REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

DEPARTEMENT D'ELECTRONIQUE

THESE

DE DOCTORAT D'ETAT EN ELECTRONIQUE

Présentée par

BENSELAMA Zoubir-Abdeslem

Chargé de cours au Département d'Electronique –USD-Blida

THEME

Pathologie du Langage Parlé Arabe

Cas des Sigmatismes Occlusifs et Constrictifs

Soutenue le : 15 / 12 / 2007

Devant le jury composé de :

M. Mourad HADADI	Professeur à l'ENP	Président
M^{me}. Mhania GUERTI	Maître de Conférences à l'ENP	Rapporteur
M. Abderrezak GUESSOUM	Professeur à l'Université de Blida	Examinateur
M. Mohamed TRABELSI	Maître de Conférences à l'ENP	Examinateur
M^{me}. Latifa HAMAMI	Maître de Conférences à l'ENP	Examinateur

DEDICACE

À la mémoire de mon Père

ملـــــــخص

دراسة الأمراض الكلامية أو عوائق نطق اللغة العربية تتفرع إلى عدة مجالات، عادة هذه العوائق الكلامية تعالج بالطرق البدائية المتداولة حاليا في المستشفيات و هي جد مرهقة للمريض و المعالج. عملنا هذا له علاقة بالتمرين على تحسين النطق للمرضى الذين يعانون من العوائق الكلامية من أجل تطوير نظام مساعدة اتخاذ القرار للطبيب. استعملنا في هذه الدراسة قاعدة بيانات صوتية مسجلة محليا بالجامعة تخص العائق المقصود في دراستنا. وهو تغيير الشين بالثاء، ثم مثلت هذه البيانات بأحسن المعاملات التي تعبر عن تردد الأذن الداخلية،ال MFCC أدمجت هذه المعلومات في برنامج بني على مصنفين HMM/GMM و ANN لحساب معدل الاعتراف و معدل الانحراف عن الوحدة اللغوية المستوية بالنسبة للوحدة اللغوية للمريض، من الميزات الأساسية لهذا النظام أن المريض يستطيع التمرن لوحده بدون تدخل مباشر للمعالج.

كلمات المفاتيح :عوائق الكلام في اللغة العربية، الشبكة العصبية ، نظام ماركوف.

RESUME

L'étude de la pathologie du langage rentre dans un cadre pluridisciplinaire. Généralement les différents défauts de prononciation sont corrigés à l'aide d'un orthophoniste qui utilise des méthodes très simplistes, parfois lentes et lassantes pour le patient.

Notre travail rentre dans un cadre d'entraînement à la bonne prononciation des personnes souffrant de défauts langagiers en vue d'élaborer un système d'aide à la décision à l'orthophoniste, en utilisant des méthodes graphiques et sonores, permettant de suivre l'évolution du patient présentant un sigmatisme en détectant précisément les phonèmes à corriger. Pour mettre en œuvre notre travail, nous avons d'abord commencé par élaborer un corpus constitué de mots en Arabe représentant la pathologie que nous voulons traiter.

Dans notre cas il s'agit du sigmatisme occlusif ou constrictif. Par la suite nous avons extrait les meilleures caractéristiques acoustiques qui s'adaptent à notre travail : les Coefficients Cepstraux d'échelle MEL en fréquences (MFCC). Ensuite nous avons appliqué deux classificateurs basés respectivement sur les HMM/GMM (*Hidden Markov Models/ Gaussian Mixture Model*) et les ANN (*Artificial Neural Networks*). Les résultats obtenus nous ont donné un taux intéressent de reconnaissance de 87% ainsi qu'un taux de déviation du phonème pathologique par rapport au phonème sain le plus proche. Notre système d'aide peut être aussi s'installé chez le patient afin de lui permettre de s'auto-corriger.

Mots clés : Pathologie du langage Arabe, sigmatismes, HMM/GMM, ANN, MFCC.

ABSTRACT

The study of the Arabic language pathology combines different fields of study. Generally, the different pronunciation defects are corrected by a speech therapist, which uses some very simple methods that are sometimes slow and burying for the patient.

Our work concerns training for a good pronunciation of persons suffering from language defects, in order to elaborate a therapist aided decision system, by using graphics and sounds, allowing the evolution of the patient presenting a sigmatism, by detecting the phonemes to be corrected. In order to carry out the work, we started by performing a speech database, that consists of different pathological words tackling the targeted speech pathology,

In our case, it is the occlusive or constrictive sigmatism. Then we extracted the best acoustical characteristics that are well suited to our work, the Mel frequency Cepstral coefficients (MFCC). We selected the best acoustic characteristics that adapt well to our work, the Mel Cepstral Coefficients, MFCC, that are used in two classifiers, the HMM/GMM and ANN. Results obtained gave us a recognition rate of 87% in addition to a pathological phonemic deviation score. The system can be also be used by the patient itself in an isolated therapy session.

Keywords: Arabic speech pathology, sigmatism, HMM/GMM, ANN, MFCC.

Table des matières

Remerciements

.

Je tiens à exprimer ma profonde gratitude et ma reconnaissance envers Madame Mhania GUERTI, Maître de Conférences à l'ENP, directrice de cette thèse. Ses idées, ses conseils et ses critiques m'ont été d'une aide précieuse pour mener à bien ce travail. Au-delà de l'aspect scientifique de nos discussions, j'ai été particulièrement sensible à ses qualitées humaines et à l'excellent climat relationnel qu'elle a su établir entre nous.

Je tiens également à adresser mes sincères remerciements à Monsieur Mourad HADADI, Professeur à Ecole Nationale Polytechnique pour l'honneur qu'il m'a fait en acceptant de présider ce jury.

Que Madame Latifa HAMAMI Maître de Conférences à l'Ecole Nationale Polytechnique, Messieurs Abderrezak GUESSOUM, Professeur à l'Université de Blida, Mohamed TRABELSI, Maître de Conférences à l'Ecole Nationale Polytechnique, trouvent ici mes plus vifs remerciements pour l'intérêt qu'ils ont manifesté pour mon travail et pour avoir accepté la charge d'examinateurs.

Je remercie vivement Monsieur Mohamed-Abdelkader BENCHERIF, Responsable du centre de Calcul à ALJOUF University, Arabie Saoudite, pour sa disponibilité.

Mes remerciements vont également à l'ensemble des collègues du département d'Electronique, et de l'USDB, en particulier, les membres du LABSET, en l'occurrence Messieurs Mohamed-Amine BENCHERCHALI, Abdelaziz FERDJOUNI et Nassim AMMOUR.

Je suis très reconnaissant à Monsieur Ramdane HEDJAR, PhD Assistant Professor King Saud University.Riyadh pour son aide précieuse, pour sa disponibilité et pour ses conseils.

Mes sincères remerciements vont aussi aux personnes qui m'ont aidé en contribuant, de près ou de loin, à l'aboutissement de ce travail. Qu'ils trouvent dans cette thèse une trace de ma reconnaissance. Je cite en particulier mes amis doctorants et jeunes docteurs des laboratoires Signal et Communications LSC (ENP), LATSI (USDB).

Ces dernières lignes sont pour ma famille ainsi que mes amis. Je tiens ici à leur exprimer toute ma reconnaissance pour tout le soutien et tous les encouragements qu'ils ont su me donner tout au long de ce travail. Que ceux que j'oublie ici veuillent bien me pardonner.

LISTE DES SYMBOLES ET DES ABREVIATIONS

API	Alphabet Phonétique International
ANN	Artificial Neural Networks
AS	Arabe Standard
CELP	Code Excited Linear Prediction
CMC	Chaînes de Markov Caché
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
HMM	Hidden Markov Models
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coefficients
LFC	Linear Frequency Cepstral
LFCC	Linear Frequency Cepstral coefficients
LPCC	Linear Predictive Coefficients Cepstral
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multy Layer Perceptron
NLAR	Non Linear Auto Regressive
NPC	Neural Predictive Coding
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
RAP	Reconnaissance Automatique de la Parole
RASTA	RelAtive SpecTrAl
RELP	Residual Excited Linear Prediction
RNA	Réseaux de Neurones Artificiel
SVM	Support Vector Machine
TAP	Traitement Automatique de la Parole
TF	Transformée de Fourier
TFI	Transformée de Fourier Inverse
TOP	Transcription Orthographique Phonétique

Table des figures

Liste des tableaux

Introduction Générale

Dans la nature humaine, parler est un besoin, l'un des plus élémentaires peut-être. Il en est le plus distinctif de par sa forme, son impact et surtout de par sa variabilité.

Les différents peuples communiquent dans leurs langues maternelles. Ils apprennent à leurs progénitures dès leurs naissances, les mots, les intonations, les phrases, les manières de dire les choses, de décrire les objets et ceci d'une part, pour uniformiser leur parler et d'autre part, pour développer leur présence et renforcer leur existence.

Les différentes langues se diffèrent, certes, par leur vocabulaire, grammaire et leur prononciation. Mais l'unicité du canal vocal et des différents articulateurs vocaux, les aligne à pied égal. Toutefois, ce moyen de communication peut être affecté de certaines anomalies que l'on définit par pathologie du langage. Celle-ci doit être traitée pour pouvoir communiquer. Généralement ces anomalies touchent surtout l'appareil articulatoire par conséquent il nécessite différentes thérapies.

Lors du traitement des pathologies langagières, les médecins essaient en premier lieu de détecter la pathologie par des méthodes non invasives, afin de préserver au maximum les conditions de fonctionnement de l'organe défectueux. Si toutefois, une intervention chirurgicale n'est pas nécessaire, et que le problème est d'ordre phonétique articulatoire, l'orthophoniste prend le patient en charge. Dans le cadre de la thérapie de prononciation par une correction assistée, il utilise ses sens, entre autres, son ouïe, sa vision ainsi que sa propre prononciation, procédant selon une méthodologie bien définie, afin d'évaluer la bonne prononciation.

Cependant, la qualité de l'évaluation dépend de l'ouïe de l'orthophoniste, cette évaluation se trouve amoindrit avec la défaillance de l'ouïe de l'orthophoniste. Dans le but d'aider ce dernier à décider de la prononciation ainsi que de l'évolution de la guérison du patient et rendre les méthodes manuelles utilisées plus automatiques, nous allons, à travers ce travail contribuer à la réalisation d'un système d'aide à l'évaluation de la prononciation d'une des pathologies du langage Arabe appelée para-sigmatisme. Celle-ci concerne le remplacement du [ʃ] et [s] par [θ] et [t]. Cette pathologie langagière d'ordre articulatoire est généralement traitée en mode phonème isolé, le patient répète le phonème jusqu'à ce qu'il s'accoutume avec les mécanismes d'articulation. La seconde étape concerne la prononciation du phonème, en question, précédé ou suivi par les voyelles, dans le but de dégager la prononciation de mots entiers. Notre système pourra être utilisé soit par l'orthophoniste ou bien par le malade lui-même. Ceci sera suivi par une proposition plus générale, il suffira d'introduire le mot pathologique ainsi que les phonèmes les plus proches.

En vue de simplifier l'utilisation de notre système d'aide, nous nous sommes basés sur un mode interactif, en présence ou en l'absence d'un orthophoniste, le patient enregistre la prononciation de mots bien ciblés, à travers un microphone, et observe la segmentation de ce qu'il a prononcé en unités phonémiques. Le système lui donne un score de vraisemblance lui indiquant le taux de réussite de la prononciation et le point de défaillance par rapport à un corpus de références, ainsi qu'un taux de déviation phonémique représenté par le résultat de la classification par un réseau de neurones MLP. Le patient peut écouter différents locuteurs, et visualiser la prononciation par graphes.

L'évaluation de la prononciation des phonèmes, ou mots mal prononcés, a été modélisée par des méthodes probabilistes à base de chaînes de Markov et de mélanges de gaussiennes ainsi que par des réseaux de neurones. Nous nous sommes inspirés, dans ce travail, des méthodes d'apprentissage de la langue Anglaise prononcée par des étrangers. Ces études sont orientées vers la détection des phonèmes présentant une prononciation incorrecte par ces personnes. Nous avons essayé de développer le caractère de similitude remarqué entre apprendre une langue en étant non natif de celle-ci et un patient présentant une pathologie. Ce système d'aide se base surtout sur les techniques de reconnaissance de formes

La première partie de cette thèse est constituée de deux chapitres donnant un état de l'art général sur le domaine de la parole. Le premier chapitre à pour intention de présenter un état de l'art sur le traitement automatique de la parole ainsi que ses applications en insistant surtout sur la reconnaissance, les notions fondamentales sur la parole et son traitement. Nous exposons tout d'abord les grands principes du traitement automatique de la langue avant de présenter les appareils phonatoire et auditif de l'être humain. Nous présentons ensuite deux des taxonomies possibles pour les sons observables dans un signal de parole, l'une étant spécifique au Français tandis que l'autre est spécifique à l'Arabe. Nous traitons enfin les problèmes de variabilité du signal de parole et énoncerons quelques unes des méthodes de représentation graphique du signal, qu'elles soient ou non dédiées à la parole et qu'elles soient reconnues ou non comme résistantes au bruit.

Le deuxième chapitre nous permet de présenter les trois grandes techniques de la reconnaissance des formes qui sont utilisées en Reconnaissance Automatique de la Parole (RAP) : l'alignement temporel, les chaines de Markov et les modèles connexionnistes. La présentation de ces derniers

sera plus approfondie et sera précédée d'une brève présentation des connaissances de la neurobiologie qui ont servi de fondement à l'établissement des techniques neuromimétiques.

La deuxième partie de cette thèse permet de présenter les causes susceptibles de produire des pathologies du langage ainsi que les principales définitions des pathologies de la parole et ceci sera matérialisé dans le chapitre trois

La troisième partie réalisée en deux chapitres nous permettant de présenter le développement de toute la chaine de reconnaissance avec une proposition du développement de notre travail en vue de la réalisation d'un système d'aide à l'orthophoniste et au patient représentant des mots pathologiques et cela en premier, dans le chapitre 4 représentant le développement du bloc d'extraction des caractéristiques du signal vocal à savoir les techniques classiques ainsi un nouveau procédé s'articulant sur la neuro predictive coding, par la suite le développement des trois grandes techniques de classification des formes qui sont utilisées en Reconnaissance Automatique de la Parole : l'alignement temporel (Dynamic Time Warping, DTW), les Chaînes de Markov et les modèles connexionnistes. Dans le chapitre 5 nous représentons notre système d'aide qui s'articule sur les chaines de Markov ainsi que les Réseaux de Neurones en dégageant les résultats pour chaque variante.

Nous terminons notre travail par des conclusions et perspectives.

Chapitre 1 :

Etat de l’art sur la parole

1. Introduction

Dans ce chapitre nous présentons un état de l'art concernant le Traitement Automatique de la Parole (TAP) ainsi que son application dans le domaine grand public, en insistant surtout sur les réalisations les plus récentes. Nous y exposerons tout d'abord les grands principes fondamentaux du Traitement Automatique de la Langue, et par la suite leurs domaines d'applications.

L'un des éléments clés des études concernant le TAP est de comprendre les mécanismes fondamentaux nécessaires à la production des sons ou de la parole, de distinguer les différents facteurs intervenants, de séparer les comportements entrelacés des acteurs principaux en allant des poumons vers les lèvres,…

L'étude de la production de la parole est un domaine pluridisciplinaire. Elle fait intervenir diverses théories de l'aérodynamique, de la réfraction des sons, de l'humidification, de la résonance, de l'informatique de l'électronique etc.., concernant :

- ♦ le comportement répétitif ou périodique des poumons dans leur infatigable mouvement de contraction et d'expansion ;
- ♦ le comportement aérodynamique du conduit vocal lors du passage de l'air à travers ses différentes constrictions comparé à un modèle multitube ;
- ♦ le degré d'humidification des parois mandibulaires ainsi que de toutes les surfaces en contact de l'air ;
- ♦ le système de synchronisation nerveux, qui contrôle les réflexes de déplacement de la mâchoire inférieure, des lèvres de la langue etc. ;
- ♦ le comportement mécanique lors de la synchronisation des mouvements afin d'éviter tout contact indésirable à la bonne production de la voix.

Nul aspect ne peut être isolé ou ignoré lors de la production vocale, si tel événement apparaît, une pathologie langagière se fait ressentir et tout le système est déséquilibré.

2. Problématique générale

Le Traitement Automatique de la Parole invoque l'analyse, la synthèse, l'identification, la reconnaissance du locuteur, l'apprentissage du langage, en prenant en compte les différents aspects liés à l'environnement lors de la production de celle-ci. Toutefois l'évolution des systèmes de RAP est alourdie par les problèmes suivants, concernant la variabilité :

♦ acoustique, due au fait que le même phonème prononcé dans différents contextes, en présence de phonèmes voisins, présente des réalisations acoustiques différentes, sans oublier la prosodie qui modifie le sens de la phrase, ainsi que les conditions environnantes qui faussent parfois les entrées au système ;

♦ du locuteur, en effet, le locuteur change de voix d'une façon non ponctuelle lorsqu'il est malade, de mauvaise humeur où content, etc.

♦ linguistique, lorsque la même requête peut être prononcée de différentes manières, avec différents mots exemples :

- donnez-moi de l'eau.
- Puis-je avoir un peu d'eau ?
- je pourrais avoir un verre d'eau ?
- je puis avoir de l'eau (erreur de prononciation **peux** est remplacé par **puis**)

♦ phonétique, lorsque le même mot est prononcé de plusieurs manières différentes par différents locuteurs ayant divers accents sociolinguistiques.

Les problèmes additionnels, ou plutôt technologiques viennent des canaux de transmission, des nouveaux systèmes embarqués qui nécessitent des systèmes de compression, de l'effet Lombard [1] qui apparaît lorsque les personnes parlent avec une voix basse dans des milieux bruités et l'augmente en conséquence.

♦ l'absence de silence, contraire au texte qui est segmenté en mots avec des espaces, la parole connaît très peu ce phénomène, prenons par exemple l'Anglais des Texans, qui peut être interprété comme un flux d'air incompréhensible, ceci est généralement dû au chevauchement des mots et non des phonèmes comme dans le premier point cité, qui peuvent induire le système de RAP soit à se planter ou à interpréter un mot à la place d'un autre.

3. Historique du Traitement Automatique de la Parole (TAP)

Le TAP est un domaine de recherche par essence pluridisciplinaire. Il utilise conjointement des notions empruntées au traitement du signal, à la linguistique (phonétique, phonologie, sémantique, pragmatique,...), au traitement de l'information ou encore à l'algorithmique. Son évolution ainsi que ses progrès sont comme suit :

♦ en 1791, W.Von Kempelen a construit une machine mécanique qui mimique la voix humaine ;

♦ en 1939, l'exploit de H. Dudley avec le « Voder », basé sur des composants électriques faisait l'analyse et la synthèse de la voix humaine. Les sons étaient analysés puis reproduits. Il faillait parfois une semaine pour faire fonctionner le système avec ces petits instruments de musique spéciaux ;

♦ en 1975 Baker et Jelinek en 1976, travaillant dans les laboratoires d'IBM, proposèrent, qu'au lieu de stocker différentes occurrences du mot en mémoire, parfois des milliers de mots avec différentes prononciations dans différentes contextes, le système utilise un modèle abstrait des unités à reconnaître pas nécessairement les mots, en intégrant les automates finis ou chaînes de Markov cachées "Hidden Markov Models, HMM ".Ces derniers utilisent les méthodes d'apprentissage proposés par Baum en 1975 ou de Viterbi en 1967, d'une façon similaire à la DTW (Dynamic Time Warping) ;

Toutefois, ce qui est intéressant dans les HMM par rapport à la comparaison dynamique par déformation temporelle est la prise en compte des différentes occurrences d'un mot intégrant la variabilité intralocuteur et interlocuteur. Ceci a permis d'aller jusqu'à l'apprentissage au niveau phonémique en intégrant les probabilités d'une séquence de mots dans un modèle linguistique, par son extension aux bigrammes, trigrammes, etc.

Le domaine du TAP a connu une montée fulgurante en termes de recherches et de produits industriels, et ceci essentiellement grâce à :

♦ D. Klatt qui en 1980 conçut l'un des meilleurs synthétiseurs de parole pour l'Anglais Américain [2] ;

♦ L. Rabiner, qui a établi, en 1989, l'une des études les plus intéressantes à nos jours, concernant les HMM, où il expose la méthodologie de travail d'apprentissage et de test des chaînes de Markov [3] ;

♦ En 1990 le système PSOLA (Pitch Synchronous OverLap and Add ..) a vu le jour au Centre National d'Etudes des Télécommunications - CNET par E. Moulines et F. Charpentier ;

♦ De 1989 à 1998, le projet DARPA (Defense Advanced Research Projects Agency) du département de défense des Etas-unis, concernait les technologies du langage humain a donné une lancée incroyable aux HMM qui par la suite sont devenus le standard des systèmes de reconnaissance.

Divers travaux concernant la Reconnaissance Automatique de la Parole (RAP) ont été réalisés en utilisant les Réseaux de Neurones (RN), qui demandent comme les HMM des bases de données énormes pour l'apprentissage ; Elles ne gèrent malheureusement pas l'information

temporelle comprise dans les phonèmes, l'allongement d'un mot ou d'un phonème serait reconnu comme différent pour les Réseaux de Neurones (RN) alors que pour les HMM, ce n'est qu'une transition vers le même état qui est réalisée [3].

Les travaux intégrant les RN et les HMM appelés approche hybride ont fait leur preuves dans différents travaux [4].

4. Reconnaissance Automatique de la parole (RAP)

La Reconnaissance Automatique de la parole est une sous discipline récente du Traitement Automatique de la Parole.

Vers les années 1950 apparut le premier système de reconnaissance de chiffres, appareil entièrement câblé et très imparfait.

Vers 1960, l'introduction des méthodes numériques et l'utilisation des ordinateurs changent la dimension des recherches. Néanmoins, les résultats demeurent modestes car la difficulté du problème avait été largement sous-estimée, en particulier en ce qui concerne la parole continue.

Vers 1970, la nécessité de faire appel à des contraintes linguistiques dans le décodage automatique de la parole avait été jusque-là considérée comme un problème d'ingénierie. La fin de la décennie 70 voit se terminer la première génération des systèmes commercialisés de reconnaissance de mots. Les générations suivantes, mettant à profit les possibilités sans cesse croissantes de la micro-informatique, posséderont des performances supérieures (systèmes multilocuteurs, parole continue).

Nous pouvons résumer en quelques dates, les grandes étapes de la reconnaissance de la parole (tableau 1.1) [5].

Tableau 1.1 : Etapes de Reconnaissance de la Parole [5].

Année	Systèmes de Reconnaissance de la Parole
1952	Reconnaissance des 10 chiffres, pour un mono locuteur, par un dispositif électronique câblé
1960	Utilisation des méthodes numériques
1965 :	Reconnaissance des phonèmes en parole continue
1968	Reconnaissance de mots isolés par des systèmes implantés sur de gros ordinateurs (jusqu'à 500 mots)

Année	Systèmes de Reconnaissance de la Parole
1969	Utilisation des informations linguistiques
1971	Lancement du projet DARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
1972	Premier appareil commercialisé de reconnaissance de mots
1976	Fin du projet DARPA ; les systèmes opérationnels sont HARPY, HEARSAY I et II et HWIM
1978	Commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés
1981	Utilisation de circuits intégrés VLSI (Very Large Scale Integration) spécifiques au traitement de la parole
1981	Système de reconnaissance de mots sur un circuit VLSI
1983	Première mondiale de commande vocale à bord d'un avion de chasse en France
1985	Commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
1986	Lancement du projet japonais ATR de Téléphone avec Traduction Automatique en temps réel
1988	Apparition des premières machines à dicter par mots isolés
1989	Recrudescence des modèles connexionnistes neuromimétiques
1990	Premières véritables applications de dialogue oral Homme-Machine
1994	IBM lance son premier système de reconnaissance vocale sur PC
1997	Lancement de la dictée vocale en continu par IBM

4.1. Reconnaissance du locuteur

Pour la reconnaissance du locuteur trois principaux aspects nous reviennent à l'esprit à savoir, Variabilité intra-locuteur, Variabilité interlocuteur et Variabilité due à l'environnement

4.1.1. Variabilité intralocuteur

La variabilité intralocuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie

des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.

Il existe un autre type de variabilité intra-locuteur lié à la phase de production de parole ou de préparation à la production de la parole. Cette variation est due aux phénomènes de coarticulation [6]. Il est possible de voir la phase de production de la parole comme un compromis entre une minimisation de l'énergie consommée pour produire des sons et une maximisation des scores d'atteinte des cibles des phonèmes tels qu'ils sont théoriquement définis par la phonétique.

Un locuteur adoptera donc un compromis qui est généralement partagé par une vaste majorité de la communauté de langage à laquelle il appartient, bien que ce compromis lui soit propre du fait de sa physionomie particulière. Ce compromis peut d'ailleurs être retrouvé à un plus haut niveau avec la notion d'idiolecte. Ce locuteur essaiera, lors d'une phase de production de la parole, d'atteindre les buts qui lui sont fixés par les différents éléments de sa phrase tout en conservant un rythme naturel de production de la parole. Les cibles peuvent alors être modifiées du fait d'un certain contexte phonétique. Ce contexte peut être antérieur, lorsque le phonème provoquant une modification se trouve avant le phonème considéré, ou postérieur lorsque le phonème perturbateur se trouve après.

La coarticulation peut enfin se produire à l'échelle d'un ou de plusieurs phonèmes adjacents. La variabilité intralocuteur est cependant beaucoup plus limitée que la variabilité interlocuteur. Il est en effet possible, malgré les problèmes énoncés précédemment, de mettre en œuvre des systèmes automatiques d'identification du locuteur, à la manière d'une personne reconnaissant une voix familière. Cette capacité est la preuve qu'une certaine constance existe dans la phase de production de la parole par un même individu.

4.1.2. Variabilité interlocuteur

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

poumons celle-ci génèrent un son à une fréquence de base, le fondamental. Cette fréquence sera différente d'un individu à l'autre et plus généralement d'un sexe à l'autre. Une voix d'homme est plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents qui sont regroupés selon leurs classes. Or, le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le sexe et l'âge. Ainsi, le conduit vocal féminin adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte qui mesure en moyenne 17 cm. Le conduit vocal d'un enfant en bas âge est bien sûr inférieur en longueur à celui d'un adulte. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant. Un même phonème pourra avoir des réalisations acoustiques très différentes. La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux. Ces différences s'observeront d'autant plus facilement qu'une communauté de langue occupera un espace géographique très vaste, sans même tenir compte de l'éventuel rayonnement international de cette communauté et donc de la probabilité qu'a la langue d'être utilisée comme seconde ou, pire, troisième langue par un individu de langue maternelle étrangère. Là aussi, la définition phonétique tout autant qu'une définition stricte d'un vocabulaire ou d'une grammaire peuvent être exposées à des incorrections.

La variabilité interlocuteur telle qu'elle vient d'être présentée permet de comprendre aisément pourquoi les méthodes de reconnaissance des formes fondées sur la quantification de concordances entre une forme à analyser et un ensemble de définitions strictes plus ou moins formelles ne peuvent être appliquées, avec un succès limité, qu'à des applications où le nombre de définitions est restreint, limitant ainsi le nombre de possibilités. D'une manière générale, la définition assez floue des différents phonèmes ou des différents mots d'une langue est la cause de nombreuses erreurs de classification dans les systèmes de Décodage Acoustico-Phonétique (DAP). Mais la variabilité interlocuteur, malgré son importance évidente, n'est pas encore la variabilité la plus importante car les différences au sein des classes phonétiques sont en nombre restreint.

4.1.3. Variabilité due à l'environnement

La variabilité liée à l'environnement peut, parfois, être considérée comme une variabilité intralocuteur mais les distorsions provoquées dans le signal de parole sont communes à toute

personne soumise à des conditions particulières. Cette variabilité peut également provoquer une dégradation du signal de parole sans que le locuteur n'ait modifié son mode d'élocution. Toutefois cette variation peut être considérée comme du bruit.

La variabilité environnementale due au locuteur peut tout d'abord être de nature physiologique. Ainsi, un système mécanique provoquant une déformation du conduit vocal provoquera immanquablement une variation dans le signal de parole produit. Ces contraintes physiques sont généralement rencontrées dans les systèmes de transport où une posture particulière, ou une accélération lors du déplacement, pourront provoquer une déformation.

Les moyens de transport peuvent également entraîner d'autres déformations du signal, d'origine psychologique. Le bruit ambiant peut ainsi provoquer une déformation du signal de parole en obligeant le locuteur à accentuer son effort vocal. Enfin, le stress et l'angoisse que certaines personnes finissent par éprouver lors de longs voyages peuvent également être mis au rang des contraintes environnementales susceptibles de modifier le mode d'élocution.

4.2. Reconnaissance phonémique

Le phonème étant l'unité de base du mot, peut paraître comme l'élément essentiel à reconnaître en vue de la reconnaissance du mot [7]. Il en a résulté que la reconnaissance phonémique ne se contente plus seulement d'exemples de prononciation des phonèmes de la langue à modéliser. Il vise plutôt à déduire un modèle applicable pour n'importe quelle voix qui sera ainsi susceptible de supporter un système multi locuteur.

Dans un système à reconnaissance phonétique, on identifie 4 étapes:

- ♦ le traitement du signal qui produit la suite des coefficients mathématiques formant le vecteur acoustique ;
- ♦ la modélisation acoustique qui produit une série d'hypothèses phonétiques pour chaque segment de parole et leur associe une probabilité statistique ;
- ♦ la modélisation lexicale qui force le reconnaisseur vocal à ne prendre en compte que les mots existants dans la langue considérée ;
- ♦ la modélisation syntaxique: qui force le reconnaisseur vocal à intégrer les contraintes syntaxiques, grammaticales ou même sémantiques de la langue considérée.

L'une des problématiques posée dans le cas réel sont les phonèmes pathologiques, ces dernières peuvent être le résultat d'une mauvaise articulation de l'appareil articulatoire ou a une déformation anatomique du conduit vocal due à des incidents, et pour corriger ces

pathologies on utilise les techniques de reconnaissance du phonème, ceci sera l'objet du développement de notre travail.

4.3. Reconnaissance hybride à la parole

La reconnaissance de la parole a vu plusieurs applications dont la principale étant la commande homme machine, cependant on retrouve aussi l'RAP associe à la reconnaissance d'image dynamique, qui devient une reconnaissance hybride, pour la lecture labiale qui est l'action d'identifier les sons prononcés par les individus de façon visuelle. En effet, pour prononcer un son précis, la bouche doit avoir une forme particulière (ouverture de la bouche, position de la langue, provenance du son, etc.). Les voyelles sont directement identifiables sur les lèvres. L'identification des consonnes est plus complexe (position de la langue, émission du souffle).

Les personnes sourdes ou malentendantes font appel à cette méthode pour comprendre ce qui est dit. Mais la lecture labiale, à elle seule ne permet pas de tout comprendre : on estime que seuls 30% du message oral émis sont perçus par ce biais. Cela dépend de la prononciation de l'interlocuteur, mais aussi de sa physionomie labiale (bouche lippue, forte barbe, paralysie faciale, etc.), et aussi de la position du locuteur (parler en montrant toujours son visage, pas à contre jour, pas dans une ambiance bruyante, etc.). Des règles pour bien communiquer existent. La lecture labiale n'est pas un "jeu de devinette" : "lire sur les lèvres" est une méthode qui s'apprend et qui fait appel à la suppléance mentale (vocabulaire). La difficulté provient des sosies labiaux (mots qui se prononcent de la même façon (verre, vers, vert): il est important de préciser au "lecteur" de quel sujet on parle. Pour les enfants, elle est souvent associée avec la LFPC (Langue Française Parlée Complétée). La LFPC est un codage manuel des sons de la langue Française (huit formes manuelles et leurs cinq emplacements près du visage). La LFPC offre à l'enfant une perception complète et sans ambiguïté du Français oral. Lorsque les personnes connaissent les signes de la langue des signes, les locuteurs complètent leur articulation avec les signes des mots prononcés : c'est le Français signé. D'autres personnes sourdes rejettent la communication orale et font plus appel à la langue des signes, pour diverses raisons car la communication orale leur demande trop d'efforts.

5. Représentations du signal Vocal

Différentes méthodes de représentation du signal, existent. Certaines ont été spécifiquement développées pour l'étude ou la compression de signaux de parole. Elles essaient, soit de

résoudre les problèmes posés par les méthodes fondées sur la seule transformée de Fourier, soit de simuler du mieux possible les caractéristiques de l'oreille humaine [8].

5.1. Problèmes posés par la Transformée de Fourier

La transformée de Fourier et l'implémentation d'algorithmique efficace qui y a été associée, la Transformée de Fourier Rapide, présente de nombreux avantages en tant que méthode d'analyse temps-fréquence. La rapidité de sa mise en œuvre l'a propulsée au rang d'élément incontournable des systèmes de traitement de signal. Mais, après la naissance de la notion de représentation temps-fréquence, qui fait suite à l'utilisation de représentations spectrographiques, les études théoriques du domaine ont permis de mettre à jour quelques inconvénients qui sont impossibles à éliminer et qui constituent ainsi les limites d'exploitation de la Transformée de Fourier [9]. Au rang de ces problèmes se trouve le compromis entre finesse d'analyse en fréquence et en temps. Le fait que la Transformée de Fourier ne prend pas en compte les dépendances temporelles implique, lorsque cette méthode est adaptée aux signaux non stationnaires, de considérer l'inégalité d'Heisenberg-Gabor. Cette inégalité postule qu'un signal ne peut être concentré sur des supports temps et fréquence qui soient, simultanément, arbitrairement petits. Une autre constatation exhibe une limitation qui dépasse le cadre de l'inégalité d'Heisenberg-Gabor et qui nous amène à nous demander ce que les transformées de tous types permettent de représenter. La théorie de Slepian-Pollack-Landau prouve en effet qu'un signal ne peut pas parfaitement confiner son énergie sur des supports finis, même s'ils sont arbitrairement grands. La Transformée de Fourier et les autres Transformées existantes ne permettent donc pas de représenter correctement un signal temporel discret, qui est déjà une approximation de la réalité. Ainsi, bien que la transformée de Fourier permette d'extraire d'un signal des connaissances a priori inaccessibles, l'information obtenue ne peut pas, théoriquement, être correcte. Ce qui pousse certains chercheurs du domaine à dire que nous serons toujours à la recherche d'une inaccessible fréquence instantanée [9]. Mais ces limites théoriques relatives aux représentations temps-fréquence ne sont pas les seuls problèmes existants. Le défaut majeur de la Transformée de Fourier pour l'étude de la parole vient de l'inévitable intermodulation source/conduit présente dans le spectre qui ne permet pas de connaître précisément la hauteur du fondamental. Cette intermodulation est due à la convolution qui est réalisée par le conduit vocal sur la fréquence fondamentale produite par les cordes vocales. La déconvolution ne pouvant pas être réalisée par une simple transformée, il a donc fallu développer une technique particulière capable de la

réaliser pour fournir ces deux informations utiles à l'analyse de la parole. L'étude des représentations temps-fréquence et les limites de la Transformée de Fourier ont donc poussé à créer des méthodes de traitement de signal plus adaptées à la parole, que ces méthodes soient spécifiques à la recherche ou qu'elles soient créées pour des applications plus industrielles avec une volonté de compression maximale du signal agrémentée d'une conservation de sa qualité subjective.

Nous allons maintenant présenter des méthodes adaptées à la parole qui sont les plus utilisées actuellement. Ces grandes méthodes sont les techniques cepstrale, le codage par prédiction linéaire, le codage par modulation et les modèles d'audition.

5.2. Représentations Cepstrales

Pour séparer les deux informations présentes dans le signal de parole que sont la fréquence fondamentale et la transformation, supposée linéaire, effectuée par le conduit vocal, il est nécessaire d'effectuer une déconvolution, a posteriori, du signal pour connaître la contribution des cordes vocales et du conduit vocal lors de la génération du signal qui a, par la suite, été observé en entrée du système. Cette déconvolution peut être effectuée grâce au cepstre. Il est à noter que le nom même de cepstre est défini à partir du mot spectre. De même, la représentation temps-fréquence associée n'est plus qualifiée de fréquentielle mais de quéfrentielle. Le cepstre est une méthode qui se fonde sur la Transformée de Fourier mais qui, grâce à une méthode efficace, permet d'isoler la Fréquence initiale du fondamental de la transformation qui a été opérée par le conduit vocal. Comme pour le calcul du spectrogramme, le signal est préaccentué puis convolué avec une fenêtre ajustée. Une première transformée de Fourier est alors calculée pour obtenir un spectre du signal, comme pour un spectrogramme. Ces coefficients sont ensuite transformés par le module du logarithme. La convolution étant un opérateur multiplicatif, ce passage par les logarithmes permet de passer les coefficients dans un espace additif. Une transformée de Fourier inverse permet alors d'obtenir un cepstre dont un coefficient représente le fondamental, les autres coefficients permettant d'obtenir le spectre de la convolution effectuée sur le fondamental. Cette méthode de calcul des cepstres est élémentaire [10]. Il existe également des méthodes itératives effectuant un lissage, ce qui permet d'obtenir des cepstres de meilleure qualité. Une extension possible des cepstres est leur passage dans un espace fréquentiel non linéaire proche de l'audition humaine. Il est ainsi possible de modifier la procédure de calcul précédente pour que les coefficients obtenus soient répartis selon une échelle Mel. Une telle procédure [11],

permet d'obtenir des coefficients cepstraux à échelle Mel, Mel Frequency Cepstral Coefficients (MFCC). Ces coefficients ont été très utilisés en RAP du fait des bons résultats qu'ils ont permis d'obtenir. Cette méthode a été comparée [11] avec d'autres, du même ordre, avec des conclusions qui, déjà, laissaient entrevoir la qualité des informations extraites par la méthode MFCC. Parmi les méthodes auxquelles les MFCC avaient été comparés se trouvaient des méthodes fondées sur la prédiction linéaire [12].

5.3. Codage prédictif linéaire (LPC)

Le Codage Prédictif Linéaire (*LPC, Linear Predictive Coding*) est une méthode de codage et de représentation de la parole [13]. Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire. Il s'agit donc de prédire le signal à un instant *n* à partir des *p* échantillons précédents.

$$X(n) = \sum_{k=1}^{p} a_k X(n-k) + e(n) \qquad (1)$$

La parole n'étant cependant pas un processus parfaitement linéaire, la moyenne ajustée que constitue la somme pondérée du signal sur *p* pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme *e(n)*. Le codage par prédiction linéaire consiste donc à déterminer les coefficients a_k qui minimisent l'erreur *e(n)*, ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage. (Éq.1) La méthode du codage par prédiction linéaire est tout autant utilisée en RAP qu'en compression pour le transfert de la voix par téléphone ou radio. Elle n'est cependant pas parfaite puisque l'erreur de prédiction peut être importante sans qu'il soit possible, par cette méthode, de la corriger. La méthode RELP (Residual Excited Linear Prediction), permet de réduire une partie de cette erreur. Le principe consiste à comparer, lors de la prédiction linéaire, le signal obtenu avec le signal original. L'erreur, obtenue par soustraction, représente la partie du signal original que le prédicteur n'arrive pas à modéliser. Dans la méthode RELP, l'erreur résiduelle est passée dans un filtre passe-bas permettant de conserver l'erreur effectuée dans la seule bande fréquentielle allant de 0 à 1000 Hz. La sortie du filtre est alors codée et passée au receveur qui peut alors reconstruire un signal à partir de la prédiction et de l'erreur observée. Pour pallier le problème de l'erreur résiduelle, d'autres méthodes fondées sur la prédiction linéaire ont été développées. Ainsi la méthode *CELP* (Code Excited Linear Prediction) permet d'effectuer une compression de la parole par codage d'une trame vis-à-vis de références stockées dans un corpus. Ainsi, une trame de parole sera codée selon une combinaison linéaire de certaines

trames du corpus et c'est cette combinaison qui sera considérée à la place de la trame dans les traitements ultérieurs. Cette méthode de codage de la parole est surtout employée pour la compression et la transmission de la parole à de faibles débits [14]. L'idée du codage prédictif linéaire n'a pas encore été abandonnée malgré son apparente simplicité et l'évident taux d'erreur introduit par l'hypothèse de linéarité de la production de la parole. Le groupe en charge de l'étude du GSM ("Groupe Spécial Mobile" devenu depuis "Global System for Mobile"), après avoir étudié différents systèmes de codage de la parole sur des critères de qualité subjective, de complexité algorithmique et de besoin en bande passante, a retenu le codage prédictif linéaire dit RPE-LPC (Regular-Pulse Excited - Linear Predictive Coding) agrémenté d'un système itératif de prédiction à long terme [15]. Cet ensemble algorithmique permet de transmettre un signal de parole de bonne qualité à des taux de transfert de 13,2 kbps. Ce choix va cependant à l'encontre des tendances actuelles de codage de la parole par des méthodes permettant de conserver une qualité objective au signal de parole lors de sa transmission.

5.4 Codage dit de Modulation par Impulsions Codées

Le codage prédictif linéaire peut provoquer des erreurs dégradant fortement la qualité du signal de parole. Il est cependant précieux car il permet de transmettre de la parole à de très faibles débits. D'autres méthodes, dites de codage par modulation (*PCM*, *Pulse Code Modulation*), ou plus exactement de modulation par impulsion et codées, permettent d'obtenir une meilleure qualité de parole mais nécessitent des débits beaucoup plus importants : l'espace nécessaire à la représentation de la parole est donc plus important que pour les méthodes présentées précédemment. Le codage par modulation n'est pas spécifique à la parole car très peu de connaissances relatives au domaine ont été prises en compte dans sa mise au point. Le principe de base consiste à quantifier le signal à représenter ou à transmettre selon un certain nombre de plages de même grandeur. Ce nombre de plages représente la qualité de la quantification. Le nombre de plages va également déterminer donc le nombre de bits nécessaire à la représentation binaire. Le codage par modulation est donc une méthode numérique qui suit le même principe que la conversion de l'analogique vers le numérique. Le codage par modulation peut d'ailleurs être facilement appliqué à un signal numérique. Ce principe de base peut être raffiné. Il est tout d'abord possible de quantifier le signal selon une échelle logarithmique plutôt que linéaire, ce qui permet d'obtenir une bonne quantification de la parole. Ensuite, plutôt que de transmettre les échantillons eux-mêmes, il est possible de

coder et de transmettre simplement la différence entre deux échantillons successifs. Les échantillons successifs d'un signal de parole étant fortement corrélés, cette technique réduit l'espace des valeurs à coder. La généralisation de ce principe sur plusieurs échantillons, qui assureraient le codage de leur successeur, permet d'obtenir une prédiction linéaire de l'échantillon suivant et une mesure de l'erreur effectuée dans la prédiction. La quantification de cette différence et sa transmission permet de définir la méthode par codage de modulation différentielle (*DPCM, Differential Pulse Code Modulation*). Enfin, la définition d'un quantificateur par prédiction linéaire dont les coefficients sont constamment adaptés, par la méthode des moindres carrés, au signal de parole transmis permet de définir une méthode réduisant encore l'erreur de prédiction. Cette méthode, différentielle et adaptative, est connue sous le nom d'Adaptive Differential Pulse Code Modulation (ADPCM). Les techniques que nous venons d'exposer sont très utilisées à l'heure actuelle dans le monde des télécommunications à des débits variant de 32 à 64 kbps. Elles continueront à l'être, notamment sur les Réseaux Numériques à Intégration de Service (RNIS), ce type de réseaux étant parfaitement adapté à des signaux définis par quantification.

5.5. Prédiction Linéaire Perceptuelle (PLP)

La méthode PLP [16, 17, 18], *Perceptual Linear Prediction* (ou *Perceptually based Linear Prediction*), est une méthode inspirée du principe de la prédiction linéaire. Elle combine ce principe à une représentation du signal qui suit l'échelle humaine de l'audition. Elle est à l'origine de toute une famille de techniques de traitement du signal de parole. Cette méthode peut être résumée en trois phases de traitements successifs. Le signal de parole est tout d'abord analysé pour obtenir un spectre suivant une échelle d'audition. Ce spectre est ensuite modifié par une interpolation et une transformée de Fourier inverse, le signal obtenu étant passé dans un filtre pour réduire la dimension du spectre et augmenter la résolution fréquentielle. Une troisième étape, qui peut être omise, permet de reconstruire un signal de parole par filtrage inverse, passage dans le domaine fréquentiel hertzien et désaccentuation.

La première étape est précisément constituée par :

- une analyse en bandes critiques selon une échelle Bark par un banc de filtres ;
- une préaccentuation des valeurs obtenues selon une courbe suivant approximativement les mêmes principes que les traitements effectués par l'oreille, avec accentuation des basses fréquences et atténuation des hautes fréquences ;
- une application de la loi de préaccentuation de Stevens.

La deuxième étape est, elle, constituée des phases suivantes :

– une interpolation des sorties des filtres du banc pour obtenir un spectre sur une échelle fréquentielle auditive ;

– une transformée de Fourier inverse qui permet de ramener le spectre obtenu dans le domaine temporel ;

– une résolution d'un ensemble d'équations linéaires pour obtenir les coefficients issus d'un filtre tout pôle d'ordre 5 (ce qui permet d'obtenir au moins deux sommets caractéristiques [16]).

Cette méthode a pour avantage de permettre une analyse et/ou un codage de la parole qui respectent le principe de la prédiction linéaire, qui suivent l'échelle fréquentielle observable dans l'oreille et, enfin, qui réduisent l'espace de représentation. Cette méthode a été, par la suite, améliorée pour résister à certaines conditions de bruit.

5.6. Rasta PLP

La méthode PLP [17, 18], dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsions spectrales linéaires, [19, 20] propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP, RASTA étant l'acronyme de *RelAtive SpecTrAl* [20]. La mise en place de ce filtrage permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication. Différentes études réalisées avec cette méthode [21, 22], ont permis de confirmer les bonnes qualités de cette méthode relativement aux distorsions et ses moindres qualités face aux bruits qualifiés d'additifs, signe de la présence de plusieurs sources sonores dans un même environnement. Pour améliorer encore la méthode PLP [23] définit la méthode J-RASTA, plus résistante aux bruits additifs que ne l'est la méthode RASTA, par adjonction d'un filtrage passe-bas dans le domaine spectral.

