynx (signal glottique) par le conduit vocal qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses.

Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre [6].

Les modèles les plus classiques de représentation du signal de parole s'inspirent du mode de production de type source –filtre (Figure 1.4). Le modèle est divisé en trois parties, la source (le voisement, la friction), le filtre (simulation des effets filtrants des conduits oral et nasal), et la radiation aux lèvres [7].

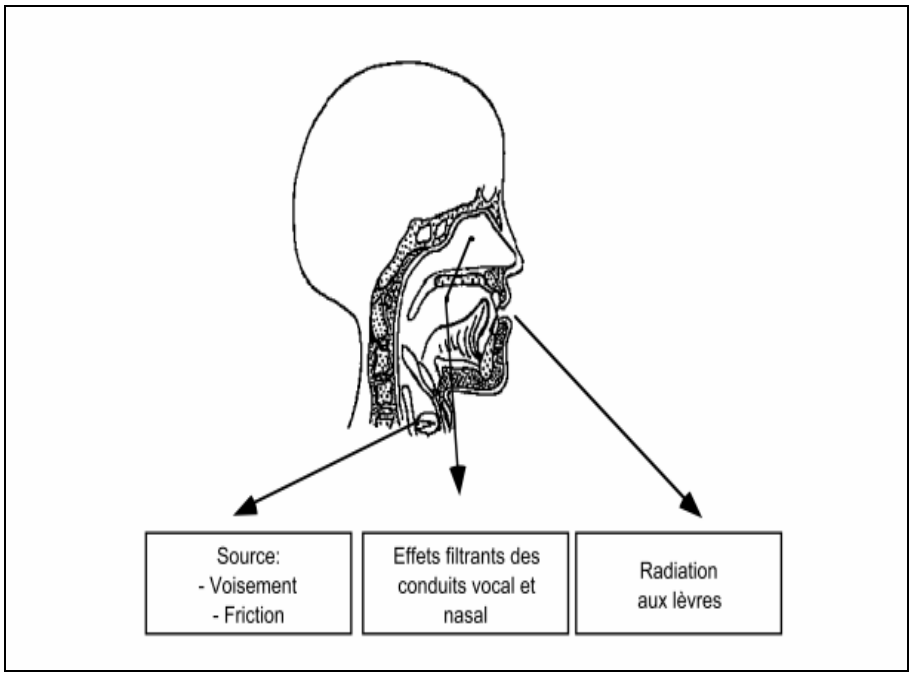


Figure 1.4 : Conceptualisation fondamentale du modèle source - filtre [7].

Le signal de source résulte de la production d'une onde acoustique au niveau de la glotte. Cette onde passe ensuite dans le conduit vocal (oral, nasal) et subit l'effet de radiation des lèvres. Les transformations du signal de source par ces différents organes peuvent être modélisées par un simple filtrage linéaire (Figure 1.5).

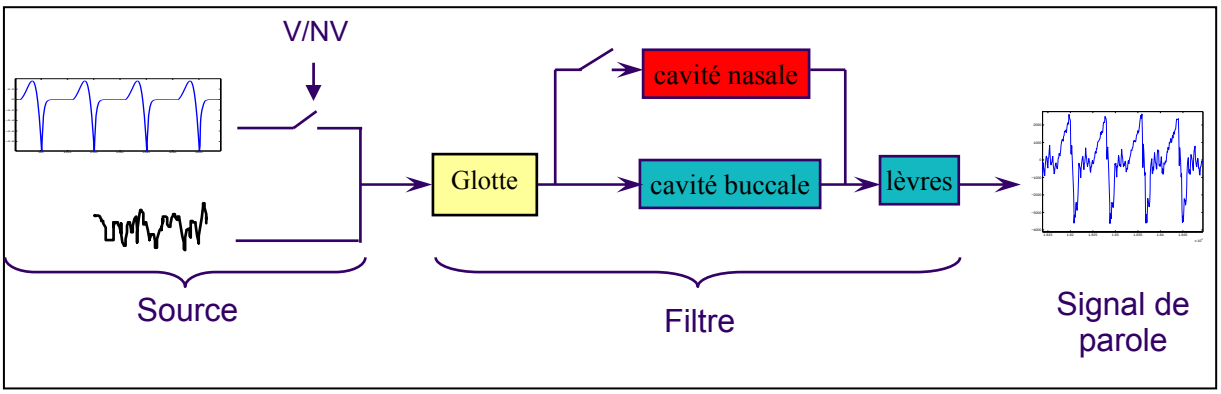


Figure 1.5 : Modèle source - filtre [9]
(V = Voisement et NV = Bruit (aspiration, frication, explosion))

Les caractéristiques acoustiques des cavités supra -glottiques peuvent être modélisées à l'aide d'un filtre linéaire AR (AutoRégressif) dont la fonction de transfert s'exprime comme suit :

$$H(z)=\frac{1}{1+\sum_{i=1}^{p} a_i z^{-i}}=\frac{1}{A(z)} \qquad (1.1)$$

où les a_i sont les coefficients de prédiction du filtre.

Pour une parole intelligible, le nombre de coefficients a_i est fixé de telle façon que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants des segments voisés.

1.3. ANALYSE ET MODELISATION DU SIGNAL DE PAROLE

L'étude de l'évolution temporelle et fréquentielle d'un signal de parole permet de mettre en évidence les caractéristiques de ce signal. Cet objectif est atteint grâce aux méthodes modernes de traitement du signal qui permettent de calculer par exemple, la transformée de Fourier d'un signal de parole pour déduire son spectre de puissance à court terme, et son spectrogramme qui représente l'évolution temporelle de ce spectre.

La quasi-stationnarité par morceaux du signal de parole est une hypothèse généralement admise qui permet de mettre en œuvre des méthodes efficaces d'analyse et de modélisation du signal stationnaire. Ces méthodes sont utilisées pour le traitement à court terme du signal de parole. Le traitement à long terme est quant à lui, assuré par le décalage temporel sur le signal, de la fenêtre de traitement à court terme. Le signal de parole est ainsi progressivement analysé ou modélisé, sur des fenêtres du signal de durée généralement comprise entre 20 à 30 ms, avec un recouvrement entre ces fenêtres qui assure la continuité temporelle des caractéristiques de l'analyse ou du modèle.

1.3.1. Analyse spectrale du signal de parole

L'analyse spectrale du signal de parole présente un intérêt majeur pour caractériser les propriétés fréquentielles des segments phonétiques de la parole. Cette analyse passe par la transformée de Fourier discrète à court terme de ce signal et qui comporte principalement les étapes suivantes [10-11] :

- un segment de 20 à 30 ms de parole est extrait du signal. Ce segment, appelé trame acoustique, est constitué de N échantillons de parole {s(0),…,s(N-1)} ;
- pour atténuer les distorsions spectrales introduites par l'extraction de la trame du signal de parole, on pondère les échantillons de cette trame par une fonction (fenêtre) de pondération. La fenêtre de Hamming par exemple est définie comme suit :

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right); & 0\le n\le N-1\\ 0 & \text{ailleurs}\end{cases} \qquad (1.2)$$

- on calcule enfin la transformée de Fourier discrète des échantillons de la trame pondérée :

$$\bar{s}(k)=\bar{s}\left(f=\frac{k}{N}\right)=\sum_{n=0}^{N-1}s(n).w(n).\exp\left(\frac{-j2\pi nk}{N}\right); \qquad 0\le k<N \qquad (1.3)$$

Le spectre de puissance (ou densité spectrale de puissance) de la transformée de Fourier est donné par :

$$S(k)=\left|\bar{s}(k)\right|^2 ; \quad 0\le k<N/2 \qquad (1.4)$$

En itérant les opérations précédentes sur des trames acoustiques extraites toutes les 5 à 10 ms du signal de parole, on obtient la description de l'évolution temporelle du spectre fréquentiel de puissance de ce signal. La représentation graphique de l'évolution temps-fréquence-énergétique est le spectrogramme. Pour représenter l'analyse en trois dimensions, l'abscisse représente le temps, l'ordonnée la fréquence et la noirceur du tracé indique l'intensité présente dans le signal au temps et à la fréquence donnés (sur une échelle logarithmique). La largeur de bande du filtre employé (qui donne la précision fréquentielle désirée) est inversement proportionnelle à la précision temporelle que l'on peut escompter. On est donc amené à utiliser en gros deux types d'analyse [5] :

- en bande étroite (typiquement 45 Hz) qui dissimule les variations temporelles rapides du signal mais révèle finement sa structure fréquentielle (on distingue chaque harmonique d'un son périodique de parole avec précision) ;
- en bande large (typiquement 300 Hz) qui, au contraire, permet de bien visualiser les évènements temporels, mais dont la résolution fréquentielle est faible.

La figure 1.6 représente le spectrogramme obtenu des trois voyelles [i], [u] et [a].

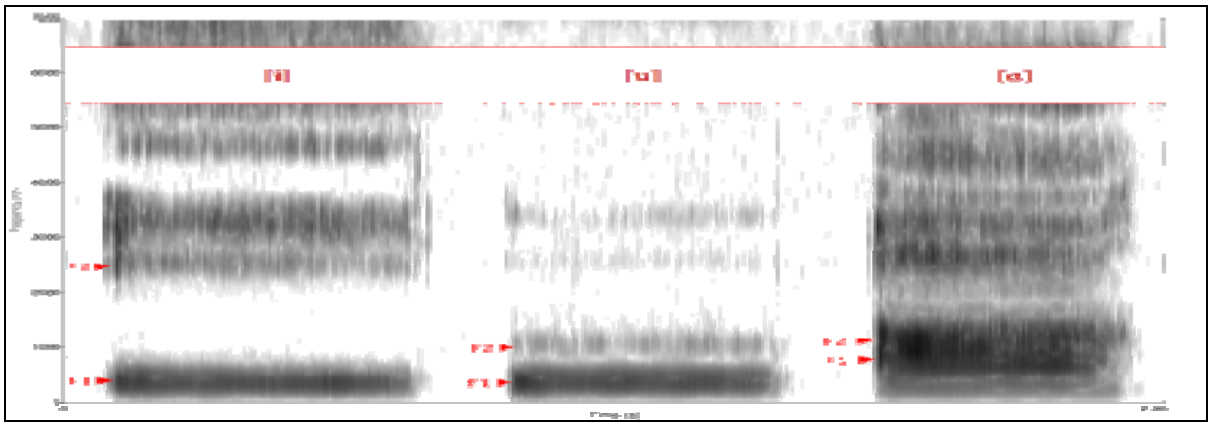


Figure 1.6 - Spectrogrammes des voyelles [i], [u] et [a] [11]

1.3.2. Modélisation de la parole par la prédiction linéaire

Suivant l'hypothèse de quasi-stationnarité à court terme du signal de parole, et dans le cadre des méthodes de modélisation des signaux stationnaires, il est possible de poser un modèle qui permet de rendre compte, d'une manière simple et relativement efficace, des interactions à court terme du processus de production de la parole. Ainsi, on peut dire que le signal de parole est le résultat de l'excitation des cavités supraglottiques par une ou deux sources acoustiques (source laryngienne, bruits d'explosion ou de friction). Ces sources vont exciter le conduit vocal dont la fonction de transfert à court terme est équivalente à celle d'un modèle de la *prédiction* linéaire.

L'analyse par prédiction linéaire LPC (Linear Prediction Coding) [10-11] se fonde sur la corrélation entre les échantillons successifs du signal vocal. L'échantillon à l'instant n, s(n), peut être prédit approximativement comme une combinaison linéaire des p échantillons précédents :

$$s(n) = \tilde{s}(n) = a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p) = \sum_{i=1}^{p} a_i s(n-i) \qquad (1.5)$$

Les coefficients de prédiction a_i sont supposés constants sur une fenêtre d'analyse du signal. En introduisant une excitation normalisée v(n) et un gain d'excitation G, on obtient :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G.v(n) \qquad (1.6)$$

où G .v(n) est identifiée à l'erreur de prédiction introduite par le modèle ou résidu d'ordre p :

$$e(n) = s(n) - \tilde{s}(n) \qquad (1.7)$$

On définit ainsi un filtre linéaire de prédiction dont la fonction de transfert est :

$$H(z)=\frac{S(z)}{GV(z)}=\frac{1}{1-\sum_{i=1}^{p} a_i.z^{-i}}=\frac{1}{A(z)} \qquad (1.8)$$

où S(z) et V(z) sont les transformées en z des signaux s(n) et v(n).

Ce filtre autorégressif (AR) tout –pôles, représenté sur la figure 1.7, peut être assimilé au modèle acoustique linéaire de production de parole formé de la concaténation de tuyaux sonores de sections différentes et variables selon les sons à produire.

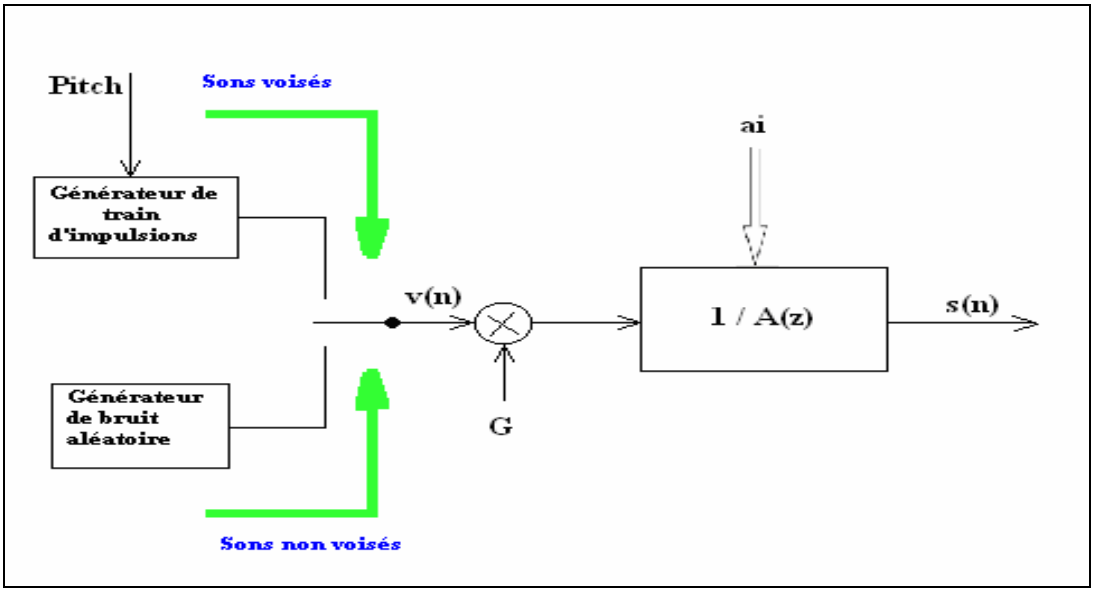


Figure 1.7 : Modèle autorégressif de prédiction linéaire de la parole [5]

Pour la modélisation du signal de parole, le nombre p de coefficients a_i est fixé de telle façon à ce que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants des segments voisés. Il prend généralement une valeur comprise entre 8 et 20.

1.3.3. Analyse cepstrale

L'analyse cepstrale permet, dans le cas d'un signal de parole, la séparation des deux composantes de ce signal qui sont : l'excitation de la source et la réponse du conduit vocal. Comme pour la modélisation de prédiction linéaire, cette analyse suppose que l'appareil de production de la parole se comporte comme un modèle source -filtre. Le signal de parole résulte donc du produit de convolution de l'excitation et de la réponse impulsionnelle du filtre :

$$s(n)=e(n)*h(n) \qquad (1.9)$$

Un traitement appelé déconvolution, ou traitement homomorphique de la convolution, permet de séparer les signaux e(n) et h(n). Ce traitement consiste à calculer d'abord la transformée en z du signal s(n), c'est-à-dire S(z) = H(z).E(z), ensuite le logarithme de S(z) et enfin la transformée en z inverse du logarithme de S(z). On appellera le signal $\hat{s}(n)$ obtenu par cette opération cepstre complexe associé au signal s(n). On a donc:

$$\hat{s}(n) = \hat{e}(n) + \hat{h}(n) \tag{1.10}$$

Si, comme le signal de parole, le signal s(n) est un signal réel, alors $\hat{s}(n)$ sera aussi réel et on pourra le calculer à partir du module de la transformée de Fourier. En pratique, le cepstre réel est obtenu en prenant le logarithme de son spectre auquel on applique ensuite une transformation de Fourier inverse, comme le résume la figure 1.8.

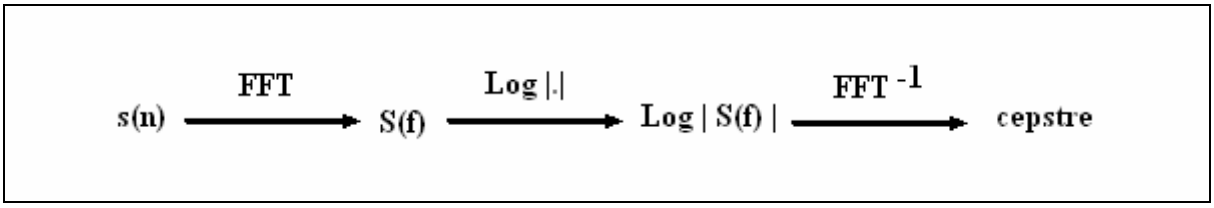


Figure 1.8 : Principe de l'analyse cepstrale [10]

Les coefficients cepstraux sont donnés par :

$$c(n) = \hat{s}(n) = TFD^{-1}(\log S(k)) = \frac{1}{K}\sum_{k=0}^{K-1} \log(|S(k)|).e^{\left(\frac{j2\pi kn}{N}\right)} \quad ; \quad 0 \le n \le K-1 \tag{1.11}$$

où S(k) représente la densité spectrale de puissance du signal s(n) de l'équation 1.4.

Pour estimer la contribution du conduit vocal dans le signal de parole, on ne conserve que les premiers échantillons du cepstre c(n) qui correspondent en particulier aux informations sur les formants. Les échantillons du cepstre d'ordre plus élevé correspondent en général aux caractéristiques de la fréquence fondamentale des cordes vocales.

1.4. TRAITEMENT AUTOMATIQUE DE L'ARABE STANDARD

L'Arabe est une langue parlée par plus de 337 millions de personnes. Elle est la langue officielle d'au moins 22 pays. C'est aussi la langue de référence pour plus de 1,3 milliard de musulmans. Comme son nom l'indique, la langue Arabe est la langue parlée à l'origine par le peuple arabe. C'est

une langue sémitique[2] (comme l'Hébreu, l'Araméen et le Syriaque). Dans le cadre de notre travail de thèse, nous parlerons de la langue Arabe en référence à ce qui est communément appelé **"l'Arabe Standard"** (AS), c'est-à-dire, la langue de communication commune à l'ensemble du Monde Arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel. La normalisation de cette variante de la langue fut généralisée par des grammairiens durant les premiers siècles de l'islam. Ces dernières années, elle connaît un regain d'intérêt, entre autres dans le domaine du traitement automatique.

L'Arabe Standard (AS) compte 34 phonèmes: 6 voyelles et 28 consonnes. Les phonèmes arabe se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques.
La graphie des lettres est différente selon leur position dans le mot. Ainsi, la lettre ب [b] est transcrite بَيْتٌ [bajtun] (une maison) en début de mot, خُبْزٌ [xubzun] (du pain) en milieu de mot, كَلْبٌ [kalbun] en fin de mot et قُرْبَ [qurba] (à proximité de) isolé en fin de mot. Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscules/majuscules n'existe pas [12]. La table des symboles utilisés et leurs équivalents en Alphabet Phonétique International (API) sont présentés dans le Tableau 1.1.

Pour les besoins de la transcription les 28 consonnes arabes ont été divisées en deux groupes:

- 14 consonnes solaires qui assimilent le «ل» de l'article ;
- 14 consonnes lunaires qui n'assimilent pas le «ل» de l'article.

Les solaires se prononcent en double, comme par exemple avec le mot « soleil » شمس(chams), au lieu de prononcer الشمس, el-chams, on prononce ech-chams, car la lettre ش (chin), est une lettre solaire.

Les lettres lunaires, se prononcent normalement et simplement pour elles-mêmes, c'est-à-dire sans les doubler. Par exemple avec le mot « lune », قمر (qamar - lune), on prononce القمر, el-qamar tout à fait normalement, parce que la lettre ق (qaf) est une lettre lunaire (Tableau 1.2).

[2] Les langues sémitiques font partie de la famille des langues afro-asiatiques, et sont parlées en Afrique septentrionale et saharienne ainsi qu'au Proche et Moyen-Orient.

Tableau 1.1 : Liste de l'alphabet arabe et leur API [13]

Nom en arabe	Position			API
	Initiale	Médiane	Finale	
ألف	ا	ـأ ، ـؤ ، ـئ	ئ، ء، ؤ، ـأ	[ʔ]
باء	بـ	ـبـ	ـب	[b]
تاء	تـ	ـتـ	ـت، ـة	[t]
ثاء	ثـ	ـثـ	ـث	[θ]
جيم	جـ	ـجـ	ـج	[ʤ]
حاء	حـ	ـحـ	ـح	[ħ]
خاء	خـ	ـخـ	ـخ	[x]
دال	د	ـد	ـد	[d]
ذال	ذ	ـذ	ـذ	[ð]
راء	ر	ـر	ـر	[r]
زاي	ز	ـز	ـز	[z]
سين	سـ	ـسـ	ـس	[s]
شين	شـ	ـشـ	ـش	[ʃ]
صاد	صـ	ـصـ	ـص	[ṣ]
ضاد	ضـ	ـضـ	ـض	[ḍ]
طاء	طـ	ـطـ	ـط	[ṭ]
ظاء	ظـ	ـظـ	ـظ	[ẓ]
عين	عـ	ـعـ	ـع	[ʕ]
غين	غـ	ـغـ	ـغ	[ɣ]
فاء	فـ	ـفـ	ـف	[f]
قاف	قـ	ـقـ	ـق	[q]
كاف	كـ	ـكـ	ـك	[k]
لام	لـ	ـلـ	ـل	[l]
ميم	مـ	ـمـ	ـم	[m]
نون	نـ	ـنـ	ـن	[n]
هاء	هـ	ـهـ	ـه	[h]
واو	و	ـو	ـو	[w]
ياء	يـ	ـيـ	ـي	[j]

Tableau 1.2 : Classification des consonnes selon les contraintes de la transcription [14]

Solaires	*Lunaires*
ن ل ظ ط ض ص ش س ز ر ذ د ث ت	ق ف غ ع خ ح ج ب أ ي و م ه ك

Les voyelles : on distingue trois voyelles courtes opposées à trois voyelles longues, la durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales et sont réparties comme suit :

- les voyelles courtes : [a], [u], [i], ces voyelles sont représentées dans un texte voyellé au dessus ou au dessous de la consonne, (ـَ , ـُ , ـِ) , exemple : تُرِكَ (turika) ;

- les voyelles longues : [aa], [uu], [ii], ces voyelles sont écrites sous forme de caractères consonantiques (ـَا , ـُو , ـِي) et sont obligatoirement représentées dans un texte écrit (sauf dans certains cas particuliers), exemple : مُسَافِرُونَ (musaafiruuna).

Les consonnes : les consonnes de l'arabe peuvent être classées suivant plusieurs critères comme suit (Tableau 1.3) :

- vibration des cordes vocales: les consonnes articulées avec une vibration des cordes vocales sont dites sonores (ou voisées), sinon elles sont dites sourdes (non voisées) ;
- le franchissement de l'air à travers le conduit vocal :
  - les fricatives qui sont caractérisées par un frottement sur les parois du conduit vocal. On distingue les fricatives non voisées comme س[s] et les fricatives voisées comme ز [z] ;
  - les occlusives qui sont caractérisées par un passage de l'air momentanément arrêté en un point quelconque de l'articulation, l'échappement de l'air s'effectue avec une petite explosion. On rencontre des dentales, des labiales et des glottales qui peuvent être aussi voisées et non voisées comme ب [b] et د [d] ;
  - une liquide caractérisée par un passage de l'air sur les côtés de la langue : ل [l] ;
  - deux nasales caractérisées par un échappement de l'air en même temps par la bouche et par le nez: م [m], ن [n] ;
  - une vibrante caractérisée par la vibration de la langue au passage de l'air: ر [r] ;
  - deux semi-consonnes (ou semi-voyelles) caractérisées par un passage rapide de l'air à travers la bouche accompagné de frottement consonantiques: ي [j], و [w].
- le mode d'articulation : suivant le mode d'articulation, on distingue les consonnes géminées et les consonnes emphatiques. Toute consonne géminée est formée par l'assemblage de deux consonnes identiques fortement articulées. La gémination est indiquée par un signe graphique spécifique appelé chadda (ـّـ). Les consonnes emphatiques ط [ṭ], ض [ḍ], ص [ṣ] , ظ [ẓ] sont caractérisées par une forte tension des différents organes du conduit vocal.

Tableau 1.3 : Classification des phonèmes selon leur point d'articulation [13]

Alphabet	Mode et lieu d'articulation
ء	Laryngale occlusive
ب	Labiale occlusive sonore
ت، ة	Dentale occlusive sourde
ث	Interdentale émise en insérant le bout de la langue entre les dents ; fricative sourde
ج	Affriquée palatale sonore
ح	Fricative laryngale sourde
خ	Vélaire fricative sourde
د	Dentale occlusive sonore
ذ	Interdentale fricative sonore émise en insérant le bout de la langue entre les dents
ر	Vibrante linguale sonore
ز	Dentale fricative sonore
س	Dentale fricative sourde
ش	Palatale fricative sourde
ص	Emphatique ; dentale fricative sonore vélarisée
ض	Emphatique ; interdentale occlusive sonore vélarisée
ط	Emphatique ; dentale occlusive sourde vélarisée
ظ	Emphatique ; interdentale fricative sonore vélarisée
ع	Laryngale fricative sonore
غ	Vélaire fricative sonore
ف	Labiodentale fricative sourde
ق	Occlusive arrière-vélaire sourde accompagnée d'une explosion glottale
ك	Palatale occlusive sourde
ل	Linguale ; sonore souvent appelée « liquide »
م	Labiale nasale sonore
ن	Dentale nasale sonore
ه	Fricative glottale sonore
و	Semi-voyelle vélaire labiale sonore
ي	Semi-voyelle palato-alvéolaire sonore

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

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

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

[an] : ـً [un] : ـٌ [in] : ـٍ

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

1.5. PARTICULARITES PHONOLOGIQUES DE L'AS

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

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

- l'itbaq : les consonnes qui ont cette qualité sont ص [ʂ], ض [ð], ط [ţ], ظ [ẓ]. Celles-ci sont pressées et produites par la langue élevée vers le palais ;
- le tafkhiim : son contraire est le tarqiiq. Il traduit une expression acoustique grasse et épaisse de certaines consonnes ;
- l'istilaa : cette qualité décrit le mouvement articulatoire que fait la langue quand elle meut vers la partie postérieure de la cavité buccale, avec ou sans tafkhiim.