6. Application du TAP

Les applications actuelles du TAP, concernent différents domaines tels que la téléphonie, les systèmes de commande vocale, les systèmes biométriques ou de sécurité etc. On peut les trouver suivant plusieurs axes et domaine dont les principaux sont [24, 25, 26] :

Téléphonie :	Automatisation de transactions téléphoniques (ex : opérations bancaires), self-service téléphonique pour l'accès à de services d'information (ex : consultation bulletins météorologiques), etc.
Automobile et navigation	Contrôle mains-libres des équipements tels que la radio, le conditionnement du système de navigation, le téléphone sans fil (ex : voice dialing), les systèmes télématiques, Commande d'appareillages, contrôle aérien automatique
Multimédia	Logiciels de dictée vocale, interaction vocale dans les logiciels pédagogiques (ex : apprentissage des langues) et ludiques (ex : jeux vidéo), etc. Aide aux personnes handicapées, rééducation assistée (ex : exercices de logopédie automatisés), Diagnostic assisté par ordinateur, choix de médicaments, comptes rendus. Commande d'appareillages divers (chirurgie...). Repérage des indices physiologiques (zézaiement, bégaiement,...) et psychologiques (émotivité, timidité, agressivité,...). Education de la voix des malentendants, commande vocale pour malades immobilisés.
Industriel	Contrôle vocal de machines, application pour la gestion de stocks, etc. Consultation par entrée vocale.
La télématique :	Demande de renseignements, réservation, consultation de bases de données. Numérotation téléphonique automatique (téléphones cellulaires,...).
Biométrie	Empreinte vocale pour accès en zone réglementée (lieu, fichier,...). Identification de suspects.
L'enseignement	Formation des pilotes, programmation, Enseignement Assisté par

et la formation Ordinateur (apprentissage des langues).

Ces domaines d'application on vu naître et ont fait développé plusieurs entreprises et sociétés, qui se sont faites connaître par leurs produits. Nous citerons dans ce qui suit quelques utilisateurs ainsi que les entreprises qui les ont réalisés. Toutes ces dernières se basent sur l'interaction homme machine et s'articule autour de ces cinq axes, Reconnaissance vocale Authentification et vérification du locuteur, Codage de la parole, Transcription automatique, Synthèse vocale.

- **Serveurs Vocaux Interactifs** (SVI) :

Ils sont conçus pour donner rapidement accès à des informations à temps réel, on les retrouve surtout lors de la demande de renseignements et des services dans :

– le transport, concernant les réservations, les possibilités de retard des transports, les conditions du trafic routier, retard les horaires des lignes, les durées du trajet ;

– consultation de compte, et cela en consultant tout type de payement ou crédit, que ce soit pour les effets de tous les jours ou évolution d'une opération tels que les remboursements des soins ;

– serveur vocal de Gestion d'Energie pour connaître la température ambiante d'une pièce, programmer son chauffage, allumer ou éteindre la lumière. Toutes ces opérations et davantage encore, deviennent accessibles à la voix en utilisant simplement son téléphone et en parlant naturellement.

– commande vocale d'environnement

Ce système autonome de contrôle d'environnement destiné aux personnes à mobilité réduite. Il intègre un algorithme de commande vocale capable de reconnaître une cinquantaine d'ordres. NEMO transmet les commandes aux appareils qu'il contrôle (télévision, lampe, porte, téléphone...) Son installation très simple et peut- être embarqué sur un fauteuil roulant par exemple. La technologie de reconnaissance mise en oeuvre est de type mono-locuteur, par conséquent indépendante de la langue [27].

Il existe un autre projet qui est entrain d'être réalisé à savoir le projet NESPOLE qui permet de faire une traduction en temps réel. Ce dernier, co-financé par l'Union Européenne et la

NSF (EU) [28], adresse la problématique de la traduction automatique de parole et ses éventuelles applications dans le domaine du commerce électronique et des services. Les langues impliquées sont l'Italien, le Français, l'Allemand et l'Anglais. Les partenaires sont : ITC/IRST de Trento (Italie), ISL Labs. de UKA (Karlsruhe, Allemagne) et CMU (Pittsburgh, USA), Aethra (une société italienne spécialisée dans le domaine de la vidéoconférence), APT (une agence de tourisme dans la région du Trentin en Italie) et le laboratoire CLIPS (Grenoble, France).

Le scénario NESPOLE! met en jeu un agent parlant italien, présent dans une agence de tourisme en Italie, et un client qui peut être n'importe où (parlant Anglais, Français ou Allemand) et utilisant un terminal de communication le plus simple possible (PC équipé d'une carte son et d'un logiciel de vidéoconférence). Ce choix correspond aux technologies disponibles aujourd'hui, mais, dans un futur proche, les mobiles de troisième génération pourraient éventuellement être utilisés comme terminaux. Le projet NESPOLE se schématise comme suit (figure 1.1).

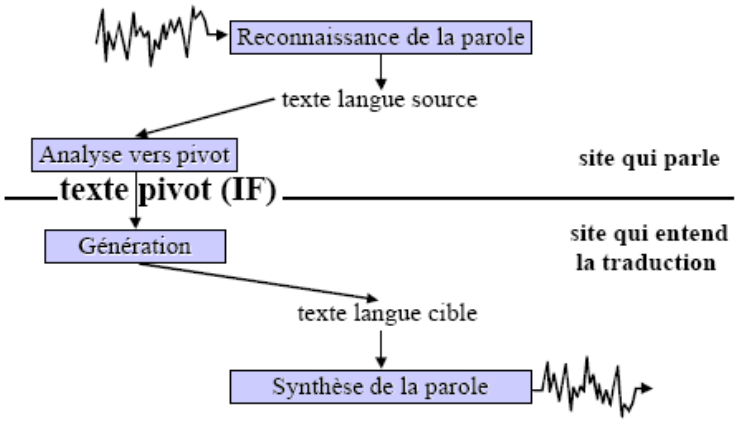


Figure 1.1 : Représentation du projet NESPOLE [28].

Le client veut organiser un voyage dans la région du Trentin en Italie, et navigue sur site Web de APT (l'agence de tourisme) pour obtenir des informations. Si le client veut en savoir plus, sur un sujet particulier, ou préfère avoir un contact plus direct, un service de traduction de parole en ligne lui permet de dialoguer, dans sa propre langue, avec un agent italien de APT. Une connexion, via un logiciel de vidéoconférence, est alors ouverte entre le client et l'agent, et la conversation médiatisée (avec service de traduction de parole) entre les deux personnes

peut alors démarrer. Dans le projet, l'accent est mis sur certains problèmes scientifiques en traduction automatique de parole : robustesse, extensibilité (extension de la couverture d'un domaine) et portabilité (passage d'un domaine à un autre).

7. Conclusion

Dans ce chapitre nous avons présenté les principales notions élémentaires qui constituent le Traitement Automatique du signal vocal en relation avec les domaines aussi divers que la phonétique, la linguistique et la reconnaissance. Nous avons aussi décrit les différentes représentations ainsi que les diverses analyses susceptibles de nous fournir de précieuses informations du signal de parole afin de rendre les données vocales plus facile a traiter et de les rendre moins encombrantes. Des applications fonctionnelles réelles réalisées ainsi que quelques projets importants qui sont entrain d'être réalisés.

Chapitre 2 :

Production de la parole

1. Introduction

Dans ce chapitre nous exposerons en premier les appareils phonatoire et auditif de l'être humain. Nous présentons par la suite les taxonomies possibles pour les sons observables dans un signal de parole. Cela va nous permettre de mettre l'accent sur les éléments essentiels qui peuvent entraîner une pathologie de la parole.

2. L'appareil phonatoire humain

Les sons de la parole se produisent lors de la phase de l'expiration grâce à un flux d'air contrôlé, en provenance des poumons et passant par la trachée-artère. Il va rencontrer sur son passage plusieurs obstacles potentiels qui vont le modifier de manière plus ou moins importante. Tous ces éléments des poumons avec les obstacles, représentent le système phonatoire

2.1. Les poumons

Les poumons sont le principal acteur de la production vocale, ils fournissent l'énergie nécessaire qui est l'air, avec un débit bien déterminé. Les muscles responsables de la respiration travaillent inlassablement en arrière plan, afin d'autoriser l'entrée ou la sortie des séquences ou flux d'air, selon une temporisation parfaite en vue d'assurer la synchronisation globale avec les autres intervenants en aval (figure 2.1).

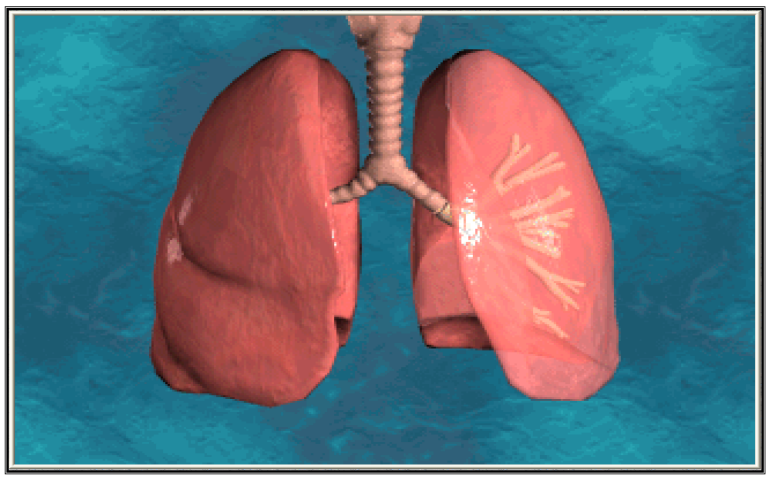


Figure 2.1 : Vue en coupe des poumons

Le mouvement des poumons dans leur infatigable geste se résume en des équilibres perpétuels de pression entre l'intérieur du corps humain et l'environnement extérieur, les différentes phases sont illustrées dans les figures 2.2 (a) et (b).

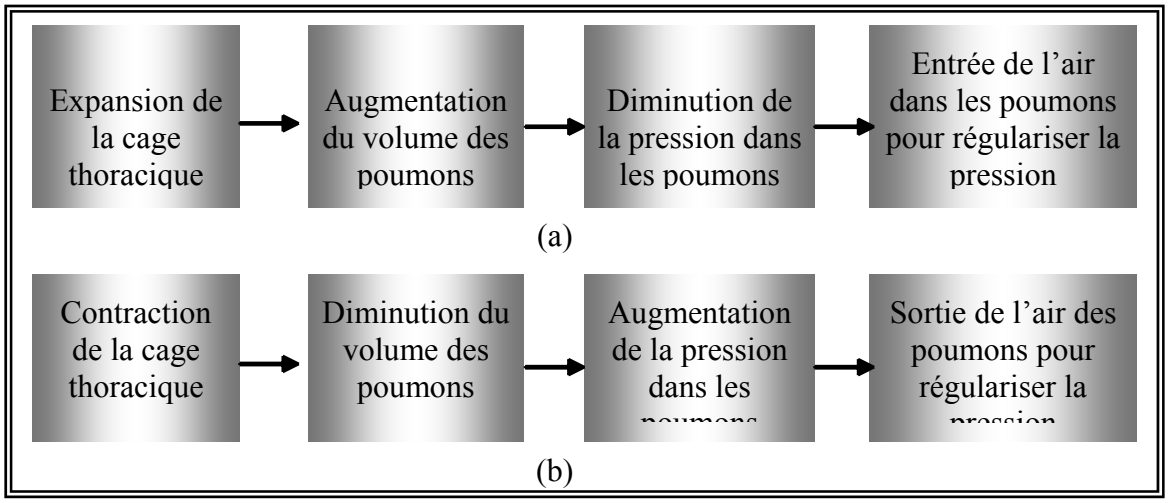


Figure 2.2 : (a) et (b) Inspiration et expiration de l'air [29]

Le réseau tubulaire très ramifié des deux poumons occupe la plus grande partie de la cage thoracique. Le poumon gauche qui se constitue du lobe supérieur et du lobe inférieur, est plus petit que le poumon droit, auquel s'ajoute encore un troisième lobe.

La surface des poumons ainsi que la paroi thoracique sont couvertes d'un pelage; les deux membranes constituent ensemble la plèvre. Elles sont superposées à plat et se confondent à la base des poumons. Entre les deux membranes, du liquide est stocké afin qu'elles puissent glisser l'une contre l'autre sans friction.

La surface intérieure des poumons est d'environ 70 cm^2 au total, ce qui représente donc environ la grandeur d'un terrain de squash ! Chaque poumon contient environ 300 millions de vésicules pulmonaires qui se groupent autour des bronchioles. Elles sont alimentées par des capillaires et forment ensemble cette grande surface qui est nécessaire afin que les poumons puissent remplir leur fonction de transporter à l'extérieur le gaz carbonique résultant lors de la combustion de substances nutritives. Un adulte absorbe normalement 400 à 500 ml d'air par inspiration. On respire à peu près 12 à 16 fois par minute. Lorsqu'on fait des efforts et quand on est physiquement actif, le volume est considérablement accru, ainsi que la fréquence respiratoire. C'est à travers la paroi de l'alvéole pulmonaire que diffusent l'oxygène et le gaz carbonique, selon des mécanismes simples d'équilibre chimique. Toutes les maladies affectant la paroi alvéolaire retentissent donc sur la qualité des échanges gazeux respiratoires [30].

2.2. La trachée artère

La trachée-artère, longue d'environ douze centimètres, touche au larynx et se ramifie au niveau de la quatrième vertèbre dorsale vers les deux bronches principales.
Les tissus élastiques et musclés de l'artère sont soutenus par 16 à 20 boucles de cartilage en forme de fer à cheval et recouverts à l'intérieur par une muqueuse avec des cils vibratiles. Ils transportent les particules de poussière qui entrent par l'air inspiré en arrière dans le pharynx (figure 2.3).

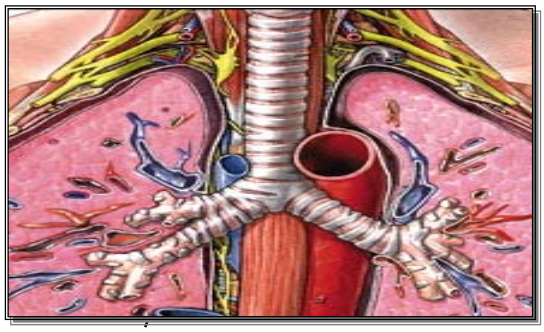


Figure 2.3 : Vue de face de la trachée artère [31]

2.3. Dispositif laryngé

Le dispositif laryngé est caractérisé par deux paramètres essentiels à savoir sa fonction physiologique et son anatomie

Sur le plan physiologique Le larynx assure trois fonctions dans la :

- ♦ respiration, puisqu'il fait partie intégrante des voies respiratoires ;
- ♦ déglutition, en fermant l'accès aux voies respiratoires sub glottiques.
- ♦ production de sons, à cause de son rôle phonatoire important, bien que non vital [32].

Anatomie du larynx se compose de deux parties essentielles, la charpente et la musculature laryngées, la charpente laryngée est composée de trois cartilages :

- ♦ le cartilage cricoïde, qui a une forme de bague, et situé dans la partie inférieure du larynx, en contact avec la trachée artère ;
- ♦ le cartilage thyroïde, qui forme la « pomme d'Adam » dans sa partie antérieure;
- ♦ le cartilage épiglottique, a une position plus centrale et supérieure.

Cette charpente est maintenue à l'aide des membranes fibreuses crico-thyroïdienne, thyroïdienne et hyoépiglottique. L'attache supérieure se fait avec le pharynx via l'os hyoïde, le cartilage thyroïde et la membrane fibreuse thyroïdienne.

La musculature laryngée a pour but de mettre en mouvement le larynx ou de modifier son ouverture pour jouer sur la production des sons. Elle est composée, entre autres :

- ♦ du muscle crico-thyroïdien, qui, en se contractant, fait basculer le cartilage thyroïde vers l'avant, ayant pour effet de tendre les cordes vocales ;
- ♦ des muscles thyro-aryténoïdiens, supérieur, interne et externe ;
- ♦ des muscles crico-aryténoïdien, latéral et postérieurs ;
- ♦ du muscle inter aryténoïde (impair).

Les trois dernières catégories de muscles ont des points de fixation en commun : les cartilages aryténoïdes, situé derrière le cartilage thyroïde, et auxquels sont attachées les cordes vocales. Leurs mouvements de translation et de rotation permettent de moduler l'ouverture et l'accolement des ces cordes.

Le larynx est innervé par des branches du nerf crânien, le nerf vague. Le nerf laryngé supérieur est mixte, mais essentiellement sensitif, alors que le nerf laryngé inférieur est moteur. Tous les muscles du larynx sont innervés, à l'exception du muscle crico-thyroïdien, (figure 2.4).

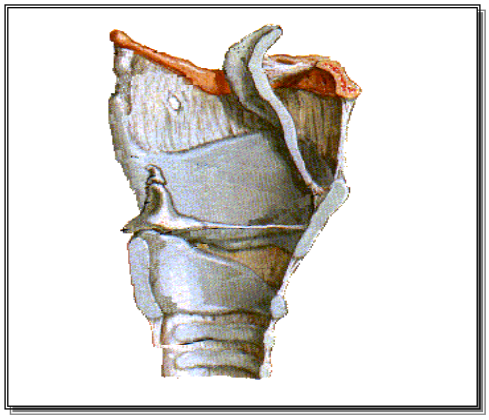


Figure 2.4 : Vue de profil du larynx

Les cartilages du larynx font des mouvements de rotation et de contraction pour permettre différentes figures géométriques des cordes vocales. Ces dernières s'ouvrent et se ferment pour permettre l'action de respirer et de modeler l'air sortant en son audible, en assurant une

pression aérodynamique par des forces élastiques. Le mouvement continuel ou périodique d'aperture et de fermeture assure la phonation et produit les sons voisés et non voisés [33].

2.4. Articulations complexes

Les vibrations des cordes vocales ne suffisent pas à produire un son intelligible, tout un système articulatoire en aval, assure la propagation de l'air vibrant ou non vibrant. Nous pouvons citer l'épiglotte, la luette, etc.

2.4.1. L'épiglotte

C'est une structure cartilagineuse reliée au larynx qui coulisse vers le haut quand les voies aériennes sont ouvertes, Elle aide à obstruer l'entrée de la trachée au moment de la déglutition. Elle descend légèrement vers le bas, afin d'entrer en contact avec le larynx qui s'élève, formant ainsi un verrou au-dessus du larynx. Il se peut que de temps à autre, lorsqu'on mange trop vite, des aliments liquides ou solides ingérés pénètrent dans le larynx avant que l'épiglotte n'ait pu se rabattre sur celui-ci. De tels cas peuvent s'avérer très dangereux du fait que les voies respiratoires peuvent se boucher et empêcher l'air de pénétrer dans les poumons .

2.4.2. La luette

La luette ou uvule est une saillie allongée mobile qui termine le voile du palais et qui contribue, lorsqu'elle se détache de la paroi pharyngale, à permettre à l'air provenant des poumons et du larynx de se diriger non seulement vers la bouche, mais également vers les fosses nasales. Lorsque la luette s'appuie sur la paroi pharyngale, elle empêche l'air de pénétrer dans les fosses nasales et ne le laisse s'échapper que par la bouche (articulations orales).

2.4.3. La langue

La langue est une masse musculaire divisée en trois parties :

- la pointe (apex) qui sert d'articulateur pour les articulations apicales, le dos pour les articulations pré médio ou post-dorsales, et la racine dans le cas des articulations radicales. Elle constitue l'articulateur principal des différents sons (figure 2.5).
- La langue permet le blocage d'air venant des poumons pour produire les consonnes occlusives, le resserrement de la cavité buccale inhérent à la production des consonnes constrictives, lorsqu'elle demeure suffisamment éloignée de la voûte du palais, elle permet la réalisation des différentes voyelles [34].

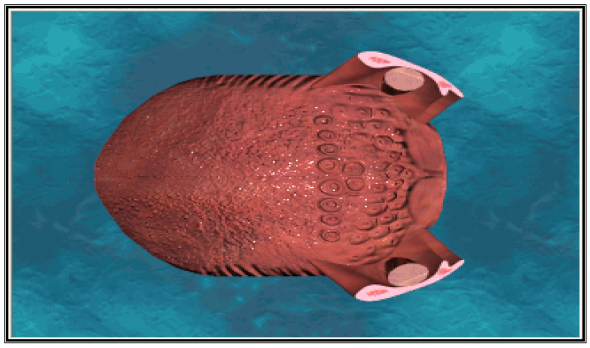


Figure 2.5 : Vue de haut de la langue

2.4.4. Les lèvres

Ce sont les parties charnues qui bordent extérieurement la bouche. Elles s'amincissent pour se joindre aux commissures. La lèvre supérieure est limitée par le nez, alors que la lèvre inférieure est limitée par le sillon mentonnier. Lorsqu'elles sont projetées et arrondies, les lèvres forment une cavité qui sert de résonateur lors de la réalisation des voyelles arrondies et des consonnes labialisées. En revanche, lorsque les lèvres sont rétractées, les voyelles sont non arrondies et les consonnes non labialisées.

La lèvre supérieure peut également agir comme lieu d'articulation par exemple, alors que la lèvre inférieure peut agir comme articulateur pour les consonnes labialisées [34]

2.4.5. Les dents

Les dents bien que pas très coopératives à la phonation, leur absence rend le système phonatoire mécaniquement défectueux, en atrophiant les ouvertures des lèvres et la prononciation des labio-dentales.

Chacun des ces organes est à la base de la production d'un son élémentaire appelé phonème. Ce dernier est la contribution distribuée du système phonatoire. La participation de chaque intervenant dépend de la langue prononcée. Les nasalisations, les roulements des [r]…. lorsqu'elles sont exagérées sont à la base des défauts langagiers et sont considérées comme pathologies nécessitant un traitement de réapprentissage de la prononciation.

La figure 2.6 illustre l'appareil phonatoire humain.

3. L'appareil auditif humain

Pour pouvoir faire l'analyse du signal vocal, il faudrait au préalable voir comment ce signal et perçu par l'oreille humaine, pour cela on fera dans se qui une description succincte de l'appareil auditif

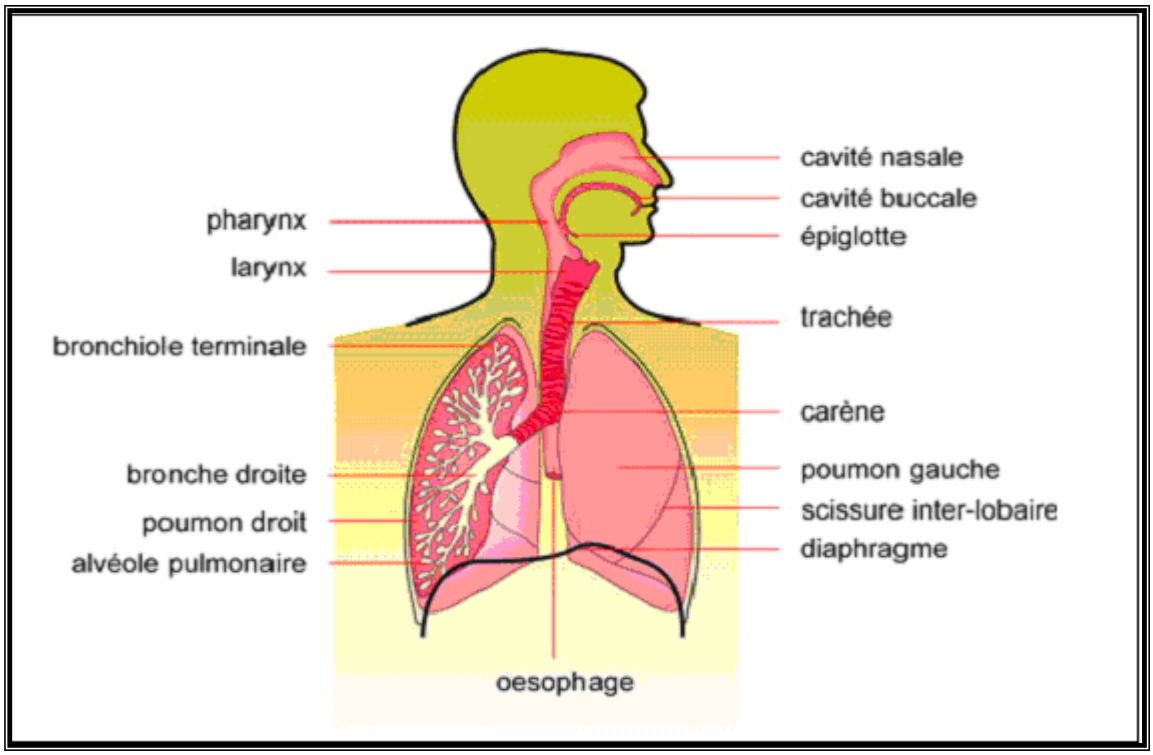


Figure 2.6 : Coupe sagittale de l'appareil phonatoire.

3.1. Description de l'appareil auditif

L'oreille est divisée en trois parties distinctes, cette division se faisant en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en un flux nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement. Une description détaillée de l'oreille permettra au lecteur de mieux appréhender les différents organes la constituant et de mieux visualiser leur répartition. Il faut noter que la présence des deux oreilles permet d'effectuer, au niveau du cerveau, des traitements plus complexes que le simple décodage d'une scène Auditive. Le positionnement des oreilles de chaque côté du crâne permet en effet de profiter des capacités

de la binauralité. Cette faculté permet de calculer la provenance d'un son en fonction du retard d'arrivée de ce son dans une oreille par rapport à l'autre. Il est à noter que cette binauralité permet à l'homme de discerner la position horizontale de l'émetteur d'un son mais pas sa position verticale. Ce principe de binauralité a été généralisé par certaines espèces animales de manière à distinguer la provenance d'un son dans un espace tridimensionnel et non plus seulement bidimensionnel. Cette généralisation pouvant être effectuée par simple désaxialisation d'une oreille par rapport à l'autre, de chaque côté du crâne (figure 2.7) [35].

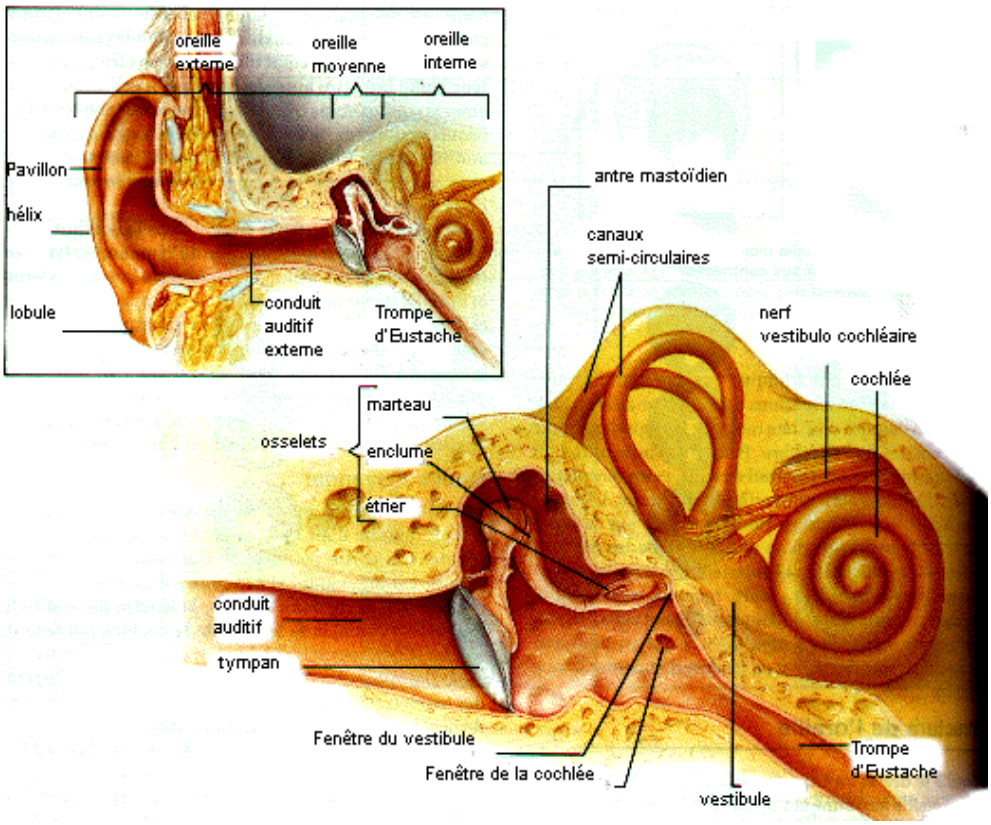


Figure 2.7 : Coupe de l'appareil auditif humain

3.2. Les courbes psycho-acoustiques

Plusieurs échelles essaient de rendre compte de la réalité perceptive de l'oreille. Elles peuvent être rapprochées des échelles de la membrane basilaire et du rang des cellules ciliées (figure 2.8). Ces échelles ne présentent pas la même morphologie. En effet, celles qui essaient de restituer le plus correctement possible les échelles de la perception humaine sont non linéaires, telles que les échelles Mel ou Bark. Les échelles qui peuvent être qualifiées de plus mathématiques sont en revanche linéaires, telle que l'échelle des fréquences. Ces différentes échelles essaient de rendre compte du mode de perception de l'homme en permettant de

distinguer les plages de plus ou moins grande importance. Ainsi les basses fréquences sont-elles perçues de manière plus fine par l'homme que les hautes fréquences. Cette différence dans la finesse de perception permet de comprendre plus facilement certaines courbes, en particulier les courbes situant l'utilisation du spectre sonore par l'homme.

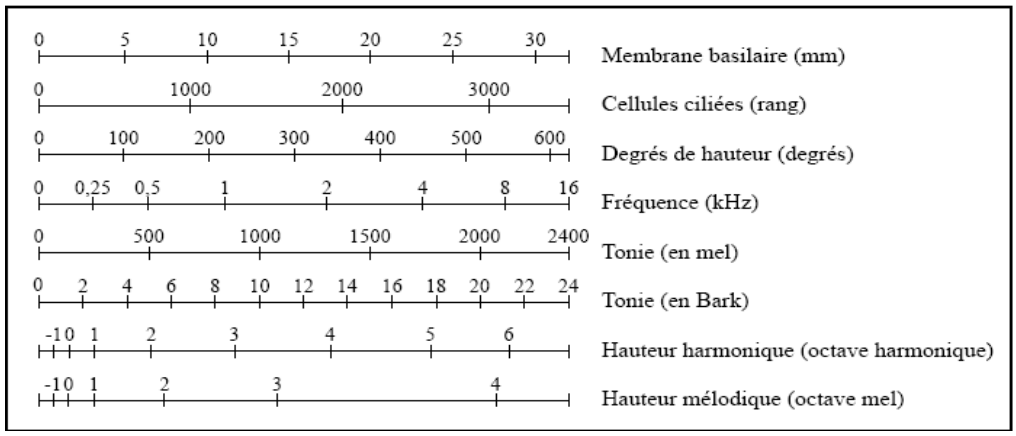


Figure 2.8 : Les échelles naturelles de la membrane basilaire [36].

L'homme est en effet très limité dans ses capacités de perception auditive vis-à-vis d'autres membres du règne animal. Il lui est ainsi impossible de distinguer des sons de plus de 20 kHz, les ultrasons, alors que certains animaux qui lui sont familiers peuvent percevoir des sons allant jusqu'à 50 kHz. De même lui est-il impossible de distinguer des sons d'une fréquence inférieure à 20-25 Hz, les infrasons. À l'intérieur de cet espace fréquentiel existe un sous-espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie en de çà de laquelle l'homme ne percevra pas un son d'une fréquence appartenant pourtant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition et il est variable en fonction de la fréquence. Inversement, il existe une limite d'énergie maximale. Cette limite ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être irrémédiablement endommagées. Cette limite s'appelle le seuil de douleur et elle aussi est variable en fonction de la fréquence. Il est intéressant de noter qu'il existe dans l'oreille deux muscles qui permettent à l'homme le transfert des vibrations du tympan à la cochlée pour limiter les dégradations qui peuvent survenir dans le cas où un bruit dépassant le seuil de douleur est perçu.

L'espace de fréquences et d'énergies ainsi défini constitue la zone d'audition à l'intérieur de laquelle l'homme peut recevoir des informations de son environnement. C'est bien sûr à l'intérieur de cet espace que se trouve le champ de la musique qui circonscrit lui-même le champ de la parole (figure 2.9).

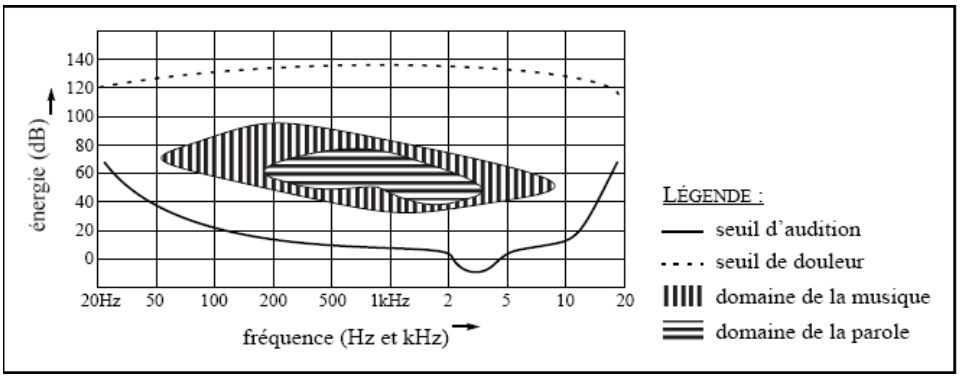


Figure 2.9 : L'aire d'audition [36].

Après avoir énoncé les caractéristiques des organes de génération et de réception de la parole, nous allons étudier les caractéristiques du traitement de la langue et les taxonomies qui ont été développées. Nous commencerons par quelques définitions avant de parler du langage arabe dont il est l'objet de notre travail.

4. Phonologie

Avant d'entamer les définitions de la phonétique articulatoire et de la phonologie, il y'a lieu de définir les éléments nécessaires à la compréhension du domaine d'étude de ces sciences.

4.1. Définitions

♦ **Phonème** : est la plus petite unité de langage parlé représenté par une articulation complète de l'appareil phonatoire. Comme par exemple les langages Arabe et Française, elles comprennent deux types de phonèmes, les voyelles et les consonnes.

♦ **Allophone** : concerne les différentes variantes d'un phonème comme [/a/] dans marche et arrête.

♦ **Les Diphones** : Un diphone est constitué d'un segment compris entre les parties stables de 2 réalisations phonémiques adjacentes et contient en son centre toute la zone de transition.

♦ une unité acoustique qui commence au milieu de la zone stable d'un phonème et se termine au milieu de la zone stable du phonème suivant.

♦ **Les Triphones** : Un triphone comprend un phonème central complet.

4.2 Domaines d'études

Il existe plusieurs domaines d'étude de la parole.

La **phonétique** étudie les sons du langage dans leur réalisation concrète appelé **'phonèmes'** indépendamment de leurs fonctions linguistiques. Il s'agit donc du 'son' en tant que 'son', sans se soucier du sens.

La **phonétique articulatoire** est une sous branche de la phonétique. Elle étudie les mouvements des organes phonatoires lors de l'émission du message, c'est-à-dire comment un être humain fait pour produire tel ou tel phonème. Quels sont les organes qui sont utilisés ? Comment sont-ils disposés et comment parviennent-ils à articuler ce phonème ?

La phonologie est la science qui étudie les phonèmes du point de vue de leur fonction dans le système de communication linguistique. Elle étudie les éléments phoniques qui distinguent, dans une même langue, deux éléments de sens différents. Elle se différencie donc de la phonétique qui étudie les éléments phoniques indépendamment de leurs fonctions dans la communication. Les principes de base de la théorie de la phonologie actuelle sont l'héritage de L.F. Saussure pour qui un **phonème**, qui est la plus petite unité, n'avait de valeur que par opposition aux autres phonèmes.

On distingue habituellement 2 grands domaines de la phonologie :

- La **phonématique**, qui étudie les unités distinctives minimales ou les phonèmes, en nombre limité dans chaque langue, les traits distinctifs ou traits pertinents qui opposent entre eux les différents phonèmes d'une même langue ;
- la **prosodie**, qui étudie les traits suprasegmentaux, c'est-à-dire les éléments phoniques qui accompagnent la réalisation de deux ou plusieurs phonèmes et qui ont aussi une fonction distinctive : l'accent, le ton, l'intonation et le rythme [37].

4.3. Le Système Phonétique de l'Arabe Standard (AS)

Dans la vie courante on distingue deux types de langages arabes, le dialectal et de le Standard (AS), ce dernier est la langue de communication commune à l'ensemble du monde arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel. La langue arabe appartient à la famille des langues sémitique. L'étude de la grammaire arabe a commencé très tôt au milieu du onzième siècle de l'hégire et a donné lieu à

d'énormes productions, avant de connaître une période de stagnation qui a duré quelques siècles. Ces dernières années, elle connaît un regain d'intérêt, entre autre dans le domaine du Traitement Automatique.

4.3.1. Historique

La recherche sur la parole a débuté depuis des siècles, les musulmans se sont intéressés à la prononciation, prenons à titre d'exemple cette décomposition phonémique de la langue arabe, il y a de cela plusieurs siècles, comme l'a mentionné El Khalil [38].

مخارج الحروف عند الخليل سبعة عشر مخرجاً. وعند سيبويه وأصحابه ستة عشر، لإسقاطهم الجوفية. وعند الفراء وتابعيه أربعة عشر، لجعلهم مخرج الذلقية واحداً. ويحصر المخارج الحلق واللسان والشفتان، ويعمها الفم .

فللحلق ثلاثة مخارج، لسبعة أحرف: فمن أقصاه الهمزة، والألف، لأن مبدأه من الحلق، ولم يذكر الخليل هذا الحرف هنا، والهاء .

ومن وسطه العين والحاء المهملتان ومن أدناه الغين والخاء .

وللسان عشرة مخارج لثمانية عشر حرفاً: فمن أقصاه مما يلي الحلق وما يحاذيه من الحنك الأعلى القاف. دونه قليلاً مثله الكاف .

ومن وسطه الحنك الأعلى الجيم والشين والياء .

ومن وسطه ووسط الحنك الأعلى الجيم والشين والياء .

ومن إحدى حافتيه وما يحاذيها من الاضراس، من اليسرى .

صعب ومن اليمنى أصعب، الضاد .

ومن رأس حافته وطرفه ومحاذيها من الحنك الأعلى من اللثة اللام .

ومن رأسه أيضاً ومحاذيه من اللثة النون .

ومن ظهره ومحاذيه من اللثة الراء .

هذا على مذهب سيبويه ، وعند الفراء وتابعيه مخرج اللثة واحد .

ومن رأسه أيضاً وأصول الثنيتين العليين الطاء والتاء والدال .

ومن رأسه أيضاً وبين أصول الثنيتين الظاء والذا ل والثاء .

ومن طرفي الثنيتين وباطن الشفة السفلى الفاء .

وللشفتين الباء والميم والواو .

والغنة من الخيشوم من داخل الأنف، هذا السادس عشر .

وأحرف المد من جو الفم وهو السابع عشر

4.3.2. L'Alphabet Phonétique International "API"

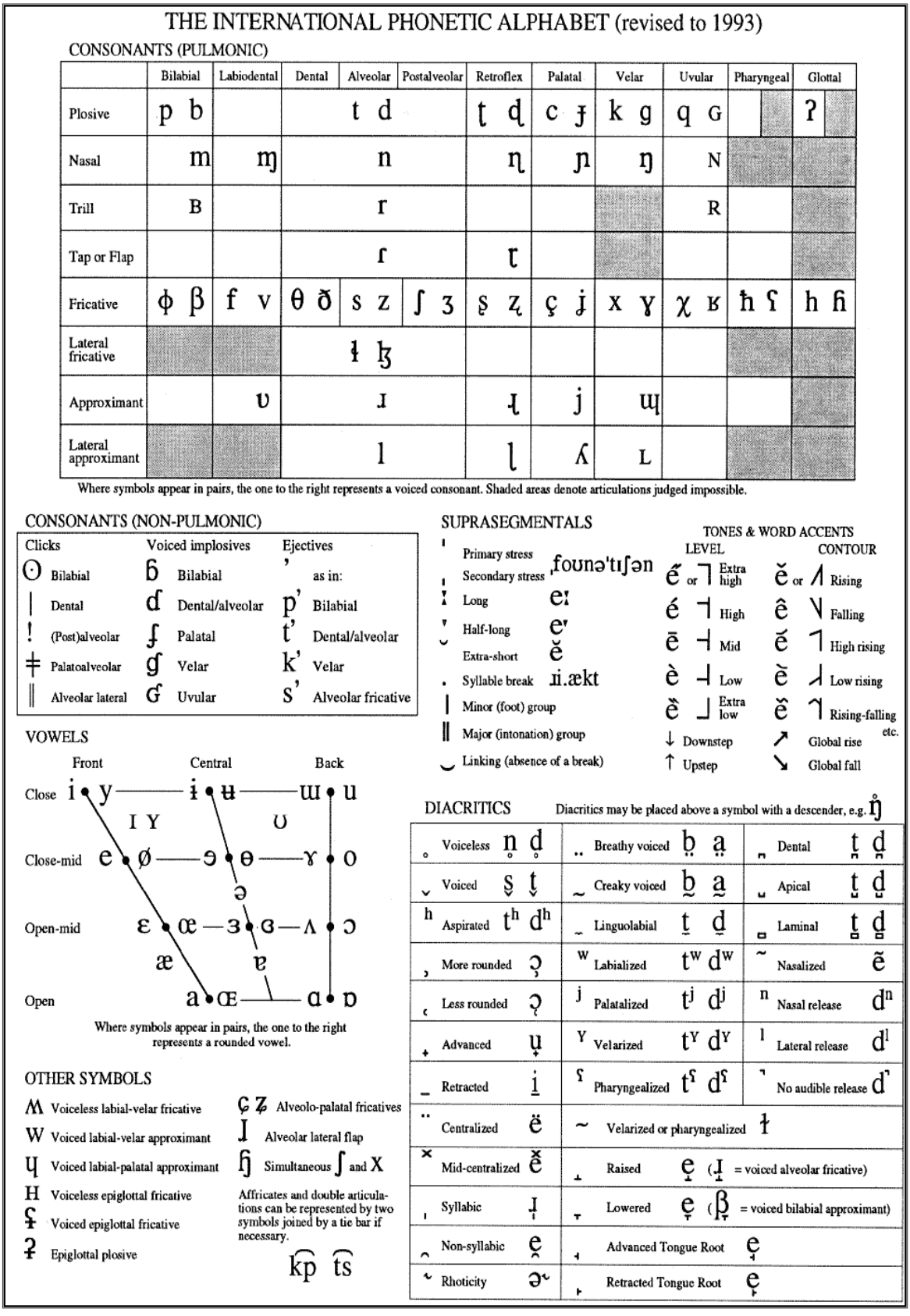
L'Alphabet Phonétique International est un alphabet universel utilisé pour la transcription phonétique des mots. Il permet d'indiquer la prononciation d'un mot, ce qui est utile lorsque

celle-ci n'est pas évidente, ou lorsqu'il s'agit d'un mot étranger au lecteur. Cette transcription se note entre crochets droits.

Il a été initialement développé par des phonéticiens Britanniques et Français sous les auspices de l'Association Phonétique Internationale, fondée à Paris en 1886 par Paul Passy (cette association et son alphabet sont plus connus sous le sigle API). La plupart des lettres sont empruntées à l'alphabet latin ou en dérivent, certaines sont d'origine grecque, et quelques caractères sont sans rapport apparent avec les lettres ordinaires. Le principe général est d'employer un symbole unique pour chaque segment sonore de la parole, en évitant les combinaisons de lettres. Des signes diacritiques peuvent être combinés avec les symboles de l'API pour transcrire des valeurs phonétiques légèrement modifiées ou des articulations secondaires. Il existe également des symboles spéciaux pour noter des phénomènes suprasegmentaux, comme les tons mélodiques [32].

Révisé en 1990 et en 1993 (tableau 2.1), l'API comprend 118 caractères principaux, 76 diacritiques et 23 marques de tons.

Tableau 2.1 : Alphabet Phonétique International [39]



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	'	as in:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

SUPRASEGMENTALS

ˈ	Primary stress	ˌfounəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
◡	Extra-short	e̯
.	Syllable break	ɹi.ækt
|	Minor (foot) group	
‖	Major (intonation) group	
‿	Linking (absence of a break)	

TONES & WORD ACCENTS

LEVEL		CONTOUR	
é᷄ or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷌ ˦˥	High rising
è ˨	Low	e᷈ ˩˧	Low rising
ȅ ˩	Extra low	ê ˥˦˥	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

etc.

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʧ ʤ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

DIACRITICS — Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	i̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɪ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ə˞	◌̙	Retracted Tongue Root	e̙			

Dans le but d'inclure les pathologies langagières, l'API a intégré un alphabet dit "Disordered speech Alphabet" (Alphabet pathologique) (tableau 2.2).

Tableau 2.2 : Alphabet phonétique pathologique

ExtIPA SYMBOLS FOR DISORDERED SPEECH
(Revised to 1997)

CONSONANTS (other than those on the IPA Chart)

	bilabial	labiodental	dentolabial	labioalv.	linguolabial	interdental	bidental	alveolar	velar	velophar.
Plosive		p̪ b̪	p̼ b̼	p͇ b͇	t̻ d̻	t̪ d̪				
Nasal			m̼	m͇	n̻	n̪				
Trill					r̻	r̪				
Fricative: central			f̼ v̼	f͇ v͇	θ̻ ð̻	θ̪ ð̪	ɧ ɧ			ɸŋ
Fricative: lateral+ central								ɬs ɮz		
Fricative: nareal	m̊							n̊	ŋ̊	
Percussive	ʬ ʬ						⫏			
Approximant: lateral					l̻	l̪				

DIACRITICS

↔ labial spreading	s̪	strong articulation	f	~ denasal	m̃
̼ dentolabial	v̼	ˌ weak articulation	vˌ	˭ nasal escape	v̊
̪ interdental/bidental	n̪	\ reiterated articulation	p\p\p	≈ velopharyngeal friction	s̴
͇ alveolar	t͇	↟ whistled articulation	s↟	↓ ingressive airflow	p↓
̻ linguolabial	d̻	→ sliding articulation	θs→	↑ egressive airflow	!↑

CONNECTED SPEECH

(.)	short pause
(..)	medium pause
(...)	long pause
f	loud speech [{*f* laʊd *f*}]
ff	louder speech [{*ff* laʊdə *ff*}]
p	quiet speech [{*p* kwaɪət *p*}]
pp	quieter speech [{*pp* kwaɪətə *pp*}]
allegro	fast speech [{*allegro* fɑ:st *allegro*}]
lento	slow speech [{ *lento* sloʊ *lento*}]
crescendo, ralentando,	etc. may also be used

VOICING

ˬ	pre-voicing	ˬz	
ˬ	post-voicing	zˬ	
(◌)	partial devoicing	(z̥)	
(◌	initial partial devoicing	(z̥	
◌)	final partial devoicing	z̥)	
(ˬ)	partial voicing	(s̬)	
(ˬ	initial partial voicing	(s̬	
ˬ)	final partial voicing	s̬)	
=	unaspirated	p=	
ʰ	pre-aspiration	ʰp	

OTHERS

(◌) indeterminate sound	(( )) extraneous noise ((2 sylls))
(V), (Pl) indeterminate vowel, plosive, etc.	¡ sublaminal lower alveolar percussive click
(Plvls) indeterminate voiceless plosive, etc.	!¡ alveolar & sublaminal click ('cluck-click')
( ) silent articulation (ʃ), (m)	* sound with no available symbol

Parmi les conséquences les plus distinguées de la transcription phonétique est que tout phonème peut être prononcé par divers locuteurs mêmes non natifs de la langue transcrite.

4.3.3. Alphabet Phonétique de l'Arabe standard

La langue Arabe standard comprend 40 phonèmes, dont 3 voyelles courtes, 3 voyelles longues plus 6 variantes vocaliques en contexte emphatique et 28 consonnes.

4.3.4. Correspondance Organes –lieux d'articulation

L'articulation complexe du système phonatoire donne lieu à une segmentation par région (figure 2.10). Cette dernière permet à un locuteur de distinguer l'emplacement exacte ou l'adresse du phonème, ce qui par la suite permet de distinguer entre :

- les phonèmes ;
- les articulateurs ayant participés à l'articulation ;

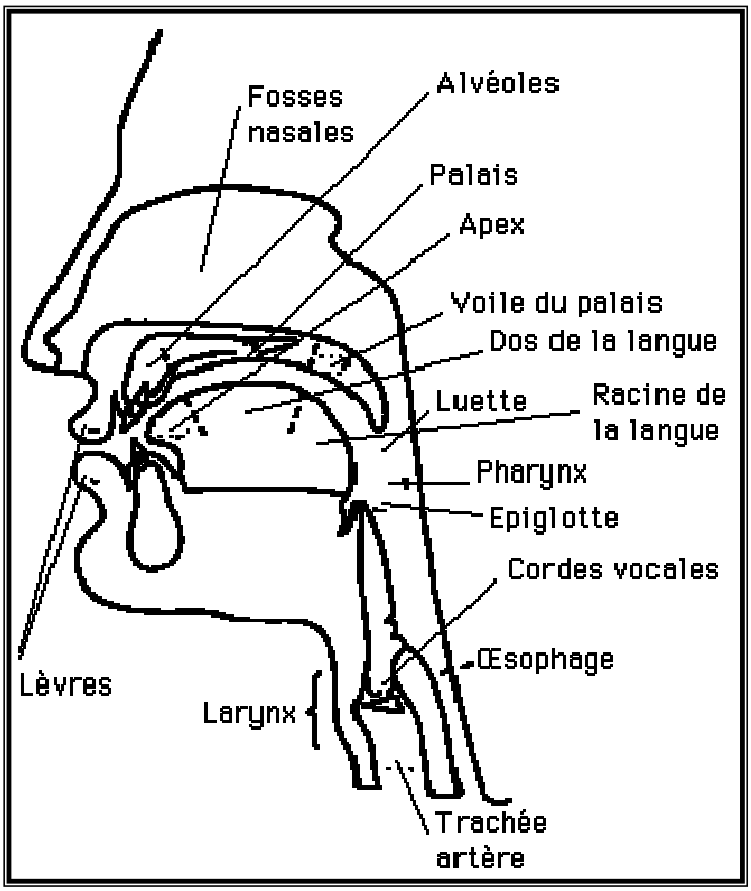


Figure 2.10 : Lieux d'articulation des phonèmes

Le tableau 2.3 montre les correspondances de la segmentation articulatoire de la figure 2.10

Tableau 2.3 : Correspondance organes - lieu d'articulation [40]

Organes	**Manifestations selon le point d'articulation**
Lèvres	Labiale
Dents	Dentale
Alvéoles des dents	Alvéolaire
Palais	Palatale
Voile du palais	Vélaire
Luette	Uvulaire
Pointe de la langue	Apicale
Dos de la langue	Dorsale
Pharynx	Pharyngale
Cordes vocales	❖ son sonore ou voisé (vibration des cordes vocales) ❖ son sourd ou non voisé (pas de vibration des cordes vocales)

La langue Arabe Standard comprend des phonèmes de différent de ceux du Français ou de l'Anglais, le tableau 2.4 illustre les différents phonèmes Arabes ainsi que leur lieux d'articulation et leur transcription phonétique.

Tableau 2.4. : Transcription phonétique des phonèmes de la langue Arabe ainsi que leurs lieux d'articulation [32]

Lieux d'articulation	Occlusif	Nasal	Vibrant roulé	Vibrant battu	Fricatif	Fricatif latéral	Spirant	Latéral
Bilabial								
	b	m						
Labio-dental					f			
Dental					**θ**			
					ð			
Alvéolaire	t				s			
	d	n	r	ɾ	z			l
Post-alvéolaire					**ʃ**			
						ʒ		
Rétroflexe								
Palatal								
							j	
Vélaire	k							
	g							
Uvulaire	q				χ			
					ʁ			
Pharyngal					ħ			
					ʕ			
Glottal	ʔ				h			

4.3.5. Segmentation d'El Khalil

El Khalil bien avant l'API, relate les lieux d'articulation des phonèmes de la langue Arabe, (tableau 2.5), cette segmentation fut utilisée par les disciples pour bien prononcer le Coran.

Tableau 2.5 : Récapitulatif des phonèmes Arabes ainsi que leurs lieux d'articulation selon El Khalil [41]

جدول توضيحي بمخارج الحروف العامة والخاصة

المخارج العامة		المخارج الخاصة		حروف كل مخرج		
5	الخيشوم		17	الغنة		
4	الشفتان	الشفتان معاً	16	و/ ب/ م		
		بطن الشفة السفلى مع طرفه	15	ف		
3	اللسان		14	س/ ص/ ز		
		حافتاه	13	ظ/ ذ / ث		
			12	ط / د / ت		
			11	ر		
			10	ن		
			9	ل		
			8	ض		
		وسطه	7	ج / ش / ي		
		أقصاه				
			6	ك		
			5	ق		
2	الحلق	أدناه	4	غ / خ		
		وسطه	3	ع / ح		
		أقصاه	2	ء / ه		
1	الجوف		1	هَاْ	حيْ	نُوْ

Pour des raisons d'apprentissage, plusieurs phonéticiens et/ou orthophonistes utilisent les lieux d'articulation ou la manière d'articuler selon des schémas illustratifs comme mentionné dans les figures 2.11. à 2.18. [41].

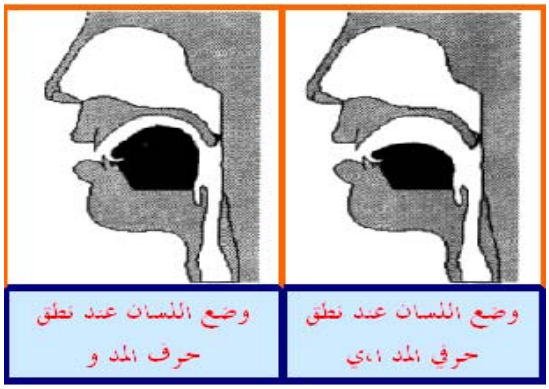


Figure 2.11 : Lieux d'articulation des phonèmes : [أ] , [و] , [ي]

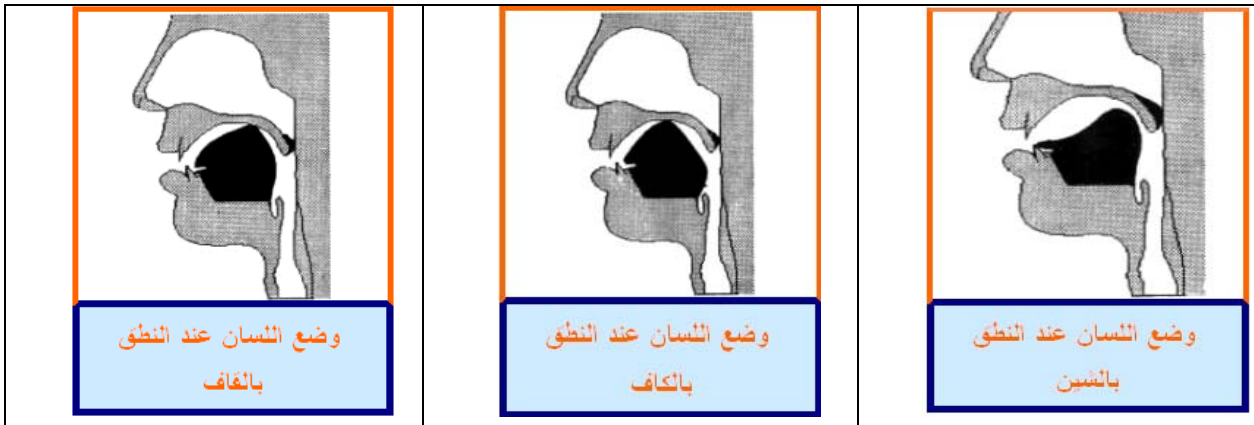


Figure 2.12 : Lieux d'articulation des phonèmes [ش] , [ك] , [ق]

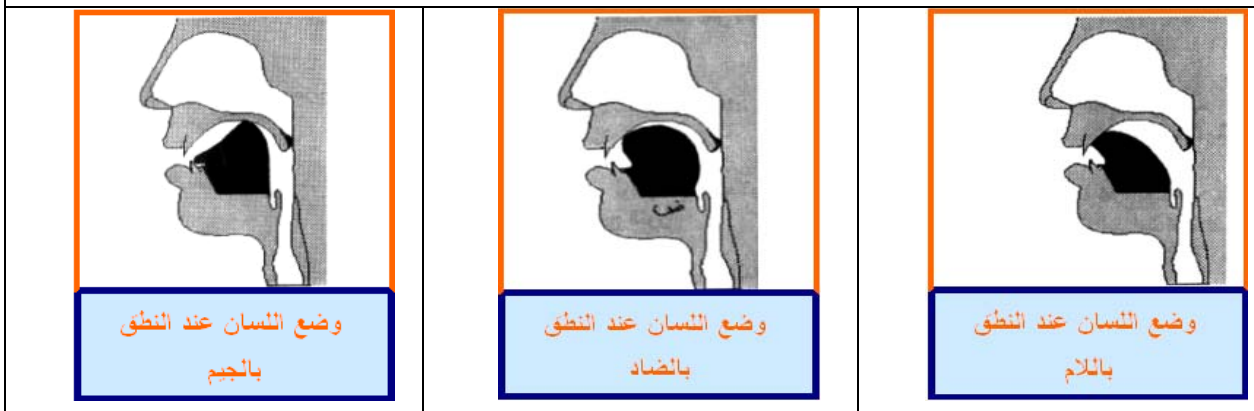


Figure 2.13. : Lieux d'articulation des phonèmes [ج] , [ض] , [ل]

Figure 2.14 : Lieux d'articulation des phonèmes [ن] , [ر] , [ت]

Figure 2.15 : Lieux d'articulation des phonèmes [ر] , [د] , [ط]

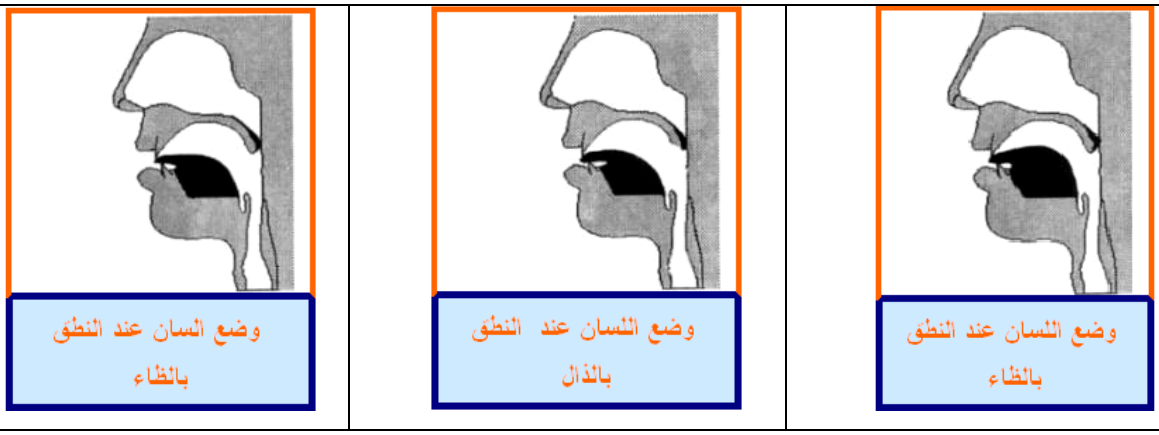


Figure 2.16 : Lieux d'articulation des phonème [ظ] , [ذ] , [ض]

Figure 2.17 : Lieux d'articulation des phonèmes [س] , [ص] , [ز]

Figure 2.18. : Position des lèvres lors de la prononciation des phonèmes

[م] , [ي] , [و] , [ف]

4.3.6. Système vocalique de l'Arabe Standard (AS)

- **Réalisation des voyelles**

La réalisation des voyelles est la classe de phonèmes issus des vibrations continues des cordes vocales, sans obstruction. Nous distinguons 3 voyelles [a], [i], [u], appelées dans l'ordre : الفتحة / الضمة / الكسرة et [A], [I], [U], représentant les voyelles longues.

La réalisation phonétique des voyelles est très variable et dépend à la fois de :

- l'origine géographique des locuteurs ;
- l'environnement consonantique et de la place de la voyelle dans le mot ;
- la place de l'accent du mot (tendance à abréger les longues non accentuées chez beaucoup d'arabophones) [42].

Afin de situer les voyelles en terme de degré de durée de voisement, celles-ci sont mentionnées dans un triangle vocalique contenant :

- deux plans représentant la durée de voisement, de la plus brève à la plus longue
- le bas des triangles représente le degré d'aperture maximale, le haut des triangles représente le degré d'aperture minimale (figure 2.19).

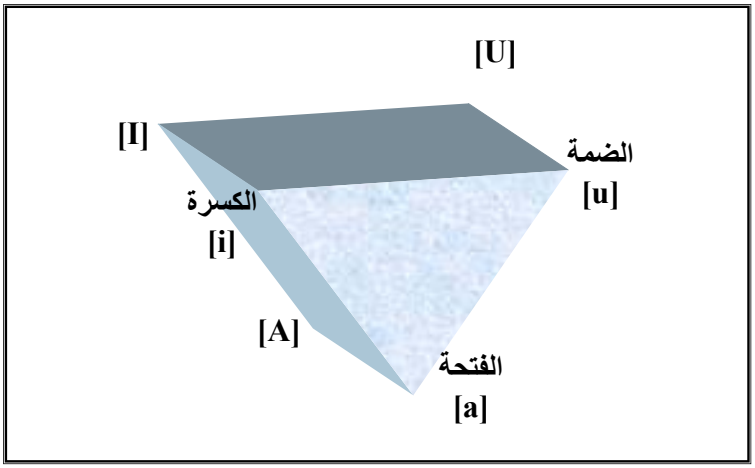


Figure 2.19. : Triangle vocalique de la langue Arabe

Les consonnes sont les éléments 2, 4, 3, 7 du système consonantique de l'AS

Toutes les consonnes peuvent être dédoublées. Ou bien géminées. On prolonge et on renforce l'articulation de cette consonne. La gémination est indiquée par un signe graphique spécifique appelé "chadda" (ّ)

La gémination joue un rôle très important en morphologie. Il est donc essentiel de bien l'entendre et de bien la réaliser

Exemples :

درس	: [darasa]	"il a étudié"
درّس	: [darrasa]	"il a enseigné"

Les consonnes emphatiques [ض],[ط],[ظ], [ص] sont caractérisées par un trait articulatoire spécifique de la "pharyngalisation" : le son produit est plus grave que pour le son non emphatique correspondant. On l'obtient en modifiant la forme du résonateur buccal dans sa partie arrière par rétraction et exhaussement de la racine de la langue.

Il est très important de bien distinguer le son emphatique du non emphatique correspondant.

Exemples :

سيف	[**s**ayf]	"épée"
صيف	[**S**ayf]	"été"

Souvent la présence d'une consonne emphatique dans un mot "influence" l'environnement, consonantique et vocalique, et c'est toute la syllabe qui est emphatisée.

4.4. Structure de la langue Arabe en succession phonétique

Dans le but d'analyser toutes les occurrences d'un phonème et de cerner tous les cas possibles de composition phonétique, le langage Arabe obéit à une structure phonétique minimale, car les autres structures complexes ne sont que la concaténation de ces formes de base, (tableau 2.6).

Tableau 2.6 : Structure phonétique minimale du langage Arabe [43]

Structure syllabique	Correspondances
[CV]	Consonne – Voyelle courte
[CVV]	Consonne – Voyelle longue
[CVC]	Consonne – Voyelle courte – Consonne

Structure syllabique	Correspondances
[CVVC]	Consonne – Voyelle longue – Consonne
[CVCC]	Consonne – Voyelle courte – Consonne – Consonne

4.5. Etudes réalisées

Notre intérêt concerne les lieux d'articulation donc, il est judicieux de s'intéresser à la manière dont ces phonèmes sont réalisés pour comprendre leurs particularités et ainsi déterminer les caractéristiques communes et discriminantes. Cette approche est fortement utilisée en segmentation phonémique [44, 45, 46, 47].

Nous pouvons distinguer, d'après les propriétés articulatoires relatives aux différents lieux d'articulation les particularités acoustiques, énergétiques, spectrales, cesptrales, etc., des phonèmes et ainsi décider des propriétés qui les discriminent, le résultat n'en sera que plus robuste.

La distinction entre phonèmes ou segmentation au niveau phonémique est une approche très répandue en Traitement Automatique de la Parole continue, car en reconnaissant les éléments de base on peut distinguer entre:

- ♦ les mots prononcés dans une même langue et reconstituer les phrases dans un contexte de parole continue, ou localiser les mots les plus significatifs dans un contexte de reconnaissance de mots isolés ou de nombres isolés comme pour le portable, la voiture, etc. [48, 49, 50] ;
- ♦ les différentes langues voire même entamer une reconnaissance directe de la langue pour engager une procédure de paiement par téléphone, par exemple, pour différents locuteurs non natifs de la langue du pays, ou pour des renseignements en différentes langues avec un aiguillage intelligent, cas du projet Raphaël, ou les locuteurs des deux côtés du système ne parlent pas la même langue [51, 52].

D'autres études comme celle portant sur la logopédie, introduit une notion phonémique très intéressante en terme de comparaison à des mots de références selon le critère de Borel-Maisonny.

5. Conclusion

Pour aborder la pathologie de la langue, il était nécessaire d'étudier tous les points qui ont trait à la parole. A travers ce chapitre, nous avons introduit la notion d'anatomie articulatoire, qui est l'une des causes de la pathologie de la parole, par une étude simplifiée.

Dans le prochain chapitre, nous allons introduire les pathologies du système phonatoire, en montrant l'impact de chaque maladie sur la production vocale. Les conséquences de ces maladies, si elles ne sont pas prises en compte à temps, cas du bégaiement, du chuintement, du schlintement ou des différents types de cancer, peuvent causer des troubles chroniques.

Chapitre 3

Pathologies du langage parlé

1. Introduction

Ce chapitre est une introduction aux pathologies relatives à la parole, ou en d'autres termes « au langage vocal » et aux pathologies subséquentes des différents éléments de l'appareil phonatoire ou des « articulateurs », lors de la production de la voix et son altération au cours de son cheminement à travers le conduit buccal ou nasal.

Lorsqu'on parle de pathologies, une référence intuitive nous fait penser à un médecin, c'est un réflexe très logique, toutefois le domaine pathologie fait intervenir différents scientifiques, par exemples l'utilisation de l'effet doppler en médecine, l'utilisation, en imagerie médicale, du scanner en 3D ainsi que des rayons X. Ces appareils sont le fruit de la technologie qui implique d'autres disciplines telle que l'électronique, la mécanique, la physique, etc.

Le traitement des pathologies langagières se situe essentiellement à détecter la zone à traiter, mesurer l'ampleur de la maladie, ou pathologie, intervenir en post ou en pré chirurgie, corriger par un orthodontiste ou par un orthophoniste. Tous ces spécialistes, diffèrent par leur degré d'intervention relatif au stade d'évolution de la pathologie et de son emplacement.

2. Définitions de certains troubles du langage

Les atteintes peuvent concerner les organes périphériques, atteintes qui gênent la production de la parole : le bec de lièvre, la division palatine, l'insuffisance vélaire, les malformations linguales, labiales ou laryngées.

Il s'agit d'anomalies consistant en des erreurs mécaniques et constantes dans l'exécution du mouvement propre à un phonème [53].

L'articulation est la capacité à articuler les sons de façon permanente et systématique, ce qui nécessite des mouvements précis de la mâchoire inférieure, de la langue, des lèvres, des joues, du voile du palais.

Le trouble d'articulation isolé est donc l'incapacité à prononcer ou à former un certain phonème correctement. C'est une erreur constante, systématique et mécanique pour un phonème donné. Cette erreur est plutôt de type praxique [54].

La Production de la voix normale est basée sur sa qualité, son intensité, son débit. Une voix pathologique présente une altération d'un ou de plusieurs de ces paramètres [55].

On peut d'ores et déjà classifier les troubles du langage en régions pathologiques, c'est ce qui montre que la voix peut être altérée ou modifiée tout le long de sa production, voire

disparaître, phénomène décrit par l'apparition d'une aphonie ou absence de voix complètement, surtout lors du cancer des cordes vocales.

3. Cordes Vocales saines

Avant de commencer l'étude des pathologies affectant la voix, il serait important de voir l'apparence de cordes vocales saines, car elles sont l'organe le plus important dans la production vocale (figure 3.1).

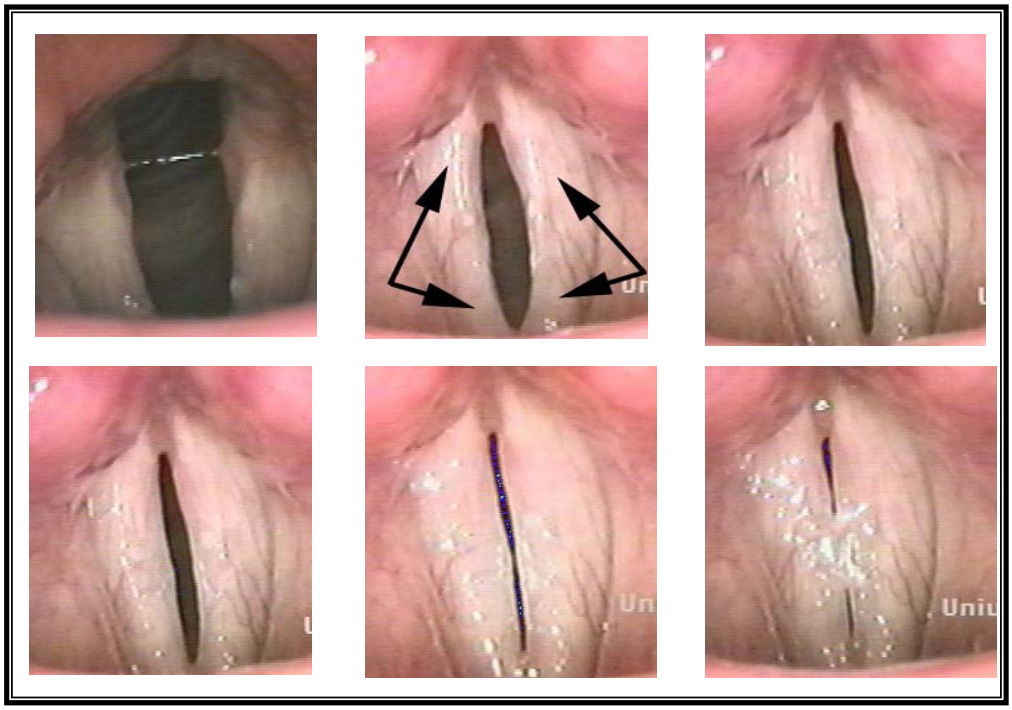


Figure 3.1 : Cordes vocales saines avec différents degrés d'aperture [56]

L'atteinte des cordes vocales par n'importe quelle maladie ou par simple irritation, lorsqu'on crie très fort pendant un événement quelconque, agit essentiellement sur leur manière de vibrer, soit en atténuant le mode vibratoire par une paralysie ou en les rendant plus enrouées ou plus âpres, (figure 3.2).

Les pathologies les plus importantes concernant les cordes vocales sont nombreuses

4. Pathologie des cordes vocales

Nous commençons notre étude sur les pathologies du langage par le lieu le plus important dans la phonation, les cordes vocales.

4.1. Les Nodules

Les nodules sont des nœuds durs en forme de pointe de flèche situés sur la partie vibratoire de contact des deux cordes vocales (figure 3.2).

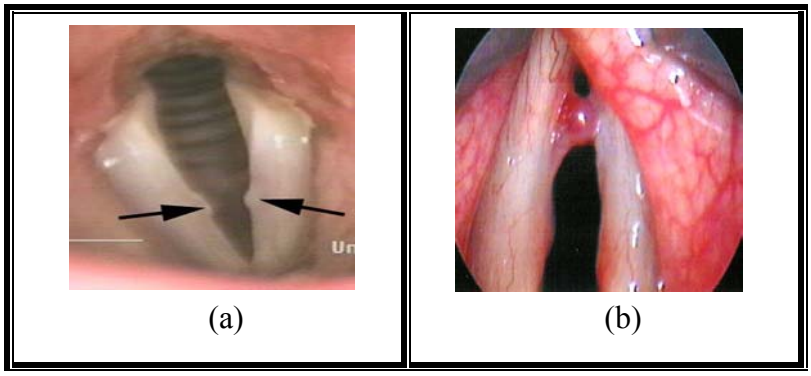


Figure 3.2 : (a, b) Nodules sur les cordes vocales

4.2. Paralysie des cordes vocales

Ce phénomène apparaît lorsque l'une des cordes Vocales, devient non élastique (figure 3.3), ou elle s'arrête presque de bouger. Cette situation est généralement traitée par intervention chirurgicale soit par une lipo-injection ou thyroplastie de la corde paralysée pour permettre une fermeture complète [56].

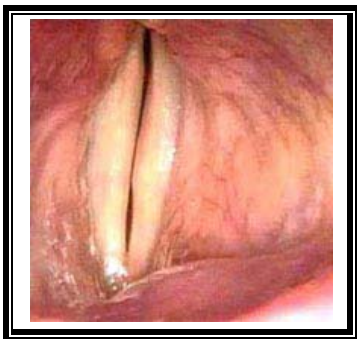


Figure 3.3 : Paralysie unilatérale des cordes vocales

4.3. Cordes vocales arquées

Lorsque les deux cordes vocales ne se ferment pas complètement, La surface restante affecte la production phonatoire, ceci peut rendre la voix très fragile [56], [57].

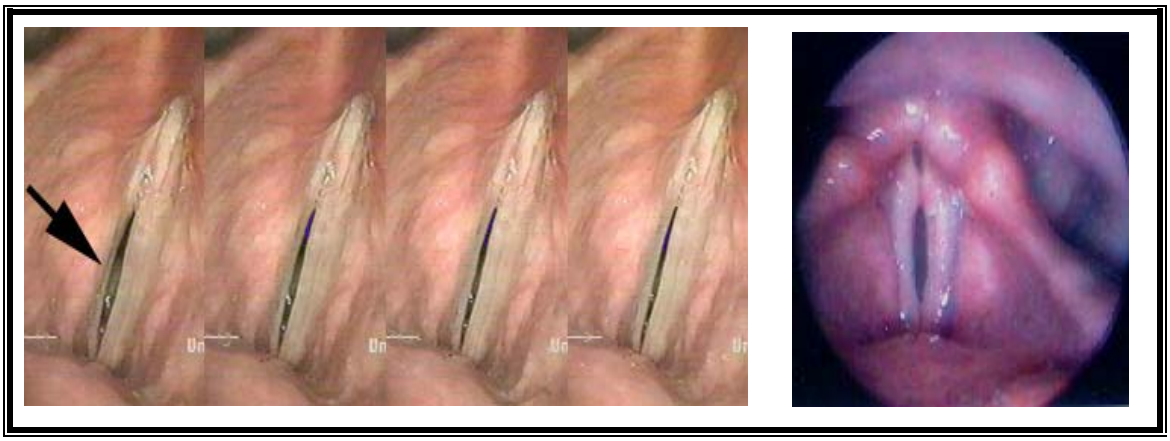


Figure 3.4 : Cordes vocales pathologiques présentant un arc à la fermeture

4.4. Polypes dans les cordes vocales

Les polypes, dérangent la fermeture et les vibrations des cordes vocales, cette situation cause l'enrouement, la fatigue des cordes et une diminution du mode musicale (figure 3.5).

4.5. Œdème de Reinke

L'oedème de Reinke est marqué par l'accumulation d'oedème et de fibrose dans la totalité de la corde vocale, (figures 3.5 a et b). L'étiologie est essentiellement le tabac associé ou non à l'alcool. Il s'agit d'une lésion bénigne mais qui peut être associée à un cancer développé ailleurs dans les Voies Aéro digestives Supérieures (VADS) dans 5 à 10 % des cas. L'accumulation de l'oedème peut conduire à un véritable ballonnement des cordes accompagnées de dyspnée. La répartition est globalement identique suivant les deux sexes, mais la population féminine consulte plus facilement du fait de la répercussion de cet oedème chronique sur la voix. En effet, le fait marquant de cette dysphonie est l'abaissement de la hauteur vocale. Ce trouble vocal est généralement bien accepté chez les hommes auxquels il donne un caractère "viril". A l'opposé chez la femme, la gêne est manifeste. La patiente est fréquemment appelée, Monsieur au téléphone.

Le traitement est microchirurgical et consiste à inciser la corde vocale sur sa face supérieure et à aspirer la glue. L'abstention tabagique prévient la récidive.

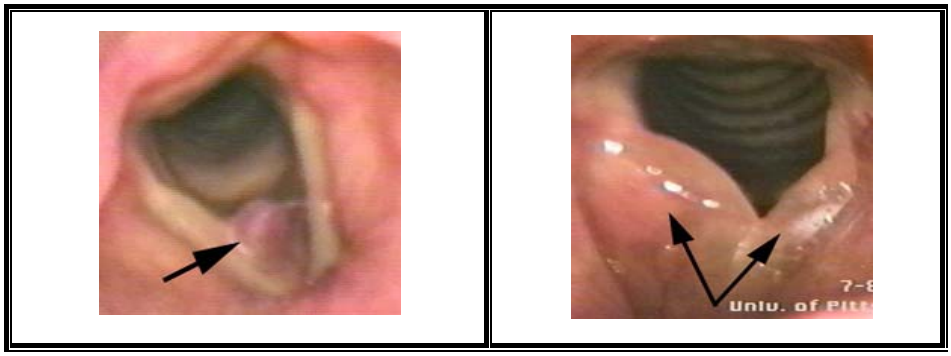


Figure 3.5 : Polype très distingué sur l'une des cordes vocales à gauche [56]

Figure 3.6 : Cordes vocales gonflées à gauche

4.6. Kyste localisé au niveau des cordes vocales

Le kyste muqueux, , est formé par l'obstruction de canal excréteur d'une glande muqueuse de la corde vocale. Macroscopiquement, on observe une voussure plus ou moins allongée. Plus le kyste est ancien et plus le liquide paraît épais. Le traitement est microchirurgical (fig 3.7 a, b)

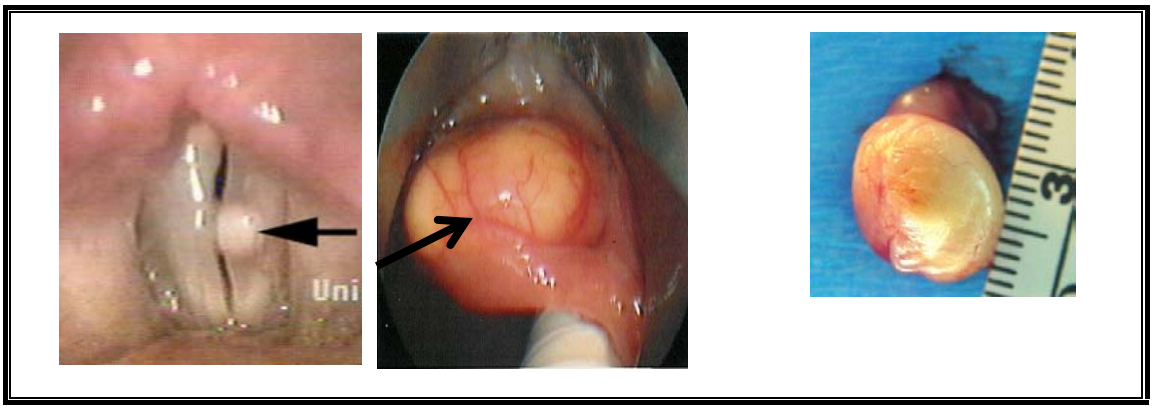


Figure 3.7 : Détail d'un kyste de différents patients

Figure 3.8 : Mesure d'un Kyste vocal après son extraction

4.7. Granulomes dans les cordes vocales

Petits amas granulomateux, inflammatoire « constitué de chair », c'est-à-dire de tissu conjonctif se développant sur la muqueuse du larynx et à ses dépens. Quelquefois ils sont observés au niveau de la trachée après une intubation du larynx et de la trachée pour une ventilation assistée, le plus souvent ayant eu lieu au cours d'une anesthésie générale ou un coma. La flèche sur la figure 3.9. Indique un tissu épais et irrégulier sur les cordes vocales [58], [59].

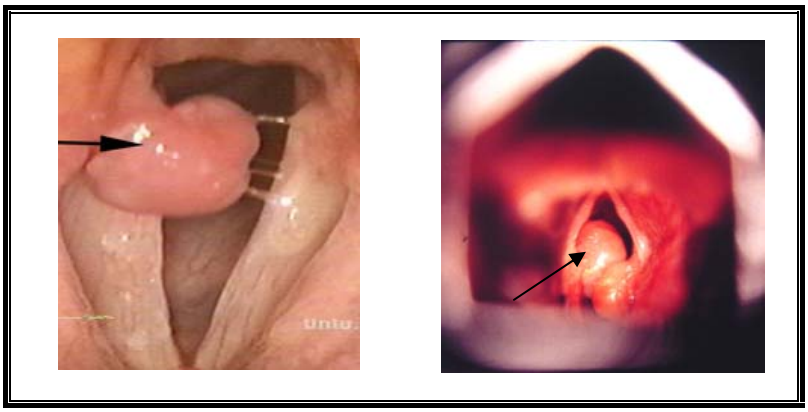


Figure 3.9 : Granulomes de différents patients

4.8. Papillomes laryngés

Les différentes flèches sur la figure 3.10 indiquent l'évolution des papillomes, dans le larynx, ceux-ci sont causés par une infection virale.

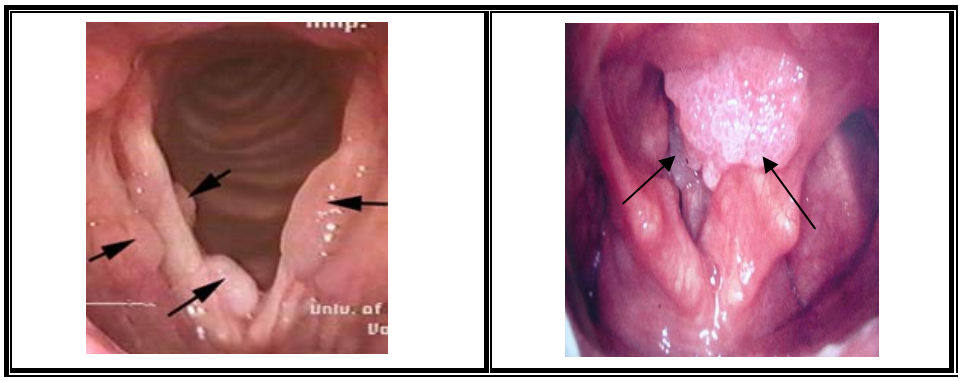


Figure 3.10 : Papillomes chez différents patients

Des études bien détaillées des pathologies indiquées sont décrites par « The National Center for Voice and Speech » et par la clinique Otolaryngology-Head & Neck Surgery [59].

4.9. Cancers des Voies Aero-digestives Supérieures "VADS"

Le cancer, maladie non bénigne, détectable après biopsie, se localise dans différents points du corps humain, entre autres dans le larynx et les différentes cavités vocales et nasales, appelées Voies Aérodigestives Supérieures (figure 3.11) [60].

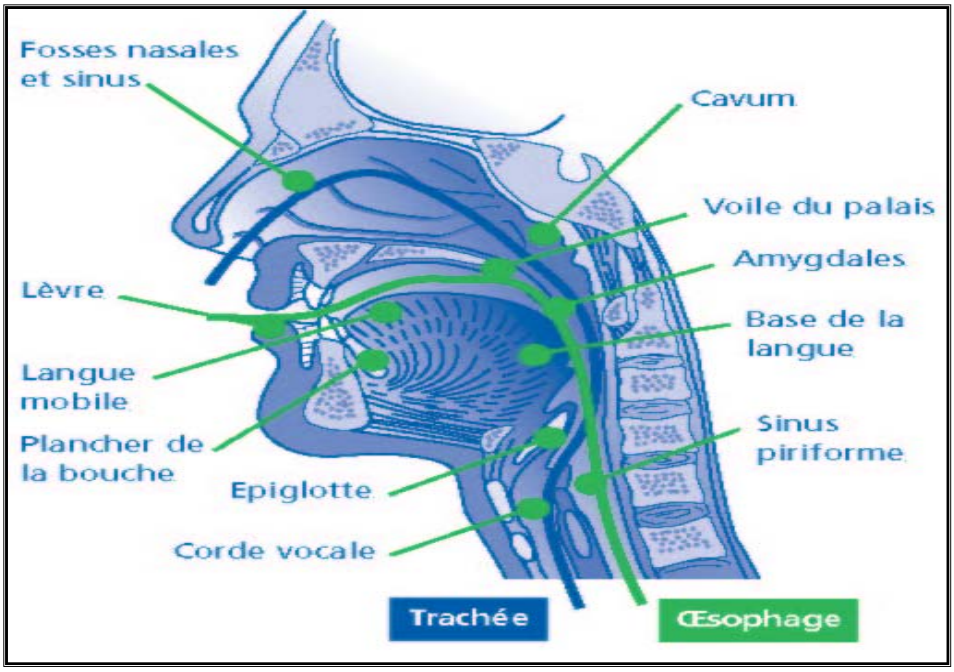


Figure 3.11 : Siéges possibles des cancers

4.10 Cancer du larynx

Ce cancer, favorisé par l'alcool et le tabac, se voit surtout chez l'homme après 50 ans.

Le signe révélateur est une dysphonie progressive : le patient se plaint d'un enrouement. La dyspnée (gène à la respiration), la dysphagie (difficulté pour avaler), sont beaucoup plus tardives.

Tout enrouement chronique nécessite un bon examen laryngologique direct.

L'ORL examine sous anesthésie locale le larynx grâce au miroir laryngé (figure 3.12). La tumeur est ainsi observée. Le médecin apprécie ensuite la mobilité du larynx et recherche des ganglions palpables. L'examen endoscopique (laryngoscopie) recherche une localisation cancéreuse oesophagienne et permet la biopsie à la pince de la lésion.

Diagnostic différentiel :

- ❖ tumeurs bénignes du larynx (polypes des cordes vocales, nodules vocaux…) ;
- ❖ tuberculose laryngée ;
- ❖ laryngite chronique (qui peut dégénérer) ;
- ❖ atteinte neurologique des cordes vocales.

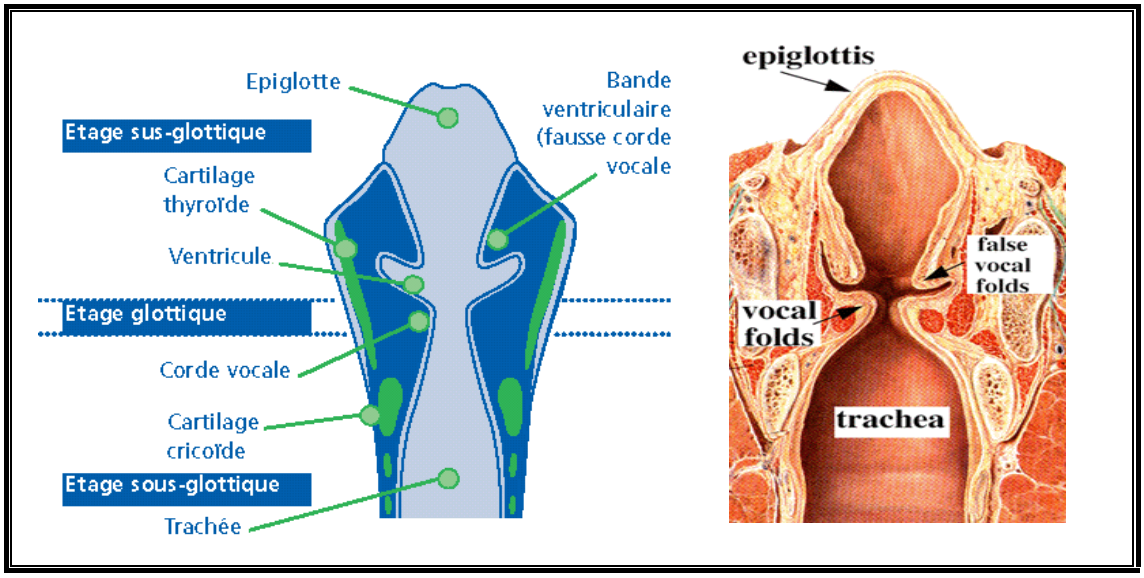


Figure 3.12 : Vue en coupe du larynx [60]

4.11 Cancer Du Pharynx (Oropharynx et Hypopharynx)

Le cancer du pharynx apparaît en général après les symptômes suivants :

- ♦ une gène ou une douleur d'un coté de la gorge;
- ♦ une sensation permanente d'un corps étranger ou d'angine traînante d'un seul côté ;
- ♦ une douleur à une oreille ;
- ♦ une difficulté à avaler, une gène à la déglutition d'un côté, parfois sans douleur ;
- ♦ une sensation de brûlure d'un côté de la gorge ;
- ♦ une modification progressive de la voix qui devient couverte, voilée ou rauque ;
- ♦ apparition d'une boule dans le cou qui correspond à un ganglion.

4.12. Cancer De La Bouche

Le cancer de la bouche apparaît en général après les symptômes suivants :

- ♦ une gène ou une douleur d'un coté de la bouche;
- ♦ une zone bourgeonnante ou creusée saignante, ne guérissant pas après un traitement d'une anomalie dentaire;

- ♦ un changement de la muqueuse persistant dans la bouche (tache rouge foncé ou blanche ressemblant à un aphte, mais à bords irréguliers);
- ♦ une gène au port d'un dentier;
- ♦ une douleur à une oreille;
- ♦ une difficulté à avaler;
- ♦ une sensation de chaud, au froid, au vinaigre, au citron.

4.13 Cancer des Cordes Vocales

Les cancers dans les zones glotttiques se révèlent par une modification progressive de la voix qui devient couverte, voilée, rauque (dysphonie). Cette modification persiste et s'aggrave progressivement. Elle est parfois précédée d'épisodes transitoires de laryngite où complique une laryngite chronique ancienne, fréquente chez les fumeurs et/ou les personnes travaillant en atmosphère chaude et sèche, ou chargée de poussières.

Les cancers des sub – glottiques siègent au niveau de l'épiglotte, par :

- ♦ une gène ou une douleur d'un seul côté de la gorge ;
- ♦ une difficulté à avaler ;
- ♦ une sensation permanente de corps étranger ou d'angine d'un seul côté ;
- ♦ une douleur à une oreille ;
- ♦ l'apparition d'une boule dans le cou qui correspond à un ganglion.

5. Pathologies des autres canaux vocaux

Nous présentons dans ce qui suit d'autres pathologies des canaux vocaux qui provoquent des anomalies sur le signal vocal

5.1 Bec de lièvre

C'est une déformation prénatale (figure 3.13), s'attaquant à la lèvre supérieure, d'origine génitale, pouvant être corrigée par une intervention chirurgicale [61].

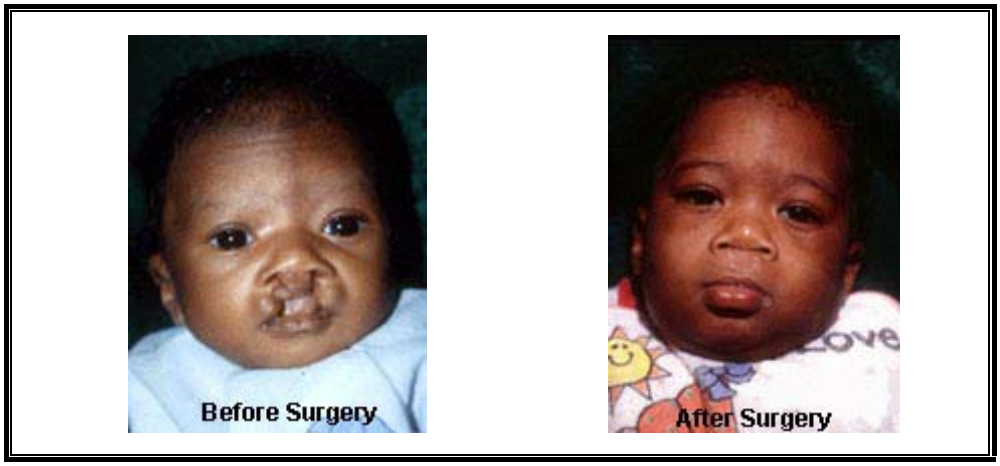


Figure 3.13 : Avant intervention / Après intervention

5.2 Palais enclavé

L'une des conséquences directes de l'absence du palais (figure 3.14) est l'hypernasalité, c'est une pathologie de la résonance de la voix, causée par le dysfonctionnement du mécanisme vélopharyngeale, celle-ci provoque un :

- ♦ nasonnement ouvert ou hyperrhinophonie : le voile du palais ne ferme pas le passage de l'air à la cavité nasale dans le cas de division palatine ou d'opérations des végétations notamment ;
- ♦ nasonnement fermé ou hyporhinophonie : pas de nasalisation pour les consonnes et les voyelles nasales [62].

Figure 3.14 : Palais totalement absent

Une intervention chirurgicale est à la base de la correction de cette pathologie.
Une étude très intéressante, portant sur les remarques d'enfants parlant l'Arabe avec un palais enclavé, est développée dans [62].