Seules les consonnes ص [ʂ], ض [ð], ط [ţ], ظ [ẓ] possèdent ces trois qualités et sont appelées consonnes *emphatiques* (ou consonnes pharyngalisées). Si nous comparons le français à l'arabe, nous constatons que la différence entre *patte* et *pâte* par exemple est rarement faite en français « standard ». En revanche, cette postériorisation a suscité beaucoup d'intérêt en ce qui concerne l'Arabe [15-17]. Du fait de sa pertinence au niveau perceptif, la modélisation de l'emphase est primordiale en synthèse de la parole à partir du texte de l'AS. Sa prise en compte passe par l'introduction de nouvelles variantes de voyelles dans les contextes emphatiques. Néanmoins, sa mise en œuvre est directement liée à la technique de synthèse utilisée. Rajouani a défini, dans son système à base de règles, un jeu de 6 voyelles brèves et longues emphatisées qui se distinguent des non-emphatisées par la valeur de leurs fréquences formantiques [18]. Dans une approche par

diphones, Ghazali [19] et Guerti [16] ont défini des unités acoustiques incluant les variantes emphatisées des voyelles avec l'ensemble des autres phonèmes. Les trajectoires des formants se trouvent ainsi préservées et fidèlement restituées.

La gémination : au niveau graphique, elle est symbolisée par le signe de la chadda qui signifie le doublement de la consonne. Sur le plan phonétique, l'opposition simple/géminée peut se résumer de la manière suivante : pour une consonne non-occlusive, l'opposition se réduit essentiellement à l'opposition temporelle brève/longue ; pour une occlusive, elle réside au niveau de la durée du silence [20]. Ce rallongement entraîne l'accentuation des propriétés de la consonne (augmentation du caractère emphatique). Une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position (les lèvres ne se referment pas après le premier [b] dans ka**bb**ara). Dans beaucoup de langues, ce phénomène permet de mettre en relief un mot dans son contexte, alors qu'il s'avère être un élément distinctif sur les plans morpho-sémantiques en langue Arabe [20] : حَضَرَ [hazḍara] ″ il a assisté″ est différente de حَضَّرَ [haḍḍara] ″il a préparé″ où la deuxième consonne est géminée.

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

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

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

1.6. PROBLEMES DE LA LANGUE ARABE EN TRAITEMENT AUTOMATIQUE

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

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

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

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

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

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

1.7. CONCLUSION

Ce chapitre a permis dans sa première partie d'introduire certains concepts de base du traitement de la parole via une caractérisation du signal de parole sur le plan physiologique, acoustique et phonétique. La deuxième partie a été consacrée à l'exposé sommaire et la description des différents sons de l'AS. Pour cela, nous avons présenté une étude brève sur le système phonétique de l'AS en présentant quelques généralités phonétiques et certaines propriétés spécifiques en donnant une description très simplifiée.

Le chapitre suivant présentera l'état de l'art de la synthèse vocale où les principales méthodes de synthèse de la parole seront présentées ainsi que les différents types d'unités de synthèse utilisables et les différents modules de traitements linguistiques et acoustiques présents dans un système de TTS.

CHAPITRE 2 :

ETAT DE L'ART SUR LA SYSTHESE

DE LA PAROLE

2.1. INTRODUCTION

La synthèse de la parole à partir du texte (TTS) est la passerelle entre le monde de l'écrit et celui de l'oral. Dés les années 80, ses premiers utilisateurs furent les personnes malvoyantes. Les systèmes de synthèse de la parole leur permettent en effet d'avoir accès aux informations écrites, sous forme vocale, apparaissant sur l'écran de leur poste de travail par exemple. Aujourd'hui, la synthèse de la parole à partir du texte est mise en œuvre dans de nouveaux types de services, en particulier pour les services téléphoniques.

Dans ce chapitre, nous allons introduire le cadre technique de notre étude : la synthèse de la parole. Le chapitre s'articule autour des principes de la synthèse vocale, des différents systèmes présents dans l'état de l'art, suivie d'une description bien détaillée des différentes techniques et méthodes utilisées.

2.2. HISTORIQUE DE LA SYNTHESE VOCALE

Dans cette partie, nous allons présenter que les grandes lignes des évolutions de la synthèse vocale. Les premières machines parlantes voient le jour avec l'abbé Mical[3] et Wolfgang von Kempelen[4] (Figure 2.1) au 18ème siècle, premières grandes simulations mécaniques des phénomènes de production de la parole humaine associant source vocale et résonateurs supraglottiques.

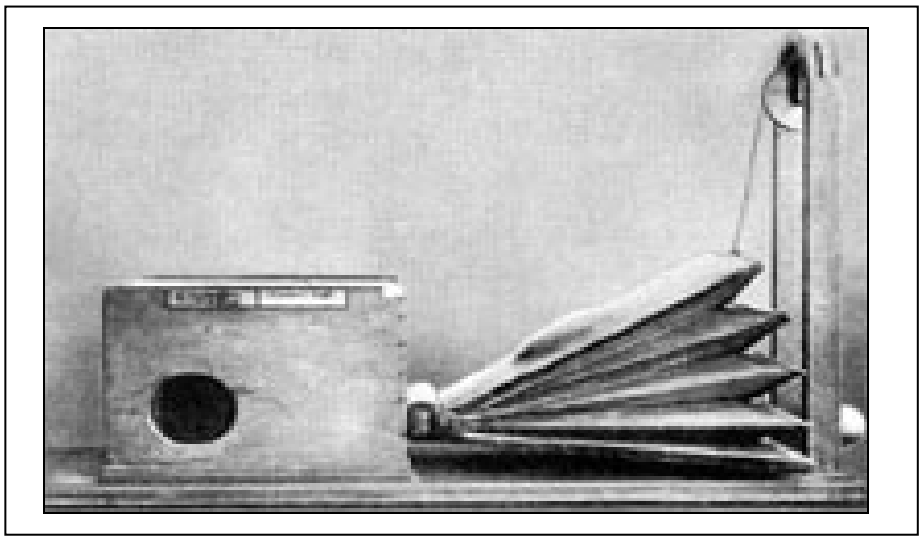


Figure 2.1 : Machine à parler de Kempelen [11]

Au 19ème siècle, la parole et la voix deviennent objets d'études scientifiques spécifiques. Inspirés par les travaux de leurs précurseurs, plusieurs chercheurs ont mis au point des machines

[3] http://www.automates-boites-musique.com/index.php?file=hismical
[4] http://www.ling.su.se/staff/hartmut/kemplne.htm

simulant le conduit vocal. Parmi ceux-ci comptent Joseph Faber avec son Euphonia (1830-40) ; Charles Wheatstone avec le perfectionnement de la machine de Kempelen ; Alexander Graham Bell pour une version simplifiée de la reconstitution de Wheatstone et R.R. Riesz avec un appareil simulant les différentes sections du conduit vocal. Ce dernier permit une meilleure compréhension de la physiologie de l'appareil phonatoire humain, de la géométrie et du rôle de ses articulateurs.

Le début du 20ème siècle vit l'apparition de l'électricité et de l'électronique ce qui autorisa des tentatives plus ambitieuses : en 1922, J. –C. Stewart fabrique une machine capable de reproduire des voyelles, des diphtongues (Voyelles complexes qui changent de timbre en cours d'émission) et quelques mots simples ; plusieurs années plus tard (1939), H. Dudley présente, à l'occasion de l'exposition universelle de New York, le VODER (Voice Operating Demonstrator)., appareil mis au point par les laboratoires Bell, fondé sur le vocodeur à canaux[5]. Mais ce n'est que dans les années cinquante que les premiers véritables synthétiseurs de la parole font leur apparition, avec, par exemple, le Pattern Playback[6], système mis au point par les laboratoires Haskins, qui se présente comme un sonagraphe fonctionnant à l'envers (un faisceau de lumière produit, après amplification, des sons à partir de la représentation de leur durée, de leur fréquence et de leur intensité).

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

2.3. APPLICATIONS DE LA SYNTHESE DE PAROLE

Les applications actuelles de synthèse de la parole à partir du texte peuvent être regroupées en cinq grands domaines [21-22] :

1. Aides pour personnes handicapées :
   - lecture d'écrans ou de documents écrits pour non-voyants ;
   - aides à la communication vocale pour personnes muets, laryngectomisés ou à infirmité motrice cérébrale ;
   - journaux vocaux, etc.
2. Outils d'Enseignement Assisté par Ordinateur (OEAO) :
   - système de dictées automatiques ;
   - système d'apprentissage des langues.

[5] http://ptolemy.eecs.berkeley.edu/~eal/audio/voder.html.
[6] http://www.haskins.yale.edu/featured/patplay.html

3. Applications industrielles :
   - serveurs d'alerte, de surveillance de sites et de supervision de réseaux ;
   - télémaintenance ;
   - fonctions d'aide dans les postes de pilotage ;
   - fonction de vérification vocale dans les postes d'édition (correction des épreuves) ou de saisie d'informations écrites (bases de données), etc.
4. Applications grand public non téléphoniques :
   - domotique (alarmes, appareils domestiques parlants, etc.) ;
   - micro-informatique (jeux et CDROMs parlants, bureautique, etc.).
5. Télématique vocale :
   - serveurs vocaux d'informations (la synthèse remplaçant la parole naturelle enregistrée pour des informations rapidement évolutives et disponibles sous forme textuelle) ;
   - serveurs de lecture vocale de FAX ou de messages électroniques (e-mails) ;
   - automatisation de services de prise de commande (vente par correspondance) ;
   - automatisation de services de renseignements (Annuaires, standards d'entreprises, etc.).

2.4. ARCHITECTURE D'UN SYSTEME DE SYNTHESE DE LA PAROLE

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

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

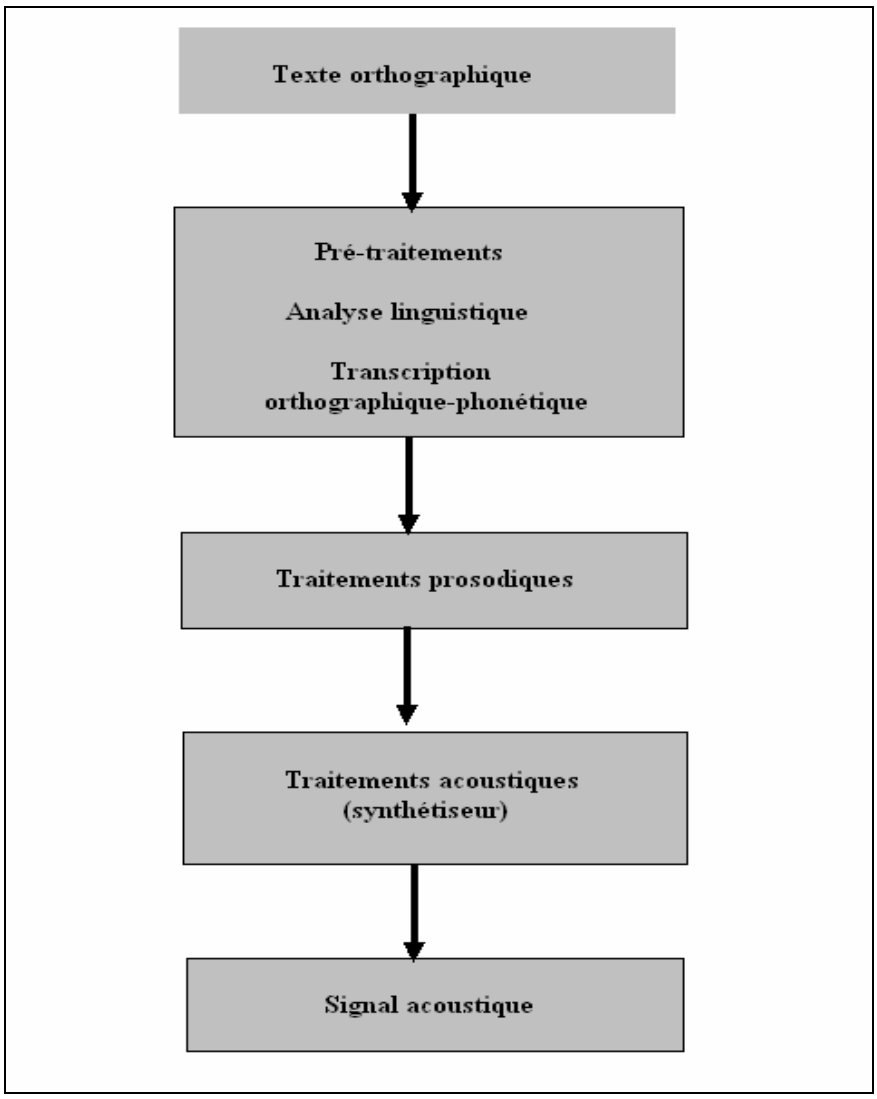


Figure 2.2 : Architecture générale d'un système de synthèse de la parole à partir du texte [12]

La figure 2.3 représente une architecture classique d'un système de synthèse de la parole à partir du texte. Elle ce compose de deux blocs de traitements cités en introduction (traitement linguistique et traitement acoustique). Le premier bloc est composé de trois modules principaux qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne symbolique, en général les phonèmes (unités acoustiques minimales), munie d'indications prosodiques caractérisant l'élocution (durée des différents sons et des pauses, évolution de la mélodie). Cette représentation phonético-prosodique est ensuite utilisée par l'étage de synthèse sonore, qui assure la génération du signal de parole.

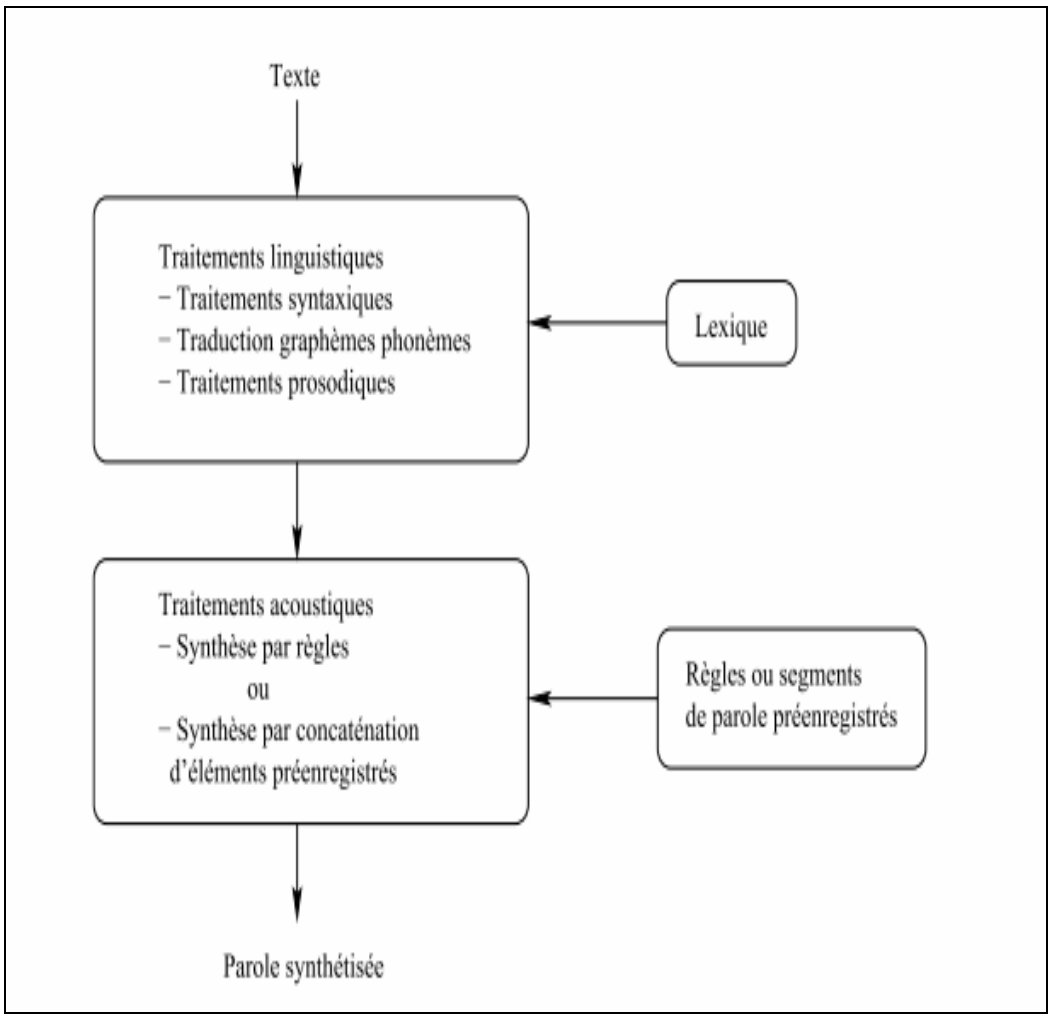


Figure 2.3 : Architecture classique d'un système de synthèse de la parole à partir du texte [3]

2.5. PRINCIPALES METHODES DE SYNTHESE

Les méthodes de synthèses se divisent d'abord quant à la taille du vocabulaire employé. Les systèmes sont dits à vocabulaire limité ou à vocabulaires illimité. Dans le premier cas, le degré de manipulation effectué sur le signal de parole pour obtenir une bonne qualité d'écoute est assez réduit : l'unité utilisée généralement est le mot. Dés que de nouvelles phrases se font nécessaires, l'explosion du vocabulaire oblige l'adoption de méthodes dont l'unité manipulée est inférieure au mot [23].

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

2.5.1. Synthèse par concaténation d'unités acoustiques

La synthèse par concaténation d'unités acoustiques ne fait pas explicitement, tout au moins dans son principe, référence à un modèle de production de la parole. Elle fonctionne en concaténant des unités acoustiques, c'est-à-dire en mettant bout à bout des signaux de parole préenregistrés. Cette

technique est la seule qui permette à ce jour de synthétiser de la parole dont le timbre est proche de celui d'un locuteur humain.

Après des tentatives infructueuses pour composer le signal de parole par concaténation d'unités de la taille du phonème [28], le diphone sera proposé comme unité minimum utilisée dans la concaténation :

> *Un diphone est un élément sonore caractéristique de la transition entre deux phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant.* [26]

La tendance suivante a été d'augmenter de plus en plus la taille de l'unité à stocker, toujours dans le même souci de préserver des transitions complexes, comme dans les groupes consonantiques ou dans le cas des semi-voyelles. Le terme de *polyson* a été proposé pour cette nouvelle unité [27]. Dans un tel système, la mémoire exigée peut devenir grande, mais le nombre de règles de concaténation est réduit.

La tendance actuelle est la synthèse par sélection d'unités. Il s'agit de la concaténation des unités de tailles différentes sélectionnées soigneusement à partir d'un dictionnaire volumineux. Le principal intérêt de cette méthode réside dans la multi représentation des données en termes de contextes linguistiques, prosodiques et acoustiques au sein des dictionnaires [28]. Cette technique a permis de produire de la parole dont l'intelligibilité et le naturel rendent possible la confusion avec une prononciation humaine. Néanmoins, elle implique un accès très rapide à plusieurs dizaines de mégaoctets de données [29].

2.5.2. Synthèse par règles

La synthèse par règles est une méthode qui a eu beaucoup de succès dans le contexte de la synthèse de la parole à partir du texte. Des règles sont utilisées pour estimer les paramètres nécessaires Cette approche est fondée sur un modèle de production du signal vocal, modèle commandé par un nombre restreint de paramètres. La synthèse se décompose alors en deux étapes : une transformation des informations phonético prosodiques, à l'aide de règles contextuelles [24], en commandes permettant de spécifier l'évolution temporelle des paramètres du modèle de synthèse; les paramètres ainsi déterminés sont utilisés pour synthétiser le signal acoustique.

Dans ce type de synthèse, les caractéristiques supra -glottiques sont modélisées à l'aide d'un filtre linéaire dont la fonction de transfert varie au cours du temps. Les paramètres utilisés pour le contrôle du filtre sont les paramètres formantiques, à savoir la fréquence centrale, la bande passante et l'amplitude des maxima significatifs de la fonction de transfert du conduit vocal. Pour obtenir une parole intelligible, il suffit de spécifier les paramètres des 3 à 4 formants les plus importants,

d'où la dénomination de synthèse par formants couramment employée pour ce type de synthèse. Une telle approche ne permet pas de restituer un signal de parole apparaissant naturel. La qualité médiocre obtenue résulte d'une part de la difficulté à modéliser suffisamment finement les trajectoires acoustiques et d'autre part de la modélisation trop grossière du signal glottique [3].

2.5.3. La synthèse par systèmes dynamiques

Cette méthode se propose à modéliser l'organisation des gestes articulatoires ou des trajectoires acoustico-visuelles inspirées par le système biologique et les théories du contrôle moteur [30]. Un autre aspect prometteur de cette méthode est l'intégration possible entre signal visuel et acoustique dans un espace articulatoire commun [31]. L'importance de la composante visuelle se fait sentir notamment dans les environnements bruités, où la récupération des cibles phonémiques est largement aidée par l'information fournie par les lèvres.

2.6. PRINCIPALES TECHNIQUES DE SYNTHESE

Les techniques de synthèse dépendent de la stratégie adoptée. Dans le cas d'une synthèse basée sur la stratégie *system-models*, la synthèse articulatoire figure comme étant la technique qui répond aux concepts de cette stratégie. Dans le cas de synthèse fondée sur la stratégie *signal-models,* deux familles de techniques sont utilisées : celles fondées sur un modèle source/filtre, et celles traitant le signal de parole directement dans le domaine temporel ou fréquentiel [21].

2.6.1. Synthèse articulatoire

La synthèse articulatoire est potentiellement considérée comme la technique la plus performante car elle reflète théoriquement le processus physiologique. Cette technique est basée sur une modélisation géométrique du conduit vocal. Elle consiste à représenter le conduit vocal comme un tube de section variable, avec des embranchements et des sections parallèles, puis à y simuler le trajet des ondes produites au niveau de la glotte. Les modèles d'écoulement d'air (mécanique des fluides), de sources et de propagation acoustique (phénomènes physiques), en association avec des modèles articulatoires (mécaniques), permettent de constituer un synthétiseur articulatoire complet, contrôlé par deux jeux de paramètres : les paramètres supra –laryngés qui commandent le modèle articulatoire, et un jeu de paramètres qui pilotent les cordes vocales (pression sub-glottique, longueur des cordes vocales et hauteur de la glotte au repos) [3,32].

La synthèse articulatoire est difficile à mettre en oeuvre. Par ailleurs, comparée aux techniques alternatives [33-34], le volume de calcul est considérablement plus élevé. C'est pourquoi la synthèse articulatoire est très rarement utilisée dans les systèmes actuels. Mais cette méthode a un grand

potentiel [35], d'une part pour sa haute qualité de synthèse, et d'autre part pour l'approfondissement des connaissances acquises jusqu'à maintenant sur la production de la parole.

2.6.2. Synthèse par formants

Les formants sont les fréquences propres du conduit vocal lors de la production d'un son voisé. Dans cette technique, le filtre du conduit vocal est composé d'un certain nombre de résonateurs similaires au nombre de formants de la parole naturelle. Les résonances formantiques naturelles du conduit vocal sont simulées par des filtres résonants du deuxième ordre caractérisés par une fréquence centrale et une largeur de bande spécifiques. Une synthèse de qualité est obtenue par la simulation des quatre premiers formants. L'implantation de ces filtres peut se faire soit de façon cascade, soit de façon parallèle, soit de façon mixte (Figure 2.4). Pour les sons voisés, ce système est excité par une onde périodique dont la forme est aussi proche que possible de l'onde glottale. Pour les sons non voisés, l'excitation est un bruit blanc.

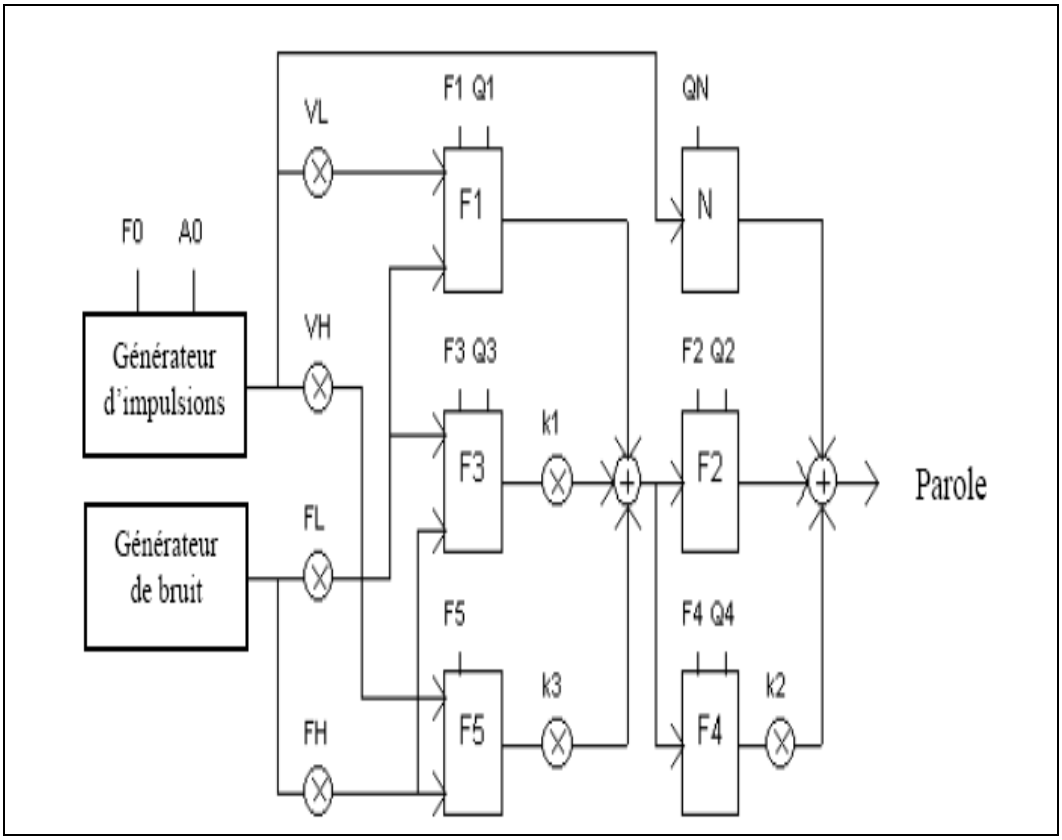


Figure 2.4 : Synthétiseur à formants qui combine les deux structures : cascade et parallèle [36].

Où :

- F_0 et A_0 sont respectivement la fréquence fondamentale et l'amplitude de la composante voisée ;
- F_n et Q_n sont respectivement les fréquences de formants et leur bande passante ;
- V_L et V_H sont respectivement l'amplitude basse et haute de la composante voisée ;

- F_L et F_H sont respectivement l'amplitude basse et haute de la composante non voisée ;
- Q_N est la valeur de la bande passante du formant nasal à 250Hz.

2.6.3. Synthèse par Prédiction Linéaire (LP)

La synthèse par prédiction linéaire est une méthode qui s'inscrit dans le cadre de la théorie source/filtre. Elle a largement été utilisée dans les systèmes par concaténation, car elle permet un codage rapide des unités à concaténer contrairement à la synthèse par formants.

La parole de synthèse produite en utilisant la synthèse LP est loin d'être parfaite. Klatt [37] a montré que la synthèse LP fondée sur la méthode d'autocorrélation ne reproduit pas correctement les fréquences et les bandes passantes des formants lors de la synthèse avec une fréquence fondamentale différente de la fréquence initiale. La synthèse du signal de parole avec sa fréquence fondamentale d'origine conduit aussi à une dégradation du signal due à l'excitation utilisée. En effet, cette dernière est de nature très simplifiée par rapport au signal d'erreur réel. Plus particulièrement, dans le cas de sons voisés, d'autres informations ne sont pas prises en compte, conduisant à la dégradation du signal.