D'autres définitions, causes et traitements sont bien traités dans le guide vocologique ou « Guide to Vocolgy » émis par le Centre National de la Voix et de la Parole, le NCVS [63].

6. Classification des pathologies

La première partie a concerné les pathologies des organes intervenant dans la production ou l'altération de la voix, celles-ci sont classées comme suit :

- **Dysfonctionnement fonctionnelle :** l'organe existe mais, il y'a eu soit un mauvais apprentissage, soit une maladie en cours d'évolution ce qui présente un symptôme de pathologie de la parole, si la détection de la pathologie n'est pas effectuée à temps [64].
- **Dysfonctionnement organique :** l'organe existe ou est absent, cas du palais enclavé ou laryngectomie, mais ne peut exécuter la tache préconçue, soit par atrophie cas de la langue trop courte, soit par surdimensionnement cas du volume du palais démesuré,

Les différents défauts émanant de ces pathologies sont classés selon le diagramme de la figure 3.15.

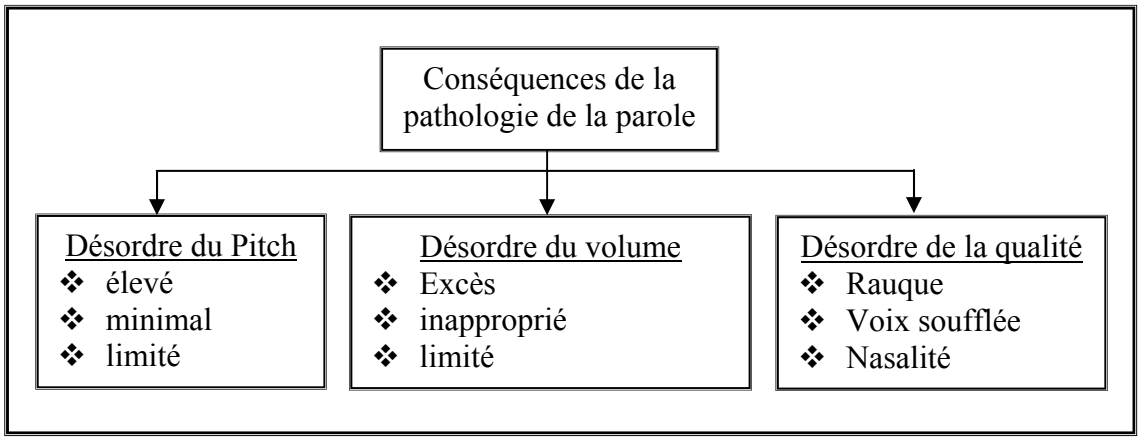


Figure 3.15 : Diagramme de classement des pathologies [65]

7. Défauts de la voix détectés par l'oreille

Nous présentons les différents défauts de la voix détectés par l'oreille humaine :

7.1. Blésement ou Zézaiement

Le blésement ou zézaiement est un défaut de prononciation qui consiste en la substitution de [ʃ] (une consonne chuintante) par [s] (une sifflante) et de [g] ou [j] (Consonnes chuintantes) par [z] (sifflante) [66].

7.2. Chuintement

Le chuintement est la prononciation du [s] et du [z] à la manière du [ʃ] et du [j] Français [66].
Exemple :

- J'ai pris l'autobu**s** jusqu'à la gare Saint-Lazare
- J'ai pris l'autobuch juchqu'à la gare Chaint-Lajare.

7.3. Rhotacisme

Le rhotacisme (terme formé à partir du grec ρ, [r]) est une modification phonétique complexe, consistant en la transformation d'un phonème en [r]; Dans d'autres langues comme le Français c'est avec le [z] que le [r] [67].

Pour la langue arabe c'est la confusion entre le [r] et le [ʁ].

Donc, au lieu de prononcer 'ريح ' [rihø].

La personne atteinte de rhotacisme prononce "غيح" [ʁihø],

7.4. Nasonnement

C'est l'altération du son de la voix; le nasonnement provient de la diminution de la résonance nasale par suite de l'obstruction du nez, de la présence de végétations adénoïdes, etc., et produit une déformation des syllabes nasales, [an], [on], [in], et des consonnes nasales, telles que [m], que l'on prononce [b] [68].

Exemple : En prononçant [pa] l'air ne doit pas passer par le nez.

Le nasonnement est l'inverse du rhume, où l'air ne peut pas passer par le nez.

7.5. Bégaiement

C'est le trouble de la communication affectant le débit et le rythme de la parole se traduisant :

- une forme clonique : répétition;
- une forme tonique : blocage;
- des troubles associés;

Si rien n'est entrepris, sur 4 enfants de 2 à 5 ans commençant à bégayer, 1 restera bègue à l'âge adulte. Il est nécessaire d'intervenir le plus tôt possible pour ne pas prendre le risque de la chronicisation [69], [70].

Les signes d'appel et manifestations du bégaiement se présentent comme suit :

- ♦ répétition de sons ou syllabes supérieures à 3 (ex: tou tou tou toupie) ;
- ♦ prolongation de sons ;
- ♦ blocage de syllabes ;
- ♦ répétitions de mots, de parties de phrases ;
- ♦ reprise d'énoncés ;
- ♦ hypertonie, blocages respiratoires lors de la prise de parole ;
- ♦ comportement ou modification du comportement : colères, retrait, timidité ; énurésie ;
- ♦ Comportement verbal: refus ou repli ;
- ♦ antécédents de bégaiement dans la famille [54].

7.6. Clichement

Le clichement est un défaut de prononciation se caractérisant par le fait d'additionner le son [ll] (double L) mouillé, positionné après certaines consonnes. Une consonne mouillée est articulée avec le son j. Par exemple l dans grisaille. Un exemple de clichement : prononcer chilluchoter au lieu de chuchoter [66].

7.7. Gammacisme

Défaut de prononciation se caractérisant par la difficulté voir l'impossibilité de prononcer les consonnes gutturales, [k] à la place de [g].

7.8. Retard de parole

Le retard de la parole est l'altération de phonèmes ou de groupes de phonèmes, par leur mise en ordre séquentielle à l'intérieur d'un même mot, le stock phonétique étant acquis. C'est la forme du mot dans son ensemble qui ne peut être reproduite.
Un parler bébé qui perdure au-delà de 4 ans est caractérisé par :

- ♦ des omissions mots raccourcis ou élidés : fleur/feur, herbe/è ;
- ♦ des inversions brouette / bourette ;
- ♦ des assimilations : lavabo lalabo ou vavabo ;
- ♦ des interversions : kiosque / kiokse ;
- ♦ des simplifications : parapluie / papui ... ;
- ♦ des substitutions : train / crain, fleur fieur ;

- des élisions de syllabes finales : pelle pè, assiette assiè...

Et de façon plus globale :

- des problèmes de perception auditive ;
- une mauvaise structuration de la perception du temps ;
- une mauvaise structuration de la chronologie des sons ;
- des difficultés motrices diverses ;
- une attention auditive labile ;
- une immaturité psychoaffective ;
- un refus de grandir ...

Fréquemment, un retard de langage et/ou un trouble d'articulation peuvent être associés au retard de parole [54].

7.9 Facio-Scapulo-Humeral (FSH)

Comme nous l'avons cité précédemment, certaines maladies ne concernent pas les cordes vocales directement, mais affectent les muscles, telle que la FSH qui est la dégénérescence de tous les muscles du corps humain, ou l'évolution du squelette osseux se développe avec quelques anomalies (Courbures de dos, déformation du bassin, etc …), tandis que les muscles d'une partie du corps s'atrophient, ceci donne lieu à une pathologie entre autres langagière qui affecte quelques phonèmes tels que le [b] et le [f] qui sont systématiquement remplacés par le [d] et le [θ] respectivement.

8. Pathologie concernée par notre travail

Nous avons ciblé notre travail sur une photologie, à savoir le sigmatisme occlusifs et constrictif, cette derniere nous a permis d'extrapoler notre méthode pour d'autre pathologies

8.1. Définition du Sigmatisme

Terme issu de la lettre grecque sigma, c'est la difficulté que présentent certaines personnes à prononcer le phonème [s]. Cette affection ne doit pas être confondue avec le zézaiement qui est un défaut de prononciation d'une personne prononçant le son [s] comme étant [z], le son [žø] comme [s] ou le son [s] comme [sø]. On dit également zozoter.

Ce défaut est généralement relié à une déviation par la langue, dans le processus d'écoulement d'air.

Il y'a deux types de sigmatismes

- **Sigmatisme latéral ou schlintement** : L'air s'échappe sur le côté de la bouche;
- **Sigmatisme interdental ou zozotement** : la langue vient buter contre les incisives supérieures ou se place entre les dents et produit une interposition linguale lors de l'émission des phonèmes [s] et [z] [66].

8.2. Sigmatisme des consonnes constrictives et occlusives

Le sigmatisme intervenant sur les consonnes constrictives peut avoir plusieurs appellations, suivant son origine :

- le sigmatisme nasal est dû à un positionnement de la langue qui rend impossible le passage de l'air par la cavité buccale ;
- le sigmatisme dorsal est également dû à un soulèvement de la langue excessif ;
- le sigmatisme occlusif est le remplacement systématique de toute consonne constrictive par la consonne occlusive dont le point d'articulation est le plus proche [44].

Ce sigmatisme concerne le remplacement de [ʃ], [j], [s], [z] par [θ] et [d] ou par [f] et [v][72] L'exemple suivant (tableau 3.1) illustre l'un des mots du corpus que nous avons pu enregistrer, et la prononciation défectueuse qui est transcrite phonétiquement.

Tableau 3.1 : Transcription Phonétique du mot présentant la prononciation pathologique

Mot Initial	Transcription phonétique saine	Transcription phonétique pathologique	Mot pathologique
شخصية	[ʃ a χ s i j a]	[θ a χ t i j a] [θ a χ ʃ θ i j a]	ثخثـيـة ثخشثيـة

9. La rééducation orthophonique

Les techniques de rééducation orthophonique utilisées sont multiples et sont en général adaptées au cas de chaque enfant : elles dépendent à la fois de la nature de son trouble, du caractère et de la problématique de l'enfant, de ses possibilités intellectuelles, d'où une prise en charge individuelle. La rééducation peut être assez technique ou davantage orientée vers la PRL (Pédagogie Relationnelle du Langage).

9.1. Au niveau articulatoire

Chaque fois qu'un phonème n'est pas correctement émis, le rééducateur sert de modèle acoustique et articulatoire à l'enfant. Il s'agit de créer de nouveaux automatismes audio kinesthésiques correspondant à une position organique correcte (le travail devant la glace permettant à l'enfant de la visualiser et d'en prendre conscience).

9.2. Bain de langage

L'orthophoniste essaie de mettre le sujet à l'aise en l'introduisant dans une sphère de langage au moyen de jeux : lotos, jeux de mémorisation, imagiers, lexidata.

9.3. Travail de perception

Cette phase concerne l'aide à la discrimination et à la mémorisation auditive, ainsi le sujet se rappelle de ce qu'il faut bien prononcer.

9.4 La Dimension Relationnelle

Les liens privilégiés qui se nouent entre l'enfant et le rééducateur, ne sont pas qu'un moyen d'appliquer la technique. Ils ont aussi pour but l'épanouissement de l'enfant. La relation duelle est gratifiante pour lui. Elle lui permet de se confier, de ressentir du plaisir dans un échange avec un adulte.

En lui montrant qu'il est capable de décider et de réaliser lui-même, en lui faisant constater ses propres progrès et en le valorisant, l'orthophoniste aide l'enfant à prendre confiance en lui, en ses capacités, et à trouver la volonté de surmonter ses échecs.

D'autre part, il est évident que la possibilité de s'exprimer oralement ne suffit pas, encore faut-il avoir envie de communiquer. Seule une situation duelle affectivement sécurisante et valorisante peut faire naître et grandir chez l'enfant ce besoin de la communication [73].

10. Etat de l'Art des logiciels de thérapie phonétique existants

L'introduction de l'outil informatique ou électronique dans l'aide à la décision, la correction par séances répétées, à l'évaluation temporelle, ainsi que le feedback visuel et vidéo, déterminent les phases de la thérapie de la pathologie langagière :

- Parole expressive :

- aphasie : Difficulté à prononcer un mot dans la bonne direction(en sens inverse)
- apraxie, dysarthrie : parole intelligible, discordance musculaire.
- dysphonie : difficulté de contrôler la hauteur ou le timbre de la voix.

♦ Compréhension de la parole

♦ Mémoire et raisonnement

- mémoire à court terme.
- incompréhension des relations entre mots
- incompétence à suivre les étapes logiques

Cette approche est taclée par différentes firmes de productions de logiciels commerciaux citées dans l'annexe 1.

Notons toutefois que les techniques utilisées dans ce type de logiciels de thérapie langagière se basent sur les techniques de reconnaissance, de distance phonémique, de biofeedback ainsi que de séances d'enregistrement, d'écoute pour une correction répétitive.

11. Conclusion

A travers ce chapitre, nous avons cité un éventail assez large de pathologies langagières, ayant trait aux variations phonétiques.

Le choix du sigmatisme occlusif, pathologie concernant les défauts de prononciation du [ʃ] et [s] prononcés comme [θ] et [z], porte essentiellement sur la disponibilité d'un corpus maladif ainsi qu'une tendance à mettre une méthodologie d'évaluation par un système d'aide, qui sera bénéfique à l'orthophoniste et au patient. Cette méthodologie sera base sur la technique de reconnaissance de formes.

Dans le chapitre suivant, différentes techniques d'analyse du signal vocal ainsi que les techniques de classification seront présentées, en vue de définir un processus permettant:

❖ d'aider l'orthophoniste en lui donnant un score de bonne ou de mal prononciation du patient.

❖ de faire travailler le patient à la maison, en s'enregistrant, en se réécoutant et en observant les diverses articulations à réaliser afin de bien prononcer.

Chapitre 4

Techniques d'Analyse du Signal vocal appliquées

1. Introduction

Dans ce chapitre nous présentons la chaîne de reconnaissance en développant l'étape de recherche des caractéristiques du signal de la parole ainsi que les techniques de classifications des formes qui sont utilisées en RAP à savoir l'alignement temporel, les chaînes de Markov et les modèles connexionnistes. La présentation des modèles connexionnistes sera précédée d'un bref rappel sur les connaissances de la neurobiologie qui ont servi de base à l'établissement des techniques neuromimétiques.

Nous allons exposer en premier lieu un état de l'art sur les différentes techniques d'analyse du signal vocal, en mettant en évidence notre choix concernant les méthodes utilisées lors de la segmentation phonémique, à travers des comparaisons avec des travaux de recherche antérieurs. Nous introduisons les différentes méthodes de paramétrage des données tels que les coefficients de prédiction linéaires, neuro-predictifs, les coefficients cepstraux, et leurs différentes variantes de génération (la transformée en cosinus, les ondelettes) et les performances subséquentes, ainsi que leur utilisation pour la segmentation et la reconnaissance phonémique, via les différentes techniques telles que les chaînes de Markov, les Réseaux de Neurones et les Multi-Gaussiennes,

Quelques techniques d'analyse serviront, dans l'élaboration de notre système d'aide interactif, en vue de l'évaluation de prononciations phonémiques pathologiques, ciblant deux objectifs essentiels :

- la correction de la prononciation ;
- le suivi de l'évolution de la maladie (appliquée comme exemple à la maladie facio-scapulo-huméral : FSH).

2. Méthodologie du Travail

L'analyse des données issues du signal de la parole est très complexe, à cause de la multitude ou la redondance de l'information vocale. En prenant en compte les informations citées auparavant où l'analyse sera orientée vers des paramètres généraux discriminants, englobant différentes occurrences de l'information du point de vue temporel, spectral, spatial, perceptuel et/ou prosodique, ….

Dans notre travail, nous ne présentons pas toutes les bases mathématiques des méthodes utilisées, mais nous essaierons de justifier notre choix, en expliquant le trajet des données à travers les différents étages de traitement.

3. La chaîne de reconnaissance

L'application de la Reconnaissance de Formes pour la Parole se traduit par le traitement de différents blocs (figure 4.1).

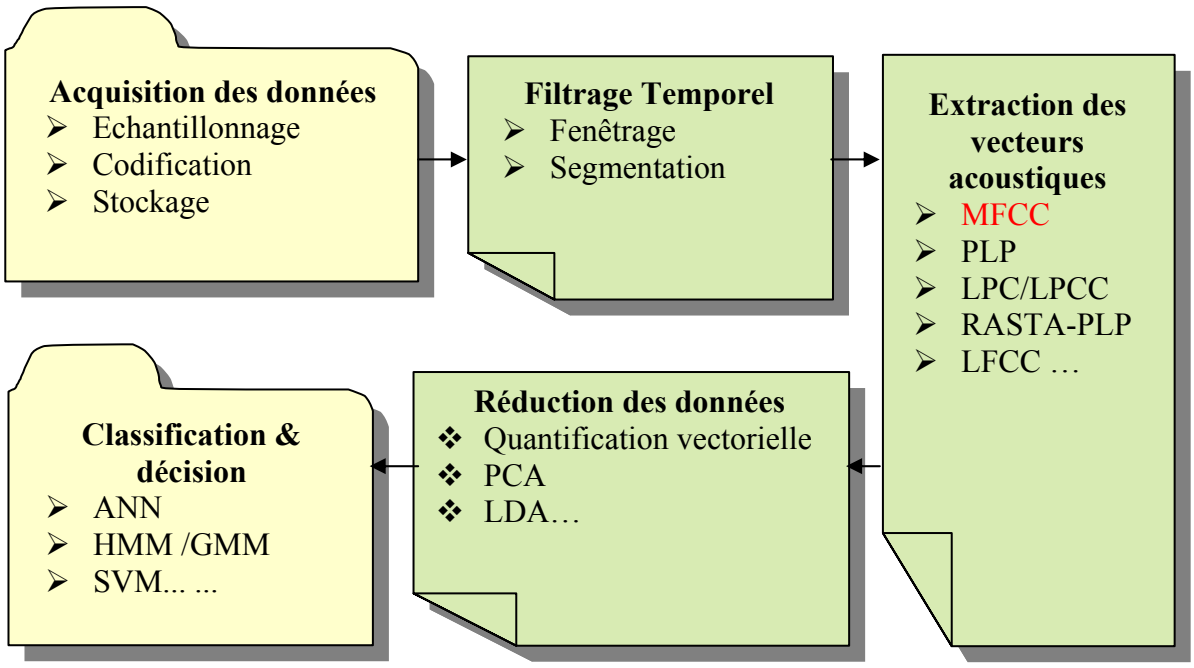


Figure 4.1 : Schéma global du processus de reconnaissance utilisé en RAP

*** Acquisition & Prétraitement du signal vocal**

L'acquisition des données présente en elle-même une chaîne de traitement, où l'erreur même infime se voit propagée sur les étages en aval. Elle incombe ainsi des erreurs relativement considérables, qui peuvent fausser l'information et donner lieu à des estimations totalement erronées.

*** Enregistrements des données sonores**

La phase d'enregistrement concerne la conversion du signal continu de la parole en une suite ou succession d'impulsions normalisées, à une fréquence de 16 KHz. Cette fréquence d'échantillonnage est suffisante pour prendre en compte la variabilité du signal vocal enregistré.

*** Segmentation et chevauchement**

Après l'acquisition et le stockage du corpus d'étude, suivant un codage bien défini, chaque mot ou phrase du signal enregistré est segmenté en fenêtres de durée fixe, obéissant aux deux contraintes suivantes :

♦ la stationnarité du signal parole (moyenne et variance constantes durant la trame ou la fenêtre temporelle d'analyse);

♦ la durée supérieure à l'inverse de la fréquence fondamentale [74].

Le tableau 4.1. Récapitule les différentes durées utilisées dans la littérature [3, 75, 76, 77, 78]

Tableau 4.1 : Durées primaires d'analyse ainsi que la durée de chevauchement

Méthodes	Durée ms	Chevauchement ms
Lawrence Rabiner	45	30
A.R El Obeid Ahmed et al	20	05
Yasuhi Tsubota et al	20	10
Alizera A. Dibazar et al	10	02.5
Georg Stemmer et al	20	10

La durée d'analyse sélectionnée dans notre travail est de 20 ms.

Après la conversion, il faut éviter les effets de bords de la segmentation fenêtrée, car les segments de 20 ms sont contigus. Lors du changement d'une fenêtre temporelle à une autre, l'apparition de transitions se fait remarquer, d'où l'on doit recourir à un chevauchement ou à un recouvrement de 10 ms (figure 4.2).

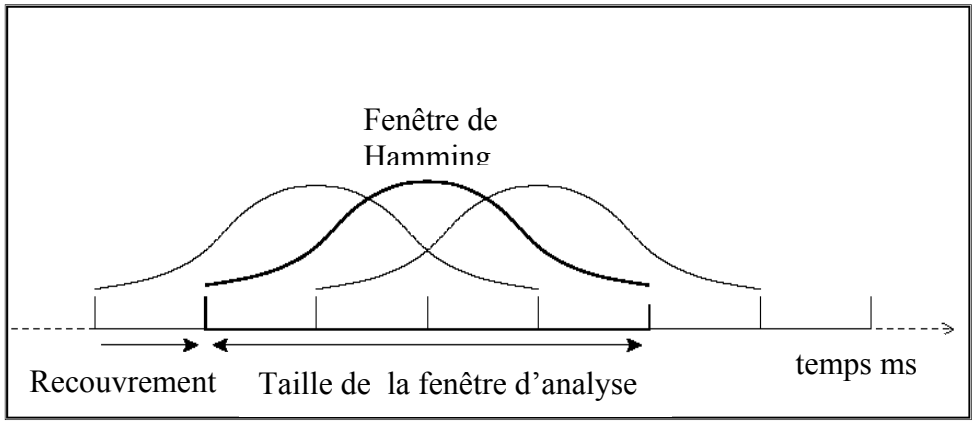


Figure 4.2 : Segmentation temporelle et Recouvrement

*** Filtrage et préaccentuation**

Pourquoi filtrer un signal si on veut récupérer toutes les informations ?

Il existe deux explications pour l'utilisation du module de préaccentuation [79] :

- ❖ la partie voisée du signal de la parole présente une accentuation spectrale approximative de (-20) dB/décade. Le filtre de préaccentuation permet de compenser cette accentuation avant d'analyser le spectre, ce qui améliore cette analyse ;
- ❖ l'audition est plus sensible dans la région du spectre autour des 1 kHz. Le filtre de préaccentuation va donc amplifier cette région centrale du spectre.

En général le filtre de préaccentuation est de la forme :

$$y'_n = y_n - \alpha . y_{n-1} \qquad (2)$$

Où n représente le n^{ième} échantillon calculé

et $\alpha \in$ [0,9 et 1].

Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Ces effets parasites sont réduits en appliquant aux échantillons de la trame une fenêtre de pondération comme par exemple la fenêtre de Hamming [61].

$$y''_n = y'_n * w_n \qquad (3)$$

$$w_n = 0.54 - 0.46.\cos\left(2.\pi.\frac{n}{N-1}\right) \quad avec \quad 0 \le n \le N-1 \qquad (4)$$

4. Paramétrisation du signal vocal

Différentes méthodes de représentation du signal, existent. Certaines ont été spécifiquement développées pour l'étude ou la compression de signaux de parole afin de diminuer le nombre d'opérations lors du traitement du signal vocal, nous allons développer dans ce qui suit les principales paramétrisations du signal vocal ainsi que l'une des plus récentes.

4.1. Le codage Prédictive Linéaire (LPC) et Coefficient Cepstrale Prediction Linéaire (LPCC)

Le codage Prédictive Linéaire ou LPC est basé sur le modèle de production de la parole, qui considère que l'appareil de production de la parole est constitué d'une source (source pseudo-périodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal) (figure 4.3). Le signal de parole peut être ainsi modélisé comme étant le signal en sortie d'un filtre H(z) dont la source d'excitation à l'entrée du filtre u(t) est soit une source de série d'impulsions quasi-périodiques, soit un bruit blanc.

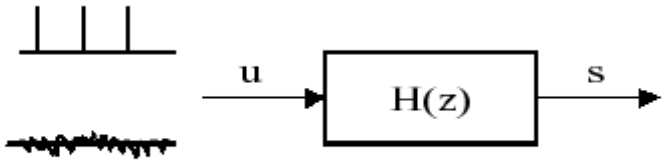


Figure 4.3 : Modèle source-filtre de production de la parole

L'analyse LPC repose sur l'hypothèse que le filtre est un filtre tous-pôles, avec cette hypothèse, le signal de la parole peut être considéré comme un signal auto ré-gressif :

$$s(n) = \sum_{k=1}^{p} a_k.s(n-k) + G.u(n) \tag{5}$$

$$H(z) = \frac{S(z)}{G.U(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{6}$$

où G est le coefficient de gain, $a_k$sont les coefficients LPC et p est l'ordre du filtre.

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance ou grâce à des méthodes fondées sur le calcul de la matrice d'auto corrélation [10].

Les coefficients LPCC (c_n) sont dérivés directement des coefficients LPC à travers le système d'équations suivant :

$$\begin{cases} c_0 = \ln G \\ c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \quad 1 \le m \le p \\ c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \quad m > p \end{cases} \tag{7}$$

où G est le coefficient de gain du modèle source-filtre.

4.2. Coefficients Cépstraux Echelle Mel (MFCC)

Les coefficients LPC ont un défaut assez remarquable c'est leur corrélation, donc pour différentes trames du signal, on obtient des coefficients très proches, ceci, n'aide pas le

système de discrimination basée sur ces coefficients, alors il y a eu recours à passer dans un domaine cepstral pour déterminer des coefficients assez discriminants et robustes au bruit qui est engendré par le conduit vocal, ou tout autre source de bruit environnant. Il y'a lieu de considérer le point suivant: Remarquer que la contribution à la perception des sons de la parole des hautes fréquences est plus faible que celle des basses fréquences, donc un changement d'échelle s'avère nécessaire. A l'origine de l'échelle Mel nous pouvons citer le psychologiste américain **Stanley Smith Stevens**, fondateur du Laboratoire de Psychoacoustique de Harvard, ainsi que Volkmann et Newman [37].

L'échelle Mel est définie par :

$$B(f) = 2595 * \log(1 + \frac{f}{700}) \qquad (8)$$

où :

f est la fréquence en Hz et

B*(f)* la fréquence suivant l'échelle de Mel.

Le calcul des coefficients cepstraux est effectué selon le schéma de la figure 4.4:

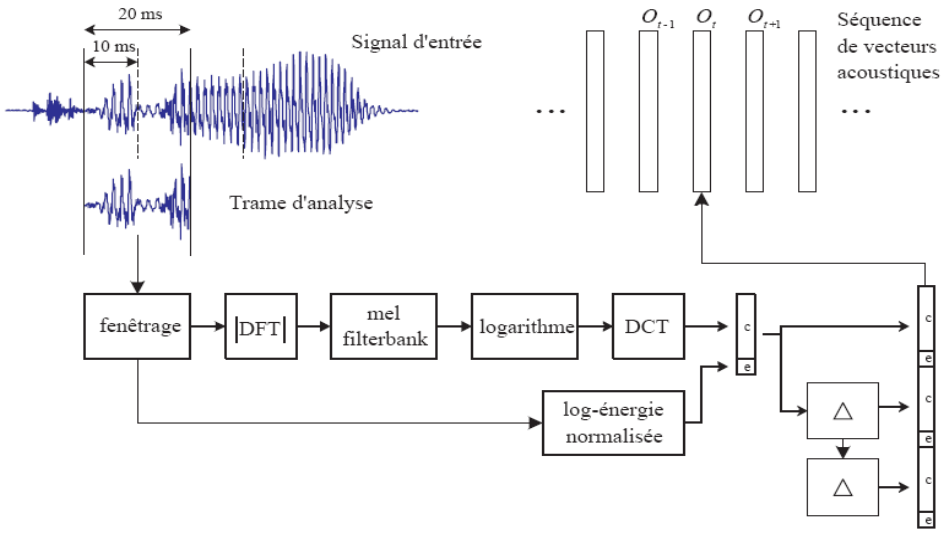


Figure 4.4 : Chaîne d'analyse du signal produisant les coefficients MFCC [61].

Soit un signal $x(n)$ discret avec $0 \leq n < N$, où N représente le nombre d'échantillons d'une fenêtre d'analyse, la Transformée de Fourier est alors définie par :

$$X(k) = \sum_{n=0}^{N-1} x(n).e^{\frac{-j.2.\pi.n.k}{N}} \quad \text{avec} \quad 0 \leq k < N \qquad (9)$$

Le spectre du signal est filtré par des filtres triangulaires (figure 4.5), dont les bandes passantes sont équivalentes en domaine fréquence Mel. Les points de frontière B*(m)* des filtres en échelle de fréquence sont calculés à partir de la formule :

$$B(m) = B(f_1) + m.\frac{B(f_h) + B(f_b)}{M+1} \qquad avec \qquad 0 \le m < M+1 \qquad (10)$$

- M désigne le nombre de filtres ;
- f_h désigne la fréquence la plus haute du signal ;
- f_b désigne la fréquence la plus basse du signal ;

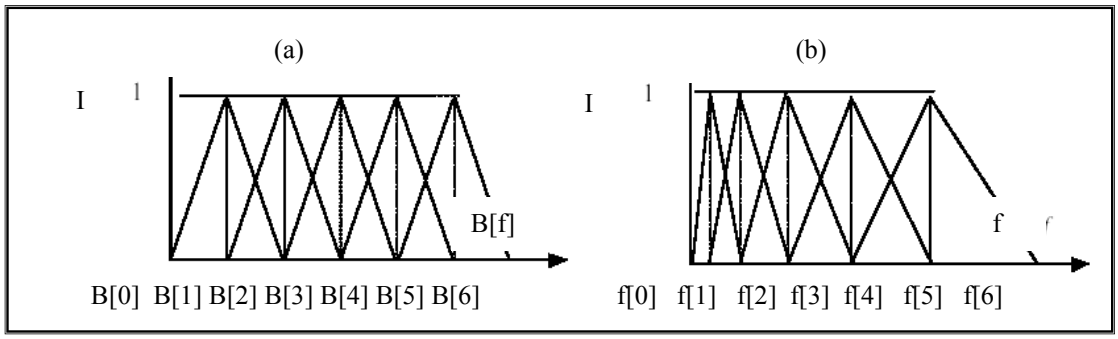


Figure 4.5 : Filtre triangulaire passe bande.
(a) en Mel fréquence B(f) (b) en fréquence f

Dans le domaine fréquentiel, les points *f(m)* discrets correspondants sont calculés d'après:

$$\mathrm{f}(m) = \left(\frac{N}{F_s}\right).B^{-1}\left[B(f_b) + m.\frac{B(f_h) + B(f_b)}{M+1}\right] \qquad (11)$$

avec $0 \le m < M+1$

et

$$B^{-1}(i) = 700.\left(10^{\frac{i}{2595}} - 1\right) \qquad (12)$$

Les coefficients des filtres sont calculés par :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & si \quad f(m) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & si \quad f(m) \le k \le f(m+1) \\ 0 & si \quad k \ge f(m+1) \end{cases} \qquad (13)$$

Ensuite on multiplie les énergies de X(k) par les coefficients $H_m[k]$ et on calcule leur logarithme :

$$E[m] = \log\left[\sum_{k=0}^{N-1} |X(k)|^2 . H_m[k]\right] \quad (14)$$

avec $0 \leq m < M$

Les coefficients MFCC de fréquence en échelle MEL sont obtenus par la transformée inverse des coefficients en sortie des filtres.

Remarque : Le nombre de MFCC est moins grand que le nombre de filtres, donc une transformée en cosinus est plutôt utilisée.

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi . n . (m + \frac{1}{2})}{M}\right) \quad (15)$$

avec $0 \leq m < M$

Ces coefficients ont été utilisés pour la reconnaissance automatique des chiffres en Anglais en conditions bruitées [49], pour la détection et reconnaissance des sons pour la surveillance médicale [70], pour la comparaison de paramètres de reconnaissance des phonèmes de la langue Arabe [75], pour la détection d'un signal de parole pathologique [77], pour la comparaison de plusieurs modélisations acoustiques pour des systèmes de reconnaissance embarqués, pour la reconnaissance/ vérification du locuteur, pour la conception d'un système hybride pour l'identification de traits phonétiques complexes de la langue Arabe, lors de la reconnaissance et la vérification de l'Anglais par des étudiants japonais dans un système d'apprentissage automatique [76].

Nous ne pouvons pas citer toutes les études établies, utilisant les MFCC, toutefois le nombre de coefficients est en général pris égal à 13, ets parfois réduit à 12, en considérant deux points essentiels :

- le premier coefficient C_0 représente l'énergie de la trame et ne peut réellement contribuer à la segmentation ou la reconnaissance;
- les coefficients de 1 à 12 représentent l'enveloppe cepstrale plus ou moins lissée, les hautes variations fréquentielles étant supprimées.

A cette étape d'extraction des caractéristique du signal vocale, le nombre de vecteurs acoustiques ainsi extrais reste énorme, par conséquent nous modélisons leur distribution par

un Mélange de Gaussiennes, car cette modélisation statistique a donné des preuves quand il s'agit d'un nombre très important de vecteurs observés.

En vue de maximiser l'information utile, la variation première et seconde des coefficients MFCC est fortement recommandée [75]. Toutefois la problématique de dimensionnement apparaît; Pour cette raison, l'adjonction de la variation des coefficients Cepstraux ΔMFCC et de leur accélération ΔΔMFCC est un point à bien évaluer judicieusement, d'un point de vue complexité de calcul et dimensionnement des matrices de covariances.

A titre d'exemple, si chaque fenêtre temporelle d'une durée de 20 ms avec un recouvrement de 10 ms avec une fréquence d'échantillonnage de 16 KHz soit (16 échantillons par ms), le nombre de paramètres à traiter par trame est mentionné dans le tableau 4.2.

Tableau 4.2 : Calcul des nombre de vecteurs acoustiques

Types Coefficients	**Nombre Coefficients**	**Nombre de vecteurs par trame de 10 ms**
MFCC	12	12 x 16 = 192
MFCC+Δ MFCC	24	12 x 24 = 288
MFCC+Δ MFCC + ΔΔ MFCC	36	12 x 36 = 432

Comme illustré dans la figure 4.5, l'utilisation du couplet (TFD, DCT) n'est pas unique, différentes variantes ont été implémentées [70]. Elles concernent l'utilisation des fonctions mentionnées dans le tableau 4.3.

Tableau 4.3 : Variante des MFCC

Fonction	**Fonction inverse**
FFT	FFT^{-1}
FFT	DCT^{-1}
DWT	DWT^{-1}
DCT	DWT^{-1}

On remarquera que dans la littérature d'autre coefficient cepstraux sont utilise a savoir LFCC (Linear Frequency Cepstral Coefficients), ces derniers sont calculés de la même manière que les MFCC, mais avec la différence que les fréquences des filtres sont uniformément réparties

sur l'échelle linéaire des fréquences, et non plus sur une échelle mel.

4.3. Codage Neuro Predictif (NPC)

Le codage NPC est une extension du codage LPC, appliquée au cas non linéaire, c'est un codeur prédictif. Il fait partie des codeurs temporels. L'intérêt principal du codage NPC est la modélisation non linéaire des signaux de parole. Cette modélisation est nécessaire pour augmenter les performances en reconnaissance Automatique de la Parole.

4.3.1. Principe d'extraction des caractéristiques Acoustiques

Le processus de production de la parole est différent selon le type de phonème prononcé voisé, non voisé, Dans le cas des phonèmes voisés le flux d'air p est un train d'impulsions de période N. Ce flux d'air est modifié par les contributions glottales g, le rayonnement aux lèvres r et celle du conduit vocal v. Le signal de parole y résultant est la convolution de p par les réponses impulsionnelles g, r, v des trois parties du processus de production de la parole (figure 4.6).

$$y = p * g * v * r \tag{16}$$

Dans le cas de la production de sons non voisés, les cordes vocales ne vibrent pas. Le flux d'air u est considéré comme un bruit blanc.

$$y = u * v * r \tag{17}$$

Aux équations (2) et (3), il faut associer d'autres modèles de production comme dans le cas des nasales. En effet, il est également nécessaire de modéliser la contribution de la cavité nasale.

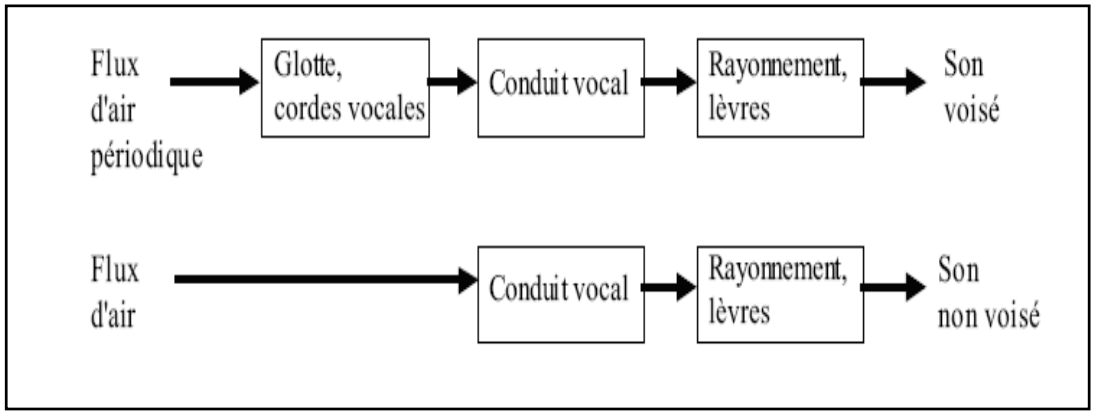


Figure 4.6 **:** Processus de production de la parole, cas des phonèmes voisés et non voisés

Pour des phonèmes ayant des caractéristiques proches, c'est-à-dire pour lesquels le processus de production est pratiquement identique, l'extraction des caractéristiques optimales

consisterait à extraire seulement celles qui sont discriminantes. Le codeur NPC qui intègre ce concept est réalisé à base d'un réseau de neurones MLP (*Multy Layer Perceptron*) à deux couches (figure 4.7). Les poids de la première couche w sont communs à tous les phonèmes tandis que les poids de la seconde couche a_i sont spécifiques à chaque classe de phonèmes [80].

Considérons 'i' et 'j' deux allophones appartenant respectivement aux classes de phonèmes différentes 'C_i'et 'C_j'. Les modèles NPC associés à ces deux allophones sont les suivants :

$$F(w,a_i)=H(a_i) \circ G(w)$$
$$F(w,a_j)=H(aj) \circ G(w) \qquad (18)$$

Les modèles NPC F(w, ai) et F(w, aj) sont différents tandis que G(w) est commune aux deux allophones i et j. Elle rassemble donc les caractéristiques communes alors que celles qui sont discriminantes sont représentées par les fonctions $H(a_i)$ et $H(a_j)$.

Nous généralisons le modèle aux allophones à tous les phonèmes, et avec les hypothèses d'apprentissage du modèle NPC. Nous pouvons considérer que les poids de la première couche 'w' modélisent les caractéristiques communes des classes de phonèmes tandis que les poids de la seconde couche 'a_i' modélisent les caractéristiques discriminantes

Notre travail est basé sur le codeur NPC à une structure de type MLP à une seule couche cachée (figure 4.7).

En effet on peut considérer les réseaux de type MLP, comme une extension naturelle au domaine non linéaire des méthodes linéaires de traitement adaptatif du signal.

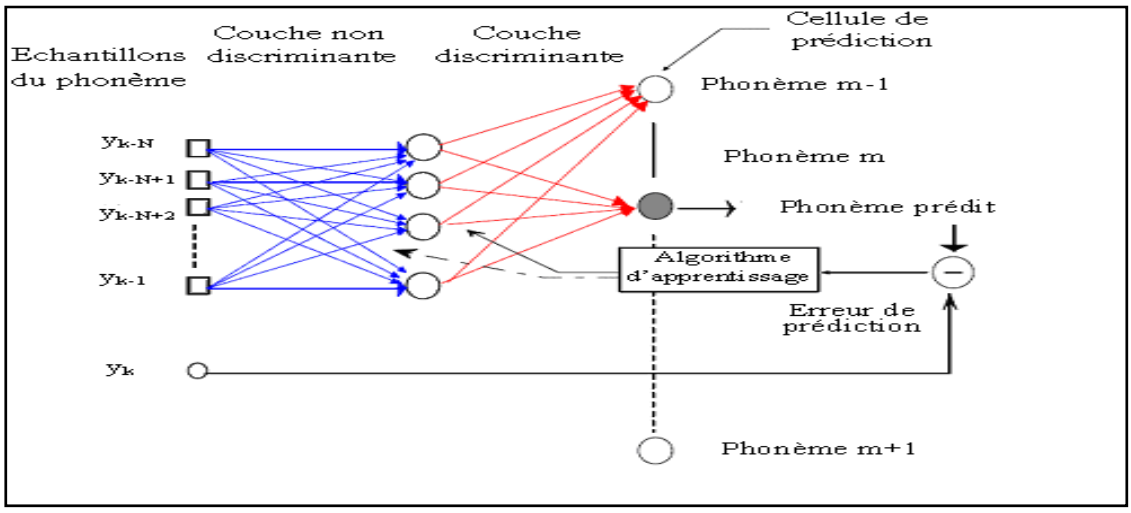


Figure 4.7 : Codeur NPC à une structure de type MLP.

4.3.2. Fonctionnement du Codeur NPC

Le fonctionnement du codeur NPC s'effectue en deux phases, après avoir pris des intervalles de temps stationnaires et les dimensions de l'entrée de notre réseau de n échantillons de l'intervalle stationnaire, et la sortie représentant la prédiction n+1 ; n étant inferieure à la dimension de l'intervalle stationnaire, cette bande de (N échantillons) balayera toute la zone stationnaire (figure 4.8).

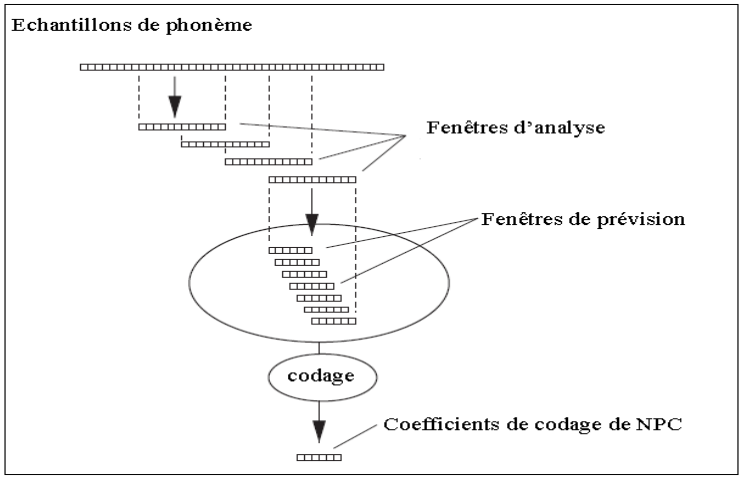


Figure 4.8 **:** Fonctionnement du codeur NPC

La phase de paramétrisation concerne le calcul des poids de la première couche. Durant cette phase, la première couche est commune à tous les exemples de phonèmes, tandis que la seconde est propre à chaque exemple (figure 4.9).

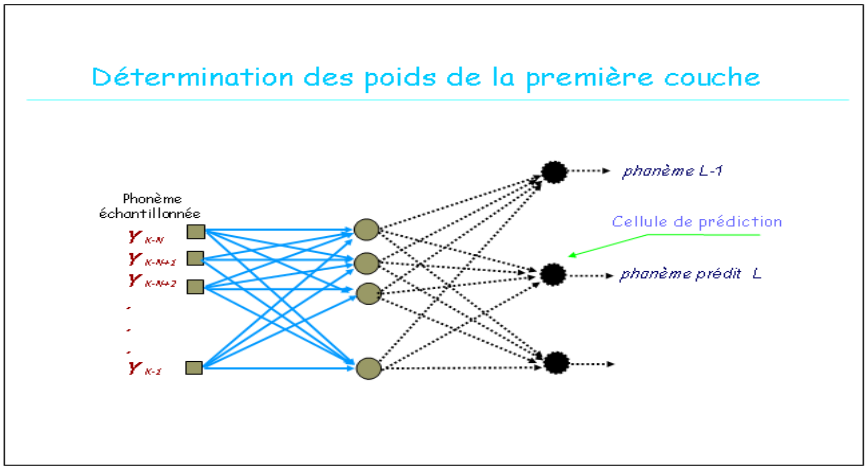


Figure 4.9 **:** Phase de paramétrisation.

La phase de codage concerne le calcul des poids de la seconde couche, car après la phase de paramétrisation, les poids de la première couche sont fixes. La phase de codage est considérée

comme une phase de test. Elle consiste donc à produire un code NPC, qui est constitué du vecteur poids de la seconde couche. Ce Calcul est effectué par prédiction successive des échantillons du phonème (figure 4.10).

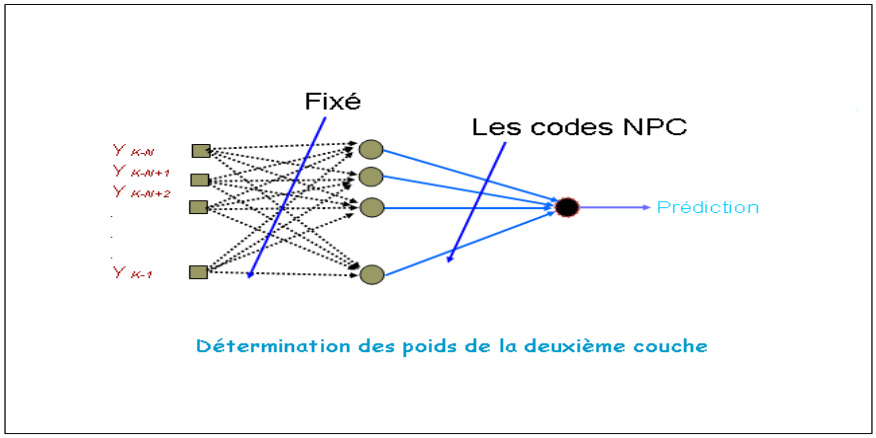


Figure 4.10 **:** Phase de codage

Les phases de Paramétrisation et de codage reposent sur le principe de la minimisation de l'erreur quadratique qui est en fait l'erreur de prédiction. Cette minimisation s'effectue par l'algorithme de rétropropagation.

4.3.3. Estimation des poids du réseau

Dans la modélisation non linéaire, Etant donnée une séquence d'échantillons $\{y_{k-i,i=1,\ldots,n}\}$ extraite d'un phonème ϕ quelconque (figure 4.7).

Le réseau effectue la prédiction de l'échantillon suivant y_k^{ϕ} en fonction des 'n' échantillons précédents.

Soit F de $R^n \xrightarrow{a} R$ la fonction réalisée par le réseau. La prédiction $\hat{y}_k^{\phi}$ s'écrit :

$$\hat{y}_k^{\phi} = F\left(\left[y_{k-1}^{\phi}, y_{k-2}^{\phi} \ldots y_{k-n}^{\phi}\right]^T\right) \qquad (19)$$

Nous désignons par x_k^{ϕ} le vecteur des échantillons précédents, avec $\hat{y}_k^{\phi} = F\left(x^{\phi}\right)$

Soit $\Omega = \lfloor w_{ij} \rfloor$ le vecteur des poids du réseau. Ces poids sont calculés de sorte à minimiser l'erreur de prédiction $E_k = y_k - \hat{y}_k$ pour toutes les séquences d'échantillons appartenant au phonème ϕ.

L'erreur quadratique de prédiction s'écrit :

$$L^{\phi} = \sum_{k=1}^{K}\left(y_k^{\phi} - F_{\Omega}\left(x_k^{\phi}\right)\right)^2 \qquad (20)$$

K étant le nombre de fenêtre de prédiction qui dépend de la longueur de la dimension de la zone stationnaire

Après minimisation de L^{ϕ} par l'algorithme de rétropropagation du gradient (erreur), F_{Ω} constitue une modélisation NLAR (Non Linear Auto-Regressive) du phonème ϕ et peut être considérée comme caractéristique de ce phonème. L'inconvénient d'une telle approche est, comme nous l'avons dit, le très grand nombre de paramètres générés.

Le codeur NPC permet de limiter arbitrairement ce nombre en spécialisant les connexions. Pour cela, nous décidons que les poids liant la fenêtre de prédiction à la première couche codent des informations non discriminantes, communes à tous les phonèmes, tandis que les poids liant la couche cachée à la couche de sortie (une cellule de prédiction) codent les informations discriminantes, propres à chaque phonème. Le vecteur de caractéristiques est donc constitué uniquement de ces derniers. Sur le plan formel, cela nous conduit à définir la fonction *F* réalisée par le réseau comme la composition de deux fonctions G_{Ω} et $H_{a\phi}$, l'une associée à la première couche, et l'autre à la seconde couche.

$$F_{\Omega} = H_{a^{\phi}} \circ G_{\Omega} \tag{21}$$

avec $\hat{y}_k^{\phi} = H_{a^{\phi}}\left(z_k^{\phi}\right)$
$z_k^{\phi} = G_{\Omega}\left(x_k^{\phi}\right)$

La répartition des informations discriminantes et non discriminantes sur les poids du réseau s'obtient en définissant deux fonctions de coût pour les deux ensembles de poids en question.

La première $L^{\phi}\left(a^{\phi}\right)$ est calculée à partir de l'erreur de prédiction moyennée sur les séquences composant un même phonème ϕ.

La deuxième $L^{\phi}\left(\Omega, a^{\phi}, ...\right)$ est calculée sur l'ensemble des séquences composant tous les phonèmes de la base de données.

4.3.4. Paramétrisation du codeur

La phase de paramétrisation est l'estimation de la fonction G_{Ω}. Elle s'obtient par minimisation du critère quadratique suivant :

$$L = \frac{1}{\left|\{\phi\}\right|}\sum_{\phi} L^{\phi}\left(a^{\phi}\right) = \frac{1}{\left|\{\phi\}\right|}\sum_{\phi=1}^{M}\sum_{k=1}^{K}\left(y_k^{\phi} - H_{a^{\phi}} \circ G_{\Omega}\left(x_k^{\phi}\right)\right)^2 \tag{22}$$

M étant le nombre de phonème de la classe

Lors de cette phase, les paramètres a^{ϕ} doivent être également estimés. En effet, le choix d'une valeur arbitraire et unique pour tous les phonèmes ($a^{\phi} = a\circ, \forall\phi$) conduirait à reporter l'information discriminante uniquement sur l'erreur de prédiction et ainsi à répartir l'information non discriminante sur l'ensemble des paramètres du réseau et non pas seulement

sur les poids [47]. L'estimation des coefficients a^{ϕ} s'obtient par minimisation de l'erreur de prédiction sur les séquences composant le phonème.

$$L^{\phi}\left(a^{\phi}\right)=\sum_{k=1}^{K}\left(y_k^{\phi}-H_{a^{\phi}}\circ G_{\Omega}\left(x_k^{\phi}\right)\right)^2 \qquad (23)$$

4.3.5. Phase de codage

La phase de codage est la phase de génération des codes. Pour un phonème quelconque ϕ et l'ensemble des séquences x_k^{ϕ} qui le composent, la première couche du réseau est utilisée comme un opérateur de changement de représentation : .

$$Z_k^{\phi}=G_{\Omega}(x_k^{\phi}) \qquad (24)$$

Le vecteur des poids étant celui obtenu par la phase de paramétrisation.

L'estimation du vecteur caractéristique a^{ϕ} s'obtient par minimisation de l'erreur de prédiction sur l'ensemble des vecteurs Z_k^{ϕ} calculés sur les séquences x_k^{ϕ} .

$$L^{\phi}=\sum_{k=1}^{K}\left(y_k^{\phi}-H_{a^{\phi}}\left(z_k^{\phi}\right)\right)^2 \qquad (25)$$

4.3.6. Codage discriminant

Une modification adéquate des fonctions de coût précédentes permet d'introduire très simplement des informations de classe d'appartenance lors de la paramétrisation du codeur. C'est ce que nous proposons d'établir dans la classification des phonèmes.

Reprenons la phase de paramétrisation. Nous ne considérons plus un vecteur de code par phonème mais un vecteur de code par classe de phonèmes. Tout se passe comme si nous devions coder les classes de phonèmes plutôt que les phonèmes eux-mêmes.

Soient $C_{1,}\ldots,C_M$ les M classes d'appartenance des phonèmes et a^{Ci} les vecteurs de code associés. Ces M vecteurs constituent les M jeux de poids de la deuxième couche du réseau qu'il convient d'estimer.

Le coût quadratique à minimiser devient naturellement (pour les poids de la deuxième couche) comme suit:

$$L^{C_i}\left(a^{C_i}\right)=\sum_{\phi\in C_i}^{M}\sum_{k=1}^{K}\left(y_k^{\phi}-H_{a^{C_i}}\circ G_{\Omega}\left(x_k^{\phi}\right)\right)^2 \qquad (26)$$

pour les paramètres de la première couche, nous avons l'équation 12 :

$$L = \frac{1}{M}\sum_{i=1}^{M} L^{C_i} \qquad (27)$$

L'étape de codage reste inchangée : nous reprenons à nouveau le codage des phonèmes les uns après les autres en minimisant la fonction de coût définie précédemment.

4.3.7. Objectif du Codage Neuro Prédictif

Dans les applications telles que la reconnaissance phonémique, le principal but de la programmation est d'extraire un maximum d'informations du signal, tout en minimisant la quantité de données. Le codeur NPC se place dans une chaîne de reconnaissance de la parole ou de phonèmes (figure. 4.11).

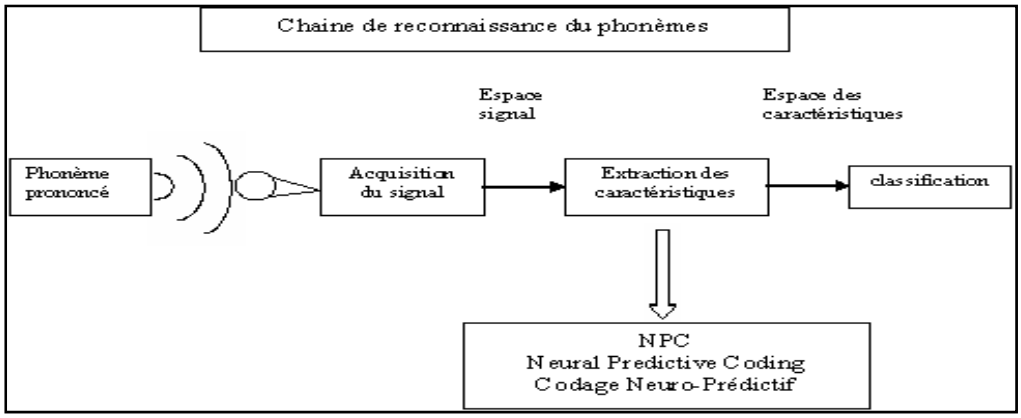


Figure 4.11 **:** Chaîne de reconnaissance des phonèmes

Nous avons présenté qualitativement le principe de fonctionnement d'une nouvelle méthode d'extraction de caractéristiques qui est le Codage Neuro-Prédictif (NPC).

Avant de passer à la classification généralement on introduit une autre étape de traitement à savoir la quantification vectorielle. Cette dernière consiste en une technique de quantification souvent utilisée dans la compression de données avec pertes de données (Lossy Data Compression) pour laquelle l'idée de base est de coder ou de remplacer par une clé des valeurs d'un espace vectoriel multidimensionnel vers des valeurs d'un sous-espace discret de plus petite dimension. Le vecteur du plus petit espace nécessite moins d'espace de stockage et les données sont donc compressées. La réduction vers un sous-espace est habituellement réalisée par une projection, ou en utilisant un dictionnaire (codebook). Dans certains cas, l'implémentation d'un **codebook** peut aussi bien servir au codage de l'entropie des valeurs discrètes qu'à la génération de valeurs codées à longueur variable et à code préfixe.

5. Techniques de Décision, Reconnaissance et Classification

Nous allons voir dans ce paragraphe les méthodes de reconnaissance de formes appliquees à la parole. Nous pouvons citer, L'alignement temporel (DTW) qui se base surtout sur la comparaison d'un signal reçu par rapport à un signal de référence, une autre qui s'articule surtout sur compression de la basse de données et cela afin de réduire le temps par apprentissage à l'utilisation en utilisant les modèle de Markov, et la troisième se basant sur les modèles connexionnistes.

5.1. Alignement temporel (Dynamic Time Warping DTW)

L'alignement temporel, plus connu sous l'acronyme de DTW, *Dynamic Time Warping*, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de références. Le signal en question est comparé avec chacune des références et est classé en fonction de sa proximité avec une des références stockées. La DTW est en fait une application dans le domaine de la reconnaissance de la parole [81] de la méthode plus générale de la programmation dynamique [82]. Elle peut ainsi être vue comme un problème de cheminement dans un graphe [83, 84]. Ce type de méthode pose deux problèmes :

- la taille de la base de références, qui doit être importante ;
- la fonction de calcul des distances, qui doit être choisie avec soin.

La taille de la base de données contenant les signaux de références est directement liée aux capacités, variables, de reconnaissance du système d'alignement temporel. Chacun de ces signaux est en effet stocké dans son état brut, sans aucune sorte compression. Ce stockage permet de disposer d'un vocabulaire dont la taille correspond au nombre de mots du vocabulaire multiplié par le nombre de locuteurs et le nombre des éventuelles répétitions des mots. Cette base de références permet d'effectuer une mise en correspondance entre le signal stocké, d'une part, et sa retranscription symbolique, d'autre part. La taille de cette base est importante et implique une charge de travail non négligeable, puisque la classification de chaque forme à analyser impose de la comparer à chaque forme de la base de références. Par conséquent, si la constitution de la base de référence est assez rapide et si le processus d'apprentissage est inexistant dans la méthode de l'alignement temporel, la phase d'utilisation nécessite une puissance de calcul non négligeable pour chaque référence du signal à analyser. (figure 4.12).

Comme le montre la figure 4.12, la forme choisie sera celle pour laquelle le chemin de mise en correspondance est le plus court. Cette taille minimale marquant le peu de différences entre la forme à analyser et celle de références.

L'autre partie importante de l'alignement temporel est la définition de la fonction de recalage qui permet de calculer, selon certaines contraintes, la distance entre la forme à comparer et la forme de référence. La forme à analyser est mise en correspondance dans le plan temporel par l'algorithme Forme 1 de la base Forme 2 de la base Base des formes de référence Forme à classer d'alignement qui essaie de trouver le plus court chemin dans le graphe ainsi constitué. Cette fonction de mise en correspondance définit une valeur pour chaque arc du graphe, ces valeurs favorisant l'axe médian qui correspond à une parfaite mise en relation de la forme à analyser et d'une forme de référence (figure 4.13).

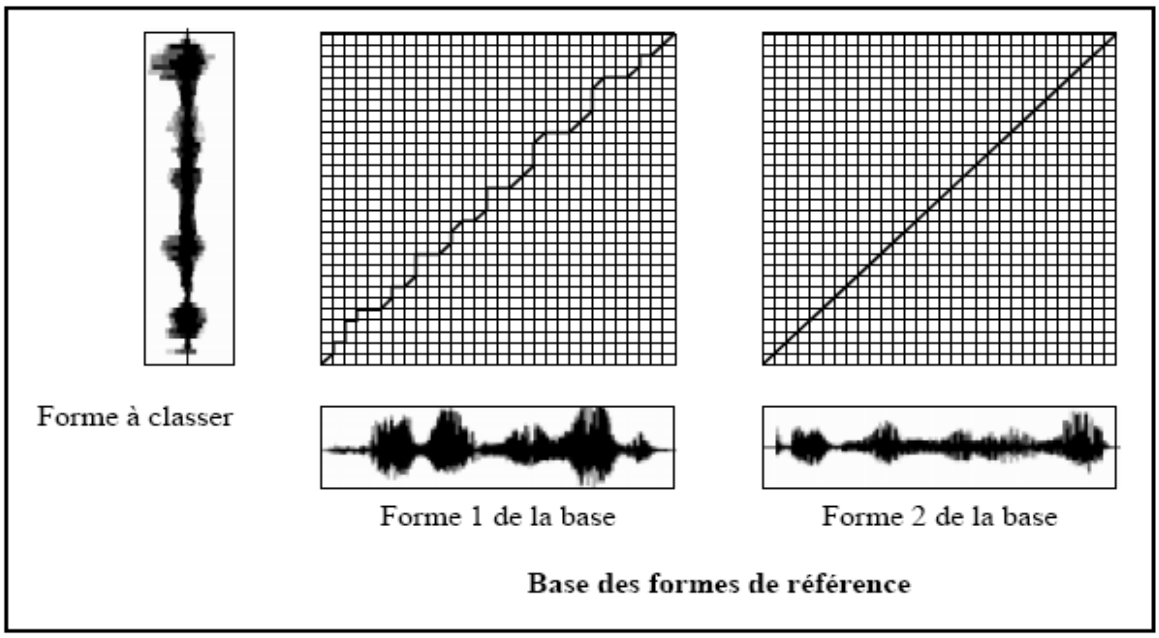


Figure 4.12 : Visualisation du cheminement de l'alignement temporel pour des formes de la base de références.

La fonction de recalage suit typiquement le schéma présenté dans la figure 4.13. La fonction *d(i,j)* est la fonction de calcul de la distance entre deux points successifs du graphe. Les valeurs α β et γ permettent de définir une partie du comportement de la fonction *d* qui peut être soit symétrique ($\alpha = \gamma$) soit asymétrique ($\alpha \neq \gamma$). Ce calcul de distance entre deux nœuds successifs du graphe n'est cependant pas suffisant pour calculer la longueur totale du chemin parcouru dans le graphe. Une fonction supplémentaire *G*, calcule une longueur totale qui permet, après le calcul de cette longueur des chemins sur toutes les formes de la base de référence, de savoir à quel mot du vocabulaire préenregistré correspond la forme à classer.

D'un point de vue mathématique, *M* et *N* étant les longueurs respectives de la forme à classer et de celle de référence.

Nous cherchons sur l'ensemble du corpus le *G(M,N)* minimal. Le calcul de cette fonction *G* répond au même principe que le principe général énoncé par Bellman pour la programmation dynamique [82] : toute sous-partie du chemin optimal est lui-même un chemin optimal. Des exemples de fonctions *d* et *G* de calcul de distance, qui peuvent être bien plus complexes que la fonction de recalage présentée en figure 4.13, pourront être trouvées dans les références [85, 86]. Les fonctions présentées peuvent analyser jusqu'à 9 chemins différents pour *d*, la fonction *G* étant de complexité égale à celle de *d*.

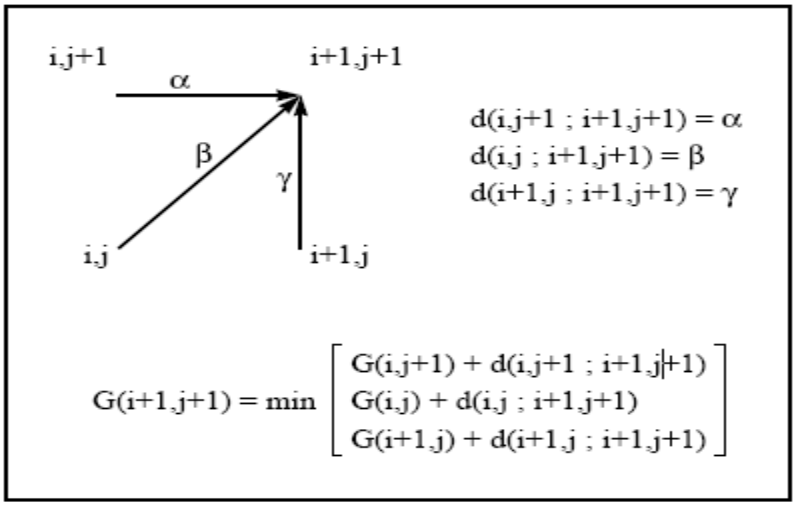


Figure 4.13 : Schéma typique d'une fonction de recalage en alignement temporel

Cette méthode de reconnaissance des formes est, initialement, bien adaptée à la reconnaissance de mots isolés mais des extensions ont été développées pour permettre de l'appliquer à la parole continue [87, 88, 89]. D'autres méthodes complémentaires ont par ailleurs été développées pour tenter de réduire la taille de la base des formes de références par sélection optimale des formes à conserver [90]. Ces méthodes reposent surtout sur une exploration statistique de la base des formes de références et permettent d'obtenir une caractérisation des différents ensembles la constituant. Ces ensembles correspondant aux différents symboles référencés dans la base. Une des techniques qu'il est possible d'employer pour ce faire est, par exemple, la méthode des plus proches voisins. Certaines méthodes permettent de réduire ce temps de calcul à l'utilisation par apprentissage a priori des coefficients qui permettent de compacter la connaissance présente dans la base de références qui devient ainsi un corpus d'apprentissage. Une première méthode mettant en œuvre ce principe de compactage de la connaissance est le modèle de Markov

5.2. Chaîne de Markov

Les chaînes de Markov trouvèrent leurs applications dans différents domaines de la physique, par exemple en expliquant les mouvements browniens, les radiations cosmiques, la radioactivité ainsi que la génétique, la parole, les fluctuations des stocks d'entreprise, les marches aléatoires, etc. (figure 4.14).

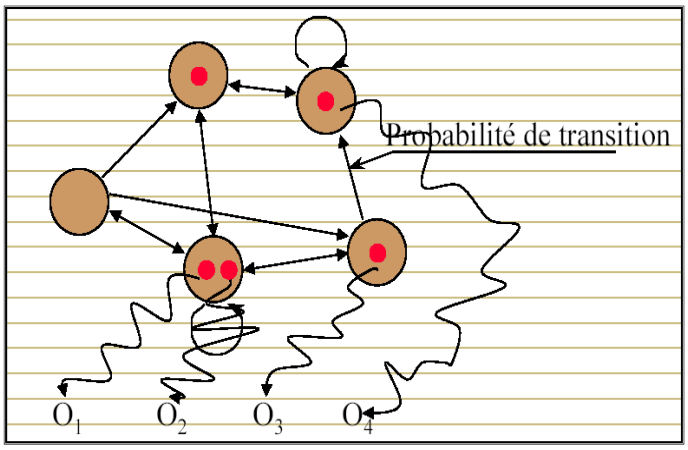


Figure 4.14 : Chaînes de Markov sous forme d'Automate probabiliste

Les chaînes de Markov sont considérées comme des exemples de processus stochastiques, incluant les processus de Markov, de Poisson, etc. [91].

L'étude des processus stochastiques a commencé au début du XX^{éme} siècle grâce à un mathématicien Russe, Markov Andreï Andreïevitch. Son étude statistique du langage l'a conduit à formuler l'hypothèse Markovienne, qui peut se résumer ainsi:
« L'évolution future d'un système ne dépend que de son état présent ».
Autrement dit, cette hypothèse implique que l'état courant du système contient toute l'information apportée par le passé. C'est donc une hypothèse très forte, mais qui semble relativement logique. En pratique, on constate que de nombreux systèmes enfreignent cette condition. Cependant, en affinant le modèle, on peut souvent le rendre markovien.

Les modèles de Markov et, plus particulièrement, les modèles de Markov à états cachés, plus connus sous le nom de HMM (Hidden Markov Models), permettent de synthétiser la connaissance contenue dans un corpus par apprentissage. Cette connaissance sera synthétisée, dans les modèles de Markov, par une représentation probabiliste au sein de plusieurs graphes,

Chaque graphe correspondant à une classe du corpus d'apprentissage qui, en RAP, peut correspondre à un phonème ou à un mot [3].

La Comparaison de la DTW avec les HMM montre que d'un point de vue général, la différence de fonctionnement entre la méthode de l'alignement temporel et les modèles de Markov n'est pas fondamentale. Dans le cas des modèles de Markov, la forme à analyser est comparée à chacune des classes constituées en graphe. Dans le cas de la DTW, par contre, la forme à analyser est comparée à chacune des formes de référence dont le rattachement à une classe, et donc à une signification symbolique, ne sera utile qu'à un stade postérieur du traitement. Le modèle de Markov étant une représentation probabiliste des formes de référence, il ne s'agit plus ici de trouver un chemin de taille minimale mais de trouver une probabilité de cheminement dont la valeur est maximale [92].

5.2.1. Approche mathématique

Un modèle de Markov caché est défini par :

- son graphe d'état ;
- sa Matrice de Transition ;
- sa Matrice d'observation ou d'émission ;
- son Vecteur d'initialisation.

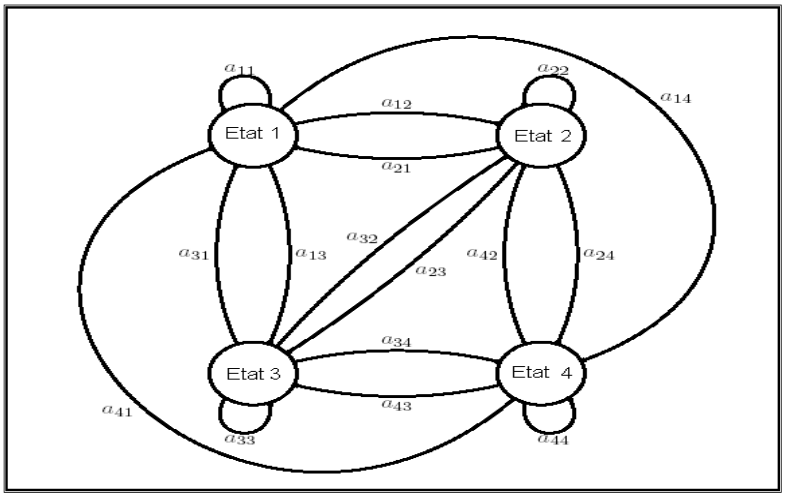


Figure 4.15. Graphe d'états d'une chaîne de Markov

La Matrice de Transition

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}_{4\times 4}$$

Dont ces éléments sont définis tels que :

$$a_{ij} = p(q_{t+1} = j / q_t = i) \qquad 1 \le i, j < N \tag{28}$$

La matrice d'observation ou d'émission, dont sa dimension dépend des observations continues ou discrètes. Ses éléments sont définis telle qu'à chaque état, une ou plusieurs observations sont associées à cet état, définis par :

$b(o_j / q_i)$: Représente la probabilité d'observer ou d'émettre l'observation j en étant à l'état q_i.

♦ **Cas discret :** Les matrices contiennent des vecteurs d'observations de dimension finie de N x M.

♦ **Cas continu :** Les matrices contiennent les paramètres des probabilités de distribution des observations par état. Celles-ci peuvent être des gaussiennes, des multi gaussiennes ou toute autre distribution propre au contexte des données.

$$\mathrm{B} = \{\ b_j(o_t)\ \}_{j=1}^{N}, \text{ tel que : } b_j(o_t) = p(o_t / q_t = j) \tag{29}$$

Le Vecteur d'initialisation donne au modèle les probabilités de transition initiale

π : vecteur de dimension N x 1, tel que : $\pi_i = p(q_1 = i)$

En résumé, le modèle des chaînes de Markov est noté comme suit :

$\lambda = (A, B, \pi)$ Avec A : N x N ; B : N x M et π : N x 1

5.2.2. Les trois problèmes fondamentaux des HMM

Différentes formulations existent en littérature, dont les principales sont [93] :

♦ Evaluation : Sachant ou ayant des vecteurs d'observation O $=\{O_1,O_2,...,O_T\}$ et $\lambda=\{A,B,\pi\}$ comment évaluer $P(O|\lambda)$? Comment trouver le modèle qui a pu générer la séquence observée ?

♦ Retirer le H de Hidden Markov Models : Ayant les vecteurs d'observation $O=\{O_1,O_2,...,O_T\}$ ainsi que les paramètres du $\lambda = \{A, B, \pi\}$ du modèle, comment trouver la séquence (cachée) optimale d'états qui explique au mieux ces observations ?

♦ Apprentissage :Sachant un corpus d'entraînement O, comment ajuster les paramètres λ du modèle pour maximiser $p(O|\lambda)$?

Prenons un exemple de m modèles, (λ_i, $\forall i \in [1, m]$) qui modélisent chacun une entité donnée (un mot ou un phonème, par exemple..), Soit O une observation dont on veut connaître l'identité $\boldsymbol{i} = \text{argMax}_i(p(O/\lambda_i))$

Soit $Q=q_1,q_2,q_3 \ldots q_T$ une séquence d'états du modèle de Markov caché pouvant « expliquer » O :

$$P(O/\lambda)= \sum_{Tous\ les\ Q} p(O,Q/\lambda) = \sum_{Tous\ les\ Q} p(O/Q,\lambda.\pi).P(Q/\lambda) \quad (30)$$

Avec :

$$p(O|Q,\lambda) = \prod_{t=1}^{T} p(o_t/q_t,\lambda)=b_{q1}(o_1)x\ b_{q2}(o_2)\ldots.\ b_{qT}(o_T) \quad (31)$$

et

$$p(Q|\lambda) =\pi_{q1}.b_{q1}(o_1)x\ b_{q2}(o_2)\ldots.\ b_{qT}(o_T) \quad (32)$$

Le problème relié à la complexité de calcul qui est de l'orddre de $(2*T-1).N^T$ multiplications et (N^T-1) additions, Donc, si on prend N=5 états, T=100 observations, on a $1.5698x10^{72}$ Multiplications et $7.8886x10^{69}$ additions.

Pour définir complètement les HMM, il y'a lieu de résoudre les 3 problèmes cités :

♦ **Prob. 1** : Etant donné une séquence d'observations, par exemple, une séquence de vecteurs acoustiques, et un modèle défini par le triplet (A, B, π): Quelle est la plus grande probabilité que le modèle donné génère la séquence observée. Ceci est résolu par l'algorithme Forward.

♦ **Prob. 2** : Etant donné une séquence d'observation et un modèle HMM, quelle est la séquence d'états qui a générée cette séquence observée. Ceci est résolu par l'algorithme de Viterbi.

♦ **Prob. 3** : Etant donné une séquence d'observations, quels sont les ajustements à faire sur les paramètres du modèle pour avoir la plus grande probabilité de générer cette séquence. C'est la phase d'apprentissage, ceci est résolu par l'algorithme backward-forward.

Il faut également définir une topologie qui va déterminer les transitions possibles entre les différents états. D'après les études déjà effectuées, un modèle de 5 états entièrement connecté est choisi pour la reconnaissance de mots, toutefois pour des phonèmes des modèles à 3 états seront suffisants, une topologie de type gauche droite, puisque que chaque état représente un état au sein d'un phonème. Chaque mot impose ainsi une suite de phonèmes déterminée [3].
La langue Arabe compte 40 phonèmes, chacun d'entre eux, peut avoir son propre modèle de Markov.

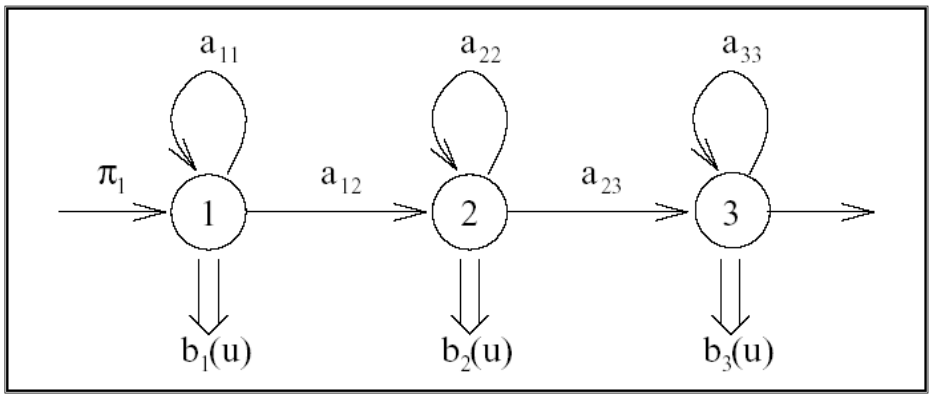


Figure 4.16 : Modèle de transitions à trois états Markoviens gauche-droite.

Remarque : Les modèles de Markov permettent de suivre une évolution temporelle globale en même temps qu'ils fournissent une évaluation locale. La figure 4.17 illustre ce principe.

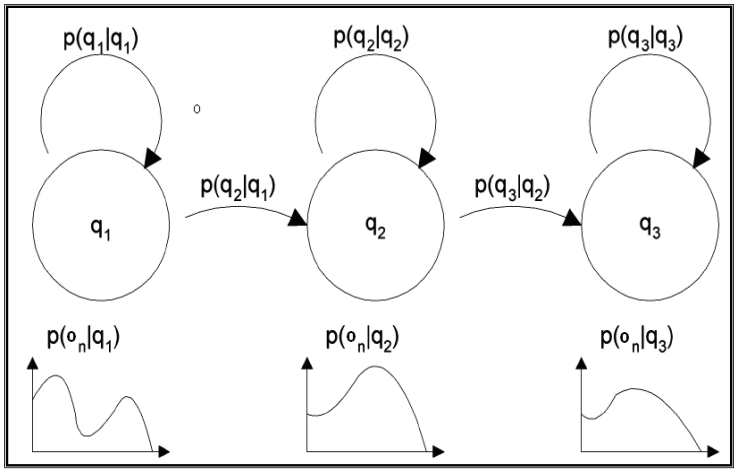


Figure 4.17 : Modèle de transitions entre états markoviens gauche-droite et leur probabilité de transition et d'émission respectives.

Notre approche s'adapte bien aux modèles de Markov. Notamment, les modèles qui sont conçus pour reconnaître n'importe quelle phrase ou séquence de mots lors du traitement de la

parole continue. Malheureusement, il n'est pas possible de construire un modèle pour chaque mot du dictionnaire, qui parfois atteignent des centaines de milliers, toutefois, dans notre contexte de pathologie langagière, nous avons ciblé un apprentissage supervisé qui compte des mots sélectionnés définissant la mauvaise prononciation. Nous modélisons chaque mot par une suite d'états.

Remarque : Les probabilités des matrices de transition et d'observation sont indépendantes du temps.

* Comment résoudre les trois problèmes ?

* Solution au prob. 1 :

a. Evaluation par l'algorithme « Forward »

Hypothèse, nous disposons de k modèles de Markov $\{\lambda_1, \lambda_2 ... \lambda_k\}$, représentant les mots ou les phonèmes à analyser ayant $\{q_1, q_2 ... q_L\}$ états possibles, et d'un ensemble d'observations $O = \{o_1, o_2, ... o_N\}$ [94].

Trouvons le modèle qui a généré cette séquence, tel que : $P(O/\lambda)$ est maximale.

« Une façon de calculer cette probabilité est d'énumérer tous les chemins du modèle :

$$P(O/\lambda) = \sum_{Q \in Modéle} p(Q, O/\lambda) \tag{33}$$

Où la somme porte sur tous les chemins Q de longueur L, dans le modèle, la complexité est toutefois considérable, alors une récurrence avant est utilisée de la manière suivante :

$$P(O/\lambda) = \sum_{l=1}^{L} p(q_l^n, O/\lambda), \forall n \in [1, N] \tag{34}$$

Chaque terme de cette somme exprime la probabilité que X soit émis par le modèle λ en passant par l'état q_l à l'instant n, qui peut se factoriser comme suit :

$$p(q_l^n, O/\lambda) = p(q_l^n, O_1^n/\lambda).p(O_{n+1}^N / q_l^n, O_1^n, \lambda) \tag{35}$$

Où O_1^n représente une séquence partielle de vecteurs d'observation $\{o_1, o_2, ... o_n\}$

Définissons une nouvelle variable, représentant la probabilité que le modèle λ ait généré la séquence partielle O_1^n, en se trouvant dans l'état q_l à l'instant n.

$$\alpha_n(l/\lambda) = p(q_l^n, O_1^n/\lambda) \tag{36}$$

$$\alpha_n(l/\lambda) = p(q_l^n, O_1^n/\lambda) = \sum_{k=1}^{L} p(q_l^{n-1}, q_l^n, O_1^{n-1}, o_n/\lambda) \tag{37}$$

$$\alpha_n(l/\lambda) = \sum_{k=1}^{L} p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda).p(q_k^{n-1}, O_1^{n-1} / \lambda) \quad (38)$$

$$\alpha_n(l/\lambda) = \sum_{k=1}^{L} p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda).\alpha_{n-1}(k/\lambda) \quad (39)$$

L'initialisation de cette forme récurrente se fait par :

$$\alpha_n(l/\lambda) = \pi\ (l) \quad (40)$$

Toutefois le terme $p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda)$ n'est pas facilement calculable, il faudrait le simplifier :

$$p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda) = p(q_l^n / q_k^{n-1}, O_1^{n-1}, \lambda).p(o_n / q_l^n, q_k^{n-1}, O_1^{n-1}, \lambda) \quad (41)$$

Les modèles de Markov sont d'ordre 1, donc les états ne dépendent que de l'état précédent la transition et sont conditionnellement indépendants du passé, alors

$$p(q_l^n / q_k^{n-1}, O_1^{n-1}, \lambda) \quad \rightarrow \quad p(q_l^n / q_k^{n-1}, \lambda) \quad (42)$$

Les observations sont conditionnellement indépendantes du passé, elles ne dépendent ni des observations du passé ni des états HMM précédents, alors

$$p(o_n / q_l^n, q_k^{n-1}, O_1^{n-1}, \lambda) \rightarrow \quad p(o_n / q_l^n, \lambda) \quad (43)$$

Alors

$$p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda) = p(q_l^n / q_k^{n-1}, \lambda).p(o_n / q_l^n, \lambda) \quad (44)$$

Cette équation définit, les probabilités d'émission $p(o_n / q_l^n, \lambda)$, celles-ci seront considérées dans notre cas multi gaussienne.

Et $p(q_l^n / q_k^{n-1}, \lambda)$ les probabilités de transition à l'intérieur du modèle.

L'équation $\alpha_n(l/\lambda) = \sum_{k=1}^{L} p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda).\alpha_{n-1}(k/\lambda)$ est réécrite de la manière suivante :

$$\alpha_n(l/\lambda) = p(o_n, q_l) \sum_{k=1}^{L} .\alpha_{n-1}(k/\lambda).p(q_l^n / q_k^{n-1}, \lambda) \quad (45)$$

En résumé cette récurrence nous permet de calculer la probabilité de toute une séquence à partir des probabilités d'émission et de transition locales.

*b. Evaluation par l'algorithme « Backward »

Une nouvelle variable est définie telle que, c'est la probabilité que le modèle λ génère le restant de la séquence $O = \{o_{n+1}, o_2, ...o_N\}$ au départ de l'état q_l.

Par la même méthode de calcul de l'algorithme forward,

$$\beta_n(l/\lambda) = p(O_{n+1}^N / q_l^n, O_1^n, \lambda) = \sum_k \beta_{n+1}(q_l^{n+1} / q_l^n, \lambda).p(o_{n+1} / q_k) \qquad (46)$$

Ou la somme sur k porte sur les successeurs possibles de q_l, l'initialisation de cette nouvelle récurrence est donnée par :

$$\beta_n(l/\lambda) = \pi_{lF}(\lambda) \qquad (47)$$

Qui représente la probabilité de rejoindre l'état final de λ à partir de q_l

Etant donné (19) et la définition de α, nous pouvons écrire :

$$P(O/\lambda) = \sum_{l=1}^{T} p(q_l^N, O_1^N / \lambda) = \sum_{l=1}^{L} \alpha_N(l/\lambda) \qquad (48)$$

La somme se limite aux états finaux possibles dans le modèle λ, et pouvant correspondre à la fin du mot ou du phonème, pour l'estimation de $P(O/\lambda)$, et donc aussi de la reconnaissance, nous avons donc :

$$P(O/\lambda) = \sum_{l=1}^{L}\sum_{n=1}^{T} \alpha_n(l/\lambda)\beta_n(l/\lambda) = \sum_{\{Finaux\}} \alpha_N(I/\lambda) = \sum_{\{Initiaux\}} \beta_N(F/\lambda) \qquad (49)$$

* Solution au problème 2 : Meilleur chemin par l'algorithme de Viterbi

Le modèle étant connu, quelle est le meilleur chemin, qui a donné la meilleure vraisemblance, donc quels sont les états qui ont réellement participé au calcul de la probabilité $P(O/\lambda)$?

Nous disposons du modèle défini par $\lambda = \{A, B, \pi\}$ et des observations $O = \{o_1, o_2, ...o_N\}$, trouvons le meilleur chemin maximisant $P(O/\lambda)$.

D'après Bellman, « Toute politique optimale est issue de sous politiques optimales »

L'approche Viterbi est plus simple à calculer, car au lieu de faire la somme sur tous les chemins, le chemin ayant une probabilité maximale est pris, donc l'équation $p(q_l^n, O/\lambda) = p(q_l^n, O_1^n / \lambda).p(O_{n+1}^N / q_l^n, O_1^n, \lambda)$, peut être réécrite comme suit :

$$\bar{p}(q_l^n, O_1^n / \lambda) = \max_k \bar{p}(q_k^{n-1}, O_1^{n-1} / \lambda).p(q_l^n, o_n / q_k^{n-1}, O_1^{n-1}, \lambda) \qquad (50)$$

Où : $\bar{p}(q_k^{n-1}, O_1^{n-1})$ représente la probabilité du meilleur chemin possible allant de l'état initial $q_{Initial}$ à l'état q_l, en ayant émis les n premiers vecteurs O_1^n de la séquence *O*.

Utilisant l'indépendance et la supposition de Markov d'ordre 1, (36) devient

$$\bar{p}(q_l^n, O_1^n / \lambda) = \max_k \bar{p}(q_k^{n-1}, O_1^{n-1} / \lambda).p(q_l^n / q_k^{n-1}, \lambda).p(o_n / q_l) \qquad (51)$$

En passant au logarithme, afin d'éviter l'underflow,

$$-\log\left[\bar{p}(q_l^n, O_1^n / \lambda)\right] = \min_k \left\{-\log\left[\bar{p}(q_k^{n-1}, O_1^{n-1} / \lambda)\right] - \log\left[p(q_l^n / q_k^{n-1}, \lambda)\right] - \log p(o_n / q_l)\right\}$$

La probabilité

$$\bar{p}(O / \lambda) = \bar{p}(q_F^N, O_1^N / \lambda) \qquad (52)$$

Où q_F^N est l'état final du modèle λ. [92]

C'est ce point qui fait que les HMM ressemblent à la DTW, il y'a dualité entre les logarithmes et les distances euclidiennes.

* Solution au Problème. 3 : Ajuster le modèle

Ayant un corpus d'entraînement O, comment ajuster les paramètres λ du modèle pour maximiser P(O/λ)? [91]. Ce stade, en général est le premier à réaliser car nous ne disposons que des vecteurs d'observation, et nous voudrions chercher un modèle qui pourrait « s'ajuster » à ces données, pour les modéliser avec la meilleure vraisemblance pour une future reconnaissance. Le choix du modèle HMM, du nombre de gaussiennes est déterminant, (tableau 4.4), il dépend de plusieurs facteurs, entre autres du :

- ♦ modèle du HMM : droite - gauche pour la parole [3] ;
- ♦ nombre de gaussiennes modélisant les observations par état.

Tableau 4.4 : Choix des paramètres des HMM/GMM

Etudes réalisées	**Nombre d'états par HMM**	**Nombre de Gaussiennes**
Lawrence Rabiner, [3]	Pour chaque mot : Mono locuteur : 5 à 8 états Multi locuteurs 8 à 10	 3 à 5 par état 9 par état
Projet Raphaël, [22] Traduction multi langue	3 états par phonème	16 gaussiennes par état
M. A. Mokhtar et al, [77]	Language arabe complet : 41 états	
A. R. Elobeid Ahmed et al, [54]	3 états par Phonème	Pas de Gaussienne mais un Dictionnaire de 128 valeurs
A.A. Dibazar et al, [58]	3 états pour le phonème [a]	3 gaussiennes par état

5.2.3. Modélisation des données par mélange de Gaussiennes GMM

Avant d'aborder l'apprentissage, quelques définitions s'avèrent utiles, dans le cas de notre étude, l'utilisation d'un nombre très important de données nous incite à modéliser leurs distributions, l'une des techniques de modélisation est l'approche multi gaussiennes, ou l'ensemble des données est écrit en fonction de diverses gaussiennes de la manière suivante :

$$p(O/\Theta) = \sum_{i=1}^{T} c_i . p_i(O/\theta_i) \quad (53)$$

$$\Theta = \{c_1, c_2, ..., c_T, \theta_1, \theta_2, ..., \theta_T\} \quad (54)$$

$$\theta_i = \{\mu_i, \sigma_i\} \quad (55)$$

Avec θ_i Représentant chaque gaussienne

On ne peut citer tous les travaux relatifs au choix des paramètres, dans le tableau 4.5, toutefois, nous présentons une comparaison entre différentes études concernant le nombre d'états par modèle et le nombre de gaussienne par état.

Tableau 4.5 : Choix du nombre de gaussiennes

Nombre d'états par mot	Nombre de gaussiennes par état
A chaque phonème du mot un état	4, 8, 16
A chaque phonème du mot 3 états	4, 8, 16
Prendre le mot le plus long comme référence	4, 8, 16

5.2.4. Apprentissage par l'Algorithme de Baum Welch «Variante de l'algorithme EM »

L'idée générale derrière cet algorithme est d'estimer les paramètres du modèle HMM, ainsi que les paramètres des mélanges de gaussiennes, ayant en main deux informations essentielles:

- les vecteurs d'observation en nombre suffisant ;
- le nombre d'états gouvernant les transitions à trouver ;
- la topologie des transitions à utiliser. (gauche - droite; Ergodique).

L'algorithme EM (Expectation – Maximisation) Maximiser l'espérance,

- calcule l'espérance d'une variable aléatoire manquante par rapport à une variable aléatoire présente ;

- maximise l'espérance trouvée en fonction des variables présentes, par une méthode récursive, jusqu'à ajuster les paramètres des chaînes de Markov notamment les probabilités de transition ainsi que les paramètres des mélanges de gaussiennes.

L'algorithme EM n'est pas démontré mathématiquement dans notre étude, toutefois pour tout détail complémentaire, se référer à [93].

Soit $\{O\}$ l'ensemble des variables aléatoires connues et $\{Y\}$ les variables aléatoires inconnues, nous supposons qu'il existe une densité de probabilité jointe $z=(O,Y)$ telle que :

$$p(z/\lambda)=p(Y/O,\theta).p(X/\lambda) \tag{56}$$

Définissons une nouvelle quantité *Q*, représentant l'espérance jointe z,

*a. Etape E de l'algorithme

$$Q(\lambda,\lambda^{t-1})=E\left[\log\ p(O,Y/\lambda)/Y,\lambda^{t-1}\right] \tag{57}$$

Où λ^{t-1} représente le modèle utilisé à l'itération *t-1* pour calculer λ à l'itération *t*.

*b. Etape M de l'algorithme

Cette valeur est alors maximisée selon λ,

Donc l'algorithme calcule le modèle

$$\lambda^{t}=\arg\max_{\lambda} Q(\lambda,\lambda^{t-1}) \tag{58}$$

Donc à chaque itération, on cherchera si le nouveau modèle apporte une amélioration à l'ajustement des données, c'est-à-dire est ce que le modèle représente les données à l'étape t mieux qu'à l'étape t-1.

Dans notre cas, nous allons définir deux nouvelles valeurs qui serviront lors de l'ajustement du modèle, telle que :

$$\gamma_i(n)=p(q_i^n/O,\lambda) \tag{59}$$

Qui représente la probabilité d'être à l'état q_i à l'instant *n*, générant la séquence *O*.

$$\gamma_i(n)=p(q_i^n/O,\lambda)=\frac{p(q_i^n,O/\lambda)}{p(O/\lambda)}=\frac{p(q_i^n,O/\lambda)}{\sum_{j=1}^{L}p(O,q_j^n/\lambda)} \tag{60}$$

Remarquons que :

$$\alpha_i(n)\beta_i(n)=p(o_1,o2,...,o_t,q_i^n/\lambda).p(o_{t+1},o_{t+2},o...,o_T/q_i^n,\lambda)=p(O,q_i^n/\lambda) \tag{61}$$

Alors (46) devient :

$$\gamma_i(n) = p(q_i^n / O, \lambda) = \frac{\alpha_i(n).\beta_i(n)}{\sum_{j=1}^{L} \alpha_j(n).\beta_j(n)} \quad (62)$$

On définit une seconde valeur telle que :

$$\xi_{i\ j}(n) = p(q_i^n, q_j^{n+1} / O, \lambda) \quad (63)$$

Qui représente la probabilité d'être à l'état i à l'instant n et de passer à l'état j à l'instant *n+1*, ceci peut être reformulé comme suit :

$$\xi_{i\ j}(n) = p(q_i^n, q_j^{n+1} / O, \lambda) \quad (64)$$

$$\xi_{i\ j}(n) = p(q_i^n, q_j^{n+1} / O, \lambda) = \frac{p(q_i^n, q_j^{n+1}, O / \lambda)}{p(O / \lambda)} = \frac{\alpha_i(n).a_{ij}.b_j(o_{n+1}).\beta_i(n)}{\sum_{j=1}^{L} p(O, q_j^n / \lambda)} \quad (65)$$

L'on peut remarquer que le terme $\sum_{n=1}^{T} \gamma_i(n)$ représente la valeur espérée d'être à l'état q_i pendant tous les instants *n* pour toutes les observations *O* donnant ainsi le nombre de transitions partant de l'état q_i .