Pour pallier ce problème, une technique dite de prédiction linéaire par impulsions multiples est mise en oeuvre. Elle consiste à construire une excitation composée de plusieurs impulsions pour chaque trame de parole analysée. La synthèse avec la combinaison de cette excitation et les coefficients LP produit un signal de parole très proche du signal naturel. Cette technique est très intéressante pour les vocodeurs et le codage de la parole à bas débit [38].

Les autres extensions de la synthèse LP sont : RELP (Residual Excited Linear Prediction) [39] et CELP (Codebook Excited Linear Prediction) [40]. La première technique utilise le signal d'erreur ou résiduel comme signal d'excitation. La deuxième technique utilise un dictionnaire de signaux d'excitation.

2.6.4. Synthèse fondée sur l'algorithme PSOLA

L'algorithme PSOLA (*Pitch Synchronous Overlap and Add*) [41] consiste à concaténer, à l'aide d'un lissage, des unités de parole pré-stockées en modifiant le pitch et la durée des segments. Cette technique est associée à la méthode de synthèse par concaténation. L'algorithme PSOLA permet la synthèse d'une parole de haute qualité.

Les différentes versions de PSOLA existantes [21] fonctionnent selon le même principe. Le segment de signal de parole naturelle est subdivisé en un ensemble de signaux dits à Court-Terme (CT) en utilisant un fenêtrage synchronisé avec le pitch (trame voisée, Figure 2.5) et à intervalles fixes (trame non voisée). Le pitch est augmenté ou diminué en agissant sur la distance entre les signaux à CT durant le processus de synthèse. La durée est gérée par suppression ou duplication des signaux à CT.

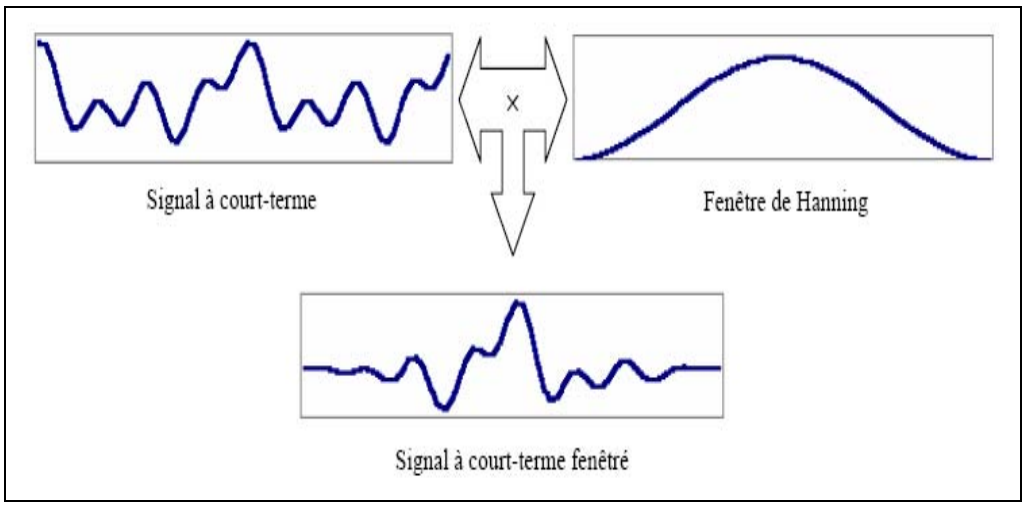


Figure 2.5 : Exemple de signal à Court-Terme [21]

Les signaux à Court-Terme (CT) sont recombinés pour produire le signal de synthèse à l'aide d'une technique d'addition/recouvrement (OverLapp-Add :OLA) (Figures 2.6 et 2.7).

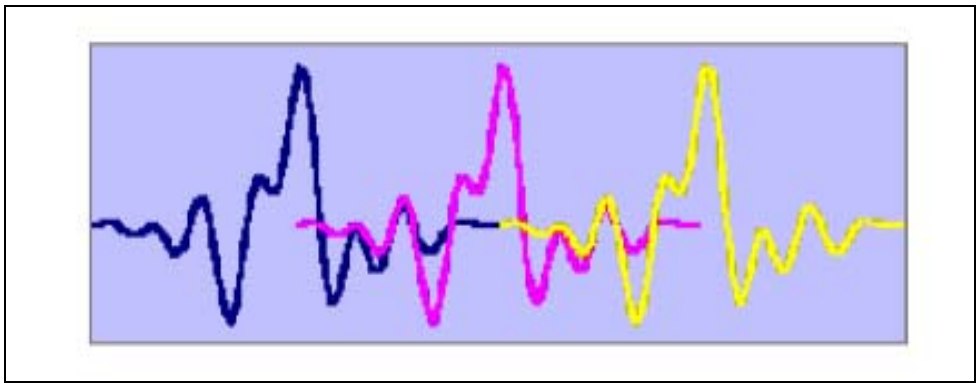


Figure 2.6 : Etape d'addition et recouvrement OLA [21]

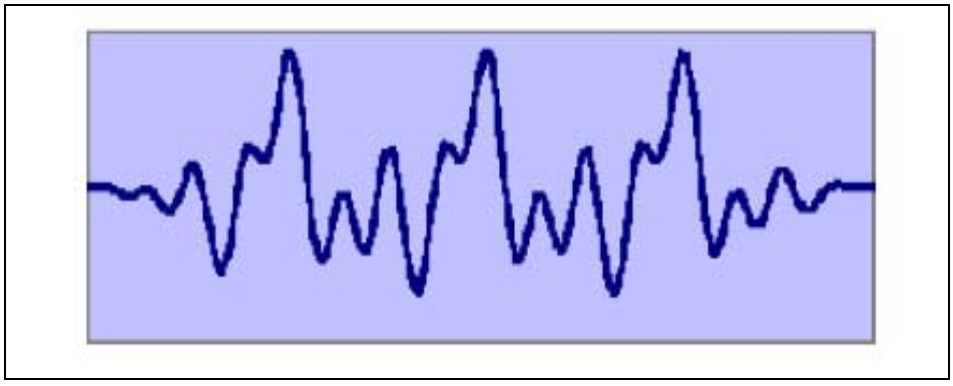


Figure 2.7 : Signal synthétisé avec PSOLA [21]

2.6.4.1. La version TD-PSOLA

La première version simple de PSOLA manipule le signal de parole dans le domaine temporel. Elle est connue sous le nom de TD-PSOLA. La qualité de synthèse obtenue est supérieure à celle obtenue par la Prédiction Linéaire (LP).

Même si la qualité de synthèse par TD-PSOLA est bonne, elle n'est pas parfaite. Le problème majeur apparaît avec l'augmentation significative des durées des sons non voisés. La répétition des signaux à CT non voisés peut avoir comme conséquence l'apparition d'une périodicité locale qui est perçue comme un son tonal. D'autres problèmes élémentaires contribuent à dégrader la qualité de synthèse. On cite à titre d'exemple l'utilisation des fenêtres de Hanning larges (comprennent plusieurs périodes fondamentales) qui peut provoquer une incohérence entre la fréquence du pitch imposée et la fréquence intrinsèque contenue dans chaque signal à court-terme. Ceci a comme conséquence le changement sélectif des amplitudes des harmoniques du pitch qui se manifeste par une réverbération dans le signal synthétique. Ce défaut peut être atténué en utilisant des petites fenêtres de Hanning. Cependant, dans ce cas, les bandes passantes des formants sont élargies. Ainsi l'estimation de l'enveloppe spectrale globale dans chaque signal à court terme est caractérisée par une faible résolution fréquentielle due à la taille courte des trames analysées [21].

2.6.4.2. La version FD-PSOLA

Pour pallier les inconvénients décrits précédemment, une variante de l'algorithme PSOLA a été proposée dans le domaine fréquentiel FD-PSOLA [42]. Dans cette approche, l'enveloppe spectrale globale est obtenue pour chaque signal à court terme en utilisant par exemple la technique LPC. L'estimation du spectre source est effectuée en divisant la Transformée de Fourier Discrète (TFD) du signal à court terme par l'enveloppe spectrale globale. Le spectre source peut être modifié pour adapter la fréquence fondamentale désirée. Cette procédure permet de pallier le problème d'incohérence décrit auparavant. L'enveloppe spectrale peut être aussi utilisée pour modifier la qualité vocale ou lisser les transitions entre les unités sonores à concaténer. Après les modifications requises, les deux spectres sont recombinés et une Transformée de Fourier Inverse (TFI) est appliquée pour générer le signal à court terme de synthèse qui sera traité comme précédemment. La technique FD-PSOLA nécessite cependant un temps de calcul considérable en comparaison avec la technique TD-PSOLA.

2.6.4.3. La version LP-PSOLA

La technique LP-PSOLA est une combinaison de la technique TD-PSOLA et de la LP. Dans cette approche, l'algorithme TD-PSOLA est appliqué au signal résiduel de la prédiction linéaire LP (ou au signal à impulsions multiples) plutôt que directement sur le signal de parole. L'avantage réside dans le fait que les distorsions spectrales au niveau des fréquences des formants sont

minimisées [43]. Étant donné que les résonances du spectre d'excitation sont souvent très larges, l'utilisation de petites fenêtres, est suffisante pour les signaux à court terme qui ne se dégradent pas malgré le manque de résolution. L'algorithme LP-PSOLA permet le lissage de l'enveloppe spectrale au niveau des transitions entre les unités concaténées.

2.6.4.4. La version MBR-PSOLA

Cette version de PSOLA, appelée Multi-Band Re-synthesis PSOLA, ou MBR-PSOLA a été développée par Dutoit et Leich [44]. Cette technique est appliquée sur une base de données contenant des segments de parole prétraités. Ce traitement consiste à re-synthétiser une base de données de diphones de parole naturelle. Ces diphones sont codés avec le modèle d'Excitation Multi Bande (MBE) [45], puis décodés avec des règles de modification.

Ainsi, un nouveau dictionnaire est généré et est utilisé par l'algorithme PSOLA. Afin d'éviter les incohérences de pitch et de phase, cet algorithme propose d'utiliser des valeurs fixes de pitch et de différences de phase entre les harmoniques. L'enveloppe spectrale originale de chaque segment est préservée. Ainsi les incohérences d'enveloppes peuvent être corrigées durant la phase de synthèse en utilisant une simple interpolation dans le domaine temporel. Ces stratégies réduisent le temps nécessaire pour la conception de la base de données des unités de parole [21].

2.7. TRAITEMENTS LINGUISTIQUES

Le bloc de traitements linguistiques (Figure 2.3, page 33) regroupe les différents modules qui permettent de transformer la forme textuelle du message à synthétiser en une chaîne de phonèmes éventuellement enrichis d'informations linguistiques et prosodiques caractérisant l'élocution. Ces différents modules sont : les prétraitements des éléments non lexicaux, l'analyse lexicale, l'analyse syntaxique, la transcription orthographique - phonémique et le traitement prosodique.

2.7.1. Prétraitement des éléments non lexicaux

Cette étape de prétraitement permet de retranscrire en toutes lettres les chaînes non orthographiques. Il peut s'agir de chiffres, de dates (20/10/95, 19 Jan. 2008) ou plus généralement de sigles composés de caractères orthographiques et numériques (vol AH2106, référence AM66). En général, on fait appel à des règles de transcription pour le traitement des quantités numériques, des dates ou des sigles standard (SNCF, PTT, etc.). Si le système de synthèse est destiné à un domaine spécifique, le lexique propre à ce domaine sera appliqué.

2.7.2. Analyse lexicale

L'analyse lexicale consiste à déterminer dans un lexique les différents lexèmes[7] composant le texte orthographique à synthétiser. Cette analyse est réalisée en trois étapes : un découpage du texte en lexèmes, une analyse morphologique et une analyse lexicale [3].

2.7.3. Analyse syntaxique

L'analyse syntaxique vise à déterminer la structure de la phrase. Elle est conduite par application de règles pouvant être de deux types. Dans certains cas, il peut s'agir d'heuristiques, résultant généralement de l'application de règles grammaticales standards (par exemple, on ne peut observer la succession de deux verbes conjugués). En complément ou à la place de ces heuristiques parfois très complexes, on utilise aussi fréquemment des règles probabilistes, exploitant des modèles de langage. Ces modèles sont fondés sur l'observation que toutes les séquences de catégories grammaticales dans une langue donnée ne sont pas équiprobables [3]. La connaissance de la catégorie syntaxique exacte est également utile pour déterminer la prononciation correcte et notamment pour désambiguïser les homographes hétérophones.

2.7.4. Transcription Orthographique-Phonétique (TOP)

Traditionnellement appelée conversion graphème-phonème, l'étape de Transcription Orthographique-Phonétique (TOP) constitue le noyau minimal, indispensable à tout système de synthèse de parole, aussi élémentaire soit-il. Cette étape repose sur l'utilisation d'un automate paramétré appliquant un ensemble de règles de réécriture, qui permettent d'associer un phonème (ou un groupe de phonèmes) à un caractère (ou un groupe de caractères) orthographique en prenant en compte le contexte gauche et le contexte droit. Ces règles sont organisées de façon hiérarchique, des règles les plus particulières aux règles les plus générales [3]. Le nombre de règles nécessaires pour effectuer la TOP dépend de la langue que l'on considère

2.7.5. Traitements prosodiques

La chaîne parlée est d'abord subdivisée en unités suprasegmentales qui facilitent le décodage du message par l'auditeur. La délimitation de ces unités est faite à l'aide de marqueurs dont la réalisation fait appel à des variations paramétriques de durée, de fréquence et d'intensité [3].

Les traitements prosodiques sont complexes et s'articulent en différents modules (insertion des pauses, durées phonétiques et fréquence fondamentale). Cependant, l'apparition des techniques de synthèse par sélection dynamique d'unités non uniformes de segments de parole ont permis d'envisager des techniques nouvelles pour la génération de la prosodie. En effet, ces approches

[7] Dans ce contexte le terme "lexème", qui représente une suite de caractères orthographique, est plus approprié que "mot".

génèrent automatiquement la prosodie sans modèle a priori puisqu'elles utilisent une caractérisation symbolique fine des unités d'un corpus de grande taille, ce qui permet de conserver la prosodie originale des segments sélectionnés.

2.7.5.1. Insertion des pauses

Les pauses correspondent aux silences, de durées variables, qui s'insèrent à la fin de chacun des groupes de souffle. L'importance de la coupure syntaxique liée à un marqueur syntaxico-prosodique détermine la durée de la pause à insérer. Ce facteur est particulièrement important pour le naturel de l'élocution [3]. La génération de pauses est absolument nécessaire à la synthèse de la parole, et le réalisme de leur durée et de leur position est indispensable à la qualité de la synthèse résultante.

2.7.5.2. Durées phonétiques

Une bonne détermination des durées est cruciale pour assurer le naturel de l'élocution. Des durées erronées produisent une parole heurtée, chaotique et parfois difficilement intelligible. Deux approches existent, pour la modélisation de la durée :

- la première basée sur des règles et une bonne analyse statistique, initiée par [24], détermine la durée en prenant en compte différents facteurs, en particulier la durée intrinsèque des sons constituant le segment et le contexte. Parmi les facteurs influençant la durée phonétique, nous pouvons citer : le contexte phonétique (certains phonèmes ont tendance à allonger les phonèmes adjacents, d'autres auront tendance à les raccourcir), la position de la syllabe porteuse dans le groupe prosodique (en français par exemple, la syllabe finale des mots est généralement allongée, d'un facteur d'autant plus important que le groupe précède une frontière syntaxique majeure), la nature du groupe prosodique (sa fonction dans la phrase), la longueur du groupe prosodique, etc. [3] ;
- la deuxième approche est basée sur des techniques d'apprentissage automatique. Celles-ci peuvent reposer sur l'utilisation de réseaux connexionnistes pour prédire la durée des syllabes et ainsi calculer les durées des phonèmes à partir de leur moyenne et de leur écart-type. Dans [46], Price et al. proposent un modèle HMM à 7 états pour détecter automatiquement les coupures prosodiques à partir de l'analyse des durées des phonèmes.

2.7.5.3. Fréquence fondamentale

Le contrôle de la fréquence fondamentale, dont l'évolution dans le temps définit le contour mélodique, est le point essentiel pour la détermination de l'intonation. L'évolution de la fréquence fondamentale pour chaque phonème est spécifiée à l'aide d'un modèle prédictif complexe, prenant en compte deux types de phénomène globaux dits de macromélodie, et locaux dits de micromélodie :

- la macromélodie a une portée supérieure à celle de la syllabe (groupe prosodique, phrase). Les facteurs influant sur la macromélodie sont la position de la syllabe dans le groupe prosodique, la fonction du groupe prosodique dans la phrase et le mode de la phrase (interrogatif, déclaratif...) ;
- la micromélodie est l'influence de l'évolution de la fréquence fondamentale par sa position dans la syllabe et par son environnement phonétique immédiat (certains phonèmes, comme les consonnes occlusives, contribuent à abaisser la fréquence fondamentale ; d'autres ont tendance à l'augmenter).

Plusieurs modèles de contours mélodiques ont été proposés. Dans [47], Fujisaki et Hirose utilisent un modèle acoustique source/filtre d'un ensemble limité de paramètres structurés entre eux pour modéliser la fréquence fondamentale. Taylor dans [48] propose une modélisation automatique de la courbe de fréquence fondamentale en terme de montées et de descentes non linéaires et de connexions.

2.8. CONCLUSION

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

D'une manière générale, nous pouvons dire que la qualité de synthèse est principalement mesurée par l'intelligibilité et le naturel de la parole. L'intelligibilité dépend essentiellement de la technique et de la méthode de synthèse utilisées. Le naturel est quant à lui associé en grande partie à l'aspect prosodique de la langue étudiée. Pour cela, le prochain chapitre sera consacré donc à l'étude de la prosodie.

CHAPITRE 3 :

ETUDE DE LA PROSODIE

3.1. INTRODUCTION

La prosodie est essentielle à la compréhension et au naturel de la parole, et est donc indispensable pour un système de reconnaissance ou de synthèse vocale. Les récentes conférences intégralement dédiées à la prosodie (International Conference on Speech Prosody 2002, 2004, 2006 & 2008) [49] sont une illustration remarquable de ses applications futures et de sa présence dans notre vie quotidienne. Le terme ″prosodie″ fait donc partie du vocabulaire de nombreuses communautés scientifiques, en sciences humaines comme en sciences de l'ingénieur. Dans ce chapitre, il est important et nécessaire de le définir, tout d'abord de façon informelle et de manière plus précise, de sorte à comprendre le rôle de la prosodie dans la communication orale. Ensuite, nous expliquons les difficultés rencontrées dans l'étude de la prosodie du point de vue de l'ingénieur, notamment pour l'extraction de paramètres et finalement, nous étudions l'intérêt d'exploiter la prosodie pour l'identification et la synthèse des langues.

3.2. DEFINITION DE LA PROSODIE

Le mot prosodie est originaire du grec Προσωδια, qui signifie ″accent, quantité, dans la prononciation″. Ces derniers l'utilisaient pour faire référence aux traits du discours, plus précisément aux tons ou accent mélodique. La mélodie de la prosodie est restée dans l'oubli jusqu'à la fin des années 1940, lorsque Firth [50] utilisa ce terme à nouveau pour décrire l'approche qu'il préconisait pour l'analyse linguistique.

La première expérience sonore prosodique est l'écoute d'une langue étrangère, dont nous ne maîtrisons aucun aspect. Ainsi, les impressions qui nous parviennent sont transmises par le chant, la force, ou le timbre de la voix, qui nous permet d'identifier un ami par exemple, même si nous ne comprenons pas ses propos. La prosodie est donc liée à l'impression musicale que fournit un locuteur lorsqu'il parle.

D'après Di Cristo [51], « La prosodie est une branche de la linguistique consacrée à la description et à la représentation formelle des éléments de l'expression orale tels que les accents, les tons, et l'intonation, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F_0), de la durée (D) et de l'intensité (I) (paramètres prosodiques physiques) » (Figure 3.1).

A travers les différentes composantes de la prosodie, nous retiendrons qu'elles ne se définissent pas uniquement à partir des caractéristiques physiques du signal, c'est à dire l'acoustique. Néanmoins, la prosodie est un des constituants de la parole qui reste accessible à chacun d'entre nous sans connaissance particulière.

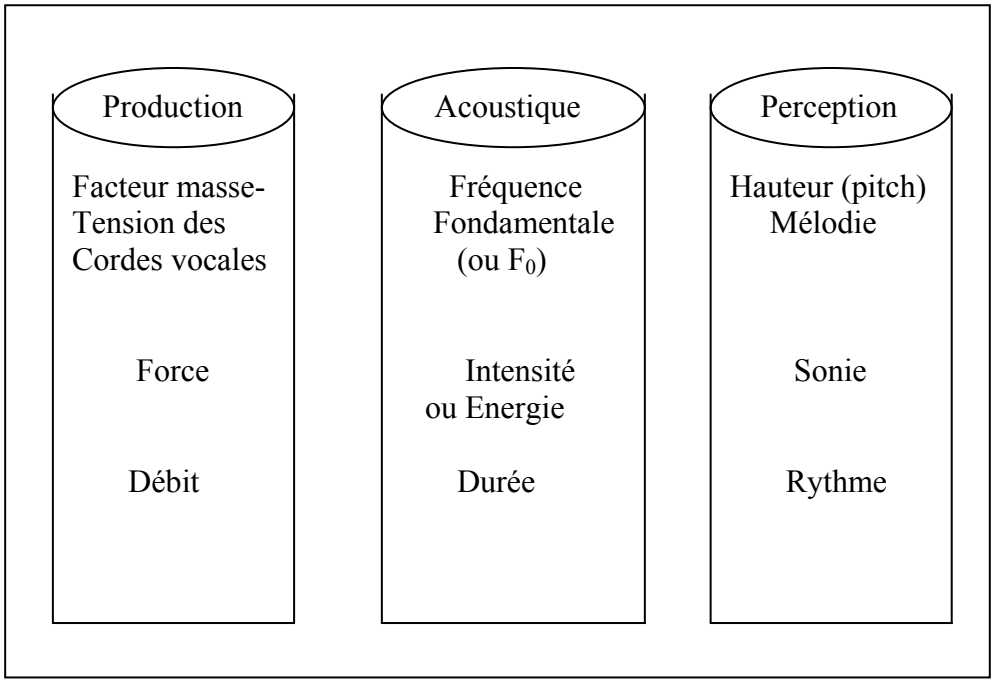


Figure 3.1 : Paramètres prosodiques dans les différents domaines de la parole [52]

La substance prosodique repose sur 4 propriétés acoustiques [53] :

- le rythme : cette notion comprend le débit de parole, la longueur et la répartition des pauses, les allongements syllabiques, la durée de divers événement sonores (syllabes, phonèmes), etc. Étant donné son lien avec tous les autres traits acoustiques de la prosodie, il est particulièrement difficile d'isoler les paramètres acoustiques qui puissent résumer cette dimension ;
- l'intonation **:** elle est souvent présentée comme le paramètre primordial de la prosodie. Elle contient le caractère chantant de la parole. Contrairement au rythme, l'intonation est définie à partir d'un seul paramètre acoustique : la fréquence fondamentale (F_0). Il est la conséquence de la vibration des cordes vocales et de la pression trans-glottique ;
- le volume sonore **:** il correspond au paramètre physique de l'intensité (I), c'est-à-dire l'énergie contenue dans le signal au cours d'un intervalle de temps donné ;
- le timbre **:** il est spécifique des instruments ou des voix. Il est perçu indépendamment de sa hauteur ou de son intensité. L'évolution du timbre est provoquée par la superposition des composantes harmoniques durant l'émission du son. Une fois encore, cette dimension est difficile à expliciter physiquement puisqu'elle s'appuie sur l'ensemble des valeurs spectrales du signal.

Ainsi, un très grand nombre de définitions de la prosodie ont été proposées, suivant notamment les contextes dans lesquels ce terme est employé. D'un point de vue acoustique, par exemple, nous trouverons une définition comme: « étude de la **durée**, de la **fréquence fondamentale** et de **l'intensité** (ou énergie) du son », ou bien pour des aspects plus linguistiques : « partie de la phonologie qui échappe à l'analyse en phonèmes et traits distinctifs, tels que le **ton**, **l'intonation**, la **durée** et **l'accent**».

3.3. EXTRACTION DES PARAMETRES PROSODIQUES

Les trois paramètres prosodiques classiquement extraits du signal acoustique sont la fréquence fondamentale, la durée et l'intensité (ou l'énergie). La première étape de l'extraction de paramètres prosodiques est généralement une analyse à CT du signal, en faisant l'hypothèse que sur une fenêtre de faible longueur le signal est quasi-stationnaire, ce qui signifie que les caractéristiques statistiques du signal évoluent peu. En général une fenêtre d'analyse de 30 ms, appelée trame, est utilisée. L'analyse est répétée à intervalles réguliers, typiquement toutes les 10 ms. Compte tenu du caractère suprasegmental[8] de la prosodie, une analyse sur 30 ms ne suffit pas. C'est pourquoi il est nécessaire de calculer d'autres paramètres à partir d'une analyse sur plusieurs trames successives afin de traduire le rythme, l'intonation et l'accentuation.

3.3.1. Fréquence fondamentale (F_0)

La fréquence fondamentale (F_0), corrélât acoustique de la fréquence laryngienne, correspond à la fréquence de vibrations des cordes vocales. C'est essentiellement la conséquence des variations de la pression trans-glottique et de la tension des cordes vocales. Autrement dit, c'est l'estimation de la fréquence laryngienne à partir du signal acoustique à un instant donné. Les algorithmes d'extraction de F_0 utilisent une représentation temporelle ou spectrale du signal.

Un algorithme d'extraction de la fréquence fondamentale se décompose en trois phases successives [54] :

- un prétraitement et un changement de représentation ;
- l'extraction du fondamental ;
- un post-traitement visant à corriger les erreurs.

Le prétraitement vise à optimiser les caractéristiques du signal en vue de l'extraction en utilisant un filtrage passe-bas, une pré-accentuation ou un filtrage non linéaire. Ensuite des transformations

[8] Employé ici dans le sens d'une propriété portée par une unité plus grande que le segment, ou liée aux relations entre segments distants.

sont appliquées pour adapter la représentation du signal au domaine du traitement fondamental (temporel, temporel à court terme, fréquentiel, …).

La deuxième phase consiste à extraire la fréquence fondamentale et dépend donc du domaine utilisé. Généralement, cela revient à optimiser une fonction de la fréquence fondamentale (fonction de coût, résultat d'une transformation, corrélation, densité de probabilité).

La phase de post-traitement a pour but de diminuer les erreurs qui sont de plusieurs types :

- les erreurs de voisement, lorsqu'une valeur de F0 a été trouvée sur une zone non voisée, ou lorsque aucune n'a été trouvée sur une zone voisée ;
- les erreurs grossières (« gross-errors » en anglais), la fréquence fondamentale correspond à une harmonique ou une sous-harmonique. Ce type d'erreur peut facilement être corrigé en tenant compte du voisinage ou en effectuant un lissage ;
- les erreurs fines, la valeur trouvée est située à plus ou moins 10 % de la valeur réelle.

Le calcul de la fréquence fondamentale se fait donc sur les sons voisés qui ont un caractère pseudo périodique. Cela concerne principalement les voyelles, mais aussi quelques consonnes. Il existe plusieurs algorithmes pour l'estimation de la fréquence fondamentale et qui ne donnent pas toujours des résultats identiques. Ainsi, de nombreuses recherches ont été menées dans le domaine de l'extraction de la fréquence fondamentale des signaux de parole. On peut citer l'ouvrage de référence de Hess [55], où un grand nombre d'algorithmes est détaillé.

Les plus performants d'entre eux, sont cependant incapables de fournir des valeurs toujours correctes de F_0 dans toutes les circonstances (sons, bruits, locuteurs, etc.). Les principaux problèmes rencontrés sont :

Les sauts d'octaves : l'analyseur fournit une valeur de F_0 qui ne correspond pas au premier harmonique. Cela peut arriver pour un spectre dont le deuxième harmonique correspond au premier formant ou dans le cas d'une insuffisance passagère de l'amplitude du fondamental.

Les non-détections : il existe une fréquence ″théorique″ que l'algorithme n'a pas détectée. Ceci arrive très souvent dans des portions peu énergétiques et / ou bruitées du signal de parole.

La finesse du détecteur : les valeurs proposées sont éloignées faiblement des valeurs théoriques.

La décision de voisement : cette décision, bien que difficile à prendre dans certaines situations (faible énergie, parole bruitée, …) serait cependant fort utile, au-delà du bon fonctionnement du détecteur, à des fins de segmentation du continuum sonore.

On recense deux grandes catégories d'algorithmes de décision :

- ceux qui opèrent dans le domaine temporel comme la technique d'AMDF (Average Magnitude Difference Function).
- Et ceux qui travaillent dans le domaine spectral : les valeurs de F_0 sont calculées à partir des maxima des spectres d'amplitude.

3.3.2. Durée

Des trois paramètres prosodiques, la durée (D) est le plus difficile à préciser, car elle n'est pas directement associable à un corrélat biologique du système phonatoire.

Avant de mesurer des durées, il faut cerner correctement les entités à mesurer. On distingue les durées des unités phonétiques, des syllabes, des phonèmes ou même la distance entre voyelles et les durées des pauses. Comme les autres paramètres, la durée de l'entité choisie est largement dépendante du locuteur et du débit de parole. Ainsi, aucune mesure ne peut donner de modèle absolu de la durée. La considération des résultats des observations devra plutôt s'orienter vers un modèle relatif qui pourra s'exprimer en termes d'allongements ou de réductions.

Chaque phonème a une durée intrinsèque et co-intrinsèque. Ces durées sont des caractéristiques des phonèmes. On se rend compte aisément que le phonème [a], pris seul, est plus long que le phonème [b], par exemple.

Les pauses en parole spontanée ne sont pas toutes des silences. On distingue les pauses silencieuses des pauses non silencieuses (qui peuvent être remplies, faux départs, répétitions, ou syllabes allongées). En situation de lecture seule, les pauses qui se traduisent acoustiquement par une absence de signal (les pauses silencieuses) sont considérées.

La durée des différentes unités constitue le phénomène central pour la prosodie. En effet, chaque variation de fréquence fondamentale ou d'intensité s'établit sur un certain laps de temps. Etudier l'organisation temporelle de la parole est incontournable. Etudier la durée, c'est observer et modéliser les durées d'unités bien déterminées.

Pour cela, la durée et la nature de ces unités ont fait l'objet de nombreuses études, principalement motivées par la nécessité de la modéliser dans des systèmes de synthèse de la parole. Dans sa thèse, Barbosa [23] consacre tous le chapitre 2 pour classer les différents modèles de prédiction des durées, selon un ordre croissant de taille des unités utilisées. Parmi les unités qui ont servi de base à ces modélisations, on en trouve principalement quatre : le phonème, la syllabe, le pied et le GIPC (Groupe-Inter-Perceptuel-Center).

3.3.3. Energie

L'énergie (E) (ou l'intensité (I)) du signal sonore de la parole est perçue comme la force de la voix. Son niveau est lié au fonctionnement des systèmes respiratoire et phonatoire et à la pression sous glottale : si la pression sous glottale augmente, l'intensité de la voix augmente également et inversement. Cette intensité est relative à l'énergie contenue dans le signal au cours d'un intervalle de temps donné. Ce terme correspond au corrélât acoustique de la pression sous glottique et d'ouverture du conduit vocal.

Sachant que l'énergie est un paramètre couramment utilisé en traitement du signal, c'est le paramètre prosodique le plus facile à calculer. L'énergie à CT d'un signal s_t échantillonné sur une fenêtre de longueur T est définie par :

$$E = \frac{1}{T}\sum_{t=1}^{T} s_t^2 \tag{3.1}$$

Étant donnée sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db} = 10 \times \log_{10}\left(\frac{1}{T}\sum_{t=1}^{T} s_t^2\right) \tag{3.2}$$

Pour un signal échantillonné de longueur infinie, on calcule l'énergie à CT sur des fenêtres glissantes. Ces fenêtres sont étroites, de l'ordre de 5 à 10 ms [54].

3.4. ANALYSE DE LA PROSODIE

L'analyse de la prosodie permet d'observer les paramètres physiques du signal et leurs attributs perceptifs à différentes échelles : **macroprosodie** à l'échelle de la phrase, **microprosodie** à l'échelle du phonème). Elle s'appuie à la fois sur les résultats de la phonétique et de la phonologie pour considérer des phénomènes relatifs qui ne sont pas liés à un phonème particulier mais plutôt à l'enchaînement des segments entre eux (éléments suprasegmentaux). C'est pourquoi les notions de

mélodie, de rythme, de durée relative et d'accentuation sont étudiées en linguistique dans le cadre de l'analyse de la prosodie [56].

3.4.1. Macroprosodie

La macroprosodie concerne la prosodie générale pour l'ensemble d'une phrase (interrogation, exclamation, etc.). C'est la silhouette que peut prendre une courbe représentant la fréquence par rapport au temps. Par exemple, pour une phrase interrogative, la courbe représentant la fréquence sera croissante vers la fin. Au contraire, pour une phrase exclamative ou un point, la fréquence va chuter (Figure 3.2).

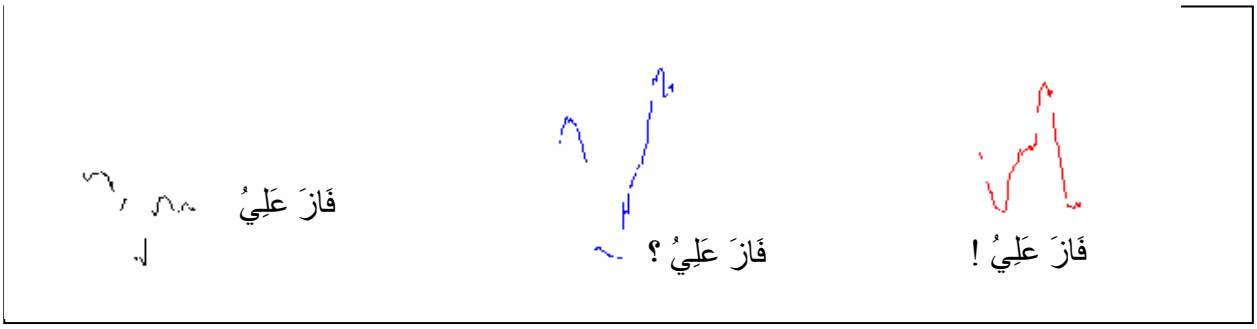


Figure 3.2 : Exemples de macroprosodie

3.4.2. Microprosodie

La microprosodie se caractérise par des variations locales des paramètres prosodiques du phonème. Elle traite généralement l'influence de ces propriétés intrinsèques.

Autrement dit, l'intonation acoustiquement représentée par la courbe de fréquence fondamentale (F_0), est la combinaison d'une composante microprosodique (ou micromélodique), directement dépendante de la nature des phonèmes utilisés par le locuteur et qui ne sont ainsi pas motivés linguistiquement (comme les segments non voisés qui ne présentent pas de valeur de F_0 ou les consonnes constrictives ou fricatives qui provoquent un affaiblissement de la courbe de F_0), et d'une composante macroprosodique, qui reflète les choix syntaxiques et pragmatiques du locuteur en termes de patrons intonatifs [57].

De nombreuses études se sont attachées, depuis les années 60, à factoriser ces deux composantes, et à extraire automatiquement le plus possible, l'information macroprosodique pertinente du signal de parole. Cette extraction peut être décomposée en deux étapes :

- **stylisation**, c'est-à-dire remplacement de la courbe de F_0 par une fonction numérique plus simple conservant la même information macroprosodique;
- **codage symbolique**, c'est-à-dire représentation par une suite de symboles constituant une discrétisation de la courbe stylisée.

3.5. INTERET DE LA PROSODIE

La prosodie apporte une valeur ajoutée réelle dans nombre de contextes liés à la parole. Les plus importantes sont dans le domaine de la :

- reconnaissance et identification des langues : deux aspects de la prosodie motivent l'utilisation de la prosodie dans le cadre de l'identification des langues : l'aspect rythmique et l'intonation. Depuis les deux dernières décennies, de nombreuses expériences montrent l'efficacité des êtres humains pour la reconnaissance des langues [58]. En ce qui concerne les paramètres prosodiques, de nombreuses expériences tentent de mettre en avant les capacités humaines à distinguer les langues en n'en gardant que les propriétés rythmiques ou intonatives [59]. Il s'agit en général de dégrader un enregistrement de parole au moyen de filtrage ou de resynthèse en ne laissant que peu d'indices aux sujets qui doivent identifier la langue. Malgré son importance, la prosodie n'a jusqu'à présent jamais été utilisée de façon efficace en reconnaissance automatique de la parole. Néanmoins, les informations prosodiques sont potentiellement très intéressantes, en particulier pour la reconnaissance de la parole spontanée [10], ainsi que pour la compréhension de phrases [54] ;
- synthèse des langues : la génération automatique d'une intonation de bonne qualité joue un rôle essentiel dans le naturel et l'intelligibilité de la parole synthétique. De nombreux logiciels de synthèse de la prosodie ont vu le jour, citons entre autres : Lacheret-Dujour et Morel [60] ; Morlec, Bailly et Aubergé [61]. Les modélisations informatiques de la prosodie ont besoin pour être opérationnelles d'un étiquetage prosodique. Cependant, la segmentation du continuum prosodique représente une opération extrêmement coûteuse en temps, et ce particulièrement pour la parole spontanée. Pour pouvoir faciliter ces opérations, des logiciels de traitement de la prosodie sont apparus (ProSig, PRAAT, etc…), ainsi que des outils et des méthodes qui automatisent les transcriptions prosodiques (Campione [62] à partir d'INTSINT défini par Hirst et DI Cristo [63]).

3.6. ETAT DE L'ART SUR LA PROSODIE EN ARABE STANDARD

L'étude de la prosodie en Arabe Standard a été suscitée par le développement d'un système de synthèse de la parole arabe à partir du texte au début des années 80. Ce qui implique que l'intonation de l'AS n'a pas bénéficié d'une longue tradition d'analyse en comparaison avec les autres langues.

Les études qui traitent de l'AS dans un contexte général indépendamment de l'influence dialectale sont rares. Al-Ani, dans une étude dédiée à la phonologie arabe [20], a cité à titre indicatif les réalisations acoustiques de l'intonation de quelques modalités de la phrase en Arabe. L'une des

premières études sur l'intonation de la langue AS est celle de Haydar et al. [64]. Cette étude s'articule autour de quatre types de phrases de la langue AS : les phrases affirmatives, exclamatives, interrogatives et celles qui expriment la négation. L'objectif de cette étude était de mettre en évidence les caractéristiques générales des courbes macromélodiques d'un corpus composé de 48 phrases.

Dans le but d'améliorer la qualité et le naturel d'un système de synthèse de la parole AS à partir du texte par diphone, Es-Skali [65] a opté pour l'hypothèse de corrélation entre la structure intonative et la structure syntaxique de la phrase. Dans un premier temps il a élaboré une étude pour analyser l'accent lexical du mot arabe en fonction des paramètres prosodiques et est arrivé à la conclusion que la fréquence fondamentale (F_0) était le paramètre acoustique qui, avec ses variations, caractérisait l'accent lexical. Ce résultat a été ensuite confirmé par Rajouani [18]. Ce dernier a classé les paramètres prosodiques par leurs corrélations avec l'accent lexical. La fréquence fondamentale figure en première position suivie par l'intensité alors que la durée figure en dernière position. Dans un deuxième temps, Es-Skali a extrapolé les règles d'accentuation sur les phrases en tenant compte de la fréquence fondamentale pour l'analyse de l'intonation de l'AS. Les résultats de cette analyse ont été élaborés sous forme de règles acoustico-linguistiques permettant de décrire les variations intonatives des phrases affirmatives et des phrases interrogatives.

Dans sa thèse, Baloul [12] stipule que la syntaxe est incontournable, mais qu'une analyse syntaxique superficielle et partielle peut être suffisante pour le calcul de la prosodie. Il propose un modèle intonatif à trois niveaux distincts (mot, tronçon et phrase) pour le calcul des valeurs de F_0 pour des phrases déclaratives et interrogatives. L'avantage du modèle prosodique proposé par Baloul réside dans sa capacité à effectuer un traitement automatique de l'intonation depuis l'acquisition du texte jusqu'à la synthèse des paramètres acoustico-prosodiques en passant par un analyseur syntaxique élémentaire.

Dans sa thèse, Zaki [21] a proposé deux approches différentes pour la génération automatique de la prosodie. La première approche a consisté en l'utilisation des réseaux de neurones pour la génération automatique de la prosodie. Deux modèles ont ainsi été élaborés. Le premier est dédié à la synthèse des contours intonatifs. Le deuxième est consacré à la prédiction de la durée segmentale. La deuxième approche s'est articulée autour du développement d'un modèle intonatif fondé sur l'approche phonologique. Dans l'implémentation finale de son modèle prosodique, il a adopté le modèle prédictif de la durée segmentale, fondé sur l'approche neuronale, et le modèle de synthèse de l'intonation, fondé sur l'approche phonologique. Le système résultant permet, à partir d'un texte

écrit, de fournir aux systèmes de synthèse un fichier prosodique complet comprenant les informations suivantes :

- transcription phonétique du texte ;
- la durée de chaque phonème ;
- la fréquence de début et de fin de chaque phonème.

Cependant son modèle intonatif fondé sur l'approche phonologique n'intègre pas les phénomènes microprosodiques. Comme perspective, il propose d'appliquer les réseaux de neurones à la modélisation des phénomènes microprosodiques. Une combinaison du modèle macro intonatif et du modèle micro intonatif, pourra aboutir à un système hybride permettant la synthèse complète des phénomènes prosodiques.

3.7. CONCLUSION

Nous avons présenté dans ce chapitre une introduction à la prosodie dans un contexte général. L'état de l'art de la prosodie de l'AS a également été présenté. La prosodie est donc essentielle à la compréhension et au naturel de la parole, et par conséquent à l'identification et la synthèse vocale. Elle est universelle mais est aussi spécifique à une langue. De plus, il n'existe pas qu'une seule prosodie. A travers les différentes composantes de la prosodie, nous retiendrons qu'elles ne se définissent pas uniquement à partir des caractéristiques physiques du signal, c'est à dire l'acoustique. Deux tendances de modélisation des variations de F_0 peuvent être distinguées. La première tendance est fondée sur une synthèse de l'intonation sans la moindre utilisation des informations syntaxiques. La deuxième tendance suppose que l'information syntaxique est indispensable pour la gestion des hauteurs mélodiques au niveau du mot ou du tronçon. Cependant, les deux tendances s'accordent sur la nécessité de l'accent lexical pour caractériser F_0 au niveau phonologique et phonétique.

CHAPITRE 4 :

DETECTION DE L'ACCENT LEXICAL PRIMAIRE EN ARABE STANDARD (AS)

4.1. INTRODUCTION

Dans ce chapitre, nous présentons une nouvelle méthode basée sur la Classification par Analyse Discriminante (CAD) en exploitant le paramètre acoustique énergie (facteur simple à calculer) afin de détecter l'Accent Primaire en AS. Ainsi, une fois la notion de syllabe et d'accent en AS définie, nous exposerons celle de la CAD puis nous présenterons le principe de la méthode choisie, suivie des résultats obtenus.

4.2. NOTIONS SUR L'ACCENT

L'accent est la mise en valeur d'une syllabe et d'une seule dans ce qui représente, dans une langue déterminée, l'unité accentuelle. Autrement dit, l'accent est le phénomène de mise en relief de certaines syllabes qui sont perçues de manière plus forte que les syllabes voisines.
Généralement, les traits phoniques utilisés pour la mise en valeur accentuelle sont l'énergie articulatoire, la hauteur mélodique et la durée, réelle ou perçue, de la syllabe accentuée. Dans bien des langues, la syllabe accentuée tend à être articulée de façon plus énergique, avec un timbre plus élevé et plus long que celui des syllabes inaccentuées voisines qui contrastent avec elle. Et c'est le degré d'énergie, de hauteur et de durée qui permet d'établir la hiérarchie des accents dans l'énoncé. La réalisation de l'accent varie d'une langue à une autre.

Pour les langues à accent fixe, la syllabe à accentuer est parfaitement définie. En Français par exemple, l'accent est toujours porté sur la dernière syllabe du mot. Pour les langues à accent libre, la syllabe accentuée n'a pas une position fixe.

Généralement les trois paramètres prosodiques se combinent dans des proportions inégales pour donner à chaque langue ses caractéristiques accentuelles particulières. En Français par exemple, les facteurs influençant l'accent sont le Fondamental (F_0) et la Durée (D) [66-68]. En Anglais ce sont le fondamental et l'Intensité (I) [63].

L'étude de l'accent en AS a longtemps été dissimulée par les grammairiens, et ce pour plusieurs raisons. D'abord, la diversité et l'influence des dialectes arabes ne permettaient pas l'ébauche d'une étude uniformisée qui soit admise par tous. Ensuite, le rôle de l'accent n'est pas évident au premier abord, au point que certains linguistes ont nié son existence [18] Leur argumentation était que le placement de l'accent, s'il existait, sur n'importe quelle syllabe du mot n'affecte aucunement le sens de celui-ci.

Ces dernières années, nous avons noté un certain engouement pour l'étude de la prosodie arabe, ce qui a consolidé l'hypothèse selon laquelle l'accent lexical existe en Arabe [69-70]. G. Bohas quant à lui, admet non seulement son existence, mais il cite des exemples où celui-ci joue un rôle distinctif [71]. Ce postulat a été ensuite confirmé par A. Rajouani. Ce dernier, a montré que la détection de l'accent primaire semble suffisante pour l'étude de l'intonation arabe et démontre à partir de ses expériences la hiérarchie (F_0, I, D) pour la langue Arabe [18].

4.3. LA SYLLABE ET L'ACCENT EN AS

Tout mot en Arabe possède un accent de mot : l'une des syllabes est prononcée avec plus d'intensité. Autrement dit, tout mot isolé en Arabe reçoit un accent qui sera porté sur la syllabe accentuée. Nous allons essayer de décrire de façon très simplifiée les différents types de syllabes et la place de l'accent dans un mot [72-74].

4.3.1. Différents types de syllabes en AS

L'Arabe Standard comporte cinq types de syllabes classées selon les traits : Ouvert/Fermé et Court/Long (Tableau 4.1). Une syllabe est dite ouverte (respectivement fermée) si elle se termine par une voyelle (respectivement une consonne). Toutes les syllabes comportent une seule voyelle et commencent toujours par une consonne suivie d'une voyelle. La syllabe [CV] peut se trouver au début, au milieu ou à la fin du mot [20]. Les types de syllabes considérées sont : [CV], [CVC], [CVV], [CVVC] et [CVCC], où [V], [VV] et [C] sont respectivement une voyelle courte, une voyelle longue et une consonne :

- chaque syllabe contient une et une seule voyelle, d'où le nombre de syllabes d'un mot est égal au nombre de ses voyelles ;
- chaque syllabe commence obligatoirement par une consonne ;
- la syllabe de type [CV] est dite syllabe courte, c'est la plus fréquente dans la langue Arabe, les autres types sont toutes longues ;
- toutes les syllabes peuvent figurer au début, au milieu ou en fin de mot ;
- les syllabes de type [CV] et [CVV] sont dites ouvertes, les syllabes de type [CVC], [CVVC] et [CVCC] sont dites fermées.

Note: Al-Ani considère que les voyelles forment les noyaux syllabiques alors que les consonnes sont les phonèmes marginaux de la syllabe [20].

Tableau 4.1 : Classification des syllabes en Arabe [20]

Syllabe	Ouverte	Fermée
Courte	[CV]	
Longue	[CVV]	[CVC], [CVVC], [CVCC]

4.3.2. Place de l'accent dans un mot en AS

Al-Ani parle de la présence de trois niveaux d'accent :

- un premier niveau ou Accent Primaire (A_cP) ;
- un deuxième niveau ou Accent Secondaire (A_cS) ;
- un troisième niveau ou Accent Faible ″weak stress″ ou Accent Tertiaire (A_cF ou A_cT).

Les règles qui régissent la place de l'accent dans un mot sont définies comme suit [20] :

- la position et la distribution de l'accent dépendent du nombre et des types de syllabes contenus dans le mot ;
- si toutes les syllabes du mot sont de type [CV] alors c'est la première syllabe du mot qui porte l'A_cP, les autres syllabes reçoivent un A_cF (exemple : دَخَلَ [**da**xala]) ;
- si dans le mot, il y a une seule syllabe longue, alors cette syllabe reçoit l'A_cP (exemple: كَافَحَ [**kaa**faħa]) ;
- si le mot contient deux syllabes longues ou plus, alors c'est la syllabe la plus proche de la fin du mot qui reçoit l'A_cP. La syllabe longue la plus proche du début du mot reçoit un A_cS, les autres syllabes reçoivent un A_cF (exemple: حَيَوَانَاتٍ [ħajawaa**naa**tin]) ;
- la dernière syllabe est exclue dans le processus d'accentuation, et ceci, quels que soient son type et sa nature.

4.4. CLASSIFICATION PAR ANALYSE DISCRIMINANTE (CAD)

Les méthodes de classification sont très utilisées, car elles permettent de grouper des objets selon leur ressemblance. Elles placent alors certains objets dans un même groupe et séparent d'autres en les plaçant dans des groupes différents.

Trois grandes familles peuvent être choisies [75] (indépendamment des méthodes syntaxiques), nous pouvons citer la méthode de :

- recherche de formes proches, par comparaison dynamique ;
- probabilités où les Modèles de Markov Cachés et les réseaux Bayésiens, sont de loin les plus utilisés en Reconnaissance Automatique de la Parole (RAP) ;
- surfaces de décision et fonctions discriminantes des formes.

Dans toutes ces méthodes, le choix de la distance ou métrique entre vecteurs formes est important. La distance Euclidienne est souvent utilisée :

$$d_E(x,y) = \sqrt{(x-y)'.(x-y)} \tag{4.1}$$

Mais la distance de Mahalanobis [76] où C est la matrice de covariance des vecteurs formes x et y est également intéressante, car elle permet de prendre en compte la corrélation entre les paramètres des formes :

$$d_M(x,y) = \sqrt{(x-y)'.C^{-1}.(x-y)} \tag{4.2}$$

4.4.1. L'Analyse Discriminante (AD)

L'Analyse Discriminante (AD) est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire, ...) d'un ensemble d'observations (individus, exemples, ...) à partir d'une série de variables prédictives (descripteurs, variables exogènes, ...). C'est un ensemble de méthodes que l'on classe généralement en deux catégories [76], la discrimination à but :

- descriptif : on cherche parmi les variables quantitatives, ou parmi les combinaisons linéaires de ces variables, celles qui permettent de séparer ou discriminer le mieux possible les différentes classes. L'objectif est donc de proposer un nouveau système de représentation, des variables latentes formées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner le plus possible les groupes d'individus. En ce sens, elle se rapproche de l'analyse factorielle, car elle permet de proposer une représentation graphique dans un espace réduit, plus particulièrement de l'analyse en composantes principales calculée sur les centres de gravité conditionnels des nuages de points avec une métrique particulière. On parle également d'analyse canonique discriminante, notamment dans les logiciels anglo-saxons ;
- décisionnel : on dispose d'un nouvel individu, pour lequel on connaît les valeurs des variables quantitatives, mais pas la classe à laquelle il appartient : il s'agit d'affecter l'individu à l'une des classes (de le classer). Il s'agit dans ce cas de construire une fonction de classement (règle d'affectation, ...) qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. En ce sens, cette technique se rapproche des techniques supervisées en apprentissage automatique telles que les arbres de décision, les réseaux de neurones,... Elle repose sur un cadre probabiliste. Elle est très séduisante dans la pratique car la fonction de classement s'exprime comme une combinaison linéaire des variables prédictives, facile à analyser et à interpréter.

Plusieurs applications de l'AD existent en médecine, dans le domaine bancaire, en biologie, en analyse d'images, en géologie (définition d'indices de prospection caractérisation géochimique de types de roches), etc. L'AD se rattache au champ plus vaste de la Reconnaissance Des Formes. Par ses objectifs, elle s'apparente également aux réseaux neuronaux.

4.4.2. Principe de l'Analyse Discriminante (AD)

L'analyse discriminante descriptive est une technique de statistique exploratoire qui travaille sur un ensemble de **n** observations décrites par **J** variables, répartis en **K** groupes. Elle vise à produire un nouveau système de représentation, constitué de combinaisons linéaires des variables initiales, qui permet de séparer au mieux les **K** catégories.

Nous disposons d'un échantillon de **n** observations réparties dans **K** groupes d'effectifs $\mathbf{n_k}$. Notons **Y** la variable définissant les groupes, elle prend ses valeurs dans $\mathbf{\{y_1, \ldots, y_K\}}$. Nous disposons de **J** variables $\mathbf{X = (X_1, \ldots, X_J)}$. Nous notons $\boldsymbol{\mu_k}$ les centres de gravité des nuages de points conditionnels, $\mathbf{W_k}$ leur matrice de variance-covariance (Les éléments de sa diagonale représentent la variance de chaque variable et les éléments en dehors de la diagonale représentent la covariance entre les variables i et j , $i \neq j$).

L'objectif de l'analyse discriminante est de produire un nouvel espace de représentation qui permet de distinguer le mieux les **K** groupes. La démarche consiste à produire une suite de variables discriminantes $\mathbf{Z_h}$, non-corrélés deux à deux, tels que des individus du même groupe projetés sur ces axes, soient les plus proches possibles les uns des autres, et que des individus de groupes différents soient les plus éloignés possibles.
La dispersion à l'intérieur d'un groupe est décrite par la matrice de variance covariance $\mathbf{W_k}$. Nous pouvons en déduire (à un facteur près) la dispersion intra-groupe :

$$W = \frac{1}{n}.\sum_k n_k \times W_k \tag{4.3}$$

L'éloignement entre les groupes, entre les centres de gravité des groupes, est traduit par la matrice de variance-covariance inter-groupes (à un facteur près) :

$$B = \frac{1}{n}.\sum n_k .(\mu_k - \mu).(\mu_k - \mu)' \tag{4.4}$$

où $\boldsymbol{\mu}$ est le centre de gravité du nuage de points, global.

La dispersion totale du nuage est obtenue par la matrice de variance-covariance totale **V**. En vertu du théorème d'Huyghens (qui est la généralisation mutlidimensionnelle de la formule de décomposition de la variance) :

$$V = B + W \tag{4.5}$$

Le premier axe factoriel sera donc défini par le vecteur directeur $\mathbf{u_1}$ tel que l'on maximise la quantité $\frac{u_1'.B.u_1}{u_1'.V.u_1}$.