Et que le terme $\sum_{n=1}^{T-1} \xi_{i\ j}(n)$ représente le nombre de transitions de l'état q_i à l état q_j pour toutes les observations *O*.

L'utilisation de l'algorithme EM pour estimer les nouveaux paramètres à chaque itération nécessite de mettre à jour les valeurs manquantes itérativement de la manière suivante :

$$\tilde{\pi}_i = \gamma_i(1) \quad (66)$$

Qui est la fréquence relative de passage à l'état q_i à l'instant 1.

$$\tilde{a}_{i\ j} = \frac{\sum_{n=1}^{N-1} \xi_{i\ j}(n)}{\sum_{n=1}^{N} \gamma_i(n)} \quad (67)$$

Qui représente le nombre de transitions de l'état q_i à l'état q_j relatif au nombre de transitions sortant de l'état q_i .

Pour le mélange de gaussiennes, les paramètres à estimer sont les moyennes et variances des gaussiennes ainsi que le taux de participation de la gaussienne à l'état q_i noté :

$$\widetilde{c}_{i\,l} = \frac{\sum_{n=1}^{N} \gamma_{i\,l}(n)}{\sum_{n=1}^{N} \gamma_{i}(n)} \tag{68}$$

Où l représente la $l^{\text{'iéme}}$ gaussienne modélisant les vecteurs d'observation à un état q_i .

$$\mu_{i\,l} = \frac{\sum_{n=1}^{N} \gamma_{i\,l}(n).o_t}{\sum_{n=1}^{N} \gamma_{i\,l}(n)} \tag{69}$$

L'équation (69) représente la moyenne de chaque gaussienne à l'état q_i

Et le terme

$$\sum_{il} = \frac{\sum_{n=1}^{N} \gamma_{i\,l}(n).(o_t - \mu_{il})(o_t - \mu_{il})^T}{\sum_{n=1}^{N} \gamma_{i\,l}(n)} \tag{70}$$

Représente la variance de chaque gaussienne à l'état q_i [66].

Cependant cette technique vu le taux de reconnaissance très appréciable, elle représente une problématique reposant sur le fait que chaque classe est représentée par un réseau de Markov, pour remédier à cela une technique de classification de données a été mise en œuvre ces dernières années de manière intensive dans de nombreux domaines de recherche [95, 96]. Cette méthode permet d'effectuer un pas supplémentaire et important sur la voie de la synthèse des connaissances grâce à l'utilisation d'un seul réseau pour représenter l'ensemble des classes présentes dans le corpus de références. Cette méthode, le connexionnisme, s'inspire très largement d'une modélisation assez fine du cerveau humain et se veut donc être une méthode neurobiologique plausible basée sur les réseaux de neurones. Cette technique sera utilisée pour réaliser une classification de phonèmes appliquée dans notre contexte à savoir les phonèmes pathologiques.

5.3. Réseaux de Neurones Artificiels (ANN)

Les Réseaux de Neurones Artificiels sont basés sur un modèle simplifié du fonctionnement du cerveau humain qui est constitué d'un ensemble de neurones et de synapses. L'information y est codée sous la forme d'un potentiel électrique. Chaque sortie de neurones correspond à une somme pondérée, grâce aux synapses, des autres sorties des neurones. Le neurone est considéré comme actif si son potentiel dépasse un seuil donné. Ceci n'est qu'un rapide aperçu

de l'origine de ces modèles, mais la littérature spécialisée la détaille abondamment [44]. Le but d'un tel réseau est donc d'apprendre à produire des sorties particulières en fonction des entrées, données. C'est au cours d'une phase d'apprentissage que le réseau va modifier ces poids des connexions, comme le fait le cerveau en modifiant les connexions synaptiques

***Perceptron multicouches**

L'unité élémentaire de ces réseaux est appelée perceptron. Il calcule la somme pondérée de ces entrées et d'un biais. Le résultat est finalement l'entrée d'une fonction d'activation non linéaire. Cette fonction non linéaire a pour but de générer des combinaisons non linéaires d'ordre élevé. La figure 4.18 schématise cette unité.

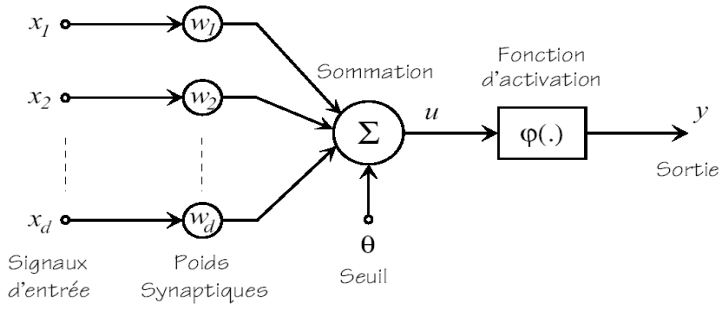


Figure 4.18 : Perceptron.

Le réseau de neurones est un rassemblement de plusieurs perceptrons en couches. L'utilisation de plusieurs couches génère un perceptron multicouche (MLP). L'architecture d'un tel réseau est présentée à la figure 4.19. L'ensemble des sorties de la première couche constitue les entrées de la couche suivante et ainsi de suite.

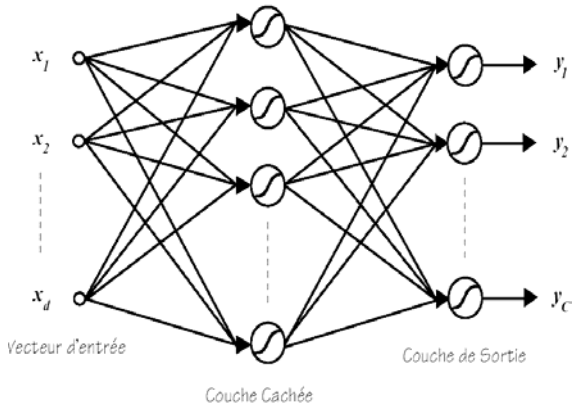


Figure 4.19 : Perceptron multicouches (MLP)

***Critère d'entraînement**

Le problème consiste maintenant à ajuster les poids du réseau selon un critère de minimisation de fonction de coût qui garantit un taux d'erreur minimum. Un des critères les plus courants est celui des moindres carrés

$$E = \sum_{n=1}^{N} \sum_{k=1}^{K} [g_k(x_n), \theta - d_k(x_n)]^2 \qquad (71)$$

Où

1. x_n (1, ..., N) représente les vecteurs présentés à l'entrée du réseau lors de l'apprentissage.
2. g_k (x_n,Θ) est la valeur observée à la sortie k du réseau étant donné le vecteur x_n à l'entrée et l'ensemble des paramètres Θ.
3. d_k (x_n) la valeur de sortie désirée pour la k éme sortie, étant donné x_n en entrée. Comme l'entraînement est de type supervisé, cette sortie désirée est supposée connue. Dans le cas de la classification, elle vaut : d_l (x_n) = δ_{kl} si x_n appartient à la classe q_k.

L'entraînement vise donc à trouver l'ensemble des paramètres Θ optimaux qui annule la dérivée de l'équation (71). Puisque la fonction g_k (x_n, Θ) est non-linéaire, l'utilisation d'une méthode itérative de descente de gradient est nécessaire. C'est l'algorithme de retro propagation qui permet l'ajustement des poids. Celui-ci calcule l'erreur sur chaque sortie et la propage depuis la couche de sortie vers la couche d'entrée sur les poids avec un certain taux d'apprentissage. Lorsque le réseau est utilisé comme classificateur, les sorties désirées pendant l'apprentissage sont :

– 1 pour le neurone correspondant à la classe du vecteur présente en entrée ;

– 0 pour tous les autres neurones.

De par le critère utilisé, l'algorithme tente de faire approcher les sorties des neurones de ces valeurs. L'apprentissage augmente la valeur de sortie correspondant à la classe correcte, tout en diminuant les valeurs des autres classes incorrectes. L'apprentissage est donc discriminant. De plus, toujours dans le cadre de la classification, il a été démontré que si un perceptron multicouche à une seule couche cachée, composée de neurones à fonction d'activation continue non linéaire, est entraîné sous les conditions suivantes [44]:

- le réseau comporte un neurone de sortie par classe, et est entraîné à produire 1 sur la sortie associée à la classe de l'objet présenté, et 0 pour les toutes les autres sorties ;
- le critère d'optimisation est celui des moindres carrés de l'erreur ;

- le nombre d'unités cachées est suffisamment grand ;
- l'apprentissage ne converge pas vers un minimum local ; alors, les sorties du perceptron multicouches peuvent être interprétées comme des approximations, au sens des moindres carrés, des probabilités a posteriori des classes.

***Méthode d'apprentissage**

Il existe plusieurs manières pour adapter les poids au cours de l'entraînement. Nous utiliserons la méthode reconnue pour être optimale dans le cas de problèmes de classification, c'est-à-dire l'apprentissage en ligne. Cette méthode consiste à modifier les valeurs des poids après présentation de chaque vecteur. Les données doivent être présentées de manière aléatoire afin d'éviter de tomber dans un minimum local. Cette méthode permet de converger rapidement vers un optimum.

6. Conclusion

Dans ce chapitre nous avons présenté les principales techniques d'extraction des caractéristiques du signal vocal, la LPC, la PLP, la RASTA-PLP, la NPC, les MFCC ainsi que les techniques de reconnaissance de formes qui sont utilisées en Reconnaissance Automatique de la Parole (RAP) : l'alignement temporel, les Chaînes de Markov et les modèles connexionnistes. La présentation des modèles connexionnistes étaient précédée d'une brève présentation des connaissances de la neurobiologie qui ont servi de base à établissement des techniques neuromimétiques. Nous avons aussi présenté les principaux détails de chaque technique utile en RAP en citant les algorithmes utilisés, ceci étant un préambule sur le background mathématique nécessaire pour la réalisation de notre travail que nous nous sommes fixés, à savoir la réalisation d'un système d'aide au orthophoniste où les malades eux- mêmes représentant des pathologies de la parole.

Chapitre 5

Analyses des corpus et évaluation des Résultats obtenues

1. Introduction

Le traitement du signal vocal est un processus très complexe en termes d'extraction de l'information utile et sa reconnaissance, car il concerne l'une des problématiques les plus influentes dans cette phase. Le corpus sur lequel s'effectuent les tests de performance est fondamental étant donnée que l'extraction des paramètres acoustiques est sujette aux différentes influences qui sont dues à l'enregistrement, au réglage du niveau sonore du microphone, au bruit environnant, les défauts de prononciation non pathologiques, la prononciation incorrecte d'un mot avec ou sans gémination et lors de la segmentation manuelle, etc.

Tous ces défauts influent directement, d'une façon décisive, sur la phase de reconnaissance. L'absence d'une grande base de données sonore Arabe nous a conduit à faire les tests de reconnaissance sur une base préenregistrée ainsi que sur des phonèmes présentant certaines caractéristique proches des sons de l'Arabe de la base TIMIT, afin de valider nos résultats en vue d'une segmentation correcte des phonèmes.

2. Description du corpus

L'un des éléments clés de toute recherche est sa base de données, dans cet ordre d'idée, nous avons essayé de développer un corpus constitué de 126 mots en Arabe Standard, destiné essentiellement à mettre en évidence l'occurrence du phonème [ʃ], ainsi qu'une deuxième base de données concernant des phonèmes isolées, en vue de son utilisation dans une nouvelle approche qui est la déviation phonémique.

2.1 Enregistrement dans un milieu Bruité

L'enregistrement du corpus ont été faits dans une salle ordinaire, ceci afin d'inclure le bruit environnant, car nous préconisons d'utiliser notre système d'aide dans un environnement en conditions normales (dans un cabinet de médecin, à la maison, dans une salle de classe, etc.).

2.2. Locuteurs

Les 126 mots du corpus ont été enregistrés par une vingtaine de locuteurs (hommes et femmes), (tableau 5.1).

Tableau 5.1 : Différents locuteurs ayant enregistré le corpus avec Fi, Feminin et Mi Masculin

Code	Sexe	Age (ans)	Observations
F1	F	40	
F2	F	27	
F3	F	8	
F4	F	31	
F5	F	32	
F6	F	28	
F6P	F	21	voix Pathologique
F7	F	27	
F8	F	35	
F9	F	35	
M1	M	22	Accent régional
M2	M	27	Accent régional
M3	M	43	
M4	M	23	Nasalisation
M5	M	30	Nasalisation
M6	M	16	
M7	M	45	
M8	M	35	
M9	M	35	
M10	M	25	
M11	M	36	
M12	M	43	
M13	M	45	
M14	M	30	
M15	M	37	

Dans le but de faire des essais sur un corpus présentant le sigmatisme occlusif, nous avons pu enregistrer un cas de sexe féminin âgée de 19 ans, « code F6P », en parole continue et en mots isolés afin de bien distinguer l'impact du sigmatisme sur les mots en débit lent, normal et rapide.

La pathologie a été revue par le Dr. Ghozali du service de rééducation fonctionnelle du CHU Frantz Fanon de Blida.

2.3. Constitution du corpus de Mot

Les différents mots sélectionnés ont été tirés du dictionnaire "Al Mawrid El Wasit" du Dr. Rouhi Balabaki et al [78]. Ces mots sont divisés en 3 classes d'occurrence (Annexe B), Ces

classes représentant les trois cas possible du positionnement du phonème pathologique par rapport aux voyelles existantes.

Classe 1 : Début de mot (Annexe B tableau B.1)

- 22 occurrences de [ʃ] suivi de [a]
- 17 occurrences de [ʃ] suivi de [o]
- 17 occurrences de [ʃ] suivi de [i]

Classe 2 : Milieu de mot (Annexe B tableau B.2)

- 53 occurrences du [ʃ] avec les combinaisons suivantes :
- [aʃa],[aʃi],[aʃo] ;
- [iʃa],[oʃi],[oʃo].

Classe 3 : Fin de mot (Annexe B tableau B.3)

- 12 occurrences du [v_iʃ] avec [v_i]≡ [a],[o],[i].

3. Description du corpus de phonèmes

Ce corpus concerne les phonèmes enregistrés localement, ainsi que quelques phonèmes de la Base de Données TIMIT, qui correspondent auditivement aux phonèmes de la langue Arabe, (Annexe B tableau B.4).

4. Présentation du cas d'étude

En vue de pouvoir réaliser la première étape de segmentation automatique des phonèmes, nous avons sélectionné des mots présentant le sigmatisme occlusif d'une façon très distinguée lors de la prononciation des phonèmes [ʃ] et [s]. Ces derniers présentent une grande ambiguïté de prononciation du fait du rapprochement de leur lieu d'articulation (tableau 5.2).

Tableau 5.2 : Mots présentant le cas du sigmatisme

Code	Mots en Arabe	TOP	Prononciations pathologiques	Transcriptions en Arabe standard
C1	شخصية	[ʃ a χ s i j a] [ʃ a χ s i j a t u n]	[θ a χ t i j a] [θ a χ ʃ θ i j a]	ثخثـيـة ثخشثيـة
C2	شمس	[ʃ a m s ø] [ʃ a m s u n]	[θ a m ʃ ø] [θ a m θ u n]	ثـمـش ثـمـث

Notons que quelques locuteurs, ou même la patiente prononçait le mot avec la nasalisation, induisant une extension du mot par la composition [t u n].

Nous avons segmenté manuellement en mots et en phonèmes les segments dont nous avons besoin. Chaque mot ou phonème a été étiqueté avec un code spécifique (tableau 5.3).

Tableau 5.3. : Mots de tests sélectionnés à partir
de la base de données enregistrées.

TOP	Locuteurs	Codification
[ʃ a χ s i j a]	F	F_Chaksiya1
	M	M_Chaksiya2
[ʃ a m s ø]	F	F_Chamsse5
	M	M_Chamsse3

Chaque mot ou phonème des mots cibles est ensuite modélisé par les HMM avec les GMM correspondantes.

Avant d'entamer le traitement des données par une modélisation markovienne, nous allons voir quelques aspects visuels des signaux sonores et faire des comparaisons, dans le but de comprendre le fait de prononcer le [θ] au lieu du [ʃ] et son impact sur les différents spectrogrammes (déplacement formantique).

5. Comparaison visuelle des différentes prononciations

A titre de comparaison visuelle, nous avons choisi quelques locuteurs (masculin et féminin) lors de la prononciation du mot C1.

Nous remarquons le défaut du [ʃ] transformé en [θ], prononcé par la Patiente F6P (figure 5.1)

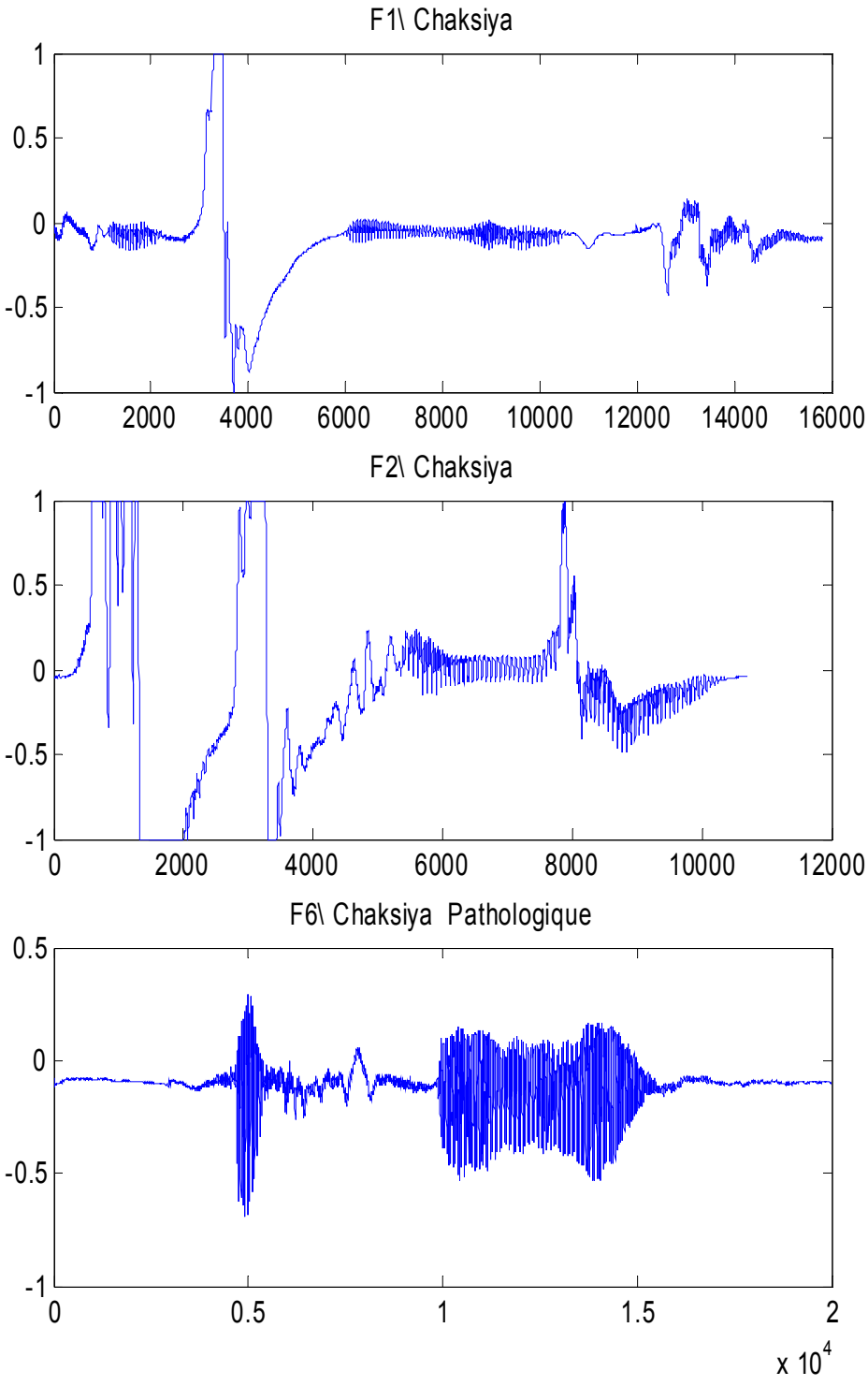


Figure 5.1 : Comparaisons visuelles des prononciations
(F1 / F2 / F6P)

La seconde comparaison est effectuée avec des locuteurs masculins de référence (figure 5.2).

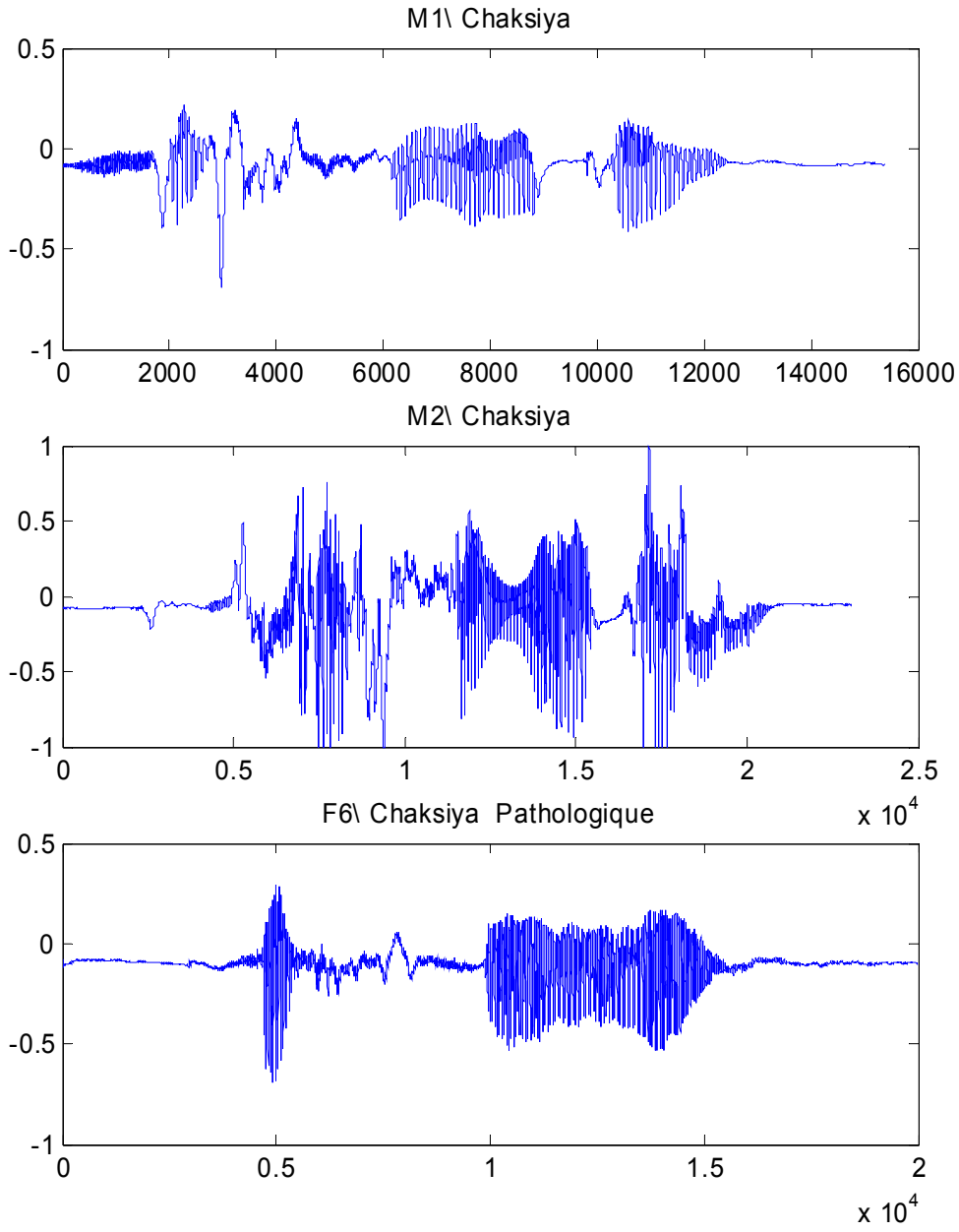


Figure 5.2 : Comparaisons visuelles des prononciations
(M1 / M2 / F6P)

Il est difficile de décider sur la pathologie visuellement, pour cela, l'utilisation de méthodes utilisant des techniques d'analyse plus avancées est impérative.

6. Spectrogrammes et transitions formantiques

Le spectrogramme est un moyen de voir le déplacement formantique et dire si le bon phonème a été prononcé ou non, ceci se remarque par une analyse visuelle comparative des fréquences propres à chaque phonème, (figures 5.3, 5.4 et 5.5).

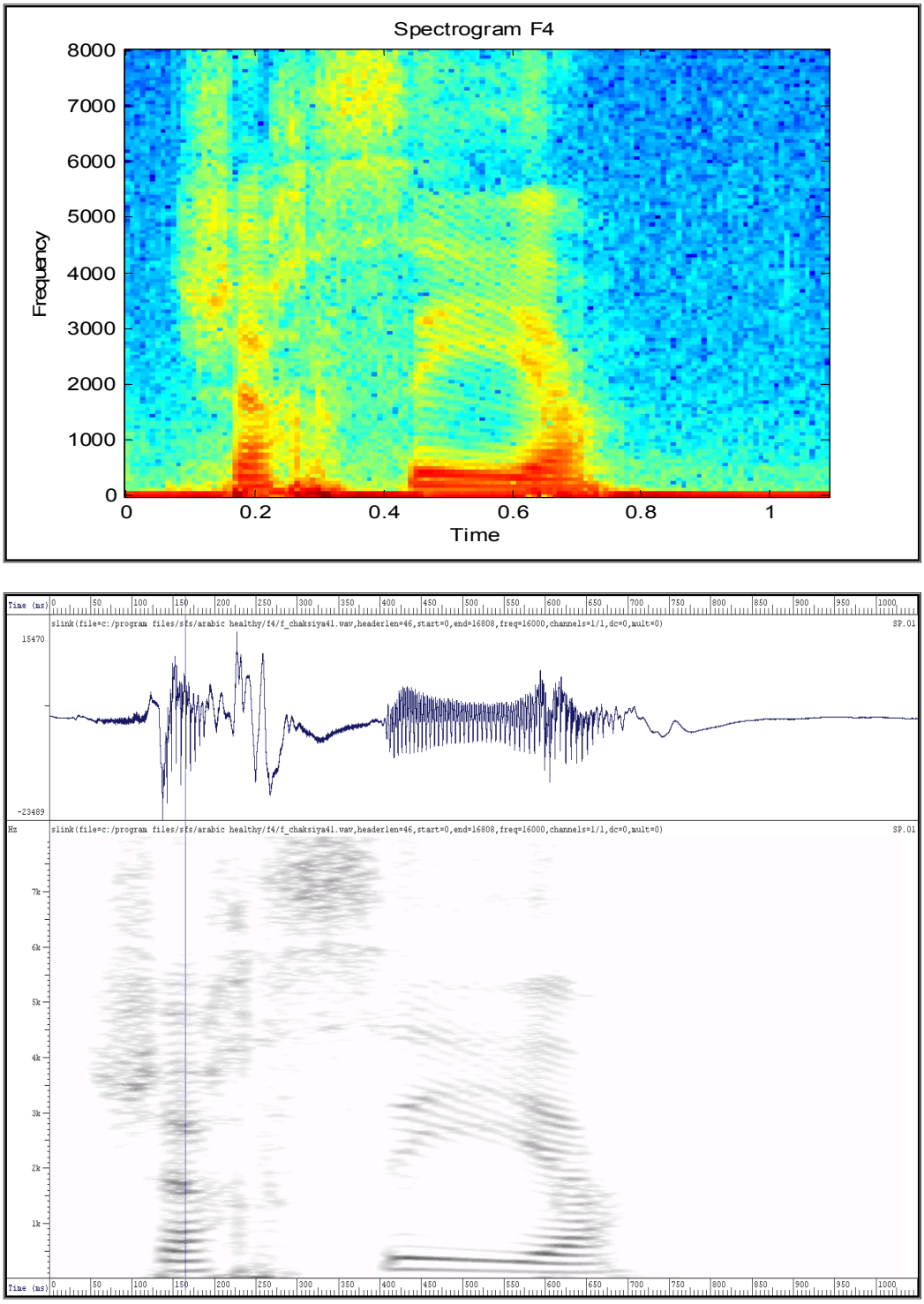


Figure 5.3 : Spectrogramme du mot « C1 »prononcé par F4.

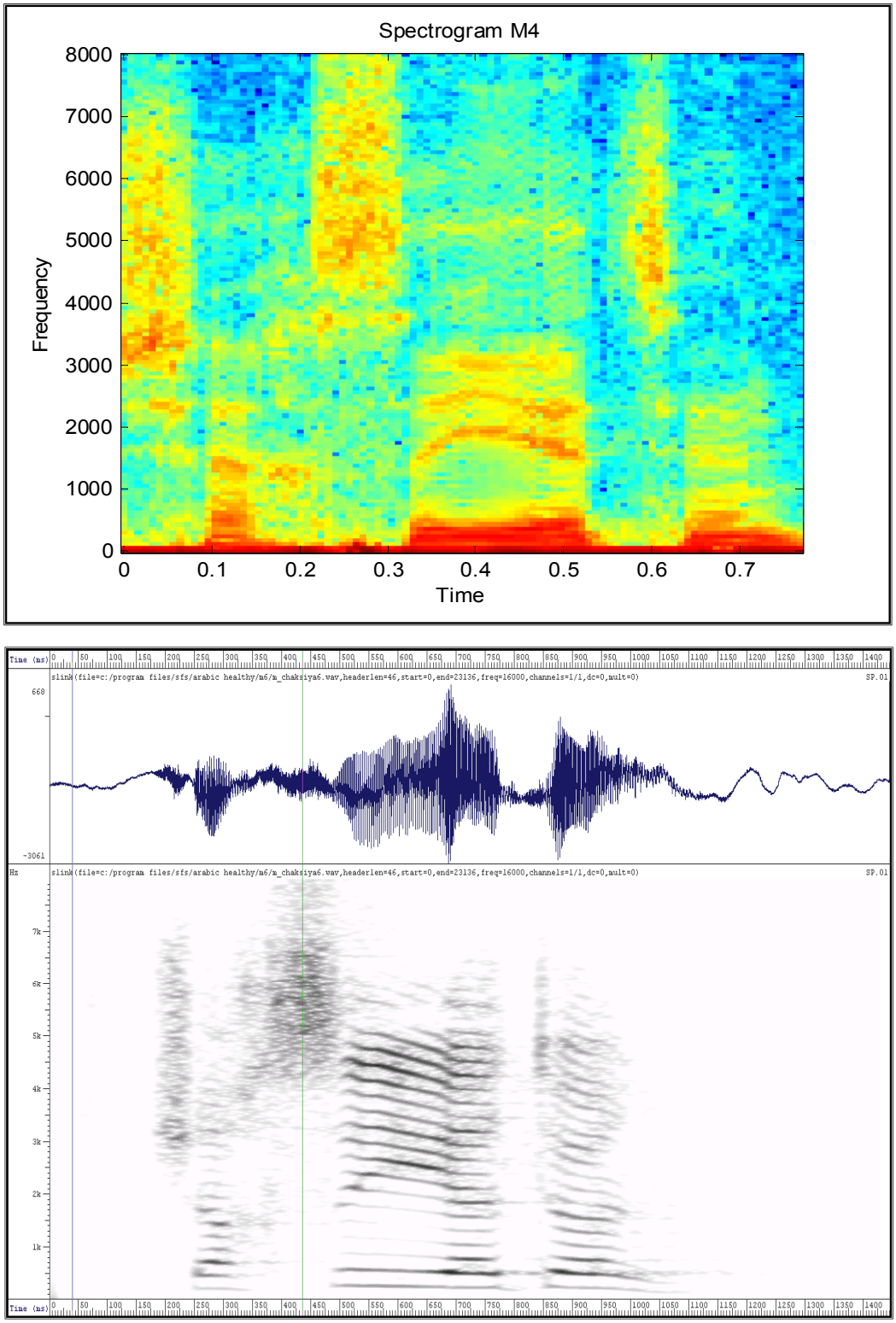
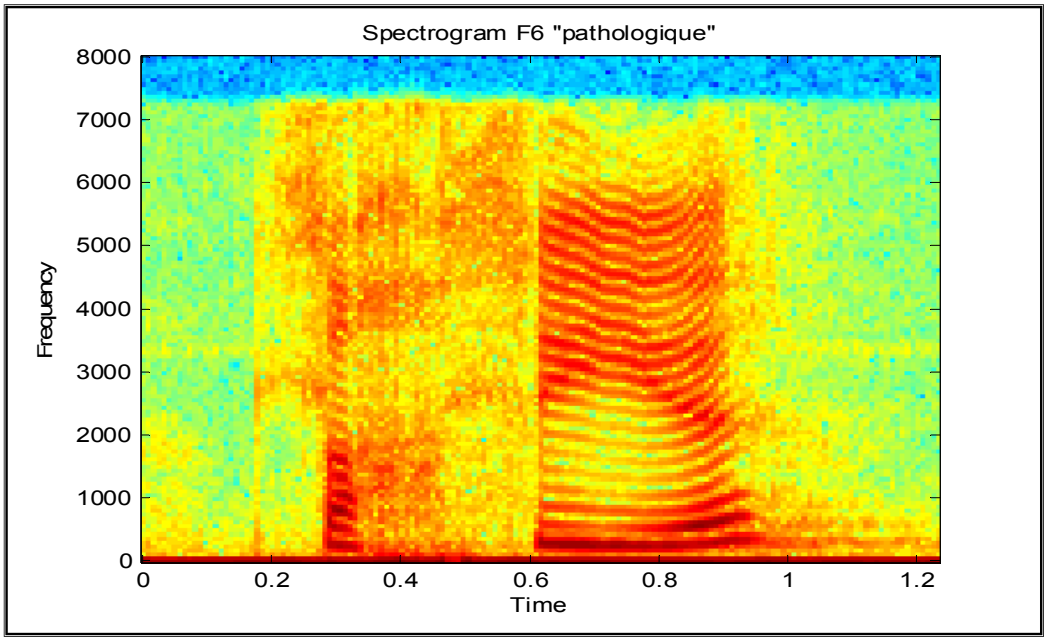


Figure 5.4 : Spectrogramme du mot « C1 »prononcé par M4.



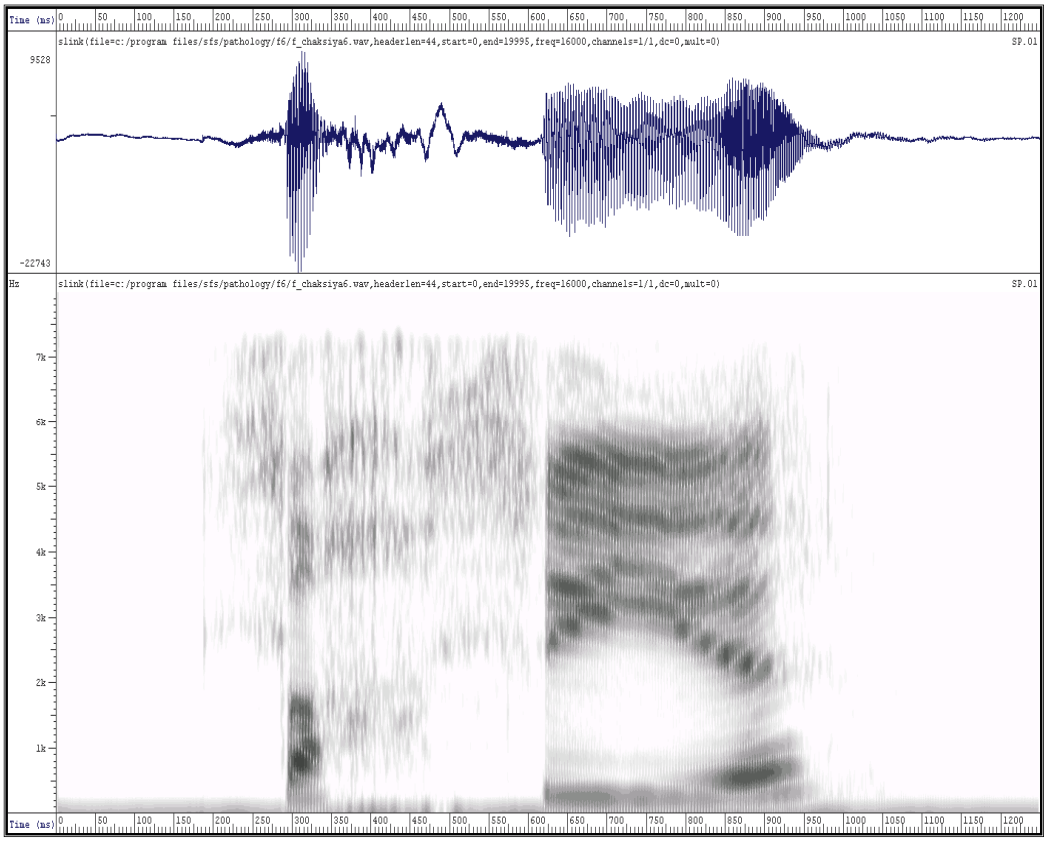


Figure 5.5 : Spectrogramme du mot « C1 »prononcé par F6P

Patiente F6 présentant la pathologie

7. Méthodologie de travail

Toutes les comparaisons présentées précédemment se basent sur une connaissance a priori, toutefois si le système d'aide doit être utilisé par un patient, les graphes précédents deviennent obsolètes, pour cela, il y a lieu d'automatiser les actions et de donner des courbes et/ou graphes d'évolution ou même des scores sur 20 ou 100, ainsi qu'un taux de déviation phonémique qui seront d'une utilité significative soit pour le thérapeute ou le patient quel que soit son âge.

La méthodologie adoptée se base sur l'utilisation de classificateurs tels que les HMM/GMM/ANN ou toute combinaison efficace, utilisant les MFCC comme paramètres acoustiques robustes, en vue d'une reconnaissance phonémique poussée et du calcul d'un score de vraisemblance ainsi qu'un score de déviation phonémique (Figure 5.6).

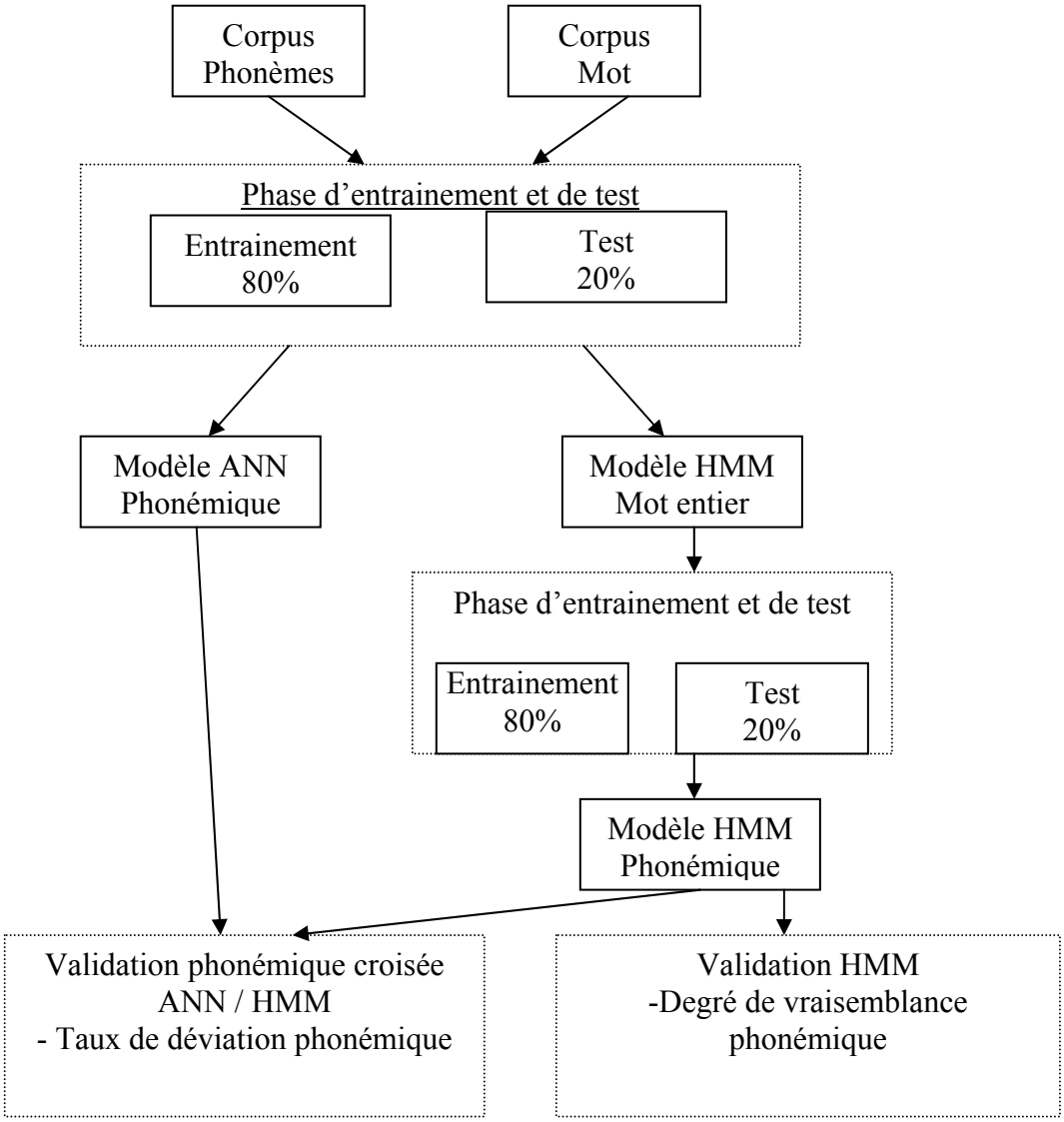


Figure 5.6 : Schéma synoptique général de la méthodologie de travail.

8. Modélisation des mots du corpus par les HMM

En premier lieu, nous avons effectué une modélisation sur 80% du corpus, les autres 20% restant ont servi aux tests. Le nombre d'états des modèles de Markov est spécifique à la longueur phonémique du mot, sachant que les comparaisons se font au sein d'une même classe et la détection du phonème mal prononcé est sujette au mot pris comme exemple de test.

Les modèles HMM / GMM sont définis par :

- ♦ le modèle du HMM, dans notre cas d'étude tous les modèles sont de type gauche-droite ;
- ♦ le nombre d'états ;
- ♦ l'état initial le plus probable ;
- ♦ le nombre de gaussiennes modélisant les données par état ;
- ♦ le nombre de paramètres représentant les données ;
- ♦ le nombre d'itérations de l'algorithme EM.

8.1. Modèles HMM du mot $\lfloor \int a\chi.s.i.j.a \rfloor$

En premier lieu, nous allons modéliser le mot $\lfloor \int a\chi.s.i.j.a \rfloor$ par deux modèles :

- ♦ le premier modèle est un HMM à 7 états en considérant chaque phonème (consonne et voyelle) comme des états séparés.
- ♦ le second modèle en insistant sur l'assimilation d'une voyelle par une consonne, ce qui réduirait les états à 4 ([ʃa][χ][si] [ja]).

Le nombre de coefficients MFCC varie entre 12 et 42, le nombre de GMM varie entre 6 et 16 pour chaque état, pour définir, ainsi, une stratégie de robustesse globale.

8.1.1. Modèle à 7 états du mot : HMM1 $\lfloor \int a\chi.s.i.j.a \rfloor$

Les paramètres globaux des modèles HMM ainsi que la matrice de transition obtenus sont mentionnées respectivement dans les tableaux 5.4. à 5.11.