La variance inter-classes sur ce premier axe factoriel $\mathbf{Z_1}$ sera maximum.

La solution de ce problème d'optimisation linéaire passe par la résolution de l'équation :

$$V^{-1}.B.u = \lambda.u \tag{4.6}$$

Cette équation peut être résolue par la méthode généralisée de la décomposition en valeurs singulières La réponse nous est directement fournie par le calcul des valeurs propres et vecteurs propres de la matrice $\mathbf{V^{-1}\ B}$. Le premier axe factoriel $\mathbf{Z_1}$ est donc obtenu à l'aide du vecteur propre $\mathbf{u_1}$ correspondant à la plus grande valeur propre λ_1. Le second axe factoriel est défini par le vecteur propre suivant, etc. L'ensemble des axes factoriels est déterminée par les valeurs propres non-nulles de la matrice $\mathbf{V^{-1}\ B}$.
Dans le cas usuel où **n > J > K**, nous obtenons **K-1** axes factoriels.

Enfin, la variance inter-classes calculée sur l'axe factoriel $\mathbf{Z_h}$, que l'on appelle également pouvoir discriminant de l'axe, est égale à la valeur propre λ_h associée.

En supposant éventuellement des probabilités d'affectation a priori données sur les différents groupes :

$$(\text{Probabilité a priori})_i = \frac{n_i}{\sum_{k=1\ldots\text{Nombre de Groupes}} n_k} \tag{4.7}$$

où n_i est le nombre de vecteurs d'entraînement dans le groupe i , l'AD calcule les probabilités a posteriori qu'une nouvelle observation appartienne à chacun de ces groupes (fonction discriminante) et attribue cette observation au groupe le plus probable. Les variantes de la méthode tiennent à la manière dont est définie et calculée cette probabilité.

Il existe plusieurs façons de vérifier la qualité d'une AD. Certaines font appel à des hypothèses probabilistes, d'autres non. Le pourcentage de bien classer, est la statistique la plus utilisée tout en

étant la plus simple. L'idée est la suivante : on a une procédure de classement, alors pourquoi ne pas l'appliquer aux observations dont on connaît le véritable groupe et vérifier ainsi si l'on effectue un bon classement.

Un classement fait entièrement de façon aléatoire donnerait en moyenne 50% de bien classé. Une façon d'obtenir un estimé plus réaliste consiste à mettre de côté une certaine proportion des observations initiales de chaque groupe, de trouver les fonctions de classification avec les autres observations puis d'effectuer le classement des observations mises de côté (échantillon test). Une autre variante consiste à mettre à part une observation à la fois et de répéter l'analyse et le classement ″n″ fois.

4.5. METHODE DE CLASSIFICATION PAR ANALYSE DISCRIMINANTE (CAD)

Dans ce qui suit, nous présentons une nouvelle approche basée sur une méthode de CAD en exploitant le paramètre acoustique énergie (facteur simple à calculer) afin de détecter l'A_cP en AS.

L'idée est la suivante : on a une procédure de classement, alors pourquoi ne pas l'appliquer aux observations dont on connaît le véritable groupe et vérifier ainsi si l'on effectue un bon classement à partir de la matrice de confusion, obtenue.

4.5.1. Corpus et Matériel utilisés

Pour cette méthode, nous disposons de 4 locuteurs arabophones (2 hommes et 2 femmes), nous leur faisons énoncer des mots arabes ayant la structure syllabique S_1 S_2 S_3 (mots de trois syllabes). Ces mots sont prononcés à l'intérieur de phrases porteuses (Tableau 4.2).

Tableau 4.2 : Phrases prononcées par les 4 locuteurs

Phrases
هَلْ تُعْتَبَرْ كَتَبَ، كِتِبِ وَ كُتُبُ كَلِمَات ؟ نَعَمْ كَتَبَ كَلِمة، كَلاَّ كِتِبِ لَيْسَتْ كَلِمَة وَ أَجَلْ كُتُبُ كَلِمَة.
هَلْ تُعْتَبَرْ عُبُثُ، َعبَثَ وَ ِعبِثِ كَلِمَات ؟ لاَ عُبُثُ لَيْسَتْ كَلِمَة، نَعَمْ عَبَثَ كَلِمَة وَ كَلاَّ عِبِثِ لَيْسَتْ كَلِمَة .
هَلْ تُعْتَبَرْ ِبرز، بُرُزُ وَ بَرَزَ كَلِمَات ؟ كَلاَّ ِبرز لَيْسَتْ كَلِمَة، لاَ بُرُزُ لَيْسَتْ كَلِمَة وَ أَجَلْ بَرَزَ كَلِمَة .
هَلْ تُعْتَبَرْ خَبَزَ، خُبُزُ وَ خِبِز كَلِمَات ؟ أَجَلْ خَبَزَ كَلِمَة، لاَ خُبُزُ لَيْسَتْ كَلِمَة وَ كَلاَّ خِبِز لَيْسَتْ كَلِمَة .
هَلْ تُعْتَبَرْ حَزَنَ، حِزِن وَ حُزُنُ كَلِمَات ؟ نَعَمْ حَزَنَ كَلِمَة، كَلاَّ حِزِن لَيْسَتْ كَلِمَة ، لاَ حُزُنُ لَيْسَتْ كَلِمَة.

L'enregistrement s'est fait dans une chambre sourde au niveau du Laboratoire Parole et Langage (LPL) à Aix en Provence (France) puis traité grâce au logiciel de transcription et d'analyse phonétique PRAAT [77].

Après avoir subi une analyse sonagraphique grâce à PRAAT, ces phrases sont ensuite, segmentées et alignées semi-automatiquement en phonèmes et à la fin, une Transcription Orthographique Phonétique (TOP) leur est faite.

PRAAT a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam. Il permet de mener des analyses phonétiques, de faire de la synthèse de la parole et de manipuler des données (analyses statistiques, construction de grammaires, etc.). Avec ce logiciel, il est possible [78] :

- d'enregistrer des fichiers audio qui pourront ensuite être analysés ;
- de transcrire, d'étiqueter et de segmenter des données audio (enregistrements effectués sous PRAAT ou provenant d'autres fichiers, au format WAV, par exemple) ;
- d'effectuer des analyses phonétiques et acoustiques au niveau segmental (spectrogramme, analyse de formants, sonagramme, etc.) et au niveau suprasegmental (pitch, courbe de F_o, intensité et durée) ;
- de manipuler et modifier le signal de parole (utilisation de filtres, modification des contours intonatifs et de la durée, etc.) ;
- de faire de la synthèse de la parole (créer des stimuli audio, synthèse articulatoire, analyse-synthèse de données modifiées, etc.) ;
- de construire des outils d'apprentissage (Réseau de neurones et élaboration de grammaires dans le cadre de la théorie de l'optimalité (Optimality Theory : OT) ;
- de faire des analyses statistiques à partir des études phonétiques (analyses de covariances, etc.).

En effet, le logiciel dispose d'une fenêtre principale permettant d'effectuer des traitements importants et de fenêtres annexes affichant le résultat des divers traitements comme la stylisation de F_0 ou le tracé du spectre (Figure 4.1).

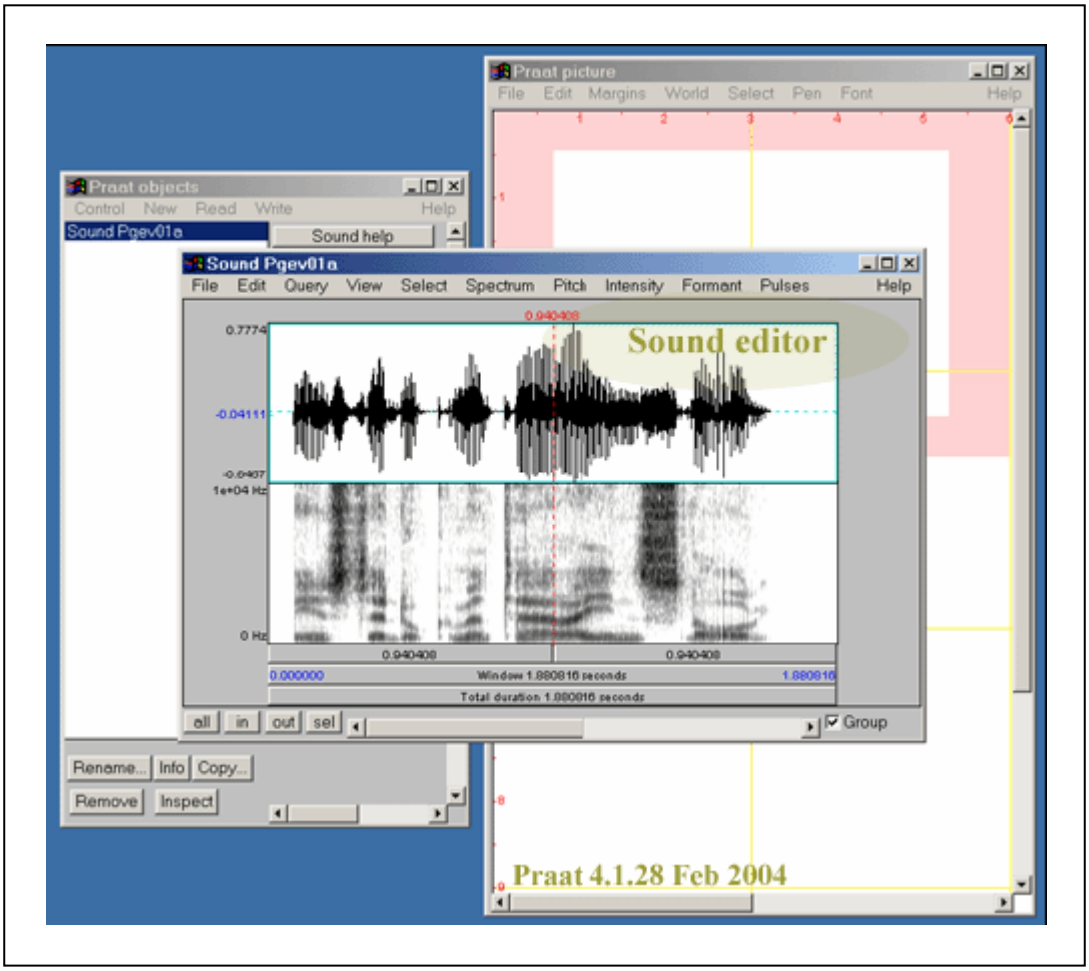


Figure 4.1 : Interface du logiciel PRAAT

Les intérêts principaux de ce logiciel, outre les traitements prosodiques standards, sont la possibilité d'éditer des ″tires″ et le langage de script intégré qui fait de lui une véritable plate-forme de développement. Les tires sont des zones de texte superposées qui contiennent des informations complémentaires. Par exemple, il est possible de disposer d'une tire orthographique contenant la transcription proprement dite, puis d'éditer des tires contenant la transcription phonétique en Alphabet Phonétique International (API), des informations prosodiques, syntaxiques, etc. En ce qui concerne les scripts, cette fonctionnalité s'adresse aux utilisateurs expérimentés, car il s'agit d'un véritable langage de programmation. L'ajout de cette composante offre à l'utilisateur de PRAAT des possibilités quasi-illimitées car il peut développer des scripts répondant exactement aux objectifs qu'il s'est fixé.

4.5.2. Méthodologie de la CAD

Afin de détecter l'Accent Lexical à l'intérieur des nos mots, nous avons alors élaboré un script sous PRAAT, dont la démarche peut être résumée dans les quelques étapes suivantes [79] :

Étape 1 :

Elle concerne l'extraction puis le calcul du spectre à moyen terme pour chaque voyelle détectée à l'intérieur du mot utilisé. Ceci est justifié par le fait qu'en Arabe, une syllabe contient une et une seule voyelle.

Le spectre d'un signal de parole x(t) dans l'intervalle de temps $[t_1, t_2]$ est :

$$X(f) = \int_{t_1}^{t_2} x(t).e^{-j2\pi ft}.dt \tag{4.8}$$

d'où sa Densité Spectrale de Puissance qui est donnée par :

$$DSP(f) = \frac{|X(f)|^2}{t_2 - t_1} = \frac{\left|\int_{t_1}^{t_2} x(t).e^{-j2\pi ft}.dt\right|^2}{t_2 - t_1} \tag{4.9}$$

où $w = 2\pi f$ est la fréquence du signal. Sachant que le module au carré de la Transformée de Fourier représente l'énergie de tout le signal entre t_1 et t_2, il est normalisé en le divisant par la durée du signal. Pour cela, la DSP a pour unité Pa^2 /s.

Pour les systèmes discrets, la Transformée de Fourier Discrète (TFD) est utilisée pour obtenir l'image spectrale du signal dans le domaine fréquentiel. Par conséquent, il est utile de définir la DSP pour un signal discret. Elle est calculée comme suit :

$$DSP(k) = \frac{|X(k)|^2}{N.(t_2 - t_1)} = \frac{\left|\sum_{n=0}^{N-1} x(n).e^{-jkn}\right|^2}{N.(t_2 - t_1)} \tag{4.10}$$

où N est le nombre d'échantillons qui dépend de la fréquence d'échantillonnage et de la durée du signal $(t_2 - t_1)$.

Le signal de parole étant en général non stationnaire. Pour obtenir la distribution d'énergie du signal entier divisé en L bandes, il est raisonnable de calculer le Spectre Moyen à Long Terme (Long-Term Average Spectrum : LTAS) ou periodogram. La méthode de calcul proposée par Welch [80] est la suivante :

- le signal est divisé en L parties, chacune est composée de N échantillons ;
- nous calculons la TFD de chaque bande de N échantillons ;
- nous utilisons l'équation (4.10) pour calculer la DSP correspondante ;
- à la fin, nous calculons le spectre moyen à long terme (LTAS) comme étant la valeur moyenne des L densités spectrales de puissance calculées.

Pour cela, le LTAS peut être calculé comme suit :

$$\mathrm{LTAS}(f) = \frac{1}{L}.\sum_{i=1}^{L} \mathrm{DSP}_i(f) \tag{4.11}$$

où DSP_i (f) est la densité spectrale de puissance de la $i^{ème}$ bande du signal. Une valeur normalisée du LTAS est calculée en unités de dB/HZ :

$$\mathrm{LTAS}_{dB}(f) = 10.\log_{10}\left(\mathrm{LTAS}(f) / P_0^2\right) \tag{4.12}$$

où $P_0=2\cdot10^{-5}$ Pa, représente la densité spectrale de puissance de référence qui correspond au seuil d'audition d'un être humain à la fréquence 1 kHz.

L'analyse par LTAS a été trouvée pour offrir des informations représentatives sur le timbre de la voix. Elle fournit des informations spectrales moyennées pour la durée et est particulièrement utile quand des caractéristiques spectrales persistantes sont à l'étude.

A noter, que le choix du paramètre L affecte la performance de cette méthode de deux façons. Premièrement, diviser le signal en petites bandes et calculer la puissance moyenne d'un long signal temporel permet d'assigner la puissance des fréquences exactes et réduit les variations du bruit dans la distribution de la puissance. Cependant, de petites bandes fournissent une résolution de fréquence plus petite puisqu'il y a moins d'échantillons dans le signal original. De cette façon, il y a une sorte de compromis entre la précision de l'attribution de la puissance de corriger les fréquences et la fréquence de résolution.

Étape 2 :

Elle permet d'effectuer une Analyse Discriminante pour classer toutes les voyelles dans une structure ordonnée et de créer la configuration appropriée.

Étape 3 :

Elle génère la matrice des confusions afin de vérifier la conformité de la classification prédictive avec la réalité. En général, cette matrice est de dimension n_k x n_k (où n_k est le nombre de groupes), en ligne figurent les appartenances réelles et en colonnes les affectations par le modèle. On peut y repérer le nombre d'affectations correctes et erronées. Le pourcentage d'affectations correctes par rapport au nombre total d'individus est

un indicateur global. Pour que le modèle présente un intérêt, il faut qu'il soit suffisamment élevé.

L'allure de la matrice de confusion est la suivante :

$$\begin{bmatrix} & S_1 & S_2 & S_3 \\ S_1 & & & \\ S_2 & & & \\ S_3 & & & \end{bmatrix}$$

où les lignes présentent les appartenances réelles et les colonnes les affectations obtenues par le modèle calculé. $S_{i=1,\ldots,3}$, les trois syllabes présentes dans les mots utilisés.

Étape 4 :

Elle permet d'estimer des valeurs pour des ajouts de voyelles qui n'étaient pas présentes dans l'échantillon de départ (ou d'apprentissage). On arrivera ainsi à estimer (ou prédire) les valeurs de nouvelles observations dans la classification ou regroupement déjà existant.

Étape 5 :

Elle génère la matrice de confusion correspondante.

4.6. METHODE DE MONTE-CARLO

Les chercheurs, les ingénieurs et bien d'autres professionnels se posent souvent la question : quel est le résultat que j'obtiens si j'exerce telle action sur un élément ? Le moyen le plus simple serait de tenter l'expérience, c'est-à-dire d'exercer l'action souhaitée sur l'élément en cause pour pouvoir observer ou mesurer le résultat. Dans de nombreux cas l'expérience est irréalisable, ou trop coûteuse. On a alors recours à la simulation.

La plupart des méthodes d'estimation et de test d'hypothèse ont des propriétés statistiques connues seulement asymptotiquement. Ceci est vrai pour les modèles non linéaires de tous types, pour les modèles d'équations simultanées linéaires. Ainsi, dans la pratique, la théorie exacte en échantillon fini est rarement valable pour interpréter des estimations ou des statistiques de test. Malheureusement, à moins que la taille de l'échantillon ne soit effectivement très grande, il est très difficile de savoir si la théorie asymptotique est suffisamment précise pour nous permettre d'interpréter nos résultats en toute confiance.

Il existe fondamentalement deux manières de gérer cette situation. La première est d'affiner les approximations asymptotiques en additionnant des termes d'ordre inférieur par rapport à la taille de

l'échantillon. La seconde approche, que nous exposons dans cette partie, consiste à examiner les propriétés en échantillon fini des estimateurs et des statistiques de test en utilisant les expériences Monte-Carlo. Le terme Monte-Carlo est employé dans de nombreuses disciplines et fait référence aux procédures où les quantités d'intérêt sont approximées en générant de nombreuses réalisations aléatoires d'un processus stochastique quelconque et en calculant une moyenne quelconque de leurs valeurs.

Autrement dit, la méthode de Monte Carlo est une méthode numérique, qui utilise des tirages aléatoires pour réaliser le calcul d'une quantité déterministe. Dans une expérience Monte Carlo, on génère un nombre de répétitions N. Chaque réplication consiste à générer un échantillon de données à partir duquel on calcule une réalisation d'un estimateur ou d'une statistique de test. Après N répétitions, on dispose de N réalisations de l'estimateur ou de la statistique d'intérêt. Cet ensemble de réalisations forme alors la distribution empirique de l'estimateur ou de la statistique, à partir de laquelle on peut calculer une quantité spécifique, comme par exemple un biais, un intervalle de confiance, ou un seuil critique.

La méthode de simulation de Monte-Carlo permet aussi d'introduire une approche statistique du risque dans une décision expérimentale d'un projet. Elle consiste à isoler un certain nombre de variables-clés du projet et à leur affecter une distribution de probabilités. Pour chacun de ces facteurs, on effectue un grand nombre de tirages aléatoires dans les distributions de probabilité déterminées précédemment, afin de déterminer la probabilité d'occurrence de chacun des résultats.

4.7. RESULTATS DE LA CAD

Afin de pouvoir interpréter les résultats, nous avons exploité la méthode du Bootstrap à notre corpus [81]. C'est une méthode non paramétrique, dont le but est de fournir des indications sur une statistique autre que sa valeur (dispersion, distribution, intervalles de confiance) afin de connaître la précision des estimations réalisées. Cette méthode s'organise autour d'une technique de rééchantillonnage, accompagnée d'un grand nombre d'itérations qui résultent de l'application de la méthode de Monte-Carlo (Figure 4.2) [82].

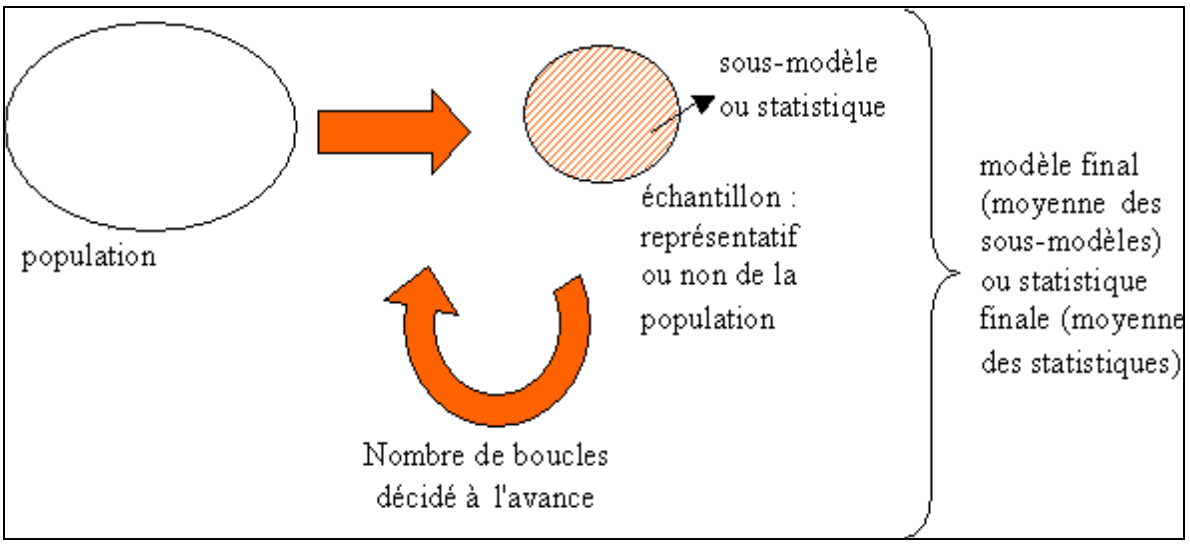


Figure 4.2 : Principe de la méthode du Bootstrap [81]

Au départ, dans notre corpus, nous avons procédé à l'apprentissage de 18 phrases parmi 20 puis à la reconnaissance des 2 phrases non incluses dans la phase d'apprentissage. Pour cela, nous avons toujours choisi d'enlever lors de l'étape d'apprentissage la même phrase mais pour deux locuteurs différents, ce qui nous donne 30 cas possibles (5 phrases x 6 combinaisons possibles).

Les données utilisées sont :

- la largeur de bande utilisée = 500 HZ, nombre de bandes utilisées = 3 ;
- H_1 : Homme (0), H_2 : Homme (1), F_1 : Femme (2), F_2 : Femme (3) ;
- 20 phrases au total (5 pour chaque locuteur) : phrase 1 à phrase 5 ;
- 18 phrases en phase d'Apprentissage (18 A) et 2 phrases en phase de Reconnaissance (2 R);
- les 2 phrases en Reconnaissance (X_1-X_2) sont les mêmes pour les deux locuteurs choisis.

La figure 4.3 représente un exemple de spectre moyen à long terme (LTAS) obtenu pour la voyelle [a] dans un mot formé de trois syllabes. On y vérifie bien que la première voyelle [a] correspondant à la syllabe accentuée S_1 (en rouge) présente des valeurs plus élevées que celles obtenues pour la deuxième syllabe S_2 (en bleu) et la troisième syllabe S_3 (en vert). Ce qui justifie, l'idée que le paramètre énergie peut être considéré comme un facteur de détection de la syllabe accentuée.

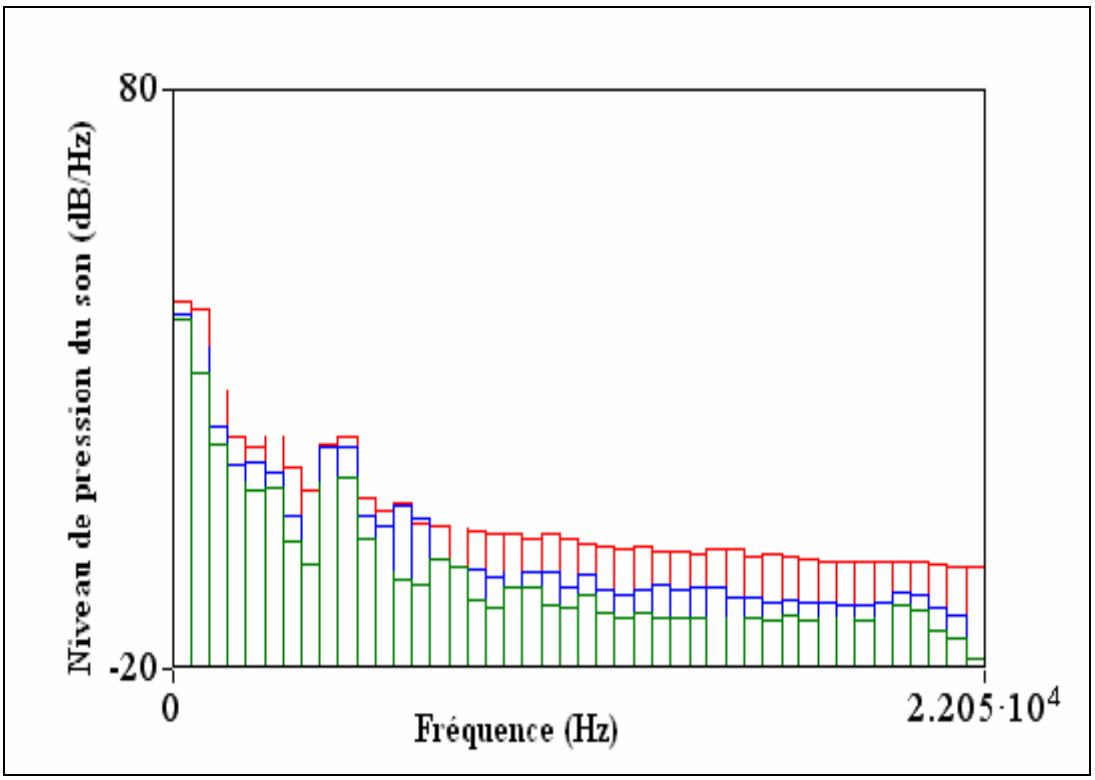


Figure 4.3 : LTAS obtenu pour un mot tri-syllabique : cas de la voyelle [a]
(S_1: rouge, S_2: bleu, S_3: vert)

La figure 4.4, montre un exemple d'ellipses de concentration obtenues après les deux étapes de classification réalisée par une analyse discriminante et de configuration. Nous notons les trois classes obtenues qui correspondent aux trois syllabes S_1, S_2 et S_3 ainsi que l'affectation de toutes les syllabes à l'intérieur de ces ellipses.

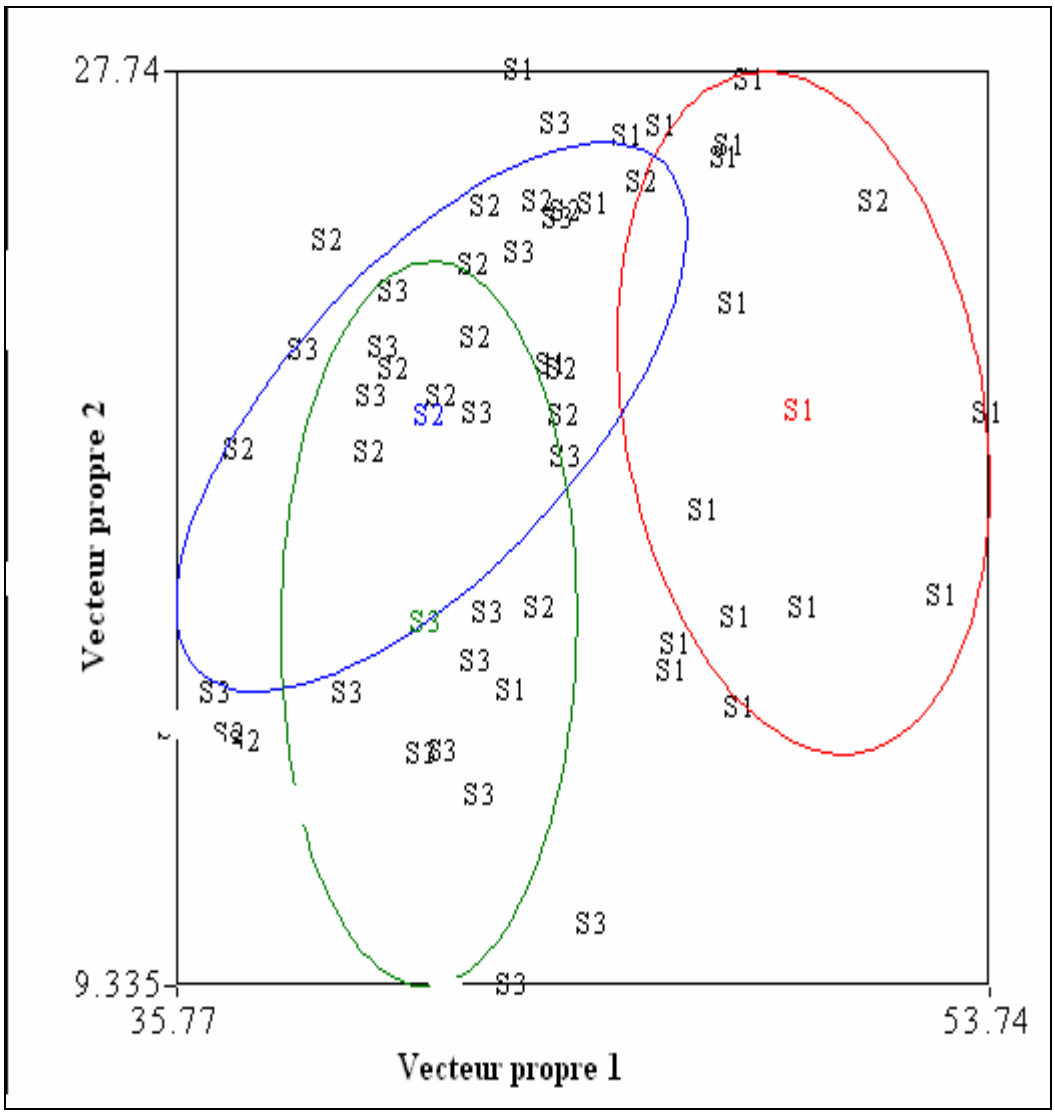


Figure 4.4 : Ellipses de concentration
(Étape d'apprentissage avec la phrase 3 enlevée pour les deux locuteurs hommes : H_1 et H_2)

La conformité de cette classification prédictive avec la réalité est alors vérifiée par le calcul de la matrice de confusion correspondante. Les calculs nous donnent la matrice de confusion suivante :

$$\begin{bmatrix} & S_1 & S_2 & S_3 \\ S_1 & 14 & 3 & 1 \\ S_2 & 2 & 12 & 4 \\ S_3 & 0 & 8 & 10 \end{bmatrix}$$

Nous remarquons que :

- pour la syllabe S_1 : 14 parmi les 18 ont été bien classées, 3 ont été classées comme étant des S_2 et une seule comme une syllabe S_3. Ce qui nous donne un pourcentage de bon classement pour cette syllabe égal à : 14 / 18 = 77.78 % ;
- pour la syllabe S_2 : 12 parmi les 18 ont été bien classées, alors que 2 ont été classées en S_1 et 4 en S_3. Ce qui nous donne un pourcentage égal à : 12 / 18 = 66.67 % ;

- pour la syllabe S_3 : 10 seulement parmi les 18 ont été bien classées, alors que 8 ont été classées en S_2. Ce qui nous donne un pourcentage égal à : 10 / 18 = 55.56 % ;
- si nous essayons d'évaluer le bon classement des trois syllabes, nous obtiendrons alors : (14+12+10) / 54 = 66.67 %.

Pour la phase de reconnaissance des deux phrases que nous n'avons pas utilisées lors de la phase d'apprentissage, nous avons obtenu la matrice de confusion suivante :

$$\begin{bmatrix} & S_1 & S_2 & S_3 \\ S_1 & 2 & 0 & 0 \\ S_2 & 2 & 0 & 0 \\ S_3 & 1 & 1 & 0 \end{bmatrix}$$

Nous observons le très bon classement de la 1ère syllabe S_1 (2 / 2 = 100 %) alors que nous remarquons le très mauvais classement des deux autres syllabes (0 %), ce qui réduit très sérieusement le bon classement total : 2 / 6 = 33.33 %.

Les résultats obtenus après exécution de notre algorithme sont représentés dans le tableau 4.3. Il est clair que nous ne pouvons rien affirmer sur le classement des voyelles et plus précisément, la 1ère voyelle correspondant à la syllabe accentuée que nous cherchons à reconnaître.

Tableau 4.3 : Matrices de confusion et les pourcentages d'affectation par paire phrase-locuteur : Phases d'apprentissage et de reconnaissance

X1 – X2	PHRASE 1		PHRASE 2		PHRASE 3		PHRASE 4		PHRASE 5	
H1 – H2	18A	15 2 1 1 11 6 1 8 9 64.8 % S1:83.33%	18A	15 3 0 3 11 4 1 8 9 64.8 % S1:83.33%	18A	14 3 1 2 12 4 0 8 10 66.7 % S1:77.78%	18A	14 3 1 3 12 3 1 8 9 64.8 % S1:77.78%	18A	13 4 1 3 11 4 2 8 8 59.3 % S1:72.22%
	2R	0 2 0 1 1 0 0 1 1 33.3 % **S1:00.0%**	2R	1 0 1 0 1 1 0 1 1 50 % **S1 : 50 %**	2R	2 0 0 2 0 0 1 1 0 33.3 % **S1: 100 %**	2R	2 0 0 1 0 1 0 1 1 50 % **S1: 100 %**	2R	2 0 0 0 1 1 0 1 1 66.7 % **S1: 100 %**
H1 – F1	18A	16 1 1 3 11 4 1 8 9 66.7 % S1:88.89%	18A	13 4 1 4 10 4 1 8 9 59.3 % S1:72.22%	18A	15 2 1 3 11 4 1 8 9 64.8 % S1:83.33%	18A	14 3 1 3 11 4 1 8 9 63 % S1:77.78%	18A	14 3 1 4 9 5 1 8 9 59.3 % S1:77.78%
	2R	0 2 0 1 0 1 0 1 1 16.7 % **S1: 00.0%**	2R	2 0 0 0 2 0 0 1 1 83.3 % **S1: 100 %**	2R	2 0 0 1 0 1 0 1 1 50 % **S1: 100 %**	2R	2 0 0 1 1 0 0 1 1 66.7 % **S1: 100 %**	2R	2 0 0 0 2 0 0 1 1 83.3 % **S1: 100 %**
H1 – F2	18A	15 2 1 3 11 4 1 8 9 64.8 % S1:83.33%	18A	13 4 1 4 10 4 1 7 10 61.1 % S1:72.22%	18A	14 3 1 3 11 4 1 7 10 64.8 % S1:77.78%	18A	14 3 1 3 11 4 1 7 10 64.8 % S1:77.78%	18A	14 3 1 4 10 4 1 7 10 63 % S1:77.78%
	2R	1 1 0 1 1 0 1 1 0 33.3 % **S1 : 50 %**	2R	2 0 0 0 2 0 0 2 0 66.7 % **S1: 100 %**	2R	2 0 0 1 1 0 0 2 0 50 % **S1: 100 %**	2R	2 0 0 1 1 0 0 2 0 50 % **S1: 100 %**	2R	2 0 0 0 2 0 0 2 0 66.7 % **S1: 100 %**
H2 – F1	18A	15 2 1 2 11 5 1 9 8 63 % S1:83.33%	18A	14 3 1 3 10 5 1 9 8 59.3 % S1:77.78%	18A	14 3 1 3 12 3 0 8 10 66.7 % S1:77.78%	18A	13 4 1 3 11 4 1 9 8 59.3 % S1:72.22%	18A	14 3 1 3 10 5 2 9 7 57.4 % S1:77.78%
	2R	0 2 0 0 1 1 0 0 2 50 % **S1:00.0 %**	2R	1 0 1 0 0 2 0 0 2 50 % **S1 : 50 %**	2R	2 0 0 1 0 1 1 0 1 50 % **S1: 100 %**	2R	2 0 0 0 1 1 0 0 2 83.3 % **S1 : 100%**	2R	2 0 0 0 1 1 0 0 2 83.3 % **S1: 100 %**
H2 – F2	18A	15 2 1 2 11 5 1 9 8 63 % S1:83.33%	18A	14 3 1 3 11 4 2 8 8 61.1 % S1:77.78%	18A	13 4 1 3 11 4 0 8 10 63 % S1:72.22%	18A	14 3 1 3 11 4 2 8 8 61.1 % S1:77.78%	18A	14 3 1 3 10 5 2 7 9 66.7 % S1:77.78%
	2R	0 2 0 0 2 0 1 0 1 50 % **S1:00.0 %**	2R	1 0 1 0 1 1 0 1 1 50 % **S1 : 50 %**	2R	2 0 0 1 1 0 1 1 0 50 % **S1: 100 %**	2R	2 0 0 0 1 1 0 1 1 66.7 % **S1: 100 %**	2R	2 0 0 0 1 1 0 1 1 66.7 % **S1: 100 %**
F1 – F2	18A	15 2 1 3 10 5 1 9 8 61.1 % S1:83.33%	18A	14 3 1 4 10 4 1 8 9 61.1 % S1:77.78%	18A	13 4 1 4 11 3 1 8 9 61.1 % S1:72.22%	18A	15 2 1 3 9 6 1 8 9 61.1 % S1:83.33%	18A	14 3 1 4 10 4 2 8 8 59.3 % S1:77.78%
	2R	1 1 0 0 2 0 1 0 1 66.7 % **S1 : 50 %**	2R	2 0 0 0 2 0 0 1 1 83.3 % **S1: 100 %**	2R	2 0 0 0 1 1 0 1 1 66.7 % **S1: 100 %**	2R	2 0 0 0 2 0 0 1 1 83.3 % **S1: 100 %**	2R	2 0 0 0 2 0 0 1 1 83.3 % **S1: 100 %**

Pour cela, nous avons procédé au calcul de la matrice de confusion totale pour chaque phrase ainsi que le pourcentage d'affectations correctes. Nous obtenons donc 18 phrases x C_2^4 = 108 phrases en Apprentissage et 2 phrases x C_2^4 = 12 phrases en Reconnaissance. Nous avons alors obtenu le tableau 4.4.

Tableau 4.4 : Matrices de confusion et les pourcentages d'affectation de chaque phrase : Phases d'apprentissage et de reconnaissance

	PHRASE 1	**PHRASE 2**	**PHRASE 3**	**PHRASE 4**	**PHRASE 5**
108 A	91 11 6 14 65 29 6 51 51	83 20 5 21 62 25 7 48 53	83 19 6 18 68 22 3 47 58	84 18 6 18 65 25 7 48 53	83 19 6 21 60 27 10 47 51
	96.3 % **S1:84.26%**	61.11 % **S1:76.85%**	64.51 % **S1:76.85%**	62.35 % **S1:77.78%**	59.88 % **S1:76.85%**
12 R	2 10 0 3 7 2 3 3 6	9 0 3 0 8 4 0 6 6	12 0 0 6 3 3 3 6 3	12 0 0 3 6 3 0 6 6	12 0 0 0 9 3 0 6 6
	41.67 % **S1:16.67%**	63.89 % **S1 : 75 %**	50 % **S1: 100 %**	66.67 % **S1: 100 %**	75 % **S1: 100 %**

D'après ce tableau, nous remarquons :

- le bon classement obtenu dans la phase d'apprentissage des différentes voyelles (> 50 %) et surtout le très bon classement de la première voyelle correspondant à la syllabe accentuée (>70 %) ;
- le mauvais classement de la phrase 1, lors de la phase de test ou de reconnaissance. On justifie ce faible résultat au fait que cette phrase a été prononcée en premier par nos 4 locuteurs ;
- le très bon pourcentage de reconnaissance de la voyelle correspondant à la syllabe accentuée (égal à 100 %) pour les trois dernières phrases lors de la phase de test.

Afin de conclure sur l'efficacité de la méthode utilisée, nous avons calculé la matrice de confusion totale correspondant au corpus testé. Soient 108 x 5 = 540 A (Apprises) et 12 x 5 = 60 R (Reconnues). Nous avons alors, obtenu le tableau 4.5.

Tableau 4.5 : Matrices de confusion et les pourcentages d'affectation totaux : Phases d'apprentissage et de reconnaissance

Phrases en Apprentissage		**Phrases en Reconnaissance**	
540 A	424 87 29 92 320 128 33 241 266	**60 R**	47 10 3 12 33 15 6 27 27
	62.35 % **S1 : 78.52 %** **S2 : 59.26 %** **S3 : 49.26 %**		59.44 % **S1 : 78.33 %** **S2 : 55 %** **S3 : 45 %**

Ce qui nous permet de conclure que la phase :

- d'apprentissage donne un bon pourcentage de classification (62.35 %). Il est clair que ce sont les taux de bonne classification obtenus pour les deux syllabes inaccentuées (S_2 et S_3) qui sont à l'origine de cet abaissement (59.26 % pour S_2 et 49.26 % pour S_3). La syllabe accentuée S_1 est classée avec un taux égal à 78.52 % ;
- de reconnaissance est très légèrement supérieure au seuil correspondant à un classement aléatoire. Cependant, nous notons le très bon classement de la syllabe S_1 (78.33 %).

Sachant que le nombre de bandes utilisées joue un rôle important sur la performance du calcul du Spectre Moyen à Long Terme (LTAS), nous avons procédé dans une deuxième étape à augmenter ce nombre. Nous avons alors constaté expérimentalement, qu'en passant des trois bandes utilisées précédemment à 6 bandes et en gardant toutes les autres hypothèses inchangées, nous obtenons de meilleurs résultats (Tableau 4.6).

Tableau 4.6 : Matrices de confusion et les pourcentages d'affectation totaux :
(cas de 6 bandes)

Phrases en Apprentissage		**Phrases en Reconnaissance**	
540 A	483 30 27 32 475 33 26 170 344	**60 R**	51 6 3 4 44 12 5 18 37
	80.37 % **S1 : 89.44 %** **S2 : 87.96 %** **S3 : 63.70 %**		73.33 % **S1 : 85 %** **S2 : 73.33 %** **S3 : 61.67 %**

D'après ces résultats, nous constatons que :

- le taux d'apprentissage est beaucoup meilleur, il est passé de 62.35 % à 80.37 %. Le taux d'apprentissage de la syllabe accentuée S_1 est passé de 78.52 % à 89.44 %. Les taux d'apprentissage des deux syllabes S_2 et S_3 n'avoisinent plus le seuil d'un classement aléatoire ;
- on ne peut plus parler de classement aléatoire pour la phase de reconnaissance. Le taux de bon classement est passé de 59.44 % à 73.33 % ;
- la syllabe accentuée S_1 présente un très bon taux de classement en phase de reconnaissance (de 78.33 % à 85 %), de même pour la syllabe S_2 (de 55 % à 73.33 %) ;
- bien que le taux de la syllabe S_3 s'est beaucoup amélioré (de 45 % à 61.67 %), nous pensons qu'il reste toujours assez faible.

4.8. VERSIONS AMELIOREES DE LA CAD

Dans un souci d'amélioration des résultats obtenus précédemment, nous allons proposer dans cette partie deux versions améliorées de la CAD.

4.8.1. Version 1 : Énergie - Fréquence fondamentale

Nous avons repris le script que nous avons développé pour la méthode de Classification par Analyse Discriminante (CAD) du paramètre acoustique Energie. En plus du calcul du spectre moyen à long terme, nous avons mesuré pour chaque voyelle correspondant à chaque syllabe, la valeur moyenne de la fréquence fondamentale. Ceci, a fait augmenté les valeurs à analyser de six (6 valeurs LTAS calculées) à 7 valeurs (6 valeurs LTAS + F_0) [83]. Nous avons procédé à l'exécution de notre nouveau script, en respectant les mêmes conditions précédentes et en appliquant toujours le principe de la méthode de Monte-Carlo et la méthode du bootstrap. Nous avons alors obtenu le tableau 4.7 suivant :

Tableau 4.7 : (Version 1) Matrices de confusion et les pourcentages d'affectation totaux

Phrases en Apprentissage		**Phrases en Reconnaissance**	
540 A	503 37 0 34 473 33 10 51 479	**60 R**	51 9 0 5 49 6 3 7 50
	89.81 % **S1 : 93.15 %** **S2 : 87.59 %** **S3 : 88.70 %**		83.33 % **S1 : 85 %** **S2 : 81.67 %** **S3 : 83.33 %**

D'après ce tableau, nous remarquons que :

- le fait d'avoir introduit le critère de la fréquence fondamentale comme élément discriminant supplémentaire, nous a fourni de très bons résultats dans la phase d'apprentissage (de 80.37% à 89.81 %) et celle de la reconnaissance aussi (de 73.33 % à 83.33 %) ;
- cette fois-ci, la classification est bonne pour les trois syllabes, c'est-à-dire que la discrimination entre les trois groupes est très nette ;
- cependant, nous notons que le taux de reconnaissance de la première syllabe est resté inchangé.

4.8.2. Version 2 : Energie - Fréquence fondamentale et Durée

En notant que la version 1, n'a pas amélioré la classification de la première syllabe, nous avons pensé à utiliser comme paramètre supplémentaire la durée segmentale de la voyelle détectée à l'intérieur de chaque syllabe. Notre nouveau script calcule donc, en plus des 6 valeurs du LTAS et

de la valeur de la fréquence fondamentale, F_0, la valeur de la durée D. Ceci nous donne 8 valeurs au total à analyser. Nous avons obtenu le tableau 4.8.

Tableau 4.8 : (Version 2) Matrices de confusion et les pourcentages d'affectation totaux

Phrases en Apprentissage		**Phrases en Reconnaissance**	
540 A	504 36 0 35 480 25 0 0 540	**60 R**	46 14 0 9 48 3 0 0 60
	94.07 % **S1 : 93.33 %** **S2 : 88.89 %** **S3 : 100 %**		85.56 % **S1 : 76.67 %** **S2 : 80 %** **S3 : 100 %**

D'après les résultats obtenus, il est clair que la version 2 ne change en rien les résultats précédents pour la syllabe accentuée S_1. Cependant, nous observons que pour la troisième syllabe, nous obtenons des taux de bon classement en phase d'apprentissage et en phase de reconnaissance, égaux à 100 %, ce qui a attribué à l'augmentation des taux de bon classement en phase d'apprentissage (de 89.81 % à 94.07 %) et en phase de reconnaissance (de 83.33 % à 85.56 %).

Nous justifions ces résultats par le fait que nous travaillons sur des mots trisyllabiques, la voyelle correspondant à la troisième syllabe (fin du mot) présente une durée segmentale plus grande que celles des deux autres syllabes, ce qui présente un caractère discriminant supplémentaire pour le classement de la troisième syllabe. Alors que pour la première et la deuxième syllabe, les durées segmentales des voyelles sont proches ce qui nous permet de justifier la non amélioration des résultats correspondants à ces deux syllabes.

Les tableaux 4.9 et 4.10, justifient nos constatations (cas simple d'un apprentissage de 18 phrases pour la reconnaissance de la $5^{ème}$ phrase pour le $2^{ème}$ locuteur homme (H_2) et la $1^{ère}$ femme (F_1). Nous y remarquons que dans la plupart des cas, la troisième syllabe S_3 présente une durée, de la voyelle correspondante, plus grande que les deux autres syllabes.

A la fin, nous pouvons dire que l'ajout de la variable Durée n'affecte en rien les performances de la méthode de la Classification par Analyse Discriminante du point de vue détection de l'accent primaire. Son intérêt réside seulement dans la détection de l'accent tertiaire du mot à analyser.

Tableau 4.9 : Valeurs calculées du paramètre Durée
(phase d'apprentissage)

Syllabe	Durée (s)	Syllabe	Durée (s)	Syllabe	Durée (s)	Syllabe	Durée (s)	Syllabe	Durée (s)
S_1	0.1004	S_3	0.1291	S_2	0.1151	S_1	0.1095	S_3	0.1835
S_2	0.0698	S_1	0.0925	S_3	0.1798	S_2	0.0790	S1	0.0912
S_3	0.1096	S_2	0.0642	S_1	0.1188	S_3	0.1466	S_2	0.0863
S_1	0.0735	S_3	0.1654	S_2	0.1240	S_1	0.0838	S_3	0.1658
S_2	0.0824	S_1	0.0831	S_3	0.1860	S_2	0.1166	S_1	0.1128
S_3	0.1170	S_2	0.1206	S_1	0.0551	S_3	0.1484	S_2	0.0940
S_1	0.1093	S_3	0.1733	S_2	0.0544	S_1	0.1094	S_3	0.1573
S_2	0.0710	S_1	0.1259	S_3	0.1459	S_2	0.0823	S_1	0.0846
S_3	0.1289	S_2	0.1341	S_1	0.0593	S_3	0.1421	S_2	0.0907
S_1	0.0936	S_3	0.1894	S_2	0.0908	S_1	0.0927	S_3	0.1407
S_2	0.0655	S_1	0.0911	S_3	0.1510	S_2	0.0898		

Tableau 4.10 : Valeurs calculées du paramètre Durée
(phase de reconnaissance)

Syllabe	Durée (s)
S1	0.1092
S2	0.0861
S3	0.1074
S1	0.0697
S2	0.0845
S3	0.1342

4.9. CONCLUSION

Dans ce chapitre, nous avons donné un aperçu général sur l'accent et la syllabe en AS, ainsi que la Classification par Analyse Discriminante. Nous avons présenté une nouvelle approche de la détection de l'Accent Primaire en AS basée sur une Classification par Analyse Discriminante (CAD) du paramètre acoustique énergie. La méthode est simple à implémenter et présente des résultats encourageants. Ensuite, dans le but d'améliorer les résultats obtenus, nous avons dans une première étape, augmenté le nombre de bandes pour le calcul du spectre moyen à long terme, ce qui nous a permis d'améliorer les résultats obtenus. Puis dans une deuxième étape, nous avons présenté deux versions améliorées de la CAD, en y ajoutant la fréquence fondamentale et la Durée segmentale.

CHAPITRE 5 :

DETERMINATION DE L'EFFET

MICROPROSODIQUE EN ARABE STANDARD (AS)

5.1. INTRODUCTION

Dans le but d'extraire l'effet microprosodique pour une éventuelle amélioration du naturel de la parole synthétisée, nous allons présenter dans ce chapitre, une nouvelle méthode de modélisation de l'effet microprosodique en AS, grâce au logiciel d'analyse de la parole PRAAT [77] et à l'algorithme MOMEL [84]. Ce dernier permet d'obtenir une représentation de la courbe mélodique, caractérisant les variations temporelles de la fréquence laryngienne, via une approximation quadratique (Quadratic Spline Function). Développé dès 1993, cet algorithme transforme la courbe discontinue issue de la détection brute de la fréquence fondamentale en une courbe continue conçue comme sa résultante intonativement pertinente. Le rôle de l'algorithme est de séparer la composante macroprosodique de la composante microprosodique, écartée comme linguistiquement non pertinente. En sortie, MOMEL génère un ensemble de points définis par un couple localisation temporelle/F_0. Ces points sont ensuite reliés par une fonction spline quadratique dont ils représentent les zones d'inflexion (les sommets et les vallées).

5.2. ETUDE DE LA MELODIE

La mélodie est le résultat des vibrations des cordes vocales qui se traduit par la variation de la fréquence fondamentale F_0 en fonction du temps. Beaucoup d'études phonétiques sur la prosodie se limitent à analyser les variations de F_0 car les trois paramètres prosodiques sont fortement liés et complexes à étudier ensemble. On peut dire de manière générale que chez l'homme adulte la gamme de variations de F_0 est située entre 100 Hz et 150 Hz et la femme adulte de 140 Hz à 250 Hz. Des variations considérables peuvent être relevées selon le locuteur, son âge, son état émotif, l'acte de parole, etc.

Au niveau global (groupe prosodique, phrase), l'évolution de la fréquence fondamentale est essentiellement influencée par la position de la syllabe dans le groupe prosodique, la fonction du groupe prosodique dans la phrase et le mode de la phrase (interrogatif, déclaratif…). On qualifie l'ensemble de ces phénomènes globaux de macromélodie.

Au niveau local (disons de taille inférieure à la syllabe), les facteurs importants sont la nature du phonème, sa position dans la syllabe et son environnement phonétique immédiat : on parle, pour cet ensemble de phénomènes locaux, de micromélodie. On remarque alors, des évolutions de F_0 caractéristiques pour des émissions de voyelles ou de consonnes voisées particulières. Les consonnes non voisées marquent une rupture dans la ligne mélodique comme pour les pauses ou les silences.

Pour tenir compte de cette hiérarchie entre contextes locaux et globaux, les systèmes de synthèse procèdent généralement en deux étapes : dans une première passe, on spécifie le contour mélodique global (macromélodie), par des modèles élémentaires au niveau de chaque groupe prosodique. On affine ensuite ce contour en appliquant les contraintes locales (la micromélodie) [85].

Si nous tentons d'établir la nature des relations qui lient les variations de la fréquence fondamentale (F_0) à l'actualisation des entités phonologiques, lexicale et supralexicale que représentent le ton, l'accent et l'intonation, il s'avère utile, voire indispensable, de connaître la nature et l'importance des contraintes de production qui affectent l'évolution de ce paramètre [86]. Ces contraintes, qui sont associées à plusieurs facteurs, engendrent des effets universels (quoi que quantitativement variables d'une langue à l'autre) dont nous décrivons, certaines d'elles, brièvement.

5.2.1. Les contraintes idiosyncrasiques

Il est clairement établi que les variations de F_0 sont en partie déterminées par les caractéristiques physiologiques des locuteurs (notamment par la masse volumique des cordes vocales) et que par conséquent la tessiture tonale peut varier d'un sujet à l'autre, et varie de toute évidence systématiquement entre les voix d'homme, de femme et d'enfant. Cette source de variations non linguistique, appelée idiosyncrasie, étant reconnue, il importe de la neutraliser avant toute interprétation linguistique, ce qui peut être réalisé par la mise en oeuvre de procédures de normalisation. Ces dernières consistent principalement à convertir, à l'aide de formules appropriées, les valeurs absolues de F_0 mesurées en Hertz (Hz) en valeurs relatives (ou en valeurs logarithmiques). Plusieurs échelles sont alors disponibles à cet effet, telles que l'échelle des demi-tons ou l'échelle ERB (Equivalent Rectangular Bandwidth) [56].