Tableau 5.4 : Paramètres du modèle HMM1

Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	Coefficients MFCC/trame
7	6	15	12

Tableau 5.5 : Matrice de transition du modèle HMM1

Transitions	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5	Etat 6	Etat 7
Etat 1	0.95699	0.043007	2.99E-08	9.74E-29	1.79E-47	3.75E-15	0
Etat 2	0	0.94985	0.04582	0.0019421	3.23E-41	0.0023881	0
Etat 3	0	0	0.95501	0.042649	3.16E-18	0.002345	0
Etat 4	0	0	0	0.95187	0.046425	0.0017044	1.33E-14
Etat 5	0	0	0	0	0.9456	0.045001	0.0094008
Etat 6	0	0	0	0	0	0.96995	0.030052
Etat 7	0	0	0	0	0	0	1

Tableau 5.6 : Paramètres du modèle HMM1

Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	coefficients MFCC/trame
7	8	22	39

Tableau 5.7 : Matrice de transition du modèle HMM1

Transitions	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5	Etat 6	Etat 7
Etat 1	0.91649	0.083515	2.62E-76	0	0	0	0
Etat 2	0	0.90831	0.082952	8.73E-03	0	0	0
Etat 3	0	0	0.93195	0.068052	0	0	0
Etat 4	0	0	0	0.93941	0.05762	0.0029655	0
Etat 5	0	0	0	0	0.92347	0.076525	4.25E-23
Etat 6	0	0	0	0	0	0.93709	0.06291
Etat 7	0	0	0	0	0	0	1

Tableau 5.8 : Paramètres du modèle HMM1

Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	coefficients MFCC/trame
7	16	31	39

Tableau 5.9. : Matrice de transition du modèle HMM1

Transitions	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5	Etat 6	Etat 7
Etat 1	0.92132	0.078675	0	0	0	0	0
Etat 2	0	0.91329	0.086709	0	0	0	0
Etat 3	0	0	0.93874	0.058347	0.0029117	0	0
Etat 4	0	0	0	0.92089	0.079108	0.00E+00	0
Etat 5	0	0	0	0	0.92257	0.077425	0
Etat 6	0	0	0	0	0	0.93671	0.063295
Etat 7	0	0	0	0	0	0	1

Tableau 5.10 : Paramètres du modèle HMM1

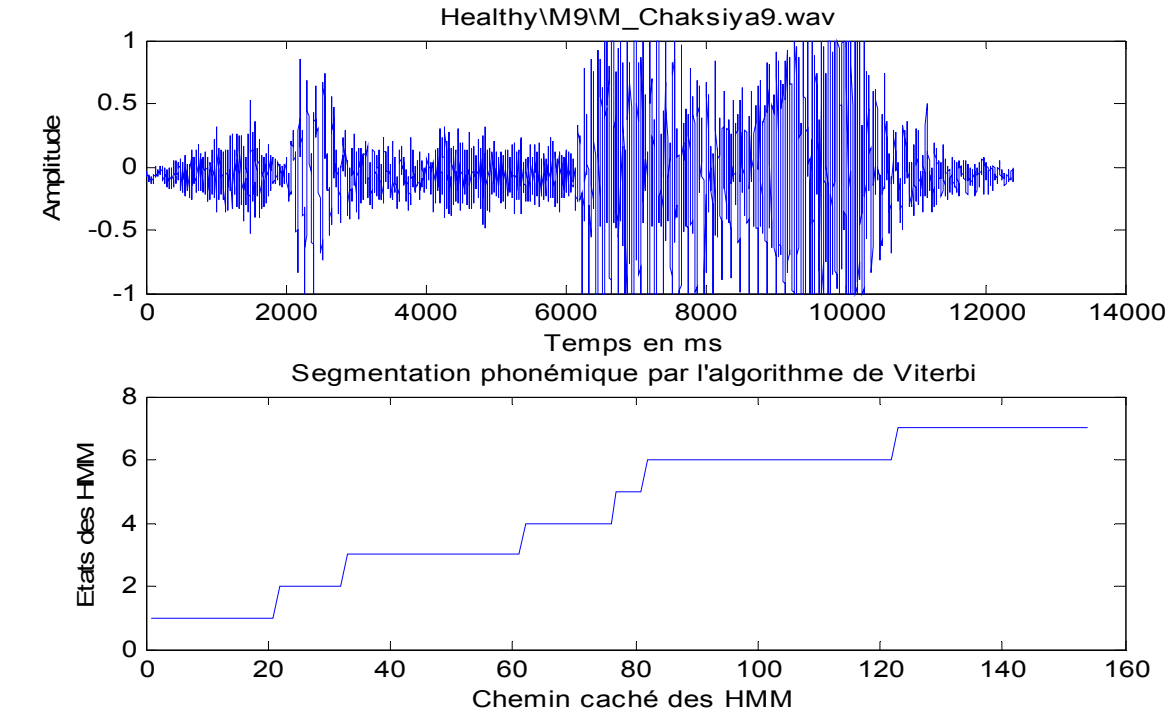
Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	coefficients MFCC/trame
7	16	26	42

Tableau 5.11 : Matrice de transition du modèle HMM1

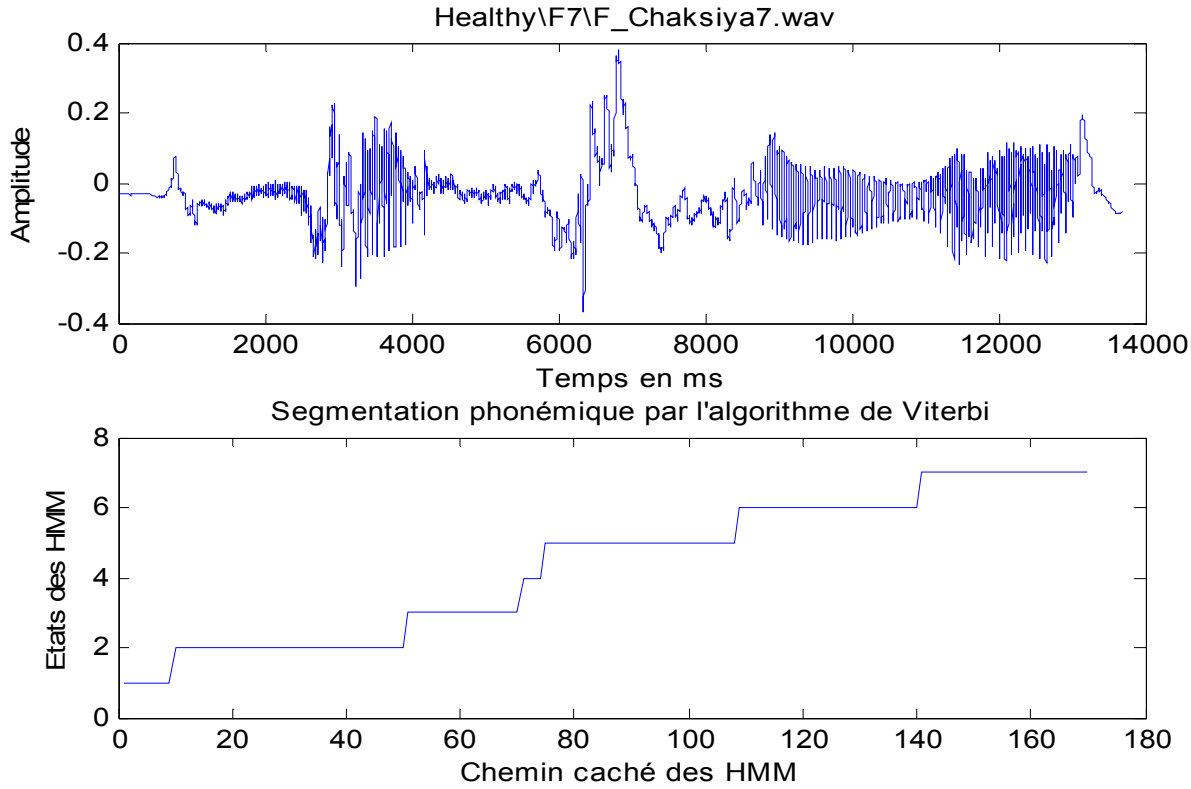
Transitions	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5	Etat 6	Etat 7
Etat 1	0.92499	0.075007	0	0	0	0	0
Etat 2	0	0.915	0.085001	0	0	0	0
Etat 3	0	0	0.93403	0.065972	0	0	0
Etat 4	0	0	0	0.92089	0.07911	3.36E-166	0
Etat 5	0	0	0	0	0.9174	0.082598	0
Etat 6	0	0	0	0	0	0.92764	0.07236
Etat 7	0	0	0	0	0	0	1

Remarquons que le modèle HMM1 s'affine du premier modèle, (tableau 5.7, 5.8.), vers un modèle de type gauche droite, voir tableaux 5.13, 5.14 par l'annulation des probabilités de transition des états non successifs, exemple de l'état 2 à l'état 4 ou l'état 5 à l'état 7, etc.

Les figures 5.7. (a & b) montrent la segmentation phonémique selon 12 coefficients MFCC, pour un locuteur et une locutrice (saine).



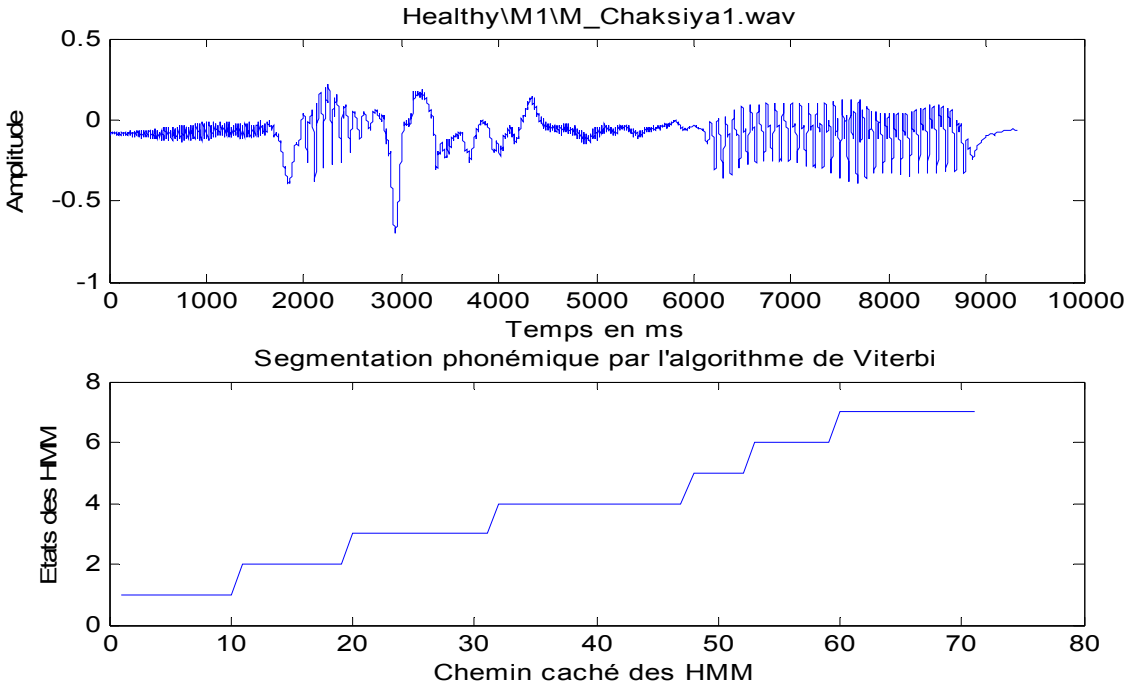
(a)



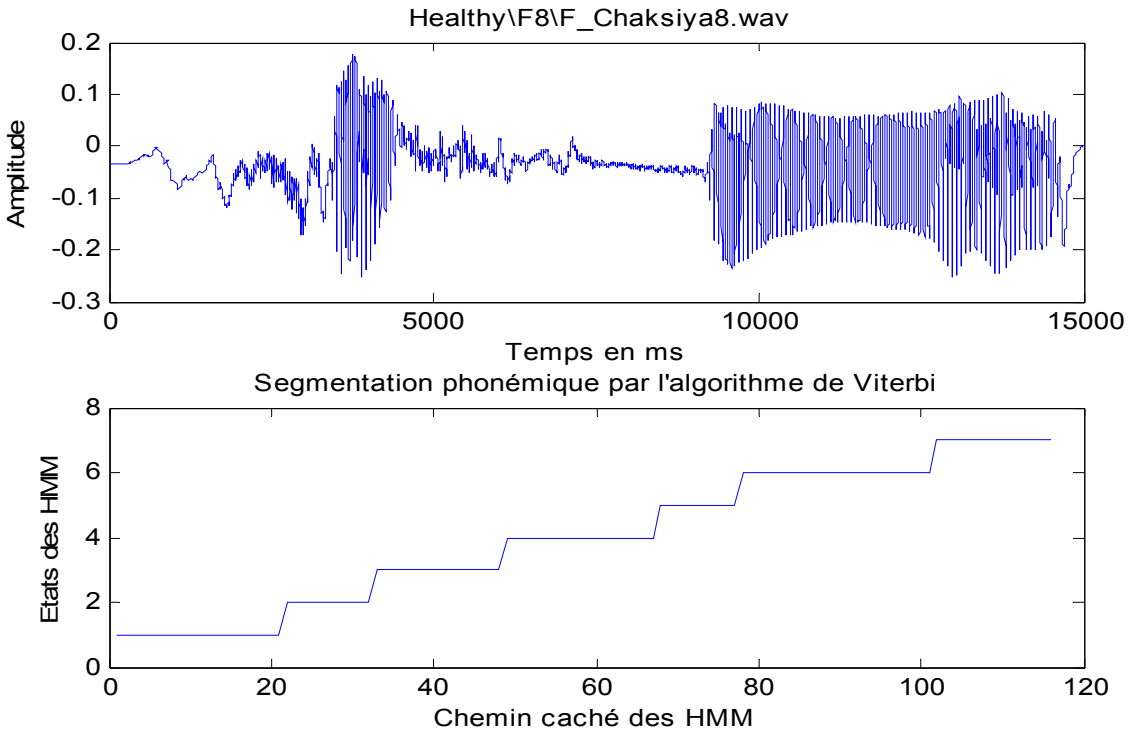
(b)

Figure 5.7 (a, b) : Segmentation phonémique pour différents locuteurs en utilisant 12 coefficients MFCC

Les figures 5.8, (a & ab) montrent la segmentation phonémique, en utilisant 39 coefficients MFCC (dérivées premières et secondes sans le log énergie de la trame), de deux locuteurs M1 et F8.



(a)



(b)

Figure 5.8 (a, b) : Segmentation phonémique pour différents locuteurs en utilisant 39 MFCC

8.1.2. Courbes de convergence de l'algorithme EM

La figure 5.9. Illustre la différence de convergence de l'algorithme EM après l'adjonction des dérivées première et seconde des MFCC, avec/sans la log-énergie des trames [79].

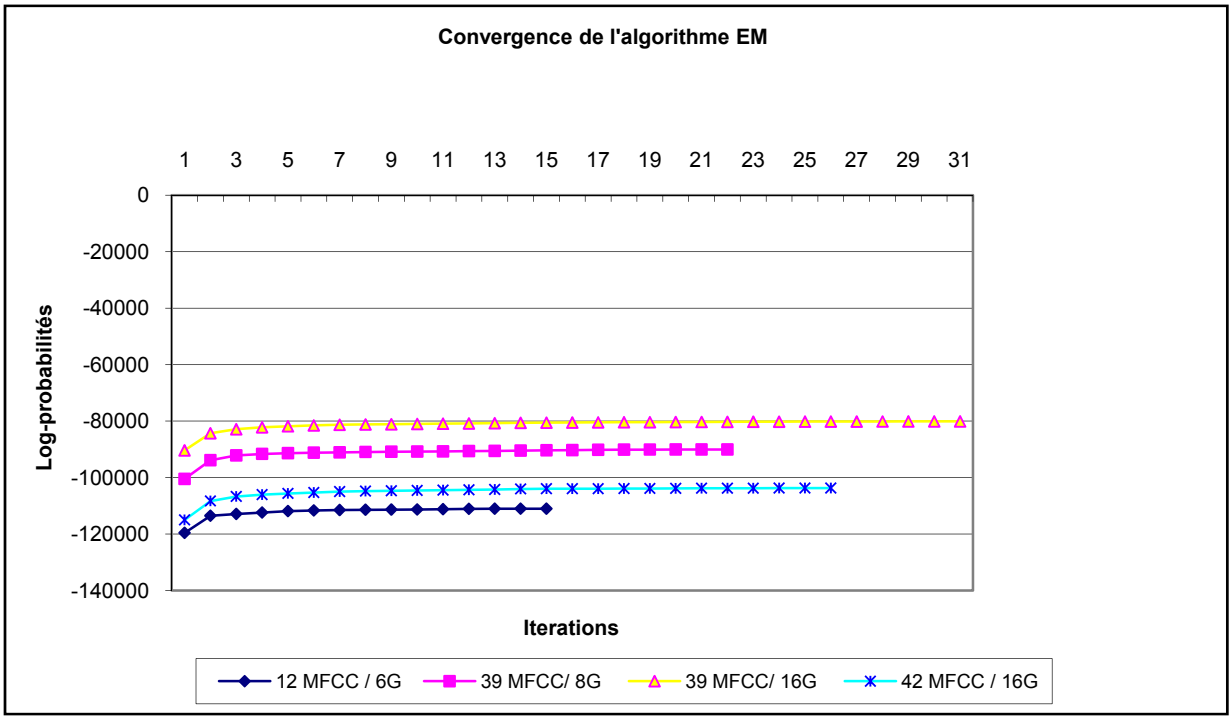


Figure 5.9 : Courbe de convergence de l'algorithme EM (HMM à 7 états) Légende :(# de Coefficients MFCC/ # de gaussiennes associées à chaque état)

Nous avons remarqué que l'augmentation du nombre de coefficients agit directement sur la matrice de transition et fait tendre notre HMM vers un modèle de type gauche droite, pour une meilleure segmentation phonémique. Ceci n'influe pas directement sur la convergence en minimisant les itérations, mais agit plutôt sur la vitesse de convergence (temps d'une itération).

8.1.2. Modèle à 4 états : HMM2 $\left[\int .a\right][\chi][s.i][j.a]$

Le deuxième modèle concerne un HMM dans lequel nous essayons de prendre l'assimilation Consonne Voyelle. Les matrices de transition obtenues sont mentionnées respectivement dans les tableaux 5.12 à 5.15.

Tableau 5.12 : Paramètres du modèle HMM2

Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	Coefficients MFCC/trame
4	4	10	12

Tableau 5.13 : Matrice de transition du modèle HMM2 :

Transitions	**Etat 1**	**Etat 2**	**Etat 3**	**Etat 4**
Etat 1	0.9743	0.018417	0.0060612	0.0012239
Etat 2	0	0.97369	0.026313	9.4388e-019
Etat 3	0	0	0.98267	0.017334
Etat 4	0	0	0	1

Tableau 5.14 : Paramètres du modèle HMM2

Nombre d'états du HMM	Mélange de gaussiennes par état	Nombre d'itérations de l'algorithme EM	coefficients MFCC/trame
4	8	35	39

Tableau 5.15 : Matrice de transition du modèle HMM2

Transitions	**Etat 1**	**Etat 2**	**Etat 3**	**Etat 4**
Etat 1	0.95223	0.047767	**0**	**0**
Etat 2	0	0.96346	0.036543	**0**
Etat 3	0	0	0.93459	0.065406
Etat 4	0	0	0	1

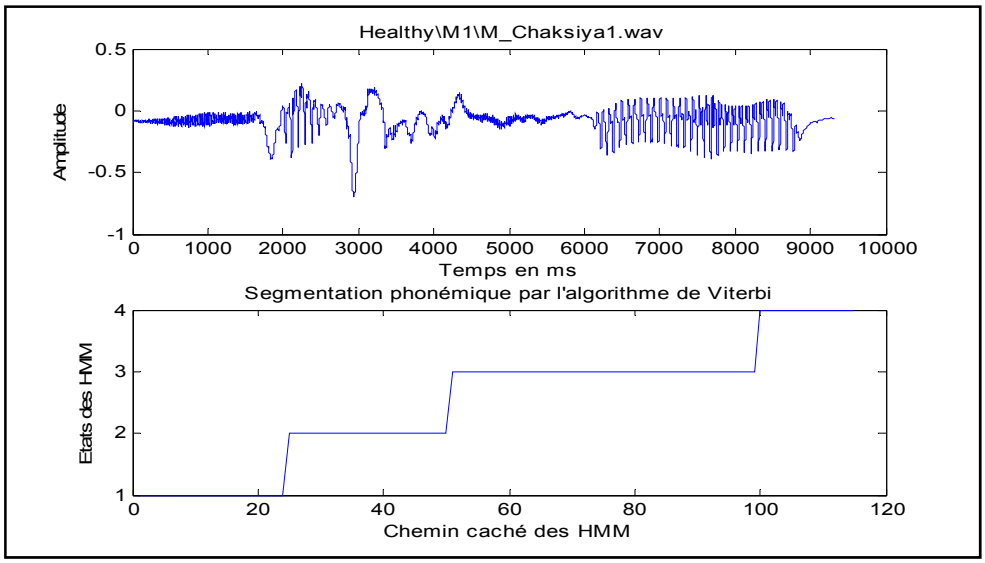
Remarquons que le modèle HMM2, dans ce cas aussi, s'affine vers un modèle de type gauche droite, par l'annulation des probabilités de transition de l'état 2 vers état 4, ou de l'état 1 à l'état 3.

Maintenant, essayons de procéder à la sélection du meilleur modèle :

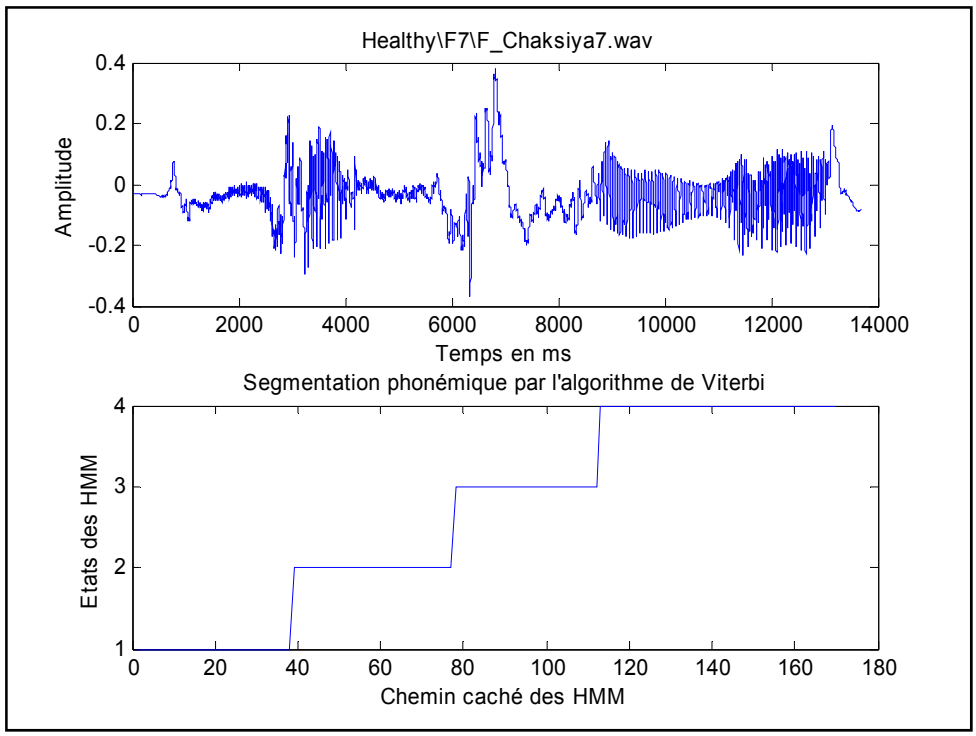
- le modèle HMM1 (7 états) à 42 coefficients, ainsi que 16 gaussiennes par état, donne la « meilleure » matrice de transition, et servira par la suite à l'extraction des phonèmes, un par état, qui seront réutilisés pour la segmentation automatique.

♦ le modèle HMM1 (4 états) à 42 coefficients, 16 gaussiennes est basé sur la supposition qu'il y a assimilation. Nous avons remarqué que les locuteurs/locutrices, n'assimilent pas tous de la même manière, donc le modèle ne peut réellement être utilisé pour la comparaison, malgré que ce soit un modèle minimal, avec moins de calcul.

Les figures 5.10. (a,b) montrent la segmentation phonémique selon 12 coefficients MFCC seulement, remarquons que les observations associées aux états 3 et 4 ne sont pas distribuées d'une manière adéquate sur tous les locuteurs.



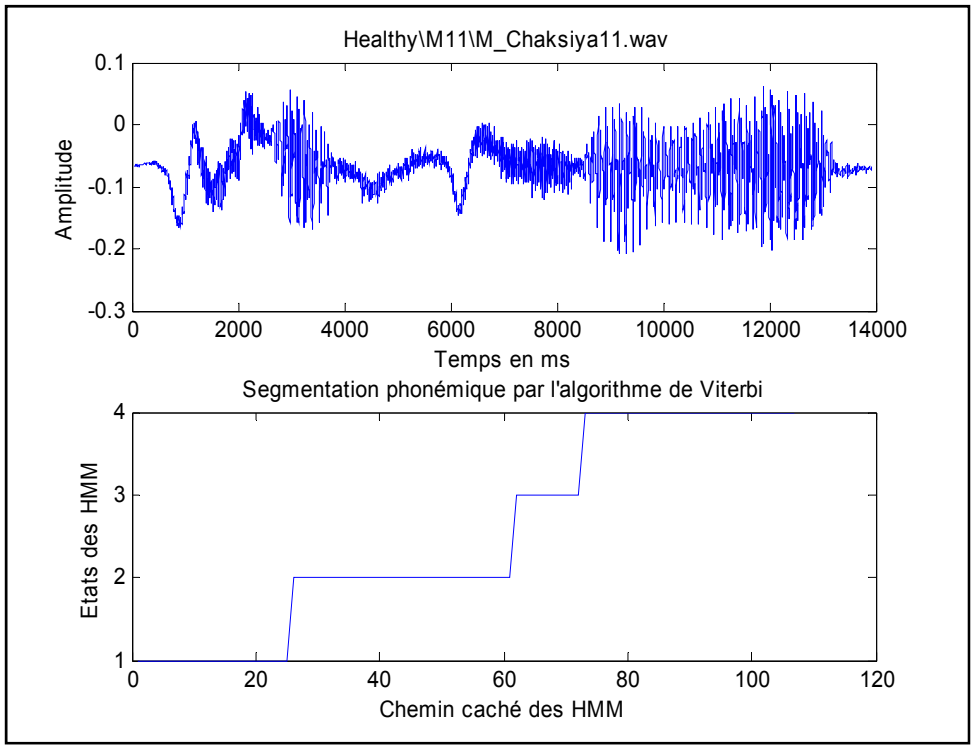
(a)



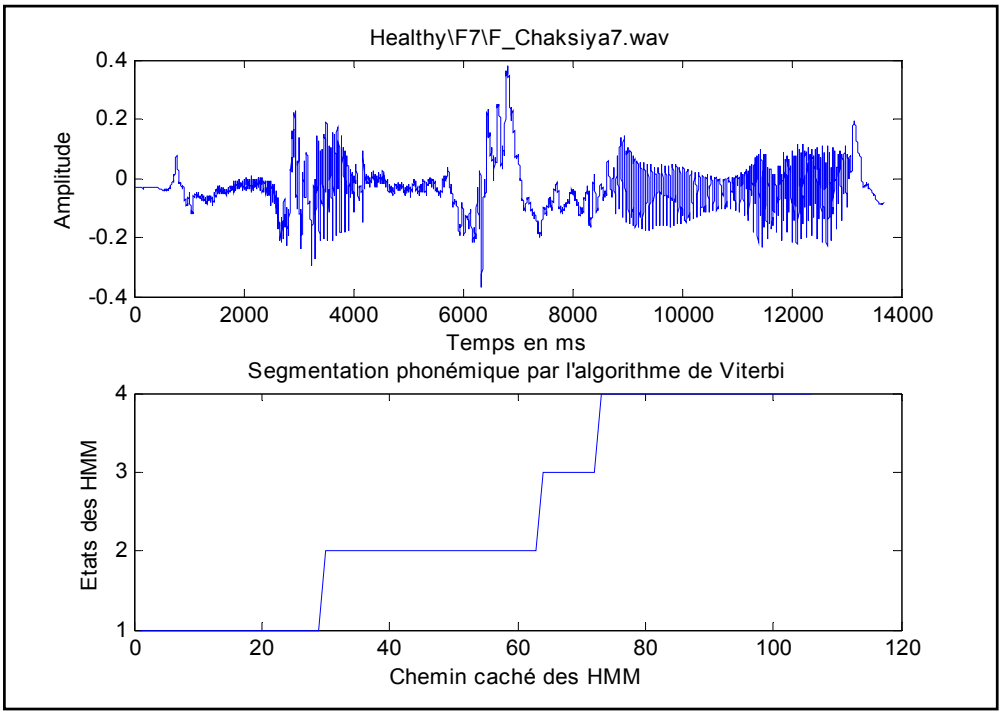
(b)

Figure 5.10 (a, b) : Segmentation phonémique, en utilisant 12 coefficients MFCC

Les 12 coefficients MFCC sont adjoints avec leurs dérivées premières et secondes en vue de voir la différence phonémique visuellement.



(a)



(b)

Figure 5.11 : Segmentation phonémique pour des locuteurs différents en utilisant 39 MFCC.

Les segmentations phonémiques de deux locuteurs (M11 et F7) ont été reproduites pour les deux modèles HMM2, en vue de voir la différence visuelle, en fonction des coefficients MFCC variant entre de 12 à 39.

8.1.4. Courbes de convergence de l'algorithme EM

La figure 5.12, Illustre la différence de convergence de l'algorithme EM après l'adjonction des dérivées première et seconde des MFCC, ainsi que du log-énergie des trames [79].

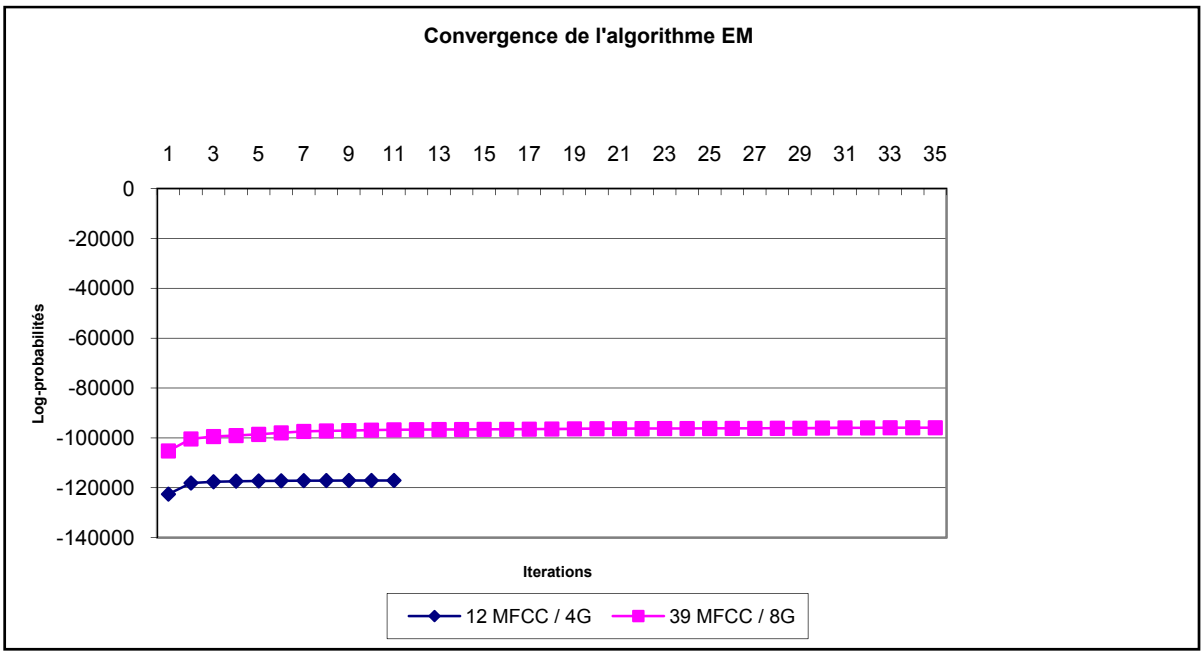


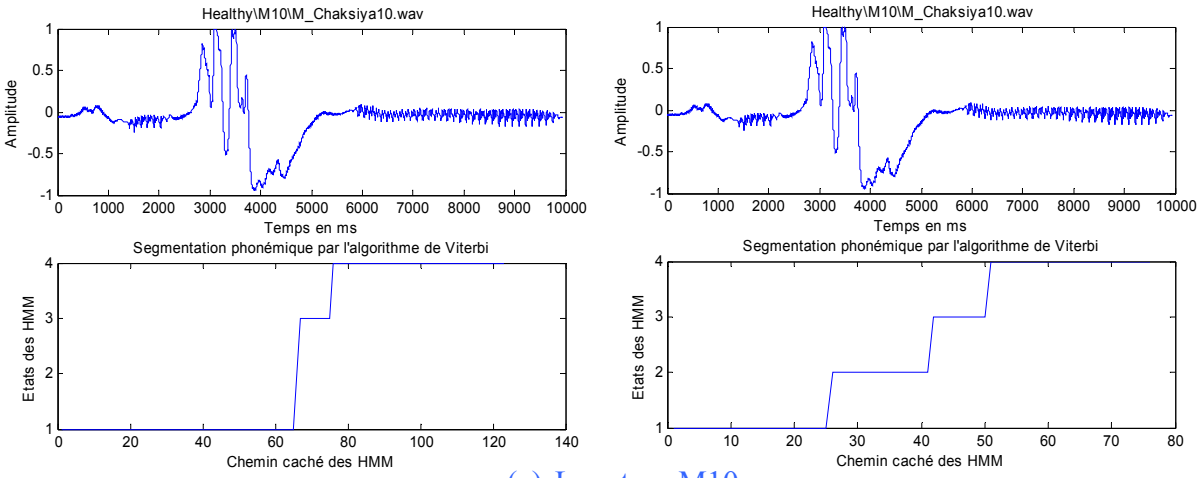
Figure 5.12 : Courbe de convergence de l'algorithme EM (HMM à 4 états)

8.1.2. Discussions

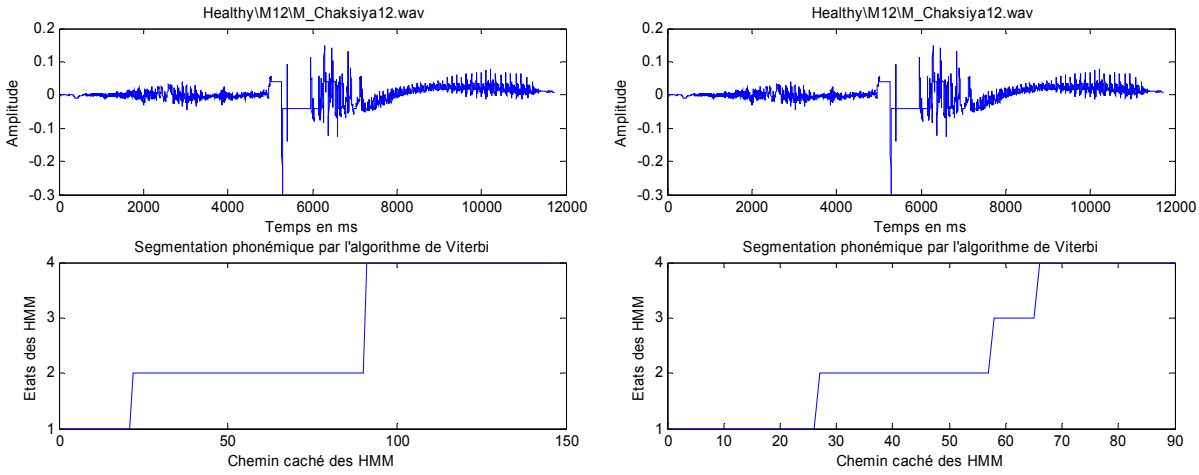
Nous avons remarqué que l'augmentation du nombre de coefficients MFCC agit directement sur la matrice de transition et fait tendre notre HMM vers un modèle de type gauche droite. Toutefois, la segmentation pour le modèle HMM2 à 4 états, n'est pas satisfaisante, selon les états demandés, car une partie des paramètres est prise par les gaussiennes adjacentes, ce qui tend a générer un HMM à 2 ou 3 états, donnant de fausses segmentations (figures 5.13 et 5.14).

Segmentation défaillante
12 MFCC /4 Gaussiennes par état

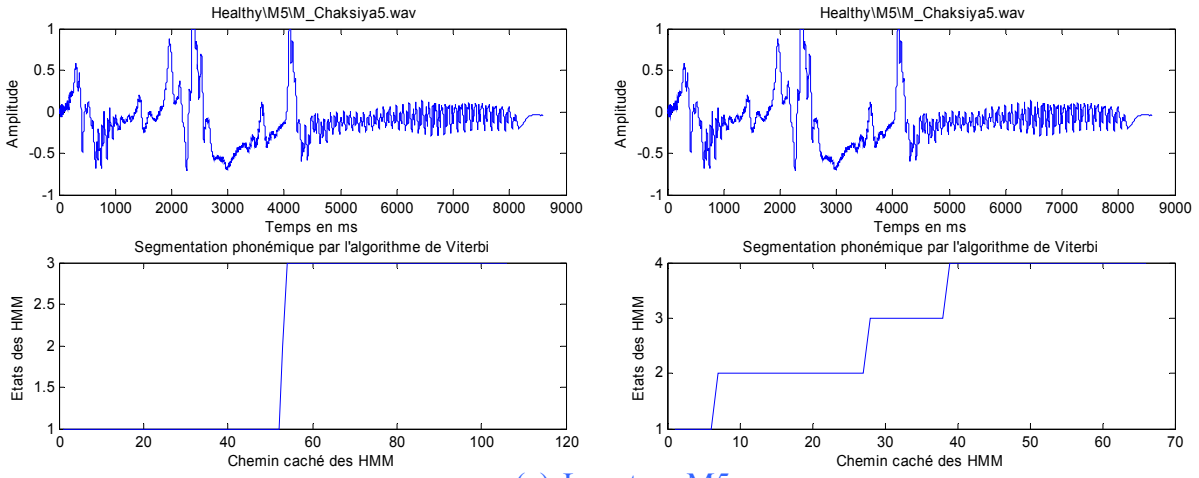
Segmentation désirée
39 MFCC /8 Gaussiennes par état



(a) Locuteur M10



(b) Locuteur M12

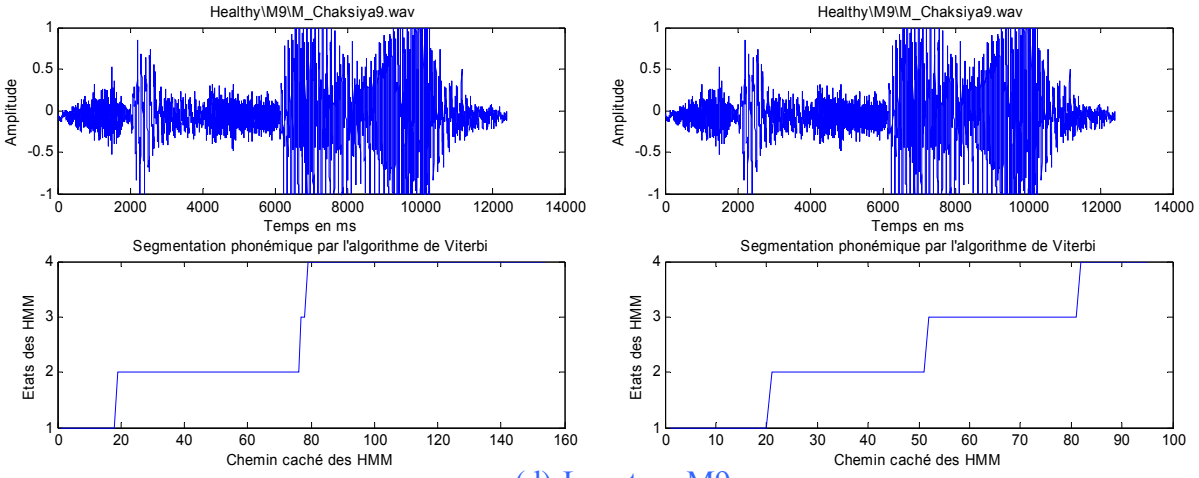


(c) Locuteur M5

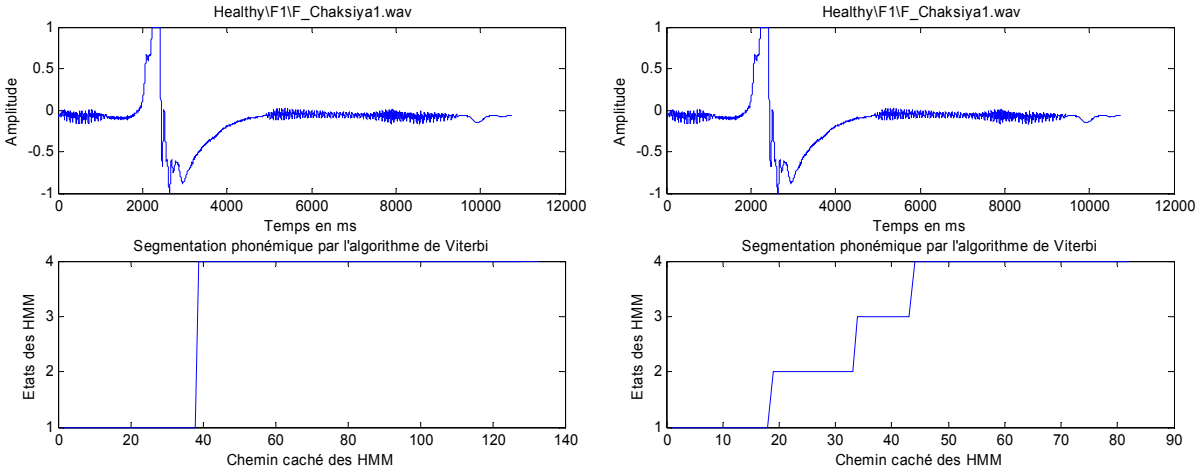
Figure 5.13 : Segmentation défaillante, cas du modèle HMM à 4 états

Segmentation défaillante
12 MFCC / 4Gaussiennes par état

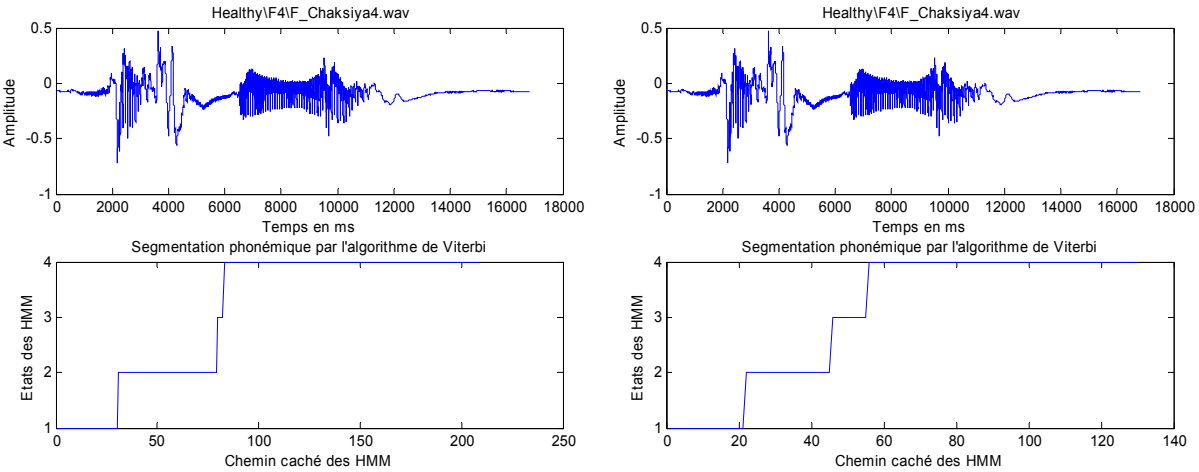
Segmentation désirée
39 MFCC / 8Gaussiennes par état



(d) Locuteur M9



(e) Locutrice F1



(f) Locutrice F4

Figure 5.14 : Segmentation défaillante, cas du modèle HMM à 4 états

9. Validation de nos travaux sur la phrase SA1 de la base TIMIT

Cette étape est une validation de notre modélisation HMM, en effet, notre choix se porte sur la phrase SA1 : «She had your dark suit in greasy wash water all year», nous allons essayer de reconnaître les 11 mots de cette phrase par notre approche (tableau 5.16 et 5.17).

Tableau 5.16. : Modèles HMM de la phrase SA1 par 80% du corpus (80 locuteurs)

Mots	Modèle HMM	Matrice de Transition	Moyenne par état	Variance par état	Mélange de gausiennes
She	**Mot1**	5*5	39*5*5	39*39*5*5	5 * 5
Had	**Mot2**	4*4	39*4*5	39*39*4*5	4 * 5
Your	**Mot3**	2*2	39*2*5	39*39*2*5	2 * 5
Dark	**Mot4**	4*4	39*4*5	39*39*4*5	4 * 5
Suit	**Mot5**	3*3	39*3*5	39*39*3*5	3 * 5
In	**Mot6**	1*1	39*1*5	39*39*1*5	1 * 5
greasy	**Mot7**	7*7	39*7*5	39*39*7*5	7 * 5
Wash	**Mot8**	4*4	39*4*5	39*39*4*5	4 * 5
water	**Mot9**	5*5	39*5*5	39*39*5*5	5 * 5
All	**Mot10**	2*2	39*2*5	39*39*2*5	2 * 5
Year	**Mot11**	3*3	39*3*5	39*39*3*5	3 * 5

Reconnaissance des mots de la phrase SA1 par les 20% du corpus restant (20 locuteurs) :

Tableau 5.17 : Taux de reconnaissance sur la base TIMIT (20 locuteurs)

Mots	Modèle HMM	Taux de reconnaissance
She	Mot1	95%
Had	Mot2	100%
Your	Mot3	100%
Dark	Mot4	100%
Suit	Mot5	100%
In	Mot6	95%
Greasy	Mot7	100%
Wash	Mot8	100%
Water	Mot9	100%
All	Mot10	100%
Year	Mot11	100%
Taux de reconnaissance Phrase compléte		**99%**

La phrase SA1 est reconnue à 99% des cas, ceci montre qu'en terme de reconnaissance, l'approche HMM est adéquate à notre problème, toutefois la base TIMIT c'est faite sous un enregistrement sans bruit, donc l'approche est « stable » si la base contient le moins de bruit possible. Cette condition ne peut se réaliser qu'avec des conditions d'enregistrement très rigoureuses. Cas très difficile à mettre en œuvre, surtout lorsque le patient ou l'orthophoniste ne travaillent que dans le cabinet du médecin ou à la maison.

Nous allons, maintenant, extraire les phonèmes à modéliser de la structure HMM1, en considérant l'une des premières raisons possibles de la difficulté de la reconnaissance.

10. Degré de Vraisemblance par rapport aux modèles HMM1 et HMM2

Nous allons voir dans ce qui suit les degrés de vraisemblance des locuteurs dont leurs parole est saine suivant les variantes des modèles des Chaînes de Markov.

10.1. Locuteurs sains via HMM1 (7 états)

En vue de comparer le degré de vraisemblance de notre corpus par rapport au modèle HMM1, tous les locuteurs ont participé dans la phase de test.

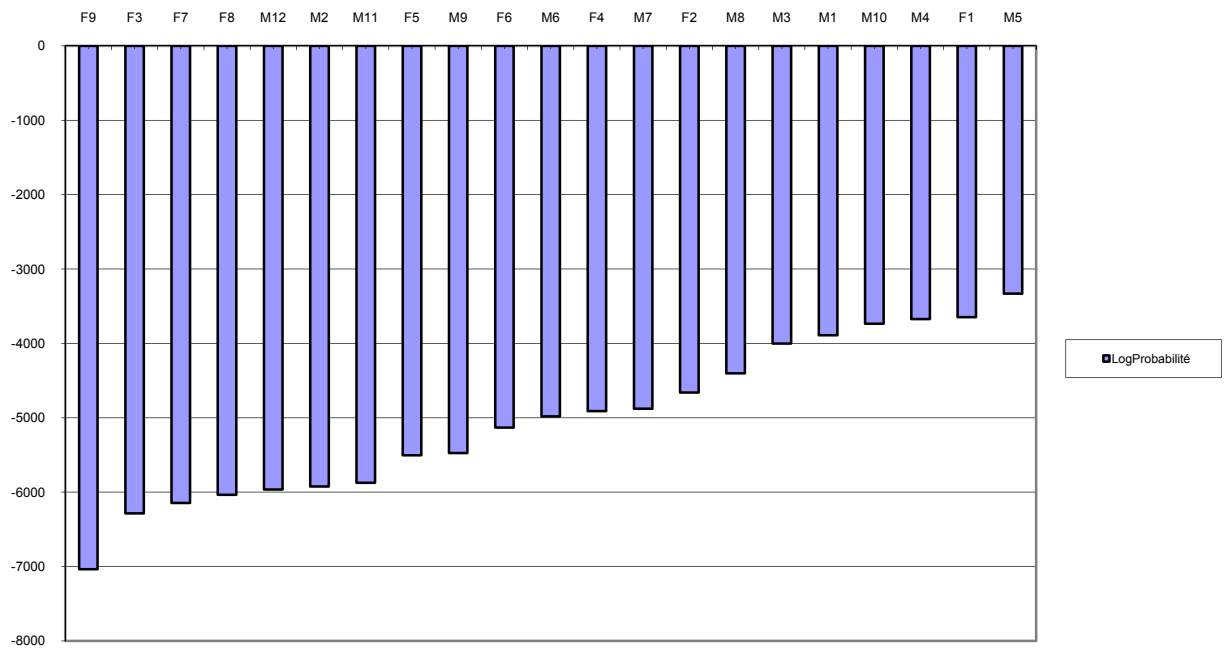


Figure 5.15 : Degré de vraisemblance que HMM1 a généré les observations acoustiques par locuteur cas de 42 MFCC / 16 Gaussiennes par état

Nous remarquons dans la figure 5.15, qu'il y a déjà dissemblance intra-classe dans le modèle HMM1. Nous pouvons dire que ce modèle est plus adapté à générer les observations du locuteur M5 et moins adapté à générer les valeurs de l'observation de F9.