5.2.2. Les contraintes interactives : effets microprosodiques

Les variations de F_0, en particulier, et celles des paramètres prosodiques physiques, en général, sont soumises à des contraintes de production, interactives, qui résultent dans ce cas d'une interaction entre ces paramètres et la prononciation du matériau segmental. Les phénomènes engendrés par ce type de contrainte sont appelés microprosodiques, car leur largeur est très limitée et n'excède pas la taille du segment phonémique. Il existe deux sortes de phénomènes microprosodiques, qui sont qualifiés par les termes : intrinsèques et co-intrinsèques [87].

Les effets intrinsèques sont dénommés ainsi parce qu'ils concernent l'influence locale que peut exercer la production de certains segments (voyelles ou consonnes) sur la configuration et les valeurs d'un paramètre prosodique physique donné.

Les phénomènes co-intrinsèques doivent leur appellation au fait qu'ils relèvent de la coarticulation, en l'occurrence de l'influence des consonnes sur les caractéristiques prosodiques des voyelles adjacentes.

Nous retiendrons la définition donnée par A. Di Cristo [88], « Par microprosodie ou faits microprosodiques, nous entendons les variations de fréquence fondamentale, de durée et d'intensité des unités segmentales, attribuable à la nature acoustique spécifique de ces unités ainsi qu'aux effets inhérents aux phénomènes de coarticulation. Nous désignerons par micromélodie l'ensemble des phénomènes microprosodiques relatifs à la seule variation de la fréquence fondamentale ».

5.3. MODELISATION MELODIQUE : MOMEL

L'algorithme de modélisation automatique (MOMEL) permet la représentation de la fréquence fondamentale par une séquence de points cibles constituée par des couples de valeurs (F_0, temps). Les points cibles correspondent aux variations locales pertinentes de la courbe mélodique et permettent, interpolés par une fonction de type spline quadratique, de retrouver le profil suprasegmental caractérisant globalement l'intonation.

5.3.1. Présentation de MOMEL

Le script PRAAT de MOMEL proposé par D. J. Hirst et R. Espesser permet d'obtenir une représentation de la courbe mélodique, caractérisant les variations temporelles de la fréquence laryngienne, via une approximation quadratique (Quadratic Spline Function). Le principe de base de cette approche est de considérer les variations de F_0 comme étant la superposition de deux phénomènes distincts, la macroprosodie (qui caractérise le choix intonatif de l'élocution) et la microprosodie (propre aux phonèmes constituant le lexique de la phrase). La macroprosodie permet donc d'avoir une approche globale de la courbe mélodique, la microprosodie se caractérisant quant à elle par des variations locales des paramètres prosodiques.

L'algorithme de modélisation automatique MOMEL, permet donc la représentation de la fréquence fondamentale par une séquence de points cibles, constituée par des couples de valeurs (F_0, temps). Les points cibles correspondent aux variations locales pertinentes de la courbe mélodique et permettent (interpolés par une fonction de type spline quadratique) de retrouver le profil suprasegmental caractérisant globalement l'intonation. La synthèse par PSOLA (Pitch

Synchronous OverLap and Add) d'un signal original, et d'un signal dont seule la courbe de F_0 a été remplacée par sa modélisation quadratique, permet l'obtention de deux signaux perceptivement semblables : rares sont les endroits où une différence peut être remarquée. L'approche qui est développée dans l'algorithme MOMEL, permet donc bien le filtrage microprosodique recherché, conservant intacte la caractéristique macroprosodique, porteuse des informations que sont l'intonation et l'accentuation [89].

5.3.2. Principe de l'algorithme MOMEL

MOMEL repose sur l'acceptation de l'hypothèse suivante : la courbe mélodique peut être approximée par morceaux par une fonction quadratique (polynôme du $2^{ème}$ degré).

Pour ce faire, une fenêtre glissante de taille A (typiquement 300 ms) parcourt de manière chronologique le signal acoustique. Sur chaque fenêtre est extraite la courbe de F_0, et les paramètres d'un polynôme du second degré sont calculés de manière à minimiser l'erreur quadratique de la distance entre la courbe de F_0 et son approximation polynomiale. Les points de la courbe initiale situés à plus de 5% (paramètre delta) au-dessous du polynôme sont supprimés. Un nouveau polynôme est alors calculé, et ainsi de suite jusqu'à accepter tous les points restants. A cet instant, le sommet de la parabole (annulation de la dérivée du polynôme), s'il appartient bien en abscisse à la fenêtre A, et en ordonnée à l'intervalle $[H_{zmin}, H_{zmax}]$ (respectivement valeurs de F_0 minimum et de F_0 maximum, acceptées) est sauvegardé : il est appelé candidat.

L'étape suivante consiste donc à extraire de ces candidats les points qui constitueront les points cibles (extrema) de la courbe finale. Pour ce faire, l'espace temporel est partitionné de manière à isoler chacun des lieux où se regroupent les points d'inflexion. Sur chaque partition sont calculés la moyenne des candidats, et l'écart type des réalisations (écart type en temps et en fréquence). Les valeurs éloignées des valeurs moyennes d'une distance supérieure à l'écart type (soit en temps, soit en fréquence) sont supprimées, et la moyenne est alors recalculée sur les points restants. Le point cible final correspond donc à cette moyenne en temps et en fréquence.

Le processus est répété sur chacun des espaces de partition, et l'on obtient donc un ensemble de points cibles, sur lesquels on effectue une régression quadratique, qui correspond à l'approximation mélodique désirée [90].

Les paramètres suivants sont modifiables dans l'appel du programme :

- H_{zmin} : valeur minimale de F_0 acceptée ;
- H_{zmax} : valeur maximale de F_0 acceptée ;

- A : taille de la fenêtre initiale (par défaut 300 ms) ;
- Delta : pourcentage maximal d'erreur acceptée par l'approximation polynomiale ;
- R : taille de la fenêtre pour le choix des partitions (par défaut 200 ms).

La figure 5.1 représente une courbe de F_0, sur laquelle ont été disposés les candidats (points noirs en gras), les points cibles (ronds en verts) et l'approximation quadratique MOMEL en rouge.

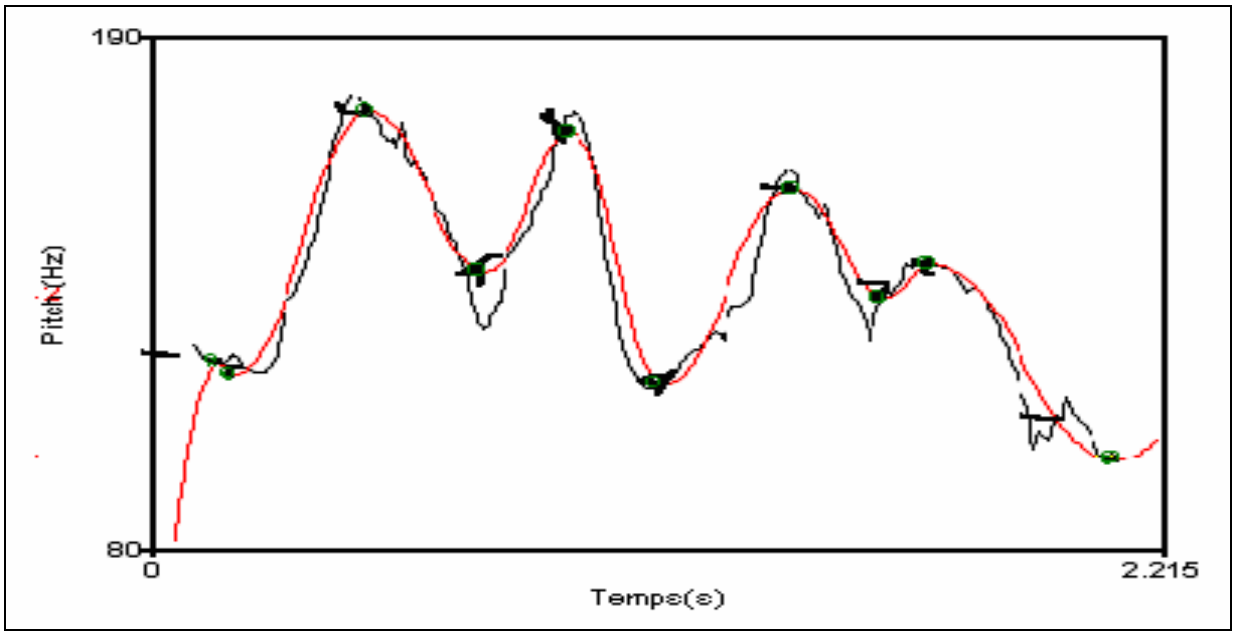


Figure 5.1 : Courbe de F_0 et sa modélisation mélodique MOMEL

5.4. METHODE D'EXTRACTION DE L'EFFET MICROPROSODIQUE (EEM)

Si la preuve n'est plus à faire de l'existence des variations microprosodiques, il demeure néanmoins que 1 'accord n'est pas unanime quant à la place qu'il convient de leur accorder dans les études prosodiques. En effet, les chercheurs ne s'accordent pas sur la nécessité de prendre en considération de façon distincte, ou de départager les rôles respectifs des variations microprosodiques et des variations macroprosodiques porteuses d'informations linguistiquement pertinentes [91].

A partir du moment où les chercheurs s'accordent pour reconnaître l'existence des variations intrinsèques et co-intrinsèques, il nous a semblé imprudent de les écarter. Nous allons chercher à évaluer leur importance en Arabe Standard, pour juger ensuite de la pertinence de les retenir ou non.

Etant admis que les effets microprosodiques qui résultent des contraintes de production sont relativement stables, il est possible de les dissocier de la composante macroprosodique (qui reflète

les variations de F_0 intentionnellement motivées) au moyen d'une méthode de calcul appropriée. C'est ainsi que l'usage de l'algorithme MOMEL [84] permet d'extraire d'une courbe brute de F_0, le profil suprasegmental qui découle de la composante macroprosodique.

La méthode que nous allons présenter ci-dessous, va nous permettre de retrouver l'information micromélodique qui a été filtrée lors de l'exécution de l'algorithme MOMEL sur la courbe mélodique réelle, pour le cas des consonnes voisées en AS.

5.4.1. Corpus et Matériel utilisés

Pour la méthode d'Extraction de l'Effet Microprosodique (EFM), nous faisons énoncer notre cinquième locuteur arabophone (1 femme), 16 phrases englobant tous les phonèmes de l'AS (Tableau 5.1).

Tableau 5.1 : Phrases prononcées par le 5ème locuteur

Phrases	
الطِفْلُ الذَكِيُّ شَرِبَ الحَلِيبَ اللَّذِيذَ.	ذَهَبَتْ جَارَةُ أَحْمَدِ إِلَى الجَزَائِرِ.
أَضَاعَتْ صَابِرَةُ نَظَارَةً فَوْقَ المِنْضَدَةِ.	وَجَدَ شَعْبَانُ خَرُوفًا.
قَرَأَتْ زَيْنَبُ عِشْرِينَ كِتَابًا خِلاَلَ العُطْلَةِ.	أُمُّ البَطَلِ عُمَرْ رَافَقَتْهُ فِي سَفَرِهِ مِنْ فَاسْ إِلَى تُونِسْ.
أَحْمَدُ يَأْكُلُ دَائِمًا فِي نَفْسِ المَطْعَمِ.	سُكَانُ المَغْرِبِ أَكْثَرُ عَدَدٍ مِنْ سُكَانِ الجَزَائِرِ بِكَثِيرٍ.
يَعْتَبِرُ جَمَالُ أُسْتَاذَ اللُّغَةِ شَخْصًا غَيْرَ مَرْغُوبٍ فِيهِ.	وَزَّعَ سَاعِي البَرِيدِ ظَرْفًا وَاحِدًا اليَوْمَ.
شَاهَدَ خَالِدٌ السَيَارَةَ المَسْرُوقَةَ أَمَامَ مَكْتَبِ البَرِيدِ.	وَضَعَ الوَلَدُ الصَّغِيرُ مِحْفَظَتَهُ فَوْقَ الطَاوِلَةِ
قَابَلَ مُصْطَفَى المُرَاسِلَ نَفْسَهُ.	تَتَسَبَبُ الغَازَاتُ المُنْبَعِثَةُ مِنَ المَصَانِعِ فِي تَلَوُثِ الهَوَاءِ.
وَضَعَ تَابِرُ العِقْدَ حَوْلَ جِسْمِ خَرُوفِهِ.	شَارَكَتْ خَوْلَةُ وَ خَدِيجَةُ فِي الحَفْلِ.

Ces phrases ont été enregistrées puis traitées grâce au logiciel de transcription et d'analyse phonétique PRAAT. Ainsi ces phrases, après avoir subi l'opération de segmentation, d'alignement semi-automatique et une TOP, vont subir les différentes étapes amenant à l'extraction de l'effet microprosodique recherché.

5.4.2. Méthodologie de l'EEM

Afin d'aboutir à l'extraction automatique de l'effet microprosodique des différentes consonnes voisées présentes dans notre corpus, nous avons suivi la démarche suivante [92] :

- l'exécution de l'algorithme MOMEL au corpus segmenté et transcrit phonétiquement afin d'en extraire les courbes mélodiques correspondantes ;
- nous procédons ensuite à la correction manuelle des courbes mélodiques modélisées par MOMEL ;
- nous exécutons alors sous PRAAT, le 1er programme que nous avons développé et qui va nous permettre de déterminer le profil microprosodique recherché ;
- une fois les données mélodiques représentant l'effet microprosodique calculées, nous exécutons sous PRAAT, le 2ème programme que nous avons développé pour modéliser le profil microprosodique pour chaque consonne voisée étudiée ;
- nous transférons alors les valeurs trouvées vers le logiciel Excel, afin de calculer les valeurs médianes et tracer les différentes courbes correspondantes.

Algorithme du 1er programme : Création du profil microprosodique

Afin d'extraire l'effet microprosodique, nous avons exécuté, pour chaque fichier son traité auparavant par MOMEL, les différentes opérations suivantes :

- lecture de la valeur maximale et la valeur minimale de F_0, déjà enregistrées dans leurs fichiers respectifs ;
- lecture des valeurs réelles de F_0 enregistrées dans le fichier dont l'extension est *.Hz ;
- lecture des valeurs cibles calculées par MOMEL à partir du fichier d'extension *.PitchTier ;
- réaliser une interpolation quadratique des valeurs cibles lues ;
- déductions à partir de Momel, des valeurs de F_0 correspondantes (F_0_Momel) ;
- calcul de l'effet microprosodique grâce au rapport : $\mathbf{mpp} = \frac{\mathbf{F_0}}{\mathbf{F_0\_Momel}}$ (mpp = microprosodic profile point) ;
- enregistrement des résultats dans un fichier d'extension *.mpp.

Algorithme du 2ème programme : Modélisation du profil microprosodique

Partant du principe que si on avait réalisé une très bonne modélisation mélodique grâce à MOMEL, le rapport mpp varie entre 0 et 1. La valeur nulle est obtenue lorsqu'on est en face d'une consonne non voisée, ce qui correspond à l'absence de F_0, alors que la valeur 1 correspondrait au cas d'une voyelle [V].

Afin de modéliser avec précision l'évolution de l'effet microprosodique d'une consonne voisée, nous avons adopté la démarche suivante : partant du principe que chaque consonne étudiée est comprise entre deux voyelles, nous procédons à l'extraction des mesures de l'effet microprosodique de F_0 aux points suivants (Figure 5.2).

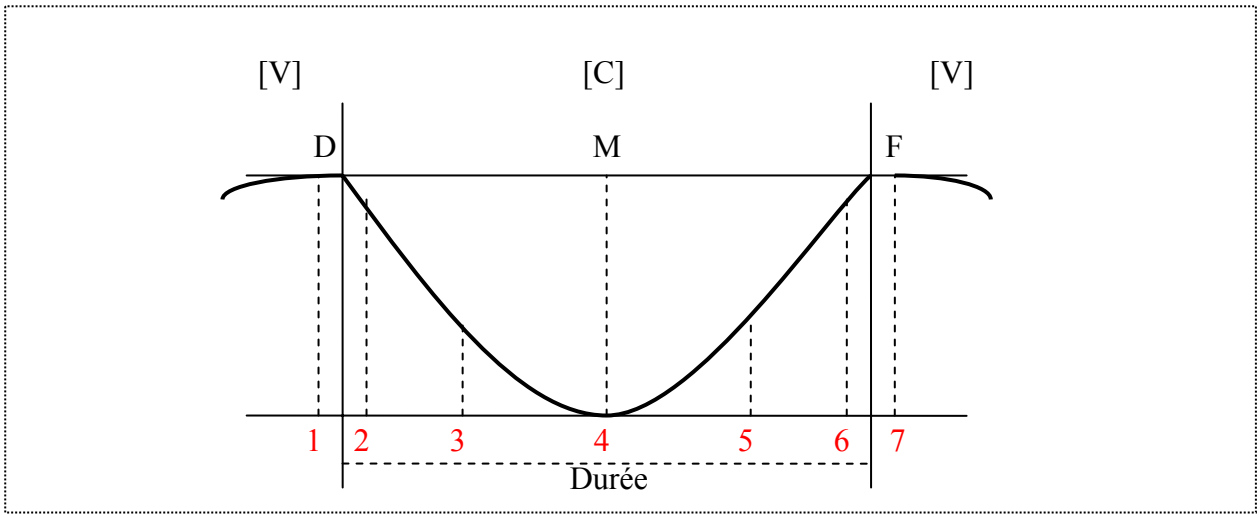


Figure 5.2 : Représentation Schématique des points retenus pour évaluer les variations microprosodiques de F_0

A noter, que nous avons considéré pour notre étude, seulement le cas où la consonne étudiée n'est ni au début d'un mot, ni géminée, ni à la fin d'un mot se terminant par une soukoun.

Une fois la consonne [C] à étudier est détectée, nous commençons par calculer la Durée totale correspondante. Nous passons ensuite à l'extraction de sept points bien précis sur cette Durée. Pour cela, nous commençons tout d'abord à lire les deux valeurs extrêmes délimitant respectivement le Début de la durée temporelle de la consonne [C], soit D, et la Fin de la durée temporelle de [C], soit F. Une fois ces deux valeurs connues, nous passons au calcul des autres points retenus pour la suite de notre modélisation.

Pour cela, nous avons procédé comme suit :

- le premier point calculé est alors le point Milieu M de la Durée totale de notre consonne [C], soit $M = \frac{F - D}{2}$, il correspond alors au point ″4″ ;
- partant du principe qu'il puisse y avoir une erreur sur la localisation des frontières extrêmes (D et F) de [C] avec les deux voyelles adjacentes, nous calculons alors deux valeurs de part et d'autre de ces deux points. Nous obtenons alors 4 nouvelles valeurs, point ″1″ = D-Δ,

point ″2″ = D+ Δ, point ″6″ = F - Δ et point ″7″ = F + Δ avec Δ un paramètre positif défini par l'utilisateur ;

- et afin de pouvoir détecter l'évolution de la courbure de l'évolution mélodique de la consonne [C], nous avons considéré deux points supplémentaires, le premier détecté à un quart de la Durée totale, soit le point ″3″ = Durée * 25 %, et le deuxième à trois quart de la Durée totale, soit le point ″5″ = Durée * 75 %.

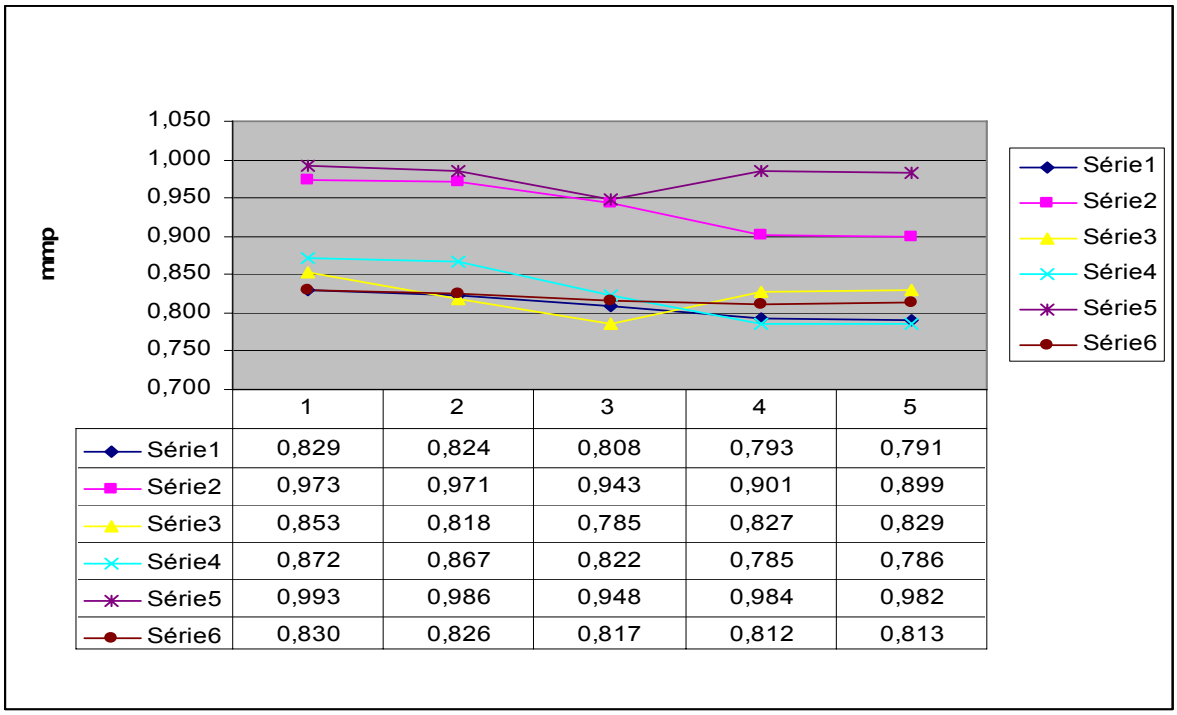
Nous avons donc procédé, pour chaque fichier son traité par le 1er programme, à l'exécution du 2ème programme, selon les étapes suivantes :

- lecture du fichier son (*.wav) ;
- lecture du fichier de transcription phonétique correspondant (*.TextGrid) ;
- lecture du fichier du profil microprosodique correspondant (*.mpp) ;
- détection de la consonne recherchée ;
- détection des points D et F correspondants ;
- calcul de la durée de la consonne détectée ;
- calcul des temps correspondants aux 7 points définis sur la figure 5.2 ;
- déduire les valeurs microprosodiques correspondant aux 7 points calculés ;
- enregistrer les résultats obtenus dans un fichier (*.dat) afin de l'exploiter par la suite par Excel.

5.5. RESULTATS DE L'EEM

Sachant que chaque consonne étudiée peut apparaître plusieurs fois tout au long du corpus utilisé, nous avons alors opté pour le calcul de la médiane de chacun des 7 points correspondant au profil microprosodiques obtenus. Ce choix est justifié par le fait que contrairement à la moyenne arithmétique qui est considérée comme une moyenne de grandeur, la médiane est plutôt considérée comme une moyenne de position et elle n'est pas influencée par les valeurs extrêmes éventuellement très grandes ou très petites.

Une fois le calcul de la médiane fait, on procède au tracé du profil correspondant. A noter, que nous n'avons pas pris en considération les deux valeurs extrêmes calculées auparavant, soient D - Δ et F +Δ, car nous avons estimé que ces deux valeurs appartenaient en réalité aux voyelles extrêmes et non pas à la consonne traitée (les valeurs obtenues correspondantes, étaient toutes égales à 1) . La figure 5.3, qui présente six trajectoires possibles, montre l'évolution du profil microprosodique du phonème [b].



	1	2	3	4	5
Série1	0,829	0,824	0,808	0,793	0,791
Série2	0,973	0,971	0,943	0,901	0,899
Série3	0,853	0,818	0,785	0,827	0,829
Série4	0,872	0,867	0,822	0,785	0,786
Série5	0,993	0,986	0,948	0,984	0,982
Série6	0,830	0,826	0,817	0,812	0,813

Figure 5.3 : Evolution du profil microprosodique du phonème [b] [92]

Nous avons ensuite procédé, pour tous les phonèmes [b] détectés, au calcul de la valeur médiane de chacun des cinq points (″2″ à ″6″) correspondants au profile microprosodique calculé. La figure 5.4, représente ainsi la courbe spécifique de l'évolution du profil microprosodique du phonème [b].

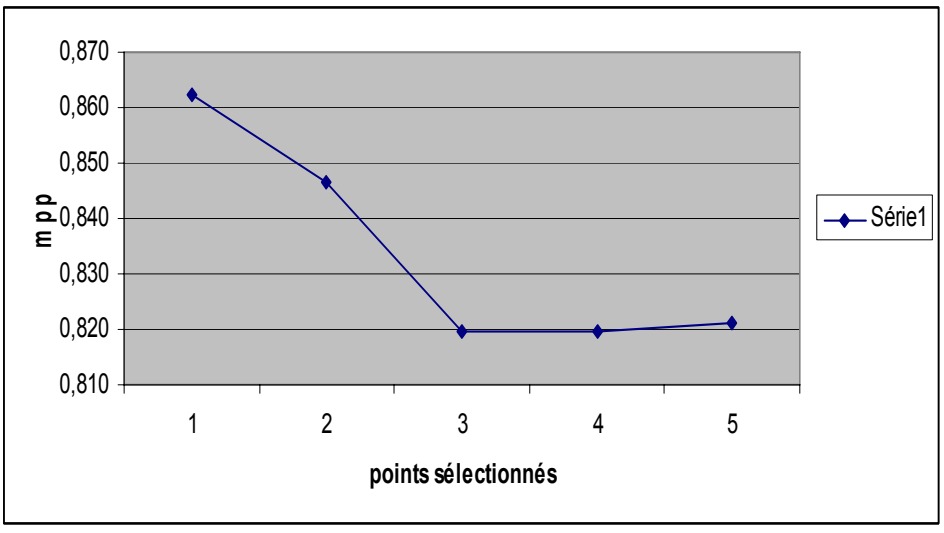


Figure 5.4 : Valeurs Médianes du profil microprosodique du phonème [b] [92]

Il est clair pour nous, que les variations microprosodiques existent bel et bien, du moment que nous obtenons un rapport (mpp) différent de 1. Cependant, nous constatons que bien qu'il y ait une

variation de la courbe micromélodique, cette dernière est très faible (l'étendue des valeurs max – min est de l'ordre de 0.045).

Nous remarquons aussi, que cette variation est toujours (pour la plupart des consonnes voisées étudiées) maximale au niveau du milieu (Point M) de la durée totale de la consonne étudiée. Cette variation devient presque nulle pour certaines consonnes (Figure 5.5).