10.2. Locuteurs sains via HMM2 (4 états)

En vue de comparer le degré de vraisemblance de notre corpus par rapport au modèle HMM2, tous les locuteurs ont participé à la phase de test.

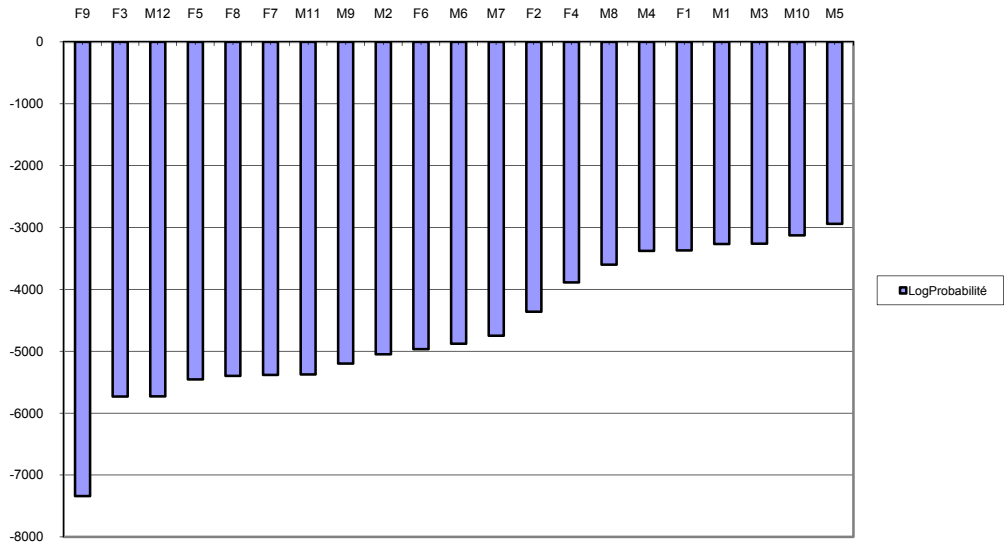


Figure 5.16 : Degré de vraisemblance que HMM2 a généré les observations acoustiques par locuteur cas de 39 MFCC / 8 Gaussiennes par état

Nous remarquons, dans la figure 5.16, qu'il y'a aussi dissemblance intra classe dans le modèle HMM2, nous pouvons dire que ce modèle est plus adapté à générer les observations du locuteur M5 et moins adapté à générer les valeurs de l'observation de F9.
Les deux modèles donnent les mêmes valeurs extrêmes en terme de dissemblance et de vraisemblance, ceci montre que la force des HMM réside dans leur aptitude de séparation de classes, ayant comme référence une base de donnes bien enregistrée.

10.3. Locutrice pathologique (F6P)

Avant de modéliser les phonèmes constituant [ʃ a χ s i j a], calculons la probabilité que le HMM1 ait généré les observations de la locutrice F6P, voir figure 5.17.

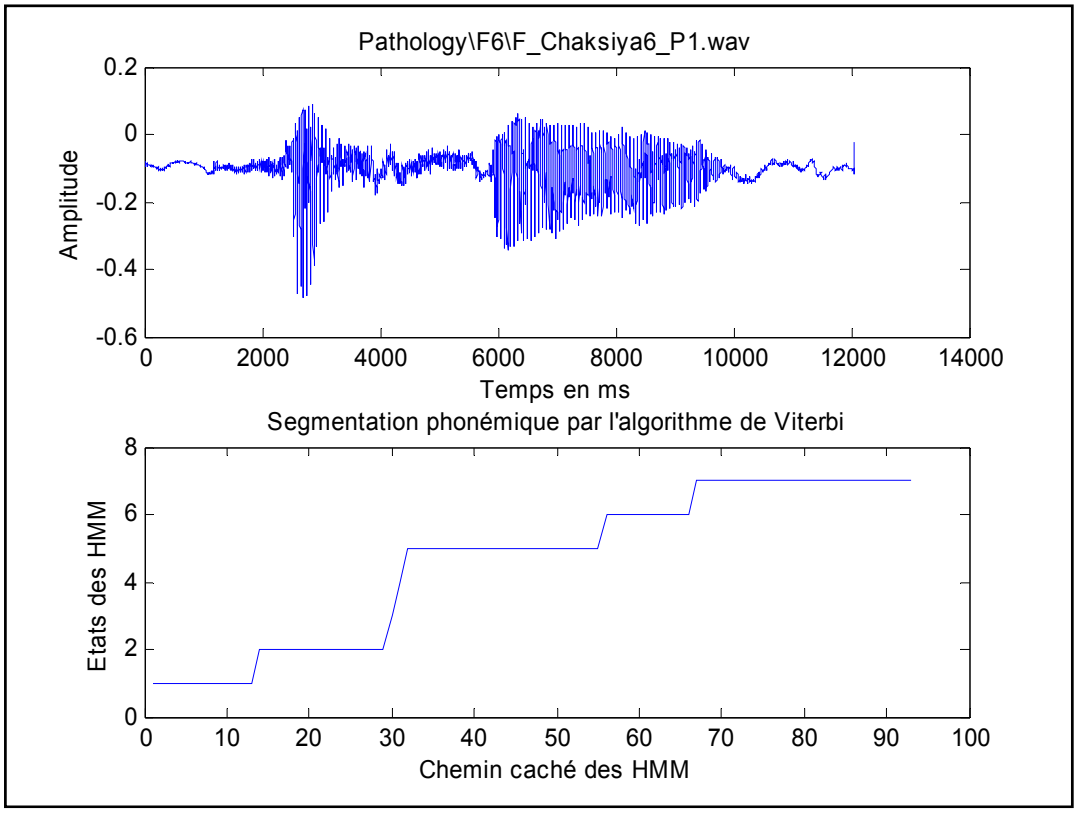


Figure 5.17 : Segmentation du mot pathologique [ʃ a χ s i j a]

Remarque : La locutrice a prononcée [θ a χ s i j a] au lieu de [ʃ a χ s i j a]

Lorsque nous essayons de segmenter manuellement le mot prononcé, le phonème [θ] correspond **auditivement** au segment automatiquement segmenté par l'algorithme de Viterbi, c'est-à-dire qu'il y a eu distinction phonémique ce qui tend à diminuer la probabilité d'entendre le [ʃ]. Le degré de vraisemblance ou loglikelihood calculée est de : **-15361**, ceci montre que la prononciation pathologique est nettement inférieure à toutes les prononciations de référence. Nous remarquons aussi, que l'on segmente [ʃa], ou une partie tronquée de [a] avec le [ʃ] au lieu de [ʃ] et [si] ou une partie tronquée de [i] concaténée avec le [s] au lieu de [s], etc..., à cause de la coarticulation qui tend à minimiser l'effet de la segmentation phonémique automatique par rapport à la segmentation manuelle qui est basée sur la vision et l'écoute, et une expérience de plusieurs dizaines ou centaines d'heures d'écoute.

11. Résultat de l'apprentissage visuel et auditif

Après quelques heures de feedback visuels et auditifs, la patiente F6P a appris a prononcer le mot C1 « **correctement** », la figure 5.18 illustre le changement dans la prononciation.

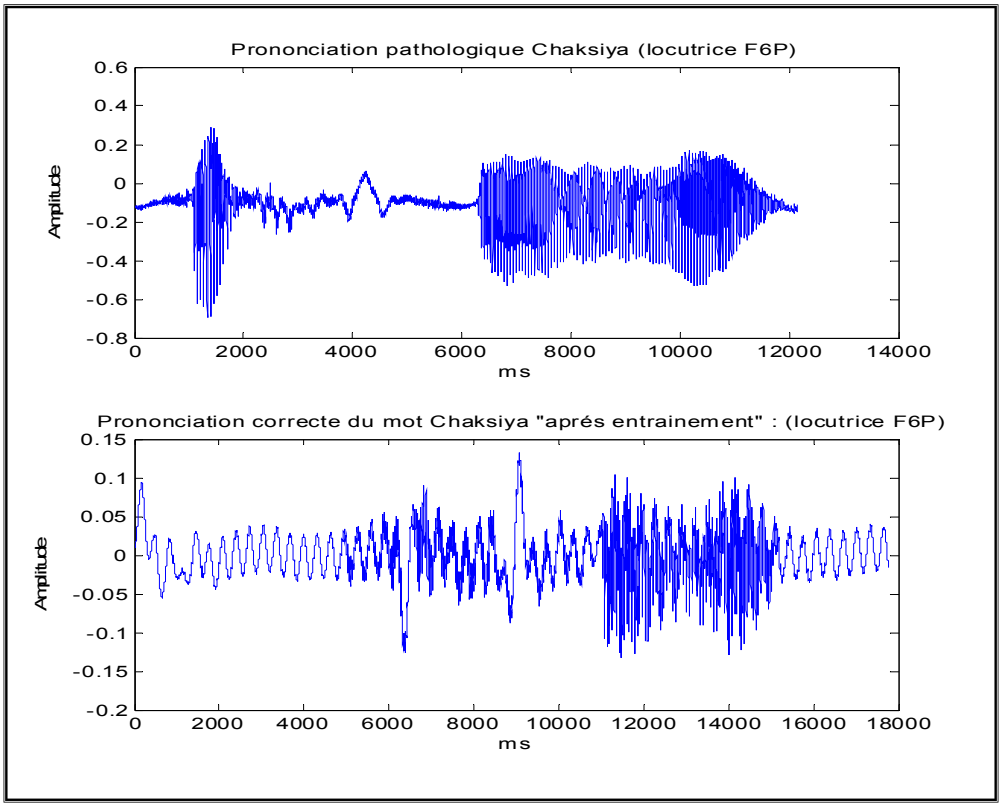


Figure 5.18 : Prononciations de la locutrice F6P, avant et après différents feedback visuels et auditifs

12. Segmentation et reconnaissance phonémique

A cette étape, les données issues du modèle HMM1 vont être utilisées par deux systèmes de reconnaissance, afin de générer un modèle basé sur les HMM élémentaires représentant les phonèmes : [ʃ],[a],[χ],[s],[i],[j],[a] en vue de calculer un degré de vraisemblance phonémique, ainsi qu'un deuxième modèle basé sur les réseaux de neurones multicouches, en vue de calculer un taux de déviation phonémique.

12.1. Modèles HMM élémentaires

La segmentation phonémique du mot C1, nous a permis de définir un nouveau modèle HMM propre au phonème [ʃ], voir tableaux 5.18 et 5.19, ainsi qu'un modèle HMM du phonème [χ], voir tableaux 5.20 et 5.21.

12.1.1. Modèle HMM du phonème [ʃ] de type gauche droite

Tableau 5.18 : Paramètres du modèle HMM du phonème [ʃ]

Nombre d'états du HMM	Mixture de gaussiennes par état	Nombre d'itérations de l'algorithme EM	Coefficients MFCC/trame
3	6	10	39

Tableau 5.19 : Matrice de transition du modèle du phonème [ʃ]

Transitions	Etat 1	Etat 2	Etat 3
Etat 1	0.8573	0.1427	5.528^{e}-116
Etat 2	0	0.8632	0.1368
Etat 3	0	0	1

12.1.2. Modèle HMM du phonème [χ] de type gauche droite

Tableau 5.20 : Paramètres du modèle HMM du phonème [χ]

Nombre d'états du HMM	Mixture de gaussiennes par état	Nombre d'itérations de l'algorithme EM	Coefficients MFCC/trame
3	6	10	39

Tableau 5.21 : Matrice de transition du modèle HMM du phonème [χ]

Transitions	Etat 1	Etat 2	Etat 3
Etat 1	0.86994	0.13006	3.9418e-253
Etat 2	0	0.87782	0.12218
Etat 3	0	0	1

12.1.3. Taux de reconnaissance et de confusion des phonèmes segmentés manuellement

En vue de prendre compte ou d'estimer le taux de fausse reconnaissance des phonèmes

segmentés manuellement, nous avons établi un comparatif intra classe, sur les phonèmes concerné par notre pathologie : [s] ,[ʃ],[χ],[θ]., voir tableau 5.22.

Tableau 5.22. : Taux de reconnaissance et de confusion

Phonème	**[s]**	**[ʃ]**	**[χ]**	**[θ]**
Taux de reconnaissance	88%	100%	96%	97%
Phonème faussement confondu avec :	[ʃ]	-	[ʃ]	[n]

Nous avons remarqué que seul le [ʃ] a un taux de reconnaissance sans confusion sur l'ensemble des occurrences, ceci est probablement dû a deux raisons :

1. la facilite de passage des post alvéolaires aux alvéolaires ;
2. la mauvaise segmentation manuelle de quelques locuteurs.

12.2. Degré de vraisemblance phonémique (HMM)

La dernière étape de reconnaissance phonémique est illustrée par les degrés de vraisemblances entre la prononciation correcte et la prononciation corrigée de la locutrice F6P.

Nous avons segmenté manuellement le mot [ʃaχsija] présentant un sigmatisme occlusif au niveau du [ʃ] et la version « correcte » prononcée par la locutrice F6P lors des séances d'enregistrement, afin de calculer le degré de vraisemblance entre le modèle HMM doublement stochastique et les prononciations réelles.

Une comparaison a été faite avec les deux premiers phonèmes du mot, car nous considérons l'occlusion initiale pour cet exemple, (tableau 5.23).

Tableau 5.23. : Vraisemblance phonémique

N°	**Phonème segmenté Manuellement**	**Prononciation avec sigmatisme**	**Prononciation correcte**
1	[θ]	-1002.6	**-917**
2	[ʃ]	-3073.4	**-2732.1**
5	[χ]	-1.0824	-1.4440
6	[χ]	-1352.0	-1830.3

D'après ces données, nous pouvons confirmer que le fait que le patient se corrige lors des sessions d'apprentissage de la prononciation tend à augmenter le degré de vraisemblance (valeurs moins négatives) entre phonèmes prononcés, ceci aidera à améliorer le seuil d'écoute et de prononciation du patient lors des sessions d'orthophonie.
Nous pouvons dire qu'à ce stade de segmentation et de reconnaissance, les phonèmes même proches au lieu d'articulation ont été classés selon le rapprochement du lieu d'articulation, (tableau 2.4).

12.3. Modèle ANN

En premier lieu, nous avons effectué une modélisation sur 80% du corpus, les 20% restant ont servi aux tests.

En vue de calculer le taux de déviation phonémique, nous avons pris des phonèmes, [ʃ],[s], [θ] et [t] pour lesquelles la malade fait des confusions, dans la base TIMIT.

12.3.1. Modélisation phonémique

Dans l'optique d'avoir un taux de reconnaissance phonémique optimal, nous avons fait l'étude et l'analyse d'un second classifieur basé sur les réseaux de neurones multicouches, différentes variantes de réseaux ont été introduites pour minimiser les temps d'entrainement et augmenter les probabilités a postériori, des phonèmes étudiés, le premier réseau est formé de 12 et 39 entrées, une couche cachée, 3 sorties, et le second réseau est formé de 12 et 39 entrées, de deux couches cachées, et de 3 sorties phonémiques.
En vue d'augmenter le taux de déviation phonémique, l'utilisation de la base TIMIT pour l'entrainement ainsi que les tests préliminaires, s'est avérée un choix très judicieux, car les phonèmes arabes utilisés dans notre pathologie en question, correspondent à des phonèmes de la base TIMIT, (Annexe B tableau B.4).
Les phonèmes concernés par notre modèle sont le [s], [ʃ],[θ].
Ce choix repose sur la déviation possible du phonème [s] vers [ʃ] ou [θ], (tableau 2.4).
L'entraînement a concerné 70% des 640 locuteurs de la base TIMIT, les 30% restant concernent les tests, à ces derniers ont été adjoints les phonèmes enregistrés localement, pour évaluer notre méthode d'enregistrement et perfectionner les manquements.

Les variantes des différents réseaux sont mentionnées dans les tableaux 5.24 à 5.25

a) Cas ou l'entrée du réseau et constitué de (12 coefficients MFCC)

Tableau 5.24: Taux de reconnaissance en % pour un Réseau à une couche cachée

Couche cachée / Phoneme	8 N	16N	32N	64N	128N
[س][s]	84	10	78	94	89
[ش][ʃ]	5	89	62	45	85
[ث][θ]	14	56	60	38	52

Tableau 5.25 : Taux de reconnaissance en % pour un Réseau à deux couches cachées

Couche cachée

Couches cachées / Phonème	16:16:3	16:32:3	32:16:3	32:32:4	16:16:3
[س][s]	83.7	87	78	80	83.7
[ش][ʃ]	84	88	88	85	84
[ث][θ]	54	48	58	62	54

b) Cas ou l'entrée du réseau et constitué de (36 coefficients MFCC)

Tableau 5.26 : Taux de reconnaissance en % pour un Réseau à une couche cachée

Couche cachée / Phonème	8 N	16N	32N	64N	128N
[س][s]	81	10	78	94	89
[ش][ʃ]	5	89	62	45	85
[ث][θ]	14	56	60	38	52

Tableau 5.27 : Taux de reconnaissance en % pour un Réseau à deux couches cachées

Couches cachées / Phonème	16:16:3	16:32:3	32:16:3	32:32:4	16:16:3
[س][s]	83.7	87	78	80	83.7
[ش][ʃ]	84	88	88	85	84
[ث][θ]	54	48	58	62	54

D'après la comparaison entre les réseaux étudiés, le choix est porté sur un parallélisme au niveau des réseaux choisis, en vue de calculer le taux de déviation comme démontré dans le paragraphe (5.13).

13. Déviation phonémique et Système d'aide à la thérapie langagière

L'approche de la déviation phonémique est une approche plus pointue que la vraisemblance, car elle utilise l'information des HMM, et effectue un comparatif autour des phonèmes probablement prononcés lors des séances de correction.

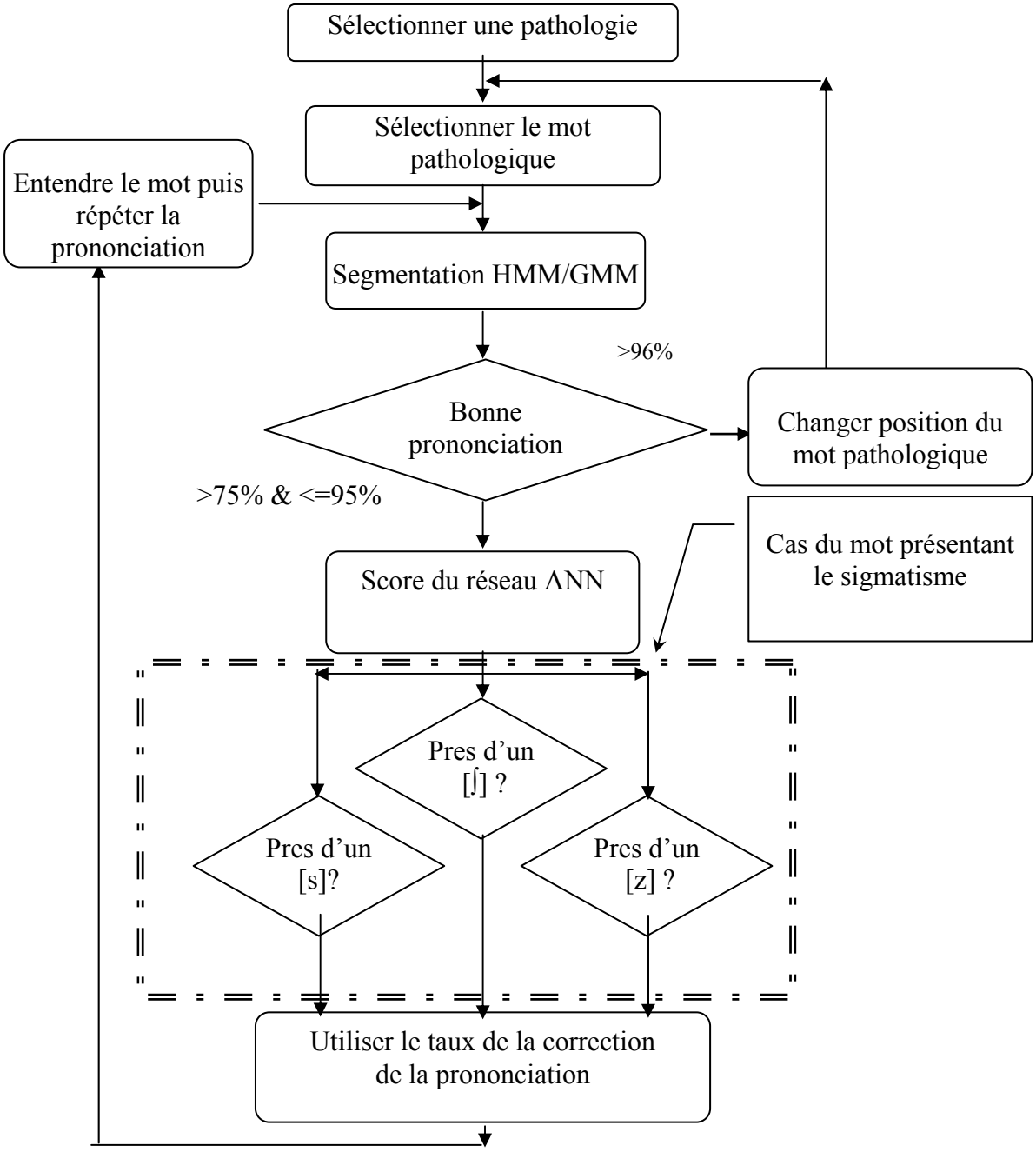


Figure 5.19 : Schéma global de la thérapie langagière

14. Conclusion

Dans ce chapitre, nous avons adoptée une méthodologie de détection du sigmatisme occlusif et constrictif par l'utilisation des coefficients MFCC connus pour leur robustesse au bruit et une modélisation stochastique basée sur les HMM / ANN à base de GMM.

Différents tests et comparaisons ont été réalisés sur un corpus enregistré ainsi que des tests sur la phrase SA1 de la base TIMIT, ceci nous a permis de déduire des règles de travail relatives aux paramètres de l'enregistrement et au choix du corpus.

La segmentation phonémique automatique obtenue nous a permis de comprendre les difficultés de la coarticulation qui tend à minimiser la séparation des phonèmes, toutefois l'utilisation hybride de la segmentation manuelle en plus de la modélisation Markovienne nous a permis de résoudre une grande partie du problème posé.

Conclusions Générales et perspectives

La Reconnaissance Automatique de la Parole a connu un essor incommensurable par l'introduction de la modélisation stochastique, depuis les études établies par J. Fergusson et L. Rabiner.

La modélisation du signal vocal est régie par le choix des vecteurs acoustiques dont les performances doivent être des plus optimales, en termes de sensitivité au bruit, de complexité de calcul, de degré de perception, etc.

Dans ce travail, nous avons développé une application qui s'insère dans le domaine du Traitement Automatique de la parole. La segmentation phonémique que nous avons faite est orientée vers l'aide à la correction de la prononciation des sigmatismes occlusifs et constrictifs. Cette approche est basée sur des "feedback" visuels et auditifs qui aident l'orthophoniste ou le patient à réentendre la prononciation incorrecte présentant la pertinence de la pathologie en lui indiquant les zones d'erreurs possibles. Ce travail fait pour une pathologie a été extrapolé pour d'autres pathologies représentant une certaine similitude dans la substitution d'un phonème par d'autres.

La détection de la défaillance de prononciation du phonème [ʃ] [ش] prononcé [s] [س] ou par [θ] [ث] ou par [t] [ت], a été traitée ainsi que différentes comparaisons ont été réalisées avec des corpus de parole saine ou de références et pathologiques enregistrées, ainsi qu'avec des données de la Base de Données TIMIT en vue de valider nos travaux pour une éventuelle extension à d'autres maladies.

Nous avons pu segmenter automatiquement le phonème pathologique, à traiter, et extrait sa position dans le mot pathologique cible. Les différents degrés de vraisemblance obtenus par rapport au modèle de référence montrent que la fiabilité des résultats est basée sur les conditions d'enregistrements, ainsi que le corpus de références qui modélise le mieux les différents modèles Markoviens des phonèmes à traiter.

A titre comparatif, la modélisation de la phrase SA1 ainsi noté dans la base TIMIT nous a donné un taux de reconnaissance de 99%, ceci montre la fiabilité de l'utilisation des HMM/GMM dans cet axe de recherche.

Notre approche est basée sur une aide à l'apprentissage à partir de graphes et de sons, suivie de la génération d'un score de vraisemblance, ou « note d'appréciation » qui aide le patient à corriger sa prononciation au fur et à mesure avec les HMM/GMM. Cette dernière nous l'avons améliorée en introduisant une notion que nous avons appelée, déviation phonémique, ceci en se basant sur un classificateur ANN.

Les résultats des expériences réalisées nous ont permis de tirer des règles relatives au choix du corpus de travail, à l'impact de la segmentation manuelle et les erreurs de reconnaissance issues des erreurs infimes d'approximation ainsi que le nombre et le type de vecteurs acoustiques à utiliser.

En termes de perspectives à ce travail, nous préconisons de modéliser d'autres sigmatismes ainsi que les schlintements, le zézaiement ainsi que d'autres défauts pathologiques détectables par le système auditif humain, selon la méthodologie mentionnée dans l'organigramme que nous avons proposé. L'intégration d'autres maladies s'effectuera par l'addition du mot pathologique, des différentes occurrences du phonème en question, ainsi que la Base de Données Correspondante.

Références Bibliographiques

[1] http://www.voiceacademy.org:8080/vaweb/glossary.html

[2] D. H. KLATT, Software for cascade parallel formant synthesizer,JASA, 67, 1980, p . 971-995.

[3] L.R. Rabiner, A tutorial on Hidden Markov Models and selected Applications in Speech recognition, Proceedings of the IEEE, Vol. 77 N°. 2, February 1989.

[4] S. Poitoux, Etude des mesures de confiance dans le traitement de la parole avec application en logopédie, Faculté polytechnique de Lausanne, Suisse 2002.

[5] Techniques de l'ingénieur, Vol. H1 940, p.3.

[6] J.P. Zerling, Articulation et coarticulation dans les groupes occlusives-voyelles en Français. Thèse de doctorat de 3ème cycle, Université de Nancy 2, France, 1979.

[7] M. Chetouani, B.Gas and J.L. Zarader, Coopération entre codeurs-Neuro prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes, Laboratoire des Instruments et systèmes d'Ile de France, Université Paris IV, France, 2004.
isir.ro**b**ot.jussieu.fr/?op=view_page&type=perso&lang=fr&id

[8] M. Cooke, S. Beet and M. Crawford Éds. Visual representations of speech signals., Collection Wiley professional computing, John Wiley and Sons,1993. 385 pp

[9] P. Flandrin. Temps-fréquence., Traité des nouvelles technologies, série Traitement du signal, Hermès, 1993. 394 pp

[10] Calliope. La parole et son traitement automatique. Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pages, 1989.

[11] S. B. Davis et P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, N° 4, pp. 357-366, 1980.

[12] D.M.Istrate, Détection et reconnaissance des sons pour la surveillance médicale, Thèse de Doctorat, pp. 95-129, CLIPS- IMAG, France, 2003.

[13] J. D. Markel et A. H. Gray, Linear prediction of speech. 288 pages, Springer-Verlag, 1976.

[14] U.S. Department of Defense. LPC-10 2400 bps voice coder. Release 1.0, 1993. (fichier readme.txt du package).
www.cs.cmu.edu/afs/cs.cmu.edu/project/

[15] J. Scourias. Overview of the global system for mobile communications. Rapport technique, 25 pages, Université de Waterloo, Waterloo (on, Canada), 1995.

[16] H. Hermansky, B. A. Hanson et H. Wakita. Low-dimensional representations of vowels based on all-pole modeling in the psychophysical domain, Speech Communication, Vol. 4, pp 181-187, 1985.

[17] H. Hermansky. Perceptual linear predictive analysis of speech. Journal of the Acoustical Society of America, Vol. 87, N° 4, pp 738-752, 1990.

[18] N. Morgan, H. Hermansky, H. Bourlard, P. Kohn and C. Wooters. Continuous speech recognition using PLP analysis with multilayer perceptrons. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 49-52, 1991.

[19] H. Hermansky, N. Morgan, A. Bayya et P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Proceedings of the European Conference on Speech Communication and Technology, pp 1367-1370, 1991.

[20] H. Hermansky, N. Morgan, A. Bayya and P. Kohn. RASTA-PLP speech analysis. Rapport technique TR-91-069, 6 pp, International Computer Science Institute, Berkeley (CA, États-Unis), 1991.

[21] H. Hermansky, N. Morgan, A. Bayya et P. Kohn. RASTA-PLP speech analysis technique. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, N° 4, pp 121-124, 1992.

[22] H. Hermansky and N. Morgan. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, Vol. 2, N° 4, pp 578–589, 1994.

[23] N. Morgan and H. Hermansky, RASTA extensions : robustness to additive and convolutional noise. ESCA Technical Research Workshop, Speech processing in adverse conditions, pp. 115-118, 1992.

[24] http://www.vecsys.fr/applications/applis-vocales.htm#siel

[25] http://www.vecsys.fr/applications/applis-vocales.htm#rentacar

[26] http://www.vecsys.fr/applications/applis-vocales.htm#edf

[27] http://www.vecsys.fr

[28] www.loria.fr/projets/JEP-TALN/actes/TALN/conf_assoc/Ecrit_Oral05.pdf

[29] http://www.sfu.ca/~saunders/l33098/L3/Respiration_02.html

[30] http://www.kehlkopfoperiert.ch/F/f-lunge.html

[31] http://cystic-fibrosis-symptom.com/lungs_trachea.html

[32] http://fr.wikipedia.org/wiki

[33] http://sprojects.mmi.mcgill.ca/larynx/notes/n_frames.html

[34] http://www.lli.ulaval.ca/labo2256/lexique

[35] D. Ducassou. Cours d'acoustique. Cours de 2ème année de médecine, Université de Nancy 1, 1991

[36] http://r.battault.free.fr/probatoire/probatoire.html

[37] http://www.geocities.com/phlaplacephntics/elements.html

[38] http://www.tadjweed.com/Makharijul7uroof.html

[39] http://www.arts.gla.ac.uk/IPA/fullchart.html

[40] http://www.linguistes.com/phonetique/phon.html

[41] http://al-zahra.net/qurannet/old_qurannet/main/quran/lesson6/tajweed-MKH.html

[42] Cantineau J., 1960, Esquisse d'une phonologie de l'arabe classique [1946], Etudes de Linguistique arabe, Paris, Klincksieck, p. 165-204.

[43] T.Saidane, M. Zrigui et M.Benahmed, La transcription orthographique-phonetique de la langue arabe», Société Tunisienne d'Electricité et du Gaz, Centre de production de Sousse; Laboratoire RIADI, Unité Monastir, Faculté des sciences de Monastir, Ecole Nationale des Sciences de l'informatique, Tunis, Tunisie, 2004.

[44] T. Koizumi, M. Mori, S. Taniguchi, and M. Maruya, Recurrent neural networks for phoneme recognition, Dept. of Information Science, Fukui University, Japan, 1996.

[45] O.Essa, Using supra-segmentals in training H Markov Models for Arabic, Computer Science Dept, University of South California, 1998.

[46] A.M.A.Ali, J.V.Der-spiegel, P. Mueller, G.Haentjens and J.Berman, An acoustic-phonetic feature-bases system for automatic phoneme recognition in continuous speech, University of Pennsylvania, 1999.

[47] A.Juneija and C.Epsy-Wilson, Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning, ECE Dept., University of Maryland, College Park, USA, 2003.

[48] Y. Lee, K.Papineni, S.Roukos, O.Emam, H.Hasny, Language model based Arabic word segmentation, Proceedings of the 41st Annual meeting of the Association for Computational Linguistics, pp 399-406, July 2003.

[49] B. LE Viet, Reconnaissance automatique de digits en anglais en conditions bruitées, Memoire de DEA d'informatique, Systèmes et Communications, INP de Grenoble, France, 2002.

[50] C. Lévy, G. Linarés, P. Nocera et J.F Bonastre, Reconnaissance de chiffres isolés embarquée dans un téléphone portable, Laboratoire Informatique d'Avignon, France, 2004.

[51] M. Akbar et J.Caelen, Parole et traduction automatique : le module de reconnaissance RAPHAEL, Université Joseph Fourier, Grenoble, France, 1998.

[52] J. Cernocky, G. Baudoin, et G. Chollet, ALISP : Quelques outils pour une analyse acoustico - phonétique de la parole indépendante de la langue, *Revue Parole*, Déc. 2000.

[53] M. Moscato et J. Wittwer, La psychologie du langage, Collection 'Que sais je ?', Presses Universitaires de France, 1978, 75 pages.

[54] Fédération nationale des orthophonistes, Semaine nationale de prévention des troubles du langage 27 au 31 Mai 2002, Fédération Nationale des Orthophonistes, SPODS, France.

[55] http://www.geneva-link.ch/ceppim/final/ORL/Laphonation.htm, Université de Genève

[56] http://www.upmc.edu, Université de Pittsburgh, Voice Center

[57] J. Koufman, Bowing of the Vocal Cords, the visible voice, Vol. 3 N°.2, April, 1994.

[58] T. Murry, C.A. Rosen, Vocal fold granuloma, University of Pittsburgh Voice Center, Department of Otolaryngology, Pittsburgh, Pennsylvania, 23 May 2001.

[59] www.ghorayeb.com, Otolaryngology-Head & Neck Surgery

[60] www.ligue-cancer.net, La ligue contre le Cancer: Information et prévention, les cancers des voies æro-digestives, Ligue Nationale contre le cancer, Paris.

[61] http://hsc.virginia.edu, Health System, University of Virginia

[62] K.Shahin, «Remarks on the speech of Arabic-Speaking Children with cleft palate, The University of British Columbia, 2002.

[63] K.Verdolini, K. DeVore, S. McCoy and J. Ostrem, Vocology Guide, National Centre for Voice and speech, 2002.

[64] http://www.aquacorpus.be/logopedie/index.php

[65] H. S. Venkatagiri, Voice is one aspect of speech production, Department of Psychology Iowa State University, 2003.

[66] www.vulgaris-medical.com

[67] http://www.infovoyager.com/francais/wikipedia/r/rh/rhotacisme.html

[68] http://dictionnaire.metronimo.com/term/53aa5da457a7acaaa2,xhtml

[69] http://www.begaiement.org/faqapb.html

[70] http://www.geneva-link.ch/ceppim/final/ORL/La_phonation.htm#trouble

[71] http://encyclopaedic.net/franc/dy/dyslalie.html

[72] http://tecfa.unige.ch/tecfa/teaching/UVLibre/tp-iish/ex-9798/logopedie/trouble1.html

[73] http://imerosendael.free.fr/orthopho.html

[74] Y.Laprie, Analyse spectrale de la parole, Cours, CRIN, Nancy, France,2002.

[75] A. R.Elobeid Ahmed, Performance Tests on Several Parametric representations for an Arabic phoneme recognition system using HMM's, Transactions on Information and communications Technologies, Vol. 20, 1998.

[76] Y. Tsubota, T. Kawahara and M. Dantsuji, Recognition and verification of English by Japanese students for computer-assisted language learning , School of informatics, Center for information and multimedia studies, Kyoto University, 2002.

[77] A.Alizera, Dibazar, S. Narayanan and T.W. Berger, Feature analysis for automatic detection of pathological speech , Biomedical Eng. Dept, Elect. Eng. Dept., University of South California, 2002.

[78] G. Stemmer, C. Hacker, E. Noth and H. Niemann, Multiple time resolutions for derivatives of Mel-Cepstral Coefficients, Université Erlangen, Nurenberg, 2001.

[79] W.J.Picone., Signal Modeling Techniques in Speech Recognition, Proceedings of IEEE, 1215-1247 p, Vol. 9, 1993

[80] B. Gas, J.L. Zarader, C. Chavy, A new approach to speech coding: the Neural Predictive Coding, Journal of Advanced Computational Intelligence, Vol 4, N°1, Janvier 2001, pp 120-127.

[81] H. Sakoe and S. Chiba. A dynamic programming approach to continuous speech recognition.Proceedings of the 7th International Conference on Acoustics, article 20-13, 6 pp, 1971.

[82] R. E. Bellman. Dynamic Programming. Princeton University Press, 1957

[83] R. E. Bellman. On a routing problem. Quaterly Journal of Applied Mathematics, Vol. 16,pp 87-90, 1958.

[84] J. S. Bridle, Optimization and search in speech and language processing. Survey of the State of the art in human language technology, NSF, CE-DG13E & OGI-CSLU éds, art. 11.7, pp 423-428, 1995.

[85] F. Itakura, Minimum production residual principle applied to speech recognition. IEEE Transaction on Acoustics, Speech an Signal Processing, Vol. 23, pp 67-72, 1975

[86] H. Sakoe et S. Chiba. Dynamic programming algorithms optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 26, N°1, pp 43-49, 1978.

[87] C. S. Myers and L. R. Rabiner, Connected digit recognition using a level building DTW algorithm. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 29, pp 351-363, 1981.

[88] C. S. Myers et L. R. Rabiner, A comparative study of several dynamic time-warping algorithms for connected-word recognition. The Bell System Technical Journal, Vol. 60, N° 7, pp 1389-1409, 1981.

[89] H. Sakoe. Two level DP-matching - A dynamic programming based pattern matching algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, N° 3,pp 588-595, 1979.

[90] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg et J. G. Wilpon. Speaker independent recognition for isolated words using clustering techniques. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, N° 4, pp 336-349, 1977.

[91] Z.A. Benselama, M. Guerti, M.A. Bencherif, Occlusive Sigmatism Correction Aided Design, *Wseas Transactions on Signal Processing*, Vol. 3, N° 6, pp 361-367, 2007.

[92] L. Buniet, Traitement automatique de la parole en milieu bruite : étude de modèles connexionnistes statiques et dynamiques, Thèse de Doctorat de l'Université Henri Poincaré - Nancy 1, spécialité informatique, 1997.

[93] P. Langlais, "Introduction VMM vs HMM", Université de Montréal, fiches de cours, (Janvier 2002).

[94] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, Traitement de la parole » Presses Polytechniques et Universitaires romandes, 2000.

[95] J. Tebelskis, Speech Recognition using Neural Networks, Thèse de Doctorat, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, 1995.

[96] Z.A. Benselama, M. Guerti, M.A. Bencherif, Arabic Speech Pathology Therapy computer Aided System**,** *Journal Computer Science,* Science Publication, New York, USA, Vol. 3 N° 9, pp 685-692, 2007,

Annexes

Tableau B.1 : Occurrences du [ʃ] en position initiale des mots

الكسرة	الضمة	الفتحة	Phonème contextuel	Phonème
		شَابٌ	ا	ش
شِيْرٌ	شُبَاكٌ	شَبَابٌ	ب	ش
شِتَائِيٌ		شَتَاتٌ	ت	ش
			ث	ش
شِجَارٌ	شُجَاعٌ	شَجَرَةٌ	ج	ش
	شُحْنَةٌ	شَحْمَةٌ	ح	ش
شِمَالٌ	شُرْطَةٌ	شَخْصِيَةٌ	خ	ش
شِدْقٌ		شَدَّةٌ	د	ش
	شُذُوذٌ	شَذّ	ذ	ش
شِرَاعٌ	شُرْطِيٌ	شَرِيعَةٌ	ر	ش
			ز	ش
			س	ش
			ش	ش
شِصٌ			ص	ش
			ض	ش
شِطْرَنْجٌ		شَطْبٌ	ط	ش
		شَظِيةٌ	ظ	ش
شِعَارْ	شُعَاعٌ	شَعْرٌ	ع	ش
	شُغُورٌ	شَغَلَ	غ	ش
شِفَاءْ	شُفْرَةٌ	شَفَاعَةٌ	ف	ش
شِقَاقْ	شُقْرَةٌ	شقاء	ق	ش
شِكْ	شُكْبٌ	شُكْرٌ	ك	ش
شِلَّةٌ	شُلٌّ	شَلالٌ	ل	ش
شِمَالْ	شُمْرَةٌ	شَمْسٌ	م	ش
شِنْجَارْ	شُنْقُبٌ	شَنِيعٌ	ن	ش
	شُكولاتة	شَوَاءٌ	و	ش
شِوَاءْ	شُورَى	شَهِيدٌ	هـ	ش
شِيكْ	شُيوعٌ	شَيْخٌ	ي	

Tableau B.2. Occurrences du [ʃ] en position médiane des mots

Phonème	Phonème contextuel	السكون	الفتحة	الضمة	الكسرة
ا	ش	إِشْتَرَطَ	أَشَعَ		أَشِعَةٌ
ب	ش	بُشْرَةٌ	بَشَاشَةٌ	بَشُوشٌ	بَشِيرٌ
ت	ش	تَشْتَرِي	تَشَاجَرَ		
ث	ش	تَشْرِيدْ	تَشَدَدَ		
ج	ش				
ح	ش	حَشْوَةٌ	حَشَدَ		حَشِيشَةٌ
خ	ش	خَشْرَمْ	خَشَبٌ	خُشُوعٌ	خَشِيَ
د	ش	دِشْ	دَشَنَ		
ذ	ش				
ر	ش	رَشْوَة	رَشَاشٌ		رَشِيقٌ
ز	ش				
س	ش				
ش	ش				
ص	ش				
ض	ش				
ط	ش				
ظ	ش				
ع	ش	عُشْبٌ	عَشَرَةٌ	عُشْرٌ	عَشِيةٌ
غ	ش	غِشْ	غَشَاشٌ		غَشِيمٌ
ف	ش	فَشْخَة	فَشَلَ		فَشِلَ
ق	ش	قِشْرَةٌ	قَشَطَ		قَشِيبٌ
ك	ش	كَشْرَة	كَشَافٌ	كُشُوفٌ	
ل	ش		لَشَكَ		
م	ش	مِشْوَاةٌ	مُشَوشٌ		مَشِيئَةٌ
ن	ش	نَشْوَةٌ	نَشَرَ	نُشُوبٌ	نَشِطٌ
و	ش		وِشَاحٌ		وَشِيكٌ
هـ	ش	وَشْوَشٌ	هَشَاشَةٌ		هَشِيمٌ
ي	ش	يَشْمَقُ		يَشُمُ	

Tableau B.3. Occurrences du [ʃ] en position finale des mots

Position finale	Phonème contextuel	Phonème
فِرَاشٌ	ش	ا
كَبْشٌ	ش	ب
مَرْتَشٍ	ش	ت
	ش	ث
	ش	ج
جَحْشٌ	ش	ح
	ش	خ
دِشٌ	ش	د
	ش	ذ
رَشْ	ش	ر
	ش	ز
	ش	س
	ش	ش
	ش	ص
	ش	ض
	ش	ط
	ش	ظ
عُشْ	ش	ع
غِشْ	ش	غ
	ش	ف
	ش	ق
	ش	ك
	ش	ل
هَامِشْ	ش	م
	ش	ن
مَدْهُوشْ	ش	و
هشْ	ش	هـ
شويش	ش	ي

Tableau B.4 : Correspondance TIMIT IPA Arabic

TIMIT	Arabic phoneme	IPA symbol	Kind	Manner of production	Place of articulation	Voiced or unvoiced
	ء	?	Semivowel	Plosive	Glottal	Unvoiced
b - bcl	ب	b	Consonant	Plosive	Bilabial	Voiced
t - tcl	ت	t	Consonant	Plosive	Dental	Unvoiced
th	ث	θ	Consonant	Fricative	Dental	Unvoiced
zh	ج	dz	Consonant	Plosive	Velar	Voiced
	ح	hh	Consonant	Fricative	Pharyngeal	Unvoiced
	خ	x	Consonant	Fricative	Velar	Unvoiced
d – dcl	د	d	Consonant	Plosive	Dental	Voiced
dh	ذ		Consonant	Fricative	Dental	Voiced
r	ر	r	Consonant	Trill	Alveolar	Voiced
z	ز	z	Consonant	Fricative	Alveolar	Voiced
s	س	s	Consonant	Fricative	Alveolar	Unvoiced
sh	ش		Consonant	Fricative	Post alveolar	Unvoiced
	ص	s	Consonant	Fricative	Coronal	Unvoiced
	ض	d̪	Consonant	Plosive	Alveolar	Voiced
	ط	t̪	Consonant	Plosive	Interdental	Unvoiced
	ظ		Consonant	Fricative	Palatal	Voiced
	ع	ʔ	Consonant	Fricative	Pharyngeal	Voiced
	غ	ɣ	Consonant	Fricative	Velar	Voiced
f	ف	f	Consonant	Fricative	Labiodental	Unvoiced
q	ق	q	Consonant	Plosive	Uvular	Unvoiced
k – kcl	ك	k	Consonant	Plosive	Velar	Unvoiced
l	ل	l	Consonant	Lateral	Post alveolar	Voiced
m	م	m	Consonant	Nasal	Bilabial	Voiced
n	ن	n	Consonant	Nasal	Dental	Voiced
h – hh	هه	h	Consonant	Fricative	Glottal	Unvoiced
w	و	w	Semivowel	Approximant	Velar	Voiced
y	ى	j	Semivowel	Approximant	Palatal	Voiced
aa	ـَ	a	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced
iy	ـِ	i	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced
uw	ـُ	u	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced
		a:	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced
		i:	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced
	ـُو	u:	Vowel	\\\\\\\\\\\\\\\	\\\\\\\\\\\\\\\	Voiced