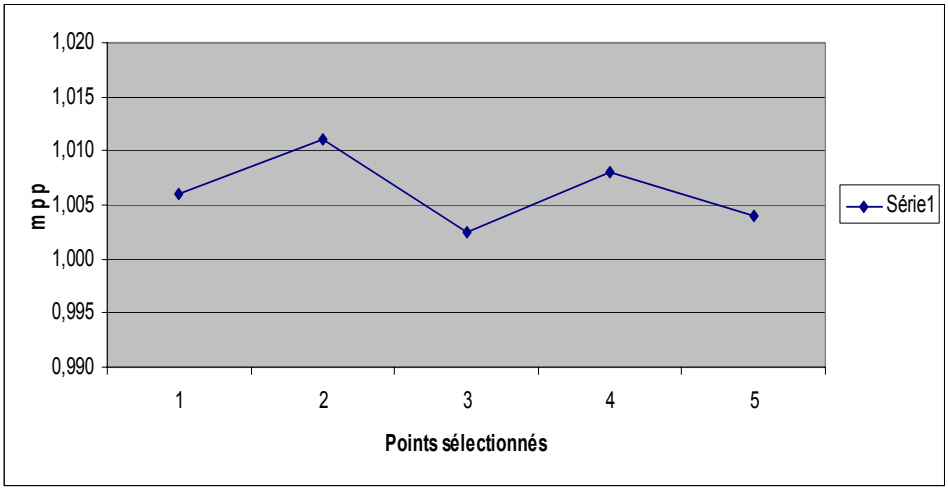


Figure 5.5 : Valeurs Médianes du profil microprosodique du phonème [n] [92]

A partir de la Figure 5.5, cas du phonème [n], les valeurs calculées varient entre 1.003 et 1.011, soit une étendue d'environ 0.008. Ceci nous permet aussi d'affirmer que dans le cas de cette nasale, l'effet microprosodique est presque absent (mpp ≈ 1).

En calculant la valeur médiane globale pour chacune des consonnes voisées étudiées, nous avons alors obtenu le Tableau 5.2 :

Tableau 5.2 : Valeurs médianes des phonèmes voisés [92]

	Phonème	**Médiane**		**Phonème**	**Médiane**
Occlusives	ب [b]	0.85	Nasales	م [m]	1.00
	د [d̪]	0.83		ن [n]	1.00
	ض [ð]	0.81	Liquide	ل [l]	0.99
Fricatives	ز [z]	0.95	Affriquée	ج [ʤ]	0.87
	ذ [ð]	0.86	Vibrante	ر [r]	0.96
	غ [ɣ]	0.95	Semi-Voyelles	و [w]	0.99
	ع [ʕ]	0.93		ي [j]	0.96
	ظ [z̪]	0.95			

A partir du tableau 5.2, plusieurs constatations ont été faites :

- les valeurs médianes calculées sont comprises entre 0.81 et 1.00, ce qui implique que la fréquence modélisée par MOMEL s'approche très fortement de la valeur réelle de F_0. Ce qui justifie encore une fois que MOMEL offre une bonne représentation des événements pertinents de la courbe de F_0 quelle que soit la langue ;
- l'effet microprosodique est presque inexistant pour le cas des nasales, des semi-voyelles et de la liquide (mpp ≈ 1) ;
- l'effet microprosodique est plus évident au niveau des occlusives que des fricatives. Nous justifions cela par le fait que ces deux types de consonnes présentent des modes d'articulations différents. Les fricatives se caractérisent par un bruit de frottement qui est obtenu par un resserrement du passage de l'air en un endroit quelconque du conduit buccal, alors que les occlusives se caractérisent entre autres par un bruit d'explosion. Celui-ci résulte d'une modalité d'articulation plus complexe que pour la friction, puisqu'elle nécessite deux étapes. Dans un premier temps, le passage de l'air est obstrué en un point quelconque du canal oral, puis, dans un deuxième temps, l'air ainsi comprimé est subitement relâché ;
- l'effet microprosodique existe mais il est très faible, ce qui n'exige pas une modélisation mathématique complexe pour modéliser sa variation ;
- les résultats obtenus viennent renforcer la théorie de plusieurs chercheurs qui maintiennent l'idée que l'effet micromélodique peut très bien être négligé, ce qui n'affecte en rien la bonne qualité de la synthèse vocale correspondante.

5.6. CONCLUSION

Dans ce chapitre, nous avons présenté une nouvelle approche qui nous permet d'extraire automatiquement l'information micromélodique du signal de parole. Pour cela, nous avons exploité les valeurs réelles de la fréquence fondamentale et les valeurs obtenues grâce à la courbe macromélodique. Ces dernières ont été obtenues grâce à la modélisation mélodique de MOMEL.

Les résultats montrent l'existence de l'effet microprosodique. Ce dernier, pourra très facilement être représenté par un simple abaissement relatif de la courbe macromélodique.

CONCLUSIONS GENERALES

ET PERSPECTIVES

Il est convenu que l'information prosodique doit être intégrée avec profit dans les systèmes de reconnaissance et de synthèse automatique de la parole. Cependant, nous savons que pour chaque type d'utilisation, des efforts importants restent nécessaires pour aboutir à un emploi optimal de ces données et de ces connaissances. Les difficultés doivent stimuler les travaux des chercheurs car l'ouverture nécessaire des systèmes de reconnaissance et de synthèse à la parole spontanée donnera une importance inévitablement plus grande aux phénomènes prosodiques.

Le but poursuivi dans cette thèse a consisté à contribuer à l'amélioration du naturel d'un système de synthèse de la parole en AS à partir du texte : TTS. Pour parvenir à ce résultat, nous avons commencé par élaborer un corpus en AS composé de phrases contenant les différents phonèmes de l'AS, prononcées par des locuteurs arabophones et enregistrées dans une chambre sourde. Ces phrases ont subi une analyse sonagraphique, une segmentation et un alignement semi-automatique et à la fin une TOP leur est appliquée.

Une fois, toutes ses étapes réalisées, nous avons proposé deux approches différentes pour la génération automatique de la prosodie :

- la première a présenté une nouvelle approche de la détection de l'Accent Primaire (A_cP) en AS par une Classification par Analyse Discriminante (CAD) du paramètre prosodique énergie. Un pourcentage de détection égal à 78% de la syllabe accentuée a été obtenu, ce qui montre l'efficacité d'une telle approche qui pourra venir renforcer les méthodes existantes basées sur le critère du fondamental. Nous notons cependant, le faible pourcentage de bonne classification obtenu lors de la phase de reconnaissance. Ceci est dû au mauvais classement des deux autres syllabes (avoisinant le cas d'un classement aléatoire). Nous avons ensuite montré que le nombre de bandes utilisées lors du calcul du spectre Moyen à long Terme (LTAS) jouait un rôle important dans l'efficacité d'une telle méthode. En fait, nous avons obtenu de meilleurs résultats lorsque nous sommes passés de 3 bandes à 6 bandes. La syllabe accentuée a présenté un très bon taux de reconnaissance, elle est passée de 78% à 85 %. Puis dans un souci d'amélioration des taux globaux des phases d'apprentissage et de reconnaissance, nous avons procédé à l'établissement de versions améliorées de la méthode proposée au départ. La première version qui réalise une CAD basée sur le critère d'énergie et de la fréquence fondamentale, nous a fourni de très bons résultats dans la phase d'apprentissage (de 80.37% à 89.81 %) et celle de la reconnaissance aussi (de 73.33 % à 83.33 %). Nous avons obtenu une bonne classification pour les trois syllabes, ce qui nous permet de dire que la discrimination entre les trois syllabes, a été très bien réalisée et qu'il

n'est plus question cette fois-ci d'effets aléatoires du moment que tous les résultats dépassent les 80%. Cependant, nous avons noté que le taux de reconnaissance de la première syllabe est resté inchangé. Alors, dans un souci d'améliorer encore les résultats obtenus, nous avons proposé une deuxième version basée sur le critère d'énergie, la fréquence fondamentale et la durée segmentale de la voyelle détectée à l'intérieur de chaque syllabe. Les résultats obtenus sont plus que satisfaisants, la phase d'apprentissage présente un pourcentage de bon classement égal à 94.07 % et celui de la phase de reconnaissance égal à 85.56 %. Les trois syllabes sont très bien classées. Mais nous avons constaté deux points très importants, la version 2 ne change en rien les résultats précédents (ceux de la version 1) pour la syllabe accentuée S_1, alors que nous observons pour la troisième syllabe, des taux de bon classement en phase d'apprentissage et en phase de reconnaissance égaux à 100 %, ce qui a attribué à l'augmentation des taux de bon classement en phase d'apprentissage et en phase de reconnaissance. Nous avons justifié ces résultats par le fait que nous avons utilisé des mots trisyllabiques où la voyelle correspondant à la troisième syllabe (fin du mot) présente une durée segmentale plus grande que celles des deux autres syllabes, ce qui a présenté un caractère discriminant supplémentaire pour le classement de la troisième syllabe. Par contre pour la première et la deuxième syllabe, les durées segmentales des voyelles sont proches ce qui nous a permis de justifier la non amélioration des résultats correspondants à ces deux syllabes ;

- la deuxième est une nouvelle méthode qui consiste en l'extraction la plus automatique possible de l'information micromélodique du signal de parole pour élaborer un modèle le plus juste possible quant à l'estimation de l'effet micromélodique afin d'en déduire son rôle dans l'amélioration du naturel du signal vocal synthétisé. Afin d'aboutir à l'extraction de cet effet microprosodique, nous avons exploité la courbe réelle de la fréquence fondamentale et celle obtenue grâce à la modélisation mélodique obtenue par l'application de l'algorithme MOMEL qui réalise une modélisation macromélodique de la courbe mélodique. Le rapport entre ces deux valeurs calculées, représente donc l'effet microprosodique que nous recherchons et dont nous avons essayé de modéliser l'allure en prenant quelques points bien particuliers. Les résultats obtenus sont très intéressants et viennent renforcer l'idée que l'effet microprosodique existe bien et qu'il joue un rôle sur l'allure de la courbe mélodique. Nous avons ainsi constaté que cet effet était plus visible au niveau des occlusives que des fricatives et qu'il était presque inexistant pour les consonnes nasales, les semi-voyelles et la liquide.

En conclusion, nous pouvons dire que la Classification par Analyse Discriminante basée sur le critère énergie est simple à implémenter et peut être un paramètre supplémentaire pour la détection de syllabes accentuées ce qui peut venir enrichir les méthodes de reconnaissance déjà existantes qui sont basées sur le critère du fondamental. Les versions proposées améliorent bien les résultats mais augmentent le nombre de variables à traiter et par conséquent le temps d'exécution de notre programme. Cependant, un modèle complet englobant les trois paramètres acoustiques reste le plus avantageux, soit une combinaison linéaire de (F_0, I, D). A noter que les résultats obtenus sont à tester sur des corpus d'Arabe plus importants. La deuxième approche est une nouvelle innovation dans le domaine du calcul de l'effet microprosodique. Elle vient confirmer l'existence de l'effet microprosodique au niveau de la fréquence fondamentale. L'analyse de la variance nous pousse à proposer seulement un abaissement relatif supplémentaire de la courbe macromélodique, si nous désirons améliorer au mieux le naturel de la parole synthétique. Cependant, des études complémentaires concernant le cas de contexte interrogatif et exclamatif ainsi que l'effet microprosodique au niveau de la durée et de l'intensité sont à prévoir pour compléter notre analyse.

REFERENCES BIBLIOGRAPHIQUES

[1] J.-M. Schaeffer and O. Ducrot, Nouveau dictionnaire encyclopédique des sciences du langage. Collection Points Essais, N° 397, 2000.

[2] M. Kob, Physiologie des lèvres et des cordes vocales. Journée d'étude « Lèvres vibrantes et cordes vocales », ENST, Paris, France, Juillet 2004.

[3] S. Rouibia, Prise en compte de critères acoustiques pour la synthèse de la parole. Thèse de Doctorat en Traitement du signal et Télécommunications, Ecole Nationale Supérieure des Télécommunications de Bretagne en habilitation conjointe avec l'université de Rennes 1, France, 2006.

[4] J. Vaissière, Phonétique et Phonologie. Cours Deug 2. Second semestre, Laboratoire de Phonétique et Phonologie, Paris III, France, 2002.

[5] J. Mariani, Analyse, synthèse et codage de la parole. Editions Hermès Science Publications, Paris, France, 2002.

[6] T. Dutoit, Traitement Automatique de la parole. Notes de cours / DEC2, Faculté Polytechnique de Mons, Première édition, Belgique, 2000.

[7] R. Boite et al., Traitement de la parole. Presses Polytechniques et Universitaires Romandes, 2000.

[8] E. Keller, Les théories de la parole dans l'éprouvette de la synthèse. Dans E. Keller, & B. Zellner (eds.), Etudes des Lettres, Vol.3 : Les défis actuels en synthèse de la parole, pp. 9-27. Université de Lausane, Suisse, 1997.

[9] T. En-Najjary, Conversion de voix pour la synthèse de la parole. Thèse de Doctorat en Traitement du signal et télécommunications, Université de Rennes 1, France, 2005.

[10] J.-P. Haton et al., Reconnaissance automatique de la parole: du signal à son interprétation. Editions Dunod, Paris, France, 2006.

[11] Calliope, La parole et son traitement automatique. Editions Masson, Paris, France, 1989.

[12] S. Baloul, Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, Thèse de Doctorat, Spécialité : Informatique, Université du Maine, Le Mans, France, 2003.

[13] N. Boukadida, Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe. Thèse de Doctorat, Université Rennes 2 (France) et Université de Tunis I (Tunisie), 2008.

[14] T. saidane, M. Zrigui & M. Ben Ahmed, La transcription orthographique-phonétique de la langue arabe. JEP - TALN - RECITAL, Fès, Maroc, 19-22 avril, 2004.

[15] S. Ghazali, La coarticulation de l'emphase en arabe. Dans Etudes de linguistique arabe, Arabica Journal, Paris, France, Vol. 28, n° 2-3, pp. 251-277, 1981.

[16] M. Guerti, Contribution à la synthèse de la parole en Arabe Standard. Actes des 16èmes Journées d'études sur la parole, Hammamet, Tunisie, pp. 290-292, 1987.

[17] A. El-Khairy, The Arabic Pharyngeal Approximant. ICPhs99, San Francisco, USA, pp. 1029-1032, 1999.

[18] A. Rajouani, Contribution à la réalisation d'un système de synthèse à partir du texte pour l'arabe. Thèse de Doctorat, Université Mohamed V de Rabat, Maroc, 1989.

[19] S. Ghazali, M. Zrigui, Z. Miled et H. Jemni, Synthèse de l'arabe standard à partir du texte par TD-PSOLA : Le traitement des processus phonologiques. Actes des 19èmes Journées d'étude sur la parole, Bruxelles, Belgique, pp. 89-93, 1992.

[20] M. El-Ani, Arabic phonology : An acoustical and physiological investigation. Mouton, The Hague, Paris, France, 1970.

[21] A. Zaki, Modélisation de la prosodie pour la synthèse de la parole arabe standard à partir du texte. Thèse de Doctorat en Automatique, productique, signal et image. Université Bordeaux I, France, 2004.

[22] C. Sorin, et F. Emerard, Domaines d'Application et Evaluation de la Synthèse de Parole à partir du Texte. Dans Fondements et Perspectives en Traitement Automatique de la Parole, Bruxelles : AUPELF-UREF, édition DUCULOT, pp.123-131, 1996.

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

[24] D.H. Klatt, Synthesis by Rule of Segmental Duration in English Sentences. Frontiers of Speech Communication Research, pp. 287-299, 1979.

[25] V. Aubergé, La synthèse de la parole : des règles aux lexiques. Thèse de Doctorat en Informatique. INPG, université Pierre Mendès, France, 1991.

[26] F. Emerard, Synthèse par diphones et traitement de la prosodie. Thèse de Doctorat, Université de Grenoble III, France, 1977.

[27] G. Bailly, T. Barbe & H. Wang, Automatic labelling of large prosodic databases : tools, methodology and links with a Text-To-Speech system. In G. Bailly & C. Benoît (Eds), Talking Machines: Theories, Models and Designs, pp. 323-333. Editors: Elsevier BV, 1992.

[28] C. Blouin, Sélection des unités pour la synthèse vocale par concaténation. Thèse de Doctorat en Sciences. Université Paris XI, Orsay, France, 2003.

[29] Z. Zemirli, Synthèse vocale de textes arabes voyellés. Thèse de Doctorat en Informatique. Université de Toulouse III, France, 2004.

[30] A. M. Liberman & I. G. Mattingly, The motor theory of speech perception revisited. International Journal of Cognitive Science (Cognition), Vol. 21, pp. 1-36, 1985.

[31] J. Robert-Ribes, P. Escudier & J. L. Schwartz, Modèles d'intégration audition-vision dans la perception des voyelles : une étude neuromimétrique. Actes du Cinquième Colloque de l'ARC, Nancy, France, pp. 85-100, 1992.

[32] S. Maeda, Un Modèle Articulatoire de la Langue avec des Composantes Linéaires. 10èmes Journées d'Etude sur la Parole, Grenoble, France, pp. 152-162, 1979.

[33] B. J. Kröger, Minimal Rules for Articulatory Speech Synthesis. Signal Processing VI : Theories and Applications, Elsevier, Amsterdam, The Netherlands, pp. 331-334,1992.

[34] M. G. Rahim & al., On the Use of Neural Networks in Articulatory Speech Synthesis. Journal of the Acoustical Society of America, Vol. 93, N°2, pp. 1109-1121, 1993.

[35] C. H. Shadle & R. I. Damper, Prospects for Articulatory Synthesis: A Position Paper. Proceedings of 4th ISCA Workshop on Speech Synthesis, Pitlochry, pp. 121-126, 2001.

[36] D. H. Klatt, Software for a Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America, Vol.67, N°3, pp. 971-995, 1980.

[37] D. H. Klatt, Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America, Vol. 82, N°3, pp. 737-793, 1987.

[38] B. S. Atal & J. R. Remde, A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates. Proceedings of IEEE - ICASSP, Paris, France, pp. 614-617, 1982.

[39] B. Fette, W. Clark & C. Jaskie, Experiments with High-Quality Low Complexity 4800 bps Residual Excited LPC (RELP). Proceedings of IEEE - ICASSP, New York City, USA, Vol.1, pp. 263-266, 1988.

[40] M. Schroeder & B. Atal, Code-Excited Linear Prediction CELP: High- Quality Speech at Very Low Bit Rates". Proceedings of IEEE-ICASSP, Tampa, Vol.10, 937-940, 1985.

[41] F. Charpentier & M. Stella, Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. Proceedings of IEEE - ICASSP, Tokyo, Japan, pp. 2015–2018, 1986.

[42] E. Moulines & F. Charpentier, Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones. Speech Communication, Vol. 9, pp. 453-467, 1990.

[43] G. de los Galanes & al., New Algorithm for Spectral Smoothing and Envelope Modification for LP-PSOLA Synthesis. Proceedings of IEEE - ICASSP, Adelaide, Australia, Vol.1, pp. 573-576, 1994.

[44] T. Dutoit & H. Leich, MBR-PSOLA : Text-To-Speech Synthesis Based on an MBE Re-Synthesis of the Segments Database. Speech Communication, Vol. 13, N° 3-4, pp.435-440, 1993.

[45] D. Griffin & J. Lim, Multiband Excitation Vocoder. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 36, N° 8, pp. 1223-1235, 1988.

[46] P.J. Price, C.W. Wightman, M. Ostendorf & J. Bear, The use of Relative Duration in Syntactic Disambigution. International Conference on Spoken Language, Vol. 1, pp. 13-16, 1990.

[47] H. Fujisaki & K. Hirose, Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences. Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'82, Vol. 7, pp. 950-953, 1982.

[48] P. Taylor, ATR Automatic Recognition of Intonation from F0 Contours Using the Rise/Fall/Connection Model. *Eurospeech*, vol. 5, pp. 789-792, 1993.

[49] http://www.isle.uiuc.edu/speechprosody2010/

[50] J.R. Firth, Papers in linguistics 1934-1951. London: Oxford University Press, UK, 1951.

[51] A. Di Cristo, Interpréter la prosodie. Actes des XXIII^{èmes} Journées d'Etude sur la Parole, Aussois, France, 2000.

[52] F. Beaugendre, Modèles de l'intonation pour la synthèse de la parole. In Fondements et perspectives en traitement automatique de la parole, Aupelf-Uref (ed.), 1996.

[53] Blanc, J.M., Traitement de la Prosodie par un Réseau Récurrent Temporel. Thèse de Doctorat en Sciences Cognitives, Université Lumière Lyon II, France, 2004.

[54] J.L. Rouas, Caractérisation et identification automatique des langues. Thèse de Doctorat en Informatique, Université Toulouse III – Paul Sabatier, France, 2005.

[55] W. Hess, Pitch Determination of Speech Signals - Algorithms and Devices. Springer Verlag, 1983.

[56] A. Di Cristo, La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, France, (TIPA) Vol. 23, pp. 67-211, 2004.

[57] A. Di Cristo and D. Hirst, Modelling French micromelody: analysis and synthesis. *Phonetica*, Vol.43, pp. 11-30, 1986.

[58] M. Barkat-Defradas, I. Vasilescu, & F. Pellegrino, Stratégies perceptuelles et identification automatique des langues. Parole, pp. 25-26, 2003.

[59] F. Ramus, & al., Language Discrimination by human Newborns and by Cotton- Top Tamarin Monkeys. Science Journal, Vol. 288, pp. 349–351, 2000.

[60] A. Lacheret-Dujour, & M. MOREL, Génération automatique de la prosodie pour la synthèse à partir du texte : le système Kali. Journées prosodie, Grenoble, France, 2001.

[61] Y. Morlec, G. Bailly, & V. Aubergé, Generating prosodic attitudes in French: data, model and evaluation. Speech Communication, 33(4), pp. 357-371, 2001.

[62] E. Campione, Etiquetage semi-automatique de la prosodie dans les corpus oraux : algorithmes et méthodologie, Thèse de Doctorat en Langage et Parole, Université de Provence, France, 2001.

[63] D.J. Hirst, & A. Di Cristo, A Survey of Intonation Systems. In D. Hirst, & A. Di Cristo (eds.), Intonation Systems: A Survey of Twenty Languages. Cambridge University Press, pp. 1-44, 1998.

[64] H. Haydar & M. Maryati, Etude de l'Intonation, la Courbe Mélodique de Phrase de l'Arabe Standard. Travaux de l'Institut de Phonétique de Starsbourg, France, Vol. 17, pp. 75-113, 1985.

[65] L. Es-Skalli, Eléments d'un Modèle Intonatif pour la Synthèse de la parole Arabe. Thèse de Troisième Cycle, Faculté des Sciences, Université Mohamed V, Rabat, Maroc, 1988.

[66] P. Delattre, La nuance de sens par l'intonation. The French Review, Vol. XII, N° 2, 1967.

[67] J. Guibert, La parole, compréhension et synthèse par les ordinateurs, Edition Presses universitaires de France, 1979.

[68] F. Carton, I. Fonagy, P.R. Leon, P. Martin, M. Rossi, R. Warren, et L. Santerre, L'accent en Français contemporain, Studia Phonetica 15, 1980.

[69] A.M. Elgendy, Aspects of Parynged Coarticulation, PhD Thesis, University of Amesterdam, the netherlands, 2001.

[70] A.N. Hanna & N.A. Ghattas, Text-To-Speech Synthesis of Arabic, Workshop on Friendly Exchanging Trough the Internet, ENSERB, Bordeaux, France, 2000.

[71] G. Bohas, Contribution à l'étude de la méthode des grammairiens arabes en morphologie et en phonologie d'après les grammairiens arabes tardifs, Thèse de Doctorat, Université de Lille 3, France, 1979.

[72] J.J. Mc Carthy, On stress and syllabifiaction. In Linguistic Inquiry, 10 (3), pp. 443-465, 1979.

[73] Z. Zemirli, S. Khabet and M. Mosteghanem, An effective model of stressing in an Arabic Text To Speech System. IEEE/ACS International Conference on Computer Systems and Applications (AICCSA), May 13-16, Jordan, pp. 700-707, 2007.

[74] G.B.M. Ghalib, An Experimental Study of Consonant Gemination in Iraqi Colloquial Arabic, PhD Thesis, University of Leeds, Department of Linguistics and Phonetics, UK, 1984.

[75] Y.M. Kiang, A comparative assessment of classification methods. In Decision Support Systems, Elsevier Science Publishers B., Amsterdam, The Netherlands, 35(4), pp. 441-454, 2003.

[76] G.J. Mc Lachlan, Discriminant Analysis and Statistical Pattern Recognition. Wiley-Interscience, 2004.

[77] P. Boersma and D. Weenink, Praat: doing phonetics by computer, (Version 5.0.32) [Computer program]. http://www.praat.org.

[78] C. Benzitoun, Outils informatisés pour la constitution et l'exploitation de corpus oraux. Séminaire doctoral de Sciences du langage : Construire une problématique de l'oral, Université de Paris X Nanterre, France, Septembre 2005. http://www.u-paris10.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHIER=4537

[79] A. Chentir, M. Guerti and D.J. Hirst, Classification by Discriminant Analysis of Energy in View of the Detection of Accented Syllables in Standard Arabic. Journal of Computer Science 4 (8), pp. 668-673, 2008. ISSN 1549-3636. http://www.scipub.org/fulltext/jcs/jcs48668-673.pdf

[80] P.D. Welch, The Use of Fast Fourier Transform for the Estimation of Power Spectra. IEEE Trans. Audio Electroacoustics, 15, pp. 70-73, 1967.

[81] B. Efron and R.J. Tibshirani, An introduction to the Bootstrap. Chapman and Hall/CRC, USA, 1994.

[82] D.P. Landau and K. Binder, A Guide to Monte Carlo Simulations in Statistical Physics. Cambridge University Press, Cambridge, UK, 2000.

[83] A. Chentir, M. Guerti and D.J. Hirst, Discriminant Analysis for Classification of Stressed Syllables in Arabic. Proceedings of ICCSE'09, International Conference of Computer Science and Engineering, London, UK, 1-3 July 2009.

[84] D.J. Hirst and R. Espesser, Automatic modelling of fundamental frequency. Travaux de l'Institut de Phonétique d'Aix, France, Vol. 15, pp. 71-85, France, 1993.

[85] G. Richard et O. Cappé, Synthèse de la parole à partir du texte. Dans Traité Informatique, H7 288. Techniques de l'ingénieur, 2004.

[86] Y. Xu and S. Xuejing, Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America, 111 (3), pp. 1399-1413, 2002.

[87] A. Di Cristo, De la microprosodie à l'intonosyntaxe. Thèse de Doctorat d'Etat, Université de Provence, Publications Université de Provence, France, 2 Tomes, 850 p., 1978.

[88] A. Di Cristo, Prolégomènes à l'Etude de l'Intonation : Micromélodie. Editions du CNRS, Paris, France, 1982.

[89] C. Astésano, R. Espesser, D.J. Hirst et J. Llisterri, Stylisation automatique de la fréquence fondamentale : une évaluation multilingue. Dans Actes du 4ème Congrès Français d'Acoustique. Marseille, France, Vol. 1, pp. 441-443, 14-18 Avril, 1997.

[90] G. Rolland, Documentation MOMEL. Institut de la Communication Parlée (ICP), Grenoble, France, Novembre 2000.

[91] S. Rogers, Effets du type de discours sur le comportement microprosodique des voyelles en français québécois (intensité et durée). Mémoire de maîtrise, Université du Québec à Chicoutimi, Canada, 1997.

[92] A. Chentir, M. Guerti and D.J. Hirst, Extraction of Arabic Standard Micromelody. Journal of Computer Science 5 (2): 86-89, 2009. ISSN 1549-3636. http://www.scipub.org/fulltext/jcs/jcs5286-89.pdf